TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele

Liset Marleen Pak

Tekstikaeve rakendamine isiksusetestidel personali värbamise eesmärgil

Magistritöö (15 EAP)

Juhendaja Anna Leontjeva, PhD

Tartu 2018

Tekstikaeve rakendamine isiksusetestidel personali värbamise eesmärgil

Lühikokkuvõte:

Käesolevas töös rakendatakse tekstikaeve meetodeid ettevõttes Psience OÜ välja töötatud isiksusetesti avatud vastuste ehk enesekohaste tekstide analüüsimiseks. Töö eesmärgiks on analüüsida tekste personali värbamise kontekstis, kasutades tõenäosusliku teemade modelleerimise algoritmi Dirichlet' peitlahutust (LDA), ja võrrelda tekstianalüüsi tulemusi testi numbriliste skooridega. Teemade modelleerimise tulemusel leiti kolm teemat, mille abil oli võimalik vastajaid nende sõnakasutuse alusel kirjeldada. Samuti leiti võrdleva analüüsi põhjal, et teemade modelleerimise edasiarendatud versioon võiks ettevõttel tulevikus aidata muuta tekstianalüüsi automaatsemaks.

Võtmesõnad:

Tekstikaeve, teemade modelleerimine, Dirichlet' peitlahutus, isiksusetestid, psühholoogia.

CERCS: P175, Informaatika, süsteemiteooria

Application of Text Mining in Personality Tests for Recruiting

Abstract:

This thesis describes how text mining methods have been applied to analyse self-description texts given in the context of personality tests that are used by Estonian HR-company Psience OÜ. The aim of the thesis is to use probabilistic topic modelling algorithm called latent Dirichlet allocation (LDA) to analyse the texts and to compare the results with the scores of the personality test. As a result of applying topic modelling techniques, three topics where found that allowed to describe the personality test takers by their word use. Also, the comparative analysis showed that refined version of topic modelling approach could help to make the text analysis more automatic in the future as it is today.

Keywords:

Text mining, topic modelling, latent Dirichlet allocation, personality tests, psychology.

CERCS: P175, Informatics, systems theory

Sisukord

1. Sissejuhatus

Andmekaeve (*data mining*) ja selle võimaluste laialdase levikuga, sh suurandmetel põhineva analüüsiga on järjest suuremat huvi äratanud ka n-ö struktureerimata andmehulkade ehk erinevate tekstikorpuste analüüs. Tekste „toodetakse“ igapäevaselt väga suurtes kogustes ning suur osa neist on meile ka Interneti vahendusel kättesaadavad, nt *online*-uudised, kodulehe tekstid, sotsiaalmeedia postitused, erinevate toodete arvustused, teenuste tagasisided, kuulutused, reklaamid jms. Tekstid kannavad endas olulist infot, kuid nende läbitöötamine on aeganõudev ja peamiselt inimressurssi kasutades ka kulukas tegevus. Võrreldes arve sisaldavate andmehulkadega, mida on mugavaim talletada andmebaasides struktureeritud kujul, on tekstid eelkõige mõeldud inimeste vahel informatsiooni vahetamiseks. Seega on informatsiooni vahetamine vaadeldav sotsiaalse protsessina. Kuivõrd struktureeritud andmeid on arvutil lihtne lugeda ja analüüsida, on teksti ja selles sisalduva tähenduse mõistmine masina jaoks ilma inimese kõrvalise abita veel võrdlemisi keerukas ülesanne.

Inimeste loomulikul keelel (nt eesti, inglise, korea vms keel) põhineva teksti analüüsi ehk tekstikaeve (*text mining*) kasutusvaldkond on lai, alustades automaatotsingusüsteemide loomisest kuni tarkade tekstikaeve rakenduste ja soovitussüsteemideni välja. Antud magistritöös on fookusesse tõstetud tekstikaeve võimalused selles osas, kuidas paremini mõista personaliotsingu ja karjäärinõustamise kontekstis erinevate inimeste isiksuseomadusi, kasutades uuritavate kirja pandud enesekohast teksti. Võrreldes inglise keelel põhineva tekstikaevega, on eesti keelel põhineva tekstikaevega tegelemine kindlasti omaette väljakutse väiksema keelekorpuse tõttu.

Kuigi isiksuse ja isiksuseomaduste uurimine kuulub peaasjalikult psühholoogia valdkonda, on tegu interdistsiplinaarselt huvipakkuva teemaga. Nii on näiteks keeleteadlased huvitunud sellest, milliste keeleliste atribuutidega inimesed enda ja teiste isiksuseomadusi kirjeldavad ja edasi annavad (Orav, 2006). Antropoloogia alal rajas Margaret Mead 1920. aastatel tee uurimaks kultuurilisi seoseid, konstruktsioone ja isiksuseomadusi. Samuti on isiksuse käsitlemise kui uurimisteema vastu tundnud huvi sotsioloogid, otsides seoseid sotsiaalsete, ühiskondlike struktuuride ja indiviidide tunnete, hoiakute ja käitumise vahel (vt nt McLeod ja Lively, 2006). Eesti keelel põhinevaid tekstikaeve meetodeid rakendavaid uurimusi on samuti tehtud mitmeid (vt ka ptk 3.1). Käesoleva magistritöö autorile teadaolevalt ei ole eesti keele loomuliku keele töötlusele põhinevalt ükski teadustöö isiksuseomadusi senini käsitlenud.

Probleemi lähtekoht

Personaliettevõttel Psience OÜ on välja töötatud ja kasutuses psühholoogiline isiksusetest PIST (Psience'i isiksuslike soodumuste test)[1], mis aitab testi täitjate kohta avada nende isiksuslikke omadusi, kogemusi ja väärtusi. Test põhineb psühholoogias väga tuntud ja laialdaselt kasutatud suure viisiku teoorial[2] (Big Five Theory) ning võtab arvesse ka uuemaid käsitlusi isiksusepsühholoogiast (Mischel jt, 2002), mis rõhutavad isiksuseomaduste avaldumise olulisust kontekstipõhiselt. Ettevõttes Psience kasutatakse PIST-testi peamiselt selleks, et värbamisprotsessis hinnata erinevate kandidaatide sobivust vastavalt täitmist vajavale ametikohale. Kuivõrd Psience pakub värbamisteenust vahendusteenusena – sobilikku personali aidatakse värvata Psience'i klientidele ehk teistele ettevõtetele –, on vahendajal oluline roll pakkuda oma kliendile potentsiaalselt kõige sobilikumat uut töötajat. Selleks võetakse ühelt poolt arvesse kandidaadi motivatsiooni ja varasemat kogemust (sh CV, motivatsioonikirja ja tööintervjuu põhjal), isikuomadusi (sh intervjuu, soovitajate tagasiside ja isiksusetesti põhjal) ning teisalt peetakse silmas ka seda, millisel ametikohal kandideerija tööle hakkab, millised on selle ettevõte ootused jm. Sobiliku kandidaadi välja valimine on seega ajamahukas ja analüütiline protsess, mis nõuab kvalifitseeritud tööjõudu.

PIST-test ja ettevõtte vajadus

Ettevõttel Psience on vajadus teostada 2012. aastal välja töötatud PIST-testi uurimuslik analüüs. PIST-testi tulemuste hindamine tugineb ühelt poolt numbrilistele väärtustele (tulenevalt vastaja väitepaari valikust ja skaalapõhiste väidete hindamisest) ja teisalt testi täitja enesekohastele kirjeldustele, mis antakse avatud vastustena. Avatud vastuseid ei ole seni võetud arvesse PIST-testi numbrilistes lõppskoorides, vaid neid kasutatakse testiraportis täiendava informatsiooni väljatoomiseks analüütilise tekstina. Antud töö peamine huvirõhk on asetatud vaba teksti analüüsimisele, kuivõrd see on testi hindamisel ajamahukaim osa. Avatud vastuste analüüsimine tekstikaeve meetodil aitaks saada ülevaadet testi täitjate isikuomadustest (lähtudes vastajate enesekohasest kirjeldusest ja sõnakasutusest), võimaldaks kõrvutada tekstianalüüsi tulemusi testi skooridega ning teha praegu n-ö käsitsi tehtavat tekstianalüüsi automaatsemaks.

[1] Edaspidi kasutatakse läbivalt terminit PIST-test.
[2] Eesti keeles kasutatakse ka viie faktori teooria mõistet.

Uurimuse eesmärgid

1. Leida ja kirjeldada PIST-testi enesekohastes tekstides peidetud mustreid vastanute isiksuseomaduste kohta, rakendades masinõppel põhinevat teemade modelleerimise algoritmi.
2. Võrrelda tekstianalüüsi tulemusi PIST-testi numbriliste skooridega.
3. Teha ettepanekuid PIST-testi kui toote edasi arendamiseks.

Uurimisküsimused

1. Millised peidetud mustrid vastanute isiksuseomaduste kohta avalduvad PIST-testi tekstilistes andmetes?
2. Milliseid sõnu teatud klastrisse koonduvad inimesed enda kohta kasutavad?
3. Millised omadused iseloomustavad PIST-testi täitnuid tekstianalüüsi ja omadusskooride analüüsi põhiselt?
4. Milliseid ettepanekuid PIST-testi arendamiseks on võimalik uurimistulemuste põhjal teha?

Käesoleva töö struktuur

Töö teises peatükis antakse ülevaade isiksusetestide kasutamisest psühholoogias ning tuuakse välja isiksusetestide kasutamise võimalused personalivaldkonnas. Kolmandas peatükis antakse ülevaade tekstikaevest, selle meetoditest ja tuuakse näiteid selle rakendusvõimalustest. Lisaks tutvustatakse tõenäosusliku teemade modelleerimise meetodit LDA (Dirichlet' peitlahutus), mis on käesolevas töös peamine analüüsimeetod. Töö neljandas peatükis kirjeldatakse lähemalt PIST-testi ja analüüsi aluseks olevaid andmeid. Viiendas peatükis kirjeldatakse töövoogu etappide kaupa. Kuuendas peatükis esitatakse tõenäosusliku teemade modelleerimise meetodil teostatud tekstianalüüsi ning teksti ja omadusskooride võrdlusanalüüsi tulemusi. Kokkuvõttes tuuakse lisaks välja PIST-testide analüüsiga seonduvalt edasised ettepanekud.

2. Isiksusetestide kasutamine ja vajadus

Käesolevas peatükis kirjeldatakse isiksusetestide peamisi teoreetilisi lähtekohti, kasutusvaldkondi ja väljakujunenud analüüsimeetodeid. Samuti tuuakse välja, miks isiksuseteste kasutatakse.

2.1 Isiksusepsühholoogia teoreetiline taust

Isiksus (*personality*) on kogum isiksuseomadustest, mida me omistame indiviidile iseloomulikuna seotuna tema mõtlemisviisi, tunnete ja nende avaldumise, käitumisega. Isiksusetestid on psühholoogia valdkonnas laialdaselt kasutuses ning nende ulatuslikum levik on seotud isiksusepsühholoogia arenguga 20. sajandil. Kõige avaramas mõttes tegeleb isiksusepsühholoogia "inimkäitumise üldiste ja kõikehaaravate seaduspärasuste uurimisega", sh inimeste individuaalsete erinevuste ja sarnasuste uurimisega (Allik, 2003: 26). Isiksuseomadused (*traits*) hõlmavad endas nii konkreetse inimese mõtteid, tundeid kui ka teatud väljendus- ja käitumisviisi, kuid nende abil saab üldistatult kirjeldada inimesi teatud sarnasuste või vastanduste alusel. Tänapäevase isiksusepsühholoogia seisukohast lähtuvalt on isiksuseomadused ajas suhteliselt püsivad (Allik, 2003: 53).

Erinevate psühholoogiliste testide, sh isiksusetestidega on kokku puutunud arvatavasti enamik inimesi. Kuigi isiksuseteste täites annab inimene subjektiivseid hinnanguid enda omaduste kohta, on korduvate uuringutega tõestatud, et need hinnangud langevad põhilises osas kokku sellega, kuidas hindavad sama inimest teda hästi tundvad lähedased isikud (Allik, 2003: 53). Kuivõrd isiksuseomadused ei ole olemuslikult füüsikaliste suuruste abil mõõdetavad, on isiksuseomaduste uurimise üheks oluliseks alustalaks kujunenud leksikaalne lähenemine. Leksikaalse hüpoteesi järgi „talletuvad kõik sotsiaalses lävimises olulised individuaalsed omadused aja jooksul keeles; mida silmatorkavama ja olulisema omadusega on tegemist, seda suurem on tõenäosus, et selle omaduse tähistamiseks on keeles omaette sõna" (Allik, 2003: 33). Seega on põhjendatud, miks kasutada keelelisi vahendeid isiksuseomaduste ja nendega seotud seaduspärasuste uurimisel.

Isiksustestide väljatöötamine sai alguse aga juba enne leksikaalse hüpoteesi formuleerimist. Isiksustestide algusaastateks saab pidada I maailmasõja perioodi, kui Ameerika psühholoog Robert Woodworth töötas koos kolleegidega välja ankeettestid, saamaks teada, millised sõdurid on sõjaväeteenistuseks kõlbulikumad oma isiksuseomaduste poolest, sh kellel on väiksem tõenäosus langeda lahinguolukorras šokiseisundisse. Erinevate isiksuseomaduste kirjeldamiseks loodi palju erinevaid teste, kuid peagi tekkis vajadus kirjeldada inimese

isiksuslikke suundumusi võimalikult lühikese omaduste loendi abil. Rakendades leksikaalset lähenemist inimeste isiksuslike eripärade uurimisel, jõuti 20. sajandi teisel poolel järeldusele, et isiksuse struktuuri on võimalik kirjeldada minimaalselt viie faktori kaudu. 1990. aastatel formuleerisid Ameerika psühholoogid Robert McCrae ja Paul Costa viie faktori teooria (1992; 1995), mis tänapäevalgi isiksuse uurimisel laialdast rakendust leiab. Antud teooria, mida tuntakse ka suure viisikuna (Big Five), jagab isiksuseomadused viide põhilisse dimensiooni (Allik, 2003: 43–44):

1. **neurootilisus** (*neuriticism*) – on seotud negatiivsete emotsioonide kogemisega (nt kurbus, hirm, viha jms);
2. **ekstravertsus** (*extraversion*) – on seotud positiivsete emotsioonidega, aktiivsusega;
3. **avatus kogemusele** (*openess to experience*) – on seotud inimese huviga ümbritseva maailma ja oma siseelu vastu;
4. **sotsiaalsus** (*agreeableness*) – on seotud usaldusega teiste inimeste vastu ning omakasupüüdmatusega;
5. **meelekindlus** (*conscientiousness*) – on seotud kontrolliga oma soovide ja impulsside üle.

Suurel viisikul põhinevad paljud isiksusetestid, millest ülemaailmselt tuntuim on ehk NEO-PI-R (Allik, 2003: 39). Samuti on suudetud tõestada, et viiel faktoril põhinev struktuur on universaalne ja kehtib kultuuride üleselt (Eysenck, 1997), kuigi sellele esineb ka teatavat kriitikat (vt Friedman ja Schustack, 2016). Eestikeelse NEO-PI-R testi hindamisel on viiefaktoriline struktuur isiksuseomaduste jaotumusel samuti leidnud kinnitust (Kallasmaa jt, 2000: 265).

