VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

MODELIAVIMO IR DUOMENŲ ANALIZĖS MAGISTRANTŪROS STUDIJŲ PROGRAMA

Magistro baigiamasis darbas

Lietuviškų tekstų klasifikavimas naudojant skiemenis

Lithuanian Texts Classification Using Syllables

Agnė Žeruolienė

Darbo vadovas doc. dr. Gediminas Murauskas
(darbo vadovo pedagoginis vardas, vadovo vardas ir pavardė)

Vilnius, 2021

Lietuviškų tekstų klasifikavimas naudojant skiemenis

Santrauka

Greitas svarbiausios informacijos išgavimas iš tekstinių dokumentų yra aktuali problema, kadangi būtent jie sudaro didžiąją dalį sukurtų nestruktūrizuotų duomenų. Lietuvių kalbos sudėtingumas ir skirtingų žodžio formų turėjimas iškeltą uždavinį dar labiau apsunkina. Šis faktas skatina ieškoti tekstą charakterizuojančių elementų, kurie yra paprastesnės struktūros nei žodis. Įvairių metodų pritaikymas panaudojant lietuvių kalbos skiemenis anksčiau nagrinėtas nebuvo.

Šiame darbe tyrinėjamos grožinės literatūros tekstų, parašytų ar išverstų į lietuvių kalbą, skiemenų savybės bei jų panaudojimo tekstų klasifikavimui galimybės. Sudaromas naujas tekstų fragmentų klasifikavimo pagal žanrą algoritmas naudojant dviejų etapų logistinę regresiją. Pradžioje skiemenų šansai modeliuojami naudojant binominę logistinę regresiją. Antrame etape sumodeliuotų šansų charakteristikos ir kiti skiemenų požymiai naudojami klasifikavimui. Sudarytas algoritmas yra palyginamas su kitais klasifikavimo algoritmais.

Raktiniai žodžiai: skiemuo, binominė logistinė regresija, klasterinė analizė, klasifikavimas, susietumo taisyklės.

Lithuanian Texts Classification Using Syllables

Summary

Rapid key information extraction from text documents is a pressing problem, as it is texts that make up the bulk of the unstructured data generated. The complexity of the Lithuanian language and having different word forms complicates the task even more. This fact encourages the search for text-characterizing elements that are simpler in structure than the word. The application of various methods using Lithuanian syllables has not been studied before.

In this work, the syllables properties of Lithuanian (or translated into Lithuanian) fiction texts are explored. The possibilities to use the syllables characteristics for texts classification are investigated. A new algorithm for classifying text fragments by genre is developed using two-stage logistic regression. Initially, syllable odds are modeled using binomial logistic regression. In the second stage, the characteristics of the odds are modeled and other syllable features are used for classification. The developed algorithm is compared with other classification algorithms.

Key words: syllable, binomial logistic regression, cluster analysis, classification, association rules.

TURINYS

Terminų ir santrumpų žodynas

Funkciniai žodžiai	(ang. stop words), įvairūs nereikšmingi žodžiai (jungtukai, ištiktukai, jaustukai, įvardžiai), kurie dažnai pasitaiko tekste, bet nesuteikia jokios informacijos apie patį tekstą ar jo temą.
Priešdėlis, prefiksas	afiksas, kuris eina prieš savarankiškų žodžių – daiktavardžių (pvz., „priemiestis“), būdvardžių (pvz., „apygeris“), veiksmažodžių („suprasti“), prieveiksmių (pvz., „paeiliui“) – šaknį.
Sangrąžos dalelytė	Morfema *-s(i)(-)* paprastai vadinama sangrąžos dalelyte, ją turi sangrąžiniai veiksmažodžiai ir kai kurie iš jų padaryti daiktavardžiai. Nepriešdėlinių veiksmažodžių sangrąžos dalelytė yra formos galinė morfema, o priešdėliniuose veiksmažodžiuose jos vieta po priešdėlių.
Skiemuo	trumpiausias garsinės kalbos vienetas, vienu kartu ištariama garsinė žodžio dalis, kurios pagrindą sudaro balsis arba dvibalsis.
Balsiai	kalbos garsai, kurių pagrindą sudaro muzikiniai tonai: a, e, i, o, u ir jų dariniai.
Priebalsiai	kalbos garsai, kuriuos tariant iškvepiamo oro srovė burnoje sutinka kliūčių.
Fonema	mažiausias kalbos vienetas, turintis skiriamąją reikšmę.

ĮVADAS

Pastaruoju metu, vis mažiau tekstinės informacijos yra pateikiama popieriniu formatu ir intensyviai didėja skaitmenizuotų rašto darbų kiekiai. Portalas „Marketingprofs.com“ skelbia, kad per dieną pasaulyje publikuojama daugiau nei 2 milijonai straipsnių. Tuo tarpu IBM įvertino, kad 80% sukuriamos informacijos yra nestruktūrizuota, ir didžiąją dalį jos sudaro tekstiniai duomenys. Norint iš skaitomo teksto gauti vertę reikia analizuoti ir suprasti, o tai reikalauja daug laiko resursų. Natūralu, kad esant tokiems dideliems duomenų srautams, be informacijų technologijų pagalbos nebeįmanoma identifikuoti aktualių įrašų.

Spartus kompiuterių mokslo, technikos vystymasis ir vis spartesnis duomenų skaitmenizavimas išplėtė kiekybinių metodų, tuo pačiu ir duomenų analizės bei statistikos, panaudojimo galimybes. Natūralios kalbos apdorojimo uždaviniai (automatinis teksto taisymas, teksto atpažinimas ir vertimas, tekstų klasifikavimas ir pan.) plačiai nagrinėjami tiek užsienyje, tiek Lietuvoje. Pavyzdžiui, daugumoje informacinio pobūdžio ir verslo svetainių sutinkame pokalbį inicijuojančią ir palaikančią kompiuterinę programą. Konsultavimo įmonės „Invesp“ atliktas tyrimas atskleidė, kad pokalbių robotai (ang. chatbot) greitina bendravimą su klientais ir gali atsakyti iki 80% kasdienių užklausų.

Nors pasaulyje yra kalbama daugiau nei 7000 skirtingų kalbų, didžiausia ir sparčiausia technologinė pažanga pasiekta anglų kalba paremtose neurolingvistinio programavimo sistemose. Lietuvių kalbos sudėtingumas ir skirtingų žodžio formų turėjimas skatina ieškoti tekstą charakterizuojančių elementų, kurie yra paprastesnės struktūros nei žodis.

Darbo tikslas – išskirti tekstų, parašytų ar išverstų į lietuvių kalbą, skiemenų ir skiemenų grupių požymius. Ištirti jų savybes bei jas panaudoti tekstų klasifikavimui. Analizei naudojamas tekstynas, kurį sudaro 5-8 klasei rekomenduojami lietuvių ir užsienio autorių grožinės literatūros kūriniai. Požymių bei požymių grupių išskyrimui bei charakteristikų skaičiavimui naudojami įvairus duomenų analizės metodai (susietumo taisyklės, klasterinė analizė ir pan.). Proceso tikslas– sumodeliuoti skiemenų patekimo į poeziją šansus; sudaryti tekstų fragmentų klasifikavimo algoritmą (atskiriantį poezijos kūrinių fragmentus nuo prozos) naudojant dviejų etapų logistinę regresiją; palyginti sudarytąjį algoritmą su kitais klasifikavimo algoritmais.

Darbe naudojami duomenys patalpinti SQLite duombazėje, jie apdorojami statistikos paketu R. Pirmoje darbo dalyje pateikiama panašia tema atliktų tyrimų apžvalga. Tiriamojoje dalyje įvairias pjūviais nagrinėjamos skiemenų ir jų sąryšių charakteristikos, sudaromi ir palyginami kūrinių fragmentų klasifikavimo modeliai: logistinė regresija, atraminių vektorių klasifikatorius, atsitiktinio miško modelis, k-artimiausių kaimynų algoritmas, sprendimų medis, išpūstų medžių metodas, naivūs Bayes klasifikatoriai, pateikiami naudotų metodų bei modelių aprašymai.

Kiekvienai skiemens pozicijai atskirai, sudaromi skiemenų šansų patekti į poeziją binominės logistinės regresijos modeliai. Fragmentų klasifikavimo modelyje empiriniai šansai pakeičiami įvertintais. Klasifikatorius pritaikomas originaliems, modeliavimo metu nenaudotiems kūriniams.

1. AUTOMATINIO TEKSTŲ KLASIFIKAVIMO DARBŲ APŽVALGA

Natūralios kalbos apdorojimo (ang. NLP – natural language processing) pritaikymas leidžia sutaupyti laiko analizuojant tekstinius duomenis, padėti priimti sprendimus ir automatizuoti verslo procesus. Daugelyje straipsnių nagrinėjami modeliai ir statistiniai metodai, kurie naudojami turinio vadybos, kontekstinės ir nuomonės paieškos, atsiliepimų apie produktą analizės, šlamšto filtravimo, teksto nuotaikos identifikavimo uždavinių sprendimui.

Teksto klasifikavimas mašininio mokymosi pagalba yra vienas iš plačiai naudojamų NLP pritaikymo būtų. Rini Wongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli (2017) atliko tyrimą, kurio tikslas – rasti tinkamiausią algoritmą automatiniam straipsnių parašytų Indonezijos kalba klasifikavimui. Palyginus Multinominį Naive Bayes (ang. Multinomial Naive Bayes), Multivariate Bernoulli Naive Bayes ir Support Vector Machine metodus nustatyta, kad geriausi rezultatai pasiekiami naudojant Multinomial Naive Bayes klasifikatorių.

Yuchul Jung su kolegomis Hogun Park ir Sung Hyon Myaeng (2006) atsižvelgdami į unikalias tekstų savybes siekė atpažinti tinklaraščių (ang. Blog) nuotaiką. Tinklaraščiuose žmonės savanoriškai gali aprašinėti savo patirtis ir mintis, to pasėkoje, tekstai atspindi ir rašytojo emocijas. Norint tekstus skirstyti pagal nuotaiką susiduriama su naujais iššūkiais, nei klasifikuojant pagal temą. Rašytojo nuotaika gali keistis teksto eigoje, tad nustatyti nuotaiką gali būti sudėtinga net rankiniu būdu. Autoriai pritaikė hibridinį statistine analize ir atraminių vektorių klasifikatoriumi (ang. SVM - Support vector machine) paremtą metodą. Gauti reikšmingi rezultatai keturiems nuotaikų tipams: linksmas, liūdnas, piktas, gąsdinantis.

Naujienos.vu.lt svetainėje (2014, https://naujienos.vu.lt/vilniaus-universitete-emocionalus-tekstas-virsta-kompiuterine-schema/) pasakojama, apie Lino Bukausko su studentais kuriamą, teksto emocionalumą išmatuoti galinčią, sentimentų analizės (angl. sentiment analysis) programą. Idėjos sumanytojai teigia, kad išanalizavus planuojamos išleisti knygos emocinį foną ir palyginus su panašių kūrinių pardavimais, galėtume prognozuoti, ar knyga bus perkama.

Prekybos pramonė susiduria su didžiule skaitmenine konkurencija, kur laimi tie, kurie tiksliausiai ir greičiausiai identifikuoja savo klientų norus ir poreikius. Vartotojų atsiliepimai yra vienas iš informacijos šaltinių, kuriuo tikslingai pasinaudoję prekybininkai gali atitikti pirkėjų lūkesčius. Vertinimai dažniausiai yra tekstinės formos duomenys, kurių viešojoje erdvėje per dieną patalpinama milijonai. Dr. Xing Fang, dr. Justin Zhan (2015) remdamiesi sentimentine jausmų ir nuomonės analize, nagrinėja atsiliepimus apie Amazon.com pardavinėjamus produktus. Atlikta sakinio ir teksto lygio klasifikacija, į teigiamų ir neigiamų žodžių sąrašus įtraukiant ir dažnai pasitaikančias rašybos klaidas. Atsitiktinio miško (ang. Random forest) modelis geriausiai veikia

sakinio lygio klasifikacijoje, tuo tarpu atsiliepimo lygio klasifikacija veikia prasčiau, dėl netikslaus kategorijos „neutralus“ atpažinimo.

Profesorius ir tyrėjas Grigori Sidorov knygoje „Syntactic n-grams in Computational Linguistics“ nagrinėja autoriaus priskyrimo (ang. AA – Authorship Attribution) uždavinį naudojant skiemenis. Naudoti duomenų masyvai Anglų ir Ispanų kalbomis. Anglų kalboje skiemenų naudojimas davė geresnius rezultatus nei žodžių krepšelio (ang. BoW – bag of words) metodas. Ispanų kalboje rezultatai naudojant skiemenis prastesni, tačiau skirtumas labai mažas 0.3% lyginant su BoW.

