KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Marius Žalinauskas

INDIVIDUALIAI KLASIFIKUOTŲ DOKUMENTŲ KLASTERIZAVIMO METODAS

Magistro darbas

Darbo vadovas:
doc. dr. Eimutis Karčiauskas

KAUNAS, 2006

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Marius Žalinauskas

INDIVIDUALIAI KLASIFIKUOTŲ DOKUMENTŲ KLASTERIZAVIMO METODAS

Magistro darbas

Kalbos konsultantė:
doc. dr. Jurgita Mikelionienė

Vadovas:
doc. dr. Eimutis Karčiauskas

Recenzentas:
doc. dr. Rimantas Butleris

Atliko:
IFM-0/2 gr. studentas
Marius Žalinauskas

KAUNAS, 2006

TURINYS

1 Įvadas 4
1.1 Darbo tikslai ir rašymo aplinkybės 4
1.2 Klasterizavimo metodas individualiai klasifikuotiems dokumentams 5
1.3 Dokumento turinys 6
2 Paieškos metodai ir jų problemos 7
3 Klasterizavimo ir klasifikavimo principai 8
4 Klasterizavimo metodai 11
4.1 Metodų ir algoritmų klasifikacija 11
4.2 Klasterizavimo procesas 12
4.3 Dokumentų reprezentacija klasterizavimo metu 12
4.4 Dokumentų panašumo matavimas 14
4.5 Klasterizavimo algoritmai 18
4.5.1 Dalinantis klasterizavimas 19
4.5.2 Hierarchinis klasterizavimas 22
4.5.3 Griežto ir negriežto klasterizavimo algoritmai 26
4.6 Skirstinių pateikimas 27
4.6.1 Vidinis pateikimas 27
4.6.2 Pateikimas išorei 27
4.7 Skirstinių validavimas 32
5 Individualiai klasifikuotų dokumentų klasterizavimo metodas 33
5.1 Dokumentų reprezentacija klasterizavimo metu 33
5.1.1 Lietuviškų dokumentų kolekcijos žodyno sudarymo problemos 33
5.1.2 Individualaus klasifikavimo principai 34
5.1.3 Pasirinktas dokumentų reprezentacijos būdas 36
5.2 Dokumentų panašumo matavimas 38
5.2.1 Euklidiniu atstumu pagrįstų panašumo metrikų trūkumai 38
5.2.2 Pasirinktas dokumentų panašumo matavimo būdas 39
5.3 Klasterizavimo algoritmas 39
6 Eksperimentai 40
6.1 Eksperimentiniai duomenys 40
6.2 Eksperimentų sąlygos ir metrikos 43
6.3 Tinkamiausio klasterizavimo algoritmo atranka 45
6.4 Metodo palyginimas su tradiciniu klasterizavimo metodu 49
7 Rezultatų apibendrinimas ir darbai ateičiai 53
8 Išvados 55
Literatūra 57
Summary 61
Priedas A – eksperimentų žurnalas 62
Priedas B – eksperimentų žurnalas 84
Priedas C – eksperimentų žurnalas 106
Priedas D – eksperimentų žurnalas 128
Priedas E – konferencijos pranešimas 150
Priedas F – konferencijos pranešimas 154

1 ĮVADAS

1.1 DARBO TIKSLAI IR RAŠYMO APLINKYBĖS

Autoriui dirbant inžineriniame projekte, kurio metu buvo bandoma patobulinti specializuotos dokumentų valdymo sistemos – redakcinės sistemos periodinei žiniasklaidai News Processor (Bernotas; Žalinauskas, 2006) – architektūrą, buvo pastebėta, kad vartotojai vengia naudotis centralizuota kategorijų sistema. Atlikus apklausą paaiškėjo, kad dauguma vartotojų egzistuojantį kategorijų sąrašą laiko neadekvačiu. Iš anksto centralizuotai sukurtos kategorijos yra per ne lyg bendros, kad pakankamai tiksliai apibrėžtų joms priskirtų straipsnių arba fotografijų turinį, be to, egzistuojančiame kategorijų sąraše neretai iš vis nėra tinkamos kategorijos[1].

Suvokiant, kad negalima kiekvienam vartotojui suteikti teisių centralizuotai kategorijų sistemai keisti, vietoje pastarosios buvo pasiūlytos individualios kategorijų sistemos, kuriomis naudodamiesi vartotojai galėtų lanksčiai ir savarankiškai organizuoti savo dokumentus. Deja, nenaudojant centralizuotos kategorijų sistemos būtų prarandamas visą dokumentų saugyklą vienijantis elementas ir tai labai komplikuotų reikalingos informacijos paiešką. Periodinėje žiniasklaidoje, ypač dienraščiuose, rašomi straipsniai yra dažnai susiję su kartą jau aprašytais faktais[2], todėl sparti ir efektyvi reikalingos medžiagos paieška yra kritiškai svarbi.

Panagrinėjus egzistuojančių paieškos, klasifikacijos ir klasterizavimo metodų veikimo principus (Žalinauskas, 2006), buvo nuspręsta sutelkti dėmesį dokumentų klasterizavimui. Klasterizavimas pasirinktas ne kaip paieškos ar klasifikacijos priešprieša, o kaip puiki priemonė automatiniam dokumentų grupavimui ir kombinavimui su paieškos bei klasifikacijos metodais atlikti.

1 Elementarus pavyzdys. Ne vienerius metus egzistuojanti Lietuvos naujienų agentūra ELTA turi dvi stambias kategorijų grupes: užsienio naujienas ir Lietuvos naujienas. Kiekvienoje iš šių grupių yra penkios kategorijos: kultūra, politika, sportas, teisėtvarka ir ūkis. Klausimas: kuriai iš šių kategorijų turėtų būti priskirta naujiena apie Indijos vandenyne įvykusį cunamį?

2 Kriminalinių straipsnių sekos pavyzdys. Įvykdžius banko apiplėšimą, išleidžiamas straipsnis, aprašantis įvykį ir lyginantis jį su panašiais (vadinasi bent kartą praeityje aprašytais) įvykdytais nusikaltimais. Sulaikius įtariamuosius, išleidžiamas antras straipsnis, pranešantis apie šį įvykį ir aprašantis praeityje įtariamųjų atliktus darbus. Įvykus pirmajam teismo posėdžiui, išleidžiamas trečias straipsnis, primenantis apie įvykdytą nusikaltimą, pateikiantis tyrimo metu atskleistus faktus ir prognozes apie nusikaltėliams gresiančias bausmes.

Šio darbo tikslas yra klasterizavimo metodo, kuris galėtų skirtingų vartotojų individualioms kategorijoms priskirtus dokumentus suburti į prasmingas ir bendrai naudojamas grupes, sukūrimas.

1.2 KLASTERIZAVIMO METODAS INDIVIDUALIAI KLASIFIKUOTIEMS DOKUMENTAMS

Tradiciniai klasterizavimo metodai nelabai tinka lietuviškų dokumentų klasterizavimui. Šiuose metoduose dokumentai paprastai yra reprezentuojami žodžių dažnumo vektoriais $d = \{w_1, w_2, ..., w_n\}$, kur w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente. Kolekcijos žodynas yra sudaromas iš dokumentų tekste esančių žodžių atmetant nereikšmingus (įvardžius, prielinksnius, jungtukus ir pan.) ir semantiškai tą pačią prasmę turinčius žodžius (sinonimus, žodžių formas). Deja, šiuo metu neegzistuoja nė vienos autoriui žinomos ir laisvai prieinamos priemonės kompaktiškiems lietuviškų dokumentų kolekcijų žodynams sudaryti.

Dokumentų klasterizavimą galima vykdyti ir su nekompaktišku žodynu, tačiau tokiu atveju tenka susitaikyti su ženkliai padidėjusiu klasterizavimo proceso imlumu skaičiuojamiesiems resursams. Lietuvių arba kitos sintetinės kalbos atveju nekompaktiškas žodynas, turintis daug skirtingų to paties žodžio formų, gali taip pat nulemti iškraipytus arba netikslius klasterizavimo rezultatus.

Siūlomame klasterizavimo metode dokumentai yra reprezentuojami jiems vartotojų individualiai priskirtomis kategorijomis – žymėmis (angl. *tags*). Kiekvienas dokumentas yra prilyginamas daugiamatės erdvės vektoriui $d = \{t_1, t_2, ..., t_m\}$, kuriame matmenų skaičius m atitinka žymių kiekį dokumentų kolekcijos *žymių* žodyne, o t_j yra j-osios, kolekcijos *žymių* žodyne esančios, žymės svoris dokumente.

Ši dokumentų reprezentacija remiasi perdirbtomis (Chang et al., 2004), (Chang; Hsu, 2005), (Kummamuru et al., 2003) ir (Golder; Huberman, 2006) darbuose esančiomis idėjomis ir jų rezultatais. Pirmuosiuose darbuose su automatiškai sugeneruojamais raktažodžiais atlikti eksperimentai leido daryti išvadą, kad ženkliai, bet kokybiškai sumažintas dokumentus reprezentuojančių savybių skaičius neturi neigiamos įtakos klasterizavimo rezultatams. Paskutiniajame darbe Golder ir Huberman atlikti tyrimai parodė, kad net ir turėdami subjektyviai susikurtas individualias klasifikacijos schemas, vartotojai dažniausiai yra linkę naudoti sutampančius terminus panašiems objektams apibūdinti. Tai taip pat leido daryti prielaidą, kad lietuviš-

kai (ar kita sintetine kalba) kalbantys vartotojai, klasifikuodami dokumentus kategorijų pavadinimams greičiausiai bus linkę naudoti pirmines žodžių formas.

Siūlomame klasterizavimo metode žymių vektoriais reprezentuojamų dokumentų panašumas yra nustatomas kosinuso koeficientu, o dokumentų kolekcijos klasterizavimas vykdomas eksperimentuojant atrinktu skaldančio K-means (angl. *bisecting K-means*) algoritmu.

Eksperimentų metu naujojo metodo galimybės buvo palygintos su tradicinio dokumentų klasterizavimo metodo, naudojančio nekompaktišką kolekcijos žodyną, galimybėmis. Atliktų eksperimentų rezultatai parodė, kad didėjant dokumentų kiekiui ir/arba skirstinių skaičiui, naujasis metodas tiksliau suformuoja kolekcijos skirstinius ir tai leidžia jį drąsiai naudoti didelėse dokumentų kolekcijose.

1.3 DOKUMENTO TURINYS

Šį dokumentą kartu su įvadu sudaro aštuoni skyriai.

Antrame skyriuje yra supažindinama su pagrindine dokumentų klasterizavimo alternatyva – paieška. Skyriuje yra trumpai apibūdinami paieškos metodai ir nurodomos esminės, visiems paieškos metodams būdingos problemos.

Trečiame skyriuje apibūdinamos klasterizavimo ir klasifikavimo veiklos, nurodomos jų taikymo sritys bei tarpusavio skirtumai.

Ketvirtame skyriuje koncentruojamasi ties klasterizavimu. Skyriuje pateikiama metodų ir algoritmų klasifikacija, nurodomi klasterizavimo proceso žingsniai, detaliai aprašomi jų aspektai ir analizuojama susijusi literatūra.

Penktame skyriuje išsamiai apibūdinama tradicinio lietuviškų dokumentų klasterizavimo problema, apibrėžiama individualios klasifikacijos sąvoka ir pateikiamos individualiai klasifikuotų dokumentų klasterizavimo metodo detalės.

Šeštame skyriuje aprašomi atlikti eksperimentai, kurių metu buvo išrinktas tinkamiausias klasterizavimo algoritmas, o naujojo metodo galimybės palygintos su tradicinio dokumentų klasterizavimo metodo galimybėmis.

Paskutiniuose, septintame ir aštuntame skyriuose, apibendrinami darbo rezultatai, numatomi darbai ateičiai ir pateikiamos išvados.

2 PAIEŠKOS METODAI IR JŲ PROBLEMOS

Norint surasti reikiamus dokumentus, šiuo metu dažniausiai yra naudojami paieškos varikliai. Juose naudojamų metodų tyrimais užsiima daug komercinių organizacijų ir akademinių įstaigų. Paieškos metodai skirstomi į (Small, 1999):

- Sintaksinės paieškos metodus. Paprasčiausi iš paieškos metodų. Paieška atliekama dokumento turinyje žodžio arba frazės lygyje. Kuo dažniau ieškomas žodis arba frazė sutinkama dokumente, tuo tinkamesniu jis yra laikomas. Šiuos metodus naudoja dauguma žiniatinklio paieškos sistemų.
- Paieškos metaduomenyse metodus. Ieškant tinkamų dokumentų paieška atliekama su dokumentais susijusiuose metaduomenyse – antraštėje, autoriaus varde bei pavardėje ir pan. Šie metodai yra nepakeičiami dirbant su binariniais arba uždaro formato dokumentais.
- Semantinės paieškos metodus. Sudėtingiausi ir intelektualiausi iš visų išvardintų metodų. Atliekant paiešką yra naudojami sinonimai ir kitos semantinės koncepcijos. Semantinių paieškų rezultatuose dažnai atsiduria dokumentai, neturintys paieškos frazės, bet atitinkantys paieškos lūkesčius savo turiniu.

Paieškos variklių efektyvumui įvertinti naudojami du dydžiai – rezultatyvumas (angl. *recall*) ir tikslumas (Davies; Cochrane, 1998):

$$Rezultatyvumas = \frac{(Rastų\ tinkamų\ dokumentų\ kiekis)}{(Visas\ tinkamų\ dokumentų\ kiekis)}$$

$$Tikslumas = \frac{(Rastų\ tinkamų\ dokumentų\ kiekis)}{(Rezultatuose\ pateiktų\ dokumentų\ kiekis)}$$

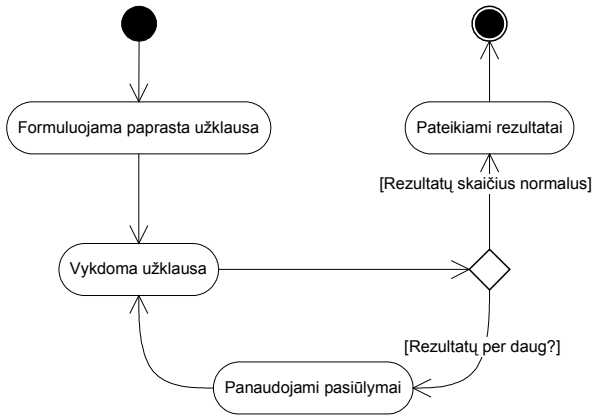
Deja, dažniausiai šie įverčiai nėra aukšti, nes kad ir kokie spartūs, sudėtingi ir intelektualūs būtų paieškos algoritmai, jiems visiems būdingos tos pačios esminės problemos (Hagen, 2000):

- Tipinis vartotojas paprastai nėra pažįstamas su dokumentų saugykla. Keletą kartų gavęs tuščią rezultatų sąrašą, jis galiausiai iki minimumo sumažina paieškos užklausą ir, vėliau gaudamas per ilgus rezultatų sąrašus, pradeda vėl po truputį ją pildyti.

- Vartotojas retai nuo pat pradžių tiksliai žino, ko ieško. Labai dažnai pradėjęs paiešką vienais raktažodžiais ir patyrinėjęs gautus rezultatus, jis iš principo pakeičia nuomonę, ko būtent reikėtų ieškoti.

Abi šios problemos paprastai yra sprendžiamos arba atliekant papildomą paiešką rezultatuose, arba tobulinant paieškos užklausą.

Paieškos užklausos tobulinimas (angl. *query refinement*)y ra labai svarbus informacijos gavybos (angl. *information retrieval*) komponentas, paieškos varikliui interaktyviai siūlantis papildomus paieškos žodžius ir frazes vykdomai užklausai patikslinti (Vélez et al., 1997). Bendri užklausos tobulinimo principai UML diagrama iliustruoti 1 paveiksle:



1 pav. Paieškos užklausos tobulinimo principas

Deja, egzistuojantys paieškos užklausos tobulinimo metodai ne visada gali būti praktiškai taikomi realiomis sąlygomis (Campbell, 2000), nes dažnai pateikia per ilgus pasiūlymų sąrašus. Atlikti eksperimentai (Efthimiadis, 2000) parodė, kad paieškos sistemas su užklausos tobulinimu naudojantys vartotojai geriausiu atveju tinkamais laiko vos trečdalį visų sistemos pateikiamų pasiūlymų. Kitaip tariant, dauguma šiuolaikinių paieškos metodų vartotojui suteikia labai mažai pasirinkimo. Tenka rinktis arba per ilgą rezultatų sąrašą, arba laiko užimantį ir netrivialų paieškos užklausos tobulinimą.

3 KLASTERIZAVIMO IR KLASIFIKAVIMO PRINCIPAI

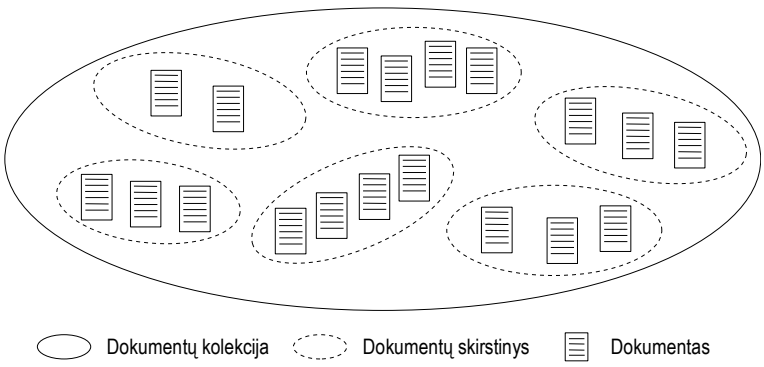
Dokumentų klasterizavimas siūlo alternatyvų būdą greitam reikalingos informacijos suradimui. Kitaip nei atliekant paieškos užklausas, klasterizuojant dokumentai į bendras grupes – skirstinius (angl. *cluster*) – patenka ne pagal ieškomų žodžių ar frazių buvimą/nebu-

vimą juose, bet pagal dokumentų savybių (dokumentų turinys taip pat gali būti laikomas savybe) panašumą (Jain et al., 1999).

Aplinkos objektų klasterizavimas yra įgimta kiekvieno žmogaus savybė. Net maži vaikai aplinkinius asmenis intuityviai skirsto į vyrus ir moteris, o gaunamus patiekalus – į saldžius, karčius, sūrius, rūgščius ir pan. Valinga klasterizavimo veikla žmonija užsiima tūkstančius metų (Kural, 1999). Klasterinė analizė yra taikoma tokiose įvairiose mokslo srityse kaip gamtos mokslai (biologija, zoologija), medicina (psichiatrija, patologijos), socialiniai mokslai (archeologija, sociologija, kriminologija), geografija, geologija, inžineriniai mokslai (struktūrų atpažinimas, kibernetika) (Anderberg, 1973). Klasterinė analizė yra išnaudojama DNR struktūrom sistematizuoti (Hartuv et al., 1999), programinės įrangos dekompozicijai atlikti (Andritsos, 2005) ir net muzikos įrašams pagal atlikėjo balso charakteristikas suskirstyti (Wei-Ho et al., 2004).

Pirmieji siūlymai panaudoti klasterizavimą informacijos gavybai pagerinti buvo pateikti 1971 metais Jardine ir Van Rijsbergen darbuose (Jardine; Van Rijsbergen, 1971). Buvo tikimasi, kad panaudojus klasterizavimą informacijos gavybos sistemų (angl. *information retrieval systems*) efektyvumas padidės, kadangi organizuojant dokumentus į skirstinius pagal jų savybių panašumą jie daugiau ar mažiau atitiks intuityvų vartotojo elgesį ieškant.

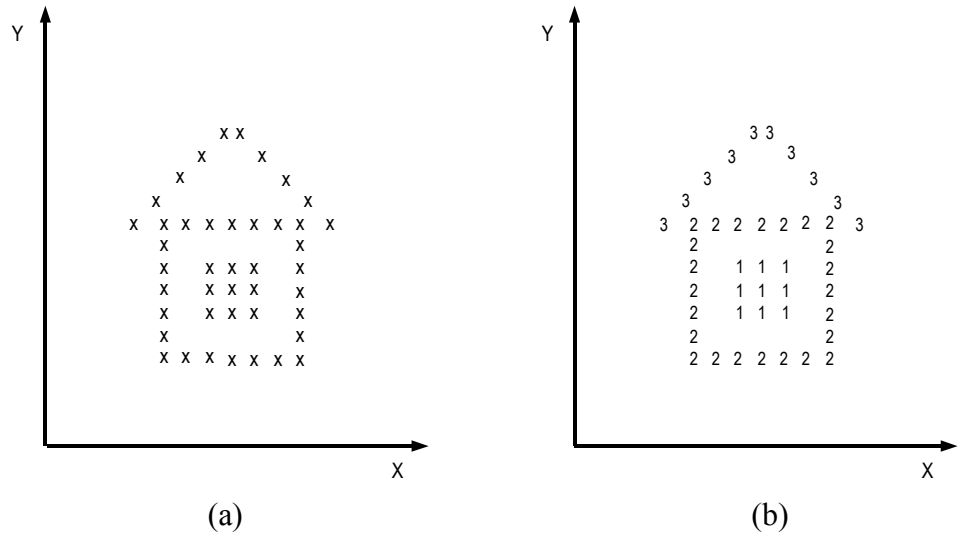
Efektyvumo padidėjimo lūkesčiai buvo grindžiami skirstinių hipoteze (angl. *cluster hypothesis*) (Jardine; Van Rijsbergen, 1971). Ši hipotezė teigia, kad tiesiogiai susiję dokumentai turi daugiau tarpusavio panašumų vienas su kitu nei su nesusijusiais dokumentais, todėl yra linkę atsidurti tuose pačiuose skirstiniuose. Kitaip tariant, hipotezė daro prielaidą, kad dokumentų kolekcijose (žr. 2 pav.) tarpusavyje susijusius dokumentus visada įmanoma suskirstyti į atskiras grupes pagal jų savybių panašumus.



2 pav. Skirstiniai dokumentų kolekcijoje

Dokumentų klasterizavimas yra labai dažnai painiojamas su dokumentų klasifikavimu[3]. Reikia pastebėti, kad nors klasterizavimas ir klasifikavimas (dar vadinamas kategorizavimu) yra teksto gavybos (angl. *text mining*) veiklos turinčios daug panašių savybių, jos taip pat turi ir keletą esminių skirtumų.

Klasifikuojant dokumentus turima reikalų su iš anksto apibrėžta kategorijų (dažnai vadinamų klasėmis) aibe. Klasifikacijos proceso metu priklausomai nuo turinio dokumentas yra priskiriamas kuriai nors iš kategorijų. Klasterizuojant dokumentus jokios iš anksto apibrėžtos kategorijų aibės nėra. Dokumentai suburiami į grupes tik pagal savo savybių panašumą. Kitaip tariant, klasterizuoti dokumentai atskleidžia natūralius užsislėpusius rinkinius, o klasifikuoti dokumentai yra „prievarta“ įspraudžiami į numatytus organizacinius rėmus (Willet, 1988). Ši idėja iliustruota 3 paveiksle, kur (a) pavaizduotas taškų išdėstymas, slepiantis (b) skaitmenimis pavaizduotus ir intuityviai atskiriamus skirstinius.



(a) (b)

3 pav. Skirstiniai iš taškų

Žvelgiant iš techninės pusės, dokumentų klasifikacija yra apmokymo su mokytoju (angl. *supervised learning*) problema. Tam kad būtų įmanoma atlikti automatinį klasifikavimą, prieš tai visoms kategorijoms būtina rankiniu būdu priskirti dokumentų rinkinius pagal kuriuos turi apsimokyti klasifikatorius–automatas. Pavyzdiniai dokumentai turi būti itin kruopš-

3 Greičiausiai šito priežastis – literatūroje, pavyzdžiui, (Jain et al., 1999), naudojamų terminų panašumas. Klasterizavimas yra dažnai vadinamas neprižiūrimu klasifikavimu (angl. *unsupervised classification*), o klasifikavimas (arba kategorizavimas) – prižiūrimu klasifikavimu (angl. *supervised classification*). Painiavą taip pat didina ir tai, kad ankstyvoje literatūroje terminai *klasifikavimas* ir *klasterizavimas* buvo laikomi sinonimais.

čiai atrenkami, nes jų tinkamumas labai stipriai nulemia vėlesnį klasifikatoriaus efektyvumą. Dokumentų klasterizavimas yra apmokymo be mokytojo (angl. *unsupervised learning*) problema. Klasterizuojant dokumentai sugrupuojami be jokių išankstinių apmokymų. Tiesiog stengiamasi, kad panašūs dokumentai atsidurtų tuose pačiuose, o nepanašūs – skirtinguose skirstiniuose. (Deboeck, 1998), (Jain et al., 1999)

Tradiciškai klasifikavimas ir klasterizavimas yra atliekamas statiškai visai dokumentų kolekcijai vienu metu[4] (Cutting et al., 1992), tačiau šias metodikas lygiai taip pat galima taikyti ir dinamiškai[5]. Paskutiniuosius du dešimtmečius labai domimasi galimybe panaudoti klasifikavimą ir klasterizavimą paieškos užklausų rezultatams struktūrizuoti (Hearst; Pedersen, 1996), (Leouski; Croft, 1996) ir yra laikoma, kad potencialiai tai gali labai pagerinti paieškos efektyvumą (Hearst; Pedersen, 1996).

4 KLASTERIZAVIMO METODAI

4.1 METODŲ IR ALGORITMŲ KLASIFIKACIJA

Literatūroje yra aprašyta begalė klasterizavimo metodų ir algoritmų. Net apsiribojus svarbiausiais ir labiausiai paplitusiais iš jų galima sudaryti kelias lygiagrečias taksonomijas.

Taksonomijos gali būti sudaromos atsižvelgiant į (Jain et al., 2004):

- Įvesties duomenų pateikimo pobūdį. Įvesties duomenys gali būti pateikiami panašumo matricomis, struktūrų (angl. *pattern*) matricomis, skaitmenimis, kategorijomis, tekstu, grafais ir kt.
- Išvesties duomenų pateikimo pobūdį. Klasterizavimo rezultatai gali būti pateikiami skirstinių rinkiniais arba skirstinių rinkinių hierarchija.
- Naudojamą tikimybinį modelį (jeigu jis yra naudojamas iš vis).
- Optimizacijos procesą.
- Klasterizavimo kryptį. Skirstiniai gali būti sudaromi apjungiant (aglomeruojant) klasterizuojamus objektus arba juos išskaidant.

4 Angliškai šis procesas įvardijamas kaip *static* arba *pre-retrieval* klasifikacija ir klasterizavimas.

5 Angliškai šis procesas įvardijamas kaip *dynamic* arba *post-retrieval* klasifikacija ir klasterizavimas.

Kadangi šiose tezėse yra svarbus greito dokumentų suradimo klausimas, labai besisiejantis su skirstinių atvaizdavimu, klasterizavimo metodai bus apžvelgiami suskirsčius algoritmus pagal išvesties duomenų (rezultatų) pateikimo pobūdį (žr. 4.5 skyrių).

4.2 KLASTERIZAVIMO PROCESAS

Nepriklausomai nuo pasirenkamo metodo, klasterizuojant yra atliekami penki žingsniai (Jain et al., 1999), (Theodoridis; Koutroumbas, 2003). Pritaikius juos dokumentų klasterizavimui, žingsnius galima išvardinti kaip:

- Dokumentų reprezentacijos parinkimas. Klasterizuojant dokumentus, būtina pasirinkti atributus, kurie taikant algoritmą reprezentuos kiekvieną dokumentą. Dažniausiai dokumentas yra reprezentuojamas daugiamatės erdvės vektoriumi $X = \{x_1, x_2, ..., x_n\}$, kur n yra klasterizuojamos dokumentų kolekcijos žodyne esantis žodžių skaičius.

- Asociacijos mato parinkimas. Šis matas apibrėžia kiek panašūs ar nepanašūs yra tarpusavyje du dokumentai. Mato parinkimas daro didelę įtaką galutiniams klasterizavimo rezultatams.

- Klasterizavimo algoritmo parinkimas. Be abejonės, norint atlikti dokumentų kolekcijos klasterizavimą, būtina pasirinkti konkretų algoritmą, atliekantį kolekcijos struktūrizavimą.

- Skirstinių reprezentavimo pasirinkimas. Dažniausiai klasterizavimo rezultatas yra daugiamačiai skirstiniai. Norint pasinaudoti atliktu struktūrizavimu, reikia pasirinkti būdus daugiamačių rezultatų projekcijoms į dvimatę arba trimatę erdvę atlikti.

- Rezultatų patikrinimas. Klasterizavimo algoritmui pateikus rezultatus juos būtina patikrinti. Paprastai tai atliekama panaudojant atitinkamus testus.

4.3 DOKUMENTŲ REPREZENTACIJA KLASTERIZAVIMO METU

Atliekant pirmąjį žingsnį klasterizavimo proceso metu reikia nuspręsti, kokios konkrečios savybės apibūdina kiekvieną dokumentą.

Dažniausiai dokumento reprezentacijai yra panaudojamas dokumento tekstas. Tokiu atveju dokumentas yra prilyginamas n-matės erdvės vektoriui $d = \{d_1, d_2, ..., d_n\}$, kur n yra klasterizuojamos dokumentų kolekcijos žodyne esantis žodžių skaičius, o w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente.

Kiekvienas svoris w_j paprastai yra apskaičiuojamas panaudojant tam reikalui plačiai dokumentų klasterizavime taikomą formulę

$$Svoris = tj_j \times idf_j \ , \tag{1}$$

kur tf_j yra j-ojo žodžio pasikartojimo dažnumas dokumente, o idf_j yra invertuotas j-ojo žodžio pasikartojimo dažnumas visoje dokumentų kolekcijoje. Invertuotas pasikartojimo dažnumas yra suformuotas taip, kad itin dažnai pasikartojantys žodžiai neužgožtų rečiau pasitaikančių, bet reikšmingų žodžių, o labai reti žodžiai taip pat neįgautų per daug reikšmės. Šis dažnumas yra apskaičiuojamas pagal formulę

$$idf_j = \log\frac{n}{n_j} \ , \tag{2}$$

kur n yra dokumentų kiekis visoje dokumentų kolekcijoje, o n_j yra dokumentų kiekis, turintis j-ąjį žodį.

Kadangi kiekvienas dokumentas gali būti skirtingo ilgio, svoriai w_j yra papildomai normalizuojami taip, kad dokumentų vektorių ilgis būtų lygus vienetui. Svorių normalizavimui naudojama formulė:

$$w_j = \frac{tf_j \times df_j}{\sqrt{\sum_{r=1}^{m} (tf_r \times df_r)^2}} \ . \tag{3}$$

Kolekcijos žodynas yra sudaromas iš dokumentų tekste esančių žodžių. Savaime suprantama, kad realiuose dokumentuose toli gražu ne visi jų turinį sudarantys žodžiai yra vienodai reikšmingi. Žodžiai gali turėti daug skirtingų formų, semantinių atitikmenų, o į tokias kalbos dalis kaip įvardžiai, prielinksniai ir pan. dėmesio nereikėtų kreipti iš vis. Dėl šios priežasties sudarant dokumentų kolekcijos žodyną atliekamas žodžių filtravimo procesas. Šio proceso metu yra:

- atmetami nereikšmingi žodžiai iš negatyvaus žodyno (angl. *stop-list dictionary*);
- remiantis sinonimų žodynais atmetami semantiškai tą pačią prasmę turintys žodžiai;
- morfologiniais analizatoriais – lemuokliais (angl. *lemmatizer*) – atrenkamos pirminės žodžių formos arba kamieno atskyrimo programomis (angl. *stemmer*) atrenkami žodžių kamienai.

Gali atrodyti, kad kolekcijos žodyno sudarymo metu yra sumažinamas dokumentų reprezentacijos pilnumas, tačiau atlikti eksperimentai parodė, kad priklausomai nuo pasirinktų klasterizavimo metodų, mažesnis dokumento vektoriaus matmenų kiekis gali net padidinti klasterizavimo efektyvumą ir tikslumą (Shaw, 1990), (Burgin, 1991), (Burgin, 1995).

Netgi labai kruopščiai atrinkus dokumentus reprezentuojančias savybes yra logiška tikėtis, kad ne visos savybės turi tokią pačią svarbą. Savybių reikšmingumui įvertinti klasterizavime yra įvedami savybių svoriai. Pavyzdžiui, dokumentų, kuriuos reprezentuoja jų turinio tekstas, atveju, šie svoriai gali atitikti vieno ar kito žodžio pasikartojimų skaičių tekste. Nors nėra vieningos nuomonės, kiek reikšmingą įnašą daro savybių svorių naudojimas dokumentų klasterizavimo atveju, vis daugiau darbų šioje srityje bando išnaudoti ne binarinius dokumentų reprezentacijos vektorius (Korpimies; Ukkinen, 1998), (Friedman et al., 2004), (Liping et al., 2005).

Dokumentų reprezentacijai yra naudojami ne tik juose esantys žodžiai, bet ir frazės (Hatzivassiloglou et al., 2000) bei žodžių deriniai (Yuan-Chao et al., 2004). Kadangi didelis dokumentus aprašančių savybių kiekis neigiamai atsiliepia dokumentų kolekcijos klasterizavimo spartai, yra bandoma savybių vektorius, pagrįstus kolekcijos žodynu, pakeisti automatiškai sugeneruotais raktažodžiais pagrįstais vektoriais (Chang et al., 2004), (Chang; Hsu, 2005), (Kummamuru et al., 2003).

4.4 DOKUMENTŲ PANAŠUMO MATAVIMAS

Apibrėžus dokumentus reprezentuojančias savybes būtina taip pat apibrėžti jų tarpusavio panašumą nusakančias matavimo priemones. Šiai dienai klasterinėje analizėje panašumo matavimo priemonių yra sukurta tikrai daug, tačiau visas jas galima suskirstyti į keturias bazines grupes (Tan et al., 2005):

- panašumo koeficientai;
- nepanašumo koeficientai;
- tikimybiniai koeficientai;
- koreliacijos koeficientai.

Tikimybinių ir koreliacijos koeficientų naudojimas dokumentų klasterizavime nėra labai paplitęs, todėl didžioji literatūros dalis koncentruojasi ties pirmosiomis grupėmis, dažnai vadinamų vienu vardu – artumo (angl. *proximity*) koeficientais (Hand et al., 2001).

Matematiškai dokumentų panašumą ar nepanašumą galima apibrėžti atstumo funkcija

$$D : C \times C \to R \quad ,$$

kur C yra klasterizuojamų dokumentų kolekcijos aibė, o R – neneigiamų realiųjų skaičių aibė. Dažniausiai naudojamos metrinės atstumo funkcijos t.y. funkcija D tenkina šias sąlygas:

1. $D(x, y) = 0, \ kai \ x = y$;
2. $D(x, y) \geq 0, \ \forall x, y \in C$;
3. $D(x, y) = D(y, x), \ \forall x, y \in C$;
4. $D(x, y) \leq D(x, z) + D(z, y), \ \forall x, y, z \in C$.

Klasterinėje analizėje atstumo funkcijas yra įprasta normalizuoti taip, kad jos grąžintų reikšmes intervale nuo *0* iki *1*. Tokiais atvejais turint nepanašumo funkciją, labai lengva paskaičiuoti panašumo reikšmę ir atvirkščiai:

$$\text{jei} \quad nep(x, y) = D(x, y) \quad ,$$

$$\text{tai} \quad pan(x, y) = 1 - D(x, y) \quad .$$

Konkretus atstumo funkcijos įgyvendinimas priklauso nuo situacijos. Pateiksime pavyzdį. Tarkime, kad n dokumentų kolekcijos C dokumentus d_i nuspręsta reprezentuoti kolekcijos žodyne (žr. 4.4 skyrių) esančiais žodžiais. Tokiu atveju kiekvieną dokumentą galima atvaizduoti kaip binarinį m-matį vektorių $d_i = \{w_1, w_2, ..., w_m\}$, kur m yra žodžių kiekis kolekcijos žodyne, o w_j nurodo ar j-asis, dokumentų kolekcijos žodyne esantis, žodis yra dokumente d_i. Šitoje situacijoje atstumo funkcija $D(d_i, d_j)$ galėtų grąžinti dydį, kuris nurodytų kiek procentaliai žodžių sutampa dokumentuose d_i ir d_j.

Atstumų funkcija gautiems panašumų įverčiams saugoti reikalinga $n \times n$ dydžio panašumų matrica $S(C)$, kurios elementai s_{ij} nurodo dviejų dokumentų d_i ir d_j panašumo laipsnį. 4 paveiksle pavaizduota 5×5 dydžio panašumų matrica atvejui, kai dokumentų skaičius n kolekcijoje C yra *5*, o dokumentų tarpusavio panašumas paskaičiuotas pagal aukščiau pateiktą atstumo funkcijos pavyzdį.

	d_1	d_2	d_3	d_4	d_5
d_1	1	0,6	0,1	0,4	0,2
d_2	0,6	1	0,5	0,2	0
d_3	0,1	0,5	1	0,9	0,3
d_4	0,4	0,2	0,9	1	0,8
d_5	0,2	0	0,3	0,8	1

4 pav. Panašumų matrica

Nors simetriška ($s_{ij} = s_{ji}$) $n \times n$ dydžio panašumų (arba nepanašumų) matrica yra paprastas ir elegantiškas būdas dokumentų artumui nustatyti, jis turi savų silpnybių. Kadangi matricos užpildymo kaina yra

$$O\left(n\frac{(n-1)}{2}\right) \quad ,$$

o reikalavimai atminčiai – $O(0,5 \times n^2)$, didelės dokumentų kolekcijos atveju darbui su matrica gali prireikti labai didelių atminties ir skaičiuojamųjų resursų[6].

Matricos užpildymo efektyvumo problema yra aktuali iki šių dienų, todėl nenuostabu, kad yra nemažai darbų šioje srityje. Vienas iš būdų efektyvumui padidinti yra aprašomas (Silverstein; Pedersen, 1997). Šiame darbe autoriai remiasi elementaria idėja, kad toli gražu ne visi dokumente esantys žodžiai yra labai reikšmingi tarpusavio artumui nustatyti. Dėl šios priežasties skaičiuodami panašumų matricos elementus, jie sutrumpina dokumentus reprezentuojančių ir į kolekcijos žodyną įtraukiamų žodžių sąrašą iki 20–50 reikšmingiausių žodžių. Labai panašiu keliu darbuose (Chang et al., 2004), (Chang; Hsu, 2005) yra nuėję kiti autoriai, kurie reprezentuodami dokumentus pasitenkina automatiškai sugeneruotais raktažodžiais.

Tiek (Silverstein; Pedersen, 1997), tiek (Chang et al., 2004) ir (Chang; Hsu, 2005) darbuose remiamasi idėja, kad jeigu dokumentą galima apibūdinti vektoriumi $d = \{w_1, w_2, ..., w_m\}$, kur w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente, tai lygiai taip pat kiekvieną kolekcijos žodyno žodį galima išreikšti vektoriumi $w = \{d_1, d_2, ..., d_n\}$, kur d_i yra žodžio w svoris i-ajame dokumente. Į kolekcijos žodyną atrinkdami tik reikšmin-

6 33 tūkstančių dokumentų kolekcijai artumų matricos sudarymui prireiktų daugiau nei 0,5 milijardo operacijų ir daugiau nei dviejų gigabaitų atminties tuo atveju, jeigu kiekvienam matricos elementui būtų išskiriama po keturis baitus (standartinis sveikajam skaičiui reikalingas atminties kiekis) atminties.

giausius žodžius arba raktažodžius, autoriai stipriai sumažina vektoriaus w matmenų kiekį. Pasinaudodami tuo, jie galiausiai klasterizuoja ne didelę kolekciją, o reikšmingiausius žodžius arba raktažodžius ir pagal gautus skirstinius vėliau atitinkamai organizuoja dokumentus.

Kiti autoriai artumų matricos dydį yra linkę mažinti kitokiais būdais. Kadangi tipiškoje panašumų matricoje dauguma jos elementų yra nuliniai arba artimi nuliui, (Smeathon et al., 1998) yra siūloma ne $n \times n$ dydžio, bet $n \times k$ dydžio matrica, kur k – eksperimentų metu parenkamas dydis. Autorių naudotoje hierarchinėje aglomeratyvioje klasterizavimo sistemoje k nurodydavo žemiausią vertingą suskaidymo į skirstinius lygį.

Paprastai klasterizuojant objektus yra skaičiuojamas jų tarpusavio nepanašumo dydis (Jain et al., 1999). Objektams, turintiems tolydžias savybes, dažniausiai naudojamos šios atstumų funkcijos (Jain et al., 1999), (Húsek et al., 2006):

- Manhetano atstumas L_1

$$D(x_i, x_j) = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right| \tag{4}$$

- Euklidinis atstumas L_2

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2} \tag{5}$$

- Čebiševo atstumas L_∞

$$D(x_i, x_j) = \max_{k=1..m} \left(\left| x_{ik} - x_{jk} \right| \right) \tag{6}$$

Visos aukščiau išvardintos atstumų funkcijos yra atskiri Minkovskio atstumo L_q

$$D(x_i, x_j) = \sqrt[q]{\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^q} \tag{7}$$

atvejai, kur m yra erdvės matmenų skaičius, o mažesni q ($1 \le q \le \infty$) atitinka didesnį atsparumą atsiskyrėliams (problema detaliau aptariama 4.5.1 skyriuje).

Nors objektų panašumui nustatyti naudojamas kosinuso koeficientas

$$\text{cosine}(x_i, x_j) = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2} \sqrt{\sum_{k=1}^{m} x_{jk}^2}} \tag{8}$$

kartu su euklidiniu atstumu (5) yra populiariausi dokumentų artumo nustatymo būdai, jų yra gerokai daugiau (Jain et al., 1999), (Húsek et al., 2006). Universalaus artumo nustatymo būdo nėra, kiekvienas jų tinka vis kitokioms klasterizavimo situacijoms.

Penki artumo nustatymo būdai (tarp jų euklidinis atstumas, kosinuso ir Žakardo[7] koeficientai) yra lyginami (Kirriemuir; Willett, 1995) darbe. Šiame darbe eksperimentais buvo nustatyta, kad hierarchiškai klasterizuojant dokumentus Žakardo ir kosinuso koeficientai leidžia sudaryti tiksliausius skirstinius.

Panašūs eksperimentai buvo atliekami (Rorvig, 1999) darbe lyginant Daiso[8], Žakardo, kosinuso, persidengimo ir asimetrijos koeficientais suformuotus skirstinius. Autorius pasinaudojo kolekcija su iš anksto į atskiras grupes suskirstytais dokumentais ir šias grupes lygino su eksperimentų metu gautais skirstiniais. Darbo rezultatuose nurodoma, kad atkuriant grupių struktūrą geriausiai pasirodė kosinuso ir persidengimo koeficientai.

4.5 KLASTERIZAVIMO ALGORITMAI

Pats populiariausias klasterizavimo algoritmų skirstymas pagal savybes yra skirstymas pagal tai sudaro skirstiniai hierarchiją ar ne. Jeigu po klasterizavimo skirstiniai sudaro hierarchiją, tai atliekantys algoritmai vadinami hierarchiniais. Jeigu po klasterizavimo skirstiniai hierarchijų nesudaro, tai atliekantys algoritmai vadinami dalinančiais (angl. *partitioning*).

Klasterinės analizės literatūroje (Jain et al., 2004) yra išvardinta daugiau algoritmų suskirstymo būdų, trumpai išvardintų šio dokumento 4.1 skyriuje.

Žemiau esančiuose skyriuose trumpai apžvelgiamos abi algoritmų grupės. Pateikiami bendri algoritmų veikimo principai, pavyzdžiai, išvardinamos silpnosios ir stipriosios pusės.

7 Angliškoje literatūroje minimas kaip *Jaccard*.

8 Angliškoje literatūroje minimas kaip *Dice*.

4.5.1 DALINANTIS KLASTERIZAVIMAS

Visi dalinančio klasterizavimo algoritmai pagrįsti vienu bendru principu. Pradedant klasterizavimą, n objektų kolekcija $C = \{x_1, x_2, ..., x_n\}$, kur x_i – vektoriumi aprašytas objektas, yra sudalinama į k pradinių skirstinių C_j, kuriems yra priskiriami visi kolekcijos objektai (pradžioje tai galima atlikti tiesiog atsitiktine tvarka). Klasterizavimo eigoje optimizuojant nuo konkretaus metodo priklausančią tikslo funkciją objektai iteratyviai perkeliami iš vieno skirstinio į kitą. Klasterizavimas baigiamas, kai iteracijos metu nebeatliekamas nė vienas objekto perkėlimas.

Dalinančio klasterizavimo algoritmai yra labai patrauklūs, nes turi neaukštus reikalavimus skaičiuojamajai technikai. Kolekcijai iš n objektų klasterizavimo kaina paprastai būna intervale nuo $O(n)$ iki $O(n \times log\ n)$ (Willet, 1988). Dalinančio klasterizavimo algoritmai yra gerokai efektyvesni už hierarchinio klasterizavimo algoritmus, aprašomus 4.5.2 skyriuje.

Galutinis dalinančio klasterizavimo rezultatas labai stipriai priklauso nuo pradinių skirstinių parinkimo. Tai, kad prieš pradedant klasterizavimą tenka priimti prielaidas apie skirstinių formą ir jų kiekį k, yra laikoma pagrindiniu dalinančio klasterizavimo algoritmų trūkumu (Jain et al., 1999), (Tan et al., 2005). Paprastai tai sukelia itin didelių sunkumų, nes toli gražu ne kiekvienoje srityje kiekį k galima numatyti iš vis. Ypač tai sunku padaryti, kai objektų skaičius labai didelis. Vykdoma nemažai tyrimų bandant surasti būdus skirstinių kiekiui parinkti atsižvelgiant į klasterizuojamus duomenis (pavyzdžiui, (Peleg; Moore, 2000), (Rasmussen, 2000), (Fred; Jain, 2002)), tačiau iš esmės ši problema kol kas laikoma neišspręsta.

Vienas seniausių ir dažniausiai naudojamų dalinančio klasterizavimo algoritmų yra K-means (Jain et al., 1999), (Tan et al., 2005).

K-means algoritmo pradžioje klasterizuojamoje objektų aibėje parenkami pradiniai k skirstinių centrai. Po to algoritmas iteratyviai vykdomas dviem žingsniais. Pirmo žingsnio metu kiekvienas aibės objektas yra priskiriamas tam skirstiniui, kuriame esantis centras yra arčiausiai objekto. Antrojo žingsnio metu kiekvienam pirmajame žingsnyje suformuotam skirstiniui apskaičiuojamas naujas skirstinio geometrinis centras c_j (centras gali nesutapti nė su vienu skirstinio objektu). Žingsniai kartojami tol, kol nenusistovi tikslo funkcijos reikšmė. Žemiau pateikiamas vaizdus algoritmas pseudoprogramavimo kalba:

```
1: Parenkami pirminiai skirstinių centrai;
2: repeat
```

```
3:     Objektai priskiriami skirstiniui su artimiausiu jiems centru;
4:     Perskaičiuojami skirstinių centrai;
5: until (centrai nebesikeičia).
```

Priskiriant objektus skirstiniui su artimiausiu jiems centru, būtina metrika, kuri pasakytų kiek arti vienas kito yra du pasirinkti taškai. Euklidinėje erdvėje paprastai tam yra panaudojamas 4.4 skyriuje aprašytas euklidinis atstumas. Dokumentų atveju paprastai yra naudojamas tame pačiame 4.4 skyriuje aprašytas kosinuso koeficientas. Žinoma, priklausomai nuo poreikių ir situacijos galima lygiai taip pat naudoti kitas artumo metrikas, pavyzdžiui, Manheteno atstumą euklidinei erdvei ir Žakardo arba Daiso koeficientus dokumentams.

Algoritmui perskaičiuojant skirstinių centrus, reikalinga tikslo funkcija, kuri iteracijų metu turi būti nuolat optimizuojama. Euklidinės erdvės atveju dažniausiai yra naudojama kvadratinių klaidų suma (angl. *sum of the squared error*), kuri literatūroje (Jain et al., 1999) dar yra vadinama kvadratinės klaidos kriterijumi:

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in C_j} \text{dist}(c_j, x_i)^2 \quad . \tag{9}$$

Kvadratinės klaidos kriterijuje esanti funkcija *dist* yra ne kas kita kaip jau šiame skyriuje minėtas euklidinis atstumas (5), c_j – skirstinio C_j centras, o x_i – objekto, esančio skirstinyje C_j, savybių vektorius. Optimizuojant SSE reikšmė yra mažinama.

K-means algoritmu klasterizuojant dokumentus yra tikslingiau naudoti ne euklidinį atstumą, o kosinuso panašumo metriką. Tarkime, kad turime dokumentų kolekciją C, kurioje kiekvienas dokumentas d_i yra aprašytas daugiamačiu savybių vektoriumi (pavyzdžiui, 4.3 skyriuje minėtu žodžių dažnių vektoriumi). Tokiu atveju klasterizavimo iteracijos tikslas yra padidinti dokumentų ir skirstinių centrų panašumą. Kitaip tariant, reikia padidinti skirstinių glaudumą (angl. *cohesion*). K-means algoritmo ir dokumentų atveju tikslo funkcijai yra naudojamas kvadratinės klaidos kriterijaus analogas – bendrasis glaudumas (angl. *total cohesion*):

$$TC = \sum_{j=1}^{k} \sum_{d_i \in C_j} \text{cosine}(c_j, d_i) \quad . \tag{10}$$

Bendrojo glaudumo aprašyme esanti *cosine* šiuo atveju yra kosinuso funkcija (8), c_j – skirstinio C_j centras, o d_i – objekto, esančio skirstinyje C_j, savybių vektorius. Optimizuojant *TC* reikšmė yra didinama.

K-means algoritmas garantuoja, kad jis turės baigtinį iteracijų skaičių, jeigu yra parenkamas baigtinis pradinių skirstinių skaičius *k*, o kiekvienos paskesnės iteracijos metu tikslo funkcija yra optimizuojama. Deja, algoritmas negarantuoja, kad jis baigsis pasiekus optimaliausią galimą tikslo funkcijos reikšmę, nes tai priklauso nuo pradinių skirstinių centrų parinkimo. Problema išsamiai aprašyta ir nuodugniai išnagrinėta (Tan et al., 2005).

Vienas iš dažniausiai praktikoje naudojamų būdų pradinių centrų problemai išspręsti yra labai elementarus. Kadangi algoritmas nėra itin reiklus skaičiuojamiesiems resursams, išbandomi keli klasterizavimo variantai, kiekvieną kartą pasirenkant vis kitus atsitiktinius pirminius centrus. Gavus rezultatus yra palyginamos kiekvienu atveju gautų tikslo funkcijų reikšmės ir atrenkamas geriausias klasterizavimo variantas.

Kitas taip pat dažnai praktikoje naudojamas būdas pradinių centrų problemai išspręsti atsitiktinai parenka tik pirmąjį tašką. Antrasis, trečiasis ir kiti centrai parenkami taip, kad jie būtų kiek galima toliau nuo prieš tai parinktų centrų. Toks taškų parinkimas praktiškai garantuoja gerai išsklaidytus pradinius centrus, tačiau kartu yra labai jautrus taškams–atsiskyrėliams, kurie gali stipriai iškraipyti natūralius objektų aibėse esančius skirstinius, todėl gali būti naudojamas tik kompaktiškose kolekcijose. Šiai problema apeiti naudojami atsiskyrėlių identifikavimo algoritmai, minimi darbuose (Jiang et al., 2001) ir (Hautamaki et al., 2005).

Įdomiai pradinių centrų problema yra išspręsta K-means plėtinyje – skaldančiame K-means (angl. *bisecting K-means*). Jis pagrįstas elementaria idėja: norint suskaidyti objektų aibę į *k* skirstinių, iš pradžių reikia suskaidyti ją į du skirstinius, tada pasirinkti ir suskirstyti prastesnį iš jų, vėl pasirinkti prastesnį iš jų ir tai tęsti tol, kol nebus gauti visi reikiami *k* skirstinių. Kadangi gauti skirstiniai nebūtinai yra labai optimalūs (jie daugiau ar mažiau optimalūs tik skirstinyje, iš kurio atsirado patys), dažnai gautų skirstinių centrai panaudojami kaip pirminiai centrai įprastam K-means algoritmui pakartoti.

Dar viena labai įdomi detalė, susijusi su skaldančiu K-means, yra tai, kad algoritmo rezultatai iš tikrųjų sudaro skirstinių rinkinių hierarchiją ir (Steinbach et al., 2000) eksperimentai rodo, kad klasterizuojant tekstinius dokumentus algoritmas tam tinka netgi labiau nei aglomeratyvus hierarchinis klasterizavimo algoritmas.

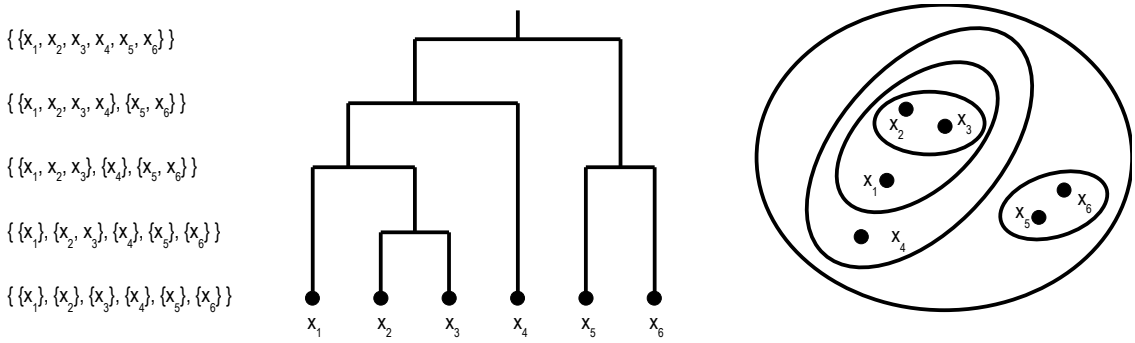
Apibendrinimui galima pasakyti, kad K-means yra paprastas ir efektyvus (netgi tais atvejais, kai jį tenka kartoti kelis kartus) klasterizavimo algoritmas, kurį galima taikyti įvairių rūšių duomenims. Jo rezultatai yra priklausomi nuo pradinių centrų parinkimo, tačiau yra daugiau ar mažiau sėkmingesnių būdų šiai problemai apeiti. Deja, K-means nelabai tinka tais at-

vejais, kai skirstiniai klasterizuojamoje aibėje nėra sferiniai, labai skiriasi jų tankumas arba jie nėra kompaktiški (yra atsiskyrėlių). Tokiais atvejais gali padėti tik specialūs būdai atsiskyrėliams identifikuoti ir pakankamai didelis pradinių skirstinių skaičius. (Tan et al., 2005)

4.5.2 HIERARCHINIS KLASTERIZAVIMAS

Hierarchinio klasterizavimo algoritmai yra antra labai svarbi klasterizavimo algoritmų grupė. Kaip ir K-means, populiariausi hierarchinio klasterizavimo algoritmai atsirado vieni pirmųjų ir yra plačiai naudojami iki šių dienų. Dauguma dabar egzistuojančių hierarchinių algoritmų yra single-link, complete-link arba Ward metodo atmainos (Jain et al., 1999). Populiariausiais jų yra laikomi single-link ir complete-link algoritmai.

Kitaip nei dalinančio klasterizavimo algoritmai, hierarchinis klasterizavimas dokumentus suburia ne į skirstinių rinkinį, o į skirstinių hierarchiją. Labai panašūs objektai yra suburiami į mažesnius skirstinius, o šie medžio principu suburiami į didesnius skirstinius, kuriuose esančių objektų bendras tarpusavio panašumas yra mažesnis. Paveiksle nr. 5 yra pateikta vaizdi hierarchinio klasterizavimo iliustracija.



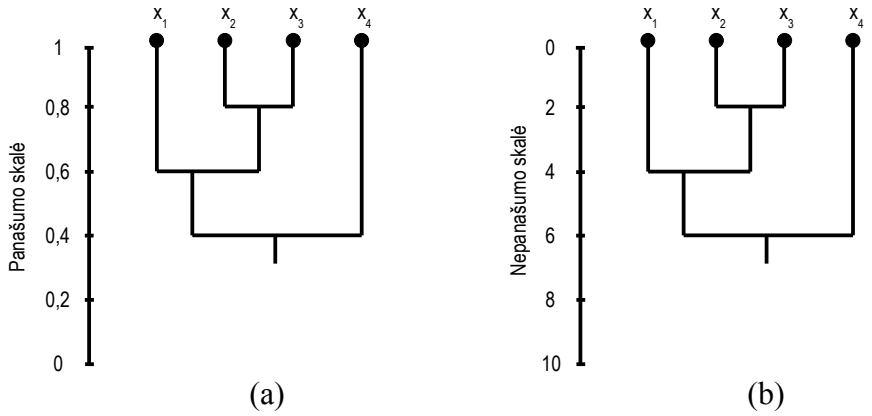
5 pav. Hierarchinio klasterizavimo iliustracija

Tarkime, kad klasterizuojama *n* objektų aibė yra apibrėžta kaip $C = \{x_1, x_2, ..., x_n\}$, kur x_i yra objektas, aprašytas daugiamačiu vektoriumi. Skirstiniai C_i pagal tarpusavio panašumą gali būti sugrupuoti į rinkinius $G_j = \{C_i \mid 1 \leq i \leq k\}$.

Rinkinys G_1, turintis *q* skirstinių, yra laikomas įterptu į skirstinių rinkinį G_2, turintį $r < q$ skirstinių, jeigu kiekvienas skirstinys rinkinyje G_1 yra kurio nors G_2 rinkinio skirstinio poaibis ir bent vienas G_1 rinkinio skirstinys yra griežtas kurio nors G_2 rinkinio skirstinio poaibis. Pavyzdžiui, rinkinys $G_1 = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$ yra įterptas į skirstinių rinkinį $G_1 = \{\{x_1, x_3,$

$x_4\}$, $\{x_2, x_5\}\}$ ir nėra įterptas į rinkinį $G_3 = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$ (pavyzdžiai paimti iš (Theodoridis; Koutroumbas, 2003)).

Hierarchiniai klasterizavimo algoritmai yra skirstomi į dvi grupes: aglomeratyvius algoritmus (angl. *agglomerative*) ir skaidančius (angl. *divisive*) algoritmus. Vienintelis jų skirtumas yra tai, kaip jie formuoja skirstinių rinkinius. Bet kuriuo atveju, tiek vienos, tiek kitos grupės algoritmų rezultatus mėgstama pateikti dendogramomis (Jain et al., 1999), (Theodoridis; Koutroumbas, 2003), (Tan et al., 2005). Jos dažniausiai vaizduojamos medžiu su šalia pateikta objekto panašumo (a) arba nepanašumo (b) lygių skale (žr. 6 pav.).



6 pav. Dendogramų su panašumo (a) ir nepanašumo (b) skalėmis pavyzdžiai

Aglomeratyvūs algoritmai yra gerokai populiaresni nei skaidantieji ir veikia būdu „iš apačios į viršų“. Pirmojo žingsnio metu sudaromas pirmasis skirstinių rinkinys G_0, sudarytas iš n skirstinių, turinčių lygiai po vieną objektą iš C kolekcijos (kitaip tariant, pradėtas algoritmas visus klasterizuojamos aibės objektus sukelia į savus skirstinius). Antrojo žingsnio metu yra sujungiami du panašiausi skirstiniai taip sudarant antrąjį rinkinį G_1, turintį $n–1$ skirstinių, ir esantį G_0 viršaibiu. Žingsniai tęsiami tol, kol nesuformuojamas rinkinys G_{n-1}, teturintis vienintelį skirstinį su visais C elementais. Algoritmo rezultatas yra skirstinių rinkinių hierarchija:

$$G_0 \subset G_1 \subset \ldots \subset G_{n-1}$$

Žemiau pateikiamas vaizdus algoritmas pseudoprogramavimo kalba:

```
1: Apskaičiuojama skirstinių artumų matrica;
```

```
2: repeat
3:     Apjungiami du artimiausi skirstiniai;
4:     Atsižvelgiant į pasikeitusį skirstinių skaičių
       perskaičiuojama skirstinių artumų matrica;
5: until (lieka tik vienas skirstinys).
```

Itin resursams reiklūs (detalės pateikiamos vėliau) skaidantieji hierarchinio klasterizavimo algoritmai veikia atvirkščiu būdu „iš viršaus į apačią“. Pirmasis skirstinių rinkinys G_0 yra sudarytas iš vienintelio skirstinio, turinčio visus C elementus. Paskutinis rinkinys G_{n-1} yra sudarytas iš n skirstinių, turinčių lygiai po vieną dokumentą iš C. Algoritmo rezultatas yra skirstinių rinkinių hierarchija:

$$G_{n-1} \subset G_{n-2} \subset \ldots \subset G_0$$

Kadangi praktikoje skaidantieji hierarchinio klasterizavimo algoritmai yra taikomi retai, toliau šiame skyriuje bus kalbama beveik vien tik apie aglomeratyvius algoritmus.

Galima pastebėti, kad svarbiausia aglomeratyvaus klasterizavimo algoritmo operacija yra dviejų skirstinių artumo nustatymas. Būdai, kuriais ši operacija yra atliekama, praktiškai vieninteliai skiria vieną aglomeratyvų algoritmą nuo kito, todėl kartu suteikia pastariesiems savo pavadinimą.

Grafo pagrindu reprezentuojamų skirstinių artumui nustatyti dažniausiai yra naudojamos single-link (taip pat žinomo kaip MIN) ir complete-link (taip pat žinomo kaip MAX) technikos. Single-link atveju dviejų skirstinių artumas prilyginamas artumui tarp dviejų artimiausių abiejų skirstinių objektų. Naudojant grafų teorijos terminologiją tai reikštų, kad dviejų skirstinių artumą apsprendžia trumpiausia briauna tarp dviejų viršūnių, esančių skirtinguose viršūnių poaibiuose. Complete-link atveju dviejų skirstinių artumas prilyginamas artumui tarp dviejų tolimiausių abiejų skirstinių objektų. Naudojant grafų teorijos terminologiją tai reikštų, kad dviejų skirstinių artumą apsprendžia ilgiausia briauna tarp dviejų viršūnių, esančių skirtinguose viršūnių poaibiuose. Nors yra atvejų, kada verta taikyti single-link algoritmą, (situacijos puikiai iliustruotos (Jain et al., 1999)), yra laikoma, kad complete-link algoritmas paprastai sukuria kompaktiškesnius skirstinius ir naudingesnes hierarchijas (Jain; Dubes, 1988).

Į skirstinius taip pat galima žiūrėti kaip į aibes su centrais. Šiuo atveju artumą tarp dviejų skirstinių galima prilyginti tiesiog artumui tarp jų centrų, tačiau dažniau yra naudojama tikslesnė technika – Ward metodas (dar vadinamas mimimum-variance metodu). Šiame metode, panašiai kaip dalinančio klasterizavimo atveju, atstumai tarp skirstinių skaičiuojami opti-

mizuojant tikslo funkciją – 4.5.1 skyriuje aprašytus kvadratinės klaidos kriterijų su euklidiniu atstumu, bendrąjį glaudumą su kosinuso kriterijumi ir kt. Tiesa, Ward metodo autoriaus amžininkai kritikavo šią techniką už polinkį į sferinius, daugiau ar mažiau panašaus dydžio skirstinius (Cormack, 1971), (Milligan et al., 1983).

1967 metais Lance ir Williams (Lance; Williams, 1967) parodė, kad egzistuoja bendra kombinatorinė lygtis, nusakanti kaip skirtingi aglomeratyvūs algoritmai keičia artumų matricos reikšmes po kiekvieno skirstinių apjungimo:

$$p(R,Q) = \alpha_A p(A,Q) + \alpha_B p(B,Q) + \beta p(A,B) + \gamma \left| p(A,Q) - p(B,Q) \right| \qquad (11)$$

Šioje lygtyje p yra artumo funkcija, Q ir R – skirstiniai, tarp kurių nustatinėjamas artumas, be to, R yra skirstinys, suformuotas apjungiant skirstinius A ir B. Koeficientai α_A, α_B, β, γ yra priklausomi nuo konkretaus klasterizavimo algoritmo ir yra apskaičiuojami naudojant parametrus m_A, m_B, m_Q, kurie nurodo objektų skaičių skirstiniuose A, B ir Q. Lentelėje nr. 1 pateikiamos Lance-Williams lygties koeficientų reikšmės šiame skyriuje paminėtiems aglomeratyvaus hierarchinio klasterizavimo algoritmams.

1 lentelė. Lance–Williams lygties koeficientai skirtingiems algoritmams

Algoritmas	α_A	α_B	β	γ
Single-link	1/2	1/2	0	–1/2
Complete-link	1/2	1/2	0	1/2
Skirstinių centrai	$\frac{m_A}{m_A + m_B}$	$\frac{m_B}{m_A + m_B}$	$\frac{-m_A m_B}{(m_A + m_B)^2}$	0
Ward metodas	$\frac{m_A + m_Q}{m_A + m_B + m_Q}$	$\frac{m_B + m_Q}{m_A + m_B + m_Q}$	$\frac{-m_B}{m_A + m_B + m_Q}$	0

Kitaip nei dalinančių klasterizavimo algoritmų atveju, hierarchinių klasterizavimo metodų algoritmai yra labai reiklūs skaičiuojamiesiems resursams. Kaip ir dalinančių algoritmų atveju, hierarchiniams algoritmams reikalinga panašumų matrica, turinti $O(0,5 \times n^2)$ reikalavimus atminčiai (tipiniu atveju, kai matrica yra simetriška), kur n yra klasterizuojamų objektų kiekis. Deja, kartu reikia saugoti ir informaciją apie kiekvieną hierarchijos lygį, todėl visas hierarchiniam klasterizavimo algoritmui reikalingos atminties poreikis išauga iki $O(n^2)$. Pana-

ši situacija yra su laiko reikalavimais. Iteratyviai kuriant skirstinių hierarchiją atliekami *n–1* ciklų, kiekvieno jų metu perskaičiuojant artumų matricą. Neoptimizuotam algoritmui reikalingos laiko sąnaudos yra $O(n^3)$, kurias optimizavus įmanoma sumažinti iki $O(n^2 \times log\ n)$ (Tan et al., 2005). Skirstantieji hierarchinio klasterizavimo algoritmai yra dar reiklesni – jau pirmojo skirstinio skaidymo metu algoritmui tenka sumokėti $O(2^n)$ laiko kainą.

Lyginant hierarchinius klasterizavimo algoritmus su dalinančiais, hierarchiniai algoritmai yra įvairiapusiškesni. Pavyzdžiui, tiek minėtas single-link, tiek minėtas complete-link algoritmai gali puikiai susitvarkyti su prastai atskirtais, skirtingo tankio, nekompaktiškais duomenimis, kai tuo tarpu K-means gerai susitvarko tik su sferiniais, gerai atskirtais, kompaktiškais duomenimis.

4.5.3 GRIEŽTO IR NEGRIEŽTO KLASTERIZAVIMO ALGORITMAI

Visi iki šiol aprašyti klasterizavimo algoritmai buvo griežti (angl. *exclusive, hard*), nes vieną objektą priskyrus vienam skirstiniui, jis jokiu būdu negalėdavo kartu atsidurti ir kitame skirstinyje. Matematiškai šie apribojimai gali būti apibrėžiama kaip:

$$C = \bigcup_{i=1}^{k} C_i$$
$$C_i \neq \emptyset\ ,\ i = 1,2,\ldots,k$$
$$C_i \cap C_j = \emptyset\ ,\ i \neq j\ \ i,j = 1,2,\ldots,k$$

Žinoma, praktikoje taip pat pasitaiko ir labai daug situacijų, kada objektai vienu metu priklauso kelioms grupėms, pavyzdžiui, asmuo universitete gali būti ne tik studentu, bet ir darbuotoju. Tokiais atvejais yra naudojami negriežto (angl. *non-exclusive, soft, fuzzy*) klasterizavimo algoritmai, kuriems negalioja paskutinė iš aukščiau nurodytų sąlygų.

Negriežto klasterizavimo metu objektai į skirstinius priskiriami kartu su svorio koeficientu, esančiu tarp *0* ir *1* taip, kad bendra objekto svorių suma būtų lygi vienetui. Deja, pastarasis apribojimas neleidžia iki galo išspręsti priklausomybės keliems skirstiniams tais atvejais, kai objekto sąryšis su grupėmis yra lygiavertis. Pavyzdžiui, jeigu vyras yra ir sūnus, ir tėvas, kuriai vyrų grupei jis priklauso labiau – sūnų ar tėvų?

Populiariausias negriežto klasterizavimo algoritmas yra c-means (dažnai vadinamas trumpiniu FCM), kurio žingsniai labai primena K-means. Nuo pastarojo algoritmo c-means skiriasi skirstinio centro ir tikslo funkcijos skaičiavimuose dalyvaujančiais koeficientais iš priklausomybės svorių matricos (Jain et al., 1999).

4.6 SKIRSTINIŲ PATEIKIMAS

Skirstinių pateikimas yra vienas svarbiausių aspektų dokumentų klasterizavime. Anastasios Tombros savo disertacijoje (Tombros, 2002) kalbėdamas apie dokumentų skirstinių pateikimą mini dvi sąvokas: vidinis pateikimas ir pateikimas išorei. Vidinis pateikimas yra susijęs su skirstinių atstovų parinkimu šitaip stengiantis sumažinti užklausose dalyvaujančių duomenų kiekį. Pateikimas išorei susijęs su skirstinių pateikimu vartotojui taip, kad pastarasis galėtų kaip galima sėkmingiau pasinaudoti skirstinių teikiamais privalumais.

4.6.1 VIDINIS PATEIKIMAS

Tombros (Tombros, 2002) mini, kad tradiciškai vidiniu skirstinių pateikimu apsiima sprendžiant efektyvumo problemas. Juo yra naudojamasi informacijos gavyboje skirstiniais pagrįstose užklausose, kurios yra atliekamos ne su visu skirstinio turiniu, o su jo atstovu. Skirstinio atstovui atitinkant užklausos keliamas sąlygas, rezultatuose pateikiamas visas skirstinys.