Oluline on siinkohal aga silmas pidada, et kuigi suure viisiku teooria järgi on isiksuseomadusi võimalik grupeerida viie faktori alla, vaadatakse inimese isiksuseomaduste hindamisel suuremat hulka spetsiifilisi soodumusi või iseloomujooni. Näiteks NEO-PI-R-tüüpi testi puhul hindab inimene kokku 30 erinevat iseloomuomadust, mis omakorda koonduvad eelpool kirjeldatud viie faktori alla (Costa jt, 1995: 124). Nii kuuluvad näiteks sotsiaalsuse faktori alla usaldus, siirus, omakasupüüdmatus, järeleandlikkus jm, meelekindluse faktori all aga hinnatakse näiteks kohusetunnet, eesmärgipärasust, asjatundlikkust (Allik, 2003: 52). Oluline on siinkohal rõhutada, et neid viit faktorit ei saa teineteisele vastandada, kuivõrd igat faktorit on võimalik vaadelda teatava skaalana, ning samuti ei kirjuta ühe omaduse tugevam avaldumine üle teisi omadusi. Seega on inimestel üldjuhul kõik viis isiksuseomaduse faktorit teatava tonaalsuse või tugevusega esindatud.

PIST-testi puhul hinnatakse kokku 27 erinevat omadust (vt ptk 4.1), mis on töötegemise kontekstis relevantsed. PIST-testil on oma originaalstruktuur, mis tähendab, et suure viisiku teooria põhist küsimustikku on kasutatud PIST-testi struktuuri kehtivuse valideerimiseks. Seega tugineb PIST-test suure viisiku teooriale vaid osaliselt. Teine oluline lähtekoht PIST-testi väljatöötamisel on Walter Mischeli sotsiaalkognitiivsel lähenemisel, mis sedastab, et inimene käitub tingimuslikult vastavalt teatud kontekstile ning seda on ka oluline inimese isiksuse hindamisel jälgida (Mischel jt, 2002). Sotsiaalkognitiivse lähenemise järgi ei pruugi teatud isiksusomadus täpselt sama moodi avalduda erinevates olukordades, vaid näiteks muidu tasane kolleeg võib oma pereringis olla väga avatud ja energiline. Lähtudes sellest, et omadusi ja käitumuslikke kontekste ei ole otstarbekas üksteisest lahus vaadata, on PIST-testi välja töötamisel pandud oluline rõhk avatud vastuste kaudu anda vastajale võimalus retrospektiivseks hinnanguks omaenda käitumisele ja mõtlemisviisile. See omakorda võimaldab numbriliste skooride ja keskmiste statistiliste näitajate kõrval rohkem avada ja mõista inimese enda kirjeldatud sisemaailma.

2.2 Isiksusetestide kasutamine

Jättes kõrvale teaduslikud uurimused ja inimeste huvi mõista paremini oma siseelu, kasutatakse professionaalses sfääris isiksusteste nii karjäärinõustamises kui ka personalivaliku kontekstis. Käesolevas magistritöös keskendutakse neist viimasele. Eelkõige aitavad isiksusetestid personalivaliku kontekstis hinnata, kuivõrd hästi võiks värbamisprotsessis osalev potentsiaalne töötaja sobida konkreetsele ametikohale oma isiksuseomaduste poolest. Ühtlasi võimaldavad isiksusetestid ka prognoosida inimese toimetulekut ja pühendumist ametikohal. Isiksusteste kasutatakse ka siis, kui soovitakse töötajat oma organisatsiooni sees teisele ametikohale paigutada. Siiski tuleks isiksusetestide tulemusi kasutada kõrvuti teiste hindamismeetodite tulemustega, mitte otsuseid langetada ainult isiksusetestide põhjal (Costa jt, 1995). Inimese isiksuseomaduste hindamine mingile ametikohale sobivuse testimiseks on oluline, sest isiksuseomadused on ajas suhteliselt püsivad, avalduvad nii suhtlemises, töö tegemise stiilis kui ka mõtlemislaadis. Samuti on isiksuseomadused vähemal või rohkemal määral seotud inimese väärtuste ja motivatsiooniga. Näiteks PIST-testi abil hinnatud isiksuslike soodumuste alusel saadakse inimese põhiprofiil, mis annab võimaluse määrata, kuidas ta tööalastes situatsioonides tõenäoliselt käitub, võttes aluseks eelduse, et inimene on oma vastustes olnud adekvaatne.

Psühholoogid on uurinud ka tööl hakkama saamise ja isiksuseomaduste vahelisi seoseid. Näiteks on leitud, et inimesed, kellel on kõrgem skoor meelekindluse faktoris, saavad tööl paremini hakkama tänu oma püsivusele ja kõrgemale vastutustundele, mis aitavad eesmärke ellu viia (Friedman ja Schustack, 2016: 191). Ettevõtjatel on täheldatud kõrgemat meelekindlust ja avatust ning madalamat neurootilisust – neil on kõrge saavutusvajadus ja innovaatiline lähenemisviis ning nad peavad stressile paremini vastu (Zhao ja Siebert, 2006).

Isiksusetestide kasutamisel isiksuseomaduste uurimisel tuleb arvestada ka selle meetodi võimalike puudustega. Üheks enim uuritumaks selliseks puuduseks on sotsiaalse soovitavuse mõju testi täitja vastustele (vt Konstabel, 2006). Sotsiaalne soovitavus isiksustestide täitmisel seisneb selles, et testi täitja võib vastuste andmisel juhinduda soovist näidata end inimesena paremas valguses, kui ta objektiivselt võttes seda on. Sotsiaalset soovitavust seostatakse ka ühiskonnas kehtivate normidega, mille kohaselt peetakse näiteks positiivsust, heatujulisust ja suhtlusvalmidust pigem headeks ning soositud omadusteks. PIST-testis on kasutatud sotsiaalset soovitavust tuvastavaid väiteid, mille eesmärk on vähendada sotsiaalselt soovitatavate vastuste mõju testi tulemustele.

3. Tekstikaeve, meetodid ja võimalused

Igapäevaselt tekib tohututes kogustes elektroonilisi andmeid, millest suur osa on ka Internetist kättesaadavad. Ühe osana talletuvatest andmehulkadest moodustavad andmed, mida on võimalik hoida struktureeritud kujul andmebaasides. Suurandmed (*big data*) ja andmekaeve (*data mining*) on seotud rohkemal määral just struktureeritud andmete ja nende analüüsimisega. Samas „toodetakse" elektrooniliselt isegi rohkem struktureerimata andmeid, mis on peamiselt kas vabateksti, video, piltide või audio. Sellist laadi struktureerimata andmeid tekib inimtegevuse käigus näiteks erinevate sotsiaalmeedia kanalite kaudu, aga ka foorumite, blogipostituste, uudisvoogude, ametlike dokumentide, teadusartiklite, e-kirjade, tekstiliste logidega jms (Gandomi ja Haider, 2015: 140). Arvutiteaduse alal on teksti kujul hoitavaid struktureerimata andmeid võimalik analüüsida tekstikaeve meetoditega.

3.1 Tekstikaeve mõiste ja lähenemisviisid

Tekstikaeve (*text mining*) on protsess tähendusliku informatsiooni saamiseks tekstist (Kwartler, 2017: 1). Seega on tekstikaeve üheks olulisimaks ülesandeks avastada tekstis sisalduvad peidetud mustrid masinõppe põhimõtetel. Tihti kasutatakse tekstikaeve ja tekstianalüütika (*text analytics*) mõisteid ka sünonüümidena (Kwartler, 2017: 2). Tekstikaeve andmeteks on loomulikul keelel põhinevad tekstid, nii nagu on harjunud neid lugema inimsilm. Peamiselt on tekstikaeve abiks seal, kus inimtegevusena on tekkinud palju struktureerimata andmeid (tekste), mida saaks aga analüüsituna kasutada vajalike otsuste langetamiseks. Näiteks võimaldab tekstikaeve rakendamine analüüsida klientidelt laekunud toodete ja teenuste tagasisidet, saades seeläbi äriliste otsuste tegemiseks sisendit (näiteks millised tooted ja mis põhjusel tarbijatele rohkem meeldivad või miks ei olda teatud toodetega rahul). Samuti on võimalik tekstikaeve abil leida huvitavaid mustreid näiteks sotsiaalmeedia tekstilistest andmetest nende tekkimisel reaalajas. Selliselt on näiteks võimalik saada kohest tagasisidet inimeste hulkadelt teatud poliitiliste vms muudatuste või otsuste osas (Zhai ja Massung, 2016: 243).

Masinatele teeb loomulikul keelel põhinevast tekstist aru saamise keeruliseks eelkõige seal esinevate tähenduste, konteksti ja meelsuse mõistmine. Tekstikaeve vaates on see keeruline ülesanne väga arusaadavatel põhjustel – tekstilised andmed on tavapäraselt loodud inimsuhtluses ja avaldatud seega mõtestamiseks teatud sotsiaalses kontekstis. Zhai ja Massung (2016: 244) on inimese kohta kasutanud ka mõistet „subjektiivne sensor" (*subjective sensor*), viidates sellega topeltvahendatusele: inimesed väljendavad teksti kaudu seda, mille

kohta nad on ümbritsevast keskkonnast vaatluse tulemusel tähelepanekuid teinud. Seega on keel ja kirjapandud tekst kui vahend andmaks edasi inimese tehtud järeldusi, tundeid ja mõtteid nii end ümbritseva keskkonna kui ka oma sisemaailma ja tundmuste osas.

Loomuliku keele töötlus (*natural language processing*) on arvutiteaduslik tehnika, mis võimaldab arvutil ehk masinal mõista loomulikus keeles esitatud teksti (Zhai ja Massung, 2016: 39). Kui loomuliku keele töötluse arenguaegadel 20. sajandi keskel toetuti automaatse masintõlke välja töötamisel veel suuresti käsitsi kirjutatud reeglitele, siis tänapäeval on end tõestanud statistiliste meetodite ja masinõppe kasutamine loomuliku keele töötluses (Zhai ja Massung, 2016: 43). Loomuliku keele töötluse peamised probleemid tulenevad aga keelelise kommunikatsiooni sellest osast, mida me jätame sõnadega otse väljendamata või täpsustamata. Inimeste vahelises suhtluses on tavaline, et esitatud lauseid ja sõnu (nii suulisel kui ka kirjalikul kujul) on võimalik mitmeti mõista. Tavalises inimsuhtluses suudab aju sellised mitmeti mõistetavused lahendada, kuna inimaju on võimeline konteksti ja tähendusi kiiresti sünteesima tänu varasemale (sotsiaalsele) kogemusele ja keeletajule. Mitmeti mõistetavus on samuti sisse kodeeritud meie sõnavarasse, näiteks esineb sõnavormidel samakujulisus ehk homonüümia[3]. Mitmeti mõistetavust toetab ka nii inimsuhtluse kaudu edasi antav iroonia, demagoogiavõtted kui ka eufemismide, metafooride jms kasutamine.

Tekstikaeve kasutusvaldkond on äärmiselt lai ning lisaks andmeteadusele leiavad tekstikaeve masinõppe meetodid üha enam kasutust ka sotsiaal- ja humanitaarvaldkonnas (vt nt Schwartz jt, 2013; Coppersmith jt, 2014; Schmidt, 2012). Kuivõrd tekstikaeve peamiseid lähenemisviise võib erinevalt liigitada (Aggarwal ja Zhai, 2012; Zhai ja Massung, 2016), on järgnevalt pakutud üks võimalus nende organiseerimiseks:

- **Klassifitseerimine** (*classification*) ja **klasterdamine** (*clustering*) – mõlema lähenemisviisi ülesandeks on dokumente teatud viisil grupeerida. Kui klassifitseerimine toetub juhendatud masinõppe meetoditele, vajades andmete treenimisel ja testimisel ühe olulise sisendina ka inimese lisatud märgendeid (*labels*), siis klasterdamine kui juhendamata meetod grupeerib andmeid nendes leiduvate seaduspärasuste alusel. (Kwartler, 2017: 181–182) Oluline on märkida, et juhendatud ja juhendamata masinõpet kasutatakse paljudes uurimisprobleemides kombineerituna.

[3] Homonüümia näiteks võib tuua omastavas käändes sõna „kuue“, mis võib viidata nii arvsõnale „kuus“ kui ka nimisõnale „kuub“.

- **Teemade modelleerimine** (*topic modeling*) on tõenäosusel põhinev lähenemine dokumentide klastriteks ehk teemadeks jaotamisel (Kwartler, 2017: 161). Teemade modelleerimist on kasutatud peamise meetodina isiksustestide vabavastuste analüüsiks ka käesolevas magistritöös. Samuti on teema modelleerimist rakendatud Elizaveta Yankovskaya (2017) ja José Santose (2011) magistritöödes.
- **Meelsusanalüüsi** (*sentiment analysis*) kasutatakse siis, kui teksti töötluse tulemusel on vaja teha järeldusi tekstina kirja pandud arvamusavalduste kohta. Meelsusanalüüsiga on võimalik teada saada teksti kirjutaja arvamuste (sh kas need kalduvad positiivse või negatiivse suunale) ja ka emotsioonide kohta. Palju rakendatakse meelsusanalüüsi näiteks toodete ja teenuste tagasiside analüüsimiseks ärilistel eesmärkidel. (Zhai ja Massung, 2016: 389–394)
- **Informatsiooni eraldamine** (*information extraction*) – on meetod erinevate infot kandvate mõtestatud elementide eraldamiseks tekstist. Informatsiooni eraldamise vajaduse näiteks oleks näiteks sündmuste kirjeldusest võtta välja uurija jaoks tähenduslikud osas: ürituse nimi, toimumise aeg, piletihind, koht jms. Informatsiooni eraldamist eestikeelsetest kohtulahenditest on kasutanud Katrin Valdson (2016) oma magistritöös.

3.2 Teemade modelleerimine LDA meetodiga

Teemade modelleerimise üldpõhimõtte seletamiseks on vaja teha vahet mõistetel „teema“ ja „võtmesõna“. Kui informatsiooni otsimise kontekstis ollakse harjutud opereerima võtmesõnapõhiste päringute kaudu (näiteks kõige tavalisem Google’i otsing), siis „teema“ on abstraktsem kategooria, mida on võimalik sõnastada (üldjuhul) semantiliselt omavahel seotud sõnade ja lausete koosesinemise pealt ning mis eraldiseisva sõnana ei pruugi tekstides esinedagi (vrd näiteks laulusõnadele antav teemakategooria „lembelaulud“). Üheks levinumaks teemade modelleerimise näiteks on ajalehe artiklite jaotumine teemade järgi gruppidesse, näiteks välis-, spordi-, kultuuri-, majandusuudisteks jne. Üldteemasid saab jagada ka väiksemateks alateemadeks, näiteks spordiuudiste all saab omakorda eristada jalgpalli-, tennise-, aga ka dopingu-uudiseid. Seega aitab dokumentide kogumit teemadeks modelleerida nendes peituvad läbivad sarnased temaatilised jooned. Sõnastatud teemad aitavad aga saada ülevaadet tekstide peamisest sisust. Teemade modelleerimise ülesanded võivad olla: info otsimine, kategoriseerimine ja korpuse uurimine (Blei, 2012a: 79).

Käesolevas magistritöös on teemade modelleerimisel kasutatud *latent Dirichlet allocation* ehk **LDA mudelit** (Blei jt, 2003), mille eestikeelse vastena on kasutatud ka Dirichlet' peitlahutust (Yankovskaya, 2017). LDA mudel kasutab tõenäosusliku teemade modelleerimise algoritmi, mis analüüsib statistiliselt sõnade kasutust dokumentides, leidmaks tekstide vahelised temaatilised seosed (Blei, 2012a: 77). LDA-meetodit kasutades analüüsitakse dokumendid generatiivse protsessi kaudu, leidmaks nendes peituvad varjatud struktuurid, mida nimetakse teemadeks. LDA mudeli rakendamisel on võimalik ühte dokumenti[4] liigitada tõenäosuse alusel mitme erineva teema alla. Kõik dokumendid kokku jagavad sama hulka teemasid.