Svetainėje vdu.lt (2019, https://www.vdu.lt/lt/vdu-mokslininkai-vysto-di-technologiju-sprendimus-lietuviu-kalbai/) skelbiama apie valstybės užsakymu Vytauto Didžiojo universiteto (VDU) mokslininkų kuriamus ir modernizuojamus dirbtinio intelekto sprendimus lietuvių kalbos supratimui. Keletas iš įgyvendinamų sprendimų: automatinis sakytinės kalbos pavertimas į tekstą (transkripcija), santraukų formavimas, įžeidžios kalbos atpažinimas naujienų portalų komentaruose ir socialiniuose tinkluose.

2. LIETUVIŲ KALBOS SKIEMENŲ SAVYBIŲ TYRIMAS

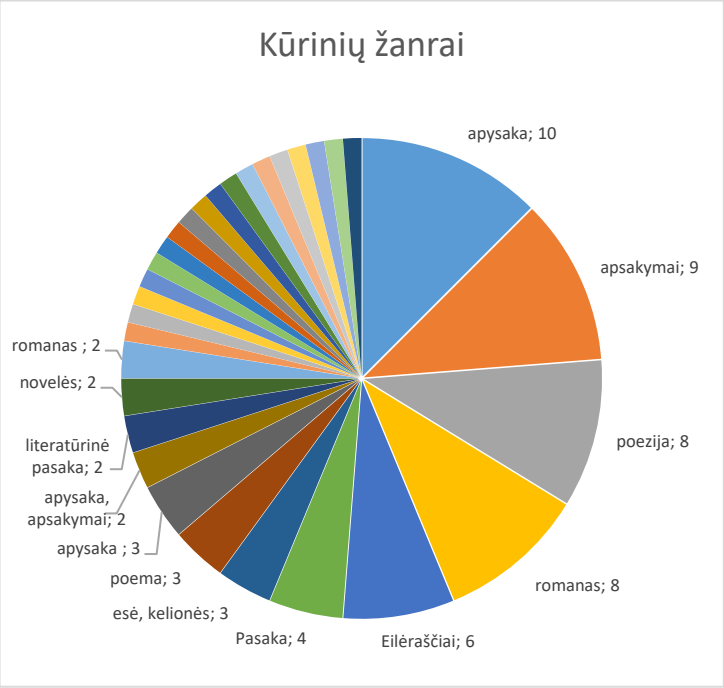
Skiemenys ir jų savybės nagrinėjamos A. Kazlauskienės knygoje „Pirminio lietuvių kalbos ritmo dėsningumai“. Tiriama įvairių skiemenų (ilgųjų, trumpųjų, kirčiuotų, nekirčiuotų) trukmė, centrai, atkarpų nuo vieno balsio (imtinai) iki kito atstumai, vertinama įtaka teksto ritmiškumui. Skiemenų vartoseną ir struktūras taip pat analizuoja kalbininkai V. Karosienė ir A. Girdenis.

Šiame skyrelyje pateikiamas duomenų aprašymas ir nagrinėjamos skirtingiems tekstams būdingos skiemenų charakteristikos. Tikrinamas Zipfo dėsnis skiemenims, nagrinėjamos skiemenų struktūros, jų medis. Išskiriamos dažniausios skiemenų poros pagal žanrus. Tiriami kalbos garsai ir žodžio dalys.

2.1. Duomenų aprašymas

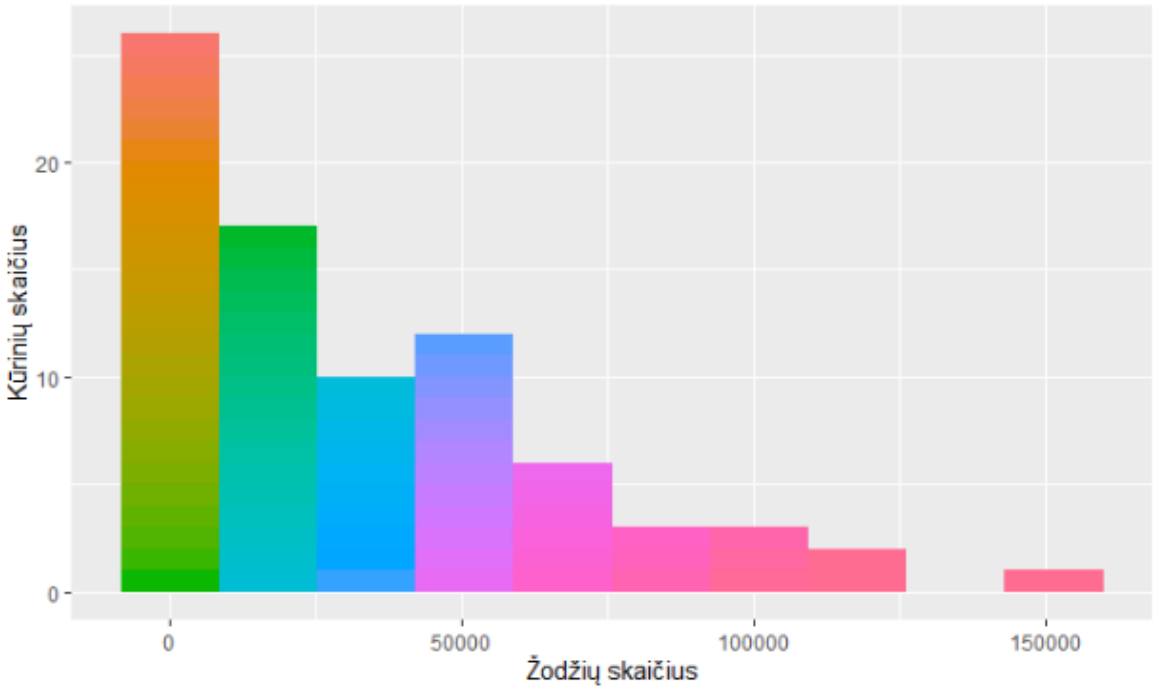
Tyrime naudojama tekstų, parašytų lietuvių kalba, duomenų aibė. Duomenų rinkinį sudaro 5–8 klasių mokiniams rekomenduojami lietuvių ir užsienio autorių grožinės literatūros kūriniai (romanai, apsakymai, poemos, eilėraščiai, pjesės), paimti iš laisvai prieinamos skaitmeninės bibliotekos (http://ebiblioteka.mkp.emokykla.lt/). Duomenų masyvą sudaro 80 kūrinių, kuriuos parašė 64 skirtingi autoriai. 20 iš jų yra versti, o likusiųjų originalo kalba – lietuvių. 1 pav. vaizduojamas kūrinių žanrų pasiskirstymas. Nagrinėjami tekstai apima 31 skirtingą žanrą. Kadangi kategorijų pakankamai daug, o 58% jų turi tik po vieną kūrinį, panašūs žanrai tolimesniame darbe bus sukonkretinti ir apjungiami.

1 pav.: kūrinių žanrų pasiskirstymas



Iš pradinių tekstų išvalytos internetinės nuorodos, standartinė informacija apie skaitmeninimo projektą, leidimo metus, autorių ir kūrinio pavadinimą. Panaikinami skyrybos ženklai, skaičiai, didžiosios raidės paverčiamos mažosiomis. Išvalyti tekstai turi nuo 61 iki 151445 žodžių. Vidutiniškai viename kūrinyje yra 32581 žodis. 2 pav. pateiktoje histogramoje matoma, kad dauguma tekstų yra panašaus ilgio, tačiau yra vienas kūrinys išsiskiriantis ypač didele apimtimi. Tai amerikiečių rašytojo Edgaro Allano Po apsakymų rinkinys „Raudonosios mirties kaukė“.

2 pav.: Tekstų pasiskirstymas pagal žodžių skaičių



2.2. Skiemenų empirinės charakteristikos

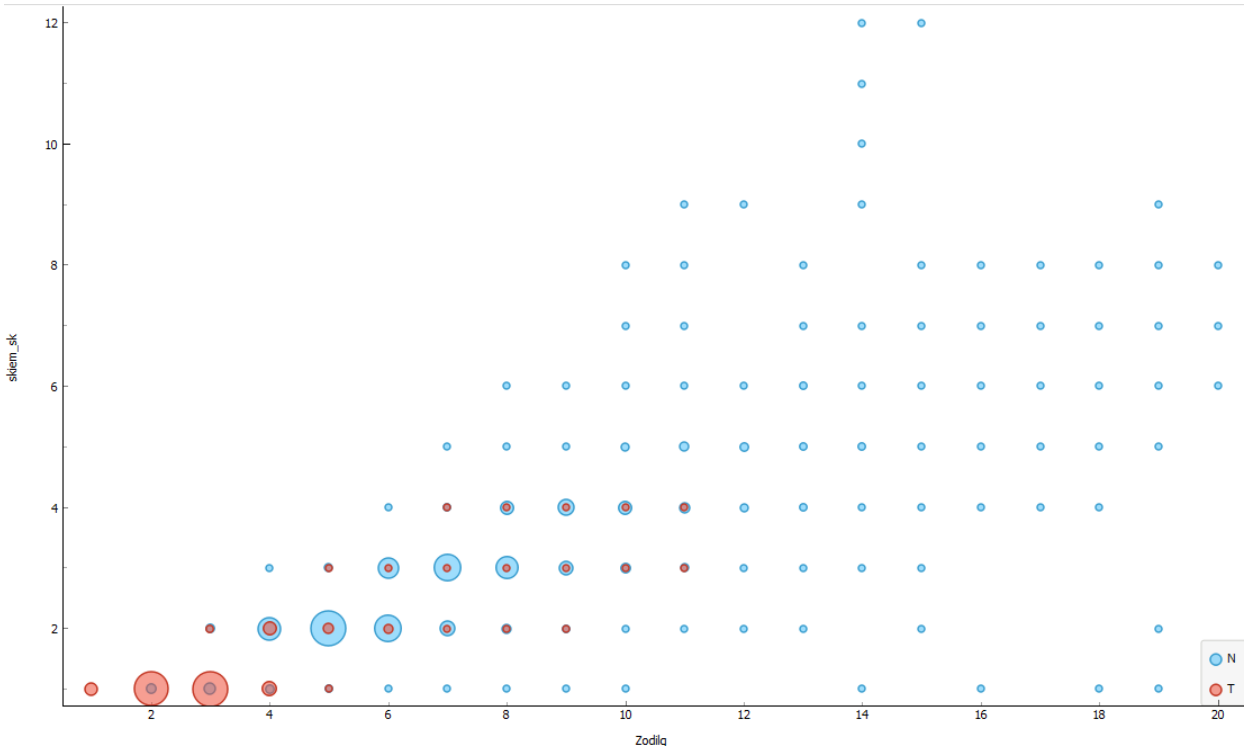
Kalbos garsai jungiasi tarpusavyje ir sudaro tam tikras atkarpas. Toks kalbos srauto junginys, kurio garsai sudaro mažiausią tarimo vienetą, yra vadinamas skiemeniu. Tekstai suskiemenuoti naudojantis python paketu „pyphen“, kuris skiemenuoja žodžius remdamasis Hunspell žodynu. Bibliotekoje naudojamos lietuvių kalbos lentelės, kurias sudarė S. Tolušis ir V. Statulevičius. Didžioji dalis neteisingai suskiemenuotų žodžių pataisyti ir duomenys patalpinti SQLite duombazėje. Gauto duomenų masyvo dimensija 18 stulpelių ir 2606483 eilutės – tiek yra žodžių tekstuose, 215551 iš jų unikalūs. Nagrinėjamuose kūriniuose iš viso 5813065 skiemenys, iš jų 9077 unikalūs. Reiškia skirtingų žodžių ir jų formų yra daugiau nei 23 kartus gausiau nei skiemenų.

Duomenų masyvo kintamieji:

- Kūrinio numeris: nagrinėjama 80 kūrinių, atskiroje lentelėje saugoma informacija apie kūrinį pagal jo numerį;
- Žodis: kiekvienas žodis rašomas naujoje eilutėje, viso 2606483 žodžiai;
- Žodžio pozicija: nurodo žodžio poziciją kūrinyje;
- Žodžio ilgis: ilgiausias žodis turi 20 raidžių, trumpiausias 1. Vidutinis žodžio ilgis – 5;
- Skiemenų skaičius: žodis turi nuo 1 iki 12 skiemenų. Tekstuose yra 8 žodžiai, turintys daugiau nei 9 skiemenis, visi jie ne iš lietuvių kalbos žodyno (pavyzdžiui: mauuuuuuuuuuuu);
- Funkcinis žodis: ar žodis yra funkcinis. Funkcinių žodžių sąrašas paimtas iš R bibliotekos „stopwords“ ;
- 12 stulpelių skiemenims: pirmame stulpelyje – pirmas žodžio skiemuo, antrame – antras ir t.t. Žodžiams, kurie turi mažiau nei 12 skiemenų, tušti stulpeliai užpildomi NULL reikšmėmis.