Klasterizavimo apžvalgoje (Jain et al., 1999) pateikiami trys būdai atstovui parinkti. Populiariausiu iš jų yra laikomas skirstinio centro (angl. *centroid*) panaudojimas. Kaip ir K-means algoritmo atveju, centras nebūtinai turi būti skirstiniui priklausantis objektas, pavyzdžiui, dokumentas. Netgi atvirkščiai, dažniausiai tam naudojamas specialus darinys. Bet kuriuo atveju, skirstinio atstovui visada keliamos dvi sąlygos (Tombros, 2002):

- atstovas turi pakankamai tiksliai apibūdinti skirstinio, kuriam atstovauja, turinį;
- atstovas turi pakankamai gerai nurodyti atstovaujamo skirstinio skirtumus nuo likusių skirstinių.

Yra keletas mokslinių darbų, nagrinėjančių kaip geriau parinkti skirstinių atstovus, tačiau kaip pastebi Tombros, esamoje literatūroje nėra atsakyta į pagrindinius klausimus, problema nepakankamai išsamiai išnagrinėta, todėl šioje srityje dar galima nuveikti labai daug.

4.6.2 PATEIKIMAS IŠOREI

Klasterizavimo rezultatų pateikimo vartotojui būdas yra labai svarbi klasterizavimo metodo savybė. Tinkamas skirstinių pateikimas dokumentų klasterizavime yra ne ką mažiau svarbesnis, nei pačių skirstinių suformavimas, nes tik remdamasis apie skirstinius pateikiama informacija vartotojas naršydamas gali surasti reikiamus dokumentus.

Turbūt patys populiariausi būdai dokumentų skirstiniams pateikti yra dažniausiai skirstinyje sutinkamų žodžių ir juose esančių dokumentų pavadinimų nurodymas. Tokio pobūdžio pateikimas yra naudojamas (Hearst; Pedersen, 1996), (Neto et al., 2000), (Roussinov; Chen, 2001) darbuose aprašomose sistemose. (Anick; Vaithyanathan, 1997) bei (Maarek et al., 2000) darbuose autoriai yra pasukę truputį kitokiu, bet ganėtinai panašiu keliu ir skirstinius bando pateikti naudodami dokumentuose esančias frazes.

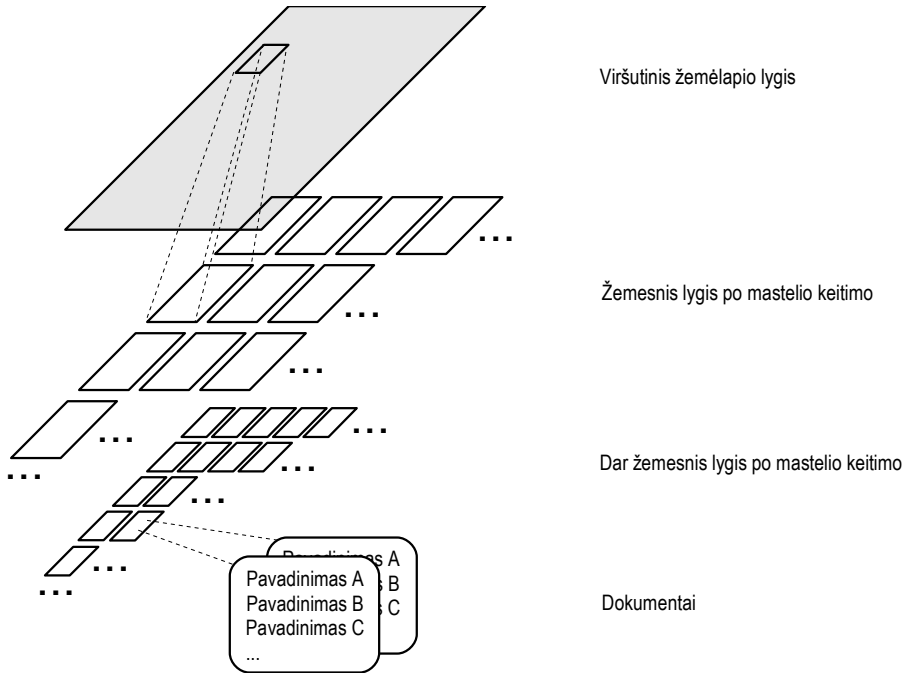
Visiškai kitu keliu yra pasukę savaime susitvarkančiais žemėlapiais (angl. *self-organizing maps*) besiremiančių klasterizavimo metodų autoriai. Savaime susitvarkantys žemėlapiai (angliškas trumpinys – SOM) yra praėjusio amžiaus paskutiniuosiuose dešimtmečiuose Kohoneno pasiūlyta ir išdirbta neuronų tinklais pagrįsta paradigma, kurioje lygiagrečiai rūpinamasi ir duomenų klasterizavimu ir klasterizavimo rezultatų pateikimu (Kohonen, 1998), (Kohonen, 2000).

SOM yra plačiai naudojami kaip būdas skaidančiam klasterizavimui atlikti įvairiose srityse, tarp jų ir tekstinių dokumentų klasterizavime. Ypač šiuo aspektu yra įdomus WEBSOM metodas, kuriame savaime susitvarkančių žemėlapių algoritmas yra panaudojamas dokumentų skirstinių projekcijai iš daugiamačių rezultatų į dvimatę plokštumą (Kaski et al., 1998), (Lagus, 2000).

Naudojant WEBSOM metodą dokumentų kolekcija yra išdėstoma grafiniame žemėlapyje[9] taip, kad pastarasis ne tik padėtų susidaryti bendrą supratimą apie jos turinį, bet ir leistų joje naršyti. Tai pasiekiama aprašant kiekvieną dokumentą reikšminių žodžių histograma ir priklausomai nuo konteksto išdėstant panašių dokumentų grupes atvaizduojamame žemėlapyje greta viena kitos. Žemėlapyje vartotojas gali greitai susidaryti įspūdį apie žodžių pasikartojimo laipsnį atskirose dokumentų kolekcijos vietose, nes skirtingo tankio zonos yra išskiriamos tamsesnėmis arba šviesesnėmis spalvomis.

Kadangi WEBSOM žemėlapiai yra organizuoti hierarchiniu principu (žr. 7 pav. – iliustraciją pagal (Lagus, 2000) pavyzdį), vartotojui spragtelėjus vieną arba kitą zoną keičiamas mastelis taip pateikiant dar detalesnį dominančios vietos vaizdą. 8 (b) paveiksle atvaizduotas žemėlapis, gautas (a) spragtelėjus ties žodžiu „neuron“ (pavyzdžiai paimti iš WEBSOM interaktyvios demonstracijos tinklapio).

9 Net kelios interaktyvios WEBSOM metodo demonstracijos yra pasiekiamos internete adresu http://websom.hut.fi/websom/ [žiūrėta 2006-04-16].

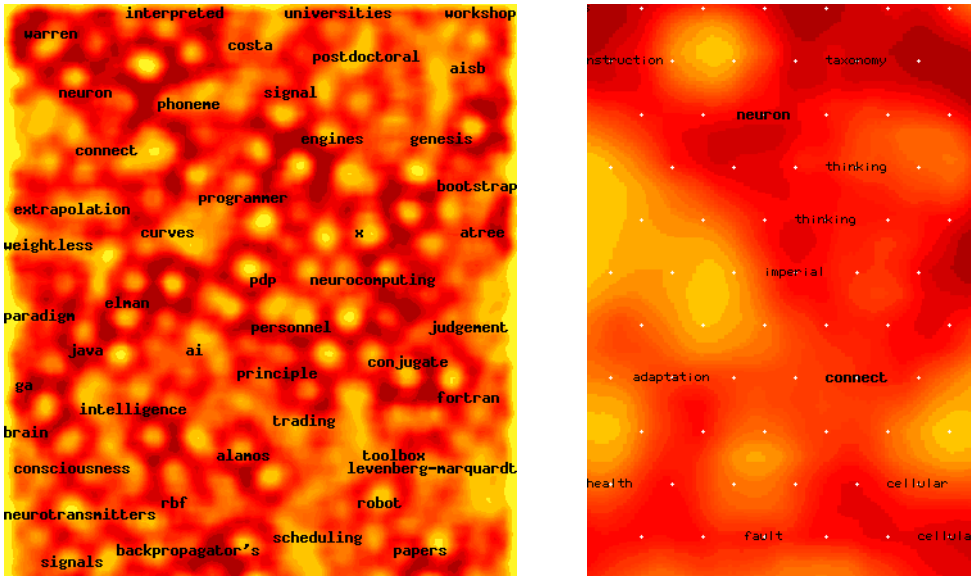


7 pav. Hierarchinė WEBSOM dokumentų žemėlapių prigimtis

Metodo autoriai neapsiriboja vien vizualia navigacija ir naršymą kolekcijoje palengvina reikšminių žodžių paieška žemėlapyje.

Tiesa, reikia pripažinti, kad bandant WEBSOM demonstracinius žemėlapius gali būti pakankamai sunku surasti reikiamus dokumentus (tiksliau – straipsnius naujienų grupėse). Dideliame žemėlapyje, sudarytame reikšminių žodžių pagrindu, nelengva pataikyti į reikiamą kontekstą iš pirmo karto. Dėl šios priežasties Rusijos tyrėjų grupė sukūrė WEBSOM atmainą TopSOM, kuriame dokumentai yra reprezentuojami ne reikšminių žodžių, o dokumento tematikos (angl. *topic*) histogramomis (Pleshko et al., 2001). TopSOM metodas taip pat pagerina skirstinių pateikimą granuliuodamas žemėlapio vaizdą šešiakampiais[10].

10 TopSOM demonstracija yra pasiekiama internete adresu http://demo.rco.ru/topsom/default.asp?LANGUAGE=RUS&INFO_SYSTEM=topdemo [žiūrėta 2006-04-16].



(a) (b)

8 pav. WEBSOM žemėlapio pavyzdys

Viena įdomi detalė jungia daugumą klasterizavimo problematiką nagrinėjančius darbų. Nepaisant to, kad skirstinių pateikimas vartotojui (ypač pastarojo požiūriu) yra esminis viso klasterizavimo proceso momentas, nemažai autorių šią problemą yra linkę sąmoningai ar nesąmoningai ignoruoti. Pagrinde yra koncentruojamasi ties dokumentų reprezentacijos, panašumo ir klasterizavimo algoritmų problemomis. Autoriai retai bando aprašyti galutinį rezultatų pateikimą, dar rečiau bandoma nustatyti, kiek naudojamas skirstinių pateikimas vartotojui yra efektyvus. Dažniausiai šios problemos sprendimas yra atidedamas išvadose ar apibendrinimuose paminint jį tarp darbų ateičiai.

Gera esamos tendencijos išimtimi yra Kural darbai (Kural, 1999) ir (Kural et al., 2001), kuriuose ji kartu su bendraautoriais pagrinde stengėsi atsakyti į klausimą, ar vartotojai gali pakankamai efektyviai surasti jiems reikiamą skirstinį pateiktuose rezultatuose. Beje, jos atliktų eksperimentų apibendrinimuose galiausiai yra padaroma išvada, kad klasterizavimas nėra tinkama ir patikima priemonė dokumentams grupėmis pateikti. Tiesa, čia pat kyla pagrįstų abejonių dėl autorės atliktų eksperimentų pilnumo – buvo išbandytas vos vienas klasterizavimo metodas ir, kas yra dar svarbiau, tik vienas skirstinių pateikimo būdas. Kural eksperimentai tikrai nebuvo pakankami tokioms išvadoms pateikti ir čia reikėtų sutikti su Tombros (Tombros, 2002) išsakyta kritika.

Kitokie klasterizavimo metodai ir skirstinių pateikimo būdai taip pat liko neišnagrinėti ir (Wu et al., 2001) darbe. Jame eksperimentų metu autorius skirstinius bandė pateikti dešimčia dažniausiai dokumentuose pasikartojančių žodžiu, penkiomis dažniausiai pasikartojančiomis žodžių poromis ir trimis labiausiai paieškos užklausą atitikusių dokumentų antraštėmis. Eksperimentų apibendrinime pažymima, kad dažniausiai vartotojams pavykdavo surasti reikiamą skirstinį, tačiau dauguma jų taip pat išreiškė savo nepasitenkinimą rezultatų pateikimo būdu.

Labai daug darbų apie atitikimą ir pateikimą yra apžvelgta puikioje Mizzaro (Mizzaro, 1997) apžvalgoje, kurioje autorius apibendrino paskutiniuosius keturių dešimtmečių tyrimus šioje srityje. Vienoje iš apibendrinamųjų išvadų Mizzaro teigia, kad pateikimas labai stipriai veikia vartotojų sprendimą apie dokumentų tinkamumą jų poreikiams. Pateikiama medžiaga leidžia daryti prielaidą, kad tinkamumą geriausiai nusako dokumento santrauka (angl. *abstract*), šiek tiek prasčiau už ją tai atlieka automatiškai paimtos dokumentų ištraukos, pavadinimai, citatos ir dažniausiai pasikartojantys reikšminiai žodžiai.

Tiesa, reikia pastebėti, kad visi (Mizzaro, 1997) apžvalgoje minimi autoriai dirbo kitokiame nei dokumentų klasterizavimas kontekste. Pagrindinė jų nagrinėjama problema – greitas trumpos esminės informacijos apie kiekvieną dokumentą, esantį paieškos rezultatuose, pateikimas. Dokumentų klasterizavimo atveju abstrakcijos lygmuo turi būti dar aukštesnis, nes rezultatuose dokumentus atstovauja skirstiniai. Būtent skirstinio pateikimas nulemia vartotojo sprendimus apie jame esančių dokumentų atitikimą poreikiams.

Deja, galiausiai tenka pripažinti, kad kaip ir vidinio pateikimo atveju, skirstinių pateikimo išorei srityje iki šiol yra atlikta nepakankamai tyrimų. Greičiausiai tai lemia dvi priežastys:

- Inertiškumas ir uždarumas. Dauguma tyrėjų turi polinkį gilintis į stabilias ir ne tokias įnoringas kaip žmonės matematines problemas taip paminant vieną iš esminių Demingo principų – orientaciją į vartotoją.
- Srities sudėtingumas ir painumas. Jau minėtas Mizzaro kitame savo darbe (Mizzaro, 1998) labai taikliai pastebi, kad atitikimas informacijos gavyboje yra daugialypė problema, iki šiol netgi neturinti nusistovėjusios vientisos terminijos.

4.7 SKIRSTINIŲ VALIDAVIMAS

Visus klasterizavimo metodus vienija bendra silpnybė – kadangi jie yra kuriami tam, kad aptiktų duomenyse struktūras, jie jas „suranda“ netgi ir tuo atveju, kai jokių realių struktūrų duomenyse nėra (Theodoridis; Koutroumbas, 2003). Dėl šios priežasties kiekvieno klasterizavimo algoritmo rezultatai turėtų būti validuojami proceso, kuris vadinamas skirstinių validavimu (angl. *cluster validity*), metu.

Skirstinių validavimo technikos yra skirstomas į tris grupes:

- išorinio kriterijaus;
- vidinio kriterijaus;
- santykinio kriterijaus.

Išorinio kriterijaus technikų atveju klasterizavimo algoritmas yra tikrinamas su specialiai tam paruoštomis duomenų kolekcijomis, kuriose yra iš anksto dirbtinai sukuriamos objektų struktūros. Klasterizavimo rezultatai yra lyginami su realiai kolekcijoje esančiomis grupėmis ir nustatomas jų atitikimo lygis. Šios technikos privalumas yra tai, kad žinant kolekcijos struktūrą, gautus algoritmo rezultatus ir tikėtinus rezultatus yra pakankamai nesudėtinga palyginti (Tombros, 2002). Technikos trūkumu yra laikoma tai, kad klasterizavimo algoritmo kokybę ji gali nusakyti tik išbandytų kolekcijų kontekste (Tombros, 2002).

Ko gero populiariausia iš išorinio kriterijaus technikų, skirtų dirbtinių kolekcijų sukūrimui ir klasterizavimo algoritmo pateiktų bei tikėtinų rezultatų palyginimui atlikti, yra Monte Carlo.

Vidinio kriterijaus technikų atveju klasterizavimo algoritmas yra tikrinamas atsižvelgiant į vidinius duomenis, pavyzdžiui, artumų matricą. Skirtingos šios grupės technikos gali būti taikomos skirstinių rinkinio validavimui bei skirstinių hierarchijos validavimui (Theodoridis; Koutroumbas, 2003), tačiau praktikoje yra taikomos retai (Tombros, 2002).

Kitaip nei aukščiau jau apibūdintos, santykinio kriterijaus grupei priklausančios technikos nėra pagrįstos statistiniu testavimu. Šių technikų esmė yra daugkartinis tiriamojo algoritmo išbandymas su ta pačia duomenų kolekcija, bet vis su kitais pradiniais parametrais, stengiantis atrasti tinkamiausius iš jų (Theodoridis; Koutroumbas, 2003), taip patį algoritmą pritempiant prie reikiamos situacijos.

5 INDIVIDUALIAI KLASIFIKUOTŲ DOKUMENTŲ KLASTERIZAVIMO METODAS

Iki šiol darbe buvo apžvelgtos esminės paieškos metodų problemos, nurodyti klasterizavimo ir klasifikacijos skirtumai, aprašyti klasterizavimo principai kartu trumpai apžvelgiant dokumentų klasterizavimo srities problemas sprendžiančius darbus. Šiame skyriuje yra koncentruojamasi ties autoriaus siūlomu dokumentų klasterizavimo metodu, besiremiančio individualia klasifikacija. Poskyriuose yra apibūdinamos konkrečios sprendžiamos problemos ir pateikiami jų sprendimo būdai. Aprašomas klasterizavimo metodas yra išbandomas ir palyginamas su tradiciniu klasterizavimo metodu eksperimentinėje dalyje (žr. 6 skyrių).

5.1 DOKUMENTŲ REPREZENTACIJA KLASTERIZAVIMO METU

5.1.1 LIETUVIŠKŲ DOKUMENTŲ KOLEKCIJOS ŽODYNO SUDARYMO PROBLEMOS

Kaip jau buvo minėta 4.3 skyriuje, klasterizuojant dažniausiai dokumento reprezentacijai yra naudojamas dokumento tekstas. Tokiais atvejais dokumentas prilyginamas n-matės erdvės vektoriui $d = \{w_1, w_2, ..., w_n\}$, kur w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente.

Kolekcijos žodynas yra sudaromas iš dokumentų tekste esančių žodžių. Savaime suprantama, kad realiuose dokumentuose toli gražu ne visi jų turinį sudarantys žodžiai yra vienodai reikšmingi. Žodžiai gali turėti daug skirtingų formų, semantinių atitikmenų, o į tokias kalbos dalis kaip įvardžiai, prielinksniai ir pan. dėmesio nereikėtų kreipti iš vis. Dėl šios priežasties sudarant dokumentų kolekcijos žodyną atliekamas žodžių filtravimo procesas. Šio proceso metu yra:

- atmetami nereikšmingi žodžiai iš negatyvaus žodyno (angl. *stop-list dictionary*);
- remiantis sinonimų žodynais atmetami semantiškai tą pačią prasmę turintys žodžiai;
- morfologiniais analizatoriais – lemuokliais (angl. *lemmatizer*) – atrenkamos pirminės žodžių formos arba kamieno atskyrimo programomis (angl. *stemmer*) atrenkami žodžių kamienai.

Žodžių filtravimo procesas itin svarbus sudarant sintetinėmis kalbomis (tarp jų ir lietuvių) parašytų dokumentų kolekcijų žodynus. Nors kitaip nei analitinėse kalbose (tipinė analitinė kalba – anglų kalba) sintetinėse nėra gausu įvairių tarnybinių ir pagalbinių žodžių, o

vienu žodžiu dažnai įmanoma pasakyti tai, kam išreikšti analitinei kalbai reikia kelių žodžių, sintetinėse kalbose yra labai daug žodžių kaitymo. Dėl naudojamų žodžių formų gausybės, nefiltruoti sintetinėmis kalbomis parašytų dokumentų kolekcijų žodynai yra gerokai didesni už nefiltruotus analitinėmis kalbomis parašytų dokumentų kolekcijų žodynus[11].

Deja, formalizuotų analizės metodų taikymas sintetinių kalbų (ypač senų ir neplačiai naudojamų) morfologijoje dažniausiai yra komplikuotas. Ne išimtis yra ir lietuvių kalba. Dėl menko kalbos paplitimo ir nepakankamo dėmesio šiai sričiai šiuo metu[12] egzistuoja tik vienas autoriui žinomas automatinis lietuvių kalbos morfologinis analizatorius (Zinkevičius, 2000) ir tik vienas elektroninis sinonimų žodynas (Tildės biuras 2006[13] dalis). Neegzistuoja nė vienos autoriui žinomos lietuvių kalbos žodžių kamieno atskyrimo programos ir nė vieno negatyvaus žodyno. Negana to, nei morfologinis analizatorius, nei sinonimų žodynas nėra laisvai platinami, todėl praktiškai nėra jokių priemonių kompaktiškiems lietuviškų dokumentų kolekcijų žodynams sudaryti.

5.1.2 INDIVIDUALAUS KLASIFIKAVIMO PRINCIPAI

Paskutiniu metu sparčiai populiarėja aprašančių terminų – žymių (angl. *tags*) – priskyrimo įvairiems objektams technika, vadinama žymėjimu (ang. *tagging*) (Golder; Huberman, 2006). Žymės yra objektus aprašantys metaduomenys. Tai raktažodžiai, atliekantys tematikos arba kategorijos vaidmenį ir yra naudojami glaustam objektų savybių apibūdinimui. Naudodamiesi žymėmis vartotojai gali organizuoti objektus su panašiomis savybėmis į tarpusavyje persidengiančius rinkinius arba atlikti 2 skyriuje aprašytą paiešką metaduomenyse.

Kitaip nei kitoje labai dažnai naudojamoje klasifikavimo formoje – hierarchiniame kategorizavime – žymėmis pagrįstos kategorijos nesudaro griežtai atskirtų aibių. Šitaip yra išvengiama dažnų situacijų, kada objekto neįmanoma priskirti vienai taksonomijos klasei, nes pastarasis turi keletui klasių būdingų požymių.

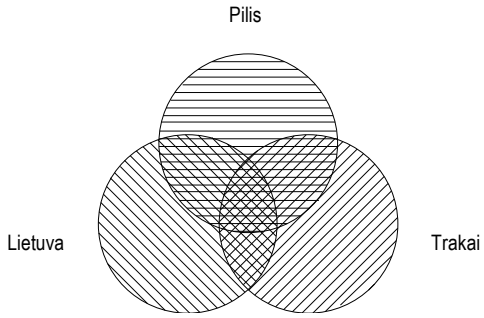
---

11 Elementarus pavyzdys. Kauno technologijos universiteto tinklapio (http://www.ktu.lt) apžvalgos puslapio angliškoje versijoje yra apytiksliai 10% daugiau žodžių negu lietuviškoje to paties puslapio versijoje, tačiau nefiltruotas lietuviškos versijos žodynas yra apytiksliai 26% didesnis už angliškąjį. Atliekant primityvų angliško žodyno išgryninimą, pastarojo dydis sumažėja visu trečdaliu ir tampa per pus mažesniu už nefiltruotą lietuvišką žodyną.

12 Paskutinį kartą šiomis problemomis domėtasi 2006 metų balandžio mėnesį.

13 Produkto aprašymas pateikiamas adresu http://www.tilde.lt/PORTAL/go/tildelt/3722/en-US/DesktopDefault.aspx [žiūrėta 2006-04-30].

Žymėjimas yra panašus į tinklinį kategorizavimą be hierarchinės sistemos. Dėl hierarchijos nebuvimo visos objektams priskirtos žymės turi vienodą svorį, o tinklinė žymėjimo prigimtis leidžia sudaryti persidengiančias klases. Šis principas yra iliustruotas 9 paveiksle, kuriame Oilerio diagrama atvaizduotos Trakų pilies fotografijai priskirtinos žymės.



9 pav. Trakų pilies fotografijos žymių iliustracija Oilerio diagrama

Žymėjimas yra laikomas neoficialaus naujosios kartos žiniatinklio standarto Web 2.0 dalimi, tačiau jo taikymas neapsiriboja vien žiniatinklio objektų (pagrinde tinklapių ir juose esančių iliustracijų) aprašymu. Pašto sistema Gmail[14] ir grupinio darbo sistema Zimbra Collaboration Suite[15] leidžia laisvai priskiriamomis žymėmis kategorizuoti elektroninius laiškus. Foto archyvo sistema Flickr[16] yra ne tik pirmoji sistema, pradėjusi naudoti žymes fotografijoms aprašyti, bet ir pirmoji sistema, kurioje buvo panaudotas žymių žemėlapis (ang. *tag cloud*) navigacijai atlikti. CiteULike[17] tinklapyje žymės yra naudojamos mokslinio turinio dokumentams aprašyti. Debian GNU/Linux distribucijoje[18] žymės yra sudėtos daugiau nei dešimčiai tūkstančių programinių paketų, taip sukuriant persidengiančias paketų kategorijas.

Pati žymėjimo idėja nėra nauja. Dokumentų saugyklos ir skaitmeninės bibliotekos neretai turi panašių priemonių jų turiniui organizuoti, tačiau tradiciškai tokiu dokumentų klasifikavimu apsiima kvalifikuotas personalas (Rowley, 2000). Žiniatinklyje yra paplitęs dinamiškas, demokratiškas ir kartu anarchiškas klasifikavimo variantas žymėmis – bendradarbiaujantis žymėjimas (angl. *collaborative tagging*). Objektai yra klasifikuojami ne specialiai tuo užsi-

14 Prieiga internetu adresu http://www.gmail.com [žiūrėta 2006-05-02].

15 Prieiga internetu http://www.zimbra.com [žiūrėta 2006-05-02].

16 Prieiga internetu http://www.flickr.com [žiūrėta 2006-05-02].

17 Prieiga internetu http://www.citeulike.org [žiūrėta 2006-05-02].

18 Prieiga internete http://www.debian.org [žiūrėta 2006-05-02].

imančios kvalifikuotų asmenų grupės, o įvairaus kontingento interneto vartotojų, žymes parenkančių pagal individualų objektų suvokimą. Priskirdami savo ir validuodami svetimas žymes, vartotojai bendromis jėgomis sukuria stebėtinai dėsningas ir stabilias (Golder; Huberman, 2006) pasirinktų objektų taksonomijas, dar vadinamas folksonomijomis arba liaudies taksonomijomis (angl. *folk taxonomy*, *folksonomy*).

Šiame darbe yra koncentruojamasi ties individualiu žymėjimu pagrįsta dokumentų klasifikacija, kurios metu vieno dokumento žymėjime dalyvauja vienintelis asmuo – jo savininkas.

5.1.3 PASIRINKTAS DOKUMENTŲ REPREZENTACIJOS BŪDAS

Kaip jau buvo minėta 4.3 skyriuje, klasterizuojant dažniausiai dokumento reprezentacijai yra naudojamas dokumento tekstas. Tokiais atvejais dokumentas prilyginamas n-matės erdvės vektoriui $d = \{w_1, w_2, ..., w_m\}$, kur w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente.

Kolekcijos žodynas yra sudaromas iš dokumentų tekste esančių žodžių. Deja, būna situacijų, kada kompaktiškam dokumentų kolekcijos žodynui sudaryti nėra priemonių. Būtent taip yra lietuvių kalbos atveju (problema detaliai aprašyta 5.1.1 skyriuje).

Dokumentų klasterizavimą galima vykdyti ir su nekompaktišku žodynu, tačiau tokiu atveju tektų susitaikyti su atitinkamai padidėjusiu klasterizavimo proceso imlumu skaičiuojamiesiems resursams, kuris priklausomai nuo dokumentų panašumo nustatymo būdo gali būti labai ženklus. Sintetinių kalbų atveju nekompaktiškas žodynas, turintis daug skirtingų to paties žodžio formų, gali taip pat nulemti iškraipytus arba netikslius klasterizavimo rezultatus[19].

Bandant apeiti nekompaktiško dokumentų kolekcijos žodyno problemą, individualiai klasifikuotų dokumentų klasterizavimo metode dokumentai yra reprezentuojami pasinaudojant jų žymėmis. Kiekvienas dokumentas yra prilyginamas daugiamatės erdvės vektoriui $d = \{t_1, t_2, ..., t_m\}$, kuriame matmenų skaičius m atitinka žymių kiekį dokumentų kolekcijos *žymių* žodyne, o t_j yra j-osios, kolekcijos *žymių* žodyne esančios, žymės svoris dokumente.

19 Elementarus pavyzdys. Bet kuris lietuvių kalbos daiktavardis atsižvelgus į jo linksnius, vienaskaitą ir daugiskaitą gali turėti bent 12 skirtingai atrodančių formų, kurių gali padaugėti dar kelis kartus panaudojus mažybines priesagas. Vargu ar egzistuoja bent vienas klasterizavimo metodas, kuris šias skirtingai atrodančias vieno žodžio formas sugebėtų priskirti tam pačiam skirstiniui.

Nors kiekviena žymė tam pačiam dokumentui gali būti priskiriama tik vieną kartą, dokumentai nėra reprezentuojami binariniais vektoriais. Stengiantis, kad labai dažnai kolekcijoje sutinkamos žymės neužgožtų rečiau pasitaikančių, bet reikšmingų žymių, o labai retos žymės taip pat neįgautų per daug reikšmės, apskaičiuojant vektorių koordinates t_j yra pasinaudojama 4.3 skyriuje aprašytu ir formulėje (2) nurodytu invertuotu pasikartojimo dažniu idf_j.

Skirtingiems dokumentams gali būti priskiriamas nevienodas žymių skaičius, todėl svoriai t_j yra papildomai normalizuojami taip, kad dokumentų vektorių ilgis būtų lygus vienetui. Tam naudojama iš (3) perdaryta formulė:

$$t_j = \frac{a_j \times df_j}{\sqrt{\sum_{r=1}^{m} (a_r \times df_r)^2}} \quad , \tag{12}$$

kur a_j (a_r) nurodo ar j-oji (r-oji), dokumentų kolekcijos *žymių* žodyne esanti, žymė yra priskirta dokumentui.

Pasirinkta dokumentų reprezentacija remiasi perdirbtomis (Chang et al., 2004), (Chang; Hsu, 2005), (Kummamuru et al., 2003) ir (Golder; Huberman, 2006) darbuose esančiomis idėjomis ir jų rezultatais.

(Chang et al., 2004), (Chang; Hsu, 2005), (Kummamuru et al., 2003) darbuose yra nagrinėjama didelių dokumentų kolekcijų klasterizavimo efektyvumo problema, kuri sprendžiama dviem etapais. Pirmo etapo metu į kolekcijos žodyną atrenkami automatiškai sugeneruoti dokumentų turinį apibūdinantys raktažodžiai. Antro etapo metu šie raktažodžiai yra klasterizuojami ir pagal gautus skirstinius atitinkamai organizuojama visa dokumentų kolekcija. Nurodytuose darbuose atliktų eksperimentų rezultatai leidžia daryti išvadas, kad tinkamai parinkus apibūdinančius raktažodžius, stipriai sumažintas kolekcijos žodynas neturi neigiamos įtakos klasterizavimo rezultatams. Tai tai pat leidžia daryti prielaidą, kad panaši situacija turėtų atsikartoti ir dokumentams, turintiems pakankamai priskirtų žymių.

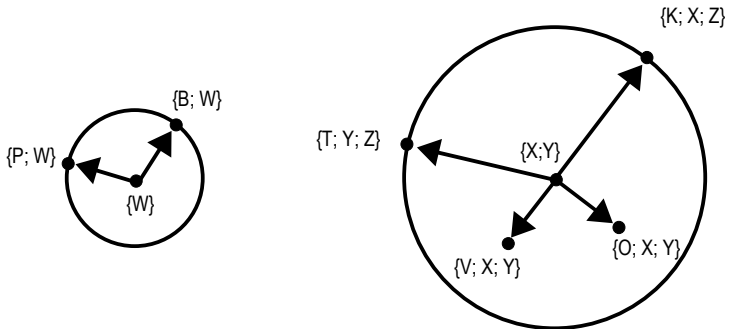
(Golder; Huberman, 2006) darbe yra tyrinėjami bendradarbiaujančio žymėjimo dėsningumai. Jo autoriai pastebi, kad nepaisant to, jog dauguma žymių yra sudedamos remiantis subjektyvia ir individualia vartotojų klasifikacijos schema, dažniausiai yra naudojami bendri terminai. Šis pastebėjimas leidžia daryti prielaidą, kad net lietuviškai (ar kita sintetine kalba) kalbantys vartotojai, žymėms greičiausiai bus linkę naudoti pirmines žodžių formas, o tai leistų išspręsti nekompaktiškų kolekcijos žodynų problemą.

5.2 DOKUMENTŲ PANAŠUMO MATAVIMAS

5.2.1 EUKLIDINIU ATSTUMU PAGRĮSTŲ PANAŠUMO METRIKŲ TRŪKUMAI

Vienas mėgiamiausių ir populiariausių matų dokumentų panašumui nustatyti yra euklidinis atstumas (5), o juo paremtas kvadratinės klaidos kriterijus (9) dokumentų klasterizavimo algoritmuose yra dažnai naudojamas kaip optimizavimo funkcija (žr. 4.5 skyrių). Deja, nei euklidinis atstumas, nei kvadratinės klaidos kriterijus nėra labai tinkami klasterizuoti aibėms, kuriuose objektai yra reprezentuojami kategoriniais atributais (Guha et al., 2000).

Pateiksime pavyzdį dalinančio klasterizavimo atvejui. Tarkime yra individualiai suklasifikuotų dokumentų kolekcija, kurioje kategorinius atributus atitinka individualūs klasifikatoriai – dokumentų žymės. Bendras žymių kiekis kolekcijos žymių žodyne gali būti labai didelis, tačiau atskiriems dokumentams vartotojai paprastai priskiria vos po keletą žymių (naudotose dokumentų kolekcijose – vidutiniškai po 4 žymes, žr. 6.1 skyrių). Kuo mažiau žymių yra priskirta dokumentui, tuo didesnė tikimybė, kad jis turi mažiau į save panašių (tokį pat ar artimą žymių rinkinį turinčių) dokumentų. Galioja ir atvirkštinė taisyklė – kuo daugiau žymių yra priskirta dokumentui, tuo didesnė tikimybė, kad jis turi daugiau į save panašių dokumentų. Kitaip tariant, žymėmis reprezentuojamų dokumentų kolekcija greičiausiai būtų sudaryta iš mažų skirstinių, turinčių dokumentus su trumpais priskirtų žymių sąrašais, ir didesnių skirstinių, turinčių dokumentus su ilgesniais priskirtų žymių sąrašais (žr. 10 pav).



10 pav. Euklidinio atstumo trūkumai žymėmis reprezentuojamoje dokumentų kolekcijoje

Deja, dalinančio klasterizavimo atveju dideliuose skirstiniuose formaliai mažiau centrui panašūs dokumentai euklidiniu atstumu yra įvertinami prasčiau, nei lygiai tiek pat forma-

liai mažiau centrui panašūs dokumentai mažesniuose skirstiniuose ir tai lemia nereikalingus pakankamai kompaktiškų didelių skirstinių suskaldymą. Ši problema vaizdžiai iliustruota 10 paveiksle.

Euklidinis atstumas nelabai tinka ir aglomeratyvaus klasterizavimo atveju. Pateiksime antrą pavyzdį. Tarkime yra keturių dokumentų kolekcija su žymių rinkiniais *{A; B; C; E}*, *{B; C; D; E}*, *{A; D}* ir *{F}*. Paprastumo dėlei dokumentus galima reprezentuoti binariniais vektoriais $d_1 = \{1; 1; 1; 0; 1; 0\}$, $d_2 = \{0; 1; 1; 1; 1; 0\}$, $d_3 = \{1; 0; 0; 1; 0; 0\}$ ir $d_4 = \{0; 0; 0; 0; 0; 1\}$. Panašumą tarp jų matuojant euklidiniu atstumu, atstumas tarp pirmųjų dviejų dokumentų yra $\sqrt{2}$ ir tai yra mažiausias atstumas tarp visų galimų porų. Šiuos dokumentus klasterizavimo metu apjungus į vieną skirstinį, naujo skirstinio centru tampa vektorius *{0,5; 1; 1; 0,5; 1; 0}*. Kitos aglomeratyvaus klasterizavimo iteracijos metu į vieną skirstinį turėtų būti apjungiami trečias ir ketvirtas dokumentai, nes atstumas tarp jų yra $\sqrt{3}$ ir jis mažesnis nei atstumai tarp pirmojo skirstinio centro ir likusių dokumentų – $\sqrt{3,5}$ su trečiuoju dokumentu ir $\sqrt{4,5}$ su ketvirtuoju dokumentu. Deja, tai reiškia, kad būtų bandoma į vieną skirstinį apjungti dokumentus su žymėmis *{A; D}* ir *{F}*, tarp kurių nėra nieko bendra.

5.2.2 PASIRINKTAS DOKUMENTŲ PANAŠUMO MATAVIMO BŪDAS

Dėl aukščiau išvardintų euklidinio atstumo ir juo paremtų optimizuojančių kriterijų trūkumų individualiai klasifikuotų dokumentų metode panašumui yra naudojama kitos dokumentų klasterizavime labai populiarios tarpusavio panašumo nustatymo priemonės – kosinuso koeficientas (8) ir juo paremtas bendrojo glaudumo kriterijus (10).

Pastarojo kriterijaus tikslas yra padidinti skirstinių glaudumą. Jis neturi kvadratinės klaidos kriterijui būdingų trūkumų ir kartu su kosinuso koeficientu yra naudojamas iš dalies šiam darbui gimininguose, bei 5.1.3 skyriuje apžvelgtuose darbuose (Chang et al., 2004), (Chang; Hsu, 2005), (Kummamuru et al., 2003).

5.3 KLASTERIZAVIMO ALGORITMAS

Ieškant tinkamiausio klasterizavimo algoritmo individualiai klasifikuotiems dokumentams buvo atlikta serija eksperimentų. Jų metu keturios 6.1 skyriuje aprašytos dokumentų kolekcijos trimis skirtingais klasterizavimo algoritmais buvo suklasterizuotos į 5, 10, 15 ir 20 skirstinių.

Naudoti šie algoritmai: aglomeratyvus hierarchinis, K-means ir skaldantis K-means. Kolekcijas klasterizuojant K-means ir skaldančio K-means algoritmais, lokalaus optimumo problema (Tan et al., 2005) spręsta pasinaudojant dažniausiai praktikoje taikomu būdu – kartojimu. Kadangi abu algoritmai nėra reiklūs skaičiuojamiesiems resursams, kiekvieno klasterizavimo atveju jie kolekcijoms buvo taikomi po tūkstantį kartų. Vėliau iš gautų rezultatų buvo atrenkami geriausi klasterizavimo variantai.

Gautų skirstinių kokybė buvo matuojama dvejomis metrikomis. Pirmoji metrika – entropija – vertino skirtingų klasių dokumentų išsisklaidymo laipsnį vieno skirstinio ribose. Antroji metrika – grynumas – vertino vienos klasės dokumentų susitelkimo laipsnį vieno skirstinio ribose. Detalus metrikų aprašymas ir naudotos formulės pateiktos 6.2 skyriuje.

Klasterizuojant individualiai klasifikuotų dokumentų kolekcijas prasčiausiai pasirodė aglomeratyvus hierarchinis, o geriausiai – skaldantis K-means algoritmas, nuo kurio itin nedaug atsiliko tradicinis K-means algoritmas. Apibendrinti entropijos ir grynumo įvertinimai pateikti šio skyriaus 2 lentelėje. Detalūs eksperimentų rezultatai pateikti 6.3 skyriuje, o eksperimentų žurnalai – prieduose A, B ir C.

2 lentelė. Galutiniai entropijos ir grynumo vidurkiai

Algoritmas	**Entropija**	**Grynumas**
Aglomeratyvus	0,598	0,639
K-means	0,391	0,778
Skald. K-means	0,391	0,782

6 EKSPERIMENTAI

6.1 EKSPERIMENTINIAI DUOMENYS

Eksperimentų metu buvo naudojamos keturios skirtingo dydžio dokumentų kolekcijos, suformuotos iš 13769 Lietuvos naujienų agentūros ELTA[20] straipsnių archyvo. Kolekcijų charakteristikų santrauka yra pateikiama 3 lentelėje nurodant dokumentų kiekį, unikalių žodžių ir žymių (atlikusių individualių kategorijų vaidmenį) skaičių bei kolekcijas sudariusių klasių kiekį. Žemiau kolekcijos vadinamos skaitiniais vardais, atitinkančiais dokumentų kiekį kolekcijoje.

20 Naujienų agentūros tinklapio prieiga internetu: http://www.elta.lt [žiūrėta 2006-05-10].

3 lentelė. Eksperimentuose naudotų dokumentų kolekcijų charakteristikų santrauka

Dokumentų kiekis	Žodžių kiekis	Žymių kiekis	Klasių kiekis
1056	36843	1379	5
2112	55352	1882	5
4224	81293	2449	5
8448	116755	2754	5

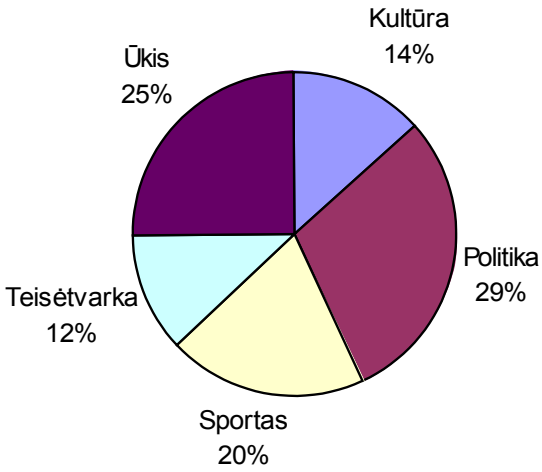
Tiek unikalių žodžių, tiek unikalių žymių skaičius nustatytas neatsižvelgiant į morfologines žodžių ypatybes t.y. skirtingos žodžio formos buvo laikomos savarankiškais dariniais.

Dokumentų klasės gautos sujungus[21] dešimt oficialių ELTA agentūros kategorijų. Eksperimentinėse kolekcijose naudotos klasės ir jų pasiskirstymas pateiktas 4 lentelėje. Skirtingose kolekcijose šis pasiskirstymas apylygis (žr. 11 pav.).

4 lentelė. Klasių pasiskirstymas dokumentų kolekcijose

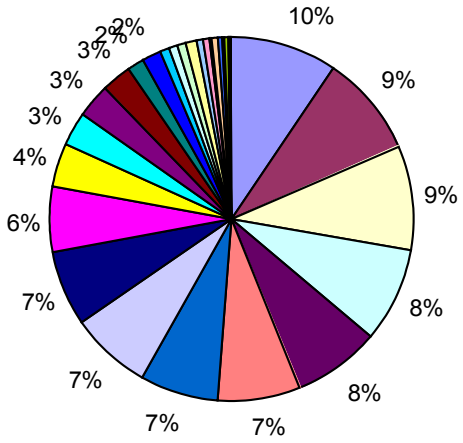
	1056	**2112**	**4224**	**8448**
Kultūra	152	286	574	1145
Politika	311	594	1240	2498
Sportas	199	457	841	1648
Teisėtvarka	122	246	495	1023
Ūkis	272	529	1074	2134

21 ELTA kategorijos: užsienio kultūra, Lietuvos kultūra, užsienio sportas, Lietuvos sportas, užsienio politika, Lietuvos politika, užsienio ūkis, Lietuvos ūkis, užsienio teisėtvarka, Lietuvos teisėtvarka apjungtos pastebėjus, kad daug archyvo straipsnių tuo pačiu metu priklauso dviem kategorijoms – ir užsienio, ir Lietuvos.



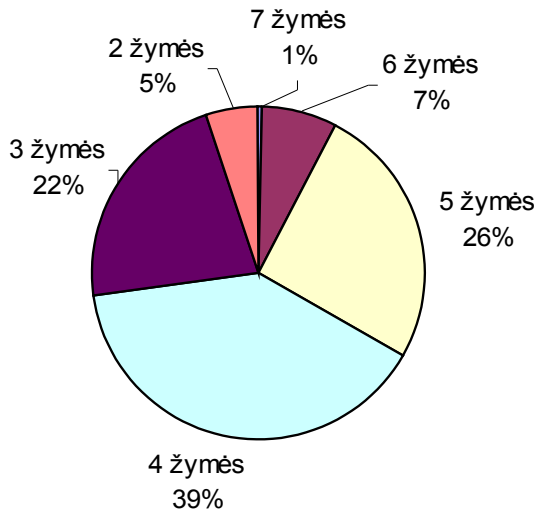
11 pav. Vidutinis klasių pasiskirstymas dokumentų kolekcijose

Visi archyvo straipsniai parašyti, priskirti oficialioms kategorijoms ir individualiai sužymėti trisdešimt dviejų ELTA reporterių pastangomis. Šešiolika iš jų yra suklasifikavę daugiau nei 90% į kolekcijas patekusių dokumentų (žr. 12 pav.). Atrinktose kolekcijose vidutiniškai vienam dokumentui autoriai būdavo priskyrę po 4 žymes (žr. 13 pav).



12 pav. Dokumentų pasiskirstymas pagal juos suklasifikavusius autorius

Dokumentų kolekcijos suformuotos atsitiktine tvarka į juos atrenkant ELTA archyvo straipsnius. Į kolekcijas nebuvo įtraukiami dokumentai, neturintys žymių arba autorių priskirti daugiau nei vienai klasei. Sudarant žymių sąrašą į jį buvo atrenkamos žymės, ne trumpesnės nei 2 simboliai ir priskirtos bent dviems dokumentams vienu metu.



13 pav. Vienam dokumentui priskirtų žymių kiekių pasiskirstymas

6.2 EKSPERIMENTŲ SĄLYGOS IR METRIKOS

Eksperimentų metu kiekviena 6.1 skyriuje aprašyta dokumentų kolekcija naudojant tris skirtingus algoritmus ir du dokumentų reprezentacijos būdus buvo klasterizuojama į 5, 10, 15 ir 20 skirstinių.

Kolekcijas klasterizuojant K-means ir skaldančio K-means algoritmais, lokalaus optimumo problema (Tan et al., 2005) spręsta pasinaudojant dažniausiai praktikoje taikomu būdu – kartojimu. Kadangi abu algoritmai nėra reiklūs skaičiuojamiesiems resursams, kiekvieno klasterizavimo atveju jie kolekcijoms buvo taikomi po tūkstantį kartų. Vėliau iš gautų rezultatų buvo atrenkami geriausi klasterizavimo variantai.

Gautų skirstinių kokybė buvo matuojama dvejomis metrikomis, kurių kiekviena pagrįsta 6.1 skyriuje aprašytomis klasėmis. Pirmoji metrika – entropija – vertino skirtingų klasių dokumentų išsisklaidymo laipsnį vieno skirstinio ribose. Antroji metrika – grynumas – vertino vienos klasės dokumentų susitelkimo laipsnį vieno skirstinio ribose.

Turint skirstinį C_r, sudarytą iš n_r dokumentų, jo entropija apskaičiuota naudojantis formule

$$E(C_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_{ri}}{n_r}\log\frac{n_{ri}}{n_r} \quad , \tag{13}$$

kur q yra klasių kiekis dokumentų kolekcijoje, o n_{ri} – i-osios klasės dokumentų, esančių skirstinyje C_r, skaičius. Bendra į k skirstinių suklasterizuotos n dokumentų kolekcijos entropija apskaičiuota susumuojant jos skirstinių entropijos vertes, prieš tai padaugintas iš svorio koeficientų, atitinkančių skirstinių dydį:

$$Entropija = \sum_{r=1}^{k}\frac{n_r}{n}E(C_r) \quad . \tag{14}$$

Mažesnė skirstinių entropija reiškia geresnius klasterizavimo rezultatus. Idealaus klasterizavimo atveju, kiekvienas skirstinys turėtų turėti tik vienos klasės dokumentų t.y. entropijos dydis turėtų būti nulinis.

Turint skirstinį C_r, sudarytą iš n_r dokumentų, jo grynumas apskaičiuotas naudojantis formule

$$G(C_r) = \frac{1}{n_r}\max(n_{ri}) \quad , \tag{15}$$

kur q yra klasių kiekis dokumentų kolekcijoje, o n_{ri} – i-osios klasės dokumentų, esančių skirstinyje C_r, skaičius. Praktiškai skirstinio grynumas yra ne kas kita, kaip didžiausios dokumentų klasės, esančios skirstinyje, dalis tarp visų skirstinio dokumentų.

Bendras į k skirstinių suklasterizuotos n dokumentų kolekcijos grynumas apskaičiuotas susumuojant jos skirstinių grynumo vertes, prieš tai padaugintas iš svorio koeficientų, atitinkančių skirstinių dydį:

$$Grynumas = \sum_{r=1}^{k}\frac{n_r}{n}G(C_r) \quad . \tag{16}$$

Didesnis skirstinių grynumas reiškia geresnius klasterizavimo rezultatus. Idealaus klasterizavimo atveju, kiekvienas skirstinys turėtų turėti tik vienos klasės dokumentų t.y. grynumo dydis turėtų būti lygus vienetui.

Specialiai eksperimentams susikurtomis programinėmis priemonėmis apdorotas 6.1 skyriuje aprašytas ELTA archyvas buvo klasterizuojamas šio dokumento autoriaus papildomai modifikuotais CLUTO[22] rinkinio įrankiais.

6.3 TINKAMIAUSIO KLASTERIZAVIMO ALGORITMO ATRANKA

Ieškant tinkamiausio klasterizavimo algoritmo individualiai klasifikuotiems dokumentams, buvo atlikta serija eksperimentų su keturiomis 6.1 skyriuje aprašytomis dokumentų kolekcijomis. Šios kolekcijos aglomeratyvaus hierarchinio, K-means ir skaldančio K-means klasterizavimo algoritmais 6.2 skyriuje nurodytomis sąlygomis buvo suklasterizuotos į 5, 10, 15 ir 20 skirstinių.

Vykdant eksperimentus dokumentai buvo reprezentuojami 5.1.3 skyriuje aprašytu būdu, o jų tarpusavio panašumas nustatomas 5.2.2 skyriuje nurodytais kosinuso koeficientu ir bendrojo glaudumo kriterijumi. Gautų skirstinių kokybė buvo matuojama dvejomis metrikomis – entropija ir grynumu. Metrikos detalizuotos 6.2 skyriuje.

Tinkamiausio algoritmo individualiai klasifikuotiems dokumentams klasterizuoti atrankos eksperimentų žurnalai yra pateikti šio dokumento prieduose A, B ir C.

5 lentelė. 1056 dokumentų kolekcijos klasterizavimo rezultatų įvertinimai

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Aglomeratyvus	0,693	0,619	0,522	0,479	0,564	0,566	0,680	0,720
K-means	**0,413**	**0,414**	0,374	**0,308**	**0,746**	**0,756**	0,780	**0,830**
Skald. K-means	0,485	0,421	**0,361**	0,320	0,691	0,740	**0,806**	0,818

6 lentelė. 2112 dokumentų kolekcijos klasterizavimo rezultatų įvertinimai

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Aglomeratyvus	0,732	0,638	0,621	0,571	0,561	0,620	0,620	0,666
K-means	**0,461**	0,442	0,399	**0,327**	**0,736**	0,721	0,777	**0,832**
Skald. K-means	0,509	**0,436**	**0,385**	0,331	0,710	**0,738**	**0,786**	**0,832**

22 Prieiga internetu http://glaros.dtc.umn.edu/gkhome/views/cluto [žiūrėta 2006-05-12].

7 lentelė. 4224 dokumentų kolekcijos klasterizavimo rezultatų įvertinimai

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Aglomeratyvus	0,661	0,561	0,541	0,500	0,611	0,689	0,689	0,737
K-means	0,445	0,414	0,382	**0,337**	0,752	**0,759**	0,786	0,829
Skald. K-means	**0,404**	**0,412**	**0,375**	0,339	**0,767**	0,755	**0,798**	**0,832**

8 lentelė. 8448 dokumentų kolekcijos klasterizavimo rezultatų įvertinimai

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Aglomeratyvus	0,703	0,613	0,568	0,541	0,554	0,633	0,643	0,664
K-means	0,452	0,398	0,358	0,338	0,736	0,781	0,804	0,819
Skald. K-means	**0,426**	**0,389**	**0,340**	**0,316**	**0,760**	**0,784**	**0,840**	**0,850**

5–8 lentelėse yra pateikti keturių kolekcijų klasterizavimo metu gautų skirstinių kokybės įvertinimai pagal (14) ir (16) formules (žr. 44 psl.). Šiose lentelėse išskirtos reikšmės nurodo geriausius įverčius skirtingais algoritmais klasterizuojant nurodytą kolekciją į k skirstinių.

Lentelėse esantys rezultatai rodo, kad klasterizuojant individualiai klasifikuotų dokumentų kolekcijas aglomeratyviu hierarchiniu algoritmu, gautų skirstinių kokybė akivaizdžiai nusileidžia skirstiniams, gautiems naudojant kitus du algoritmus. Sprendžiant pagal vidutinius kokybės įvertinimus 9 lentelėje ir santykinius palyginimus su geriausiais vidutiniais rezultatais 10 lentelėje, aglomeratyvus algoritmas sukuria maždaug 46–60% didesnę entropiją ir 16–23% mažesnį grynumą turinčius skirstinius.

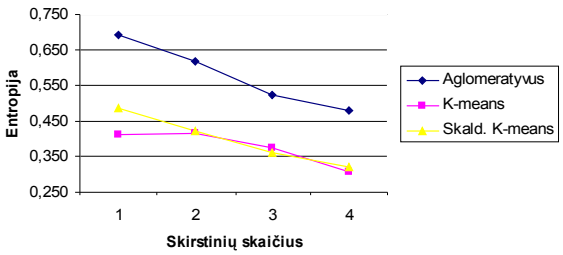
9 lentelė. Klasterizavimo rezultatų įvertinimų vidurkiai

	Entropija					**Grynumas**				
	5	**10**	**15**	**20**	**Vid.**	**5**	**10**	**15**	**20**	**Vid.**
Aglomeratyvus	0,697	0,608	0,563	0,523	0,598	0,573	0,627	0,658	0,697	0,639
K-means	**0,443**	0,417	0,378	0,328	**0,391**	**0,743**	**0,754**	0,787	0,828	0,778
Skald. K-means	0,456	**0,415**	**0,365**	**0,327**	**0,391**	0,732	**0,754**	**0,808**	**0,833**	**0,782**

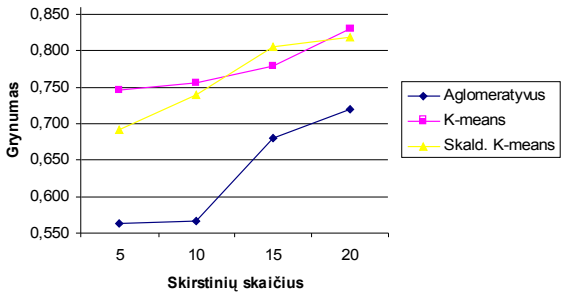
10 lentelė. Santykiniai klasterizavimo rezultatų palyginimai su geriausiais vidutiniais rezultatais

	Entropija				Grynumas			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Aglomeratyvus	57,34%	46,51%	54,25%	59,94%	22,88%	16,84%	18,56%	16,33%
K-means	0,00%	0,48%	3,56%	0,31%	0,00%	0,00%	2,60%	0,60%
Skald. K-means	2,93%	0,00%	0,00%	0,00%	1,48%	0,00%	0,00%	0,00%

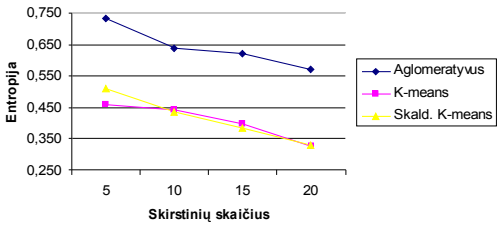
Kitų dviejų algoritmų – K-means ir skaldančio K-means – rezultatai yra gerokai artimesni. Tai neblogai matosi 14–21 paveiksluose esančiuose grafikuose ir ypač gerai – 22, 23 paveiksluose esančiuose grafikuose.



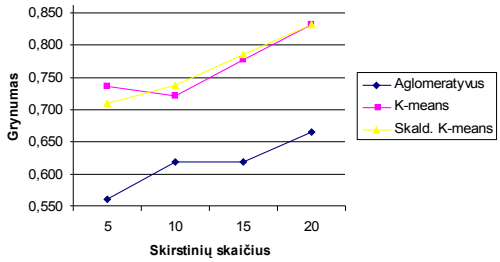
14 pav. 1056 dokumentų kolekcijos klasterizavimo entropijos įvertinimai



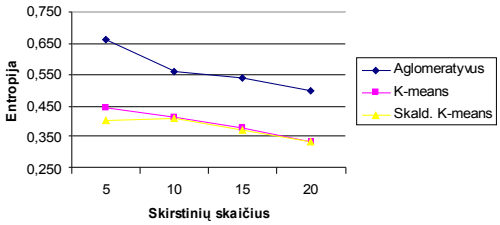
15 pav. 1056 dokumentų kolekcijos klasterizavimo grynumo įvertinimai



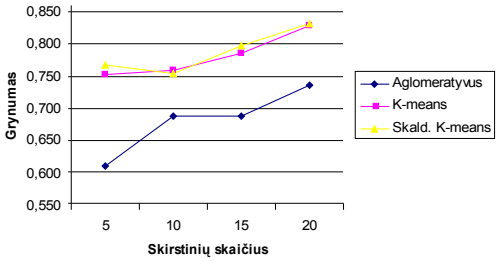
16 pav. 2112 dokumentų kolekcijos klasterizavimo entropijos įvertinimai



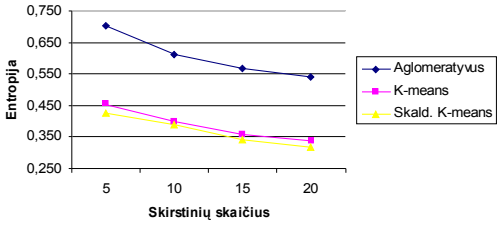
17 pav. 2112 dokumentų kolekcijos klasterizavimo grynumo įvertinimai



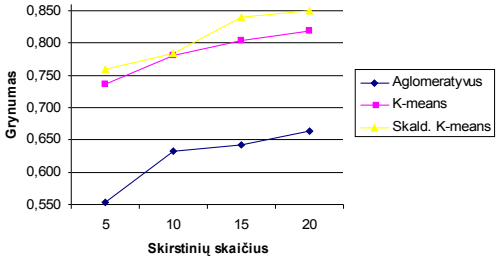
18 pav. 4224 dokumentų kolekcijos klasterizavimo entropijos įvertinimai



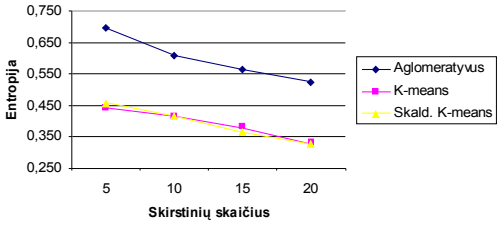
19 pav. 4224 dokumentų kolekcijos klasterizavimo grynumo įvertinimai



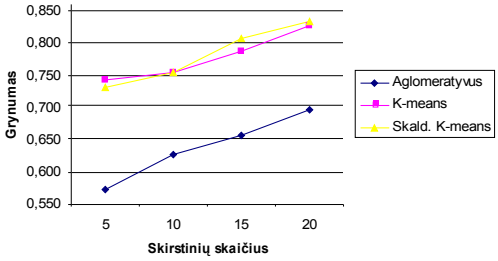
20 pav. 8448 dokumentų kolekcijos klasterizavimo entropijos įvertinimai



21 pav. 8448 dokumentų kolekcijos klasterizavimo grynumo įvertinimai



22 pav. Klasterizavimo entropijos vidurkiai



23 pav. Klasterizavimo grynumo vidurkiai

Santykiniai palyginimai su geriausiais vidutiniais rezultatais 10 lentelėje rodo, kad skaldančio K-means ir tradicinio K-means klasterizavimo rezultatų kokybė tarpusavyje skiriasi vos 0,3–3,6% entropijos ir 0,6–2,6% grynumo. Kita vertus, vidutiniai kokybės įvertinimai 9 lentelėje rodo, kad skaldantis K-means algoritmas vidutiniškai sukuria kokybiškesnius skirstinius, o 5–8 lentelėse esančiuose rezultatuose matosi, kad skaldantis K-means algoritmas kokybiškesnius rezultatus statistiškai pateikia beveik dvigubai dažniau nei tradicinis K-means algoritmas (20:11, 10:6 entropijos atveju ir 10:5 grynumo atveju). Dėl šių priežasčių tinkamesniu algoritmu individualiai klasifikuotiems dokumentams klasterizuoti pripažintas skaldantis K-means.

6.4 METODO PALYGINIMAS SU TRADICINIU KLASTERIZAVIMO METODU

Atlikus seriją eksperimentų, šiame skyriuje vienas tradicinis dokumentų klasterizavimo metodas yra palyginamas su 5 skyriuje apibrėžtu individualiai klasifikuotų dokumentų klasterizavimo metodu. Eksperimentai buvo atlikti su keturiomis 6.1 skyriuje aprašytomis dokumentų kolekcijomis, klasterizuojant jas į 5, 10, 15 ir 20 skirstinių. Eksperimentų tikslas – įvertinti naujojo klasterizavimo metodo galimybes lyginant jas su tradicinio klasterizavimo metodo, naudojančio nekompaktišką kolekcijos žodyną (žr. 5.1.1 skyrių), galimybėmis.

Metodo, kuris šiame skyriuje yra vadinamas tradiciniu, pasirinkimą įtakojo (Steinbach et al., 2000) darbe pateikti populiariausių dokumentų klasterizavimo metodų lyginamųjų eksperimentų rezultatai. Šiame metode dokumentai yra reprezentuojami žodžių pasikartojimo dažnių vektoriais, aprašytais 4.3 skyriuje, dokumentų tarpusavio panašumas matuojamas 4.4 skyriuje, (8) formulėje nurodytu kosinuso koeficientu, dokumentai klasterizuojami skaldančio K-means algoritmu, aprašytu 4.5.1 skyriuje, optimizuojant tame pačiame skyriuje aprašytą, (10) formulėje nurodytą bendrojo glaudumo kriterijų. Klasterizuojant tradiciniu metodu buvo naudojami nekompaktiški (žr. 5.1.1 skyrių) kolekcijų žodynai.

Naujojo ir tradicinio metodais gautų skirstinių kokybė buvo matuojama dvejomis metrikomis – entropija ir grynumu. Abi metrikos detalizuotos 6.2 skyriuje. Lyginamųjų eksperimentų žurnalai yra pateikti šio dokumento prieduose C ir D.

11 lentelė. Metodų rezultatų įvertinimas klasterizuojant 1056 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	0,485	0,421	0,361	0,320	0,691	0,740	**0,806**	0,818
Tradicinis	**0,348**	**0,301**	**0,355**	**0,306**	**0,811**	**0,842**	0,792	**0,834**

12 lentelė. Metodų rezultatų įvertinimas klasterizuojant 2112 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	0,509	0,436	0,385	0,331	0,710	0,738	0,786	0,832
Tradicinis	**0,355**	**0,339**	**0,308**	**0,271**	**0,774**	**0,813**	**0,825**	**0,853**

13 lentelė. Metodų rezultatų įvertinimas klasterizuojant 4224 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	0,404	0,412	0,375	0,339	**0,767**	0,755	0,798	**0,832**
Tradicinis	**0,395**	**0,359**	**0,305**	**0,297**	0,738	**0,786**	**0,836**	0,824

14 lentelė. Metodų rezultatų įvertinimas klasterizuojant 8448 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	0,426	0,389	0,340	0,316	**0,760**	**0,784**	**0,840**	**0,850**
Tradicinis	**0,417**	**0,358**	**0,301**	**0,291**	0,723	0,783	0,831	0,833

11–14 lentelėse pateikti keturių kolekcijų klasterizavimo metu gautų skirstinių įvertinimai pagal (14) ir (16) formules (entropija ir grynumas, žr. 44 psl.). Šiose lentelėse išskirtos reikšmės nurodo geriausius įverčius skirtingais algoritmais klasterizuojant nurodytą kolekciją į k skirstinių.

Lentelėse esantys rezultatai rodo, kad nors statistiškai dažniau (25:7, 16:0 entropijos atveju ir 9:7 grynumo atveju) kokybiškesni skirstiniai yra gaunami klasterizuojant tradiciniu metodu, didėjant dokumentų ir skirstinių skaičiui naujasis metodas ima vytis ir lenkti tradicinį klasterizavimo metodą.

15 lentelė. Santykiniai metodų rezultatų palyginimai klasterizuojant 1056 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	39,37%	39,87%	1,69%	4,58%	14,80%	12,11%	0,00%	1,92%
Tradicinis	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,74%	0,00%

16 lentelė. Santykiniai metodų rezultatų palyginimai klasterizuojant 2112 dokumentų kolekciją

	Entropija				**Grynumas**			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	43,38%	28,61%	25,00%	22,14%	8,27%	9,23%	4,73%	2,46%
Tradicinis	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

17 lentelė. Santykiniai metodų rezultatų palyginimai klasterizuojant 4224 dokumentų kolekciją

	Entropija				Grynumas			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	2,28%	14,76%	22,95%	14,14%	0,00%	3,94%	4,55%	0,00%
Tradicinis	0,00%	0,00%	0,00%	0,00%	3,78%	0,00%	0,00%	0,96%

18 lentelė. Santykiniai metodų rezultatų palyginimai klasterizuojant 8448 dokumentų kolekciją

	Entropija				Grynumas			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	2,16%	8,66%	12,96%	8,59%	0,00%	0,00%	0,00%	0,00%
Tradicinis	0,00%	0,00%	0,00%	0,00%	4,87%	0,13%	1,07%	2,00%

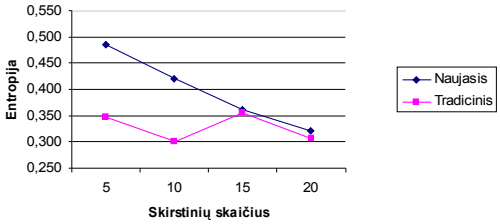
15–18 lentelėse gerai matosi, kad individualiai klasifikuotų dokumentų klasterizavimo metodas tradiciniam metodui pagrinde nusileidžia pagal skirstinių entropijos įvertinimą. Deja, skirtingų klasių dokumentų išsisklaidymo laipsnis vieno skirstinio ribose naujajame metode yra didesnis nei tradicinio metodo sukurtuose skirstiniuose ir tai greičiausiai lemia tai, kad dalis dokumentų turi mažą (žr. 6.1 skyrių) priskirtų žymių kiekį. Kita vertus, skirstinių grynumo atžvilgiu naujasis metodas tradiciniam metodui nusileidžia labai nedaug, o didėjant dokumentų ir skirstinių skaičiui jį ima lenkti.

19 lentelė. Metodų rezultatų įvertinimo vidurkiai

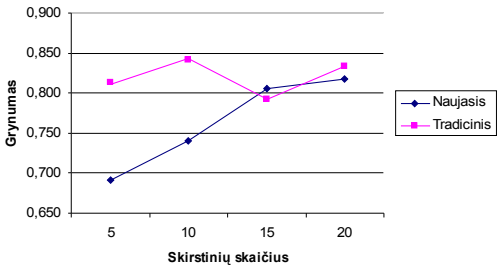
	Entropija					Grynumas				
	5	**10**	**15**	**20**	**Vid.**	**5**	**10**	**15**	**20**	**Vid.**
Naujasis	0,456	0,415	0,365	0,327	0,391	0,732	0,754	0,808	0,833	0,782
Tradicinis	**0,379**	**0,339**	**0,317**	**0,291**	**0,332**	**0,762**	**0,806**	**0,821**	**0,836**	**0,806**

20 lentelė. Santykiniai metodų rezultatų palyginimai pagal vidutines įvertinimo reikšmes

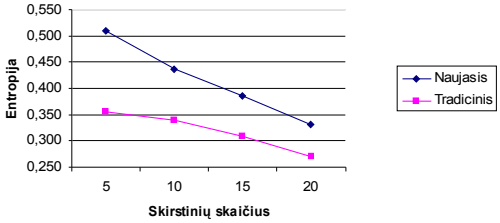
	Entropija				Grynumas			
	5	**10**	**15**	**20**	**5**	**10**	**15**	**20**
Naujasis	20,40%	22,18%	15,13%	12,10%	3,87%	6,42%	1,64%	0,36%
Tradicinis	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%



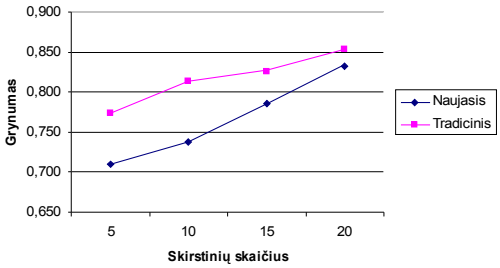
24 pav. Metodų entropijos įvertinimas klasterizuojant 1056 dokumentų kolekciją



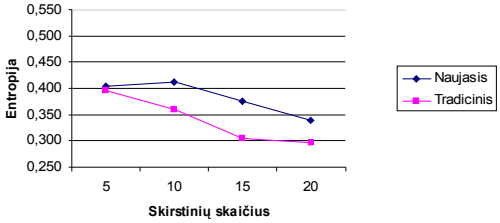
25 pav. Metodų grynumo įvertinimas klasterizuojant 1056 dokumentų kolekciją



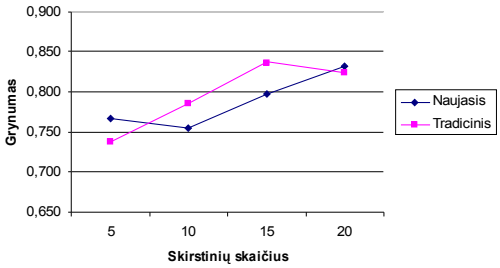
26 pav. Metodų entropijos įvertinimas klasterizuojant 2112 dokumentų kolekciją



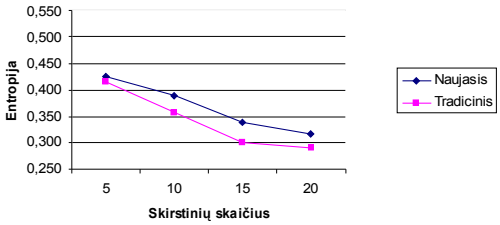
27 pav. Metodų grynumo įvertinimas klasterizuojant 2112 dokumentų kolekciją



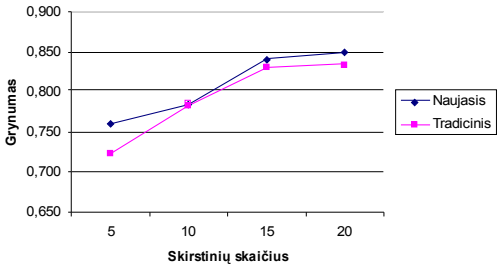
28 pav. Metodų entropijos įvertinimas klasterizuojant 4224 dokumentų kolekciją



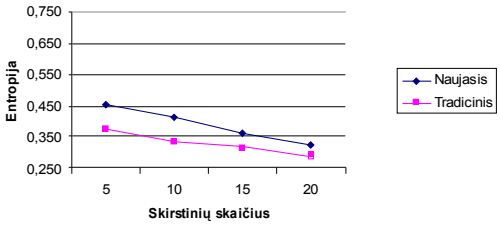
29 pav. Metodų grynumo įvertinimas klasterizuojant 4224 dokumentų kolekciją



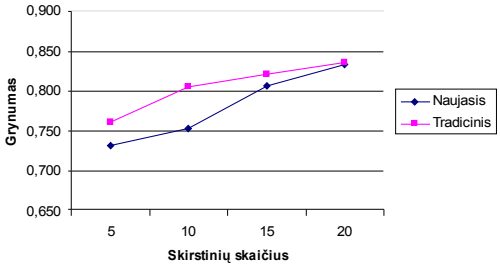
30 pav. Metodų entropijos įvertinimas klasterizuojant 8448 dokumentų kolekciją



31 pav. Metodų grynumo įvertinimas klasterizuojant 8448 dokumentų kolekciją



32 pav. Metodų klasterizavimo entropijos vidurkiai



33 pav. Metodų klasterizavimo grynumo vidurkiai

24–23 paveiksluose esančiose diagramose gerai matosi tendencijos, ne taip pastebimos šiame skyriuje esančiose lentelėse. Diagramos rodo, kad didėjant skirstinių skaičiui abu metodai gerina klasterizavimo metu gaunamų skirstinių kokybę tiek entropijos, tiek grynumo atžvilgiu, tačiau naujasis metodas tai atlieka ženkliai sparčiau. Tai leidžia daryti prielaidą, kad esant pakankamai dideliam skirstinių skaičiui, naujasis metodas abiem kokybės aspektais aplenktų tradicinį klasterizavimo metodą, naudojantį nekompaktišką kolekcijos žodyną. Šią prielaidą papildomai sutvirtina 8448 dokumentus turinčios kolekcijos klasterizavimo rezultatai.