Teemade modelleerimisel kasutab LDA-mudel järjestamata sõnade esinemissageduse ehk *bag-of-words* meetodit, mille kohaselt pole oluline sõnade järjekord lauses ega tekstis. Kwartleri järgi on *bag-of-words* meetodid on tekstikaeves laialdaselt kasutusel, kuna selle taga on võrdlemisi lihtne ja efektiivselt rakendatav loogika – iga sõna või sõnaühendit (N-gramme) võetakse kui üht dokumendi tunnust nende esinemissageduse järgi. Seega ei ole sõnade järjekorda ega grammatilisi sõnatüüpe (nt nimi-, tegu- omadussõna) *bag-of-words* analüüsi puhul arvesse võetud või eristatud. Tavapärane väljund selle analüüsimeetodi kasutamisel on dokument-termin maatriks, kus iga rida representeerib üht dokumenti vaadeldavas korpuses, ning veergudes on unikaalsed sõnad või moodustatud sõnaühendid (N-grammid). Maatriksi sisuks on sõna esinemissagedus vastava dokumendi kohta. Ühtlasi võimaldab *bag-of-words* meetodi rakendamine talletada struktureerimata tekstides olevad andmed maatriksisse struktureeritud kujul. (Kwartler, 2017: 20–25) Tüüpiliselt on sellised maatriksid aga väga hõredad (*sparse*), st neis esineb sagedusloendites palju nulle (0), kuivõrd mõnd sõna võib esineda ainult ühes dokumendis ja ülejäänutes puudub.

LDA-meetod võtab arvesse mitut olulist eeldust (Blei, 2012a):

- Sõnade järjekord tekstis ei oma tähtsust (*bag-of-words*).
- Korpuses esinevate dokumentide järjekord ei oma tähtsust.
- Teemade arv on fikseeritud (antakse mudelile ette). Teemad tuginevad koosesinevatele sõnadele ja sõnakasutuse mustritele dokumentides.

[4] Dokumendina saab mõtestada erinevaid tekste sõltuvalt uurimisprobleemist, nt võib dokument olla ajalehe artikkel, *tweet*, kasutaja kommentaar, ametlik dokument, kohtulahend vms. Käesolevas töös vastab dokumendi mõistele ühe inimese enesekohane tekst.

- Igas korpuses asuvas dokumendis on teemad esitatud erineval määral. See tähendab, et üks dokument võib korraga kuuluda kahe või rohkema teema alla, aga erineva tõenäosusega (nt 10% tõenäosusega esimese, 25% teise ja 65% kolmanda teema alla).

Teema on seega tõenäosusjaotus teatud sõnade vahel (Blei, 2012b). Teemade jaotamisel on aluseks võetud dokumentides esinevad sõnad, kuid ühe teema alla kuuluvad kõik sõnad ei pruugi esineda ühes dokumendis koos. Seega koonduvad teemade moodustamisel valitud sõnad teatud tõenäosusega konkreetse teema alla. Teemade sõnastamist mudel ise ei paku. Mudeli koostajal tuleb LDA-mudeli treenimisel ette anda nii teemade arv kui ka lõpptulemusena sõnastada teemad neis koosesinevate tõenäosuslike sõnade alusel. Oluline on ka see, et LDA-mudeli puhul ei pea olema vaadeldavad dokumendid eelnevalt (teemade järgi) märgendatud.

LDA põhimõtte selgitamiseks defineeritakse esiteks järgmised mõisted (Blei jt, 2003: 995):

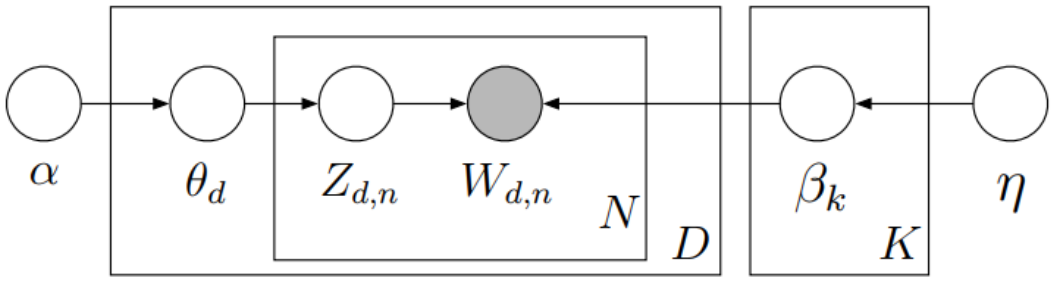
- **sõna** *w* on peamine uurimisalune üksus, mis on korpusele vastavas sõnastikus indekseeritud $\{1,\ldots,V\}$;
- **dokument**, mis koosneb N arvust sõnadest, on tähistatud $d = (w_1, w_2,\ldots, w_N)$, kus w_n märgib sõna kohal *n*;
- **Korpus** on kogum M arvust dokumentidest ning on tähistatud $D = \{d_1, d_2,\ldots d_M\}$.

Blei järgi on dokumendis esinevad sõnad vaadeldavad tunnused (*observed variables*), samal ajal kui teemade struktuur koosneb peidetud tunnustest. LDA-mudeli ülesandeks on jõuda teemade kui peidetud tunnusteni generatiivse protsessi kaudu, kasutades aposterioorset jaotust (*posterior distribution*). (Blei, 2012a)

LDA graafiline representatsioon on kujutatud joonisel 1 koos vastava märgistusega:

- D – dokumentide kogum,
- N – sõnade arv dokumendis *d*,
- K – teemade arv kogu korpuse kohta,
- θ_d – teemade proportsioon dokumendis *d*,
- Z_{dn} – sõnale *n* dokumendis *d* omistatud teema,
- W_{dn} – sõna *n* dokumendis *d* ehk vaadeldav tunnus,
- β_k – sõnade jaotus teemas *k*,
- α – dokumendis esindatud teemade proportsiooni eeljaotus (Dirichlet' parameeter),

- η – teemade eeljaotus sõnade kohta (Dirichlet' parameeter).



Joonis 1. LDA graafiline representatsioon (Blei, 2012a: 81).

Joonis 1 kujutab LDA generatiivset protsessi. Halliga varjutatult on märgitud tipp (*node*), mis tähistab vaadeldavat üksust ehk sõna (W_{dn}). Varjutamata ehk tühjad ringid märgivad peidetud tippe, millest moodustub teemade struktuur. Ristkülikutega piiratult on kujutatud generatiivse protsessi osad, mida mudeldamise käigus tsükliliselt korratakse. Olulised on ka LDA sõltuvuslikud seosed (joonisel märgitud nooltega). Näiteks teema omistamine dokumendis esinevale sõnale (Z_{dn}) sõltub teemade jaotusest dokumendis (θ_d). Samuti sõltub vaadeldav konkreetne sõna (W_{dn}) nii omistatud teemast (Z) kui ka kõikidest teemadest (β_K). (Blei, 2012a)

Generatiivse protsessi kaudu üritatakse etteantud dokumentide korpuse põhjal vastata küsimusele: „Milline võiks olla teemade peidetud struktuur, mis on genereerinud vaadeldavad dokumendid?" Selleks võetakse aluseks mudelile etteantud teemade arv K ja valitakse sõnade tõenäosused iga teema kohta (β_k). Seejärel võetakse aluseks dokumentide arv korpuses ning valitakse teemade jaotus dokumendis (θ_d), kirjeldamaks neid teemasid, mis antud dokumenti tõenäosuslikult iseloomustavad. Viimasena valitakse iga sõna kohta igas dokumendis (W_{dn}) kuuluvus teema alla (Z_{dn}). (Blei 2012b) Protsessi tulemusena saame kirjeldada teemasid neid iseloomustavate tõenäosuslike sõnade alusel, samuti vaadelda dokumendi tõenäosuslikku kuuluvust (erinevate) teemade all.

Lihtsaks näiteks siinkohal oleks taaskord ajalehe artiklites peituvate teemade analüüs, kus LDA mudel võiks valitud artiklites esinevate sõnade põhjal moodustada kolm teemakategooriat, mille analüüsi läbiviija saab mõtestada ja vastavalt uurimisprobleemile

pealkirjastada (vt joonis 2). Vastavalt toodud näitele on sõnastatud teemadeks: sise-, spordi- ja reisiuudised.

Joonis 2. Teemakategooriate näide.

Olgu meil fiktiivne näitekorpusest valitud artikkel pealkirjaga „Peaminister sõitis perega suusareisile“. LDA-põhise teemade modelleerimisel jaotuks see artikkel teatud osakaalude alusel kolme teema vahel (nt 50% reisiuudis, 30% spordiuudis ja 20% siseuudis). Samal ajal kui artikkel „Jalgpalli maailmameister on Basiilia“ paigutuks peaaegu 100%-liselt spordiuudiste alla.

3.3 LDA-mudeli kasutamine enesekohaste tekstide analüüsimisel

Käesolevas magistritöös on peamise tekstikaeve meetodina kasutatud LDA-d kui tõenäosusliku teema modelleerimise algoritmi mitmel olulisel põhjusel:

- Selle abil on võimalik leida dokumentides peidetud mustreid.
- Seda on võimalik teostada juhendamata meetodil ja eelnevalt märgendamata korpusel.
- Klastrite arvu käsitsi määramine pole probleemiks, kuivõrd PIST-testi teoreetilised alused lubavad eeldada klastrite arvu, mis jääb alla kümne (vastavalt suure viisiku teooriale viis klastrit või PIST-testi struktuurist tulenevalt kaheksa omadust).

Automaatselt teostatava teema modelleerimise meetodi kasuks otsustati käesoleva magistritöö puhul ka seetõttu, et sellel on mitmeid sarnasusi sotsiaalteadustes kasutusel oleva kontentanalüüsiga (Kalmus, 2015), mis oleks sellise tekstikorpuse analüüsimiseks üsna tavapärane lähenemine. Sisuliselt on kontentanalüüs teksti kodeerimine teatud kokkulepitud reeglistiku alusel. Samas on n-ö käsitsi tehtava kontentanalüüsi näol tegu väga töömahuka meetodiga, mis nõuab tööga alustamiseks ka konkreetset ja üheselt mõistetavat juhendit ning teostamiseks mitut kodeerijat. Teema modelleerimise sõnade esinemissageduse põhine lähenemine läheb kokku ka kontentanalüüsi põhimõttega kirjeldada teksti eksplitsiitset sisu

ehk sisu, mida on sõnaselgelt väljendatud, mitte aga ridade vahelt välja loetud (Kalmus, 2015).

LDA-mudeli looja David Blei järgi annab mudel raamistiku, mille abil uurida ja analüüsida tekste ilma, et me peaksime teemad enne ise määratlema või dokumenti nende alusel kodeerima asuma. Mudeli tugevuseks on selle algoritm, mis võimaldab peidetud struktuuride kaudu uurida ja mõista paremini tekstikorpust. (Blei, 2012b) Käesolevas magistritöös on huvi pakkunud hinnata LDA-mudeli võimalusi väiksemal tekstikorpuse uurimisel.

3.4 Isiksusekaeve ja selle ohud

Isiksusekaeve (*personality mining*)[5] on seotud tekstikaevega selles osas, et seoseid isiksuseomadustega on võimalik leida tekstianalüüsi kaudu. Isiksusekaeve ülesandeks on väljatöötatud mudeli alusel ennustada inimese isiksuseomadusi või isiksustüübilist kuuluvust. Interneti ja sotsiaalmeedia leviku, suurandmete kasutusele võtmise ning arvutite infotöötluse mahtude suurenemisega on uurijate huviorbiiti tõusnud ka isiksuse uurimine digitaalselt salvestunud andmete kaudu. Inimeste käitumist sotsiaalmeedia ja Interneti keskkonnas laiemalt saab analüüsida nii erinevate infohulkade talletumise (klikid, *like*'id jms), tekstiliste andmete (postitused, kommentaarid, arvustused) kui ka erinevate metaandmete kaudu (aktiivsusnäitajad jms). Eelkõige pakub sellist laadi info huvi ettevõtetele, kes soovivad inimestele edastada oma reklaami võimalikult personaalselt mõjuvana. Sotsiaalmeedia kanalite kaudu isiksusekaevele lähenemine pakub aga uurimuslikult põnevat sissevaadet inimeste käitumismustritele digitaalses kontekstis (Friedman ja Schustack, 2016: 36).

Isiksusekaevet on kasutatud erinevates uurimustes. Näiteks on isiksusekaevet rakendatud uurimaks õpilaste hakkama saamist e-kursustel ja kohandamaks e-õppekeskkondi vastavalt õpilaste isiksustüüpidele või kognitiivsetele stiilidele (vt Zakrzewska, 2009; Jovanovic jt, 2012). Isiksusekaeve üks kasutusvaldkondi suhestatuna psühholoogiaga on stressi ja depressiooniilmingute jälgimine inimestel (Coppersmith jt, 2014; De Choudhury jt, 2013). Samuti palju uurimusi on isiksuseomaduste ja sotsiaalmeedia andmete, sh tekstiliste andmete vaheliste seoste leidmiseks (Schwartz jt 2013, Jaidka jt, 2018).

[5] Kasutatakse ka mõistet isiksuse ennustamine (*personality prediction*).

Lisaks on võimalik tuua mitmeid näiteid isiksusekaeve põhimõtteid kasutavatest ärimaailma toodetest:

- IBMi loodud Watson™ platvorm võimaldab Personlity Insights teenusena[6] kasutada API-t, mis aitab analüüsida struktureerimata teksti ning selle põhjal koostada isiksusele vastav kirjeldus koos inimese isiksuseomaduste jaotuse ja tarbimiseelistustega.
- HR-valdkonna näide on ettevõttelt Textio[7], kes aitab tekstide põhjal ennustada tööotsijate ja ettevõtete sobivust ning pakub tekstide kirjutamise ajal analüütikat sobilikemate sõnade kasutamiseks.
- Crystal[8] analüüsib avalike andmete alusel inimese kuuluvust DISC-profiili alla ning annab mh soovitusi, mis sõnastuse ja võtetega sellele inimesele kirja teel läheneda.

Kindlasti on tekstikaeve ja isiksuseomaduste analüüsi ühendavaid teenused oma haaret turul laiendamas, kuna juba praegu toimub suur osa inimeste tarbimisest ja suhtlusest *online*-keskkonnas. Kuivõrd masinõppe meetoditel on palju huvitavaid uurimuslikke kasutuskohti, on võimalik andmetega ka pahatahtlikel eesmärkidel manipuleerida ja sekkuda inimeste otsustusprotsessi, nagu näitas 2018. aasta alguses ilmsiks tulnud Facebooki ja Cambridge Analytica skandaalne juhtum (Cadwalladr ja Graham-Harrison, 2018), mille kohaselt lekkisid üle 50 miljoni Facebooki kasutaja andmed, kellest osade kohta olid olemas ka nende isiksustesti tulemused. Ühtlasi ilmnes, et nendest andmetest lähtuvalt mõjutati 2016. aastal Ameerika Ühendriikide presidendi valimisi. Sellest tulenevalt on oluline rõhutada, et isiksus on indiviididele taandatuna personaalne ja tundlik teema ning selle isiksuse uurimine mistahes vahendi või keskkonna kaudu nõuab eetilist lähenemist. Igasugused manipulatsioonid võivad huvi pakkuda eksperimenteerimiseks, kuid eksperimentidel võivad olla ka tõsised ja pöördumatud tagajärjed.

