Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

„Airbnb“ paslaugų žaliųjų vartotojų pasitenkinimo vertinimas panaudojant teksto tyrybą

Baigiamasis magistro studijų projektas

Gabrielė Monkuvienė

Projekto autorė

Doc. Dr. Aušra Rūtelionė

Vadovė

Doc. Dr. Tomas Ruzgas

Vadovas

Kaunas, 2021

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

„Airbnb“ paslaugų žaliųjų vartotojų pasitenkinimo vertinimas panaudojant teksto tyrybą

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Gabrielė Monkuvienė
Projekto autorė

Doc. Dr. Aušra Rūtelionė
Vadovė

Doc. Dr. Tomas Ruzgas
Vadovas

Prof. Dr. Jurgita Bruneckienė
Recenzentė

Doc. Dr. Vytautas Janilionis
Recenzentas

Kaunas, 2021

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Gabrielė Monkuvienė

„Airbnb“ paslaugų žaliųjų vartotojų pasitenkinimo vertinimas panaudojant teksto tyrybą

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Gabrielė Monkuvienė

Patvirtinta elektroniniu būdu

Monkuvienė, Gabrielė. „Airbnb“ paslaugų žaliųjų vartotojų pasitenkinimo vertinimas panaudojant teksto tyrybą. Magistro studijų baigiamasis projektas.

Vadovė doc. dr. Aušra Rūtelionė; Kauno technologijos universitetas, Ekonomikos ir verslo fakultetas.

Vadovas doc. dr. Tomas Ruzgas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika.

Reikšminiai žodžiai: didieji duomenys, tvarumas, Airbnb, teksto tyryba, klasterinė analizė, sentimentų analizė.

Kaunas, 2021. 75 p.

Santrauka

Šio darbo tyrimo objektas yra „Airbnb” apgyvendinimo paslaugų žalųjų vartotojų atsiliepimai. „Airbnb“ yra optimalus tyrimo objektas, kuris yra skirtas išanalizuoti tvarų vartotojų elgesį apgyvendinimo sektoriuje. Skirtingi tyrėjai bandė geriau suprasti savybes, lemiančias šio tipo apgyvendinimo vartotojų elgesį ir jų pasitenkinimo lygį. Daugumoje tyrimų buvo skiriamas ribotas dėmesys apgyvendinimo paslaugų tvarumo atributams. Pirmoje darbo dalyje atlikta literatūros analizė parodo, kad apgyvendinimo paslaugų vartotojų internetinių atsiliepimų reikšmė jų pasitenkinimui yra labai svarbi įvairiuose literatūros šaltiniuose.

Atlikus mokslinės literatūros analizę, nustatyta pagrindinė darbo problema – kokie atributai lemia žaliųjų „Airbnb“ apgyvendinimo paslaugų vartotojų pasitenkinimą? Antroje darbo dalyje buvo pasiūlyta žaliųjų „Airbnb” vartotojų pasitenkinimo vertinimo metodika bei jos realizacija. Tai apima duomenų tvarkymą ir paruošimą tolimesnei analizei, asociacijų analizę, duomenų klasterizavimą Ward‘o, k-vidurkių ir k-medoidų metodais ir sentimentų analizę. Naudojama programinė įranga R, bei programavimo kalba Python.

Trečioje darbo dalyje atlikus sentimentų analizę pastebėta, kad teigiamą emociją žaliesiems “Airbnb” vartotojams kelia tokie dalykai kaip grožis, draugiškumas, taikumas, švara, skanus maistas, šviežumas, šiluma bei tvarumas. Neigiamas emocijas vartotojams kelia įvairios problemos, triukšmas, šaltis, kvapas, stresas, trukdymas, šiukšlės, įvairūs vabalai ir panašiai. Pastebėta, kad žalieji “Airbnb” vartotojai yra linkę palikti teigiamas emocijas sukeliančius atsiliepimus daug labiau nei neigiamus.

Monkuviene, Gabriele. Assessing the satisfaction of „Airbnb“ services green users using text mining. Master‘s Final Degree Project.

Supervisor doc. dr. Ausra Rutelione, Faculty of Economics and Business, Kaunas University of Technology.

Supervisor doc. dr. Tomas Ruzgas; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics.

Keywords: Big data, sustainability, Airbnb, text mining, cluster analysis, sentiment analysis.

Kaunas, 2021. 75 p.

Summary

The object of this thesis is the feedback of green users of „Airbnb“ accommodation services. „Airbnb“ is the optimal object of research to analyze sustainable consumer behavior in the accommodation sector. Different researchers have tried to better understand the characteristics that determine the behavior of users of this type of accommodation and their level of satisfaction. Most studies have paid limited attention to the sustainability attributes of accommodation services. In the first part the analysis of the literature shows that the significance of online feedback of accommodation service users for their satisfaction is very important in various sources.

An analysis of the scientific literature has identified the main problem of the thesis - what attributes determine the satisfaction of green users of „Airbnb“ accommodation services? In the second part, the methodology of green „Airbnb“ users satisfaction assessment and its implementation were proposed. It includes data processing and preparation for further analysis, association analysis, clustering by Ward, k-means and k-medoid methods and sentiment analysis. In this thesis R software and Python programming language are used.

In the third part, the analysis of sentiments revealed that things as beauty, friendliness, peace, cleanliness, delicious food, freshness, warmth and sustainability evoke a positive emotion for green „Airbnb“ users. Negative emotions to users are caused by various problems, noise, cold, smell, stress, disturbance, garbage, various bugs and others. It has been observed that green „Airbnb“ users tend to leave positive emotional feedback much more than negative ones.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Dalijimosi ekonomikos apgyvendinimo platformos pakeitė turistų apgyvendinimo sektorių bei pakeitė turizmo vartojimo modelius. Pastaruoju metu internetinės vartotojų apžvalgos daro vis didesnę įtaką vartotojų sprendimams, užsisakant apgyvendinimo paslaugas internetu. Vis labiau auga susidomėjimas ištirti vartotojų patirtį, analizuojant jų internetines apžvalgas, panaudojant teksto tyrybos ir sentimentų analizės metodus. Gebėjimas apdoroti didelę nuomonių įvairovę iš vartotojų atsiliepimų yra vienas iš svarbiausių verslo įrankių, norint suprasti priežastis, kaip ir kodėl vartotojas reaguoja į paslaugas bei produktus.

Temos aktualumas. „Airbnb“ yra dalijimosi ekonomikos verslo modelio paradigma, kuriuo šiuo metu naudojasi daugiau nei 300 milijonų vartotojų. „Airbnb“ platformoje tiek šeimininkai, tiek svečiai gali rašyti atsiliepimus. Šiose apžvalgose svečiai rašo apie savo patirtį ir apžvelgia teigiamus ir neigiamus viešnagės aspektus. Analizuojant tokias vartotojų apžvalgas, buvo sukurtos įvairios naujos įžvalgos ne tik apie „Airbnb“, bet ir apie pačią dalijimosi ekonomiką apskritai. Vis dėlto tyrimų, kuriais būtų bandyta ištirti konkrečiai žaliųjų „Airbnb“ vartotojų patirtį, vis dar nėra, ypač nėra tokių, kurie naudotųsi didžiųjų duomenų tyrybos metodais.

Pastebėta, kad yra atlikta gana nemažai mokslinių tyrimų viešbučių vartotojų pasitenkinimo vertinimo tema, tačiau būtent privataus būsto nuomos atvejai kai analizuojama klientų pasitenkinimą lemiantys atributai yra nagrinėta rečiau [1, 2, 4, 5, 6], ypač retai nagrinėjami žaliųjų „Airbnb“ vartotojų atsiliepimai [3, 8].

Šiame darbe bus tiriama **problema** – kokie atributai lemia žaliųjų „Airbnb“ apgyvendinimo paslaugų vartotojų pasitenkinimą?

Šio tyrimo **tikslas** - išnagrinėti žaliųjų „Airbnb“ vartotojų atsiliepimus, naudojant teksto tyrybos ir sentimentų analizės metodus ir įvertinti atributus, darančius įtaką vartotojų pasitenkinimui.

Šio darbo tikslui pasiekti iškeliami tokie **uždaviniai**:

1. atlikti mokslinės literatūros analizę apie žaliųjų „Airbnb“ paslaugų vartotojų atributus, darančius įtaką vartotojų pasitenkinimui;
2. parinkti atributus, lemiančius žaliųjų „Airbnb“ vartotojų pasitenkinimą, ir metodus, kurie yra labiausiai tinkami tokio pobūdžio tyrimams;
3. sukurti metodiką „Airbnb“ apgyvendinimo paslaugų vartotojų pasitenkinimo vertinimo modeliams;
4. pritaikyti sukurtą metodiką bei programines priemones realių pasirinktų duomenų analizei;

1. Literatūros apžvalga

Šiame skyriuje aptariami dalijimosi ekonomikos principai apgyvendinimo sektoriuje, taip pat kalbama apie žaliuosius „Airbnb“ vartotojus, bei jų pasitenkinimo vertinimą panaudojant didžiuosius duomenis. Taip pat, šiame skyriuje aptariami įvairių mokslininkų darbai skirti įvertinti apgyvendinimo įstaigų klientų pasitenkinimą, panaudojant didžiuosius duomenis. Taip pat aprašomi pagrindiniai matematiniai metodai, kurie yra naudojami norint išspręsti darbo uždavinius ir problemą, tokie kaip teksto tyryba, segmentų analizė, bei atliekama programinės įrangos apžvalga.

1.1. Dalijimosi ekonomikos principai apgyvendinimo sektoriuje

Dalijimosi ekonomika (angl. *sharing economy*) – tai tokia rinka, kuri apima daugybės ekonomikos sektorių, tokių kaip apgyvendinimo, transporto, laisvalaikio ir maisto sektorių dalijimąsi prekėmis ir paslaugomis [6]. Galima sakyti, kad dalijimosi ekonomika koordinuoja tam tikrų prekių bei paslaugų įsigijimą ir jų platinimą gaunant atlygį arba tam tikras kitokio pobūdžio kompensacijas [7]. Šiuo metu informacinės technologijos leidžia visiems įsitraukti į verslumą bei teikti visų rūšių produktus ir paslaugas vos vienu mygtuko paspaudimu [1].

Dalijimosi ekonomika gali padidinti prieigą prie prekių ir paslaugų, taip pat prisideda prie vartotojų gerovės gerinimo bei socialinių išlaidų mažinimo. Dalijimosi ekonomika sumažina investicijų poreikį į išteklius bei infrastruktūrą [3]. Pasak autorių M. Cheng‘o, X. Jin‘o būtent dalijimosi ekonomikos fenomeną lemia žmonių noras būti tvariais bei mėgautis veikla ir ekonomine nauda [4].

C. Midgett‘as, S. Bendickson‘as, J. Muldoon‘as ir J. Solomon‘as savo straipsnyje teigia, kad dalijimosi ekonomikai yra būdingi šie bruožai [8]:

- dalijimosi ekonomika yra orientuota į rinką;
- dalijimosi ekonomika paprastai apima „didelio poveikio“ kapitalą;
- dalijimosi ekonomika yra pagrįsta socialiniais mainais;

Dalijimosi ekonomika nėra pagrįsta tik dalijimusi, bet taip pat yra pagrįsta ir informacinių technologijų naudojimu, tam kad neefektyvumas rinkoje būtų panaudojamas, bei kad būtų galima konkuruoti dėl geresnės kainos. Informacinių technologijų naudojimas suteikia daugiau galimybių tiems, kurie nori išnuomoti savo būstą, automobilį ar bet kokį kitą daiktą. Jeigu nebūtų informacinių technologijų, dalyvauti dalijimosi ekonomikoje būtų labai mažai galimybių, arba praktiškai neįmanoma [8].

Daugelis skaitmeninių platformų vartotojų, kuriose plėtojama dalijimosi ekonomika mano, kad bendras vartojimas reiškia dideles pajamas vietos gyventojams, o būtent tai savo ruožtu leidžia prisidėti gerinant bendruomenės gerovę [3].

M. Petruzzi, C. Marques ir V. Sheppard savo tyrime išskiria 12 skirtingų dalijimosi ekonomikos savybių: socialiniai ryšiai, bendros nuosavybės jausmas, priklausomas, panašumas į realų dalijimąsi, socialinė reprodukcija, singuliarumas, tvarumas, nepakankamai panaudoti ištekliai, abipusis ryšys, pinigų svarba, pinigų svarba ir skaičiavimas [6].

P2P („peer-to-peer“) yra tinklo modelis, kuriame tiesiogiai tarp vartotojų vyksta keitimasis resursais. P2P yra novatoriškas paslaugų reiškinys, kuris leidžia paprastiems žmonėms, pasiūlyti svetingumą

(išnuomodami savo laisvus miegamuosius kambarius ar neužimtas patalpas) turistams. Būtent todėl atrodo, kad dalijimosi ekonomika transformuoja bei trikdo jau labai seniai įsitvirtinusią verslo praktiką. Dalijimosi ekonomikai augant ir vis didėjant verslo modelių įvairovei, vartotojų elgsena vis keičiasi. Vartotojai vis labiau nori dalyvauti P2P ekonomikoje, prekiaudami visais įmanomais ištekliais su svetimais žmonėmis, įskaitant ir savo būsto nuomą. Vartojimo kontekste paslaugos, teikiamos per apgyvendinimą P2P, gali būti suvokiamos kaip kitokios nei viešbučio paslaugos, o tai gali lemti skirtingus vartotojų lūkesčius ir paslaugų vertinimą. Tai yra ypatingai svarbu, kadangi literatūroje dažniausiai yra daroma tokia prielaida, kad vartotojai naudojasi paslaugų teikėjų ir klientų santykiais kaip pačių paslaugų vertinimo pagrindu. Be to, makroaplinkos požiūriu dalijimosi ekonomikos gimimas buvo susietas su požiūrio ir elgesio su vartojimo praktika pokyčiais apskritai, atsirandančio dėl įvairių visuomenės ir ekonominių spaudimų, tokių kaip bendruomenės troškimas, tvari vartojimo forma ir kt. taupumas, kurį palengvina socialinių tinklų ir mobiliųjų technologijų pažanga. Tai rodo, kad apgyvendinimas P2P gali skirtis vartotojų poreikius, palyginti su viešbučiais. Būtent todėl tai, kas lemia vartotojų pasitenkinimą bei jų tolimesnį ketinimą naudotis apgyvendinimo paskaugomis P2P, gali gerokai skirtis nuo susijusių su viešnage viešbutyje [12].

„Fairbnb.coop“ neseniai atvyko į rinką, neva norėdama išspręsti daugelį neigiamų socialinių problemų, susijusių su „Airbnb“, tokių kaip daugelio ilgalaikių nuomų pašalinimas iš rinkos ir nekilnojamojo turto vertės padidėjimas. „Fairbnb.coop“ siekia sudaryti tokias sąlygas visoms suinteresuotosioms šalims, tokioms kaip svečiai, šeimininkai ir kaimynai, bendradarbiauti su savivaldybėmis, siekiant užtikrinti, kad tam tikra dalis pelno būtų investuota atgal į tos vietos projektus. „Fairbnb.coop“ tvirtina, kad tai sujungia šeimininkus ir svečius prasmingoms, tvarioms ir socialiai teigiamoms kelionių ir mainų galimybėms, tuo pačiu suteikiant gyventojams demokratinę galimybę kartu formuoti operacijas [6].

„Airbnb“ yra vienas gerai žinomas P2P („peer-to-peer“) ekonomikos atstovas, prekiaujantis apgyvendinimu tarp asmenų (paprastai nepažįstamų žmonių) per internetinę platformą, siūlančią privataus kambario / buto rezervavimo paslaugą už tam tikrą mokestį. Tūkstančiai „Airbnb“ šeimininkų laukia svetimų žmonių iš viso pasaulio, kad galėtų išnuomoti savo namus [53].

1.1.1. „Airbnb“atvejis

„Airbnb“ platforma paprastai yra laikoma dalijimosi ekonomikos pavyzdžiu. „Airbnb“ – tai platforma, kuri buvo įkurta 2008 m. San Fransiske. Ši platforma siūlo prieigą prie milijonų vietų apsistoti beveik visame pasaulyje, pradedant apartamentais bei vilomis, baigiant pilimis bei nameliais medžiuose [1]. „Airbnb“ įžengimas į turizmo sektoriaus rinką bei klestėjimas pasaulyje smarkiai sumažino tradicinių viešbučių pajamas. Ši internetinė platforma, kurios pagalba galima dalintis namais ne tik, kad palengvino keliones šalies viduje, bet taip pat palengvino ir patį išvykimą į užsienį bei žinoma turizmą [18]. Kaip tipiškas dalijimosi ekonomikos pavyzdys, „Airbnb“ sujungia šeimininkus bei svečius dalinantis visu namu ar tik jo dalimi trumpalaikei nuomai [1].

Spartus „Airbnb“ augimas su savo išskirtiniu veikimo modeliu suteikia ne tik alternatyvią apgyvendinimo patirtį savo vartotojams, bet ir meta labai didelį iššūkį tradicinių viešbučių sektoriaus sukurtoms teorijoms ir praktikai. „Airbnb“ vadina save kaip atskirą interneto platformą, kuri suteikia žmonėms unikalią kelionių patirtį. Dėl šios priežasties vis daugėja tyrimų, kuriuose mokslininkai pradėjo tyrinėti atributus, darančius įtaką „Airbnb“ vartotojų patirčiai [4].

Pasak autorių L. Serrano, A. Ariza-Montes‘o, M. Nader‘io, A. Sianes‘o ir R. Law trumpalaikė nuoma yra viena iš ekonominės veiklos rūšių, kurios plėtrai daugiausiai įtakos turi dalijimosi ekonomika. „Airbnb“ platforma vertinama turistų kaip pirmaujanti visame pasaulyje apgyvendinimui skirta platforma. „Airbnb“ palyginti su tradicinėmis apgyvendinimo paslaugomis suteikia tam tikrą pridėtinę vertę, kadangi savo svečiams suteikia autentiškus išgyvenimus [3]. Būtent vienas iš svarbiausių „Airbnb“ vartotojų elgsenos veiksnių yra jų rūpestis tvarumu bei įmonės įsipareigojimas remti vietos ekonomiką ir kartu sumažinti neigiamą poveikį aplinkai [13].

Akademinis susidomėjimas „Airbnb“ platforma yra labai akivaizdus, kadangi vis daugiau mokslo disciplinų tyrinėja būtent jo sėkmės modelį. Tyrėjai iš skirtingų akademinių sričių vis bando suprasti „Airbnb“ poveikį turizmo sektoriui [3]. Pažangus turizmas bei jame esanti dalijimosi ekonomika dabar keičia žmonių gyvenimą ir yra viena iš didžiausių turizmo sektoriaus naujovių [7].

„Airbnb“ ir kitos trumpalaikės nuomos P2P („peer-to-peer“) kompanijos atstovauja platesnei dalijimosi ekonomikos daliai. Šių paslaugų augimas dalijimosi ekonomikoje yra transformuojanti turizmo apgyvendinimo pramonės naujovė [14]. Dalijimosi ekonomika būsto nuomos sektoriuje sparčiai augo, užtikrindama gana pigų būstą bei namų aplinką, taip pat ir tiesioginį bendravimą su vietos bendruomene. Nors „Airbnb“ yra vyraujantis trumpalaikės nuomos pavyzdys apgyvendinimo sektoriuje, tačiau kiti „Airbnb“ konkurentai, įskaitant „HomeAway“, „HouseTrip“ ir „FlipKey“ dalijasi apgyvendinimo rinka, sutelkdami dėmesį į klientus, kuriems reikalinga būsto nuoma atostogoms [16].

Pasak autoriaus I. Tussyadiah‘o, dalijimosi ekonomika įsibėgėja svetingumo rinkoje su apgyvendinimo kategorija P2P („peer-to-peer“). Šis P2P modelis keliautojams tampa vis perspektyvesniu pasirinkimu, todėl labai svarbu vis geriau suprasti veiksnius, kurie lemia klientų pasitenkinimą naudojant būtent šią apgyvendinimo rūšį [12].

1.2. Paslaugų vartotojų pasitenkinimo samprata ir dimensijos: „Airbnb“ atvejis

Yra daugybė efektyvių modelių ir teorijų, kurie apibrėžia ir tyrinėja vartotojų pasitenkinimą. Autoriai Liang‘o, Choi‘o ir Joppe savo tyrime išskiria vieną pagrindinių teorijų – Oliver‘io laukiamumo ir nepatvirtinimo teorija (angl. *expectancy-disconfirmation theory*), kurią vėliau Kristensen‘as ir kt. išplėtė į lūkesčių patvirtinimo teoriją (angl. *expectancy-confirmation theory*). Pasak Oh‘o bei Parks‘o [53] yra išskiriamos dar aštuonios teorijos ar sąvokos, kurios tyrinėja vartotojų pasitenkinimą. Tyrėjai sutaria dėl platesnio pasitenkinimo apibrėžimo. Pavyzdžiui, Fang‘as ir kt. [53] priėmė Holmes‘o siūlomą pasitenkinimo apibrėžimą, nurodydamas jį kaip su praeityje susijusios patirties bei mainų vertinimo rezultatą. Šis apibrėžimas yra panašus į Kim‘o, kuris teigia, kad pasitenkinimas suvokiamas kaip požiūris, atsirandantis dėl paslaugos ir kokybės palyginimo, kurio vartotojas tikisi gauti iš sandorio po pirkimo. Visgi, autoriai išskiria du pasitenkinimo tipus, tai sandoriais pagrįstas pasitenkinimas ir patirtimi pagrįstas pasitenkinimas. Pripažįstama, kad vartotojai skirtinguose procesuose gali taikyti skirtingus vertinimo kriterijus [53].

Vartotojų pasitenkinimas apibrėžiamas kaip skirtumas tarp vartotojo lūkesčių ir įgytos prekės ar paslaugos suvokimo rezultatas. Vartotojas yra patenkintas preke arba paslauga, jeigu ji atitiko jo lūkesčius [51]. Pasitenkinimo paslaugomis jausmas nebūtinai reiškia, kad paslaugos buvo aukščiausios kokybės, tai labiau reiškia, kad pasiektas priimtinas ir atitinkantis lūkesčius standartas.

Pasitenkinimo sąvoka yra sudėtinga [52]:

- Pasitenkinimas nėra statiškas ir yra kintantis bėgant laikui.
- Pasitenkinimas yra sudėtinis procesas, apimantis patirtis prieš, per ir po atskaitos taško.
- Pasitenkinimas socialiniame kontekste kyla, nuolat kinta, bei gali nenuspėjamas paslaugų vartotojams.
- Pasitenkinimo priežasčių nustatymas yra pakankamai sudėtingas, kartais gali būti lengviau nustatyti nepasitenkinimo priežastis.

Pasitenkinimo sąvoka paaiškinama nepatvirtinimo teorija, pagal kurią vartotojo pasitenkinimas paslauga yra susiejamas nepatvirtinimo patirtimi ir jos dydžiu, o nepatvirtinimas yra susijęs su pradiniais lūkesčiais kurių turėjo asmuos. Jeigu paslauga, kuri yra suteikta gerokai viršija vartotojo lūkesčius, jo pasitenkinimas bus didelis ir atvikščiai. Pasitenkinimą galima išmatuoti trimis skirtingais lygmenimis, tokais kaip bendrasis pasitenkinimas paslauga, pasitenkinimas veiklos sričių darbu ir pasitenkinimas procesais veiklos srityse [51].

Vartotojų pasitenkinimas - pati pagrindinė paslaugų rinkodaros ir valdymo sąvoka. Paslaugų, kurios sukelia vartotojams pasitenkinimą bei norą sugrįžti teikimas yra pripažintas kritiniu sėkmės veiksniu bei konkurencinio pranašumo šaltiniu įvairioms paslaugų įmonėms [12].

Po prekių ar paslaugų vartojimo, pasidalijus emocijomis ir patirtimi, bendrasis žmonių pasitenkinimas yra pagrįstas sąveika ir socialine komunikacija. Būtent dėl šios priežasties socialinė sąveika yra labai svarbus „Airbnb“ paslaugų vartotojų patirties atributas [2]. Akademinė bendruomenė daugelį metų bando nustatyti, kokie kritiniai atributai, lemia žaliųjų „Airbnb“ vartotojų pasitenkinimą [3].

Mokslininkų G. Dominici ir R. Guzzo teigimu, paslaugų kokybė ir vartotojų pasitenkinimas yra pagrindiniai veiksniai siekiant konkurencinio pranašumo bei norint išlaikyti savo vartotojus. Vartotojo pasitenkinimas tai yra rezultatas, kai vartotojas suvokia gaunamą vertę, gautą sandorio metu, kai vertė yra lygi kliento suvokiamai paslaugų kokybei, palyginti su verte, kurios yra tikimasi iš sandorių su konkuruojančiais pardavėjais [15].

Pasak autoriaus I. P. Tussyadiah‘o nustatyta, kad atributas „vieta“ nėra statistiškai reikšminga ir nedaro didelės įtakos „Airbnb“ vartotojų pasitenkinimui, tačiau tuo tarpu tokie atributai kaip „malonumas“, „patogumai“ bei „sąnaudų taupymas“ pagal svarbą buvo vertinami teigiamai, tai reiškia, kad būtent šie atributai yra reikšmingi vartotojų pasitenkinimui [12]. Kitų autorių teigimu, pagrindiniai veiksniai, kurie daro įtaką „Airbnb“ klientų pasitenkinimui yra šeimininkas, kambarys / namas, vieta bei kaimynystė [1]. Pasak autoriaus Y. Luo, aplinka bei paslaugų kokybė įvairiuose restoranuose yra svarbūs atributai, kurie stipriai lemia klientų pasitenkinimą [10].

Rinkodaros specialistai vis dažniau atkreipia dėmesį į personalizavimo svarbą, kadangi tai veikia vartotojų pasitenkinimą bei jų lojalumą. Apgyvendinimo paslaugų sektoriuje vartotojai teigiamai vertina apgyvendinimą, kuriame yra teikiamos paslaugos pagal jų vardus, pageidavimus ir jų kitą asmeninę informaciją. „Airbnb“ platformos vartotojai tikisi, kad su jais bus elgiamasi unikaliai. Kadangi „Airbnb“ apgyvendinimo platforma neteikia standartizuotų kambarių bei paslaugų, vartotojai iš šeimininkų gali susilaukti ypatingos bei visiškai netikėtos patirties [2].

Įvairūs atlikti tyrimai turizmo sektoriuje rodo stiprią priklausomybę tarp vartotojų pasitenkinimo bei jų lojalumo arba ketinimo apsilankyti dar kartą ir taip pat rekomenduoti kelionės tikslą kitiems žmonėms. Pasak autorių C.V. Priporas‘o, N. Stylos‘o, L. N. Vedanthachari ir P. Santiwatana paslaugų

kokybė daro tikrai reikšmingą ir tiesioginę įtaką apgyvendinimo paslaugų klientų pasitenkinimui. Paslaugų kokybė, klientų pasitenkinimas bei lojalumas yra esminiai šio verslo sėkmės elementai, kuriuos būtina nuolat stebėti [5].

Autoriai T.N. Akarsu, P. Foroudi, T.C. Melewar'o savo tyrime teigia, jog vartotojų suvokiama vertė daro teigiamą ir tiesioginę įtaką tolimesniems elgesio ketinimams. Todėl, autorių teigimu vartotojai, suvokiantys didenę „Airbnb" vertę, greičiausiai turi stipresnį ryšį ir su savo suvoktu autentiškumu bei „Airbnb" patirtimi [17].

Nustatyta, kad klientų aptarnavimo patirtis bei jų elgesys po paslaugos vartojimo priklauso ne tik nuo siūlomų paslaugų tam tikrų savybių, bet priklauso ir nuo klientų socialumo stiprumo. Teikiant paslaugas, kurios yra susijusios su klientų bei darbuotojų sąveika, tokie klientai, kurie yra labiau socialūs, dažniausiai labiau vertina santykius su paslaugų teikėjais bei tokie santykiai yra vertinami kaip sėkmingesni [19].

Taigi, bendrovėms, stipriai orientuotoms į vartotoją, vartotojo pasitenkinimas yra ne tik tikslas, bet taip pat ir pagrindinis veiksnys, kuris lemia jų sėkmę rinkoje, todėl labai svarbu atkreipti dėmesį į vartotojų pasitenkinimą lemiančius veiksnius, norint pirmauti rinkoje.

1.3. Žaliųjų vartotojų koncepcija

Visoje turizmo pramonėje yra pastebima tvarumo tendencija, kuri daro poveikį visai industrijai, kadangi keliautojų nuomonė bei norai tampa vis labiau orientuoti į aplinką bei jos tausojimą. Tvarus turizmas pasižymi darnaus vystymosi koncepcijos pritaikymu turizmo pramonei bei turizmo plėtrai.

Kelerius pastaruosius dešimtmečius turistų reikalavimai ekologiškam turizmui vis auga. Tai reiškia, kad ekologiška įmonių veikla tampa būtina, siekiant išlaikyti įmonių konkurencingumą ar užimti didesnę rinkos dalį. Daugelis turizmo įmonių stengėsi išrasti naujus ekologiškus produktus, kad pritrauktų ir skirtų daugiau turistų, tačiau ekologiško turizmo įmonėms tenka žymiai didesnės investicijos ir ilgesnis atsipirkimo laikotarpis nei tradicinėms. Be to, kai kurios vietos valdžios institucijos daugiau dėmesio skyrė ekonominiams rodikliams nei poveikiui aplinkai [54].

Verta paminėti, kad vietos valdžios institucijų elgesys negali tiesiogiai paveikti turistų sprendimo pirkti, tačiau tai gali netiesiogiai paskatinti turistus priimti ekologiško turizmo modelį, skatinant turizmo įmones parduoti ekologiškus produktus. P. He, Y. He ir F. Xu tyrime pabrėžiama, kad turizmo įmonių prekės ženklo naudos didinimas ir (arba) ekologiškas turistų pasirinkimas yra nepaprastai naudingas būdas skatinti suinteresuotąsias šalis imtis ekologiško turizmo. Pasak autorių, norint plėtoti tvarų turizmą, svarbu padidinti vartotojų norą mokėti už ekologiškus produktus [54].

