TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Kermo Saarse

Tervisesündmuste üldistamine sõnavektorite abil

Magistritöö (15 EAP)

Juhendaja: Sulev Reisberg, PhD

Tartu 2024

Tervisesündmuste üldistamine sõnavektorite abil

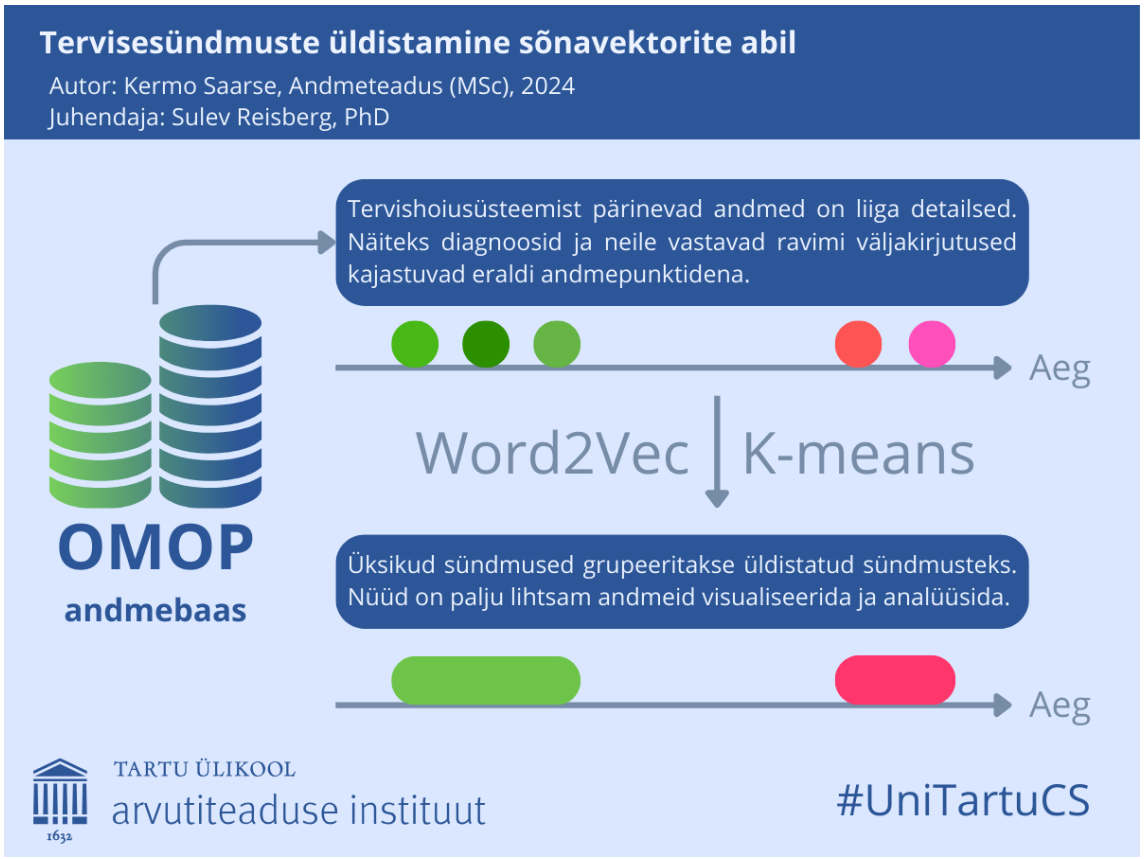
Lühikokkuvõte

Tervishoiusüsteemis võib iga arstivisiit tekitada mitmeid andmepunkte. Sama terviseprobleemiga võivad olla seotud mitmed diagnoosid, ravimi väljakirjutused ja mõõtmised, mis kõik on erinevad sündmused. Andmete detailsus teeb aga nende analüüsimise keerukaks. Selles töös agregeeritakse üksteisega seotud tervisesündmused üldistatud sündmusteks, kasutades word2vec mudelit ja K-means klasterdamist ning demonstreeritakse seda OMOP CDM formaadis andmestikul. Tulemused näitavad, et word2vec suudab edukalt tuvastada sisult sarnaseid tervisesündmusi. Mida rohkem tekitada klastreid, seda ühetaolisem on klastrite sisu, kuid seda rohkem on ka üksteisega sarnaseid klastreid. Üldistamise tulemusena vähenes oluliselt patsiendi andmestikus olevate sündmuste arv.

Võtmesõnad: Word2vec, K-means, OMOP CDM, ICD10

CERCS: P176 - Tehisintellekt, B110 - Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Visuaalne kokkuvõte



Generalizing Healthcare Events Using Word Vectors

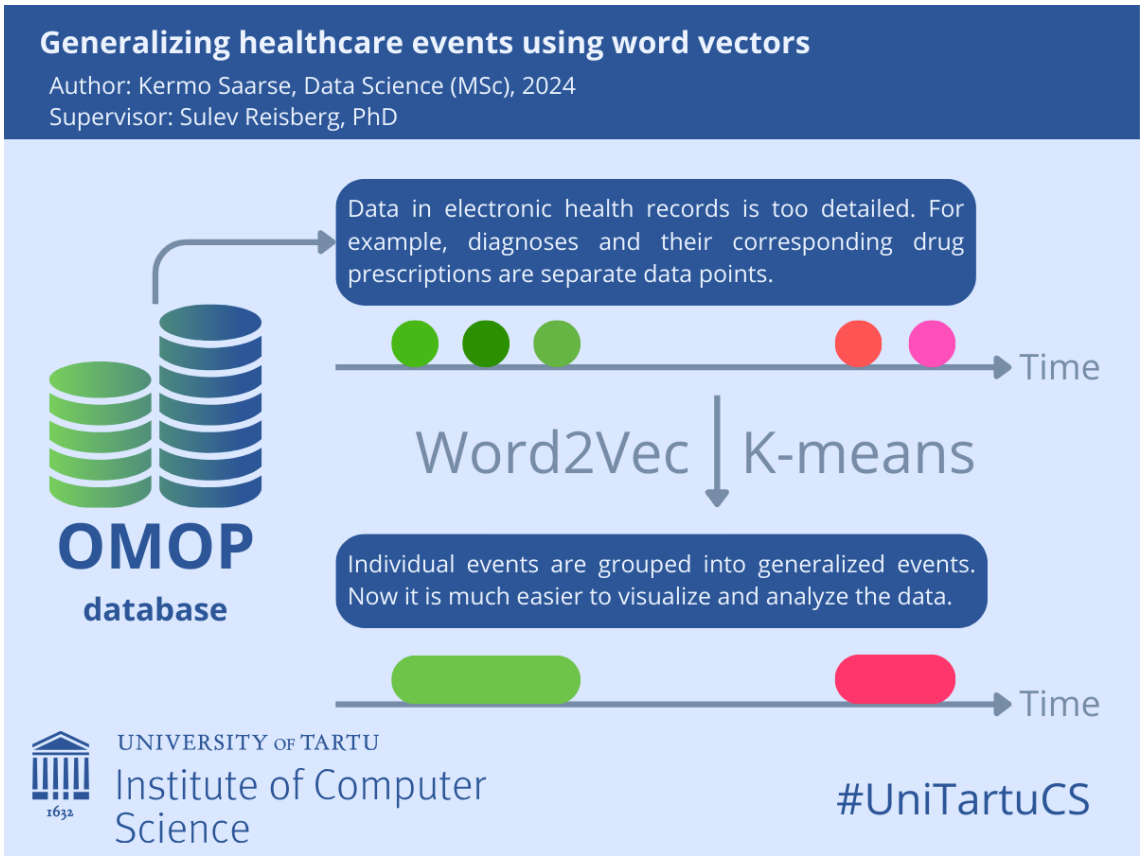
Abstract

In the electronic health record, each visit to doctor could generate multiple data points. The same health issue could be linked to multiple diagnoses, drug prescriptions and measurements that are all separate events. Such a high resolution of the data makes its analysis difficult. In this thesis, word2vec model and K-means clustering are used to aggregate related health events into generalised events in an OMOP CDM dataset. It is shown that word2vec can successfully identify related events. As the number of clusters grows, each cluster becomes more homogenous, but there will also be a higher number of similar clusters. As a result of generalization, the number of events in a patient's dataset decreased significantly.

Keywords: Word2vec, K-means, OMOP CDM, ICD10

CERCS: P176 - Artificial intelligence, B110 - Bioinformatics, medical informatics, biomathematics, biometrics

Visual abstract



Sisukord

1 Sissejuhatus

Kui inimene pöördub perearsti poole oma terviseprobleemiga, tekitab see andmeid. Perearst võib mõõta vererõhku, kirjutada saatekirja mingi proovi võtmiseks, laboris saadakse mõõtmistulemused ning lõpuks määrab perearst diagnoosi ja kirjutab ravimi välja. Kõik need sündmused kajastuvad tervishoiusüsteemis eraldi andmepunktidena. Andmete analüüsi seisukohalt oleks aga kasulikum teada, millal inimesel haigus algas ja millal lõppes, kuid selleks oleks vaja andmeid üldistada. Seega, kõik need eri sündmusi kajastavad andmepunktid oleks vaja üldistada üheks sündmuseks, kui need on seotud sama terviseprobleemiga. Kui inimene aga mitme eri haigusega perearsti (või mitme eriarsti) vastuvõtul käib, võib olla keeruline eri haigustega seotud andmepunkte üksteisest eristada.

Keeletöötlusest tuntud algoritmi word2vec [Mik+13] idee on selles, et samas lauses tekstis järjestikku paiknevad sõnad on omavahel seotud. Algoritm leiab igale sõnale numbrilise vektori, mis väljendab sõna tähendust erinevates kontekstides. Sarnase tähendusega sõnade vektorid on samuti sarnased ning ka vastupidi.

Ka ühe patsiendi terviseandmeid võib vaadelda kui lauset, milles iga sõna on tervisesündmus. Sellisele „tekstile“ on võimalik rakendada word2vec algoritmi ning tulemuseks on unikaalne vektor iga sündmuse kohta. Sarnases kontekstis toimuvad sündmused (nt. haiguse diagnoos ning selle vastu loodud ravimi välja kirjutamine) võivad saada sarnase vektori, kui need esinevad tihti üksteisele lähedal. Seepärast pakub huvi küsimus, kas tervisesündmusi on võimalik word2vec'i abil üldistada.

Töö uurimisküsimused on järgnevad:

1. Kuidas leida sarnaseid tervisesündmusi word2vec'i abil?
2. Kuidas sarnaseid tervisesündmusi üldistada ja leida üldistustele nimesid?

Töö ülejäänud osa koosneb järgmistest peatükkidest. Peatükis 2 antakse ülevaade ajaliselt järjestatud sündmuste analüüsist ning word2vec algoritmist selles valdkonnas. Peatükk 3 kirjeldab töö sisendandmete struktuuri ning nendega läbi viidud andmeanalüüsi. Tulemusi kirjeldatakse peatükis 4 ning arutelu on peatükis 5.

Töö käigus loodud kood word2vec mudeli treenimiseks ja kasutamiseks on saadaval GitLab'i projektina, mille link on lisas I.

2 Kirjanduse ülevaade

Mitmes valdkonnas on vaja analüüsida andmeid, mis kirjeldavad suurt hulka ajalises järjestuses toimunud sündmuseid. See peatükk kirjeldab mõningaid analüüsimeetodeid selliste andmete jaoks. Samuti on on kirjeldatud word2vec algoritmi tööpõhimõtet ning tuuakse näiteid selle kasutusest sündmuste järjendi analüüsis.

2.1 Ilma word2vec'ita sündmuste üldistamise meetodid

EventThread on mudel [Guo+18], mis sobib mistahes valdkonnast pärit sündmuste järjendi analüüsiks, kuigi see töötati välja meditsiinivaldkonna vajadusi arvestades. Igal südmusel on üksus (ingl. *entity*), millega sündmus toimub. Selleks võib olla patsient, kellele määratakse diagnoos või ka auto, millel viiakse läbi hooldust. Igale sündmusele arvutatakse TF-IDF skoor. TF-IDF on meetod sõna tähtsuse hindamiseks dokumendis milles need esinevad. Sõnad, mis esinevad sagedasti antud dokumendis, kuid harva teistes dokumentides (nt. erialased terminid) saavad kõrgema skoori kui sõnad, mis on sagedased igas dokumendis (nt. sidesõnad). Antud töös olid dokumentideks sündmuste järjendid ning sõnadeks sündmused. Väga madala TF-IDF skooriga sündmused loetakse müraks ning eemaldatakse andmestikust. Seejärel kõrvutatakse eri üksustega toimunud sündmuste seeriad selliselt, et sündmuste täpne toimumise aeg ei ole oluline, loeb vaid järjekord. Sündmuste seeriad jaotatakse fikseeritud pikkusega (või sündmuste arvuga) intervallideks ning nende põhjal moodustatakse 3-mõõtmeline maatriks (tensor) X, mille mõõtmeteks on sündmuste arv intervallis (M), intervallide arv (T) ning üksuste arv (N). Tensori X abil leitakse sündmuste lõimed (ingl *thread*) ehk sagedasti korduvad sündmuste toimumise mustrid. Nende leidmiseks eraldatakse tensor X uuteks tensoriteks λ, A, B ja C, mille mõõtmed on vastavalt $K \times K \times K$, $N \times K$, $T \times K$ ja $M \times K$. Sealjuures on K lõimede arv ning tensorite λ, A, B ja C korrutamine annab algse tensori X. Iga maatriksi B element B_{tk} on tõenäosus, et lõim k toimub intervallis t. Elemendid A_{ik} ja C_{jk} on lõime k seos vastavalt i-nda üksuse ja intervalli j-nda sündmuse vahel. Autorid rakendasid EventThread mudelit auto hooldusajaloo sündmuste peal ning leidsid, et täissünteetilist õli kasutavad autod on üldiselt paremas seisukorras kui poolsünteetilist õli kasutavad autod.