[6] Vt teenuse demolehte: https://personality-insights-demo.ng.bluemix.net/.
[7] https://textio.com/products/
[8] https://www.crystalknows.com/about

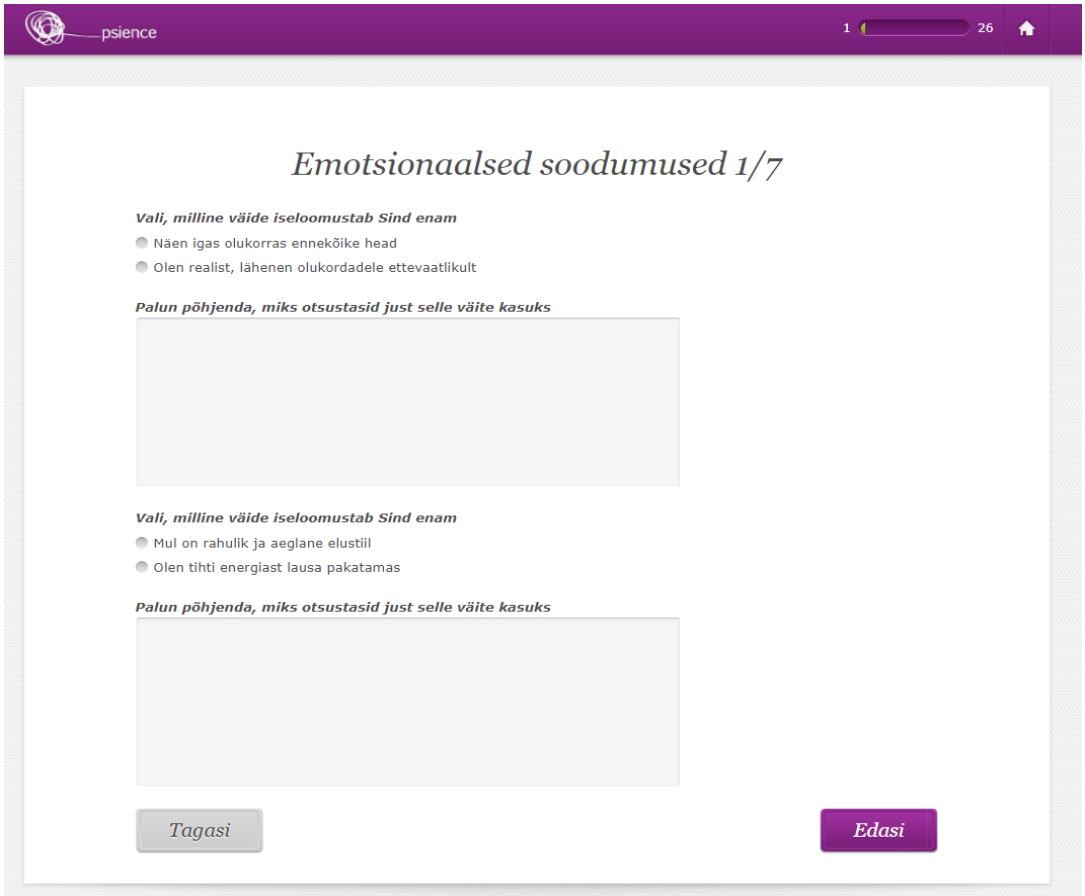
4. PIST-test ja andmed

Järgnevalt kirjeldatakse analüütilise osa aluseks olevaid andmeid, nende kogumist ja seda, kuidas on tagatud vastajate anonüümsus.

4.1 PIST-testi kirjeldus

Nagu eelpool kirjeldatud, on Psience'i isiksuslike soodumuste test ehk PIST-test ettevõtte Psience intellektuaalomand ning see on välja töötatud kasutamiseks ettevõtte teenusena koos sinna juurde kuuluva analüüsiga. PIST-testi täitmiseks suunatakse vastaja PsienceReports veebikeskkonda.

Test koosneb kahest struktuursest osast. Esimeses osas peab testi täitja valima etteantud väitepaari seast end enim iseloomustava väite kasuks ning tal on võimalus väidete valikut ka avatud kommentaariga põhjendada (vt Joonis 1). Näiteks optimismi mõõtva väitepaari alla kuuluvad kaks väidet: 1) „Näen igas olukorras ennekõike head"; 2) „Olen realist, lähenen olukordadele ettevaatlikult". Seda, millist omadust mis väitepaar mõõdab, testi vastaja ise ei näe.



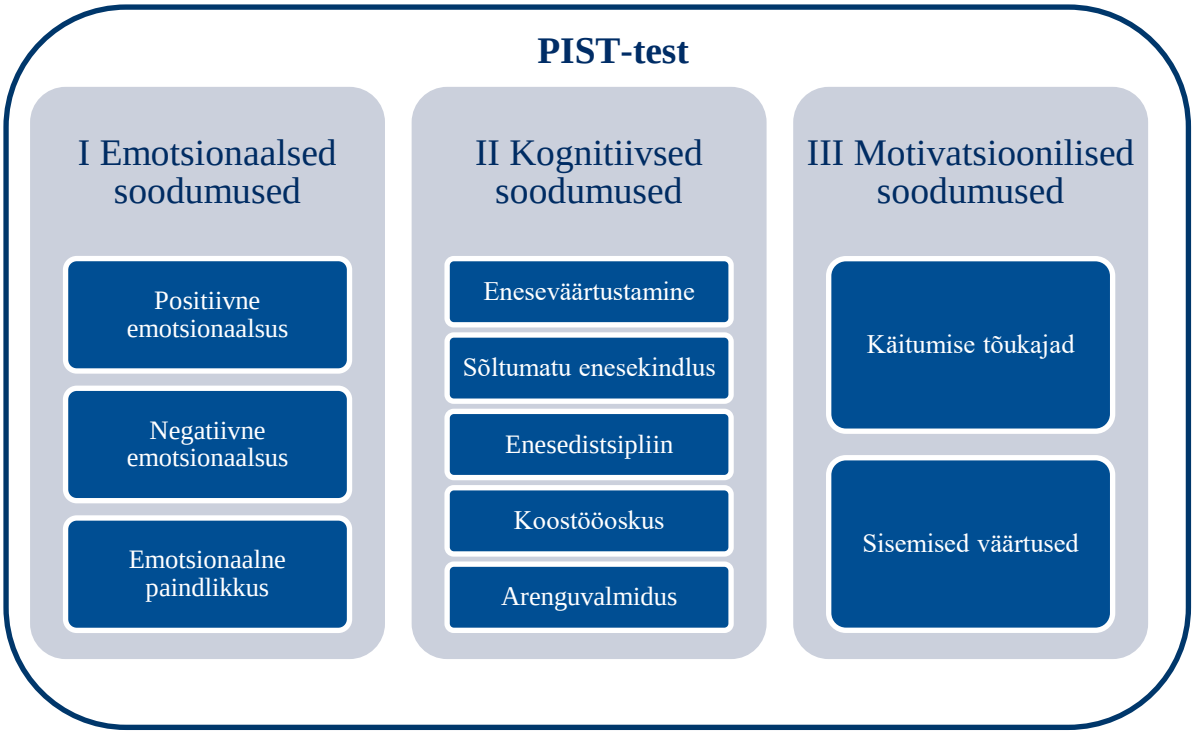
Joonis 3. Kuvatõmmis PIST-testi 1. osast – väidete paarid.

Testi teises osas tuleb erinevaid väiteid hinnata 5-pallisel skaalal, kusjuures osad skaalaväited on esitatud pööratud kujul, nt optimismi pööratud väide on testis esitatud järgmiselt: „Ma ei loodagi, et asjad lähevad minu tahtmist mööda“. Skaalaväidete juures saab vastaja soovi korral samuti oma valikut avatult kommenteerida (vt Joonis 2).

Joonis 4. Kuvatõmmis PIST-testi 2. osast – skaalaväited.

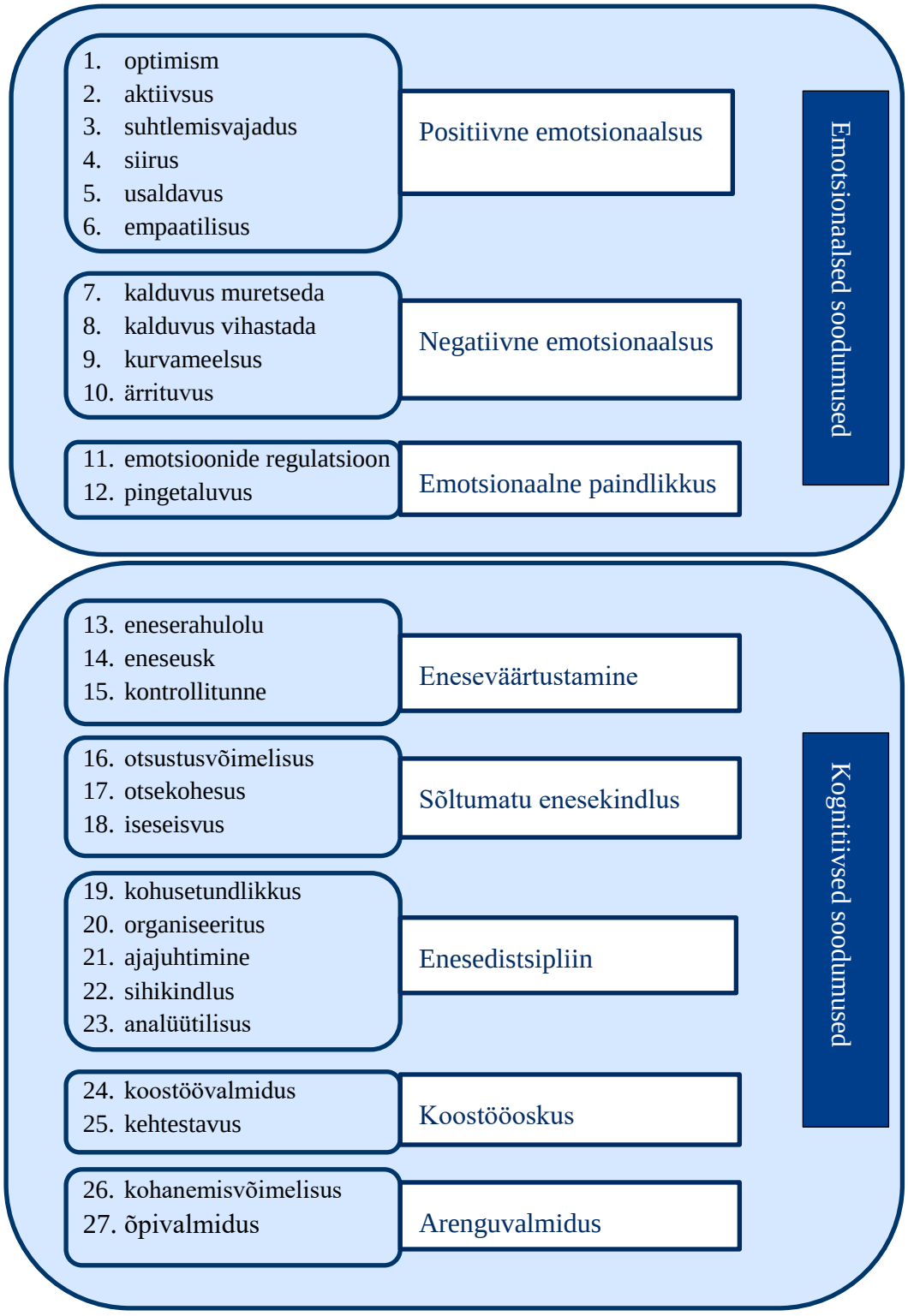
Testi tulemused saab testi analüüsija (Psience'i töötaja) PsienceReports veebikeskkonnast XML-failina alla laadida. Veebist alla laetud toorandmed lisatakse analüüsifaili, mille abil arvutatakse välja testi skoorid. Testi tulemused esitatakse veebiraporti kujul nii testi täitjale kui ka selle ettevõtte esindajale, kelle juurde testi täitnu on avaldanud soovi tööle kandideerida. Oluline on siinkohal märkida, et täidetud testi põhjal arvutatakse tulemus ainult väidete valiku ja skaalapõhiste väidete alusel, mis on testi loogikale vastavalt talletatud numbriliste andmetena. Avatud vastuseid ehk enesekohaseid tekste ei võeta numbriliste kaaludena arvesse erinevate isiksuslike omaduste avaldumise skoorides, vaid neid kasutakse tekstiraportis täiendava informatsiooni välja toomiseks vabatekstina. Avatud kommentaaride analüüs on osutunud PIST-testide hindamise juures kõige ajamahukamaks protsessiks ning seda oli seni tehtud ainult manuaalselt.

PIST-testi sisuline loogika on ülesehitatud erinevaid omadusi mõõtvatele kolmele suuremale moodulile: emotsionaalsed soodumused, kognitiivsed soodumused ja motivatsioonilised soodumused, mis omakorda jagunevad spetsiifilisteks alammooduliteks (vt joonis 3).



Joonis 5. PIST-testi sisuline struktuur.

Antud magistritöös keskendutakse PIST-testi kontekstis peamiselt emotsionaalsete ja kognitiivsete soodumuste alla kuuluvate omadusrühmadega (kokku kaheksa omadust), kuna nende kohta on andmestikus vaba teksti kujul andmed. Joonisel 3 välja toodud kolm moodulit jagunevad omakorda nn teise tasandi kirjeldavateks omadusteks. Kokku on vastavaid omadusi 27 (vt joonis 4).



Joonis 6. PIST-testi tekstianalüüsi all vaadeldavad omadused ja nende sisuline jaotumus.

4.2 Andmete kirjeldus

Käesoleva analüüsi aluseks on 119 inimese PIST-testi tulemusskoorid ja enesekohased tekstid, mis on antud avatud vastustena. Vastanute seast 37 on mehed ja 82 naised. Andmed on kogutud perioodil september 2013 kuni mai 2018. Testi on täidetud peamiselt kahes kontekstis: 1) tööle kandideerimisel – testi on täitnud inimesed, kes on olnud motiveeritud kandideerimaks uuele ametikohale; 2) karjäärinõustamisel – inimene ise on soovinud karjäärivalikute tegemisel enda isiksuseomaduste kohta rohkem teada saada. Oluline on siinkohal välja tuua ka see, et ettevõtte tavapärases praktikas täidavad PIST-testi just eelkõige lõppvooru kandidaadid, nt konkursi viis tugevaimat kandidaati, kelle vahel tehakse lõplik valik tööle võtmiseks.
Andmete kasutamiseks käesoleva magistritöö analüüsis küsiti luba testi täitjatelt e-posti vahendusel. Oma andmete kasutamist ei lubanud üks inimene, kelle vastused käesoleva töö andmestikust ka välja jäeti.

Kõik analüüsitavad tekstid on eestikeelsed. Kogu analüüsitava tekstikorpuse mahuks on 59 738 sõna. Kuigi n-ö käsitsi tehtavaks kontentanalüüsiks on selline tekstikorpus üsna mahukas[9], on masinõppe mudelite treenimise seisukohalt selline tekstihulk siiski võrdlemisi väike ning antud töö analüüsi tulemustest on seda ka näha.

PIST-testi vabavastuste analüüsimisel on võetud eeldusteks järgmised aspektid:

- Vastajad on kirjeldanud end adekvaatselt (pole esitlenud end kellegi teisena).
- Testi täitmise olukord on loonud eeldused, kus inimene on motiveeritud ennast kirjeldama lähtuvalt sisemisest analüüsist: tööle kandideerimisel pole põhjust kirjeldada end liiga tagasihoidlikult, kuid samas pole mõtet peita endale väga omaseid jooni, mis nagunii välja tuleksid või mida soovitajad võiksid inimese kohta esmajoones öelda.

