TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Birgit Sõrmus

Nimisõnade klasterdamine vastavalt neid kirjeldavatele omadussõnadele

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur, PhD

Tartu 2021

Nimisõnade klasterdamine vastavalt neid kirjeldavatele omadussõnadele

Lühikokkuvõte:

Omadussõnade kasutus annab lisainformatsiooni nendega seotud nimisõnade kohta. See võimaldab kokku grupeerida sarnaste omadustega nimisõnu. Antud töö eesmärgiks on kasutada kolme erinevat meetodit, et klasterdada nimisõnu vastavalt nendele omadussõnadele, millega neid tekstides kirjeldatakse. Klasterdamiseks on kasutatud Jaccardi sarnasust koos spektraalklasterdusega, mittenegatiivset maatriksi faktorisatsiooni ning Dirichlet' peitlahutust. Klasterdamise tulemusena saadakse nimisõnade grupid ning analüüsitakse klastritesse kuuluvate sõnade seotust ning seda, millised omadused milliste sõnade jaoks on keelekasutuses olulised.

Võtmesõnad:
loomuliku keele töötlus, semantiline sarnasus, klasterdamine, Jaccardi sarnasus, teemade modelleerimine, Dirichlet' peitlahutus, mittenegatiivne maatriksi faktorisatsioon

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Clustering of Nouns Using the Adjectives that Describe Them

Abstract:

The usage of adjectives gives additional information about the nouns they describe. That allows the grouping of similar nouns. The aim of this thesis is to use three different methods to cluster nouns using adjectives that describe them in texts. For clustering, the models used are Spectral Clustering on data of Jaccard similarities, Non-Negative Matrix Factorization and Latent Dirichlet Allocation. The resulting clusters are analysed to determine the relatedness of nouns in clusters and which adjectives are significant for which nouns in language use.

Keywords:
natural language processing, semantic similarity, clustering, Jaccard similarity, topic modelling, latent Dirichlet allocation, non-negative matrix factorization

CERCS: P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus

Sõnade klasterdamine on keeleteaduses oluline ülesanne, sest see võimaldab paremini uurida ja mõista keelt ja keelekasutust. Sõnu on võimalik klasterdada erinevatel alustel, näiteks saab sõnu klassifitseerida vastavalt nende sõnaliikidele. See töö käsitleb täpsemalt nimisõnade ja omadussõnade klasterdamist vastavalt sellele, millised omadussõnad kirjeldavad milliseid nimisõnu.

Omadussõnu kasutatakse keeltes, et anda lisainformatsiooni nende sõnade kohta, mille kirjeldamiseks neid kasutatakse. Vastavalt sellele, milliseid omadussõnu kasutatakse nimisõnade kirjeldamiseks, saab nimisõnu jaotada loogilistesse klastritesse. Näiteks omadussõna 'maitsev' kasutatakse eeldatavasti söögi või joogi kirjeldamiseks. Selle töö eesmärk on jaotada nimisõnad klastritesse vastavalt sellele, milliste omadussõnadega neid kirjeldatakse.

Sõnu on võimalik klasterdada semantiliselt tähenduslikesse gruppidesse käsitsi, kuid see on aeganõudev protsess. Selle töö käigus kasutatakse sõnade klasterdamiseks erinevaid masinõppemeetodeid. Nimisõnade klasterdamisel on võimalik leida semantiliselt tähenduslikke klastreid, mis annavad informatsiooni keelekasutuse kohta.

Selle töö käigus kogutakse kokkukäivate nimisõna ja omadussõna paaride andmestik ning kasutades erinevaid masinõppe meetodeid jaotatakse saadud andmestikus sõnad klastritesse. Eesmärk on leida semantiliselt seotud nimisõnade gruppe vastavalt nimisõnu kirjeldavatele omadussõnadele ning uurida, millised omadused mingeid nimisõnu kirjeldavad.

Esimeses peatükis antakse teoreetiline ülevaade semantilise sarnasuse ja seose mõistetest ning erinevatest sõnapaaride kogumise meetoditest. Lisaks sellele kirjeldatakse paaride kogumist ja saadud andmestiku eeltöötlust. Teises peatükis antakse ülevaade erinevatest kasutatud masinõppe algoritmidest ning kirjeldatakse mudelite loomist. Kolmandas peatükis analüüsitakse saadud sõnade klastreid ja hinnatakse tulemusi.

1 Keeleteaduse meetodid ja andmestiku eeltöötlus

Uurides nimisõnadega kokkukäivaid omadussõnu saab nimisõnu jaotada erinevatesse semantiliselt tähenduslikesse klastritesse. Sellisel viisil sõnade klasterdamiseks on vaja suurt hulka tekste, kust saab leida nimisõna ja omadussõna paare. Lisaks sellele peaksid paarid olema sagedaselt esinevad ning nimisõnade klasterdamisel peaks iga kasutatavat nimisõna kirjeldama ka mitu erinevat omadussõna.

Selles peatükis on antud ülevaade semantilise sarnasuse ja semantilise seose mõistetest ning kirjeldatud erinevaid võimalusi vajalike paaride leidmiseks. Lisaks sellele on antud ülevaade paaride leidmiseks kasutatud meetoditest ning andmestiku eeltöötlusest.

1.1 Semantiline sarnasus ja seos

Üldiselt eristatakse kirjanduses üksteisest semantilist sarnasust ja semantilist seost. Semantilise seotuse näol on tegemist üldisema mõistega, mis hindab sõnade omavahelist seotust erinevatel alustel [Res95]. Semantiline sarnasus on konkreetne semantilise seose juhtum, mis võtab vaid taksonoomilisi seoseid arvesse. See tähendab, et sarnasust arvutatakse sünonüümsete sõnade ning sõna ülem- ja alammõiste põhjal.

Erinevates allikates on siiski erinevusi nende kahe mõiste tähenduse osas. Feng jt kirjelduse kohaselt on kahe sõna vaheline semantiline seos määratav vastavalt sellele, kui sarnased on nende tähendused ehk kui sünonüümsed sõnad on [FBEJ17]. Li jt kirjeldasid semantilist sarnasust kui mõõtu, kus kahe sõna vaheline sarnasus on leitav vastavalt nendega seotud mõistetele [LBM03].

Selles töös on nimisõnu klasterdatud sarnastesse gruppidesse vastavalt sellele, milliste omadussõnadega neid on tekstides kirjeldatud. Seetõttu ei leita selles töös sõnade gruppe vastavalt sünonüümsusele, vaid ühes grupis on sõnad, mis on seotud neid kirjeldavate omadussõnade abil. Selle abil on võimalik näiteks ühte gruppi määrata sõnad 'juuli' ja 'august', sest mõlema puhul on tegemist suvekuudega, millest räägitakse sarnastes kontekstides. Lisaks sellele võib aga ka samasse klastrisse sattuda näiteks sõna 'pliit', mis ei ole nii sarnane, kuid võib omadussõnu vaadates olla seotud, sest tihti räägitakse mõistest 'kuum pliit' ning ilmale viidates saab rääkida mõistetest 'kuum juuli' või 'kuum august'.

1.2 Nimisõna ja omadussõna paaride leidmine

Kuna antud töö eesmärk on klasterdada nimisõnu omadussõnade abil, on esimene oluline samm leida kokkukäivaid nimisõnade ja omadussõnade paare tekstidest. Selleks on

erinevaid võimalusi, millest mõningaid on järgnevalt pikemalt kirjeldatud.

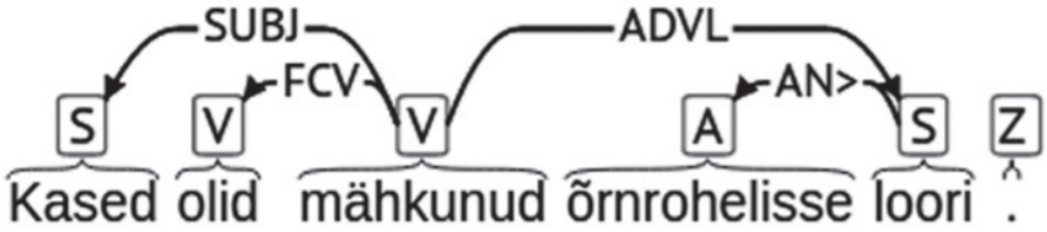
1.2.1 Paaride leidmine vastavalt paiknemisele lauses

Kokkukäivaid sõnapaare on võimalik ilma keerulisemaid algoritme kasutamata leida vastavalt sõnade asukohale lauses. Selleks on vaja arvesse võtta seda, kuidas otsitavad sõnad üksteise suhtes lauses paiknevad. Seda on võimalik teha süntaksi abil. Süntaks ehk lauseõpetus käsitleb lausete ehitust ning lause ja selle osade rolle [Ere13]. Laused ning fraasid lausetes järgivad tüüpjuhtudel sarnaseid struktuure, mis võimaldab määrata seoseid üksteise suhtes.

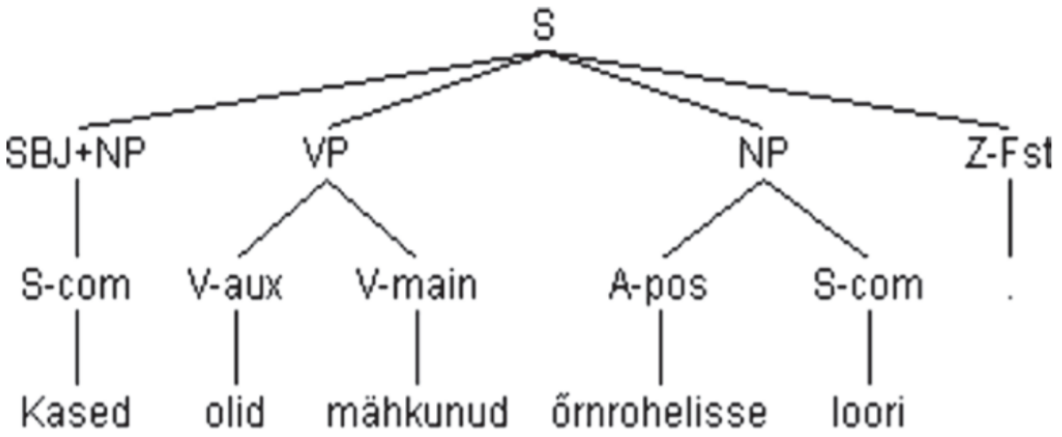
Lausete struktuuri tundmine võimaldab sõnade paiknemise abil leida erineva rolliga sõnu lausest või mitmeid kokkukäivaid sõnu. Selle töö jaoks on vaja leida nimisõnu ning omadussõnu, mis esinevad nende nimisõnade täiendina ehk kirjeldavad neid. Eesti keele puhul on teada, et üldiselt esinevad nimisõnu kirjeldavad omadussõnad vahetult nende ees.