7 REZULTATŲ APIBENDRINIMAS IR DARBAI ATEIČIAI

Tradiciniai klasterizavimo metodai nelabai tinka lietuviškų dokumentų klasterizavimui. Šiuose metoduose dokumentai paprastai yra reprezentuojami žodžių dažnumo vektoriais $d = \{w_1, w_2, ..., w_n\}$, kur w_j yra j-ojo, dokumentų kolekcijos žodyne esančio, žodžio svoris dokumente. Kolekcijos žodynas yra sudaromas iš dokumentų tekste esančių žodžių, atmetant nereikšmingus (įvardžius, prielinksnius, jungtukus ir pan.) bei semantiškai tą pačią prasmę turinčius žodžius (sinonimus, žodžių formas). Deja, šiuo metu neegzistuoja nė vienos autoriui žinomos ir laisvai prieinamos priemonės: sinonimų žodyno, morfologinio analizatoriaus arba kamieno atskyrimo programos kompaktiškiems lietuviškų dokumentų kolekcijų žodynams sudaryti (žr. 5.1.1 skyrių).

Dokumentų klasterizavimą galima vykdyti ir su nekompaktišku žodynu, tačiau tokiu atveju tektų susitaikyti su ženkliai padidėjusiu klasterizavimo proceso imlumu skaičiuojamiesiems resursams. Lietuvių arba kitos sintetinės kalbos atveju nekompaktiškas žodynas, turintis daug skirtingų to paties žodžio formų, gali taip pat nulemti iškraipytus arba netikslius klasterizavimo rezultatus.

Šiame darbe buvo sukurtas ir aprašytas metodas individualiai klasifikuotiems dokumentams klasterizuoti. Sukurtasis metodas skirtas tiems atvejams, kai tradicinių metodų pilnavertiškai taikyti neįmanoma. Pavyzdžiui, naujasis metodas gali būti taikomas skyriaus pradžioje aprašytu atveju, kai nėra priemonių kompaktiškam kolekcijos žodynui sudaryti, bet dokumentai yra aprašyti raktažodžiais arba turi jiems priskirtas individualias kategorijas – žymes (angl. *tags*). Metodas taip pat gali būti taikomas klasterizuojant binarinius objektus, pavyzdžiui, nuotraukas foto archyve, jeigu pastarieji yra aprašyti papildomais metaduomenimis.

Siūlomame klasterizavimo metode dokumentai yra reprezentuojami jiems vartotojų individualiai priskirtomis kategorijomis – žymėmis. Kiekvienas dokumentas yra prilyginamas daugiamatės erdvės vektoriui $d = \{t_1, t_2, ..., t_m\}$, kuriame matmenų skaičius m atitinka žymių kiekį dokumentų kolekcijos *žymių* žodyne, o t_j yra j-osios, kolekcijos *žymių* žodyne esančios, žymės svoris dokumente. Kadangi praktiškai žymės yra raktažodžiai, atliekantys tematikos arba kategorijos vaidmenį (Golder; Huberman, 2006), žymes šiame metode galima pakeisti ir dokumentus apibūdinančiais raktažodžiais.

Siūlomame klasterizavimo metode žymių vektoriais reprezentuojamų dokumentų panašumas yra nustatomas kosinuso koeficientu, o dokumentų kolekcijos klasterizavimas vykdomas eksperimentuojant atrinktu skaldančio K-means (angl. *bisecting K-means*) algoritmu.

Šiame darbe naujojo metodo galimybės buvo eksperimentais palygintos su tradicinio dokumentų klasterizavimo metodo, naudojančio nekompaktišką kolekcijos žodyną, galimybėmis. Atliktų eksperimentų rezultatai parodė (žr. 6.4 skyrių), kad didėjant dokumentų kiekiui ir/arba skirstinių skaičiui, naujasis metodas tiksliau nei tradicinis suformuoja kolekcijos skirstinius ir tai leidžia jį drąsiai naudoti didelėse dokumentų kolekcijose. Deja, esant mažam skirstinių kiekiui, klasterizavimo rezultatus neigiamai veikia santykinai nedidelis dokumentus reprezentuojančių savybių – žymių – skaičius, todėl šiais atvejais naujasis metodas nusileidžia tradiciniam klasterizavimui vidutiniškai penktadaliu didesne entropija.

Gali būti, kad pastarąjį metodo trūkumą įmanoma ištaisyti pakeitus dokumentų panašumui nustatyti naudojamą kosinuso koeficientą priemone, aprašyta (Guha et al., 2000). Pastarajame darbe kategorinius atributus turinčių objektų panašumo lygis nustatomas atsižvelgiant į bendrų kaimynų kiekį, objektus kaimynais laikant tuo atveju, jeigu jų panašumas pagal įprastines metrikas viršija eksperimentų metu parinktą slenkstį.

Guha ir jo kolegų aprašytos metrikos privalumas yra tame, kad nustatant panašumą, atsižvelgiama ne tik į tiesiogiai du objektus vienijančias savybes, bet ir jų kontekstą. Kita ver-

tus, ši metrika buvo taikoma binariniais vektoriais reprezentuojamiems objektams, turintiems fiksuotą kategorinių atributų sąrašą, todėl metrikos pritaikymas kintančio ilgio sąrašo atvejui reikalauja papildomos analizės ir eksperimentų.

Individualiai klasifikuotų dokumentų klasterizavimo metode liko nepaliestas skirstinių pateikimo klausimas. Labiausiai tai įtakojo psichologiniai šios problemos aspektai. Klausimas galėtų būti išspręstas pasinaudojant 6.4 skyriuje bei prieduose C ir D aprašytų eksperimentų rezultatais, tačiau tai pareikalautų atskiro, psichologinėmis metrikomis pagrįsto tyrimo. Šiame darbe eksperimentuose naudotos klasterizavimo metodų tikslumo įvertinimo metrikos – entropija ir grynumas – tam netinka. Jos gali statistiškai įvertinti skirstiniuose esančių dokumentų klasių fragmentacijos ir triukšmo lygį, tačiau negali pasakyti ar suformuoti skirstiniai yra prasmingi.

Prasmingumo matavimus atliekant subjektyviai ir panagrinėjus 6.4 skyriuje aprašytų eksperimentų žurnalus C, D prieduose, galima pastebėti, kad naujuoju metodu suformuotuose skirstiniuose dokumentus siejantys ir iš kitų dokumentų masės išskiriantys žodžiai geriau apibūdina juose esančių dokumentų turinį, nei tai tai daro tradiciniu metodu suformuoti skirstiniai[23]. Šis subjektyvus pastebėjimas leidžia daryti prielaidą, kad net tais atvejais, kai naujasis metodas statistiniais tikslumo įverčiais nusileidžia tradiciniam metodui, jis vis tiek suformuoja prasmingesnius skirstinius.

8 IŠVADOS

- Dokumentų valdymo sistemose centralizuotai iš anksto sukurtą klasifikatorių vartotojai dažnai laiko neadekvačiu. Jame esančios kategorijos paprastai per ne lyg bendros, kad pakankamai tiksliai apibrėžtų joms priskirtų dokumentų turinį, be to, jame neretai iš vis nėra tinkamos kategorijos. Centralizuotą klasifikatorių galima pakeisti kiekvieno vartotojo individualiai susikuriamais klasifikatoriais, tačiau nenaudojant centralizuotos kategorijų sistemos būtų prarandamas visą dokumentų saugyklą vienijantis elementas ir tai apsunkintų reikalingos informacijos paiešką.

23 Puikus pavyzdys – 1056 dokumentų kolekcijos klasterizavimo į penkis skirstinius eksperimentas (žr. pirmuosius priedo C ir D puslapius). Geriausias naujuoju metodu gautas skirstinys entropija ir grynumu ženkliai nusileidžia geriausiam tradiciniu metodu gautam skirstiniui (0,225:0,022 ir 0,915:0,994 atitinkamai). Kita vertus, pirmąjį skirstinį apibūdinantys ir išskiriantys žodžiai: „krepšinis“, „čempionatas“, „lietuviai“ ir „įvairybės“ apie jame esančių dokumentų tematiką pasako gerokai daugiau, nei antrąjį skirstinį apibūdinantys ir išskiriantys žodžiai: „taškų“, „pelnė“, „min“ ir „taškus“.

- Sprendžiant vientiso klasifikatoriaus neturinčios dokumentų saugyklos problemą, galima panaudoti dokumentų klasterizavimą. Deja, tradicinių metodų panaudojimas lietuviškiems dokumentams klasterizuoti yra keblus, nes trūksta laisvai prieinamų priemonių kompaktiškiems kolekcijų žodynams sudaryti. Klasterizavimą galima vykdyti ir neturint kompaktiškų žodynų, tačiau tai gali lemti ne tik padidėjusias skaičiuojamųjų resursų sąnaudas, bet ir netikslius arba iškraipytus klasterizavimo rezultatus.

- Šiame darbe lietuviškų dokumentų klasterizavimo problemai spręsti buvo sukurtas naujas klasterizavimo metodas, kuriame dokumentų reprezentacijai panaudoti aprašantys metaduomenys – tematikos arba kategorijos vaidmenį atliekantys raktažodžiai – žymės.

- Žymėmis reprezentuojamų dokumentų panašumui nustatyti netinka euklidinis atstumas, nes šis Minkovskio atstumo atvejis yra linkęs laikyti panašiais netgi tuos dokumentus, kurie neturi tarpusavyje bendrų žymių. Tam labiau tinka kosinuso koeficientas.

- Eksperimentais nustatyta, kad kosinuso koeficientu matuojant žymėmis reprezentuojamų dokumentų panašumą, klasterizavimui labiausiai tinka skaldantis K-means algoritmas, optimizuojantis bendrojo glaudumo (angl. *total coherence*) kriterijų.

- Eksperimentais nustatyta, kad didėjant dokumentų kiekiui ir/arba skirstinių skaičiui, sukurtasis metodas skirstinius formuoja tiksliau nei tradicinis klasterizavimo metodas, naudojantis nekompaktišką kolekcijos žodyną, tačiau esant mažam skirstinių kiekiui, klasterizavimo rezultatus neigiamai veikia santykinai nedidelis dokumentus reprezentuojančių savybių skaičius, todėl šiais atvejais naujasis metodas nusileidžia tradiciniam klasterizavimui didesne entropija.

- Suformuotų skirstinių tikslumo įvertinimo eksperimentai leidžia daryti išvadą, kad sukurtasis metodas gali išspręsti vientiso klasifikatoriaus neturinčios dokumentų saugyklos problemą. Be to, naujasis metodas labiau nei tradicinis tinka didelėms, žymėmis arba raktažodžiais aprašytų dokumentų kolekcijoms klasterizuoti.

LITERATŪRA

ANDERBERG, M. R. *Cluster Analysis for Applications*. New York, Academic Press, 1973.

ANDRITSOS, P.; TZERPOS, V. Information-Theoretic Software Clustering. *IEEE Transactions on Software Engineering*, 2005, Nr. 31/2, p. 150–65.

ANICK, P. G.; VAITHYANATHAN, S. Exploiting Clustering and Phrases for Context-Based Information Retrieval. *ACM SIGIR Conference on Research and Development in Information Retrieval: konferencijos pranešimų medžiaga. ACM*, 1997, p. 314–323.

BERNOTAS, M.; ŽALINAUSKAS, M. Dokumentų valdymo modelis žiniasklaidos redakcijos sistemoje. *Informacinės technologijos'2006: konferencijos pranešimų medžiaga. KTU*, 2006, p. 437–440.

BURGIN, R. The Effect of Indexing Exhaustivity on Retrieval Performance. *Information Processing & Management*, 1991, Nr. 27/6, p. 623–628.

BURGIN, R. The Retrieval Effectiveness of Five Clustering Algorithms as a Function of Indexing Exhaustivity. *Journal of the American Society for Information Science*, 1995, Nr. 46/8, p. 562–572.

CAMPBELL, I. The Ostensive Model of Developing Information Needs. *Daktaro disertacija*. University of Glasgow, Glasgow, 2000.

CHANG, H-C.; HSU, C-C.; DENG, Y-W. Unsupervised Document Clustering Based on Keyword Clusters. *IEEE International Symposium on Communications and Information Technologies: konferencijos pranešimų medžiaga. IEEE*, 2004, p. 1198–1203.

CHANG, H-C.; HSU, C-C. Using Topic Keyword Clusters for Automatic Document Clustering. *IEICE Transactions on Information and Systems*, 2005, Nr. E88/8, p. 1852–1860.

CORMACK, R. M. A Review of Classification. *Journal of the Royal Statistical Society*, 1971, Nr. A/134, p. 321–367.

CUTTING, D. R.; KARGER, D. R.; PEDERSEN, J. O.; TUKEY, J. W. Scatter/Gather: A Cluster-Based Approach To Browsing Large Document Collections. *ACM SIGIR Conference on Research and Development in Information Retrieval: konferencijos pranešimų medžiaga. ACM*, 1992, p. 318–329.

DAVIES, J.; COCHRANE, R. Knowledge Discovery and Delivery. *British Telecommunications Engineering*, 1998, Nr. 17, p. 25–35.

DEBOECK, G. Financial Applications of Self-Organizing Maps. *Neural Network World*, 1998, Nr. 8/2, p. 213–241.

EFTHIMIADIS, E. N. Interactive query expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 2000, Nr. 51/11, p. 989–1003.

FRED, A. L. N.; JAIN, A. K. Data Clustering Using Evidence Accumulation. *Proceedings of the International Conference on Pattern Recognition: konferencijos pranešimų medžiaga. IEEE Computer Society Press*, 2002, p. 276–280.

FRIEDMAN, M.; SCHNEIDER, M.; LAST, M.; ZAAFRANY, O.; KANDEL, A. A New Approach for Fuzzy Clustering of Web Documents. *IEEE International Conference on Fuzzy Systems: konferencijos pranešimų medžiaga. IEEE*, 2004, p. 377–381.

GOLDER, S. A.; HUBERMAN, B. A. Usage Patterns of Collaborative Tagging Systems.

Journal of Information Science, 2006, Nr. 32/2, p. 198–208.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, 2000, Nr. 25/5, p. 345–366.

HAGEN, P. Must Search Stink? *Tyrimų ataskaita.* Cambridge, JAV, Forrester Research, 2000.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of Data Mining*. MIT Press, 2001.

HARTUV, E.; SCHMITT, A.; LANGE, J.; MEIER-EWERT, S.; LEHRACH, H.; SHAMIR, R. Algorithm for Clustering cDNAs for Gene Expression Analysis. *Proceedings of the Annual International Conference on Computational Molecular Biology: konferencijos pranešimų medžiaga. ACM*, 1999, p. 188–197.

HATZIVASSILOGLOU, V.; GRAVANO, L.; MAGANTI, A. Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering. *ACM SIGIR Conference on Research and Development in Infornation Retrieval: konferencijos pranešimų medžiaga. ACM*, 2000, p. 224–231.

HAUTAMAKI, V.; CHEREDNICHENKO, S.; KARKKAINEN, I.; KINNUNEN, T.; FRANTI, P. Improving K-means by Outlier Removal. *Image Analysis. 14th Scandinavian Conference, SCIA 2005: konferencijos pranešimų medžiaga. Springer-Verlag*, 2005, p. 978–87.

HEARST, M. A.; PEDERSEN, J. O. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *ACM SIGIR Conference on Research and Development in Information Retrieval: konferencijos pranešimų medžiaga. ACM*, 1996, p. 76–84.

HÚSEK, D.; POKORNÝ, J.; ŘEZANKOVÁ, H.; SNÁŠEL, V. *Data Clustering: From Documents to the Web*. Skyrius iš knygos *Web Data Management Practices: Emerging Techniques and Technologies*. Vakali, A.; Pallis, G. (red.). Idea Group, Inc., 2006.

YUAN-CHAO L.; XIAO-LONG W.; BING-QUAN L. A Feature Selection Algorithm for Document Clustering Based on Word Co-Occurrence Frequency. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics: konferencijos pranešimų medžiaga. IEEE*, 2004, p. 2963–2968.

JAIN, A. K.; MURTY M. N.; FLYNN P. J. Data Clustering: A Review. *ACM Computing Surveys*, 1999, Nr. 31/3, p. 264–323.

JAIN, A. K.; TOPCHY, A.; LAW, M. H. C.; BUHMANN, J. M. Landscape of clustering algorithms. *International Conference on Pattern Recognition: konferencijos pranešimų medžiaga. IEEE*, 2004, p. 260–263.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Prentice Hall, 1988.

JARDINE, N.; VAN RIJSBERGEN C. J. The Use of Hierarchic Clustering in Information Retrieval. *Information Storage and Retrieval*, 1971, Nr. 7/5, p. 217–240.

JIANG, M. F.; TSENG, S. S.; SU, C. M. Two-Phase Clustering Process for Outliers Detection. *Pattern Recognition Letters*, 2001, Nr. 22/6–7, p. 691–700.

KASKI, S.; HONKELA, T.; LAGUS, K.; KOHONEN, T. WEBSOM – Self-Organizing Maps of Document Collections. *Neurocomputing*, 1998, Nr. 21/1–3, p. 101–117.

KIRRIEMUIR, J. W.; WILLETT, P. Identification of Duplicate and Near-Duplicate Full-Text Records in Database Search-Outputs Using Hierarchic Cluster Analysis. *Program*, 1995, Nr. 29/3, p. 241–256.

KOHONEN, T. Self-Organizing Map. *Neurocomputing*, 1998, Nr. 21/1–3, p. 1–6.

KOHONEN, T. *Self-Organizing Maps, Third Edition*. Springer, 2000.

KORPIMIES, K.; UKKINEN, E. Term Weighting in Query-Based Document Clustering. *Advances in Databases and Information Systems. Second East European Symposium, ADBIS'98.: konferencijos pranešimų medžiaga. Springer–Verlag*, 1998, p. 151–153.

KUMMAMURU, K.; DHAWALE, A.; KRISHNAPURAM, R. Fuzzy Co-Clustering of Documents and Keywords. *Proceedings of the 12th IEEE International Conference on Fuzzy Systems: konferencijos pranešimų medžiaga. IEEE*, 2003, p. 772–777.

KURAL, Y.; ROBERTSON, S. E.; JONES, S. Deciphering Cluster Representations. *Information Processing and Management*, 2001, Nr. 37/4, p. 593–601.

KURAL, Y. Clustering information retrieval search outputs. *Daktaro disertacija*. City University, London, 1999.

LAGUS, K. Text Mining With the WEBSOM. *Daktaro disertacija*. Helsinki University of Technology, Helsinki, 2000.

LANCE, G. N.; WILLIAMS, W. T. A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems.. *Computer Journal*, 1967, Nr. 9, p. 373–380.

LEOUSKI, A. V.; CROFT, W. B. An Evaluation of Techniques for Clustering Search Results. *Tyrimų ataskaita.* Department of Computer Science, University of Massachusetts, Amherst, 1996.

LIPING J. ; NG, M. K.; JUN, X.; HUANG, J. Z. Subspace Clustering of Text Documents with Feature Weighting K-means Algorithm. *Advances in Knowledge Discovery and Data Mining: konferencijos pranešimų medžiaga. Springer–Verlag*, 2005, p. 802–812.

MAAREK, Y. S.; FAGIN, R.; BEN-SHAUL, I. Z.; PELLEG, D. Ephemeral Document Clustering for Web Applications. *Tyrimų ataskaita.* IBM Research, 2000.

MILLIGAN, G. W.; SOON, S. C.; SOKOL, L. M. The Effect of Cluster Size, Dimensionality, and Thenumber of Cluster on Recovery of True Cluster Structure.. *IEEE Transactions on PatterRecognition and Machine Intelligence*, 1983, Nr. 5/1, p. 40–47.

MIZZARO, S. Relevance: The Whole History. *Journal of the American Society for Information Science*, 1997, Nr. 48/9, p. 810–832.

MIZZARO, S. How many relevances in information retrieval? *Interacting With Computers*, 1998, Nr. 10/3, p. 305–322.

NETO, J. L.; KAESTNER, C. A. A.; SANTOS, A. D.; FREITAS, A. A. Document Clustering and Summarization. *International Conference on the Practical Application of Knowledge Discovery and Data Mining: konferencijos pranešimų medžiaga. Practical Application Company*, 2000, p. 41–56.

PELLEG D.; MOORE, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Proceedings of the International Conference on Machine Learning: konferencijos pranešimų medžiaga. Stanford University, Standord*, 2000, p. 727–734.

ПЛЕШКО В. В.; ЕРМАКОВ А. Е.; ЛИПИНСКИЙ Г. В. TopSOM: визуализация информационных массивов с применением самоорганизующихся тематических карт. *Информационные технологии*, 2001, Nr. 8, p. 8–11.

RASMUSSEN, C. E. The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems*, 2000, Nr. 12, p. 554–560.

RORVIG, M. Images of Similarity: A Visual Exploration of Optimal Similarity Metrics and Scaling Properties of TREC Topic-Document Sets. *Journal of the American Society for Information Science*, 1999, Nr. 50/8, p. 639–651.

ROUSSINOV, D. G.; CHEN, H. Information navigation on the web by Clustering and Summarizing Query Results. *Information Processing and Management*, 2001, Nr. 37/6, p. 789–816.

ROWLEY, J. Knowledge Organisation for a New Millennium: Principles and Processes. *Journal of Knowledge Management*, 2000, Nr. 4/3, p. 217–23.

SHAW, W. M. Subject Indexing and Citation Indexing. I. Clustering Structure in the Cystic Fibrosis Document Collection. *Information Processing & Management*, 1990, Nr. 26/6, p. 693–703.

SILVERSTEIN, C.; PEDERSEN, J. O. Almost-Constant-Time Clustering of Arbitrary Corpus Subsets. *ACM SIGIR Conference on Research and Development in Information Retrieval: konferencijos pranešimų medžiaga. ACM*, 1997, p. 60-66.

SMALL, D. J. A Model-Driven Architecture for Enterprise Document Management, Supporting Discovery and Reuse. *Daktaro disertacija*. University of Leeds, Leeds, 1999.

SMEATON, A.F.; BURNETT, M.; KELLEDY, F.; QUINN, Q. An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts. *Proceedings of the 20th BCS-IRSG Colloquium: konferencijos pranešimų medžiaga. Springer-Verlag*, 1998, p. 1–9.

STEINBACH M.; KARYPIS, G.; KUMAR, V. A Comparison of Document Clustering Techniques. *Tyrimų ataskaita.* University of Minessota, 2000.

TAN, P-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. Addison Wesley, 2005.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Second Edition*. San Diego, Academic Press, 2003.

TOMBROS, A. The Effectiveness of Query-Based Hierarchic Clustering of Documents for Information Retrieval. *Daktaro disertacija*. University of Glasgow, Glasgow, 2002.

VÉLEZ, B.; WIESS, R.; SHELDON, M. A.; GIFFORD, D. K. Fast and Effective Query Refinement. *ACM SIGIR Conference on Research and Development in Information Retrieval: konferencijos pranešimų medžiaga. ACM*, 1997, p. 6–15.

WEI-HO, T.; RODGERS, D.; HSIN-MIN W. Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics. *Computer Music Journal*, 2004, Nr. 28/3, p. 68–78.

WILLETT, P. Recent Trends in Hierarchic Document Clustering: a Critical Review. *Information Processing and Management*, 1988, Nr. 24/5, p. 577–597.

WU, M.; FULLER, M.; WILKINSON, R. Using Clustering and Classification Approaches Ininteractive Retrieval. *Information Processing and Management*, 2001, Nr. 37/3, p. 459–484.

ZINKEVIČIUS, V. Lemuoklis – morfologinei analizei. *Darbai ir dienos*, 2000, Nr. 24, p. 245–273.

ŽALINAUSKAS, M. Dokumentų klasterizavimas. *Informacinės technologijos 2006: konferencijos pranešimų medžiaga. VUKHF*, 2006, p. 207–211.

ŽALINAUSKAS, M. Clustering Method for Personally Classified Documents. *Master thesis*. Kaunas technology university, Kaunas, 2006.

SUMMARY

Traditional clustering methods, where documents are represented by term frequency vectors, are not very suitable for Lithuanian document clustering as there is no any freely available morphological analyzer or stemmer to make compact term dictionaries. It is still possible though to cluster Lithuanian documents using loose term dictionaries, but as Lithuanian is a highly synthetic language significant increase in resources and possibly inaccurate or distorted results must be taken into account.

In this master thesis a clustering method for personally classified documents is developed to overcome shortcomings of traditional document clustering stated above. In a new method documents are represented by tag frequency vectors, pair-wise similarities are measured by cosine coefficient and clustering itself is performed using experimentally selected bisecting K-means algorithm.

Experiments comparing developed method with traditional document clustering using loose term dictionary showed that former copes better with large document collections and/or large cluster number. At the same time subjective clustering estimation showed that even when new method demonstrates larger entropy and lower purity values, it still overcomes traditional method by clustering sense.

PRIEDAS A – EKSPERIMENTŲ ŽURNALAS

Šiame priede pateikiamas individualiai suklasifikuotų dokumentų kolekcijų klasterizavimo aglomeratyviu algoritmu eksperimento žurnalas. Klasterizavimo metu dokumentų reprezentacijai naudotos individualiai vartotojų priskirtos žymės.

AGLOMERATYVUS, 1056 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 1056 **Skirstinių:** 5

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,693, **Grynumas:** 0,564

Skirstinys	Dydis[24]	VPan[25]	VNuok[26]	IPan[27]	INuok[28]	Entropija	Grynumas
0	2	1,000	0,000	0,000	0,000	0,000	1,000
1	2	1,000	0,000	0,000	0,000	0,000	1,000
2	221	0,093	0,071	0,004	0,004	0,383	0,837
3	348	0,025	0,014	0,005	0,005	0,732	0,560
4	475	0,018	0,010	0,005	0,005	0,814	0,436

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	2	0	0	0
1	0	0	0	2	0
2	22	2	5	7	185
3	195	69	60	15	9
4	91	49	207	123	5

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	ugnikalnis	100,0%	savivaldybės	0,0%				
Išskiria:	ugnikalnis	50,0%	krepšinis	4,5%				
Skirstinys 1 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	moksleiviai	100,0%	kalėdos	0,0%				
Išskiria:	moksleiviai	50,0%	krepšinis	4,5%				
Skirstinys 2 **Dydis:** 221 **VPan:** 0,093 **IPan:** 0,004								
Sieja:	krepšinis	28,7%	čempionatas	20,5%	rezultatai	17,7%	lietuviai	7,0%
Išskiria:	krepšinis	15,8%	čempionatas	11,3%	rezultatai	9,0%	lietuviai	3,7%
Skirstinys 3 **Dydis:** 348 **VPan:** 0,025 **IPan:** 0,005								
Sieja:	aukos	14,2%	irakas	6,6%	energetika	6,4%	rinkimai	6,3%
Išskiria:	krepšinis	9,2%	aukos	7,7%	čempionatas	6,1%	rezultatai	4,5%
Skirstinys 4 **Dydis:** 475 **VPan:** 0,018 **IPan:** 0,005								
Sieja:	įvairybės	12,5%	finansai	5,6%	ekonomika	5,5%	įmonės	5,0%
Išskiria:	krepšinis	10,1%	čempionatas	7,7%	rezultatai	7,1%	įvairybės	6,0%

24 Skirstiniui priskirtų dokumentų skaičius.

25 Vidutinis panašumas tarp skirstinyje esančių dokumentų (**V**idinis **Pan**ašumas).

26 Skirstinyje esančių dokumentų panašumų standartinis nuokrypis (**V**idinis **Nuok**rypis).

27 Vidutinis panašumas tarp skirstinyje esančių dokumentų ir likusių kolekcijos dokumentų (**I**šorinis **Pan**ašumas).

28 Skirstinyje esančių dokumentų panašumų su likusiais kolekcijos dokumentais standartinis nuokrypis (**I**šorinis **Nuok**rypis).

AGLOMERATYVUS, 1056 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 1056 **Skirstinių:** 10

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,619, **Grynumas:** 0,566

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	2	1,000	0,000	0,000	0,000	0,000	1,000
1	2	1,000	0,000	0,000	0,000	0,000	1,000
2	170	0,030	0,016	0,004	0,005	0,874	0,312
3	127	0,059	0,031	0,009	0,006	0,556	0,717
4	171	0,043	0,019	0,005	0,005	0,640	0,579
5	115	0,064	0,039	0,005	0,004	0,525	0,574
6	134	0,054	0,033	0,007	0,005	0,725	0,425
7	94	0,283	0,124	0,008	0,006	0,000	1,000
8	104	0,063	0,036	0,006	0,005	0,616	0,692
9	129	0,047	0,028	0,006	0,005	0,762	0,442

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	2	0	0	0
1	0	0	0	2	0
2	23	40	51	53	3
3	22	2	5	7	91
4	49	4	99	18	1
5	66	44	4	0	1
6	19	5	57	52	1
7	0	0	0	0	94
8	72	16	7	6	3
9	57	9	49	9	5

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	ugnikalnis	100,0%	savivaldybės	0,0%				
Išskiria:	ugnikalnis	50,0%	krepšinis	4,5%				
Skirstinys 1 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	moksleiviai	100,0%	kalėdos	0,0%				
Išskiria:	moksleiviai	50,0%	krepšinis	4,5%				
Skirstinys 2 **Dydis:** 170 **VPan:** 0,030 **IPan:** 0,004								
Sieja:	transportas	18,4%	muzika	4,4%	kinas	4,4%	aviacija	4,0%
Išskiria:	transportas	10,5%	krepšinis	6,2%	čempionatas	4,7%	rezultatai	4,6%
Skirstinys 3 **Dydis:** 127 **VPan:** 0,059 **IPan:** 0,009								
Sieja:	rezultatai	13,2%	olimpiada	9,1%	ledo ritulys	8,7%	čempionatas	8,5%
Išskiria:	olimpiada	6,5%	ledo ritulys	6,1%	futbolas	5,2%	krepšinis	5,1%
Skirstinys 4 **Dydis:** 171 **VPan:** 0,043 **IPan:** 0,005								
Sieja:	finansai	16,9%	ekonomika	12,6%	es	9,0%	palestiniečiai	7,5%
Išskiria:	finansai	10,0%	krepšinis	7,0%	ekonomika	6,1%	čempionatas	5,0%
Skirstinys 5 **Dydis:** 115 **VPan:** 0,064 **IPan:** 0,005								
Sieja:	aukos	34,4%	irakas	15,1%	gaisras	5,2%	išpuolis	5,2%
Išskiria:	aukos	16,7%	irakas	7,1%	krepšinis	6,2%	čempionatas	4,5%
Skirstinys 6 **Dydis:** 134 **VPan:** 0,054 **IPan:** 0,007								
Sieja:	įvairybės	32,1%	įmonės	12,0%	prekyba	8,2%	vokietija	5,3%
Išskiria:	įvairybės	15,4%	krepšinis	6,8%	įmonės	5,7%	čempionatas	5,2%
Skirstinys 7 **Dydis:** 94 **VPan:** 0,283 **IPan:** 0,008								
Sieja:	krepšinis	47,1%	čempionatas	18,4%	rezultatai	11,7%	lietuviai	10,3%
Išskiria:	krepšinis	25,9%	čempionatas	6,7%	lietuviai	5,2%	įvairybės	4,0%
Skirstinys 8 **Dydis:** 104 **VPan:** 0,063 **IPan:** 0,006								
Sieja:	rinkimai	21,6%	seimas	18,5%	baltarusija	12,5%	parlamentas	5,9%

Išskiria:	rinkimai	11,3%	seimas	10,4%	baltarusija	6,7%	krepšinis	6,4%
Skirstinys 9 **Dydis:** 129 **VPan:** 0,047 **IPan:** 0,006								
Sieja:	energetika	20,6%	nafta	8,3%	rusija	7,3%	dujos	6,1%
Išskiria:	energetika	11,2%	krepšinis	6,9%	nafta	5,4%	čempionatas	4,3%

AGLOMERATYVUS, 1056 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 1056 **Skirstinių:** 15

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,522, **Grynumas:** 0,680

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	2	1,000	0,000	0,000	0,000	0,000	1,000
1	2	1,000	0,000	0,000	0,000	0,000	1,000
2	91	0,062	0,028	0,009	0,007	0,648	0,637
3	82	0,075	0,035	0,005	0,004	0,727	0,463
4	110	0,034	0,014	0,003	0,005	0,745	0,418
5	115	0,064	0,039	0,005	0,004	0,525	0,574
6	71	0,090	0,054	0,007	0,005	0,667	0,662
7	94	0,283	0,124	0,008	0,006	0,000	1,000
8	104	0,063	0,036	0,006	0,005	0,616	0,692
9	89	0,081	0,035	0,006	0,005	0,380	0,820
10	36	0,289	0,102	0,008	0,004	0,211	0,917
11	60	0,111	0,052	0,006	0,005	0,430	0,800
12	58	0,079	0,031	0,006	0,005	0,439	0,810
13	52	0,119	0,075	0,007	0,006	0,634	0,673
14	82	0,073	0,034	0,008	0,006	0,606	0,634

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	2	0	0	0
1	0	0	0	2	0
2	20	2	5	6	58
3	38	4	26	14	0
4	23	37	3	46	1
5	66	44	4	0	1
6	10	7	47	3	4
7	0	0	0	0	94
8	72	16	7	6	3
9	11	0	73	4	1
10	2	0	0	1	33
11	0	3	48	7	2
12	47	2	2	6	1
13	8	3	5	35	1
14	11	2	52	17	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	ugnikalnis	100,0%	savivaldybės	0,0%				
Išskiria:	ugnikalnis	50,0%	krepšinis	4,5%				
Skirstinys 1 **Dydis:** 2 **VPan:** 1,000 **IPan:** 0,000								
Sieja:	moksleiviai	100,0%	kalėdos	0,0%				
Išskiria:	moksleiviai	50,0%	krepšinis	4,5%				
Skirstinys 2 **Dydis:** 91 **VPan:** 0,062 **IPan:** 0,009								
Sieja:	čempionatas	15,9%	futbolas	15,8%	ispanija	10,1%	tenisas	3,4%
Išskiria:	futbolas	10,3%	ispanija	6,9%	krepšinis	4,0%	čempionatas	3,0%

Skirstinys 3	**Dydis:** 82	**VPan:** 0,075	**IPan:** 0,005					
Sieja:	palestiniečiai	18,8%	paukščių gripas	13,6%	hamas	12,8%	izraelis	9,1%
Išskiria:	palestiniečiai	10,0%	paukščių gripas	7,5%	hamas	7,2%	krepšinis	5,8%
Skirstinys 4	**Dydis:** 110	**VPan:** 0,034	**IPan:** 0,003					
Sieja:	muzika	9,2%	kinas	9,2%	pasienis	6,1%	žemės	4,4%
Išskiria:	krepšinis	5,7%	kinas	5,4%	muzika	5,3%	čempionatas	4,1%
Skirstinys 5	**Dydis:** 115	**VPan:** 0,064	**IPan:** 0,005					
Sieja:	aukos	34,4%	irakas	15,1%	gaisras	5,2%	išpuolis	5,2%
Išskiria:	aukos	16,7%	irakas	7,1%	krepšinis	6,2%	čempionatas	4,5%
Skirstinys 6	**Dydis:** 71	**VPan:** 0,090	**IPan:** 0,007					
Sieja:	energetika	31,6%	nafta	13,0%	dujos	10,5%	rusija	7,9%
Išskiria:	energetika	16,4%	nafta	7,8%	dujos	6,5%	krepšinis	6,1%
Skirstinys 7	**Dydis:** 94	**VPan:** 0,283	**IPan:** 0,008					
Sieja:	krepšinis	47,1%	čempionatas	18,4%	rezultatai	11,7%	lietuviai	10,3%
Išskiria:	krepšinis	25,9%	čempionatas	6,7%	lietuviai	5,2%	įvairybės	4,0%
Skirstinys 8	**Dydis:** 104	**VPan:** 0,063	**IPan:** 0,006					
Sieja:	rinkimai	21,6%	seimas	18,5%	baltarusija	12,5%	parlamentas	5,9%
Išskiria:	rinkimai	11,3%	seimas	10,4%	baltarusija	6,7%	krepšinis	6,4%
Skirstinys 9	**Dydis:** 89	**VPan:** 0,081	**IPan:** 0,006					
Sieja:	finansai	30,6%	ekonomika	22,4%	es	12,1%	paslaugos	2,4%
Išskiria:	finansai	17,5%	ekonomika	11,0%	krepšinis	6,2%	es	4,5%
Skirstinys 10	**Dydis:** 36	**VPan:** 0,289	**IPan:** 0,008					
Sieja:	olimpiada	23,1%	ledo ritulys	22,2%	rezultatai	18,3%	turinas	17,4%
Išskiria:	olimpiada	13,2%	ledo ritulys	12,5%	turinas	9,7%	krepšinis	5,5%
Skirstinys 11	**Dydis:** 60	**VPan:** 0,111	**IPan:** 0,006					
Sieja:	transportas	36,4%	aviacija	8,5%	statyba	7,4%	streikas	4,0%
Išskiria:	transportas	19,8%	krepšinis	5,2%	aviacija	4,8%	statyba	4,2%
Skirstinys 12	**Dydis:** 58	**VPan:** 0,079	**IPan:** 0,006					
Sieja:	iranas	17,3%	tyrimas	15,3%	adamkus	11,0%	branduolinė programa	8,3%
Išskiria:	tyrimas	9,1%	iranas	8,6%	adamkus	6,6%	krepšinis	5,8%
Skirstinys 13	**Dydis:** 52	**VPan:** 0,119	**IPan:** 0,007					
Sieja:	įvairybės	55,8%	britanija	11,8%	prancūzija	6,0%	kalėdos	4,5%
Išskiria:	įvairybės	24,1%	britanija	5,6%	krepšinis	5,2%	čempionatas	4,2%
Skirstinys 14	**Dydis:** 82	**VPan:** 0,073	**IPan:** 0,008					
Sieja:	įmonės	23,6%	prekyba	14,4%	japonija	10,4%	vokietija	9,4%
Išskiria:	įmonės	12,7%	prekyba	8,5%	krepšinis	6,8%	japonija	6,4%

AGLOMERATYVUS, 1056 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 1056 **Skirstinių:** 20

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,479, **Grynumas:** 0,720

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	2	1,000	0,000	0,000	0,000	0,000	1,000
1	2	1,000	0,000	0,000	0,000	0,000	1,000
2	53	0,068	0,029	0,004	0,004	0,642	0,623
3	68	0,065	0,035	0,006	0,004	0,713	0,574
4	41	0,122	0,057	0,005	0,004	0,720	0,488
5	115	0,064	0,039	0,005	0,004	0,525	0,574
6	71	0,090	0,054	0,007	0,005	0,667	0,662
7	94	0,283	0,124	0,008	0,006	0,000	1,000
8	59	0,067	0,019	0,008	0,006	0,759	0,525
9	89	0,081	0,035	0,006	0,005	0,380	0,820
10	36	0,289	0,102	0,008	0,004	0,211	0,917
11	60	0,111	0,052	0,006	0,005	0,430	0,800

12	58	0,079	0,031	0,006	0,005	0,439	0,810
13	52	0,119	0,075	0,007	0,006	0,634	0,673
14	36	0,225	0,078	0,008	0,004	0,211	0,917
15	41	0,161	0,074	0,005	0,004	0,489	0,756
16	57	0,062	0,026	0,003	0,004	0,489	0,737
17	32	0,212	0,059	0,011	0,009	0,318	0,844
18	28	0,222	0,079	0,008	0,006	0,347	0,821
19	54	0,089	0,039	0,009	0,006	0,668	0,537

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	2	0	0	0
1	0	0	0	2	0
2	13	33	2	4	1
3	39	16	7	5	1
4	7	2	20	12	0
5	66	44	4	0	1
6	10	7	47	3	4
7	0	0	0	0	94
8	16	2	4	6	31
9	11	0	73	4	1
10	2	0	0	1	33
11	0	3	48	7	2
12	47	2	2	6	1
13	8	3	5	35	1
14	33	0	0	1	2
15	31	2	6	2	0
16	10	4	1	42	0
17	4	0	1	0	27
18	0	1	23	4	0
19	11	1	29	13	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 2	**VPan:** 1,000	**IPan:** 0,000					
Sieja:	ugnikalnis	100,0%	savivaldybės	0,0%				
Išskiria:	ugnikalnis	50,0%	krepšinis	4,5%				
Skirstinys 1	**Dydis:** 2	**VPan:** 1,000	**IPan:** 0,000					
Sieja:	moksleiviai	100,0%	kalėdos	0,0%				
Išskiria:	moksleiviai	50,0%	krepšinis	4,5%				
Skirstinys 2	**Dydis:** 53	**VPan:** 0,068	**IPan:** 0,004					
Sieja:	pasienis	13,3%	žemės	9,5%	drebėjimas	9,5%	sulaikymas	8,5%
Išskiria:	pasienis	7,5%	drebėjimas	5,4%	žemės	5,4%	krepšinis	5,3%
Skirstinys 3	**Dydis:** 68	**VPan:** 0,065	**IPan:** 0,006					
Sieja:	seimas	36,2%	lenkija	11,7%	aukos	2,5%	liberalai	2,4%
Išskiria:	seimas	20,4%	krepšinis	6,0%	lenkija	5,7%	čempionatas	4,3%
Skirstinys 4	**Dydis:** 41	**VPan:** 0,122	**IPan:** 0,005					
Sieja:	paukščių gripas	29,7%	žemės ūkis	20,0%	aplinkosauga	10,7%	artimieji	5,2%
Išskiria:	paukščių gripas	15,9%	žemės ūkis	9,8%	aplinkosauga	5,7%	krepšinis	5,3%
Skirstinys 5	**Dydis:** 115	**VPan:** 0,064	**IPan:** 0,005					
Sieja:	aukos	34,4%	irakas	15,1%	gaisras	5,2%	išpuolis	5,2%
Išskiria:	aukos	16,7%	irakas	7,1%	krepšinis	6,2%	čempionatas	4,5%
Skirstinys 6	**Dydis:** 71	**VPan:** 0,090	**IPan:** 0,007					
Sieja:	energetika	31,6%	nafta	13,0%	dujos	10,5%	rusija	7,9%
Išskiria:	energetika	16,4%	nafta	7,8%	dujos	6,5%	krepšinis	6,1%
Skirstinys 7	**Dydis:** 94	**VPan:** 0,283	**IPan:** 0,008					
Sieja:	krepšinis	47,1%	čempionatas	18,4%	rezultatai	11,7%	lietuviai	10,3%
Išskiria:	krepšinis	25,9%	čempionatas	6,7%	lietuviai	5,2%	įvairybės	4,0%
Skirstinys 8	**Dydis:** 59	**VPan:** 0,067	**IPan:** 0,008					

Sieja:	tenisas	7,4%	čempionatas	6,9%	rankinis	6,9%	dviračiai	6,9%
Išskiria:	krepšinis	5,7%	tenisas	5,0%	rankinis	4,8%	dviračiai	4,6%
Skirstinys 9	**Dydis:** 89	**VPan:** 0,081	**IPan:** 0,006					
Sieja:	finansai	30,6%	ekonomika	22,4%	es	12,1%	paslaugos	2,4%
Išskiria:	finansai	17,5%	ekonomika	11,0%	krepšinis	6,2%	es	4,5%
Skirstinys 10	**Dydis:** 36	**VPan:** 0,289	**IPan:** 0,008					
Sieja:	olimpiada	23,1%	ledo ritulys	22,2%	rezultatai	18,3%	turinas	17,4%
Išskiria:	olimpiada	13,2%	ledo ritulys	12,5%	turinas	9,7%	krepšinis	5,5%
Skirstinys 11	**Dydis:** 60	**VPan:** 0,111	**IPan:** 0,006					
Sieja:	transportas	36,4%	aviacija	8,5%	statyba	7,4%	streikas	4,0%
Išskiria:	transportas	19,8%	krepšinis	5,2%	aviacija	4,8%	statyba	4,2%
Skirstinys 12	**Dydis:** 58	**VPan:** 0,079	**IPan:** 0,006					
Sieja:	iranas	17,3%	tyrimas	15,3%	adamkus	11,0%	branduolinė programa	8,3%
Išskiria:	tyrimas	9,1%	iranas	8,6%	adamkus	6,6%	krepšinis	5,8%
Skirstinys 13	**Dydis:** 52	**VPan:** 0,119	**IPan:** 0,007					
Sieja:	įvairybės	55,8%	britanija	11,8%	prancūzija	6,0%	kalėdos	4,5%
Išskiria:	įvairybės	24,1%	britanija	5,6%	krepšinis	5,2%	čempionatas	4,2%
Skirstinys 14	**Dydis:** 36	**VPan:** 0,225	**IPan:** 0,008					
Sieja:	rinkimai	34,7%	baltarusija	25,0%	parlamentas	13,8%	ukraina	6,3%
Išskiria:	rinkimai	16,8%	baltarusija	12,9%	parlamentas	7,8%	krepšinis	5,6%
Skirstinys 15	**Dydis:** 41	**VPan:** 0,161	**IPan:** 0,005					
Sieja:	palestiniečiai	32,6%	hamas	24,0%	izraelis	17,0%	parama	5,9%
Išskiria:	palestiniečiai	16,9%	hamas	13,1%	izraelis	8,9%	krepšinis	5,3%
Skirstinys 16	**Dydis:** 57	**VPan:** 0,062	**IPan:** 0,003					
Sieja:	muzika	19,0%	kinas	16,3%	teatras	7,6%	mirtis	6,1%
Išskiria:	muzika	10,5%	kinas	9,0%	krepšinis	5,2%	teatras	4,2%
Skirstinys 17	**Dydis:** 32	**VPan:** 0,212	**IPan:** 0,011					
Sieja:	futbolas	37,2%	ispanija	23,8%	čempionatas	11,5%	anglija	5,2%
Išskiria:	futbolas	21,8%	ispanija	14,3%	rusija	3,5%	rezultatai	3,3%
Skirstinys 18	**Dydis:** 28	**VPan:** 0,222	**IPan:** 0,008					
Sieja:	įmonės	51,1%	japonija	24,6%	verslas	5,9%	įsigijimas	3,2%
Išskiria:	įmonės	25,2%	japonija	13,5%	krepšinis	5,5%	čempionatas	3,9%
Skirstinys 19	**Dydis:** 54	**VPan:** 0,089	**IPan:** 0,009					
Sieja:	prekyba	27,3%	vokietija	15,7%	kinija	6,6%	pramonė	5,4%
Išskiria:	prekyba	16,6%	krepšinis	6,5%	vokietija	5,4%	čempionatas	4,7%

AGLOMERATYVUS, 2112 DOKUMENTŲ, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 2112 **Skirstinių:** 5

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,732, **Grynumas:** 0,561

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	391	0,099	0,074	0,006	0,004	0,525	0,762
1	730	0,020	0,012	0,005	0,005	0,760	0,547
2	442	0,027	0,016	0,005	0,004	0,678	0,613
3	297	0,031	0,025	0,004	0,004	0,844	0,475
4	247	0,042	0,030	0,004	0,004	0,937	0,296

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	33	298	5	28	27
1	50	41	58	399	182
2	32	26	101	12	271
3	141	19	59	22	56
4	26	73	22	68	58

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 391	**VPan:** 0,099	**IPan:** 0,006					
Sieja:	krepšinis	33,2%	čempionatas	25,3%	rezultatai	7,9%	lietuviai	6,7%
Išskiria:	krepšinis	18,7%	čempionatas	13,5%	įvairybės	4,0%	lietuviai	3,4%
Skirstinys 1	**Dydis:** 730	**VPan:** 0,020	**IPan:** 0,005					
Sieja:	rusija	12,3%	energetika	7,9%	finansai	6,6%	es	6,4%
Išskiria:	krepšinis	9,6%	čempionatas	7,1%	rezultatai	6,0%	energetika	5,1%
Skirstinys 2	**Dydis:** 442	**VPan:** 0,027	**IPan:** 0,005					
Sieja:	aukos	13,7%	irakas	11,1%	rinkimai	9,9%	izraelis	5,1%
Išskiria:	krepšinis	8,9%	čempionatas	6,8%	aukos	6,8%	irakas	5,9%
Skirstinys 3	**Dydis:** 297	**VPan:** 0,031	**IPan:** 0,004					
Sieja:	įvairybės	37,1%	britanija	10,8%	kinas	4,9%	prancūzija	3,7%
Išskiria:	įvairybės	17,4%	krepšinis	7,2%	čempionatas	6,4%	britanija	5,1%
Skirstinys 4	**Dydis:** 247	**VPan:** 0,042	**IPan:** 0,004					
Sieja:	ledo ritulys	10,5%	seimas	10,2%	turinas	9,8%	olimpiada	9,7%
Išskiria:	krepšinis	7,0%	ledo ritulys	6,4%	turinas	5,8%	olimpiada	5,6%

AGLOMERATYVUS, 2112 DOKUMENTŲ, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 2112 **Skirstinių:** 10

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,638, **Grynumas:** 0,620

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	182	0,041	0,033	0,003	0,004	0,875	0,374
1	339	0,023	0,012	0,007	0,005	0,890	0,431
2	179	0,265	0,113	0,012	0,006	0,076	0,978
3	297	0,031	0,025	0,004	0,004	0,844	0,475
4	231	0,051	0,031	0,006	0,004	0,484	0,792
5	160	0,073	0,036	0,006	0,004	0,345	0,844
6	65	0,273	0,072	0,007	0,004	0,049	0,985
7	212	0,058	0,035	0,010	0,009	0,748	0,580
8	184	0,051	0,030	0,005	0,003	0,587	0,598
9	258	0,039	0,024	0,005	0,004	0,703	0,624

Pasiskirstymas klasėse

Skirstinys	**Kultūra**	**Sportas**	**Teisėtvarka**	**Ūkis**	**Politika**
0	25	9	22	68	58
1	37	34	41	81	146
2	0	175	0	2	2
3	141	19	59	22	56
4	13	6	9	183	20
5	0	1	8	135	16
6	1	64	0	0	0
7	33	123	5	26	25
8	7	2	60	5	110
9	25	24	41	7	161

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 182	**VPan:** 0,041	**IPan:** 0,003					
Sieja:	seimas	19,2%	eismas	8,9%	sąlygos	8,8%	keliai	8,6%
Išskiria:	seimas	10,3%	krepšinis	6,2%	sąlygos	5,1%	keliai	5,0%
Skirstinys 1	**Dydis:** 339	**VPan:** 0,023	**IPan:** 0,007					
Sieja:	rusija	16,1%	prekyba	6,3%	iranas	5,7%	paukščių gripas	5,1%
Išskiria:	krepšinis	7,2%	čempionatas	5,1%	rusija	5,0%	rezultatai	5,0%

Skirstinys 2	**Dydis:** 179	**VPan:** 0,265	**IPan:** 0,012					
Sieja:	krepšinis	38,3%	čempionatas	24,8%	rezultatai	12,8%	lietuviai	7,6%
Išskiria:	krepšinis	17,9%	čempionatas	9,5%	įvairybės	4,5%	rezultatai	4,1%
Skirstinys 3	**Dydis:** 297	**VPan:** 0,031	**IPan:** 0,004					
Sieja:	įvairybės	37,1%	britanija	10,8%	kinas	4,9%	prancūzija	3,7%
Išskiria:	įvairybės	17,4%	krepšinis	7,2%	čempionatas	6,4%	britanija	5,1%
Skirstinys 4	**Dydis:** 231	**VPan:** 0,051	**IPan:** 0,006					
Sieja:	energetika	22,4%	įmonės	15,4%	nafta	6,4%	transportas	6,2%
Išskiria:	energetika	12,9%	įmonės	8,7%	krepšinis	7,5%	čempionatas	5,9%
Skirstinys 5	**Dydis:** 160	**VPan:** 0,073	**IPan:** 0,006					
Sieja:	finansai	26,3%	ekonomika	22,1%	es	17,1%	zona	2,4%
Išskiria:	finansai	13,7%	ekonomika	11,5%	es	7,6%	krepšinis	7,5%
Skirstinys 6	**Dydis:** 65	**VPan:** 0,273	**IPan:** 0,007					
Sieja:	ledo ritulys	22,3%	turinas	21,8%	olimpiada	21,6%	rezultatai	13,9%
Išskiria:	ledo ritulys	12,5%	turinas	12,1%	olimpiada	11,8%	krepšinis	6,3%
Skirstinys 7	**Dydis:** 212	**VPan:** 0,058	**IPan:** 0,010					
Sieja:	futbolas	22,6%	čempionatas	10,0%	vokietija	9,9%	italija	9,5%
Išskiria:	futbolas	16,2%	italija	6,2%	vokietija	5,3%	rusija	4,2%
Skirstinys 8	**Dydis:** 184	**VPan:** 0,051	**IPan:** 0,005					
Sieja:	irakas	26,6%	aukos	23,4%	bušas	5,7%	avarija	2,7%
Išskiria:	irakas	14,1%	aukos	10,5%	krepšinis	7,4%	čempionatas	6,1%
Skirstinys 9	**Dydis:** 258	**VPan:** 0,039	**IPan:** 0,005					
Sieja:	rinkimai	17,4%	izraelis	9,8%	baltarusija	9,3%	palestiniečiai	8,2%
Išskiria:	rinkimai	9,1%	krepšinis	7,9%	čempionatas	6,0%	izraelis	5,7%

AGLOMERATYVUS, 2112 DOKUMENTŲ, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 2112 **Skirstinių:** 15

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,621, **Grynumas:** 0,620

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	182	0,041	0,033	0,003	0,004	0,875	0,374
1	76	0,221	0,082	0,013	0,006	0,076	0,974
2	145	0,057	0,030	0,006	0,005	0,494	0,786
3	158	0,044	0,022	0,006	0,004	0,826	0,373
4	133	0,076	0,047	0,006	0,005	0,602	0,669
5	160	0,073	0,036	0,006	0,004	0,345	0,844
6	65	0,273	0,072	0,007	0,004	0,049	0,985
7	212	0,058	0,035	0,010	0,009	0,748	0,580
8	184	0,051	0,030	0,005	0,003	0,587	0,598
9	125	0,064	0,037	0,005	0,004	0,726	0,576
10	86	0,154	0,082	0,007	0,004	0,455	0,802
11	103	0,484	0,134	0,020	0,005	0,060	0,981
12	181	0,033	0,017	0,007	0,006	0,863	0,481
13	130	0,077	0,058	0,006	0,004	0,764	0,554
14	167	0,036	0,017	0,004	0,004	0,862	0,413

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	25	9	22	68	58
1	0	74	0	2	0
2	10	5	4	114	12
3	23	10	7	59	59
4	8	2	28	6	89
5	0	1	8	135	16
6	1	64	0	0	0

7	33	123	5	26	25
8	7	2	60	5	110
9	17	22	13	1	72
10	3	1	5	69	8
11	0	101	0	0	2
12	14	24	34	22	87
13	72	3	15	15	25
14	69	16	44	7	31

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 182	**VPan:** 0,041	**IPan:** 0,003					
Sieja:	seimas	19,2%	eismas	8,9%	sąlygos	8,8%	keliai	8,6%
Išskiria:	seimas	10,3%	krepšinis	6,2%	sąlygos	5,1%	keliai	5,0%
Skirstinys 1	**Dydis:** 76	**VPan:** 0,221	**IPan:** 0,013					
Sieja:	eurolyga	34,3%	krepšinis	16,5%	moterys	12,1%	telekomas	4,2%
Išskiria:	eurolyga	22,6%	moterys	6,9%	įvairybės	3,6%	telekomas	2,9%
Skirstinys 2	**Dydis:** 145	**VPan:** 0,057	**IPan:** 0,006					
Sieja:	įmonės	32,3%	transportas	13,9%	automobiliai	6,4%	aviacija	5,7%
Išskiria:	įmonės	18,7%	transportas	6,9%	krepšinis	6,8%	čempionatas	5,2%
Skirstinys 3	**Dydis:** 158	**VPan:** 0,044	**IPan:** 0,006					
Sieja:	iranas	12,6%	paukščių gripas	11,9%	prekyba	11,4%	žemės ūkis	10,5%
Išskiria:	paukščių gripas	7,6%	krepšinis	7,0%	iranas	6,1%	žemės ūkis	6,0%
Skirstinys 4	**Dydis:** 133	**VPan:** 0,076	**IPan:** 0,006					
Sieja:	rinkimai	26,7%	baltarusija	18,0%	ukraina	10,6%	prezidentas	7,0%
Išskiria:	rinkimai	13,4%	baltarusija	9,9%	krepšinis	7,3%	čempionatas	5,9%
Skirstinys 5	**Dydis:** 160	**VPan:** 0,073	**IPan:** 0,006					
Sieja:	finansai	26,3%	ekonomika	22,1%	es	17,1%	zona	2,4%
Išskiria:	finansai	13,7%	ekonomika	11,5%	es	7,6%	krepšinis	7,5%
Skirstinys 6	**Dydis:** 65	**VPan:** 0,273	**IPan:** 0,007					
Sieja:	ledo ritulys	22,3%	turinas	21,8%	olimpiada	21,6%	rezultatai	13,9%
Išskiria:	ledo ritulys	12,5%	turinas	12,1%	olimpiada	11,8%	krepšinis	6,3%
Skirstinys 7	**Dydis:** 212	**VPan:** 0,058	**IPan:** 0,010					
Sieja:	futbolas	22,6%	čempionatas	10,0%	vokietija	9,9%	italija	9,5%
Išskiria:	futbolas	16,2%	italija	6,2%	vokietija	5,3%	rusija	4,2%
Skirstinys 8	**Dydis:** 184	**VPan:** 0,051	**IPan:** 0,005					
Sieja:	irakas	26,6%	aukos	23,4%	bušas	5,7%	avarija	2,7%
Išskiria:	irakas	14,1%	aukos	10,5%	krepšinis	7,4%	čempionatas	6,1%
Skirstinys 9	**Dydis:** 125	**VPan:** 0,064	**IPan:** 0,005					
Sieja:	izraelis	23,4%	palestiniečiai	20,1%	gaisras	4,9%	tenisas	4,5%
Išskiria:	izraelis	12,9%	palestiniečiai	11,3%	krepšinis	6,5%	čempionatas	4,7%
Skirstinys 10	**Dydis:** 86	**VPan:** 0,154	**IPan:** 0,007					
Sieja:	energetika	40,4%	nafta	15,3%	dujos	9,6%	pasaulis	9,2%
Išskiria:	energetika	21,0%	nafta	8,8%	krepšinis	6,7%	čempionatas	5,5%
Skirstinys 11	**Dydis:** 103	**VPan:** 0,484	**IPan:** 0,020					
Sieja:	krepšinis	35,2%	čempionatas	33,6%	rezultatai	17,6%	lietuviai	8,9%
Išskiria:	čempionatas	14,0%	krepšinis	12,9%	rezultatai	6,3%	įvairybės	4,5%
Skirstinys 12	**Dydis:** 181	**VPan:** 0,033	**IPan:** 0,007					
Sieja:	rusija	19,3%	kinija	8,8%	terorizmas	8,5%	sprogimas	5,8%
Išskiria:	krepšinis	5,8%	terorizmas	5,7%	rusija	5,6%	kinija	5,3%
Skirstinys 13	**Dydis:** 130	**VPan:** 0,077	**IPan:** 0,006					
Sieja:	įvairybės	50,8%	britanija	18,3%	didžioji	3,4%	anekdotai	3,0%
Išskiria:	įvairybės	22,6%	britanija	8,8%	krepšinis	7,0%	čempionatas	5,9%
Skirstinys 14	**Dydis:** 167	**VPan:** 0,036	**IPan:** 0,004					
Sieja:	kinas	13,7%	prancūzija	9,3%	pasienis	6,4%	sulaikymas	5,7%
Išskiria:	kinas	8,4%	krepšinis	6,3%	čempionatas	5,6%	prancūzija	4,2%

AGLOMERATYVUS, 2112 DOKUMENTŲ, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 2112 **Skirstinių:** 20

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,571, **Grynumas:** 0,666

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	87	0,077	0,041	0,006	0,004	0,613	0,632
1	76	0,221	0,082	0,013	0,006	0,076	0,974
2	145	0,057	0,030	0,006	0,005	0,494	0,786
3	136	0,054	0,027	0,009	0,008	0,856	0,441
4	133	0,076	0,047	0,006	0,005	0,602	0,669
5	160	0,073	0,036	0,006	0,004	0,345	0,844
6	65	0,273	0,072	0,007	0,004	0,049	0,985
7	160	0,034	0,018	0,003	0,004	0,902	0,363
8	102	0,047	0,022	0,003	0,003	0,789	0,382
9	125	0,064	0,037	0,005	0,004	0,726	0,576
10	86	0,154	0,082	0,007	0,004	0,455	0,802
11	103	0,484	0,134	0,020	0,005	0,060	0,981
12	181	0,033	0,017	0,007	0,006	0,863	0,481
13	130	0,077	0,058	0,006	0,004	0,764	0,554
14	22	0,967	0,010	0,002	0,002	0,000	1,000
15	76	0,183	0,072	0,013	0,010	0,376	0,829
16	71	0,091	0,046	0,005	0,004	0,633	0,690
17	65	0,108	0,049	0,005	0,004	0,735	0,569
18	74	0,111	0,042	0,006	0,004	0,468	0,703
19	110	0,068	0,039	0,006	0,004	0,421	0,827

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	19	0	3	55	10
1	0	74	0	2	0
2	10	5	4	114	12
3	32	60	5	22	17
4	8	2	28	6	89
5	0	1	8	135	16
6	1	64	0	0	0
7	25	9	22	46	58
8	32	1	39	5	25
9	17	22	13	1	72
10	3	1	5	69	8
11	0	101	0	0	2
12	14	24	34	22	87
13	72	3	15	15	25
14	0	0	0	22	0
15	1	63	0	4	8
16	4	10	4	4	49
17	37	15	5	2	6
18	1	0	52	2	19
19	6	2	8	3	91

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 87	**VPan:** 0,077	**IPan:** 0,006					
Sieja:	paukščių gripas	22,6%	prekyba	21,6%	žemės ūkis	18,0%	turkija	3,4%
Išskiria:	paukščių gripas	13,8%	prekyba	10,8%	žemės ūkis	10,0%	krepšinis	7,0%
Skirstinys 1	**Dydis:** 76	**VPan:** 0,221	**IPan:** 0,013					
Sieja:	eurolyga	34,3%	krepšinis	16,5%	moterys	12,1%	telekomas	4,2%
Išskiria:	eurolyga	22,6%	moterys	6,9%	įvairybės	3,6%	telekomas	2,9%

Skirstinys 2	**Dydis:** 145	**VPan:** 0,057	**IPan:** 0,006					
Sieja:	įmonės	32,3%	transportas	13,9%	automobiliai	6,4%	aviacija	5,7%
Išskiria:	įmonės	18,7%	transportas	6,9%	krepšinis	6,8%	čempionatas	5,2%
Skirstinys 3	**Dydis:** 136	**VPan:** 0,054	**IPan:** 0,009					
Sieja:	vokietija	18,4%	laureatai	14,1%	krepšinis	9,9%	europa	7,8%
Išskiria:	laureatai	10,6%	vokietija	10,2%	rusija	4,1%	europa	4,0%
Skirstinys 4	**Dydis:** 133	**VPan:** 0,076	**IPan:** 0,006					
Sieja:	rinkimai	26,7%	baltarusija	18,0%	ukraina	10,6%	prezidentas	7,0%
Išskiria:	rinkimai	13,4%	baltarusija	9,9%	krepšinis	7,3%	čempionatas	5,9%
Skirstinys 5	**Dydis:** 160	**VPan:** 0,073	**IPan:** 0,006					
Sieja:	finansai	26,3%	ekonomika	22,1%	es	17,1%	zona	2,4%
Išskiria:	finansai	13,7%	ekonomika	11,5%	es	7,6%	krepšinis	7,5%
Skirstinys 6	**Dydis:** 65	**VPan:** 0,273	**IPan:** 0,007					
Sieja:	ledo ritulys	22,3%	turinas	21,8%	olimpiada	21,6%	rezultatai	13,9%
Išskiria:	ledo ritulys	12,5%	turinas	12,1%	olimpiada	11,8%	krepšinis	6,3%
Skirstinys 7	**Dydis:** 160	**VPan:** 0,034	**IPan:** 0,003					
Sieja:	seimas	30,0%	paslaugos	9,6%	kaunas	6,9%	paroda	3,4%
Išskiria:	seimas	16,7%	krepšinis	6,2%	paslaugos	5,5%	čempionatas	4,9%
Skirstinys 8	**Dydis:** 102	**VPan:** 0,047	**IPan:** 0,003					
Sieja:	pasienis	13,0%	sulaikymas	11,7%	kontrabanda	7,1%	danija	4,8%
Išskiria:	pasienis	7,2%	sulaikymas	6,5%	krepšinis	6,2%	čempionatas	5,1%
Skirstinys 9	**Dydis:** 125	**VPan:** 0,064	**IPan:** 0,005					
Sieja:	izraelis	23,4%	palestiniečiai	20,1%	gaisras	4,9%	tenisas	4,5%
Išskiria:	izraelis	12,9%	palestiniečiai	11,3%	krepšinis	6,5%	čempionatas	4,7%
Skirstinys 10	**Dydis:** 86	**VPan:** 0,154	**IPan:** 0,007					
Sieja:	energetika	40,4%	nafta	15,3%	dujos	9,6%	pasaulis	9,2%
Išskiria:	energetika	21,0%	nafta	8,8%	krepšinis	6,7%	čempionatas	5,5%
Skirstinys 11	**Dydis:** 103	**VPan:** 0,484	**IPan:** 0,020					
Sieja:	krepšinis	35,2%	čempionatas	33,6%	rezultatai	17,6%	lietuviai	8,9%
Išskiria:	čempionatas	14,0%	krepšinis	12,9%	rezultatai	6,3%	įvairybės	4,5%
Skirstinys 12	**Dydis:** 181	**VPan:** 0,033	**IPan:** 0,007					
Sieja:	rusija	19,3%	kinija	8,8%	terorizmas	8,5%	sprogimas	5,8%
Išskiria:	krepšinis	5,8%	terorizmas	5,7%	rusija	5,6%	kinija	5,3%
Skirstinys 13	**Dydis:** 130	**VPan:** 0,077	**IPan:** 0,006					
Sieja:	įvairybės	50,8%	britanija	18,3%	didžioji	3,4%	anekdotai	3,0%
Išskiria:	įvairybės	22,6%	britanija	8,8%	krepšinis	7,0%	čempionatas	5,9%
Skirstinys 14	**Dydis:** 22	**VPan:** 0,967	**IPan:** 0,002					
Sieja:	sąlygos	25,6%	keliai	25,1%	būklė	24,7%	eismas	22,6%
Išskiria:	sąlygos	13,1%	keliai	12,8%	būklė	12,4%	eismas	11,0%
Skirstinys 15	**Dydis:** 76	**VPan:** 0,183	**IPan:** 0,013					
Sieja:	futbolas	40,7%	italija	20,3%	čempionatas	11,7%	statistika	6,6%
Išskiria:	futbolas	24,9%	italija	12,2%	rezultatai	4,2%	rusija	3,2%
Skirstinys 16	**Dydis:** 71	**VPan:** 0,091	**IPan:** 0,005					
Sieja:	iranas	30,3%	branduolinė programa	9,8%	atomas	7,3%	jt	6,5%
Išskiria:	iranas	14,8%	branduolinė programa	5,6%	krepšinis	5,6%	atomas	4,2%
Skirstinys 17	**Dydis:** 65	**VPan:** 0,108	**IPan:** 0,005					
Sieja:	kinas	30,0%	prancūzija	16,3%	dviračiai	10,2%	lenktynės	8,5%
Išskiria:	kinas	17,5%	prancūzija	7,4%	dviračiai	5,7%	krepšinis	5,6%
Skirstinys 18	**Dydis:** 74	**VPan:** 0,111	**IPan:** 0,006					
Sieja:	aukos	43,6%	avarija	7,8%	lėktuvas	7,6%	katastriša	7,2%
Išskiria:	aukos	19,5%	krepšinis	6,7%	čempionatas	5,4%	lėktuvas	4,6%
Skirstinys 19	**Dydis:** 110	**VPan:** 0,068	**IPan:** 0,006					
Sieja:	irakas	41,4%	bušas	12,0%	vizitas	4,4%	kačinskis	3,9%
Išskiria:	irakas	21,5%	krepšinis	7,0%	bušas	6,8%	čempionatas	5,7%