[Hua+13] keskendub haigla sündmuste logist patsientide ravitrajektooride üldistamisele. Artiklis toodi välja, et tavalised protsessikaeve meetodid annavad meditsiinilistel andmetel ebaselgeid tulemusi, kuna tervishoiu protsessid on palju keerukamad kui tavalised äriprotsessid. Näiteks kui patsient saab korraga kahe või enama haiguse ravi, kajastuvad ka sündmuste logis erinevate haigustega seotud sündmused läbisegi. Töö eesmärk oli eri patsientidega ja samaaegselt toimuvad sündmuste jadad üldistada etappideks, mis on kõigil patsientidel sarnased. Iga sündmust haigla logis iseloomustab kolm atribuuti: patsiendi ID, tegevus ning toimumise aeg. Sündmuste jadad jagati mitmeks intervalliks ning igas intervallis leiti mingi hulk tegevusi, millel oleks võimalikult suur ühisosa iga patsiendiga läbi viidud tegevustega selles intervallis. Töö tulemusena avastati, et Hiina tervishoiuministeeriumi ning kliiniliste ekspertide väljatöötatud ravijuhised erinevad

reaalselt läbiviidavatest raviteekondadest.

2.2 Word2vec

Word2vec on keeletöötlusest tuntud mudel, mille eesmärk on anda igale tekstikorpuses olevale sõnale vektor, mis võtab arvesse kõiki kontekste ja tähendusi, milles see sõna võib esineda. Algoritm liigub fikseeritud pikkusega aknaga üle teksti. Igal akna positsioonil treenitakse kahe kihiga närvivõrku. Mudelil on kaks erinevat arhitektuuri. CBOW (*Continuous Bag-of-Words*) arhitektuur annab närvivõrgu sisendiks akna äärtes olevad sõnad ning soovitud väljundiks on akna keskel olev sõna. *Skip-gram* arhitektuur on vastupidine. Treenitud närvivõrgu esimesel kihil on kaalude maatriks, mille igast reast saab ühe sõna vektor.

2.3 Word2vec'i kasutamine sündmuste üldistamiseks

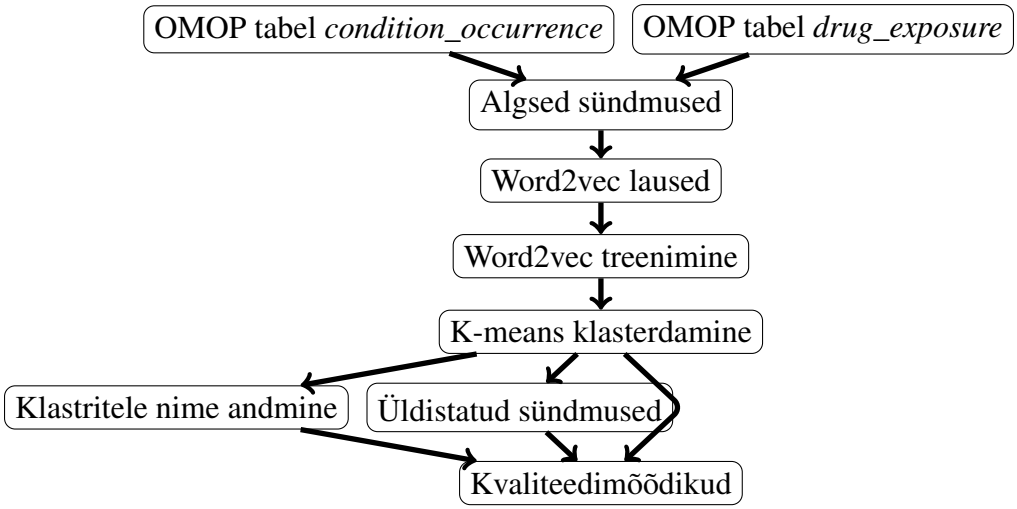
Mitmed uurijad on kasutanud word2vec mudelit väljaspool keeletöötlust nii, et lausete ja sõnade asemel on mingid muud järjendid, mis tähistavad sündmuste jadasid.

Mudelis trace2vec [DBD18] on sõnadeks mistahes valdkonnast pärit sündmused ning lauseteks sündmuste jadad. Igal lausel on oma ID, mida käsitletakse samuti sõnana ning mis pannakse iga word2vec konteksti akna lõppu. Seega saab lisaks igale sõnale ka iga lause omale vektori. Saadud lausete vaktoreid on võimalik klasterdada. Töö autorid treenisid trace2vec mudeli viiest Hollandi piirkonnast pärit sündmuste logide peal. Igal lausel on küljes piirkonna silt ning kui klasterdada lausete vektoreid viie klastriga, on võimalik mõõta klasterduse vastavust piirkondadele.

Word2vec'i on tervishoiusündmuste uurimiseks kasutanud [Sii23]. Selles töös olid sisendandmeteks raviarvetelt tulevad teenuste koodid ajalises järjestuses. Koode käsitleti sõnadena ning samalt raviarvelt pärinev koodide jada moodustas ühe lause. Koodidele leiti sõnavektorid kahe keeletöötlusest pärit masinõppemudeliga: word2vec ja BERT [Dev+19]. Kui word2vec'ist saadavad sõnavektorid on staatilised, s.t need ei sõltu sõna kontekstist, siis BERT'i sõnavektorid on kontekstist sõltuvad. Peale mudelite treenimise esimest faasi peideti osad teenuste koodid lausetest ära ning mudelitel lasti puuduolevat koodi ennustada. Selles ülesandes andis word2vec paremaid tulemusi. Põhjuseks võis olla see, et erinevalt BERT'ist ei vaatle word2vec sõnade järjekorda lauses. Teises faasis kasutati mõlemat mudelit teatud teenuste koodidele vastavate tunnuste ennustamiseks. Word2vec'i puhul oli selleks vaja sõnavektorite peal treenida juhendatud masinõppemudel, kasutati K-lähima naabri mudelit ja tugivektor-masinat. Selles faasis osutus BERT edukamaks.

3 Andmed ja metoodika

See peatükk selgitab andmete päritolu, nende formaati ning analüüsi. Joonisel 1 on kokkuvõttev skeem.



Joonis 1. Andmete analüüsi skeem

3.1 Andmed

Andmed pärinevad [Oja+23] käigus loodud andmebaasist, mis sisaldab 10% Eesti elanike terviseandmeid OMOP CDM [SI] formaadis. OMOP CDM on terviseandmete analüütikas kasutatav standardne andmebaasi formaat, milles on defineeritud tabelite ja veergude nimed ning eri tabelite vahelised seosed. Tervishoiusüsteemist pärinevad andmed on teistes formaatides, kuid OMOP CDM kujule viiakse andmed üksnes teaduslikuks uurimistööks.

Andmestik koosneb sündmustest, milleks on muuhulgas diagnoosid, ravimi väljakirjutused ja ravimi väljaostmised. Iga sündmus toimub mingi patsendiga ja kindlal kuupäeval ning pärineb mõnest dokumendist, näiteks digiretseptist või haigusloo kokkuvõttest ehk epikriisist. Otse tervishoiusüsteemist pärinevates andmetes on igal sündmusel kood, mis kirjeldab sündmuse tüüpi. Näiteks on igal diagnoosil ICD10 [Org19] kood. ICD10 on rahvusvaheline standard, mis seab erinevatele haigustele ja seisunditele vastavusse tähtede ja numbrite kombinatsiooni, näiteks ägeda bronhiidi kood on J20. Kuna koodid võivad pärineda mitmetest erinevatest klassifikaatoritest, mitte ainult ICD10-st, seati andmete OMOP CDM kujule viimsel kõikidele sündmustele vastavusse uus kood, mis pärineb OHDSI sõnastikust [Ser] ja mida nimetatakse kontseptsiooniks (ingl *concept ID*). Ägedale bronhiidile vastab kontseptsioon 260139.

Sündmused võeti tabelitest *condition_occurrence* (diagnoosid, 20 351 014 rida) ja *drug_exposure* (ravimi väljakirjutamised ja väljaostmised, 16 159 423 rida). Nendes tabelites vastab iga rida ühele sündmusele. Mõlemast tabelist jäeti välja sellised sündmused, millele ei õnnestunud sobivat kontseptsiooni leida. Seega tabelist *condition_occurrence*

eemaldati read, millel veeru *condition_concept_id* väärtus oli 0, alles jäi 20 333 065 rida. Tabelist *drug_exposure* jäeti välja sellised read, millel veeru *drug_concept_id* väärtus oli 0, alles jäi 15 654 848 rida. Kokku on andmestikus seetõttu 35 987 913 sündmust. Tabelis 1 on näide andmestikust.

Tabel 1. Näide sündmuste andmestikust.

Patsiendi ID	Kontseptsioon	Algne kood	Dokumendi ID	Kuupäev
1	435524	G47	1	2013-04-28
1	435524	G47	2	2013-04-28
1	19044885	1068083	2	2013-04-28
1	435524	G47	3	2013-05-09
1	19044885	1068083	3	2013-05-09
1	435524	G47	4	2013-07-02
1	435524	G47	5	2013-07-02
1	19044885	1068083	5	2013-07-02
2	257011	J06.9	6	2015-07-27
2	257011	J06.9	7	2015-07-27
2	4320791	J30	8	2017-09-15
2	4320791	J30	9	2017-09-25
2	4320791	J30	10	2017-09-25

Tabelis 1 on kahe patsiendi (1 ja 2) andmed. Selle veerus „Algne kood“ on kolm erinevat ICD10 koodi: G47 „Unehäired“, J06.9 „Täpsustamata ägedad ülemiste hingamisteede nakkused“ ning J30 „Vasomotoorne ja allergiline riniit“. Selles veerus leiduv kood 1068083 ei pärine ICD10-st.

3.2 Metoodika

See alapeatükk selgitab andmetega läbi viidud juhendamata masinõppe etappe. Samuti kirjeldab see parameetreid, millega masinõppe algoritme katsetati.

3.2.1 Word2vec treenimine

Word2vec mudelile anti laused, kus sõnadeks on sama patsiendi sündmuste kontseptsioonid ajalises järjestuses. Näiteks tabelist 1 pärit andmed annavad kaks lauset:

1. 435524 435524 19044885 435524 19044885 435524 435524 19044885
2. 257011 257011 4320791 4320791 4320791

Word2vec treeniti CBOW meetodil 10 epohhiga ning vektori pikkus oli 100. Epohhide arv valiti piisavalt väike, et oleks võimalik mõistliku aja jooksul (alla 10 minuti) mudelit treenida. Töö käigus oli vajalik word2vec'i korduvalt treenida, et katsetada erinevate sisendandmete ning parameetritega. Suur epohhide arv teeks katsetamise ebapraktiliselt aeganõudvaks. 100-mõõtmelised sõnavektorid ning CBOW arhitektuur on selles töös

kasutatud word2vec implementatsiooni vaikeväärtused ning nende eri väärtuste mõju käesolevas töös ei uuritud. Treenitud mudeli pealt kontrolliti, kas 2. tüüpi diabeedi ja migreeni kontseptsiooniga kõige sarnasemad kontseptsioonid on loogilised, kasutades n-mõõtmeliste sõnavektorite A ja B vahelist koosinus-kaugust, mille valem on

$$S_C(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3.2.2 Sündmusteta ajavahemiku tokeni kasutamine

Word2vec'i kasutamisel tekib küsimus, et kuidas valida ühte lauset. Kui andmetes esineb 5 aastane paus, ei tohiks viimane sündmus enne pausi ega esimene sündmus peale pausi omavahel põhjuslikult seotud olla. Nende järjestikune esinemine samas lauses võib word2vec'i jaoks vale tähenduse tekitada, kuna neid ei ole võimalik eristada sellisest sündmuste paarist, mille vahele jäi mõni päev. Üks lahendus oleks asetada aastatepikkuse pausi kohale lausete piirid, kuid siis tekib küsimus, et kui pikk peab paus olema, et sellest saaks lause lõpp.