1.2.2 Süntaktiline analüüs

Süntaktiline analüüs ehk parsimine on lause süntaktilise struktuuri kindlaks tegemine [JM19b]. Seda saab teha, määrates lausele vastava süntaksipuu, kus on määratud sõnade- või fraasidevahelised seosed. Vastavalt sellele määratakse kas sõltuvusstruktuur või fraasistruktuur. Joonisel 1 on toodud näide sõltuvusstruktuuri ning joonisel 2 näide fraasistruktuuri visualisatsioonist.



Joonis 1. Näide võimalikust lause sõltuvusstruktuuri visualisatsioonist [MM16].



Joonis 2. Näide võimalikust lause fraasistruktuuri visualisatsioonist [NdV04].

Sõltuvusstruktuuri määramisel tulevad esile seosed, mille abil saab leida, millised nimisõnad ja omadussõnad käivad lauses kokku ning selle alusel luua otsitavad paarid. Seetõttu oleks selles töös otstarbekam kasutada sõltuvusstruktuuri. Lisaks sellele on süntaktilist analüüsi võimalik rakendada erinevates keeletöötluse ülesannetes, näiteks grammatika kontrollimine, semantiline analüüs ja küsimustele vastamine [JM19b].

1.2.3 Grammatikad

Nagu Jurafsky ja Martin kirjeldavad oma raamatus [JM19a], on kontekstivabad grammatikad kogum reeglitest ja sõnavarast. Reeglid määravad grammatika juures selle, kuidas sümboleid kokku grupeerida. Reeglites kasutatavad sümbolid jagunevad terminalideks ja mitteterminalideks. Terminalid on siin sõnavara hulka kuuluvad sümbolid ning mitteterminalid on sümbolid, mille abil väljendatakse reegleid. Grammatika poolt defineeritud keel on hulk sõnedest, mille saab tuletada selles määratud reeglite abil. Kirjutades sobivad reeglid, saab grammatikaid kasutada ka tekstist nimisõna ja omadussõna paaride leidmiseks. Joonisel 3 on näide lihtsatest grammatikareeglitest, mille abil on võimalik leida tekstist fraase, mis sisaldavad nimisõna ja sellele eelnevat suvalist arvu omadussõnu. Selles näites on eeldus, et nimisõnadel on juures märgend ’noun’ ning omadussõnadel märgend ’adj’, mis on grammatikas terminalideks.

$$S \rightarrow A \;\; NP$$
$$NP \rightarrow A \;\; NP \;\mid\; N$$
$$A \rightarrow \textit{adj}$$
$$N \rightarrow \textit{noun}$$

Joonis 3. Näide võimalikust grammatikareeglite esitusest, et leida nimisõnad ja neile eelnev suvaline arv omadussõnu.

Võrreldes süntaksianalüüsiga on grammatikate kasutamisel keerulisem leida kõiki kokkukäivaid paare. Süntaktiline analüüs määrab täpselt iga sõna jaoks lauses selle ülema. Grammatikad eeldavad aga kasutaja poolt kirjutatud reegleid, mille abil ei ole võimalik kogu keelekasutust kirjeldada. Kui ei ole soov kõiki paare leida, siis grammatikad võimaldavad lihtsasti täpsustada reegleid, mille kaudu tekstist sõnu otsida.

1.3 Paaride kogumine tekstidest

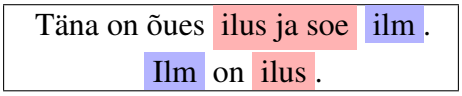
Omadussõna ja nimisõna paaride leidmiseks on kasutatud tekste eesti keele koondkorpusest [ekk], mis on kogumik eestikeelseid terviktekste erinevatest valdkondadest. Kuna konkreetselt nimisõnade ja omadussõnade leidmiseks on vaja teada sõnaliike, oli vaja kasutada morfoloogiliselt analüüsitud tekste. Morfoloogilist analüüsi on võimalik erinevaid tööriistu kasutades ise teostada, kuid kuna tegemist on aeganõudva protsessiga, on selles töös kasutatud eelnevalt EstNLTK v1.6.6-ga [LOST20] analüüsitud koondkorpust.

Sobilike sõnapaaride leidmiseks on erinevaid võimalusi. Töö käigus proovis autor paare leida väiksemast andmestikust nii EstNLTK [LOST20] grammatikaid kasutades kui ka sõnade paiknemise kaudu. EstNLTK abil tehtud morfoloogiline analüüs sisaldab sõnaliiki määravat atribuuti *part-of-speech*, mis võimaldab leida lähedal paiknevaid nimisõnu ja omadussõnu.

Grammatikaid kasutades oli näha, et see võttis palju rohkem aega kui paaride leidmine sõnade paiknemist kasutades. Seetõttu on kasutatud paaride leidmist vastavalt sõnade paiknemisele lauses. Selle käigus vaatab programm kõik sõnad lauses lihtsa tsükli abil läbi ning leiab vastavalt reeglitele sõnapaarid.

Antud töös kasutatud andmestiku loomisel vaadati igale nimisõnale kolme eelnevat ning kahte järgnevat sõna. Kui kolm eelnevat sõna on kõik omadussõnad või on kaks koma või sidesõnaga eraldatud nimisõna, luuakse sealt vastavalt kolm või kaks paari ning need salvestatakse. Kui ainult üks või kaks eelnevat sõna on omadussõnad, leitakse

ning salvestatakse sealt vastavalt üks või kaks paari. Kahe järgneva sõna puhul leitakse paar sellisel juhul, kui nimisõnale järgnev sõna on verb olema ning sellele järgnev sõna on omadussõna. Joonisel 4 on toodud näited paaride leidmisest tekstist. Sinisega on tähistatud leitud nimisõna ja punasega kontrollitud potentsiaalsed omadussõnad. Esimesel juhul leitakse paarid 'ilus ilm' ja 'soe ilm', teisel juhul paar 'ilus ilm'.



Joonis 4. Näited lausest paaride leidmisest vastavalt sõnade paiknemisele

1.4 Lõpliku andmestiku moodustamine

Esialgses andmestikus on unikaalseid paare kokku 3070458. Paaride leidmisel kasutatud meetodi tulemusena esineb esialgses andmestikus väiksema esinemisarvuga paaride hulgas valepositiivseid sõnapaare. Valepositiivsete paaride eemaldamiseks uuris autor juhuslikke näiteid erinevate esinemisarvudega paaride hulgast ning otsustas vastavalt lausele, kus paar esines, kas antud nimisõna ja omadussõna paar esines koos või mitte.

Uuritud sai näiteid, mis esinesid 1–10 korda, kusjuures iga lävendi juures uuris autor 10 juhuslikku näidet. Kõikide juhuslike näidete puhul, mis esinesid tekstides 1–9 korda, esines valepositiivseid paare. 10 oli vähim lävend, mille puhul valepositiivseid näiteid uuritud paaride hulgas ei leidunud, mistõttu valis autor selle lävendiks, mille järgi moodustada lõplik andmestik.

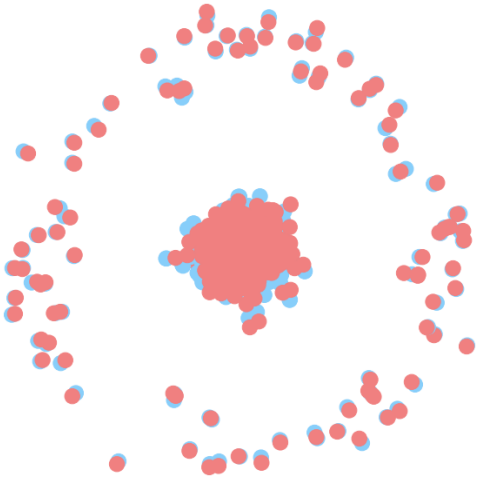
Valepositiivsed paarid said tekkida erinevatel põhjustel. Üks võimalus oli see, kui oli kasutatud sõnu, mis saavad tekstis käituda nii omadussõna kui nimisõnana. Näiteks leidis programm kaks korda paari 'kohalik pall'. Üks lausetest, millest see paar leiti, oli "Õhtul oli õnneks vaba aega, siis saime **kohalikega palli** mängida.", kust on näha, et sõnad 'kohalik' ja 'pall' on kõrvuti ning vastavad otsinguks määratud reeglitele. Lauset lugedes on näha, et sõna 'kohalik' käitub antud kontekstis nimisõnana ning seetõttu paar ei ole korrektne. Kuna EstNLTK aga määras sõna 'kohalik' selles lauses omadussõnaks, tuvastati see paar. Sarnane näide oli veel näiteks paar 'haige lennuk', mis esines näiteks lauses "Kindlustussumma eest on vajadusel võimalik **haige lennukiga** koju tuua, samuti minnakse järele surnukehale või korraldatakse matus välismaal.". Teine levinud viga oli mitmest sõnast koosnev omadussõnafraas. Selliste fraasidega ei osanud programm arvestada, sest soov oli leida üksikuid omadussõnu. Selle idee järgi oli ka programm koostatud. Sellest tekkis näiteks paar 'vaheline rahukõnelus' lausest "Eile aktiviseerusid Marylandis jätkuvad Iisraeli ja Palestiina liidrite **vahelised rahukõnelused** veelgi.". 'Vaheline' ja 'rahukõnelus' on küll korrektselt tuvastatud kui omadussõna ja nimisõna, kuid paarina ei ole see tähenduslik. Lausest on näha, et tegelikult tähenduslik fraas sellest

oleks ’Iisraeli ja Palestiina liidrite vahelised rahukõnelused’. Tabelis 1 vastab igale reale esimeses tulbas uuritud lävend ning iga lävendi kohta on välja toodud valepositiivsete paaride arv ja üks valesti leitud paari näide.