AGLOMERATYVUS, 4224 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 4224 **Skirstinių:** 5

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,661, **Grynumas:** 0,611

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	542	0,136	0,089	0,006	0,004	0,193	0,935
1	1035	0,015	0,013	0,004	0,004	0,940	0,361
2	852	0,025	0,014	0,005	0,004	0,498	0,790
3	695	0,026	0,020	0,005	0,004	0,695	0,588
4	1097	0,020	0,014	0,004	0,004	0,734	0,561

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	21	5	5	507	4
1	147	114	374	246	154
2	48	673	57	38	36
3	409	177	47	25	37
4	615	104	89	25	264

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 542	**VPan:** 0,136	**IPan:** 0,006					
Sieja:	krepšinis	32,9%	čempionatas	25,9%	rezultatai	9,9%	lietuviai	8,7%
Išskiria:	krepšinis	17,3%	čempionatas	13,7%	lietuviai	4,5%	įvairybės	3,2%
Skirstinys 1	**Dydis:** 1035	**VPan:** 0,015	**IPan:** 0,004					
Sieja:	įvairybės	23,6%	rezultatai	6,7%	ledo ritulys	5,6%	olimpiada	5,6%
Išskiria:	įvairybės	13,3%	krepšinis	8,3%	čempionatas	4,4%	ledo ritulys	4,0%
Skirstinys 2	**Dydis:** 852	**VPan:** 0,025	**IPan:** 0,005					
Sieja:	ekonomika	14,0%	finansai	12,6%	įmonės	7,7%	žemės ūkis	5,0%
Išskiria:	ekonomika	8,2%	finansai	7,4%	čempionatas	6,1%	rezultatai	5,8%
Skirstinys 3	**Dydis:** 695	**VPan:** 0,026	**IPan:** 0,005					
Sieja:	energetika	13,7%	rinkimai	12,8%	baltarusija	10,0%	seimas	5,5%
Išskiria:	energetika	7,5%	krepšinis	7,2%	rinkimai	6,7%	baltarusija	5,7%
Skirstinys 4	**Dydis:** 1097	**VPan:** 0,020	**IPan:** 0,004					
Sieja:	aukos	19,5%	irakas	8,8%	iranas	6,8%	rusija	3,9%
Išskiria:	aukos	11,8%	krepšinis	7,3%	čempionatas	6,2%	rezultatai	5,5%

AGLOMERATYVUS, 4224 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 4224 **Skirstinių:** 10

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,561, **Grynumas:** 0,689

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	153	0,173	0,087	0,004	0,003	0,420	0,686
1	268	0,084	0,053	0,005	0,004	0,513	0,716
2	852	0,025	0,014	0,005	0,004	0,498	0,790
3	548	0,026	0,021	0,004	0,004	0,580	0,737
4	767	0,016	0,014	0,004	0,004	0,862	0,480
5	499	0,041	0,029	0,004	0,003	0,715	0,429
6	147	0,189	0,090	0,006	0,004	0,092	0,966
7	445	0,024	0,020	0,005	0,004	0,655	0,681
8	157	0,162	0,070	0,010	0,008	0,251	0,904
9	385	0,195	0,115	0,008	0,005	0,157	0,948

Pasiskirstymas klasėse					
Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	105	46	0	0	2
1	6	6	6	192	58
2	48	673	57	38	36
3	404	35	47	25	37
4	141	108	368	54	96
5	207	26	47	5	214
6	5	142	0	0	0
7	303	32	42	20	48
8	9	0	2	142	4
9	12	5	3	365	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 153	**VPan:** 0,173	**IPan:** 0,004					
Sieja:	būklė	15,7%	palestiniečiai	13,0%	izraelis	12,8%	sąlygos	11,8%
Išskiria:	būklė	8,7%	sąlygos	6,4%	keliai	6,2%	palestiniečiai	6,2%
Skirstinys 1	**Dydis:** 268	**VPan:** 0,084	**IPan:** 0,005					
Sieja:	rezultatai	18,0%	ledo ritulys	15,0%	olimpiada	14,0%	turinas	13,7%
Išskiria:	ledo ritulys	8,8%	olimpiada	8,3%	turinas	8,1%	krepšinis	5,9%
Skirstinys 2	**Dydis:** 852	**VPan:** 0,025	**IPan:** 0,005					
Sieja:	ekonomika	14,0%	finansai	12,6%	įmonės	7,7%	žemės ūkis	5,0%
Išskiria:	ekonomika	8,2%	finansai	7,4%	čempionatas	6,1%	rezultatai	5,8%
Skirstinys 3	**Dydis:** 548	**VPan:** 0,026	**IPan:** 0,004					
Sieja:	rinkimai	21,1%	baltarusija	14,9%	seimas	9,0%	parlamentas	6,3%
Išskiria:	rinkimai	11,5%	baltarusija	8,4%	krepšinis	6,4%	seimas	5,3%
Skirstinys 4	**Dydis:** 767	**VPan:** 0,016	**IPan:** 0,004					
Sieja:	įvairybės	37,8%	britanija	8,8%	kinas	2,5%	muzika	2,3%
Išskiria:	įvairybės	21,2%	krepšinis	7,2%	čempionatas	5,4%	rezultatai	5,3%
Skirstinys 5	**Dydis:** 499	**VPan:** 0,041	**IPan:** 0,004					
Sieja:	aukos	39,5%	irakas	14,2%	sprogimas	4,5%	pakistanas	2,8%
Išskiria:	aukos	22,4%	irakas	6,6%	krepšinis	6,6%	rezultatai	4,7%
Skirstinys 6	**Dydis:** 147	**VPan:** 0,189	**IPan:** 0,006					
Sieja:	energetika	40,5%	dujos	14,1%	nafta	13,1%	pasaulis	8,6%
Išskiria:	energetika	21,0%	dujos	7,9%	nafta	7,2%	krepšinis	5,8%
Skirstinys 7	**Dydis:** 445	**VPan:** 0,024	**IPan:** 0,005					
Sieja:	iranas	24,5%	branduolinė programa	8,3%	rusija	5,9%	jt	4,4%
Išskiria:	iranas	15,3%	čempionatas	5,9%	branduolinė programa	5,7%	krepšinis	5,6%
Skirstinys 8	**Dydis:** 157	**VPan:** 0,162	**IPan:** 0,010					
Sieja:	futbolas	50,0%	čempionatas	10,9%	ispanija	8,5%	statistika	6,4%
Išskiria:	futbolas	31,3%	ispanija	4,2%	rezultatai	3,3%	krepšinis	2,9%
Skirstinys 9	**Dydis:** 385	**VPan:** 0,195	**IPan:** 0,008					
Sieja:	krepšinis	41,0%	čempionatas	22,7%	rezultatai	12,8%	lietuviai	9,7%
Išskiria:	krepšinis	21,7%	čempionatas	9,4%	lietuviai	4,7%	įvairybės	3,4%

AGLOMERATYVUS, 4224 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 4224 **Skirstinių:** 15

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,541, **Grynumas:** 0,689

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	153	0,173	0,087	0,004	0,003	0,420	0,686
1	326	0,022	0,012	0,004	0,004	0,731	0,620
2	385	0,195	0,115	0,008	0,005	0,157	0,948

3	176	0,104	0,070	0,005	0,004	0,608	0,585
4	178	0,083	0,043	0,005	0,003	0,437	0,820
5	499	0,041	0,029	0,004	0,003	0,715	0,429
6	147	0,189	0,090	0,006	0,004	0,092	0,966
7	445	0,024	0,020	0,005	0,004	0,655	0,681
8	157	0,162	0,070	0,010	0,008	0,251	0,904
9	399	0,046	0,027	0,006	0,004	0,385	0,845
10	222	0,086	0,058	0,005	0,004	0,259	0,910
11	92	0,254	0,108	0,006	0,004	0,089	0,967
12	275	0,039	0,021	0,006	0,004	0,627	0,691
13	337	0,039	0,031	0,005	0,003	0,832	0,439
14	430	0,016	0,010	0,004	0,004	0,844	0,512

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	105	46	0	0	2
1	202	26	45	21	32
2	12	5	3	365	0
3	6	6	3	103	58
4	11	146	11	3	7
5	207	26	47	5	214
6	5	142	0	0	0
7	303	32	42	20	48
8	9	0	2	142	4
9	29	337	15	4	14
10	202	9	2	4	5
11	0	0	3	89	0
12	8	190	31	31	15
13	75	50	148	4	60
14	66	58	220	50	36

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 153	**VPan:** 0,173	**IPan:** 0,004					
Sieja:	būklė	15,7%	palestiniečiai	13,0%	izraelis	12,8%	sąlygos	11,8%
Išskiria:	būklė	8,7%	sąlygos	6,4%	keliai	6,2%	palestiniečiai	6,2%
Skirstinys 1	**Dydis:** 326	**VPan:** 0,022	**IPan:** 0,004					
Sieja:	seimas	19,0%	paulauskas	6,9%	lenkija	6,6%	vyriausybė	3,5%
Išskiria:	seimas	10,9%	krepšinis	5,8%	paulauskas	4,4%	čempionatas	3,8%
Skirstinys 2	**Dydis:** 385	**VPan:** 0,195	**IPan:** 0,008					
Sieja:	krepšinis	41,0%	čempionatas	22,7%	rezultatai	12,8%	lietuviai	9,7%
Išskiria:	krepšinis	21,7%	čempionatas	9,4%	lietuviai	4,7%	įvairybės	3,4%
Skirstinys 3	**Dydis:** 176	**VPan:** 0,104	**IPan:** 0,005					
Sieja:	ledo ritulys	23,1%	rezultatai	17,2%	zubrus	10,1%	tenisas	7,7%
Išskiria:	ledo ritulys	13,2%	zubrus	6,1%	krepšinis	5,5%	tenisas	4,4%
Skirstinys 4	**Dydis:** 178	**VPan:** 0,083	**IPan:** 0,005					
Sieja:	žemės ūkis	24,5%	transportas	21,5%	paukščių gripas	17,2%	aviacija	7,8%
Išskiria:	žemės ūkis	13,9%	transportas	10,7%	paukščių gripas	9,6%	krepšinis	5,4%
Skirstinys 5	**Dydis:** 499	**VPan:** 0,041	**IPan:** 0,004					
Sieja:	aukos	39,5%	irakas	14,2%	sprogimas	4,5%	pakistanas	2,8%
Išskiria:	aukos	22,4%	irakas	6,6%	krepšinis	6,6%	rezultatai	4,7%
Skirstinys 6	**Dydis:** 147	**VPan:** 0,189	**IPan:** 0,006					
Sieja:	energetika	40,5%	dujos	14,1%	nafta	13,1%	pasaulis	8,6%
Išskiria:	energetika	21,0%	dujos	7,9%	nafta	7,2%	krepšinis	5,8%
Skirstinys 7	**Dydis:** 445	**VPan:** 0,024	**IPan:** 0,005					
Sieja:	iranas	24,5%	branduolinė programa	8,3%	rusija	5,9%	jt	4,4%
Išskiria:	iranas	15,3%	čempionatas	5,9%	branduolinė programa	5,7%	krepšinis	5,6%

Skirstinys 8	**Dydis:** 157	**VPan:** 0,162	**IPan:** 0,010					
Sieja:	futbolas	50,0%	čempionatas	10,9%	ispanija	8,5%	statistika	6,4%
Išskiria:	futbolas	31,3%	ispanija	4,2%	rezultatai	3,3%	krepšinis	2,9%
Skirstinys 9	**Dydis:** 399	**VPan:** 0,046	**IPan:** 0,006					
Sieja:	ekonomika	25,6%	finansai	21,8%	es	7,6%	prekyba	6,3%
Išskiria:	ekonomika	14,3%	finansai	12,0%	krepšinis	6,8%	čempionatas	5,6%
Skirstinys 10	**Dydis:** 222	**VPan:** 0,086	**IPan:** 0,005					
Sieja:	rinkimai	36,6%	baltarusija	21,8%	parlamentas	10,0%	prezidentas	7,1%
Išskiria:	rinkimai	19,2%	baltarusija	11,2%	krepšinis	5,9%	parlamentas	5,4%
Skirstinys 11	**Dydis:** 92	**VPan:** 0,254	**IPan:** 0,006					
Sieja:	olimpiada	39,3%	turinas	38,5%	rezultatai	4,2%	biatlonas	3,7%
Išskiria:	olimpiada	21,9%	turinas	21,5%	krepšinis	5,4%	čempionatas	2,5%
Skirstinys 12	**Dydis:** 275	**VPan:** 0,039	**IPan:** 0,006					
Sieja:	įmonės	32,0%	laureatai	8,0%	verslas	6,9%	paslaugos	6,4%
Išskiria:	įmonės	19,1%	čempionatas	5,1%	rezultatai	4,8%	laureatai	4,8%
Skirstinys 13	**Dydis:** 337	**VPan:** 0,039	**IPan:** 0,005					
Sieja:	įvairybės	50,7%	britanija	16,3%	didžioji	2,5%	latvija	2,3%
Išskiria:	įvairybės	25,3%	britanija	7,7%	krepšinis	6,7%	čempionatas	4,8%
Skirstinys 14	**Dydis:** 430	**VPan:** 0,016	**IPan:** 0,004					
Sieja:	dviračiai	7,2%	kinas	6,9%	muzika	6,0%	lenktynės	5,8%
Išskiria:	krepšinis	6,1%	dviračiai	5,0%	čempionatas	4,9%	rezultatai	4,9%

AGLOMERATYVUS, 4224 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 4224 **Skirstinių:** 20

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,500, **Grynumas:** 0,737

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	430	0,016	0,010	0,004	0,004	0,844	0,512
1	326	0,022	0,012	0,004	0,004	0,731	0,620
2	109	0,171	0,071	0,005	0,003	0,114	0,963
3	296	0,050	0,030	0,006	0,004	0,632	0,682
4	130	0,085	0,042	0,004	0,004	0,679	0,446
5	272	0,265	0,151	0,013	0,006	0,172	0,941
6	147	0,189	0,090	0,006	0,004	0,092	0,966
7	445	0,024	0,020	0,005	0,004	0,655	0,681
8	157	0,162	0,070	0,010	0,008	0,251	0,904
9	399	0,046	0,027	0,006	0,004	0,385	0,845
10	222	0,086	0,058	0,005	0,004	0,259	0,910
11	92	0,254	0,108	0,006	0,004	0,089	0,967
12	275	0,039	0,021	0,006	0,004	0,627	0,691
13	337	0,039	0,031	0,005	0,003	0,832	0,439
14	44	0,960	0,023	0,003	0,002	0,000	1,000
15	113	0,222	0,085	0,011	0,006	0,118	0,965
16	203	0,076	0,049	0,006	0,003	0,361	0,842
17	46	0,733	0,097	0,009	0,002	0,000	1,000
18	76	0,219	0,120	0,006	0,003	0,294	0,868
19	102	0,122	0,057	0,005	0,003	0,486	0,784

Pasiskirstymas klasėse

Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	66	58	220	50	36
1	202	26	45	21	32
2	105	2	0	0	2
3	36	23	30	5	202
4	6	6	3	57	58

5	10	4	2	256	0
6	5	142	0	0	0
7	303	32	42	20	48
8	9	0	2	142	4
9	29	337	15	4	14
10	202	9	2	4	5
11	0	0	3	89	0
12	8	190	31	31	15
13	75	50	148	4	60
14	0	44	0	0	0
15	2	1	1	109	0
16	171	3	17	0	12
17	0	0	0	46	0
18	1	66	8	1	0
19	10	80	3	2	7

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 430	**VPan:** 0,016	**IPan:** 0,004					
Sieja:	dviračiai	7,2%	kinas	6,9%	muzika	6,0%	lenktynės	5,8%
Išskiria:	krepšinis	6,1%	dviračiai	5,0%	čempionatas	4,9%	rezultatai	4,9%
Skirstinys 1	**Dydis:** 326	**VPan:** 0,022	**IPan:** 0,004					
Sieja:	seimas	19,0%	paulauskas	6,9%	lenkija	6,6%	vyriausybė	3,5%
Išskiria:	seimas	10,9%	krepšinis	5,8%	paulauskas	4,4%	čempionatas	3,8%
Skirstinys 2	**Dydis:** 109	**VPan:** 0,171	**IPan:** 0,005					
Sieja:	palestiniečiai	25,9%	izraelis	25,5%	hamas	14,4%	šaronas	9,5%
Išskiria:	palestiniečiai	13,1%	izraelis	12,6%	hamas	7,6%	krepšinis	5,3%
Skirstinys 3	**Dydis:** 296	**VPan:** 0,050	**IPan:** 0,006					
Sieja:	aukos	36,1%	sprogimas	6,0%	stichija	4,9%	gaisras	4,9%
Išskiria:	aukos	17,1%	krepšinis	6,4%	rezultatai	4,7%	čempionatas	4,3%
Skirstinys 4	**Dydis:** 130	**VPan:** 0,085	**IPan:** 0,004					
Sieja:	tenisas	17,2%	turnyras	15,0%	sulaikymas	10,5%	kontrabanda	9,8%
Išskiria:	tenisas	9,8%	turnyras	8,6%	kontrabanda	5,7%	sulaikymas	5,6%
Skirstinys 5	**Dydis:** 272	**VPan:** 0,265	**IPan:** 0,013					
Sieja:	krepšinis	33,5%	čempionatas	33,2%	rezultatai	15,0%	lietuviai	10,9%
Išskiria:	čempionatas	16,3%	krepšinis	13,9%	lietuviai	5,1%	rezultatai	4,0%
Skirstinys 6	**Dydis:** 147	**VPan:** 0,189	**IPan:** 0,006					
Sieja:	energetika	40,5%	dujos	14,1%	nafta	13,1%	pasaulis	8,6%
Išskiria:	energetika	21,0%	dujos	7,9%	nafta	7,2%	krepšinis	5,8%
Skirstinys 7	**Dydis:** 445	**VPan:** 0,024	**IPan:** 0,005					
Sieja:	iranas	24,5%	branduolinė programa	8,3%	rusija	5,9%	jt	4,4%
Išskiria:	iranas	15,3%	čempionatas	5,9%	branduolinė programa	5,7%	krepšinis	5,6%
Skirstinys 8	**Dydis:** 157	**VPan:** 0,162	**IPan:** 0,010					
Sieja:	futbolas	50,0%	čempionatas	10,9%	ispanija	8,5%	statistika	6,4%
Išskiria:	futbolas	31,3%	ispanija	4,2%	rezultatai	3,3%	krepšinis	2,9%
Skirstinys 9	**Dydis:** 399	**VPan:** 0,046	**IPan:** 0,006					
Sieja:	ekonomika	25,6%	finansai	21,8%	es	7,6%	prekyba	6,3%
Išskiria:	ekonomika	14,3%	finansai	12,0%	krepšinis	6,8%	čempionatas	5,6%
Skirstinys 10	**Dydis:** 222	**VPan:** 0,086	**IPan:** 0,005					
Sieja:	rinkimai	36,6%	baltarusija	21,8%	parlamentas	10,0%	prezidentas	7,1%
Išskiria:	rinkimai	19,2%	baltarusija	11,2%	krepšinis	5,9%	parlamentas	5,4%
Skirstinys 11	**Dydis:** 92	**VPan:** 0,254	**IPan:** 0,006					
Sieja:	olimpiada	39,3%	turinas	38,5%	rezultatai	4,2%	biatlonas	3,7%
Išskiria:	olimpiada	21,9%	turinas	21,5%	krepšinis	5,4%	čempionatas	2,5%
Skirstinys 12	**Dydis:** 275	**VPan:** 0,039	**IPan:** 0,006					
Sieja:	įmonės	32,0%	laureatai	8,0%	verslas	6,9%	paslaugos	6,4%
Išskiria:	įmonės	19,1%	čempionatas	5,1%	rezultatai	4,8%	laureatai	4,8%

Skirstinys 13	**Dydis:** 337	**VPan:** 0,039	**IPan:** 0,005					
Sieja:	įvairybės	50,7%	britanija	16,3%	didžioji	2,5%	latvija	2,3%
Išskiria:	įvairybės	25,3%	britanija	7,7%	krepšinis	6,7%	čempionatas	4,8%
Skirstinys 14	**Dydis:** 44	**VPan:** 0,960	**IPan:** 0,003					
Sieja:	sąlygos	25,8%	keliai	25,3%	būklė	24,6%	eismas	22,3%
Išskiria:	sąlygos	13,2%	keliai	12,8%	būklė	12,2%	eismas	10,5%
Skirstinys 15	**Dydis:** 113	**VPan:** 0,222	**IPan:** 0,011					
Sieja:	eurolyga	36,2%	krepšinis	27,0%	žalgiris	6,5%	moterys	5,0%
Išskiria:	eurolyga	23,0%	krepšinis	4,9%	čempionatas	4,9%	žalgiris	4,2%
Skirstinys 16	**Dydis:** 203	**VPan:** 0,076	**IPan:** 0,006					
Sieja:	irakas	39,2%	aukos	18,2%	pakistanas	5,3%	išpuolis	4,3%
Išskiria:	irakas	20,1%	krepšinis	6,2%	aukos	5,5%	čempionatas	4,8%
Skirstinys 17	**Dydis:** 46	**VPan:** 0,733	**IPan:** 0,009					
Sieja:	ledo ritulys	47,7%	zubrus	20,8%	rezultatai	19,5%	kasparaitis	11,7%
Išskiria:	ledo ritulys	25,9%	zubrus	11,7%	kasparaitis	6,5%	krepšinis	5,3%
Skirstinys 18	**Dydis:** 76	**VPan:** 0,219	**IPan:** 0,006					
Sieja:	žemės ūkis	51,0%	paukščių gripas	35,8%	pienas	1,8%	prekyba	0,9%
Išskiria:	žemės ūkis	27,4%	paukščių gripas	19,0%	krepšinis	5,1%	čempionatas	4,2%
Skirstinys 19	**Dydis:** 102	**VPan:** 0,122	**IPan:** 0,005					
Sieja:	transportas	43,3%	aviacija	15,3%	automobiliai	8,4%	streikas	7,3%
Išskiria:	transportas	22,2%	aviacija	8,4%	krepšinis	5,1%	automobiliai	4,3%

AGLOMERATYVUS, 8448 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 8448 **Skirstinių:** 5

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,703, **Grynumas:** 0,554

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	1205	0,107	0,079	0,005	0,003	0,370	0,856
1	1411	0,028	0,020	0,005	0,004	0,936	0,299
2	2086	0,016	0,011	0,004	0,004	0,686	0,660
3	1061	0,038	0,026	0,004	0,003	0,443	0,805
4	2685	0,011	0,008	0,004	0,004	0,845	0,372

Pasiskirstymas klasėse

Skirstinys	**Sportas**	**Politika**	**Ūkis**	**Kultūra**	**Teisėtvarka**
0	1032	65	59	28	21
1	422	395	313	110	171
2	131	184	1376	155	240
3	13	854	116	41	37
4	50	1000	270	811	554

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 1205	**VPan:** 0,107	**IPan:** 0,005					
Sieja:	krepšinis	35,6%	čempionatas	23,7%	rezultatai	9,2%	lietuviai	8,2%
Išskiria:	krepšinis	20,5%	čempionatas	12,3%	lietuviai	4,1%	aukos	2,9%
Skirstinys 1	**Dydis:** 1411	**VPan:** 0,028	**IPan:** 0,005					
Sieja:	energetika	13,0%	rezultatai	8,0%	iranas	7,2%	ledo ritulys	7,1%
Išskiria:	krepšinis	8,6%	energetika	7,2%	čempionatas	5,4%	ledo ritulys	4,7%
Skirstinys 2	**Dydis:** 2086	**VPan:** 0,016	**IPan:** 0,004					
Sieja:	ekonomika	12,4%	finansai	11,7%	es	7,3%	prekyba	6,0%
Išskiria:	krepšinis	9,0%	ekonomika	7,7%	finansai	7,5%	rezultatai	4,9%
Skirstinys 3	**Dydis:** 1061	**VPan:** 0,038	**IPan:** 0,004					
Sieja:	rinkimai	15,1%	baltarusija	11,6%	palestiniečiai	9,1%	izraelis	8,5%
Išskiria:	rinkimai	7,2%	krepšinis	6,6%	baltarusija	5,9%	palestiniečiai	5,2%

Skirstinys 4	**Dydis:** 2685	**VPan:** 0,011	**IPan:** 0,004					
Sieja:	įvairybės	16,8%	aukos	14,3%	irakas	8,8%	britanija	5,4%
Išskiria:	įvairybės	9,8%	krepšinis	8,7%	aukos	7,9%	čempionatas	6,8%

AGLOMERATYVUS, 8448 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 8448 **Skirstinių:** 10

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,613, **Grynumas:** 0,633

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	1205	0,107	0,079	0,005	0,003	0,370	0,856
1	393	0,117	0,070	0,004	0,003	0,559	0,682
2	838	0,025	0,018	0,004	0,003	0,878	0,436
3	1870	0,011	0,009	0,004	0,004	0,824	0,419
4	265	0,237	0,092	0,006	0,003	0,077	0,977
5	815	0,039	0,029	0,004	0,003	0,637	0,503
6	668	0,045	0,035	0,004	0,004	0,330	0,877
7	308	0,162	0,087	0,007	0,004	0,334	0,860
8	707	0,047	0,030	0,005	0,003	0,421	0,820
9	1379	0,016	0,010	0,005	0,004	0,780	0,577

Pasiskirstymas klasėse

Skirstinys	**Sportas**	**Politika**	**Ūkis**	**Kultūra**	**Teisėtvarka**
0	1032	65	59	28	21
1	3	268	88	11	23
2	162	365	48	96	167
3	42	590	227	784	227
4	259	3	0	3	0
5	8	410	43	27	327
6	10	586	28	30	14
7	1	27	265	11	4
8	5	63	580	27	32
9	126	121	796	128	208

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 1205	**VPan:** 0,107	**IPan:** 0,005					
Sieja:	krepšinis	35,6%	čempionatas	23,7%	rezultatai	9,2%	lietuviai	8,2%
Išskiria:	krepšinis	20,5%	čempionatas	12,3%	lietuviai	4,1%	aukos	2,9%
Skirstinys 1	**Dydis:** 393	**VPan:** 0,117	**IPan:** 0,004					
Sieja:	palestiniečiai	17,6%	izraelis	16,5%	būklė	12,6%	keliai	9,9%
Išskiria:	palestiniečiai	9,0%	izraelis	8,2%	būklė	7,1%	krepšinis	5,6%
Skirstinys 2	**Dydis:** 838	**VPan:** 0,025	**IPan:** 0,004					
Sieja:	iranas	20,1%	branduolinė programa	8,5%	tenisas	4,0%	jt	3,5%
Išskiria:	iranas	10,7%	krepšinis	6,8%	branduolinė programa	5,3%	čempionatas	4,7%
Skirstinys 3	**Dydis:** 1870	**VPan:** 0,011	**IPan:** 0,004					
Sieja:	įvairybės	31,3%	britanija	8,8%	rusija	3,1%	kinas	2,5%
Išskiria:	įvairybės	19,9%	krepšinis	7,7%	rezultatai	6,3%	čempionatas	5,9%
Skirstinys 4	**Dydis:** 265	**VPan:** 0,237	**IPan:** 0,006					
Sieja:	ledo ritulys	23,4%	rezultatai	19,1%	turinas	18,6%	olimpiada	18,4%
Išskiria:	ledo ritulys	12,9%	turinas	10,4%	olimpiada	10,2%	krepšinis	5,6%
Skirstinys 5	**Dydis:** 815	**VPan:** 0,039	**IPan:** 0,004					
Sieja:	aukos	35,6%	irakas	18,4%	sprogimas	6,2%	išpuolis	2,8%
Išskiria:	aukos	18,4%	irakas	9,5%	krepšinis	6,6%	čempionatas	5,1%
Skirstinys 6	**Dydis:** 668	**VPan:** 0,045	**IPan:** 0,004					

Sieja:	rinkimai	25,3%	baltarusija	24,1%	seimas	10,2%	parlamentas	6,1%
Išskiria:	baltarusija	13,2%	rinkimai	12,3%	krepšinis	6,2%	seimas	5,3%
Skirstinys 7	**Dydis:** 308	**VPan:** 0,162	**IPan:** 0,007					
Sieja:	energetika	39,7%	dujos	13,4%	nafta	12,1%	kaina	7,2%
Išskiria:	energetika	20,6%	dujos	7,5%	nafta	6,5%	krepšinis	6,1%
Skirstinys 8	**Dydis:** 707	**VPan:** 0,047	**IPan:** 0,005					
Sieja:	ekonomika	27,3%	finansai	23,6%	es	10,9%	statistika	4,1%
Išskiria:	ekonomika	14,9%	finansai	12,8%	krepšinis	6,9%	čempionatas	5,3%
Skirstinys 9	**Dydis:** 1379	**VPan:** 0,016	**IPan:** 0,005					
Sieja:	įmonės	11,6%	prekyba	8,4%	transportas	7,4%	žemės ūkis	5,3%
Išskiria:	krepšinis	8,5%	įmonės	7,4%	prekyba	5,0%	rezultatai	4,3%

AGLOMERATYVUS, 8448 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 8448 **Skirstinių:** 15

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,568, **Grynumas:** 0,643

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	307	0,112	0,064	0,004	0,003	0,321	0,873
1	918	0,017	0,012	0,005	0,004	0,914	0,412
2	501	0,076	0,049	0,010	0,007	0,539	0,750
3	1870	0,011	0,009	0,004	0,004	0,824	0,419
4	265	0,237	0,092	0,006	0,003	0,077	0,977
5	815	0,039	0,029	0,004	0,003	0,637	0,503
6	668	0,045	0,035	0,004	0,004	0,330	0,877
7	308	0,162	0,087	0,007	0,004	0,334	0,860
8	707	0,047	0,030	0,005	0,003	0,421	0,820
9	398	0,119	0,062	0,009	0,005	0,274	0,882
10	86	0,935	0,091	0,003	0,002	0,039	0,988
11	461	0,047	0,025	0,005	0,003	0,266	0,907
12	306	0,507	0,150	0,018	0,006	0,014	0,997
13	173	0,200	0,099	0,006	0,003	0,198	0,931
14	665	0,022	0,012	0,004	0,003	0,932	0,307

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	3	268	3	11	22
1	120	101	378	119	200
2	376	64	22	19	20
3	42	590	227	784	227
4	259	3	0	3	0
5	8	410	43	27	327
6	10	586	28	30	14
7	1	27	265	11	4
8	5	63	580	27	32
9	351	1	37	8	1
10	0	0	85	0	1
11	6	20	418	9	8
12	305	0	0	1	0
13	0	161	7	3	2
14	162	204	41	93	165

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 307	**VPan:** 0,112	**IPan:** 0,004					
Sieja:	palestiniečiai	30,1%	izraelis	28,2%	hamas	11,4%	šaronas	7,1%

Išskiria:	palestiniečiai	16,1%	izraelis	14,8%	hamas	6,0%	krepšinis	5,6%
Skirstinys 1 **Dydis:** 918 **VPan:** 0,017 **IPan:** 0,005								
Sieja:	prekyba	11,4%	žemės ūkis	10,3%	paukščių gripas	7,2%	stichija	5,8%
Išskiria:	krepšinis	8,1%	žemės ūkis	7,1%	prekyba	6,7%	paukščių gripas	5,0%
Skirstinys 2 **Dydis:** 501 **VPan:** 0,076 **IPan:** 0,010								
Sieja:	futbolas	26,4%	čempionatas	25,2%	italija	7,0%	ispanija	6,9%
Išskiria:	futbolas	17,7%	čempionatas	7,5%	ispanija	4,0%	italija	3,8%
Skirstinys 3 **Dydis:** 1870 **VPan:** 0,011 **IPan:** 0,004								
Sieja:	įvairybės	31,3%	britanija	8,8%	rusija	3,1%	kinas	2,5%
Išskiria:	įvairybės	19,9%	krepšinis	7,7%	rezultatai	6,3%	čempionatas	5,9%
Skirstinys 4 **Dydis:** 265 **VPan:** 0,237 **IPan:** 0,006								
Sieja:	ledo ritulys	23,4%	rezultatai	19,1%	turinas	18,6%	olimpiada	18,4%
Išskiria:	ledo ritulys	12,9%	turinas	10,4%	olimpiada	10,2%	krepšinis	5,6%
Skirstinys 5 **Dydis:** 815 **VPan:** 0,039 **IPan:** 0,004								
Sieja:	aukos	35,6%	irakas	18,4%	sprogimas	6,2%	išpuolis	2,8%
Išskiria:	aukos	18,4%	irakas	9,5%	krepšinis	6,6%	čempionatas	5,1%
Skirstinys 6 **Dydis:** 668 **VPan:** 0,045 **IPan:** 0,004								
Sieja:	rinkimai	25,3%	baltarusija	24,1%	seimas	10,2%	parlamentas	6,1%
Išskiria:	baltarusija	13,2%	rinkimai	12,3%	krepšinis	6,2%	seimas	5,3%
Skirstinys 7 **Dydis:** 308 **VPan:** 0,162 **IPan:** 0,007								
Sieja:	energetika	39,7%	dujos	13,4%	nafta	12,1%	kaina	7,2%
Išskiria:	energetika	20,6%	dujos	7,5%	nafta	6,5%	krepšinis	6,1%
Skirstinys 8 **Dydis:** 707 **VPan:** 0,047 **IPan:** 0,005								
Sieja:	ekonomika	27,3%	finansai	23,6%	es	10,9%	statistika	4,1%
Išskiria:	ekonomika	14,9%	finansai	12,8%	krepšinis	6,9%	čempionatas	5,3%
Skirstinys 9 **Dydis:** 398 **VPan:** 0,119 **IPan:** 0,009								
Sieja:	krepšinis	30,2%	eurolyga	24,8%	taurė	6,7%	statistika	3,9%
Išskiria:	eurolyga	17,1%	krepšinis	7,4%	čempionatas	5,4%	taurė	4,0%
Skirstinys 10 **Dydis:** 86 **VPan:** 0,935 **IPan:** 0,003								
Sieja:	keliai	26,0%	sąlygos	25,3%	būklė	24,1%	eismas	22,2%
Išskiria:	keliai	13,3%	sąlygos	12,9%	būklė	11,9%	eismas	10,5%
Skirstinys 11 **Dydis:** 461 **VPan:** 0,047 **IPan:** 0,005								
Sieja:	įmonės	27,7%	transportas	18,9%	aviacija	8,7%	automobiliai	6,0%
Išskiria:	įmonės	15,3%	transportas	9,9%	krepšinis	6,2%	aviacija	5,5%
Skirstinys 12 **Dydis:** 306 **VPan:** 0,507 **IPan:** 0,018								
Sieja:	krepšinis	37,8%	čempionatas	31,2%	rezultatai	18,3%	lietuviai	10,2%
Išskiria:	krepšinis	14,9%	čempionatas	13,0%	rezultatai	5,2%	lietuviai	4,0%
Skirstinys 13 **Dydis:** 173 **VPan:** 0,200 **IPan:** 0,006								
Sieja:	iranas	47,9%	branduolinė programa	24,6%	atomas	7,4%	derybos	5,0%
Išskiria:	iranas	23,7%	branduolinė programa	13,7%	krepšinis	5,6%	čempionatas	4,3%
Skirstinys 14 **Dydis:** 665 **VPan:** 0,022 **IPan:** 0,004								
Sieja:	tenisas	7,1%	sulaikymas	6,0%	turnyras	5,8%	dviračiai	5,6%
Išskiria:	krepšinis	6,4%	tenisas	4,6%	čempionatas	4,2%	turnyras	3,7%

AGLOMERATYVUS, 8448 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: aglomeratyvus **Dokumentų:** 8448 **Skirstinių:** 20

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,541, **Grynumas:** 0,664

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	307	0,112	0,064	0,004	0,003	0,321	0,873
1	1402	0,010	0,007	0,004	0,004	0,845	0,399
2	501	0,076	0,049	0,010	0,007	0,539	0,750
3	166	0,287	0,104	0,006	0,004	0,041	0,988

4	633	0,018	0,013	0,005	0,004	0,982	0,280
5	815	0,039	0,029	0,004	0,003	0,637	0,503
6	396	0,033	0,022	0,003	0,003	0,889	0,361
7	308	0,162	0,087	0,007	0,004	0,334	0,860
8	707	0,047	0,030	0,005	0,003	0,421	0,820
9	398	0,119	0,062	0,009	0,005	0,274	0,882
10	86	0,935	0,091	0,003	0,002	0,039	0,988
11	461	0,047	0,025	0,005	0,003	0,266	0,907
12	306	0,507	0,150	0,018	0,006	0,014	0,997
13	173	0,200	0,099	0,006	0,003	0,198	0,931
14	99	0,659	0,171	0,009	0,003	0,119	0,960
15	468	0,038	0,029	0,004	0,003	0,507	0,782
16	285	0,070	0,040	0,006	0,004	0,552	0,705
17	269	0,057	0,026	0,004	0,003	0,840	0,357
18	212	0,178	0,089	0,006	0,003	0,170	0,948
19	456	0,037	0,024	0,004	0,004	0,390	0,844

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	3	268	3	11	22
1	29	560	197	418	198
2	376	64	22	19	20
3	164	0	0	2	0
4	117	92	177	102	145
5	8	410	43	27	327
6	66	124	36	27	143
7	1	27	265	11	4
8	5	63	580	27	32
9	351	1	37	8	1
10	0	0	85	0	1
11	6	20	418	9	8
12	305	0	0	1	0
13	0	161	7	3	2
14	95	3	0	1	0
15	13	30	30	366	29
16	3	9	201	17	55
17	96	80	5	66	22
18	2	201	2	3	4
19	8	385	26	27	10

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 307	**VPan:** 0,112	**IPan:** 0,004					
Sieja:	palestiniečiai	30,1%	izraelis	28,2%	hamas	11,4%	šaronas	7,1%
Išskiria:	palestiniečiai	16,1%	izraelis	14,8%	hamas	6,0%	krepšinis	5,6%
Skirstinys 1	**Dydis:** 1402	**VPan:** 0,010	**IPan:** 0,004					
Sieja:	britanija	12,7%	įvairybės	8,3%	rusija	5,0%	vizitas	3,7%
Išskiria:	krepšinis	8,7%	britanija	8,0%	čempionatas	6,9%	rezultatai	6,7%
Skirstinys 2	**Dydis:** 501	**VPan:** 0,076	**IPan:** 0,010					
Sieja:	futbolas	26,4%	čempionatas	25,2%	italija	7,0%	ispanija	6,9%
Išskiria:	futbolas	17,7%	čempionatas	7,5%	ispanija	4,0%	italija	3,8%
Skirstinys 3	**Dydis:** 166	**VPan:** 0,287	**IPan:** 0,006					
Sieja:	turinas	39,3%	olimpiada	38,8%	rezultatai	5,3%	slidinėjimas	3,4%
Išskiria:	turinas	21,8%	olimpiada	21,4%	krepšinis	5,4%	čempionatas	2,9%
Skirstinys 4	**Dydis:** 633	**VPan:** 0,018	**IPan:** 0,005					
Sieja:	stichija	12,0%	drebėjimas	6,2%	žemės	5,7%	latvija	5,2%
Išskiria:	stichija	9,2%	krepšinis	7,2%	drebėjimas	4,7%	žemės	4,4%
Skirstinys 5	**Dydis:** 815	**VPan:** 0,039	**IPan:** 0,004					
Sieja:	aukos	35,6%	irakas	18,4%	sprogimas	6,2%	išpuolis	2,8%

Išskiria:	aukos	18,4%	irakas	9,5%	krepšinis	6,6%	čempionatas	5,1%
Skirstinys 6 **Dydis:** 396 **VPan:** 0,033 **IPan:** 0,003								
Sieja:	sulaikymas	11,0%	dviračiai	10,6%	pasienis	9,0%	kontrabanda	8,6%
Išskiria:	sulaikymas	6,4%	dviračiai	6,4%	krepšinis	6,0%	pasienis	5,3%
Skirstinys 7 **Dydis:** 308 **VPan:** 0,162 **IPan:** 0,007								
Sieja:	energetika	39,7%	dujos	13,4%	nafta	12,1%	kaina	7,2%
Išskiria:	energetika	20,6%	dujos	7,5%	nafta	6,5%	krepšinis	6,1%
Skirstinys 8 **Dydis:** 707 **VPan:** 0,047 **IPan:** 0,005								
Sieja:	ekonomika	27,3%	finansai	23,6%	es	10,9%	statistika	4,1%
Išskiria:	ekonomika	14,9%	finansai	12,8%	krepšinis	6,9%	čempionatas	5,3%
Skirstinys 9 **Dydis:** 398 **VPan:** 0,119 **IPan:** 0,009								
Sieja:	krepšinis	30,2%	eurolyga	24,8%	taurė	6,7%	statistika	3,9%
Išskiria:	eurolyga	17,1%	krepšinis	7,4%	čempionatas	5,4%	taurė	4,0%
Skirstinys 10 **Dydis:** 86 **VPan:** 0,935 **IPan:** 0,003								
Sieja:	keliai	26,0%	sąlygos	25,3%	būklė	24,1%	eismas	22,2%
Išskiria:	keliai	13,3%	sąlygos	12,9%	būklė	11,9%	eismas	10,5%
Skirstinys 11 **Dydis:** 461 **VPan:** 0,047 **IPan:** 0,005								
Sieja:	įmonės	27,7%	transportas	18,9%	aviacija	8,7%	automobiliai	6,0%
Išskiria:	įmonės	15,3%	transportas	9,9%	krepšinis	6,2%	aviacija	5,5%
Skirstinys 12 **Dydis:** 306 **VPan:** 0,507 **IPan:** 0,018								
Sieja:	krepšinis	37,8%	čempionatas	31,2%	rezultatai	18,3%	lietuviai	10,2%
Išskiria:	krepšinis	14,9%	čempionatas	13,0%	rezultatai	5,2%	lietuviai	4,0%
Skirstinys 13 **Dydis:** 173 **VPan:** 0,200 **IPan:** 0,006								
Sieja:	iranas	47,9%	branduolinė programa	24,6%	atomas	7,4%	derybos	5,0%
Išskiria:	iranas	23,7%	branduolinė programa	13,7%	krepšinis	5,6%	čempionatas	4,3%
Skirstinys 14 **Dydis:** 99 **VPan:** 0,659 **IPan:** 0,009								
Sieja:	ledo ritulys	49,3%	rezultatai	20,1%	zubrus	15,3%	kasparaitis	15,0%
Išskiria:	ledo ritulys	26,4%	zubrus	8,6%	kasparaitis	8,3%	krepšinis	5,3%
Skirstinys 15 **Dydis:** 468 **VPan:** 0,038 **IPan:** 0,004								
Sieja:	įvairybės	54,4%	kinas	10,1%	muzika	6,4%	anekdotai	2,9%
Išskiria:	įvairybės	25,2%	kinas	6,0%	krepšinis	5,1%	rezultatai	4,2%
Skirstinys 16 **Dydis:** 285 **VPan:** 0,070 **IPan:** 0,006								
Sieja:	prekyba	25,1%	žemės ūkis	24,4%	paukščių gripas	14,2%	turgus	2,9%
Išskiria:	žemės ūkis	14,1%	prekyba	13,0%	paukščių gripas	7,9%	krepšinis	6,3%
Skirstinys 17 **Dydis:** 269 **VPan:** 0,057 **IPan:** 0,004								
Sieja:	tenisas	15,6%	turnyras	13,9%	serbija	6,1%	popiežius	5,6%
Išskiria:	tenisas	9,0%	turnyras	8,0%	krepšinis	5,4%	čempionatas	3,7%
Skirstinys 18 **Dydis:** 212 **VPan:** 0,178 **IPan:** 0,006								
Sieja:	baltarusija	49,0%	rinkimai	25,2%	lukašenka	6,7%	prezidentas	6,1%
Išskiria:	baltarusija	25,4%	rinkimai	9,8%	krepšinis	5,7%	čempionatas	4,3%
Skirstinys 19 **Dydis:** 456 **VPan:** 0,037 **IPan:** 0,004								
Sieja:	seimas	22,0%	parlamentas	15,0%	rinkimai	9,1%	ukraina	8,7%
Išskiria:	seimas	12,4%	parlamentas	8,8%	krepšinis	6,1%	čempionatas	4,8%

PRIEDAS B – EKSPERIMENTŲ ŽURNALAS

Šiame priede pateikiamas individualiai suklasifikuotų dokumentų kolekcijų klasterizavimo K-means algoritmu eksperimento žurnalas. Klasterizavimo metu dokumentų reprezentacijai naudotos individualiai vartotojų priskirtos žymės.

K-MEANS, 1056 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 1056 **Skirstinių:** 5

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,413, **Grynumas:** 0,746

Skirstinys	Dydis[29]	VPan[30]	VNuok[31]	IPan[32]	INuok[33]	Entropija	Grynumas
0	188	0,126	0,082	0,002	0,003	0,088	0,968
1	189	0,050	0,033	0,003	0,004	0,726	0,545
2	193	0,041	0,022	0,003	0,003	0,331	0,865
3	255	0,039	0,018	0,004	0,004	0,362	0,843
4	223	0,038	0,024	0,003	0,003	0,549	0,516

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	6	182
1	9	15	54	103	8
2	167	9	2	14	1
3	17	0	215	19	4
4	115	98	1	5	4

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 188	**VPan:** 0,126	**IPan:** 0,002					
Sieja:	krepšinis	30,0%	čempionatas	21,5%	rezultatai	18,0%	lietuviai	7,6%
Išskiria:	krepšinis	15,9%	čempionatas	11,4%	rezultatai	8,7%	įvairybės	4,1%
Skirstinys 1	**Dydis:** 189	**VPan:** 0,050	**IPan:** 0,003					
Sieja:	įvairybės	41,4%	transportas	12,2%	britanija	6,4%	vokietija	5,7%
Išskiria:	įvairybės	23,1%	transportas	6,8%	krepšinis	6,1%	čempionatas	4,4%
Skirstinys 2	**Dydis:** 193	**VPan:** 0,041	**IPan:** 0,003					
Sieja:	rinkimai	16,1%	seimas	11,0%	palestiniečiai	9,8%	baltarusija	9,0%
Išskiria:	rinkimai	8,7%	krepšinis	6,1%	seimas	5,9%	palestiniečiai	5,5%
Skirstinys 3	**Dydis:** 255	**VPan:** 0,039	**IPan:** 0,004					
Sieja:	energetika	11,8%	ekonomika	11,6%	finansai	9,7%	es	8,1%
Išskiria:	krepšinis	7,2%	energetika	6,9%	ekonomika	6,6%	finansai	5,5%
Skirstinys 4	**Dydis:** 223	**VPan:** 0,038	**IPan:** 0,003					
Sieja:	aukos	30,6%	irakas	12,1%	rusija	8,0%	iranas	4,7%
Išskiria:	aukos	17,6%	krepšinis	6,5%	irakas	6,4%	čempionatas	4,7%

29 Skirstiniui priskirtų dokumentų skaičius.

30 Vidutinis panašumas tarp skirstinyje esančių dokumentų (**V**idinis **Pan**ašumas).

31 Skirstinyje esančių dokumentų panašumų standartinis nuokrypis (**V**idinis **Nuok**rypis).

32 Vidutinis panašumas tarp skirstinyje esančių dokumentų ir likusių kolekcijos dokumentų (**I**šorinis **Pan**ašumas).

33 Skirstinyje esančių dokumentų panašumų su likusiais kolekcijos dokumentais standartinis nuokrypis (**I**šorinis **Nuok**rypis).

K-MEANS, 1056 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 1056 **Skirstinių:** 10

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,414, **Grynumas:** 0,756

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	48	0,206	0,088	0,006	0,004	0,178	0,917
1	135	0,192	0,104	0,004	0,004	0,048	0,985
2	87	0,122	0,077	0,003	0,004	0,275	0,897
3	92	0,096	0,049	0,006	0,004	0,638	0,663
4	107	0,080	0,032	0,004	0,004	0,327	0,850
5	98	0,073	0,042	0,004	0,004	0,466	0,776
6	136	0,070	0,039	0,003	0,003	0,428	0,544
7	139	0,059	0,027	0,004	0,003	0,479	0,799
8	110	0,055	0,032	0,003	0,003	0,521	0,709
9	96	0,049	0,028	0,002	0,003	0,760	0,479

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	4	44
1	0	0	0	2	133
2	2	4	0	78	3
3	16	9	61	3	3
4	8	0	91	0	8
5	76	3	2	15	2
6	74	62	0	0	0
7	8	6	111	10	4
8	78	22	3	7	0
9	46	16	4	28	2

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 48	**VPan:** 0,206	**IPan:** 0,006					
Sieja:	rezultatai	25,1%	ledo ritulys	17,5%	olimpiada	16,3%	turinas	15,6%
Išskiria:	ledo ritulys	9,8%	olimpiada	9,0%	turinas	8,8%	rezultatai	6,4%
Skirstinys 1	**Dydis:** 135	**VPan:** 0,192	**IPan:** 0,004					
Sieja:	krepšinis	37,2%	čempionatas	27,2%	lietuviai	9,6%	rezultatai	8,6%
Išskiria:	krepšinis	20,1%	čempionatas	14,9%	lietuviai	5,3%	įvairybės	3,9%
Skirstinys 2	**Dydis:** 87	**VPan:** 0,122	**IPan:** 0,003					
Sieja:	įvairybės	77,3%	muzika	3,8%	kalėdos	1,6%	vatikanas	1,4%
Išskiria:	įvairybės	41,6%	krepšinis	5,4%	čempionatas	3,9%	rezultatai	3,7%
Skirstinys 3	**Dydis:** 92	**VPan:** 0,096	**IPan:** 0,006					
Sieja:	energetika	34,9%	rusija	25,9%	nafta	7,5%	dujos	5,9%
Išskiria:	energetika	20,5%	rusija	9,1%	krepšinis	6,1%	čempionatas	4,4%
Skirstinys 4	**Dydis:** 107	**VPan:** 0,080	**IPan:** 0,004					
Sieja:	ekonomika	33,5%	finansai	24,3%	es	13,9%	dviračiai	1,7%
Išskiria:	ekonomika	19,2%	finansai	13,5%	es	5,9%	krepšinis	5,9%
Skirstinys 5	**Dydis:** 98	**VPan:** 0,073	**IPan:** 0,004					
Sieja:	rinkimai	28,3%	baltarusija	19,4%	ukraina	7,6%	prezidentas	6,8%
Išskiria:	rinkimai	15,0%	baltarusija	11,0%	krepšinis	5,1%	čempionatas	4,1%
Skirstinys 6	**Dydis:** 136	**VPan:** 0,070	**IPan:** 0,003					
Sieja:	aukos	42,0%	irakas	17,4%	sprogimas	4,6%	išpuolis	3,9%
Išskiria:	aukos	23,1%	irakas	9,1%	krepšinis	5,9%	čempionatas	4,2%
Skirstinys 7	**Dydis:** 139	**VPan:** 0,059	**IPan:** 0,004					
Sieja:	transportas	18,9%	įmonės	16,1%	vokietija	7,0%	žemės ūkis	6,9%
Išskiria:	transportas	11,0%	įmonės	8,2%	krepšinis	6,1%	čempionatas	4,4%
Skirstinys 8	**Dydis:** 110	**VPan:** 0,055	**IPan:** 0,003					
Sieja:	palestiniečiai	21,0%	iranas	13,1%	hamas	12,1%	izraelis	9,8%

Išskiria:	palestiniečiai	11,6%	hamas	6,8%	iranas	6,7%	krepšinis	5,5%
Skirstinys 9	**Dydis:** 96	**VPan:** 0,049	**IPan:** 0,002					
Sieja:	seimas	36,6%	britanija	9,4%	kinas	7,2%	tyrimas	4,8%
Išskiria:	seimas	19,6%	krepšinis	5,2%	britanija	4,1%	kinas	3,8%

K-MEANS, 1056 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 1056 **Skirstinių:** 15

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,374, **Grynumas:** 0,780

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	89	0,310	0,135	0,009	0,005	0,038	0,989
1	73	0,167	0,091	0,004	0,004	0,486	0,767
2	51	0,165	0,063	0,004	0,004	0,249	0,863
3	48	0,162	0,071	0,009	0,006	0,360	0,792
4	51	0,148	0,067	0,004	0,003	0,221	0,922
5	62	0,136	0,065	0,005	0,004	0,277	0,903
6	55	0,107	0,051	0,002	0,003	0,291	0,873
7	94	0,107	0,050	0,005	0,004	0,417	0,606
8	55	0,104	0,062	0,004	0,004	0,476	0,545
9	92	0,096	0,034	0,005	0,004	0,272	0,880
10	89	0,089	0,048	0,005	0,004	0,641	0,640
11	87	0,087	0,037	0,006	0,004	0,363	0,839
12	61	0,084	0,046	0,005	0,006	0,714	0,590
13	54	0,067	0,028	0,001	0,002	0,551	0,556
14	87	0,060	0,029	0,004	0,003	0,308	0,874

Pasiskirstymas klasėse

Skirstinys	**Politika**	**Teisėtvarka**	**Ūkis**	**Kultūra**	**Sportas**
0	0	0	0	1	88
1	6	8	0	56	3
2	0	7	44	0	0
3	0	1	0	9	38
4	47	2	1	1	0
5	56	2	1	1	2
6	48	3	0	4	0
7	37	57	0	0	0
8	0	0	1	24	30
9	9	0	81	1	1
10	15	13	57	3	1
11	3	2	73	9	0
12	7	3	13	2	36
13	4	20	0	30	0
14	76	4	1	6	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 89	**VPan:** 0,310	**IPan:** 0,009					
Sieja:	krepšinis	29,7%	čempionatas	28,1%	rezultatai	22,9%	lietuviai	10,7%
Išskiria:	čempionatas	13,9%	krepšinis	12,1%	rezultatai	10,1%	lietuviai	5,5%
Skirstinys 1	**Dydis:** 73	**VPan:** 0,167	**IPan:** 0,004					
Sieja:	įvairybės	80,3%	britanija	7,7%	kalėdos	1,6%	vokietija	1,1%
Išskiria:	įvairybės	43,2%	krepšinis	5,4%	čempionatas	3,8%	rezultatai	3,7%
Skirstinys 2	**Dydis:** 51	**VPan:** 0,165	**IPan:** 0,004					
Sieja:	transportas	50,2%	aviacija	10,1%	aplinkosauga	5,3%	sąlygos	5,0%
Išskiria:	transportas	27,2%	aviacija	5,4%	krepšinis	5,1%	čempionatas	3,7%

Skirstinys 3	**Dydis:** 48	**VPan:** 0,162	**IPan:** 0,009					
Sieja:	eurolyga	29,7%	krepšinis	22,5%	moterys	7,3%	laureatai	6,5%
Išskiria:	eurolyga	17,9%	rezultatai	4,2%	laureatai	4,0%	moterys	3,8%
Skirstinys 4	**Dydis:** 51	**VPan:** 0,148	**IPan:** 0,004					
Sieja:	palestiniečiai	36,2%	hamas	20,8%	izraelis	16,8%	vyriausybė	4,3%
Išskiria:	palestiniečiai	19,5%	hamas	11,3%	izraelis	9,0%	krepšinis	5,1%
Skirstinys 5	**Dydis:** 62	**VPan:** 0,136	**IPan:** 0,005					
Sieja:	rinkimai	38,1%	baltarusija	24,8%	ukraina	9,2%	parlamentas	5,6%
Išskiria:	rinkimai	20,4%	baltarusija	13,9%	krepšinis	4,9%	ukraina	4,1%
Skirstinys 6	**Dydis:** 55	**VPan:** 0,107	**IPan:** 0,002					
Sieja:	seimas	51,3%	adamkus	9,0%	liberalai	5,8%	paulauskas	2,9%
Išskiria:	seimas	26,7%	krepšinis	5,0%	adamkus	4,8%	čempionatas	3,6%
Skirstinys 7	**Dydis:** 94	**VPan:** 0,107	**IPan:** 0,005					
Sieja:	aukos	57,2%	sprogimas	6,3%	gaisras	4,6%	incidentas	3,9%
Išskiria:	aukos	31,9%	krepšinis	5,8%	čempionatas	4,1%	rezultatai	3,6%
Skirstinys 8	**Dydis:** 55	**VPan:** 0,104	**IPan:** 0,004					
Sieja:	olimpiada	24,6%	turinas	23,5%	kinas	10,4%	rezultatai	5,7%
Išskiria:	olimpiada	13,6%	turinas	13,3%	kinas	5,7%	krepšinis	5,3%
Skirstinys 9	**Dydis:** 92	**VPan:** 0,096	**IPan:** 0,005					
Sieja:	ekonomika	34,4%	finansai	27,5%	es	15,6%	mokesčiai	1,7%
Išskiria:	ekonomika	19,2%	finansai	15,2%	es	6,7%	krepšinis	5,8%
Skirstinys 10	**Dydis:** 89	**VPan:** 0,089	**IPan:** 0,005					
Sieja:	energetika	38,0%	rusija	14,9%	nafta	8,6%	dujos	6,8%
Išskiria:	energetika	21,8%	krepšinis	5,9%	nafta	5,0%	čempionatas	4,2%
Skirstinys 11	**Dydis:** 87	**VPan:** 0,087	**IPan:** 0,006					
Sieja:	įmonės	23,3%	prekyba	18,8%	žemės ūkis	14,5%	paukščių gripas	12,7%
Išskiria:	įmonės	11,9%	prekyba	11,2%	žemės ūkis	8,5%	paukščių gripas	7,4%
Skirstinys 12	**Dydis:** 61	**VPan:** 0,084	**IPan:** 0,005					
Sieja:	futbolas	40,8%	italija	8,1%	dviračiai	5,1%	čempionatas	5,0%
Išskiria:	futbolas	24,2%	krepšinis	5,7%	rusija	3,3%	rezultatai	3,3%
Skirstinys 13	**Dydis:** 54	**VPan:** 0,067	**IPan:** 0,001					
Sieja:	muzika	17,8%	sulaikymas	10,0%	pasienis	8,3%	koncertas	6,2%
Išskiria:	muzika	9,0%	sulaikymas	5,2%	krepšinis	4,8%	pasienis	4,1%
Skirstinys 14	**Dydis:** 87	**VPan:** 0,060	**IPan:** 0,004					
Sieja:	irakas	22,5%	iranas	21,4%	branduolinė programa	6,0%	derybos	4,2%
Išskiria:	iranas	11,9%	irakas	10,3%	krepšinis	5,7%	čempionatas	4,1%

K-MEANS, 1056 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 1056 **Skirstinių:** 20

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,308, **Grynumas:** 0,830

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	76	0,358	0,151	0,011	0,005	0,076	0,974
1	43	0,248	0,085	0,007	0,004	0,069	0,977
2	57	0,241	0,102	0,004	0,004	0,275	0,895
3	44	0,184	0,083	0,009	0,007	0,272	0,841
4	47	0,173	0,064	0,004	0,004	0,237	0,872
5	53	0,166	0,071	0,006	0,004	0,347	0,868
6	40	0,160	0,062	0,007	0,006	0,237	0,900
7	49	0,147	0,072	0,005	0,005	0,106	0,959
8	50	0,142	0,065	0,004	0,003	0,364	0,820
9	42	0,141	0,083	0,004	0,003	0,597	0,619
10	46	0,141	0,053	0,007	0,005	0,722	0,543
11	68	0,127	0,061	0,006	0,004	0,476	0,779

12	53	0,115	0,052	0,002	0,003	0,234	0,906
13	65	0,113	0,034	0,006	0,004	0,000	1,000
14	60	0,113	0,047	0,006	0,004	0,331	0,817
15	46	0,109	0,055	0,005	0,004	0,130	0,957
16	50	0,102	0,047	0,005	0,004	0,482	0,520
17	65	0,101	0,048	0,007	0,005	0,323	0,862
18	44	0,081	0,037	0,001	0,002	0,333	0,864
19	50	0,071	0,027	0,003	0,003	0,715	0,500

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	2	74
1	0	0	0	1	42
2	1	2	0	51	3
3	0	0	0	7	37
4	0	6	41	0	0
5	46	1	1	3	2
6	3	0	1	0	36
7	47	2	0	0	0
8	41	3	0	6	0
9	4	1	26	11	0
10	13	1	25	3	4
11	9	3	53	2	1
12	48	2	0	3	0
13	0	0	65	0	0
14	10	49	0	1	0
15	44	1	1	0	0
16	26	23	0	1	0
17	2	1	56	6	0
18	3	2	1	38	0
19	11	25	2	12	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 76	**VPan:** 0,358	**IPan:** 0,011					
Sieja:	čempionatas	35,1%	krepšinis	34,2%	rezultatai	13,3%	lietuviai	12,3%
Išskiria:	čempionatas	18,7%	krepšinis	14,6%	lietuviai	6,5%	įvairybės	4,0%
Skirstinys 1	**Dydis:** 43	**VPan:** 0,248	**IPan:** 0,007					
Sieja:	rezultatai	26,0%	ledo ritulys	18,1%	olimpiada	16,9%	turinas	16,2%
Išskiria:	ledo ritulys	10,1%	olimpiada	9,3%	turinas	9,1%	rezultatai	6,7%
Skirstinys 2	**Dydis:** 57	**VPan:** 0,241	**IPan:** 0,004					
Sieja:	įvairybės	88,7%	kalėdos	1,8%	vokietija	1,2%	britanija	0,8%
Išskiria:	įvairybės	47,3%	krepšinis	5,3%	čempionatas	3,8%	rezultatai	3,7%
Skirstinys 3	**Dydis:** 44	**VPan:** 0,184	**IPan:** 0,009					
Sieja:	eurolyga	34,7%	krepšinis	25,5%	laureatai	6,8%	moterys	4,0%
Išskiria:	eurolyga	21,4%	krepšinis	4,6%	čempionatas	4,3%	laureatai	4,2%
Skirstinys 4	**Dydis:** 47	**VPan:** 0,173	**IPan:** 0,004					
Sieja:	transportas	48,2%	aviacija	11,3%	sąlygos	5,6%	eismas	5,0%
Išskiria:	transportas	25,3%	aviacija	6,1%	krepšinis	5,1%	čempionatas	3,6%
Skirstinys 5	**Dydis:** 53	**VPan:** 0,166	**IPan:** 0,006					
Sieja:	rinkimai	42,7%	baltarusija	27,1%	ukraina	6,3%	parlamentas	6,2%
Išskiria:	rinkimai	22,7%	baltarusija	14,9%	krepšinis	4,8%	čempionatas	3,9%
Skirstinys 6	**Dydis:** 40	**VPan:** 0,160	**IPan:** 0,007					
Sieja:	futbolas	49,7%	italija	11,2%	dviračiai	6,2%	čempionatas	6,1%
Išskiria:	futbolas	28,9%	krepšinis	5,5%	italija	4,6%	dviračiai	3,5%
Skirstinys 7	**Dydis:** 49	**VPan:** 0,147	**IPan:** 0,005					
Sieja:	irakas	66,8%	aukos	5,5%	išpuolis	5,2%	sirija	3,3%
Išskiria:	irakas	37,0%	krepšinis	5,4%	čempionatas	3,9%	rezultatai	3,2%
Skirstinys 8	**Dydis:** 50	**VPan:** 0,142	**IPan:** 0,004					

Sieja:	palestiniečiai	31,7%	hamas	22,6%	izraelis	18,3%	vatikanas	3,7%
Išskiria:	palestiniečiai	16,5%	hamas	12,3%	izraelis	9,8%	krepšinis	5,1%
Skirstinys 9	**Dydis:** 42	**VPan:** 0,141	**IPan:** 0,004					
Sieja:	paukščių gripas	37,8%	žemės ūkis	34,3%	prancūzija	6,0%	kontrolė	2,6%
Išskiria:	paukščių gripas	20,5%	žemės ūkis	18,2%	krepšinis	5,1%	čempionatas	3,6%
Skirstinys 10	**Dydis:** 46	**VPan:** 0,141	**IPan:** 0,007					
Sieja:	es	68,9%	finansai	3,5%	parama	2,8%	ekonomika	2,3%
Išskiria:	es	38,1%	krepšinis	5,6%	čempionatas	4,0%	rezultatai	3,9%
Skirstinys 11	**Dydis:** 68	**VPan:** 0,127	**IPan:** 0,006					
Sieja:	energetika	42,8%	rusija	13,9%	nafta	10,3%	dujos	8,2%
Išskiria:	energetika	24,1%	nafta	5,9%	krepšinis	5,8%	dujos	4,8%
Skirstinys 12	**Dydis:** 53	**VPan:** 0,115	**IPan:** 0,002					
Sieja:	seimas	51,6%	adamkus	9,1%	liberalai	4,4%	pranešimas	4,0%
Išskiria:	seimas	26,9%	krepšinis	5,0%	adamkus	4,8%	čempionatas	3,5%
Skirstinys 13	**Dydis:** 65	**VPan:** 0,113	**IPan:** 0,006					
Sieja:	ekonomika	40,1%	finansai	28,8%	paslaugos	3,2%	rūpyba	2,5%
Išskiria:	ekonomika	21,0%	finansai	14,5%	krepšinis	5,7%	čempionatas	4,1%
Skirstinys 14	**Dydis:** 60	**VPan:** 0,113	**IPan:** 0,006					
Sieja:	aukos	56,0%	gaisras	9,6%	katastriša	5,1%	rusija	3,1%
Išskiria:	aukos	26,9%	krepšinis	5,8%	gaisras	5,4%	čempionatas	4,1%
Skirstinys 15	**Dydis:** 46	**VPan:** 0,109	**IPan:** 0,005					
Sieja:	iranas	38,4%	branduolinė programa	11,8%	derybos	6,9%	gynyba	5,6%
Išskiria:	iranas	21,1%	branduolinė programa	6,8%	krepšinis	5,4%	čempionatas	3,9%
Skirstinys 16	**Dydis:** 50	**VPan:** 0,102	**IPan:** 0,005					
Sieja:	sprogimas	23,3%	incidentas	16,0%	afganistanas	9,5%	žemės	7,1%
Išskiria:	sprogimas	13,3%	incidentas	8,6%	krepšinis	5,4%	afganistanas	5,2%
Skirstinys 17	**Dydis:** 65	**VPan:** 0,101	**IPan:** 0,007					
Sieja:	įmonės	40,2%	prekyba	14,4%	vokietija	10,1%	verslas	5,2%
Išskiria:	įmonės	22,3%	prekyba	7,5%	krepšinis	6,0%	čempionatas	4,3%
Skirstinys 18	**Dydis:** 44	**VPan:** 0,081	**IPan:** 0,001					
Sieja:	muzika	22,4%	kinas	20,9%	koncertas	7,8%	lietuva	5,5%
Išskiria:	muzika	11,3%	kinas	10,7%	krepšinis	4,7%	koncertas	4,0%
Skirstinys 19	**Dydis:** 50	**VPan:** 0,071	**IPan:** 0,003					
Sieja:	britanija	28,3%	sulaikymas	11,0%	pasienis	9,1%	kaunas	6,5%
Išskiria:	britanija	13,7%	sulaikymas	6,0%	krepšinis	5,0%	pasienis	4,7%

K-MEANS, 2112 DOKUMENTŲ, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 2112 **Skirstinių:** 5

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,461, **Grynumas:** 0,736

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	406	0,111	0,072	0,003	0,003	0,107	0,970
1	284	0,057	0,039	0,003	0,004	0,796	0,507
2	335	0,038	0,023	0,003	0,003	0,636	0,490
3	533	0,034	0,016	0,003	0,003	0,281	0,891
4	549	0,019	0,013	0,003	0,003	0,616	0,681

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	5	394	2	2	3
1	144	44	65	10	21
2	18	6	140	7	164
3	19	1	6	475	32

4	96	12	32	35	374

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 406	**VPan:** 0,111	**IPan:** 0,003					
Sieja:	krepšinis	30,9%	čempionatas	24,9%	rezultatai	12,5%	lietuviai	6,7%
Išskiria:	krepšinis	17,0%	čempionatas	13,6%	rezultatai	5,6%	įvairybės	4,6%
Skirstinys 1	**Dydis:** 284	**VPan:** 0,057	**IPan:** 0,003					
Sieja:	įvairybės	54,9%	britanija	10,4%	olimpiada	6,7%	turinas	6,3%
Išskiria:	įvairybės	30,5%	krepšinis	6,9%	čempionatas	5,3%	britanija	4,9%
Skirstinys 2	**Dydis:** 335	**VPan:** 0,038	**IPan:** 0,003					
Sieja:	aukos	29,4%	irakas	21,6%	terorizmas	3,5%	ispanija	3,5%
Išskiria:	aukos	16,6%	irakas	12,3%	krepšinis	7,2%	čempionatas	5,9%
Skirstinys 3	**Dydis:** 533	**VPan:** 0,034	**IPan:** 0,003					
Sieja:	finansai	10,5%	energetika	9,5%	ekonomika	9,2%	es	8,6%
Išskiria:	krepšinis	8,5%	čempionatas	7,0%	finansai	5,9%	energetika	5,4%
Skirstinys 4	**Dydis:** 549	**VPan:** 0,019	**IPan:** 0,003					
Sieja:	rinkimai	12,8%	baltarusija	7,8%	seimas	7,1%	iranas	6,8%
Išskiria:	krepšinis	8,1%	rinkimai	6,9%	čempionatas	6,6%	baltarusija	4,6%