Žalieji vartotojai yra tie, kurie gerai suvokia savo elgesio poveikį aplinkai. Žaliųjų turistų bruožas yra tas, kad jiems labiau patinka rinktis produktus, kurie yra ekologiški, jei yra jautrūs ir gerbia vietos kultūrą, yra rūpestingi, entuziastingi ieškant naujos patirties, dalyviai, o ne žiūrovai. Šiai grupei žmonių taip pat yra būdinga viduriniosios klasės socialinė padėtis, patiriamos didelės išlaidos ir aukštasis išsilavinimas. Jų gyvenimo būdas lemia ekologišką vartojimo elgesį, kuris taip pat žinomas kaip ekologiškas vartotojas. Teoriškai žalioji rinkodara žino kaip vartotoją, kuris nusipirko ekologinių ženklų produktą dėl savo aplinkos supratimo. Žalieji turistai taip pat turi daugybę pageidavimų, susijusių su numatoma paslauga, ypač renkantis apgyvendinimą [55].

„Žaliasis“ vartotojas (angl. *green user*) yra sutelkęs savo dėmesį į visų tvarumo aspektų įtraukimą į savo turizmo patirtį. Tai yra visiškai naujas segmentas rinkoje, kurį turizmo pramonės dalyviai pradeda naudoti [8]. Būtent susidomėjimas žaliaisiais apgyvendinimo paslaugų vartotojais pateisina bendrą turizmo bei svetingumo sektoriaus tvarumo tendenciją. Dėl šios tendencijos vis dažniau atsiranda į ekologiškumą linkusių vartotojų, kurie sutelkia dėmesį į savo patirtį naudojantis apgyvendinimo paslaugomis [3]. P. He, Y. He ir F. Xu savo tyrime patvirtino, kad aplinką tausojantys vartotojai išreiškia didesnį palankumą ekologiškiems produktams ir didesnį jų pasitenkinimą [54].

Nuo 2000-ųjų metų pradžios visą vartotojo elgsenos evoliuciją žymi du reiškiniai, tokie kaip eksponentinis interneto naudojimo augimas bei vartotojų rūpestis tvarumu. Įvairūs socialiniai tinklai, skaitmeninės kelionių platformos bei interneto svetainės galiausiai pasiekė tokį vystymosi lygmenį, kuris tai pavertė turistų bei apgyvendinimo sektoriaus įvairių specialistų prioritetiniais bendravimo kanalais. Visos šios informacinės bei ryšių technologijos yra informacijos šaltiniai, kurie turi labai didelę reikšmę visam sprendimų priėmimo procesui. Būtent turizmo sektoriuje susirūpinimas dėl ekonominės veiklos padarinių aplinkai yra ypač didelis, labiausiai tai pastebima viešbučių sektoriuje, todėl kuo toliau, tuo labiau stengiamasi eiti tvarumo link. [3].

Pasak autorių D. Guttentag‘o, S. Smith‘o, L. Potwarka‘so ir M. Havitz‘o tvarumas yra įvardijamas kaip labai svarbi motyvacinė priemonė „Airbnb“ paslaugų vartotojams [14]. Jeigu turizmo apgyvendinimo paslaugas teikiantys asmenys nori aukšto lygio svečių pasitenkinimo bei didelio sugrįžtančių svečių skaičiaus, jie turi nuolat stengtis, kad klientams nekiltų jokių abejonių dėl jų teikiamų tvarių iniciatyvų [8].

Autoriaus I. P. Tussyadiah‘o teigimu, tvarumas neigiamai veikia svečių pasitenkinimą privačiuose kambariuose, tačiau yra nereikšmingas būtent tiems, kurie apsistoja visame name arba visame bute. Tai gali reikšti, kad dauguma „Airbnb“ svečių apgyvendinimą pasirenka ne dėl ekologinių priežasčių [12]. Autorių M. Cheng‘o ir X. Jin‘o teigimu, tvarumo faktorius nėra kritinis vartotojo pirkimo sprendimo elementas.

Autoriai L. Serrano, A. Montes‘as, M. Nader‘is ir kiti teigia, kad būtent žalieji „Airbnb“ vartotojai yra glaudžiai susiję su dalijimosi ekonomika. Žaliųjų vartotojų pasirinkimai yra grindžiami didesnio tvarumo ieškojimu, tai yra, jie visada apsvarsto būtent ekologiniu ir socialiniu požiūriu tvaresnę alternatyvą tradicinės ekonomikos siūlomam vartojimo modeliui. Tvarumo paradigmą puikiai atitinka dalijimosi ekonomika, taip prisidėdama prie nuolatinio socialinių išlaidų mažinimo bei vartotojų gerovės gerinimo [3].

„Airbnb“ yra optimalus tyrimo objektas, kuris yra skirtas norint išanalizuoti ekologišką vartotojų elgesį apgyvendinimo sektoriuje. Daugelyje tyrimų buvo bandoma geriau suprasti atributus, lemiančius tokio tipo apgyvendinimo vartotojų elgesį bei jų pasitenkinimo lygį [4]. Daugelyje tyrimų buvo skiriamas ribotas dėmesys į apgyvendinimo tvarumo atributus. Išsami Cheng‘o ir Jin‘o gautų rezultatų analizė rodo, kad tvarumo faktorius nėra kritinis vartotojo pirkimo sprendimo elementas. Taigi reikia kurti naujus tyrimus ir giliau suvokti žaliųjų „Airbnb“ vartotojų ketinimus renkantis apgyvendinimą [4].

Empirinių tyrimų šiuo klausimu dalijimosi ekonomikos srityje trūkumas galbūt yra dėl turimos informacijos apie vartotojų elgseną išsisklaidymo ir nevienalytiškumo. Kitame skyriuje bus patikrinta, ar didžiųjų duomenų analizė yra naudinga priemonė, padedanti įveikti apribojimus, susijusius su klasikiniais analizės metodais.

1.4. Didžiųjų duomenų panaudojimas vertinant vartotojų pasitenkinimą dalinimosi ekonomikos atveju

Didžiulis socialinių tinklų bei jų vartotojų sukurto turinio augimas įkvėpė kurti vadinamąją didžiųjų duomenų analitiką, tam kad būtų galima suprasti bei išspręsti tam tikras realaus gyvenimo problemas [24]. Duomenys tai yra tam tikri objektyviai egzistuojantys faktai, taip pat gali būti vaizdai arba net garsai, kurie gali būti labai naudingi tam tikriems uždaviniams spręsti. Duomenys galiausiai virsta tam tikra informacija, tuomet kai jiems yra suteikiamas kažkoks kontekstas bei jie būna susiejami su tam tikra problema arba jos sprendimu. [34].

Didžiųjų duomenų panaudojimas greitai pateko į turizmo tyrimų sritį [22]. Didesnis duomenų analizės, kaip verslo pagrindo, suvokimas tapo vis labiau paplitęs, kai įžengėme į „didžiųjų duomenų" amžių. Didieji duomenys paprastai apibrėžiami per tris skiriamąsias ypatybes, tai yra apimtis (angl. *Volume*), sparta (angl. *Velocity*) bei įvairovė (angl. *Variety*). Apimtis apibūdina duomenų kiekį, sparta nurodo duomenų apdorojimo greitį, o įvairovė nurodo duomenų tipą. Didžiulis duomenų kiekis bei jų įvairovė kelia didelį iššūkį įprastam požiūriui į duomenų analizę, kadangi duomenų kiekis bei jų struktūra gerokai viršija bet kokius žmogaus rankinio apdorojimo metodus [4].

Skirtingai nuo klasikinių metodų, kurie daugiausia yra pagrįsti dedukciniu samprotavimu, didžiųjų duomenų analizė yra grindžiama indukciniais samprotavimais bei dideliu statistinių duomenų apdorojimo kiekiu, nereikalaujant jokios išankstinės teorijos [20]. Trumpai tariant, didžiųjų duomenų analizė papildo įprastas duomenų analizės metodikas, kurios dažniausiai yra naudojamos moksliniuose tyrimuose, siekiant įveikti jų ribotumą. Naudojant didžiuosius duomenis galima apdoroti įvairius didelės apimties duomenis ir gauti reikiamą informaciją, tai padeda akademikams ir specialistams geriau suprasti skirtingus kintamuosius, darančius lemiamą įtaką turizmo sektoriaus sėkmei [3].

Didžiųjų duomenų analizė turi tikrai daug privalumų turizmo sektoriaus tyrimų kontekste, kadangi turizmo sektorius pasižymi ganėtinai intensyviu duomenų kaupimu [21]. Nors didžiųjų duomenų analizė buvo įvardinta kaip nauja tyrimų paradigma daugelyje mokslo disciplinų, yra labai nedaug pritaikymų svetingumo bei turizmo srityje iki galo panaudoja jos galimybes [9]. Tyrėjai, kurie analizuoja turizmo sritį, vis dažniau naudoja didžiuosius duomenis kartu su turinio analize [3]. Naudojant didžiuosius duomenis bei teksto analizę galima geriau suprasti svečių patirtį, taip pat ir klientų pasitenkinimą [16].

Būtent turizmo sritį galima vadinti pavyzdine didžiųjų duomenų analizės studijų sritimi, kadangi joje yra pateikiamos gairės visiems naujiems tyrimų metodams bei didžiųjų duomenų analizės metodų pagalba leidžiama analizuoti netradicinius duomenų šaltinius bei apdoroti didžiulį gaunamos informacijos kiekį. Didieji duomenys suteikia naujų galimybių turizmo tyrimams, tokių kaip turistų vartojimo bei jų elgesio modelių vizualizavimą ir panašiai [3]. Turistų socialiniuose tinkluose paskelbti internetiniai komentarai sukuria milijonus duomenų, kurie, atliekant teksto tyrybą bei sentimentų analizę, leidžia mums vis geriau suprasti jų bei kitų turizmo sektoriaus žaidėjų suvokimą bei jų pasitenkinimą [22].

Pasak autorių A. Alaei, S. Becken ir B. Stantic, didžiųjų duomenų metodų naudojimas gali stipriai pagelbėti turizmo sektoriaus tyrimams atrasti dinamiką, kuri būtų pagrįsta dideliais tarpusavyje susijusiais duomenų rinkiniais bei gauti kiek įmanoma daugiau įžvalgų iš skirtingų didžiųjų duomenų aspektų. Turizmo sektoriaus tyrimai gali toliau pereiti į naują sritį, kurioje tik teorija grindžiami

metodai bei duomenimis paremta praktika gali padėti vienas kitam vis labiau suprasti ar konkrečiai paaiškinti tam tikrus vykstančius reiškinius, taip pat dar galėtų padėti suvokti ir naujas teorijų dimensijas [22].

Nors didžiųjų duomenų analizė gana stipriai išpopuliarėjo atliekant turizmo sektoriaus analizės tyrimus, tik nedaugelyje šios srities tyrimų buvo pasirinkta ši metodika, vis mažiau tyrimų šią metodiką pritaikė būtent dalijimosi ekonomikos kontekste, pavyzdžiui tokia populiari platforma - „Airbnb“ [3].

Turistų socialiniuose tinkluose paskelbti internetiniai komentarai sukuria milijonus duomenų taškų, kurie, naudojant teksto tyrybą ir sentimentų analizę, leidžia geriau suprasti jų ir kitų turizmo pramonės žaidėjų suvokimą ir pasitenkinimą. Nagrinėjant būtent žaliųjų vartotojų internetinius atsiliepimus, išrenkami tik tie atsiliepimai, kuriuose naudojami žaliesiems vartotojams būdingi terminai. Žalieji vartotojai yra tie, kurie vartoja žodžius, susijusius su tvariu gyvenimo būdu. Atsižvelgiant į literatūros apžvalgą būvo pastebėta, kad dažniausiai žalieji vartotojai savo atsiliepimuose vartoja tokius žodžius kaip “tvaru”, “tvarumas” bei “ekologiškas” [3].

1.5. „Airbnb“ paslaugų žaliųjų vartotojų pasitenkinimo vertinimo tyrimai

Šiame poskyryje aprašomi moksliniai tyrimai apie „Airbnb“ paslaugų vartotojų pasitenkinimo vertinimą panaudojant duomenų analizės metodus, tokius kaip teksto tyryba bei sentimentų analizė, taip pat regresinė analizė bei faktorinė analizė.

L. Serrano , A. Ariza-Montes , M. Nader, A. Sianes ir R. Law atliko tyrimą [3], kuriame vertino ir analizavo „Airbnb“ žaliųjų vartotojų pasitenkinimą. Pirmiausia tyrimu buvo siekiama nustatyti „žaliųjų turistų“ (angl. *green users*) pageidavimus trumpalaikės nuomos apgyvendinimo tinklalapyje „Airbnb“, susijusius su jų vertinamais atributais. Taip pat, šiuo tyrimu buvo siekiama suprasti, kaip šių „žaliųjų turistų“ emocijos, kurios yra susijusios būtent su tokiais požymiais, lemia jų galutinį įmonės vertinimą. Šiam tyrimui buvo naudojama duomenų bazė, kurią sudarė internetiniai atsiliepimų komentarai iš 83 pasaulio miestų. Iš viso buvo gauti 176 852 704 komentarai. Tinklalapyje „Inside Airbnb“ yra renkama informacija kasmet bei pateikiama visomis kalbomis. Taigi, duomenys buvo sutvarkyti, tai yra buvo pašalinti dublikatai bei komentarai, kurie nerašyti anglų kalba. Galutinę duomenų bazę sudarė 13 181 297 unikalūs komentarai. Tyrime buvo taikyti didžiųjų duomenų analizės metodai: teksto tyryba, sentimentų analizė.

Gauti tyrimo rezultatai:

1. Žalieji „Airbnb“ vartotojai aiškiai teikia didelę pirmenybę tvarumui. Jų pasirinkimas yra grindžiamas kuo didesnio tvarumo ieškojimu, tai yra, jie visada apsvarsto kiek įmanoma ekologiškai bei socialiai tvaresnę alternatyvą tradicinės ekonomikos siūlomam vartojimo modeliui.
2. Tokie atributai kaip „tvarumas“ ir „šeimininkas“ yra pagrindiniai žaliųjų „Airbnb“ platformos vartotojų patirties tarpininkai, tai yra aspektai, kurie yra susiję su emociniu atributu „džiaugsmas“.
3. Žaliųjų vartotojų požiūris į „Airbnb“ savybes bei jų emocinis intensyvumas, kuris yra susijęs su jų parašytais komentarais, atsispindi apskritai paskirtose skaitinėse vertėse ir kiekviename iš šešių konkrečių atributų (švara, vieta, bendravimas, pridėtinė vertė, aprašymo tikslumas bei registracija).

4. Visi su tvarumu susiję elementai vaidina pagrindinį vaidmenį išreiškiant žaliųjų „Airbnb“ vartotojų nuomones ir vertinimus.

Autoriai savo tyrime apibrėžė apribojimus ir nurodymus kitiems ateities tyrimams. Šiame tyrime buvo nagrinėjami tik internetinių apžvalgų komentarai iš miestų, kurie buvo įtraukti į „Inside Airbnb“ svetainę. Taigi, ateities tyrimams reikėtų išplėsti imtį ir į kitus miestus, kuriuose veikia „Airbnb“ paslaugos bei naudoti kitus patikimus duomenų šaltinius. Toliau, kiti tekstai, kurie nebuvo parašyti anglų kalba, nebuvo įtraukti į šį tyrimą, taip neįtraukiant kitų kultūrinių bei demografinių perspektyvų, kurios tyrimui galėtų suteikti daugiau gylio gautoms išvadoms. Taip pat, autoriai teigia, kad išanalizavus žaliųjų „Airbnb“ vartotojų apžvalgų komentarus, būtų galima įtraukti informaciją, kuri būtų gauta iš aprašymų, kuriuos pateikia nuosavybės šeimininkai. Tokia informacija kartu su kitais pamatiniais kintamaisiais, tokiais kaip kaina už naktį ir geografinė padėtis, galėtų padėti pagerinti tyrimo rezultatus.

Mokslininkai C. K. H. Lee, Y. K. Tse, M. Zhang‘as ir J. Ma atliko tyrimą [1], kurio tikslas buvo ištirti atributus, kurie turi įtakos „Airbnb“ platformos klientų patirčiai, analizuojant pateiktas internetines apžvalgas iš Londone esančių vartotojų. Šiame straipsnyje buvo analizuojamos 169 666 apžvalgos, kurias paskelbė „Airbnb“ vartotojai, kurie viešėjo Londone 2011–2015 m. Tyrime buvo taikoma teksto tyryba, klasterinė analizė, bei naudojamas grafinės vizualizacijos.

Gauti tyrimų rezultatai:

1. Turistai paprastai naudoja panašius atributų rinkinius, norėdami įvertinti apgyvendinimo paslaugų patirtį. Pavyzdžiui, įprastai atributai apima nakvynės vietą, patogumus bei kainą. Tačiau pastebima, kad „Airbnb“ paslaugoms yra papildomų svarbių atributų, tokių kaip šeimininkas, švara bei jaukumas.
2. Įdomus pastebėjimas yra tas, kad saugumas internetinėse apžvalgose minimas rečiau, palyginti su švara ir jaukumu.
3. Remiantis atlikta sezonine analize, rezultatai parodo, kad svečiai vasarą labiau linkę vertinti rajoną nei kitais sezonais. Daugelis svečių mano, kad vieta yra geresnė, jei ji yra šalia parkų, kuriuose galima rasti tvenkinių, kuriuose galima maudytis bei yra vietų, kuriose vasarą būtų galima degintis.
4. Priešingai, būtent žiemos metu svečiai tiki, kad galimybė mėgautis karštu dušu švariame vonios kambaryje yra didelis privalumas. Neigiami komentarai, kurie yra susiję su žiemos patogumais, yra tai, kad nepakanka karšto vandens dušui. Taip pat pasitaikė tokių neigiamų komentarų, kurie teigė, kad durys arba langai turėjo būti uždaryti, tam kad būtų šilta, bei nustatytos ribotos valandos šildymui per dieną.

Autorių teigimu, kai kurie žmonės gali teigti, kad išvados, kurios yra gautos remiantis istoriniais 2011–2015 m. duomenimis, ateityje gali būti nebeaktualios, kadangi pasaulinė svetingumo pramonės rinka yra labai dinamiška. Tačiau autoriai mano, kad kai kurios šio tyrimo išvados išliks svarbios ir formuojant apgyvendinimo paslaugas ateityje. Ateities tyrimams autoriai siūlo analizuoti asmenines „Airbnb“ šeimininkų nuotraukas. Remiantis nuotraukomis, šeimininkų amžių, lytį bei jų emocijas būtų galima išskirti kaip atributus, naudojant veido atpažinimo technologijas. Taip pat siūloma išsiplėsti ir į kitas pasaulio šalis bei tirti bet kokius geografiškai išsisklaidžiusių klientų elgesio pokyčius.

J. Li, S. Hudson ir K. K. F. So atliko tyrimą [2], kurio tikslas buvo ištirti daugialypę „Airbnb“ klientų patirties struktūrą bei ištirti šios patirties įtaką vartotojų elgesio rezultatams. Dimensijoms tirti buvo naudojama daugiafazė metodika, naudojant apklausos klausimyną. Duomenys buvo surinkti iš 561 „Airbnb“ naudotojo iš JAV. Siekiant įvertinti skalės patikimumą ir pagrįstumą, buvo atlikta tiriamoji faktorinė analizė. Atlikus šį tyrimą buvo sukurta patikima ir galiojanti keturių dimensijų „Airbnb“ klientų patirties skalė.

Gauti tyrimų rezultatai:

1. Pagrindiniai „Airbnb“ apgyvendinimo elementai, tokie kaip švara, namų atmosfera bei namų patogumai, buvo svarbūs „Airbnb“ klientams.
2. Individualizuotas aptarnavimas yra kritinė „Airbnb“ vartotojų patirties dalis. Individualus aptarnavimas yra ne tik pagrindinė motyvacija, norint pritraukti klientus, bet ir akcentas jų visai viešnagei.
3. „Airbnb“ vartotojams buvo ypač svarbios dvi dimensijos - autentiškumas bei socialinė sąveika. Keliautojai reikalauja unikalios patirties, kuri būtų susijusi su prasmingu bendravimu su vietos gyventojais.

Kaip ir dauguma tyrimų šia tematikai, taip ir šis tyrimas turi tam tikrų apribojimų. Pirmiausia, šį tyrimą sudarė „Airbnb“ vartotojai JAV. Būsimi tyrimai galėtų išnagrinėti JAV ir kitų šalių „Airbnb“ vartotojų patirties skirtumus. Jeigu būtų atliekama daugiau tokio tipo tyrimų skirtinguose regionuose, tokie tyrimai būtų naudingi tiek „Airbnb“ šeimininkams, tiek ir viešbučių vadovams. Antra, apriboti šio tyrimo rezultatus gali ir demografinės respondentų charakteristikos, kadangi daugiau nei pusė respondentų šiame tyrime buvo nuo 21 iki 30 metų, o būtent tokia jauna imtis visos populiacijos negali pilnai atstovauti. Galiausiai, autoriai siūlo, remiantis šiuo tyrimu palyginti klientų „Airbnb“ ir viešbučių patirtį. Kadangi vis daugiau klientų renkasi „Airbnb“, o ne tradicinius viešbučius, rezultatai ypač turėtų dominti viešbučių vadovus. Būsimi tyrimai taip pat galėtų įvertinti veiksnius, kurie galėtų paveikti „Airbnb“ klientų patirtį.

Autoriai M. Cheng ir X. Jin [4] savo tyrime tyrė atributus, turinčius įtakos „Airbnb“ vartotojų patirčiai. Tyrime buvo analizuojamos klientų internetines apžvalgos - klientų komentarai. Duomenų rinkinys buvo gautas iš „Inside Airbnb“ svetainės, iš viso buvo gauti 181 263 klientų internetinių apžvalgų komentarai. Tiriamos „Airbnb“ apgyvendinimo paslaugos Sidnėjaus mieste. Duomenų analizei buvo tokie didžiųjų duomenų analizės metodai kaip teksto tyryba, regresinė analizė bei sentimentų analizė.

Gauti tyrimo rezultatai:

1. „Airbnb“ vartotojai yra linkę vertinti savo patirtį remdamiesi ankstesnių viešnagių viešbutyje gairėmis.
2. Trys pagrindiniai duomenyse nustatyti atributai yra „vieta“, „patogumai“ ir „priegloba“. Keista, bet „kaina“ nėra įvardijama kaip pagrindinis atributas.
3. Šis tyrimas parodė, kad „Airbnb“ vartorojai mielai skaito šeimininkų pateiktas instrukcijas, o želgiant į viešbučių kontekstą, retai kada pastebima, kad viešbučiai turėtų, ar kad jiems reikėtų išankstinių nurodymų.
4. Taip pat pastebima, kad geras bendravimas vaidina svarbų vaidmenį formuojant pradinį vartotojų pasitikėjimą.

5. Analizė parodo teigiamą šališkumą „Airbnb“ vartotojų komentaruose, o neigiamas nuotaikas dažniausiai sukelia „triukšmas“.

Šis tyrimas taip pat atveria daugybę galimybių ateities tyrimams. Pirmiausia, naudojant regresinę analizę įtraukiant kitus kintamuosius, tokius kaip reitingas bei savybių aprašymai, būtų galima pagerinti šio tyrimo išvadas, taip pat suteikti daugiau įžvalgų. Antra, atliekant duomenų tvarkymo ir valymo procesą buvo pastebėta, kad tikrai nemaža dalis vartotojų savo komentarus paskelbė ir gimtąja, ir anglų kalba. Taip pat, nors tyrimas nustatė daugiau mikro skirtumų tarp „Airbnb“ ir viešbučių, šiame etape nėra gauta pakankamai įrodymų, kad būtų galima pasiūlyti, kokiu tiksliai mastu galima apibendrinti išvadas Sidnėjaus kontekste.

C.V. Priporas, N. Stylos, L. N. Vedanthachari ir P. Santiwatana savo straipsnyje [5] nagrinėjo „Airbnb“ paslaugų kokybę, klientų pasitenkinimą ir jų lojalumą. Tyrimas vyko klausimyno pagrindu. Klausimynas buvo išplatintas 202 tarptautiniams turistams Pukete, Tailande. Šis miestas buvo pasirinktas todėl, nes tai yra yra viena iš populiariausių turistinių vietų visame pasaulyje. Tyrimui buvo naudota faktorinė analizė ir trajektorijos analizė.

1. Egzistuoja teigiamas paslaugų kokybės, klientų pasitenkinimo ir lojalumo ryšys ir taip pat, pasitenkinimas iš dalies priklauso nuo paslaugos kokybės ir lojalumo.
2. Paslaugų kokybės bei pasitenkinimo įtaka vartotojų lojalumui gali nepriklausyti nuo konkretaus nakvynės tipo.

Autoriai taip pat turi ir keletą pastebėjimų būsimiems ateities tyrimams. „Airbnb“ galima vadinti didžiausiu tinklu, nuomojančiu privačią nuosavybę turistams, tačiau tai nėra vienintelis nuomos tinklas. Be to, ateities tyrimai galėtų apimti apgyvendinimą, kurį reklamuoja įvairūs socialiniai tinklai. Tolimesni tyrimai taip pat galėtų būti atliekami ir su kitų tipų „Airbnb“ apgyvendinimu, pavyzdžiui, svetingumu vietoje, kadangi būtent šis tyrimas buvo sutelktas tik į nuotolinį svetingumą. Taip pat, autoriai siūlo būsimiems paslaugų kokybės tyrimams svetingumo industrijoje, palyginti svečių apgyvendinimo „Airbnb“ patirtį su viešbučių patirtimi.

Siekiant geriau suprasti vartotojų elgesio ypatybes dalijimosi ekonomikoje, Iis P. Tussyadiah atliko tyrimą [6], kuriame buvo nagrinėjami įvairūs veiksniai, darantys įtaką svečių pasitenkinimui apgyvendinimo įstaigose naudojant P2P apgyvendinimo modelį (angl. peer-to-peer), bei darančius įtaką tolimesniems klientų ketinimams naudotis P2P apgyvendinimo paslaugomis. P2P yra toks tinklo modelis, kuriame tiesiogiai tarp vartotojų vyksta keitimasis turimais resursais. Duomenys analizei surinkti remiantis internetine apklausa, kurioje dalyvavo 644 keliautojai, gyvenantys JAV. Tyrime buvo taikoma faktorinė analizė.

Gauti tyrimo rezultatai:

1. Svečių pasitenkinimą įtakoja tokie malonumo veiksniai, kaip piniginė nauda (vertė) bei apgyvendinimo patogumai.
2. Būsimą klientų ketinimą naudotis P2P apgyvendinimo paslaugomis taip pat nulėmė malonumas ir vertė.
3. Taip pat buvo pastebėta, kad tvarumas neigiamai veikia svečių, kurie nuomojasi privačius kambarius, pasitenkinimą.

Kadangi šiame tyrime buvo tiriami tik tie keliautojai, kurie gyvena JAV, būsimi tyrimai turėtų pakartoti tyrimą skirtingose šalyse, tam kad būtų galima ir toliau išbandyti sukurtą modelį. Atsižvelgiant į tai, kad P2P apgyvendinimo platformos pirmą kartą buvo pristatytos būtent JAV, galima teigti, kad šio tyrimo išvados atspindi rinkos modelio elgesį augimo etape. Būsimiems tyrimams yra būtina ištirti galimus skirtingose vietose gyvenančių P2P svečių skirtumus atsižvelgiant ir į tvarumo vaidmenį formuojant jų pasitenkinimą bei tolimesnius ketinimus. Būsimi tyrimai taip pat turėtų išnagrinėti viešbučių patirties veiksnius, kurie gali paskatinti vartotojus palankiau žiūrėti į P2P apgyvendinimą.

Taigi, išnagrinėjus mokslinius straipsnius „Airbnb“ paslaugų vartotojų pasitenkinimo vertinimo tematika buvo pastebėta, kad viešbučių klientų pasitenkinimo vertinimas mokslinėje literatūroje yra tiriamas gana dažnai, tačiau privataus „Airbnb“ sektoriaus vartotojų pasitenkinimas buvo tirtas ganėtinai retai ir mokslinių tyrimų šią tematika yra labai mažai. Visgi galima pastebėti, kad dažniausiai analizei naudotos programinės priemonės buvo R, SAS ir SPSS. Dažniausiai klientų pasitenkinimas buvo tiriamas iš internetinių atsiliepimų, tačiau buvo ir tokių tyrimų, kurie pasinaudojo anketomis ir klausimynais. Didžiųjų duomenų kontekste daugiausia tyrimuose buvo naudojama teksto tyryba bei sentimentų analizė, taip pat regresinė analizė, tačiau anketiniams duomenims ir klausimynams analizuoti buvo naudojama ir faktorinė analizė ir trajektorijų analizė. Nagrinėjant būtent žaliųjų vartotojų internetinius atsiliepimus, išrenkami tik tie atsiliepimai, kuriuose naudojami žaliesiems vartotojams būdingi terminai. Žalieji vartotojai yra tie, kurie vartoja žodžius, susijusius su tvariu gyvenimo būdu. Atsižvelgiant į literatūros apžvalgą būvo pastebėta, kad dažniausiai žalieji vartotojai savo atsiliepimuose vartoja tokius žodžius kaip “tvaru” , “tvarumas” bei “ekologiškas” Panaudojant teksto tyrybą ir sentimentų analizę buvo ištirti atributai, darantys įtaką žaliųjų “Airbnb” vartotojų pasitenkinimui, tokie kaip švara, namų atmosfera bei namų patogumai. Šie atributai vartotojus įtakojo teigiamai, tačiau pasitaikė ir tokių atributų, kurie vartotojus veikia neigiamai, t.y. triukšmas, šaltis bei nemalonus kvapas. Buvo pateikta ir tam tikrų rekomendacijų ateities tyrimams, tokių kaip labiau išplėtoti regresinę analizę, pasirenkant daugiau kintamųjų, taip pat išplėsti tyrimo geografinę teritoriją, kuri apimtų ne tik vieną miestą, taip pat svarbu išplėsti tyrimą ir skirtingose kultūrose. Buvo pasiūlyta į tyrimą įtraukti daugiau informacijos, kuri būtų gauta iš aprašymų, kuriuos pateikia nuosavybės šeimininkai.