Naudojantis duomenų tyrybos sistema „Orange“ nubraižyta 3 pav. vaizduojama žodžio ilgio, skiemenų skaičiaus ir žodžių skaičiaus priklausomybės sklaidos diagrama (ang. scatter plot). Raudoni taškai reiškia, kad žodis yra funkcinis, mėlyni – priešingai. Žymenų dydis atspindi kiekį žodžių, turinčių grafike vaizduojamas savybes (nurodytą ilgį ir skiemenų skaičių). Iš grafiko matome, kad lietuvių kalboje nėra žodžių, kurie turėtų vieną raidę ir būtų ne funkciniai. Taip pat, net būdami pakankamai ilgi (11 – 12 raidžių ilgio), dažniausiai lietuvių kalboje pasikartojantys žodžiai turi nedaug skiemenų. Nagrinėjamuose tekstynuose gausiausia 2 – 3 raidžių ilgio ir vieno skiemens funkcinių žodžių. Taip pat, dviskiemeniai 5 raidžių ilgio lietuviški žodžiai. Vizualizacija padeda identifikuoti išskirtis. Žodžiai, turintys daugiau nei 9 skiemenis ir daug raidžių, bet tik vieną skiemenį, tekstyne reti ir nėra taisyklingi lietuvių kalbos žodžiai.

3 pav.: žodžio ilgio ir skiemenų skaičiaus sklaidos diagrama



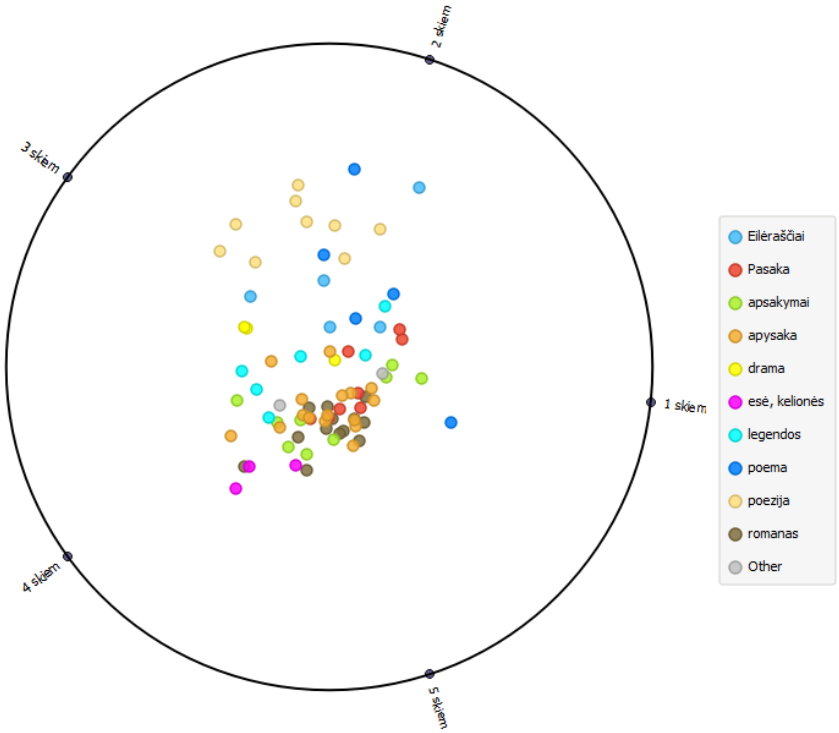
Kiekvienam kūriniui sudaroma procentinė skiemenų skaičiaus žodyje pasiskirstymo lentelė (žr. 1 lentelė).

1 lentelė: skiemenų skaičius žodyje pagal kūrinį

kurinys	1 skiem	2 skiem	3 skiem	4 skiem	5 skiem	6 skiem	7 skiem	8 skiem
1	32,48%	35,42%	21,06%	8,88%	1,90%	0,25%	0,01%	0,00%
2	30,15%	31,89%	25,34%	9,76%	2,52%	0,34%		
3	32,27%	32,08%	22,99%	9,51%	2,71%	0,42%	0,03%	0,00%
4	21,74%	47,58%	24,40%	5,78%	0,46%	0,04%		
5	36,05%	35,00%	21,27%	7,17%	0,49%			
6	32,40%	28,64%	23,39%	12,29%	2,84%	0,41%	0,03%	
7	23,19%	39,21%	28,96%	8,28%	0,33%	0,01%		0,01%
...	...	...	...	...	...	...	...	...
80	28,18%	32,46%	24,60%	11,68%	2,58%	0,44%	0,06%	0,01%

Lentelė pavaizduota daugiamačių duomenų vizualizavimui 2D erdvėje pritaikytame „Radviz“ grafike (žr. 4 pav.). Kūriniai nuspalvinti pagal žanrus. Nors grupes atskirti gana sunku, tačiau galima pastebėti, kad „poezija“ išsiskiria iš kitų grupių didesne, mažai skiemenų turinčių žodžių, koncentracija. Tarp jų išsibarstę eilėraščiai ir poemos. Panašiausią skiemenų ilgių pasiskirstymą tekste, turi „romanų“ kategorija.

4 pav.: skiemenų skaičiaus žodyje Radviz vizualizacija



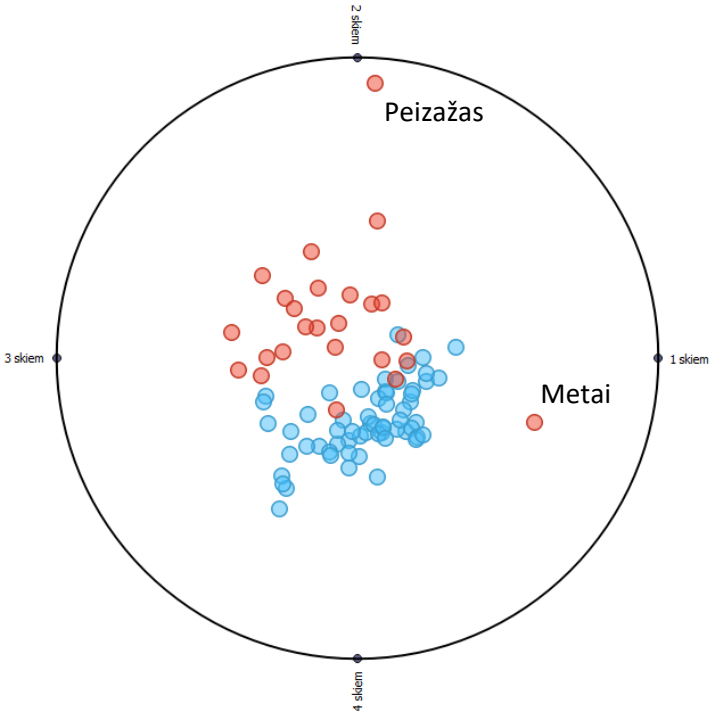
Kūriniai suskirstomi į dvi grupes: poezija ir proza. Poezijai priskiriami žanrams „poezija“, „eilėraščiai“, „poema“ priklausantys tekstai ir sudaroma procentinė skiemenų skaičiaus žodyje pasiskirstymo lentelė (žr. 2 lentelė). Tarp grupių ryškiausiai skiriasi antro, penkto ir šešto skiemenų procentinė dalis.

2 lentelė: skiemenų skaičius žodyje pagal žanrą

žanras	**1 skiem**	**2 skiem**	**3 skiem**	**4 skiem**	**5 skiem**	**6 skiem**	**7 skiem**	**8 skiem**
proza	29,92%	**33,35%**	23,24%	10,54%	**2,51%**	**0,40%**	0,04%	0,003%
poezija	29,07%	**40,05%**	22,76%	7,36%	**0,69%**	**0,06%**	0,004%	0,002%

„Radviz“ vizualizacijoje (žr. 5 pav.) poezijos kūriniai (žymima raudonai) šiek tiek atsiskiria didesne, mažai skiemenų turinčių žodžių koncentracija. Yra keletas poezijos kūrinių, kurie išsiskiria dominuojančiu skiemenų skaičiumi žodyje nuo likusių tekstų.

5 pav.: skiemenų skaičiaus žodyje Radviz vizualizacija



Žodžių debesis yra vienas iš tekstinių duomenų sijonės metodų, kuris padeda išryškinti dažniausiai vartojamus žodžius. Naudojant R biblioteką „wordcloud“ pavaizduoti dažniausiai pasikartojančio pirmo ir antro skiemens porų debesys (žr. 6 pav). Vizualizacijai naudojami ne funkciniai, ilgesni nei vieno skiemens žodžiai, išvalyti nuo priešdėlio ir po jo einančios sangrąžos dalelytės „si“. Kuo šriftas didesnis, tuo skiemuo dažnesnis.

6 pav.: pirmo ir antro skiemens poros debesys

Poezija Proza

Balsių ir priebalsių dažnumas

Balsiai sudaro skiemens pagrindą. Tariant šiuos garsus oras be kliūčių eina pro atvirą burną. Lietuvių kalbos balsiai žymimi šiomis raidėmis: a, ą, e, ę, ė, i, į, y, o, u, ų, ū. Priebalsiai yra tokie

garsai, kuriuos tariant iškvepiamo oro srovė burnoje sutinka kliūčių (P. Kniūkšta, 2004). Lietuvių kalbos priebalsiai: b, c, ch, č, d, dz, dž, f, , h, j, k, l, , n, p, r, s, š, t, v, z, ž. Nagrinėjamuose tekstuose priebalsiniai vienetai C (lot. consonans) 1,09 karto dažnesni už balsius V (lot. vocalis). Atitinkamai 52,1% priebalsių ir 47,9% balsių. Panašius rezultatus C:V = 1,2:1 gavo G. Raškinis ir A. Kazlauskienė publikacijoje „Bendrinės Lietuvių kalbos garsų dažnumas“ (2009).

Žymint priebalsius C, o balsius V galime identifikuoti įvairaus sudėtingumo skiemenų struktūras. Tekstyne sutinkami 188 unikalūs skiemens struktūros variantai. Struktūros nustatymas leido identifikuoti žodžius, kuriuose yra nelietuviškų raidžių, dalį neteisingai suskiemenuotų žodžių ir netaisyklingus žodžius. Jei juos eliminuotume, liktų 71 unikali skiemenų struktūra.

Pačios dažniausios ir atitinkamų struktūrų procentinė dalis bendrai ir pagal pasirinktus žanrus pateikta 3-oje ir 4-oje lentelėse. Prozoje 0.68 procentiniais punktais retesnė CVVC struktūra, tačiau panašiai tiek pat procentinių punktų dažniau sutinkamos struktūros V arba CVVV.

3 lentelė: dažniausios skiemenų struktūros bendrai

Skiemens struktūra	**Skiemenų skaičius**	**Procentinė dalis**
CV	2892593	49,8%
CVC	1039150	17,9%
CVV	621921	10,7%
VC	319003	5,5%
CVVC	174858	3,0%
CCV	157257	2,7%
V	149168	2,6%
CVVV	85888	1,5%
CVCC	81147	1,4%
CCVV	77395	1,3%
CCVC	65934	1,1%

4 lentelė: dažniausios skiemenų struktūros pagal žanrą

Skiemens struktūra	**Proza**	**Poezija**
CV	50,03%	44,20%
CVC	17,77%	20,65%
CVV	10,65%	11,82%
VC	5,50%	5,10%
CVVC	**2,98%**	**3,66%**
CCV	2,70%	2,77%
V	**2,65%**	**2,06%**
CVVV	**1,48%**	**1,38%**
CVCC	1,37%	1,97%
CCVV	1,32%	1,56%
CCVC	1,12%	1,39%

VCC	0,33%	0,37%
VV	0,33%	0,47%

Iš skiemenų struktūrų sudaromi žodžiai, tad į skiemenį galima žiūrėti kaip į medžio lapą. Medžiai ypač naudingi kai dirbama su hierarchiniais duomenimis. Naudojantis R bibliotekomis „treemap“ ir „data.tree“ sudarytas skiemenų medis reprezentuojantis žodžius. Iš viso nagrinėjamuose tekstuose sutinkamos 11292 unikalios žodžių struktūros. 7 paveiksle pavaizduota dalis medžio: atitinkamos struktūros žodžių skaičius ir pavyzdinis žodis.