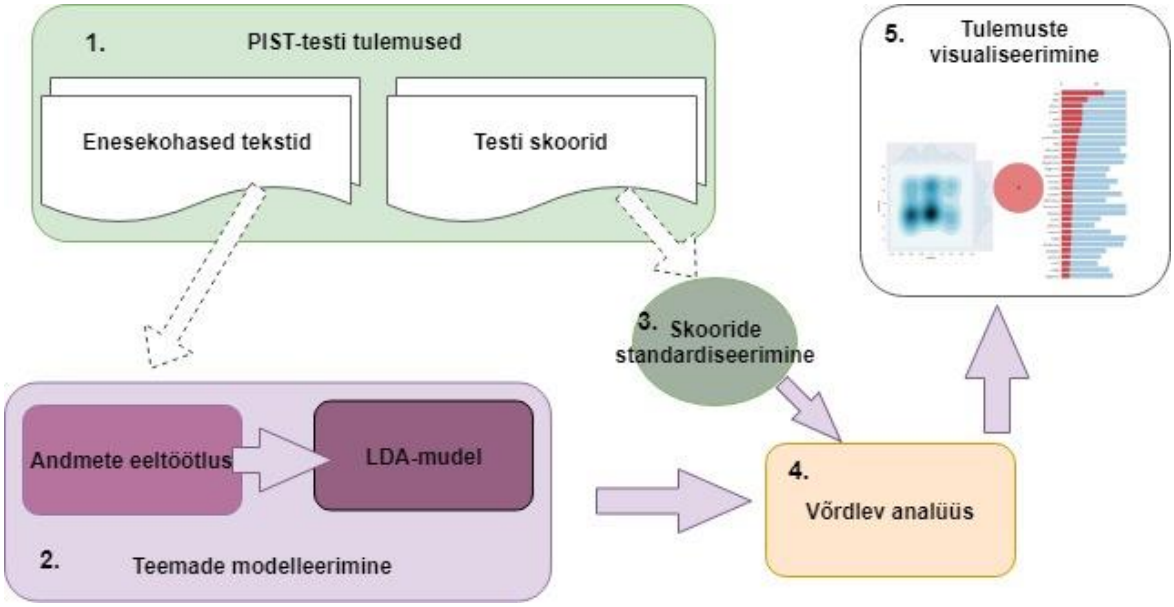
Lisaks PIST-testi enesekohastele tekstidele kasutatakse võrdleva ülevaate saamiseks ka testi numbrilisi skoore. Skoorid on avaldatud kaheksa testiga mõõdetava omaduse kohta: positiivne emotsionaalsus, negatiivne emotsionaalsus, emotsionaalne paindlikkus, eneseväärtustamine, enesedistsipliin, sõltumatu enesekindlus, koostööoskus, arenguvalmidus (vt Joonis 6). Iga omaduse maksimumskoor on 100, miinimum 0 punkti.

[9] Võrdluseks: tavalises Wordi dokumendis oleks antud tekstikorpuse maht *ca* 150 lehekülge (Times New Roman, 12 p, reavahe 1,5).

5. Töövoog

Antud uurimuses on valitud analüüsimise aluseks probleemiks, kuidas tekste ehk dokumente automaatselt teemadeks ehk klastriteks liigitada. Käesoleva uurimuse kontekstis mõistetakse mõiste „dokument“ all ühe inimese kokku koondatud avatud vastuseid. Seega vastab ühele dokumendile kõik ühe inimese PIST-testi täitmisel kirjutatud avatud vastused.

Koodi kirjutamisel on kasutatud Pythoni versiooni 3.5.5 ning kood vormistati Jupyter Notebook'i keskkonnas[10]. Andmete eeltöötluses on kasutatud ka Pythoni eesti keelele tugineva loomuliku keele töötluse EstNLTK teeki versiooniga 1.4.1 (Orasmaa jt, 2016). LDA-mudeli loomisel kasutati Pythoni tarkvarapaketti gensim (Řehůřek ja Sojka, 2010). Käesoleva töö aluseks olev kood on kättesaadav ka GitHubi repositooriumis[11]. Järgnevalt on kirjeldatud käesoleva analüüsi etappe vastavalt joonisel 7 toodud nummerdusele.



Joonis 7. Töövoo skeem.

1. PIST-testi tulemused

PIST-testi tulemused laaditi alla testi veebikeskkonnast PsienceReports. Allalaetud failist eemaldati vastajate isikuandmed ning koostati kaks eraldi CSV-faili: 1) ainult tekstiliste andmetega (enesekohased tekstid) fail, kus üks rida vastas ühele inimesele; 2) punktskooridega fail, kus üks rida vastas ühele inimesele.

[10] http://jupyter.org/
[11] https://github.com/lisetmarleen/textmining

2. Teemade modelleerimine

2.1. Andmete eeltöötlus

Tekstianalüüs teostati teemade modelleerimise meetodil, kasutades tõenäosusliku teemade modelleerimisele põhinevat LDA-mudelit. Enne LDA-mudeli loomist teostati andmete eeltöötlus, sh moodustati andmefaili ühest reast üks dokument. Eeltöötlus on eelkõige vajalik selleks, et masinõppe mudelil oleks võimalik tulemuslikumalt tekstilistest andmetest mustreid leida. Selleks lihtsustatakse teksti teatud võtete alusel. Käesoleva analüüsi eeltöötluse osadeks olid järgnevad etapid, mida korrati igas dokumendis:

- numbrite eemaldamine,
- muudeti kõik tähemärgid väiketähtedeks (*lowercase*),
- etteantud sõnastiku alusel teatud sõnade asendamine (nt levinumad kirjavead „vüi“ asendati „või“, ühtlustati erinevad kirjapildid „aegajalt“ asendati „aeg-ajalt“ jms),
- sõnad lemmatiseeriti ehk viidi algvormi, kasutades EstNLTK `Text` klassi muutujat 'lemmas'. Lemmatiseerimise tulemusel saame näiteks käändsõnast „arvamustel“ sõna „arvamus“.
- Lemmatiseerimise tulemusel pakuti osade sõnade algvormina kaht varianti, mis olid eraldatud püstkriipsuga (nt „kurvastama|kurvastanud“). Selle olukorra lahendamiseks jäeti korpusesse ainult püstkriipsust vasakul pool olev sõna (*ma*-tegevusnimi).
- Lisaks eemaldati tekstidest stopp-sõnad. Stopp-sõnadeks on sellised sõnad, mida tekstides esineb sageli, kuid mis annavad vähe infot teksti olemuse või sisu kohta. Tavapäraselt on stopp-sõnadeks keeles laialt kasutatud sõnad nagu sidesõnad („ja“, „ehk“, „kui“, „kuid“, „aga“), asesõnad („mina“, „see“), lühendid („n-ö“, „jne“, „jms“). Tulenevalt teksti kontekstist võidakse eemaldada ka vaadeldavates tekstides levinumad sõnad, näiteks antud analüüsis on stopp-sõnana eemaldatud ka sõna „olukord“[12], mis esines kõikides klastrites sageduselt esimesel või teisel kohal.

Teksti eeltöötluse tulemusel muudeti esialgne tekst:

Valisin koostöö punkti, kuna olemuslikult mulle meeldib suhelda, teemasid meeskonnas põhjalikult arutada ja ühiste eesmärkide nimel töötada. Samas mulle meeldib ja sobib ka üksi töötamine, eriti süvenemist vajavate ülesannete puhul. Minu meelest on koostöö oluline, kuna selle läbi sünnivad huvitavad ja väärtuslikumad otsused ja tulemused.

[12] Sõna „olukord“ sagedasemad kasutused PIST-testi enesekohastes testides olid näiteks selle sõna esinemine lauses vastusena väitevalikule: „oleneb/sõltub olukorrast“, aga ka erinevate olukordade kirjeldustes.

kompaktsemaks:

valima koostöö punkt olemuslikult meeldima suhtlema teema meeskond põhjalikult arutama ühine eesmärk nimel töötama meeldima sobima üksi töötamine eriti süvenemine vajav ülesanne puhul meelest koostöö oluline läbi sündima huvitav väärtuslikum otsus tulemus.

LDA-mudeli loomiseks viidi iga dokument esmalt järjendi kujule ning seejärel moodustati Gensim teegi abil bigrammid ehk lemmapaarid, mis asetsevad korpuses kõrvuti ning mida sellisel kujul esineb terves korpuses vähemalt neli korda. Bigrammide moodustamisega sooviti leevendada LDA *bag-of-word* aluspõhimõtet, mis ei arvesta sõnade järjekorraga ning mille tõttu kannatab sõnade konteksti uurimine. Näiteks oli moodustatud bigrammide seas lemmapaarid: „usaldama_kontroll“, „paha_tuju“, „eesmärk_saavutamine“ jms. Küll aga ei jõudnud ükski bigramm teemasid iseloomustavate sõnaloendite alla (vt ptk 6.1).

Keskmine sõnade arv ühe dokumendi kohta enne tekstide eeltöötlust oli 502 sõna. Keskmine sõnade arv ühe dokumendi kohta pärast tekstide eeltöötlust oli 282 sõna.

2.2. LDA mudeli loomine

Olemasolevate tekstide alusel loodi esmalt gensimi `Dictionary()` funktsiooni abil sõnastik, kus igale unikaalsele sõnale määratakse unikaalne numbriline indeks (id). Sõnastiku abil luuakse dokumentide järjend ehk koprus dokument-termin maatriksina. Korpus salvestati *bag-of-words* kujul. Iga järjendi kujul olev dokument-vektor sisaldab loendit ennikutest, kus moodustuvad paarid: sõna id (sõnastiku alusel) ja sõna sagedus antud dokumendis. Oluline on see, et dokumendi vektoris sisalduvad ainult need sõnad/terminid, mis vastavas dokumendis esinevad. Edasi loodi tarkvarapaketi gensim abil LDA-mudel dokument-termini maatriksi, sõnastiku ja etteantud teemade ehk klastrite arvu alusel (antud analüüsis on valitud klastreid kolm). LDA-mudeli paremaks interpretatsiooniks visualiseeriti tulemused LDAvis teegi abil (vt joonised 8–11).

3. Testi skooride standardiseerimine

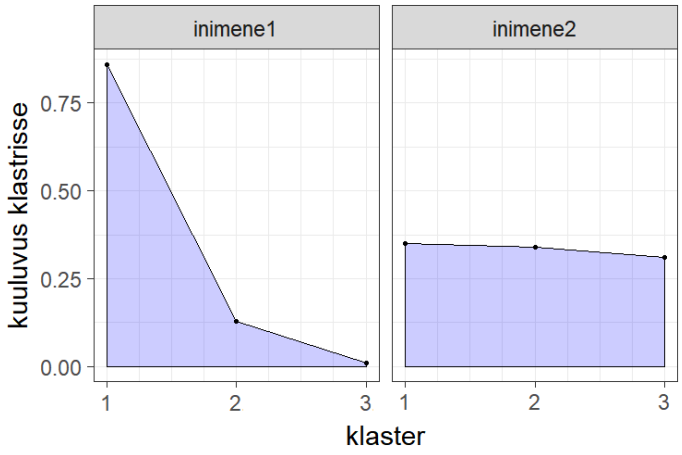
Kuna käesoleva analüüsi eesmärgiks oli võrrelda tekstianalüüsi tulemusi testi numbriliste skooridega, tuli erinevate inimeste testitulemuste paremaks võrdluseks skoorid esmalt standardiseerida. Skoorid standardiseeriti nii ühe tunnuse piires kui ka ühe vastaja piires. Standardiseerimisel ühe tunnuse piires võeti aluseks tunnuse (nt positiivne emotsionaalsuse skoorid ühes veerus) miinimum ja maksimumväärtused ning iga väärtus standardiseeriti järgneva valemi (1) alusel:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1),$$

kus z_i märgib standardiseeritud väärtust ja x_i vaadeldavat väärtust kohal i. Standardiseerimise tulemusena teisendati väärtused vahemikku 0–1. Seejärel summeeriti ühe vastaja piires skoorid ja jagati iga skoor tema üldsummaga. Standardiseeritud skooride alusel on võimalik vastajaid omavahel võrrelda võrdväärsemalt.

4. Võrdleva analüüsi teostamine

Tekstianalüüsi ja standardiseeritud skooride võrdlusest parema ülevaate saamiseks kasutati Pythoni Seaborn teeki ning jooniste 15–16 kujutamiseks R-i. Võrdleva analüüsi teostamise lähtekohaks valiti suurima tõepära printsiip (Lepik, 2017: 47), mille järgi võeti iga vastaja kohta tema kõige suurema tõenäosusega klaster ning kõrgeim standardiseeritud omadusskoor. Siinkohal tuleb arvesse võtta, et selle printsiibi alusel valiku tegemine limiteerib analüüsi tulemusi selles osas, et kaotame infot inimese klastrisse tõenäosusliku kuulumise ja omadusskooride jaotuste kohta, kuna valik tehakse ainult maksimumnäitajate järgi. Allolev joonis 8 illustreerib kahe erineva inimese tõenäosuslikku kuuluvust klastritesse. Rakendades suurima tõepära printsiipi, valitakse nii inimene1 kui ka inimene2 puhul tema klastri kuuluvuseks klaster 1, sest seda iseloomustab kõrgeim tõenäosusnäitaja (vastavalt 0,86 ja 0,35 punkti). Samas näeme, et inimene2 puhul on klastritesse kuuluvuse tõenäosus palju ühtlasemalt jagunenud kui inimene1 puhul, kelle 1. klastri kuuluvus eristub selgelt. Seega suurima tõepära printsiibi kasutamisel peame arvestama, et see vähendab inimese erisuste väljatoomist.



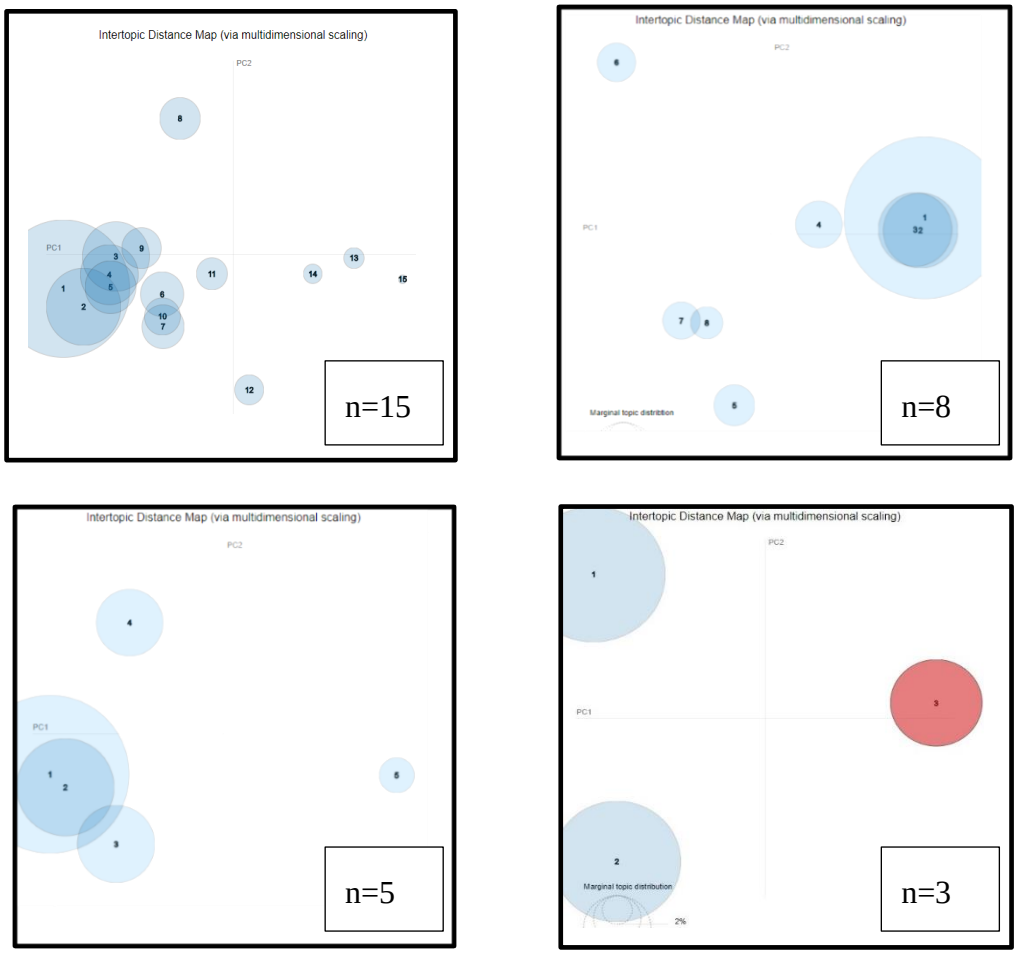
Joonis 8. Näide kahe erineva inimese klastrite kuuluvuse jaotusest.