K-MEANS, 2112 DOKUMENTŲ, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 2112 **Skirstinių:** 10

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,442, **Grynumas:** 0,721

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	260	0,196	0,099	0,006	0,004	0,031	0,992
1	87	0,190	0,070	0,006	0,004	0,068	0,977
2	133	0,156	0,080	0,004	0,004	0,450	0,812
3	151	0,078	0,038	0,003	0,003	0,343	0,848
4	267	0,059	0,027	0,004	0,003	0,251	0,895
5	262	0,055	0,025	0,005	0,003	0,458	0,794
6	252	0,050	0,029	0,003	0,003	0,559	0,504
7	242	0,045	0,030	0,003	0,004	0,318	0,880
8	248	0,038	0,024	0,004	0,006	0,831	0,379
9	205	0,036	0,020	0,004	0,003	0,947	0,293

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	1	258	0	0	1
1	2	85	0	0	0
2	108	6	11	3	5
3	4	0	16	3	128
4	4	0	2	239	22
5	6	3	17	208	28
6	6	3	113	3	127
7	15	6	4	4	213
8	94	79	53	12	10
9	42	17	29	57	60

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 260	**VPan:** 0,196	**IPan:** 0,006					
Sieja:	krepšinis	41,8%	čempionatas	25,9%	rezultatai	9,4%	lietuviai	8,2%
Išskiria:	krepšinis	23,8%	čempionatas	12,2%	įvairybės	4,3%	lietuviai	4,2%
Skirstinys 1	**Dydis:** 87	**VPan:** 0,190	**IPan:** 0,006					
Sieja:	olimpiada	21,4%	turinas	20,1%	ledo ritulys	18,8%	rezultatai	15,0%

Išskiria:	olimpiada	12,1%	turinas	11,3%	ledo ritulys	10,6%	krepšinis	6,4%
Skirstinys 2	**Dydis:** 133	**VPan:** 0,156	**IPan:** 0,004					
Sieja:	įvairybės	87,7%	britanija	2,4%	anekdotai	1,4%	rekordas	0,5%
Išskiria:	įvairybės	47,3%	krepšinis	6,3%	čempionatas	4,8%	rezultatai	3,2%
Skirstinys 3	**Dydis:** 151	**VPan:** 0,078	**IPan:** 0,003					
Sieja:	iranas	25,7%	izraelis	19,3%	palestiniečiai	17,3%	jt	5,2%
Išskiria:	iranas	13,9%	izraelis	10,5%	palestiniečiai	9,7%	krepšinis	6,4%
Skirstinys 4	**Dydis:** 267	**VPan:** 0,059	**IPan:** 0,004					
Sieja:	finansai	22,9%	ekonomika	19,9%	es	13,0%	prekyba	5,5%
Išskiria:	finansai	13,0%	ekonomika	11,5%	krepšinis	7,4%	čempionatas	6,0%
Skirstinys 5	**Dydis:** 262	**VPan:** 0,055	**IPan:** 0,005					
Sieja:	rusija	23,1%	įmonės	14,5%	transportas	13,2%	energetika	5,2%
Išskiria:	rusija	8,7%	įmonės	8,2%	transportas	8,0%	krepšinis	7,7%
Skirstinys 6	**Dydis:** 252	**VPan:** 0,050	**IPan:** 0,003					
Sieja:	aukos	36,1%	irakas	25,5%	bušas	3,0%	stichija	2,5%
Išskiria:	aukos	19,5%	irakas	13,8%	krepšinis	6,8%	čempionatas	5,5%
Skirstinys 7	**Dydis:** 242	**VPan:** 0,045	**IPan:** 0,003					
Sieja:	rinkimai	25,4%	baltarusija	16,9%	seimas	15,6%	prezidentas	6,5%
Išskiria:	rinkimai	13,2%	baltarusija	9,5%	seimas	8,9%	krepšinis	6,4%
Skirstinys 8	**Dydis:** 248	**VPan:** 0,038	**IPan:** 0,004					
Sieja:	futbolas	35,1%	kaunas	7,5%	muzika	4,0%	statistika	3,9%
Išskiria:	futbolas	20,7%	krepšinis	7,1%	kaunas	4,4%	įvairybės	3,2%
Skirstinys 9	**Dydis:** 205	**VPan:** 0,036	**IPan:** 0,004					
Sieja:	britanija	15,7%	prancūzija	13,7%	ispanija	11,7%	paukščių gripas	9,9%
Išskiria:	prancūzija	7,5%	britanija	7,3%	krepšinis	6,7%	paukščių gripas	5,9%

K-MEANS, 2112 DOKUMENTŲ, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 2112 **Skirstinių:** 15

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,399, **Grynumas:** 0,777

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	167	0,300	0,137	0,012	0,005	0,023	0,994
1	81	0,208	0,074	0,006	0,004	0,041	0,988
2	78	0,202	0,090	0,004	0,004	0,117	0,962
3	99	0,177	0,075	0,011	0,006	0,166	0,949
4	129	0,132	0,072	0,004	0,003	0,357	0,860
5	86	0,126	0,076	0,003	0,002	0,395	0,826
6	113	0,127	0,067	0,005	0,004	0,261	0,903
7	160	0,095	0,043	0,005	0,003	0,024	0,994
8	147	0,091	0,047	0,007	0,006	0,686	0,633
9	117	0,085	0,052	0,004	0,004	0,799	0,419
10	199	0,065	0,033	0,004	0,003	0,668	0,477
11	227	0,066	0,030	0,004	0,003	0,237	0,907
12	178	0,050	0,026	0,005	0,004	0,704	0,607
13	166	0,039	0,019	0,003	0,002	0,444	0,783
14	160	0,028	0,015	0,002	0,002	0,719	0,619

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	1	166	0	0	0
1	1	80	0	0	0
2	0	0	2	75	1
3	2	94	1	1	1
4	111	2	7	2	7
5	2	0	6	7	71

6	5	4	2	0	102
7	0	0	0	159	1
8	5	93	11	11	27
9	31	0	49	13	24
10	10	1	80	13	95
11	3	0	3	206	15
12	8	6	33	23	108
13	4	4	26	2	130
14	99	7	25	17	12

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 167	**VPan:** 0,300	**IPan:** 0,012					
Sieja:	čempionatas	36,9%	krepšinis	34,4%	rezultatai	13,3%	lietuviai	9,7%
Išskiria:	čempionatas	18,9%	krepšinis	14,4%	lietuviai	4,8%	įvairybės	4,4%
Skirstinys 1	**Dydis:** 81	**VPan:** 0,208	**IPan:** 0,006					
Sieja:	olimpiada	21,2%	turinas	21,1%	ledo ritulys	19,8%	rezultatai	15,8%
Išskiria:	turinas	11,9%	olimpiada	11,8%	ledo ritulys	11,2%	krepšinis	6,3%
Skirstinys 2	**Dydis:** 78	**VPan:** 0,202	**IPan:** 0,004					
Sieja:	transportas	40,7%	eismas	12,0%	keliai	10,2%	sąlygos	9,8%
Išskiria:	transportas	21,8%	eismas	6,3%	krepšinis	6,0%	keliai	5,5%
Skirstinys 3	**Dydis:** 99	**VPan:** 0,177	**IPan:** 0,011					
Sieja:	eurolyga	29,5%	krepšinis	24,3%	moterys	11,6%	lietuvos	4,5%
Išskiria:	eurolyga	19,5%	moterys	6,8%	krepšinis	3,9%	įvairybės	3,6%
Skirstinys 4	**Dydis:** 129	**VPan:** 0,132	**IPan:** 0,004					
Sieja:	įvairybės	81,3%	kinas	5,9%	anekdotai	1,8%	kinija	1,4%
Išskiria:	įvairybės	41,4%	krepšinis	6,4%	čempionatas	5,2%	rezultatai	3,3%
Skirstinys 5	**Dydis:** 86	**VPan:** 0,126	**IPan:** 0,003					
Sieja:	izraelis	35,1%	palestiniečiai	33,0%	hamas	8,3%	šaronas	3,8%
Išskiria:	izraelis	18,4%	palestiniečiai	17,9%	krepšinis	5,9%	čempionatas	4,8%
Skirstinys 6	**Dydis:** 113	**VPan:** 0,127	**IPan:** 0,005					
Sieja:	rinkimai	43,4%	baltarusija	26,9%	prezidentas	6,7%	ukraina	5,0%
Išskiria:	rinkimai	23,2%	baltarusija	15,0%	krepšinis	6,1%	čempionatas	5,3%
Skirstinys 7	**Dydis:** 160	**VPan:** 0,095	**IPan:** 0,005					
Sieja:	energetika	34,5%	įmonės	22,0%	nafta	8,6%	dujos	5,2%
Išskiria:	energetika	19,6%	įmonės	12,1%	krepšinis	6,9%	čempionatas	5,6%
Skirstinys 8	**Dydis:** 147	**VPan:** 0,091	**IPan:** 0,007					
Sieja:	futbolas	43,2%	ispanija	17,6%	prancūzija	5,8%	statistika	5,0%
Išskiria:	futbolas	27,1%	ispanija	8,9%	krepšinis	6,9%	rezultatai	3,2%
Skirstinys 9	**Dydis:** 117	**VPan:** 0,085	**IPan:** 0,004					
Sieja:	britanija	57,4%	pasienis	7,2%	sulaikymas	4,9%	tyrimas	4,1%
Išskiria:	britanija	32,1%	krepšinis	6,3%	čempionatas	5,2%	pasienis	4,0%
Skirstinys 10	**Dydis:** 199	**VPan:** 0,065	**IPan:** 0,004					
Sieja:	aukos	38,8%	irakas	27,1%	paukščių gripas	3,6%	sprogimas	1,9%
Išskiria:	aukos	20,5%	irakas	14,3%	krepšinis	6,7%	čempionatas	5,4%
Skirstinys 11	**Dydis:** 227	**VPan:** 0,066	**IPan:** 0,004					
Sieja:	finansai	27,9%	ekonomika	22,6%	es	14,1%	prekyba	6,8%
Išskiria:	finansai	15,8%	ekonomika	12,7%	krepšinis	7,1%	es	6,3%
Skirstinys 12	**Dydis:** 178	**VPan:** 0,050	**IPan:** 0,005					
Sieja:	rusija	34,4%	iranas	22,7%	jt	6,2%	atomas	2,8%
Išskiria:	rusija	13,2%	iranas	13,0%	krepšinis	7,3%	čempionatas	5,9%
Skirstinys 13	**Dydis:** 166	**VPan:** 0,039	**IPan:** 0,003					
Sieja:	seimas	34,3%	bušas	5,7%	lenkija	5,2%	paulauskas	4,6%
Išskiria:	seimas	19,0%	krepšinis	6,2%	čempionatas	5,1%	rezultatai	3,2%
Skirstinys 14	**Dydis:** 160	**VPan:** 0,028	**IPan:** 0,002					
Sieja:	kaunas	15,7%	muzika	14,0%	vilnius	11,4%	paroda	4,2%
Išskiria:	kaunas	8,1%	muzika	7,4%	vilnius	6,1%	krepšinis	6,0%

K-MEANS, 2112 DOKUMENTŲ, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 2112 **Skirstinių:** 20

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,327, **Grynumas:** 0,832

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	183	0,270	0,129	0,011	0,005	0,021	0,995
1	77	0,233	0,089	0,011	0,006	0,188	0,935
2	88	0,208	0,064	0,006	0,004	0,000	1,000
3	82	0,205	0,073	0,006	0,004	0,041	0,988
4	67	0,201	0,082	0,003	0,003	0,096	0,970
5	84	0,186	0,073	0,008	0,008	0,274	0,905
6	79	0,178	0,070	0,005	0,004	0,272	0,899
7	135	0,154	0,073	0,004	0,003	0,427	0,807
8	95	0,150	0,072	0,005	0,004	0,231	0,916
9	70	0,116	0,048	0,004	0,004	0,890	0,329
10	74	0,108	0,048	0,001	0,002	0,166	0,946
11	111	0,103	0,042	0,005	0,003	0,300	0,847
12	137	0,096	0,049	0,004	0,003	0,074	0,978
13	148	0,095	0,034	0,005	0,004	0,025	0,993
14	123	0,085	0,039	0,005	0,003	0,436	0,780
15	83	0,080	0,045	0,005	0,004	0,668	0,566
16	86	0,064	0,033	0,002	0,003	0,651	0,605
17	131	0,067	0,033	0,005	0,004	0,673	0,664
18	118	0,039	0,020	0,002	0,001	0,670	0,669
19	136	0,032	0,015	0,003	0,003	0,604	0,676

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	1	182	0	0	0
1	3	72	0	1	1
2	0	0	0	88	0
3	1	81	0	0	0
4	1	0	1	0	65
5	3	76	2	1	2
6	2	0	3	3	71
7	109	0	13	5	8
8	3	4	1	0	87
9	23	17	3	7	20
10	1	0	1	2	70
11	2	0	94	0	15
12	0	0	2	134	1
13	0	0	0	147	1
14	10	0	1	96	16
15	5	5	25	1	47
16	6	2	52	23	3
17	15	5	12	12	87
18	79	8	15	8	8
19	18	5	20	1	92

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 183	**VPan:** 0,270	**IPan:** 0,011					
Sieja:	čempionatas	37,7%	krepšinis	33,6%	rezultatai	12,6%	lietuviai	8,8%
Išskiria:	čempionatas	20,2%	krepšinis	14,3%	įvairybės	4,4%	lietuviai	4,2%
Skirstinys 1	**Dydis:** 77	**VPan:** 0,233	**IPan:** 0,011					

Sieja:	eurolyga	37,1%	krepšinis	24,6%	moterys	5,2%	lietuvos	5,1%
Išskiria:	eurolyga	23,6%	čempionatas	6,0%	krepšinis	3,6%	įvairybės	3,4%
Skirstinys 2	**Dydis:** 88	**VPan:** 0,208	**IPan:** 0,006					
Sieja:	energetika	53,8%	nafta	12,9%	dujos	7,9%	pasaulis	5,5%
Išskiria:	energetika	30,1%	nafta	7,3%	krepšinis	6,5%	čempionatas	5,3%
Skirstinys 3	**Dydis:** 82	**VPan:** 0,205	**IPan:** 0,006					
Sieja:	olimpiada	21,1%	turinas	20,9%	ledo ritulys	19,7%	rezultatai	15,6%
Išskiria:	turinas	11,8%	olimpiada	11,7%	ledo ritulys	11,1%	krepšinis	6,3%
Skirstinys 4	**Dydis:** 67	**VPan:** 0,201	**IPan:** 0,003					
Sieja:	izraelis	39,3%	palestiniečiai	34,1%	hamas	8,6%	šaronas	4,0%
Išskiria:	izraelis	20,7%	palestiniečiai	18,3%	krepšinis	5,9%	čempionatas	4,8%
Skirstinys 5	**Dydis:** 84	**VPan:** 0,186	**IPan:** 0,008					
Sieja:	futbolas	68,5%	statistika	6,1%	čempionatas	5,4%	anglija	2,6%
Išskiria:	futbolas	41,7%	krepšinis	7,0%	rezultatai	2,8%	rusija	2,5%
Skirstinys 6	**Dydis:** 79	**VPan:** 0,178	**IPan:** 0,005					
Sieja:	irakas	79,8%	vyriausybė	3,1%	aukos	2,8%	išpuolis	1,4%
Išskiria:	irakas	43,8%	krepšinis	6,1%	čempionatas	5,0%	rusija	2,9%
Skirstinys 7	**Dydis:** 135	**VPan:** 0,154	**IPan:** 0,004					
Sieja:	įvairybės	78,8%	britanija	12,0%	anekdotai	1,4%	didžioji	0,9%
Išskiria:	įvairybės	41,6%	krepšinis	6,3%	čempionatas	5,2%	britanija	5,2%
Skirstinys 8	**Dydis:** 95	**VPan:** 0,150	**IPan:** 0,005					
Sieja:	rinkimai	41,4%	baltarusija	32,2%	prezidentas	5,9%	ukraina	4,8%
Išskiria:	rinkimai	21,2%	baltarusija	18,0%	krepšinis	6,0%	čempionatas	5,2%
Skirstinys 9	**Dydis:** 70	**VPan:** 0,116	**IPan:** 0,004					
Sieja:	prancūzija	45,9%	kinas	12,9%	dviračiai	11,0%	lenktynės	6,8%
Išskiria:	prancūzija	24,7%	kinas	6,7%	dviračiai	6,0%	krepšinis	5,4%
Skirstinys 10	**Dydis:** 74	**VPan:** 0,108	**IPan:** 0,001					
Sieja:	seimas	62,1%	paulauskas	9,6%	nepasitikėjimas	1,8%	partijos	1,8%
Išskiria:	seimas	31,9%	krepšinis	5,6%	paulauskas	5,0%	čempionatas	4,6%
Skirstinys 11	**Dydis:** 111	**VPan:** 0,103	**IPan:** 0,005					
Sieja:	aukos	62,6%	stichija	6,2%	katastriša	4,0%	avarija	3,0%
Išskiria:	aukos	32,4%	krepšinis	6,4%	čempionatas	5,2%	stichija	3,5%
Skirstinys 12	**Dydis:** 137	**VPan:** 0,096	**IPan:** 0,004					
Sieja:	transportas	27,8%	įmonės	16,6%	eismas	8,2%	keliai	7,0%
Išskiria:	transportas	15,5%	įmonės	7,6%	krepšinis	6,4%	čempionatas	5,2%
Skirstinys 13	**Dydis:** 148	**VPan:** 0,095	**IPan:** 0,005					
Sieja:	finansai	41,8%	ekonomika	30,0%	es	3,0%	euras	1,9%
Išskiria:	finansai	23,3%	ekonomika	16,2%	krepšinis	6,8%	čempionatas	5,5%
Skirstinys 14	**Dydis:** 123	**VPan:** 0,085	**IPan:** 0,005					
Sieja:	prekyba	34,8%	žemės ūkis	14,9%	es	12,8%	paukščių gripas	11,7%
Išskiria:	prekyba	20,4%	žemės ūkis	8,5%	paukščių gripas	6,8%	krepšinis	6,8%
Skirstinys 15	**Dydis:** 83	**VPan:** 0,080	**IPan:** 0,005					
Sieja:	ispanija	36,7%	terorizmas	19,7%	afganistanas	9,2%	sprogimas	8,8%
Išskiria:	ispanija	17,6%	terorizmas	10,9%	afganistanas	5,5%	krepšinis	5,1%
Skirstinys 16	**Dydis:** 86	**VPan:** 0,064	**IPan:** 0,002					
Sieja:	pasienis	19,3%	paslaugos	16,4%	sulaikymas	14,7%	kontrabanda	8,3%
Išskiria:	pasienis	10,3%	paslaugos	8,4%	sulaikymas	7,8%	krepšinis	5,8%
Skirstinys 17	**Dydis:** 131	**VPan:** 0,067	**IPan:** 0,005					
Sieja:	iranas	28,7%	rusija	27,3%	jt	7,5%	branduolinė programa	5,2%
Išskiria:	iranas	15,9%	rusija	8,3%	krepšinis	6,9%	čempionatas	5,6%
Skirstinys 18	**Dydis:** 118	**VPan:** 0,039	**IPan:** 0,002					
Sieja:	kaunas	18,3%	muzika	17,9%	vilnius	11,6%	koncertas	5,5%
Išskiria:	muzika	9,2%	kaunas	9,1%	vilnius	5,9%	krepšinis	5,8%
Skirstinys 19	**Dydis:** 136	**VPan:** 0,032	**IPan:** 0,003					
Sieja:	lenkija	11,8%	adamkus	10,2%	bušas	6,2%	kačinskis	4,3%
Išskiria:	krepšinis	6,6%	adamkus	5,9%	čempionatas	5,4%	lenkija	5,1%

K-MEANS, 4224 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 4224 **Skirstinių:** 5

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,445, **Grynumas:** 0,752

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	822	0,087	0,062	0,002	0,002	0,085	0,976
1	668	0,039	0,025	0,002	0,002	0,598	0,490
2	782	0,028	0,018	0,003	0,003	0,301	0,890
3	1141	0,024	0,014	0,003	0,002	0,499	0,772
4	808	0,023	0,019	0,002	0,003	0,748	0,579

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	2	1	14	802	3
1	327	49	9	0	283
2	696	14	35	8	29
3	103	881	46	10	101
4	112	128	468	21	79

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 822	**VPan:** 0,087	**IPan:** 0,002					
Sieja:	krepšinis	27,3%	čempionatas	21,2%	rezultatai	16,8%	lietuviai	7,5%
Išskiria:	krepšinis	14,8%	čempionatas	11,4%	rezultatai	8,2%	lietuviai	4,0%
Skirstinys 1	**Dydis:** 668	**VPan:** 0,039	**IPan:** 0,002					
Sieja:	aukos	33,4%	irakas	15,2%	eismas	4,6%	sprogimas	3,1%
Išskiria:	aukos	18,9%	irakas	7,7%	krepšinis	6,2%	čempionatas	4,9%
Skirstinys 2	**Dydis:** 782	**VPan:** 0,028	**IPan:** 0,003					
Sieja:	rinkimai	16,5%	iranas	11,5%	baltarusija	10,6%	seimas	6,1%
Išskiria:	rinkimai	9,3%	krepšinis	6,5%	iranas	6,3%	baltarusija	6,0%
Skirstinys 3	**Dydis:** 1141	**VPan:** 0,024	**IPan:** 0,003					
Sieja:	ekonomika	11,2%	energetika	9,7%	finansai	9,3%	rusija	9,0%
Išskiria:	krepšinis	7,7%	ekonomika	6,6%	čempionatas	6,0%	energetika	5,7%
Skirstinys 4	**Dydis:** 808	**VPan:** 0,023	**IPan:** 0,002					
Sieja:	įvairybės	46,1%	britanija	11,7%	transportas	7,9%	aviacija	2,3%
Išskiria:	įvairybės	26,5%	krepšinis	6,2%	britanija	5,8%	čempionatas	4,8%

K-MEANS, 4224 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 4224 **Skirstinių:** 10

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,414, **Grynumas:** 0,759

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	196	0,147	0,068	0,005	0,003	0,069	0,980
1	596	0,131	0,082	0,004	0,003	0,052	0,985
2	280	0,080	0,049	0,004	0,003	0,555	0,543
3	379	0,072	0,037	0,003	0,003	0,649	0,633
4	340	0,064	0,034	0,003	0,003	0,288	0,894
5	456	0,053	0,033	0,003	0,002	0,501	0,544
6	502	0,049	0,031	0,004	0,003	0,425	0,813
7	506	0,047	0,026	0,004	0,003	0,521	0,771
8	487	0,040	0,028	0,003	0,003	0,295	0,891
9	479	0,014	0,009	0,001	0,002	0,740	0,522

Pasiskirstymas klasėse					
Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	0	0	1	192	3
1	1	0	8	587	0
2	107	152	0	0	21
3	29	83	240	6	21
4	304	7	17	2	10
5	196	4	8	0	248
6	55	408	9	7	23
7	46	390	15	31	24
8	434	4	24	8	17
9	68	25	250	8	128

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 196	**VPan:** 0,147	**IPan:** 0,005					
Sieja:	rezultatai	22,0%	ledo ritulys	17,0%	olimpiada	17,0%	turinas	16,3%
Išskiria:	olimpiada	9,8%	ledo ritulys	9,7%	turinas	9,4%	krepšinis	5,8%
Skirstinys 1	**Dydis:** 596	**VPan:** 0,131	**IPan:** 0,004					
Sieja:	krepšinis	34,2%	čempionatas	26,3%	rezultatai	8,9%	lietuviai	8,7%
Išskiria:	krepšinis	19,2%	čempionatas	14,6%	lietuviai	4,6%	įvairybės	3,5%
Skirstinys 2	**Dydis:** 280	**VPan:** 0,080	**IPan:** 0,004					
Sieja:	transportas	28,0%	irakas	18,9%	eismas	10,0%	keliai	8,2%
Išskiria:	transportas	15,9%	irakas	8,0%	krepšinis	5,7%	eismas	5,4%
Skirstinys 3	**Dydis:** 379	**VPan:** 0,072	**IPan:** 0,003					
Sieja:	įvairybės	63,6%	britanija	11,3%	žemės ūkis	8,0%	paukščių gripas	7,0%
Išskiria:	įvairybės	35,0%	krepšinis	5,7%	britanija	4,8%	čempionatas	4,4%
Skirstinys 4	**Dydis:** 340	**VPan:** 0,064	**IPan:** 0,003					
Sieja:	iranas	27,8%	palestiniečiai	13,7%	izraelis	13,0%	branduolinė programa	7,0%
Išskiria:	iranas	15,2%	palestiniečiai	7,5%	izraelis	6,9%	krepšinis	5,7%
Skirstinys 5	**Dydis:** 456	**VPan:** 0,053	**IPan:** 0,003					
Sieja:	aukos	51,3%	sprogimas	4,9%	terorizmas	3,8%	pakistanas	3,1%
Išskiria:	aukos	28,9%	krepšinis	5,9%	čempionatas	4,6%	rezultatai	4,0%
Skirstinys 6	**Dydis:** 502	**VPan:** 0,049	**IPan:** 0,004					
Sieja:	energetika	24,2%	rusija	19,0%	įmonės	14,7%	nafta	5,7%
Išskiria:	energetika	14,4%	įmonės	8,4%	krepšinis	6,5%	rusija	6,4%
Skirstinys 7	**Dydis:** 506	**VPan:** 0,047	**IPan:** 0,004					
Sieja:	ekonomika	28,8%	finansai	25,5%	es	11,2%	statistika	4,4%
Išskiria:	ekonomika	17,1%	finansai	15,1%	krepšinis	6,5%	čempionatas	5,1%
Skirstinys 8	**Dydis:** 487	**VPan:** 0,040	**IPan:** 0,003					
Sieja:	rinkimai	26,6%	baltarusija	18,5%	seimas	10,8%	prezidentas	7,0%
Išskiria:	rinkimai	14,4%	baltarusija	10,2%	seimas	6,1%	krepšinis	5,5%
Skirstinys 9	**Dydis:** 479	**VPan:** 0,014	**IPan:** 0,001					
Sieja:	vilnius	7,9%	sulaikymas	7,7%	kaunas	4,8%	muzika	4,7%
Išskiria:	krepšinis	5,4%	čempionatas	4,2%	sulaikymas	4,2%	rezultatai	3,8%

K-MEANS, 4224 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 4224 **Skirstinių:** 15

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,382, **Grynumas:** 0,786

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	352	0,234	0,131	0,008	0,004	0,079	0,974
1	177	0,171	0,070	0,005	0,003	0,000	1,000
2	167	0,164	0,090	0,003	0,003	0,248	0,892
3	231	0,118	0,058	0,004	0,004	0,261	0,909
4	253	0,104	0,061	0,004	0,004	0,448	0,688

5	312	0,092	0,042	0,003	0,002	0,455	0,654
6	306	0,090	0,048	0,003	0,002	0,432	0,817
7	207	0,087	0,051	0,003	0,002	0,549	0,739
8	211	0,087	0,044	0,004	0,003	0,288	0,891
9	287	0,090	0,038	0,007	0,005	0,248	0,909
10	252	0,071	0,046	0,004	0,003	0,478	0,790
11	411	0,057	0,030	0,004	0,003	0,312	0,876
12	322	0,038	0,018	0,003	0,003	0,523	0,506
13	331	0,032	0,019	0,003	0,002	0,407	0,795
14	402	0,016	0,008	0,002	0,002	0,743	0,560

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	1	0	8	343	0
1	0	0	0	177	0
2	4	149	0	0	14
3	210	2	5	9	5
4	3	174	3	0	73
5	102	2	4	0	204
6	24	8	250	4	20
7	153	6	18	25	5
8	4	188	13	1	5
9	15	1	7	261	3
10	199	14	10	4	25
11	26	360	11	0	14
12	163	149	4	2	4
13	263	4	14	0	50
14	73	16	225	15	73

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 352	**VPan:** 0,234	**IPan:** 0,008					
Sieja:	krepšinis	35,7%	čempionatas	31,5%	lietuviai	13,5%	rezultatai	13,1%
Išskiria:	krepšinis	17,1%	čempionatas	16,3%	lietuviai	7,9%	įvairybės	3,5%
Skirstinys 1	**Dydis:** 177	**VPan:** 0,171	**IPan:** 0,005					
Sieja:	rezultatai	22,1%	ledo ritulys	17,9%	olimpiada	17,5%	turinas	17,2%
Išskiria:	ledo ritulys	10,2%	olimpiada	10,0%	turinas	9,9%	krepšinis	5,7%
Skirstinys 2	**Dydis:** 167	**VPan:** 0,164	**IPan:** 0,003					
Sieja:	transportas	36,8%	eismas	13,6%	keliai	11,2%	sąlygos	10,5%
Išskiria:	transportas	19,6%	eismas	7,0%	keliai	6,0%	sąlygos	5,6%
Skirstinys 3	**Dydis:** 231	**VPan:** 0,118	**IPan:** 0,004					
Sieja:	rinkimai	39,8%	baltarusija	29,3%	parlamentas	8,5%	prezidentas	7,3%
Išskiria:	rinkimai	21,5%	baltarusija	16,3%	krepšinis	4,9%	čempionatas	4,4%
Skirstinys 4	**Dydis:** 253	**VPan:** 0,104	**IPan:** 0,004					
Sieja:	energetika	45,2%	dujos	10,5%	nafta	10,3%	pasaulis	5,1%
Išskiria:	energetika	25,5%	dujos	5,9%	nafta	5,7%	krepšinis	5,7%
Skirstinys 5	**Dydis:** 312	**VPan:** 0,092	**IPan:** 0,003					
Sieja:	aukos	62,4%	sprogimas	5,5%	irakas	2,8%	avarija	2,5%
Išskiria:	aukos	34,7%	krepšinis	5,7%	čempionatas	4,5%	rezultatai	3,9%
Skirstinys 6	**Dydis:** 306	**VPan:** 0,090	**IPan:** 0,003					
Sieja:	įvairybės	76,8%	britanija	11,0%	kinas	3,0%	didžioji	0,9%
Išskiria:	įvairybės	41,2%	krepšinis	5,4%	čempionatas	4,3%	britanija	4,3%
Skirstinys 7	**Dydis:** 207	**VPan:** 0,087	**IPan:** 0,003					
Sieja:	izraelis	26,0%	palestiniečiai	25,5%	hamas	11,4%	šaronas	6,7%
Išskiria:	izraelis	13,9%	palestiniečiai	13,8%	hamas	6,2%	krepšinis	5,4%
Skirstinys 8	**Dydis:** 211	**VPan:** 0,087	**IPan:** 0,004					
Sieja:	prekyba	30,9%	žemės ūkis	27,3%	paukščių gripas	21,1%	aplinkosauga	2,1%
Išskiria:	prekyba	17,0%	žemės ūkis	15,8%	paukščių gripas	12,1%	krepšinis	5,7%
Skirstinys 9	**Dydis:** 287	**VPan:** 0,090	**IPan:** 0,007					

Sieja:	futbolas	38,6%	eurolyga	9,1%	krepšinis	7,7%	laureatai	6,9%
Išskiria:	futbolas	25,1%	eurolyga	5,0%	laureatai	4,5%	rezultatai	3,6%
Skirstinys 10	**Dydis:** 252	**VPan:** 0,071	**IPan:** 0,004					
Sieja:	iranas	41,9%	rusija	18,6%	branduolinė programa	11,4%	jt	6,9%
Išskiria:	iranas	24,1%	branduolinė programa	6,8%	krepšinis	5,9%	čempionatas	4,6%
Skirstinys 11	**Dydis:** 411	**VPan:** 0,057	**IPan:** 0,004					
Sieja:	ekonomika	32,0%	finansai	26,9%	es	9,4%	statistika	4,5%
Išskiria:	ekonomika	18,2%	finansai	15,1%	krepšinis	6,2%	čempionatas	4,9%
Skirstinys 12	**Dydis:** 322	**VPan:** 0,038	**IPan:** 0,003					
Sieja:	įmonės	30,8%	seimas	24,0%	verslas	5,6%	paulauskas	5,3%
Išskiria:	įmonės	16,0%	seimas	13,5%	krepšinis	5,7%	čempionatas	4,4%
Skirstinys 13	**Dydis:** 331	**VPan:** 0,032	**IPan:** 0,003					
Sieja:	irakas	36,9%	bušas	7,9%	terorizmas	5,8%	kariai	3,0%
Išskiria:	irakas	18,2%	krepšinis	5,9%	čempionatas	4,6%	bušas	4,4%
Skirstinys 14	**Dydis:** 402	**VPan:** 0,016	**IPan:** 0,002					
Sieja:	vilnius	9,1%	prancūzija	8,4%	muzika	6,0%	lenkija	5,5%
Išskiria:	krepšinis	5,7%	vilnius	4,6%	čempionatas	4,3%	rezultatai	4,0%

K-MEANS, 4224 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 4224 **Skirstinių:** 20

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,337, **Grynumas:** 0,829

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	177	0,299	0,131	0,013	0,010	0,000	1,000
1	88	0,286	0,113	0,005	0,004	0,131	0,955
2	218	0,242	0,135	0,014	0,007	0,125	0,954
3	99	0,221	0,099	0,004	0,003	0,217	0,889
4	168	0,191	0,078	0,005	0,004	0,118	0,958
5	159	0,174	0,094	0,003	0,003	0,240	0,899
6	159	0,176	0,067	0,007	0,006	0,193	0,931
7	165	0,142	0,058	0,004	0,004	0,184	0,933
8	212	0,126	0,061	0,004	0,003	0,235	0,920
9	175	0,127	0,058	0,008	0,004	0,179	0,931
10	307	0,090	0,048	0,003	0,002	0,438	0,811
11	200	0,088	0,054	0,003	0,002	0,483	0,755
12	261	0,081	0,037	0,004	0,003	0,299	0,851
13	270	0,076	0,040	0,006	0,003	0,363	0,830
14	240	0,072	0,048	0,004	0,003	0,428	0,825
15	237	0,070	0,034	0,004	0,003	0,404	0,797
16	233	0,063	0,032	0,004	0,003	0,208	0,931
17	263	0,034	0,021	0,003	0,002	0,553	0,688
18	304	0,029	0,019	0,002	0,002	0,492	0,757
19	286	0,029	0,017	0,002	0,002	0,747	0,409

Pasiskirstymas klasėse

Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	0	0	0	177	0
1	0	0	3	84	1
2	1	0	9	208	0
3	0	88	11	0	0
4	6	161	1	0	0
5	2	143	1	0	13
6	0	1	3	148	7
7	154	1	2	0	8
8	195	1	5	7	4

9	1	0	10	163	1
10	26	9	249	3	20
11	151	11	33	2	3
12	35	2	2	0	222
13	36	224	5	2	3
14	198	7	14	5	16
15	4	189	36	0	8
16	8	217	2	3	3
17	181	4	22	2	54
18	230	9	47	2	16
19	12	6	117	35	116

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 177	**VPan:** 0,299	**IPan:** 0,013					
Sieja:	rezultatai	53,3%	krepšinis	14,5%	čempionatas	13,2%	ledo ritulys	8,5%
Išskiria:	rezultatai	27,2%	ledo ritulys	5,3%	rusija	3,8%	įvairybės	3,2%
Skirstinys 1	**Dydis:** 88	**VPan:** 0,286	**IPan:** 0,005					
Sieja:	olimpiada	43,3%	turinas	41,5%	rezultatai	3,5%	biatlonas	2,3%
Išskiria:	olimpiada	23,8%	turinas	22,8%	krepšinis	5,2%	čempionatas	3,9%
Skirstinys 2	**Dydis:** 218	**VPan:** 0,242	**IPan:** 0,014					
Sieja:	lietuviai	33,0%	čempionatas	32,5%	krepšinis	23,7%	statistika	1,7%
Išskiria:	lietuviai	22,7%	čempionatas	13,5%	krepšinis	5,9%	įvairybės	3,6%
Skirstinys 3	**Dydis:** 99	**VPan:** 0,221	**IPan:** 0,004					
Sieja:	žemės ūkis	51,0%	paukščių gripas	40,3%	prekyba	0,8%	išmokos	0,7%
Išskiria:	žemės ūkis	27,7%	paukščių gripas	21,9%	krepšinis	5,1%	čempionatas	4,0%
Skirstinys 4	**Dydis:** 168	**VPan:** 0,191	**IPan:** 0,005					
Sieja:	energetika	48,9%	dujos	13,0%	nafta	12,7%	pasaulis	6,2%
Išskiria:	energetika	26,8%	dujos	7,3%	nafta	7,0%	krepšinis	5,6%
Skirstinys 5	**Dydis:** 159	**VPan:** 0,174	**IPan:** 0,003					
Sieja:	transportas	35,5%	eismas	14,2%	keliai	11,6%	sąlygos	10,9%
Išskiria:	transportas	18,6%	eismas	7,3%	keliai	6,2%	sąlygos	5,8%
Skirstinys 6	**Dydis:** 159	**VPan:** 0,176	**IPan:** 0,007					
Sieja:	futbolas	65,3%	statistika	7,3%	čempionatas	5,5%	anglija	3,7%
Išskiria:	futbolas	39,3%	krepšinis	6,0%	rezultatai	2,9%	įvairybės	2,4%
Skirstinys 7	**Dydis:** 165	**VPan:** 0,142	**IPan:** 0,004					
Sieja:	irakas	76,8%	aukos	5,2%	išpuolis	2,8%	bagdadas	2,8%
Išskiria:	irakas	42,4%	krepšinis	5,5%	čempionatas	4,3%	rezultatai	3,2%
Skirstinys 8	**Dydis:** 212	**VPan:** 0,126	**IPan:** 0,004					
Sieja:	rinkimai	38,9%	baltarusija	31,4%	parlamentas	7,8%	prezidentas	7,6%
Išskiria:	rinkimai	20,4%	baltarusija	17,4%	krepšinis	5,2%	čempionatas	4,4%
Skirstinys 9	**Dydis:** 175	**VPan:** 0,127	**IPan:** 0,008					
Sieja:	krepšinis	35,2%	eurolyga	23,9%	laureatai	9,4%	moterys	4,2%
Išskiria:	eurolyga	14,6%	krepšinis	8,8%	laureatai	5,7%	čempionatas	5,1%
Skirstinys 10	**Dydis:** 307	**VPan:** 0,090	**IPan:** 0,003					
Sieja:	įvairybės	76,3%	britanija	11,8%	kinas	2,7%	didžioji	0,9%
Išskiria:	įvairybės	41,0%	krepšinis	5,4%	britanija	4,7%	čempionatas	4,3%
Skirstinys 11	**Dydis:** 200	**VPan:** 0,088	**IPan:** 0,003					
Sieja:	palestiniečiai	26,2%	izraelis	25,9%	hamas	12,7%	vyriausybė	7,7%
Išskiria:	palestiniečiai	14,0%	izraelis	13,5%	hamas	6,9%	krepšinis	5,3%
Skirstinys 12	**Dydis:** 261	**VPan:** 0,081	**IPan:** 0,004					
Sieja:	aukos	56,0%	stichija	5,1%	incidentas	4,6%	avarija	4,3%
Išskiria:	aukos	28,4%	krepšinis	5,9%	čempionatas	4,6%	rezultatai	4,1%
Skirstinys 13	**Dydis:** 270	**VPan:** 0,076	**IPan:** 0,006					
Sieja:	ekonomika	44,0%	prekyba	12,9%	es	11,2%	statistika	5,6%
Išskiria:	ekonomika	25,4%	prekyba	6,6%	krepšinis	6,4%	čempionatas	5,0%
Skirstinys 14	**Dydis:** 240	**VPan:** 0,072	**IPan:** 0,004					
Sieja:	iranas	42,4%	branduolinė programa	12,4%	rusija	11,4%	jt	7,7%
Išskiria:	iranas	23,7%	branduolinė programa	7,3%	krepšinis	5,8%	čempionatas	4,5%

Skirstinys 15	**Dydis:** 237	**VPan:** 0,070	**IPan:** 0,004					
Sieja:	finansai	67,5%	muzika	3,6%	aplinkosauga	2,5%	es	1,9%
Išskiria:	finansai	37,7%	krepšinis	5,7%	čempionatas	4,4%	rezultatai	4,0%
Skirstinys 16	**Dydis:** 233	**VPan:** 0,063	**IPan:** 0,004					
Sieja:	įmonės	55,7%	verslas	8,0%	paslaugos	7,3%	japonija	1,8%
Išskiria:	įmonės	32,2%	krepšinis	5,8%	čempionatas	4,5%	verslas	4,1%
Skirstinys 17	**Dydis:** 263	**VPan:** 0,034	**IPan:** 0,003					
Sieja:	terorizmas	20,9%	pakistanas	9,2%	afganistanas	8,6%	bušas	7,0%
Išskiria:	terorizmas	12,0%	krepšinis	5,8%	pakistanas	4,9%	afganistanas	4,8%
Skirstinys 18	**Dydis:** 304	**VPan:** 0,029	**IPan:** 0,002					
Sieja:	seimas	37,8%	paulauskas	7,5%	adamkus	6,5%	lenkija	4,2%
Išskiria:	seimas	20,7%	krepšinis	5,3%	čempionatas	4,1%	paulauskas	4,1%
Skirstinys 19	**Dydis:** 286	**VPan:** 0,029	**IPan:** 0,002					
Sieja:	vilnius	10,2%	sulaikymas	9,5%	dviračiai	8,9%	pasienis	6,4%
Išskiria:	krepšinis	5,2%	sulaikymas	5,0%	dviračiai	4,9%	vilnius	4,8%

K-MEANS, 8448 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 8448 **Skirstinių:** 5

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,452, **Grynumas:** 0,736

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	1600	0,084	0,061	0,002	0,002	0,087	0,976
1	1312	0,037	0,023	0,003	0,003	0,543	0,704
2	1646	0,026	0,017	0,002	0,002	0,507	0,605
3	2137	0,025	0,015	0,003	0,002	0,402	0,832
4	1753	0,015	0,013	0,002	0,002	0,724	0,548

Pasiskirstymas klasėse

Skirstinys	**Sportas**	**Politika**	**Ūkis**	**Kultūra**	**Teisėtvarka**
0	1562	6	5	21	6
1	16	923	275	49	49
2	3	996	15	40	592
3	31	196	1779	74	57
4	36	377	60	961	319

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 1600	**VPan:** 0,084	**IPan:** 0,002					
Sieja:	krepšinis	27,6%	čempionatas	20,9%	rezultatai	17,4%	lietuviai	7,6%
Išskiria:	krepšinis	14,9%	čempionatas	11,2%	rezultatai	8,7%	lietuviai	4,1%
Skirstinys 1	**Dydis:** 1312	**VPan:** 0,037	**IPan:** 0,003					
Sieja:	rinkimai	19,4%	baltarusija	12,5%	transportas	7,7%	palestiniečiai	7,6%
Išskiria:	rinkimai	10,9%	baltarusija	7,0%	krepšinis	6,5%	čempionatas	5,0%
Skirstinys 2	**Dydis:** 1646	**VPan:** 0,026	**IPan:** 0,002					
Sieja:	aukos	27,1%	irakas	12,6%	iranas	11,4%	branduolinė programa	3,1%
Išskiria:	aukos	15,5%	irakas	6,8%	krepšinis	6,7%	iranas	6,2%
Skirstinys 3	**Dydis:** 2137	**VPan:** 0,025	**IPan:** 0,003					
Sieja:	energetika	11,4%	ekonomika	11,1%	finansai	9,4%	rusija	9,2%
Išskiria:	krepšinis	7,8%	energetika	6,8%	ekonomika	6,6%	čempionatas	6,0%
Skirstinys 4	**Dydis:** 1753	**VPan:** 0,015	**IPan:** 0,002					
Sieja:	įvairybės	45,0%	britanija	9,8%	seimas	7,3%	kinas	2,3%
Išskiria:	įvairybės	25,6%	krepšinis	6,1%	britanija	4,6%	čempionatas	4,4%

K-MEANS, 8448 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 8448 **Skirstinių:** 10

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,398, **Grynumas:** 0,781

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	901	0,166	0,101	0,005	0,004	0,078	0,977
1	508	0,106	0,055	0,003	0,003	0,240	0,919
2	675	0,083	0,039	0,006	0,005	0,085	0,976
3	674	0,081	0,049	0,004	0,003	0,640	0,672
4	615	0,075	0,035	0,003	0,003	0,192	0,937
5	541	0,069	0,054	0,003	0,002	0,601	0,656
6	804	0,054	0,033	0,003	0,002	0,516	0,577
7	914	0,037	0,027	0,002	0,002	0,577	0,679
8	1427	0,030	0,017	0,003	0,002	0,302	0,889
9	1389	0,012	0,008	0,002	0,002	0,624	0,615

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	880	4	1	16	0
1	15	467	4	11	11
2	659	1	1	8	6
3	35	114	453	26	46
4	1	576	8	16	14
5	1	36	355	32	117
6	1	310	6	23	464
7	25	61	12	621	195
8	17	75	1268	44	23
9	14	854	26	348	147

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 901	**VPan:** 0,166	**IPan:** 0,005					
Sieja:	krepšinis	43,4%	čempionatas	22,1%	lietuviai	11,4%	rezultatai	9,9%
Išskiria:	krepšinis	24,9%	čempionatas	9,6%	lietuviai	6,4%	įvairybės	2,8%
Skirstinys 1	**Dydis:** 508	**VPan:** 0,106	**IPan:** 0,003					
Sieja:	rinkimai	43,5%	baltarusija	29,5%	prezidentas	7,2%	parlamentas	6,5%
Išskiria:	rinkimai	23,7%	baltarusija	16,3%	krepšinis	5,2%	čempionatas	4,4%
Skirstinys 2	**Dydis:** 675	**VPan:** 0,083	**IPan:** 0,006					
Sieja:	futbolas	24,9%	rezultatai	16,0%	ledo ritulys	12,6%	olimpiada	8,7%
Išskiria:	futbolas	15,2%	ledo ritulys	7,9%	krepšinis	6,8%	olimpiada	5,2%
Skirstinys 3	**Dydis:** 674	**VPan:** 0,081	**IPan:** 0,004					
Sieja:	energetika	34,8%	rusija	28,6%	dujos	8,5%	nafta	8,5%
Išskiria:	energetika	20,6%	rusija	11,6%	krepšinis	6,4%	dujos	5,1%
Skirstinys 4	**Dydis:** 615	**VPan:** 0,075	**IPan:** 0,003					
Sieja:	iranas	27,3%	izraelis	15,7%	palestiniečiai	14,5%	branduolinė programa	7,8%
Išskiria:	iranas	14,7%	izraelis	8,5%	palestiniečiai	7,8%	krepšinis	5,7%
Skirstinys 5	**Dydis:** 541	**VPan:** 0,069	**IPan:** 0,003					
Sieja:	transportas	29,4%	eismas	11,8%	keliai	9,5%	sąlygos	8,8%
Išskiria:	transportas	16,5%	eismas	6,4%	krepšinis	5,7%	keliai	5,4%
Skirstinys 6	**Dydis:** 804	**VPan:** 0,054	**IPan:** 0,003					
Sieja:	aukos	54,0%	sprogimas	5,5%	terorizmas	3,6%	incidentas	2,9%
Išskiria:	aukos	30,3%	krepšinis	5,9%	čempionatas	4,5%	rezultatai	4,0%
Skirstinys 7	**Dydis:** 914	**VPan:** 0,037	**IPan:** 0,002					
Sieja:	įvairybės	68,0%	kinas	3,5%	muzika	2,7%	sulaikymas	2,3%
Išskiria:	įvairybės	37,3%	krepšinis	5,6%	čempionatas	4,0%	rezultatai	3,9%
Skirstinys 8	**Dydis:** 1427	**VPan:** 0,030	**IPan:** 0,003					
Sieja:	ekonomika	19,0%	finansai	15,8%	es	11,0%	prekyba	6,4%

Išskiria:	ekonomika	11,0%	finansai	8,9%	krepšinis	6,8%	čempionatas	5,2%
Skirstinys 9	**Dydis:** 1389	**VPan:** 0,012	**IPan:** 0,002					
Sieja:	irakas	16,8%	seimas	16,4%	vizitas	4,1%	bušas	3,9%
Išskiria:	seimas	9,6%	irakas	7,7%	krepšinis	6,3%	čempionatas	4,8%

K-MEANS, 8448 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 8448 **Skirstinių:** 15

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,358, **Grynumas:** 0,804

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	876	0,177	0,102	0,005	0,003	0,050	0,985
1	308	0,165	0,098	0,003	0,003	0,207	0,903
2	423	0,130	0,072	0,005	0,003	0,044	0,988
3	370	0,117	0,063	0,006	0,006	0,314	0,886
4	334	0,112	0,053	0,004	0,003	0,160	0,940
5	373	0,110	0,059	0,003	0,003	0,293	0,885
6	487	0,106	0,055	0,004	0,003	0,262	0,910
7	378	0,105	0,062	0,003	0,003	0,232	0,915
8	562	0,092	0,059	0,004	0,003	0,579	0,694
9	490	0,053	0,033	0,003	0,003	0,536	0,676
10	824	0,051	0,028	0,003	0,003	0,270	0,891
11	719	0,049	0,029	0,003	0,002	0,500	0,623
12	741	0,047	0,033	0,002	0,002	0,238	0,919
13	546	0,038	0,024	0,004	0,003	0,663	0,619
14	1017	0,015	0,009	0,002	0,002	0,683	0,538

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	863	1	0	12	0
1	0	1	278	0	29
2	418	0	0	4	1
3	328	8	7	15	12
4	0	314	2	1	17
5	2	330	5	28	8
6	14	443	4	13	13
7	0	346	6	8	18
8	4	99	390	25	44
9	3	22	331	124	10
10	1	65	734	11	13
11	0	244	6	21	448
12	7	23	8	681	22
13	3	55	338	38	112
14	5	547	25	164	276

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 876	**VPan:** 0,177	**IPan:** 0,005					
Sieja:	krepšinis	43,1%	čempionatas	23,6%	lietuviai	11,1%	rezultatai	9,8%
Išskiria:	krepšinis	24,7%	čempionatas	10,8%	lietuviai	6,1%	įvairybės	2,8%
Skirstinys 1	**Dydis:** 308	**VPan:** 0,165	**IPan:** 0,003					
Sieja:	transportas	35,0%	eismas	15,3%	keliai	12,1%	sąlygos	11,4%
Išskiria:	transportas	18,5%	eismas	8,0%	keliai	6,5%	sąlygos	6,1%
Skirstinys 2	**Dydis:** 423	**VPan:** 0,130	**IPan:** 0,005					
Sieja:	rezultatai	22,7%	ledo ritulys	20,5%	turinas	15,7%	olimpiada	15,6%
Išskiria:	ledo ritulys	11,8%	turinas	9,0%	olimpiada	8,9%	krepšinis	5,8%

Skirstinys 3	**Dydis:** 370	**VPan:** 0,117	**IPan:** 0,006					
Sieja:	futbolas	64,9%	čempionatas	6,3%	statistika	5,8%	anglija	2,5%
Išskiria:	futbolas	39,9%	krepšinis	6,2%	rezultatai	2,9%	rusija	2,4%
Skirstinys 4	**Dydis:** 334	**VPan:** 0,112	**IPan:** 0,004					
Sieja:	irakas	77,6%	aukos	4,6%	išpuolis	2,6%	bušas	1,8%
Išskiria:	irakas	43,4%	krepšinis	5,5%	čempionatas	4,2%	rezultatai	3,4%
Skirstinys 5	**Dydis:** 373	**VPan:** 0,110	**IPan:** 0,003					
Sieja:	palestiniečiai	30,2%	izraelis	29,7%	hamas	13,5%	šaronas	6,4%
Išskiria:	palestiniečiai	16,2%	izraelis	15,8%	hamas	7,4%	krepšinis	5,4%
Skirstinys 6	**Dydis:** 487	**VPan:** 0,106	**IPan:** 0,004					
Sieja:	rinkimai	39,6%	baltarusija	32,0%	prezidentas	8,4%	parlamentas	5,1%
Išskiria:	rinkimai	20,8%	baltarusija	17,7%	krepšinis	5,2%	čempionatas	4,4%
Skirstinys 7	**Dydis:** 378	**VPan:** 0,105	**IPan:** 0,003					
Sieja:	iranas	51,5%	branduolinė programa	14,8%	jt	7,4%	derybos	5,7%
Išskiria:	iranas	27,9%	branduolinė programa	8,3%	krepšinis	5,5%	čempionatas	4,2%
Skirstinys 8	**Dydis:** 562	**VPan:** 0,092	**IPan:** 0,004					
Sieja:	energetika	39,9%	rusija	20,0%	dujos	10,7%	nafta	10,1%
Išskiria:	energetika	22,7%	dujos	6,2%	krepšinis	6,1%	rusija	6,0%
Skirstinys 9	**Dydis:** 490	**VPan:** 0,053	**IPan:** 0,003					
Sieja:	prekyba	32,6%	žemės ūkis	24,5%	paukščių gripas	16,6%	popiežius	2,4%
Išskiria:	prekyba	18,0%	žemės ūkis	14,3%	paukščių gripas	9,6%	krepšinis	5,8%
Skirstinys 10	**Dydis:** 824	**VPan:** 0,051	**IPan:** 0,003					
Sieja:	ekonomika	31,8%	finansai	25,0%	es	12,3%	statistika	3,1%
Išskiria:	ekonomika	18,2%	finansai	13,8%	krepšinis	6,3%	es	4,9%
Skirstinys 11	**Dydis:** 719	**VPan:** 0,049	**IPan:** 0,003					
Sieja:	aukos	48,5%	sprogimas	5,6%	terorizmas	4,8%	incidentas	3,8%
Išskiria:	aukos	25,2%	krepšinis	6,0%	čempionatas	4,6%	rezultatai	4,2%
Skirstinys 12	**Dydis:** 741	**VPan:** 0,047	**IPan:** 0,002					
Sieja:	įvairybės	74,9%	kinas	4,2%	muzika	3,2%	britanija	2,5%
Išskiria:	įvairybės	40,0%	krepšinis	5,4%	čempionatas	4,2%	rezultatai	3,8%
Skirstinys 13	**Dydis:** 546	**VPan:** 0,038	**IPan:** 0,004					
Sieja:	įmonės	49,5%	verslas	10,4%	vokietija	5,4%	japonija	3,7%
Išskiria:	įmonės	29,2%	krepšinis	6,1%	verslas	5,6%	čempionatas	4,7%
Skirstinys 14	**Dydis:** 1017	**VPan:** 0,015	**IPan:** 0,002					
Sieja:	seimas	23,7%	britanija	5,6%	adamkus	4,9%	sulaikymas	4,8%
Išskiria:	seimas	13,3%	krepšinis	5,8%	čempionatas	4,4%	rezultatai	4,0%

K-MEANS, 8448 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: K-means **Dokumentų:** 8448 **Skirstinių:** 20

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,338, **Grynumas:** 0,819

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	580	0,260	0,141	0,010	0,005	0,054	0,983
1	257	0,188	0,074	0,004	0,003	0,243	0,918
2	301	0,170	0,100	0,003	0,003	0,183	0,914
3	361	0,169	0,078	0,005	0,003	0,030	0,992
4	313	0,160	0,076	0,008	0,005	0,188	0,936
5	298	0,149	0,072	0,004	0,003	0,173	0,943
6	340	0,135	0,065	0,006	0,006	0,314	0,882
7	337	0,122	0,060	0,003	0,002	0,256	0,908
8	495	0,110	0,065	0,005	0,003	0,545	0,719
9	297	0,093	0,056	0,006	0,004	0,423	0,825
10	502	0,090	0,043	0,002	0,002	0,218	0,928
11	374	0,085	0,039	0,004	0,003	0,276	0,885
12	419	0,081	0,041	0,004	0,003	0,408	0,819

13	534	0,074	0,036	0,004	0,002	0,381	0,762
14	729	0,059	0,030	0,004	0,003	0,230	0,911
15	370	0,055	0,027	0,001	0,002	0,157	0,949
16	474	0,049	0,028	0,004	0,003	0,300	0,888
17	338	0,037	0,028	0,003	0,002	0,893	0,331
18	595	0,021	0,014	0,001	0,001	0,632	0,526
19	534	0,022	0,014	0,002	0,002	0,690	0,543

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	570	0	0	10	0
1	5	236	3	5	8
2	0	0	275	0	26
3	358	0	0	3	0
4	293	3	2	13	2
5	0	281	7	3	7
6	300	7	3	18	12
7	2	306	5	17	7
8	2	82	356	23	32
9	15	245	3	21	13
10	5	14	6	466	11
11	1	5	331	33	4
12	1	343	8	29	38
13	0	119	2	6	407
14	1	49	664	7	8
15	0	351	4	11	4
16	2	16	421	12	23
17	87	112	12	41	86
18	5	39	15	313	223
19	1	290	17	114	112

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 580	**VPan:** 0,260	**IPan:** 0,010					
Sieja:	čempionatas	36,3%	krepšinis	33,4%	rezultatai	13,2%	lietuviai	12,3%
Išskiria:	čempionatas	19,4%	krepšinis	14,2%	lietuviai	6,3%	rezultatai	3,1%
Skirstinys 1	**Dydis:** 257	**VPan:** 0,188	**IPan:** 0,004					
Sieja:	baltarusija	63,9%	rinkimai	18,8%	lukašenka	5,2%	opozicija	2,8%
Išskiria:	baltarusija	35,0%	rinkimai	6,7%	krepšinis	5,4%	čempionatas	4,2%
Skirstinys 2	**Dydis:** 301	**VPan:** 0,170	**IPan:** 0,003					
Sieja:	transportas	35,2%	eismas	15,5%	keliai	12,5%	sąlygos	11,5%
Išskiria:	transportas	18,6%	eismas	8,0%	keliai	6,7%	sąlygos	6,2%
Skirstinys 3	**Dydis:** 361	**VPan:** 0,169	**IPan:** 0,005					
Sieja:	rezultatai	23,7%	ledo ritulys	21,6%	turinas	16,5%	olimpiada	16,3%
Išskiria:	ledo ritulys	12,3%	turinas	9,4%	olimpiada	9,2%	krepšinis	5,8%
Skirstinys 4	**Dydis:** 313	**VPan:** 0,160	**IPan:** 0,008					
Sieja:	eurolyga	33,3%	krepšinis	30,4%	laureatai	6,9%	žalgiris	3,8%
Išskiria:	eurolyga	21,3%	krepšinis	6,7%	čempionatas	5,0%	laureatai	4,1%
Skirstinys 5	**Dydis:** 298	**VPan:** 0,149	**IPan:** 0,004					
Sieja:	iranas	54,7%	branduolinė programa	16,8%	jt	6,8%	derybos	6,2%
Išskiria:	iranas	29,2%	branduolinė programa	9,3%	krepšinis	5,4%	čempionatas	4,2%
Skirstinys 6	**Dydis:** 340	**VPan:** 0,135	**IPan:** 0,006					
Sieja:	futbolas	66,2%	čempionatas	6,5%	statistika	5,8%	taurė	3,0%
Išskiria:	futbolas	40,4%	krepšinis	6,2%	rezultatai	2,9%	rusija	2,4%
Skirstinys 7	**Dydis:** 337	**VPan:** 0,122	**IPan:** 0,003					
Sieja:	izraelis	30,9%	palestiniečiai	29,0%	hamas	13,7%	šaronas	7,0%
Išskiria:	izraelis	16,3%	palestiniečiai	15,3%	hamas	7,3%	krepšinis	5,3%
Skirstinys 8	**Dydis:** 495	**VPan:** 0,110	**IPan:** 0,005					
Sieja:	energetika	41,8%	rusija	17,4%	dujos	11,5%	nafta	10,7%

Išskiria:	energetika	23,6%	dujos	6,6%	nafta	6,0%	krepšinis	6,0%
Skirstinys 9 **Dydis:** 297 **VPan:** 0,093 **IPan:** 0,006								
Sieja:	rinkimai	40,3%	parlamentas	22,6%	ukraina	10,7%	prezidentas	8,6%
Išskiria:	rinkimai	19,3%	parlamentas	13,6%	krepšinis	4,7%	čempionatas	4,6%
Skirstinys 10 **Dydis:** 502 **VPan:** 0,090 **IPan:** 0,002								
Sieja:	įvairybės	85,2%	kinas	4,1%	britanija	2,7%	anekdotai	1,1%
Išskiria:	įvairybės	45,4%	krepšinis	5,4%	čempionatas	4,1%	rezultatai	3,8%
Skirstinys 11 **Dydis:** 374 **VPan:** 0,085 **IPan:** 0,004								
Sieja:	prekyba	36,9%	žemės ūkis	26,3%	paukščių gripas	17,5%	kainos	2,2%
Išskiria:	prekyba	20,6%	žemės ūkis	15,3%	paukščių gripas	10,1%	krepšinis	5,8%
Skirstinys 12 **Dydis:** 419 **VPan:** 0,081 **IPan:** 0,004								
Sieja:	irakas	64,5%	britanija	11,1%	aukos	4,1%	išpuolis	2,6%
Išskiria:	irakas	36,8%	krepšinis	5,8%	čempionatas	4,4%	britanija	4,2%
Skirstinys 13 **Dydis:** 534 **VPan:** 0,074 **IPan:** 0,004								
Sieja:	aukos	57,4%	sprogimas	5,9%	incidentas	4,9%	katastriša	4,3%
Išskiria:	aukos	29,8%	krepšinis	5,9%	čempionatas	4,5%	rezultatai	4,1%
Skirstinys 14 **Dydis:** 729 **VPan:** 0,059 **IPan:** 0,004								
Sieja:	ekonomika	34,5%	finansai	27,0%	es	11,2%	statistika	2,6%
Išskiria:	ekonomika	19,7%	finansai	14,8%	krepšinis	6,2%	čempionatas	4,7%
Skirstinys 15 **Dydis:** 370 **VPan:** 0,055 **IPan:** 0,001								
Sieja:	seimas	53,0%	adamkus	10,7%	paulauskas	7,6%	partijos	1,7%
Išskiria:	seimas	28,0%	adamkus	5,6%	krepšinis	5,1%	paulauskas	4,0%
Skirstinys 16 **Dydis:** 474 **VPan:** 0,049 **IPan:** 0,004								
Sieja:	įmonės	53,9%	verslas	9,7%	paslaugos	7,2%	vokietija	3,5%
Išskiria:	įmonės	31,4%	krepšinis	5,9%	verslas	5,1%	čempionatas	4,6%
Skirstinys 17 **Dydis:** 338 **VPan:** 0,037 **IPan:** 0,003								
Sieja:	prancūzija	28,3%	dviračiai	15,3%	lenktynės	11,4%	miloševičius	7,4%
Išskiria:	prancūzija	14,4%	dviračiai	9,0%	lenktynės	6,7%	krepšinis	5,3%
Skirstinys 18 **Dydis:** 595 **VPan:** 0,021 **IPan:** 0,001								
Sieja:	sulaikymas	10,6%	pasienis	9,2%	vilnius	8,8%	muzika	8,0%
Išskiria:	sulaikymas	5,5%	krepšinis	5,1%	pasienis	4,9%	kontrabanda	4,3%
Skirstinys 19 **Dydis:** 534 **VPan:** 0,022 **IPan:** 0,002								
Sieja:	bušas	8,3%	drebėjimas	7,5%	žemės	7,0%	karikatūros	6,0%
Išskiria:	krepšinis	5,7%	čempionatas	4,4%	drebėjimas	4,4%	bušas	4,2%

PRIEDAS C – EKSPERIMENTŲ ŽURNALAS

Šiame priede pateikiamas individualiai suklasifikuotų dokumentų kolekcijų klasterizavimo skaldančių K-means algoritmu eksperimento žurnalas. Klasterizavimo metu dokumentų reprezentacijai naudotos individualiai vartotojų priskirtos žymės.

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 5

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,485, **Grynumas:** 0,691

Skirstinys	Dydis[34]	VPan[35]	VNuok[36]	IPan[37]	INuok[38]	Entropija	Grynumas
0	201	0,111	0,078	0,002	0,003	0,225	0,915
1	194	0,050	0,032	0,003	0,003	0,743	0,423
2	187	0,048	0,028	0,003	0,003	0,458	0,583
3	250	0,039	0,018	0,004	0,004	0,527	0,760
4	216	0,034	0,019	0,003	0,003	0,470	0,736

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	1	10	0	6	184
1	10	16	80	82	6
2	109	76	1	1	0
3	29	8	190	15	8
4	159	12	1	43	1

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 201	**VPan:** 0,111	**IPan:** 0,002					
Sieja:	krepšinis	29,6%	čempionatas	21,2%	rezultatai	17,8%	lietuviai	7,5%
Išskiria:	krepšinis	15,8%	čempionatas	11,3%	rezultatai	8,6%	įvairybės	4,1%
Skirstinys 1	**Dydis:** 194	**VPan:** 0,050	**IPan:** 0,003					
Sieja:	įvairybės	38,8%	transportas	11,4%	įmonės	7,0%	britanija	6,0%
Išskiria:	įvairybės	21,6%	transportas	6,4%	krepšinis	6,2%	čempionatas	4,5%
Skirstinys 2	**Dydis:** 187	**VPan:** 0,048	**IPan:** 0,003					
Sieja:	aukos	34,0%	irakas	14,1%	iranas	5,7%	sprogimas	3,6%
Išskiria:	aukos	19,3%	irakas	7,5%	krepšinis	6,3%	čempionatas	4,5%
Skirstinys 3	**Dydis:** 250	**VPan:** 0,039	**IPan:** 0,004					
Sieja:	energetika	12,1%	ekonomika	11,9%	rusija	10,6%	finansai	9,9%
Išskiria:	krepšinis	7,2%	energetika	7,1%	ekonomika	6,8%	finansai	5,7%
Skirstinys 4	**Dydis:** 216	**VPan:** 0,034	**IPan:** 0,003					
Sieja:	rinkimai	15,3%	seimas	10,4%	palestiniečiai	9,3%	baltarusija	8,6%
Išskiria:	rinkimai	8,2%	krepšinis	6,1%	seimas	5,6%	palestiniečiai	5,2%

34 Skirstiniui priskirtų dokumentų skaičius.

35 Vidutinis panašumas tarp skirstinyje esančių dokumentų (**V**idinis **Pan**ašumas).

36 Skirstinyje esančių dokumentų panašumų standartinis nuokrypis (**V**idinis **Nuok**rypis).

37 Vidutinis panašumas tarp skirstinyje esančių dokumentų ir likusių kolekcijos dokumentų (**I**šorinis **Pan**ašumas).

38 Skirstinyje esančių dokumentų panašumų su likusiais kolekcijos dokumentais standartinis nuokrypis (**I**šorinis **Nuok**rypis).