7 pav.: skiemenų struktūrų medis žodžiams

	levelName	zodziu_skaicius	pavyzdinis_zodis
1	zodziu strukturos	NA	
2	¦--CVC	219169	ten
3	¦ ¦--CV	69045	tikro
4	¦ ¦ ¦--CV	30743	randasi
5	¦ ¦ ¦ ¦--CV	7060	dėbtelėjo
6	¦ ¦ ¦ ¦ ¦--CV	747	sergėdamasi
7	¦ ¦ ¦ ¦ ¦ °--CV	58	persirėdysime
8	¦ ¦ ¦ ¦ ¦--CVC	362	bandavodamas
9	¦ ¦ ¦ ¦ °--CVV	78	persikėlimui
10	¦ ¦ ¦ ¦--CVC	3091	rogvolodas
11	¦ ¦ ¦ ¦ ¦--CV	81	piktadėjysta
12	¦ ¦ ¦ ¦ ¦--CVV	56	piktadėjysčių
13	¦ ¦ ¦ ¦ °--CVC	54	piktadėjiškas
14	¦ ¦ ¦ ¦--CVV	1141	piktadėjai
15	¦ ¦ ¦ ¦ °--CV	127	varžydamiesi
16	¦ ¦ ¦ ¦--CVVV	269	kaštoniniai
17	¦ ¦ ¦ ¦--CVVC	264	tikrutėlius
18	¦ ¦ ¦ °--CVVVC	54	kastuvėliais
19	¦ ¦ ¦--CVC	14411	ginvilos
20	¦ ¦ ¦ °--... 4 nodes w/ 1 sub	NA	
21	¦ ¦ °--... 7 nodes w/ 13 sub	NA	
22	¦ °--... 15 nodes w/ 75 sub	NA	
23	°--... 36 nodes w/ 1372 sub	NA	

Priešdėliai

Lietuvių kalbos žinyne (2007) išskiriami dvejų rūšių priešdėliai:

1) dalelyčių kilmės be-, te-, tebe-, ne-, nebe-;
2) prielinksninių kilmės at- (ata-, ato-), ant- (anta-), ap- (api-, apy-), į- (in-, im-), iš-, nu- (nuo-), pa- (po-), par-, per-, pra- (pro-), pri- (prie-), prieš-, su- (są-, sam-, san-), už- (užu-, užuo-, ūž-)

Žodis gali turėti kelis priešdėlius, einančius vienas po kito. Po priešdėlio gali būti sutinkama sangrąžos dalelytė „si“. Pagal duomenis, pateiktus smulkesnėje žanrų lentelėje (žr. 5 lentelė) ir į „poezija“, „proza“ suskirstytų kūrinių lentelėje (žr. 6 lentelė), galima daryti išvadą, kad poezijoje priešdėliai ir sangrąžos dalelytė vartojama rečiau, tačiau vienskiemenių ne funkcinių žodžių yra daugiau.

5 lentelė: priešdelius ir dalelytę „si“ turinčių žodžių dalis žanruose

Žanras	Sangrąžos dalelytė %	priešdėlis %
Novelės	3,02%	16,06%
literatūrinė pasaka	2,95%	18,07%
Mikrodrama	2,82%	14,59%
Apysaka	2,82%	17,73%
Apsakymas	2,72%	17,66%
Romanas	2,66%	17,36%
Pjesės	1,42%	16,48%
Eilėraščiai	1,17%	13,02%
Poezija	0,97%	13,41%
poezija, eksperimentų knyga	0,73%	10,63%

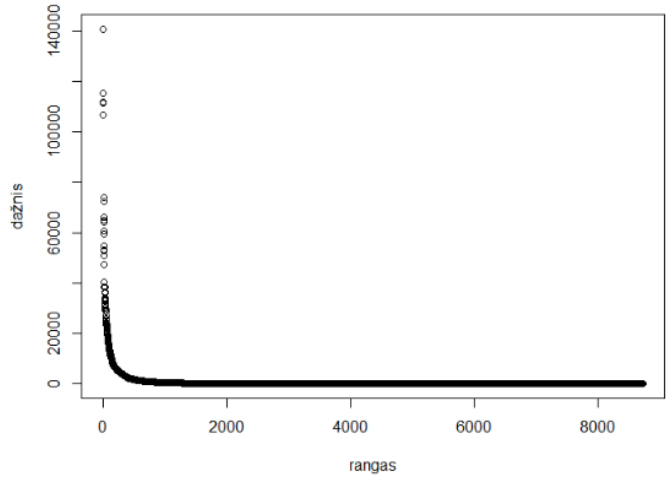
6 lentelė: vienskiemenių ne funkcinių, priešdėlius, dalelytę „si“ turinčių žodžių dalis

Žanras	Priešdėlis %	Sangrąžos dalelytė %	Vienskiemeniai ne funkciniai žodžiai %
Proza	18,51%	2,61%	3,73%
Poezija	14,35%	1,60%	4,95%

Zipfo dėsnis

Nestandartines dažniausių žodžių savybes pirmasis užrašė amerikiečių lingvistas George Kingsley Zipf. Zipfo dėsnis teigia, kad žodžio dažnis yra atvirkščiai praporcingas jo rangui dažnių lentelėje. Tai reiškia, kad antras dažniausias žodis turėtų būti vidutiniškai per pus, trečias – tris kartus retesnis, už dažniausią. Tikriname ar tekstyne naudojamų skiemenų rangų dažniui galioja Zipfo dėsnis, su parametrais a, b, c. Funkcijoje $f\_r = \frac{c}{(r+b)^a}$ parametras *f_r* reiškia rangą *r* turinčio skiemens dažnį. Patikrinsime ar Zipfo dėsnis galioja skiemenims. Paimami visi unikalūs nagrinėjamuose kūriniuose vartojami skiemenys ir jų dažniai. Duomenys išrūšiuojami dažnio mažėjimo tvarka ir sugeneruojamas rango stulpelis. Pasak Zipfo dėsnio, didėjant rangui dažnis staigiai krenta. Dažnio ir rango priklausomybės grafikas elgiasi panašiai pagal dėsnį, tik pirmieji dažniai nekrenta taip staigiai, kaip turėtų.

8 pav.: dažnio ir rango priklausomybės grafikas

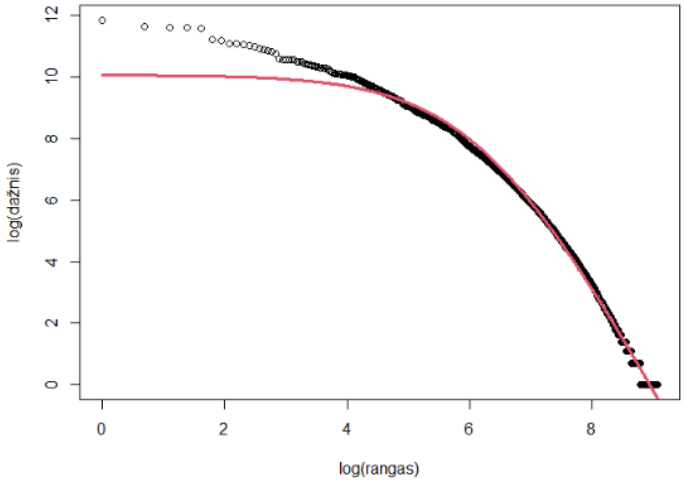


Funkcija logaritmuojama ir mažiausių kavadratų metodu įverinami nežinomi parametrai a, b, c. Visi gauti parametrai yra statitiškai reikšmingi. Gautas modelis:

$$\ln(f) = \ln(2.213 * 10^{14}) - 3.666 * \ln(r + 524.6)$$

$$s.e. \qquad (1.516 * 10^{13}) \quad (7.628 * 10^{-3}) \quad (4.587)$$

9 pav.: dažnio ir rango logaitmo priklausomybės ir įvertinto modelio atitikimas



Rezultatai gali būti vertinami dvejopai. Yra teigiama, kad svarbiausia, jog funkcija su įvertintais parametrais gerai atitiktų žemesnių dažnių elementus. Tada galima teigti, kad Zipfo dėsnis galioja ne tik žodžiams, bet ir skiemenims. Kita vertus, pasak griežto Zipfo dėsnio, parametras $a \approx 1$, o parametras $b \approx 2,7$. Parametrą b pasiūlė įtraukti Mandelbrotas ir išvedė šio dėsnio apibendrinimą Zipfo-Mandelbroto dėsnį, labiau tinkantį dažnio pasiskirstymui kalboje. Tuomet, atsižvelgus į įvertintus parametrus, griežtas dėsnis skiemenims negalioja.

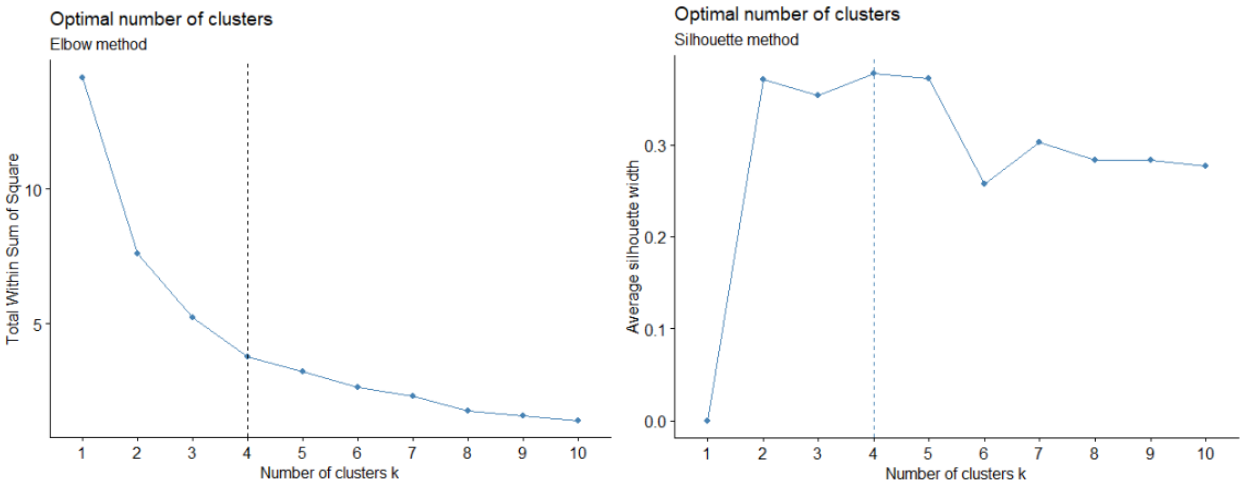
Kūrinių klasterizavimas

Patikrinsim ar pasirinkti skiemenų požymiai atskiria kūrinių grupes pagal žanrą. Kūriniai klasterizuojami remiantis šiais požymiais:

- Vidutinis skiemenų skaičius žodyje
- Standartinis skiemenų skaičiaus nuokrypis
- Funkcinių žodžių dalis tekste
- Vidutinis balsių skaičius žodyje
- Standartinis balsių skaičiaus nuokrypis
- Vidutinis priebalsių skaičius žodyje
- Standartinis priebalsių skaičiaus nuokrypis
- Žodžių, turinčių priešdėlį dalis
- Žodžių, turinčių sangrąžos dalelytę „si“ dalis

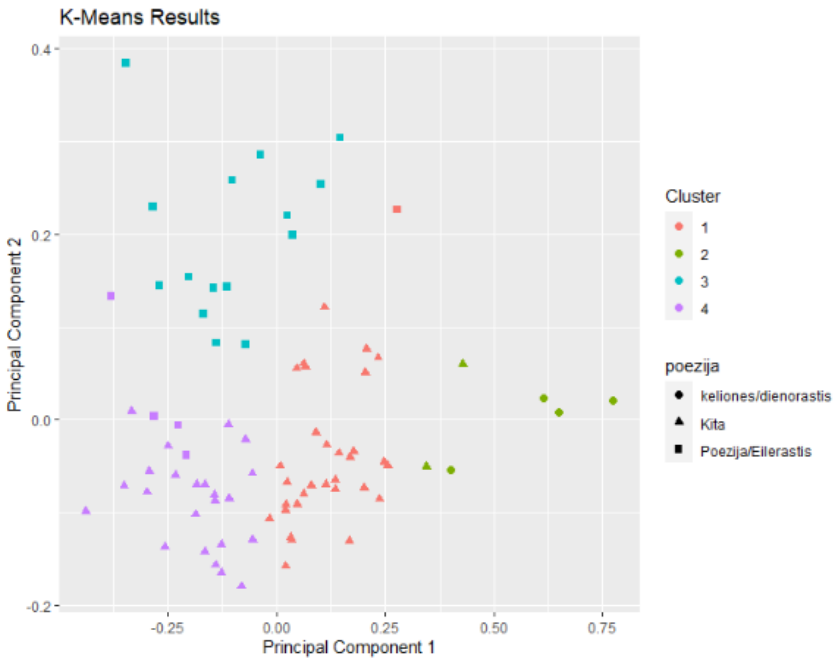
Optimalus klasterių skaičius 4, nustatytas naudojant Elbow ir Silhouette metodus (žr. 10 pav.).