Lähemalt tutvustatakse võrdleva analüüsi tulemusi järgmises peatükis.

6. Analüüs

LDA masinõppel põhinevat mudelit kasutakse juhtudel, kui soovitakse tekstikorpusesse võetud dokumentide struktuuri uurida, sh leida ja mõtestada dokumentides olevate sõnade esinemissageduse põhjal tekstide peidetud mustreid. Kuivõrd käesoleva uurimuse puhul on vabateksti andmed seotud numbriliste testiskooridega, saab LDA-mudeli alusel loodud teemasid ehk klastreid kõrvutada ka PIST-testi isiksuseomaduste tulemustega. Kuna andmestikus representeerib iga dokument ühe inimese ehk testi täitja vastuseid, saab analüüsi tulemuste põhjal leida seoseid ka vastanute isiksuseomaduste vahel eeldusel, et sarnaseid inimesi iseloomustab ka sarnane sõnakasutus.

Käesolevas töös on LDA-mudel treenitud 119 dokumendiga 3 teema alusel. Teemade arv tuli LDA-mudeli treenimisel ise ette anda ning optimaalne teemade ehk klastrite arv leiti tulemuste visuaalsel analüüsil: kolm on suurim klastrite arv, kus on selgelt eristuvad teemad ega esine (täielikke) ülekatteid (vt joonis 9).



Joonis 9. Optimaalne klastrite ehk teemade arv leiti LDA-mudeli korduval treenimisel. Joonisel on toodud klastrite arvud (n). Optimaalseimaks osutus klastrite arv, kus n=3.

6.1 Tekstianalüüsi tulemused

Nagu eelnevalt kirjeldatud, koondati tekstianalüüsi teostamiseks ühe inimese kõik PIST-testi täitmisel antud avatud vastused ühte dokumenti. Seega saab analüüsifaasis dokumentide struktuuri ja sarnasusi üle kanda teatud mõttes ka inimeste sarnasusele, kuivõrd tekstide sisuks on testi täitjate enesekohane kirjeldus. Samas, kuna andmete eeltöötluse käigus on palju konteksti andvaid sõnu eemaldatud (sh erituse vormid), tuleb tekstianalüüsi põhjal tehtud üldistuste tegemisel olla ettevaatlik. Näiteks fraas „Alati see ei õnnestu“ on eeltöötluse tulemusena eraldi sõnade „alati“ ja „õnnestuma“ kujul. Järgnevalt kirjeldatakse LDA-mudeli abil teostatud ja teegi LDAvis abil kujutatud tekstianalüüsi tulemusi.

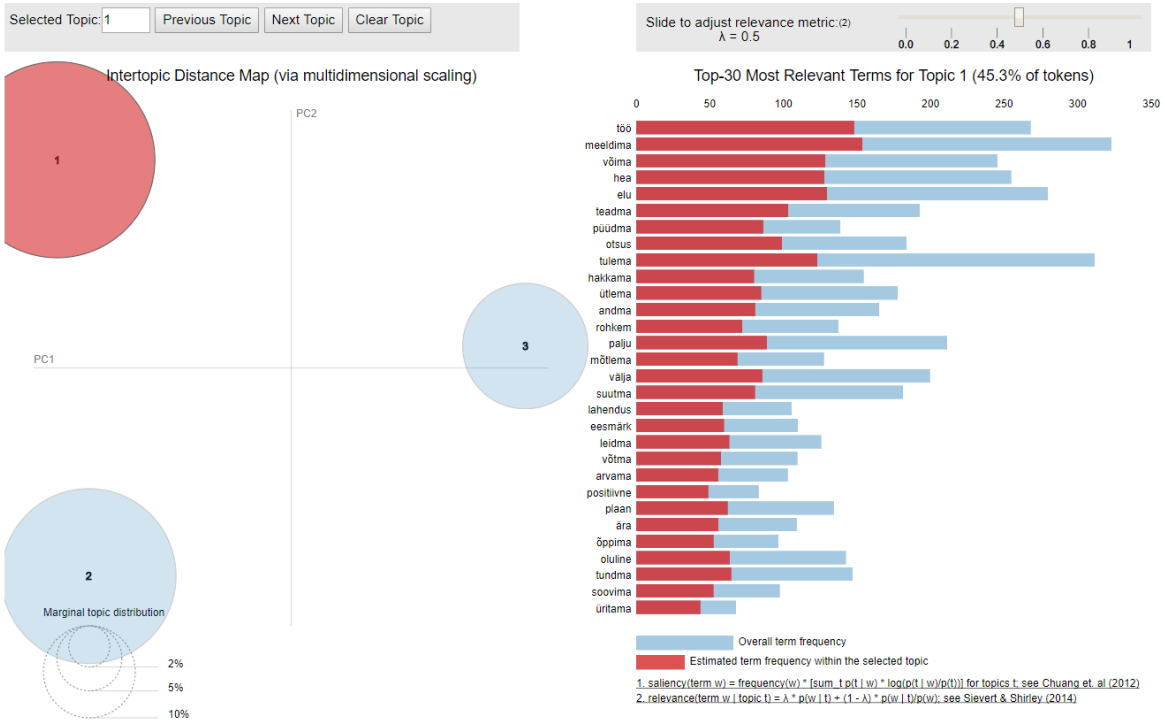
Alloleval joonistel 10–12 on kujutatud tekstianalüüsi tulemusel moodustatud kolme teemat ehk klastrit koos iga teema alla kuuluva 30 kõige relevantsema sõnaga (vt parema ülevaate saamiseks klastrite ja sõnade jaotust Lisa 1). Viidatud joonised on kuvatõmmised LDAvis abil loodud interaktiivsetest graafikutest.

Teemade modelleerimise tulemuseks on kolm eraldiseisvat teemat, mis ei ole teineteise suhtes ülekattega, kuid mis erinevad teineteisest suuruse poolest – neist kõige esimene teema on suurim ja kolmas väikseim. Vastava teema selekteerimisel värvub graafiku vasakul pool valitud teemat kujutav ring punaseks ning paremale poole moodustuvad horisontaalsed tulbad sõnadest, mis aitavad valitud teemat mõista. Sinisega kujutatud tulp näitab sõna esinemissagedust terves korpuses ning punane, ülekattev tulp näitab sama sõna esinemissagedust vastava teema all. Teemade sõnastamiseks on abiks relevantsus- (*relevance*) ehk olulisusnäitaja (lambda), mis on käesolevas analüüsis seadistatud väärtusele $\lambda=0.5$ (λ-parameetri suurus on vahemikus 0…1). See olulisusnäitaja on lisatud teeki LDAvis just sellel eesmärgil, et LDA-mudeli abil leitud tõenäosuslike sõnade kuuluvuse uurimise kõrval oleks lisavõimalus teemade paremaks interpreteerimiseks, kuivõrd tavapäraselt kipuvad korpuses enimkasutatud sõnad sattuma teema kõige tõenäolisemate sõnade sekka ega pruugi nii selgelt teemade lõikes eristuda (Sievert ja Shirley, 2014: 63–64). Kui lambda suurus on seadistatud väärtusele $\lambda=1$, siis järjestatakse graafikul paremal pool olevad sõnad selle alusel, kui suur on sõnade tõenäosused antud teema alla kuulumiseks. Vähendades lambda suurust, on võimalik järjestada sellele teemale iseloomulikumad sõnad, mitte puhtalt sõnade tõenäosuse kuuluvuse alusel antud teemasse, vaid sõna tõenäosuse antud teemas esinemise osas (punane tulp) jagatuna selle sõna üleüldise esinemissagedusega korpuses (sinine tulp). Ehk teisisõnu toimub sõnade järjestamine lambda suuruse vähendamisel punaste ja siniste

tulpade omavahelise suhte alusel. Seega võimaldab lambda suuruse vähendamine vähendada üle korpuse enam levinud sõnade sattumist teemat iseloomustavate sõnade sekka (mis raskendab üldjuhul teemade interpretatsiooni) ja selle asemel suurendada antud teemale iseloomulikemate ehk antud teemas sagedasemalt esinevate sõnade osakaalu. (Sievert ja Shirley, 2014)

Esimene teema – „eesmärgipärased tegutsejad“

Teema modelleerimise meetodi oluliseks etapiks on teemades esinevate sõnade põhjal sõnastada vastav teema. Seda etappi võiks ka nimetada pealkirjastamiseks, kuna sõnade põhjal antakse teemale üldistav nimetus. Kuna antud analüüsi aluseks on võrdlemisi väike tekstikorpus ning eri teemade all esines ka samu sõnu (näiteks „töö“, „meeldima“, „suutma“), mis pigem lähendasid teemasid, ei teinud see teemade sõnastamist lihtsamaks. Siiski on võimalik joonisel 10 kujutatud parempoolses tulbas esinevate sõnade pealt teha teema kohta teatavaid üldistusi (vt sõnade loendit ka Lisa 1).



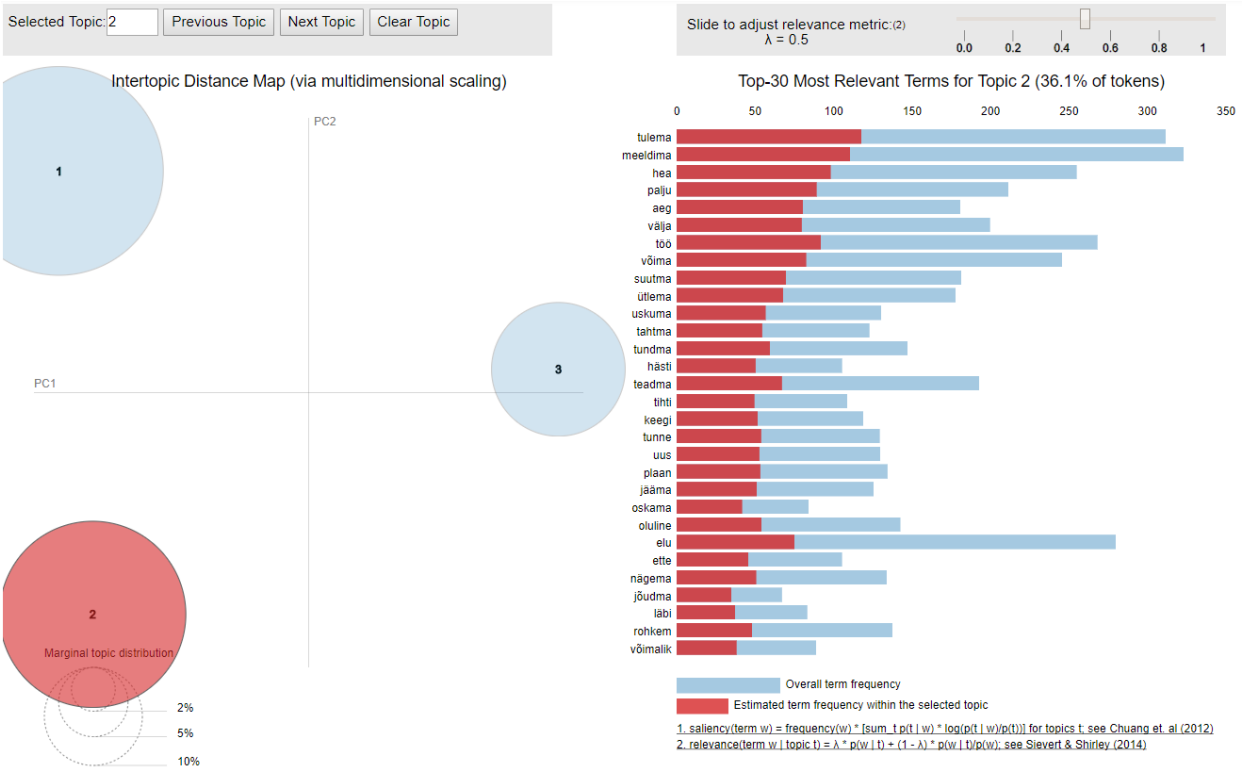
Joonis 10. LDAvis tulemused 1. teema kohta.

Esiteks paistab selle teema all silma tegusõnade võrdlemisi suur esindatus ja sealhulgas proaktiivseid tegevusi väljendavad tegusõnad nagu „tulema“, „võtma“, „andma“, „hakkama“, „õppima“ jne. Siit võib omakorda üldistada, et sellesse klastrisse kuuluvad inimesed on aktiivsemad tegutsejad. Samas nimisõnade loend: „eesmärk“, „plaan“, „töö“, „otsus“, „lahendus“ jätavad mulje eesmärgipärasusest, mistõttu on esimese teema nimetuseks

pakutud „eesmärgipärased tegutsejad“. Samuti ei saa märkamata jätta sõna „positiivsus“ esinemist antud teema all, mis võiks viidata optimistlikumale meelelaadile.

Teine teema – „sisekaemuslikud“

Ka selle teema all (vt joonis 11) paistab silma esinduslik valik tegusõnu, kuid erinevalt esimesest klastrist on siin proportsionaalselt rohkem esindatud seisundile ja eneseanalüüsile viitav tegusõnade komplekt nagu „tundma“, „jääma“, „tahtma“, „uskuma“, „teadma“ ning nende juurde sobituv nimisõna „tunne“. Samuti kumab sõnadest „uskuma“, „võimalik“, „ette“, „uus“ teatav tulevikku vaatav või veendumuslik suund, kuid selle üldistuse tegemisel tuleks olla pigem ettevaatlik, kuivõrd me ei tea nende sõnade tegelikku konteksti. Seega võrreldes esimese klastriga tunduvad teise klastri esindajad passiivsemad, ei kiirusta tegutsemisele, vaid on rohkem oma sisemaailmaga tegelevad ehk sisekaemuslikud inimesed.

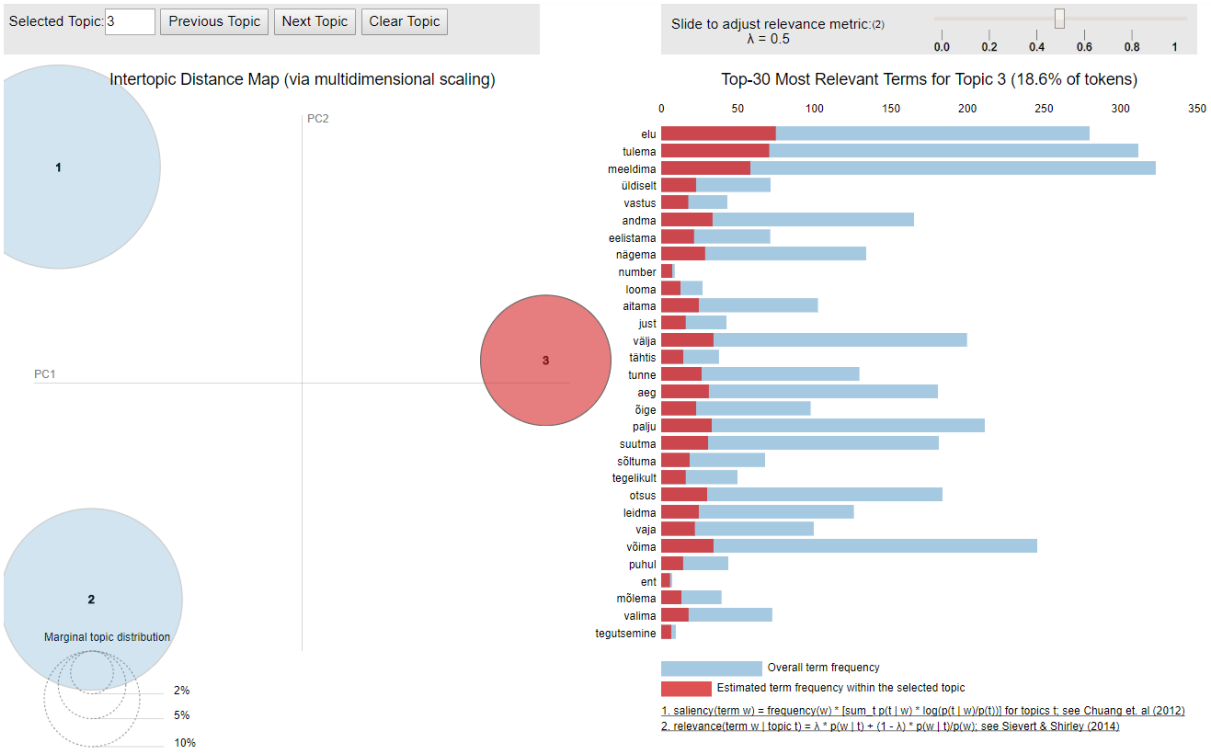


Joonis 11. LDAvis tulemused 2. teema kohta.