Töös valitud lahenduseks on käsitleda kõiki ühe patsiendi sündmusi ühe lausena, kuid sündmuste vahele paigutatakse spetsiaalsed tokenid, mis väljendavad pausi pikkust. Pausi pikkuse väljendamine päevades võib sõnade arvu liiga suureks ajada. Samuti ei ole oluline, kas kaks sündmust toimusid näiteks 5 või 6 aastase vahega. Sõnade arvu minimaalsena hoidmiseks kasutati süsteemi, milles token $\#N$ tähendab, et kahe sündmuse vaheline päevade arv jääb fibonnaci arvude F_{N-1} ja F_N vahele. Vähem kui 14 päeva pikkused pausid jäeti ilma tokenita. Kui tabelis 1 toodud sündmustest moodustada word2vec laused koos aja tokenitega, oleks laused järgmised:

1. 435524 435524 19044885 #6 435524 19044885 #9 435524 435524 19044885
2. 257011 257011 #15 4320791 #6 4320791 4320791

Word2vec'i katsetati nii aja tokenitega kui ka ilma, et näha kuidas nende kasutamine tulemust mõjutab.

3.2.3 Erineva akna pikkusega katsetamine

Word2vec'i algoritmil on parameeter, mida nimetatakse akna suuruseks. See on suurim kahe sõna vaheline kaugus sõnade arvu mõttes, mille puhul loetakse neid kõrvuti esinevaks. Selles töös katsetati järgmiste akna pikkustega: 10, 20, 30, 40, 50, 60, 70, 80, 90 ja 100.

3.2.4 K-means klasterdamine

Selleks, et mitut algset sündmust oleks võimalik üldistada üheks sündmuseks, oleks neid vaja grupeerida. Word2vec'i abil luuakse nii kontseptsioonidele kui ka aja tokenitele vastavad 100-mõõtmelised sõnavektorid. Sündmuste kontseptsioonidele vastavatele vektoritele on võimalik rakendada mõnda klasterdusalgoritmi. Selles töös rakendatakse K-means klasterdamist. Klasterdamise tulemusel määratakse iga kontseptsioon ja seeläbi kõik selle kontseptsiooniga sündmused mingisse klastrisse. Selles töös katsetati järgmiste klastrite arvudega: 100, 200, 300, 400, 500, 600, 700, 800, 900 ja 1000. Tabelis 2 on näide klasterdamise tulemusest.

Tabel 2. Näide K-means klasterdamise tulemusest. Veeru „kontseptsioon" väärtused on võetud tabeli 1 samanimelisest veerust.

Kontseptsioon	Klastri ID
435524	45
19044885	26
257011	187
4320791	187

3.2.5 Klastritele nime andmine

Algses andmestikus saab sündmuseid kirjeldada kontseptsioonide järgi. Kui aga mitu sündmust üldistatakse klasterdamise teel üheks sündmuseks, oleks ka sellele vaja kirjeldust. Kuna üldistatud sündmus võib koosneda suurest hulgast algse andmestiku sündmusest, võib iga üksiku kontseptsiooni kirjelduse kokku ühendamine anda liiga pika nimetuse. Selles töös valiti lähenemine, mille kohaselt antakse igale klastrile nimetus. Peale K-means klasterdamist tuleb läbi vaadata sündmuste andmestik ning iga klastri puhul lugeda kokku, mis kontseptsioonid sellesse klastrisse kuuluvad ning mitu sündmust nende kontseptsioonidega leidub. Üks võimalus on valida klastri sündmuste seas kõige sagedamini esinev kontseptsioon ning selle järgi nimi anda, kuid siis on oht, et üks kontseptsioon ei pruugi piisavalt klastri sisu selgitada. Liiga paljude kontseptsioonide kasutamine klastri nimetuses võib anda liiga pika nimetuse. Otsustati, et klastrid saavad nime minimaalselt kahe ja maksimaalselt viie sagedaseima kontseptsiooni järgi. Nime andvate kontseptsioonide arv sõltub sellest, kui mitu kontseptsiooni kokku moodustavad vähemalt 20% kõikidest vastava klastriga sündmustest.

Näide: leiame et kõik sündmused, mille kontseptsioon on klastris 187. Nende seas viis kõige sagedasemat kontseptsiooni on:

1. 15.69% sündmustest on kontseptsiooniga 257011 (ingl *Acute upper respiratory infection*).
2. 10.6% sündmustest on kontseptsiooniga 40484028 (ingl *Dental caries extending into dentin*).
3. 6.39% sündmustest on kontseptsiooniga 25297 (ingl *Acute pharyngitis*).
4. 4.4% sündmustest on kontseptsiooniga 141095 (akne).
5. 4.2% sündmustest on kontseptsiooniga 257007 (ingl *Allergic rhinitis*).

Kuna esimesed kaks kontseptsiooni moodustavad vähemalt 20% kõikidest sündmustest, saab klastri nimeks „Acute upper respiratory infection: 15.69%; Dental caries extending into dentin: 10.6%“.

3.2.6 Üldistatud sündmuste leidmine

Üldistatud sündmuseks loetakse järjestikuste sündmuste jada, milles

- kõik sündmused on sama patsiendiga
- kõik sündmused on samast klastrist kontseptsiooniga
- iga kahe järjestikuse sündmuse vahele jääv aeg on vähem kui 30 päeva

Vähem kui 30 päeva kahe järjestikuse sündmuse vahel peab olema seetõttu, et mitte lugeda kõiki sama klastriga sündmusi patsiendi andmestikus sama üldistatud sündmuse alla. Kui agregeerida mitu üksteisele ajaliselt lähedal toimunud sündmust sama üldistuse alla, on andmetes vähem müra. Samas võiksid üldistatud sündmused kajastada ka mitu kuud või aastat kestnud pause sündmuste vahel. Seetõttu otsustati lugeda maksimaalseks sündmustevaheliseks ajaks sama üldistuse siseselt 1 kuu ehk ligikaudu 30 päeva. Kui võtta aluseks tabelis 1 toodud näiteandmed ning tabelis 2 toodud klasterdamise tulemuse, saadakse tulemuseks üldistatud sündmused, mis on toodud tabelis 3.

Tabel 3. Näide üldistatud sündmustest

Patsiendi ID	Klastri ID	Klastri Nimetus	Algus	Lõpp
1	26	zopiclone 7.5 MG Oral Tablet: 40.96%; amoxicillin 875 MG / clavulanate 125 MG Oral Tablet: 22.72%	2013-04-28	2013-05-09
1	45	Sleep disorder: 11.96%; Anxiety disorder: 5.82%; Depressive disorder: 5.66%	2013-04-28	2013-05-09
1	45	Sleep disorder: 11.96%; Anxiety disorder: 5.82%; Depressive disorder: 5.66%	2013-07-02	2013-07-02
1	26	zopiclone 7.5 MG Oral Tablet: 40.96%; amoxicillin 875 MG / clavulanate 125 MG Oral Tablet: 22.72%	2013-07-02	2013-07-02
2	187	Acute upper respiratory infection: 15.69%; Dental caries extending into dentin: 10.6%	2015-07-27	2015-07-27
2	187	Acute upper respiratory infection: 15.69%; Dental caries extending into dentin: 10.6%	2017-09-15	2017-09-25

3.2.7 Tulemuste hindamine

Et valida optimaalseid parameetreid word2vec ja k-means klasterdamise jaoks, on vaja kvaliteedimõõdikuid. Katsetati kolme erineva parameetri erinevaid väärtusi:

- **aja tokenite olemasolu** word2vec lausetes
- word2vec **akna suurus**
- K-means **klastrite arv**

Arvutati järgmised kvaliteedimõõdikud:

- **Klastrite arv dokumendi kohta**. Iga unikaalse tervisedokumendi kohta, millest andmed pärinesid, loeti kokku, mitmesse eri klastrisse selle dokumendiga sündmused satuvad ning leiti keskmine üle kõigi dokumentide. Mõõdik näitab, kui paljudes eri klastrites sama dokumendi sündmused keskmiselt on. Arvestati vaid selliseid sündmusi, mille puhul patsient oli sündmuse kuupäeval noorem kui 40 aastane. Põhjus on selles, et vanemad inimesed käivad arsti juures sagedamini mitme haigusega korraga ning sellest tulenevalt võib ka arstivisiidist pärinev dokument käsitleda väga erinevaid haigusi.
- **Klastrite arv ICD10 koodi kohta**. Mõõdiku arvutamine toimub sarnaselt eelmisele, kuid dokumentide asemel on ICD10 koodid. Patsiendi vanust sündmuse hetkel ei arvestatud, kuid kasutati vaid selliseid sündmusi, millel tabeli 1 veerus „Kood“ on ICD10 tüüpi kood. ICD10 koode loeti samaks, kui need erinevad vaid peale punkti esinevate numbrite poolest. Näiteks koodid I42.0 „Dilateeruv kardio(müo)paatia“ ja

I42.1 „Ummistav e obstruktiivne hüpertroofiline kardio(müo)paatia“ loeti üheks ja samaks koodiks. Sarnaselt eelmisele mõõdikule mõõdab ka see mõõdik üksteisega sarnaste klastrite hulka, kuid teeb seda ICD10 koodide tasemel.

- **ICD10 koodide arv klastri kohta**. Seda arvutatakse eelmisele mõõdikule vastupidiselt. Mõõdik näitab, kui ühtlane on keskmiselt ühe klastri sisu. Mida suurem see on, seda rohkem erinevad samasse klastrisse kuuluvad sündmused üksteisest ning seda keerulisem on klastri sisu üldistada.
- **Klastrite arv ICD10 peatüki kohta**. Mõõdiku arvutamine toimub samamoodi nagu klastrite arvu leidmine ICD10 koodi kohta, kuid üksikute koodide asemel loetakse ICD10 peatükke. Näiteks koodid I30 „Äge perikardiit“ ja I42 „Kardio(müo)paatia“ kuuluvad mõlemad peatükki I00-I99 „Vereringeelundite haigused“. See mõõdik mõõdab samuti üksteisega sarnaste klastrite hulka, kuid ICD10 peatükkide abil võib see olla ülevaatlikum kui üksikute koodide abil.
- **ICD10 peatükkide arv klastri kohta** Seda arvutatakse eelmisele mõõdikule vastupidiselt. Ka see mõõdik näitab, kui ühtlane on klastrite sisu, kuid peatükkide abil on klastrit kergem üldistada kui üksikute koodide abil.
- **3 kõige sagedasema kontseptsiooni osakaal klastrist**. Loetakse kokku iga klastri ja kontseptsiooni kombinatsiooni esinemise sagedus sündmustes. Iga klastri kohta leitakse kolm kõige sagedasemat kontseptsiooni ning leitakse nende osakaal kõikidest sama klastriga sündmustest. See mõõdik näitab samuti klastri sisu puhtust, kuid mõõdab seda teisel viisil. Klastrisse võib küll kuuluda palju eri kontseptsioone, kuid klastrisse kuuluvate sündmuste seas võib mõni kontseptsioon esineda sagedamini kui teised. Sellest tulenevalt on klastrile nime leidmisel võimalik ignoreerida kontseptsioone, mis ülejäänud sama klastri kontseptsioonidega kokku ei sobi, kuid mis moodustavad klastri sündmuste hulgas väga väikese osa.
- **Kõige sagedasema ICD10 peatüki osakaal klastrist**. Arvutatakse sarnaselt eelmisele, kuid kontseptsiooni asemel on ICD10 peatükk. Arvestatakse vaid selliseid sündmusi, millel on ICD10 kood olemas. ICD10 peatükkide abil on klastri sisu kergemini üldistatav kui üksikute kontseptsioonidega.
- **Klastrite osakaal kontseptsioonidest**. Iga patsiendi kohta loetakse kokku, mitu erinevat kontseptsiooni tema sündmuste seas on ning mitmesse erinevasse klastrisse need sündmused langevad. Arvutatakse klastrite arvu suhe kontseptsioonide arvu ning võetakse keskmine üle kõigi patsientide. Mõõdiku eesmärk on hinnata, kui palju lihtsatamaks patsiendi andmed muutuvad peale klasterdamist. Mida väiksem see on, seda parem.
- **Üldistatud sündmuste osakaal algsetest sündmustest**. Iga patsiendi kohta loetakse kokku, mitu sündmust on tema andmestikus algselt ning mitu üldistatud sündmust tekib peale klasterdamist ja üldistamist. Arvutatakse üldistatud sündmuste arvu suhe algsete sündmuste arvu. Näiteks kui patsiendil oli algselt andmetes 300 sündmust ja

pärast üldistamist oli tema andmetes 150 sündmust, siis on mõõdiku väärtus selle patsiendi puhul on 150/300=0.5. Mõõdiku lõpliku väärtuse arvutamiseks võetakse keskmine üle kõigi patsientid. Ka selle mõõdiku eesmärk on hinnata, kui palju lihtsatamaks patsiendi andmed muutuvad peale klasterdamist. Mida väiksem see on, seda parem.