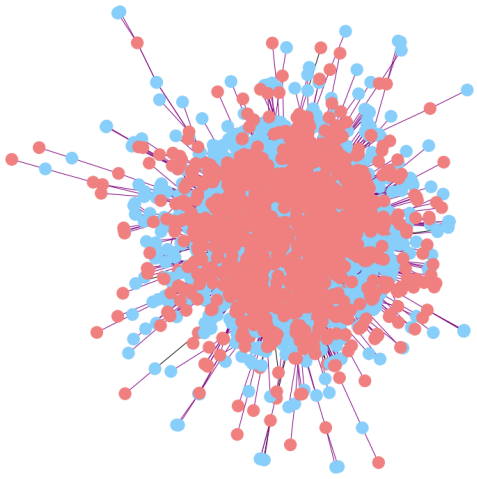
Tabel 1. Tekstist leitud valepositiivsete paaride arv ning näited iga uuritud lävendi kohta

	Valepositiivsete arv	Valesti leitud paari näide	Lause valepositiivse paariga
1	5	läinud meeleavaldus	Vene noorukite laamendamiseks üle **läinud meeleavaldus** USA saatkonna ees pärast Eesti ühinemist Iraagi-vastase koalitsiooniga.
2	3	kohalik pall	Õhtul oli õnneks vaba aega, siis saime **kohalikega palli** mängida.
3	2	haige lennuk	Kindlustussumma eest on vajadusel võimalik **haige lennukiga** koju tuua, samuti minnakse järele surnukehale või korraldatakse matus välismaal.
4	3	maksev laserplaadikogu	Hagi tagamiseks arestis täitevbüroo möödunud suvel korraks ka Jüri Makarovi umbes pool miljonit krooni **maksva laserplaadikogu** rohkem kui 20 000 plaadiga.
5	4	vaheline rahukõnelus	Eile aktiviseerusid Marylandis jätkuvad Iisraeli ja Palestiina liidrite **vahelised rahukõnelused** veelgi.
6	1	kehtiv sõiduk	On ju auto pikkust silma järgi küllalt raske täpselt hinnata ja praegu **kehtivas sõiduki** tehnilises passis sellekohast märget pole.
7	3	pakatav nägu	Karini muidu elurõõmust **pakatavalt näolt** võis lugeda hirmu.
8	2	olnud maakler	Info vahendajaks **olnud maakler** muutub tulevikus peamiselt kliendinõustajaks.
9	2	oodatud koha	**Kohale** on **oodatud** kõikide kõrgkoolide ja tudengiorganisatsioonide esindajad.
10	0	-	-

Lisaks eemaldati paarid, milles leidunud sõnad ei olnud teistega seotud. Selleks uuriti vaid paare, mis jäid alles, kui eemaldati vähem kui 10 korda esinenud paarid. Ülejäänud andmetega mitteseotud paaride leidmiseks esitas autor paarid graafina ning uuris graafi seotud alamkomponente. Tulemustest oli näha, et tekkis üks seotud alamkomponent, mis sisaldas 126288 paari. Lisaks neile leidus 157 paari, milles olevad sõnad paigutusid teistesse alamkomponentidesse ning ei olnud seega suurema osa sõnadega seotud. Kuna sellised sõnad ei ole klasterdamisel kasulikud, sest need ei ole muude andmetega seotud, eemaldas autor need 157 paari andmestikust. Nii eemaldati näiteks paarid ’suhkruvaba näts’ ja ’suhkruvaba närimiskumm’. Nendest sõnadest oli tekkinud kolmetipuline alamgraaf, sest omadussõna ’suhkruvaba’ kirjeldas nii sõna ’näts’ kui ka ’närimiskumm’, kuid ükski neist kolmest sõnast ei olnud ülejäänud sõnadega paaris. Joonistel 5 ja 6 on visualiseeritud 10000 levinumat paari graafina. Joonistelt on samuti näha, et tekkinud on üks suur ning mitu väiksemat alamkomponenti, mis ei ole omavahel seotud.



Joonis 5. 10000 levinuma paari visualisatsioon



Joonis 6. 10000 levinuma paari suurima alamkomponendi visualisatsioon

Lisaks selgus töö käigus, et problemaatiliseks osutusid ka sellised paarid, kus nimisõnu kirjeldab liiga väike arv omadussõnu. Seetõttu sai loodud andmestikust enne klasterdamist eemaldatud paarid, kus olevaid nimisõnu leidus vähem kui kolme erineva omadussõnaga paaris. Niimoodi eemaldati sõnu, mida on väheste paaride tõttu keeruline loogilistesse klastritesse jaotada. Eemaldati näiteks nimisõna ’võrdõiguslikkus’, mis esines paaris ’sooline võrdõiguslikkus’ 615 korda, kuid millest muus kontekstis ei räägitud. Samuti eemaldati näiteks sõna ’kesktee’, mis esines vaid paaris ’kuldne kesktee’ 181 korda. Selle tagajärjel jäi klasterdamiseks alles 5817 nimisõna, 7358 omadussõna ja 115295 paari.

2 Mudelite loomine

Selles peatükis on antud teoreetiline ülevaade kasutatud masinõppe meetoditest ning lisaks on räägitud, kuidas kirjeldatud mudelid praktiliselt loodi.

2.1 Masinõppe meetodid

Pärast andmete eeltöötlemist saab nende peal rakendada ülesande lahendamiseks sobilikke masinõppe meetodeid. Antud bakalaureusetöös kasutatakse nimisõnade klastritesse jagamiseks sarnasusmaatriksi järgi klasterdamist, mittenegatiivse maatriksi faktorisatsiooni ning Dirichlet' peitlahutust.

2.1.1 Sarnasusmõõdud

Üks selles töös kasutatud sõnade gruppidesse jagamise meetoditest on klasterdamine vastavalt kahe sõna omavahelisele sarnasusele. Sõnadevahelisi sarnasusi saab leida erinevatel viisidel, kuid selles töös kasutati selleks Jaccardi sarnasust.

Jaccardi sarnasus on sarnasusmõõt, mille töötas esimesena välja Paul Jaccard oma töös [Jac12] võrreldes taimeliike erinevates piirkondades. Kui uurimisel on kaks erinevat hulka A ja B, siis nendevahelist Jaccardi sarnasust saab arvutada järgneva valemi abil:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

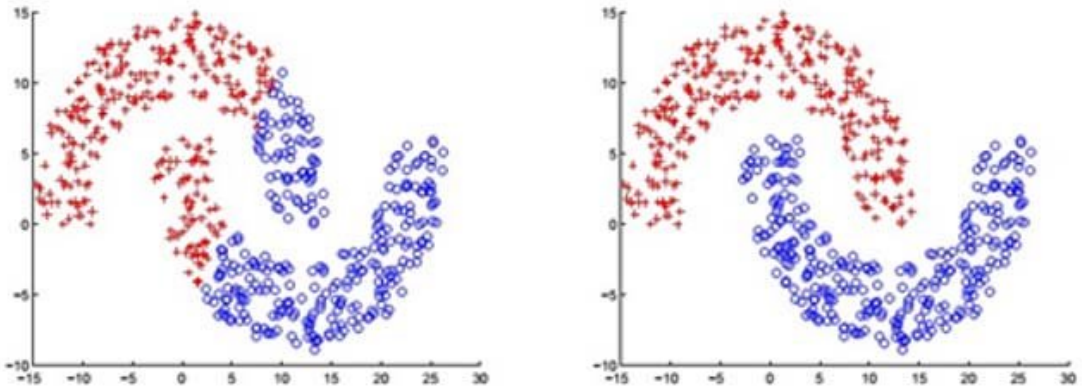
kus on $|A \cap B|$ kahe hulga ühisosa suurus ning $|A \cup B|$ on kahe hulga ühendi suurus.

2.1.2 Klasterdamine

Esimene viis sõnade grupeerimiseks on nende klasterdamine mingi klasterdusalgoritmi abil. Klasterdamine on andmete jaotamine gruppidesse ehk klastritesse. Klastreid kirjeldatakse üldiselt kui gruppe andmepunktidest, mis on omavahel sarnased ning erinevad andmepunktidest teistes klastrites. Klasterdusalgoritme leidub erinevaid, kuid selles töös on kasutatud spektraalklasterdust. Täpsema ülevaate klasterdamisest saab Xu ja Wunschi raamatust [XI09].

Spektraalklasterdus (inglise keeles *Spectral Clustering*) on Ng jt [NJW01] poolt kirjeldatud kui klasterdusalgoritmi, mis kasutab andmestiku põhjal loodud sarnasusmaatriksi omavektoreid, et andmepunkte klastritesse jagada. Klasterdamiseks kasutatakse

K-keskmiste (inglise keeles *K-Means*) meetodit. Nagu on näha joonisel 7, saab spektraalklasterduse algoritmiga ka ebakorrapäraseid kujusid klasterdada edukamalt kui tavalise K-keskmiste meetodiga.



Joonis 7. K-keskmiste meetodi ja spektraalklasterduse võrdlus andmepunktide klasterdamisel [AHA17].