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 10

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,421, **Grynumas:** 0,740

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	142	0,178	0,101	0,004	0,004	0,046	0,986
1	76	0,158	0,088	0,004	0,004	0,498	0,776
2	63	0,130	0,077	0,005	0,004	0,412	0,746
3	91	0,090	0,049	0,006	0,004	0,691	0,615
4	98	0,086	0,033	0,004	0,004	0,260	0,888
5	83	0,074	0,041	0,004	0,004	0,727	0,554
6	110	0,072	0,036	0,003	0,003	0,287	0,891
7	148	0,063	0,036	0,003	0,003	0,467	0,534
8	125	0,059	0,031	0,003	0,004	0,568	0,632
9	112	0,055	0,031	0,003	0,004	0,452	0,759

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	2	140
1	7	6	1	59	3
2	0	2	0	14	47
3	18	10	56	4	3
4	9	0	87	1	1
5	10	3	46	21	3
6	98	7	1	3	1
7	79	67	0	2	0
8	2	7	79	36	1
9	85	20	2	5	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 142	**VPan:** 0,178	**IPan:** 0,004					
Sieja:	krepšinis	37,1%	čempionatas	26,6%	lietuviai	9,4%	rezultatai	8,4%
Išskiria:	krepšinis	20,3%	čempionatas	14,5%	lietuviai	5,1%	įvairybės	3,9%
Skirstinys 1	**Dydis:** 76	**VPan:** 0,158	**IPan:** 0,004					
Sieja:	įvairybės	78,0%	britanija	9,1%	tyrimas	2,1%	kalėdos	1,6%
Išskiria:	įvairybės	41,9%	krepšinis	5,4%	britanija	4,5%	čempionatas	3,9%
Skirstinys 2	**Dydis:** 63	**VPan:** 0,130	**IPan:** 0,005					
Sieja:	rezultatai	23,1%	ledo ritulys	16,1%	olimpiada	15,0%	turinas	14,4%
Išskiria:	ledo ritulys	9,0%	olimpiada	8,3%	turinas	8,2%	rezultatai	5,6%
Skirstinys 3	**Dydis:** 91	**VPan:** 0,090	**IPan:** 0,006					
Sieja:	energetika	37,8%	rusija	20,9%	nafta	8,1%	dujos	6,4%
Išskiria:	energetika	22,0%	rusija	6,1%	krepšinis	6,0%	nafta	4,7%
Skirstinys 4	**Dydis:** 98	**VPan:** 0,086	**IPan:** 0,004					
Sieja:	ekonomika	31,0%	finansai	26,9%	es	15,2%	paslaugos	1,9%
Išskiria:	ekonomika	16,9%	finansai	14,9%	es	6,5%	krepšinis	5,8%
Skirstinys 5	**Dydis:** 83	**VPan:** 0,074	**IPan:** 0,004					
Sieja:	prekyba	24,4%	žemės ūkis	20,2%	paukščių gripas	18,5%	prancūzija	4,2%
Išskiria:	prekyba	14,0%	žemės ūkis	11,6%	paukščių gripas	10,6%	krepšinis	5,6%
Skirstinys 6	**Dydis:** 110	**VPan:** 0,072	**IPan:** 0,003					
Sieja:	rinkimai	24,3%	seimas	19,0%	baltarusija	15,6%	prezidentas	4,8%
Išskiria:	rinkimai	12,6%	seimas	10,2%	baltarusija	8,5%	krepšinis	5,6%
Skirstinys 7	**Dydis:** 148	**VPan:** 0,063	**IPan:** 0,003					
Sieja:	aukos	41,5%	irakas	16,4%	sprogimas	4,3%	išpuolis	3,7%
Išskiria:	aukos	23,3%	irakas	8,6%	krepšinis	5,9%	čempionatas	4,2%
Skirstinys 8	**Dydis:** 125	**VPan:** 0,059	**IPan:** 0,003					
Sieja:	transportas	23,5%	įmonės	18,7%	vokietija	6,0%	muzika	4,8%

Išskiria:	transportas	13,2%	įmonės	9,1%	krepšinis	5,7%	čempionatas	4,1%
Skirstinys 9 **Dydis:** 112 **VPan:** 0,055 **IPan:** 0,003								
Sieja:	palestiniečiai	20,0%	iranas	12,5%	hamas	11,5%	izraelis	9,3%
Išskiria:	palestiniečiai	11,2%	hamas	6,5%	iranas	6,4%	krepšinis	5,6%

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 15

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,361, **Grynumas:** 0,806

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	98	0,290	0,119	0,007	0,005	0,062	0,980
1	70	0,180	0,093	0,004	0,004	0,443	0,800
2	54	0,169	0,084	0,006	0,004	0,261	0,852
3	53	0,169	0,070	0,006	0,004	0,313	0,887
4	53	0,139	0,067	0,004	0,003	0,215	0,925
5	56	0,128	0,061	0,005	0,005	0,308	0,804
6	51	0,127	0,051	0,007	0,006	0,357	0,863
7	79	0,108	0,054	0,006	0,004	0,570	0,722
8	87	0,097	0,044	0,004	0,004	0,279	0,885
9	61	0,093	0,046	0,002	0,003	0,298	0,869
10	69	0,095	0,050	0,005	0,004	0,635	0,667
11	97	0,088	0,033	0,004	0,004	0,262	0,887
12	55	0,084	0,042	0,003	0,004	0,568	0,582
13	93	0,084	0,041	0,005	0,004	0,430	0,656
14	72	0,052	0,020	0,002	0,003	0,496	0,694

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	2	96
1	6	5	0	56	3
2	0	0	0	8	46
3	47	1	1	2	2
4	49	2	1	1	0
5	45	11	0	0	0
6	1	3	1	2	44
7	12	4	57	2	4
8	1	7	77	2	0
9	53	4	0	4	0
10	4	3	46	13	3
11	9	0	86	1	1
12	32	19	1	3	0
13	31	61	0	1	0
14	18	2	2	50	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 98 **VPan:** 0,290 **IPan:** 0,007								
Sieja:	krepšinis	46,9%	čempionatas	20,3%	lietuviai	11,7%	rezultatai	9,7%
Išskiria:	krepšinis	26,3%	čempionatas	8,3%	lietuviai	6,5%	įvairybės	3,9%
Skirstinys 1 **Dydis:** 70 **VPan:** 0,180 **IPan:** 0,004								
Sieja:	įvairybės	80,9%	britanija	7,1%	tyrimas	2,1%	kalėdos	1,6%
Išskiria:	įvairybės	43,5%	krepšinis	5,4%	čempionatas	3,8%	rezultatai	3,7%
Skirstinys 2 **Dydis:** 54 **VPan:** 0,169 **IPan:** 0,006								
Sieja:	rezultatai	24,2%	ledo ritulys	16,9%	olimpiada	15,7%	turinas	15,0%
Išskiria:	ledo ritulys	9,4%	olimpiada	8,7%	turinas	8,5%	rezultatai	6,0%

Skirstinys 3	**Dydis:** 53	**VPan:** 0,169	**IPan:** 0,006					
Sieja:	rinkimai	42,2%	baltarusija	29,0%	parlamentas	6,2%	ukraina	5,3%
Išskiria:	rinkimai	22,4%	baltarusija	16,2%	krepšinis	4,8%	čempionatas	3,9%
Skirstinys 4	**Dydis:** 53	**VPan:** 0,139	**IPan:** 0,004					
Sieja:	palestiniečiai	35,8%	hamas	20,6%	izraelis	16,6%	vyriausybė	4,3%
Išskiria:	palestiniečiai	19,3%	hamas	11,2%	izraelis	8,9%	krepšinis	5,2%
Skirstinys 5	**Dydis:** 56	**VPan:** 0,128	**IPan:** 0,005					
Sieja:	irakas	58,6%	aukos	6,3%	išpuolis	4,6%	žemės	4,5%
Išskiria:	irakas	32,4%	krepšinis	5,5%	čempionatas	3,9%	rezultatai	3,2%
Skirstinys 6	**Dydis:** 51	**VPan:** 0,127	**IPan:** 0,007					
Sieja:	futbolas	38,4%	čempionatas	15,3%	italija	5,4%	rankinis	4,8%
Išskiria:	futbolas	23,3%	krepšinis	5,9%	rusija	3,5%	įvairybės	3,2%
Skirstinys 7	**Dydis:** 79	**VPan:** 0,108	**IPan:** 0,006					
Sieja:	energetika	39,7%	rusija	18,1%	nafta	9,0%	dujos	7,1%
Išskiria:	energetika	22,8%	krepšinis	5,9%	nafta	5,2%	rusija	4,6%
Skirstinys 8	**Dydis:** 87	**VPan:** 0,097	**IPan:** 0,004					
Sieja:	transportas	29,3%	įmonės	23,3%	aviacija	5,9%	vokietija	5,1%
Išskiria:	transportas	16,5%	įmonės	11,5%	krepšinis	5,6%	čempionatas	4,0%
Skirstinys 9	**Dydis:** 61	**VPan:** 0,093	**IPan:** 0,002					
Sieja:	seimas	47,8%	adamkus	8,4%	liberalai	5,5%	pranešimas	3,7%
Išskiria:	seimas	25,0%	krepšinis	5,0%	adamkus	4,5%	čempionatas	3,6%
Skirstinys 10	**Dydis:** 69	**VPan:** 0,095	**IPan:** 0,005					
Sieja:	prekyba	27,3%	žemės ūkis	22,6%	paukščių gripas	20,6%	prancūzija	4,0%
Išskiria:	prekyba	15,6%	žemės ūkis	12,9%	paukščių gripas	11,8%	krepšinis	5,5%
Skirstinys 11	**Dydis:** 97	**VPan:** 0,088	**IPan:** 0,004					
Sieja:	ekonomika	31,1%	finansai	27,1%	es	15,3%	paslaugos	1,9%
Išskiria:	ekonomika	17,0%	finansai	14,9%	es	6,5%	krepšinis	5,8%
Skirstinys 12	**Dydis:** 55	**VPan:** 0,084	**IPan:** 0,003					
Sieja:	iranas	31,6%	branduolinė programa	10,7%	sulaikymas	7,7%	pasienis	6,3%
Išskiria:	iranas	16,3%	branduolinė programa	5,9%	krepšinis	5,2%	sulaikymas	4,2%
Skirstinys 13	**Dydis:** 93	**VPan:** 0,084	**IPan:** 0,005					
Sieja:	aukos	45,6%	sprogimas	8,2%	incidentas	7,0%	gaisras	6,0%
Išskiria:	aukos	22,7%	krepšinis	5,8%	sprogimas	4,7%	čempionatas	4,2%
Skirstinys 14	**Dydis:** 72	**VPan:** 0,052	**IPan:** 0,002					
Sieja:	muzika	12,9%	kinas	12,0%	vizitas	7,7%	ministras	6,8%
Išskiria:	muzika	6,8%	kinas	6,4%	krepšinis	5,1%	čempionatas	3,7%

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 20

Suklasterizuota dokumentų: [1048 iš 1056], **Entropija:** 0,320, **Grynumas:** 0,818

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	63	0,455	0,148	0,014	0,005	0,087	0,968
1	35	0,271	0,078	0,012	0,006	0,000	1,000
2	41	0,211	0,061	0,004	0,004	0,163	0,927
3	36	0,210	0,112	0,006	0,004	0,133	0,944
4	68	0,187	0,096	0,004	0,004	0,408	0,824
5	43	0,182	0,076	0,006	0,005	0,069	0,977
6	49	0,172	0,054	0,006	0,004	0,000	1,000
7	54	0,167	0,068	0,006	0,004	0,349	0,870
8	49	0,158	0,068	0,004	0,004	0,167	0,939
9	34	0,153	0,077	0,005	0,003	0,220	0,912
10	46	0,148	0,050	0,008	0,006	0,065	0,978
11	48	0,133	0,055	0,003	0,003	0,126	0,958
12	51	0,131	0,048	0,005	0,004	0,404	0,784

13	46	0,126	0,075	0,004	0,005	0,745	0,391
14	70	0,124	0,060	0,006	0,004	0,462	0,771
15	49	0,113	0,073	0,003	0,004	0,695	0,551
16	66	0,100	0,048	0,007	0,005	0,324	0,848
17	90	0,091	0,040	0,005	0,004	0,380	0,700
18	53	0,067	0,027	0,002	0,002	0,561	0,717
19	57	0,056	0,022	0,004	0,003	0,663	0,544

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	2	61
1	0	0	0	0	35
2	0	3	38	0	0
3	0	0	0	2	34
4	4	5	0	56	3
5	42	1	0	0	0
6	0	0	49	0	0
7	47	2	1	2	2
8	46	2	0	1	0
9	31	0	2	1	0
10	0	0	0	1	45
11	46	1	0	1	0
12	9	0	40	1	1
13	3	18	0	7	18
14	10	4	54	2	0
15	5	4	27	13	0
16	1	1	56	8	0
17	27	63	0	0	0
18	6	5	4	38	0
19	31	13	1	12	0

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 63	**VPan:** 0,455	**IPan:** 0,014					
Sieja:	krepšinis	37,3%	čempionatas	31,3%	rezultatai	14,1%	lietuviai	14,1%
Išskiria:	krepšinis	16,2%	čempionatas	15,0%	lietuviai	7,7%	įvairybės	4,0%
Skirstinys 1	**Dydis:** 35	**VPan:** 0,271	**IPan:** 0,012					
Sieja:	eurolyga	37,3%	krepšinis	31,3%	moterys	4,3%	žalgiris	3,7%
Išskiria:	eurolyga	23,1%	krepšinis	6,9%	čempionatas	4,3%	rusija	3,5%
Skirstinys 2	**Dydis:** 41	**VPan:** 0,211	**IPan:** 0,004					
Sieja:	transportas	51,8%	aviacija	12,2%	sąlygos	6,0%	keliai	4,5%
Išskiria:	transportas	27,2%	aviacija	6,5%	krepšinis	5,1%	čempionatas	3,6%
Skirstinys 3	**Dydis:** 36	**VPan:** 0,210	**IPan:** 0,006					
Sieja:	ledo ritulys	27,6%	rezultatai	27,2%	zubrus	9,5%	tenisas	8,0%
Išskiria:	ledo ritulys	15,3%	rezultatai	6,5%	zubrus	5,4%	krepšinis	5,4%
Skirstinys 4	**Dydis:** 68	**VPan:** 0,187	**IPan:** 0,004					
Sieja:	įvairybės	82,6%	britanija	5,2%	tyrimas	2,2%	kalėdos	1,7%
Išskiria:	įvairybės	44,4%	krepšinis	5,4%	čempionatas	3,8%	rezultatai	3,7%
Skirstinys 5	**Dydis:** 43	**VPan:** 0,182	**IPan:** 0,006					
Sieja:	irakas	70,0%	aukos	5,8%	išpuolis	5,5%	sirija	3,5%
Išskiria:	irakas	38,6%	krepšinis	5,4%	čempionatas	3,8%	rezultatai	3,2%
Skirstinys 6	**Dydis:** 49	**VPan:** 0,172	**IPan:** 0,006					
Sieja:	ekonomika	72,0%	paslaugos	2,7%	rūpyba	2,0%	euras	2,0%
Išskiria:	ekonomika	40,5%	krepšinis	5,4%	čempionatas	3,9%	rezultatai	3,8%
Skirstinys 7	**Dydis:** 54	**VPan:** 0,167	**IPan:** 0,006					
Sieja:	rinkimai	41,0%	baltarusija	28,2%	ukraina	7,6%	parlamentas	6,0%
Išskiria:	rinkimai	21,7%	baltarusija	15,7%	krepšinis	4,8%	čempionatas	3,9%
Skirstinys 8	**Dydis:** 49	**VPan:** 0,158	**IPan:** 0,004					
Sieja:	palestiniečiai	36,8%	hamas	21,2%	izraelis	17,1%	vyriausybė	4,4%

Išskiria:	palestiniečiai	19,7%	hamas	11,5%	izraelis	9,1%	krepšinis	5,1%
Skirstinys 9 **Dydis:** 34 **VPan:** 0,153 **IPan:** 0,005								
Sieja:	iranas	45,7%	branduolinė programa	15,5%	derybos	9,0%	korėja	3,6%
Išskiria:	iranas	24,0%	branduolinė programa	8,6%	krepšinis	5,2%	derybos	4,7%
Skirstinys 10 **Dydis:** 46 **VPan:** 0,148 **IPan:** 0,008								
Sieja:	futbolas	40,5%	čempionatas	16,1%	rankinis	5,1%	dviračiai	5,1%
Išskiria:	futbolas	24,6%	krepšinis	5,9%	rusija	3,4%	įvairybės	3,2%
Skirstinys 11 **Dydis:** 48 **VPan:** 0,133 **IPan:** 0,003								
Sieja:	seimas	54,3%	adamkus	9,6%	liberalai	6,2%	paulauskas	3,1%
Išskiria:	seimas	28,3%	adamkus	5,1%	krepšinis	4,9%	čempionatas	3,5%
Skirstinys 12 **Dydis:** 51 **VPan:** 0,131 **IPan:** 0,005								
Sieja:	finansai	52,4%	es	22,1%	mokesčiai	4,0%	bankai	3,1%
Išskiria:	finansai	28,1%	es	9,0%	krepšinis	5,4%	čempionatas	3,9%
Skirstinys 13 **Dydis:** 46 **VPan:** 0,126 **IPan:** 0,004								
Sieja:	olimpiada	29,0%	turinas	27,8%	sulaikymas	6,1%	pasienis	6,1%
Išskiria:	olimpiada	15,7%	turinas	15,3%	krepšinis	5,1%	čempionatas	3,7%
Skirstinys 14 **Dydis:** 70 **VPan:** 0,124 **IPan:** 0,006								
Sieja:	energetika	42,3%	rusija	15,5%	nafta	9,9%	dujos	7,9%
Išskiria:	energetika	24,0%	krepšinis	5,8%	nafta	5,7%	dujos	4,6%
Skirstinys 15 **Dydis:** 49 **VPan:** 0,113 **IPan:** 0,003								
Sieja:	žemės ūkis	37,8%	paukščių gripas	31,0%	prancūzija	5,5%	kontrolė	2,4%
Išskiria:	žemės ūkis	20,7%	paukščių gripas	16,7%	krepšinis	5,1%	čempionatas	3,7%
Skirstinys 16 **Dydis:** 66 **VPan:** 0,100 **IPan:** 0,007								
Sieja:	įmonės	41,7%	prekyba	14,0%	vokietija	7,7%	verslas	6,8%
Išskiria:	įmonės	23,3%	prekyba	7,3%	krepšinis	6,0%	čempionatas	4,3%
Skirstinys 17 **Dydis:** 90 **VPan:** 0,091 **IPan:** 0,005								
Sieja:	aukos	48,0%	sprogimas	7,1%	gaisras	5,9%	afganistanas	4,2%
Išskiria:	aukos	24,2%	krepšinis	5,8%	čempionatas	4,1%	rezultatai	4,0%
Skirstinys 18 **Dydis:** 53 **VPan:** 0,067 **IPan:** 0,002								
Sieja:	muzika	18,7%	kinas	17,5%	mirtis	8,3%	aplinkosauga	6,5%
Išskiria:	muzika	9,6%	kinas	9,1%	krepšinis	4,9%	mirtis	4,4%
Skirstinys 19 **Dydis:** 57 **VPan:** 0,056 **IPan:** 0,004								
Sieja:	vizitas	13,7%	ministras	10,1%	latvija	8,1%	vatikanas	7,2%
Išskiria:	vizitas	7,0%	ministras	5,7%	krepšinis	5,4%	vatikanas	4,1%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 5

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,509, **Grynumas:** 0,710

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	459	0,093	0,064	0,003	0,003	0,208	0,930
1	346	0,041	0,022	0,004	0,003	0,495	0,775
2	412	0,036	0,020	0,004	0,003	0,662	0,587
3	402	0,031	0,026	0,002	0,003	0,739	0,540
4	488	0,029	0,017	0,003	0,003	0,486	0,703

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	9	427	5	2	16
1	26	4	12	268	36
2	14	10	30	242	116
3	217	10	77	15	83
4	16	6	121	2	343

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 459	**VPan:** 0,093	**IPan:** 0,003					
Sieja:	krepšinis	28,8%	čempionatas	23,2%	rezultatai	13,2%	lietuviai	6,4%
Išskiria:	krepšinis	15,8%	čempionatas	12,6%	rezultatai	6,6%	įvairybės	4,7%
Skirstinys 1	**Dydis:** 346	**VPan:** 0,041	**IPan:** 0,004					
Sieja:	finansai	22,1%	ekonomika	16,2%	es	12,9%	prekyba	7,9%
Išskiria:	finansai	13,0%	ekonomika	9,2%	krepšinis	7,5%	es	6,2%
Skirstinys 2	**Dydis:** 412	**VPan:** 0,036	**IPan:** 0,004					
Sieja:	rusija	21,0%	energetika	14,2%	transportas	8,3%	įmonės	7,2%
Išskiria:	rusija	8,9%	energetika	8,3%	krepšinis	8,2%	čempionatas	6,7%
Skirstinys 3	**Dydis:** 402	**VPan:** 0,031	**IPan:** 0,002					
Sieja:	įvairybės	50,7%	britanija	8,2%	seimas	6,4%	kinas	2,7%
Išskiria:	įvairybės	28,1%	krepšinis	6,9%	čempionatas	5,3%	britanija	3,6%
Skirstinys 4	**Dydis:** 488	**VPan:** 0,029	**IPan:** 0,003					
Sieja:	aukos	19,0%	irakas	12,0%	rinkimai	11,9%	baltarusija	5,7%
Išskiria:	aukos	10,8%	krepšinis	7,7%	irakas	6,6%	rinkimai	6,4%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 10

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,436, **Grynumas:** 0,738

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	264	0,191	0,098	0,006	0,004	0,015	0,996
1	88	0,182	0,077	0,005	0,004	0,115	0,955
2	153	0,132	0,070	0,003	0,003	0,543	0,739
3	167	0,096	0,046	0,006	0,004	0,618	0,659
4	165	0,074	0,044	0,006	0,006	0,753	0,582
5	178	0,068	0,038	0,004	0,003	0,366	0,848
6	273	0,048	0,027	0,003	0,003	0,263	0,908
7	276	0,046	0,027	0,003	0,003	0,540	0,525
8	316	0,043	0,022	0,004	0,003	0,493	0,769
9	227	0,026	0,014	0,002	0,002	0,666	0,445

Pasiskirstymas klasėse

Skirstinys	**Kultūra**	**Sportas**	**Teisėtvarka**	**Ūkis**	**Politika**
0	1	263	0	0	0
1	4	84	0	0	0
2	113	2	20	6	12
3	4	6	11	110	36
4	25	96	20	5	19
5	7	0	11	151	9
6	9	3	9	4	248
7	9	2	119	1	145
8	14	1	34	243	24
9	96	0	21	9	101

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 264	**VPan:** 0,191	**IPan:** 0,006					
Sieja:	krepšinis	42,3%	čempionatas	25,5%	rezultatai	9,4%	lietuviai	8,0%
Išskiria:	krepšinis	24,4%	čempionatas	11,9%	įvairybės	4,3%	lietuviai	4,1%
Skirstinys 1	**Dydis:** 88	**VPan:** 0,182	**IPan:** 0,005					
Sieja:	olimpiada	21,9%	turinas	20,5%	ledo ritulys	19,2%	rezultatai	15,3%
Išskiria:	olimpiada	12,3%	turinas	11,5%	ledo ritulys	10,8%	krepšinis	6,3%
Skirstinys 2	**Dydis:** 153	**VPan:** 0,132	**IPan:** 0,003					

Sieja:	įvairybės	76,6%	britanija	13,0%	anekdotai	1,3%	didžioji	1,0%
Išskiria:	įvairybės	41,0%	krepšinis	6,4%	britanija	5,9%	čempionatas	5,2%
Skirstinys 3	**Dydis:** 167	**VPan:** 0,096	**IPan:** 0,006					
Sieja:	energetika	33,2%	rusija	32,4%	nafta	7,7%	dujos	5,1%
Išskiria:	energetika	19,6%	rusija	13,1%	krepšinis	7,2%	čempionatas	5,9%
Skirstinys 4	**Dydis:** 165	**VPan:** 0,074	**IPan:** 0,006					
Sieja:	futbolas	41,9%	ispanija	16,1%	kaunas	7,7%	statistika	4,3%
Išskiria:	futbolas	25,9%	ispanija	7,9%	krepšinis	7,4%	kaunas	4,6%
Skirstinys 5	**Dydis:** 178	**VPan:** 0,068	**IPan:** 0,004					
Sieja:	transportas	23,1%	įmonės	18,6%	eismas	6,8%	aviacija	6,0%
Išskiria:	transportas	13,0%	įmonės	9,2%	krepšinis	6,6%	čempionatas	5,4%
Skirstinys 6	**Dydis:** 273	**VPan:** 0,048	**IPan:** 0,003					
Sieja:	rinkimai	21,3%	baltarusija	12,1%	iranas	10,0%	izraelis	9,8%
Išskiria:	rinkimai	11,4%	krepšinis	7,0%	baltarusija	6,8%	čempionatas	5,7%
Skirstinys 7	**Dydis:** 276	**VPan:** 0,046	**IPan:** 0,003					
Sieja:	aukos	34,9%	irakas	24,2%	bušas	2,8%	stichija	2,5%
Išskiria:	aukos	19,3%	irakas	13,3%	krepšinis	6,9%	čempionatas	5,6%
Skirstinys 8	**Dydis:** 316	**VPan:** 0,043	**IPan:** 0,004					
Sieja:	finansai	21,7%	ekonomika	16,6%	es	12,4%	prekyba	6,5%
Išskiria:	finansai	12,1%	ekonomika	9,1%	krepšinis	7,3%	čempionatas	6,0%
Skirstinys 9	**Dydis:** 227	**VPan:** 0,026	**IPan:** 0,002					
Sieja:	seimas	27,7%	muzika	7,1%	kinas	5,5%	vilnius	5,3%
Išskiria:	seimas	14,9%	krepšinis	6,1%	čempionatas	5,0%	muzika	3,6%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 15

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,385, **Grynumas:** 0,786

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	172	0,286	0,133	0,012	0,005	0,044	0,988
1	82	0,204	0,073	0,006	0,004	0,071	0,976
2	92	0,195	0,068	0,006	0,004	0,089	0,967
3	101	0,173	0,077	0,011	0,006	0,225	0,921
4	144	0,144	0,072	0,004	0,003	0,486	0,785
5	90	0,123	0,073	0,003	0,003	0,264	0,900
6	114	0,101	0,057	0,004	0,004	0,240	0,886
7	161	0,091	0,034	0,005	0,004	0,088	0,975
8	153	0,083	0,046	0,006	0,007	0,699	0,627
9	162	0,076	0,042	0,004	0,003	0,334	0,864
10	154	0,074	0,049	0,004	0,004	0,292	0,890
11	148	0,073	0,037	0,004	0,003	0,432	0,764
12	147	0,061	0,031	0,005	0,003	0,676	0,653
13	199	0,042	0,023	0,004	0,003	0,725	0,533
14	188	0,033	0,017	0,002	0,002	0,632	0,452

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	1	170	0	0	1
1	2	80	0	0	0
2	0	0	0	89	3
3	5	93	1	1	1
4	113	2	14	6	9
5	2	2	5	0	81
6	0	0	12	1	101
7	1	1	0	157	2

8	22	96	12	4	19
9	6	0	5	140	11
10	9	5	1	2	137
11	7	1	113	0	27
12	7	7	21	16	96
13	23	0	49	106	21
14	84	0	12	7	85

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 172	**VPan:** 0,286	**IPan:** 0,012					
Sieja:	čempionatas	40,5%	krepšinis	29,3%	rezultatai	13,6%	lietuviai	9,8%
Išskiria:	čempionatas	22,2%	krepšinis	10,8%	lietuviai	4,9%	rezultatai	4,5%
Skirstinys 1	**Dydis:** 82	**VPan:** 0,204	**IPan:** 0,006					
Sieja:	olimpiada	21,1%	turinas	21,0%	ledo ritulys	19,7%	rezultatai	15,2%
Išskiria:	turinas	11,8%	olimpiada	11,8%	ledo ritulys	11,1%	krepšinis	6,3%
Skirstinys 2	**Dydis:** 92	**VPan:** 0,195	**IPan:** 0,006					
Sieja:	energetika	54,1%	nafta	12,6%	dujos	7,7%	pasaulis	5,8%
Išskiria:	energetika	30,4%	nafta	7,1%	krepšinis	6,5%	čempionatas	5,3%
Skirstinys 3	**Dydis:** 101	**VPan:** 0,173	**IPan:** 0,011					
Sieja:	krepšinis	34,1%	eurolyga	29,0%	laureatai	5,2%	moterys	4,8%
Išskiria:	eurolyga	18,9%	krepšinis	8,1%	čempionatas	6,3%	įvairybės	3,5%
Skirstinys 4	**Dydis:** 144	**VPan:** 0,144	**IPan:** 0,004					
Sieja:	įvairybės	78,1%	britanija	12,2%	anekdotai	1,3%	didžioji	1,0%
Išskiria:	įvairybės	41,7%	krepšinis	6,4%	britanija	5,3%	čempionatas	5,2%
Skirstinys 5	**Dydis:** 90	**VPan:** 0,123	**IPan:** 0,003					
Sieja:	izraelis	37,0%	palestiniečiai	30,9%	hamas	7,8%	šaronas	3,6%
Išskiria:	izraelis	19,8%	palestiniečiai	16,7%	krepšinis	5,9%	čempionatas	4,8%
Skirstinys 6	**Dydis:** 114	**VPan:** 0,101	**IPan:** 0,004					
Sieja:	irakas	67,5%	bušas	6,1%	aukos	2,4%	afganistanas	1,9%
Išskiria:	irakas	37,3%	krepšinis	6,3%	čempionatas	5,1%	rusija	3,0%
Skirstinys 7	**Dydis:** 161	**VPan:** 0,091	**IPan:** 0,005					
Sieja:	finansai	41,6%	ekonomika	28,7%	es	3,6%	euras	1,7%
Išskiria:	finansai	23,8%	ekonomika	15,7%	krepšinis	6,8%	čempionatas	5,6%
Skirstinys 8	**Dydis:** 153	**VPan:** 0,083	**IPan:** 0,006					
Sieja:	futbolas	43,7%	ispanija	17,1%	kaunas	7,1%	statistika	4,2%
Išskiria:	futbolas	26,8%	ispanija	8,4%	krepšinis	6,9%	kaunas	4,1%
Skirstinys 9	**Dydis:** 162	**VPan:** 0,076	**IPan:** 0,004					
Sieja:	transportas	25,0%	įmonės	18,3%	eismas	7,4%	keliai	6,2%
Išskiria:	transportas	14,1%	įmonės	8,9%	krepšinis	6,6%	čempionatas	5,3%
Skirstinys 10	**Dydis:** 154	**VPan:** 0,074	**IPan:** 0,004					
Sieja:	rinkimai	33,1%	baltarusija	24,8%	prezidentas	9,8%	ukraina	4,5%
Išskiria:	rinkimai	17,1%	baltarusija	14,1%	krepšinis	6,3%	čempionatas	5,4%
Skirstinys 11	**Dydis:** 148	**VPan:** 0,073	**IPan:** 0,004					
Sieja:	aukos	55,0%	stichija	5,4%	katastriša	3,2%	sprogimas	3,1%
Išskiria:	aukos	28,8%	krepšinis	6,5%	čempionatas	5,3%	rezultatai	3,3%
Skirstinys 12	**Dydis:** 147	**VPan:** 0,061	**IPan:** 0,005					
Sieja:	rusija	31,7%	iranas	25,2%	jt	7,5%	branduolinė programa	4,6%
Išskiria:	iranas	14,0%	rusija	11,0%	krepšinis	7,0%	čempionatas	5,7%
Skirstinys 13	**Dydis:** 199	**VPan:** 0,042	**IPan:** 0,004					
Sieja:	prekyba	23,0%	žemės ūkis	11,0%	es	10,4%	paukščių gripas	9,6%
Išskiria:	prekyba	12,9%	krepšinis	6,8%	žemės ūkis	6,1%	paukščių gripas	5,7%
Skirstinys 14	**Dydis:** 188	**VPan:** 0,033	**IPan:** 0,002					
Sieja:	seimas	30,1%	muzika	7,9%	kinas	6,2%	prancūzija	5,5%
Išskiria:	seimas	16,0%	krepšinis	6,0%	čempionatas	4,9%	muzika	4,0%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 20

Suklasterizuota dokumentų: [2107 iš 2112], **Entropija:** 0,331, **Grynumas:** 0,832

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	45	0,339	0,108	0,005	0,004	0,113	0,956
1	167	0,297	0,134	0,012	0,005	0,023	0,994
2	49	0,274	0,133	0,007	0,004	0,000	1,000
3	56	0,221	0,141	0,004	0,002	0,777	0,446
4	84	0,215	0,065	0,006	0,004	0,000	1,000
5	95	0,190	0,077	0,011	0,006	0,100	0,968
6	135	0,153	0,074	0,004	0,003	0,261	0,911
7	80	0,150	0,078	0,003	0,003	0,141	0,950
8	100	0,128	0,037	0,005	0,004	0,070	0,980
9	124	0,117	0,053	0,007	0,007	0,518	0,758
10	86	0,109	0,057	0,005	0,003	0,581	0,686
11	120	0,108	0,062	0,005	0,004	0,294	0,892
12	120	0,093	0,054	0,004	0,004	0,279	0,883
13	96	0,087	0,041	0,004	0,003	0,274	0,896
14	92	0,080	0,040	0,002	0,003	0,680	0,511
15	95	0,078	0,038	0,002	0,002	0,186	0,937
16	142	0,078	0,037	0,004	0,003	0,410	0,754
17	146	0,072	0,036	0,006	0,003	0,386	0,815
18	129	0,071	0,033	0,005	0,003	0,580	0,713
19	146	0,030	0,015	0,002	0,001	0,702	0,616

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	2	43	0	0	0
1	1	166	0	0	0
2	0	49	0	0	0
3	6	0	16	25	9
4	0	0	0	84	0
5	2	92	0	1	0
6	123	2	4	3	3
7	1	0	3	0	76
8	1	0	1	98	0
9	6	94	6	2	16
10	6	0	6	59	15
11	6	4	1	2	107
12	3	0	10	1	106
13	2	0	5	86	3
14	13	0	47	29	3
15	3	0	2	1	89
16	5	0	107	0	30
17	6	0	2	119	19
18	6	2	18	11	92
19	90	5	17	8	26

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 45	**VPan:** 0,339	**IPan:** 0,005					
Sieja:	olimpiada	44,9%	turinas	42,0%	slidinėjimas	2,6%	rezultatai	1,4%
Išskiria:	olimpiada	24,2%	turinas	22,6%	krepšinis	5,8%	čempionatas	4,8%
Skirstinys 1	**Dydis:** 167	**VPan:** 0,297	**IPan:** 0,012					
Sieja:	čempionatas	41,0%	krepšinis	29,3%	rezultatai	13,1%	lietuviai	9,8%
Išskiria:	čempionatas	22,5%	krepšinis	10,7%	lietuviai	4,9%	įvairybės	4,4%

Skirstinys 2	**Dydis:** 49	**VPan:** 0,274	**IPan:** 0,007					
Sieja:	ledo ritulys	34,6%	rezultatai	20,5%	zubrus	14,1%	tenisas	9,5%
Išskiria:	ledo ritulys	19,3%	zubrus	8,0%	krepšinis	6,2%	tenisas	5,3%
Skirstinys 3	**Dydis:** 56	**VPan:** 0,221	**IPan:** 0,004					
Sieja:	britanija	20,5%	keliai	18,0%	eismas	17,5%	sąlygos	17,3%
Išskiria:	keliai	9,7%	sąlygos	9,3%	eismas	9,0%	būklė	8,8%
Skirstinys 4	**Dydis:** 84	**VPan:** 0,215	**IPan:** 0,006					
Sieja:	energetika	52,7%	nafta	13,7%	dujos	8,4%	pasaulis	5,9%
Išskiria:	energetika	29,1%	nafta	7,7%	krepšinis	6,4%	čempionatas	5,2%
Skirstinys 5	**Dydis:** 95	**VPan:** 0,190	**IPan:** 0,011					
Sieja:	krepšinis	34,4%	eurolyga	29,9%	moterys	5,4%	laureatai	4,6%
Išskiria:	eurolyga	19,5%	krepšinis	8,2%	čempionatas	6,3%	įvairybės	3,5%
Skirstinys 6	**Dydis:** 135	**VPan:** 0,153	**IPan:** 0,004					
Sieja:	įvairybės	83,7%	kinas	4,6%	britanija	2,3%	anekdotai	1,4%
Išskiria:	įvairybės	44,5%	krepšinis	6,3%	čempionatas	5,1%	rezultatai	3,2%
Skirstinys 7	**Dydis:** 80	**VPan:** 0,150	**IPan:** 0,003					
Sieja:	izraelis	38,5%	palestiniečiai	32,1%	hamas	8,1%	šaronas	3,7%
Išskiria:	izraelis	20,5%	palestiniečiai	17,3%	krepšinis	5,9%	čempionatas	4,8%
Skirstinys 8	**Dydis:** 100	**VPan:** 0,128	**IPan:** 0,005					
Sieja:	finansai	73,1%	ekonomika	2,9%	es	2,4%	euras	1,8%
Išskiria:	finansai	40,8%	krepšinis	6,4%	čempionatas	5,2%	rezultatai	3,3%
Skirstinys 9	**Dydis:** 124	**VPan:** 0,117	**IPan:** 0,007					
Sieja:	futbolas	47,2%	ispanija	20,9%	statistika	4,2%	čempionatas	4,2%
Išskiria:	futbolas	29,1%	ispanija	10,9%	krepšinis	6,4%	rezultatai	3,1%
Skirstinys 10	**Dydis:** 86	**VPan:** 0,109	**IPan:** 0,005					
Sieja:	transportas	49,6%	aviacija	14,2%	vokietija	13,1%	automobiliai	4,5%
Išskiria:	transportas	27,8%	aviacija	8,1%	krepšinis	6,5%	čempionatas	5,3%
Skirstinys 11	**Dydis:** 120	**VPan:** 0,108	**IPan:** 0,005					
Sieja:	rinkimai	37,4%	baltarusija	28,1%	prezidentas	10,5%	ukraina	4,2%
Išskiria:	rinkimai	19,4%	baltarusija	15,9%	krepšinis	5,8%	čempionatas	5,3%
Skirstinys 12	**Dydis:** 120	**VPan:** 0,093	**IPan:** 0,004					
Sieja:	irakas	66,0%	bušas	5,3%	aukos	2,3%	prancūzija	2,0%
Išskiria:	irakas	36,9%	krepšinis	6,4%	čempionatas	5,2%	rusija	3,0%
Skirstinys 13	**Dydis:** 96	**VPan:** 0,087	**IPan:** 0,004					
Sieja:	įmonės	55,7%	verslas	13,3%	japonija	2,9%	google	2,6%
Išskiria:	įmonės	30,0%	verslas	7,3%	krepšinis	6,3%	čempionatas	5,2%
Skirstinys 14	**Dydis:** 92	**VPan:** 0,080	**IPan:** 0,002					
Sieja:	žemės ūkis	23,7%	paukščių gripas	22,0%	pasienis	12,4%	sulaikymas	10,3%
Išskiria:	žemės ūkis	12,1%	paukščių gripas	11,7%	pasienis	6,7%	krepšinis	5,9%
Skirstinys 15	**Dydis:** 95	**VPan:** 0,078	**IPan:** 0,002					
Sieja:	seimas	52,2%	adamkus	9,8%	paulauskas	8,0%	nepasitikėjimas	1,5%
Išskiria:	seimas	27,4%	krepšinis	5,8%	adamkus	5,1%	čempionatas	4,7%
Skirstinys 16	**Dydis:** 142	**VPan:** 0,078	**IPan:** 0,004					
Sieja:	aukos	55,6%	stichija	5,5%	katastriša	3,2%	sprogimas	3,1%
Išskiria:	aukos	29,0%	krepšinis	6,5%	čempionatas	5,3%	rezultatai	3,3%
Skirstinys 17	**Dydis:** 146	**VPan:** 0,072	**IPan:** 0,006					
Sieja:	ekonomika	30,2%	prekyba	18,2%	es	14,9%	statistika	6,3%
Išskiria:	ekonomika	16,4%	prekyba	10,0%	krepšinis	7,3%	čempionatas	5,9%
Skirstinys 18	**Dydis:** 129	**VPan:** 0,071	**IPan:** 0,005					
Sieja:	rusija	28,3%	iranas	28,1%	jt	8,4%	branduolinė programa	5,1%
Išskiria:	iranas	15,6%	rusija	8,9%	krepšinis	6,9%	čempionatas	5,6%
Skirstinys 19	**Dydis:** 146	**VPan:** 0,030	**IPan:** 0,002					
Sieja:	kaunas	15,5%	muzika	14,5%	vilnius	9,8%	koncertas	4,6%
Išskiria:	kaunas	7,7%	muzika	7,5%	krepšinis	5,9%	vilnius	5,0%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 5

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,404, **Grynumas:** 0,767

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	826	0,087	0,062	0,002	0,002	0,088	0,975
1	729	0,031	0,023	0,002	0,003	0,708	0,588
2	733	0,031	0,022	0,002	0,002	0,523	0,513
3	1070	0,027	0,015	0,003	0,002	0,411	0,824
4	863	0,025	0,016	0,002	0,003	0,341	0,864

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	3	1	15	805	2
1	58	168	429	18	56
2	331	8	18	0	376
3	102	882	43	8	35
4	746	14	67	10	26

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 826	**VPan:** 0,087	**IPan:** 0,002					
Sieja:	krepšinis	27,2%	čempionatas	21,2%	rezultatai	16,7%	lietuviai	7,4%
Išskiria:	krepšinis	14,8%	čempionatas	11,4%	rezultatai	8,2%	lietuviai	4,0%
Skirstinys 1	**Dydis:** 729	**VPan:** 0,031	**IPan:** 0,002					
Sieja:	įvairybės	42,8%	britanija	10,1%	transportas	10,1%	eismas	3,5%
Išskiria:	įvairybės	24,1%	krepšinis	6,1%	transportas	5,5%	britanija	4,8%
Skirstinys 2	**Dydis:** 733	**VPan:** 0,031	**IPan:** 0,002					
Sieja:	aukos	34,9%	irakas	16,6%	terorizmas	3,5%	sprogimas	3,3%
Išskiria:	aukos	19,7%	irakas	8,7%	krepšinis	6,2%	čempionatas	4,9%
Skirstinys 3	**Dydis:** 1070	**VPan:** 0,027	**IPan:** 0,003					
Sieja:	ekonomika	11,7%	energetika	10,0%	finansai	9,9%	rusija	8,0%
Išskiria:	krepšinis	7,6%	ekonomika	6,9%	čempionatas	6,0%	energetika	5,9%
Skirstinys 4	**Dydis:** 863	**VPan:** 0,025	**IPan:** 0,002					
Sieja:	rinkimai	15,1%	iranas	9,9%	baltarusija	9,7%	izraelis	5,9%
Išskiria:	rinkimai	8,4%	krepšinis	6,6%	baltarusija	5,3%	iranas	5,3%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 10

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,412, **Grynumas:** 0,755

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	451	0,183	0,104	0,005	0,003	0,078	0,976
1	193	0,148	0,071	0,005	0,003	0,137	0,953
2	260	0,087	0,052	0,006	0,006	0,527	0,765
3	350	0,074	0,044	0,003	0,002	0,528	0,757
4	351	0,060	0,034	0,003	0,003	0,353	0,860
5	448	0,057	0,035	0,004	0,003	0,478	0,779
6	492	0,039	0,028	0,003	0,003	0,306	0,884
7	414	0,037	0,035	0,002	0,002	0,751	0,411
8	631	0,038	0,021	0,003	0,003	0,396	0,830
9	631	0,036	0,025	0,002	0,002	0,485	0,507

Pasiskirstymas klasėse					
Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	2	0	9	440	0
1	1	0	7	184	1
2	20	5	17	199	19
3	37	15	265	6	27
4	302	8	24	3	14
5	53	349	16	3	27
6	435	8	26	3	20
7	35	159	170	1	49
8	55	524	32	2	18
9	300	5	6	0	320

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 451	**VPan:** 0,183	**IPan:** 0,005					
Sieja:	krepšinis	43,2%	čempionatas	23,8%	rezultatai	10,5%	lietuviai	10,1%
Išskiria:	krepšinis	24,9%	čempionatas	11,0%	lietuviai	5,3%	įvairybės	3,4%
Skirstinys 1	**Dydis:** 193	**VPan:** 0,148	**IPan:** 0,005					
Sieja:	rezultatai	21,5%	ledo ritulys	17,5%	olimpiada	17,5%	turinas	16,8%
Išskiria:	olimpiada	10,1%	ledo ritulys	10,0%	turinas	9,6%	krepšinis	5,7%
Skirstinys 2	**Dydis:** 260	**VPan:** 0,087	**IPan:** 0,006					
Sieja:	futbolas	50,2%	ispanija	9,5%	statistika	5,9%	čempionatas	4,8%
Išskiria:	futbolas	31,1%	krepšinis	6,3%	ispanija	4,9%	rezultatai	3,2%
Skirstinys 3	**Dydis:** 350	**VPan:** 0,074	**IPan:** 0,003					
Sieja:	įvairybės	71,5%	britanija	14,5%	kinas	2,5%	didžioji	1,6%
Išskiria:	įvairybės	38,6%	britanija	6,5%	krepšinis	5,5%	čempionatas	4,3%
Skirstinys 4	**Dydis:** 351	**VPan:** 0,060	**IPan:** 0,003					
Sieja:	iranas	27,4%	palestiniečiai	13,7%	izraelis	13,1%	branduolinė programa	7,0%
Išskiria:	iranas	15,0%	palestiniečiai	7,6%	izraelis	6,9%	krepšinis	5,7%
Skirstinys 5	**Dydis:** 448	**VPan:** 0,057	**IPan:** 0,004					
Sieja:	energetika	26,2%	rusija	20,5%	įmonės	15,7%	nafta	6,2%
Išskiria:	energetika	15,5%	įmonės	8,9%	rusija	7,0%	krepšinis	6,4%
Skirstinys 6	**Dydis:** 492	**VPan:** 0,039	**IPan:** 0,003					
Sieja:	rinkimai	25,8%	baltarusija	19,1%	seimas	11,0%	prezidentas	6,8%
Išskiria:	rinkimai	13,8%	baltarusija	10,6%	seimas	6,2%	krepšinis	5,9%
Skirstinys 7	**Dydis:** 414	**VPan:** 0,037	**IPan:** 0,002					
Sieja:	transportas	27,4%	eismas	9,8%	keliai	8,0%	sąlygos	7,5%
Išskiria:	transportas	15,1%	krepšinis	5,5%	eismas	5,1%	keliai	4,4%
Skirstinys 8	**Dydis:** 631	**VPan:** 0,038	**IPan:** 0,003					
Sieja:	ekonomika	22,8%	finansai	19,8%	es	9,8%	žemės ūkis	7,3%
Išskiria:	ekonomika	13,3%	finansai	11,5%	krepšinis	6,6%	čempionatas	5,1%
Skirstinys 9	**Dydis:** 631	**VPan:** 0,036	**IPan:** 0,002					
Sieja:	aukos	38,9%	irakas	17,3%	sprogimas	2,9%	pakistanas	2,5%
Išskiria:	aukos	21,8%	irakas	8,7%	krepšinis	6,1%	čempionatas	4,7%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 15

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,375, **Grynumas:** 0,798

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	166	0,194	0,079	0,005	0,004	0,079	0,976
1	448	0,184	0,105	0,005	0,003	0,071	0,978
2	185	0,160	0,070	0,005	0,003	0,021	0,995
3	228	0,116	0,058	0,004	0,003	0,222	0,925
4	191	0,100	0,060	0,003	0,003	0,293	0,885

5	192	0,096	0,057	0,003	0,002	0,437	0,807
6	249	0,094	0,052	0,006	0,006	0,488	0,791
7	332	0,078	0,046	0,003	0,002	0,471	0,792
8	328	0,079	0,032	0,004	0,003	0,122	0,963
9	317	0,068	0,036	0,004	0,002	0,395	0,763
10	274	0,065	0,041	0,004	0,003	0,268	0,883
11	271	0,052	0,037	0,004	0,003	0,536	0,712
12	385	0,043	0,036	0,002	0,002	0,770	0,387
13	315	0,045	0,025	0,005	0,003	0,623	0,679
14	340	0,026	0,017	0,002	0,002	0,557	0,685

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	3	162	1	0	0
1	1	0	9	438	0
2	0	0	1	184	0
3	211	1	5	7	4
4	169	3	8	0	11
5	155	8	21	2	6
6	17	5	14	197	16
7	28	11	263	4	26
8	4	316	3	0	5
9	68	3	4	0	242
10	242	2	4	0	26
11	43	193	28	0	7
12	18	149	112	1	105
13	48	214	25	4	24
14	233	6	74	4	23

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 166	**VPan:** 0,194	**IPan:** 0,005					
Sieja:	energetika	48,8%	dujos	13,1%	nafta	12,8%	pasaulis	6,7%
Išskiria:	energetika	26,6%	dujos	7,3%	nafta	7,1%	krepšinis	5,6%
Skirstinys 1	**Dydis:** 448	**VPan:** 0,184	**IPan:** 0,005					
Sieja:	krepšinis	43,0%	čempionatas	24,0%	rezultatai	10,5%	lietuviai	10,0%
Išskiria:	krepšinis	24,7%	čempionatas	11,1%	lietuviai	5,3%	įvairybės	3,4%
Skirstinys 2	**Dydis:** 185	**VPan:** 0,160	**IPan:** 0,005					
Sieja:	rezultatai	21,6%	ledo ritulys	17,6%	olimpiada	17,5%	turinas	16,8%
Išskiria:	olimpiada	10,1%	ledo ritulys	10,0%	turinas	9,7%	krepšinis	5,7%
Skirstinys 3	**Dydis:** 228	**VPan:** 0,116	**IPan:** 0,004					
Sieja:	rinkimai	36,5%	baltarusija	29,9%	prezidentas	9,8%	parlamentas	9,2%
Išskiria:	rinkimai	19,1%	baltarusija	16,5%	krepšinis	5,2%	prezidentas	5,0%
Skirstinys 4	**Dydis:** 191	**VPan:** 0,100	**IPan:** 0,003					
Sieja:	iranas	50,3%	branduolinė programa	14,0%	jt	9,7%	atomas	5,4%
Išskiria:	iranas	27,1%	branduolinė programa	7,8%	krepšinis	5,4%	jt	5,3%
Skirstinys 5	**Dydis:** 192	**VPan:** 0,096	**IPan:** 0,003					
Sieja:	izraelis	27,5%	palestiniečiai	27,0%	hamas	12,8%	šaronas	7,0%
Išskiria:	izraelis	14,5%	palestiniečiai	14,4%	hamas	6,9%	krepšinis	5,3%
Skirstinys 6	**Dydis:** 249	**VPan:** 0,094	**IPan:** 0,006					
Sieja:	futbolas	50,5%	ispanija	9,9%	statistika	5,9%	čempionatas	4,8%
Išskiria:	futbolas	31,2%	krepšinis	6,3%	ispanija	5,1%	rezultatai	3,1%
Skirstinys 7	**Dydis:** 332	**VPan:** 0,078	**IPan:** 0,003					
Sieja:	įvairybės	74,7%	britanija	12,0%	kinas	2,6%	didžioji	1,0%
Išskiria:	įvairybės	40,1%	krepšinis	5,4%	britanija	4,9%	čempionatas	4,3%
Skirstinys 8	**Dydis:** 328	**VPan:** 0,079	**IPan:** 0,004					
Sieja:	ekonomika	39,6%	finansai	30,9%	statistika	3,6%	es	3,4%
Išskiria:	ekonomika	22,9%	finansai	17,3%	krepšinis	6,1%	čempionatas	4,8%
Skirstinys 9	**Dydis:** 317	**VPan:** 0,068	**IPan:** 0,004					

Sieja:	aukos	54,0%	incidentas	4,1%	stichija	4,1%	sprogimas	3,8%
Išskiria:	aukos	28,0%	krepšinis	5,8%	čempionatas	4,6%	rezultatai	4,1%
Skirstinys 10	**Dydis:** 274	**VPan:** 0,065	**IPan:** 0,004					
Sieja:	irakas	60,5%	bušas	4,3%	aukos	4,2%	terorizmas	2,7%
Išskiria:	irakas	33,9%	krepšinis	5,7%	čempionatas	4,5%	rezultatai	3,4%
Skirstinys 11	**Dydis:** 271	**VPan:** 0,052	**IPan:** 0,004					
Sieja:	žemės ūkis	28,3%	paukščių gripas	21,4%	prekyba	11,7%	es	10,6%
Išskiria:	žemės ūkis	16,8%	paukščių gripas	12,7%	krepšinis	5,9%	prekyba	5,1%
Skirstinys 12	**Dydis:** 385	**VPan:** 0,043	**IPan:** 0,002					
Sieja:	transportas	24,7%	eismas	9,1%	keliai	7,7%	sąlygos	7,6%
Išskiria:	transportas	13,1%	krepšinis	5,4%	eismas	4,6%	čempionatas	4,2%
Skirstinys 13	**Dydis:** 315	**VPan:** 0,045	**IPan:** 0,005					
Sieja:	įmonės	44,6%	rusija	20,2%	verslas	6,4%	prekyba	2,4%
Išskiria:	įmonės	27,8%	krepšinis	6,4%	rusija	5,6%	čempionatas	5,0%
Skirstinys 14	**Dydis:** 340	**VPan:** 0,026	**IPan:** 0,002					
Sieja:	seimas	33,7%	paulauskas	6,8%	adamkus	6,0%	lenkija	4,2%
Išskiria:	seimas	18,4%	krepšinis	5,3%	čempionatas	4,1%	paulauskas	3,7%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 20

Suklasterizuota dokumentų: [4221 iš 4224], **Entropija:** 0,339, **Grynumas:** 0,832

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	79	0,346	0,094	0,005	0,004	0,042	0,987
1	285	0,289	0,145	0,011	0,005	0,072	0,975
2	109	0,228	0,110	0,007	0,003	0,000	1,000
3	164	0,198	0,078	0,005	0,004	0,064	0,982
4	171	0,156	0,066	0,009	0,005	0,099	0,971
5	144	0,142	0,115	0,002	0,002	0,458	0,653
6	178	0,129	0,056	0,006	0,003	0,219	0,921
7	225	0,118	0,058	0,004	0,003	0,224	0,924
8	218	0,116	0,055	0,006	0,006	0,369	0,853
9	182	0,109	0,061	0,003	0,003	0,245	0,907
10	182	0,103	0,058	0,003	0,002	0,402	0,830
11	192	0,102	0,038	0,004	0,003	0,148	0,953
12	286	0,101	0,049	0,003	0,002	0,468	0,794
13	156	0,097	0,053	0,004	0,003	0,515	0,718
14	294	0,075	0,037	0,004	0,002	0,410	0,738
15	253	0,073	0,043	0,004	0,003	0,173	0,933
16	237	0,060	0,043	0,004	0,003	0,571	0,700
17	282	0,050	0,025	0,005	0,003	0,579	0,695
18	282	0,034	0,020	0,002	0,002	0,469	0,745
19	302	0,019	0,011	0,001	0,001	0,618	0,679

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	0	0	1	78	0
1	0	0	7	278	0
2	0	0	0	109	0
3	2	161	1	0	0
4	0	1	3	166	1
5	2	47	1	0	94
6	3	164	3	0	8
7	208	1	4	7	5
8	16	3	5	186	8

9	165	3	3	0	11
10	151	8	17	2	4
11	4	183	2	0	3
12	25	10	227	3	21
13	25	112	2	0	17
14	71	2	4	0	217
15	236	0	3	0	14
16	40	166	22	3	6
17	48	196	9	3	26
18	210	5	53	0	14
19	34	11	205	6	46

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 79	**VPan:** 0,346	**IPan:** 0,005					
Sieja:	olimpiada	44,5%	turinas	42,6%	rezultatai	3,6%	slidinėjimas	1,5%
Išskiria:	olimpiada	24,4%	turinas	23,4%	krepšinis	5,2%	čempionatas	3,9%
Skirstinys 1	**Dydis:** 285	**VPan:** 0,289	**IPan:** 0,011					
Sieja:	čempionatas	37,6%	krepšinis	30,8%	rezultatai	13,9%	lietuviai	12,9%
Išskiria:	čempionatas	20,2%	krepšinis	12,2%	lietuviai	6,9%	įvairybės	3,5%
Skirstinys 2	**Dydis:** 109	**VPan:** 0,228	**IPan:** 0,007					
Sieja:	ledo ritulys	29,3%	rezultatai	24,1%	zubrus	11,9%	tenisas	11,6%
Išskiria:	ledo ritulys	16,5%	zubrus	6,9%	tenisas	6,7%	krepšinis	5,6%
Skirstinys 3	**Dydis:** 164	**VPan:** 0,198	**IPan:** 0,005					
Sieja:	energetika	49,2%	dujos	13,2%	nafta	12,8%	pasaulis	6,4%
Išskiria:	energetika	26,8%	dujos	7,4%	nafta	7,1%	krepšinis	5,6%
Skirstinys 4	**Dydis:** 171	**VPan:** 0,156	**IPan:** 0,009					
Sieja:	krepšinis	36,4%	eurolyga	28,7%	laureatai	6,9%	moterys	4,2%
Išskiria:	eurolyga	18,6%	krepšinis	9,7%	čempionatas	5,1%	laureatai	4,1%
Skirstinys 5	**Dydis:** 144	**VPan:** 0,142	**IPan:** 0,002					
Sieja:	eismas	19,9%	keliai	17,3%	sąlygos	16,2%	būklė	15,5%
Išskiria:	eismas	9,9%	keliai	9,0%	sąlygos	8,4%	būklė	7,7%
Skirstinys 6	**Dydis:** 178	**VPan:** 0,129	**IPan:** 0,006					
Sieja:	ekonomika	61,9%	statistika	9,6%	zona	3,7%	euro	3,5%
Išskiria:	ekonomika	35,0%	krepšinis	6,0%	čempionatas	4,7%	rezultatai	4,2%
Skirstinys 7	**Dydis:** 225	**VPan:** 0,118	**IPan:** 0,004					
Sieja:	rinkimai	37,0%	baltarusija	29,4%	prezidentas	9,6%	parlamentas	9,0%
Išskiria:	rinkimai	19,4%	baltarusija	16,2%	krepšinis	5,2%	prezidentas	4,9%
Skirstinys 8	**Dydis:** 218	**VPan:** 0,116	**IPan:** 0,006					
Sieja:	futbolas	53,6%	ispanija	9,5%	statistika	4,9%	čempionatas	4,8%
Išskiria:	futbolas	32,8%	krepšinis	6,0%	ispanija	4,7%	rezultatai	3,1%
Skirstinys 9	**Dydis:** 182	**VPan:** 0,109	**IPan:** 0,003					
Sieja:	iranas	52,0%	branduolinė programa	14,2%	jt	9,2%	atomas	4,9%
Išskiria:	iranas	28,0%	branduolinė programa	7,9%	krepšinis	5,4%	jt	4,9%
Skirstinys 10	**Dydis:** 182	**VPan:** 0,103	**IPan:** 0,003					
Sieja:	izraelis	27,5%	palestiniečiai	27,1%	hamas	12,4%	šaronas	7,3%
Išskiria:	palestiniečiai	14,5%	izraelis	14,4%	hamas	6,7%	krepšinis	5,3%
Skirstinys 11	**Dydis:** 192	**VPan:** 0,102	**IPan:** 0,004					
Sieja:	finansai	76,2%	ekonomika	1,8%	euras	1,7%	es	1,6%
Išskiria:	finansai	42,8%	krepšinis	5,6%	čempionatas	4,4%	rezultatai	3,9%
Skirstinys 12	**Dydis:** 286	**VPan:** 0,101	**IPan:** 0,003					
Sieja:	įvairybės	78,3%	britanija	12,4%	didžioji	1,0%	anekdotai	0,8%
Išskiria:	įvairybės	42,0%	krepšinis	5,4%	britanija	5,0%	čempionatas	4,3%
Skirstinys 13	**Dydis:** 156	**VPan:** 0,097	**IPan:** 0,004					
Sieja:	transportas	54,0%	aviacija	15,0%	automobiliai	3,9%	vokietija	3,8%
Išskiria:	transportas	29,4%	aviacija	8,4%	krepšinis	5,5%	čempionatas	4,3%
Skirstinys 14	**Dydis:** 294	**VPan:** 0,075	**IPan:** 0,004					
Sieja:	aukos	56,5%	stichija	4,3%	sprogimas	4,1%	pakistanas	3,5%
Išskiria:	aukos	29,2%	krepšinis	5,8%	čempionatas	4,5%	rezultatai	4,1%

Skirstinys 15	**Dydis:** 253	**VPan:** 0,073	**IPan:** 0,004					
Sieja:	irakas	60,4%	bušas	4,5%	aukos	4,3%	terorizmas	2,8%
Išskiria:	irakas	33,6%	krepšinis	5,7%	čempionatas	4,5%	rezultatai	3,3%
Skirstinys 16	**Dydis:** 237	**VPan:** 0,060	**IPan:** 0,004					
Sieja:	žemės ūkis	28,6%	paukščių gripas	25,1%	es	11,0%	prekyba	10,3%
Išskiria:	žemės ūkis	16,6%	paukščių gripas	14,9%	krepšinis	5,8%	čempionatas	4,5%
Skirstinys 17	**Dydis:** 282	**VPan:** 0,050	**IPan:** 0,005					
Sieja:	įmonės	40,7%	rusija	25,0%	verslas	7,4%	prekyba	2,7%
Išskiria:	įmonės	24,6%	rusija	7,9%	krepšinis	6,4%	čempionatas	5,0%
Skirstinys 18	**Dydis:** 282	**VPan:** 0,034	**IPan:** 0,002					
Sieja:	seimas	37,5%	paulauskas	7,5%	adamkus	6,6%	lenkija	4,2%
Išskiria:	seimas	20,3%	krepšinis	5,2%	čempionatas	4,1%	paulauskas	4,0%
Skirstinys 19	**Dydis:** 302	**VPan:** 0,019	**IPan:** 0,001					
Sieja:	vilnius	13,3%	muzika	8,8%	kaunas	8,4%	koncertas	6,6%
Išskiria:	vilnius	6,4%	krepšinis	5,2%	muzika	4,5%	kaunas	4,3%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 5

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,426, **Grynumas:** 0,760

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	1611	0,083	0,061	0,002	0,002	0,087	0,976
1	1519	0,026	0,018	0,002	0,003	0,448	0,791
2	1682	0,025	0,017	0,002	0,002	0,508	0,588
3	2134	0,025	0,015	0,003	0,002	0,419	0,821
4	1502	0,024	0,019	0,002	0,003	0,686	0,605

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	1573	5	5	21	7
1	18	1201	20	88	192
2	2	989	16	38	637
3	23	206	1752	89	64
4	32	97	341	909	123

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 1611	**VPan:** 0,083	**IPan:** 0,002					
Sieja:	krepšinis	27,5%	čempionatas	20,8%	rezultatai	17,4%	lietuviai	7,6%
Išskiria:	krepšinis	14,9%	čempionatas	11,2%	rezultatai	8,6%	lietuviai	4,1%
Skirstinys 1	**Dydis:** 1519	**VPan:** 0,026	**IPan:** 0,002					
Sieja:	rinkimai	20,5%	baltarusija	13,1%	palestiniečiai	8,3%	izraelis	7,8%
Išskiria:	rinkimai	11,3%	baltarusija	7,2%	krepšinis	6,3%	čempionatas	4,8%
Skirstinys 2	**Dydis:** 1682	**VPan:** 0,025	**IPan:** 0,002					
Sieja:	aukos	26,8%	irakas	12,6%	iranas	11,4%	branduolinė programa	3,1%
Išskiria:	aukos	15,2%	irakas	6,8%	krepšinis	6,7%	iranas	6,2%
Skirstinys 3	**Dydis:** 2134	**VPan:** 0,025	**IPan:** 0,003					
Sieja:	energetika	11,5%	ekonomika	11,0%	finansai	9,5%	rusija	9,5%
Išskiria:	krepšinis	7,8%	energetika	6,8%	ekonomika	6,5%	čempionatas	6,0%
Skirstinys 4	**Dydis:** 1502	**VPan:** 0,024	**IPan:** 0,002					
Sieja:	įvairybės	39,6%	transportas	10,5%	britanija	9,3%	eismas	4,4%
Išskiria:	įvairybės	22,5%	krepšinis	6,2%	transportas	5,8%	čempionatas	4,5%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 10

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,389, **Grynumas:** 0,784

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	894	0,172	0,101	0,005	0,003	0,080	0,977
1	476	0,104	0,058	0,003	0,002	0,586	0,676
2	602	0,090	0,056	0,004	0,003	0,593	0,688
3	752	0,071	0,037	0,005	0,005	0,110	0,968
4	541	0,060	0,044	0,003	0,002	0,458	0,756
5	863	0,051	0,028	0,003	0,003	0,344	0,856
6	968	0,039	0,028	0,002	0,003	0,271	0,904
7	1083	0,039	0,026	0,002	0,002	0,519	0,506
8	1005	0,029	0,016	0,003	0,002	0,419	0,825
9	1264	0,023	0,019	0,002	0,002	0,558	0,700

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	873	5	1	15	0
1	2	322	96	26	30
2	6	106	414	29	47
3	728	3	1	11	9
4	0	409	11	24	97
5	2	76	739	24	22
6	16	875	8	40	29
7	0	548	5	34	496
8	6	63	829	57	50
9	15	91	30	885	243

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 894	**VPan:** 0,172	**IPan:** 0,005					
Sieja:	krepšinis	43,1%	čempionatas	23,4%	lietuviai	11,3%	rezultatai	9,6%
Išskiria:	krepšinis	24,8%	čempionatas	10,8%	lietuviai	6,3%	įvairybės	2,8%
Skirstinys 1	**Dydis:** 476	**VPan:** 0,104	**IPan:** 0,003					
Sieja:	palestiniečiai	19,5%	izraelis	18,8%	būklė	10,0%	eismas	9,8%
Išskiria:	palestiniečiai	10,4%	izraelis	9,9%	krepšinis	5,4%	būklė	5,4%
Skirstinys 2	**Dydis:** 602	**VPan:** 0,090	**IPan:** 0,004					
Sieja:	energetika	39,3%	rusija	22,9%	dujos	9,5%	nafta	9,3%
Išskiria:	energetika	23,0%	rusija	7,8%	krepšinis	6,2%	dujos	5,5%
Skirstinys 3	**Dydis:** 752	**VPan:** 0,071	**IPan:** 0,005					
Sieja:	futbolas	24,4%	rezultatai	15,1%	ledo ritulys	11,9%	olimpiada	9,0%
Išskiria:	futbolas	15,1%	ledo ritulys	7,4%	krepšinis	6,9%	olimpiada	5,6%
Skirstinys 4	**Dydis:** 541	**VPan:** 0,060	**IPan:** 0,003					
Sieja:	iranas	45,4%	branduolinė programa	12,6%	jt	6,3%	derybos	5,0%
Išskiria:	iranas	25,0%	branduolinė programa	7,1%	krepšinis	5,6%	čempionatas	4,3%
Skirstinys 5	**Dydis:** 863	**VPan:** 0,051	**IPan:** 0,003					
Sieja:	ekonomika	32,3%	finansai	26,9%	es	12,4%	statistika	3,0%
Išskiria:	ekonomika	18,8%	finansai	15,4%	krepšinis	6,3%	es	5,1%
Skirstinys 6	**Dydis:** 968	**VPan:** 0,039	**IPan:** 0,002					
Sieja:	rinkimai	27,5%	baltarusija	21,7%	seimas	11,9%	prezidentas	6,3%
Išskiria:	rinkimai	14,4%	baltarusija	12,1%	seimas	6,7%	krepšinis	5,9%
Skirstinys 7	**Dydis:** 1083	**VPan:** 0,039	**IPan:** 0,002					
Sieja:	aukos	40,5%	irakas	17,9%	sprogimas	4,4%	terorizmas	3,7%
Išskiria:	aukos	22,4%	irakas	9,4%	krepšinis	6,1%	čempionatas	4,6%
Skirstinys 8	**Dydis:** 1005	**VPan:** 0,029	**IPan:** 0,003					
Sieja:	įmonės	14,5%	prekyba	12,8%	transportas	12,6%	žemės ūkis	9,0%

Išskiria:	įmonės	7,7%	prekyba	6,9%	transportas	6,8%	krepšinis	6,6%
Skirstinys 9	**Dydis:** 1264	**VPan:** 0,023	**IPan:** 0,002					
Sieja:	įvairybės	54,3%	britanija	9,9%	kinas	2,9%	muzika	2,2%
Išskiria:	įvairybės	29,6%	krepšinis	5,7%	čempionatas	4,4%	britanija	4,2%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 15

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,340, **Grynumas:** 0,840

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	119	0,580	0,249	0,003	0,002	0,341	0,790
1	564	0,269	0,143	0,010	0,005	0,064	0,979
2	376	0,158	0,077	0,005	0,003	0,057	0,984
3	349	0,142	0,071	0,008	0,005	0,195	0,934
4	369	0,112	0,060	0,003	0,003	0,303	0,881
5	469	0,111	0,057	0,004	0,003	0,264	0,908
6	413	0,103	0,055	0,006	0,006	0,337	0,872
7	570	0,097	0,058	0,004	0,003	0,554	0,714
8	519	0,063	0,043	0,003	0,003	0,584	0,576
9	549	0,055	0,034	0,003	0,003	0,301	0,867
10	832	0,054	0,028	0,003	0,003	0,288	0,885
11	680	0,053	0,031	0,003	0,002	0,420	0,754
12	1042	0,029	0,024	0,002	0,002	0,411	0,825
13	971	0,030	0,016	0,003	0,002	0,372	0,850
14	626	0,027	0,016	0,002	0,002	0,377	0,837

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	0	1	94	0	24
1	552	0	0	12	0
2	370	2	0	4	0
3	326	3	4	14	2
4	2	325	5	27	10
5	16	426	4	12	11
6	360	18	4	10	21
7	4	97	407	25	37
8	1	299	11	29	179
9	0	476	7	10	56
10	2	65	736	15	14
11	0	144	7	16	513
12	9	72	22	860	79
13	3	46	825	54	43
14	3	524	8	57	34

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 119	**VPan:** 0,580	**IPan:** 0,003					
Sieja:	eismas	29,6%	keliai	23,4%	sąlygos	22,1%	būklė	20,9%
Išskiria:	eismas	15,0%	keliai	12,1%	sąlygos	11,4%	būklė	10,5%
Skirstinys 1	**Dydis:** 564	**VPan:** 0,269	**IPan:** 0,010					
Sieja:	čempionatas	37,3%	krepšinis	31,4%	rezultatai	13,7%	lietuviai	12,9%
Išskiria:	čempionatas	20,1%	krepšinis	12,6%	lietuviai	6,7%	rezultatai	3,4%
Skirstinys 2	**Dydis:** 376	**VPan:** 0,158	**IPan:** 0,005					
Sieja:	rezultatai	23,4%	ledo ritulys	21,4%	turinas	16,3%	olimpiada	16,1%
Išskiria:	ledo ritulys	12,2%	turinas	9,3%	olimpiada	9,1%	krepšinis	5,8%

Skirstinys 3	**Dydis:** 349	**VPan:** 0,142	**IPan:** 0,008					
Sieja:	krepšinis	34,9%	eurolyga	30,2%	laureatai	6,7%	žalgiris	3,5%
Išskiria:	eurolyga	19,5%	krepšinis	9,0%	čempionatas	5,1%	laureatai	4,1%
Skirstinys 4	**Dydis:** 369	**VPan:** 0,112	**IPan:** 0,003					
Sieja:	palestiniečiai	31,0%	izraelis	29,4%	hamas	13,5%	šaronas	6,4%
Išskiria:	palestiniečiai	16,7%	izraelis	15,6%	hamas	7,4%	krepšinis	5,4%
Skirstinys 5	**Dydis:** 469	**VPan:** 0,111	**IPan:** 0,004					
Sieja:	rinkimai	41,2%	baltarusija	32,0%	prezidentas	6,8%	parlamentas	5,9%
Išskiria:	rinkimai	21,7%	baltarusija	17,6%	krepšinis	5,1%	čempionatas	4,4%
Skirstinys 6	**Dydis:** 413	**VPan:** 0,103	**IPan:** 0,006					
Sieja:	futbolas	59,2%	čempionatas	5,9%	statistika	5,2%	dviračiai	3,5%
Išskiria:	futbolas	36,5%	krepšinis	6,3%	rezultatai	3,0%	rusija	2,5%
Skirstinys 7	**Dydis:** 570	**VPan:** 0,097	**IPan:** 0,004					
Sieja:	energetika	40,6%	rusija	21,0%	dujos	9,8%	nafta	9,6%
Išskiria:	energetika	23,7%	rusija	6,7%	krepšinis	6,1%	dujos	5,7%
Skirstinys 8	**Dydis:** 519	**VPan:** 0,063	**IPan:** 0,003					
Sieja:	iranas	41,9%	branduolinė programa	13,2%	jt	6,2%	derybos	5,1%
Išskiria:	iranas	22,3%	branduolinė programa	7,4%	krepšinis	5,5%	čempionatas	4,2%
Skirstinys 9	**Dydis:** 549	**VPan:** 0,055	**IPan:** 0,003					
Sieja:	irakas	57,8%	terorizmas	6,9%	bušas	4,4%	aukos	3,8%
Išskiria:	irakas	33,0%	krepšinis	5,8%	čempionatas	4,4%	rezultatai	3,6%
Skirstinys 10	**Dydis:** 832	**VPan:** 0,054	**IPan:** 0,003					
Sieja:	ekonomika	32,9%	finansai	26,9%	es	12,3%	statistika	2,7%
Išskiria:	ekonomika	19,2%	finansai	15,3%	krepšinis	6,3%	es	4,9%
Skirstinys 11	**Dydis:** 680	**VPan:** 0,053	**IPan:** 0,003					
Sieja:	aukos	50,8%	sprogimas	5,7%	stichija	4,1%	incidentas	3,9%
Išskiria:	aukos	26,2%	krepšinis	5,9%	čempionatas	4,5%	rezultatai	4,1%
Skirstinys 12	**Dydis:** 1042	**VPan:** 0,029	**IPan:** 0,002					
Sieja:	įvairybės	61,2%	britanija	10,7%	kinas	3,4%	muzika	2,5%
Išskiria:	įvairybės	33,1%	krepšinis	5,6%	britanija	4,5%	čempionatas	4,3%
Skirstinys 13	**Dydis:** 971	**VPan:** 0,030	**IPan:** 0,003					
Sieja:	įmonės	15,2%	prekyba	13,1%	transportas	12,4%	žemės ūkis	9,1%
Išskiria:	įmonės	8,1%	prekyba	7,1%	transportas	6,6%	krepšinis	6,5%
Skirstinys 14	**Dydis:** 626	**VPan:** 0,027	**IPan:** 0,002					
Sieja:	seimas	37,5%	adamkus	8,7%	paulauskas	5,4%	vizitas	5,2%
Išskiria:	seimas	20,8%	krepšinis	5,5%	adamkus	4,9%	čempionatas	4,2%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 20