- **Klastri nime täpsus**. Klastrite nimetused koosnevad 2-5 kontseptsiooni nimetusest. Leitakse, kui suur osa üldistatud sündmustest sisaldab vähemalt ühte algset sündmust, mille kontseptsioon sisaldub vastava klastri nimetuses. Mõõdiku eesmärk on hinnata, kui hästi klastrite nimetused üldistatud sündmusi iseloomustavad. Mida suurem see on, seda parem.

Lisaks kvaliteedimõõdikutele analüüsitakse ICD10 koodidega sündmuste jagunemist klastrite vahel peatükkide kaupa. Erinevatesse ICD10 peatükkidesse (mida on kokku 21) kuuluvad koodid on üksteisest väga erinevad ning see annab ühe võimaluse klasterduse tulemuste analüüsimiseks. Kõigist 35 987 913-st sündmusest omavad ICD10 koodi 20 108 782. Seega ei anna nende klasterduse tulemus ülevaadet teistest sündmustest, kuid võimaldab siiski hinnata klasterduse kvaliteeti.

3.3 Eetikakomitee luba

Käesolev töö viidi läbi Eesti bioeetika ja inimuuringute nõukogu loa nr. 1.1-12/3833 alusel.

4 Tulemused

Selles peatükis kirjeldatakse eri andmeanalüüsi etappide tulemusi. Alapeatükk 4.1 kirjeldab word2vec'ist saadud vektorite sarnasust kahe kontseptsiooni näitel. Alapeatükk 4.2 annab ülevaate klasterduse kvaliteedimõõdikutest erinevatel word2vec ja K-means parameetrite väärtustel. Alapeatükis 4.3 vaadeldakse ICD10 koodidega sündmuste jaotumist klastrite vahel. Alapeatükis 4.4 on toodud ühe patsiendi sündmuste klasterdamise tulemus ning ülevaade tema üldistatud sündmustest.

4.1 Sarnased kontseptsioonid

Tabel 4 koosneb 2. tüüpi diabeedile (kontseptsioon 201826) vastavale word2vec vektorile 10-st kõige sarnasemast vektorist, kui word2vec treenida ilma aja tokeniteta. Tabelis 5 on need 10 vektorit sellisel juhul, kui aja tokeneid word2vec treenimisel kasutada.

Tabel 4. 10 kõige sarnasemat word2vec vektorit 2. tüüpi diabeedi vektorile. Word2vec akna suurus on 20, aja tokeneid ei kasutatud.

Kontseptsioon	Nimetus	Koosinus sarnasus
201826	Type 2 diabetes mellitus	1
442604	Hypertensive heart disease	0.61
443735	Coma due to diabetes mellitus	0.51
40480694	High hemoglobin A1c level	0.51
443732	Disorder due to type 2 diabetes mellitus	0.45
201254	Type 1 diabetes mellitus	0.43
316139	Heart failure	0.43
1503297	metformin	0.41
43185378	Insulin Lispro 100 UNT/ML Injectable Suspension	0.38
443731	Renal disorder due to type 2 diabetes mellitus	0.36
44032735	exenatide 2 MG Injectable Suspension	0.36

Tabel 5. 10 kõige sarnasemat word2vec vektorit 2. tüüpi diabeedi vektorile. Word2vec akna suurus on 20, aja tokeneid kasutati.

Kontseptsioon	Nimetus	Koosinus sarnasus
201826	Type 2 diabetes mellitus	1
442604	Hypertensive heart disease	0.62
443735	Coma due to diabetes mellitus	0.54
40480694	High hemoglobin A1c level	0.49
1525220	pioglitazone 45 MG Oral Tablet	0.45
443732	Disorder due to type 2 diabetes mellitus	0.45
201254	Type 1 diabetes mellitus	0.44
316139	Heart failure	0.43
44032735	exenatide 2 MG Injectable Suspension	0.4
1503297	metformin	0.4
321822	Peripheral vascular disorder due to diabetes mellitus	0.38

Tabel 6 koosneb migreenile (kontseptsioon 318736) vastavale word2vec vektorile 10-st kõige sarnasemast vektorist, kui word2vec treenida ilma aja tokeniteta. Tabelis 7 on need 10 vektorit sellisel juhul, kui aja tokeneid word2vec treenimisel kasutada.

Tabel 6. 10 kõige sarnasemat word2vec vektorit migreeni vektorile. Word2vec akna suurus on 20, aja tokeneid ei kasutatud.

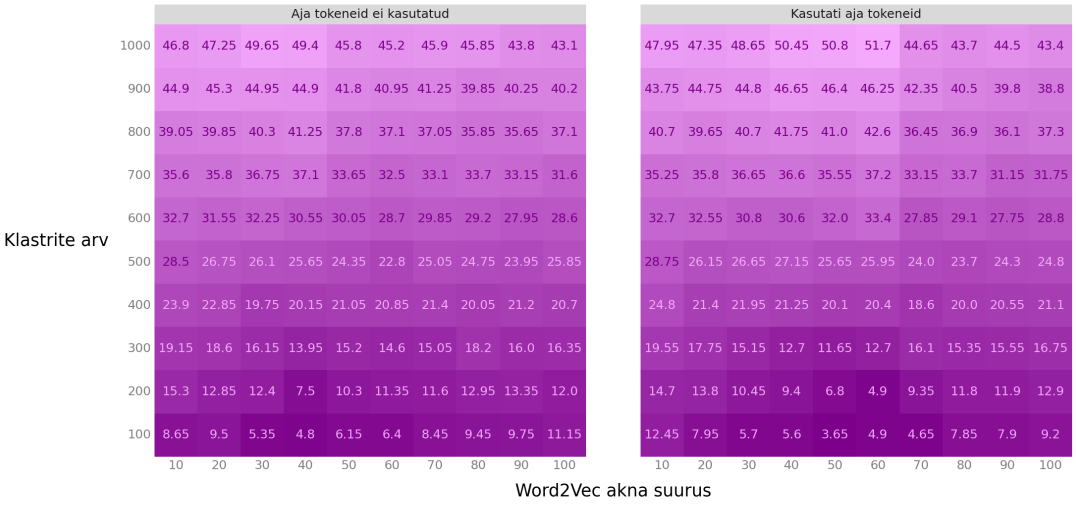
Kontseptsioon	Nimetus	Koosinus sarnasus
318736	Migraine	1
378735	Migraine without aura	0.94
381549	Migraine with aura	0.88
1189459	frovatriptan 2.5 MG Oral Tablet	0.77
19071336	zolmitriptan 2.5 MG Disintegrating Oral Tablet	0.74
19079711	sumatriptan 100 MG Oral Tablet	0.72
1140650	sumatriptan 50 MG Oral Tablet	0.71
1116031	zolmitriptan	0.7
40724912	rizatriptan 10 MG Oral Strip	0.63
1189458	frovatriptan	0.62
1140643	sumatriptan	0.6

Tabel 7. 10 kõige sarnasemat word2vec vektorit migreeni vektorile. Word2vec akna suurus on 20, aja tokeneid kasutati.

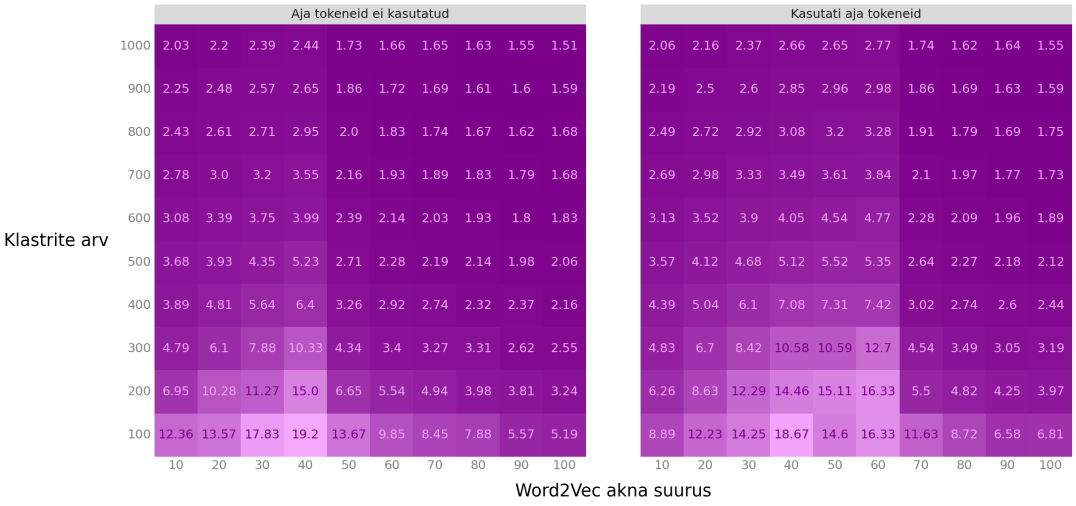
Kontseptsioon	Nimetus	Koosinus sarnasus
318736	Migraine	1
378735	Migraine without aura	0.92
381549	Migraine with aura	0.86
19071336	zolmitriptan 2.5 MG Disintegrating Oral Tablet	0.7
1189459	frovatriptan 2.5 MG Oral Tablet	0.67
1140650	sumatriptan 50 MG Oral Tablet	0.64
1116031	zolmitriptan	0.61
19079711	sumatriptan 100 MG Oral Tablet	0.6
40724912	rizatriptan 10 MG Oral Strip	0.57
1189458	frovatriptan	0.57
4105620	Complicated migraine	0.55

4.2 Klasterduse kvaliteedimõõdikud

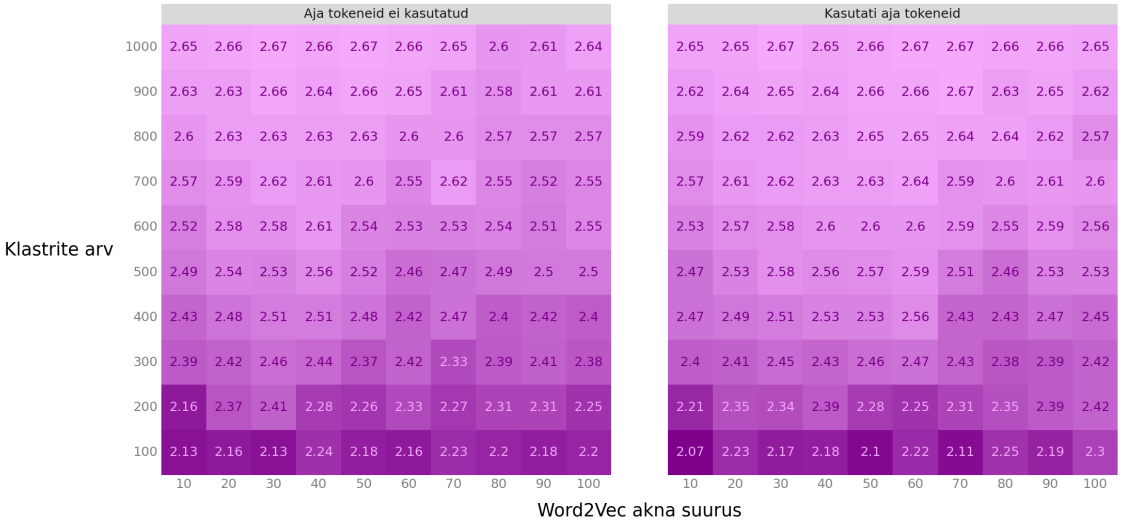
Joonised 2, 3 ja 4 näitavad kolme kvaliteedimõõdiku väärtuseid erinevatel klastrite arvu ning word2vec akna suuruse väärtustel.



Joonis 2. Klastrite arv ICD10 peatüki kohta.