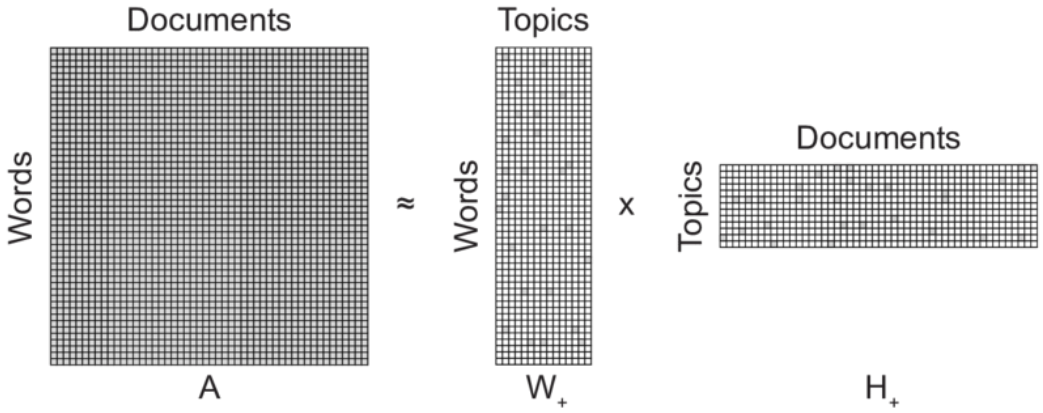
2.1.3 Mittenegatiivne maatriksi faktoriseerimine

Koren jt kirjeldavad maatriksi faktoriseerimist kui meetodit lahutada maatriks kaheks väiksemamõõtmeliseks maatriksiks [KBV09]. Nende töös kirjeldatu järgi on seda kasutatud soovitussüsteemide korral, et kasutajatele tooteid ja teenuseid pakkuda. Taolistes süsteemides soovitatakse teatud tooteid kasutajatele vastavalt eelnevalt tarbitud toodetele ehk kasutajad ja tooted jaotatakse gruppidesse vastavalt sellele, kuidas need on seotud. Analoogselt saab ka nimisõnu ja omadussõnu jaotada gruppidesse, sest süsteem leiab sõnadevahelisi seoseid ning grupeerib rohkem seotud sõnad kokku. Maatriksi faktorisatsiooni algoritme on erinevaid, kuid kuna käesolevas töös on andmetes vaid mittenegatiivsed arvud, sest sõnapaare ei saa tekstis leiduda negatiivne arv kordi, on kasutatud mittenegatiivset maatriksi faktorisatsiooni.

Mittenegatiivne maatriksi faktoriseerimine ehk NMF (inglise keeles *Non-Negative Matrix Factorisation*) sai tuntuks selle nime all tänu Lee ja Seungi tööle [LS99]. Nende töös kirjeldati, et NMF on algoritm, mis jaotab etteantud maatriksi V ligikaudselt kahe maatriksi korrutiseks, nii et

$$V \approx WH,$$

mille visualisatsioon on näha joonisel 8. Algoritmi nimi tuleneb sellest, et ükski kolmest maatriksist ei sisalda negatiivseid väärtuseid.



Joonis 8. Kontseptuaalne illustratsioon NMF abil teostatud maatriksi lahutusest [KBB17].

2.1.4 Dirichlet' peitlahutus

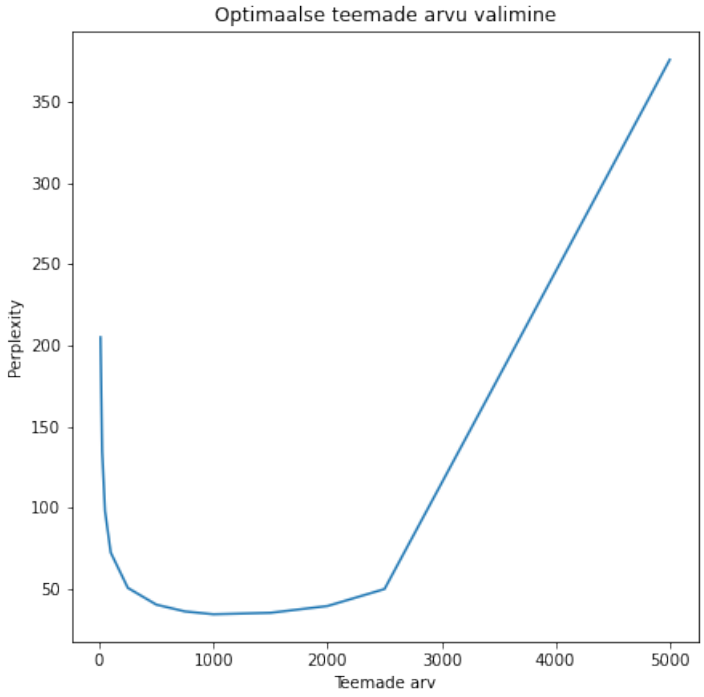
Teemade modelleerimine on juhendamata masinõppe valdkonda kuuluv meetod, mida rakendatakse suurele hulgale tekstidele ehk dokumentidele, et leida seoseid tekstidesse kuuluvate sõnade abil. Algoritm leiab teemad, mis on määratud leitud sõnade ja nendevaheliste seoste abil. Vastavalt teemadesse määratud sõnadele on võimalik leida dokumentide jaotus teemade vahel. Täpsemalt on võimalik lugeda teemade modelleerimise kohta Boyd-Graber jt raamatust [BHM17]. Selles töös on dokumentideks nimisõnad ning teemade määramiseks kasutatavad sõnad on nimisõnu kirjeldavad omadussõnad.

Teemade modelleerimise valdkonda kuuluv Dirichlet' peitlahutus ehk LDA (inglise keeles *Latent Dirichlet Allocation*) on generatiivne ja tõenäosuslik mudel, mille eesmärk on etteantud dokumendid jagada teemadesse [BNJ01]. Tegemist on juhendamata masinõppega, seega mudelile ei ole ette antud teemasid, kuid määratud on teemade arv. Teemade sisu määratakse mudeli poolt vastavalt leitud seostele. Seoste määramisel käiakse läbi igale dokumendile vastavad sõnad ning määratakse need ühte teemasse. Vastavalt sellele, kui suur osa sõnadest määratakse mingisse teemasse, saab ka määrata, kui suure tõenäosusega võib dokument igasse teemasse kuuluda.

2.2 Mudelite treenimine

Töö käigus klasterdas autor sõnu erinevate meetodite abil. Esimesena kasutati Jaccardi sarnasust, mille tulemustele rakendati spektraalklasterdust. Teisena kasutas autor NMF algoritmi ning seejärel LDA algoritmi.

Kõigi kolme mudeli puhul oli tulemuseks saadud klastrite arv 1000. Klastrite arvu 1000 määras autor LDA mudeli tulemuste järgi ning võrreldavuse jaoks kasutas sama arvu ka teiste meetodite juures. Optimaalse arvu leidmiseks treenis autor erinevate klastrite arvudega mudelid ning arvutas iga mudeli puhul *perplexity* skoori [BNJ01]. Madal skoor viitab sellele, et mudelil on hea üldistusvõime, seega klastrite arvuks sai valitud see, millega treenitud mudel andis madalaima *perplexity* skoori. Joonisel 9 on näidatud saavutatud skoorid erinevate klastrite arvu juures. Ülejäänud parameetrid jäid mudelitel muutmata.



Joonis 9. LDA *perplexity* skoor erinevate klastrite arvude juures

2.2.1 Spektraalklasterdus Jaccardi sarnasuste põhjal

Esimesena leidis autor sõnade klastreid vastavalt Jaccardi sarnasustele nimisõnade vahel. Kahe nimisõna omavahelise Jaccardi sarnasuse leidmiseks võeti arvesse pärast eeltöötlust alles jäänud paare. Arvutatud nimisõnapaaride sarnasuste põhjal loodi ruutmaatriks, kus iga väärtus väljendab vastavas reas ning veerus olevate nimisõnade omavahelist sarnasust. Saadud sarnasusmaatriksile rakendati spektraalklasterdust, mis jagas sõnad 1000 erinevasse klastrisse.

2.2.2 Mittenegatiivne maatriksi faktoriseerimine

Teisena rakendas autor andmetele NMF mudelit. Selleks moodustati andmetest maatriks, kus ridadele vastasid nimisõnad ning tulpadele omadussõnad. Iga väärtus maatriksis oli vastava rea nimisõna ning vastava tulba omadussõna paari esinemisarv. NMF mudel jaotab etteantud maatriksi kaheks maatriksiks, millest üks vastab nimisõnade ning teine omadussõnade jaotusele klastrite vahel.

2.2.3 Dirichlet' peitlahutus

Kolmanda meetodina kasutas autor sõnade klasterdamiseks LDA mudelit. LDA mudel sai treenimiseks ette sama andmestiku, mida kasutati NMF mudeli puhul. LDA mudelilt on võimalik pärast treenimist küsida nimisõnade jaotust klastrite vahel. Saadud jaotuses vastab igale nimisõnale tõenäosusvektor, mis näitab, kui suure tõenäosusega nimisõna igasse klastrisse kuulub. Lisaks saab LDA mudelilt omadussõnade jaotuse, milles on iga klastri kohta antud kõikidele omadussõnadele vastavad väärtused. See väärtus näitab, kui määrav on iga omadussõna selle klastri jaoks, mis tähendab, et mida suurem väärtus, seda rohkem kirjeldab see sõna antud klastrit.

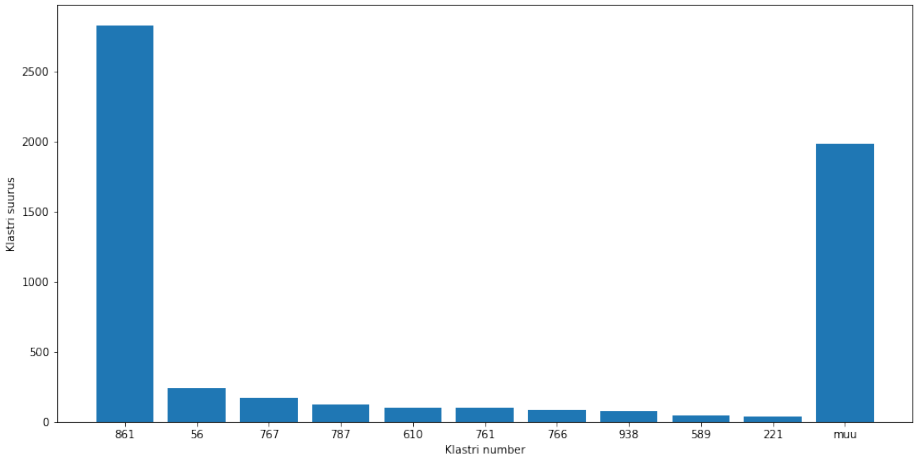
3 Tulemuste analüüs

Antud peatükis hinnatakse tulemusi peamiselt kolmel alusel. Esiteks on vaadatud saavutatud klastrite suurusi, sest kui suurem osa sõnu on paigutatud ühte klastrisse, ei ole tegemist informatiivsete tulemustega. Teiseks on LDA ja NMF meetodite korral uuritud klastreid, mis on tugevalt kirjeldatud ühe omadussõna poolt. Kuna LDA ja NMF väljastavad ka omadussõnadele vastavad jaotused, on võimalik leida klastrid, kus on ühel omadussõnal palju kõrgem väärtus kui teistel. Samuti on võimalik uurida üldiselt omadussõnadele vastavaid klastreid. Kolmandaks on uuritud, kuidas on klastrite vahel jaotunud sõnad, mis võiks omavahel olla semantiliselt seotud. Selleks kasutas autor kuid, nädalapäevi, aastaaegu, spordialasid ning ameteid. Spordialadest uuriti täpsemalt sõnu 'jalgpall', 'korvpall' ja 'tennis'. Elukutsetest uuriti sõnu 'laulja', 'näitleja', 'kunstnik', 'tantsija' ja 'artist', mis on tihti omavahel seotud. NMF ja LDA meetodite puhul uuris autor lisaks veel värvide jaotust klastrite vahel.