10 pav.: klasterių skaičiaus parinkimas



11 paveiksle vaizduojamas nagrinėjamų tekstų suskirstymas į 4 grupes. Požymiai, kurių klasterių centrai ypatingai panašūs, išimami iš klasterinės analizės. Gautoje pirmoje grupėje dominuoja romanai ir apysakos. Antrame klasteryje identifikuoti visi „esė, kelionės“ ir „Dokumentinė literatūra; dienoraštis“ žanro kūriniai. Trečioje grupėje visi poezijos kūriniai ir didžioji dalis eilėraščių. Ketvirtame klasteryje daugiausia romanų, apysakų ir apsakymų.

11 pav.: Kūrinių klastereriai



Gauti klasterių centrai atspindi identifikuotas žanrų savybes (žr. 7 lentelė). Klasteris, kuriame dominuoja poezija turi mažesnį vidurinį skiemenų skaičių ir jo variaciją. Taip pat, mažiau priešdėlių ir dalelyčių „si“.

7 lentelė: klasterių centrai

skiem_sk_mean	skiem_sk_sd	stop_word_dalis	balses_mean	balses_sd	priebalses_mean	priebalses_sd	si_dalis	priesdeliai2
2.24	1.07	0.30	2.69	1.37	2.94	1.54	0.03	0.18
2.41	1.16	0.27	2.93	1.48	3.18	1.66	0.02	0.18
2.14	0.88	0.24	2.63	1.22	2.94	1.39	0.01	0.13
2.10	1.01	0.33	2.52	1.29	2.75	1.45	0.02	0.17

Ne ten priskirti poezijos kūriniai:

- J. Erlickas „Bilietas iš dangaus arba Jono Grigo kelionė greituoju traukiniu“
- H. Radauskas „Eilėraščiai“
- A. Mickevičius „Gražina“
- K. Donelaitis „Metai“
- J. Vaičiūnaitė „Eilėraščiai“

Pirmi keturi kūriniai priskirti į 4-tą klasterį, o paskutinis į pirmą. Pagal minėtų tekstų charakteristikas, viena iš klaidingo priskyrimo priežasčių – didesnė funkcinių žodžių dalis. Klasterinė analizė prodo, kad atskirti tekstus įmanoma. Panaudoti skiemenų požymiai veikia neblogai.

3. TEKSTŲ FRAGMENTŲ KLASIFIKAVIMAS

Tekstyne yra nedaug poezijos kūrinių. Tam, kad klasifikatorių sudarymo ir testavimo procesas būtų patikimesnis, į duomenų rinkinį įtraukti papildomi penki poezijos kūriniai. Parinkta atsitiktinė 2000 fragmentų imtis – naudojama 1000 poezijos ir 1000 prozos fragmentų, po 150 vienas po kito einančių žodžių. Atsitiktinai parenkamas stebėjimo pradžios taškas su salyga, kad jis nėra tarp paskutinių 149 kūrinio žodžių. Taip išvengiama mažesnių nei 150 žodžių stebėjimų. 75% bendros imties naudojama modeliavimui, 25% modelio testavimui. Modeliavimui naudojamoms kūrinių ištraukoms suskaičiuoti požymiai, nagrinėti ankstesniame skyrelyje. Taip pat, įvertinamos papildomos imties elementus klasifikuoti padedančios charakteristikos.

3.1. Papildomos skiemenų sekų charakteristikos

Susietumo taisyklės

Susietumo taisyklių (ang. association rules) analizė atskleidžia objektų tarpusavio sąsajas. Šią techniką prekybininkai gali pritaikyti prekių – komplektų identifikavimui, ir tikslingai panaudoję informaciją, padidinti pardavimus. Medicinoje, nustatymas, kokie simptomai dažniausiai būna kartu, gali padėti parinkti vaistus. Yra trys būdai matuoti susietumus:

1) Parama (ang. support). Lygi stebėjimų, kuriuose sutinkamas objektas ar komplektas ir visų stebėjimų santykiui.
2) Pasitikėjimas (ang. confidence). Lygus tikimybei, kad objektas X bus perkamas, kai perkamas Y. Šio mato trūkumas tas, kad atsižvelgiama tik į objekto X populiarumą.

$$Confidence\{X \to Y\} = \frac{Support\{X, Y\}}{Support\{X\}}$$

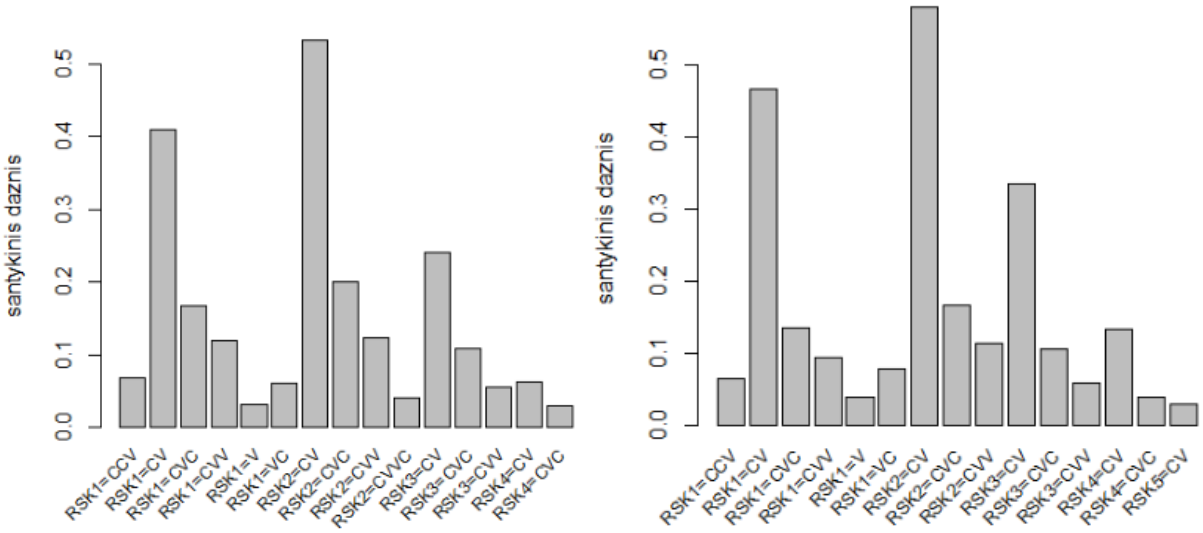
3) Pakėlimas (ang. lift). Rodo kaip tikėtina, kad objektas Y bus perkamas, kai perkamas objektas X, atsižvelgiant į objekto Y populiarumą.

$$Lift\{X \to Y\} = \frac{Support\{X, Y\}}{Support\{X\} * Support\{Y\}}$$

Susietumo taisyklių paieškos technika pritaikoma poezijos ir prozos tekstams, siekiant identifikuoti šiems žanrams būdingas skiemenų poras.

Paveikslėlyje nr. 12 pateikiamos dažniausios poezijos ir prozos skiemenų struktūros pagal skiemens poziciją. Atitinkamose pozicijose esančių dažniausių struktūrų pasiskirstymas poezijoje ir prozoje labai panašus, ypač pirmojo skiemens. Tačiau yra keli ryškesni skirtumai: tik poezijoje išskirtas antrojo skiemens struktūros CVVC dažnumas, prozoje CV struktūros ketvirtas ir penktas skiemuo kur kas dažnesni nei poezijoje.

12 pav.: dažniausios poezijos ir prozos skiemenų struktūros



Poezija Proza

Lentelėse nr. 8 ir 9 pateikiamos susietumo taisyklės, turinčios didžiausią pakėlimo „lift“ parametrą, prozoje ir poezijoje. Dominuojanti struktūra – CV. Šių taisyklių charakteristikos vėliau panaudojamos klasifikavimo modelyje.

8 lentelė: susietumo taisyklės poezijoje

Lhs		rhs	support	confidence	coverage	lift	count
{RSK4=CV}	=>	{RSK3=CV}	0.04152826	0.6618857	0.06274235	2.760920	3138
{RSK1=CV,RSK4=CV}	=>	{RSK3=CV}	0.02550190	0.6563351	0.03885500	2.737767	1927
{RSK1=CV,RSK4=CVC}	=>	{RSK3=CV}	0.01107685	0.6423638	0.01724389	2.679488	837
{RSK2=CV,RSK4=CV}	=>	{RSK3=CV}	0.02926035	0.6329802	0.04622633	2.640347	2211
{RSK1=CV,RSK2=CV,RSK4=CV}	=>	{RSK3=CV}	0.01793206	0.6284787	0.02853248	2.621570	1355
{RSK4=CVC}	=>	{RSK3=CV}	0.01868639	0.6176728	0.03025290	2.576495	1412

9 lentelė: susietumo taisyklės prozoje

lhs		rhs	support	confidence	coverage	lift	count
{RSK1=CV,RSK2=CV,RSK5=CV}	=>	{RSK4=CV}	0.01109832	0.8091908	0.01371533	6.033713	810
{RSK2=CV,RSK3=CV,RSK5=CV}	=>	{RSK4=CV}	0.01034473	0.7980973	0.01296175	5.950994	755
{RSK2=CV,RSK5=CV}	=>	{RSK4=CV}	0.01855201	0.7964706	0.02329278	5.938865	1354
{RSK3=CV,RSK5=CV}	=>	{RSK4=CV}	0.01414009	0.7853881	0.01800395	5.856229	1032
{RSK5=CV}	=>	{RSK4=CV}	0.02345720	0.7781818	0.03014359	5.802495	1712
{RSK1=CV,RSK5=CV}	=>	{RSK4=CV}	0.01430451	0.7750557	0.01845610	5.779185	1044
{RSK4=CVC}	=>	{RSK3=CV}	0.02606051	0.6655003	0.03915927	1.977400	1902
{RSK3=CVV,RSK4=CV}	=>	{RSK2=CV}	0.01072838	0.7760159	0.01382495	1.337412	783
{RSK1=CVV,RSK3=CV}	=>	{RSK2=CV}	0.01441412	0.7673231	0.01878494	1.322431	1052

Skiemenų struktūrų variacija

Allenas Wilcoxas (1973) publikavo darbą, pristatantį šešis indeksus kokybinių ar kategorinių duomenų variacijai išmatuoti:

- DM (ang. deviation from the mode) – nuokrypis nuo modos

$$DM = 1 - \frac{\sum_{i=1}^{k}(f_m - f_i)}{N(K-1)}$$

  Čia f_m – modalinis dažnis, f_i – i-osios grupės dažnis, K – kategorijų skaičius, N – imties dydis.

- ADA (ang. average deviation analog) – vidutinio nuokrypio analogas

$$ADA = 1 - \frac{\sum_{i=1}^{k}\left|f_i - \frac{N}{K}\right|}{2\frac{N}{K}(K-1)}$$

- MDA (ang. mean difference analog) – vidurkio skirtumo analogas

$$MDA = 1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|f_i - f_j|}{N(K-1)}$$

- VA (ang. variance analog) – dispersijos analogas

$$VA = 1 - \frac{\sum_{i=1}^{k}\left(f_i - \frac{N}{K}\right)^2}{\frac{N^2(K-1)}{K}}$$

- HREL indeksas

$$HREL = -\frac{\sum_{i=1}^{k}\frac{f_i}{N}\log_2\frac{f_i}{N}}{\log_2 K}$$

- B indeksas

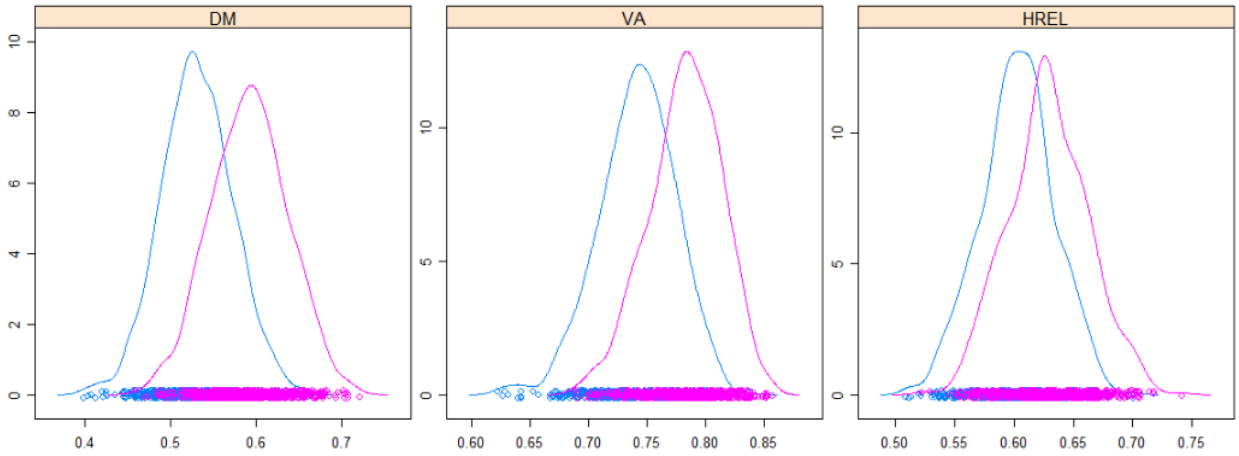
$$B = 1 - \sqrt{1 - \left(\sqrt[K]{\prod_{i=1}^{k}\frac{f_i K}{N}}\right)^2}$$

Indeksų savybės:

- Galimos indekso reikšmės patenka į intervalą [0;1]
- 0 – visi stebėjimai priklauso vienai kategorijai.
- 1 – stebėjimai po lygiai pasidalinę per visas kategorijas.
- Kuo didesnis dažnių skirtumas, tuo indeksas mažesnis.
- Dažnių pasiskirstymui artėjant prie tolygiojo skirstinio (ang. Uniform distribution), indekso reikšmė didėja.