1.6. Tyrimų metodų ir programinės įrangos apžvalga

Šiame poskyryje bus išanalizuoti dažniausiai tyrimuose naudoti duomenų analizės metodai bei naudojama programinė įranga. Išanalizavus mokslinę literatūrą apie klientų pasitenkinimą naudojantis „Airbnb“ paslaugomis, tinkamiausi metodai duomenų analizei yra teksto analizė bei sentimentų analizė, regresinė analizė bei faktorinė analizė, taip pat atliekamos įvairios duomenų grafinės vizualizacijos. Pastebėta, kad populiariausios programinės įrangos minėtiems didžiųjų duomenų analizės metodams yra R, SAS, SPSS.

1.6.1. Faktorinė analizė

Faktorinė analizė atsirado 1900-ųjų pradžioje, kai C. Spearman domėjosi žmogaus sugebėjimais bei plėtojo dviejų faktorių teoriją. Tai galiausiai ir paskatino teorijų bei matematinių veiksnių analizės principų kūrimą [30]. Faktorinė analizė - stebimų kintamųjų suskirstymas į įvairias grupes, kurias kartu suvienytų tam tikras faktorius, kuris nebūtų tiesiogiai stebimas, taip pat atsižvelgiant į kintamųjų tarpusavio koreliaciją. Pereinant nuo daug kintamųjų prie tam tikrų faktorių, informacija tampa daug

labiau koncentruota, taip ji tampa labiau aprėpiama. Faktoriai dažniausiai neturi jokio kiekybinio mato ir jie negali būti išmatuoti, bet juos galima įsivaizduoti kaip tam tikras požymių grupes, kurios vienytų įvairias kategorijas [27].

Faktorinė analizė turi tokią prielaidą, jog visi stebimi atsitiktiniai dydžiai, priklausomi nuo mažesnio kiekio kintamųjų, kurie yra nestebimi. Pats faktorinės analizės modelis apibrėžia tiesinę priklausomybę tarp faktorių bei tam tikrų stebimų kintamųjų. Faktorinės analizės pagalba galima sumažinti duomenų dimensiją, išskiriant mažesnį kiekį faktorių, darančių įtaką stebimiems kintamiesiems [29]. Faktorinė analizė plačiai naudojama daugelyje sričių, tokių kaip socialiniai mokslai, ekonomika, medicina bei geografija, dėl didelės informacinių technologijų pažangos [30].

Faktorinė analizė remiasi tokiomis pačiomis prielaidomis kaip ir daugialypė tiesinė regresinė analizė, kadangi ji priklauso bendrojo tiesinio modelio kategorijai. Pagrindinės taikomos prielaidos - kintamųjų tiesinė priklausomybė, duomenys yra intervaliniai arba jiems artimi, kitamųjų parinkimas yra tinkamas, nėra kintamųjų multikolinearumo [27].

Faktorine analize siekiama [27]:

- Sumažinti didelį kintamųjų skaičių, pereinant prie kiek įmanoma mažesnio bendrų faktorių kiekio.
- Pašalinti skalės sudedamąsias dalis, patenkančias į kelis faltorius ir taip pat patvirtinti skalę, kuri yra naudojama bei parodyti, jog skalės sudedamosios dalys priklauso tam pačiam faktoriui.
- Sudaryti tarpusavyje nekoreliuotus, dar kitaip vadinamus ortogonalius faktorius, kurie toliau būtų naudojami regresinėje analizėje, tam kad būtų išvengta multikolinearumo.

Lengviau yra sutelkti dėmesį į tam tikrus pagrindinius faktorius, o ne atsižvelgti į per didelį kiekį kintamųjų, kurie gali būti nereikšmingi, todėl faktorinė analizė yra naudinga, norint kintamuosius suskirstyti į prasmingas kategorijas. Galima rasti daugelį kitų faktorinės analizės panaudojimo būdų, tokių kaip duomenų transformavimas, hipotezių tikrinimas, žemėlapių sudarymas bei mastelio keitimas [30].

Išskiriami keturi faktorinės analizės etapai [27]:

1. Pirmiausia patikrinamas duomenų tinkamumas faktorinei analizei.
2. Išskiriami faktoriai, nustatomas jų skaičius.
3. Faktoriai sukami ir interpretuojami.
4. Skaičiuojami faktorių reikšmių įverčiai.

Faktorinė analizė skaičiuodama naudoja matricinę algebrą. Pagrindinė faktorinės analizės statistika yra koreliacijos koeficientas, kuris nustato priklausomybę tarp dviejų kintamųjų. Faktorinė analizė nėra atliekama, kol nėra apskaičiuota kiekviena įmanoma koreliacija tarp kintamųjų. Yra ištiriama, ar kintamieji turi kokių nors bendrų bruožų bei tuomet apskaičiuojama koreliacija arba kovariacijos matrica. Paprastai faktorinė analizė yra atliekama panaudojant koreliacijų matricą, kuri pateikia standartizuotus duomenis, būtent todėl faktorinė analizė yra rekomenduojama kintamiesiems, kurie nėra prasmingai palyginami [30].

Faktorinės analizės modelis, kuris sieja k kintamųjų $X_1, X_2, \dots, X_k$ su m bendrųjų latentinių, kitaip dar vadinamų nepastebėtų, nepastebimų faktorių $F_1, F_2, \dots, F_m$ ir tam tikru charakteringuoju latentiniu faktoriumi e_i yra aprašomas tokia lygčių sistema [27]:

$$Xi = \sum_{j=1}^{m} \lambda_{ij} F_j + e_i \tag{1}$$

čia $i = 1, \dots, k$, $m < k$, tai reiškia, kad bendrųjų faktorių turime mažiau nei kintamųjų. Daugikliai žymimi λ_{ij} ir jie yra vadinami faktorių svoriais. Daromos prielaidos, jog:

1. kintamieji, kurie yra stebimi, yra pasiskirstę pagal normalųjį skirstinį, tai yra $X_i \sim N(\mu_i, \sigma_i^2)$,
2. bendrieji faktoriai nėra koreliuoti bei jų dispersija $DF_j = 1$,
3. charakteringieji faktoriai e_i nėra koreliuoti ir jų dispersija $De_j = \tau_i$,
4. bendrieji faktoriai F_j ir charakteringieji faktoriai e_i nėra koreliuoti, čia $i = 1, \dots, k$, $j = 1, \dots, m$

kintamųjų, kurie yra stebimi, dispersijos užrašomos:

$$DX_i = \sigma_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{im}^2 + \tau_i = h_i^2 + \tau_i. \tag{2}$$

Dydis $h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ yra kintamojo X_i bendrumas, o dydis τ_i yra specifiškumas. Pagal tai kuo didesnis yra h_i^2, palyginus su σ_i^2, tuo daugiau informacijos apie kinamąjį galima išsaugoti, pereinant nuo pačių pradinių kintamųjų prie bendrų faktorių. Faktorinės analizės uždavinys yra žinant kintamųjų X_i reikšmes, padaryti išvadas apie bendruosius faktorius, kurie salygotų kintamųjų elgseną, tai reiškia nustatyti faktorių svorius λ_{ij}, specifinių dispersijų τ_i bei taip pat bendrųjų faktorių $F_1, F_2, \dots, F_m$ įverčių reikšmes [27].

1.6.2. Regresinė analizė

Regresinė analizė yra toks statistinis procesas, kuris sugeba įvertinti skirtumus bei ryšius tarp kintamųjų, bei tų ryšių stiprumą. Regresinės analizės pagalba yra įvertinami kintamieji, kurie yra nepriklausomi ir iš jų sudaroma regresinė lygtis, kuri yra aprašoma [28]:

$$Y = C + b_1 X + b_2 Z + b_3 W + e \tag{3}$$

čia e – liekamoji paklaida, C - konstanta, b_1, b_2, b_3 – lygties koeficientai.

Įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ yra gaunami panaudojus imties duomenis. Regresijos lygtis Y reikšmei, kuri yra apytikslė, aprašoma tokia lygtimi:

$$\hat{Y} = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W \tag{4}$$

Ši regresinė lygtis yra naudojama kokybinei bei kiekybinei kintamųjų priklausomybių analizei [28]. Jeigu turime duomenis, kurie yra tinkami regresinei analizei atlikti, tai dar nereiškia, kad būtinai pavyks sudaryti modelį, kuris būtų tikrai tinkamas. Regresinio modelio tinkamumą parodo tokie rodikliai kaip [28]:

– Apibrėžtumo koeficientas (R^2) yra modeliuojamų ir taip pat stebimų priklausomojo kintamojo reikšmių skirtumų matas, kuris yra labai svarbi regresijos modelio charakteristika bei yra privaloma visuose tyrymų aprašymuose. Šio koeficiento įgyjamos reikšmės yra intervale [0,1]. Kuo apibrėžtumo koeficientas didesnis, tuo modelis yra tinkamesnis.
– Koreguotasis apibrėžtumo koeficientas yra tam tikra apibrėžtumo koeficiento alternatyva, kai modelyje yra nedaug stebinių, tačiau yra daug regresorių. Šis koeficientas yra apskaičiuojamas pagal formulę:

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{p-1}{n-p} \quad (5)$$

čia $\bar{R}^2 -$ koreguotasis apibrėžtumo koeficientas, $R^2 -$ apibrėžtumo koeficientas, $n -$ tiriamos imties dydis, $p -$aiškinamieji kintamieji [28].

– ANOVA *p* reikšmė yra tam tikras dydis, kuris parodo, ar modelyje yra regresorių,susijusių su priklausomu kintamuoju. Jei *p* reikšmė didesnė už skaičių 0,05, tai modelio tinkamumas yra labai abejotinas. Jeigu $p < 0{,}05$, tai galima teigti, kad modelis yra visai geras ir galima tirti toliau.
– T (Stjudento) kriterijai atskiriems regresoriams padeda nuspręsti, ar tam tikrą regresorių galima pašalinti iš modelio. Jeigu kriterijaus p reikšmė yra mažiau už 0,05 reikšmingumo lygmenį, tai regresorių galima vadinti statistiškai reikšmingu, jeigu $p > 0{,}05$, tai regresorius nėra statistiškai reikšmingas [31].

Standartinės regresijos paklaidos didumas priklauso būtent nuo *Y* reikšmių didumo. Ji yra naudojama, kai norima palyginti tarpusavyje keletą regresijos modelių, kurie būtų skirti to paties *Y* prognozavimui [32].

Daugiamatės daugialypės tiesinės regresijos pagrindu galima vadinti daugialypę tiesinę regresiją.

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (6)$$

čia β_0 populiacijos Y -atkarpa, $\beta_1, \beta_2, \ldots, \beta_k$ - populiacijos nuolydžiai, Y - priklausomasis (atsako) kintamasis, $X_1, X_2, \ldots, X_k$ - nepriklausomieji (aiškinamieji) kintamieji, ε – atsitiktinė paklaida [33].

Yra išskiriami tokie pagrindiniai daugialypės tiesinės regresinės analizės uždaviniai:

– Regresijos funkcijos koeficientų intervalinių bei taškinių įverčių radimas.
– Hipotezių apie regresijos funkcijos koeficientus tikrinimas.
– Optimaliosios regresijos lygties sudarymas.
– Modelio prielaidų tikrinimas.
– Prognozavimo paklaidų įvertinimas.

Išskiriami tam tikri reikalavimai, kuriuos turi atitikti atsitiktinės paklaidos ε_i , parodančios kiek stebėtoji *Y* reikšmė skiriasi nuo tos reikšmės, kuri yra gaunama prognozuojant pagal regresijos lygtį. Pagrindines tiesinės regresinės analizės prielaidos būtų šios [33]:

– Atsitiktinės paklaidos ε_i yra atsitiktiniai dydžiai, kurie yra pasiskirstę pagal normalųjį skirstinį.
– Visų ε_i vidurkiai lygūs nuliui, $E\varepsilon_i = 0$.

- Homoskedastiškumo prielaida. Visų ε_i dispersijos yra lygios, $D\varepsilon_3 = \sigma^2$.
- Visi ε_i yra nepriklausomi.
- Duomenyse nėra jokių išskirčių.
- Nėra stiprios koreliacijos tarp nepriklausomų kintamųjų $X_1, X_2, \dots, X_k$.

Visgi negalima pačios regresijos prilyginti priežastingumui. Taikant regresinę analizę, ne visada vieno kintamojo pakitimai gali sąlygoti kito kintamojo pakitimus. Kartais būna taip, kad taikant regresinę analizę, galima rasti kažkokį ryšį tarp tam tikrų kintamųjų, kurie neturi tarpusavyje nieko bendro [33].

1.6.3. Duomenų vizualizavimas

Didžiųjų duomenų vizualizavimas suteikia galimybę tyrėjui stebėti pagal kokia tendencijas duomenys grupuojasi, taip pat padeda atskirų taškų tarpusavio artimumo įvertinimui, galima racionaliai priimti tam tikrus sprendimus. Norint gauti kiek įmanoma daugiau naujos informacijos apie analizuojamus duomenis, kartais bandoma sujungti keletą skirtingais principais pagrindžiamus tam tikrus vizualizavimo metodus [34].

Pagrindinė vizualizavimo idėja yra ta, kad norima duomenis pateikti būtent tokia forma, kuri leistų tyrėjui juos lengviau suprasti, padėtų daryti tam tikras išvadas. Vizualizavimas padeda lengviau suvokti sudėtingesnes duomenų aibes, taip pat padeda nustatyti dominančius jų poaibius [34].

Vizualizavimo metodai padalinti į dvi kryptis:

1. Tiesioginio vizualizavimo metodai. Naudojantis šiais metodais, kiekvienas daugiamačio objekto parametras pateikiamas įvairiomis vizualiomis formomis. Šie metodai yra skirstomi į geometrinius arba simbolinius, taip pat gali būti ir hierarchinio vizualizavimo metodai.
2. Projekcijos metodai, dar kitaip vadinami matmenų skaičiaus mažinimo (angl. *dimension reduction*) metodais. Projekcijos metodais transformavus daugiamačius duomenis į dvimatę arba trimatę vaizdo erdvę bei atlikus vizualizavimą, galima paprasčiau suvokti įvairių duomenų struktūrą bei tam tikrus sąryšius tarp jų. Transformuojant duomenis į tam tikrą mažesnio skaičiaus matmenų erdvę, dažnai pasitaiko duomenų iškraipymai bei įvairios paklaidos. Išskiriami tiesinės bei netiesinės projekcijos metodai.

Pagrindiniai duomenų vizualizavimo tikslai, kurie yra išskiriamti - tiriamoji analizė (angl. *explorative analysis*), kai atlikus duomenų vizualizavimą iškeliamos hipotezės, bei patvirtinančioji analizė (angl. *confirmative analysis*), kai po duomenų vizualizavimo yra patvirtinamos jau esančios hipotezės [34].

1.6.4. Tyrimuose naudojama programinė įranga bei jos apžvalga

Šiame poskyryje bus apžvelgiama tyrimuose naudotos programinės įrangos. Pagal atliktą literatūros analizę galima teigti, kad daugiausia naudojamos programinės įrangos, norint išanalizuoti „Airbnb“ apgyventinimo klientų pasitenkinimą yra R, SAS, Python bei SPSS.

Programinė įranga SAS:

Programinė įranga SAS supaprastina prieigą prie duomenų bei supaprastina jų paruošimą, pagerina duomenų kokybę bei užtikrina optimalius procesus, tam kad padidėtų atliekamos analizės projektų našumas. Naudojant SAS ar kitus atviruosius šaltinius, tokius kaip „Python“, „R“ ar „Lua“, galima

kurti naujoves neprarandant energijos, greičio bei svarbiausia - saugumo. Vieninga kodų bazė yra taupanti laiką bei pastangas [35].

SAS teksto analizė naudoja natūralios kalbos apdorojimą bei pažangias kalbines technikas, tam kad automatiškai išanalizuotų didelės apimties turinį. Analizuojant tekstą yra sukuriami tam tirki metaduomenys, lingvistiniai modeliai bei sąvokų apibrėžimai. Juos galima pritaikyti automatiškai dideliems dokumentų rinkiniams, jeigu norima greitai bei tiksliai klasifikuoti, aptikti temą, taip pat įvertinti nuotaikas bei semantinę įžvalgą. Būtent tai padeda palengvinti įvairių dokumentų klasifikavimą bei atrasti kiek įmanoma aiškesnes sąsajas tarp terminų bei dokumentų, taip pat padeda grupuoti dokumentus į įvairias kategorijas bei atrasti kalbines taisykles, kurios leidžia pasidaryti įvairių įžvalgų [23].

Programinė įranga R:

R yra programavimo kalba bei tuo pačiu yra ir aplinka statistiniams skaičiavimams bei grafikai atlikti. R teikia gana platų statistinių, tokių kaip tiesinio bei netiesinio modeliavimo, laiko eilučių analizės, klasifikavimo, grupavimo ir kt., bei grafinių metodų didelę įvairovę. Viena iš stipriausių R pusių yra tai, jog gana lengvai yra sukuriami gerai suprojektuoti matematiniai simboliai bei formulės. Tikrai labai rūpestingai yra atsižvelgiama į įvairius grafikos dizaino pasirinkimus, taip pat vartotojui leidžiama išlaikyti visišką kontrolę [36].

R yra integruotas programinės įrangos paketas į kurį įeina:

- veiksminga duomenų tvarkymo bei jų saugojimo įranga,
- operatorių rinkinys matricų skaičiavimams,
- didelis, nuoseklus bei integruotas tarpinių duomenų analizės įrankių rinkinys,
- grafinės priemonės duomenų analizei bei atvaizdavimui,
- puikiai išvystyta, paprasta bei labai efektyvi programavimo kalba, kuri apima sąlyginius elementus, vartotojo nustatytas rekursines funkcijas bei įvairias įvesties ir išvesties galimybes.

Programinė įranga SPSS:

Programinis paketas SPSS (angl. *Statistical Package for the Social Sciences*) yra vienas iš labiausiai paplitusių statistinės informacijos apdorojimo paketų, kuris yra tinkamas ir pradedančiajam, ir taip pat yra tinkamas ir patyrusiam vartotojui. SPSS pasižymi dideliu statistinių analizės metodų pasirinkimu, taip pat duomenų analizės rezultatų vizualizavimo priemonėmis bei jų įvairove. SPSS programinis paketas pasižymi lengvai įvaldoma dialogine sąsaja. Šis paketas yra plačiai taikomas sociologijoje, biologijoje, psichologijoje, rinkodaroje, medicinoje ir taip pat įvairiuose kokybės valdymo procesuose [27].

Plačiausiai socialiniuose tyrimuose taikomi vienmačiai bei daugiamačiai statistikos metodai ir modeliai yra požymių dažnių analizė, faktorinė analizė, binarinė logistinė regresija, klausimynų patikimumo analizė bei taip pat sprendimų medžių sudarymas [27].

SPSS leidžia vartotojams atlikti visus duomenų analizės procesus, tokius kaip [37]:

- įkelti duomenis pasinaudojant įvairiais šaltiniais,
- paruošti duomenis analizei, t.y. atlikti reikalingas transformacijas, jeigu reikia sukurti naujus kintamuosius, duomenis apjungti ir t.t.,

- pasinaudojant statistiniais metodais išanalizuoti duomenis bei gauti reikšmingus rezultatus,
- gautus rezultatus pateikti grafiškai bei įvairiomis lentelėmis,
- gautus rezultatus eksportuoti į įvairius formatus.

Programavimo kalba Python:

Programavimo kalba Python yra paprastai interpretuojama ir orientuota į tikslą programavimo kalba. Ji pasižymi gana paprasta bei paprastai išmokstama sintakse, kuri sumažina programų palaikymo kaštus. Dėl pakankamai lengvos sintaksės, nesunku pradėti naudotis programavimo kalba Python tiek pradedančiajam programuotojui, tiek pažengusiam. Python oficialioje svetainėje galima rasti daug įvairių naudojimosi vadovų, kurie yra skirti mokymuisi. Python bendruomenė nuolat rengia įdomius susitikimus bei konferencijas, taip pat aktyviai bendradarbiauja patobulinant programinį kodą. Visą reikiamą informaciją galima rasti Python dokumentacijoje. Python sukurta pagal „OSI“, ji turi laisvai prieinamo šaltinio licenciją, kuri leidžia šia kalba nemokamai naudotis bei platinti net ir komerciniais tikslais. Python licencija administruojama „Python Software Foundation“.

1.7. Baigiamojo projekto temos ir uždavinių pagrindimas

Išnagrinėjus mokslinę literatūrą „Airbnb“ paslaugų vartotojų pasitenkinimo vertinimo tematika buvo pastebėta, kad viešbučių klientų pasitenkinimo vertinimas mokslinėje literatūroje yra tiriamas gana dažnai, tačiau „Airbnb“ vartotojų pasitenkinimas buvo tirtas ganėtinai retai ir mokslinių tyrimų šia tematika yra labai mažai. Tyrimų, kuriais būtų bandyta ištirti žaliųjų „Airbnb“ vartotojų pasitenkinimą, vis dar beveik nėra, ypač tų, kurie naudotųsi didžiųjų duomenų tyrybos metodais. Pagrindinis šio darbo tikslas yra išnagrinėti žaliųjų „Airbnb“ vartotojų pasitenkinimą, išanalizuojant internetines apžvalgas, kurios yra paskelbtos „Inside Airbnb“, naudojant teksto tyrybos ir sentimentų analizės metodus. Šiam baigiamojo projekto tikslui pasiekti reikia išspręsti baigiamojo projekto uždavinius, kurie yra apibrėžti įžangoje.

2. Tyrimo metodai

Šioje darbo dalyje pateikiama darbo metodika, kuri buvo siūloma „Airbnb“ paslaugų vartotojų pasitenkinimo vertinimui nustatyti. Ši metodika apima duomenų paruošimą tolimesnei analizei, teksto tyrybą bei sentimentų analizę. Ši metodika yra realizuota panaudojant Python programavimo kalbą bei programinę įrangą R.

2.1. Teksto analitika

Teksto analitika yra gana dažnai naudojama daugelyje pramonės šakų. Ji yra laikoma laisvai integruotu įrankiu bei tam tikrų metodų rinkiniu, kuris yra sukurtas gauti, išvalyti, išskleisti, tvarkyti, analizuoti bei interpretuoti informaciją gautą iš daugybės duomenų šaltinių. Pastaruoju metu teksto analitikos metodai buvo labai dažnai naudojama atrandant tam tikras tendencijas tekstiniuose duomenyse. Pavyzdžiui, teksto analitika, panaudojant socialinių tinklų duomenis, buvo naudojama nusikaltimų prevencijai bei sukčiavimui nustatyti ir t.t. Teksto analitikos programos taip pat yra labai populiarios verslo aplinkoje. Šios programos duoda pačių inovatyviausių rezultatų ir dar gilesnių įžvalgų.

Yra išskiriamos septynios teksto analizės praktikos sritys [23]:

1. Paieškos bei informacijos gavyba (angl. *Search and information retrieval*) - tekstinių dokumentų, įskaitant raktinių žodžių paiešką bei paieškos sistemas, gavimas ir saugojimas.
2. Dokumentų klasterizavimas (angl. *Document clustering*) - terminų, fragmentų, pastraipų ar dokumentų grupavimas ir jų kategorizavimas, panaudojant įvairius duomenų tyrybos grupavimo metodus.
3. Dokumentų klasifikavimas (angl. *Document classification*) - fragmentų, pastraipų ar dokumentų grupavimas ir jų kategorizavimas, panaudojant duomenų tyrybos klasifikavimo metodus, remiantis modeliais su apmokytais pavyzdžiais.
4. Žiniatinklio tyryba (angl. *Web mining*) - internetinė duomenų ir teksto tyryba, ypatingą dėmesį skiriant žiniatinklio dydžiui ir tarpusavio ryšiui.
5. Informacijos gavimas (angl. *Information extraction*) - atitinkamų faktų ir ryšių identifikavimas ir išskyrimas iš nestruktūrizuoto teksto.
6. Natūralios kalbos apdorojimas (angl. *Natural Language Processing*) - kalbos aptikimas tekste.
7. Koncepcijos išgavimas (angl. *Concept extraction*) - žodžių bei frazių grupavimas į semantiškai panašias grupes.

2.1.1. Teksto tyryba

Teksto tyryba yra vienas iš teksto analizės pogrupių, orientuotas į duomenų tyrybos metodų taikymą tekstinės informacijos srityje [23]. Teksto tyrybos pagalba automatiškai išanalizuojamas tekstinių dokumentų korpusas ir taip atrandama anksčiau paslėpta informacija. Teksto tyryba taip pat dar gali būti naudojama norint nustatyti pagrindinę teksto temą.

Teksto tyryba turi daug privalumų [26]:

1. Padeda greitai bei labai efektyviai iš daugybės duomenų išgauti reikiamą bei naudingą informaciją.
2. Padeda puikiai numatyti tam tikrus ateities aspektus, remiantis pateiktais stebėjimais bei statistika.

3. Padeda kurti modelius iš pateiktų duomenų, kurie parodo tendencijų didėjimą arba mažėjimą. Puikiai taikoma versle bei ekonomikoje.
4. Teksto tyrybos programinė įranga padeda saugumo agentūroms, stebint bei analizuojant tekstinius duomenis, surinktus iš interneto šaltinių tinklaraščių ir kt.

Teksto tyryba dar yra skirstoma į dvi grupes, tokias kaip tiriamoji analizė bei sentimentų analizė. Tiriamoji analizė apima tokias tyrimų sritis kaip temos išskyrimas, klasterinė analizė bei kt. Sentimentų analizė gali būti traktuojama kaip klasifikavimo analizė. Sentimentų analizė yra naudinga norint identifikuoti nuotaiką visame dokumente. Nuotaka gali būti teigiama arba neigiama, taip pat gali būti neutrali arba neklasifikuojama [23].

Teksto tyryba yra vienas iš teksto analizės pogrupių, orientuota į duomenų tyrybos metodų taikymą tekstinės informacijos srityje [23]. Norint išanalizuoti klientų internetinius atsiliepimus yra taikomi teksto tyrybos metodai. Šios analizės tikslas yra sudaryti žaliųjų vartotojų pasitenkinimą atspindinčių žodžių sąrašą, kuris bus naudojamas kitame tyrimo etape norint ištirti žaliųjų vartotojų pasitenkinimui turinčius įtakos atributus. Šiame darbe yra analizuojami tik tie klientų atsiliepimai, kurie yra parašyti anglų kalba.

Duomenų tyryba apima tokius tyrybos metodus kaip klasterizavimas, klasifikavimas ir t.t. Kadangi teksto tyryba yra iteracinis procesas, jo metu analizė yra nuolat kartojama panaudojant skirtingus nustatymus bei įtraukiant arba neįtraukiant tam tikrus terminus, tam kad būtų pasiekiami geresni rezultatai. Šio žingsnio rezultatas yra dokumentų grupės (klasteriai), temų sąrašai arba taisyklės, kurios atsako į klasifikavimo problemą.

Teksto tyryba susideda iš tokių žingsnių [23]:

1. Duomenų rinkimas (angl. *Data Collection*). Pirmasis bet kokio teksto tyrybos žingsnis yra surinkti analizei reikalingus tekstinius duomenis.
2. Teksto gramatinis nagrinėjimas ir transformavimas.
3. Teksto filtravimas.
4. Duomenų tyryba.

Teksto tyryba yra labai plačiai naudojama. Ji gali gali būti naudojama ieškant informacijos, stebint temą, apibendrinant, suskirstant kategorijas, kaupiant grupes, vizualizuojant informaciją bei atsakant į įvairius klausimus. Teksto tyryba paprastai apima informacijos arba teksto skirstymą į tam tikras kategorijas, teksto klasterizavimą, subjekto arba sąvokos išskyrimą, klasifikavimo sukūrimą bei formulavimą.

Teksto tyryba reikalinga norint konvertuoti tekstą į duomenis, kurie vėliau būtų analizuojami naudojant kitas duomenų tyrybos technologijas. Teksto tyryba apima statistinius, lingvistinius bei mašininio mokymosi metodus, reikalingus tekstinei informacijai, kuri yra reikalinga tolimesnei duomenų analizei [26].

Tradiciniai mašininio mokymo ir statistiniai metodai, kurie yra skirti išmokti nežinomus teksto duomenų modelius, dabar keičiami žymiai pažangesniais metodais, apjungiančiais natūralios kalbos apdorojimą (NLP) bei pačią kalbotyrą [23].

Pagrindiniai teksto tyrybai naudojami programinės įrangos R paketai ir jų bibliotekos:

- Paketas ***tm***. Ji yra naudojama teksto tyrybos operacijoms, pvz., skaičių, specialiųjų simbolių, skyrybos ženklų ir nereikšminių žodžių pašalinimui.
- Paketas **fpc.** Įvairūs grupavimo ir grupių patvirtinimo metodai.
- Paketas ***wordcloud***. Tai grafikas, kuriame atvaizduojamas žodžių debesis.
- Paketas ***RcolorBrewer***. Ji naudojama spalvų paletėms, kurios yra naudojamos įvairiuose grafikuose.
- Paketas ***syuzhet***. Ji reikalinga sentimentų balams išgauti ir emocijų klasifikavimui.
- Paketas ***ggplot2***. Jis naudojamas grafikams braižyti.
- Paketas ***NLP***. Jis naudojamas naturalios kalbos aptikimui.
- Paketas ***Pacman***. Jis yra R paketo valdymo įrankis, kuris sujungia bazinės bibliotekos funkcijų funkcionalumą į intuityviai pavadintas funkcijas.
- Paketas **tidyverse**. Apima paketus, kurie greičiausiai naudojami kasdienėje duomenų analizėje.
- Paketas **cluster**. Klasterinės analizės metodai.
- Paketas **ClusterR**. Siluetui nubraižyti, bei apskaičiuoti.
- Paketas **factoextra**. Teikia keletą lengvai naudojamų funkcijų, kad būtų galima išgauti ir vizualizuoti daugiamatės duomenų analizės rezultatus, įskaitant „PCA“ (pagrindinių komponenčių analizę) ir pan.
- Paketas **stringr**. Funkcijų rinkinys, sukurtas tam, kad kuo paprasčiau būtų dirbti su žodžiais.