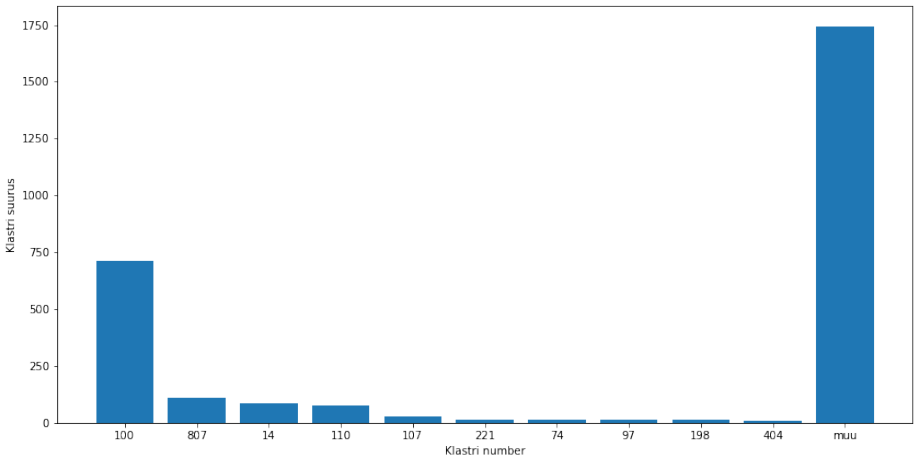
3.1 Spektraalklasterdus Jaccardi sarnasuste põhjal

Nimisõnade omavaheliste Jaccardi sarnasuste järgi klasterdas autor sõnad 1000 erinevasse klastrisse. Saadud tulemustest oli näha, et tekkis üks suur klaster, mis sisaldas 2835 sõna. See tähendab, et see klaster sisaldas 48,7% kõikidest andmestikus esinenud nimisõnadest ning ei ole seega informatiivne klaster. 745 klastrit sisaldas vaid üht sõna, mis samuti ei ole tulemusena informatiivne, sest ei anna teavet sõnade seotuse kohta. Joonisel 10 on näha, kui suur osa sõnadest oli kümnes kõige suuremas klastris. Ülejäänud klastrid on kokku grupeeritud 'muu' alla. Numbrid klastrite kõrval näitavad mudeli poolt määratud klastri indeksit. Selline suur klaster tekkis ilmselt seetõttu, et kasutatud Jaccardi sarnasuse valem võttis arvesse vaid seda, millised omadussõnad iga nimisõna kirjeldavad. See tähendab, et arvesse ei võetud sõnade üldist sagedust ega ka konkreetsete paaride sagedust. Selle tõttu mõjutavad sarnasust väga levinud omadussõnad, mis kirjeldavad paljusid erinevaid nimisõnu, kuid ei ole informatiivsed.

Tekkinud probleemi lahendamiseks proovis autor kahte erinevat asja. Esiteks jäeti andmestikku alles ainult need sõnad, mis olid tekkinud suures klastris. Sellele väiksemale hulgale andmestikust rakendati uuesti spektraalklasterdust. Tulemusena oli näha, et tekkinud klastritest suurim sisaldas 715 sõna ning suuruselt teine sisaldas 110 sõna. See näitab juba tulemuse paranemist, sest väiksemad klastrid on suurema tõenäosusega semantiliselt tähenduslikud ning sisaldavad omavahel seotud sõnu. Siiski oli näha, et 676 klastrit sisaldas vaid üht sõna, mis ei ole informatiivne tulemus. Joonisel 11 on näha, kuidas sõnad on jaotunud klastrite vahel teises mudelis. Analoogselt esialgsele mudelile vastavale joonisele on siin välja toodud kümne suurima klastri nimisõnade arv ning ülejäänud on grupeeritud 'muu' alla.



Joonis 10. Sõnade jaotus erinevate klastrite vahel esimeses mudelis.

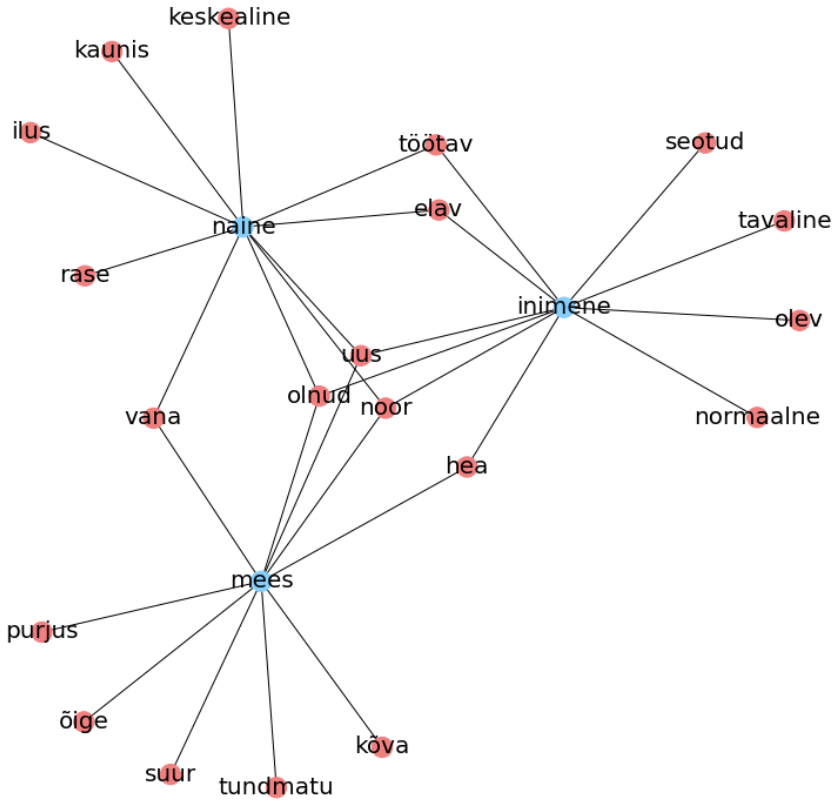


Joonis 11. Sõnade jaotus erinevate klastrite vahel teises mudelis.

Selle meetodi puhul ei ole võimalik hinnata omadussõnade jaotust klastrite vahel, sest Jaccardi sarnasustest loodud andmestik ning selle põhjal leitud klastrid ei säilita informatsiooni omadussõnade kohta.

Pärast kahte klasterdamist uuris autor täpsemalt näiteid klastrite hulgast. Esimese mudeli

tulemuste hulgast suvalisi näiteid uurides tulid välja näiteks klastrid [’inimene’, ’mees’, ’naine’] ning [’automaatkast’, ’automaatkäigukast’, ’käsikast’]. Teise mudeli juhuslike näidete hulgas olid näiteks klastrid [’bussifirma’, ’kinnisvarafirma’, ’kütusefirma’] ning [’eluase’, ’elupaik’]. Nende näidete puhul on näha, et ühes klastris on tegemist semantiliselt seotud sõnadega. Joonisel 12 on illustreeritud klastri [’inimene’, ’mees’, ’naine’] seotus. Iga klastrisse kuuluva nimisõna jaoks on toodud välja 10 omadussõna, millega see esines kõige rohkem paaris. Joonisel on nimisõna ja omadussõna ühendatud, kui see paar esines vastava nimisõna jaoks 10 levinuma paari seas.



Joonis 12. Klastri [’inimene’, ’mees’, ’naine’] ning iga klastrisse kuuluva sõna 10 levinuima paari visualisatsioon

Uurides eelnevalt välja toodud semantiliselt seotud sõnu, oli esimese mudeli puhul näha, et kõik aastaajad ja nädalapäevad olid paigutunud tekkinud suurde klastrisse. Spordialadest olid ’jalgpall’ ja ’korvpall’ samuti kõige suuremas klastris. Sõna ’tennis’ oli semantiliselt vähetähenduslikus klastris 28 teise sõnaga. Seal olid veel näiteks sõnad ’pitsa’, ’serblane’ ning ’vabadussõda’. Mainitud elukutsetest kõik peale sõna ’tantsija’ olid analoogselt eelnevatega paigutunud samasse tekkinud suurde klastrisse. Sõna ’tantsija’

oli klastris 128 teise sõnaga, mis sisaldas paljusid erinevatele ametitele või rahvustele viitavaid sõnu, aga ka muid sõnu, näiteks sõnu seotud hoonetega, nagu 'kultuurimaja' või 'tselluloositehas'. Kuude klasterdamine õnnestus esialgsel mudelil kõige paremini, tekkis järgnev klaster: ['august', 'detsember', 'jaanuar', 'juuli', 'juuni', 'märts', 'november', 'september', 'veebruar'].

Teise mudeli peal, mis jaotas vaid suurde klastrisse kogunenud sõnu, oli näha edasist klasterdumist. Aastaaegadest tekkisid klastrid ['sügis', 'tali', 'talv'] ja ['kevad', 'suvi']. Nädalapäevadest tekkisid klastrid ['kolmapäev', 'teisipäev'] ja ['laupäev', 'neljapäev', 'pühapäev', 'reede'], kuid sõna 'esmaspäev' jäi endiselt sellest mudelist tekkinud suurimasse klastrisse. Uuritud ametid ning spordialad jäid ka edasisel klasterdamisel saadud 715-sõnalisse klastrisse.