Kolmas teema – „paindlikud“

Kolmanda teema (vt joonis 12) sõnastamisel aitab kaasa analüüsi aluseks olevate tekstide konteksti parem tundmine. Kuivõrd avatud vastuseid antakse PIST-testi täitmisel olukorras, kus tuleb kahe iseloomustava väite vahel valik teha ja valikut kommenteerida, kumab mitmetes vastustes läbi vastaja kahevahel olek – võimalusel langetaks vastaja valiku mõlema väite kasuks, kuid valida tuleb üks väide. Sellised väite valimisel kahevahel olekut märkivad

sõnad, mis kolmanda teema all esinevad, võiksid olla näiteks „vastus“, „number“, „mõlema“, „valima“, „sõltuma“. Arvestada tuleb siin ka sellega, et stopp-sõnade eemaldamisega on võetud välja ase- ja sidesõnadega väljendatavad kahevahel olemise sõnad ja väljendid, nagu „nii ja naa“, „see ega teine“ jms. Lisaks sellele iseloomustab seda teemat suund koostööle sõnade „aitama“ ja „tegutsema“ läbi. Seega on teemat üldistavaks nimetuseks valitud „paindlikud“.



Joonis 12. LDAvis tulemused 3. teema kohta.

Tekstianalüüsi tulemusena saab üldistavalt teemade kohta kokku võtta järgmist: 1. teema koondab välimisi tegevusi (tavaliselt väljenduvad liikumises), nagu „tulema“, „võtma“, „hakkama“ ning 2. teema pigem sisemisi tegevusi, nagu „tundma“, „uskuma“. 3. teemale iseloomulikumad sõnad pakuvad PIST-testi kontekstis eelkõige huvi selle poolest, et peegeldab neid inimesi, kes testi täitmisel tõid välja teatud omaduste avaldumist erinevates olukordades või pehmendasid oma väite valikut, tuues välja oma kahevahel olekut. See annab võimaluse LDA-mudelit tulevikus täiustada selliselt, et teksti analüüsi tulemusi saab automaatselt arvesse võtta ka numbrilise skoori kaaluna.

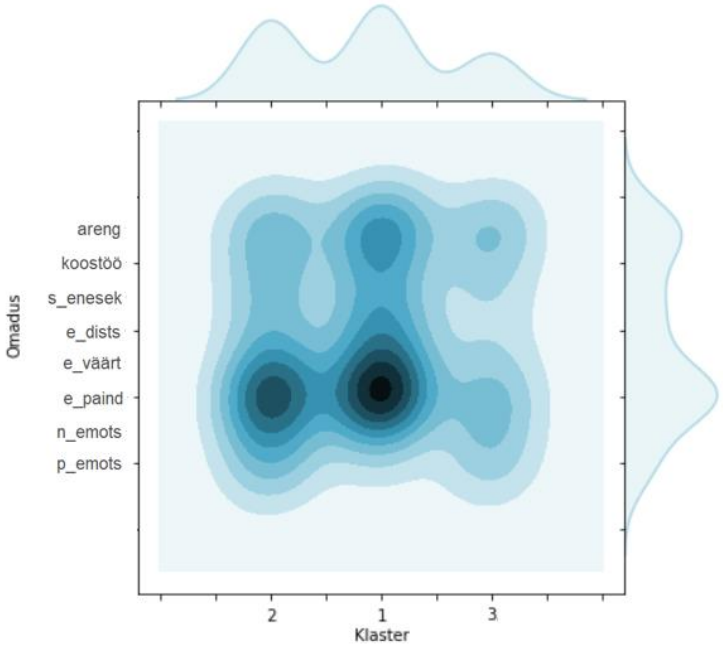
6.2 Võrdlustulemused

Teemade modelleerimise abil teostati PIST-testi avatud vastuste ehk tekstianalüüs. Samas esitatakse ettevõtte Psience tavapärases praktikas testi tulemused numbriliste skooridena ja

tekstilise osa tulemused skooridesse ei panusta. Järgevalt esitatakse tekstianalüüsi ja testiskooride võrdleva analüüsi tulemused vastajate kohta ülevaatlikult.

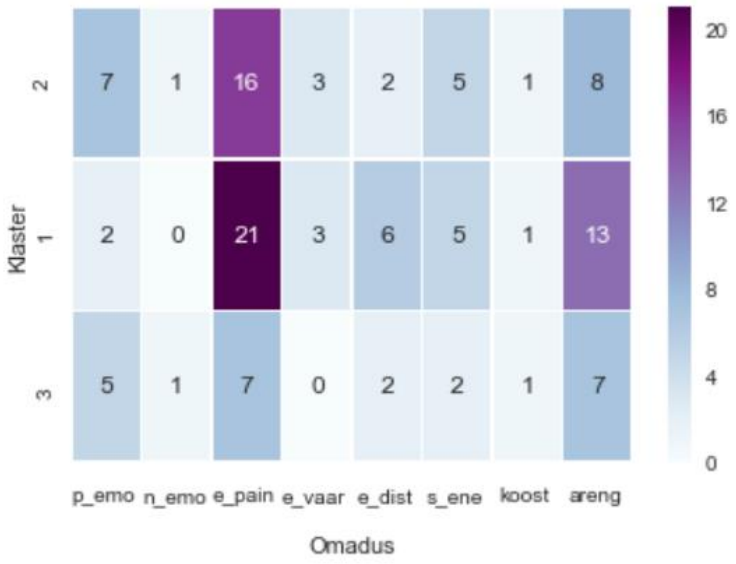
Klastrite ehk teemade ja testi abil mõõdetavate isiksuseomaduste skooride kokku viimiseks koostati maatrikspõhimõttel tabel. Iga inimese kohta valiti suurima tõepära printsiibil teda enim iseloomustav klaster (valikus klastrid 1–3) ning standardiseeritud testiskooridest suurima osakaaluga omadus (valikus 8 omadust). Siinkohal tasub aga arvestada ka suurima tõepära printsiibi kasutamise miinustega, kuivõrd selle alusel on eemaldatud kõik teised skoorid ja jäetud võrdluste tegemise aluseks ainult kõrgeim omadusskoor. Samas võib olla olukordi, kus ühe vastaja puhul on kahe kõrgeima omadusskoori vahe on minimaalne, näiteks 0,001 punkti.

Allolevatel joonistel 13–16 vastavad omaduste lühenditele järgmised kirjeldused: p_emots= positiivne emotsionaalsus, n_emots=negatiivne emotsionaalsus, e_paind=emotsionaalne paindlikkus, e_väärt= eneseväärtustamine, e_dists=enesedistsipliin, s_enesek=sõltumatu enesekindlus, koostöö=koostööoskus, areng=arenguvalmidus (vt ka ptk 4.1). Vastajate klastritesse jaotuvuse ja kõrgeima skooriga omaduse seostest ülevaate saamiseks kuvati andmed graafikul, mis näitab lisaks tihedusele ka marginaalseid sagedusjaotusi (vt Joonis 13). Graafiku x-teljelt on võimalik näha, et kõige rohkem inimesi kuulub esimesse klastrisse (graafiku keskel), suuruse poolest teisel kohal on klaster 2 ja väikseim klaster on klaster number 3. Sama tulemust peegeldas ka LDAvis abil kujutatud joonis (vt nt Joonis 12). Y-teljel kujutatud omaduste jaotus annab aimu, et kõrgemad skoorid on olnud emotsionaalse paindlikkuse osas, kuivõrd see omadus on tumedamalt esil kõikide klastrite all. Mõnevõrra kõrgem on ka arenguvalmiduse esindatus (jällegi graafikul tumedam koondumine). Samuti võimaldab graafik saada ülevaadet klastrite ja kõrgeima skooriga omaduste omavahelisest jaotumusest. Näiteks näeme, et emotsionaalne paindlikkus on enim seotud 1. klastri alla kuulujatega (graafikul kõige tumedam punkt ehk tihedaim vastajate kuuluvus), kuid ka osaliselt klastriga 2. See omakorda tekitab vajaduse hinnata PIST-testi emotsionaalse paindlikkuse väited selle alusel, kas nende väidetega on olemuslikult kergem nõustuda. Sellisel juhul võib tekkida vajadus need väited ümber sõnastada, et testi kvaliteeti parandada.



Joonis 13. Vastajate sagedusjaotus klastrite kuuluvuse ja kõrgeima omadusskoori järgi.

Jooniselt 13 on võimalik saada aimu vastajate üldise jaotumuse osas tema suurima tõenäosusega klastrite ja kõrgeima omadusskoori vahel. Järgnevalt antakse ülevaade, kuidas vastajad loenduspõhiselt jaotuvad klastrite ja kõrgeima omadusskoori järgi (joonis 14). Tulemused pakuvad huvitavat võrdlust eelmises alapeatükis välja toodud tekstianalüüsile.



Joonis 14. Vastajate klastritesse kuuluvus (y-telg) koos kõrgeima omadusskoori (x-telg) ja vastajate loendusega (arvud lahtrites).

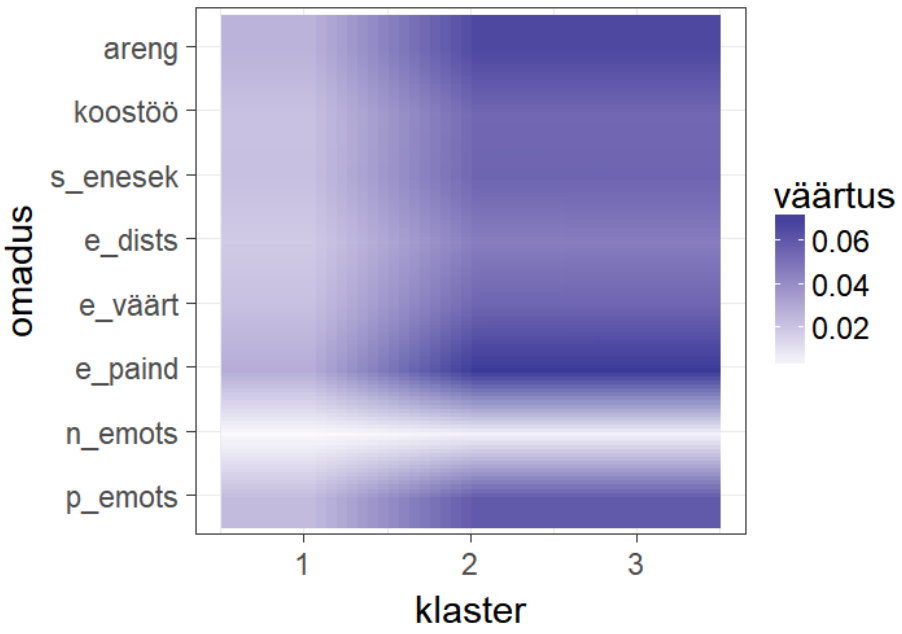
Kõrvutades tekstianalüüsi tulemusi ning klastrikuuluvust koos kõrgeima omadusskoori ülevaatega näeme, et esimese klastri, mille teema on eelnevalt sõnastatud „eesmärgipärased tegutsejad“, inimesi iseloomustab kõrgem emotsionaalne paindlikkus (21 vastajat). Kuivõrd emotsionaalse paindlikkuse omadus viitab suutlikkusele oma emotsioone reguleerida ja ka keerulistes olukordades rahulikuks jääda, viitab teksti ja skooride analüüsi võrdlemine sellele, et sellesse klastrisse kuuluvad heade juhiomadustega inimesed. Ühelt poolt iseloomustab neid eesmärgipärane tegutsemine ja teatav aktiivsus, kuid samas suudavad nad vajalikul hetkel oma emotsioone talitseda. Samuti ilmnes tekstianalüüsist sõna „õppimine“ kasutamine 1. klastris, mis läheb hästi kokku kõrgema arenguvalmiduse skooriga (13 vastajat). Kui võtta arvesse taustateadmist, et PIST-testi täidavad tihti just lõppvooru kandidaadid, saame ülevaate, et värbamisprotsessis jõuavad kaugemale need (nt juhikohale pürgivad) kandidaadid, kes ei paista silma mitte oma ranguse ja kontrollivajadusega, vaid on ühelt poolt tegutsejad ja avatud õppimisele ning teisalt on vajalikul hetkel valmis säilitama enesedistsipliini. Samas aga sõna „positiivsus“, mis esines 1. klastri tekstianalüüsis, ei lähe kokku omadusskooridega, kuivõrd 1. klastris oli vaid kaks inimest, kelle enim avalduv omadus oli seotud positiivse emotsionaalsusega. Siinkohal tuleks põhjuseid jällegi sügavamalt vaadata – kas näiteks esines sõna „positiivsus“ selles klastris mõnes spetsiifilises või äraspidises kontekstis (nt „pole eriti positiivne“) või oli sellesse klastrisse kuulujatel positiivse emotsionaalsuse skoor küll kõrge, aga mitte alati kõige kõrgem võrreldes teiste skooridega (arvestades, et analüüs on teostatud suurima tõepära printsiibist lähtudes).

Teise klastrisse kuulujaid üldistati teemakategooriaga „sisekaemuslikud“. Võrreldes esimese klastriga, paistsid sellesse klastrisse kuulujad silma oma analüütilisema ja sissepoole hoiakuga. Samas kirjeldas neid ka teatav tulevikku vaatav või veendumuslik suund. Numbriliste tulemuste analüüs kinnitab sellele klastrile iseloomulikku emotsionaalset paindlikkust (16 vastajat) ja positiivset emotsionaalsust (7 vastajat), kuid ka kõrgemat arenguvalmidust (8 vastajat). Lisades siia juurde sõltumatu enesekindluse omaduse (5 vastajat), näitab see analüütilisemat laadi inimeste otsustusvõimelisuse avaldumist. Seega võiks tekstianalüüsi ja numbriliste skooride võrdluses üldistada, et sellesse klastrisse kuuluvad inimesed, keda iseloomustab hea eneseanalüüsivõime.