2.1.2. Teksto gramatinis nagrinėjimas ir transformavimas

Surinkus analizei reikalingus tekstinius duomenis, sekantis žingsnis yra ištraukti, išvalyti bei sukurti žodžių žodyną iš dokumentų naudojant natūralios kalbos apdorojimą (angl. *Natural Language Processing*). Tai apima sakinių identifikavimą, kalbos dalių nustatymą taip pat žodžių kilmės nustatymą [23]. Natūralios kalbos apdorojimas apima kelių sričių, tokių kaip dirbtinio intelekto, matematikos bei informacijos mokslo metodus, tai yra būdas, kompiuteriui suprasti natūralią kalbą ir atlikti tam tikras užduotis [41]. Šis žingsnis taip pat apima ištrauktų žodžių analizavimą, nereikalingų žodžių pašalinimą (angl. *stop words*) bei rašybos tikrinimą [23].

Šiame gramatinio nagrinėjimo žingsnyje nustatoma vartotojo atsiliepimo kalba ir pašalinami įrašai, kurie nėra parašyti anglų kalba. Buvo panaudota kalbos atpažinimo biblioteka CLD (angl. Compact Language Detector), kuri yra sukurta Google inžinierių. Ši biblioteka naudoja neuroninių tinklų modelius bei aptinka daugiau nei 80 kalbų. CLD2 pateikia teksto vektorių su kiekvieno atsiliepimo atpažinta kalba. CLD3 yra neuroninio tinklo modelis, kuris veikia tokiu principu, kad norint nustatyti įvesties teksto kalbą, atliekamas perdavimas į priekį per tinklą.

Toliau atliekamas duomenų valymas, kuris prasideda nuo transformacijų. Visas tekstas verčiamas į mažąsias raides, tokiu būdu palengvinami tolimesni veiksmai. Šiam žingsniui atlikti yra naudojama R programinės įrangos funkcija *content_transformer*.

Kitas teksto apdorojimo žingsnis yra nereikalingų žodžių pašalinimas (angl. *stop words*). Nereikalingi žodžiai yra dar kitaip vadinami atmetiniais. Tai įvairūs funkciniai žodžiai, tokie kaip: jungtukai, jaustukai, kreipiniai, ištiktukai, klausiamieji žodeliai, įvardžiai, taip pat kai kurie veiksmažodžiai arba būdvardžiai bei kiti įvairūs žodžiai. Tai yra ypač paplitę žodžiai, kurie yra mažai naudingi, todėl tokie žodžiai yra pašalinami [23].

Siekiant dar labiau sumažinti unikalių tekstų skaičių yra naudojamas šaknies išskyrimo algoritmas, dar kitaip vadinamas žodžio kamieno išskyrimu, kurio metu yra nupjaunama žodžio galūnė. Skirtingos to paties žodžio formos paprastai yra problemiškos analizuojant teksto duomenis, nes jų rašyba ir reikšmė skiriasi. Žodžio kamieno išskyrimas yra terminas, kuris yra vartojamas kalbų morfologijoje, apibūdinantis linksniuotų (ar kartais išvestinių) žodžių pertvarkymą į jų žodžio kamieną. Žodžio kamienas neturi būti identiškas morfologinei žodžio šakniai. Paprastai pakanka, kad susiję žodžiai sisisietų su tuo pačiu kamienu, net jei šis kamienas savaime nėra galiojanti šaknis.

Anglų kalbai dažniausiai naudojamas Porter‘io metodas, būtent jis ir buvo naudojamas tyrime žodžių kamienų išskyrimui. Porter‘io kamieno išskyrimo metodas sukurtas Martin‘o Porter‘io Kembridžo universitete 1980 m. Šis metodas remiasi idėja, kad galūnės anglų kalboje (maždaug 1200) daugiausia sudaro mažesnių ir paprastesnių priesagų derinį. Kiekviename etape, jei priesagos taisyklė sutapo su žodžiais, tada prie šios taisyklės pridedamos sąlygos tikrinamos, koks būtų kamienas, jei ši priesaga būtų pašalinta, taisyklės apibrėžtu būdu. Pavyzdžiui, tokia situacija gali būti balsių simbolių skaičius, po kurio eina priebalsių simbolis, kurio kamiene turi būti daugiau nei vienas taikytinoms taisyklėms [42]. Porterio algoritmas yra svarbus, kadangi jame pateikiamas gana paprastas požiūris į dviejų ar daugiau informacijos, tekstų, idėjų ir kt. rinkinių sujungimą į vieną, kuris, atrodo veikia gerai ir yra pritaikomas įvairioms kalboms.

Pati svarbiausia užduotis po gramatinio nagrinėjimo yra teksto transformavimas. Šis žingsnis yra susijęs su skaitiniu teksto vaizdavimu taikant tiesinės algebros metodus. Šio uždavinio rezultatas yra dokumento terminų matrica (angl. *term-by-document matrix*). Matricos matmenis lemia dokumentų ir terminų skaičius. Šis žingsnis gali apimti kiekvieno dokumento matmenų matricos dimensijos mažinimą naudojant matricos skaidymo singuliariosiomis reikšmėmis metodą (angl. *Singular value decomposition(SVD)*) [23]. Dokumento terminų matrica yra matematinė matrica, kuri apibūdina terminų, atsirandančių dokumentų rinkinyje, dažnumą. Tai yra tokia matrica, kur kiekviena eilutė reiškia vieną dokumentą, o kiekvienas stulpelis reiškia vieną terminą (žodį). Kiekviena vertė matricoje nurodo to termino pasirodymų tame dokumente skaičių (dažnumą).

Tiesinėje algebroje SVD yra vadinama faktorizacija:

$$A = U\Sigma V^T \tag{7}$$

kur A – pradinė matrica, U ir V – unitariosios matricos, Σ diagonalinė matrica, kurios elementai yra neneigiami.

Paprasčiausiu matricos, kurios matmenys 2×2 atveju, matrica A tenkina sąlygą:

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = \begin{vmatrix} u_{1,1} & u_{1,2} \\ u_{2,1} & u_{2,2} \end{vmatrix} \begin{vmatrix} \sigma_{1,1} & 0 \\ 0 & \sigma_{2,2} \end{vmatrix} \begin{vmatrix} v_{1,1} & v_{1,2} \\ v_{2,1} & v_{2,2} \end{vmatrix} \tag{8}$$

Kai,

$$UU^T = VV^T = I, \quad \sigma_{1,1}, \sigma_{2,2} \geq 0 \tag{9}$$

2.1.3. Teksto filtravimas

Turint didelį tekstinį duomenų rinkinį, greičiausiai jame yra daug tokių terminų, kurie nėra svarbūs nei norint atskirti nei apibendrinti dokumentus. Norint pašalinti nereikšmingus terminus, tai reikia atlikti rankiniu būdu. Tai dažniausiai yra viena iš daugiausiai laiko reikalaujančių ir subjektyvių užduočių visuose teksto tyrybos etapuose. Tam reikia turėti pakankamai daug žinių apie tiriamą sritį.

Analizei nesvarbūs dokumentai be terminų filtravimo yra ieškomi naudojant raktinius žodžius, dar kitaip vadinamus prasminiais žodžiais (angl. *keywords*). Prasminiai žodžiai – tai yra žodžiai, kurie apibūdina teksto turinį. Dokumentai yra filtruojami, jeigu juose nėra kai kurių terminų, arba filtruojami remiantis vienu iš kitų dokumento kintamųjų, tokių kaip kategorija, data ir kt. Dokumentų filtravimas arba terminų filtravimas keičia dokumento terminų matricą.

Sutvarkius pirminius tekstus bei iš jų sudarius žodžių žodyną, toliau tiriami dažniausiai pasitaikantys žodžiai. Jeigu žodis pasitaiko dažnai, bet nėra reikšmingas jis yra įtraukiamas į išmetamų žodžių sąrašą ir duomenų tvarkymo operacijos vėl pakartojamos. Nereikšmingų žodžių sąrašas yra labai svarbus elementas visame teksto apdorojimo procese. Visose kalbose yra tam tikrų žodžių, kurie tekste pasitaiko ganėtinai dažnai, tačiau jokios vertingos informacijos apie patį tekstą nesuteikia, todėl juos reikia pašalinti.

2.1.4. Grafinės vizualizacijos

Vienas iš duomenų analizės etapų yra duomenų vizualizavimas, tai grafinis informacijos pateikimas. Vizualizacija yra labai naudinga norint gauti įvairių įžvalgų iš pirmų rankų, remiantis analizės rezultatais [1]. Tyrėjui yra lengviau suvokti bei susisteminti gautą įvairią informaciją, taip pat apibendrinti rezultatus, kuomet duomenys yra pateikti vaizdiškai.

Pagrindinės teksto tyrybos grafinės vizualizacijos gaunamos naudojant programinės įrangos R paketus *ggplot2* ir *wordcloud*. Paketas *ggplot2* yra R paketas, kuris yra skirtas duomenų vizualizavimui. Tai gali žymiai pagerinti grafikos kokybę bei estetiką ir padaryti ją kur kas efektyvesne. Šis paketas leidžia sukurti beveik bet kokio tipo diagramas. Anotacija yra vienas iš pagrindinių duomenų vizualizavimo žingsnių. Tai leidžia išryškinti pagrindinę diagramos mintį, *ggplot2* šiam tikslui siūlo nemažai funkcijų, kurios leidžia pridėti įvairiausių tekstų ir formų [44].

Wordcloud yra R paketas, sukuriantis gražius žodžių debesis, atvaizduojantis dokumentų skirtumus ir panašumus. Žodžių debesis yra vienas iš populiariausių vizualizavimo būdų norint išanalizuoti kokybinius duomenis. Žodžių debesis yra dažniausiai naudojamas grafikas, norint aiškiai vizualizuoti kalbą ar dokumentų rinkinį [43]. Šis grafikas yra labai naudingas, norint greitai suvokti ryškiausius terminus. Tokie grafikai plačiai naudojami žiniasklaidoje bei gerai suprantami visuomenėje. Kuo žodis didesnis ir ryškesnis, tuo daugiau kartųš jis panaudojamas tekste.

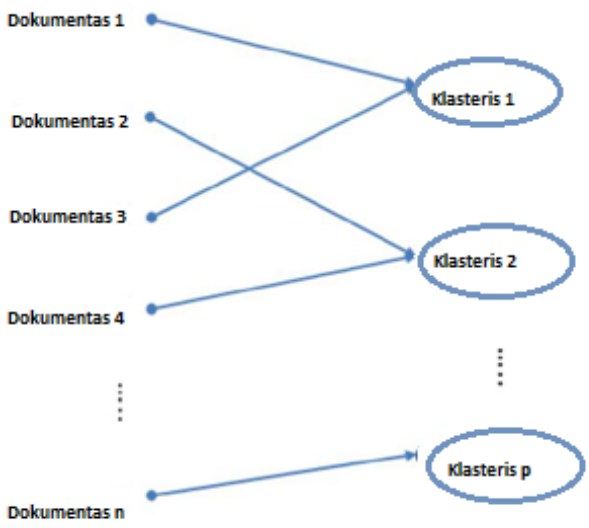
2.2. Duomenų tyryba

Šis duomenų tyrybos žingsnis apima tradicinius duomenų tyrybos algoritmus, tokius kaip klasterizavimas, klasifikavimas, asociacijų analizė ir kitus. Teksto tyryba yra tam tikras iteracinis procesas, o jo metu analizė nuolat kartojama panaudojant įvairius nustatymus bei pašalinant arba įtraukiant tam tikrus terminus, tam kad būtų pasiekti kuo geresni rezultatai. Šio žingsnio rezultatas gali būti dokumentų grupės (klasteriai), temų sąrašai arba taisyklės, kurios atsako į klasifikavimo problemą.

2.2.1. Klasterinė analizė

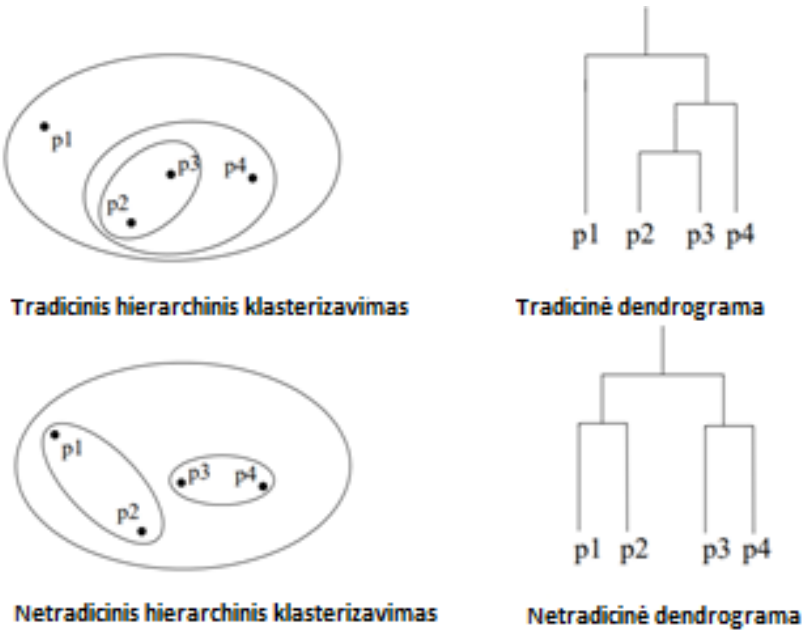
Klasterinė analizė yra populiarus metodas, kurį duomenų analitikai naudoja daugybėje verslo programų. Grupuojant duomenų rinkinio skaidinius įrašai į grupes sudaromi taip, kad grupės subjektai būtų panašūs, o tarp grupių - skirtingi. Klasterinės analizės tikslas yra gauti tokius klasterius, kurie turi vertę sprendžiamos problemos atžvilgiu, tačiau šis tikslas ne visada yra pasiekiamas. Dėl šios priežasties yra daug konkuruojančių klasterizavimo algoritmų. Klasterizavimo procesas suskirsto dokumentus į nepersidengiančias grupes [23].

Kiekvienas tekstinis dokumentas gali būti klasifikuojamas daugiau nei vienoje temoje. Tai yra vienas iš pagrindinių skirtumų tarp grupavimo ir bendro teksto klasifikavimo procesų, nors klasterizavimas suteikia teksto klasifikavimo sprendimą, kai grupės yra viena kitą išskiriančios. Teksto tyrybos kontekste klasterizavimas suskirsto dokumentų rinkinį į vienas kitą išskiriančias grupes pagal panašias temas. Daugumoje verslo programų, kurios yra susijusios su dideliu kiekiu tekstinių duomenų, dažnai yra sunku atskirti kiekvieną klasterį rankiniu būdu skaitant ir atsižvelgiant į visą klasterio tekstą [23].



1 pav. Teksto grupavimo procesas kiekvieną dokumentą priskiriant tik vienam klasteriui

Yra išskiriamos pagrindinės klasterinės analizės sąvokos tokios kaip panašumas bei skirtingumas, dar kitaip apibrėžiamas kaip atstumas tarp objektų. Atstumas tarp objektų nurodo, kiek objektai yra nutolę vienas nuo kito, o objektų panašumas parodo artimumą. Objektai, kurie yra panašūs priklauso tam pačiam klasteriui, o nutolę objektai priklauso skirtingiems klasteriams. Tam, kad būtų paprasčiau suvokti dviejų kintamųjų, kurių duomenys sudaro kelias homogenines grupes paskirstymo į klasterius reiškinį, nubraižoma šių kintamųjų sklaidos (angl. scatter) diagrama. Dažniausiai klasterių struktūra nėra aiški ir klasteriai tapusavyje persidengia, o kintamųjų skaičius būna labai didelis, tai analizė būna atliekama n-matėje erdvėje. Klasterinės analizės metodai skirstomi į du pagrindinius tipus - hierarchinius metodus ir nehierarchinius metodus [23].



2 pav. Hierarchinio klasterizavimo pavyzdžiai

Hierarchinis klasterizavimas dažniausiai yra pateikiamas grafiškai, panaudojant diagramą, kuri yra vadinama dendrograma. Dendrograma parodo kiekvieno klasterio sąryšį bei kokia tvarka klasteriai buvo sujungti. Geriausiai apibrėžtos grupės yra tos, kurios turi didžiausią atskyrimą. Daugelis glaudžiai nutolusių šakos taškų rodo, kad trūksta klasterių (atstumo). Hierarchinis klasterizavimas taip pat gali būti pateiktas lizdo tipo klasterine diagrama (1 pav.).

Hierarchiniuose klasterizavimo metoduose objektai laikomi vienu dideliu klasteriu, kuris susidaro iš mažesnių klasterių, o mažesni iš dar mažesnių ir t.t. Taikant būtent šiuos metodus nustatoma klasterių hierarchija tarpusavyje, dar kitaip vadinama klasterių tarpusavio priklausomybių struktūra [45].

Nehierarchiniai metodai yra taikomi, kai klasterių skaičius yra žinomas arba pasirenkamas.

2.2.2. Tyrime naudojami klasterizavimo metodai

Šiame tyrime bus naudojami klasterizavimo metodai – Ward‘o (angl. *Ward‘s*), k-vidurkių (angl. *k-means*) ir k-medoidų (angl. *k-medoids*) metodai.

Ward‘o metodas remiasi klasterių vidinės sklaidos minimizavimo principu, o tikslo funkcija apskaičiuojama kaip klasterių vidinių kvadratinių nuokrypių suma. Ward‘o metodas yra linkęs apjungti klasterius su nedideliais stebėjimų skaičiais. Šis metodas yra mažiau jautrus „triukšmui“ ir išskirtims, taip pat šis metodas yra greitesnis palyginus su kitais metodais [23].

Ward‘o metodas yra įdomus tuo, kad jis ieško klasterių daugiamatėje Euklido erdvėje. Duomenų lentelės bendrą kvadratų sumą galima apskaičiuoti pagal pradinius kintamuosius arba atstumo matricą tarp stebinių, taip nustatant ryšį tarp atstumų ir kvadratų sumos (arba dispersijos). Atstumas tarp klasterių yra apskaičiuojamas pagal formulę:

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{(\frac{1}{n_K} + \frac{1}{n_L})} \quad (10)$$

čia $\bar{x}$ yra klasterio vektoriaus vidurkis, o n yra stebinių skaičius.

Kiekviename etape klasterių pora, kurios atstumas tarp grupių yra mažiausias, sujungiamos. Kiekviename etape atsižvelgiama į visų įmanomų klasterių porų sąjungą ir sujungiami du klasteriai, kurių sujungimas lemia minimalų informacijos praradimą.

K-vidurkių metodas (angl. *k-means*) – tai nehierarchinis klasterizavimo metodas, kurio metu iš anksto pasirenkamas klasterių skaičius, tačiau pats klasterizavimas yra duomenų egzistuojančių struktūrų paieška, o nustačius klasterių skaičių iš anksto, struktūra yra sukuriama dirbtinai. Šis metodas yra naudojamas, kai turimų objektų skaičius yra ganėtinai didelis ir objektų atstumų matrica gaunama labai didelė. K-vidurkių klasterizavimo metodo etapai [45]:

1. Pirmiausia objektai yra suskirstomi į *k* pradinių klasterių,
2. apskaičiuojamas Euklido atstumas tarp kiekvieno objekto iki klasterių centrų. Objektai suskirstomi į artimiausius klasterius, tuomet pakartotinai perskaičiuojami klasterių centrai,
3. antras žingsnis toliau kartojamas tol, kol daugiau nebelieka perskirstymų.

Euklido atstumas apskaičiuojamas pagal formulę:

$$\|X - Y\| = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (11)$$

čia *m* – požymių skaičius.

Šio klasterizavimo metodo algoritmo trūkumas yra tas, kad būtina žinoti klasterių skaičių iš anksto.

Naudojant šį klasterizavimo metodą yra labai svarbu paleisti algoritmą ne vieną kartą, o daugiau, tik kiekvieną kartą keisti pradinę konfiguraciją ir tada pasirinkti tokį rezultatą, kuris pateikia bendrą mažiausią visų dokumentų grupių atstumų sumą. Reikia paleisti algoritmą vis didesniam skaičiui pradinių konfigūracijų, kol rezultatas reikšmingai nepasikeis. Beje, ši procedūra negarantuoja optimalaus sprendimo.

Tyrime naudojami metodai klasterių skaičiui nustatyti:

- Alkūnės metodas. Klasterio dispersijos sumos kreivėje pasirenkamas posūkio taškas.
- Pagrindinių komponenčių analizė

Klasterizavimo algoritmai veikia matematinėje erdvėje, kurios matmenis yra lygus žodžių skaičiui korpuse, žinoma to neįmanoma vizualizuoti. Norint tai išspręsti, matematikai išrado dimensijos mažinimo metodą, tokį kaip pagrindinių kompončių analizę (PCA), kuri sumažina dimensijų skaičių iki 2 (būtent šio tyrimo atveju) ir tokiu būdu sumažintų dimensijų dėka užfiksuojama kuo daugiau kintamumo tarp klasterių [46]. Norint nuspręsti tinkamiausią klasterių skaičių nubraižomas *clusplot* grafikas. Rezultatas nubraižomas panaudojant programinės įrangos R bibliotekos *cluster* funkciją *clusplot()*.

Pagrindinė klasterių skaidymo metodų idėja yra apibrėžti klasterius taip, kad būtų sumažinta bendra klasterių kvadratų suma :

$$min(\sum_{k=1}^{k} W(C_k)) \tag{12}$$

Čia C_k k-tasis klasteris, o $W(C_k)$ – variacija klasterio viduje.

Bendra klasterio kvadrato suma (angl. *within-cluster sum of square*) rodo klasterių kompaktiškumą, kadangi siekiama, kad jis būtų kuo mažesnis.

K-medoidų metodas (angl. *k-medoids*) – tai toks padalijimo metodas, kai kiekvieną klasterį vaizduoja vienas iš klasteryje esančių objektų. Šis metodas dar kitaip vadinamas k-vidurinių taškų klasterizavimo metodas. Šis metodas gali būti taikomas įvairiausiems duomenims. Metodas pradedamas nuo pradinio medoidų rinkinio ir iteratyviai pakeičiamas vienas iš medoidų vienu iš ne medoidų, jei tai pagerina bendrą susidariusio klasterio atstumą.

Šis metodas taip pat kaip ir k-vidurkių klasterizavimo metodas padalija *n* objektų į klasterius taip, kad būtų minimali objektų atstumo iki klasterių centrų kvadratų suma. K-medoidų metodo trūkumas yra tas, kad jis gerai tinka tik sferiniams klasteriams, yra ganėtinai jautrus išskirtims. Naudojant šį metodą reikia nurodyti norimą klasterių skaičių *k* ir pradinius klasterių centrus, nuo jų parinkimo dažniausiai priklauso klasterizavimo rezultatas. Klasterio medoidas apibrėžiamas kaip klasterio objektas, kurio vidutinis nepanašumas su visais klasterio objektais yra minimalus, tai yra labiausiai centre esantis klasterio taškas.

Skirtingai nuo k-vidurkių algoritmo, k-medoidai pasirenka faktinius duomenų taškus kaip centrus (medoidus) ir taip leidžia geriau suprasti klasterių centrus nei k-vidurkių metodu, kur klasterio centras nebūtinai yra vienas iš įvesties duomenų taškų. Be to, k-medoidai gali būti naudojami su savavališkomis nepanašumo priemonėmis, kai tuo tarpu k-vidurkių algoritme efektyviems sprendimams paprastai reikalingas Euklido atstumas. Kadangi k-medoidai sumažina porinių skirtumų sumą, o ne kvadratinių Euklido atstumų sumą, ji yra tvirtesnė triukšmui ir pašaliniams tikslams nei k-vidurkių metodas.

Grupavimo aplink medoidus algoritmas (angl. *Partitioning Around Medoids (PAM)*), vadinamas buvo vienas iš pirmųjų pristatytų k-medoidų algoritmų. Jis bando nustatyti *n* objektų *k* skaidinius. Po pirminio atsitiktinio k-medoidų pasirinkimo algoritmas pakartotinai bando geriau pasirinkti medoidus [47]. Grupavimo aplink medoidus algoritmas yra programinės įrangos R pakete *cluster*. Kadangi klasterių skaičius parenkamas savarankiškai, tai klasterių skaičiau tinkamumą galima įvertinti tokiais metodais kaip silueto (angl. *silhouette*) metodas.

Silueto metodas nurodo duomenų klasterių nuoseklumą ir patvirtinimą. Šis metodas pateikia glaustą grafinį vaizdą, kaip gerai klasifikuotas kiekvienas objektas. Silueto vertė yra toks matas, kuris parodo kiek objektas yra panašus į jo klasterį, palyginti su kitais klasteriais. Šis dydis svyruoja nuo −1 iki +1, kuo šis matas yra didesnis, tuo objektas yra geriau pritaikytas prie jo pačio klasterio ir blogai prie kaimyninių klasterių. Jeigu dauguma objektų turi didelę silueto mato vertę, galima teigti, kad grupavimo konfigūracija yra tinkama, o jeigu vertė yra maža arba neigiama, grupių konfigūracijoje gali būti arba per daug, arba per mažai klasterių. Silueto matas apskaičiuojamas naudojant bet kokią

atstumo metriką, pvz., Euklido ar Manheteno atstumą. Siluetui nubraižyti, bei apskaičiuoti, buvo naudojama R paketas *clusterR*. Optimalus klasterių *k* skaičius yra tas, kuris maksimaliai padidina vidutinį siluetą per galimų *k* reikšmių diapazoną.

Manheteno atstumas išreiškiamas matuojant dimensijų skirtumų sumą, jis yra labai panašus į Euklido atstumą. Apskaičiuojamas pagal formulę:

$$d(x, y) = \sum_i |x_i - y_i| \quad (13)$$

Atlikus klasterinę analizę, kiekvieną klasterį galima apibūdinti terminų rinkiniu, kuris tam tikru mastu atskleidžia klasterio temą. Tokio pobūdžio analizė padeda verslui suprasti visą rinkinį ir gali padėti teisingai klasifikuoti klientus pagal klientų skundų ar atsakymų bendrąsias temas [23].

2.3. Sentimentų analizė

Dėl socialinių tinklų vartotojų sukurto turinio apimties bei nestruktūrizuoto pobūdžio nuomonės analizė ir sentimentų analizė, tai yra vadinamoji teksto analitika, atlieka svarbų vaidmenį didžiųjų duomenų analizėje. Sentimentų analizės technologijos leidžia išgauti nuomones iš nestruktūrizuotų žmogaus sukurtų dokumentų, tai gali būti labai puiki priemonė daugeliui verslo analitikos (angl. business intelligence) užduočių, taip pat įskaitant reputacijos valdymą, ryšius su visuomene, viešųjų nuomonių sekimą bei rinkos tendencijų prognozavimą [24]. Sentimentų analizė - tai vienas iš duomenų tyrybos metodų, kuris yra naudojamas norint išgauti bei apdoroti subjektyvią informaciją, kurią vartotojai pateikia apžvalgos komentaruose, norint suprasti vartotojų emocijas susijusias su konkrečia tema [3].

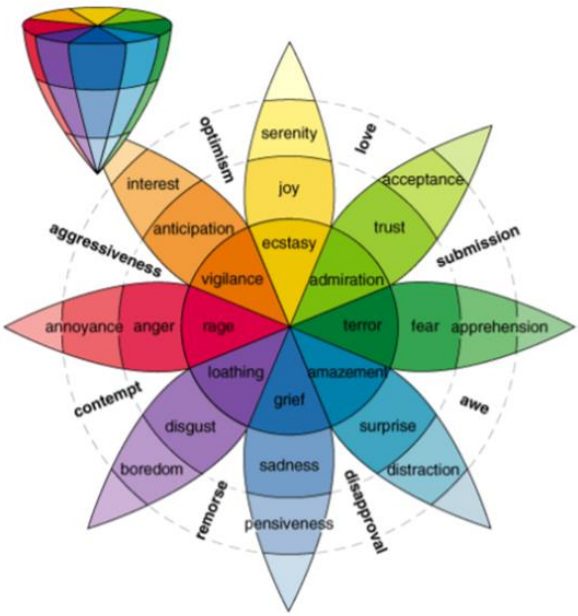
Mašininio mokymosi algoritmą galima pritaikyti ir sentimentų analizei, žiūrint į ta, tikrą užduotį kaip į klasifikavimo. Tekstai, kurie yra priskiriami vienai iš trijų kategorijų yra surenkami bei užkoduojami įvairiuose skaitmeniniuose vektoriuose. Sentimentų analizė dažniausiai yra labai priklausoma nuo tam tikrų teigiamų bei neigiamų sąlygų, kadangi nuomonė pateikiama kaip labai trumpas tekstas, todėl informacijos dažniausiai nepakanka norint atskirti teigiamas ir neigiamas nuomones [25].

Pagrindinis semantinių požiūrių privalumas yra tas, kad klaidas yra palyginti lengva ištaisyti, pridedant tiek žodžių, kiek tik reikia. Mašininio mokymosi metoduose klaidas ištaisyti yra sudėtingiau, ir tai dažniausiai įmanoma tik išplėtus tekstų rinkinį bei permokant modelį. Kita vertus, mokymosi principų pranašumas yra tas, kad gana lengva ir greita sukurti sentimentų / nuomonės analizės variklį, mokomą rinkti pažymėtus tekstus. Dėl šios priežasties yra nesudėtinga sukurti tam tikroje srityje pritaikytus klasifikatorius [23].

Sentimentų analizė yra metodas, kuris yra naudojamas norint klasifikuoti teksto emocijas. Yra du pagrindiniai budai, kaip sprendžiami sentimentų analizes uždaviniai. Pirmasis būdas tai žodynais pagrįstas metodas, kuomet semantinė teksto orientacija yra vertinama pagal žodžius sakinyje ir pagal jų nešamą emocinį ivertį. Antrasis būdas, tai teksto klasifikavimo metodas, kuris dar yra įvardijamas kaip mašininio mokymosi (angl. *machine-learning*) metodas. Jis yra pagristas sukurta klasifikavimo sistema, panaudojant jau pažymetus tekstus arba sakinius.