Visiems matams atitinkamai yra funkcijos R bibliotekoje „qualvar“. Indeksai DM, VA ir HREL poezijos ir prozos fragmentuose skiriasi. Poezijoje indeksai linkę būti didesni, reiškia struktūrų dažniai yra pasiskirstę tolygiau nei prozoje (žr. 13 pav.)

13 pav.: DM, VA ir HREL indeksai poezijoje ir prozoje



3.2. Logistinė regresija

Modelis

Logistinė regresija arba logit modelis, yra tiesinės regresijos transformacija, kuri leidžia tikimybiškai modeliuoti dvireikšmius (binary) kintamuosius. Modelis sudaromas priklausomo kintamojo tikimybių santykio, dar vadinamo šansu, logaritmui.

Tarkime, atsitiktinis dydis Y eksperimento metu įgyja reikšmę: 1 – įvykio „sėkmė“, 0 – įvykio „nesėkmė“ atveju. Nepriklausomų kintamųjų vektorių pažymime $x = (x_1, \dots, x_n)^T$. Tuomet logistinės regresijos modelis atrodo taip:

$$logit(P(Y = 1|x)) = ln\frac{P(Y = 1|x)}{P(Y = 0|x)} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n = \beta^T x$$

Kadangi $P(Y = 0|x) = 1 - P(Y = 1|x)$, „sėkmės“ tikimybė

$$P(Y = 1|x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}} = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

Dvireikšmio kintamojo modeliavimui tiesinė regresija nėra tinkama, nes įvertinę parametrus $\beta = (\beta_0, \beta_1, \dots, \beta_n)$, galime gauti įvertį, nepriklausantį intervalui [0; 1].

Koeficientų interpretacija

Tikimybių santykis $\frac{P(Y=1)}{P(Y=0)}$ gali įgyti reikšmes tarp 0 ir +∞. Artimos 0 reikšmės indikuoja mažą, o artimos +∞ didelę Y=1 tikimybę. Logistinėje regresijoje padidinus x_i vienu vienetu, tikimybių santykio logaritmas *logit* pasikeičia dydžiu β_i. Įvykio „sėkmė“ šansas dauginamas iš e^{β_i}. Panašu į tiesinę regresiją, tačiau vietoje pridedamo pokyčio gaunamas dauginamasis pokytis. Šis pokytis vadinamas šansų santykiu:

$$\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_i(x_i+1)+\ldots+\beta_n x_n}}{e^{\beta_0+\beta_1 x_1+\ldots+\beta_i x_i+\ldots+\beta_n x_n}} = e^{\beta_i}$$

Kiek pasikeičia $P(Y = 1)$, padidinus i-tąjį kintamąjį vienetu, priklausys nuo esamos x_i reikšmės. Jei koeficientas β prie nepriklausomo kintamojo teigiamas, kintamojo reikšmei didėjant, įvykio „sėkmė“ tikimybė didėja. Ir atvirkščiai, jei koeficientas neigiamas, didėjant kintamojo reikšmei, įvykio „sėkmė“ tikimybė mažėja. Tuo tarpu įvykio „nesėkmė“ – didėja.

- Jei $\beta_i > 0$, tai $e^{\beta_i} > 1$ ir šansai didėja.
- Jei $\beta_i < 0$, tai $e^{\beta_i} < 1$ ir šansai mažėja.

Į sudaromą logistinės regresijos modelį įtraukiami:

- kiekvienam imties elementui apskaičiuoti klasterinėje kūrinių analizėje naudoti požymiai (skiemenų skaičiaus, balsių, priebalsių vidurkiai ir standartiniai nuokrypiai; funkcinių, turinčių sangrąžos dalelytę, priešdėlį žodžių dalis)
- Kadangi poezijos kūriniuose esantys žodžiai linkę turėti mažiau skiemenų, o prozos daugiau, įtraukiama 2, 5, 6 skiemenis turinčių žodžių dalis.
- Vidutinis ilgųjų, trumpųjų balsių ir priebalsių, dvibalsių, dvigarsių, pusbalsių skaičius skiemenyje.
- Skirtingų struktūrų skaičius, struktūrų CVVC, V arba CVVV dalis, susietumo taisyklių pagalba identifikuotos pirmo ir antro skiemens taisyklės, būdingos poezijai ir prozai.
- Struktūrų variacijų indeksai DM, VA, HREL
- Stebėjime esančių skiemenų šansų patekti į poeziją, vidurkis ir standartinis nuokrypis.

Modelyje sėkme laikomas patekimo į poeziją įvykis. Norėdami atrinkti kovariantes į logistinės regresijos modelį, atliekame pažingsninį parinkimą. Yra trys galimi metodo pritaikymo būdai:

1. Pažingsninė kintamųjų įtraukimo procedūra (ang. forward stepwise regression), kai pradedant modeliu be kintamųjų, pažingsniui pridedamas kintamasis, kurio įtraukimas statistiškai reikšmingiausiai pagerina modelį. Procesas kartojamas, kol nebėra kintamųjų darančių statistiškai reikšmingą įtaką modelio tinkamumui.
2. Pažingsninė kintamųjų išmetimo procedūra (ang. backward stepwise regression), kai į modelį įtraukiamos visos kovariantės ir panašiu statistinio reikšmingumo principu pažingsniui išimami mažiausią įtaką turintys nepriklausomi kintamieji, kol gaunamas galutinis modelis.
3. Dvikryptė pažingsninė procedūra (ang. bidirectional stepwise regression), kai naudojama išmetimo ir įtraukimo procedūrų kombinaciją. Įtraukiant kiekvieną naują reikšmingiausią kintamąjį yra patikrinama ar neatsirado kovariančių, kurias reikia išmesti.

Naudojant pažingsninę kintamųjų išmetimo procedūra gautas geriausiai modeliavimo duomenis atitinkantis ir testavimo stebėjimus prognozuojantis modelis. Modelio deviacijos (663.76) ir laisvės laipsnių (1483) santykis 0.45<1 reiškia modelis gerai tinka duomenims. Lentelėje nr. 10 pateiktos atrinktos kovariantės.

10 lentelė: atrinktos kovariantės

Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	16.44629	6.04769	2.719	0.006539
X5_skiem	-60.52478	13.30138	-4.550	5.36e-06
X6_skiem	-116.69778	34.58443	-3.374	0.000740
skiem_sk_mean	-11.61776	2.29726	-5.057	4.25e-07
skiem_sk_sd	-9.98589	2.93789	-3.399	0.000676
stop_word_dalis	-15.31416	3.47056	-4.413	1.02e-05
priesdeliai_dalis	-12.20690	2.87947	-4.239	2.24e-05
balses_skiem_sd	6.37092	1.57389	4.048	5.17e-05
priebalses_skiem_mean	5.55770	1.07976	5.147	2.64e-07
trm_balsiai_skiem_mean	-12.79316	2.72935	-4.687	2.77e-06
ilg_balsiai_skiem_mean	7.92623	4.01095	1.976	0.048138
VA	40.35770	8.54993	4.720	2.36e-06
HREL	-37.50672	7.21238	-5.200	1.99e-07
SK1cvSK3cvSK4cv	-0.31860	0.08933	-3.566	0.000362
SK2cvSK3cvSK4cv	-0.27606	0.07899	-3.495	0.000475
SK1cvSK2cvSK3cvSK4cv	0.62642	0.13513	4.636	3.56e-06
sansu_mean	0.36505	0.12111	3.014	0.002578

Modeliavimo imtyje, iš 753 poezijos kūrinių taikant logistinę regresiją teisingai klasifikuoti 696. Iš 747 prozos kūrinių 680 priskirtas teisingas žanras. Bendrai teisingai klasifikuota 91.7% modeliavimo imties fragmentų.

Modelio pritaikymo testavimo imčiai rezultatai pateikiami 2×2 lentele (žr. 11 lentelė). Pirmas stulpelis skirtas poezijos fragmentams, iš jų 231 – nustatytas teisingas teigiamas rezultatas, 16 – neteisingai neigiamas. Antras stulpelis skirtas prozos fragmentams, iš jų 34 – gautas neteisingai teigiamas rezultatas, 219 – teisingai neigiamas rezultatas. Jautrumas 231/(231 + 16) = 93.5%, tai tikimybė, kad poezija bus priskirta į poezijos klasę. Specifiškumas 219/(219 + 34) = 86.6%, tai tikimybė, kad proza priskiriame į teisingą klasę. Bendrai teisingai klasifikuota 90% testavimo imties fragmentų.

11 lentelė: testavimo imties klaidų matrica

	Stebėjimas - poezija	Stebėjimas - proza
Prognozė - poezija	231	34

Prognozė - proza	16	219

3.3. Skiemenų šansų modeliavimas

Binominės regresijos modelis gali būti naudojamas įvykio šansų prognozavimui. Priklausomas modelio kintamasis susijęs su kiekvieno scenarijaus sėkmių ir nesėkmių skaičiaus santykiu. Logistinės regresijos modelis yra specialus binominės regresijos modelio atvejis, kur kiekvienos unikalios aiškinančiųjų komponenčių grupės dydis lygus vienam.

Į logistinę regresiją įtrauktas skiemenų šansų patekti į poezijos klasę vidurkis yra reikšmingas. Modelio sudarymui šansai išskaičiuoti iš imties, tačiau norint pritaikyti klasifikatorių visiškai naujiems duomenims to negalėtume padaryti, kol neturėsime skiemenų šansų charakteristikų. Dėl šios priežasties, binominės logistinės regresijos modeliu toliau modeliuojami skiemenų šansai. Į nepriklausomų kintamųjų sąrašą įtraukiama kiek įmanoma daugiau charakteristikų, aprašančių tą skiemenį. Visos kovariantės suvedamos į binarines išraiškas. Gauti 194 binariniai nepriklausomi kintamieji: ar skiemuo yra paskutinis skiemuo, ar tai skiemuo iš žodžio, turinčio priešdėlį, kuris tai priešdėlis, skiemenų struktūrų binariniai kintamieji, trumpų balsių, ilgų balsių, priebalsių, dvibalsių, dvigarsių, pusbalsių skaičius skiemenyje, ar tai sangrąžos dalelytė si, ar skiemuo priklauso dažniausioms poezijos ir prozos pirmo ir antro skiemens poroms. Sudaromi 8 modeliai, kiekvienoje skiemens pozicijoje esantiems skiemenims atskirai. Reikšmingi pirmų penkių modelių koeficientai pateikiami 12-oje lentelėje.