Teisena proovis autor tekkinud suure klastri probleemi lahendamiseks kasutada kaaluga Jaccardi sarnasuse valemit ning rakendada sellele TF-IDF tehnikat [Jon72], et vähendada sagedaste omadussõnade kaalu. See ei andnud võrreldes esialgse klasterdamisega paremaid tulemusi. Saadud klastrite hulgas oli üks klaster, mis sisaldas 3000 nimisõna, ning mitmed uuritud väiksemad klastrid olid sarnaselt jaotatud.

Tulemused näitavad, et mõningaid semantiliselt seotud nimisõnade gruppe oli võimalik lihtsa Jaccardi sarnasuste abil klasterdamisega leida. Samas oli selle lihtsuse tõttu ka palju probleeme sõnade jagunemise ning eriti suurte klastrite tekkimisega, sest osa informatsioonist nimisõna ja omadussõna paaride kohta ei ole Jaccardi sarnasuste leidmisel arvesse võetud. Samuti ei ole võimalik teada, miks sõnad mingisse klastrisse paigutati, sest Jaccardi sarnasuste arvutamine ning nende põhjal klasterdamine ei säilita informatsiooni omadussõnade kohta.

3.2 Mittenegatiivne maatriksi faktorisatsioon

NMF mudel jaotab etteantud maatriksi kaheks väiksemaks, millest üks vastab nimisõnade jaotusele ning teine omadussõnade jaotusele klastrite vahel. Nende maatriksite abil on võimalik sõnu klastritesse jagada. Selles töös sai iga nimisõna määratud sellesse klastrisse, mille väärtus oli vastava nimisõna jaoks suurim. Esimesena uuris autor klastrite suuruseid. Klastrite suuruseid uurides oli näha, et oli kadunud jaotuse probleem, mis esines Jaccardi sarnasuste põhjal klasterdamisel. Kõige rohkemate nimisõnadega klaster sisaldas 186 nimisõna, mis on 3,2% kõikidest nimisõnadest. See näitab, et sõnad on ühtlasemalt jaotunud.

Teisena uuris autor omadusõnade jaotust. Omadussõnadest uuriti määravaid omadussõnu ning erinevatele omadussõnadele vastavaid klastreid. Määrav omadussõna selles kontekstis on omadussõna, mis on selles klastris suurima väärtusega ning mille väärtus selles klastris oli vähemalt 100 korda suurem järgmisest sõnast. Määravate omadussõnadega

klastrite uurimisel oli näha, et selliseid klastreid oli kokku 60, kusjuures unikaalseid omadussõnu nende hulgas oli 55. See tähendab, et 5 omadussõna olid kahe klastri jaoks määravad omadussõnad. Nendeks sõnadeks olid 'viimane', 'vaba', 'uus', 'eelmine' ja 'oluline'. Nendest nelja sõna puhul oli näha, et üks klaster, mida need sõnad kirjeldasid, oli pikem, kuid teine oli tühi klaster, mis ei sisaldanud ühtki nimisõna.

Selline olukord sai tekkida, sest iga nimisõna määrati vaid ühte klastrisse, mille jaoks väärtus oli kõige suurem ehk mis oli kõige kirjeldavam klaster selle sõna jaoks. Selle tagajärjel võib juhtuda, et mõni klaster ei ole kõige kirjeldavam ühegi nimisõna jaoks, nagu oli näha nendest näidetest.

Alternatiivselt saab uurida ka iga klastri jaoks suurimate väärtustega nimisõnu, mida autor nende näidete puhul ka tegi. See näitas, et nendes klastrites ka kõige suuremate väärtustega nimisõnade puhul olid väärtused siiski väiksed ning sõnad olid seetõttu ka vähem semantiliselt seotud. Üldiselt oli ühe määrava omadussõnaga klastrite näidetest näha, et eriti rohkem nimisõnu sisaldavate klastrite hulgas oli väga levinud omadussõnu. Näiteks 'uus', 'vana' ning 'suur' olid kõik ühe rohkem kui 100 sõna sisaldava klastri kirjeldavaks omadussõnaks. See tulemus on loogiline, sest tegemist on laialdaselt kasutatavate omadussõnadega, mis saavad esineda paaris väga paljude erinevate nimisõnadega.

Lisaks uuriti üldiselt omadussõnadele vastavaid klastreid. Tabelis 2 on välja toodud suurimate väärtustega nimisõnad igas klastris, mis vastasid valitud omadussõnadele. Igale omadussõnale on vastavaks valitud klaster, kus selle väärtus oli suurem kui kõigil teistel omadussõnadel. Kui klastreid, kus uuritav omadussõna oli suurima väärtusega, leidus mitu, valiti nendest selline, milles oli uuritava omadussõna väärtus suurim. Omadussõnadeks on valitud sageduselt esimesed kaks, kümnes ja sajas omadussõna [sag], milleks on vastavalt 'suur', 'uus', 'endine' ja 'märg'. Lisaks on valitud 100. sõna vastandsõnana omadussõna 'kuiv' ning värvidest omadussõna 'valge'. Punaselt on märgitud nimisõnad, mis esinesid küll vastavas klastris, kuid ei esinenud paaris uuritud omadussõnaga. Selline olukord sai tekkida, sest kuigi valitud omadussõnad olid nendes klastrites suurimate väärtustega, ei olnud need tihti ainsad määravad omadussõnad, vaid leidus ka teisi suurte väärtustega omadussõnu. Siniselt on märgitud omadussõnad, mis olid uuritavas klastris määravad. Nii on näiteks näha, et omadussõna 'uus' ei olnud uuritud klastris ainuke määrav sõna, kuigi see oli määravaks omadussõnaks kahe teise klastri puhul. Tegemist on laialdaselt kasutatava omadussõnaga, mistõttu määras NMF selle oluliseks sõnaks mitmes klastris.

Tabel 2. Valitud omadussõnadele vastavate klastrite suurimate väärtustega nimisõnad NMF meetodil

suur	uus	endine	valge	märg	kuiv
osa	leping	president	maja	riie	nahk
tänu	harjumus	peaminister	laev	asfalt	trenn
hulk	bakter	kommunist	hoone	haud	köha
tõenäosus	toime	nõukogu	mees	lumi	jalg
kogus	organism	kolleeg	värv	lapp	fakt
huvi	kõrvalmõju	esimees	suhkur	särk	hein

Autor uuris lisaks ühte klastrit määravatele omadussõnadele ning eelnevalt kirjeldatud omadussõnadele ka klastreid, kus olid kirjeldavateks omadussõnadeks värvid. Värvide puhul on tegu semantiliselt seotud sõnadega ning intuitiivselt võiksid need ühte klastrisse kuuluda. Selle kontrollimiseks uuris autor iga värvi kohta kolme kõige suurema väärtusega klastrit. Vaadeldud tulemustest oli näha, et värvid kirjeldasid enamasti erinevaid klastreid, vaid mõnel üksikul juhul oli kahe või kolme värvi puhul klastri osas kattuvus. Uuritud klastrite puhul oli märgata, et enamasti ei olnud värvid nende jaoks kõige kirjeldavamad omadussõnad. Sellest saab järeldada, et NMF puhul värvide kokku koondumist näha ei olnud.

Lisaks uuris autor tulemustes varasemalt kirjeldatud nimisõnagruppide jaotust. Võrreldes Jaccardi sarnasustest saadud klastritega on siin näha suuremat varieeruvust, kuna kasutada oli ka rohkem informatsiooni. Kuude hulgas oli näha, et kaheksa kuud jaotusid klastrite [’jaanuar’, ’juuni’, ’veebruar’] ja [’reede’, ’nädalavahetus’, ’laupäev’, ’pühapäev’, ’rong’, ’nädalalõpp’, ’pankrot’, ’trumm’, ’kuul’, ’detsember’, ’kuulujutt’, ’oktoober’, ’september’, ’tera’, ’november’, ’välismaa’, ’august’, ’liig’] vahel. Kui neist kahest esimene sisaldab vaid kuid, siis teises on näha ka semantiliselt kaugemaid sõnu. Uurides omadussõnu nende klastrite jaoks, oli teise väljatoodud klastri puhul näha, et määravaks omadussõnaks selles oli sõna ’läinud’. Sõna ’juuli’ oli määratud klastrisse, kus määravaks omadussõnaks oli ’kuum’. Ülejäänud kuud olid määratud ühte suuremasse klastrisse.

Nagu on eelnevalt kirjeldatust näha, siis ühte klastrisse määrati lisaks viiele kuule ka kolm päeva ning sõnad ’nädalavahetus’ ja ’nädalalõpp’. See tulemus on loogiline, sest erinevad ajale viitavad sõnad on semantiliselt rohkem seotud ning kuudest ja päevadest räägitakse tihti ka sarnastes kontekstides. Lisaks tuli välja 90-sõnaline klaster, millel oli samuti üks määrav omadussõna ning selleks oli ’eelmine’. Selles klastris oli mitu päevale ja aastaajale viitavat sõna, aga muuhulgas ka sõnad nagu ’nädal’, ’kuu’, ’valimine’, ’hooaeg’, ’semester’, ’olümpia’. Uurides elukutsete klastreid tuli välja 118-sõnaline

klaster, kus lisaks kolmele eelnevalt mainitud elukutsele oli peamiselt veel mitmeid inimestele viitavaid sõnu. Nende hulgas olid näiteks sõnad ’naine’, ’õpetaja’, ’spetsialist’, ’peategelane’, ’inglane’. Sellel klastril oli samuti määrav omadussõna, milleks oli ’noor’.

Lisaks eelnevalt kirjeldatud sõnadele uuris autor juhuslikke näiteid saadud klastritest. Kui Jaccardi sarnasuste põhjal klasterdamises oli üheks saadud klastriks [’inimene’, ’mees’, ’naine’], siis NMF mudeliga sõnade grupeerimisel tekkis klaster [’inimene’, ’mees’]. Sarnaselt eelneva lõigu lõpus kirjeldatud klastrile on selles esimeseks kirjeldavaks omadussõnaks ’noor’. Selles klastris ei ole see aga ainsaks määravaks omadussõnaks, lisaks sellele on järgnevate kirjeldavate omadussõnade seas muuhulgas näiteks sõnad ’elav’ ja ’vana’. Juhuslikult uuritud klastrite hulgas oli ka näiteks [’õhtutund’, ’öötund’, ’ärkaja’] ning [’kontroll’, ’dieet’, ’järgimine’, ’julgeolekumeede’, ’kvaliteedinõue’, ’turvanõue’].