Terminas sentimentų analizė gali būti naudojamas norint nurodyti daugybę skirtingų, tačiau susijusių problemų. Dažniausiai jis naudojamas norint automatiškai nustatyti teksto, kuris yra teigiamas, neigiamas ar neutralus, valentingumą ar poliškumą. Tačiau apskritai, tai reiškia, kad reikia nustatyti

savo požiūrį į konkretų tikslą ar temą. Čia požiūris gali reikšti vertinamąjį sprendimą, pavyzdžiui, teigiamą ar neigiamą, arba emocinį ar afektinį požiūrį, pvz., nusivylimą, džiaugsmą, pyktį, liūdesį, jaudulį ir pan.



3 pav. Aštuonių pagrindinių emocijų rinkinys, siūlomas Plutchik‘o (1980m.)

Šiame darbe bus nagrinėjamos 8 pagrindinės emocijos, pasiūlytas Robert‘o Plutchik‘o, tokios kaip pyktis (angl. *anger*), numatymas (angl. *anticipation*), pasibjaurėjimas (angl. *disgust*), baimė (angl. *fear*), džiaugsmas (angl. *joy*), liūdesys (angl. *sadness*), nuostaba (angl. *surprise*), pasitikėjimas (angl. *trust*) [50]. 3 paveiksle parodyta, kaip Plutchik‘as išdėstė šias emocijas ant rato taip, kad priešingos emocijos atrodytų diametraliai priešingos viena kitai. Arčiau centro esantys žodžiai yra didesnio intensyvumo nei tolimesni.

Sentimentų analizė yra teksto klasifikavimo procesas į vieną iš trijų kategorijų: teigiamą, neutralų bei neigiamą. Teigiamas reiškia nuomonę, kurioje kažkas yra apibūdinamas teigiamomis išraiškomis, tokiomis kaip geras bei puikus. Neutralus reiškia tą, kurį apibūdina objektyviai, nei teigiamai nei neigiamai, arba tiek su teigiamu tiek su neigiamu mišiniu. Neigiamas apibūdinimas tai kai kažkas yra apibūdinamas neigiamomis išraiškomis, tokiomis kaip blogas, baisus bei prastas. Neutralus gali būti suskirstytas į du tipus: vienas be sentimentalios išraiškos ir taip pat vienas su teigiamų bei neigiamų išraiškų mišiniu [25].

Sentimento analizės modeliuose pagrindinis dėmesys skiriamas poliškumui (teigiamam, neigiamam, neutraliam), bet taip pat ir jausmams bei emocijoms (piktiems, laimingiems, liūdniems ir pan.), taip pat skubumui (skubiems, neskubiems) ir net ketinimams (suinteresuoti, neįdomūs).

Automatiškai klasifikuoti natūralia kalba parašytą tekstą į teigiamą ar neigiamą sukeliantį jausmą, nuomonę ar subjektyvumą kartais būna taip stipriai komplikuota, kad net skirtingi žmonių anotatoriai nesutaria dėl klasifikavimo, kurį reikia priskirti tam tikram tekstui. Asmeninės interpretacijos gali skirtis nuo kitų žmonių interpretacijų, kadangi tam įtakos turi net ir kultūriniai veiksniai bei kiekvieno

žmogaus asmeninė patirtis. Žinoma, taip pat daug kas priklauso ir nuo teksto kokybės - kuo trumpesnis tekstas ir kuo blogiau parašytas, tuo sunkesnė užduotis išanalizuoti tekstą tai tampa, kaip socialinių tinklų žinučių atveju [23].

Tipinės sentimentų analizės užduotys [24]:

1. Dokumentų, kurie yra susiję su konkrečia tema arba tikslu paieška;
2. Surinktų dokumentų išankstinis apdorojimas, pvz., dokumentų žymėjimas vienu žodžiu ir iš jų reikiamos informacijos išskyrimas;
3. Nuotaikos nustatymas.

Sentimentų analizės rezultatai taip pat gali būti pavaizduoti skaitine skale, siekiant geriau išreikšti teigiamo ar neigiamo nuotaikos stiprumo laipsnį tekste. Šiame tyrime sentimentų balams generuoti naudojamas programinės įrangos R paketas *Syuzhet*, kuriame yra keturi sentimentų žodynai ir siūlomas būdas, kaip pasiekti NLP grupėje Stanforde sukurtą nuotaikos išgavimo įrankį.

Programinės įrangos R paketo *Syuzhet* pagalba analizuojamas tekstas paverčiamas sakinių vektoriumi. Toliau šis vektorius nusiunčiamas į funkciją *get_sentiment()*, kuri įvertina kiekvieno žodžio ar sakinio nuotaiką. Naudojant syuzhet‘o metodą, nuotaikos balų skalė yra dešimtainė ir svyruoja nuo -1 (labiausiai neigiamas) iki +1 (labiausiai teigiamas). Numatytasis žodynas geriausiai pritaikomas grožinei literatūrai, kadangi terminai tam žodynui buvo paimti iš 165 000 žmonių užkoduotų sakinių rinkinio, kuris buvo paimtas iš ganėtinai nedidelio šiuolaikinių romanų korpuso.

Funkcija *get_sentiment()* iš viso priima du argumentus: sakinių arba žodžių vektorių ir metodą. Pasirinktas metodas nustato, kuris iš keturių galimų nuotaikų išskyrimo būdų bus naudojamas. Pagal numatytuosius nustatymus *Syuzhet* pakete yra keturi metodai - *bing, afinn, nrc* ir *stanford*. Kiekvienas metodas pateikia šiek tiek skirtingus rezultatus, kadangi naudoja skirtingą skalę. *Nrc* metodo rezultatas yra daugiau nei tik skaitinis balas, todėl jam reikalingas papildomas interpretavimas [48].

Šiame darbe atliekama emocijų klasifikacija. Emocijų klasifikacija yra paremta NRC žodžių ir emocijų asociacijos leksika (dar žinoma kaip „EmoLex“). Tai yra anglų kalbos žodžių sąrašas ir jų sąsajos su aštuoniomis pagrindinėmis emocijomis (pykčiu, baime, laukimu, pasitikėjimu, nuostaba, liūdesiu, džiaugsmu ir pasibjaurėjimu) ir dviem jausmais - neigiamu ir teigiamu. [49].

Tyrime naudojama programinės įrangos R funkcija *get_nrc_sentiments()*, kuri pateikia duomenų lentelę (angl. *data frame*) su kiekviena eilute, vaizduojančia sakinį iš pradinio failo. Stulpelių duomenis (pyktį, numatymą, pasibjaurėjimą, baimę, džiaugsmą, liūdesį, nuostabą, pasitikėjimą, neigiamą ir teigiamą jausmą) galima pasiekti atskirai arba rinkiniais. Funkcija *get_nrc_sentiments()*, kviečia NRC nuotaikų žodyną apskaičiuoti aštuonių skirtingų emocijų buvimą ir atitinkamą jų valentingumą tekstiniame faile [48].

3. Tyrimo rezultatai

Šioje darbo dalyje yra pateikiami atliktų tyrimų rezultatai. Modeliai sudaryti panaudojant R programinę įrangą ir *Python* programavimo kalbą ir jos paketus bei *Jupyter Notebook* programinę įrangą. Tyrime naudojamą duomenų rinkinį sudarė „Airbnb“ vartotojų atsiliepimai iš 108 pasaulio miestų. Vartotojų atsiliepimai apėmė laikotarpį nuo 2009m. iki 2021m. Tyrimas yra atliekamas remiantis 2 skyriuje pateikta metodologija bei jos programine realizacija. Žaliųjų „Airbnb“ vartotojų atsiliepimai ištraukti iš internetinio tinklapio „Inside Airbnb“. Duomenis sudaro atsiliepimai iš įvairių miestų, kuriuose yra galimybė naudotis „Airbnb“ apgyvendinimo paslaugomis.

3.1. Tyrime naudojamo duomenų rinkinio paruošimas analizei

Teksto tyryba pradedama nuo duomenų paruošimo analizei. Duomenų rinkinį sudarė 35001365 vartotojų internetiniai atsiliepimai. Duomenys gauti iš tinklalapio „Inside Airbnb“, kuriame kasmet yra renkama informacija bei pateikiama visomis kalbomis. Duomenų rinkinį sudarė 6 kintamieji: sąrašo ID (*listing_id*), atsiliepimo ID (*id*), atsiliepimo data (*date*), vartotojo ID (*reviewer_id*), vartotojo vardas (*reviewer_name*) ir atsiliepimas (*comments*).

Atlikus duomenų nuskaitymą, iš duomenų rinkinio išfiltruoti su tvarumu susiję internetiniai atsiliepimai. Atsižvelgiant į literatūros analizę, labiausiai su tvarumi asocijuojasi žodžiai – *“sustainable”, “sustainability”* ir *“organic”,* todėl iš visų „Airbnb“ vartotojų internetinių atsiliepimų išfiltruojami tik tie, kuriuose yra bent vienas iš šių žodžių. Išfiltruoti duomenys išsaugomi kaip atskiras duomenų rinkinys *filtered.csv*, su šiuo duomenų rinkiniu bus tęsiama tolimesnė analizė. Žemiau pateikiama iškarpa iš išfiltruoto duomenų failo.

listing_id	id	date	reviewer_	reviewer_	comments
41125	1199585	30/04/2012	1374194	Simon	Apartment is exactly like the photos, big and airy. The local area is full of life and interest, lots of great cafes and restaurants with an organic supermarket (in case you're into such things) just around the corner.
49552	55193005	29/11/2015	38381620	Ed	If you come to Amsterdam, this is YOUR PLACE TO STAY! Joanna is charming, focused on her customer's needs, knowledgable, and a consummate professional. Her three level apartment is comfortable, well furnished, and PRIVATE. If you want to stay in and cook... you can! If you want to have a few friends stop by... you can! The location cannot be beat and the neighborhood is outstanding. Convenient to organic shopping, wine, fun distractions. We WILL BE BACK.
58211	86025998	14/07/2016	623299	Tiffany	Wow. This is simply the best Airbnb experience I've had thus far. Amazing apartment, perfect

4 pav. Žaliųjų “Airbnb” vartotojų atsiliepimų duomenų rinkinio iškarpa

Šį duomenų rinkinį sudarė 25725 stebėjimai. Daugiausia tvarių vartotojų atsiliepimų buvo pastebėta Australijos „Airbnb“ duomenų rinkinyje, jų iš viso buvo 5779 ir tai yra daugiau nei 22 proc. visų vartotojų atsiliepimų, o mažiausiai Singapūre, iš viso parašyti 5 atsiliepimai, kuriuose minimas bent vienas su tvarumu susijęs žodis.

Sekančiame žingsnyje nustatoma vartotojo atsiliepimo kalba. Pašalinami įrašai, kurie nėra parašyti anglų kalba. Buvo panaudota kalbos atpažinimo biblioteką CLD (angl. *Compact Language Detector*), ši biblioteka naudoja neuroninių tinklų modelius.

values	
cld2	chr [1:25725] "en" "en" "en" "en" "en" "en" "en" "en" "en" "en" "en" "en"…
cld3	chr [1:25725] "en" "en" "en" "en" "en" "en" "en" "en" "en" "en" "en" "en"…

5 pav. CLD2 ir CLD3 kabos atpažinimo modelio rezultatai

Pašalinus komentarus, kurie nebuvo parašyti anglų kalba, liko 25440 stebėjimai. Taigi, 98,89 proc. žaliųjų „Airbnb“ vartotojų atsiliepimai buvo parašyti anglų kalba.

Toliau atliekamas duomenų valymas. Teksto duomenų valymas prasideda nuo transformacijų. Visas tekstas verčiamas į mažąsias raides, taip palengvinami tolimesni veiksmai. Tam yra naudojama R programinės įrangos funkcija *content_transformer*. Tekstuose taip pat kartais pasitaiko ir nuorodų, kurios analizei neneša jokios informacijos, todėl jos yra pašalinamos. Pašalinami tarpai, ne angliškos abėcėles raidės bei įvairūs simboliai, tokie kaip šauktukai, klaustukai, kableliai, apostrofai, taškai, skaičiai ir pan. Tai atliekama naudojant funkciją *tm_map ()*, norint pakeisti specialiuosius simbolius, jie keičiami tarpu. Pašalinami vartotojų atsiliepimų dublikatai. Taip pat ištrinami *whitespace* įrašai. Tokie įrašai susidaro kai vartotojas prideda daug tarpo simbolių tekste. Pašalinami populiariausi vardai, panaudojus sąrašą iš internetinės svetainės [39]. Sąrašą sudarė 18239 populiariausi žmonių vardai.

Ištraukiamas, išvalomas bei sukuriamas žodžių žodynas iš dokumentų naudojant NLP (angl. *Natural Language Processing*). Tai žingsnis, apimantis ištrauktų žodžių analizavimą, nereikalingų žodžių (angl. *stop words*) pašalinimą bei rašybos tikrinimą. Nėra vieno universalaus nereikalingų žodžių, kuriuos naudoja visi NLP įrankiai, sąrašo. Funkcija *tm_map()* palaiko kelias kalbas, tokias kaip anglų, prancūzų, vokiečių, italų ir ispanų. Internetinėje svetainėje rastas populiariausių nereikalingų žodžių sąrašas, visi išvardinti žodžiai pašalinami [40].

Sutvarkius pirminius tekstus bei iš jų sudarius žodžių žodyną, toliau tiriami dažniausiai pasitaikantys žodžiai. Jeigu žodis pasitaiko dažnai, bet nėra reikšmingas jis yra įtraukiamas į išmetamų žodžių sąrašą ir duomenų tvarkymo operacijos vėl pakartojamos. Tokiu principu papildomai pašalinami tokie tekstui reikšmės nesuteikiantys žodžiai, kaip – *“a”, ”minutes” ,”day” ,”stay” , ”karls”, ”apartment”, ”Airbnb”, ”just”, “lots”, “need”, “feel”, “time”, “area”, “like”, “night”, “want”, “staying”, “stayed”, “away”, “late”, “definitely”, “walk”, “home”, “house”, “really”, “thank”, “provided”, “make”, “easy”, “highly”, “use”, “arrived”, “living”, “touches”, “come”, “felt”, “right”, “morning”, “book”, “check”, “guests”, “person”, “didnt”, “return”, “set”, “don’t”, “got”, “its”, “bit”, “week”, “able”, “went”, “youre”, “asked”, “plus”.* Šie žodžiai buvo atrinkti rankiniu būdu, išsifiltravus visus žodžius, kurie duomenų rinkinyje kartojasi dažniau nei 1000 kartų. Išrinkus iš duomenų rinkinio žodžius, kurie pasikartoja rečiau, pvz. 500 kartų, buvo gautas didelės apimties duoemnų rinkinys, kuriame didelę dalį užima rečiau pasikartojantys žodžiai, todėl tolimesnei analizei buvo pasirinktas duoemnų rinkinys, kuris susidaro iš žodžių, kurie atsiliepimuose pasitaiko dažniau nei 1000 kartų.

Siekiant dar labiau sumažinti unikalių tekstų skaičių yra naudojamas šaknies išskyrimo algoritmas, dar kitaip vadinamas žodžio kamieno išskyrimu, kurio metu yra nupjaunama žodžio galūnė. Anglų kalbai dažniausiai naudojamas Porter‘io metodas, būtent jis ir buvo naudojamas žodžių kamienų išskyrimui. Šis metodas realizuotas panaudojus Python programavimo kalbą ir jos paketus bei Jupyter Notebook programinę įrangą panaudojant NLTK biblioteką, kurioje yra paketas pavadinimu „PorterStemmer“. Galiausiai sudaromas pilnų pradinių žodžių žodynas, priskiriant jiems iš nukirptų žodžių atitikmenį ir sugeneruojant galūnę. Tam, kad tuštiems žodžiams algoritmas nepriskirtų reikšmės jie yra praleidžiami.

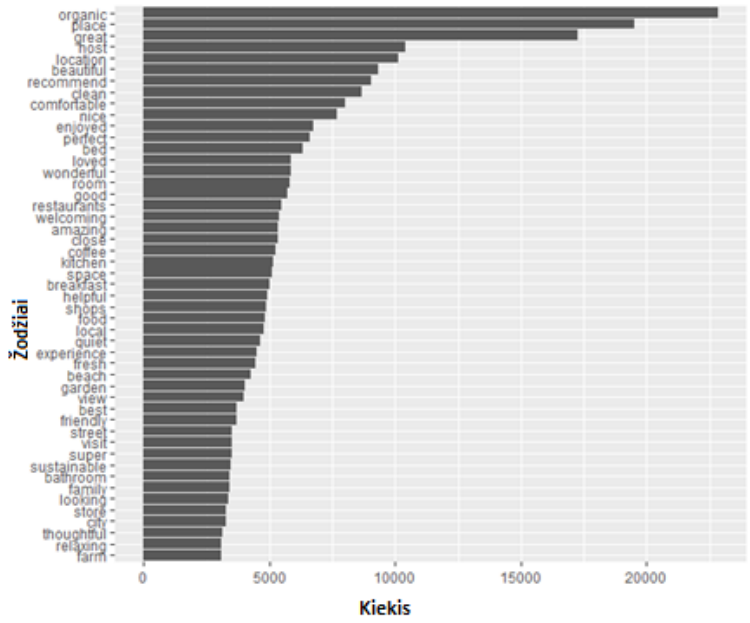
Svarbiausia užduotis po gramatinio išnagrinėjimo yra teksto transformavimas. Šio uždavinio rezultate gaunama dokumento terminų matrica (angl. *term-by-document matrix).* Buvo gauta matrica, kuri turi

7211 eilutes ir 25440 stulpelius, iš viso dokumento terminų matricą sudaro 183447840 reikšmės. Dokumento terminų matrica yra apibendrinta naudojant SVD metodą, kuris sukuria statistinius tekstinių dokumentų vaizdus. Šie SVD balai vėliau gali būti įtraukti kaip skaitiniai įvadai į skirtingų tipų modelius, tokius kaip klasterio ar nuspėjamieji modeliai. Panaudojant programinės įrangos R funkciją *TermDocumentMatrix()* iš teksto tyrybos paketo, galima sukurti dokumento matricą - lentelę, kurioje yra nurodomas žodžių dažnis. Lentelėje žemiau išvardinami 10 populiariausių žodžių visame tekste.

1 lentelė. 10 populiariausių žodžių tekste

Nr.	Žodis	Dažnumas
1.	organic	22838
2.	place	19532
3.	great	17266
4.	host	10384
5.	location	10136
6.	beautiful	9329
7.	recommend	9056
8.	clean	8653
9.	comfortable	8029
10.	nice	7663

Iš duomenų rinkinio išfiltravus visus žodžius, kurie klientų atsiliepimuose pasikartoja daugiau nei 500 kartų, buvo gautas atsiliepimuose naudotų žodžių sąrašas (žr. 1 priedą). Buvo gauti 397 žaliųjų "Airbnb" vartotojų dažniausiai vartojamų žodžių sąrašas. Taigi, pagal gautą atsiliepimuose naudojamų žodžių dažnių lentelę galima matyti, kad dažniausiai panaudojamas žodis žaliųjų "Airbnb" vartotojų atsiliepimuose buvo "*organic*", jį visuose žaliųjų vartotojų atsiliepimuose galima buvo pamatyti 22838 kartus. Antras pagal dažnumą žodis buvo *"place",* kuris pasikartojo 19532 kartus.



6 pav. Dažniausiai pasikartojantys žodžiai tekste

Toliau rezultatai atvaizduojami grafiškai (6 pav.). Panaudojus programinės įrangos R funkciją *ggplot()*, nubraižomas grafikas, kurime atvaizduojami dažniausiai žaliųjų "Airbnb" vartotojų atsiliepimuose naudojami žodžiai. Atrinkti daugiau nei 3000 kartų pasikartojantys žodžiai. Grafike žemiau pateikti rezultatai išrikiuoti nuo dažniausiai pasikartojančių žodžių iki rečiausiai.

Pagal gautą grafiką matome, kad žodis *"organic"* vartotojų atsiliepimuose pasikartoja maždaug 7 kartus dažniau nei žodžiai *"farm"*, *"relaxing"* bei *"thoughtful"*. Žodžiai *"organic", "place"* ir *"great"* labai aiškiai išsiskiria iš kitų atsiliepimuose vartojamų žodžių. Galima teigti, kad žalieji „Airbnb" vartotojai daugiausiai savo atsiliepimuose akcentuoja ekologiškumą, vietą, šeimininką, lokaciją, švarą ir patogumus. Iš viso pateikiami 49 žodžiai, kurie pasikartoja dažniau nei 3000 kartų.

Toliau pateikiama dažniausiai pasikartojančių žodžių asociacijos. Ši metodika gali būti efektyviai panaudojama, kai norima išanalizuoti kurie žodžiai dažniausiai pasitaiko kartu su dažniausiai pateikiamais žodžiais atsiliepimuose, o tai padeda suprasti šių žodžių kontekstą. Panaudojus programinės įrangos R funkciją *findAssocs(),* kai skaitinis vektorius (angl. *corlimit)* lygus 0,1, buvo gautos trijų populiariausių žodžių asociacijos. Norint pamatyti daugiau žodžių galima nustatyti žemesnį skaitinį vektorių, arba aukštesnį, jei norima pamatyti mažiau žodžių. Buvo pastebėta, kad didžiausia *corlimit* skaitinio vektoriaus reikšmė su pasirinktais žodžiais yra 0,17, taigi buvo pasirinta taikyti 0,1, tam, kad būtų galima pamatyti daugiau žodinių asociacijų. Paveikslėlyje žemiau pateikiamos asociacijos.

7 pav. Dažniausiai pasikartojančių žodžių asocijacijos

Pagal gautus žodinių asocijacijų rezultatus, pastebima, kad su žodžiu *"organic"* 12 proc. iš visų porų su šiuo žodžiu, pasitiko žodis *"coffee",* 11 proc. žodis "fresh" ir 10 proc. – *"fruit"*. Pastebimos žodžio *"organic"* tam tikros asocijacijos su maistu, kadangi tiek kava tiek vaisiai yra maisto produktai, o šviežumas dažniausiai apibūdina maistą. Tokiu pačiu principu išnagrinėjamas ir kitas žodis *"place"*. 14 proc. asocijacijų su žodžiu *"eat",* 13 proc. su žodžiu *"recommend",* bei pastebima 10 proc. su žodžiu *"clean",* o tai reiškia, kad net 10 proc. visų žodinių asocijacijų su vieta sudarė švara. Su žodžiu *"great"* dižiausios asocijacijos pastebimos su žodžiais *"location", "restaurants", "shop"* bei *"neighborhood",* o tai reiškia, kad dažniausiai vartotojai su žodžiu puikus įvardina viešnagės vietą, restoranus, parduotuves bei kaiminystę. Taip pat buvo išnagrinėtos žodžio *"recommend"* asociacijos ir buvo pastebėta, kad daugiausiai su šiuo žodžiu asociacijų turi tokie žodžiai kaip *„highly",* kas reiškia labai rekomenduoja, taip pat *"place", "staying"* ir *"apartment"*. Tai reiškia, kad vartotojai yra linkę rekomenduoti vietą bei apartamentus.

Toliau sutvarkytas duomenų rinkinys atvaizduojamas grafiškai žodžių debesyje, panaudojant programinės įrangos R paketą *wordcloud*. Šis grafikas naujos informacijos mums nesuteikia ir yra tiesiog kitoks būdas grafiškai pavaizduoti žodžių dažnumą. Žodžių debesis yra vienas iš populiariausių būdų vizualizuoti ir analizuoti kokybinius duomenis. Tai grafinis vaizdas, kuris yra

sudarytas iš raktinių žodžių, esančių tekste, kur kiekvieno žodžio dydis parodo jo dažnumą tame tekste.

8 pav. Žodžių debesis

Pagal gautą žodžių debesies grafiką, galima aiškiai matyti, kad žodžiai *"organic", "place"* ir *"great"* labai aiškiai išsiskiria iš kitų atsiliepimuose vartojamų žodžių. Tai buvo galima matyti ir dažniausiai pasitaikiusių žodžių grafike (8 pav.). Kuo grafike pavaizduotas žodis didesnis ir ryškesnis, tuo dažniau tekste jis yra naudojamas. Taip pat gana ryškiai išsiskiria tokie žodžiai kaip "*location", "host", "beautiful", "perfect"*. Galima daryti išvadą, kad žalieji "Airbnb" vartotojai labiausiai pabrėžia ekologiją savo atsiliepimuose, tačiau taip pat komentaruose dažniausiai rašo ir apie tokius dalykus kaip vieta, šeimininkas, švara, patogumai ir kt.

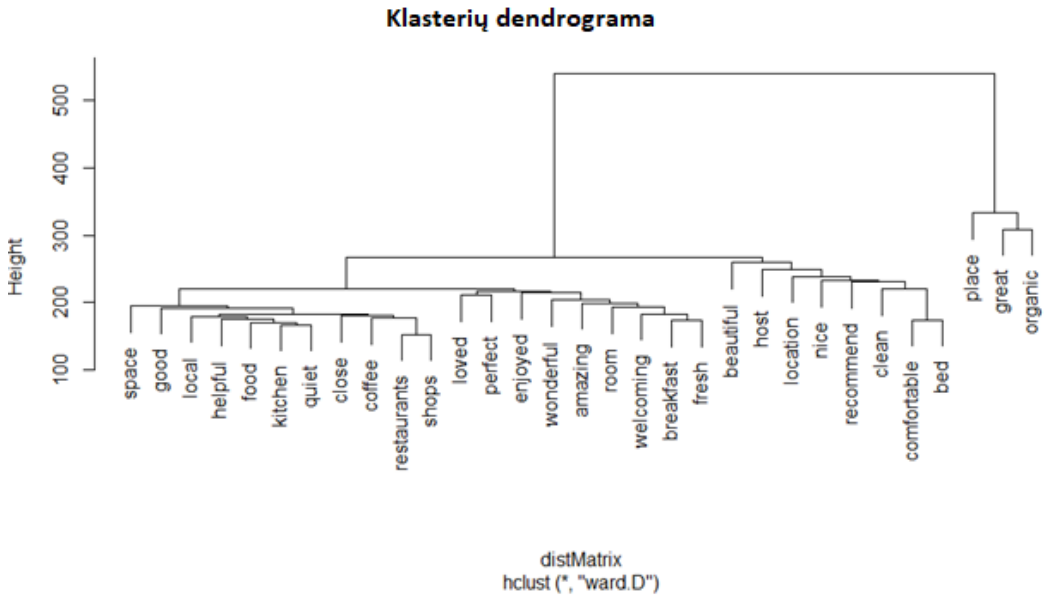
3.2. Klasterinė analizė

Šioje darbo dalyje bus aptarti klasterinės analizės rezultatai. Tyrimas buvo atliekamas panaudojant 3 klasterizavimo metodus - Ward'o (angl. Ward's), k-vidurkių (angl. k-means) ir k-medoidų (angl. k-medoids). Tyrimui naudojama programine įranga R.

3.2.1. Ward'o klasterizavimo metodo rezultatai

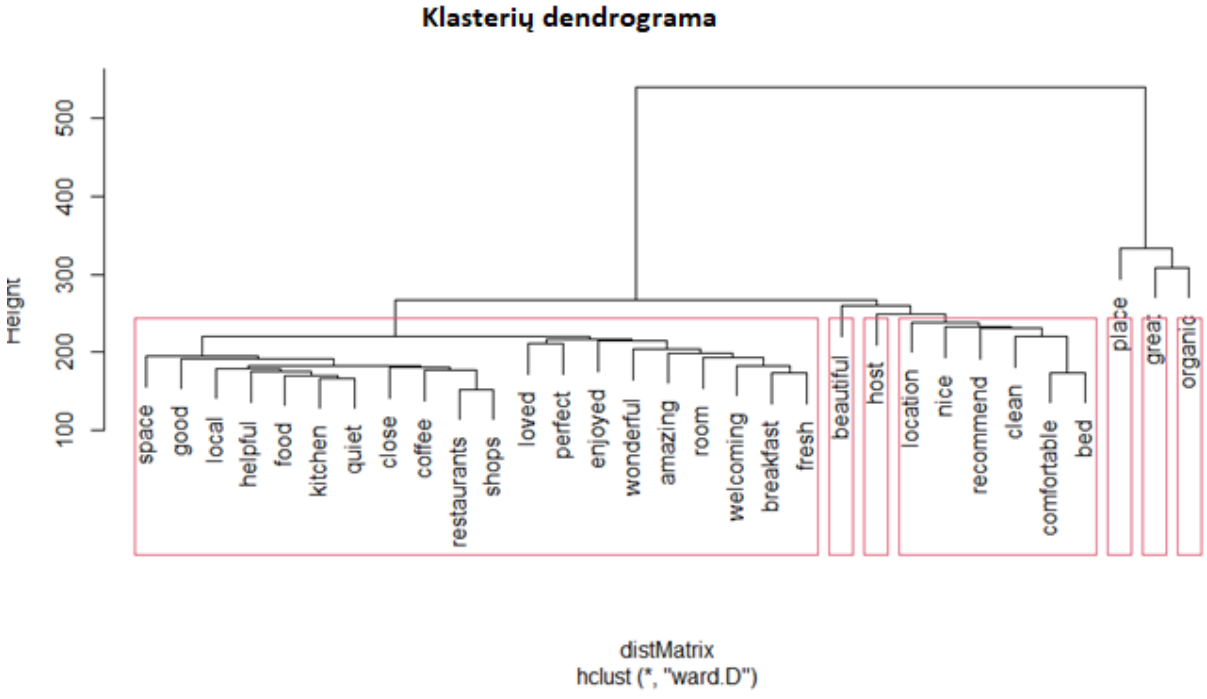
Pirmiausia buvo bandyta rasti žodžių grupes panaudojant hierarchinį klasterizavimo Ward'o metodą. Retos sąvokos iš dokumento terminų matricos buvo pašalinamos tam, kad klasterių grafikas nebūtų perkrautas žodžiais, tam buvo pabnaudota funkcija *removeSparseTerms(),* kai *sparse*=0,85. Retumas (angl. *sparse*) - skaičius didžiausiam leidžiamam retumui diapazone. Pašalinimui buvo pasirinkti visi žodžiai, kurie tyrime pasitaikė rečiau nei 0,85, tai reiškia, kad liko tik terminai, esantys tarp 15 proc. dažniausiai pasitaikančių terminų. Tada atstumai tarp terminų apskaičiuojami panaudojant

programinės įrangos R funkciją *dist()*. Po to terminai sugrupuojami panaudojant funkciją *hclust()* ir dendrogramą.