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,316, **Grynumas:** 0,850

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	167	0,312	0,098	0,005	0,004	0,110	0,964
1	555	0,276	0,143	0,010	0,005	0,056	0,982
2	218	0,227	0,128	0,006	0,003	0,000	1,000
3	338	0,150	0,071	0,008	0,005	0,127	0,962
4	283	0,138	0,113	0,002	0,002	0,475	0,650
5	356	0,120	0,060	0,003	0,003	0,252	0,910
6	465	0,112	0,057	0,004	0,003	0,230	0,923
7	502	0,111	0,064	0,005	0,003	0,526	0,717
8	404	0,106	0,055	0,006	0,006	0,264	0,906
9	278	0,093	0,051	0,004	0,003	0,484	0,763
10	419	0,092	0,049	0,005	0,003	0,389	0,811
11	542	0,084	0,041	0,003	0,002	0,421	0,815
12	455	0,078	0,051	0,003	0,002	0,344	0,826

13	425	0,074	0,032	0,004	0,003	0,203	0,934
14	368	0,074	0,042	0,004	0,003	0,344	0,853
15	452	0,072	0,041	0,003	0,003	0,258	0,874
16	635	0,057	0,033	0,003	0,002	0,440	0,720
17	420	0,052	0,028	0,004	0,003	0,329	0,871
18	516	0,036	0,020	0,002	0,002	0,274	0,897
19	650	0,015	0,009	0,001	0,001	0,468	0,782

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	161	2	0	4	0
1	545	0	0	10	0
2	218	0	0	0	0
3	325	2	2	8	1
4	0	5	91	3	184
5	2	324	5	17	8
6	14	429	3	8	11
7	1	93	360	15	33
8	366	15	2	7	14
9	1	33	212	5	27
10	0	58	340	8	13
11	4	47	10	442	39
12	0	376	6	6	67
13	2	8	397	9	9
14	1	9	314	36	8
15	0	395	3	1	53
16	0	158	4	16	457
17	1	25	366	13	15
18	2	463	7	29	15
19	5	56	12	508	69

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 167	**VPan:** 0,312	**IPan:** 0,005					
Sieja:	olimpiada	42,5%	turinas	41,7%	rezultatai	4,6%	slidinėjimas	3,4%
Išskiria:	olimpiada	23,4%	turinas	23,0%	krepšinis	5,3%	čempionatas	4,0%
Skirstinys 1	**Dydis:** 555	**VPan:** 0,276	**IPan:** 0,010					
Sieja:	čempionatas	37,4%	krepšinis	31,6%	rezultatai	13,8%	lietuviai	12,7%
Išskiria:	čempionatas	20,1%	krepšinis	12,7%	lietuviai	6,6%	rezultatai	3,4%
Skirstinys 2	**Dydis:** 218	**VPan:** 0,227	**IPan:** 0,006					
Sieja:	ledo ritulys	37,3%	rezultatai	25,3%	zubrus	9,2%	kasparaitis	9,0%
Išskiria:	ledo ritulys	21,0%	krepšinis	5,7%	rezultatai	5,5%	zubrus	5,3%
Skirstinys 3	**Dydis:** 338	**VPan:** 0,150	**IPan:** 0,008					
Sieja:	krepšinis	35,4%	eurolyga	30,6%	laureatai	6,2%	žalgiris	3,5%
Išskiria:	eurolyga	19,8%	krepšinis	9,1%	čempionatas	5,1%	laureatai	3,7%
Skirstinys 4	**Dydis:** 283	**VPan:** 0,138	**IPan:** 0,002					
Sieja:	eismas	20,3%	keliai	16,8%	sąlygos	16,4%	būklė	15,5%
Išskiria:	eismas	10,2%	keliai	8,8%	sąlygos	8,6%	būklė	7,8%
Skirstinys 5	**Dydis:** 356	**VPan:** 0,120	**IPan:** 0,003					
Sieja:	palestiniečiai	31,2%	izraelis	29,7%	hamas	13,6%	šaronas	6,4%
Išskiria:	palestiniečiai	16,8%	izraelis	15,7%	hamas	7,4%	krepšinis	5,3%
Skirstinys 6	**Dydis:** 465	**VPan:** 0,112	**IPan:** 0,004					
Sieja:	rinkimai	40,8%	baltarusija	31,7%	prezidentas	7,1%	parlamentas	6,2%
Išskiria:	rinkimai	21,4%	baltarusija	17,4%	krepšinis	5,1%	čempionatas	4,4%
Skirstinys 7	**Dydis:** 502	**VPan:** 0,111	**IPan:** 0,005					
Sieja:	energetika	42,2%	rusija	18,1%	nafta	10,7%	dujos	10,7%
Išskiria:	energetika	24,1%	dujos	6,2%	nafta	6,1%	krepšinis	6,0%
Skirstinys 8	**Dydis:** 404	**VPan:** 0,106	**IPan:** 0,006					
Sieja:	futbolas	59,7%	čempionatas	6,0%	statistika	4,6%	dviračiai	3,7%

Išskiria:	futbolas	36,7%	krepšinis	6,3%	rezultatai	3,0%	rusija	2,4%
Skirstinys 9 **Dydis:** 278 **VPan:** 0,093 **IPan:** 0,004								
Sieja:	transportas	56,7%	aviacija	14,3%	vokietija	6,6%	automobiliai	6,2%
Išskiria:	transportas	31,1%	aviacija	8,1%	krepšinis	5,6%	čempionatas	4,3%
Skirstinys 10 **Dydis:** 419 **VPan:** 0,092 **IPan:** 0,005								
Sieja:	ekonomika	54,5%	es	13,2%	statistika	5,9%	darbas	4,2%
Išskiria:	ekonomika	31,1%	krepšinis	6,2%	čempionatas	4,7%	es	4,7%
Skirstinys 11 **Dydis:** 542 **VPan:** 0,084 **IPan:** 0,003								
Sieja:	įvairybės	78,0%	britanija	12,7%	didžioji	1,4%	anekdotai	1,0%
Išskiria:	įvairybės	41,9%	krepšinis	5,4%	britanija	5,3%	čempionatas	4,2%
Skirstinys 12 **Dydis:** 455 **VPan:** 0,078 **IPan:** 0,003								
Sieja:	iranas	45,8%	branduolinė programa	13,8%	jt	6,8%	derybos	5,2%
Išskiria:	iranas	24,7%	branduolinė programa	7,8%	krepšinis	5,6%	čempionatas	4,3%
Skirstinys 13 **Dydis:** 425 **VPan:** 0,074 **IPan:** 0,004								
Sieja:	finansai	71,8%	euras	3,3%	es	2,4%	ekonomika	2,0%
Išskiria:	finansai	40,4%	krepšinis	5,7%	čempionatas	4,4%	rezultatai	3,8%
Skirstinys 14 **Dydis:** 368 **VPan:** 0,074 **IPan:** 0,004								
Sieja:	prekyba	30,0%	žemės ūkis	28,3%	paukščių gripas	20,5%	kainos	2,3%
Išskiria:	žemės ūkis	16,2%	prekyba	15,7%	paukščių gripas	11,8%	krepšinis	5,7%
Skirstinys 15 **Dydis:** 452 **VPan:** 0,072 **IPan:** 0,003								
Sieja:	irakas	63,6%	terorizmas	7,9%	bušas	4,8%	aukos	3,9%
Išskiria:	irakas	35,9%	krepšinis	5,7%	čempionatas	4,4%	terorizmas	4,0%
Skirstinys 16 **Dydis:** 635 **VPan:** 0,057 **IPan:** 0,003								
Sieja:	aukos	53,4%	sprogimas	6,0%	incidentas	4,6%	katastriša	3,9%
Išskiria:	aukos	27,7%	krepšinis	5,9%	čempionatas	4,5%	rezultatai	4,1%
Skirstinys 17 **Dydis:** 420 **VPan:** 0,052 **IPan:** 0,004								
Sieja:	įmonės	50,6%	verslas	12,4%	paslaugos	8,7%	japonija	3,1%
Išskiria:	įmonės	28,3%	verslas	6,6%	krepšinis	5,8%	paslaugos	4,8%
Skirstinys 18 **Dydis:** 516 **VPan:** 0,036 **IPan:** 0,002								
Sieja:	seimas	41,8%	adamkus	9,7%	paulauskas	6,0%	vizitas	5,6%
Išskiria:	seimas	22,9%	krepšinis	5,4%	adamkus	5,4%	čempionatas	4,1%
Skirstinys 19 **Dydis:** 650 **VPan:** 0,015 **IPan:** 0,001								
Sieja:	muzika	9,7%	vilnius	8,6%	kinas	7,7%	kaunas	5,4%
Išskiria:	krepšinis	5,3%	muzika	5,1%	vilnius	4,1%	čempionatas	4,0%

PRIEDAS D – EKSPERIMENTŲ ŽURNALAS

Šiame priede pateikiamas tradicinio dokumentų kolekcijų klasterizavimo skaldančiu K-means algoritmu eksperimento žurnalas. Klasterizavimo metu dokumentų reprezentacijai naudotas dokumentų tekstas.

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 5

Suklasterizuota dokumentų: [1056 iš 1056], **Entropija:** 0,348, **Grynumas:** 0,811

Skirstinys	Dydis[39]	VPan[40]	VNuok[41]	IPan[42]	INuok[43]	Entropija	Grynumas
0	174	0,076	0,032	0,008	0,003	0,022	0,994
1	248	0,037	0,012	0,012	0,003	0,283	0,899
2	233	0,036	0,010	0,012	0,003	0,133	0,961
3	193	0,033	0,008	0,011	0,003	0,603	0,565
4	208	0,025	0,007	0,012	0,004	0,700	0,611

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	1	0	173
1	11	3	223	6	5
2	224	2	3	3	1
3	65	109	2	16	1
4	11	8	43	127	19

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 174	**VPan:** 0,076	**IPan:** 0,008						
Sieja:		taškų	4,0%	pelnė	3,0%	min	2,7%	taškus	2,5%
Išskiria:		taškų	2,5%	pelnė	1,9%	min	1,7%	taškus	1,6%
Skirstinys 1	**Dydis:** 248	**VPan:** 0,037	**IPan:** 0,012						
Sieja:		proc	10,4%	mln	4,7%	naftos	3,2%	litų	3,1%
Išskiria:		proc	8,8%	mln	4,1%	naftos	2,9%	litų	2,5%
Skirstinys 2	**Dydis:** 233	**VPan:** 0,036	**IPan:** 0,012						
Sieja:		seimo	5,5%	hamas	2,1%	palestiniečių	1,7%	kad	1,7%
Išskiria:		seimo	5,4%	hamas	2,1%	palestiniečių	1,6%	proc	1,5%
Skirstinys 3	**Dydis:** 193	**VPan:** 0,033	**IPan:** 0,011						
Sieja:		žuvo	4,5%	policijos	2,4%	buvo	2,3%	žmonės	1,9%
Išskiria:		žuvo	4,1%	policijos	2,1%	proc	1,8%	žmonės	1,5%
Skirstinys 4	**Dydis:** 208	**VPan:** 0,025	**IPan:** 0,012						
Sieja:		paukščių	4,4%	gripo	2,2%	lietuvos	1,9%	kultūros	1,2%
Išskiria:		paukščių	5,0%	gripo	2,5%	proc	2,1%	kultūros	1,3%

39 Skirstiniui priskirtų dokumentų skaičius.

40 Vidutinis panašumas tarp skirstinyje esančių dokumentų (**V**idinis **Pan**ašumas).

41 Skirstinyje esančių dokumentų panašumų standartinis nuokrypis (**V**idinis **Nuok**rypis).

42 Vidutinis panašumas tarp skirstinyje esančių dokumentų ir likusių kolekcijos dokumentų (**I**šorinis **Pan**ašumas).

43 Skirstinyje esančių dokumentų panašumų su likusiais kolekcijos dokumentais standartinis nuokrypis (**I**šorinis **Nuok**rypis).

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 10

Suklasterizuota dokumentų: [1056 iš 1056], **Entropija:** 0,301, **Grynumas:** 0,842

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	29	0,156	0,058	0,014	0,004	0,366	0,724
1	106	0,122	0,043	0,011	0,003	0,000	1,000
2	42	0,123	0,035	0,013	0,003	0,000	1,000
3	91	0,067	0,022	0,014	0,004	0,150	0,956
4	82	0,062	0,021	0,011	0,003	0,041	0,988
5	107	0,053	0,017	0,013	0,004	0,148	0,953
6	138	0,052	0,019	0,013	0,003	0,362	0,862
7	113	0,041	0,011	0,013	0,003	0,184	0,929
8	178	0,035	0,008	0,012	0,003	0,565	0,596
9	170	0,027	0,006	0,012	0,004	0,609	0,706

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	21	8	0
1	0	0	0	0	106
2	42	0	0	0	0
3	87	1	1	1	1
4	0	0	1	0	81
5	2	0	102	1	2
6	6	6	119	5	2
7	105	0	6	2	0
8	56	106	0	15	1
9	13	9	22	120	6

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 29 **VPan:** 0,156 **IPan:** 0,014								
Sieja:	paukščių	36,1%	gripo	17,9%	viruso	2,4%	virusas	1,4%
Išskiria:	paukščių	24,6%	gripo	12,2%	viruso	1,6%	proc	1,1%
Skirstinys 1 **Dydis:** 106 **VPan:** 0,122 **IPan:** 0,011								
Sieja:	taškų	6,2%	pelnė	3,9%	taškus	3,0%	24	2,2%
Išskiria:	taškų	3,9%	pelnė	2,3%	taškus	1,8%	proc	1,0%
Skirstinys 2 **Dydis:** 42 **VPan:** 0,123 **IPan:** 0,013								
Sieja:	hamas	17,7%	palestiniečių	15,6%	izraelio	9,0%	gazos	1,5%
Išskiria:	hamas	12,4%	palestiniečių	10,9%	izraelio	6,3%	proc	1,1%
Skirstinys 3 **Dydis:** 91 **VPan:** 0,067 **IPan:** 0,014								
Sieja:	seimo	18,8%	rinkimų	3,5%	partijos	3,1%	baltarusijos	3,0%
Išskiria:	seimo	15,8%	rinkimų	2,5%	partijos	2,3%	baltarusijos	2,2%
Skirstinys 4 **Dydis:** 82 **VPan:** 0,062 **IPan:** 0,011								
Sieja:	sek	3,9%	pasaulio	2,9%	min	2,5%	vietą	1,9%
Išskiria:	sek	2,8%	pasaulio	1,6%	proc	1,3%	ledo	1,1%
Skirstinys 5 **Dydis:** 107 **VPan:** 0,053 **IPan:** 0,013								
Sieja:	naftos	11,1%	dujų	6,8%	jav	6,0%	dolerių	4,7%
Išskiria:	naftos	9,7%	dujų	5,6%	dolerių	3,5%	jav	3,0%
Skirstinys 6 **Dydis:** 138 **VPan:** 0,052 **IPan:** 0,013								
Sieja:	proc	17,6%	litų	7,3%	mln	5,3%	es	2,8%
Išskiria:	proc	13,5%	litų	6,1%	mln	3,5%	tūkst	1,3%
Skirstinys 7 **Dydis:** 113 **VPan:** 0,041 **IPan:** 0,013								
Sieja:	al	4,4%	es	2,4%	gynybos	2,3%	irake	2,3%
Išskiria:	al	3,9%	gynybos	2,1%	irake	1,9%	irano	1,6%
Skirstinys 8 **Dydis:** 178 **VPan:** 0,035 **IPan:** 0,012								
Sieja:	žuvo	4,9%	policijos	2,7%	buvo	2,1%	žmonės	2,1%

Išskiria:	žuvo	4,4%	policijos	2,3%	proc	1,7%	žmonės	1,6%
Skirstinys 9 **Dydis:** 170 **VPan:** 0,027 **IPan:** 0,012								
Sieja:	lietuvos	2,4%	kultūros	1,7%	teatro	1,5%	vilniaus	1,5%
Išskiria:	proc	1,9%	kultūros	1,8%	teatro	1,6%	muzikos	1,4%

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 15

Suklasterizuota dokumentų: [1056 iš 1056], **Entropija:** 0,355, **Grynumas:** 0,792

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	17	0,320	0,119	0,012	0,003	0,276	0,882
1	25	0,193	0,056	0,014	0,004	0,342	0,760
2	98	0,134	0,044	0,011	0,003	0,000	1,000
3	30	0,133	0,033	0,012	0,004	0,291	0,867
4	45	0,134	0,037	0,014	0,004	0,066	0,978
5	57	0,092	0,026	0,013	0,003	0,000	1,000
6	51	0,089	0,038	0,014	0,004	0,481	0,784
7	66	0,084	0,029	0,014	0,004	0,279	0,879
8	77	0,067	0,023	0,014	0,003	0,204	0,922
9	70	0,060	0,016	0,011	0,003	0,127	0,957
10	65	0,055	0,014	0,015	0,003	0,249	0,908
11	81	0,045	0,011	0,011	0,003	0,357	0,840
12	96	0,048	0,013	0,014	0,003	0,455	0,698
13	136	0,042	0,009	0,012	0,003	0,497	0,625
14	142	0,027	0,004	0,014	0,004	0,890	0,437

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	1	0	1	15
1	0	0	19	6	0
2	0	0	0	0	98
3	1	3	26	0	0
4	44	1	0	0	0
5	57	0	0	0	0
6	1	6	40	2	2
7	5	0	58	3	0
8	3	0	71	0	3
9	0	0	1	2	67
10	59	0	3	2	1
11	0	2	9	68	2
12	67	1	26	2	0
13	46	85	0	4	1
14	28	23	19	62	10

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 17 **VPan:** 0,320 **IPan:** 0,012								
Sieja:	ledo	5,7%	capitals	5,6%	rangers	4,5%	rungtynių	3,6%
Išskiria:	capitals	3,4%	ledo	3,1%	rangers	2,7%	rungtynių	1,8%
Skirstinys 1 **Dydis:** 25 **VPan:** 0,193 **IPan:** 0,014								
Sieja:	paukščių	39,2%	gripo	19,4%	viruso	2,6%	virusas	1,6%
Išskiria:	paukščių	25,8%	gripo	12,8%	viruso	1,7%	proc	1,1%
Skirstinys 2 **Dydis:** 98 **VPan:** 0,134 **IPan:** 0,011								
Sieja:	taškų	6,5%	pelnė	4,1%	taškus	3,2%	24	2,2%
Išskiria:	taškų	4,1%	pelnė	2,4%	taškus	1,9%	24	1,0%

Skirstinys 3	**Dydis:** 30	**VPan:** 0,133	**IPan:** 0,012					
Sieja:	dujų	30,7%	gazprom	6,7%	keliai	3,2%	vietomis	3,2%
Išskiria:	dujų	19,7%	gazprom	4,4%	keliai	2,1%	vietomis	2,1%
Skirstinys 4	**Dydis:** 45	**VPan:** 0,134	**IPan:** 0,014					
Sieja:	seimo	38,0%	partijos	4,4%	frakcijos	2,9%	liberalų	2,1%
Išskiria:	seimo	26,7%	partijos	2,7%	frakcijos	2,1%	liberalų	1,5%
Skirstinys 5	**Dydis:** 57	**VPan:** 0,092	**IPan:** 0,013					
Sieja:	hamas	13,1%	palestiniečių	11,8%	al	8,5%	izraelio	6,9%
Išskiria:	hamas	9,7%	palestiniečių	8,7%	al	5,9%	izraelio	5,1%
Skirstinys 6	**Dydis:** 51	**VPan:** 0,089	**IPan:** 0,014					
Sieja:	litų	28,6%	mln	14,0%	pieno	1,4%	tūkst	1,1%
Išskiria:	litų	21,6%	mln	8,6%	pieno	1,0%	jav	0,7%
Skirstinys 7	**Dydis:** 66	**VPan:** 0,084	**IPan:** 0,014					
Sieja:	proc	35,3%	tūkst	2,4%	procento	2,3%	palyginti	1,8%
Išskiria:	proc	24,2%	procento	1,7%	palyginti	1,2%	euro	1,2%
Skirstinys 8	**Dydis:** 77	**VPan:** 0,067	**IPan:** 0,014					
Sieja:	naftos	16,0%	jav	8,8%	dolerių	6,6%	mlrd	4,5%
Išskiria:	naftos	13,3%	dolerių	4,8%	jav	4,7%	mlrd	3,2%
Skirstinys 9	**Dydis:** 70	**VPan:** 0,060	**IPan:** 0,011					
Sieja:	pasaulio	4,3%	sek	3,2%	vietą	1,9%	užėmė	1,7%
Išskiria:	pasaulio	2,5%	sek	2,1%	užėmė	1,3%	proc	1,2%
Skirstinys 10	**Dydis:** 65	**VPan:** 0,055	**IPan:** 0,015					
Sieja:	baltarusijos	5,9%	rinkimų	5,4%	prezidento	4,0%	prezidentas	3,0%
Išskiria:	baltarusijos	4,9%	rinkimų	4,4%	prezidento	2,9%	prezidentas	2,0%
Skirstinys 11	**Dydis:** 81	**VPan:** 0,045	**IPan:** 0,011					
Sieja:	lietuvos	4,1%	kultūros	3,9%	vilniaus	3,7%	teatro	3,7%
Išskiria:	kultūros	3,0%	teatro	3,0%	muzikos	2,4%	vilniaus	2,0%
Skirstinys 12	**Dydis:** 96	**VPan:** 0,048	**IPan:** 0,014					
Sieja:	es	9,3%	europos	3,3%	irano	2,1%	gynybos	1,9%
Išskiria:	es	6,9%	irano	1,9%	gynybos	1,6%	europos	1,6%
Skirstinys 13	**Dydis:** 136	**VPan:** 0,042	**IPan:** 0,012					
Sieja:	žuvo	6,3%	žmonės	2,6%	policijos	2,3%	žmonių	2,1%
Išskiria:	žuvo	5,2%	žmonės	1,8%	sužeisti	1,7%	policijos	1,7%
Skirstinys 14	**Dydis:** 142	**VPan:** 0,027	**IPan:** 0,014					
Sieja:	kad	1,7%	buvo	1,2%	už	1,1%	metų	1,0%
Išskiria:	proc	2,1%	es	1,2%	seimo	1,0%	lietuvos	0,9%

SKALDANTIS K-MEANS, 1056 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 1056 **Skirstinių:** 20

Suklasterizuota dokumentų: [1056 iš 1056], **Entropija:** 0,306, **Grynumas:** 0,834

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	16	0,355	0,110	0,013	0,003	0,145	0,938
1	38	0,249	0,073	0,016	0,004	0,000	1,000
2	25	0,193	0,056	0,014	0,004	0,342	0,760
3	26	0,189	0,051	0,016	0,003	0,000	1,000
4	35	0,150	0,038	0,014	0,003	0,000	1,000
5	28	0,146	0,031	0,012	0,005	0,191	0,929
6	45	0,134	0,037	0,014	0,004	0,066	0,978
7	69	0,115	0,034	0,013	0,004	0,000	1,000
8	30	0,102	0,030	0,013	0,003	0,000	1,000
9	44	0,102	0,043	0,014	0,004	0,498	0,773
10	54	0,098	0,033	0,014	0,004	0,295	0,870
11	63	0,067	0,016	0,011	0,003	0,051	0,984
12	62	0,060	0,016	0,014	0,003	0,208	0,919

13	65	0,056	0,012	0,011	0,003	0,214	0,923
14	62	0,057	0,015	0,015	0,003	0,258	0,903
15	65	0,056	0,015	0,014	0,004	0,368	0,846
16	100	0,053	0,010	0,012	0,003	0,531	0,590
17	69	0,053	0,015	0,015	0,004	0,529	0,681
18	71	0,042	0,008	0,013	0,003	0,621	0,648
19	89	0,033	0,005	0,013	0,003	0,707	0,629

Pasiskirstymas klasėse

Skirstinys	Politika	Teisėtvarka	Ūkis	Kultūra	Sportas
0	0	0	0	1	15
1	0	0	0	0	38
2	0	0	19	6	0
3	0	0	26	0	0
4	35	0	0	0	0
5	1	1	26	0	0
6	44	1	0	0	0
7	0	0	0	0	69
8	30	0	0	0	0
9	1	5	34	3	1
10	4	0	47	3	0
11	0	0	1	0	62
12	57	0	3	2	0
13	0	1	3	60	1
14	56	0	3	2	1
15	4	0	55	2	4
16	36	59	0	4	1
17	16	1	47	5	0
18	13	46	4	8	0
19	14	8	4	56	7

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 16	**VPan:** 0,355	**IPan:** 0,013					
Sieja:	ledo	5,8%	capitals	5,7%	rangers	4,6%	rungtynių	3,7%
Išskiria:	capitals	3,4%	ledo	3,2%	rangers	2,7%	rungtynių	1,8%
Skirstinys 1	**Dydis:** 38	**VPan:** 0,249	**IPan:** 0,016					
Sieja:	taškų	4,0%	taškus	3,6%	24	3,1%	29	2,6%
Išskiria:	taškus	1,7%	taškų	1,6%	29	1,3%	24	1,3%
Skirstinys 2	**Dydis:** 25	**VPan:** 0,193	**IPan:** 0,014					
Sieja:	paukščių	39,2%	gripo	19,4%	viruso	2,6%	virusas	1,6%
Išskiria:	paukščių	25,8%	gripo	12,8%	viruso	1,7%	proc	1,1%
Skirstinys 3	**Dydis:** 26	**VPan:** 0,189	**IPan:** 0,016					
Sieja:	naftos	45,3%	barelių	6,4%	jav	2,5%	dolerio	1,9%
Išskiria:	naftos	30,5%	barelių	4,4%	barelį	1,3%	dolerio	1,2%
Skirstinys 4	**Dydis:** 35	**VPan:** 0,150	**IPan:** 0,014					
Sieja:	hamas	20,9%	palestiniečių	18,4%	izraelio	7,4%	gazos	1,6%
Išskiria:	hamas	14,2%	palestiniečių	12,5%	izraelio	4,7%	proc	1,1%
Skirstinys 5	**Dydis:** 28	**VPan:** 0,146	**IPan:** 0,012					
Sieja:	dujų	31,2%	gazprom	7,0%	keliai	3,4%	vietomis	3,3%
Išskiria:	dujų	19,7%	gazprom	4,6%	keliai	2,2%	vietomis	2,2%
Skirstinys 6	**Dydis:** 45	**VPan:** 0,134	**IPan:** 0,014					
Sieja:	seimo	38,0%	partijos	4,4%	frakcijos	2,9%	liberalų	2,1%
Išskiria:	seimo	26,7%	partijos	2,7%	frakcijos	2,1%	liberalų	1,5%
Skirstinys 7	**Dydis:** 69	**VPan:** 0,115	**IPan:** 0,013					
Sieja:	taškų	5,4%	pelnė	3,4%	min	2,5%	13	2,0%
Išskiria:	taškų	2,8%	pelnė	1,6%	krepšinio	1,2%	min	1,1%
Skirstinys 8	**Dydis:** 30	**VPan:** 0,102	**IPan:** 0,013					
Sieja:	al	28,8%	irako	5,5%	sirijos	2,2%	šiitų	1,7%

Išskiria:	al	20,2%	irako	3,2%	sirijos	1,6%	šaronas	1,2%
Skirstinys 9 **Dydis:** 44 **VPan:** 0,102 **IPan:** 0,014								
Sieja:	litų	32,8%	mln	13,6%	pieno	1,2%	proc	1,0%
Išskiria:	litų	23,6%	mln	7,8%	pieno	0,8%	jav	0,7%
Skirstinys 10 **Dydis:** 54 **VPan:** 0,098 **IPan:** 0,014								
Sieja:	proc	38,0%	procento	2,8%	tūkst	2,5%	palyginti	2,3%
Išskiria:	proc	24,4%	procento	2,0%	palyginti	1,5%	ketvirtį	0,8%
Skirstinys 11 **Dydis:** 63 **VPan:** 0,067 **IPan:** 0,011								
Sieja:	pasaulio	4,6%	sek	3,6%	vietą	2,0%	varžybose	1,7%
Išskiria:	pasaulio	2,7%	sek	2,3%	proc	1,2%	varžybose	1,2%
Skirstinys 12 **Dydis:** 62 **VPan:** 0,060 **IPan:** 0,014								
Sieja:	irano	4,1%	gynybos	2,7%	rusijos	2,7%	iranas	2,2%
Išskiria:	irano	3,3%	gynybos	2,0%	iranas	1,9%	nato	1,6%
Skirstinys 13 **Dydis:** 65 **VPan:** 0,056 **IPan:** 0,011								
Sieja:	teatro	4,8%	kultūros	4,4%	muzikos	3,9%	vilniaus	3,5%
Išskiria:	teatro	3,6%	kultūros	3,2%	muzikos	2,9%	lietuvių	1,8%
Skirstinys 14 **Dydis:** 62 **VPan:** 0,057 **IPan:** 0,015								
Sieja:	rinkimų	5,8%	baltarusijos	5,4%	prezidento	4,0%	prezidentas	3,2%
Išskiria:	rinkimų	4,6%	baltarusijos	4,4%	prezidento	2,9%	prezidentas	2,2%
Skirstinys 15 **Dydis:** 65 **VPan:** 0,056 **IPan:** 0,014								
Sieja:	dolerių	9,0%	jav	8,5%	mlrd	4,7%	prekybos	2,1%
Išskiria:	dolerių	7,1%	jav	4,4%	mlrd	3,4%	google	1,7%
Skirstinys 16 **Dydis:** 100 **VPan:** 0,053 **IPan:** 0,012								
Sieja:	žuvo	9,2%	žmonės	3,0%	sužeisti	2,6%	policijos	2,5%
Išskiria:	žuvo	7,4%	sužeisti	2,1%	žmonės	2,0%	policijos	1,7%
Skirstinys 17 **Dydis:** 69 **VPan:** 0,053 **IPan:** 0,015								
Sieja:	es	17,7%	europos	6,5%	darbo	2,5%	euro	1,3%
Išskiria:	es	14,4%	europos	4,1%	darbo	1,5%	euro	1,1%
Skirstinys 18 **Dydis:** 71 **VPan:** 0,042 **IPan:** 0,013								
Sieja:	žemės	2,3%	pareigūnai	2,0%	buvo	1,8%	policijos	1,7%
Išskiria:	proc	1,5%	žemės	1,4%	pareigūnai	1,1%	policijos	0,9%
Skirstinys 19 **Dydis:** 89 **VPan:** 0,033 **IPan:** 0,013								
Sieja:	kad	1,4%	metų	1,2%	jis	1,1%	yra	1,0%
Išskiria:	proc	1,7%	es	0,9%	lietuvos	0,8%	popiežius	0,7%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 5

Suklasterizuota dokumentų: [2112 iš 2112], **Entropija:** 0,355, **Grynumas:** 0,774

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	203	0,127	0,040	0,010	0,003	0,000	1,000
1	248	0,037	0,012	0,010	0,003	0,057	0,984
2	483	0,031	0,011	0,010	0,003	0,204	0,934
3	516	0,027	0,007	0,010	0,002	0,244	0,915
4	662	0,017	0,004	0,010	0,003	0,773	0,399

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	0	203	0	0	0
1	3	244	0	1	0
2	8	4	6	451	14
3	11	5	5	23	472
4	264	1	235	54	108

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 203	**VPan:** 0,127	**IPan:** 0,010					
Sieja:	taškų	6,3%	pelnė	4,0%	taškus	3,0%	14	2,6%
Išskiria:	taškų	4,0%	pelnė	2,3%	taškus	1,8%	proc	1,2%
Skirstinys 1	**Dydis:** 248	**VPan:** 0,037	**IPan:** 0,010					
Sieja:	sek	3,6%	pasaulio	2,6%	min	2,2%	futbolo	1,7%
Išskiria:	sek	3,0%	pasaulio	1,6%	proc	1,6%	futbolo	1,3%
Skirstinys 2	**Dydis:** 483	**VPan:** 0,031	**IPan:** 0,010					
Sieja:	proc	10,7%	mln	3,7%	litų	2,9%	naftos	2,6%
Išskiria:	proc	9,0%	mln	3,1%	naftos	2,3%	litų	2,2%
Skirstinys 3	**Dydis:** 516	**VPan:** 0,027	**IPan:** 0,010					
Sieja:	seimo	4,5%	kad	1,9%	prezidento	1,4%	prezidentas	1,4%
Išskiria:	seimo	4,5%	prezidento	1,3%	proc	1,3%	rinkimų	1,1%
Skirstinys 4	**Dydis:** 662	**VPan:** 0,017	**IPan:** 0,010					
Sieja:	buvo	1,6%	žuvo	1,5%	žmonių	1,2%	policijos	1,2%
Išskiria:	proc	2,6%	žuvo	1,6%	policijos	1,3%	taškų	1,1%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 10

Suklasterizuota dokumentų: [2112 iš 2112], **Entropija:** 0,339, **Grynumas:** 0,813

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	202	0,128	0,040	0,010	0,003	0,000	1,000
1	71	0,127	0,049	0,011	0,003	0,080	0,972
2	69	0,102	0,029	0,011	0,003	0,094	0,971
3	244	0,049	0,020	0,012	0,003	0,338	0,873
4	202	0,044	0,016	0,011	0,003	0,093	0,975
5	196	0,042	0,012	0,012	0,003	0,225	0,918
6	247	0,037	0,012	0,010	0,003	0,046	0,988
7	233	0,034	0,011	0,012	0,003	0,457	0,738
8	282	0,028	0,007	0,010	0,002	0,708	0,532
9	366	0,020	0,005	0,010	0,003	0,659	0,607

Pasiskirstymas klasėse

Skirstinys	**Kultūra**	**Sportas**	**Teisėtvarka**	**Ūkis**	**Politika**
0	0	202	0	0	0
1	2	0	0	0	69
2	0	1	0	1	67
3	12	4	5	213	10
4	1	1	1	197	2
5	8	0	6	2	180
6	2	244	0	1	0
7	5	3	2	51	172
8	34	0	150	21	77
9	222	2	82	43	17

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 202	**VPan:** 0,128	**IPan:** 0,010					
Sieja:	taškų	6,2%	pelnė	4,0%	taškus	3,0%	14	2,6%
Išskiria:	taškų	3,9%	pelnė	2,3%	taškus	1,7%	proc	1,2%
Skirstinys 1	**Dydis:** 71	**VPan:** 0,127	**IPan:** 0,011					
Sieja:	seimo	47,5%	partijos	4,5%	paulauskas	2,0%	frakcijos	2,0%
Išskiria:	seimo	32,0%	partijos	2,6%	frakcijos	1,3%	paulauskas	1,3%
Skirstinys 2	**Dydis:** 69	**VPan:** 0,102	**IPan:** 0,011					

Sieja:	izraelio	17,0%	palestiniečių	16,8%	hamas	12,4%	gazos	1,9%
Išskiria:	palestiniečių	11,6%	izraelio	11,5%	hamas	8,6%	gazos	1,3%
Skirstinys 3	**Dydis:** 244	**VPan:** 0,049	**IPan:** 0,012					
Sieja:	proc	18,6%	litų	6,9%	mln	4,7%	procento	2,9%
Išskiria:	proc	13,7%	litų	5,3%	mln	3,0%	procento	2,4%
Skirstinys 4	**Dydis:** 202	**VPan:** 0,044	**IPan:** 0,011					
Sieja:	naftos	10,0%	dujų	6,4%	jav	6,3%	dolerių	3,9%
Išskiria:	naftos	8,4%	dujų	5,3%	jav	3,0%	dolerių	2,8%
Skirstinys 5	**Dydis:** 196	**VPan:** 0,042	**IPan:** 0,012					
Sieja:	al	3,9%	irano	3,2%	jav	2,9%	irake	2,5%
Išskiria:	al	3,0%	irano	2,7%	irake	2,2%	jt	2,1%
Skirstinys 6	**Dydis:** 247	**VPan:** 0,037	**IPan:** 0,010					
Sieja:	sek	3,6%	pasaulio	2,5%	min	2,2%	futbolo	1,7%
Išskiria:	sek	3,0%	pasaulio	1,6%	proc	1,6%	futbolo	1,3%
Skirstinys 7	**Dydis:** 233	**VPan:** 0,034	**IPan:** 0,012					
Sieja:	es	6,0%	baltarusijos	5,2%	europos	3,5%	rinkimų	3,0%
Išskiria:	baltarusijos	5,3%	es	4,6%	rinkimų	2,9%	prezidento	2,1%
Skirstinys 8	**Dydis:** 282	**VPan:** 0,028	**IPan:** 0,010					
Sieja:	žuvo	4,6%	žmonių	2,8%	žmonės	2,3%	pranešė	2,2%
Išskiria:	žuvo	4,1%	paukščių	1,9%	žmonių	1,9%	žmonės	1,9%
Skirstinys 9	**Dydis:** 366	**VPan:** 0,020	**IPan:** 0,010					
Sieja:	lietuvos	2,2%	kino	1,4%	metų	1,1%	vilniaus	1,1%
Išskiria:	proc	2,0%	kino	1,6%	kultūros	1,1%	teatro	0,9%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 15

Suklasterizuota dokumentų: [2112 iš 2112], **Entropija:** 0,308, **Grynumas:** 0,825

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	29	0,246	0,094	0,007	0,003	0,339	0,828
1	193	0,134	0,039	0,011	0,003	0,000	1,000
2	68	0,133	0,050	0,011	0,003	0,048	0,985
3	74	0,120	0,031	0,013	0,003	0,105	0,959
4	69	0,102	0,029	0,011	0,003	0,094	0,971
5	113	0,066	0,029	0,012	0,003	0,448	0,814
6	143	0,064	0,027	0,013	0,003	0,394	0,825
7	126	0,056	0,014	0,010	0,003	0,029	0,992
8	129	0,056	0,023	0,011	0,003	0,000	1,000
9	191	0,043	0,012	0,012	0,003	0,185	0,937
10	189	0,038	0,013	0,012	0,003	0,317	0,852
11	154	0,033	0,009	0,009	0,003	0,263	0,896
12	164	0,035	0,009	0,011	0,003	0,238	0,921
13	245	0,031	0,007	0,010	0,002	0,677	0,555
14	225	0,024	0,006	0,011	0,003	0,766	0,409

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	1	0	4	24	0
1	0	193	0	0	0
2	1	0	0	0	67
3	0	0	0	71	3
4	0	1	0	1	67
5	9	4	3	92	5
6	12	0	3	118	10
7	1	125	0	0	0

8	0	129	0	0	0
9	6	0	4	2	179
10	5	0	1	22	161
11	138	1	0	6	9
12	3	2	3	151	5
13	24	0	136	16	69
14	86	2	92	26	19

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 29	**VPan:** 0,246	**IPan:** 0,007					
Sieja:	keliai	8,6%	kelių	8,5%	dangos	8,4%	vietomis	6,0%
Išskiria:	keliai	4,8%	dangos	4,7%	kelių	4,5%	vietomis	3,4%
Skirstinys 1	**Dydis:** 193	**VPan:** 0,134	**IPan:** 0,011					
Sieja:	taškų	6,4%	pelnė	4,0%	taškus	2,9%	14	2,6%
Išskiria:	taškų	4,0%	pelnė	2,2%	taškus	1,7%	proc	1,2%
Skirstinys 2	**Dydis:** 68	**VPan:** 0,133	**IPan:** 0,011					
Sieja:	seimo	47,7%	partijos	4,6%	frakcijos	2,1%	paulauskas	2,0%
Išskiria:	seimo	31,7%	partijos	2,6%	frakcijos	1,4%	paulauskas	1,3%
Skirstinys 3	**Dydis:** 74	**VPan:** 0,120	**IPan:** 0,013					
Sieja:	naftos	24,7%	dujų	17,7%	gazprom	5,6%	jav	2,9%
Išskiria:	naftos	17,5%	dujų	12,6%	gazprom	4,1%	kubinių	1,5%
Skirstinys 4	**Dydis:** 69	**VPan:** 0,102	**IPan:** 0,011					
Sieja:	izraelio	16,8%	palestiniečių	16,7%	hamas	12,9%	gazos	1,9%
Išskiria:	palestiniečių	11,5%	izraelio	11,4%	hamas	9,0%	gazos	1,3%
Skirstinys 5	**Dydis:** 113	**VPan:** 0,066	**IPan:** 0,012					
Sieja:	litų	25,6%	mln	13,4%	proc	2,9%	tūkst	2,1%
Išskiria:	litų	19,7%	mln	9,0%	pieno	1,0%	jav	0,9%
Skirstinys 6	**Dydis:** 143	**VPan:** 0,064	**IPan:** 0,013					
Sieja:	proc	23,5%	procento	6,2%	es	3,0%	euro	2,8%
Išskiria:	proc	15,7%	procento	5,1%	euro	2,4%	palyginti	1,6%
Skirstinys 7	**Dydis:** 126	**VPan:** 0,056	**IPan:** 0,010					
Sieja:	sek	6,5%	pasaulio	3,6%	tšk	2,6%	užėmė	2,5%
Išskiria:	sek	4,6%	pasaulio	1,9%	tšk	1,9%	užėmė	1,7%
Skirstinys 8	**Dydis:** 129	**VPan:** 0,056	**IPan:** 0,011					
Sieja:	futbolo	4,3%	rungtynes	3,1%	rungtynių	2,5%	klubo	2,0%
Išskiria:	futbolo	3,3%	rungtynes	2,2%	capitals	1,6%	rungtynių	1,6%
Skirstinys 9	**Dydis:** 191	**VPan:** 0,043	**IPan:** 0,012					
Sieja:	al	3,6%	irano	3,2%	jav	2,8%	jt	2,4%
Išskiria:	irano	2,7%	al	2,7%	jt	2,2%	irake	2,1%
Skirstinys 10	**Dydis:** 189	**VPan:** 0,038	**IPan:** 0,012					
Sieja:	baltarusijos	7,0%	rinkimų	4,2%	prezidento	3,4%	es	2,7%
Išskiria:	baltarusijos	6,7%	rinkimų	3,9%	prezidento	2,8%	ukrainos	1,3%
Skirstinys 11	**Dydis:** 154	**VPan:** 0,033	**IPan:** 0,009					
Sieja:	kino	4,8%	lietuvos	3,5%	kultūros	2,7%	teatro	2,6%
Išskiria:	kino	4,1%	teatro	2,2%	kultūros	2,2%	muzikos	1,9%
Skirstinys 12	**Dydis:** 164	**VPan:** 0,035	**IPan:** 0,011					
Sieja:	dolerių	5,7%	jav	5,6%	mlrd	4,4%	akcijų	2,4%
Išskiria:	dolerių	4,7%	mlrd	3,6%	jav	2,5%	akcijų	2,2%
Skirstinys 13	**Dydis:** 245	**VPan:** 0,031	**IPan:** 0,010					
Sieja:	žuvo	5,4%	žmonių	3,0%	žmonės	2,6%	paukščių	2,5%
Išskiria:	žuvo	4,7%	paukščių	2,2%	žmonės	2,0%	žmonių	2,0%
Skirstinys 14	**Dydis:** 225	**VPan:** 0,024	**IPan:** 0,011					
Sieja:	policijos	2,8%	vsat	1,5%	buvo	1,3%	pareigūnai	1,2%
Išskiria:	policijos	2,6%	vsat	1,8%	proc	1,6%	pareigūnai	0,9%

SKALDANTIS K-MEANS, 2112 DOKUMENTŲ, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 2112 **Skirstinių:** 20

Suklasterizuota dokumentų: [2112 iš 2112], **Entropija:** 0,271, **Grynumas:** 0,853

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	24	0,411	0,093	0,012	0,003	0,000	1,000
1	27	0,273	0,092	0,007	0,002	0,261	0,852
2	74	0,225	0,073	0,016	0,004	0,000	1,000
3	70	0,130	0,049	0,011	0,003	0,000	1,000
4	124	0,132	0,030	0,013	0,004	0,000	1,000
5	74	0,120	0,031	0,013	0,003	0,105	0,959
6	69	0,102	0,029	0,011	0,003	0,094	0,971
7	82	0,077	0,030	0,012	0,002	0,203	0,915
8	59	0,076	0,028	0,011	0,002	0,663	0,441
9	109	0,069	0,030	0,012	0,003	0,409	0,835
10	132	0,070	0,028	0,013	0,003	0,322	0,871
11	123	0,058	0,014	0,010	0,003	0,000	1,000
12	121	0,052	0,014	0,012	0,003	0,178	0,934
13	103	0,050	0,010	0,011	0,003	0,034	0,990
14	94	0,047	0,015	0,010	0,003	0,575	0,702
15	160	0,043	0,014	0,012	0,003	0,273	0,881
16	133	0,038	0,010	0,009	0,003	0,254	0,895
17	151	0,037	0,010	0,011	0,003	0,172	0,947
18	187	0,036	0,008	0,010	0,002	0,412	0,668
19	196	0,024	0,004	0,012	0,003	0,758	0,556

Pasiskirstymas klasėse

Skirstinys	Kultūra	Sportas	Teisėtvarka	Ūkis	Politika
0	0	24	0	0	0
1	0	0	4	23	0
2	0	74	0	0	0
3	0	0	0	0	70
4	0	124	0	0	0
5	0	0	0	71	3
6	0	1	0	1	67
7	1	0	0	6	75
8	26	0	14	19	0
9	8	4	2	91	4
10	7	0	3	115	7
11	0	123	0	0	0
12	3	0	5	0	113
13	1	102	0	0	0
14	7	0	66	13	8
15	3	0	1	15	141
16	119	0	0	6	8
17	1	2	2	143	3
18	1	0	125	0	61
19	109	3	24	26	34

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 24 **VPan:** 0,411 **IPan:** 0,012								
Sieja:	capitals	7,8%	ledo	5,2%	new	4,6%	rangers	4,2%
Išskiria:	capitals	4,7%	ledo	2,8%	rangers	2,5%	new	2,4%
Skirstinys 1 **Dydis:** 27 **VPan:** 0,273 **IPan:** 0,007								
Sieja:	keliai	8,9%	kelių	8,9%	dangos	8,7%	vietomis	6,2%
Išskiria:	keliai	4,9%	dangos	4,8%	kelių	4,7%	vietomis	3,4%

Skirstinys 2	**Dydis:** 74	**VPan:** 0,225	**IPan:** 0,016					
Sieja:	26	3,9%	taškų	3,5%	23	3,1%	22	2,9%
Išskiria:	26	1,9%	taškų	1,3%	23	1,2%	22	1,1%
Skirstinys 3	**Dydis:** 70	**VPan:** 0,130	**IPan:** 0,011					
Sieja:	seimo	48,1%	partijos	4,5%	paulauskas	1,9%	frakcijos	1,9%
Išskiria:	seimo	32,3%	partijos	2,5%	frakcijos	1,3%	paulauskas	1,2%
Skirstinys 4	**Dydis:** 124	**VPan:** 0,132	**IPan:** 0,013					
Sieja:	taškų	6,4%	pelnė	4,1%	min	2,8%	14	2,5%
Išskiria:	taškų	3,5%	pelnė	2,0%	dvitaškiai	1,3%	min	1,2%
Skirstinys 5	**Dydis:** 74	**VPan:** 0,120	**IPan:** 0,013					
Sieja:	naftos	24,7%	dujų	17,7%	gazprom	5,6%	jav	2,9%
Išskiria:	naftos	17,5%	dujų	12,6%	gazprom	4,1%	kubinių	1,5%
Skirstinys 6	**Dydis:** 69	**VPan:** 0,102	**IPan:** 0,011					
Sieja:	izraelio	16,8%	palestiniečių	16,7%	hamas	12,9%	gazos	1,9%
Išskiria:	palestiniečių	11,5%	izraelio	11,4%	hamas	9,0%	gazos	1,3%
Skirstinys 7	**Dydis:** 82	**VPan:** 0,077	**IPan:** 0,012					
Sieja:	irano	8,4%	jt	6,3%	iranas	5,2%	kosovo	3,3%
Išskiria:	irano	5,9%	jt	4,6%	iranas	3,9%	kosovo	2,5%
Skirstinys 8	**Dydis:** 59	**VPan:** 0,076	**IPan:** 0,011					
Sieja:	paukščių	19,9%	gripo	10,2%	sveikatos	3,1%	laipsnių	2,9%
Išskiria:	paukščių	15,1%	gripo	7,8%	laipsnių	2,2%	sveikatos	2,1%
Skirstinys 9	**Dydis:** 109	**VPan:** 0,069	**IPan:** 0,012					
Sieja:	litų	25,7%	mln	13,9%	proc	3,0%	tūkst	2,1%
Išskiria:	litų	19,6%	mln	9,4%	pieno	1,0%	jav	0,9%
Skirstinys 10	**Dydis:** 132	**VPan:** 0,070	**IPan:** 0,013					
Sieja:	proc	24,5%	procento	5,7%	es	3,3%	euro	3,1%
Išskiria:	proc	16,0%	procento	4,4%	euro	2,5%	palyginti	1,7%
Skirstinys 11	**Dydis:** 123	**VPan:** 0,058	**IPan:** 0,010					
Sieja:	sek	6,6%	pasaulio	3,5%	tšk	2,7%	užėmė	2,4%
Išskiria:	sek	4,6%	tšk	2,0%	pasaulio	1,8%	užėmė	1,6%
Skirstinys 12	**Dydis:** 121	**VPan:** 0,052	**IPan:** 0,012					
Sieja:	al	7,9%	irake	6,2%	jav	5,3%	irako	5,3%
Išskiria:	al	6,2%	irake	5,5%	irako	4,4%	dž	2,3%
Skirstinys 13	**Dydis:** 103	**VPan:** 0,050	**IPan:** 0,011					
Sieja:	futbolo	7,6%	komandos	3,4%	klubo	3,1%	rinktinės	2,5%
Išskiria:	futbolo	6,3%	komandos	2,6%	klubo	2,4%	treneris	2,0%
Skirstinys 14	**Dydis:** 94	**VPan:** 0,047	**IPan:** 0,010					
Sieja:	vsat	4,6%	policijos	4,3%	pareigūnai	2,5%	pasieniečiai	1,8%
Išskiria:	vsat	3,8%	policijos	2,7%	pareigūnai	1,5%	pasieniečiai	1,5%
Skirstinys 15	**Dydis:** 160	**VPan:** 0,043	**IPan:** 0,012					
Sieja:	baltarusijos	8,7%	rinkimų	4,0%	prezidento	3,7%	es	2,9%
Išskiria:	baltarusijos	8,1%	rinkimų	3,4%	prezidento	2,9%	lukašenka	1,4%
Skirstinys 16	**Dydis:** 133	**VPan:** 0,038	**IPan:** 0,009					
Sieja:	kino	5,3%	lietuvos	4,0%	teatro	3,1%	kultūros	3,0%
Išskiria:	kino	4,2%	teatro	2,5%	kultūros	2,2%	muzikos	1,9%
Skirstinys 17	**Dydis:** 151	**VPan:** 0,037	**IPan:** 0,011					
Sieja:	dolerių	5,5%	jav	5,3%	mlrd	4,7%	akcijų	2,7%
Išskiria:	dolerių	4,3%	mlrd	3,7%	akcijų	2,4%	jav	2,1%
Skirstinys 18	**Dydis:** 187	**VPan:** 0,036	**IPan:** 0,010					
Sieja:	žuvo	6,8%	žmonės	2,8%	žmonių	2,6%	sužeisti	2,1%
Išskiria:	žuvo	5,5%	žmonės	2,0%	sužeisti	1,8%	proc	1,5%
Skirstinys 19	**Dydis:** 196	**VPan:** 0,024	**IPan:** 0,012					
Sieja:	metų	1,3%	kad	1,2%	jis	1,1%	moteris	1,0%
Išskiria:	proc	1,4%	lietuvos	1,3%	moteris	1,2%	britų	1,0%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 5

Suklasterizuota dokumentų: [4224 iš 4224], **Entropija:** 0,395, **Grynumas:** 0,738

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	370	0,118	0,043	0,009	0,002	0,000	1,000
1	463	0,034	0,013	0,008	0,003	0,047	0,987
2	932	0,029	0,011	0,009	0,003	0,230	0,920
3	984	0,026	0,008	0,010	0,002	0,217	0,919
4	1475	0,014	0,004	0,009	0,003	0,825	0,358

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	0	0	0	370	0
1	0	1	5	457	0
2	30	857	32	3	10
3	904	59	9	2	10
4	306	157	528	9	475

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 370 **VPan:** 0,118 **IPan:** 0,009								
Sieja:	taškų	7,0%	pelnė	3,8%	taškus	3,1%	19	2,3%
Išskiria:	taškų	4,5%	pelnė	2,1%	taškus	1,8%	atkovojo	1,1%
Skirstinys 1 **Dydis:** 463 **VPan:** 0,034 **IPan:** 0,008								
Sieja:	sek	4,7%	min	3,1%	pasaulio	2,5%	vietą	1,9%
Išskiria:	sek	3,7%	min	1,6%	proc	1,5%	pasaulio	1,4%
Skirstinys 2 **Dydis:** 932 **VPan:** 0,029 **IPan:** 0,009								
Sieja:	proc	11,5%	mln	3,7%	litų	3,4%	naftos	2,3%
Išskiria:	proc	10,0%	mln	3,1%	litų	2,8%	naftos	2,0%
Skirstinys 3 **Dydis:** 984 **VPan:** 0,026 **IPan:** 0,010								
Sieja:	seimo	5,5%	kad	2,1%	baltarusijos	1,6%	rinkimų	1,5%
Išskiria:	seimo	5,6%	proc	1,7%	hamas	1,5%	rinkimų	1,5%
Skirstinys 4 **Dydis:** 1475 **VPan:** 0,014 **IPan:** 0,009								
Sieja:	žuvo	1,9%	buvo	1,7%	žmonių	1,3%	pranešė	1,1%
Išskiria:	proc	2,8%	žuvo	2,1%	seimo	1,3%	policijos	1,3%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 10

Suklasterizuota dokumentų: [4224 iš 4224], **Entropija:** 0,359, **Grynumas:** 0,786

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	52	0,267	0,091	0,006	0,002	0,255	0,885
1	367	0,119	0,043	0,009	0,002	0,000	1,000
2	148	0,113	0,051	0,011	0,003	0,114	0,966
3	410	0,053	0,021	0,011	0,003	0,259	0,907
4	329	0,047	0,015	0,011	0,003	0,079	0,979
5	334	0,044	0,020	0,011	0,003	0,152	0,952
6	450	0,035	0,013	0,008	0,003	0,020	0,996
7	599	0,028	0,010	0,011	0,002	0,403	0,766
8	609	0,025	0,008	0,009	0,002	0,601	0,534
9	926	0,015	0,004	0,009	0,003	0,740	0,559

Pasiskirstymas klasėse					
Skirstinys	**Politika**	**Ūkis**	**Kultūra**	**Sportas**	**Teisėtvarka**
0	0	46	1	0	5
1	0	0	0	367	0
2	143	2	2	0	1
3	14	372	14	1	9
4	322	4	1	1	1
5	9	318	4	2	1
6	0	1	1	448	0
7	459	125	6	2	7
8	235	21	27	1	325
9	58	185	518	19	146

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 52 **VPan:** 0,267 **IPan:** 0,006								
Sieja:	keliai	13,8%	vietomis	7,8%	kelių	6,4%	rajonuose	5,2%
Išskiria:	keliai	7,6%	vietomis	4,3%	kelių	3,3%	rajonuose	2,8%
Skirstinys 1 **Dydis:** 367 **VPan:** 0,119 **IPan:** 0,009								
Sieja:	taškų	7,0%	pelnė	3,8%	taškus	3,1%	19	2,3%
Išskiria:	taškų	4,5%	pelnė	2,1%	taškus	1,8%	atkovojo	1,1%
Skirstinys 2 **Dydis:** 148 **VPan:** 0,113 **IPan:** 0,011								
Sieja:	seimo	54,3%	partijos	2,0%	paulauskas	1,8%	darbo	1,4%
Išskiria:	seimo	37,4%	proc	1,2%	paulauskas	1,1%	partijos	1,0%
Skirstinys 3 **Dydis:** 410 **VPan:** 0,053 **IPan:** 0,011								
Sieja:	proc	21,1%	litų	8,6%	mln	4,9%	palyginti	2,0%
Išskiria:	proc	15,1%	litų	6,5%	mln	3,0%	palyginti	1,6%
Skirstinys 4 **Dydis:** 329 **VPan:** 0,047 **IPan:** 0,011								
Sieja:	hamas	6,9%	baltarusijos	6,7%	izraelio	6,1%	rinkimų	6,0%
Išskiria:	hamas	6,0%	baltarusijos	5,5%	izraelio	5,2%	rinkimų	4,9%
Skirstinys 5 **Dydis:** 334 **VPan:** 0,044 **IPan:** 0,011								
Sieja:	naftos	10,9%	dujų	10,2%	jav	8,3%	dolerių	4,1%
Išskiria:	naftos	9,6%	dujų	9,0%	jav	4,4%	dolerių	3,2%
Skirstinys 6 **Dydis:** 450 **VPan:** 0,035 **IPan:** 0,008								
Sieja:	sek	4,8%	min	3,2%	pasaulio	2,5%	vietą	1,9%
Išskiria:	sek	3,8%	min	1,6%	proc	1,5%	pasaulio	1,5%
Skirstinys 7 **Dydis:** 599 **VPan:** 0,028 **IPan:** 0,011								
Sieja:	es	5,1%	irano	3,1%	jt	2,0%	užsienio	1,8%
Išskiria:	es	4,0%	irano	3,1%	jt	2,1%	proc	1,7%
Skirstinys 8 **Dydis:** 609 **VPan:** 0,025 **IPan:** 0,009								
Sieja:	žuvo	5,7%	al	2,9%	žmonių	2,3%	žmonės	2,3%
Išskiria:	žuvo	5,3%	al	2,3%	proc	1,9%	žmonės	1,8%
Skirstinys 9 **Dydis:** 926 **VPan:** 0,015 **IPan:** 0,009								
Sieja:	paukščių	2,5%	lietuvos	1,5%	metų	1,2%	buvo	1,0%
Išskiria:	paukščių	3,3%	proc	2,1%	gripo	1,3%	jav	1,2%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 15

Suklasterizuota dokumentų: [4224 iš 4224], **Entropija:** 0,305, **Grynumas:** 0,836

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	51	0,277	0,088	0,006	0,002	0,258	0,882
1	111	0,260	0,073	0,015	0,004	0,000	1,000
2	144	0,115	0,052	0,011	0,003	0,045	0,986
3	256	0,112	0,035	0,012	0,003	0,016	0,996
4	148	0,092	0,027	0,010	0,003	0,025	0,993

5	242	0,068	0,026	0,012	0,003	0,365	0,847
6	190	0,061	0,033	0,010	0,003	0,000	1,000
7	169	0,054	0,021	0,010	0,002	0,649	0,491
8	236	0,049	0,019	0,011	0,003	0,144	0,958
9	272	0,045	0,014	0,008	0,002	0,030	0,993
10	338	0,044	0,020	0,011	0,003	0,164	0,947
11	348	0,038	0,017	0,011	0,003	0,353	0,859
12	482	0,032	0,012	0,011	0,002	0,241	0,902
13	602	0,025	0,008	0,009	0,002	0,620	0,560
14	635	0,016	0,004	0,009	0,003	0,555	0,737

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	0	45	1	0	5
1	0	0	0	111	0
2	142	2	0	0	0
3	0	1	0	255	0
4	147	1	0	0	0
5	15	205	15	0	7
6	0	0	0	190	0
7	0	57	28	1	83
8	226	3	3	1	3
9	0	1	1	270	0
10	11	320	4	2	1
11	11	299	20	1	17
12	435	36	5	0	6
13	205	30	29	1	337
14	48	74	468	9	36

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 51	**VPan:** 0,277	**IPan:** 0,006					
Sieja:	keliai	13,9%	vietomis	7,8%	kelių	6,4%	rajonuose	5,1%
Išskiria:	keliai	7,6%	vietomis	4,3%	kelių	3,3%	dangos	2,7%
Skirstinys 1	**Dydis:** 111	**VPan:** 0,260	**IPan:** 0,015					
Sieja:	taškų	6,2%	pelnė	3,4%	taškus	3,3%	22	2,1%
Išskiria:	taškų	2,8%	taškus	1,4%	pelnė	1,3%	kamuolių	1,0%
Skirstinys 2	**Dydis:** 144	**VPan:** 0,115	**IPan:** 0,011					
Sieja:	seimo	54,5%	partijos	2,0%	paulauskas	1,8%	frakcijos	1,4%
Išskiria:	seimo	37,2%	paulauskas	1,2%	proc	1,1%	frakcijos	1,0%
Skirstinys 3	**Dydis:** 256	**VPan:** 0,112	**IPan:** 0,012					
Sieja:	taškų	5,1%	pelnė	2,8%	16	2,6%	14	2,6%
Išskiria:	taškų	2,6%	pelnė	1,2%	proc	1,2%	14	1,1%
Skirstinys 4	**Dydis:** 148	**VPan:** 0,092	**IPan:** 0,010					
Sieja:	hamas	17,7%	izraelio	15,4%	palestiniečių	14,6%	fatah	1,5%
Išskiria:	hamas	12,6%	izraelio	10,8%	palestiniečių	10,4%	proc	1,1%
Skirstinys 5	**Dydis:** 242	**VPan:** 0,068	**IPan:** 0,012					
Sieja:	proc	29,3%	procento	3,6%	palyginti	3,2%	euro	2,8%
Išskiria:	proc	19,8%	procento	2,7%	palyginti	2,4%	euro	2,2%
Skirstinys 6	**Dydis:** 190	**VPan:** 0,061	**IPan:** 0,010					
Sieja:	futbolo	4,6%	rungtynių	4,1%	capitals	3,0%	min	2,7%
Išskiria:	futbolo	3,4%	rungtynių	2,7%	capitals	2,3%	ledo	1,6%
Skirstinys 7	**Dydis:** 169	**VPan:** 0,054	**IPan:** 0,010					
Sieja:	paukščių	22,3%	gripo	8,9%	vsat	2,1%	pareigūnai	2,1%
Išskiria:	paukščių	17,8%	gripo	7,1%	vsat	1,7%	h5n1	1,4%
Skirstinys 8	**Dydis:** 236	**VPan:** 0,049	**IPan:** 0,011					
Sieja:	baltarusijos	14,1%	rinkimų	7,8%	prezidento	4,2%	opozicijos	2,5%
Išskiria:	baltarusijos	12,3%	rinkimų	6,4%	prezidento	3,0%	opozicijos	2,2%
Skirstinys 9	**Dydis:** 272	**VPan:** 0,045	**IPan:** 0,008					

Sieja:	sek	7,5%	pasaulio	4,2%	tšk	2,6%	užėmė	2,4%
Išskiria:	sek	5,3%	pasaulio	2,3%	tšk	1,9%	užėmė	1,7%
Skirstinys 10	**Dydis:** 338	**VPan:** 0,044	**IPan:** 0,011					
Sieja:	naftos	10,8%	dujų	10,2%	jav	8,1%	dolerių	4,1%
Išskiria:	naftos	9,6%	dujų	9,1%	jav	4,2%	dolerių	3,3%
Skirstinys 11	**Dydis:** 348	**VPan:** 0,038	**IPan:** 0,011					
Sieja:	litų	17,9%	mln	8,0%	es	3,6%	proc	2,7%
Išskiria:	litų	16,6%	mln	6,1%	es	1,7%	jav	1,1%
Skirstinys 12	**Dydis:** 482	**VPan:** 0,032	**IPan:** 0,011					
Sieja:	irano	4,2%	jt	2,8%	jav	2,4%	reikalų	1,9%
Išskiria:	irano	4,1%	jt	2,8%	proc	1,9%	iranas	1,8%
Skirstinys 13	**Dydis:** 602	**VPan:** 0,025	**IPan:** 0,009					
Sieja:	žuvo	5,7%	žmonių	2,4%	žmonės	2,3%	policijos	2,3%
Išskiria:	žuvo	5,3%	policijos	1,9%	proc	1,8%	žmonės	1,8%
Skirstinys 14	**Dydis:** 635	**VPan:** 0,016	**IPan:** 0,009					
Sieja:	lietuvos	1,7%	kultūros	1,5%	muzikos	1,3%	kino	1,3%
Išskiria:	proc	1,9%	kultūros	1,8%	muzikos	1,7%	kino	1,6%

SKALDANTIS K-MEANS, 4224 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 4224 **Skirstinių:** 20

Suklasterizuota dokumentų: [4224 iš 4224], **Entropija:** 0,297, **Grynumas:** 0,824

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	44	0,405	0,075	0,011	0,002	0,000	1,000
1	45	0,338	0,055	0,005	0,001	0,000	1,000
2	110	0,263	0,072	0,015	0,004	0,000	1,000
3	90	0,140	0,049	0,013	0,002	0,224	0,900
4	138	0,121	0,053	0,011	0,003	0,074	0,978
5	239	0,121	0,034	0,012	0,003	0,017	0,996
6	143	0,095	0,027	0,010	0,003	0,026	0,993
7	141	0,093	0,033	0,012	0,002	0,090	0,972
8	222	0,073	0,028	0,012	0,003	0,342	0,860
9	154	0,060	0,022	0,010	0,002	0,614	0,513
10	195	0,057	0,021	0,011	0,003	0,140	0,959
11	226	0,056	0,027	0,011	0,003	0,318	0,876
12	259	0,047	0,014	0,008	0,002	0,031	0,992
13	250	0,046	0,023	0,011	0,003	0,183	0,940
14	182	0,043	0,010	0,010	0,003	0,000	1,000
15	273	0,043	0,014	0,011	0,003	0,122	0,960
16	226	0,032	0,009	0,008	0,002	0,233	0,920
17	315	0,031	0,010	0,011	0,003	0,558	0,514
18	415	0,029	0,008	0,009	0,002	0,451	0,740
19	557	0,016	0,003	0,010	0,002	0,772	0,503

Pasiskirstymas klasėse

Skirstinys	Politika	Ūkis	Kultūra	Sportas	Teisėtvarka
0	0	0	0	44	0
1	0	45	0	0	0
2	0	0	0	110	0
3	8	81	0	0	1
4	135	2	0	0	1
5	0	1	0	238	0
6	142	1	0	0	0
7	137	3	1	0	0
8	12	191	13	0	6

9	0	53	22	0	79
10	187	3	2	1	2
11	4	198	14	1	9
12	0	1	1	257	0
13	8	235	4	2	1
14	0	0	0	182	0
15	262	0	5	0	6
16	8	6	208	2	2
17	162	138	7	3	5
18	87	4	17	0	307
19	88	112	280	1	76