3.3 Dirichlet’ peitlahutus

LDA mudeli klasterdamise tulemusena saab iga nimisõna kohta tõenäosusvektori, mis näitab iga klastri ehk teema kohta, kui suure tõenäosusega see nimisõna sinna kuulub. Analoogselt NMF meetodiga sai ka LDA puhul iga nimisõna määratud sellesse klastrisse, millele vastav väärtus oli suurim. Samuti väljastab LDA mudel iga teema kohta omadussõnajaotuse.

Sarnaselt eelneva kahe meetodiga uuriti esimesena klastrite suurust. Suuruste jaotus oli sarnane NMF meetodiga, mis tähendab, et nimisõnad olid ühtlasemalt jaotunud erinevate klastrite vahel. Saadud klastritest kolm sisaldasid rohkem kui 100 sõna, kusjuures suurim klaster sisaldas 408 nimisõna, mis on 7,0% kõikidest nimisõnadest. See tähendab, et sarnaselt NMF meetodiga oli ka LDA puhul näha, et kadunud oli probleem, mis esines Jaccardi sarnasuste põhjal klasterdamise korral.

Sarnaselt NMF meetodiga võimaldab LDA mudel samuti uurida omadussõnade jaotust klastrite vahel. Eelkõige uuriti määravaid omadussõnu ning erinevatele omadussõnadele vastavaid klastreid.

Määrava omadussõnaga klastreid oli LDA mudeli tulemustes kokku 46. Sarnaselt NMF tulemustele oli näha, et vaadeldud 46 klastri hulgas olid levinud omadussõnad sellised, mis kirjeldasid paljude nimisõnadega klastreid. LDA puhul olid omadussõnad ’uus’, ’suur’ ja ’hea’, mis olid määravaks klastritel, mis sisaldasid rohkem kui 100 sõna, millest ’uus’ ja ’suur’ olid ka NMF mudeli puhul suurimate klastrite määravateks omadussõnadeks.

Lisaks määravate omadussõnade uurimisele uuriti LDA tulemuste puhul samuti üldiselt omadussõnadele vastavaid klastreid. Uuritud omadussõnad on samad, mis olid kirjeldatud NMF meetodi puhul. Tabelis 3 on välja toodud iga uuritud omadussõna kohta klaster, mille jaoks oli see suurima väärtusega omadussõna, ning nimisõnad, mis kuulusid suurima tõenäosusega sellesse klastrisse. Sarnaselt NMF meetodile on punaseks värvitud

nimisõnad, mille tõenäosus vastavasse klastrisse kuuluda oli suur, kuid mis ei olnud vastava omadussõnaga paaris. Siniseks on värvitud omadussõnad, mis olid ainsaks määravaks sõnaks selle klastri puhul.

Tabel 3. Valitud omadussõnadele vastavate klastrite suurimate väärtustega nimisõnad LDA meetodil

suur	uus	endine	valge	märg	kuiv
tänu	neer	tippsportlane	triiksärk	haud	suutlikkus
slämm	tulija	töökaaslane	pulber	hiis	köha
tõenäosus	elurajoon	komissar	vares	teekate	raev
töökoormus	tuumajaam	kasvandik	põll	plekk	hein
vanker	ravikindlustusseadus	alluv	suhkur	lapp	siider
õhin	põhimäärus	raamatupidaja	sukkpüksid	t-särk	viha

Lisaks tabelis toodud klastritele tuli sõnale ’kuiv’ vastavaid klastreid uurides välja klaster, kus kümme suurima väärtusega omadussõna olid ’ilus’, ’soe’, ’külm’, ’halb’, ’vihmane’, ’kuum’, ’kuiv’, ’hea’, ’jahe’ ja ’palav’. Nendest esimese omadussõna skooriks (*pseudo-count*) oli 1737.54 ja kümnenda omadussõna skooriks oli 432.88, seega vastava klastri omadussõnad olid ühtlasemalt jaotunud ning ühel omadussõnal ei olnud oluliselt suuremat mõju. Vaadates nimisõnu, oli näha, et teistest sõnadest suurema tõenäosusega paigutusid sellesse klastrisse ’ilm’, ’juuni’, ’jaanipäev’, ’kliima’, ’hommik’, ’suvi’ ja ’suvepäev’. Omadussõnale ’kuiv’ vastavaid klastreid uurides selgus selle klastri osas, et sõnad ’juuni’, ’jaanipäev’, ’hommik’ ja ’suvepäev’ sõnaga ’kuiv’ paaris ei esinenud, kuid on siiski sõnad, mida on võimalik koos kasutada. See näitab, et LDA mudel oli võimeline leidma seoseid, mida andmestikus otseselt välja toodud ei olnud.

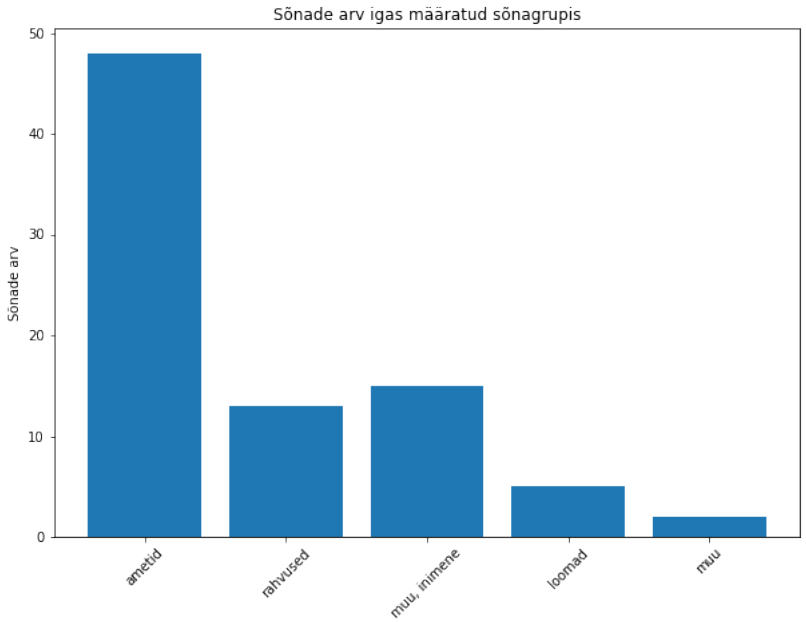
LDA puhul uuris autor samuti värve sisaldavaid klastreid. Sarnaselt NMF mudelile oli märgata, et värvid jaotusid peamiselt erinevatesse klastritesse. Kui vaadata iga värvi jaoks ainult seda klastrit, mille jaoks oli selle väärtus suurim, ei olnud ühtki värvipaari, mille klastrid oleksid kattunud. Erinevalt NMF mudelist oli LDA puhul näha, et kõik põhivärvid peale värvide ’lilla’ ja ’oranž’ olid nende jaoks vaadeldud klastrites kõige kirjeldavamad omadussõnad. Nendes klastrites olevate nimisõnade uurimine näitas, et enamasti sisaldasid need klastrid selliseid sõnu, mida peamiselt kirjeldatakse just nende värvustega. Tabelis 4 on igale värvile vastava klastri kohta viis nimisõna, mille tõenäosus sellesse klastrisse kuuluda oli suurim. Tabelist on välja jäetud sõnad ’lilla’ ja ’oranž’, sest need ei olnud oma klastrites kõige suurema väärtusega omadussõnad, ning ’valge’, sest see on välja toodud eelmises tabelis. Punaseks on värvitud nimisõnad, mis selle värviga paaris ei esinenud.

Tabel 4. Värvidele vastavate klastrite suurimate väärtustega nimisõnad LDA meetodil

kollane	roosa	punane	pruun	roheline	sinine	hall	must
liidrisärk	elevant	latern	riis	konn	vilkur	kardinal	prügikott
muutkond	põrsas	tellis	katk	aas	teksased	argipäev	pipar
kaart	ärileht	rist	kääbus	mehike	silmaalune	tsoon	nahktagi
kääbus	majandusleht	kurat	nahktagi	spargel	paviljon	habe	lammas
liblikas	õis	rätt	vitamiin	vöönd	ingel	kaabu	masendus

Sarnaselt eelneva kahe meetodiga uuris autor lisaks eelnevalt kirjeldatud sõnagruppe. Selle jaoks uuriti klastreid, mis tekkisid, kui iga nimisõna paigutati sellesse klastrisse, millesse kuulumise tõenäosus oli suurim. Vaadeldes kuid ning päevi sisaldavaid klastreid tuli välja järgnev klaster: [’reede’, ’sügis’, ’nädalavahetus’, ’kolmapäev’, ’teisipäev’, ’laupäev’, ’pühapäev’, ’nädalalõpp’, ’detsember’, ’oktoober’, ’september’, ’november’]. Selle puhul on selgelt näha, et tegemist on semantiliselt tähendusliku klastriga, sest kokku on grupeeritud erinevad päevadele ning kuudele vastavad sõnad, eriti just sellised päevad ja kuud, mis esinevad vastavalt kas nädala või aasta lõpus. Kuid vaadeldes tulid samuti ka välja klastrid [’suvepäev’, ’pirukas’, ’pann’, ’pliit’, ’võileib’, ’august’, ’juuli’, ’armulugu’] ning [’ilm’, ’juuni’]. Neist esimese puhul oli kõige kirjeldavam omadussõna ’kuum’ ning teise puhul olid kõige kirjeldavamad omadussõnad ’ilus’, ’soe’ ja ’külm’. Kuigi eriti esimese klastri puhul ei ole kõik sõnad semantiliselt seotud, on omadussõnu uurides näha, et inimeste keelekasutuses on need seotud, sest neid kirjeldatakse sarnaste omadussõnadega. Päevi sisaldavaid klastreid uurides oli näha lisaks ka klastrit [’aasta’, ’hooaeg’, ’sajand’, ’esmaspäev’].