9 pav. Ward‘o klasterizavimo metodo dendrograma

Toliau dendrograma atvaizduojama su išskirtais 7 klasteriais, kuriuos galima pamatyti dendrogramos viduje apvestus raudonai. Šis grafikas atvaizduojamas panaudojant programinės įrangos R funkciją *rect.hclust()* nurodžius klasterių skaičių – 7.



10 pav. Ward‘o klasterizavimo metodo dendrograma su išskirtais 7 klasteriais

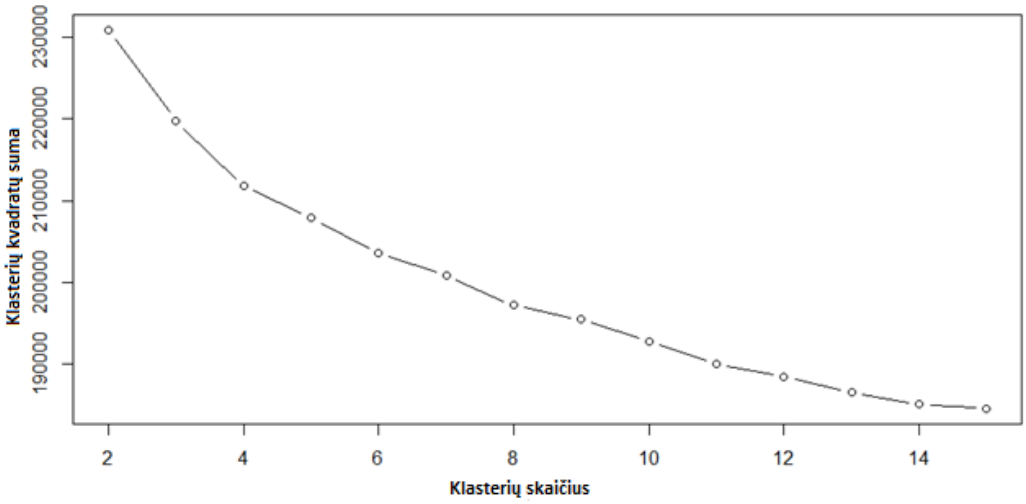
Pagal gautą dendrogramą matome atsiskyrusius 7 klasterius.

Pirmąjį klasterį sudaro tokie žodžiai kaip: *“space”, “good”, “local”, “helpful”, “food”, “kitchen”, “quiet”, “close”, “coffee”, “restaurants”, “shops”, “loved”, “perfect”, “enjoyed”, “wonderful”, “amazing”, “room”, “welcoming”, “breakfast” ir “fresh”*. Visi šie žodžiai susiję su kambariu ir jo patogumais, taip pat su maistu. Šis klasteris yra didžiausias, jį sudaro 20 žodžių. Antrąjį klasterį sudaro tik vienas žodis *“beautiful”*. Trečiąjį žodį sudaro klasteris *“host”*. Ketvirtasis klasteris susideda iš žodžių *“location”, “nice”, “recommend”, “clean”, “comfortable”* ir *“bed”*. Penktąjį klasterį sudaro vienas žodis *“place”,* šeštąjį – *“great”,* o septintąjį – *“organic”*. Paskutiniai trys klasteriai iš visų kitų išsiskiria gana aiškiai, kadangi būtent šie trys žodžiai yra populiariausi vartotojų atsiliepimuose.

3.2.2. K-vidurkių klasterizavimo metodo rezultatai

Prieš tai atliktame hierarchiniame klasterizavime Ward‘o metodu, klasterių skaičius nebuvo nurodytas iš anksto, o buvo nustatytas pažvelgus į dendogramą, kai algoritmas buvo atlikęs savo darbą. K-vidurkių algoritmas reikalauja, kad būtų apibrėžtas klasterių skaičius iš karto.

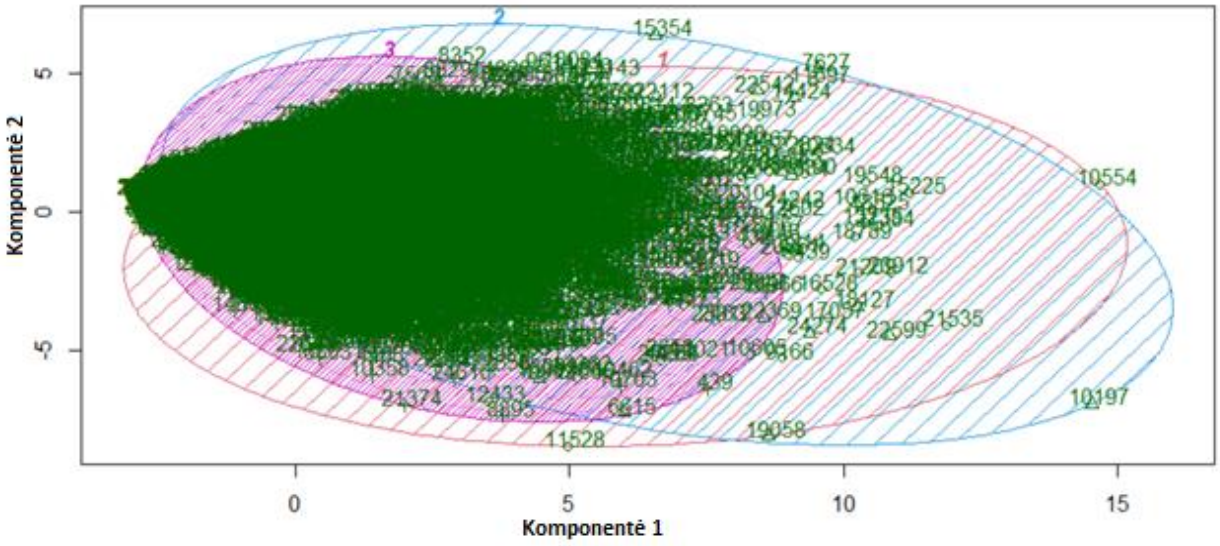
Šiuo metodu yra bandoma sumažinti atstumą tarp klasterio taškų ir klasterio centroido. Tiesą sakant, sumažinamas kiekis yra klasterio viduje esančių kvadratų suma tarp kiekvieno taško ir vidurkio. Visgi galima tikėtis, kad šis dydis bus maksimalus, kai k = 1, o tada vis mažės, kai klasterių skaičius didės, iš pradžių staigiai, o vėliau ne taip staigiai, kai *k* pasiekia optimalią vertę. Taigi, šiuo tikslu nubraižoma kvadratų sumos grupėse grafikas, remiantis alkūnės metodu. Buvo pasirinkta atvaizduoti klasterių dispersijų sumas nuo 2 iki 15 klasterių.



11 pav. Klasterio dispersijos sumos kreivė (Alkūnės metodas)

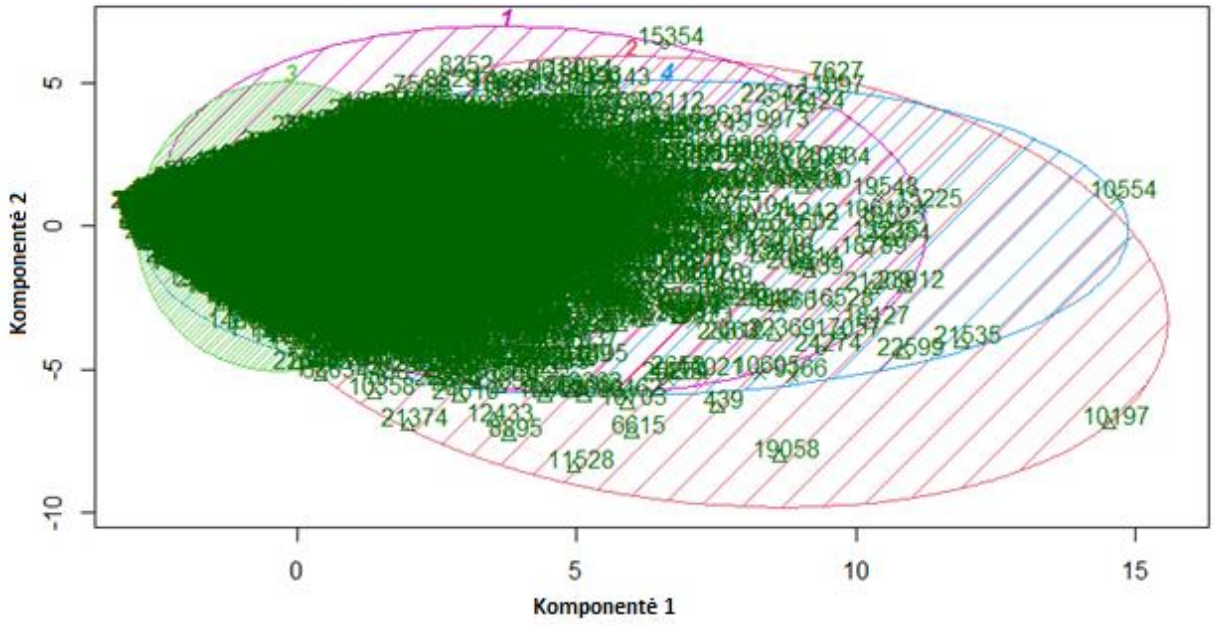
Šis grafikas parodo, kad nėra tokio aiškiai matomo klasterių skaičiaus, kuriam dispersijos suma susilygintų, kitaip tariant grafike nėra aiškios „alkūnės“. Būtent todėl šiuo atveju šis metodas nelabai nepadeda. Galima truputį įžvelgti, kad optimalus klasterių skaičius turėtų būti 3 arba 4. Tyrime toliau nagrinėjami būtent 3 ir 4 klasterių skaičius ir lyginami tarpusavyje.

Toliau nubraižomas *clusplot* grafikas. Rezultatas nubraižomas panaudojant programinės įrangos R bibliotekos *cluster* funkciją *clusplot()*. Sumažinus dimensijų skaičių iki 2, atvaizduojamas 3 klasterių pasiskirstymas. Klasteriai grafike pavaizduojami elipsės formomis.



12 pav. Pagrindinių komponenčių grafikas (k=3)

Pagal gautą pagrindinių komponenčių grafiką, matoma, kad visi 3 klasteriai persidengia tarpusavyje, o taip yra todėl, kad visuose klasteriuose yra pasikartojančių žodžių. Toliau pavaizduojamas analogiškas grafikas tik pasirenkant 4 klasterius.



13 pav. Pagrindinių komponenčių grafikas (k=4)

Pagal gautą PCA grafiką, kai klasterių skaičius yra lygus 4, matoma, kad visi 4 klasteriai persidengia tarpusavyje, visuose klasteriuose yra nemažai pasikartojančių žodžių.

Tam, kad galutinai įsitikinti koks yra tinkamiausias klasterių skaičius, panaudojant programinės įrangos R funciją *cat()*, išvardiname top 5 žodžius klasteriuose. Pirmiausia, pradedama nuo didesnio klasterio skaičiaus – 4.

2 lentelė. Populiausi 5 žodžiai esantys klasteriuose (k=4)

Klasterio Nr.	**Žodžiai esantys klasteryje**
Klasteris Nr. 1	„organic“ „place“ „host“ „beautiful“ „great“
Klasteris Nr. 2	„place“ „organic“ „great“ „recommend“ „clean“
Klasteris Nr. 3	„organic“ „comfortable“ „location“ „nice“ „bed“
Klasteris Nr. 4	„great“ „organic“ „place“ „location“ „host“

Pagal gautus rezultatus, pastebime, kad visi klasteriai yra labai panašūs ir persidengiantys, kas jau buvo pastebėta grafikuose, todėl bandoma perskaičiuoti rezultatus su 3 klasteriais.

3 lentelė. Populiausi 5 žodžiai esantys klasteriuose (k=3)

Klasterio Nr.	**Žodžiai esantys klasteryje**
Klasteris Nr. 1	„great“ „organic“ „place“ „recommend“ „clean“
Klasteris Nr. 2	„organic“ „host“ „place“ „beautiful“ „location“
Klasteris Nr. 3	„great“ „organic“ „place“ „location“ „host“

Toliau atvaizduojami atstumai tarp klasterio taškų ir centro.

```
  great local organic restaurants comfortable location place shops clean helpful  host kitchen loved perfect welcoming
1 0.494 0.238   0.923       0.234       0.339    0.406 2.537 0.220 0.452   0.248 0.433   0.248 0.271   0.336     0.217
2 0.314 0.161   0.880       0.159       0.284    0.333 0.369 0.143 0.298   0.163 0.371   0.157 0.215   0.236     0.207
3 2.278 0.251   0.946       0.419       0.416    0.647 0.716 0.353 0.401   0.257 0.530   0.334 0.253   0.279     0.221
  amazing beautiful coffee  room   bed breakfast enjoyed  good recommend wonderful  food fresh  nice quiet close space
1   0.287     0.432  0.244 0.289 0.303     0.212   0.342 0.308     0.499     0.263 0.242 0.196 0.401 0.231 0.259 0.203
2   0.194     0.364  0.172 0.191 0.206     0.196   0.240 0.180     0.303     0.218 0.152 0.173 0.254 0.157 0.170 0.181
3   0.200     0.318  0.304 0.318 0.362     0.185   0.295 0.325     0.432     0.240 0.289 0.165 0.397 0.231 0.319 0.272
```

```
  great local organic restaurants comfortable location place shops clean helpful  host kitchen loved perfect welcoming
1 0.312 0.136   0.846       0.108       0.176    0.264 0.363 0.102 0.238   0.130 0.327   0.097 0.212   0.203     0.178
2 0.494 0.231   0.915       0.219       0.305    0.387 2.570 0.204 0.423   0.239 0.409   0.225 0.271   0.331     0.212
3 0.459 0.276   1.017       0.424       0.724    0.666 0.466 0.375 0.560   0.317 0.586   0.438 0.235   0.371     0.319
4 2.490 0.228   0.923       0.349       0.339    0.575 0.751 0.279 0.361   0.222 0.488   0.272 0.248   0.250     0.200
  amazing beautiful coffee  room   bed breakfast enjoyed  good recommend wonderful  food fresh  nice quiet close space
1   0.182     0.316  0.133 0.100 0.107     0.175   0.199 0.120     0.267     0.195 0.127 0.154 0.168 0.121 0.136 0.145
2   0.284     0.423  0.230 0.244 0.264     0.208   0.325 0.283     0.490     0.256 0.232 0.190 0.367 0.218 0.249 0.190
3   0.245     0.533  0.355 0.591 0.626     0.273   0.420 0.461     0.476     0.320 0.288 0.246 0.627 0.317 0.328 0.341
4   0.192     0.290  0.265 0.239 0.283     0.171   0.255 0.259     0.396     0.213 0.257 0.150 0.322 0.200 0.291 0.239
```

14 pav. Atstumai tarp klasterių taškų ir centro, kai k=3 ir k=4

Taigi, atlikus klasterizavima k-vidurkių metodu, buvo pasirinkta suskirstyti žodžius į 4 klasterius. Pastebėta, kad tiek į 3 tiek į 4 klasterius terminai pasiskirsto panašiai, tačiau atstumai tarp klasterių taškų ir klasterio centro mažesni yra kai suskirstoma į 4 klasterius (žr. 12 pav.). Beveik visuose klasteriuose yra žodžiai *„place“* ir *„organic“*, o tai yra du dažniausiai pasitaikę žodžiai žaliųjų “Airbnb” vartotojų atsiliepimuose. Visgi, nėra lengva nustatyti kokias konkrečiai temas paliečia kiekvienas klasteris, kadangi jie tarpusavyje yra labai panašūs ir yra vienas terminas *„organic“*, kuris kartojasi visuose klasteriuose.

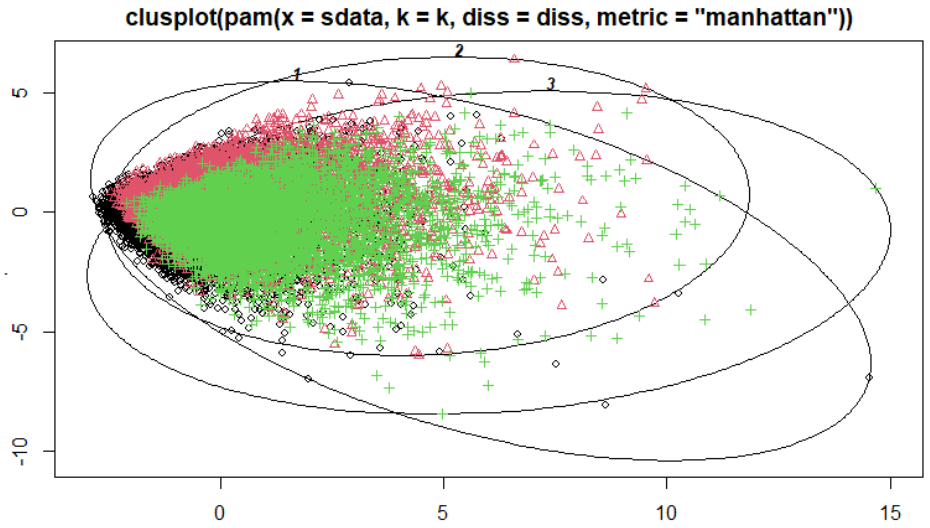
3.2.3. K-medoidų klasterizavimo metodo rezultatai

K-medoidų klasterizavimo metodas yra patikimesnis triukšmui ir išskirtims, nei k-vidurkių klasterizavimo metodas, todėl toliau bus atliekamas klasterizavimas panaudojant būtent šį metodą.

Pirmiausia atliekamas duomenų skaidymasis aplink medoidus, įvertinant klasterių skaičių. Šioje vietoje naudojamas PAM (angl. *Partitioning Around Medoids*) grupavimo aplink medoidus algoritmas. Po pirminio atsitiktinio k medoidų pasirinkimo algoritmas pakartotinai bando geriau pasirinkti medoidus. Tam, kad būtų pagerinta klasterių kokybė, bandoma nustatyti klasterių skaičių diapazoną *krange* = 3:7, panaudojant funkciją *pamk()*. Naudojamas atstumo matas – Manheteno (angl. *manhattan*). Gautas rezultatas – 3 klasteriai.

Klasterio Nr.	**Žodžiai esantys klasteryje**
Klasteris Nr. 1	„organic“
Klasteris Nr. 2	„great“„organic” „location” „place”
Klasteris Nr. 3	„organic“ „comfortable“ „place“ „clean“ „host“ „recommend“

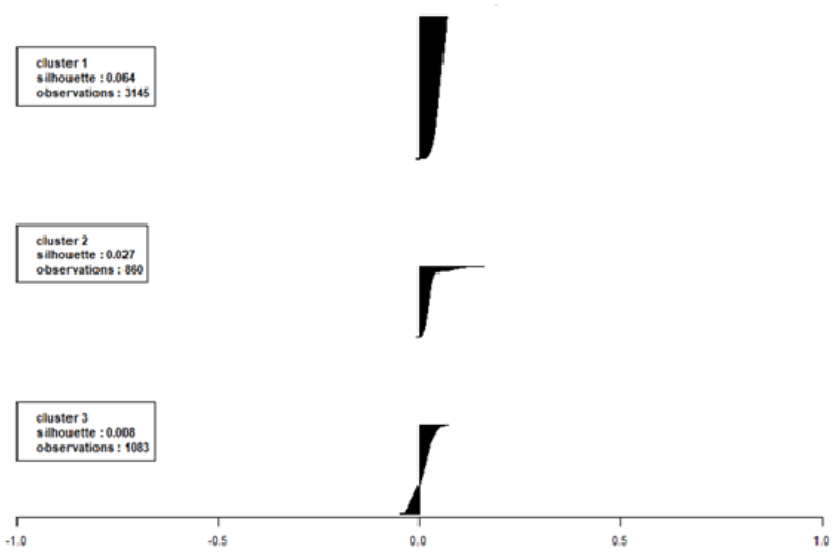
Šis metodas taip pat kaip ir k-vidurkių klasterizavimo metodas padalija n objektų į klasterius taip, kad būtų minimali objektų atstumo iki klasterių centrų kvadratų suma. Skirtingai nuo k-vidurkių algoritmo, k-medoidai pasirenka faktinius duomenų taškus kaip centrus (medoidus) ir taip leidžia geriau suprasti klasterių centrus nei k-vidurkių metodu, kur klasterio centras nebūtinai yra vienas iš įvesties duomenų taškų. Toliau nubraižomas *clusplot* grafikas atvaizduojantis duomenų pasidalinimą į tris klasterius.



15 pav. Clusplot grafikas

Kadangi klasterių skaičius parenkamas savarankiškai, tai klasterių skaičiau tinkamumą galima įvertinti tokiais metodais kaip silueto (angl. *silhouette*) metodas. Naudojant šį metodą ganamas glaustas grafinis vaizdas, kuriame matyti kaip gerai klasifikuotas kiekvienas objektas. Silueto vertė parodo kiek objektas yra panašus į jo klasterį, palyginti su kitais klasteriais. Silueto matas svyruoja nuo −1 iki +1, kur didelė reikšmė rodo, kad objektas yra gerai pritaikytas prie jo pačio klasterio ir

blogai prie kaimyninių klasterių. Jei dauguma objektų turi didelę vertę, grupavimo konfigūracija yra tinkama. Jei daugelio taškų vertė yra maža arba neigiama, grupių konfigūracijoje gali būti per daug arba per mažai klasterių. Siluetas buvo apskaičiuojamas naudojant mahalanobio (angl. *mahalanobis*) atstumo metriką.



16 pav. Silueto analizės grafikas

Pagal gautus silueto analizės rezutatus, buvo pastebėta, kad su įvairiais atstumo matais, nėra gaunami tinkami rezultatai, kadangi silueto matas visais atvejais vidutiniškai buvo intervale nuo -0,1 iki 0,05, jeigu daugelio taškų vertė yra maža arba neigiama, grupių konfigūracijoje gali būti per daug arba per mažai klasterių. Buvo bandyta keisti klasterių kiekį į didesnį ir mažesnį, tačiau silueto matas išliko labai mažas net ir su kitais atstumo matais. K-medoidų metodo trūkumas yra tas, kad jis gerai tinka tik sferiniams klasteriams, yra ganėtinai jautrus išskirtims, todėl šiuo atveju šis metodas nepasiteisino.

3.3. Sentimentų analizė

Šioje darbo dalyje atliekama duomenų rinkinio sentimentų analizė. Nagrinėjama žaliųjų "Airbnb" vartotojų atsiliepimų nuotaika bei emocijos. Sentimentų analizė yra įprasta technika, kuri yra naudojama norint klasifikuoti teksto emocijas. Sentimentų analizė yra teksto klasifikavimo procesas į vieną iš trijų kategorijų: teigiamą, neutralų bei neigiamą. Sentimentų analizės rezultatai taip pat pateikiami ir skaitine skale, siekiant geriau išreikšti teigiamo ar neigiamo nuotaikos stiprumo laipsnį tekste.

Šiame tyrime apskaičiuojami sentimentų balai, naudojantis programinės įrangos R paketu *Syuzhet*, kuriame yra keturi sentimentų žodynai. Paketo *Syuzhet* pagalba analizuojamas tekstas paverčiamas sakinių vektoriumi. Šioje darbo dalyje analizuojamas *Syuzhet* metodas. Toliau šis vektorius nusiunčiamas į funkciją *get_sentiment()*, kuri įvertina kiekvieno žodžio ar sakinio nuotaiką.

```
> head(syuzhet_vector)
[1]  2.05  6.70 12.25  8.10  4.10  4.75
```

17 pav. Syuzhet'o vektorius

Patikrinus Syuzhet‘o vektorių, pirmojo elemento vertė yra 2,05. Tai reiškia, kad visų prasmingų žodžių, esančių pirmame atsakyme (eilutėje) teksto faile, sentimentų balų suma yra 2,05. Nuotaikos balų skalė, naudojant syuzhet‘o metodą, yra dešimtainė ir svyruoja nuo -1 (nurodant labiausiai neigiamą) iki +1 (nurodant labiausiai teigiamą).

Toliau atvaizduojami syuzhet vektoriaus statistinės charakteristikos, panaudojant funkciją *summary()*.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.20    4.35    6.20    6.79    8.55   32.70
```

18 pav. Syuzhet‘o vektoriaus skaitinės charakteristikos

Pagal gautus rezultatus, matoma, kad syuzhet‘o vektoriaus suvestinė statistika rodo medianą lygią 6,20, kuri yra didesnė už nulį ir gali būti interpretuojama kaip bendras visų atsakymų vidutinis nusiteikimas yra teigiamas.

Tada atliekama tokia pati likusių dviejų metodų *bing* ir *afinn* analizė ir patikrinami jų atitinkami vektoriai. Žemiau pateikiamas bing‘o metodo vektorius ir skaitinės charakteristikos.

```
> head(bing_vector)
[1]  2  6 15 11  6  5
> summary(bing_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-19.000   5.000   7.000   7.524  10.000  35.000
```

19 pav. Bing‘o metodo vektorius ir jo skaitinės charakteristikos

Patikrinus bingo‘o vektorių, pirmojo elemento vertė yra 2. Tai reiškia, kad visų prasmingų žodžių, esančių pirmame atsakyme tekstiniame faile, sentimentų balų suma yra 2. Bing‘o metodo reikšmės priklauso dvejetainei skalei, kai -1 rodo neigiamą, o +1 rodo teigiamą nuotaiką. Pagal gautas statistines reikšmes, matoma, kad bingo‘o vektoriaus suvestinė statistika rodo medianos reikšmę 7, kuri yra didesnė už nulį ir gali būti interpretuojama kaip bendras visų atsakymų vidutinis nusiteikimas yra teigiamas.

Toliau pateikiamas afinn‘o metodo vektorius ir jo skaitinės charakteristikos.

```
> head(afinn_vector)
[1]  6 16 34 29  7 10
> summary(afinn_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -36.00   11.00   15.00   16.69   21.00   72.00
```

20 pav. Afinn‘o metodo vektorius ir jo skaitinės charakteristikos

Patikrinus affin‘o vektorių, pirmojo elemento vertė yra 6. Tai reiškia, kad visų prasmingų žodžių, esančių pirmame atsakyme tekstiniame faile, sentimentų balų suma yra 6. Afinn‘o metodo reikšmės yra sveikieji skaičiai nuo -5 iki +5. Pagal gautas skaitines charakteristikas, matoma, kad affin‘o

vektoriaus suvestinė statistika rodo medianos reikšmę lygią 15, kuri yra didesnė už nulį ir gali būti interpretuojama kaip bendras visų atsakymų vidutinis nusiteikimas yra teigiamas.

Taigi, visų trijų metodų vektorių statistika parodo, kad sentimento balų vertė yra didesnė nei 0 (mediana daugiau už 0) ir gali būti aiškinama, kad bendras visų atsakymų vidurkis yra teigiamas. Kadangi šiuose skirtinguose metoduose naudojamos skirtingos skalės, prieš jų palyginimą geriau konvertuoti jų išvestį į bendrą skalę. Šią pagrindinę skalės konversiją galima lengvai atlikti naudojant programinės įrangos R integruotą *sign()* funkciją, kuri visus teigiamus skaičius paverčia 1, visus neigiamus skaičius –1 ir visi nuliai lieka 0.

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    1    1    1    1
[2,]    1    1    1    1    1    1
[3,]    1    1    1    1    1    1
```

21 pav. Normuota trijų vektorių skalė

Pastebima, kad kiekvienos eilutės (vektoriaus) pirmasis elementas yra 1, tai parodo, kad visi trys metodai apskaičiavo teigiamą nuotaikos balą už pirmąjį atsakymą (eilutę) tekste.

Toliau naudojama programinės įrangos R funkcija *get_nrc_sentiments()*, kuri pateikia duomenų lentelę (angl. *data frame*) su kiekviena eilute, vaizduojančia sakinį iš pradinio failo. Duomenų lentelėje atvaizduojami dešimt stulpelių (po vieną stulpelį kiekvienai iš aštuonių emocijų, po vieną stulpelį teigiamiems jausmams ir vieną neigiamiems jausmams. Žemiau pateikiama programinės įrangos R funkcijos *head()* išvestis:

	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	0	0	0	0	0	0	0	0	0	2
2	0	2	0	0	2	0	0	2	1	10
3	1	4	0	0	4	0	0	7	2	12
4	0	2	0	0	3	0	0	1	0	8
5	0	3	0	0	4	0	0	4	0	6
6	0	3	0	0	1	0	0	1	0	4
7	1	6	0	1	7	1	3	9	1	11
8	0	7	0	0	6	0	2	6	1	13
9	0	1	0	0	4	1	1	5	0	8
10	0	0	0	0	0	0	0	0	0	3

22 pav. Pirmų 10 komentarų emocijų klasifikavimas

Taigi, atsiliepimuose esantys žodžiai suklasifikuojami pagal emocijas. Tyrime išskiriamos 8 emocijų rušių, tokių kaip pyktis (angl. *anger*), numatymas (angl. *anticipation*), pasibjaurėjimas (angl. *disgust*), baimė (angl. *fear*), džiaugsmas (angl. *joy*), liūdesys (angl. *sadness*), nuostaba (angl. *surprise*), pasitikėjimas (angl. *trust*), bei dveji jausmai - neigiamas (angl. *negative*) ir teigiamas (angl. *positive*). Pagal gautus emocijų klasifikavimo rezultatus, jau pirmuose 10 komentarų pastebima teigiamų emocijų tendencija. Vos keliuose internetiniuose atsiliepimuose yra žodžių, kurie klasifikuojami kaip neigiami, o labiausiai dominuoja teigiami komentarai.