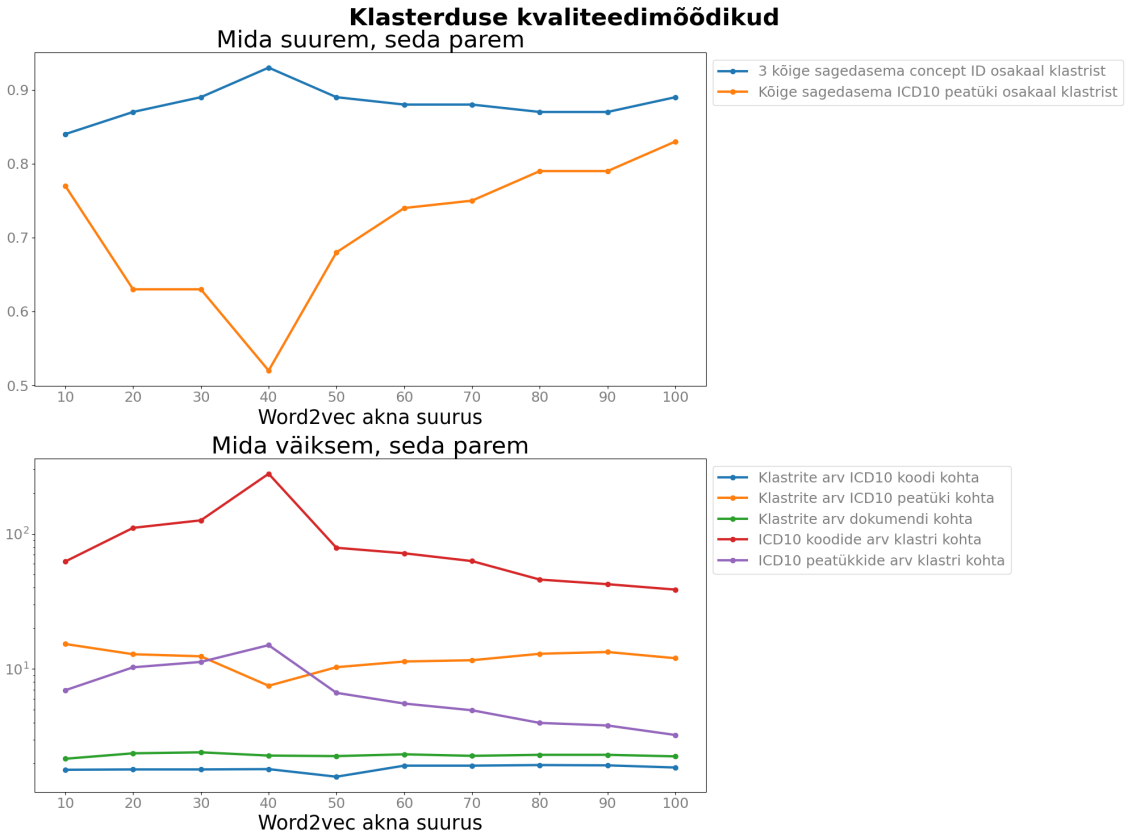


Joonis 3. ICD10 peatükkide arv klastri kohta.



Joonis 4. Klastrite arv dokumendi kohta alla 40 aastastel patsientidel.

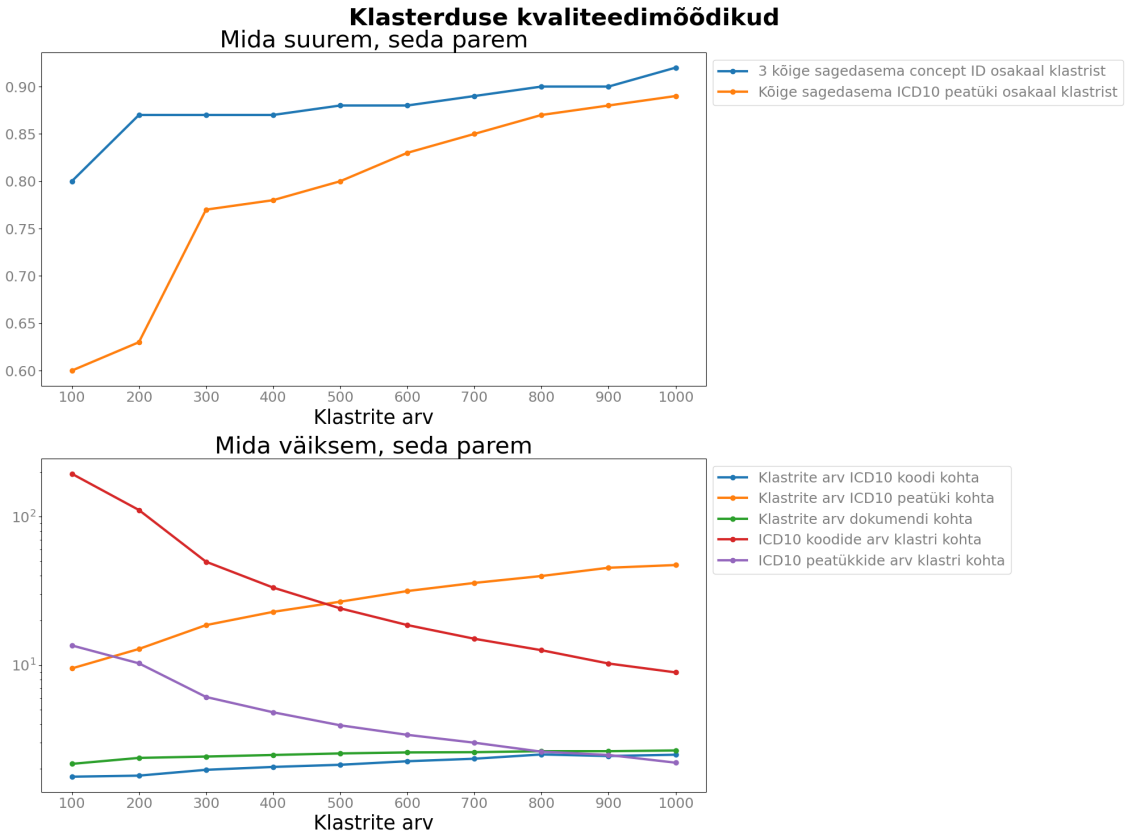
Joonistel 2 ja 3 on näha 200 klastri real mõõdikute suuremat kõikumist erinevatel akna suurustel kui suuremate klastrite arvudega ning aja tokenid ei paista suurt mõju omavat. Seetõttu otsustati korraga mitme mõõdiku ühele graafikule kandmiseks koos word2vec akna suurusega fikseerida klastrite arv 200 peale ning aja tokeneid mitte kasutada. Tulemus on joonisel 5.



Joonis 5. Klasterduse kvaliteedimõõdikute sõltuvus word2vec akna suurusest 200 klastriga. Alumise joonise y-teljel on logaritmiline skaala.

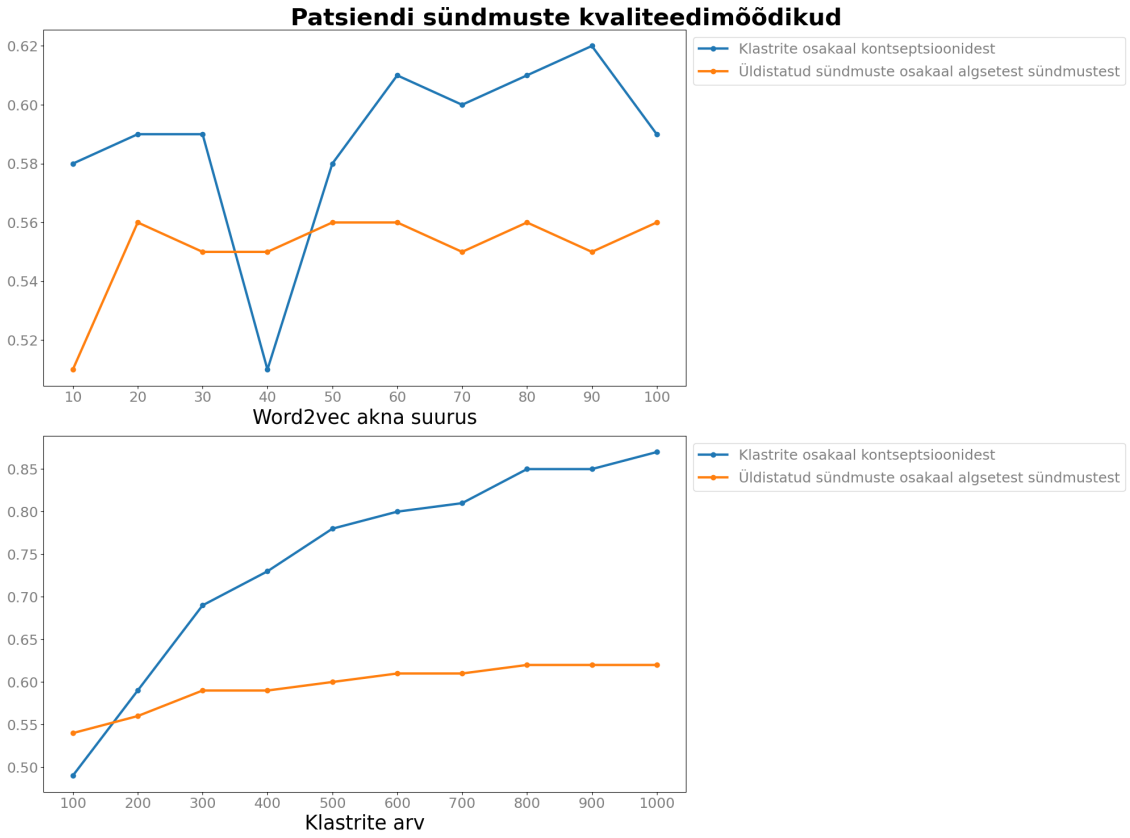
Joonistelt 2, 3 ja 5 on näha, et mõne mõõdiku väärtuses esineb maksimum/miinimum, kui fikseeritud klastrite arvu juures on word2vec akna suurus 40. Mõõdikutel „ICD10 koodide arv klastri kohta“ ja „ICD10 peatükkide arv klastri kohta“ esineb selles punktis maksimum ning mõõdikutel „klastrite arv ICD10 peatükkide kohta“ ja „kõige sagedasema ICD10 peatüki osakaal klastrist“ esineb selles punktis miinimum.

Joonisel 6 on samade mõõdikute sõltuvus klastrite arvust, kui word2vec akna suurus fikseerida 20 peale ning aja tokeneid mitte kasutada.



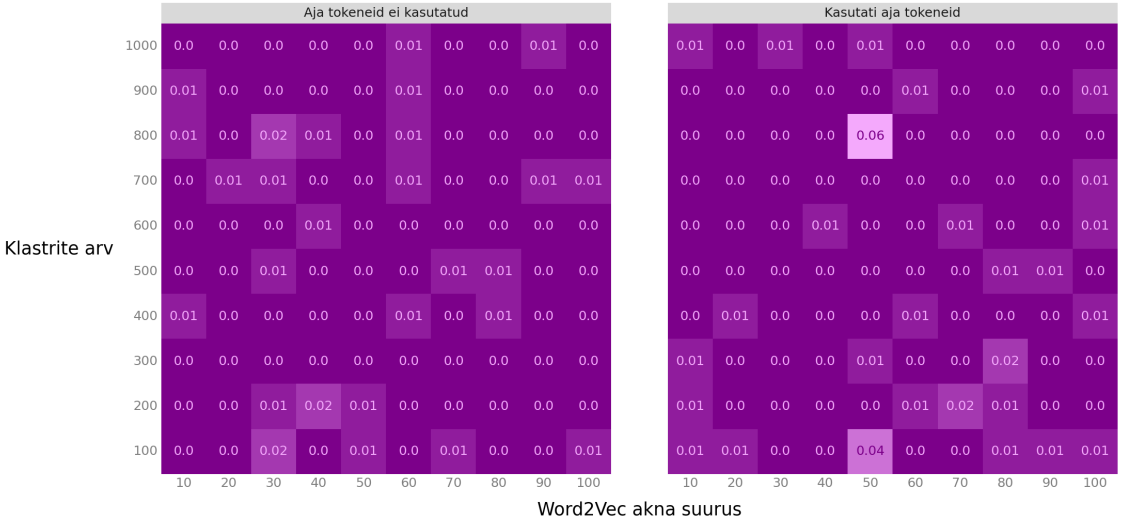
Joonis 6. Klasterduse kvaliteedimõõdikute sõltuvus klastrite arvust word2vec akna suurusega 20. Alumise joonise y-teljel on logaritmiline skaala.

Joonisel 7 on kahe kvaliteedimõõdiku, „klastrite osakaal kontseptsioonidest“ ning „üldistatud sündmuste osakaal algsetest sündmustest“ seos word2vec akna suuruse ning klastrite arvuga.



Joonis 7. Ülemisel joonisel on mõõdikute sõltuvus word2vec akna suurusest 200 klastriga. Alumisel joonisel on mõõdikute sõltuvus klastrite arvust word2vec akna suurusega 20. Alumise joonise y-teljel on logaritmiline skaala. Kummagil juhul ei kasutatud aja tokeneid word2vec treenimisel.

Mõõdiku „klastri nime täpsus“ väärtused erinevatel parameetrite väärtustel on joonisel 8.