12 lentelė: pirmų penkių modelių koeficientai

Kintamasis	**Modelis 1**	**Modelis 2**	**Modelis 3**	**Modelis 4**	**Modelis 5**
(Intercept)	0.097824	-0.138290	-0.477739	-0.75963	-1.13900
paskutinis_skiemuo_10	-0.117420	0.207195	0.437530	0.89179	0.33863
ant	-1.012190	-0.675977	-	-	-
ap	-0.167457	-0.168273	-	-	-
at	-0.430555	-0.384983	-	-	-
be	0.130866	0.106190	-	-	-
į	-0.869963	-0.641625	-	-	-
im	-0.927104	-1.164100	-	-	-
in	-1.196437	-1.142876	-	-	-
iš	-0.175704	-0.136741	-	-	-
ne	0.106993	0.050136	-	-	-
neap	-0.441309	-0.440898	-	-	-
nebe	0.225820	0.230419	-	-	-
neį	-0.510323	-0.498578	-	-	-
neiš	-0.447743	-0.447101	-	-	-
nenu	0.390941	0.408405	-	-	-

nepa	-0.253629	-0.240029	-	-	-
neprie	1.915668	2.077791	-	-	-
nesą	-2.073316	-2.198585	-	-	-
nete	-0.905710	-0.905151	-	-	-
neuž	0.482460	0.483026	-	-	-
nuo	-0.317408	-	-	-	-
pa	-0.215615	-0.186919	-	-	-
par	-0.610828	-0.291582	-	-	-
per	-0.804267	-0.377245	-	-	-
po	-0.711871	-0.624477	-	-	-
pra	-0.133021	-	-	-	-
pri	-0.534280	-0.328710	-	-	-
prie	-0.975343	-0.656304	-	-	-
pro	-0.348661	-0.240736	-	-	-
są	-0.607217	-	-	-	-
SKcv_CCVVV	0.642977	0.232105	-	-	-
SKcv_CCVVVC	-0.495354	-	-	-	-
trm_balsiai_1	-0.340848	-0.125829	-	-0.90210	-0.43650
trm_balsiai_2	-0.383174	-	-	-1.55792	-
ilg_balsiai_1	-0.031434	-	-	-0.64748	-
dvibalsiai_1	-0.032130	-0.078133	-	-1.68378	-0.30666
pusbalsiai_1	0.119115	0.122880	-	0.13409	-0.18559
priebalsiai_1	0.194550	0.139651	-	0.18938	-0.22983
priebalsiai_2	0.194832	0.373351	-	0.61675	-
priebalsiai_3	0.433805	0.530179	-	-	-
pirmas_antras_skiem_aki	-1.022643	-1.141198	-0.935561	-0.54179	-
pirmas_antras_skiem_aša	1.651791	1.533235	1.459878	0.70469	-
pirmas_antras_skiem_ati	-1.547757	-1.643622	-1.234849	-1.12473	-
pirmas_antras_skiem_dainuo	2.123503	2.417082	2.370033	1.85571	-
pirmas_antras_skiem_dary	-0.738828	-0.739811	-0.433448	-	0.71492
pirmas_antras_skiem_debe	1.884685	1.960679	1.933945	1.46398	-
pirmas_antras_skiem_dide	-0.174488	-	-	-2.50809	-
pirmas_antras_skiem_eže	1.343106	1.247427	1.046675	1.83121	-
pirmas_antras_skiem_galė	-1.398887	-1.398138	-1.169461	-0.71029	-
pirmas_antras_skiem_gali	-0.871829	-0.871181	-0.679068	0.47275	2.40735
pirmas_antras_skiem_galvo	-1.118702	-0.901058	-0.836463	-0.66817	-
pirmas_antras_skiem_gedi	2.127915	2.203910	2.557181	2.63246	-
pirmas_antras_skiem_girdė	-0.347396	-	-	-	-
pirmas_antras_skiem_gyve	-0.632987	-0.266775	-	-	-
pirmas_antras_skiem_kalbė	-1.018973	-0.776142	-0.796694	-0.66871	-
pirmas_antras_skiem_kamba	-0.802331	-	-	-	-
pirmas_antras_skiem_kara	-1.388406	-1.295641	-1.327496	-0.70004	-
pirmas_antras_skiem_kiekvie	-0.887716	-0.533646	-0.552328	-	-
pirmas_antras_skiem_krūti	0.501144	1.005668	0.954823	-	-

pirmas_antras_skiem_lietu	1.721726	2.031003	1.952556	1.48553	-
pirmas_antras_skiem_maty	-0.787896	-0.806257	-0.624879	-	-
pirmas_antras_skiem_mėly	1.754174	1.955090	1.868044	1.52179	-
pirmas_antras_skiem_mėnu	1.788576	2.115320	1.783466	-	-
pirmas_antras_skiem_mergai	-1.240750	-0.947039	-0.973846	-	-
pirmas_antras_skiem_mote	-0.938336	-0.937776	-0.905053	-	-
pirmas_antras_skiem_moti	-0.269791	-0.269173	-0.385539	-	-
pirmas_antras_skiem_norė	-0.993608	-1.062745	-0.890569	-0.49331	-
pirmas_antras_skiem_raudo	0.753695	1.045091	1.053269	1.22105	1.98376
pirmas_antras_skiem_reika	-1.229501	-0.958476	-0.750508	-	0.89611
pirmas_antras_skiem_rody	-2.500515	-2.274007	-2.563754	-2.10333	-
pirmas_antras_skiem_rude	2.742347	2.742906	2.275945	2.86619	-
pirmas_antras_skiem_saky	-0.349382	-0.322832	-	-0.24985	-
pirmas_antras_skiem_sako	-1.629027	-1.689360	-1.026953	-0.92672	-
pirmas_antras_skiem_sida	0.888224	0.964219	1.162986	1.40149	-
pirmas_antras_skiem_stebė	-1.392537	-1.412019	-1.211292	-0.97521	-
pirmas_antras_skiem_stovė	-0.757439	-0.787550	-0.912614	-	-
pirmas_antras_skiem_šešė	1.408100	1.358265	1.287538	-	-
pirmas_antras_skiem_tėvy	3.094794	3.371145	3.285402	-	-
pirmas_antras_skiem_tikė	-1.368544	-1.335679	-0.990742	-0.68297	-
pirmas_antras_skiem_turė	-1.010309	-1.025678	-0.875269	-	2.02357
pirmas_antras_skiem_vaike	1.345630	1.624990	1.523849	-	-
pirmas_antras_skiem_vaka	0.878064	0.882204	0.857838	0.59107	-
pirmas_antras_skiem_valan	-0.769909	-0.878408	-0.756174	-0.97586	-
pirmas_antras_skiem_vande	0.401821	0.718746	0.648466	1.12471	-
pirmas_antras_skiem_vasa	1.935918	1.961015	1.934211	1.84517	3.30640
pirmas_antras_skiem_vokie	1.646336	1.599199	1.538286	-	-
pirmas_antras_skiem_žiūrė	-0.410992	-	-	-	-
nepar	-	-0.848659	-	-	-
neper	-	-1.387655	-	-	-
su	-	-0.092682	-	-	-
te	-	0.129763	-	-	-
už	-	0.166873	-	-	-
SKcv_CCCV	-	-0.757965	-	-	-
SKcv_CCCVC	-	-1.474013	-	-	-
SKcv_CCV	-	-0.218739	-	-	2.71098
SKcv_CCVC	-	-0.207172	-	-	-
SKcv_CCVV	-	-0.332139	-	-	-
SKcv_CCVVC	-	-0.292514	-	-	-
ilg_balsiai_2	-	-1.200756	-	-	-
pusbalsiai_2	-	0.873800	-	1.66494	1.85591
dalelyte_si_1	-	-0.150238	-0.454043	0.85308	0.32839
pirmas_antras_skiem_kuni	-	0.207488	0.552238	0.58991	1.62529
pirmas_antras_skiem_žino	-	-0.681954	-0.672730	-	-

SKcv_CVV	-	-	-	1.13268	-
SKcv_CVVC	-	-	-	0.54873	-
SKcv_CVVCC	-	-	-	1.15345	-
SKcv_CVVV	-	-	-	2.50762	-
SKcv_CVVVC	-	-	-	1.80165	-
SKcv_V	-	-	-	1.52997	-

Gautus modelius pritaikome testavimo duomenims ir prognozuojame kiekvieno skiemens šansus patekti į poeziją. Sekančiu žingsniu suskaičiuojamas įvertintų šansų vidurkis kiekvienam imties stebėjimui, logistinės regresijos modelyje empirinis šansų vidurkis pakeičiamas įvertintu ir patikrinamas modelio veikimas. 13-oje lentelėje pateikiama testavimo imties klaidų matrica.

13 lentelė: gauta testavimo imties klaidų matrica

	Stebėjimas – Poezija	**Stebėjimas – Proza**
Prognozė – Poezija	213	18
Prognozė – Proza	34	235

Bendrai teisingai klasifikuota 89.6% testavimo imties kūrinių. Klasifikavimo jautrumas – 86.2%, specifiškumas – 92.9%.

Duomenims pritaikomi kiti klasifikavimo algoritmai. Lentelėje nr. 14 pateikiami testavimo imties klasifikavimo rezultatai. Geriausi rezultatai gauti naudojant išpūstų medžių klasifikatorių.

14 lentelė: kitų klasifikavimo algoritmų rezultatai

Metodas	**Tikslumas**	**Jautrumas**	**Specifiškumas**
Atraminių vektorių klasifikatorius (ang. SVM – Support vector machine)	89%	84,6%	93,2%
Atsitiktinio miško modelis (ang. random forest model)	89%	92,3%	85,7%
k-artimiausių kaimynų algoritmas (ang. KNN – k-nearest neighbors)	79,8%	90,2%	69,6%
Sprendimų medis (ang. Decision trees)	83%	82,1%	83,9%
Išpūstų medžių metodas (ang. boosted classification trees)	89,4%	88,2%	90,7%
Naivūs Bayes klasifikatoriai (ang. Naive Bayes classifier)	84%	78,7%	91,4%

3.4. Originalių tekstų klasifikavimas

Paimti papildomi keturi kūriniai, kurių fragmentai nebuvo naudojami modeliavimui. Du kūriniai priklauso poezijos žanrui, du – prozai. Parinkta atsitiktinė 400 fragmentų imtis – naudojama 200 poezijos ir 200 prozos fragmentų, po 150 vienas po kito einančių žodžių. Apskaičiuojamos reikalingos charakteristikos, aštuoniais binominės logistinės regresijos modeliais įvertinami fragmentuose esančių skiemenų patekimo į poeziją šansai ir pritaikomas logistinės regresijos modelis fragmentų klasifikavimui. Rezultatai pateikiami lentelėje nr. 15.

15 lentelė: originalių tekstų klasifikavimo rezultatai

	Stebėjimas – Poezija	**Stebėjimas – Proza**
Prognozė – Poezija	173	9
Prognozė – Proza	27	191

Bendrai teisingai klasifikuota 91% originalios testavimo imties fragmentų. Klasifikavimo jautrumas – 86.5%, specifiškumas – 95.5%. Galima daryti išvadą, kad modelio rezultatai nepablogėjo klasifikuojant originalius tekstus.

IŠVADOS

Šiame darbe nagrinėtos skiemenų savybės, įvertintos įvairios nestandartinės charakteristikos. Nustatyta, kad yra skiemenų požymių, kurie skiriasi tarp kūrinių ir jų žanrų. Jie panaudojami skiemenų patekimo į poeziją šansų modeliavimui ir tekstų klasifikatoriaus pagal žanrą sudarymui.

Nagrinėjant skiemenų skaičiaus ir žodžio ilgio priklausomybę, nustatyta, kad dažniausiai pasikartojantys tiek funkciniai, tiek ne funkciniai žodžiai turi nedaug skiemenų. Nagrinėjamuose tekstynuose gausiausia 2 – 3 raidžių ilgio žodžių ir vieno skiemens funkcinių žodžių. Atlikus skiemenų skaičiaus pasiskirstymo tekstuose analizę, paaiškėjo, kad dviejų skiemenų žodžiai sudaro didesnę dalį poezijos tekstų nei prozos, tuo tarpu 5 – 6 skiemenų ilgio žodžiai santykinai dažnesni prozoje. Prozoje mažesnis vienskiemenių ne funkcinių žodžių tankumas, nei poezijoje. Skiemenims patikrintas Zipfo dėsnio, teigiančio, kad žodžio dažnis yra atvirkščiai proporcingas jo rangui dažnių lentelėje galiojimas.

Taip pat, išsiaiškinta, kad apie 50% tekstyne vartojamų skiemenų struktūra yra CV, nubraižytas skiemenų struktūrų medis žodžiams. Poezijoje ir prozoje skiriasi skiemenų struktūrų CVVC, V arba CVVV dalis, tačiau galutiniame modelyje ši savybė nebuvo reikšminga. Ieškant papildomų žanrus skiriančių skiemenų struktūrų savybių, pasinaudota susietumo taisyklių metodu. Atrasta reikšmingų, poezijai ar prozai būdingų pirmo ir antro skiemens struktūrų porų. Skiemenų struktūroms apskaičiuoti šeši kategorinių duomenų variacijos indeksai.