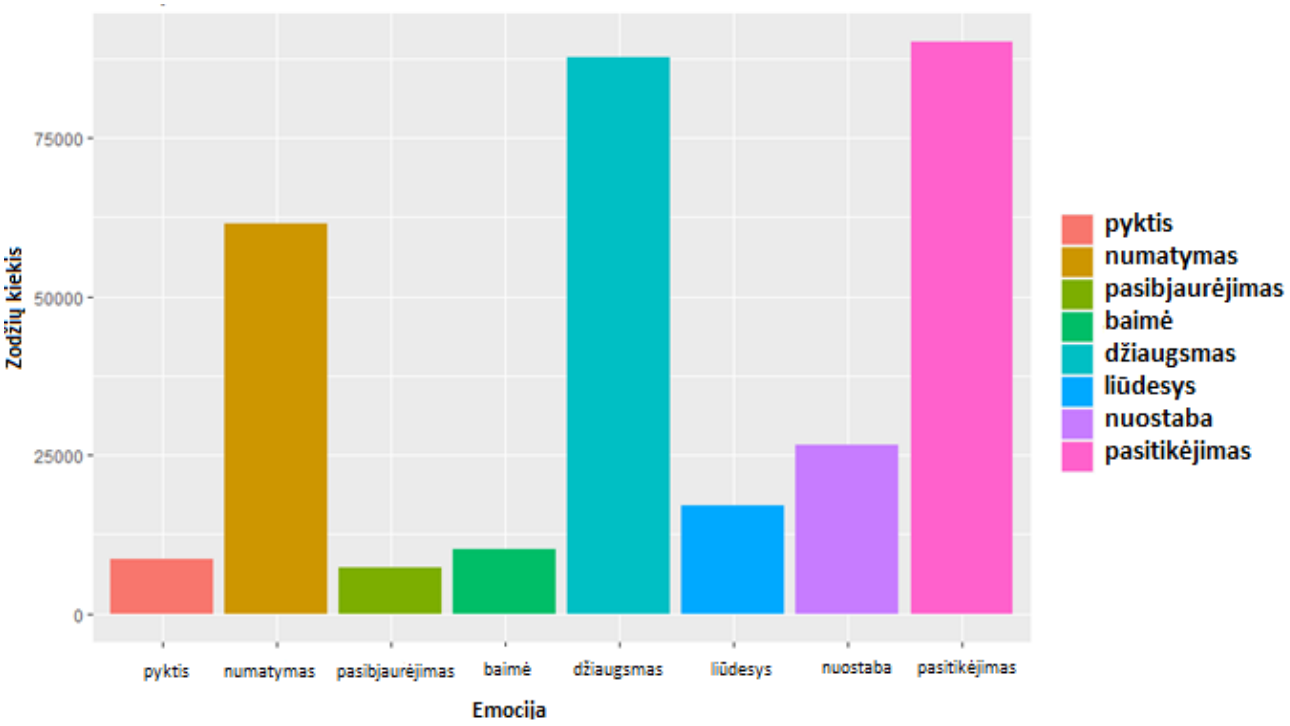
Išvestis rodo, kad pirmoje teksto eilutėje yra:

- Nulis pasitaikiusių žodžių, susijusių su pykčio, pasibjaurėjimo, baimės, liūdesio ir netikėtumo emocijomis;

- Nulis pasitaikiusių žodžių, susijusių su džiaugsmo ir pasitikėjimo emocijomis;
- Iš viso nulis neigiamus jausmus sukeliančių žodžių;
- Du teigiamus jausmus sukeliantys žodžiai.

Kitas žingsnis - sukurti du diagramų grafikus, kurie padėtų vizualiai analizuoti šiame internetinių atsiliepimų tekste pateiktas emocijas. Pirmiausia atliekama keletas duomenų transformavimo ir išvalymo veiksmų prieš braižant diagramas. Funkcija *rowSums()* apskaičiuoja stulpelių sumas eilutėse kiekvienam grupavimo kintamojo lygiui.

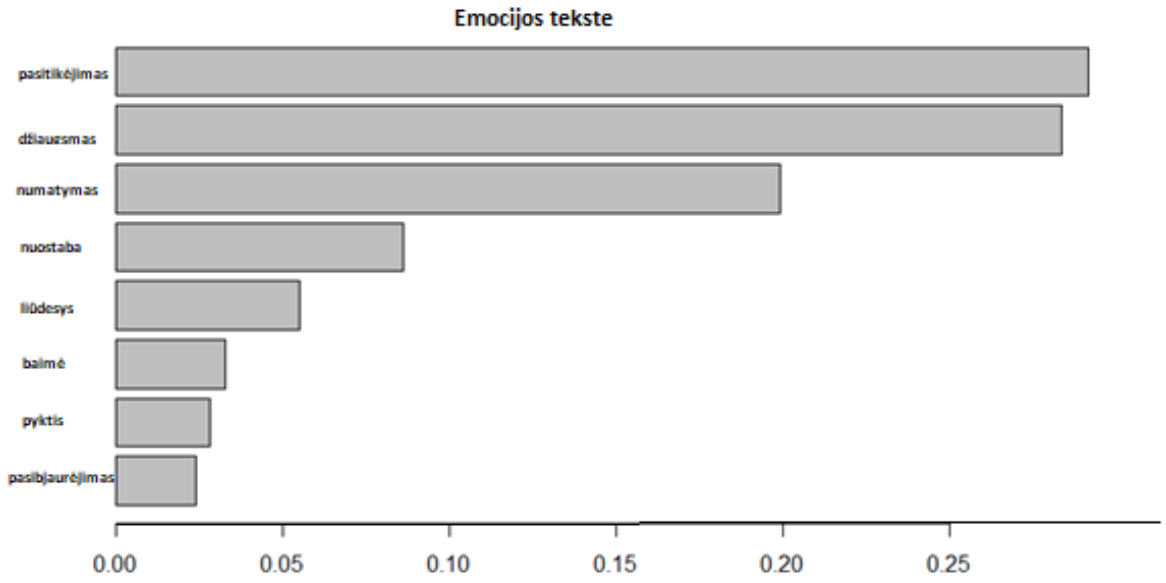
Grafikas žemiau parodo bendrą žodžių skaičių tekste, susietą su kiekviena iš aštuonių emocijų.



23 pav. Su kiekviena emocija susijusių žodžių skaičius tekste

Ši juostinė diagrama parodo, kad žodžiai, susiję su teigiama „pasitikėjimo“ emocija, tekste atsirado maždaug 90.000 kartų, o žodžiai, susiję su neigiama „pasibjaurėjimo“ emocija – apie 7000 kartų. Grafike aiškiai išsiskiria trys emocijos – numatymas, džiaugsmas ir pasitikėjimas, jos žaliųjų “Airbnb” vartotojų internetiniuose atsiliepimuose įžvelgiamos dažniausiai.

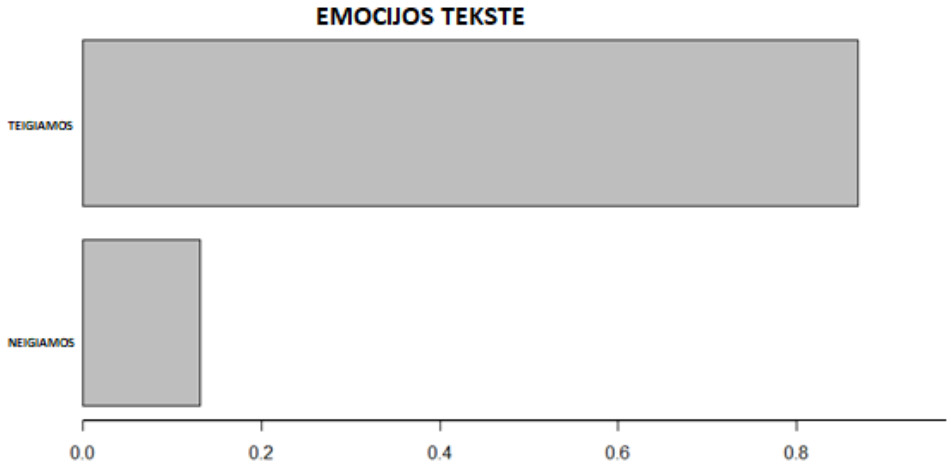
Tam, kad būtų galima turėti gilesnį supratimą apie visas apklausos atsakyme kylančias emocijas, šie skaičiai palyginami procentais nuo viso prasmingų žodžių skaičiaus.



24 pav. Žodžių, susijusių su kiekviena nuotaika, skaičius, išreikštas procentais

Ši diagrama leidžia daug paprasčiau palyginti žodžių, kurie yra susiję su kiekviena emocija užimamą dalį tekste. Galima pastebėti, kad pasitikėjimo emocija turi ilgiausią juostą ir tai parodo, kad žodžiai, kurie yra susiję su šia teigiama emocija sudaro šiek tiek daugiau nei 30% visų šiame tekste esančių prasmingų žodžių. Pažvelgus į pasibjaurėjimo emocijos juostą pastebima, kad ši emocija turi trumpiausią juostą, kuri parodo, kad su šia neigiama emocija susiję žodžiai sudaro apie 2,5% visų prasmingų žodžių tekste. Apskritai žodžiai, kurie yra susiję su teigiamomis pasitikėjimo ir džiaugsmo emocijomis, sudaro apie 60% prasmingų teksto žodžių. Mažąją teksto dalį sudaro neigiamas emocijas sudarantys žodžiai.

Toliau nubraižoma juostinė diagrama, kuri atvaizduoja teigiamų ir neigiamų emocijų dalį procentais.



25 pav. Žodžių, susijusių su teigiama ir neigiama emocija skaičius, išreikštas procentais

Pagal gautą diagramą matoma, kad apie 15 proc. visų prasmingų žodžių šiame tekste sudaro neigiamą emociją sukeliantys žodžiai, o apie 85 proc. žodžių sukelia teigiamą emociją. Iš to galima daryti išvadą, kad žalieji "Airbnb" vartotojai yra linkę vartoti teigiamas emocijas sukeliančius žodžius savo internetiniuose atsiliepimuose.

Toliau, panaudojant programinės įrangos R funkciją *count()* išvardinamos emocijos ir jas sukeliantys žodžiai. Iš viso buvo atrinkti 1645 žodžiai, kurie sukelia teigiamus arba neigiamus jausmus.

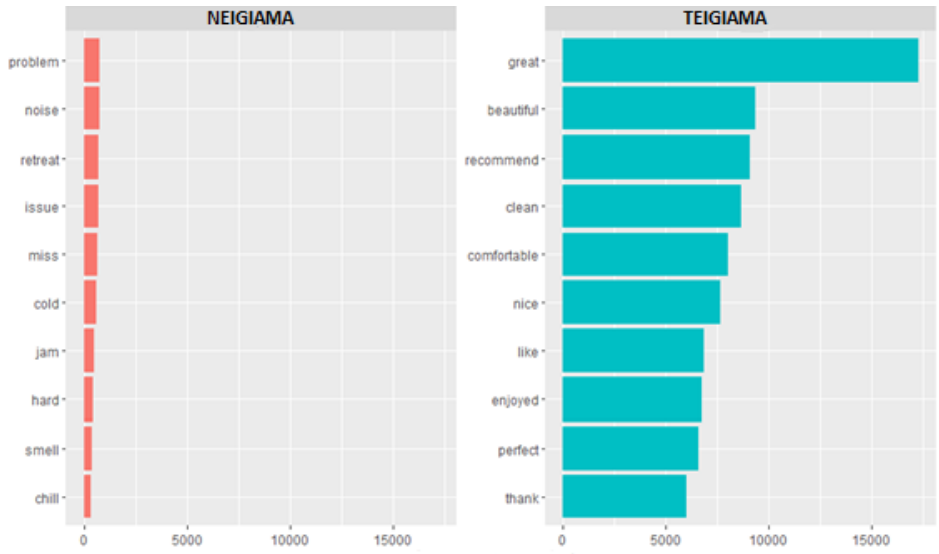
4 lentelė. 10 dažniausiai pasikartojančių žodžių, kurie sukelia emociją

Nr.	**Žodis**	**Emocija**	**Žodžių skaičius tekste**
1.	great	teigiama	17266
2.	beautiful	teigiama	9329
3.	recommend	teigiama	9056
4.	clean	teigiama	8653
5.	comfortable	teigiama	8029
6.	nice	teigiama	7663
7.	like	teigiama	6838
8.	enjoyed	teigiama	6745
9.	perfect	teigiama	6585
10.	thank	teigiama	5974

Pagal gautus rezultatus, matome, kad daugiausia teigiamų emocijų sukėlęs žodis yra *"great"*, jis tekste pasikartojo 17266 kartus. Toliau daugiau nei devynis tūkstančius kartų pasikartojo teigiamą emociją sukeliantis žodis *"beautiful*". Taip pat sąraše pastebimas žodis *"clean",* kuris tekste pasikartojo 8653 kartus ir jis suteikia teigiamą emociją, tai galima daryti išvadą, kad žalieji "Airbnb" vartotojai savo atsiliepimuose akcentuoja švarą ir yra ja patenkinti, taip pat, daugiau nei aštuonis tūkstančius kartų pasikartojo žodis "*comfortable*", kas parodo, kad vartotojams teigiamas emocijas sukėlė patogumai. Taip pat pastebimi tokie žodžiai kaip *"like"* ir *"enjoyed"*, taip pat daugiau nei šešis tūkstančius kartų buvo pavartotas žodis *"perfect",* kas parodo, kad vartotojai paslaugomis buvo labai patenkinti. 5974 vartotojai savo atsiliepimuose panaudojo teigiamą emociją sukeliantį žodį *"thank",* o tai parodo, kad vartotojai yra dėkingi ir net 9056 buvo pakartotas žodis *"recommend"*, tai parodo, kad žalieji vartotojai rekomenduoja "Airbnb" paslaugas.

Neigiamas emocijas sukeliantys žodžiai šioje lentelėje neatvaizduojami, kadangi jų visame tekste buvo procentaliai nedaug ir visi šie žodžiai atsidūrė lentelės gale.

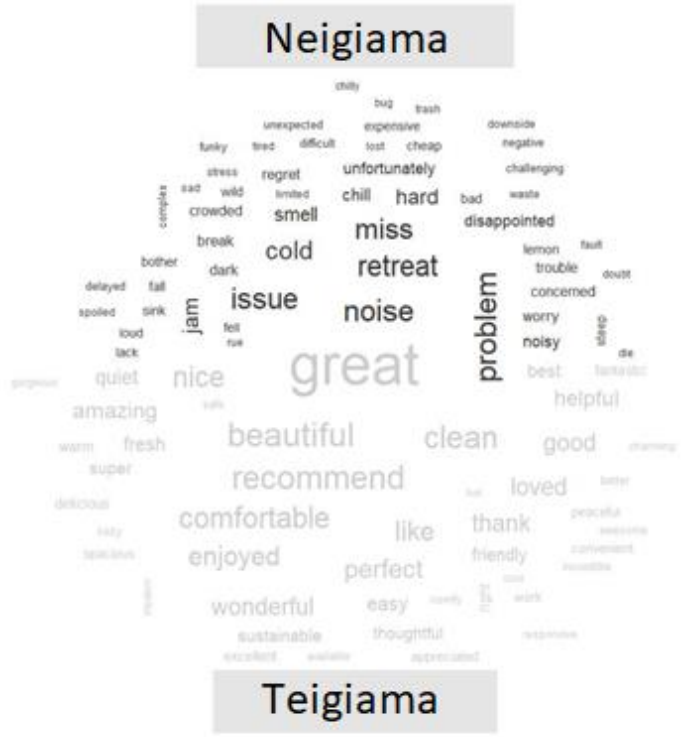
Toliau, panaudojus funkciją *ggplot()* rezultatai atvaizduojami grafiškai.



26 pav. Žodžiai, prisidedantys prie teigiamos ir neigiamos emoocijos

Pagal gautą grafiką, matoma, kad žalieji "Airbnb" vartotojai nepatenkinti yra tokiais dalykais kaip problemos (angl. *problem*), triukšmas (angl. *noise*), sutrikimais (ang. *issue*), kvapu (angl. *smell*) ir pan. Salyginai neigiamas emocijas keliančių žodžių vartotojai naudoja labai mažai, daugiausiai dominuoja teigiamas emocijas keliantys komentarai.

Toliau nubražomas žodžių debesies grafikas, kuriame žodžiai pasiskirsto į teigiamas emocijas keliančius žodžius ir į neigiamas.



27 pav. Dažniausiai pasitaikantys žodžiai suskirstyti į teigiamus ir neigiamus

Žodžių debesyje žodžio dydis reiškia jo dažnumą. Juodai parašyti žodžiai yra teigiamą emociją keliantys, o pilkai – neigiamą. Šiame debesyje pastebima, kad teigiamą emociją vartotojams kelia

tokie dalykai kaip gražumas, draugiškumas, šviežumas, taikumas bei tvarumas. Neigiamas emocijas vartotojams kelia problemos, triukšmas, šaltis, kvapas įvairūs vabalai ir pan.

3.4. Apibendrinimas

Naudojant didžiųjų duomenų analizę, atliekant teksto tyrybą bei sentimentų analizę, buvo gauta reikšmingi rezultatai apie žaliųjų „Airbnb“ vartotojų pageidavimus, įskaitant panašumus ir skirtumus su anksčiau literatūroje nustatytomis išvadomis. Šis tyrimas įrodė akivaizdų pozityvų nusistatymą žaliųjų „Airbnb“ vartotojų komentaruose. Kai žalieji vartotojai yra pilni teigiamų emocijų, jie yra linkę teigiamai vertinti savo patirtį. Būtent ši išvada parodo, kad naujausioje akademinėje literatūroje, skirtoje daugumos „Airbnb“ vartotojų komentarams, nurodoma ta pati teigiama atsiliepimų kryptis.

Vietos faktorius yra tas atributas, kuris turi didžiausią svorį žaliųjų „Airbnb“ vartotojų atsiliepimuose, nes jis skirtas visiems „Airbnb“ vartotojams, tai parodo ir ankstesni tyrimai. Šis tyrimas taip pat nustatė, kad šeimininkas vaidina ganėtinai svarbų vaidmenį žaliųjų „Airbnb“ vartotojų patirčiai, o būtent ši išvada skiriasi nuo ankstesnių tyrimų.

Tyrimas dar kartą patvirtina tvarumo ir šeimininko vaidmens svarbumą žaliesiems „Airbnb“ vartotojams. „Airbnb“ šeimininkai turėtų sutelkti dėmesį į savo ir namų aprašymo pateikimą kuo labiau linkusį į tvarumą, kadangi būtent tai yra aktualu žaliesiems „Airbnb” vartotojams. Skatinant tvarų vartojimą, privatus sektorius vaidina pagrindinį vaidmenį, kuriant su tvaria plėtra susijusią verslo kultūrą, kuri stipriai keičia gamybos modelius, įtraukiant įvarius tvarumo kriterijus.

Išvados

1. Atlikus literatūros analizę „Airbnb“ paslaugų vartotojų pasitenkinimo vertinimo tematika buvo pastebėta, kad viešbučių klientų pasitenkinimo vertinimas mokslinėje literatūroje yra tiriamas daug dažniau nei privataus būto nuomos paslaugos ir mokslinių tyrimų šią tematika yra labai mažai. Tyrimų, kuriais būtų bandyta ištirti žaliųjų „Airbnb“ vartotojų pasitenkinimą, yra vos keletas. Pagrindiniai moksliniuose tyrimuose naudoti metodai buvo teksto tyryba ir sentimentų analizė, taip pat faktorinė analizė bei regresinė analizė.

2. Antroje darbo dalyje buvo pasiūlyta žaliųjų “Airbnb” vartotojų pasitenkinimo vertinimo metodika bei jos realizacija. Tai apima duomenų tvarkymą ir paruošimą tolimesnei analizei, asociacijų analizę, duomenų klasterizavimą Ward‘o, k-vidurkių ir k-medoidų metodais ir sentimentų analizę.

3. Trečioje darbo dalyje buvo atliktas internetinių atsiliepimų klasterizavimas trimis metodais, taip pat atlikta teksto tyryba bei sentimentų analizė. Gauti rezultatai:
   - Pagal gautus atsiliepimuose naudojamų žodžių dažnių rezultatus galima matyti, kad dažniausiai panaudojamas žodis žaliųjų “Airbnb” vartotojų atsiliepimuose buvo *“organic”,* jį visuose žaliųjų vartotojų atsiliepimuose galima buvo pamatyti 22838 kartus. Tai reiškia, kad žalieji vartotojai savo viešnagės metu labiausiai atkreipia dėmesį į ekologiškumą.
   - Pastebėta, kad žalieji „Airbnb” vartotojai daugiausiai savo atsiliepimuose naudoja tokius žodžius kaip ekologiškumas, vieta, šeimininkas, vietovė, švara ir patogumai, tai parodo, kad būtent šie atributai yra svarbiausi žaliesiems „Airbnb” vartotojams.
   - Pagal gautus žodinių asocijacijų rezultatus, pastebima, kad su žodžiu *“organic”* 12 proc. iš visų porų su šiuo žodžiu pasitaiko žodis *“coffee”,* 11 proc. žodis *“fresh”* ir 10 proc. – *“fruit”.* Tai reiškia, kad žaliesiems „Airbnb” vartotojams yra aktualu ekologiški maisto produktai, tokie kaip kava bei vaisiai.
   - Taip pat buvo išnagrinėtos žodžio *“recommend”* asociacijos ir buvo pastebėta, kad daugiausiai su šiuo žodžiu asociacijų turi tokie žodžiai kaip *„highly“,* kas reiškia labai rekomenduoja, taip pat *“place”, “staying”* ir *“apartment”.* Tai reiškia, kad vartotojai yra linkę rekomenduoti vietą bei apartamentus.
   - Atlikus klasterizavimą trimis metodais, pastebėta, kad geriausiai į klasterius suskirsto Ward‘o metodas, tiek k-vidurkių, tiek k-medoidų klasterizavimo rezultatuose klasteriai labai stipriai persidengia.
   - Atlikus sentimentų analizę pastebėta, kad apie 15 proc. visų prasmingų žodžių šiame tekste sudaro neigiamą emociją sukeliantys žodžiai, o apie 85 proc. žodžių sukelia teigiamą emociją. Iš to galima daryti išvadą, kad žalieji “Airbnb” vartotojai yra linkę vartoti teigiamas emocijas sukeliančius žodžius savo internetiniuose atsiliepimuose.
   - Buvo pastebėta, kad žalieji vartotojai savo atsiliepimuose daugiausiai naudoja pasitikėjimo emociją sukeliančius žodžius. Antroje vietoje – džiaugsmo emociją sukeliančius.
   - Atlikus tyrimą pastebėta, kad teigiamą emociją vartotojams kelia tokie dalykai kaip grožis, draugiškumas, šviežumas, taikumas, švara,skanus šviežias maistas, šiluma bei tvarumas. Neigiamas emocijas vartotojams kelia problemos, triukšmas, šaltis, kvapas, stresas, trukdymas, šiukšlės, įvairūs vabalai ir pan.
   - Vietos faktorius yra tas atributas, kuris turi didžiausią svorį žaliųjų „Airbnb“ vartotojų atsiliepimuose, kadangi jis yra skirtas visiems „Airbnb“ vartotojams.

Rekomendacijos

Šis tyrimas turi keletą vadybinių patarimų, kurie yra skirti svetingumo ir apgyvendinimo sektoriaus specialistams ir tyrėjams. Šio tyrimo rezultatai gilinasi į veiksnius, kurie labiausiai lemia žaliųjų vartotojų elgesio modelį, kuris padeda apgyvendinimo sektoriui pagerinti jo siūlomas paslaugas, taip padidinant vartotojų pasitenkinimą. Apgyvendinimo platformoms reikėtų atsižvelgti į šiame tyrime nustatytus atributus bei įdiegti rekomendacijų sistemas potencialiems žaliesiems „Airbnb“ vartotojams. Derinant iš anksto nustatytas paieškos kategorijas, tokias kaip kaina, vieta, patirtis bei patogumai, taip pat pateikiant vartotojų internetinės apžvalgos komentarus ir jų suteiktus įvertinimus, būsimi vartotojai galėtų priimti greitesnius ir patikimesnius sprendimus.

Tyrimas dar kartą patvirtina tvarumo ir būsto šeimininko vaidmens svarbumą žaliesiems „Airbnb“ vartotojams. Skatinant tvarų vartojimą, privatus sektorius vaidina pagrindinį vaidmenį, norint sukurti su tvaria plėtra susijusią verslo kultūrą, keičiančią gamybos modelius, įtraukiant įvairius tvarumo kriterijus. Taigi, nors „Airbnb“ platformoje yra tvarių savybių, kurios yra prieinamos, svarbu įdiegti jos paslaugų klasifikavimo sistemą su tvarumo kriterijais, kad būtų lengviau vartotojams pasirinkti. „Airbnb“ šeimininkai turėtų sutelkti dėmesį į savo ir namų aprašymo pateikimą, kuris būtų kuo labiau linkęs į tvarumą, kadangi tai yra labai aktualu žaliesiems vartotojams.

Literatūros sąrašas

1. KAR HANG LEE C., KEI TSE Y., ZHANG M., Ma J., 2019. Analysing online reviews to investigate customer behaviour in the sharing economy. The case of Airbnb. *Information Technology & People.* [interaktyvus] Vol. 33 No. 3. 2020 pp. 945-961 [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://www.emerald.com/insight/content/doi/10.1108/ITP-10-2018-0475/full/html.
2. JING L., HUDSON S., and KAM FUNG SO K., 2019. Exploring the customer experience with Airbnb. *INTERNATIONAL JOURNAL OF CULTURE* [interaktyvus], TOURISM AND HOSPITALITY RESEARCH VOL. 13 NO. 4 2019, pp. 410-429. [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://www.emerald.com/insight/content/doi/10.1108/IJCTHR-10-2018-0148/full/html
3. SERRAONO L., ARIZA-MONTES A, NADERB M., SIANESC A., LAWD R., 2020. Exploring preferences and sustainable attitudes of Airbnb green users in the review comments and ratings: a text mining approach. *JOURNAL OF SUSTAINABLE TOURISM* [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://doi.org/10.1080/09669582.2020.1838529 .
4. CHENG M., JIN X., 2019. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*. 76 (2019), 58-70 [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0278431917307491
5. VASILIOS PRIPORAS C., STYLOS N., NARASIMHAN VEDANTHACHARI L., SANTIWATANA P., 2017. *International Journal of Tourism Research* [interaktyvus], 19 (6) . pp. 693-704. ISSN 1099-2340 [žiūrėta 2021 m. vasario 7 d.]. DOI:10.1002/jtr.2141.
6. PETRUZZI M., MARQUES C., SHEPPARD V., 2021. TO SHARE OR TO EXCHANGE: An analysis of the sharing economy characteristics of Airbnb and Fairbnb.coop. *International Journal of Hospitality Management* [interaktyvus] 92 (2021) 102724 [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://www.sciencedirect.com/science/article/pii/S0278431920302760
7. YI J., YUAN G., YOO C., 2021. The effect of the perceived risk on the adoption of the sharing economy in the tourism industry: The case of Airbnb. *Information Processing and Management* [interaktyvus] 57 (2020) 102108 [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0306457319301347
8. BENDICKSON J.S., MULDOON J., SOLOMON S.J., 2017. The Sharing Economy and Sustainability: A Case for Airbnb. *Journal Small Business Institute* [interaktyvus], Vol. 13, No. 2, 51-71 ISSN: 1994-1150/69 [žiūrėta 2021 m. vasario 5 d.]. Prieiga per: https://sbij.org/index.php/SBIJ/article/view/265.
9. LUO Y., TANG R., 2019. Understanding hidden dimensions in textual reviews on Airbnb: An application of modified latent aspect rating analysis (LARA). *International Journal of Hospitality Management* [interaktyvus] 80 (2019) 144–154 [žiūrėta 2021 m. vasario 8 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0278431918306856
10. LUO Y., 2018. What Airbnb Reviews can Tell us? An Advanced Latent Aspect Rating Analysis Approach. *Graduate Theses and Dissertations* [interaktyvus] 16403 [žiūrėta 2021 m. vasario 8 d.]. Prieiga per: https://lib.dr.iastate.edu/etd/16403
11. GUNTER U., 2018. What makes an Airbnb host a superhost? Empirical evidence from San Francisco and the Bay Area. *Tourism Management* [interaktyvus] 66 (2018) 26e37 [žiūrėta 2021 m. vasario 9 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S026151771730242X.