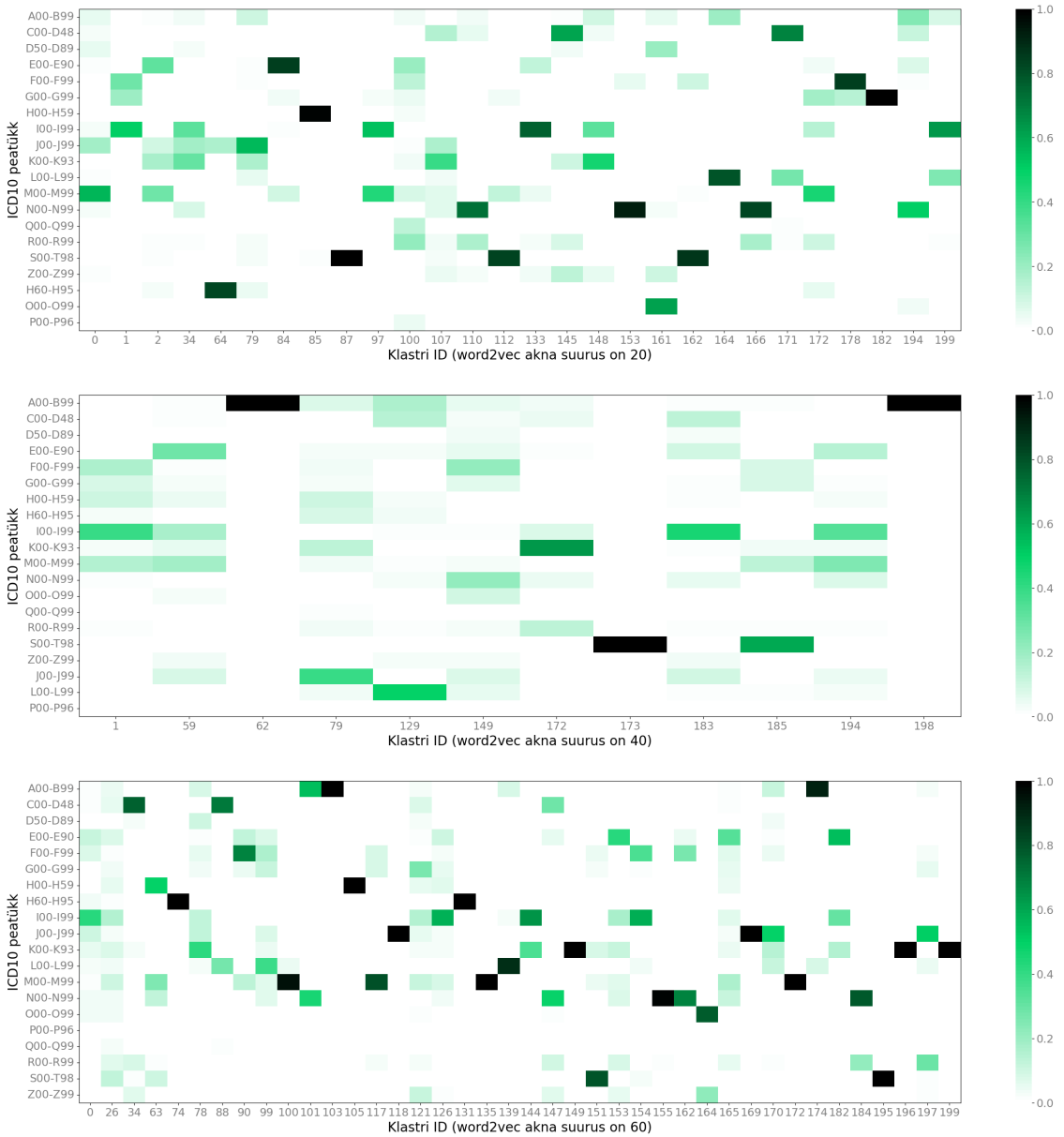


Joonis 8. Klastri nime täpsus.

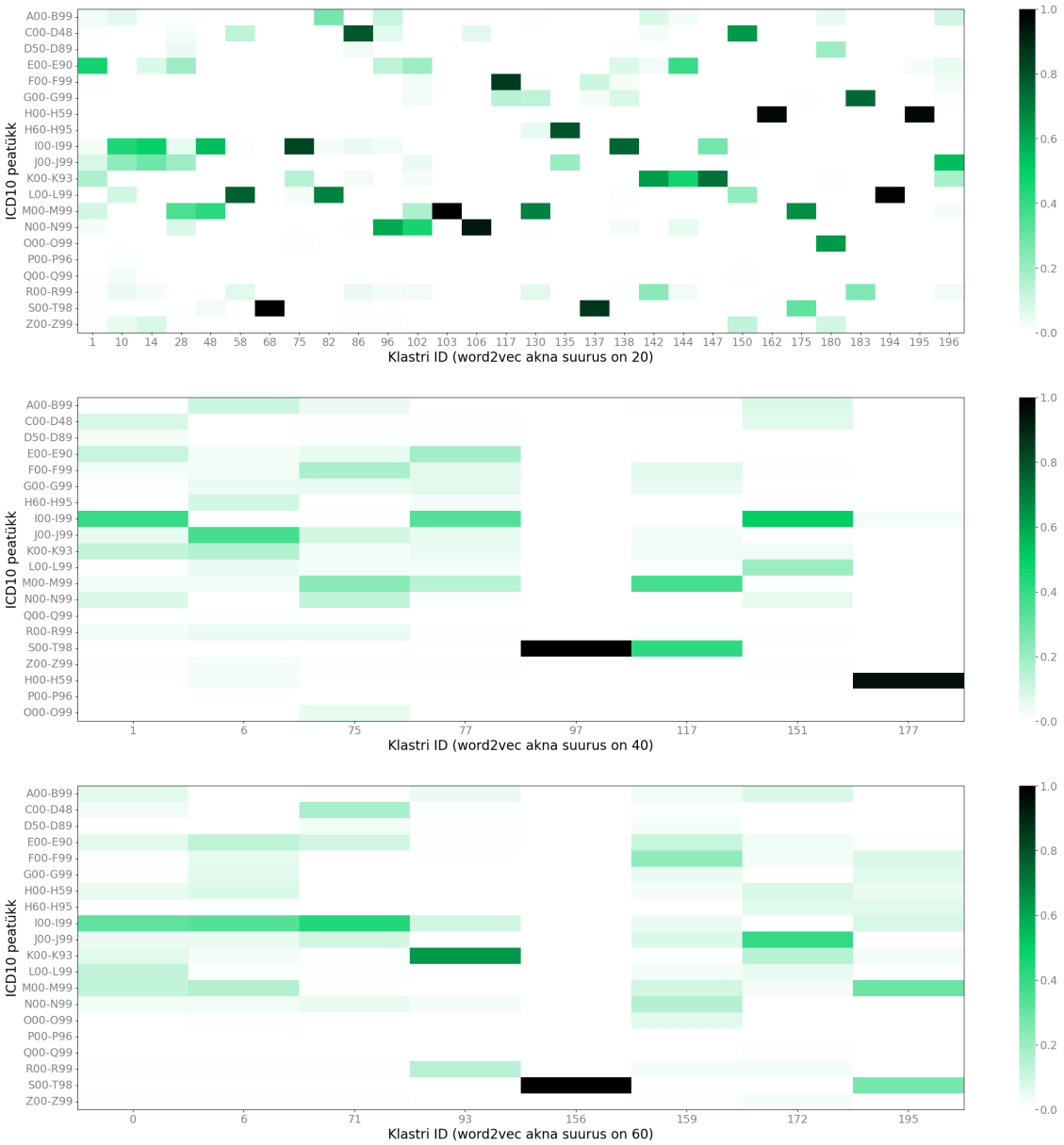
Jooniselt 8 on näha, et mõõdiku „klastri nime täpsus“ (joonisel 8) maksimum väärtus üle kõikide katsetatud parameetrite väärtuste oli 6% ning enamasti on väärtused nullilähedased. Sellel joonisel ei tähenda väärtus 0.0 siiski võrdumist nulliga, vaid tuleneb kahe komakohani ümardamisest.

4.3 Klastrite sisu ICD10 peatükkide järgi

Joonisel 9 on ICD10 koodiga sündmuste klasterduse tulemus 200 klastri ning kolme erineva word2vec akna suurusega selliselt, et aja tokeneid ei kasutatud. Joonisel 10 on sama tulemus siis, kui aja tokeneid kasutati.



Joonis 9. Sündmuste jaotumine klastritesse ICD10 peatükkide järgi erinevate word2vec akna suurustega. Klastrite koguarv on 200 ning word2vec treenimisel ei kasutatud aja tokeneid. Sellel joonisel illustreerib iga ruut selliste sündmuste arvu, mis on kuuluvad vastava veeru klastrisse ning mille kood kuulub vastava rea ICD10 peatükki. Iga veeru sündmuste arvud on normaliseeritud 0 ja 1 vahele selliselt, et veeru summa on 1.

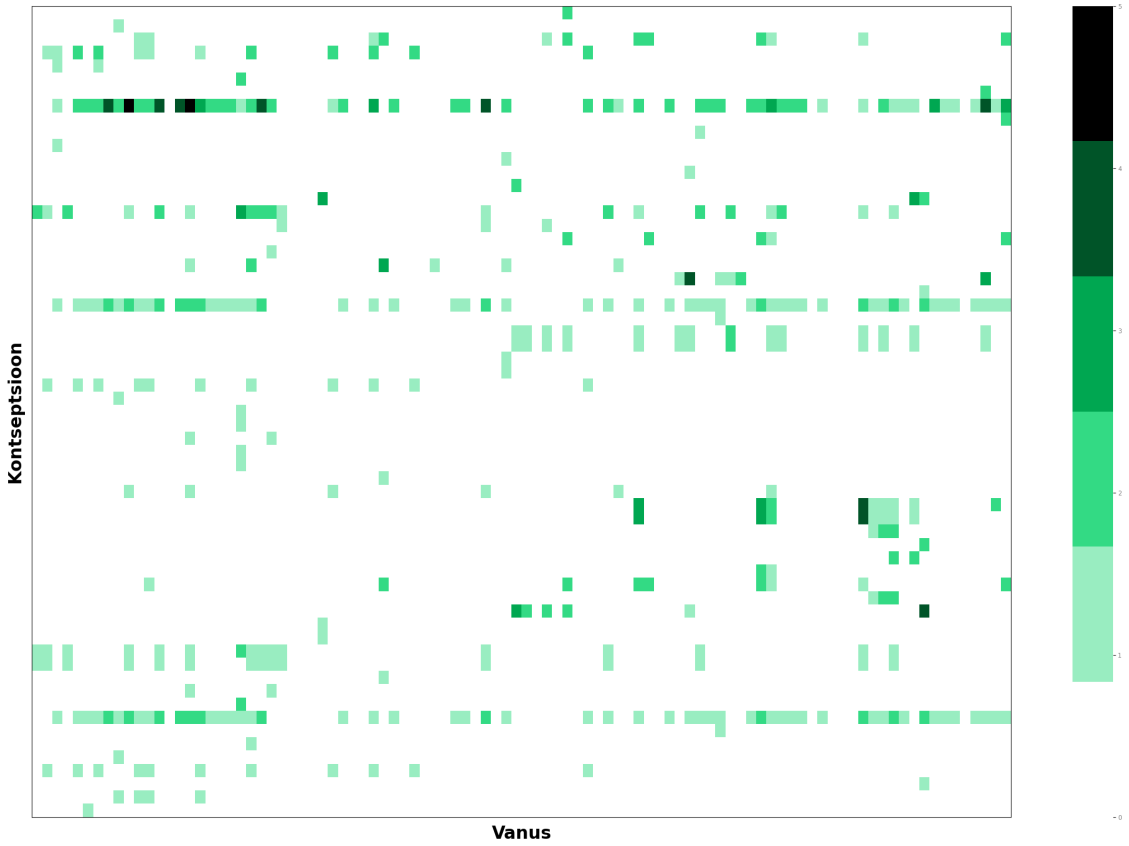


Joonis 10. Sündmuste jaotumine klastritesse ICD10 peatükkide järgi. Joonis koostati samal viisil nagu joonis 9 ainult et sel korral kasutati word2vec treenimisel aja tokeneid

Jooniselt 9 järgi jagunevad ICD10 koodiga sündmused 28 klastri vahel kui word2vec aken on suurusega 20 ja aja tokeneid mitte kasutada. Akna suurusel 40 on ICD10 koodiga sündmuseid sisaldavaid klastreid 12. Kui word2vec aken on suurusega 60, on selliseid klastreid 40. Kui aja tokeneid kasutada, siis joonise 10 järgi on word2vec akna suurustel 20, 40 ja 60 ICD10 koodiga sündmuseid sisaldavaid klastreid vastavalt 30, 8 ja 8.

4.4 Ühe patsiendi sündmuste üldistamine

Selles alapeatükis on toodud ühe patsiendi südmuste klasterdamise ning ülsistamise tulemused. See patsient valiti seetõttu, et tema andmestikus on mitmeid kontseptsioone, mis korduvad palju kordi kogu andmestiku ajaperioodi jooksul. Joonisel 11 on toodud selle patsiendi andmed. 580 sündmuse seas on 61 erinevat kontseptsiooni.



Joonis 11. Ühe patsiendi andmed. Iga veerg tähistab ühte kuud ning iga rida on üks kontseptsioon. Ruudu värv näitab, mitu vastava rea kontseptsiooniga sündmust sellele veerule vastaval kuul toimus.