Uurides klastreid, mis sisaldavad varasemalt kirjeldatud elukutseid, oli näha, et kõik viis mainitud ametit olid paigutatud ühte 83-sõnalisse klastrisse. Suurem osa sellesse klastrisse kuuluvatest sõnadest on erinevad ametid, kuid lisaks sellele leidus ka muid inimestele viitavaid sõnu. Selle klastri puhul on tegemist semantiliselt tähendusliku klastriga, sest on näha, et mudel oskas kokku grupeerida inimestele viitavaid sõnu, mis on omavahel rohkem seotud kui suure hulga teiste sõnadega andmestikus. Selle klastri kolm kõige kirjeldavamat omadussõna on ’noor’, ’andekas’ ja ’maailmakuulus’. Joonisel 13 on välja toodud, mitu sõna oli igas määratud grupis. Grupid määrati käsitsi vastavalt klastris olevatele sõnadele ning nende loogilisele jaotusele. Grupid on määratud selliselt, et oleks näha semantilist seotust saadud klastris. Grupis ’muu, inimene’ on kõik inimestele viitavad sõnad, mis ei ole ametid või rahvused, ning grupis ’muu’ on sõnad, mis teistesse gruppidesse ei sobinud.



Joonis 13. Sõnade arv igas määratud sõnagrupis ühes LDA mudeli poolt leitud klastris

3.4 Järeldused

Tulemuste analüüsist oli näha, et Jaccardi sarnasuste põhjal klasterdamine andis kõige halvemaid tulemusi. See tuleneb sellest, et Jaccardi sarnasuste arvutamine kasutab vähem informatsiooni andmete kohta. Samuti ei väljastanud klasterdamise tulemus informatsiooni omadussõnade kohta, mis tähendab, et tulemuseks sai nimisõnade klastrid, kuid ei olnud teavet selle kohta, miks need kokku on grupeeritud.

Nii LDA kui ka NMF meetod töötasid võrreldes Jaccardi sarnasuste põhjal klasterdamisega paremini. Jaccardi sarnasuste põhjal loodud klastritest oli näha, et üks klaster sisaldas 48.7% sõnadest. Nii suure klastri näol ei ole tegemist informatiivsete tulemustega. LDA ja NMF mudelitest saadud tulemustes oli näha, et sõnad olid jaotunud ühtlasemalt ning seega see probleem oli lahenenud. LDA ja NMF mudelitest saadud suurimates klastrites oli vastavalt 7,0% ja 3,2% kõikidest nimisõnadest.

Erinevalt Jaccardi sarnasuste põhjal klasterdamisest on LDA ja NMF puhul võimalik võrrelda ka omadussõnade jaotust klastrite vahel. Ühe määrava omadussõnaga klastrite puhul oli näha sarnaseid tulemusi. Mõlemal juhul oli näha, et suurematel ühe omadusega klastritel oli määrav omadussõna mõni levinum omadussõna, millega saab kirjeldada laia hulka erinevaid nimisõnu. Nii LDA kui ka NMF puhul oli näha, et sõnad ’uus’

ja ’suur’ olid määravad omadussõnad klastritel, millesse kuulus üle 100 sõna. Uurides üldiselt omadussõnadele vastavaid nimisõnade jaotusi oli näha, et mõlemate meetodite puhul suutis mudel suure tõenäosusega klastrisse paigutada sõnu, mis nende omadustega loogiliselt kokku käivad. LDA puhul oli veel näha, et klastrisse, mille jaoks kirjeldavad omadussõnad olid muuhulgas ’ilus’, ’soe’, ’külm’, ’vihmane’ ja ’kuiv’, kuulusid suurima tõenäosusega sõnad, millele saab viidata ilma kirjeldavate omadussõnade abil, isegi kui need esinesid paaris vaid mõne klastri jaoks kirjeldava omadussõnaga. See näitab, et LDA on võimeline leidma loogilisi seoseid, mida andmestikus otseselt ei leidu. Analoogseid seoseid leidis ka NMF mudel ja LDA mudel teiste klastrite hulgas. Nendest oli näha, et kuigi leiti loogilisi seoseid, leidsid mudelid ka seoseid paaride vahel, mida tegelikkuses koos ei kasutata. LDA tulemused olid paremad värvidele vastavaid klastreid uurides. Kui NMF puhul ei olnud värvide osas loogilisi tulemusi näha, siis LDA mudel leidis peaaegu kõikidele värvidele vastavad klastrid, millesse kuulusid suurima tõenäosusega enamasti nimisõnad, mida kirjeldatakse nende värvidega.

Nimisõnade klasterdumist võrreldes oli samuti näha, et Jaccardi sarnasuste põhjal klasterdamine andis kõige halvemaid tulemusi. Jaccardi klastrite hulgas oli väiksemaid semantiliselt seotud sõnade klastreid, kuid kuna peaaegu pooled sõnad paigutusid ühte klastrisse, ei ole üldiselt tegemist informatiivsete tulemustega. LDA tulemused olid nimisõnade osas kõige paremad, sest kõige selgemalt oli näha semantiliselt seotud sõnade kokku klasterdumist.

Üldiselt võib väita, et LDA poolt saadud tulemused olid kõige paljulubavamad. Tulemusi oleks võimalik veelgi parandada, kui paaride leidmiseks kasutaks täpsemaid meetodeid (näiteks süntaktilist analüüsi) ning rohkem tekste.

Kokkuvõte

Töö käigus loodi andmestik omavahel seotud nimisõna ja omadussõna paaridest, mis sisaldas informatsiooni selle kohta, mitu korda iga paar tekstides koos esines. Saadud andmestiku põhjal klasterdati nimisõnu kolme erineva mudeliga ning analüüsiti saadud tulemusi.

Tulemustest oli näha, et Jaccardi sarnasuste põhjal klasterdamine andis kõige halvemaid tulemusi. Jaccardi klastrite juures oli probleemiks see, et klasterdamiseks kasutati informatsiooni vaid paaride leiduvuse kohta, mitte selle kohta, mitu korda iga paari leidus. Lisaks ei olnud tulemustel juures teavet selle kohta, miks kindlad klastrid tekkisid. Tulemustes oli üks suur klaster, mis sisaldas 48,7% kõigist nimisõnadest, mis ei ole informatiivne tulemus.

Üldiselt oli parimaid tulemusi näha LDA mudeli klastrite hulgas. Nimisõnade klastreid uurides oli LDA mudeli puhul näha rohkem semantiliselt seotud klastreid kui ülejäänud kahe mudeli tulemustes. Lisaks nimisõnade klastritele võimaldavad NMF ja LDA uurida ka omadussõnade jaotust. Uuriti ühe määrava omadussõnaga klastreid ning üldiselt erinevatele omadussõnadele vastavaid klastreid. Ühe määrava omadussõnaga klastreid uurides olid LDA ja NMF mudelite tulemused võrreldavad. Uurides erinevatele omadussõnadele vastavaid klastreid, oli LDA puhul märgata paremaid tulemusi. LDA puhul oli näha loogilisi värvidele vastavaid klastreid ning samuti oli LDA võimeline leidma loogilisi seoseid, mida andmestikus otseselt välja toodud pole.

Tulemuste edasiseks parandamiseks oleks võimalik kasutada täpsemaid meetodeid sõnapaaride leidmiseks ning kasutada suuremaid andmestikke, et klasterdatavas andmestikus oleks rohkem sõnapaare.

Viidatud kirjandus

[AHA17] Abdelkarim Ben Ayed, Mohamed Ben Halima, and Adel M. Alimi. Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering. In *2017 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2017, Naples, Italy, July 9-12, 2017*, pages 1–6. IEEE, 2017.

[BHM17] Jordan L. Boyd-Graber, Yuening Hu, and David M. Mimno. Applications of topic models. *Found. Trends Inf. Retr.*, 11(2-3):143–296, 2017.

[BNJ01] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press, 2001.

[ekk] Eesti keele koondkorpus. `https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et`. (10.12.2020).

[Ere13] Mati Erelt. *Eesti keele lauseõpetus: Sissejuhatus. Öeldis*. Tartu: Tartu Ülikooli eesti keele osakond, 2013.

[FBEJ17] Yue Feng, Ebrahim Bagheri, Faezeh Ensan, and Jelena Jovanovic. The state of the art in semantic relatedness: a framework for comparison. *Knowl. Eng. Rev.*, 32:e10, 2017.

[Jac12] Paul Jaccard. The distribution of the flora in the alpine zone. In *The New Phytologist*, volume 11, pages 37–50. 1912.

[JM19a] Dan Jurafsky and James H. Martin. Constituency grammars. In *Speech and Language Processing (3rd ed. draft)*, chapter 12, pages 203–231. 2019.

[JM19b] Dan Jurafsky and James H. Martin. Constituency parsing. In *Speech and Language Processing (3rd ed. draft)*, chapter 13, pages 232–245. 2019.

[Jon72] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[KBB17] Da Kuang, P. Jeffrey Brantingham, and Andrea L. Bertozzi. Crime topic modeling. *CoRR*, abs/1701.01505, 2017.

[KBV09] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[LBM03] Yuhua Li, Zuhair Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, 15(4):871–882, 2003.

[LOST20] Sven Laur, Siim Orasmaa, Dage Särg, and Paul Tammo. Estnltk 1.6: Remastered Estonian NLP pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 7152–7160. European Language Resources Association, 2020.

[LS99] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. In *Nature*, volume 401, pages 788–791. 1999.

[MM16] Kadri Muischnek and Kaili Müürisep. Eesti sõltuvuspuude pank ja selle keeleteoreetilised lähted. In *Emakeele Seltsi aastaraamat*, volume 62, pages 122–145. 2016.

[NdV04] Joakim Nivre, Koenraad de Smedt, and Martin Volk. Treebanking in Northern Europe: A white paper. *Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram*, 2000–2004.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 849–856. MIT Press, 2001.

[Res95] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 448–453. Morgan Kaufmann, 1995.

[sag] Eesti kirjakeele sagedussõnastik. `https://www.cl.ut.ee/ressursid/sagedused/index.php?lang=et`. (04.05.2021).

[XI09] Rui Xu and Donald C. Wunsch II. *Clustering*. John Wiley Sons, Inc., 2009.

Lisad

I. Repositoorium

Töös kasutatud kood on kättesaadav lingilt:
`https://github.com/birgitsormus/noun-adj-classification`

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Birgit Sõrmus**,
*(*autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Nimisõnade klasterdamine vastavalt neid kirjeldavatele omadussõnadele,
   *(*lõputöö pealkiri)

   mille juhendaja on Sven Laur,
   *(*juhendaja nimi)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Birgit Sõrmus
07.05.2021