12. TUSSYADIAH I.P., 2016. Factors of Satisfaction and Intention to Use Peer-to-Peer Accommodation. *International Journal of Hospitality Management* [interaktyvus], 55, 70-80 [žiūrėta 2021 m. vasario 9 d.]. DOI: 10.1016/j.ijhm.2016.03.005
13. ANTONIDES G., 2017. Sustainable Consumer Behaviour: A Collection of Empirical Studies. *Sustainability* [interaktyvus] 2017, 9, 1686. [žiūrėta 2021 m. vasario 9 d.]. DOI:10.3390/su9101686
14. GUTTENTAG D., SMITH S., POTWARKA L. HAVITZ M., 2017. Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. *Journal of Travel Research* [interaktyvus] 1-18 [žiūrėta 2021 m. vasario 10 d.]. Prieiga per: https://journals.sagepub.com/doi/10.1177/0047287517696980
15. DOMINICI G., GUZZO R., 2010. Customer Satisfaction in the Hotel Industry: A Case Study from Sicily. *International Journal of Marketing Studies* [interaktyvus] Vol. 2, No. 2 [žiūrėta 2021 m. vasario 10 d.]. Prieiga per: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1961959
16. JU Y., BACK K.J., CHOI Y., LEE J.S., 2019. Exploring Airbnb service quality attributes and their asymmetric effects on customer satisfaction. *International Journal of Hospitality Management* [interaktyvus] 77 (2019) 342–352. [žiūrėta 2021 m. vasario 15 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S027843191730909X
17. AKARSU T.N., FOROUDI P., MELEWAR TC., 2020. What makes Airbnb likeable? Exploring the nexus between service attractiveness, country image, perceived authenticity and experience from a social exchange theory perspective within an emerging economy context. *International Journal of Hospitality Management* [interaktyvus] 91 (2020) 102635. [žiūrėta 2021 m. vasario 15 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0278431920301870
18. LI Y., LI B., WANG G., YANG S., 2021. The effects of consumer animosity on demand for sharing-based accommodations: Evidence from Airbnb. *Decision Support Systems* [interaktyvus] 140 (2021) 113430 [žiūrėta 2021 m. kovo 1 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0167923620301858
19. RUAN Y., 2020. Perceived host-guest sociability similarity and participants' satisfaction: Perspectives of airbnb guests and hosts. *Journal of Hospitality and Tourism Management* [interaktyvus] 45 (2020) 419–428 [žiūrėta 2021 m. kovo 3 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S1447677020302151
20. OLTEANU A., CASTILLO C., DIAZ F., KICIMAN E., 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front. Big Data 2:13 [žiūrėta 2021 m. kovo 3 d.]. DOI: 10.3389/fdata.2019.00013.
21. HLEE S., LEE H., KOO C., 2018. Hospitality and Tourism Online Review Research: A Systematic Analysis and Heuristic-SystematicModel. *Sustainability* [interaktyvus], 10, 1141. 13 [žiūrėta 2021 m. kovo 3 d.]. DOI:10.3390/su10041141 www.mdpi.
22. ALAEI A., BECKEN S., STANIC B., 2019. Sentiment analysis in tourism: Capitalising on Big Data. *Journal of Travel Research*. [žiūrėta 2021 m. kovo 4 d.]. Prieiga per: https://doi.org/10.1177/0047287517747753
23. JANILIONIS V. 2019. Išklausyto modulio „Didžiųjų duomenų rinkinių tyrybos metodai" paskaitų medžiaga
24. XIANG Z., SCHWARTZ Z., GERDES JR. J.H., UYSAL M., 2015. What can big data and text analytics tell us about hotel guestexperience and satisfaction? *International Journal of Hospitality Management* [interaktyvus] 44 (2015) 120–130 [žiūrėta 2021 m. kovo 4 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0278431914001698
25. Jo T., 2019. Text Mining. Concepts, Implementation and Big Data Challenge. ISSN 2197-6503 [žiūrėta 2021 m. kovo 8 d.]. Prieiga per: https://doi.org/10.1007/978-3-319-91815-0

26. HASHIMI H., HAFEZ A., MATHKOUR H., 2015. Selection criteria for text mining approaches. Computers in Human Behavior. [žiūrėta 2021 m. kovo 4 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0747563214007201
27. PUKĖNAS K., 2009. Kokybinių duomenų analizė SPSS programa: mokomoji knyga. Kaunas: Lietuvos kūno kultūros akademija. ISBN 9955-622-18-0.
28. ČEKANAVIČIUS V. ir MURAUSKAS G., Taikomoji regresinė analizė socialiniuose tyrimuose. Vilnius: Vilniaus universitetas, 2014. ISBN 9786094593000.
29. KAVALIAUSKAS M. 2020. Išklausyto modulio „ Daugiamatės statistinės analizės modeliai“ paskaitų medžiaga “ paskaitų medžiaga.
30. YONG A.G., PEARCE S., 2013. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology 2013, Vol. 9(2), p. 79-94. [žiūrėta 2021 m. kovo 9 d.]. Prieiga per: https://www.tqmp.org/RegularArticles/vol09-2/p079/
31. ČEKANAVIČIUS, V., MURAUSKAS, G. Statistika ir jos taikymai II. Vilnius, 2006. [žiūrėta 2021 m. kovo 15 d.]
32. JANILIONIS, V., Mokomoji medžiaga „Mokymai apie kiekybinių ir kokybinių HSM tyrimų duomenų analizės metodus“: KORELIACINĖS IR REGRESINĖS ANALIZĖS PAGRINDAI [žiūrėta 2021 m. kovo 15 d.]. Prieiga per LiDA: https://www.lidata.eu/index.php?file=files/mokymai/Janilionis_III/jan_III.html&course_file=jan_III_turinys.html
33. JANILIONIS V. 2020. Išklausyto modulio „Daugiamatės statistinės analizės modeliai“ paskaitų medžiaga.
34. DZEMYDA G, KURASOVA O., ŽILINSKAS J., 2008. Daugiamačių duomenų vizualizavimo metodai. Vilnius: Mokslo aidai, 2008. ISBN 978-9986-680-42-0.
35. ABOUT SAS. [interaktyvus] [žiūrėta 2021 m. kovo 15 d.]. Prieiga per: https://www.sas.com/en_us/company-information.html
36. What is R? [žiūrėta 2021 m. kovo 15 d.]. Prieiga per: https://www.r-project.org/about.html
37. JANILIONIS V., MORKEVIČIUS V., RAULECKAS R. Mokomoji medžiaga: STATISTINĖ KIEKYBINIŲ DUOMENŲ ANALIZĖ SU SPSS IR STATA. [interaktyvus] [žiūrėta 2021 m. kovo 15 d.]. Prieiga per LiDA: http://www.lidata.eu/index.php?file=files/mokymai/stat/stat.html&course_file=stat_turinys.html
38. ABOUT PYTHON [interaktyvus] [žiūrėta 2021 m. kovo 20 d.]. Prieiga per: https://www.python.org/about/
39. List of popular names [interaktyvus] [žiūrėta 2021 m. kovo 25 d.]. Prieiga per: https://www.usna.edu/Users/cs/roche/courses/s15si335/proj1/files.php%3Ff=names.txt.html
40. STOP WORDS [interaktyvus] [žiūrėta 2021 m. kovo 25 d.]. Prieiga per: http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words?fbclid=IwAR0g1QgPGvIRJ7dKHiPcSEHUjcVN3ShXXFju7uqoH4nAneg85L9DzO7wHe4
41. ZHANG F., FLEYEH H.,, WANG X., LU M., 2019. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction* [interaktyvus] 99 (2019) 238–248 [žiūrėta 2021 m. balandžio 8 d.]. Prieiga per: sciencedirect.com/science/article/abs/pii/S0926580518306137
42. ALI N.H., IBRAHIM N.A., 2012. Porter Stemming Algorithm for Semantic Checking. ResearchGate. [žiūrėta 2021 m. balandžio 9 d.]. Prieiga per: https://www.researchgate.net/profile/Noraida-Haji-Ali/publication/260385215_Porter_Stemming_Algorithm_for_Semantic_Checking/links/5584e9d708ae7bc2f448474f/Porter-Stemming-Algorithm-for-Semantic-Checking.pdf

43. Top 10 R Packages For Natural Language Processing (NLP) [interaktyvus]. [žiūrėta 2021 m. balandžio 9 d.]. Prieiga per: https://analyticsindiamag.com/top-10-r-packages-for-natural-language-processing-nlp/
44. GGPLOT2, [interaktyvus] [žiūrėta 2021 m. balandžio 9 d.]. Prieiga per: https://www.r-graph-gallery.com/ggplot2-package.html
45. STABINGIENĖ L., 2014. EKONOMETRIKA. [žiūrėta 2021 m. balandžio 9 d.]. Prieiga per: http://www.ilab.lt/stabingiene/sk2_1.html
46. A gentle introduction to cluster analysis using R. [interaktyvus] [žiūrėta 2021 m. balandžio 15 d.]. Prieiga per: https://eight2late.wordpress.com/2015/07/22/a-gentle-introduction-to-cluster-analysis-using-r/
47. VELMURUGAN T., SANTHANAM T., 2010. Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. *Journal of Computer Science* [interaktyvus] 6 (3): 363-368, [žiūrėta 2021 m. balandžio 21 d.]. Prieiga per: https://www.researchgate.net/publication/47554407_Computational_Complexity_between_K-Means_and_K-Medoids_Clustering_Algorithms_for_Normal_and_Uniform_Distributions_of_Data_Points
48. Introduction to the Syuzhet Package, 2020. [interaktyvus] [žiūrėta 2021 m. balandžio 21 d.]. Prieiga per: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html?
49. NRC Word-Emotion Association Lexicon. [interaktyvus] [žiūrėta 2021 m. balandžio 21 d.]. Prieiga per: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
50. MOHAMMAD S., 2015. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. Emotion Measurement [interaktyvus] [žiūrėta 2021 m. balandžio 25 d.]. Prieiga per: https://www.saifmohammad.com/WebDocs/emotion-survey.pdf
51. GRIGALIŪNAITĖ V., PILELIENĖ L. Vartotojų pasitenkinimo Kauno miesto picerijomis vertinimas. ISSN 2029-9370. *Regional Formation and Development Studies* [interaktyvus], No. 2 (7) [žiūrėta 2021 m. gegužės 15 d.]. Prieiga per: https://core.ac.uk/download/pdf/233176137.pdf
52. EUROPOS VARTOTOJŲ PASITENKINIMO VALDYMO VADOVAS, 2010. Lietuvos Respublikos vidaus reikalų ministerija [interaktyvus] [žiūrėta 2021 m. gegužės 15 d.]. Prieiga per: https://vakokybe.vrm.lt/index.php?id=525
53. LIANG L. J., CHOI H. C., JOPPE M. Exploring the relationship between satisfaction, trust and switching intention, repurchase intention in the context of Airbnb. *International Journal of Hospitality Management.* [interaktyvus] [žiūrėta 2021 m. gegužės 15 d.]. Prieiga per: https://www.sciencedirect.com/science/article/abs/pii/S0278431916302389
54. HE P., HE Y., Xu F., 2018. Evolutionary Analysis of Sustainable Tourism. *ResearchGate.* [interaktyvus] [žiūrėta 2021 m. gegužės 15 d.]. Prieiga per: http://eprints.lincoln.ac.uk/id/eprint/32134/1/Annals%202018.pdf
55. YUNIATI N., PRIYANTO S.H., SUHARTI L., KUSUMA L., 2020. Loyalty of Green Tourist : Mediating Role of Satisfaction. *E-Journal of Tourism* [interaktyvus] Vol.7. No.1. (2020): 114-125 [žiūrėta 2021 m. gegužės 15 d.]. Prieiga per: https://pdfs.semanticscholar.org/012a/b1fba3fedd18a624a7a963edaa7f3710e6f2.pdf

Priedai

1 priedas. Dažniau nei 500 kartų pasikartojančių žodžių sąrašas

Žodis	Dažnumas	Žodis	Dažnumas	Žodis	Dažnumas	Žodis	Dažnumas
organic	22838	gave	1848	attention	1032	heat	707
place	19532	amenities	1830	bottle	1032	value	704
great	17266	sure	1818	took	1018	retreat	704
host	10384	block	1816	information	1014	ride	703
location	10136	couple	1797	shampoo	993	deck	701
beautiful	9329	wine	1790	pool	993	ocean	701
recommend	9056	flat	1771	train	990	holiday	696
clean	8653	milk	1753	downtown	985	surprise	692
comfortable	8029	light	1747	wish	981	issue	691
nice	7663	new	1723	year	974	ive	687
enjoyed	6745	cozy	1680	try	972	totally	681
perfect	6585	access	1680	overall	971	green	681
bed	6305	available	1676	homemade	971	dining	677
loved	5853	spot	1669	met	968	situated	675
wonderful	5829	outside	1658	privacy	967	luxurious	674
room	5809	towels	1658	transport	966	rest	673
good	5722	surrounded	1650	spent	964	selection	669
restaurants	5479	open	1623	stunning	946	bring	668
welcoming	5371	large	1619	going	943	size	666
amazing	5337	bars	1618	huge	943	chickens	666
close	5325	hospitality	1614	greeted	940	allowed	663
coffee	5241	station	1612	sit	939	bay	660
kitchen	5134	extra	1607	options	935	ideal	659
space	5091	friends	1574	real	930	miss	656
breakfast	5003	explore	1574	supplies	930	world	654
helpful	4899	incredible	1564	island	930	owners	653
shops	4857	better	1548	husband	918	mountains	652
food	4826	expected	1541	cabin	917	suite	651
local	4788	fridge	1524	linen	916	month	645
quiet	4616	eat	1513	vegetables	912	subway	645
experience	4478	comfy	1505	bakery	905	second	644
fresh	4452	shared	1498	prepared	905	line	644
beach	4275	quite	1457	wifi	902	choice	644
garden	4038	care	1439	hours	895	mention	644
view	3957	safe	1424	exactly	884	directions	641

best	3714	near	1422	watch	877	children	633
friendly	3699	studio	1420	details	875	checkin	632
street	3519	stocked	1419	washing	875	described	629
visit	3500	car	1403	getting	870	sound	629
super	3487	extremely	1397	respond	865	tourist	626
sustainable	3452	windows	1383	fun	864	addition	625
bathroom	3406	gorgeous	1381	pretty	861	read	620
family	3388	building	1350	hotel	859	stars	616
looking	3344	modern	1347	plan	858	river	603
store	3289	leave	1345	busy	855	daughter	602
city	3263	quality	1329	photos	854	forward	602
thoughtful	3144	old	1326	fabulous	852	cold	592
relaxing	3087	hot	1324	completely	849	bright	588
farm	3085	especially	1323	balcony	840	future	578
park	2899	corner	1319	bike	838	healthy	578
neighborhood	2805	road	1318	pick	838	favorite	576
fruit	2776	trees	1310	neighbourhood	834	attractions	570
communication	2708	quick	1306	tour	834	created	570
warm	2705	sleep	1305	pleasant	833	simple	569
market	2682	dog	1304	longer	832	appointed	567
small	2632	far	1293	let	832	toilet	567
including	2552	cosy	1289	suggestions	830	animals	563
things	2517	theres	1284	perfectly	829	quickly	562
cooking	2505	metro	1276	easily	827	kept	561
fantastic	2434	produce	1266	think	823	base	561
way	2424	bus	1265	showed	821	immediately	559
grocery	2370	meals	1239	came	821	called	558
short	2361	responsive	1234	start	808	added	556
delicious	2350	design	1232	fully	800	answer	556
cafes	2342	charming	1217	airport	793	point	553
excellent	2322	floor	1215	toiletries	789	giving	552
accommodation	2314	mins	1214	steps	789	village	551
property	2285	questions	1214	juice	783	glass	549
spacious	2283	tips	1205	sweet	780	chat	541
absolutely	2282	cool	1205	table	775	oil	541
work	2276	truly	1198	inside	771	slept	538
distance	2261	awesome	1195	different	764	ground	538
shower	2249	air	1188	furnished	762	actually	534
cottage	2249	soap	1179	centre	758	youll	533

natural	2244	main	1172	possible	754	vegan	532
kind	2227	know	1170	hidden	753	hear	530
appreciated	2219	stop	1153	center	752	entire	530
people	2215	outdoor	1149	unit	752	bonus	528
tea	2215	happy	1140	snacks	751	tranquil	525
water	2207	interesting	1140	hand	744	connection	521
town	2198	meet	1130	run	744	note	521
decorated	2130	generous	1111	getaway	743	impressed	520
bedroom	2115	dinner	1108	problem	742	inviting	520
offered	2094	say	1106	noise	742	hiking	518
travel	2087	bath	1106	heart	742	sunset	515
peaceful	2081	central	1104	flowers	740	wasnt	514
supermarket	2072	kids	1100	chocolate	739	knowledge	512
drive	2070	evening	1087	drink	737	loft	511
nearby	2057	wait	1086	birds	736	pleasure	510
convenient	2033	high	1078	basic	731	buy	509
private	2029	weekend	1076	play	730	wed	508
trip	2025	cute	1076	listing	730	talk	508
equipped	2023	stylish	1073	machine	728	plants	507
products	2017	end	1066	taking	723	spotless	505
eggs	1978	delightful	1066	public	723	cat	501
big	1947	taste	1063	style	719	environment	501
door	1924	pictures	1060	spend	715	mind	500
plenty	1905	treat	1058	hill	713		
bread	1879	free	1050	soon	712		
left	1850	life	1042	items	712		

2 priedas. Programinės įrangos R kodas

```
install.packages("tm")  # for text mining
install.packages("SnowballC") # for text stemming
install.packages("wordcloud") # word-cloud generator
install.packages("RColorBrewer") # color palettes
install.packages("syuzhet") # for sentiment analysis
install.packages("ggplot2") # for plotting graphs
install.packages("ROAuth")
install.packages("NLP")
install.packages("Rcpp")
install.packages("fpc")
install.packages("readxl")
install.packages("tidyverse")
install.packages("ClusterR")
install.packages("factoextra")
install.packages("tidytext")
install.packages("dplyr")

library(ROAuth)
library(NLP)
library(Rcpp)
library(tm)
library(SnowballC)
library(fpc)
library(RColorBrewer)
library(wordcloud)
library(ggplot2)
library(syuzhet)
library(readxl)
library(tidyverse)
library(cluster)
library(ClusterR)
library(factoextra)
library(stringr)
library(tidytext)
library(dplyr)
library(reshape2)

temp = list.files(path = "C:/Users/gbajo/Desktop/magistras/duomenys", pattern = "*.csv", full.names = T)
dataframe = data.frame()

skaicius = 0
for (i in temp) {
  tempdata=read.csv(i)
  skaicius=nrow(tempdata)+skaicius
  tempdata=subset(tempdata, grepl("(sustainable|sustainability|organic)", tempdata$comments))
  tempdata$source=i
  dataframe= rbind(dataframe, tempdata)
}
skaicius
nrow(dataframe)

stats <- dataframe %>% count(source)
stats <- stats[order(stats$n, decreasing = TRUE),]
```

```
setwd("C:/Users/gbajo/Desktop/magistras")
write.csv(dataframe, 'filtered.csv')


dataframe <- read.csv(file = 'C:/Users/gbajo/Desktop/magistras/filtered.csv')

if (!require("pacman")) install.packages("pacman") # for package manangement
pacman::p_load("cld2")
pacman::p_load("cld3")
library(cld2)
library(cld3)

cld2 = cld2::detect_language(dataframe$comments)
cld3 = cld3::detect_language(dataframe$comments)
dataframe <- dataframe  %>% filter(cld2 == "en" & cld3 == "en")
myCorpus <- Corpus(VectorSource(dataframe$comments))
# Transform to lowercase
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
# Remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
removeURL <- function(x) gsub("https[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
removeURL <- function(x) gsub("#[A-Za-z0-9]+|@[A-Za-z0-9]+|\\w+(?:\\.\\w+)*/\\S+", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
remove3Dots <- function(x) gsub("[:alpha:]*.", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(remove3Dots))
removeNoEngLettersANDspaces <- function(x) gsub("[^[:alnum:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNoEngLettersANDspaces))
myStopwords <- c(stopwords("en"))
# Remove mystopwords from corpus
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
stop <- read.delim('C:/Users/gbajo/Desktop/magistras/stopwords.txt')
textStopword <-
c("a","about","above","across","after","afterwards","again","against","all","almost","alone","along","already","also","al
though","always","am","among","amongst","amoungst","amount","an","and","another","any","anyhow","anyone","any
thing","anyway","anywhere","are","around","as","at","back","be","became","because","become","becomes","becoming
","been","before","beforehand","behind","being","below","beside","besides","between","beyond","bill","both","bottom
","but","by","call","can","cannot","cant","co","computer","con","could","couldnt","cry","de","describe","detail","do","
done","down","due","during","each","eg","eight","either","eleven","else","elsewhere","empty","enough","etc","even","
ever","every","everyone","everything","everywhere","except","few","fifteen","fify","fill","find","fire","first","five","for
","former","formerly","forty","found","four","from","front","full","further","get","give","go","had","has","hasnt","have
","he","hence","her","here","hereafter","hereby","herein","hereupon","hers","herself","him","himself","his","how","ho
wever","hundred","i","ie","if","in","inc","indeed","interest","into","is","it","its","itself","keep","last","latter","latterly","
least","less","ltd","made","many","may","me","meanwhile","might","mill","mine","more","moreover","most","mostly"
,"move","much","must","my","myself","name","namely","neither","never","nevertheless","next","nine","no","nobody",
"none","noone","nor","not","nothing","now","nowhere","of","off","often","on","once","one","only","onto","or","other",
"others","otherwise","our","ours","ourselves","out","over","own","part","per","perhaps","please","put","rather","re","sa
me","see","seem","seemed","seeming","seems","serious","several","she","should","show","side","since","sincere","six"
,"sixty","so","some","somehow","someone","something","sometime","sometimes","somewhere","still","such","system"
,"take","ten","than","that","the","their","them","themselves","then","thence","there","thereafter","thereby","therefore","
therein","thereupon","these","they","thick","thin","third","this","those","though","three","through","throughout","thru",
"thus","to","together","too","top","toward","towards","twelve","twenty","two","un","under","until","up","upon","us","v
ery","via","was","we","well","were","what","whatever","when","whence","whenever","where","whereafter","whereas",
```

```
"whereby","wherein","whereupon","wherever","whether","which","while","whither","who","whoever","whole","whom
","whose","why","will","with","within","without","would","yet","you","your","yours","yourself","yourselves")
# Remove mystopwords from corpus
myCorpus <- tm_map(myCorpus, removeWords, c(textStopword))
stopnames <- scan("names.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))
stopnames <- scan("names2.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))
stopnames <- scan("names4.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))
stopnames <- scan("names5.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))
stopnames <- scan("names6.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))
stopnames <- scan("names7.txt", character(), quote = "")
myCorpus <- tm_map(myCorpus, removeWords, c(stopnames))


myCorpus <- tm_map(myCorpus, stripWhitespace)
myCorpus <- tm_map(myCorpus, removePunctuation)

removeDup <- function(x) unique(x)
myCorpus <- tm_map(myCorpus, removeDup)

myCorpus2 <- tm_map(myCorpus, stemDocument)

##################
# WHITESPACE
##################
# remove extra whitespace
myCorpus <- tm_map(myCorpus, stripWhitespace)
# keep a copy of corpus to use later as a dictionary for stem completion
myCorpusCopy <- myCorpus
# myCorpus2 - stemed corpus


uniquedf <- data.frame()
uniquedf <- data.frame(text = sapply(myCorpus, as.character), stringsAsFactors = FALSE)

# issaugojame Pythonui
write.csv(uniquedf, 'myCorpus.csv')

# Python rezultatus nuskaitome (jau stem completed)
unstemmed <- read.csv("C:/Users/gbajo/Desktop/magistras/teksto tyryba/unstemmed.csv")

myCorpus <- Corpus(VectorSource(unstemmed$X0))

textStopword <- c("a","minutes","day","stay","karls","apartment","airbnb", "just","lots", "need",
"feel","time","area","like","night", "want", "staying","stayed","away","late","definitely", "walk","home","house",
"really", "thank", "provided", "make", "easy", "highly", "use", "arrived", "living", "touches", "come", "felt", "right",
"morning", "book", "check", "guests", "person", "didnt", "return", "set", "dont", "got", "its", "bit", "week", "able", "went",
"youre", "asked", "plus")
# Remove mystopwords from corpus2
myCorpus <- tm_map(myCorpus, removeWords, c(textStopword))
```

```
tdm <- TermDocumentMatrix(myCorpus, control=list(bounds = list(global = c(5,Inf))))
tdm
findFreqTerms(tdm, lowfreq=1000)
termFrequency <- rowSums(as.matrix(tdm))
termFrequency <- subset(termFrequency, termFrequency>=1000)
df <- data.frame(term=names(termFrequency), freq=termFrequency)
write.csv(df, 'termFrequency.csv')
ggplot(df, aes(x=reorder(term, freq), y=freq)) + geom_bar(stat="identity") +
  xlab("Terms") + ylab("Count") + coord_flip()
barplot(termFrequency, las=2)
findAssocs(tdm,terms = c("organic", "place", "great"), 0.1)
findAssocs(tdm,terms = c("recommend"), 0.1)
m <- as.matrix(tdm)
# calculate the frequency of words and sort it descendingly by frequency
wordFreq <- sort(rowSums(m), decreasing=TRUE)
# colors
pal <- brewer.pal(9, "BuGn")
pal <- pal[-(1:4)]
# word cloud
set.seed(375) # to make it reproducible
grayLevels <- gray( (wordFreq+10) / (max(wordFreq)+10) )
wordcloud(words=names(wordFreq), freq=wordFreq, min.freq=3,
        random.order=F, colors=pal)

#####################################################
### Clutering Words
#####################################################

tdm2 <- removeSparseTerms(tdm, sparse=0.85)
m2 <- as.matrix(tdm2)
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward.D")
plot(fit)
# cut tree into 7 clusters
rect.hclust(fit, k=7)
(groups <- cutree(fit, k=7))
groups[groups==3]

#####################################################
# k- means Algorithm

m3 <- t(m2)
k<- 4
kmeansResult <- kmeans(m3, k)
# cluster centers
clusplot(as.matrix(m3), kmeansResult$cluster, color=T, shade=T, labels=2, lines=0)

#kmeans – determine the optimum number of clusters (elbow method)
#look for “elbow” in plot of summed intra-cluster distances (withinss) as fn of k
wss <- 2:15
for (i in 2:15) wss[i] <- sum(kmeans(m3,i)$withinss)
plot(2:15, wss[2:15], type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")

# set a fixed random seed
```

```
set.seed(122)
# k-means clustering of tweets
k<- 4
kmeansResult <- kmeans(m3, k)
# cluster centers
round(kmeansResult$centers, digits=3)
#
# To make it easy to fnd what the clusters are about,
# we then check the top three words in every
# cluster.
for (i in 1:k) {
  cat(paste("cluster ", i, ": ", sep=""))
  s <- sort(kmeansResult$centers[i,], decreasing=T)
  cat(names(s)[1:5], "\n")

}


set.seed(122)

####################
##### k-medoids
####################

m3 <- t(m2)

# change back to one graph per page
layout(matrix(1))

# partitioning around medoids with estimation of number of clusters
pamResult <- pamk(m3, krange=3:7, metric="manhattan")

# number of clusters identified
(k<-pamResult$nc)
pamResult <- pamResult$pamobject

# print cluster medoids
for (i in 1:k) {
  cat(paste("cluster", i, ": "))
  cat(colnames(pamResult$medoids)[which(pamResult$medoids[i,]==1)], "\n")
}
# set layout to two graphs per page matrix 2x1
# layout(matrix(c(1,2),2,1)) # set to two graphs per page
# plot clustering result
plot(pamResult, color=F, labels=4, lines=0, cex=.8, col.clus=1,
     col.p=pamResult$clustering)
# change back to one graph per page
layout(matrix(1))

cl_f = Clara_Medoids(m3, clusters =  3, distance_metric = 'mahalanobis', samples = 5, sample_size = 0.2, swap_phase =
TRUE, verbose = F)
Silhouette_Dissimilarity_Plot(cl_f, silhouette = TRUE)

#   SENTIMENTAI
```

```
syuzhet_vector <- get_sentiment(unstemmed$X0, method="syuzhet")
# see the first row of the vector
head(syuzhet_vector)
# see summary statistics of the vector
summary(syuzhet_vector)

# bing
bing_vector <- get_sentiment(unstemmed$X0, method="bing")
head(bing_vector)
summary(bing_vector)
#affin
afinn_vector <- get_sentiment(unstemmed$X0, method="afinn")
head(afinn_vector)
summary(afinn_vector)

#compare the first row of each vector using sign function
rbind(
  sign(head(syuzhet_vector)),
  sign(head(bing_vector)),
  sign(head(afinn_vector))
)

d<-get_nrc_sentiment(unstemmed$X0)
head (d,10)
td<-data.frame(t(d))

#The function rowSums computes column sums across rows for each level of a grouping variable.

td_new <- data.frame(rowSums(td[2:25440]))
#Transformation and cleaning
names(td_new)[1] <- "count"
td_new <- cbind("sentiment" = rownames(td_new), td_new)
rownames(td_new) <- NULL
td_new2<-td_new[1:8,]
#Plot One - count of words associated with each sentiment
quickplot(sentiment, data=td_new2, weight=count, geom="bar", fill=sentiment, ylab="count")+ggtitle("Survey sentiments")

barplot(
  sort(colSums(prop.table(d[, 1:8]))),
  horiz = TRUE,
  cex.names = 0.7,
  las = 1,
  main = "Emotions in Text", xlab="Percentage"
)

barplot(
  sort(colSums(prop.table(d[, 9:10]))),
  horiz = TRUE,
  cex.names = 0.7,
  las = 1,
  main = "Emotions in Text", xlab="Percentage"
)

dtext <- tibble(txt=unstemmed$X0)
```

```
tidy_text <- dtext %>%
  unnest_tokens(word, txt)
bing_word_counts <- tidy_text %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
bing_word_counts
bing_word_counts %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)

tidy_text %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("gray20", "gray80"),
                   max.words = 100)
```

3 priedas. Programos Python kodas

```
import numpy as np
import pandas as pd
import nltk

from collections import defaultdict
import tqdm
from nltk.tokenize import sent_tokenize, word_tokenize

nltk.download('punkt')

df = pd.read_csv('myCorpus.csv')
df_full= df
df = df.head(10)
df = df_full

porter = nltk.stem.PorterStemmer()

word_dic = defaultdict(lambda : [])
stemmed_s = []
for idx, row in tqdm.tqdm(df.iterrows(), total=len(df)):
    token_words=word_tokenize(row["text"])
    stem_sentence = []
    for word in token_words:
        stemmed=porter.stem(word)
        stem_sentence.append(stemmed)
        stem_sentence.append(" ")
        word_dic[stemmed].append(word)
    stemmed_s.append("".join(stem_sentence))

def most_common(lst):
    return max(set(lst), key=lst.count)
most_common(word_dic["apart"])

word_dic_new = defaultdict(lambda : [])
for key, value in word_dic.items():
    word_dic_new[key]=most_common(value)

word_dic_new['apart']

new_array=np.array(stemmed_s)

stemmed_df = pd.DataFrame(new_array)

unstemmed_f=[]
for idx, row in tqdm.tqdm(stemmed_df.iterrows(), total=len(stemmed_df)):
    token_words=word_tokenize(row[0])
    unstemmed_sentence=[]
    for word in token_words:
        unstemmed=word_dic_new[word]
        unstemmed_sentence.append(unstemmed)
        unstemmed_sentence.append(" ")
    unstemmed_f.append("".join(unstemmed_sentence))
```

```
best_array=np.array(unstemmed_f)

unstemmed_df = pd.DataFrame(best_array)

unstemmed_df.to_csv('unstemmed.csv')

porter.stem("sustainable")

word_dic_new["sustain"]
```