Kõik kontseptsioonid klasterdati selliselt, et word2vec akna suurus oli 20, aja tokeneid word2vec treenimisel ei kasutatud ning K-means klastreid oli 200. Sellisel viisil jagunesid patsiendi sündmused 34 klastri vahel. 25 klastrit 34-st on sellised, millesse kuuluvad sündmused patsiendi andmestikus on kõik sama kontseptsiooniga. 25 klastrit 34-st on sellised, kus mõni klastri sisus oleva kontseptsiooni nimetus on ka osa klastri nimest. Klasterduse tulemus on näidatud tabelis 8.

Tabel 8. Joonisel 11 toodud patsiendi sündmuste klasterduse tulemus. Word2vec akna suurus on 20, aja tokeneid word2vec treenimisel ei kasutatud ning K-means klastreid on 200. Klastri nimetuse lõpus olev sulgudes number näitab K-means algoritmi poolt antud ID-d, nendele viitab joonis 12. Veerus „N“ on sündmuste arv selle kontseptsiooniga. Privaatsuse huvides on välja jäetud 18 klastit, milles on vähem kui 5 sündmust.

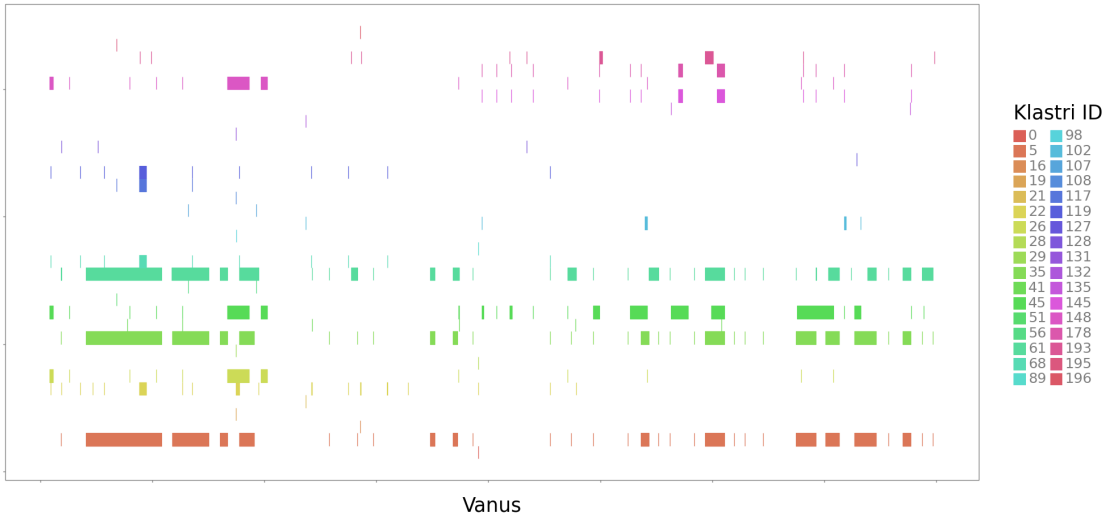
Klastri nimetus	Kontseptsiooni nimetus	Kood	N
Acute bronchitis: 16.08%; Viral disease: 12.41% (22)	Atopic dermatitis	133834	19
	Viral disease	440029	9
	Acute bronchitis	260139	2
	Disease due to Enterovirus	438068	1
	Upper respiratory tract infection due to Influenza	46273463	1
zopiclone 7.5 MG Oral Tablet: 40.96%; amoxicillin 875 MG / clavulanate 125 MG Oral Tablet: 22.72% (26)	zopiclone 7.5 MG Oral Tablet	19044885	15
	amoxicillin 875 MG / clavulanate 125 MG Oral Tablet	19115197	2
Sleep disorder: 11.96%; Anxiety disorder: 5.82%; Depressive disorder: 5.66% (45)	Sleep disorder	435524	32
	Moderate recurrent major depression	4077577	18
	Somatic syndrome present	4088489	16
	Mixed anxiety and depressive disorder	4338031	14
	Brief depressive adjustment reaction	440698	12
	Moderate major depression	4307111	5
	Somatic syndrome absent	4088609	5
	Panic disorder	436074	3
	Non-organic enduring personality change	4105190	2
Essential hypertension: 51.01%; Osteoarthritis of knee: 7.86% (61)	Essential hypertension	320128	120
	Rotator cuff syndrome	4164922	4
	Arthropathy	73553	1
	Disorder of soft tissue	376208	1
Open wound of finger: 5.56%; Alcohol dependence: 5.31%; Contusion of upper limb: 4.61%; Corneal foreign body: 4.46%; Contusion of finger: 3.76% (102)	Alcohol dependence	435243	8
	Alcohol abuse	433753	2
	Alcohol withdrawal syndrome	375519	1
	Acute alcoholic intoxication in alcoholism	433735	1

Jätkub järgmisel lehel

Tabel 8 (jätkub eelmiselt lehelt)

Klastri nimetus	Kontseptsiooni nimetus	Kood	N
Hyperplasia of prostate: 35.59%; Chronic prostatitis: 7.75% (135)	Sleep apnea	313459	2
	bupropion hydrochloride 150 MG - Extended Release Oral Tablet	40222065	1
	bupropion	750982	1
Nerve root disorder: 16.79%; Intervertebral disc disorder: 8.41% (193)	Intervertebral disc disorder	75344	18
	Nerve root disorder	4216397	15
	Chronic pain	436096	9
	Polyneuropathy	4174262	3
	Transient cerebral ischemia	373503	2
Diazepam 10 MG/ML Oral Solution: 48.22%; diazepam 5 MG Oral Tablet: 26.09% (5)	Diazepam 10 MG/ML Oral Solution	35407482	66
diazepam: 86.44%; prednisolone: 13.56% (35)	diazepam	723013	66
amoxicillin: 52.13%; zopiclone: 47.87% (148)	zopiclone	19044883	17
clonazepam 2 MG Oral Tablet: 100.0% (145)	clonazepam 2 MG Oral Tablet	798877	15
clonazepam: 100.0% (178)	clonazepam	798874	15
mometasone: 65.86%; amitriptyline: 22.68% (68)	mometasone	905233	11
Mometasone 0.05 MG Nasal Spray: 44.76%; Mometasone 0.001 MG/MG Topical Cream: 41.91% (119)	Mometasone 0.001 MG/MG Topical Solution	43641669	11
telmisartan: 14.61%; ramipril: 12.97% (41)	Serum cholesterol raised	4041554	6
Allergic contact dermatitis: 12.84%; Inflammatory dermatosis: 8.09% (117)	Inflammatory dermatosis	45766714	4

Tabelis 8 kirjeldatud klasterduse põhjal loodi üldistatud sündmused, mis on näidatud joonisel 12.



Joonis 12. Tabelis 8 kirjeldatud klasterduse põhjal loodud üldistatud sündmused.

Algsest 580-st sündmusest jäi alles 249 üldistatud sündmust, millest omakorda 140 on sellised, millele vastab üks algse andmestiku sündmus.

5 Arutelu

See peatükk arutleb peatükis 4 toodud tulemuste üle. Alapeatükid 5.1, 5.2, 5.3 ja 5.4 arutlevad vastavalt alapeatükkides 4.1, 4.2, 4.3 ja 4.4 toodud tulemuse tähenduse üle. Alapeatükk 5.5 toob välja selle töö tugevused ja nõrkused ning alapeatükk 5.6 pakub välja edasiseid uurimissuundi.

5.1 Sarnased kontseptsioonid

2. tüüpi diabeedile 10 kõige sarnasema sõnavektoriga kontseptsiooni on toodud tabelites 4 (saadud ilma aja tokeniteta) ja 5 (saadud koos aja tokenitega). Mõlemas tabelis leidub mitmeid ühiseid jooni. 10 kõige sarnasema sõnavektori seas on 4 kontseptsiooni, mille nimetuses esineb sõna „diabeet“. Esimesed 3 lähimat kontseptsiooni on mõlemas tabelis samad, nende seas ka kõrge A1C tase hemoglobiinis (kontseptioon 40480694), mis on 2. tüüpi diabeedi indikaator [Meda]. Tabelites on ka 2. tüüpi diabeedi ravimid metformiin [Medb] (kontseptsioon 1503297) ning eksenatiid [Medc] (kontseptsioon 44032735). Leidub veel üks 2. tüüpi diabeedi ravim, mis on kummagis tabelis erinev. Tabelis 4 on selleks insuliin (kontseptsioon 43185378) ning tabelis 5 on selleks pioglitasoon [Medc] (kontseptsioon 1525220). Kokkuvõttes on mõlemas tabelis adekvaatsed kontseptsioonid, kuigi sama kontseptsioon võib asuda erinevatel positsioonidel ning mõned kontseptsioonid on erinevad. Erinevus nende kahe tabeli vahel võib olla põhjustatud aja tokenite olemasolust word2vec treenimisandmetes, kuid samahästi võib see ka olla vektorite juhusliku initsialiseerimise tagajärg.

Sarnane olukord on tabelites 6 ja 7. Tabelites on 10 lähima kontseptsiooni seas vastavalt 2 ja 3 kontseptsiooni, mille nimetuses on sõna „migreen“. Ülejäänud kontseptsioonid on eri vormid migreeni ravimitest frovatriptaan [Medd] (kontseptsioonid 1189459 ja 189458), sumatriptaan [Mede] (kontseptsioonid 19079711, 1140650 ja 1140643), zolmitriptaan [Medf] (kontseptsioonid 19071336 ja 1116031) ja risatriptaan [Medg] (kontseptsioon 40724912).

5.2 Klasterduse kvaliteedimõõdikud

Joonistelt 2 ja 3 on näha, et sündmuste vahelise aja tokenite olemasolu word2vec treeningandmetes ei oma märgatavat mõju mõõdikute väärtustele. Mõlema joonise vasakul ja paremal paneelil samades kohtades olevad väärtused on üksteisele väga lähedal. Joonise 4 järgi varieerub klastrite arv dokumendi kohta väga väikestes piirides (2.07 - 2.67). Jooniselt 6 ilmneb, et mõõdikute „ICD10 koodide arv klastri kohta“ ja „ICD10 peatükkide arv klastri kohta“ väärtus langeb klastrite arvu suurenedes. Põhjus on selles, et kui klastrite arvu suurendada, on igas klastris vähem kontseptsioone ning seega ka vähem sündmusi. Samas mõõdikute „klastrite arv ICD10 koodi kohta“, „klastrite arv ICD10 peatüki kohta“ ja „klastrite arv dokumendi kohta“ väärtus tõuseb klastrite arvu suurenedes, kuna sama ICD10 kood, peatükk või dokument võib sattuse mitmesse eri klastrisse. Tõusva trendiga on ka

mõõdikud „3 kõige sagedasema concept ID osakaal klastrist“ ning „kõige sagedasema ICD10 peatüki osakaal klastrist“, kuna rohkema arvu klastrite korral on klastrid väiksemad ning nende sisu ühetaolisem.

Kokkuvõtteks võib öelda, et mida rohkem klastreid, seda kergem on klastreid iseloomustada mõne enamlevinud kotseptsiooni järgi, kuid seda rohkem on ka sarnaseid klastreid.

Mõõdik „üldistatud sündmuste osakaal algsetest sündmustest“ on joonise 7 järgi suhteliselt stabiilne 200 klastri juures word2vec akna suurustel 20-100, esineb vaid juhuslikke kõikumisi. Klastrite arvu suurenedes tõuseb mõõdiku väärtus aeglases tempos. Võimalik, et seda mõõdikut mõjutaks rohkem alapeatükis 3.2.6 toodud üldistatud sündmuse definitsioonis olev kahe algse sündmuse vahele jääv päevade arv, mis eraldab kahte üldistatud sündmust, kuigi seda ei katsetatud selles töös. Mõõdik „klastrite osakaal kontseptsioonides“ on nii klastrite arvu kui ka akna suuruse suhtes oluliselt tundlikum. Mida rohkem on klastreid, seda rohkematesse klastritesse jagunevad ka iga patsiendi sündmused ja seda keerukam on pilt üldistatud sündmustest.

Mõõdiku „klastri nime täpsus“ väärtused on joonise 8 kohaselt enamasti nullilähedased. Selle mõõdiku puudus on, et võrreldakse üksikuid kontseptsioone. Selle kohaselt loetakse kontseptsioonid zopikloon (19044883) ja zopiklooni 7.5 mg suukaudne tablett (19044885) erinevateks.

ICD10-l põhinevate mõõdikute puuduseks see, et need ei kajasta kogu andmestikku, sest kõik sündmused ei oma ICD10 koodi. ICD10 eelis on selles, et selle aluseks on hierarhiline süsteem, mille järgi kuuluvad koodid üksteist välistavatesse peatükkidesse ja alampeatükkidesse. On võimalik võrrelda, kui hästi vastavad peatükid klastritele. Ka ravimite toimeaineid kirjeldavad ATC koodid [Dru22] moodustavad sarnase hierarhia ning seega saaks ICD10-l põhinevad mõõdikud laiendada sündmustele, millel on ATC kood. See ei ole siiski jätkusuutlik lahendus, kuna on ka selliseid sündmuste kategooriaid, millel pole ei ICD10 ega ATC koodi. Isegi kui kõik erinevad sündmuste klassifikaatorid moodustaks hierarhilise süsteemi, tuleks vastavad mõõdikud iga klassifikaatori jaoks eraldi programmeerida. Kontseptsioonid on aga olemas igal sündmusel ning ka need moodustavad hierarhia, kuid selles võib ühel koodil olla mitu erinevat vanemkoodi.