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 44	**VPan:** 0,405	**IPan:** 0,011					
Sieja:	capitals	8,3%	ledo	5,4%	rungtynių	4,0%	rangers	3,7%
Išskiria:	capitals	4,9%	ledo	2,9%	rangers	2,1%	rungtynių	1,9%
Skirstinys 1	**Dydis:** 45	**VPan:** 0,338	**IPan:** 0,005					
Sieja:	keliai	14,6%	vietomis	8,2%	kelių	6,6%	rajonuose	5,2%
Išskiria:	keliai	7,9%	vietomis	4,4%	kelių	3,3%	rajonuose	2,7%
Skirstinys 2	**Dydis:** 110	**VPan:** 0,263	**IPan:** 0,015					
Sieja:	taškų	6,2%	pelnė	3,3%	taškus	3,3%	22	2,1%
Išskiria:	taškų	2,8%	taškus	1,4%	pelnė	1,3%	kamuolių	1,0%
Skirstinys 3	**Dydis:** 90	**VPan:** 0,140	**IPan:** 0,013					
Sieja:	dujų	44,9%	gazprom	11,5%	kubinių	3,3%	rusijos	3,2%
Išskiria:	dujų	31,3%	gazprom	8,2%	kubinių	2,3%	metrų	1,3%
Skirstinys 4	**Dydis:** 138	**VPan:** 0,121	**IPan:** 0,011					
Sieja:	seimo	55,2%	partijos	2,0%	paulauskas	1,9%	darbo	1,5%
Išskiria:	seimo	37,2%	paulauskas	1,2%	proc	1,1%	frakcijos	1,0%
Skirstinys 5	**Dydis:** 239	**VPan:** 0,121	**IPan:** 0,012					
Sieja:	taškų	5,3%	pelnė	2,9%	14	2,7%	16	2,6%
Išskiria:	taškų	2,7%	pelnė	1,3%	proc	1,2%	14	1,1%
Skirstinys 6	**Dydis:** 143	**VPan:** 0,095	**IPan:** 0,010					
Sieja:	hamas	17,7%	izraelio	15,6%	palestiniečių	14,5%	fatah	1,5%
Išskiria:	hamas	12,5%	izraelio	10,9%	palestiniečių	10,2%	fatah	1,1%
Skirstinys 7	**Dydis:** 141	**VPan:** 0,093	**IPan:** 0,012					
Sieja:	irano	15,4%	jt	8,1%	iranas	6,5%	saugumo	2,9%
Išskiria:	irano	11,1%	jt	5,8%	iranas	4,8%	tatena	1,9%
Skirstinys 8	**Dydis:** 222	**VPan:** 0,073	**IPan:** 0,012					
Sieja:	proc	29,7%	procento	3,8%	palyginti	3,6%	euro	2,5%
Išskiria:	proc	19,5%	procento	2,8%	palyginti	2,7%	euro	1,8%
Skirstinys 9	**Dydis:** 154	**VPan:** 0,060	**IPan:** 0,010					
Sieja:	paukščių	23,5%	gripo	9,5%	pareigūnai	2,2%	vsat	2,1%
Išskiria:	paukščių	18,2%	gripo	7,4%	vsat	1,6%	h5n1	1,5%
Skirstinys 10	**Dydis:** 195	**VPan:** 0,057	**IPan:** 0,011					
Sieja:	baltarusijos	16,5%	rinkimų	7,7%	prezidento	3,8%	opozicijos	2,8%
Išskiria:	baltarusijos	13,7%	rinkimų	5,9%	prezidento	2,5%	opozicijos	2,3%
Skirstinys 11	**Dydis:** 226	**VPan:** 0,056	**IPan:** 0,011					
Sieja:	litų	28,5%	mln	10,4%	proc	3,6%	pieno	1,5%
Išskiria:	litų	22,8%	mln	6,8%	pieno	1,2%	jav	0,9%
Skirstinys 12	**Dydis:** 259	**VPan:** 0,047	**IPan:** 0,008					
Sieja:	sek	7,9%	pasaulio	4,1%	užėmė	2,4%	tšk	2,3%
Išskiria:	sek	5,5%	pasaulio	2,2%	tšk	1,7%	užėmė	1,6%
Skirstinys 13	**Dydis:** 250	**VPan:** 0,046	**IPan:** 0,011					
Sieja:	naftos	16,3%	jav	12,1%	dolerių	5,0%	dolerio	2,9%
Išskiria:	naftos	14,0%	jav	6,8%	dolerių	3,8%	dolerio	2,5%
Skirstinys 14	**Dydis:** 182	**VPan:** 0,043	**IPan:** 0,010					
Sieja:	futbolo	7,8%	klubo	3,0%	komandos	2,4%	fc	2,3%
Išskiria:	futbolo	6,5%	klubo	2,4%	fc	1,8%	komandos	1,8%

Skirstinys 15	**Dydis:** 273	**VPan:** 0,043	**IPan:** 0,011					
Sieja:	al	10,0%	irako	9,4%	irake	7,1%	jav	3,2%
Išskiria:	irako	8,3%	al	8,3%	irake	6,3%	karių	2,0%
Skirstinys 16	**Dydis:** 226	**VPan:** 0,032	**IPan:** 0,008					
Sieja:	kultūros	5,3%	lietuvos	4,5%	muzikos	4,4%	kino	3,0%
Išskiria:	kultūros	4,3%	muzikos	3,6%	kino	2,4%	teatro	2,3%
Skirstinys 17	**Dydis:** 315	**VPan:** 0,031	**IPan:** 0,011					
Sieja:	es	15,6%	europos	5,9%	užsienio	1,6%	reikalų	1,5%
Išskiria:	es	14,4%	europos	4,3%	proc	1,1%	reikalų	0,9%
Skirstinys 18	**Dydis:** 415	**VPan:** 0,029	**IPan:** 0,009					
Sieja:	žuvo	5,9%	žmonių	3,2%	žmonės	3,2%	policijos	2,2%
Išskiria:	žuvo	4,8%	žmonės	2,4%	žmonių	2,0%	proc	1,7%
Skirstinys 19	**Dydis:** 557	**VPan:** 0,016	**IPan:** 0,010					
Sieja:	kad	1,3%	buvo	1,3%	savo	1,1%	metų	1,1%
Išskiria:	proc	1,7%	lietuvos	1,2%	seimo	1,2%	es	1,2%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 5 SKIRSTINIAI

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 5

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,417, **Grynumas:** 0,723

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	1544	0,049	0,026	0,006	0,002	0,010	0,998
1	722	0,039	0,017	0,010	0,003	0,155	0,952
2	1252	0,033	0,015	0,009	0,002	0,332	0,872
3	1870	0,025	0,008	0,009	0,002	0,200	0,933
4	3060	0,012	0,003	0,008	0,002	0,853	0,342

Pasiskirstymas klasėse

Skirstinys	**Sportas**	**Politika**	**Ūkis**	**Kultūra**	**Teisėtvarka**
0	1541	0	2	1	0
1	6	20	687	3	6
2	12	42	1092	79	27
3	18	1745	73	17	17
4	71	691	280	1045	973

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 1544	**VPan:** 0,049	**IPan:** 0,006					
Sieja:	taškų	3,9%	min	2,9%	pelnė	2,8%	taškus	2,6%
Išskiria:	taškų	2,6%	min	1,8%	pelnė	1,8%	taškus	1,7%
Skirstinys 1	**Dydis:** 722	**VPan:** 0,039	**IPan:** 0,010					
Sieja:	naftos	11,9%	dujų	11,2%	jav	6,1%	dolerių	3,1%
Išskiria:	naftos	10,2%	dujų	9,6%	jav	2,8%	gazprom	2,5%
Skirstinys 2	**Dydis:** 1252	**VPan:** 0,033	**IPan:** 0,009					
Sieja:	proc	16,7%	litų	6,2%	mln	3,9%	es	2,8%
Išskiria:	proc	13,7%	litų	5,4%	mln	2,7%	es	1,4%
Skirstinys 3	**Dydis:** 1870	**VPan:** 0,025	**IPan:** 0,009					
Sieja:	seimo	6,1%	kad	1,9%	rinkimų	1,8%	baltarusijos	1,7%
Išskiria:	seimo	6,2%	rinkimų	1,7%	hamas	1,7%	proc	1,7%
Skirstinys 4	**Dydis:** 3060	**VPan:** 0,012	**IPan:** 0,008					
Sieja:	buvo	1,7%	žuvo	1,6%	žmonių	1,3%	pranešė	1,0%
Išskiria:	proc	3,1%	žuvo	1,9%	seimo	1,4%	taškų	1,3%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 10 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 10

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,358, **Grynumas:** 0,783

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	100	0,251	0,091	0,005	0,002	0,249	0,880
1	736	0,108	0,042	0,009	0,002	0,000	1,000
2	314	0,100	0,042	0,010	0,003	0,071	0,981
3	326	0,080	0,025	0,010	0,002	0,077	0,979
4	816	0,049	0,020	0,010	0,002	0,253	0,911
5	676	0,041	0,019	0,010	0,003	0,139	0,959
6	852	0,034	0,014	0,008	0,002	0,025	0,994
7	1441	0,024	0,008	0,010	0,002	0,358	0,827
8	1336	0,021	0,006	0,008	0,002	0,580	0,522
9	1851	0,013	0,003	0,008	0,002	0,725	0,560

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	0	0	88	1	11
1	736	0	0	0	0
2	0	308	2	1	3
3	2	319	4	0	1
4	7	35	743	23	8
5	5	14	648	3	6
6	847	0	2	3	0
7	7	1191	204	19	20
8	2	550	29	58	697
9	42	81	414	1037	277

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 100	**VPan:** 0,251	**IPan:** 0,005					
Sieja:	keliai	14,6%	vietomis	7,9%	kelių	7,2%	dangos	5,6%
Išskiria:	keliai	8,1%	vietomis	4,4%	kelių	3,7%	dangos	3,1%
Skirstinys 1	**Dydis:** 736	**VPan:** 0,108	**IPan:** 0,009					
Sieja:	taškų	7,0%	pelnė	3,8%	taškus	3,4%	18	2,2%
Išskiria:	taškų	4,5%	pelnė	2,1%	taškus	1,9%	atkovojo	1,1%
Skirstinys 2	**Dydis:** 314	**VPan:** 0,100	**IPan:** 0,010					
Sieja:	seimo	51,2%	partijos	2,8%	frakcijos	1,7%	paulauskas	1,4%
Išskiria:	seimo	35,8%	partijos	1,5%	proc	1,2%	frakcijos	1,1%
Skirstinys 3	**Dydis:** 326	**VPan:** 0,080	**IPan:** 0,010					
Sieja:	palestiniečių	16,4%	hamas	16,3%	izraelio	15,6%	gazos	1,4%
Išskiria:	palestiniečių	11,8%	hamas	11,8%	izraelio	11,2%	proc	1,1%
Skirstinys 4	**Dydis:** 816	**VPan:** 0,049	**IPan:** 0,010					
Sieja:	proc	23,1%	litų	7,8%	mln	4,8%	procento	2,2%
Išskiria:	proc	17,1%	litų	5,8%	mln	3,0%	procento	1,7%
Skirstinys 5	**Dydis:** 676	**VPan:** 0,041	**IPan:** 0,010					
Sieja:	naftos	12,9%	dujų	12,0%	jav	7,0%	dolerių	3,7%
Išskiria:	naftos	11,2%	dujų	10,6%	jav	3,5%	dolerių	2,9%
Skirstinys 6	**Dydis:** 852	**VPan:** 0,034	**IPan:** 0,008					
Sieja:	sek	5,3%	min	3,4%	pasaulio	2,3%	vietą	2,0%
Išskiria:	sek	4,1%	min	1,7%	proc	1,5%	futbolo	1,4%
Skirstinys 7	**Dydis:** 1441	**VPan:** 0,024	**IPan:** 0,010					
Sieja:	es	3,3%	baltarusijos	2,6%	irano	2,4%	prezidento	1,8%
Išskiria:	baltarusijos	2,7%	irano	2,5%	es	2,3%	prezidento	1,7%
Skirstinys 8	**Dydis:** 1336	**VPan:** 0,021	**IPan:** 0,008					
Sieja:	žuvo	4,9%	žmonių	2,4%	al	2,3%	pranešė	1,8%

Išskiria:	žuvo	4,8%	proc	1,9%	al	1,8%	žmonių	1,6%
Skirstinys 9 **Dydis:** 1851 **VPan:** 0,013 **IPan:** 0,008								
Sieja:	lietuvos	1,9%	paukščių	1,9%	metų	1,2%	buvo	0,9%
Išskiria:	paukščių	2,6%	proc	2,0%	kino	1,2%	seimo	1,1%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 15 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 15

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,301, **Grynumas:** 0,831

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	104	0,325	0,107	0,010	0,002	0,081	0,971
1	98	0,259	0,090	0,005	0,002	0,239	0,888
2	215	0,245	0,074	0,014	0,004	0,000	1,000
3	305	0,102	0,043	0,010	0,003	0,072	0,980
4	583	0,089	0,033	0,010	0,003	0,008	0,998
5	322	0,080	0,025	0,010	0,002	0,069	0,981
6	473	0,047	0,024	0,010	0,003	0,348	0,869
7	680	0,047	0,021	0,011	0,003	0,237	0,909
8	534	0,046	0,020	0,010	0,002	0,150	0,951
9	332	0,045	0,016	0,009	0,002	0,656	0,524
10	667	0,041	0,019	0,010	0,003	0,129	0,961
11	707	0,032	0,011	0,008	0,002	0,029	0,993
12	792	0,029	0,011	0,010	0,002	0,281	0,888
13	1355	0,020	0,006	0,008	0,002	0,593	0,526
14	1281	0,013	0,004	0,008	0,002	0,548	0,745

Pasiskirstymas klasėse

Skirstinys	**Sportas**	**Politika**	**Ūkis**	**Kultūra**	**Teisėtvarka**
0	101	0	0	3	0
1	0	0	87	1	10
2	215	0	0	0	0
3	0	299	2	1	3
4	582	0	1	0	0
5	2	316	3	0	1
6	10	9	411	24	19
7	0	34	618	25	3
8	0	508	14	3	9
9	2	1	94	61	174
10	2	15	641	3	6
11	702	0	3	2	0
12	3	703	62	14	10
13	2	542	44	54	713
14	27	71	154	954	75

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0 **Dydis:** 104 **VPan:** 0,325 **IPan:** 0,010								
Sieja:	ledo	7,1%	capitals	5,3%	new	4,5%	rangers	4,3%
Išskiria:	ledo	4,0%	capitals	3,2%	rangers	2,5%	new	2,3%
Skirstinys 1 **Dydis:** 98 **VPan:** 0,259 **IPan:** 0,005								
Sieja:	keliai	14,8%	vietomis	8,0%	kelių	7,2%	dangos	5,4%
Išskiria:	keliai	8,1%	vietomis	4,4%	kelių	3,7%	dangos	3,0%
Skirstinys 2 **Dydis:** 215 **VPan:** 0,245 **IPan:** 0,014								
Sieja:	taškų	6,1%	taškus	3,9%	pelnė	3,5%	atkovojo	2,0%
Išskiria:	taškų	2,7%	taškus	1,7%	pelnė	1,4%	nba	1,1%

Skirstinys 3	**Dydis:** 305	**VPan:** 0,102	**IPan:** 0,010					
Sieja:	seimo	51,7%	partijos	2,7%	frakcijos	1,7%	paulauskas	1,4%
Išskiria:	seimo	35,9%	partijos	1,4%	frakcijos	1,2%	proc	1,2%
Skirstinys 4	**Dydis:** 583	**VPan:** 0,089	**IPan:** 0,010					
Sieja:	taškų	4,9%	pelnė	2,6%	16	2,4%	14	2,3%
Išskiria:	taškų	2,6%	proc	1,2%	pelnė	1,1%	krepšinio	1,1%
Skirstinys 5	**Dydis:** 322	**VPan:** 0,080	**IPan:** 0,010					
Sieja:	hamas	16,5%	palestiniečių	16,5%	izraelio	15,5%	gazos	1,4%
Išskiria:	hamas	12,0%	palestiniečių	11,9%	izraelio	11,0%	proc	1,1%
Skirstinys 6	**Dydis:** 473	**VPan:** 0,047	**IPan:** 0,010					
Sieja:	litų	28,6%	mln	11,7%	proc	2,5%	tūkst	1,8%
Išskiria:	litų	23,4%	mln	8,1%	pieno	1,2%	jav	0,9%
Skirstinys 7	**Dydis:** 680	**VPan:** 0,047	**IPan:** 0,011					
Sieja:	proc	24,3%	es	4,2%	euro	2,9%	procento	2,8%
Išskiria:	proc	18,2%	euro	2,6%	procento	2,4%	es	2,2%
Skirstinys 8	**Dydis:** 534	**VPan:** 0,046	**IPan:** 0,010					
Sieja:	irano	9,8%	jt	4,9%	iranas	3,6%	jav	2,7%
Išskiria:	irano	8,3%	jt	4,2%	iranas	3,1%	tatena	1,7%
Skirstinys 9	**Dydis:** 332	**VPan:** 0,045	**IPan:** 0,009					
Sieja:	paukščių	17,1%	gripo	8,0%	vsat	3,4%	pareigūnai	2,4%
Išskiria:	paukščių	14,0%	gripo	6,6%	vsat	2,8%	pareigūnai	1,3%
Skirstinys 10	**Dydis:** 667	**VPan:** 0,041	**IPan:** 0,010					
Sieja:	naftos	13,0%	dujų	12,3%	jav	6,8%	dolerių	3,7%
Išskiria:	naftos	11,3%	dujų	10,8%	jav	3,4%	gazprom	2,9%
Skirstinys 11	**Dydis:** 707	**VPan:** 0,032	**IPan:** 0,008					
Sieja:	sek	6,0%	pasaulio	3,5%	futbolo	2,7%	min	2,7%
Išskiria:	sek	4,5%	pasaulio	2,2%	futbolo	2,0%	proc	1,5%
Skirstinys 12	**Dydis:** 792	**VPan:** 0,029	**IPan:** 0,010					
Sieja:	baltarusijos	7,5%	rinkimų	5,0%	prezidento	3,4%	es	2,2%
Išskiria:	baltarusijos	8,0%	rinkimų	4,9%	prezidento	3,0%	opozicijos	1,3%
Skirstinys 13	**Dydis:** 1355	**VPan:** 0,020	**IPan:** 0,008					
Sieja:	žuvo	4,8%	žmonių	2,4%	al	2,2%	policijos	1,9%
Išskiria:	žuvo	4,8%	proc	1,9%	al	1,7%	žmonių	1,7%
Skirstinys 14	**Dydis:** 1281	**VPan:** 0,013	**IPan:** 0,008					
Sieja:	lietuvos	1,8%	kino	1,7%	metų	1,4%	kultūros	1,3%
Išskiria:	kino	2,3%	proc	1,9%	kultūros	1,6%	muzikos	1,4%

SKALDANTIS K-MEANS, 8448 DOKUMENTAI, 20 SKIRSTINIŲ

Klasterizavimo parametrai

Algoritmas: Skald, K-means **Dokumentų:** 8448 **Skirstinių:** 20

Suklasterizuota dokumentų: [8448 iš 8448], **Entropija:** 0,291, **Grynumas:** 0,833

Skirstinys	Dydis	VPan	VNuok	IPan	INuok	Entropija	Grynumas
0	102	0,335	0,103	0,010	0,002	0,060	0,980
1	90	0,297	0,076	0,005	0,002	0,091	0,967
2	213	0,248	0,074	0,014	0,004	0,000	1,000
3	195	0,120	0,047	0,012	0,002	0,209	0,913
4	488	0,108	0,034	0,011	0,003	0,009	0,998
5	301	0,102	0,043	0,010	0,003	0,062	0,983
6	316	0,082	0,025	0,010	0,002	0,050	0,987
7	411	0,073	0,031	0,011	0,002	0,246	0,903
8	347	0,069	0,029	0,010	0,002	0,137	0,957
9	311	0,062	0,021	0,008	0,002	0,013	0,997
10	412	0,053	0,026	0,010	0,003	0,319	0,881
11	419	0,050	0,019	0,011	0,003	0,110	0,969
12	311	0,048	0,017	0,009	0,002	0,640	0,537

13	467	0,044	0,023	0,010	0,003	0,121	0,966
14	541	0,035	0,012	0,010	0,002	0,179	0,937
15	513	0,032	0,008	0,009	0,003	0,018	0,996
16	546	0,033	0,012	0,011	0,003	0,517	0,597
17	483	0,026	0,008	0,007	0,002	0,210	0,928
18	817	0,026	0,007	0,008	0,002	0,426	0,758
19	1165	0,013	0,003	0,009	0,002	0,807	0,475

Pasiskirstymas klasėse

Skirstinys	Sportas	Politika	Ūkis	Kultūra	Teisėtvarka
0	100	0	0	2	0
1	0	0	87	0	3
2	213	0	0	0	0
3	0	14	178	0	3
4	487	0	1	0	0
5	0	296	1	1	3
6	1	312	2	0	1
7	0	18	371	21	1
8	0	332	8	2	5
9	310	0	0	1	0
10	5	8	363	21	15
11	1	406	5	4	3
12	1	1	89	53	167
13	3	6	451	2	5
14	0	507	3	9	22
15	511	0	1	1	0
16	2	200	326	9	9
17	3	12	18	448	2
18	1	168	11	18	619
19	10	218	219	553	165

Skirstinio viduje siejančios ir nuo kitų skirstinių išskiriančios savybės

Skirstinys 0	**Dydis:** 102	**VPan:** 0,335	**IPan:** 0,010					
Sieja:	ledo	6,8%	capitals	5,3%	new	4,5%	rangers	4,3%
Išskiria:	ledo	3,8%	capitals	3,2%	rangers	2,5%	new	2,3%
Skirstinys 1	**Dydis:** 90	**VPan:** 0,297	**IPan:** 0,005					
Sieja:	keliai	15,1%	vietomis	8,3%	kelių	7,2%	dangos	5,4%
Išskiria:	keliai	8,3%	vietomis	4,5%	kelių	3,6%	dangos	3,0%
Skirstinys 2	**Dydis:** 213	**VPan:** 0,248	**IPan:** 0,014					
Sieja:	taškų	6,1%	taškus	4,0%	pelnė	3,5%	atkovojo	1,9%
Išskiria:	taškų	2,7%	taškus	1,8%	pelnė	1,4%	nba	1,0%
Skirstinys 3	**Dydis:** 195	**VPan:** 0,120	**IPan:** 0,012					
Sieja:	dujų	45,4%	gazprom	12,3%	rusijos	4,0%	kubinių	3,2%
Išskiria:	dujų	32,2%	gazprom	8,9%	kubinių	2,3%	rusijos	1,4%
Skirstinys 4	**Dydis:** 488	**VPan:** 0,108	**IPan:** 0,011					
Sieja:	taškų	5,4%	pelnė	2,8%	14	2,7%	16	2,6%
Išskiria:	taškų	2,8%	pelnė	1,2%	proc	1,2%	14	1,1%
Skirstinys 5	**Dydis:** 301	**VPan:** 0,102	**IPan:** 0,010					
Sieja:	seimo	51,4%	partijos	2,9%	frakcijos	1,7%	paulauskas	1,5%
Išskiria:	seimo	35,5%	partijos	1,5%	frakcijos	1,2%	proc	1,2%
Skirstinys 6	**Dydis:** 316	**VPan:** 0,082	**IPan:** 0,010					
Sieja:	hamas	16,4%	palestiniečių	16,4%	izraelio	15,4%	gazos	1,4%
Išskiria:	hamas	11,8%	palestiniečių	11,8%	izraelio	10,9%	proc	1,1%
Skirstinys 7	**Dydis:** 411	**VPan:** 0,073	**IPan:** 0,011					
Sieja:	proc	34,0%	procento	4,6%	palyginti	3,2%	padidėjo	2,1%
Išskiria:	proc	22,3%	procento	3,4%	palyginti	2,4%	padidėjo	1,4%
Skirstinys 8	**Dydis:** 347	**VPan:** 0,069	**IPan:** 0,010					
Sieja:	irano	14,6%	jt	7,1%	iranas	5,2%	tatena	3,0%

Išskiria:	irano	10,9%	jt	5,3%	iranas	4,0%	tatena	2,4%
Skirstinys 9 **Dydis:** 311 **VPan:** 0,062 **IPan:** 0,008								
Sieja:	sek	15,7%	min	3,2%	žaidynių	2,9%	žiemos	2,8%
Išskiria:	sek	10,6%	žaidynių	2,0%	žiemos	1,9%	olimpinių	1,5%
Skirstinys 10 **Dydis:** 412 **VPan:** 0,053 **IPan:** 0,010								
Sieja:	litų	31,0%	mln	12,1%	proc	2,1%	pieno	1,7%
Išskiria:	litų	24,5%	mln	8,0%	pieno	1,3%	jav	0,9%
Skirstinys 11 **Dydis:** 419 **VPan:** 0,050 **IPan:** 0,011								
Sieja:	baltarusijos	15,4%	rinkimų	9,3%	prezidento	3,9%	opozicijos	2,5%
Išskiria:	baltarusijos	13,1%	rinkimų	7,5%	prezidento	2,6%	lukašenka	2,0%
Skirstinys 12 **Dydis:** 311 **VPan:** 0,048 **IPan:** 0,009								
Sieja:	paukščių	18,2%	gripo	8,5%	vsat	3,5%	pareigūnai	2,5%
Išskiria:	paukščių	14,6%	gripo	6,8%	vsat	2,9%	h5n1	1,4%
Skirstinys 13 **Dydis:** 467 **VPan:** 0,044 **IPan:** 0,010								
Sieja:	naftos	21,3%	jav	10,4%	dolerių	4,4%	mlrd	3,0%
Išskiria:	naftos	17,9%	jav	5,5%	dolerių	3,1%	dolerio	2,3%
Skirstinys 14 **Dydis:** 541 **VPan:** 0,035 **IPan:** 0,010								
Sieja:	al	8,8%	irako	8,7%	irake	7,0%	jav	3,6%
Išskiria:	irako	8,0%	al	7,7%	irake	6,7%	karių	2,1%
Skirstinys 15 **Dydis:** 513 **VPan:** 0,032 **IPan:** 0,009								
Sieja:	futbolo	6,5%	pasaulio	3,5%	lietuvos	1,8%	čempionato	1,8%
Išskiria:	futbolo	5,6%	pasaulio	2,1%	proc	1,5%	klubo	1,2%
Skirstinys 16 **Dydis:** 546 **VPan:** 0,033 **IPan:** 0,011								
Sieja:	es	16,4%	europos	7,6%	lietuvos	1,6%	euro	1,5%
Išskiria:	es	14,8%	europos	5,7%	euro	1,2%	ecb	0,9%
Skirstinys 17 **Dydis:** 483 **VPan:** 0,026 **IPan:** 0,007								
Sieja:	kino	5,5%	lietuvos	4,5%	kultūros	4,0%	muzikos	3,3%
Išskiria:	kino	4,6%	kultūros	3,3%	muzikos	2,8%	teatro	2,3%
Skirstinys 18 **Dydis:** 817 **VPan:** 0,026 **IPan:** 0,008								
Sieja:	žuvo	6,7%	žmonių	3,5%	žmonės	3,0%	policijos	2,1%
Išskiria:	žuvo	5,6%	žmonės	2,3%	žmonių	2,2%	sužeisti	1,7%
Skirstinys 19 **Dydis:** 1165 **VPan:** 0,013 **IPan:** 0,009								
Sieja:	kad	1,6%	metų	1,3%	buvo	1,2%	jis	1,1%
Išskiria:	proc	1,8%	seimo	1,4%	es	1,3%	taškų	1,2%

PRIEDAS E – KONFERENCIJOS PRANEŠIMAS

BERNOTAS, M.; ŽALINAUSKAS, M. Dokumentų valdymo modelis žiniasklaidos redakcijos sistemoje. *Informacinės technologijos'2006: konferencijos pranešimų medžiaga. KTU*, 2006, p. 437–440.

DOKUMENTŲ VALDYMO MODELIS ŽINIASKLAIDOS REDAKCIJOS SISTEMOJE

Marijus Bernotas[1], Marius Žalinauskas[2]

[1]Šiaulių universitetas, Informacinių technologijų katedra, Vilniaus g. 141, Šiauliai

[2]UAB „Informacijos alėja", Dvaro g. 55, Šiauliai

Šiuolaikinėse organizacijose dokumentai atlieka labai svarbų vaidmenį, o jų nuolatinis daugėjimas stimuliuoja dokumentų valdymo poreikį. Informacinių technologijų taikymas dokumentams valdyti leidžia efektyviau organizuoti ir plėtoti veiklą, išlaikyti ir stiprinti įvairių organizacijų pozicijas rinkoje. Ypač tai jaučia organizacijos, dalyvaujančios žinių ekonomikoje. Straipsnyje apžvelgiami dokumentų valdymo modeliai, kurių savybės panaudojami žiniasklaidai skirtoje redakcinėje sistemoje „News Processor" bei analizuojami jos taikymo ypatumai.

1 Įvadas

Daugumos šiuolaikinių organizacijų veiklos pagrindas yra darbas su elektroniniais dokumentais. Organizacijoms plečiantis, kartu didėja ir dokumentų skaičius, todėl darosi sunkiau pasiekti, išrinkti ir panaudoti reikiamą informaciją. Su tais pačiais dokumentais dirbant keliems žmonėms, atsiranda keletas dokumento kopijų ir prarandama vis daugiau laiko reikalingos informacijos paieškai. Dėl organizacijos dokumentų valdymo strategijos ir darbo procesų politikos nebuvimo žmogiškieji ištekliai panaudojami neoptimaliai.

Ypač tai jaučia žiniasklaidos darbuotojai, kurie dėl minėtų faktorių praranda nemažai laiko neesminiams uždaviniams spręsti, nukenčia leidybos procesas ir leidinio kokybė, o norint leidinius perkelti į virtualią interneto erdvę tenka procesus iš dalies pakartoti ir neretai visą perkėlimo darbą atlikti rankiniu būdu. Minėtoms problemoms spręsti kuriamos specializuotos dokumentų valdymo sistemos (DVS), kurios įvardijamos kaip redakcinės sistemos, skirtos žiniasklaidai. Straipsnyje analizuojami dokumentų valdymo ypatumai šiose sistemose, dokumentų versijų kūrimo būdai ir darbų srauto valdymas.

2 Dokumentų valdymo sistemų modeliai

Analizuojant daugelį taikomųjų dokumentų valdymo sistemų, galima išskirti du pagrindinius modelius – integruotas dokumentų valdymo sistemas ir dokumento modeliu grįstas valdymo sistemas. Paanalizuosime kiekvienos jų taikymo ypatumus.

Šiuo metu labiausiai paplitusios integruotų dokumentų valdymo sistemos, kuriose su kiekvienu dokumentu elgiamasi kaip su „juodąja dėže" [7]. Pats paprasčiausias tokios sistemos pavyzdys – bet kuri tipinė laikmenų sistema, kuri sudaroma remiantis įprastu hierarchiniu modeliu. Šios sistemos turi principinių trūkumų [2]:

- Vienas dokumentas – viena vieta. Nesvarbu, kad dažno dokumento turinys turėtų priklausyti kelioms šakoms, hierarchinėje sistemoje jį įmanoma laikyti tik vienoje vietoje ir tai labai trukdo sisteminti dokumentus.
- Ribota paieška ir komplikuotas dalinimasis. Dažniausiai dokumentai į sistemą įkeliami naudojantis vienais sisteminimo kriterijais, o ieškomi naudojantis kitais. Deja, taikant hierarchinį modelį, dokumentų paiešką galima atlikti tik vienu pjūviu. Negana to, dokumentų sisteminimas pagal kriterijus, savaime suprantamus vienam žmogui, gali visiškai neatitikti kito žmogaus supratimo.

Norint pašalinti prieš tai minėtus trūkumus, siūloma tobulinti vartotojo sąsają, pavyzdžiui, išnaudoti erdvę [3], automatizuoti komponentus arba naudoti lankstesnes informacines struktūras – metaduomenis. Kaip parodė tyrimai [1], trimatės erdvės taikymas didesnio efektyvumo dokumentų valdymui nesuteikia. Daugelis autorių [3, 4, 5, 6] metaduomenis laiko vienintele aprašymo priemone dokumentams, kurie neturi struktūros arba kurių struktūra yra nežinoma.

Dokumentų valdymo sistemos, kuriose dokumentai laikomi „juodosiomis dėžėmis", turi vieną trūkumą – jose esančius dokumentus ne taip paprasta panaudoti naujiems dokumentams kurti. Šio trūkumo neturi dokumento modeliu pagrįstos DVS, kuriose naujas dokumentas gali būti pusiau automatiškai sugeneruotas iš jau turimų dokumentų [7]. Visos priemonės dokumentams valdyti imamos iš integruotų DVS modeliu pagrįstose sistemose yra papildomų žinių apie dokumento struktūrą ir semantinę elementų prasmę. Tiesa, reikia pabrėžti, kad tokių sistemų potencialas atsiskleidžia tik naudojantis dokumentais, kurių struktūra atskirta nuo turinio,

pavyzdžiui, SGML (angl. Standard Generalized Markup Language) arba vartojant kurį nors kitą griežtą jos dialektą.

Dokumento modeliu pagrįstos DVS dažnai skirstomos į [8]:

- besiremiančias hipertekstinių dokumentų modeliu;
- paieška besiremiančias sistemas.

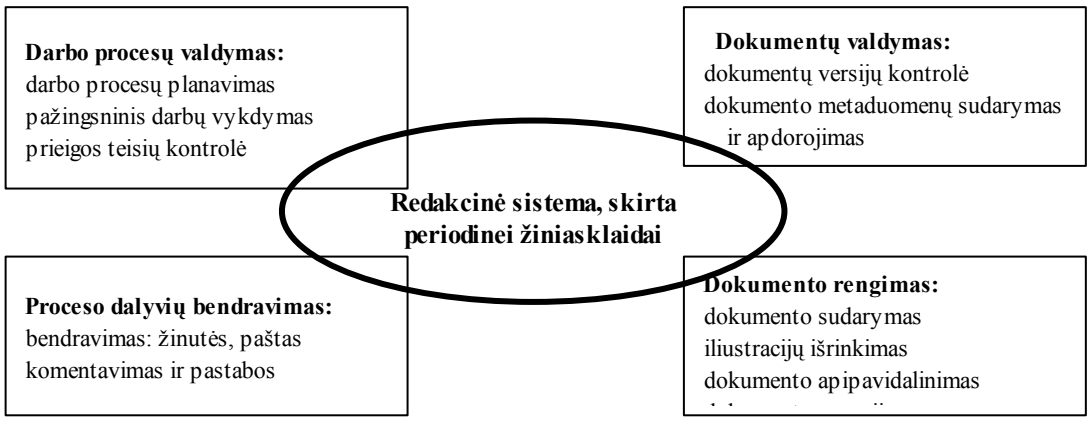
Hipertekstinių dokumentų modeliu besiremiančiose sistemose loginiai ryšiai tarp dokumentų ir dokumentų kolekcijų sukuriami naudojant nuorodas. Taip atsiranda vienalytė sistema, kurioje galima naršyti intuityviai. Deja, tokia sistema turi trūkumų: keliaujant nuorodomis, didelės apimties sistemoje labai lengva pasimesti, sunku sukurti apsaugą nuo nekorektiškų ryšių.

Didelėse, nepažįstamose, heterogeninėse dokumentų kolekcijose hipertekstinio modelio paprastai neužtenka. Tokiose sistemose lengva patekti į aklavietes – dokumentus, kuriuose nėra reikšmingų nuorodų. Paprastai tokioms situacijoms išspręsti pasitelkiama aukšto lygio paieška, kuri padeda arba surasti reikalingus dokumentus, arba patekti į dokumentus su reikiamomis nuorodomis. Įvairių paieškos metodų tyrimais užsiima daug komercinių ir akademinių įstaigų. Pačius metodus galima suskirstyti į [8]:

- Sintaksinės paieškos metodus. Tai paprasčiausi iš paieškos metodų. Paieška atliekama įrašant žodį arba frazę. Šie metodai taikomi daugelyje žiniatinklio paieškos sistemų.
- Paieškos metodus metaduomenyse. Metaduomenyse paprastai ieškoma binarinių arba uždaro formato dokumentų.
- Semantinės paieškos metodus. Sudėtingiausi iš metodų. Paieškai naudojami sinonimai ir kitos semantinės koncepcijos. Semantinių paieškų rezultatas – dažnai randami dokumentai, neturintys paieškos frazės, bet atitinkantys paieškos lūkesčius pagal turinį.

3 Redakcinės sistemos dokumentų valdymo modelis ir jo taikymo ypatumai

Kuriant šiuolaikinių dokumentų valdymą, skirtą periodinei žiniasklaidai, būtina atkreipti dėmesį į redakcinei sistemai keliamus kelis bendrus reikalavimus, kurie būdingi grupinio darbo sistemai, darbo procesų valdymo sistemai, turinio valdymo sistemai, ir specifinius, kurie būdingi leidybos procesui: darbų srautas, kelių asmenų sinchroninis dokumento redagavimas, dokumentų versijų kūrimas(žr. 1 pav.).



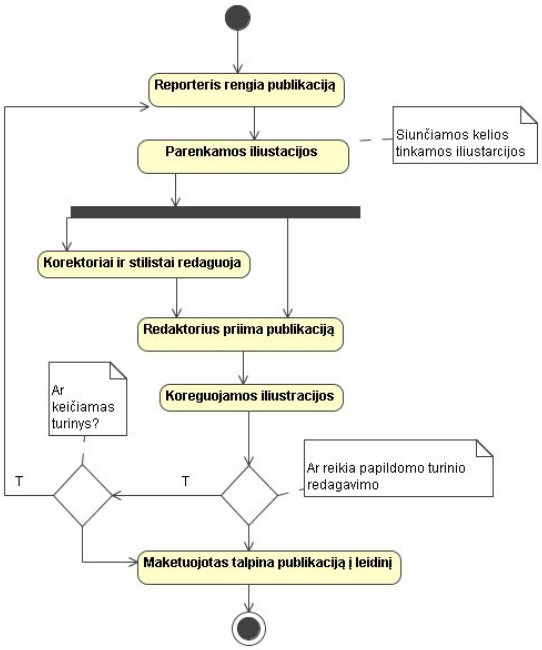
1 pav. Redakcinės sistemos, skirtos periodinei žiniasklaidai, struktūra

Šių reikalavimų aktualumas yra pakankamai akivaizdus:

- Kitaip nei dauguma informacinių technologijų produktų, kurie yra orientuojami į vieno individo poreikius, dokumentų valdymo sistemos yra išimtis – jos skirtos žmonių grupėms. Beje, toks darbo pobūdis susijęs su papildomais sunkumais, pavyzdžiui, dirbant su tais pačiais dokumentais, būtina užtikrinti vientisumą.
- Nors nuorodomis pagrįstų sistemų galimybės ir kokybė labai priklauso nuo asmens, sudariusio ryšius, naršymas jose pasižymi paprastumu ir intuityvumu. Daugeliu atvejų darbas su tokiomis sistemomis būna efektyvesnis nei su sistemomis, pagrįstomis vartotojų apibrėžta paieška, o žymėjimo kalbos tapo įprasta priemone vienareikšmiam dokumentų aprašymui.
- Žiniatinklio technologijos yra senos ir išbandytos. Nauji standartai tik didina jų patrauklumą.

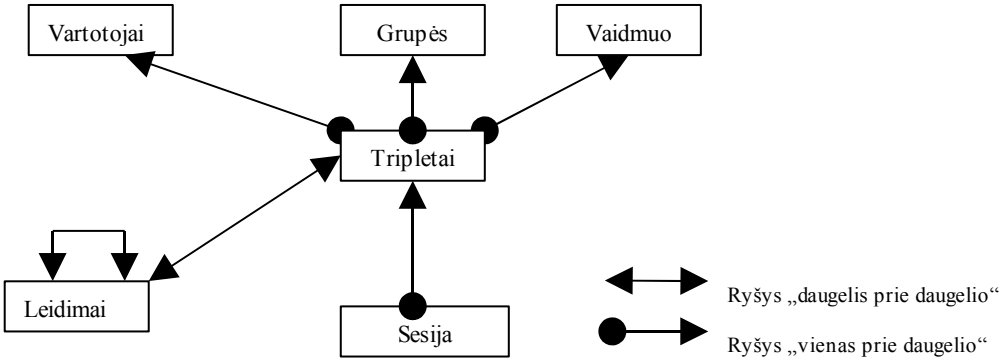
Minėti reikalavimai įgyvendinti žiniasklaidai skirtoje redakcinėje sistemoje „News Processor“. Jau keletą metų šia sistema naudojasi laikraščių redakcijos, todėl pastebėti jos taikymo ypatumai, kuriuos lemia leidybos proceso specifika.

Darbų srauto sudarymas – tai dokumento kelio tarp proceso vykdytojų ir atliekamų leidybos veiksmų visuma. Kiekvienas, rengdamas dokumentą, atlieka savo užduotį. Kiekvienai žiniasklaidos redakcijai būdingas savitas darbų srautas. Sudaryti vientisą darbų srautą praktiškai neįmanoma, nes jis nuolat kinta, nors proceso dalyvių hierarchija grindžiama griežtu darbų paskirstymu, pavyzdžiui, redaktorius publikaciją redaguoti gali siųsti arba reporteriui, arba korektoriui-stilistui.



2 pav. Supaprastintas darbų srautas

Periodinių leidinių redakcijose dokumentą rengia kelios žmonių grupės – reporteriai, korektoriai-stilistai, redaktoriai, fotografai, dizaineriai ir maketuotojai, kurie redakcinėje sistemoje gali dirbti pagal suteikiamas visam trejetui (tripletui) „vartotojas–grupė–vaidmuo“ teises. Vartotojas su tuo pačiu vaidmeniu skirtingose grupėse ar leidimų grupėje gali turėti skirtingas teises.

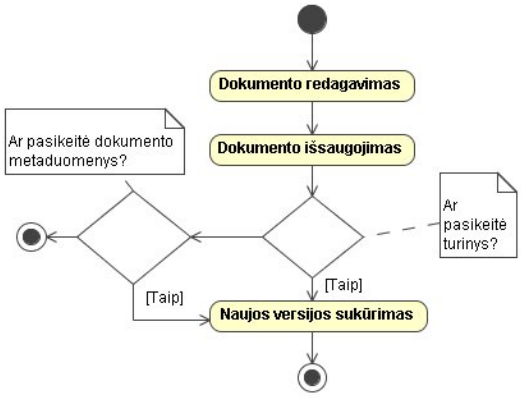


3 pav. Prieigos teisių kontrolė

Be darbų srauto valdymo ir kontrolės negalima išsiversti bet kokiame procese.

Išsaugoma kiekvieno suredaguoto dokumento ir modifikuota, ir senoji versija. Taigi, sistemoje saugomi visi dokumento pakeitimai ir galima pažiūrėti, kaip kito dokumento versijos bėgant laikui. Dokumentų versijos gali būti kuriamos automatiškai arba rankiniu būdu. Automatinis dokumento versijos sukūrimas patogesnis vartotojo požiūriu, nes automatiškai sukuriama nauja dokumento versija, jei keičiamas turinys. Tačiau tai sudėtingai realizuojama tam tikrose situacijose, pavyzdžiui, jei vienu metu dokumentą redaguoja keli žmonės, tai būtų išsaugomi tik vėliausiai darbą baigusio darbuotojo pakeitimai. To galima išvengti dviem būdais: naują dokumento versiją sukurti rankiniu būdu (pasirinkus komandas *check-in, check-out*) arba naują versiją sukurti,

kai pakeičiamas arba dokumento turinys ir/arba jo metaduomenys (žr. 4 pav.): iliustracijų skaičius, autorius, tipas, publikacijos būsena ir pan.



4 pav. Dokumento versijos sukūrimas

4 Išvados

Šiandien jau yra pasiektas toks lygis, kad kiekviena lyderio pozicijos siekianti organizacija gali sutaupyti nemažai laiko ir išteklių pasirinkdama tinkamas informacines sistemas. Ne išimtis ir periodinių leidinių redakcijos, kurios kasdien dirba su dideliais elektroninių dokumentų kiekiais. Neefektyvus dokumentų valdymas lemia dubliavimąsi, sudėtingą paiešką ir nuo vartotojo priklausančią dokumento versijų sistemą. Šiai problemai spręsti buvo pasirinkta integruoto ir dokumento modeliu grįsto DVS modelių sintezė redakcinei sistemai realizuoti. Praktinis jos taikymas žiniasklaidos procesuose parodė, kad reikia tobulinti dokumentų versijų kūrimo sistemą, kuri turėtų automatinio ir rankinio versijų kūrimo metodų ypatybių, bei prie jos būtų taikoma metaduomenų sistema formuojant naują dokumento versiją. Taip pat pastebėta, kad redakcijose darbų srautas kinta ir priklauso tik nuo vartotojo sprendimo tuo metu, todėl reikalinga lankstesnė darbų planavimo sistema.

Literatūros sąrašas

[1] **Cockburn A., McKenzie B.** 3D or not 3D? Evaluating the effect of the third dimension in a document management system. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, 434–441.

[2] **Dourish P., Edwards W. K., LaMarca A., Lamping J., Petersen K., Salisbury M., Terry D. B. and Thornton J.** Extending Document Management Systems with User-Specific Active Properties. *ACM Transactions on Information Systems*, 2000, 18/2, 140–170.

[3] **Foo J.** DocPlayer: Design Insights from Applying the Non-Hierarchical Media-Player model to Document Management, 2003. [Žiūrėta 2004 11 14] Prieiga internete:<http://www.ep.liu.se/exjobb/ida/2003/017/exjobb.pdf>.

[4] **Jones J.** An Alternative Document Management Model: Query-based Access to Hierarchical File systems, 2002. [Žiūrėta 2005 12 14] Prieiga internete: <http://innovexpo.itee.uq.edu.au/2002/projects/s354151/thesis.pdf>.

[5] **Lyytikäinen V.** Contextual and Structural Metadata in Document Management, 2004. [Žiūrėta 2005 11 14] Prieiga internete:<http://selene.lib.jyu.fi:8080/vaitos/studies/studcomp/9513917835.pdf>.

[6] **Manola F.** Towards a Web Object Model, 1998. [Žiūrėta 2005 12 14] Prieiga internete: <http://www.objs.com/OSA/wom.htm>.

[7] **Rezgui Y., Debras P.** An Integrated Approach for a Model Based Document Production and Management. *Electronic Journal of Information Technology in Construction,* 1996, 1, 1–21.

[8] **Small D. J.** A Model-Driven Architecture for Enterprise Document [Management, Supporting Discovery and Reuse, 1999. [Žiūrėta 2005 10 18] Prieiga internete: <http://www.comp.leeds.ac.uk/research/pubs/theses/small.ps.gz>.

A model for document management in the editorial system for periodical process

The documents play a very significant role in the modern organizations, while their constant increase stimulates the demand for their management. The appliance of modern technologies for the document management allows to organize and develop the activity more effectively, maintain and strengthen the positions of various organizations in the market. This is particularly relevant to the organizations participating in the knowledge economy. The article reviews the models for the document management, where their constituent elements are applied in the editorial system "News Processor" and the aspects of their appliance are being analyzed.

PRIEDAS F – KONFERENCIJOS PRANEŠIMAS

ŽALINAUSKAS, M. Dokumentų klasterizavimas. *Informacinės technologijos 2006: konferencijos pranešimų medžiaga. VUKHF*, 2006, p. 207–211.

DOKUMENTŲ KLASTERIZAVIMAS

Marius Žalinauskas
Kauno technologijos universitetas

Šiuolaikinėse organizacijose dokumentai atlieka labai svarbų vaidmenį. Jie yra nuolat kaupiami, todėl didėjant dokumentų kiekiui įprastos saugyklos, besiremiančios elementarios paieškos arba hierarchijos principais, vis labiau ir labiau lėtina reikalingos informacijos suradimą. Pranešime nagrinėjamas vienas iš galimų šios problemos sprendimo kelių – dokumentų klasterizavimas (ang. clustering), apžvelgiami dokumentų klasterizavimo metodai; į klasterizavimą pažvelgiama alternatyvių problemos sprendimo būdų fone.

1 Įžanga

Tobulėjant programinei ir techninei įrangai didėja organizacijų ir pavienių asmenų galimybės kuriant, renkant, skleidžiant ir saugant elektroninius dokumentus. Per pastaruosius dešimtmečius smarkiai nukritus elektroninių laikmenų kainoms mums tapo įprasta kaupti viską iš eilės, todėl nenuostabu, kad prireikus surasti reikalingos informacijos daugeliui tenka susidurti su visa jūra nežinomos vertės tekstinių ir netekstinių duomenų.

Tradiciškai šias problemas sprendusios priemonės, tokios kaip rankiniu būdu žmonių kataloguotos dokumentų bibliotekos ir panašiais būdais organizuotos failų hierarchijos, nebegali pakankamai greitai ir pigiai susidoroti su jas užgriuvusiu duomenų kiekiu. Tenka tobulinti paieškos metodus, kurti algoritmus ir įrankius automatiniam dokumentų sistematizavimui atlikti.

2 Paieška

Dažniausiai šiuo metu norint surasti reikiamus dokumentus yra naudojami paieškos varikliai. Juose naudojamų metodų tyrimais užsiima daug komercinių organizacijų ir akademinių įstaigų. Pačius metodus galima suskirstyti į [9]:

- Sintaksinės paieškos metodus. Paprasčiausi iš paieškos metodų. Paieška atliekama žodžio arba frazės lygyje. Šiuos metodus naudoja dauguma žiniatinklio paieškos sistemų.
- Paieškos metaduomenyse metodus. Paieška metaduomenyse paprastai naudojama atliekant binarinių arba uždaro formato dokumentų paiešką.
- Semantinės paieškos metodus. Sudėtingiausi iš metodų. Paieškoje naudojami sinonimai ir kitos semantinės koncepcijos. Semantinių paieškų rezultatuose dažnai atsiduria dokumentai, neturintys paieškos frazės, bet atitinkantys paieškos lūkesčius savo turiniu.

Dažnas sudėtingesnis paieškos variklis leidžia atlikti paiešką ankstesnės paieškos rezultatuose t.y. leidžia atlikti paieškos užklausos tobulinimą[44].

Deja, kad ir kokie spartūs, sudėtingi ir intelektualūs bebūtų paieškos algoritmai, visi jie kenčia nuo tų pačių esminių problemų [5]:

- Tipinis vartotojas paprastai nėra pažįstamas su dokumentų saugykla. Keletą kartų gavęs tuščią rezultatų sąrašą, jis galiausiai iki minimumo sumažina paieškos užklausą ir vėliau gaudamas per ilgus rezultatų sąrašus pradeda vėl po truputį ją pildyti.
- Vartotojas retai nuo pat pradžių tiksliai žino ko ieško. Labai dažnai pradėjęs paiešką vienais raktažodžiais ir patyrinėjęs gautus rezultatus jis iš principo pakeičia nuomonę ko būtent reikėtų ieškoti.

Kitaip tariant, dauguma šiuolaikinių paieškos metodų vartotojui suteikia labai mažai pasirinkimo. Tenka rinktis arba per ilgą rezultatų sąrašą, arba įsivelti į ilgą ir netrivialų paieškos užklausos tobulinimą [1], [3].

3 Klasterizavimas ir klasifikavimas

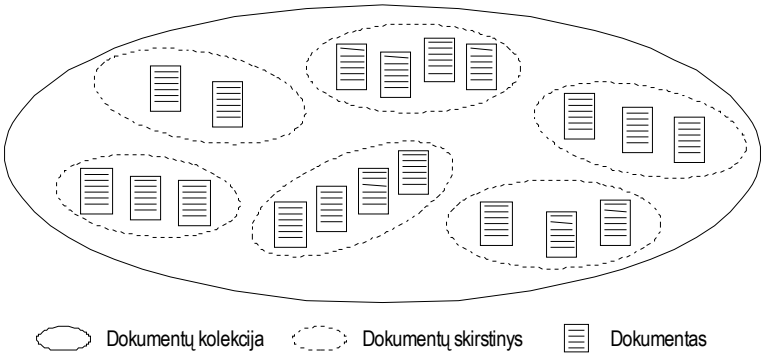
Dokumentų klasterizavimas siūlo alternatyvų būdą greitam reikalingos informacijos suradimui. Kitaip nei atliekant paieškos užklausas, klasterizuojant dokumentai į bendras grupes – skirstinius (*ang. cluster*) – patenka ne pagal ieškomų žodžių ar frazių buvimą/nebuvimą juose, bet pagal dokumentų turinio panašumą.

Pirmieji siūlymai panaudoti klasterizavimą informacijos gavybai pagerinti buvo pateikti 1971 metais Jardine ir Van Rijsbergen darbuose [6]. Buvo tikimasi, kad panaudojus klasterizavimą informacijos gavybos

44 Informacijos gavyboje (*ang. information retrieval, IR*) šis procesas įvardijamas terminu q*uery refinement*.

sistemų (*ang. information retrieval systems*) efektyvumas padidės, kadangi organizuojant dokumentus į skirstinius pagal jų turinio panašumą jie daugiau ar mažiau atitiks intuityvų vartotojo elgesį ieškant.

Efektyvumo padidėjimo lūkesčiai buvo grindžiami skirstinių hipoteze (*ang. cluster hypothesis*) [6]. Ši hipotezė teigia, kad tiesiogiai susiję dokumentai turi daugiau tarpusavio panašumų vienas su kitu nei su nesusijusiais dokumentais, todėl jie yra linkę atsidurti tuose pačiuose skirstiniuose. Kitaip tariant, hipotezė daro prielaidą, kad dokumentų kolekcijose tarpusavyje susijusius dokumentus visada įmanoma suskirstyti į atskiras grupes pagal jų turinio panašumus.



1 pav. Skirstiniai dokumentų kolekcijoje.

Dokumentų klasterizavimas yra labai dažnai painiojamas su dokumentų klasifikavimu. Reikia pastebėti, kad nors klasterizavimas ir klasifikavimas (dar vadinamas kategorizavimu) yra teksto gavybos (*ang. text mining*) veiklos turinčios daug panašių savybių, jos taip pat turi ir keletą esminių skirtumų.

Klasifikuojant dokumentus turima reikalų su iš anksto apibrėžta kategorijų (dažnai vadinamų klasėmis) aibe. Proceso metu priklausomai nuo turinio dokumentas yra priskiriamas kuriai nors iš klasių. Klasterizuojant dokumentus jokios iš anksto apibrėžtos kategorijų aibės nėra. Dokumentai suburiami į grupes tik pagal savo turinio panašumą. Kitaip tariant, klasterizuoti dokumentai atskleidžia natūralius turinyje užsislėpusius rinkinius, o klasifikuoti dokumentai yra „prievarta“ įspraudžiami į numatytus organizacinius rėmus [11].

Žvelgiant iš techninės pusės, dokumentų klasifikacija yra apmokymo su mokytoju (*ang. supervised learning*) problema. Tam kad būtų įmanoma atlikti automatinį klasifikavimą, prieš tai visoms kategorijoms būtina rankiniu būdu priskirti dokumentų rinkinius pagal kuriuos turi apsimokyti klasifikatorius–automatas. Parinktų pavyzdinių dokumentų tinkamumas labai stipriai nulemia vėlesnį klasifikatoriaus efektyvumą. Dokumentų klasterizavimas yra apmokymo be mokytojo (*ang. unsupervised learning*) problema. Klasterizuojant dokumentai sugrupuojami be jokių išankstinių apmokymų. Tiesiog stengiamasi, kad turiniu panašūs dokumentai atsidurtų tuose pačiuose, o turiniu nepanašūs – skirtinguose skirstiniuose. [2]

4 Klasterizavimo metodai

Klasterizavimo metodai gali būti suskirstyti į dvi bazines grupes: dalinančius (*ang. partitioning*) ir hierarchinius. Nepriklausomai nuo pasirenkamo metodo klasterizuojant yra atliekami šie žingsniai [10]:

- Dokumento reprezentacijos parinkimas. Klasterizuojant dokumentus būtina pasirinkti atributus, kurie taikant algoritmą reprezentuos kiekvieną dokumentą. Dažniausiai dokumentas yra reprezentuojamas daugiamatės erdvės vektoriumi $X = \{x_1, x_2, ..., x_n\}$, kur n yra klasterizuojamos dokumentų kolekcijos žodyne esantis žodžių skaičius.
- Asociacijos mato parinkimas. Šis matas apibrėžia kiek panašūs ar nepanašūs yra tarpusavyje du dokumentai. Mato parinkimas labai įtakoja galutinį klasterizavimo rezultatą.
- Klasterizavimo metodo parinkimas. Be abejonės, norint atlikti dokumentų kolekcijos klasterizavimą būtina pasirinkti konkretų algoritmą, atliekantį kolekcijos struktūrizavimą.
- Skirstinių reprezentavimo pasirinkimas. Dažniausiai klasterizavimo rezultatas yra daugiamačiai skirstiniai. Norint pasinaudoti atliktu struktūrizavimu reikia pasirinkti būdus daugiamačių rezultatų projekcijoms į dvimatę arba trimatę erdvę atlikti.
- Rezultatų patikrinimas. Klasterizavimo algoritmui pateikus rezultatus juos būtina patikrinti. Paprastai tai atliekama panaudojant atitinkamus testus.

Žemiau yra trumpai aprašomi baziniai klasterizavimo metodai.

4.1 Dalinantis klasterizavimas

Visi dalinančio klasterizavimo metodai pagrįsti vienu bendru principu. Pradedant klasterizavimą, *n*

dokumentų kolekcija $D = \{d_1, d_2, ..., d_n\}$, kur d_i – daugiamačiu vektoriumi aprašytas dokumentas, yra sudalinama į k pradinių skirstinių S_i. Nė vienas skirstinys negali būti tuščias, nė vienas jų negali turėti bendrų dokumentų, o pradiniai dokumentai jiems gali būti priskiriami tiesiog atsitiktine tvarka. Klasterizavimo eigoje minimizuojant nuo konkretaus metodo priklausančią kriterijaus funkciją dokumentai iteratyviai perkeliami iš vieno skirstinio į kitą. Klasterizavimas baigiamas, kai iteracijos metu nebeatliekamas nė vienas dokumento perkėlimas. Matematiniai apribojimai nurodyti (1) lygtyse.

$$D = \bigcup_{i=1}^{k} S_i$$
$$S_i \neq \emptyset \,, i = 1,2,\ldots,k \tag{1}$$
$$S_i \cap S_j = \emptyset \,, i \neq j \;\; i,j = 1,2,\ldots,k$$

Dalinančio klasterizavimo metodai yra labai patrauklūs, nes juos įgyvendinantys algoritmai turi neaukštus reikalavimus skaičiuojamajai technikai. Kolekcijai iš n dokumentų klasterizavimo kaina paprastai būna intervale nuo $O(n)$ iki $O(n \cdot log\, n)$ [11]. Dalinančio klasterizavimo metodai yra gerokai efektyvesni už hierarchinio klasterizavimo metodus.

Pagrindinis dalinančio klasterizavimo metodų trūkumas yra tai, kad prieš atliekant klasterizavimą tenka priimti prielaidas apie tikėtiną skirstinių formą. Negana to, taip pat iš anksto reikia nurodyti ir pageidaujamą skirstinių kiekį k. Paprastai tai sukelia itin didelių sunkumų, nes toli gražu ne kiekvienoje srityje kiekį k galima numatyti iš vis. Ypač tai sunku padaryti, kai dokumentų skaičius labai didelis. Vykdoma nemažai tyrimų skirstinių kiekį bandant parinkti atsižvelgiant į klasterizuojamus duomenis (pavyzdžiui, [7], [8], [4]), tačiau iš esmės ši problema kol kas laikoma neišspręsta.

Vienas dažniausiai naudojamų dalinančio klasterizavimo algoritmų yra K-vidurkių (*ang. K-means*) algoritmas. Gavęs pradinį k skirstinių rinkinį su atsitiktinai jiems priskirtais duomenų taškais (dokumentų klasterizavimo atveju tai – dokumentai) algoritmas iteratyviai mažindamas kvadratų sumų funkciją (2) perkelia dokumentus iš vieno skirstinio S_i į kitą. (2) funkcijoje x_n yra vektorius, apibūdinantis n-ąjį duomenų tašką.

$$J = \sum_{i}^{k} \sum_{n \in S_i} |x_n - \mu_i|^2 \tag{2}$$

Algoritmas iteratyviai vykdomas dviem žingsniais. Pirmo žingsnio metu apskaičiuojamas kiekvieno skirstinio centroidas μ_i. Antro žingsnio metu kiekvienas duomenų taškas yra priskiriamas tam skirstiniui, kuriame esantis centroidas yra arčiausiai duomenų taško. Žingsniai kartojami tol, kol nenusistovi kriterijaus funkcijos (2) reikšmė J.

K-vidurkių algoritmas garantuoja, kad jis turės baigtinį iteracijų skaičių, jeigu yra parenkamas baigtinis pradinių skirstinių skaičius, o kiekvienos paskesnės iteracijos metu kvadratų suma J nėra padidinama. Deja, algoritmas negarantuoja, kad jis baigsis pasiekus minimalų galimą J. K-vidurkių algoritmo rezultatai labai priklauso nuo pirminių skirstinių sudarymo, todėl praktikoje dažniausiai išbandoma keletas pradinių atsitiktinai sudarytų skirstinių ir parenkamas geriausias rezultatas.

K-vidurkiai turi ir daugiau trūkumų. Vieną didžiausių iš jų lemia pati algoritmo prigimtis. Kadangi algoritmas yra pagrįstas euklidiniais atstumais tarp taškų, netiesiogiai yra daroma prielaida apie kompaktišką ir sferinę skirstinių prigimtį. Deja pakankamai dažnai skirstiniai būna nei kompaktiški, nei sferiniai, todėl toli nuo pagrindinių taškų sankaupų atsiskyrę taškai gali labai stipriai iškraipyti centroidų parinkimo skaičiavimus. Vienintelis būdas išvengti šių iškraipymų yra arba tinkamas pradinių skirstinių parinkimas, arba algoritmo taikymas gausioms aibėms [11].

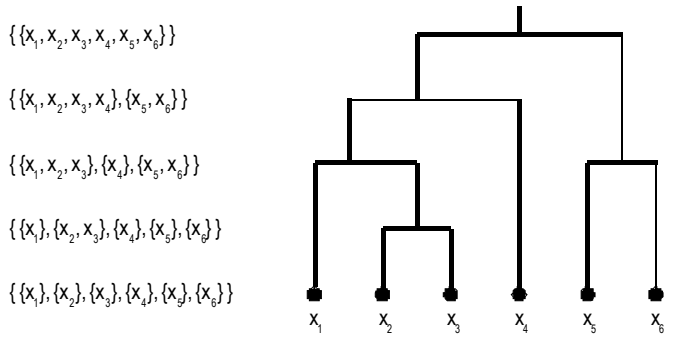
4.2 Hierarchinis klasterizavimas

Kitaip nei dalinančio klasterizavimo metodai, hierarchinis klasterizavimas dokumentus suburia ne į skirstinių rinkinį, o į skirstinių hierarchiją. Labai panašūs objektai (mūsų atveju tai – dokumentai) yra suburiami į mažesnius skirstinius, o šie medžio principu suburiami į didesnius skirstinius, kuriuose esančių objektų panašumas yra mažesnis.

Tarkime, kad klasterizuojama n dokumentų kolekcija yra apibrėžta kaip $D = \{d_1, d_2, ..., d_n\}$, kur d_i yra dokumentas, aprašytas daugiamačiu vektoriumi. Kaip ir dalinančio klasterizavimo atveju k skirstinių rinkinys gali būti apibrėžtas kaip $R = \{S_1, S_2, ..., S_k\}$, jeigu yra tenkinamos lygtyse (1) aprašytos sąlygos.

Skirstinių rinkinys R_1, turintis m skirstinių, yra laikomas įterptu į skirstinių rinkinį R_2, turintį $r < m$ skirstinių, jeigu kiekvienas skirstinys rinkinyje R_1 yra kurio nors R_2 rinkinio skirstinio poaibis ir bent vienas R_1 rinkinio skirstinys yra griežtas kurio nors R_2 rinkinio skirstinio poaibis. Pavyzdžiui, rinkinys $R_1 = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$ yra įterptas į skirstinių rinkinį $R_1 = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$ ir nėra įterptas į rinkinį $R_3 = \{\{x_1, x_4\}, \{x_3\}, \{x_2,$

$x_5\}\}$ (pavyzdžiai paimti iš [10]). Paveikslėlyje nr. 2 yra pateikta vaizdi hierarchinio klasterizavimo iliustracija.



2 pav. Hierarchinio klasterizavimo iliustracija.

Hierarchiniai klasterizavimo algoritmai dokumentų kolekcijai struktūrizuoti atlieka *n* žingsnių. Patys algoritmai yra skirstomi į dvi grupes: aglomeratyvius algoritmus (*ang. agglomerative*) ir dalomuosius (*ang. divisive*) algoritmus. Vienintelis jų skirtumas yra tai, kaip jie formuoja skirstinių rinkinius.

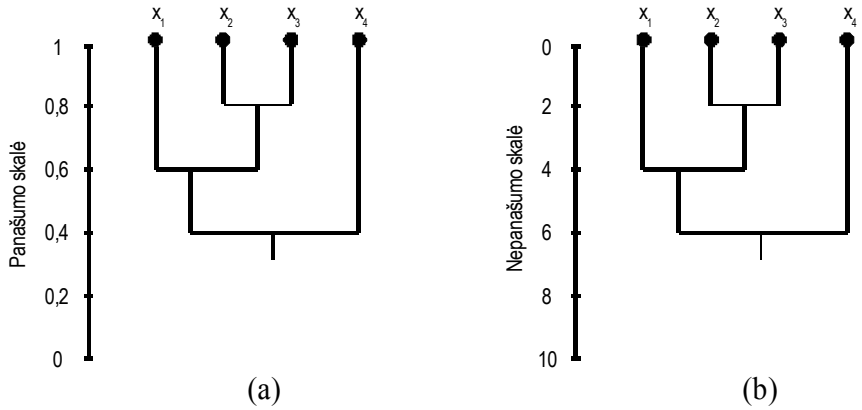
Aglomeratyvūs algoritmai veikia „iš apačios į viršų“ būdu. Pirmojo žingsnio metu sudaromas pirmasis skirstinių rinkinys R_0, sudarytas iš *n* skirstinių, turinčių lygiai po vieną dokumentą iš *D* kolekcijos. Antrojo žingsnio metu yra sudaromas antras rinkinys R_1, turintis *n–1* skirstinių, ir esantis R_0 viršaibiu. Žingsniai tęsiami tol, kol nesuformuojamas rinkinys R_{n-1}, turintis vienintelį skirstinį su visais *D* elementais. Algoritmo rezultatas yra skirstinių rinkinių hierarchija:

$$R_0 \subset R_1 \subset \ldots \subset R_{n-1}$$

Dalomieji hierarchinio klasterizavimo algoritmai veikia atvirkščiu būdu „iš viršaus į apačią“. Pirmasis skirstinių rinkinys R_0 yra sudarytas iš vienintelio skirstinio, turinčio visus *D* elementus. Paskutinis rinkinys R_{n-1} yra sudarytas iš *n* skirstinių, turinčių lygiai po vieną dokumentą iš *D*. Algoritmo rezultatas yra skirstinių rinkinių hierarchija:

$$R_{n-1} \subset R_{n-2} \subset \ldots \subset R_0$$

Hierarchinių klasterizavimo metodų rezultatai dažnai pateikiami dendogramomis (pav.) [10]. Jos dažniausiai vaizduojamos medžiu su šalia pateikta objekto panašumo (a) arba nepanašumo (b) lygių skale (žr. 3 pav.).



3 pav. Dendogramų su panašumo (a) ir nepanašumo (b) skalėmis pavyzdžiai.

Kitaip nei dalinančiųjų klasterizavimo metodų atveju, hierarchinių klasterizavimo metodų algoritmai yra labai reiklūs skaičiuojamiesiems resursams. Pavyzdžiui, dauguma aglomeratyvių algoritmų skaičiavimams atlikti naudoja matricas, laikomas atmintyje. Dėl šios priežasties tipinis *n* dokumentų kolekciją klasterizuojantis algoritmas turi $O(n^2)$ reikalavimus atminčiai (kur laikoma panašumų matrica) ir $O(n^3)$ reikalavimus laikui, nes į panašumų matricos elementus yra kreipiamasi *n–1* kartų. Dalomieji hierarchinio klasterizavimo algoritmai yra dar reiklesni – jau pirmojo žingsnio metu algoritmui tenka sumokėti $O(2^n)$ kainą.

Vieni labiausiai paplitusių hierarchinio klasterizavimo metodų yra *single-link*, *complete-link*, *group-average* ir *Ward* kvadratų sumų metodas.

5 Apibendrinimas

Šiame pranešime pateikta dokumentų klasterizavimo apžvalga supažindino su pagrindinėmis klasterizavimo sąvokomis, metodų ypatybėmis, jų žingsniais ir klasterizavimui alternatyviais dokumentų paieškos bei sistematizavimo būdais.

Metodų efektyvumo apžvalga atskleidė, kad iš pirmo požiūrio labiausiai informacijos gavybai ir dokumentų kolekcijų struktūrizavimui tinkantys hierarchinio klasterizavimo algoritmai yra labai imlūs skaičiavimo resursams. Šiuos algoritmus geriausia taikyti negausioms kolekcijoms (pavyzdžiui paieškos rezultatams) arba statiškai visai dokumentų kolekcijai vienu metu klasterizuoti (tuo labiau, kad pati jų prigimtis lemia, kad nedideli dokumentų kolekcijos pakeitimai lemia nežymius skirstinių hierarchijos pakitimus [6]). Dalinančio klasterizavimo metodai dažniausiai taiko euristinius algoritmus, todėl reikalauja visa eile mažesnių resursų. Algoritmų prigimtis lemai, kad juos verta taikyti labai gausioms ir kompaktiškoms dokumentų kolekcijoms.

Literatūros sąrašas

[1] Campbell, I. The Ostensive Model of Developing Information Needs. PhD thesis, University of Glasgow, Glasgow, 2000.

[2] Deboeck, G. Financial Applications of Self-Organizing Maps. Neural Network World, 8 (2) 1998: pp. 213-241.

[3] Efthimiadis, E.N. Interactive query expansion: a user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science, 51 (11) 2000: pp. 989-1003.

[4] Fred, A.L.N., Jain, A.K. Data Clustering Using Evidence Accumulation. Proceedings of the International Conference on Pattern Recognition, 16 2002: pp. 276–280.

[5] Hagen, P. Must Search Stink? Cambridge, USA, Forrester Research, 2000.

[6] Jardine, N., Van Rijsbergen C.J. The use of hierarchic clustering in information retrieval. Information storage and retrieval, 7 (5) 1971: pp. 217-240.

[7] Pelleg D., Moore, A. X-means: Extending K-means with efficient estimation of the number of clusters. Proceedings of the International Conference on Machine Learning, 17 2000: pp. 727-734.

[8] Rasmussen, C.E. The Infinite Gaussian Mixture Model. In Solla, S. A., T. K. Leen and K. R. Müller, eds. Advances in Neural Information Processing Systems, MIT Press 2000: pp. 554–560.

[9] Small, D.J. A Model-Driven Architecture for Enterprise Document Management, Supporting Discovery and Reuse. PhD thesis, University of Leeds, Leeds, 1999.

[10] Theodoridis, S., Koutroumbas, K. Pattern Recognition, Second Edition. San Diego, Academic Press, 2003: 640 p.

[11] Willett, P. Recent trends in hierarchic document clustering: a critical review. Information Processing and Management, 24 (5) 1988: pp. 577-597.

Document clustering

In modern-day organizations documents play a significant role. They are continually collected and as their number increases ordinary simple search and hierarchies based repositories are more and more slowing down information mining and retrieval. This paper investigates one of possible ways for such kind problem solving – document clustering. Various clustering methods and other alternative problem solving approaches are reviewed.