Priešdėlių ir po jų sutinkamos sangrąžos dalelytės identifikavimas atskleidė, kad prozoje šios žodžio dalys yra sutinkamos dažniau. Plečiant skiemenų savybių įvairovę apskaičiuotas vidutinis skiemenyje esančių trumpų balsių, ilgų balsių, priebalsių, dvibalsių, dvigarsių, pusbalsių skaičius. Taip pat, apskaičiuoti empiriniai skiemens patekimo į poeziją šansai.

Kūrinių fragmentų imčiai sudarytas logistinės regresijos modelis. Naudojant pažingsninį reikšmingų kovariančių parinkimą sudarytas 91.7% tikslumu modeliavimo ir 90% tikslumu testavimo imtį klasifikuojantis modelis. Šiems duomenims pritaikyti kiti klasifikavimo metodai:

- Atraminių vektorių klasifikatorius (ang. SVM – support vector machine)
- Atsitiktinio miško modelis (ang. random forest model)
- K-artimiausių kaimynų algoritmas (ang. KNN – k-nearest neighbors)
- Sprendimų medis (ang. decision trees)
- Išpūstų medžių metodas (ang. boosted classification trees)
- Naivūs Bayes klasifikatoriai (ang. naive Bayes classifier)

Gauti panašūs, bet šiek tiek prastesni rezultatai.

Sumodeliavus skiemenų šansus patekti į poeziją naudojant žingsninę binominę logistinę regresiją gauti 8 modeliai. Kiekvienas jų naudojamas atitinkamai pirmo, antro, .., aštunto skiemens

šansų prognozavimui. Į fragmentų logistinės regresijos modelį įdėjus įvertintų šansų vidurkius, vietoje empirinių prognozės pasikeitė labai nežymiai. Klasifikatorių pritaikius kūriniams, kurių fragmentai nebuvo naudojami modeliavime, gautas 91% tikslumas. Reiškia modelis gerai veikia ir ant originalių tekstų.

Gauti klasifikavimo rezultatai rodo, kad sudarytas klasifikavimo algoritmas gali būti naudojamas praktiškai. Tačiau, norint pasiekti didesnį naudojamų modelių tikslumą, reiktų modeliuose įtraukti daugiau šansų charakteristikų. Darbo tęsiniui reiktų panaudoti daugiau skaitmenizuotų kūrinių (tiek modelių testavimui, tiek validavimui).

LITERATŪRA

1. Wongso R., Luwinda F. A., Trisnajaya B. Ch., Rusli O., News Article Text Classification in Indonesian Language, 2017 https://www.researchgate.net/publication/320401938_News_Article_Text_Classification_in_Indonesian_Language
2. Jung Y., Park H., Myaeng S. H., A Hybrid Mood Classification Approach for Blog Text, 2006 https://www.researchgate.net/publication/221419598_A_Hybrid_Mood_Classification_Approach_for_Blog_Text
3. Vilniaus universitetas, Vilniaus universitete emocionalus tekstas virsta kompiuterine schema, Naujienos.vu 2014 https://naujienos.vu.lt/vilniaus-universitete-emocionalus-tekstas-virsta-kompiuterine-schema/
4. Dr. Fang X., dr. Zhan J., Sentiment analysis using product review data, 2015 https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2
5. Sidorov G., Syntactic n-grams in Computational Linguistics, Springer 2019 https://www.springer.com/gp/book/9783030147709
6. Vytauto Didžiojo universitetas, VDU mokslininkai vysto DI technologijų sprendimus lietuvių kalbai, vdu 2019 https://www.vdu.lt/lt/vdu-mokslininkai-vysto-di-technologiju-sprendimus-lietuviu-kalbai/
7. Kazlauskienė A., Pirminio lietuvių kalbos ritmo dėsningumai, VDU 2015 http://talpykla.elaba.lt/elaba-fedora/objects/elaba:8285639/datastreams/MAIN/content
8. Čekanavičius V., Murauskas G., Statistika ir jos taikymai III, TEV 2009 http://stat.vadoveliai.lt/files/LogReg_R.pdf
9. Dobrovolskis B., Lietuvių kalbos žinynas, Šviesa 2007
10. Piaseckienė K., The Statisical Methods in the Analysis of Lithuanian Language Complexity, 2014 https://www.mii.lt/files/doc/lt/doktorantura/apgintos_disertacijos/mii_dis_san_2014_piaseckiene.pdf
11. Gombin J., Indices of Qualitative variation, 2018 https://cran.r-project.org/web/packages/qualvar/vignettes/wilcox1973.html
12. Ng A., Association Rules and the Apriori Algorithm, 2016 https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html
13. Bastrakova E., Garcia S., Zipf's and Heap's Law, 2016

https://rstudio-pubs-static.s3.amazonaws.com/215309_736f5cc00eea4bb9be5a8c566da2beb6.html

14. Piantadosi S., Zipf's word frequency law in natural language: A critical review and future directions, 2015 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592/

4. PRIEDAI

1 Priedas

Naudoti R paketai:
library(dplyr)
library(ggplot2)
library(RSQLite)
library(gapminder)
library(wordcloud)
library(SnowballC)
library(tm)
library(quanteda)
library(FactoMineR)
library(stringr)
library(treemap)
library(data.tree)
library(DescTools)
library(base)
library(readxl)
library(factoextra)
library(cluster)
library(useful)
library(rhandsontable)
library(fpc)
library(themes)
library(gganimate)
library(MASS)
library(arules)
library(sqldf)
library(Matrix)
library(arulesViz)
library(e1071)
library(rpart)
library(rpart.plot)
library(class)
library(tidyverse)
library(caret)
library(xgboost)
library(caTools)

Programiniai kodai:

1. Zipf‘o dėsnis

```
model <- nls(log(count)~log(c)-a*log(rank+b), data=word_count, start = list(a
= 1, b = 0, c = 1))
plot(log(word_count$rank), log(word_count$count), xlab="log(rangas)",
ylab="log(dažnis)")
lines(log(word_count$rank),predict(model), col=2, lwd = 3)

#Zipfo dėsnis kai vertinama tik eksponentė
model <- lm(log(count)~log(rank), data=word_count)
summary(model)
```

```
Call:
lm(formula = log(count) ~ log(rank), data = word_count)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6333  -0.3137  -0.0276   0.5019   0.7378

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.48857    0.05346   420.7   <2e-16 ***
log(rank)   -2.46534    0.00657  -375.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6121 on 8727 degrees of freedom
Multiple R-squared:  0.9416,  Adjusted R-squared:  0.9416
F-statistic: 1.408e+05 on 1 and 8727 DF,  p-value: < 2.2e-16
```

2. Žodžių debesys

```
#debesis pagal zanra
debesiui <- tbl(db, sql("SELECT poezija, pirmas_antras_skiem, COUNT(*) FROM
duomenysfin
WHERE trecias_skiemuo!='' AND antras_skiemuo!=''
GROUP BY poezija, pirmas_antras_skiem
order by poezija,count(*) desc"))
debesiui <- as.data.frame(debesiui)

wordcloud(debesiui$pirmas_antras_skiem[debesiui$poezija==1],debesiui$`COUNT(*
)`[debesiui$poezija==1], max.words = 40, colors=brewer.pal(8,
"Dark2"),random.color=T)
wordcloud(debesiui$antras_trecias_skiem[debesiui$poezija==1],debesiui$`COUNT(
*)`[debesiui$poezija==1], max.words = 40, colors=brewer.pal(8,
"Dark2"),random.color=T)

wordcloud(debesiui$pirmas_antras_skiem[debesiui$poezija==0],debesiui$`COUNT(*
)`[debesiui$poezija==0], max.words = 40, colors=brewer.pal(8,
"Dark2"),random.color=T)
wordcloud(debesiui$antras_trecias_skiem[debesiui$poezija==0],debesiui$`COUNT(
*)`[debesiui$poezija==0], max.words = 40, colors=brewer.pal(8,
"Dark2"),random.color=T)
```

3. Žingsninė logistinė regresija

```
out <- glm(cbind(sekme, nesekme)~., family = binomial, data=unikalus_01)
summary(out)
nothing <- glm(cbind(sekme, nesekme) ~ 1,family=binomial, data=unikalus_01)
summary(nothing)
backwards = step(out) # Backwards selection is the default
summary(backwards)
forwards = step(nothing,scope=list(lower=formula(nothing),upper=formula(out)),
direction='forward')
summary(forwards)
step.model <- out %>% stepAIC(trace = FALSE)
coef(step.model)
```

4. Susietumo taisyklės

```
asocS2S3<-sqldf("select * from pilna_imtis_train where poezija=1 and
skiem_sk>1 and stopai='N';")
asocS2S3_N<-
data.frame(as.factor(asocS2S3$SK1cv),as.factor(asocS2S3$SK2cv),as.factor(asoc
```

```
S2S3$SK3cv),as.factor(asocS2S3$SK4cv),as.factor(asocS2S3$SK5cv),as.factor(aso
cS2S3$SK6cv),as.factor(asocS2S3$SK7cv),as.factor(asocS2S3$SK8cv))
asocS2S3_N_beNO<-data.frame(factor(asocS2S3$SK1cv,exclude='NO'),
factor(asocS2S3$SK2cv, exclude='NO'),factor(asocS2S3$SK3cv, exclude='NO'),
factor(asocS2S3$SK4cv, exclude='NO'),factor(asocS2S3$SK5cv, exclude='NO'),
factor(asocS2S3$SK6cv,exclude='NO'),factor(asocS2S3$SK7cv,exclude='NO'),facto
r(asocS2S3$SK8cv, exclude='NO'))
names(asocS2S3_N_beNO)<-
c("RSK1","RSK2","RSK3","RSK4","RSK5","RSK6","RSK7","RSK8")
asocS2S3_N_beNO<-as(asocS2S3_N_beNO, "transactions")
summary(asocS2S3_N_beNO)
itemFrequency(asocS2S3_N_beNO)
itemFrequencyPlot(asocS2S3_N_beNO, support = 0.03, cex.names=0.8)
rules_beNO <- apriori(asocS2S3_N_beNO, parameter = list(support = 0.01,
confidence = 0.6))
rules_beNO
summary(rules_beNO)
head(inspect(rules_beNO))
rules.sorted_beNO <- sort(rules_beNO, by="lift")
rulesV_beNO <- subset(rules.sorted_beNO, subset =  lift > 2)
```

5. Kategorinių duomenų variacijos indeksai

```
library(qualvar)
train$DM <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, DM)
train$MDA <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, MDA)
train$ADA <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, ADA)
train$VA <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, VA)
train$HREL <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, HREL)
train$B <- tapply(skiem_strukt_dazniai_test$`count(*)`,
skiem_strukt_dazniai_test$stebejimas, B)
```

6. Klasifikavimo algoritmų pritaikymas

```
#SVM
svmfit = svm(poezija ~ ., data = train_01_SVM, type = 'C-classification',
kernel = "linear",, scale = FALSE)#, cost = 10, scale = FALSE)
svmfit2 = svm(poezija ~ X5_skiem + X6_skiem + skiem_sk_mean + skiem_sk_sd +
stop_word_dalis + balses_sd + priebalses_mean + priesdeliai2  +
avg_trm_balsiai_skiem + avg_avg_ilg_balsiai_skiem + VA + HREL +
SK1cvSK3cvSK4cv + SK2cvSK3cvSK4cv + SK1cvSK2cvSK3cvSK4cv + sansu_mean_V3,
data = train_01_SVM, type = 'C-classification', kernel = "linear")
print(svmfit2)

#Fit Random Forest Model
library(randomForest)
rf = randomForest(as.factor(poezija) ~ ., ntree = 80, data = train_01_SVM)
plot(rf)
print(rf)

#KNN
library(class)
pr <-knn(train_01_SVM[,2:42],test_01_SVM[,2:42],cl=train_01_SVM$poezija,k=13)
#create confusion matrix
tab <- table(pr,test_01_SVM$poezija)

#Decision trees
library(rpart)
```

```
library(rpart.plot)
fit <- rpart(poezija~., data = train_01_SVM, method = 'class')
rpart.plot(fit, extra = 106)
predict_unseen <-predict(fit, test_01_SVM, type = 'class')
table_mat <- table(test_01_SVM$poezija, predict_unseen)

#Boosted classification trees
library(tidyverse)
library(caret)
library(xgboost)
set.seed(123)
model <- train(as.factor(poezija) ~., data = train_01_newm, method =
"xgbTree",trControl = trainControl("cv", number = 10))
# Best tuning parameter
model$bestTune

#Naive Bayes classifier
library(caTools)
library(caret)
classifier_cl <- naiveBayes(as.factor(poezija) ~ ., data = train_01_SVM)
# Predicting on test data
y_pred <- predict(classifier_cl, newdata = test_01_SVM[,2:42])
```