5.3 Klastrite sisu ICD10 peatükkide järgi

Joonistelt 9 ja 10 ilmneb oluline erinevus klastrite arvus, mis ICD10 koodiga sündmusi sisaldavad, kui word2vec akna suurus on 60. Ilma aja tokeniteta jagunevad sellised sündmused 40 klastri vahel, aja tokeneid kasutades aga 8 klastri vahel. Arvestades, et nende jooniste aluseks olev klastrite koguarv on 200, jagunevad ICD10 koodiga sündmused võrdlemisi väikese arvu klastrite vahel. Ühest küljest näitab see, et ICD10 koodiga sündmused (st. peamiselt diagnoosid ning haiguslikud seisundid) on selgelt eristuvad muud tüüpi sündmustest, näiteks ravimi väljakirjutustest. Teisest küljest võiks eeldada, et diagnoos ning sellele vastav ravimi väljakirjutus võiks kuuluda samasse klastrisse ning

kui mõni klaster koosneb ainult ravimitest, ei ole selge, milleks neid ravimeid kasutatakse. Kõikidelt jooniste 9 ja 10 alamjoonistelt on näha, et osades klastrites leidub domineeriv ICD10 peatükk ning osade klastrite sisu on hajutatud üle mitme peatüki.

5.4 Ühe patsiendi sündmuste üldistamine

Tabelis 8 on näha mitut patsiendi sündmuste klasterdusel tekkinud klastrit, kuhu on langenud üksteisega seotud kontseptsioonid.

Klaster 102 on näide klastrist, kus kontseptsioonide vahel on ilmne seos: kõikide selles olevate konsteptsioonide nimetustes esineb sõna „alkohol“. Klastri nimes on ka mõningad vigastustele viitavad nimetused. Põhjus võib olla selles, et vigastused võivad olla sagedane alkoholijoobe tulemus. Klaster 193 on näide anatoomiliselt lähedastest kontesptsioonidest, selles on selgroo vaheketaste haigus (75344) ning närvijuurte haigus (4216397). Klastris 45 on peamiselt vaimsed häired, nagu depressioon (kontseptsioonid 4077577, 4338031, 440698, 4307111), paanikahäire (436074) ja isiksuse muutus (4105190).

Need 25 klastrit, mille sees oli vaid üks kontseptsioon, sisaldasid enamasti ravimeid. Nende seas on mitmeid kontseptsioonide paare, mis kuuluvad eri klastritesse ning millest üks kirjeldab ravimit üldiselt ja teine kirjeldab selle mingit spetsiifilist vormi. Näiteks on eri klastrites diasepaam (723013) ja diasepaami 10 mg/ml suukaudne lahus (35407482), samuti kontseptsioonid zopikloon (19044883) ja zopiklooni 7.5 mg suukaudne tablett (19044885).

Klastrite sisu põhjal võib öelda, et word2vec koos K-means klasterdusega suudab paigutada sarnaseid kontseptsioone ühte klastrisse, kuigi mõnes klastris oli ka kontseptsioone, mille vahel otsest seost ei ole. Näiteks klastris 26 on uinuti zopikloon [Dru] (19044885) ja antibiootikum amoksitsilliin [Medh] (19115197). Samuti oli sarnaseid kontseptsioone, mis kuulusid eri klastritesse. Klastrite nimed on loodud kogu andmestiku põhjal, mitte ainult uuritava patsiendi andmestiku põhjal. Seetõttu sisaldavad mitmed klastri nimed klastri sisus olevaid või nendega sarnaseid kontseptsioone, kuid mõne klastri nimi ei sobi sisuga kokku.

Joonisel 12 olevate üldistatud sündmuste seas on palju selliseid, mis hõlmavad vaid ühte sündmust algsest andmestikust. Jooniselt võib märgata kolme klastrit (5, 35 ja 61), mis moodustavad sarnase mustri andmestiku esimeses pooles. Klastrite 5 ja 35 puhul on see ilmne, kuna mõlemad sisaldavad ainult ravimit diasepaam. Samamoodi moodustavad sarnase mustri andmestiku esimeses pooles klastrid 26, 45 ja 148. Klastrid 26 ja 148 sisaldavad mõlemad zopiklooni kontseptsiooni, mis on uinuti. Klaster 45 sisaldab unehäire kontseptsiooni ning see võib seletada, miks selle klastri üldistatud sündmused sarnanevad klastrite 26 ja 148 üldistatud sündmustega. Klastrid 145 ja 178 moodustavad kogu patsiendi andmestiku ajatelje ulatuses samasuguse mustri, kuna mõlemad klastrid sisaldavad ainult klonasepaami eri vormide kontseptsioone.

5.5 Töö tugevused ja nõrkused

Käesolev töö rakendas terviseandmetel word2vec mudelit, et tuvastada sarnaseid sündmusi ja seejärel neid üldistada. Tegemist on esimese sellise tööga Eesti terviseandmetel ja ka rahvusvahelisest kirjandusest leiab sellist lähenemist vähe. Töö tugevuseks on see, et iga etapp on käsitletav omaette tulemusena, mida saaks tulevases uurimistöös edasi arendada.

Töö üheks puuduseks on, et käsitleti ainult sündmuse toimumise fakti. Kui andmestikus esineb kontseptsioon 40480694, tähendab see kõrget A1C mõõtmise taset hemoglobiinis, kuid täpse mõõtetulemuse kohta see teavet ei anna. Kuna selles töös töötati välja uus metoodika, on tulemusi keeruline varasemate töödega võrrelda.

5.6 Edasine uurimistöö

Edasises uurimistöös võiks ka mõõtetulemused word2vec sisendisse kaasata. Üheks võimalikuks lahenduseks oleks kasutada fibonnaci numbrite süsteemi, mida selles töös kasutati pausi pikkuse väljendamiseks.

Selles töös võeti sündmused vaid kahest OMOP'i tabelist, *condition_occurrence* (diagnoosid) ja *drug_exposure* (ravimid). Tulevases töös saaks sama metoodikat rakendada ka teistele tabelitele, näiteks *observation* (vaatlused) ja *procedure* (teostatud protseduurid).

Edasises töös saaks ka katsetada muid viise algsetest sündmustest üldistatud sündmuste loomiseks ning ka üldistustele nime leidmiseks.

6 Kokkuvõte

Selle töö eesmärk oli leida viis detailsete tervishoiusüsteemist pärit andmete üldistamiseks word2vec algoritmi abil. Töötati välja metoodika, mille abil saab üksikuid kindla kuupäevaga sündmuseid, näiteks diagnoose ja ravimi väljakirjutusi agregeerida algus- ja lõpukuupäevaga üldistatud sündmusteks. Samuti kirjeldati viisi, kuidas üldistatud sündmustele nimesid anda.

Treenides word2vec mudeli ajaliselt järjestatud tervisesündmuste peal, saab iga kontseptsioon vektori ning vähemalt 2. tüüpi diabeedi ja migreeni puhul näidati, et lähedased haigused ja ravimid saavad sarnase vektori. Word2vec'ist saadud vektorid klasterdati K-means algoritmiga ning selle käigus selgus, et mida rohkem on klastreid, seda sarnasemad on ühe klastri sees olevad kontseptsioonid, kuid seda rohkem on üksteisega sarnaseid klastreid. Word2vec'i treenimist katsetati ka selliselt, et sündmuste vahele lisati nendevahelise pausi pikkust väljendavad tokenid, kuid leiti, et klasterduse kvaliteeti see oluliselt ei mõjuta. Klastritesse määratud sündmuste alusel defineeriti üldistatud sündmuse mõiste. Üldistatud sündmuste arv patsiendi andmestikus oli keskmiselt 50% kuni 65% algsete sündmuste arvust. Töötati ka välja meetod klastritele nime andmiseks, kuid selliste nimede kvaliteedi üle on selle töö põhjal keeruline hinnangut anda.

Klastrite sisu analüüsiti ICD10 peatükkide kaupa ning selgus, et osad klastritest sisaldasid suuremas osas ühte peatükki ning teiste sisu oli rohkem hajutatud. Detailselt võrreldi ühe konkreetse patsiendi andmestikku ning selle üldistatud versiooni. Ilmnes, et klasterduse käigus tekkis mitmeid loogilisi kontseptsioonide klastreid, kuid leidus ka sarnaseid kontseptsioone mis sattusid eri klastritesse.

7 Tänusõnad

See magistritöö on läbi viidud uuringute RITA1/02-96 ja PRG1844 raames.

Viited

[Mik+13] Tomas Mikolov *et al. Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL]. URL: https://doi.org/10.48550/arXiv.1301.3781.

[Guo+18] Shunan Guo *et al.* „EventThread: Visual Summarization and Stage Analysis of Event Sequence Data“. *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), lk. 56–65. DOI: 10.1109/TVCG.2017.2745320.

[Hua+13] Zhengxing Huang *et al.* „Summarizing clinical pathways from event logs“. *Journal of Biomedical Informatics* 46.1 (2013), lk. 111–127. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2012.10.001. URL: https://www.sciencedirect.com/science/article/pii/S1532046412001554.

[DBD18] Pieter De Koninck, Seppe vanden Broucke ja Jochen De Weerdt. „act2vec, trace2vec, log2vec, and model2vec: Representation Learning for Business Processes“. Teoses: *Business Process Management*. Toim. Mathias Weske *et al.* Cham: Springer International Publishing, 2018, lk. 305–321. ISBN: 978-3-319-98648-7.

[Sii23] Õie Renata Siimon. „Patient Treatment Trajectories Using Vector Embeddings“. Magistritöö. Tartu Ülikool, 2023.

[Dev+19] Jacob Devlin *et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[Oja+23] Marek Oja *et al.* „Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned“. *JAMIA Open* 6.4 (detsember 2023), ooad100. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooad100. eprint: https://academic.oup.com/jamiaopen/article-pdf/6/4/ooad100/54025819/ooad100.pdf. URL: https://doi.org/10.1093/jamiaopen/ooad100.

[SI] Observational Health Data Sciences ja Informatics. *Standardized Data: The OMOP Common Data Model*. https://www.ohdsi.org/data-standardization/. Vaadatud 08.07.2024.

[Org19] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. https://icd.who.int/browse10/2019/en. Vaadatud 19.12.2023. 2019.

[Ser] Odysseus Data Services. *Athena OHDSI vocabularies repository*. https://athena.ohdsi.org/search-terms/start. Vaadatud 17.03.2024.

[Meda] MedlinePlus. *Hemoglobin A1C (HbA1c) Test*. https://medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/. Vaadatud 20.04.2024.

[Medb] MedlinePlus. *Metformin*. https://medlineplus.gov/druginfo/meds/a696005.html. Vaadatud 20.04.2024.

[Medc] MedlinePlus. *Pioglitazone*. https://medlineplus.gov/druginfo/meds/a699016.html. Vaadatud 08.05.2024.

[Medd] MedlinePlus. *Frovatriptan*. https://medlineplus.gov/druginfo/meds/a604013.html. Vaadatud 22.04.2024.

[Mede] MedlinePlus. *Sumatriptan*. https://medlineplus.gov/druginfo/meds/a601116.html. Vaadatud 22.04.2024.

[Medf] MedlinePlus. *Zolmitriptan*. https://medlineplus.gov/druginfo/meds/a601129.html. Vaadatud 22.04.2024.

[Medg] MedlinePlus. *Rizatriptan*. https://medlineplus.gov/druginfo/meds/a601109.html. Vaadatud 08.05.2024.

[Dru22] WHO Collaborating Centre for Drug Statistics Methodology. *Anatomical Therapeutic Chemical*. https://www.whocc.no/atc/structure_and_principles/. Vaadatud 19.12.2023. 2022.

[Dru] Drugs.com. *Zopiclone*. https://www.drugs.com/zopiclone.html. Vaadatud 02.05.2024.

[Medh] MedlinePlus. *Amoxicillin*. https://medlineplus.gov/druginfo/meds/a685001.html. Vaadatud 02.05.2024.

Lisad

I. Word2vec'i kood

Kood, mis on mõeldud word2vec mudeli treenimiseks ning selle abil sarnaste kontseptsioonide leidmiseks, on kättesaadav GitLab'i projektina:

https://gitlab.cs.ut.ee/kermos/word2vec-similar-concepts

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Kermo Saarse**

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

   Tervisesündmuste üldistamine sõnavektorite abil

   mille juhendaja on Sulev Reisberg

   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.

Kermo Saarse
pp.kk.aaaa