Kolmanda klastri tekstianalüüsi tulemusel leiti seos teatava paindliku iseloomuga, mis tekitas vastajates tihti soovi mõlema etteantud väitega teatud määral nõustuda. Lisaks iseloomustas seda klastrit suund koostööle, mis on samuti paindlikkusega seotud – koostööaltimad inimesed peavad olema ühteaegu ka paindlikud. Võrdlus omadusskooridega näitab, et selles

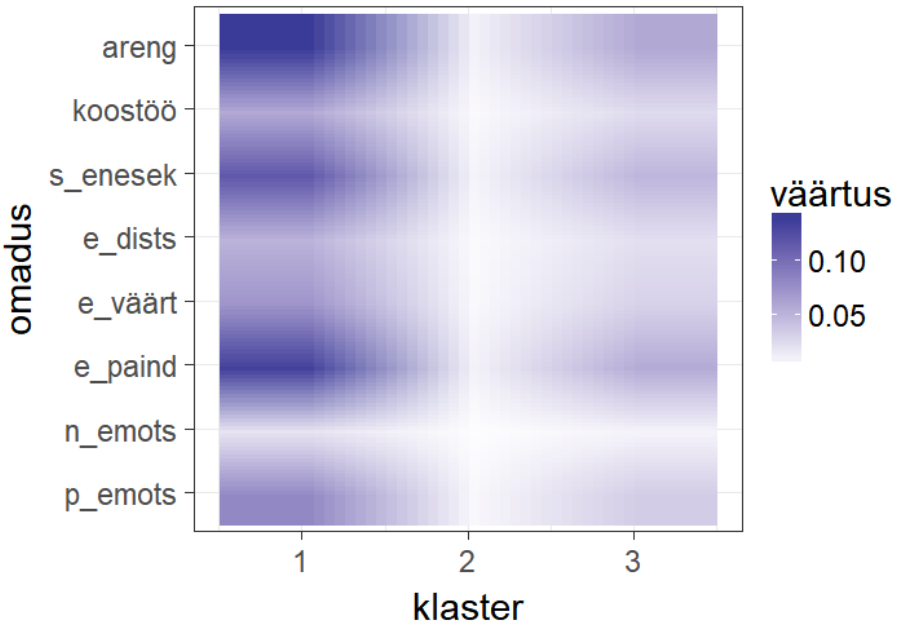
klastris on kõrgematest omadustest esindatud arenguvalmidus (7 vastajat), emotsionaalne paindlikkus (suudab oma emotsioone reguleerida ja on väljenduslaadilt adekvaatne; 7 vastajat) ning positiivne emotsionaalsus (5 vastajat). Samas on huvitav, et koostöö omadusskoorid selles klastris silma ei paista, mistõttu tuleks uurida, mis laadi koostööd testiga mõõdeti ning kuidas vastajad ise seda, näiteks „aitama“ sõna kaudu väljendasid.

Eelnevalt toodi välja suurima tõepära printsiibil põhineva võrdlusanalüüsi tulemused, mis aitas võrrelda eri klastreid selle järgi, milline oli nendesse kuulujate kõrgeim omadusskoor. Järgnevalt näitlikustatakse andmestikust valitud kahe representatiivse inimese põhjal nende standardiseeritud kõikide skooride ja klastritesse jaotumise seoseid graafiliselt. Joonistel 15 on kujutatud 1. näiteisiku ülevaatlik iseloomustus. Jooniselt on värvi tonaalsuse järgi võimalik saada infot nii selle kohta, kuidas on jaotunud tema klastritesse kuulumise tõenäosused, kui ka milline on tema testi tulemusskooride jaotumine (skaalal esindab „väärtus“ tõenäosuste korrutist). Esiteks näeme, et tema negatiivse emotsionaalsuse skoor on olnud väikseim (hele horisontaalne joon). Emotsionaalne paindlikkus on omadusena kõige tugevamalt väljendunud (tume horisontaalne joon) ja mõnevõrra tugevamalt ka arenguvalmidus. Teised omadused paistavad aga olevat üsna ühtlaselt jaotunud, kuivõrd värvide tonaalsus kõigub pigem vähe. Samuti on tema kuulumine 2. või 3. klastrisse peaaegu võrdne (vastavalt 0,4181 ja 0,4184 punkti). Seega iseloomustab seda inimest omadusena enim emotsionaalne paindlikkus, kuid kuuluvus 2. ja 3. klastri vahel on peaaegu eristamatu.



Joonis 15. 1. näiteisiku klastritesse kuulumise tõenäosuste ja testi skooride jaotumine.

Teise näiteisiku puhul on võimalik jooniselt 16 välja lugeda, et tema kõrgemad omadused on arenguvalmidus, emotsionaalne paindlikkus ja sõltumatu enesekindlus, mis on suhteliselt sarnase tugevusega joonisel värvunud. Tema madalaima skooriga omadus on olnud negatiivne emotsionaalsus, kuid ka enesedistsipliin ja koostööoskus on joonisel heledamalt kujutatud, mis annavad tunnistust nende omaduste tagasihoidlikumale avaldumisele. Samuti on 2. näiteisiku puhul võimalik selgemini välja lugeda tema tõenäosuslikku klastrite kuuluvust. Enim on ta seotud 1. klastriga ja vähim 2, klastriga.



Joonis 16. 2. näiteisiku klastritesse kuulumise tõenäosuste ja testi skooride jaotumine.

Kahe näiteisiku põhjal tehtud graafilised ülevaated nende klastritesse kuuluvuse ja omadusskooride jaotumise osas andsid huvitava võrdluspildi sellest, mis suurima tõepära printsiibi rakendamisel tulemuste interpreteerimisel kaduma läks. 1. näiteisiku puhul nägime tema suhteliselt võrdset jaotumist 2. („sisekaemuslikud“) ja 3. klastri („paindlikud“) vahel ning enim avalduvale emotsionaalse paindlikkuse omadusele vastukaaluks madalama omadusena negatiivset emotsionaalsust. 2. näiteisiku graafikult on küll selgemini võimalik tabada tema kuulumine 1. klastri („eesmärgipärased tegutsejad“) alla, kuid samuti näeme, et tema kõrgeimad ja madalamad omadused koonduvad komplektina kokku: vastavalt iseloomustab teda suurem enesekindlus, arengusoov ja emotsionaalne stabiilsus, ning vähem kirjeldab teda negatiivsete emotsioonide valdamine, enesedistsipliin ja suund koostööle. Selliselt kompleksemalt väljajoonistuvate omaduste ja klastrite järgi on konkreetset inimest võimalik ka terviklikumalt iseloomustada.

7. Kokkuvõte

Isiksustestide kasutamine on personalivaliku kontekstis üsna tavapärane praktika, kuid ettevõttes Psience välja töötatud ja kasutatud isiksusetesti (PIST-test) näitel võib testi tulemuste hindamine tekstiliste andmete pealt olla ajamahukas ettevõtmine. Käesolevas magistritöös analüüsiti teemade modelleerimise meetodi abil ettevõtte Psience'i isiksustestide täitnute enesekohaseid tekste ning võrreldi tulemusi testi tulemusskooridega. Teemade modelleerimisel kasutatava LDA-mudeli rakendamine aitas leida tekstis peidetud mustreid ja kõrvutatult PIST-testi tulemusskooridega teha üldistavaid järeldusi nii testi kvaliteedi kui ka vastajaid kirjeldavate omaduste kohta.

Enesekohaste tekstide peidetud mustrid avaldusid kolme teemasse ehk klastrisse koondumises ja nende klastrite alla kuuluvate iseloomulike sõnade esinemises. Tekstianalüüsi tulemusel kirjeldati kolme teemat ehk vastavalt kolme inimeste rühma: „eesmärgipärased tegutsejad", „sisekaemuslikud" ja „paindlikud". Kõigi kolme klastri puhul leiti ka seoseid standardiseeritud omadusskooride järgi kuuluvusega. Samas esines võrdluses ka teatavaid vasturääkivusi, mis nõuavad aga edasist süvendatud uurimist. Näiteks ilmnes 1. klastri iseloomulikema sõnade seas sõna „positiivne", kuid selles klastris oli vähe inimesi, kelle kõrgeim omadusskoor oleks olnud positiivses emotsionaalsuses. Vastajate omadusskooride võrdlemisel klastritesse kuulumisega leiti, et kõrgeimad skoorid kõigis klastrites olid emotsionaalse paindlikkuse osas, mis omakorda tekitab vajaduse hinnata seda omadust mõõtvaid väiteid (st kas nende väidetega on tavapärasest kergem nõustuda nende sõnastuse triviaalsuse vms tõttu).

PIST-testi kui toote edasi arendamise plaanis olid kõige huvitavamad aga mustrid, mis avaldusid 3. teema analüüsimisel. Nimelt viitas sõnade analüüs, et sellesse klastrisse koonduvad tõenäolisemalt need vastajad, kellel on teatav paindlikkus isiksuseomaduste avaldumisel erinevates kontekstides ning mida nad ise väljendasid tihti ka teatud kahevahel olekut peegeldava sõnastuse läbi. See omakorda aga lubab teemade modelleerimise edasiarendatud kasutamist PIST-testi enesekohastel tekstidel selles plaanis, et teha tekstianalüüsi automaatsemaks ja töötada välja meetod, kuidas teksti analüüs saaks panustada ka testi numbrilistesse tulemustesse, mida seni veel pole rakendatud.

Antud uurimuse peamiseks probleemiks oli väikesemahuline tekstikorpus, mis ei võimalda masinõppe mudeli kasutamisest täiel määral kasu saada. Näiteks sattus kõikide teemade alla ka sarnaseid sõnu, nagu „töö", „meeldima", „suutma", mis pigem lähendavad teemasid. Samuti tuleb arvestada LDA-mudeli kasutamisel sellega, et teemasid iseloomustavad sõnade

kontekst ei pruugi adekvaatselt avalduda, nt kui tekstide eeltöötluse faasis on eemaldatud eituse vormid, olulised sidesõnad vms. Siiski andis LDA-mudeli kasutamine võimaluse leida tekstides peituvaid mustreid, mis n-ö käsitsi analüüsimisel oleks väga ajamahukas tegevus.

Peamised ettepanekud PIST-testi tulevikuvõimaluste osas on järgmised:

1. Kasutada testi võrdlevas analüüsis lisaks skooridele ja tekstidele ka teisi kirjeldavaid andmeid, nagu vastaja sugu, vanus, hariduslik taust, praegune töökoht ning lisada juurde kandideerimisel olev positsiooni nimetus ning tulemus kandideerimisprotsessis (kas osutus valituks või mitte) jms. Selliselt oleks võimalik teha tuleviku tarbeks oluliselt ülevaatlikumat personalivaliku analüüsi, mis tulemuste interpreteerimisel aitaks tõsta ka ettevõtte personaliteenuste müügipotentsiaali.
2. Töötada välja mudel, mis võtab automaatsemalt arvesse ka tekstilisi vastuseid ja arvestab nendega testi tulemusskoorides, aidates seeläbi vähendada testi analüüsiks kuluvaid inimtunde.
3. Teostada testi kvaliteedi hindamine (nt faktoranalüüsil vms meetodil), võttes arvesse käesoleva uurimuse tulemusi. Näiteks võiks hinnata emotsionaalse paindlikkuse väiteid selles osas, et kas ehk pole nendega liiga kerge nõustuda ja seega saadakse selle omaduse alla pea alati kõrgemaid skoore.

8. Viidatud kirjandus

Aggarwal, C. C. ja Zhai, C. (Eds.). (2012). *Mining Text Data*. New York, NY: Springer Science & Business Media.

Allik, J. (2003). Isiksus ja seadumused. J. Allik, A. Realo, K. Konstabel (toim). *Isiksusepsühholoogia* (23–65). Tartu: Tartu Ülikooli Kirjastus.

Blei, D. M, Ng, A. Y ja Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M. (2012a). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M. (2012b). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, *2*(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ (vaadatud 20.05.2018).

Cadwalladr, C. ja Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian* 17.03.2018. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (vaadatud 17.05.2018).

Coppersmith, G., Dredze, M., ja Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 51-60).

Costa, P. T., Jr ja McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological assessment Resources

Costa, P. T., Jr, McCrae, R. R. ja Kay, G. G. (1995). Persons, places, and personality: Career assessment using the Revised NEO Personality Inventory. *Journal of Career Assessment*, *3*(2), 123–139.

De Choudhury, M., Counts, S., Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47-56). ACM.

Eysenck, H. J. (1997). *Rebel with a cause: The autobiography of Hans Eysenck*. New Brunswick: Transaction Publishers.

Friedman, H. S., ja Schustack, M., W. (2016). *Personality: Classic Theories and Modern Research*. Boston: Pearson

Gandomi, A. ja Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Jaidka, K., Guntuku, S. C., ja Ungar, L. (2018). Facebook vs. Twitter: Differences in self-disclosure and trait prediction. *Proceedings of the International AAAI Conference on Web and Social Media*.

Jovanovic, M., Vukicevic, M., Milovanovic, M., ja Minovic, M. (2012). Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, *5*(3), 597-610.

Kallasmaa, T., Allik, J., Realo, A. ja McCrae, R. R. **(**2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the Five-Factor Model. *European Journal of Personality, 14*, 256-278.

Kalmus, V. (2015). Standardiseeritud kontentanalüüs. K. Rootalu, V. Kalmus, A. Masso, ja T. Vihalemm (toim), *Sotsiaalse analüüsi meetodite ja metodoloogia õpibaas*. http://samm.ut.ee/kontentanalyys (vaadatud 08.05.2018).

Konstabel, K. (2006). *The Structure and Validity of Self- and Peer-Reported Personality Traits*. Doktoriväitekiri. Tartu Ülikool, psühholoogia osakond. Tartu: Tartu Ülikooli Kirjastus.

Kwartler, T. (2017). *Text mining in practice with R*. Hoboken, NJ: John Wiley & Sons.

Lepik, N. (2017). Tõenäosusteooria ja statistika II. Loengukonspekt. https://courses.ms.ut.ee/MTMS.02.050/2017_fall/uploads/Main/tntstat2_16.pdf (vaadatud 20.05.2018).

McLeod, J. D. ja Lively, K. J. (2006). Social structure and personality. J. Delamater (eds). *Handbook of Social Psychology* (pp. 77-102). Boston: Springer.

Mischel, W., Shoda, Y., ja Mendoza-Denton, R. (2002). Situation-behaviour profiles as a locus of consistency in personality. *Current Directions in Psychological Science, 11*, 50–54.

Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S. ja Kaalep, H.-J. (2016). EstNLTK - NLP Toolkit for Estonian. N. Calzolari, K. Choukri, T. Declerck, et al (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Paris, France: European Language Resources Association (ELRA).

Orav, H. (2006). *Isiksuseomaduste sõnavara semantika eesti keeles*. Doktoriväitekiri. Tartu Ülikool, eesti ja soome-ugri keeleteaduse osakond. Tartu: Tartu Ülikooli Kirjastus.

Řehůřek, R., ja Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (pp. 45–50). Valletta, Malta: ELRA.

Santos, J. (2011). *Analysis of Retweeting Behavior Using Topic Models*. University of Tartu, Institute of Computer Science.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ (vaadatud 20.05.2018)

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS one 8*(9), e73791.

Sievert, C. ja Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Zakrzewska, D. (2009). Cluster Analysis in Personalized E-Learning Systems. N.T. Nguyen, E. Szczerbicki (eds), *Intelligent Systems for Knowledge Management* (pp. 229-250). Heidelberg, Berlin: Springer

Zhai, C. ja Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, NY: Association for Computing Machinery and Morgan & Claypool Publishers.

Zhao, H. ja Seibert, S. E. (2006). The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of Applied Psychology, 91*(2), 259–271.

Valdson, K. (2016). *Mustripõhine informatsiooni eraldamine Eesti kohtulahenditest*. Magistritöö. Tartu Ülikool, arvutiteaduse instituut.

Yankovskaya, E. (2017). *Extraction and Classification of App Features from App Reviews*. Master's Thesis. University of Tartu, Computer Science Curriculum.

Lisad

I. LDA abil modelleeritud teemade ja sõnade jaotus

Allolevas tabelis 1 on välja toodud teemad ja nende alla kuuluvad sõnad, mis on vastavalt kujutatud joonistel 9–11.

Tabel 1. Sõnade jaotus vastavalt teemadele.

Teema 1	**Teema 2**	**Teema 3**
töö	tulema	elu
meeldima	meeldima	tulema
võima	hea	meeldima
hea	palju	üldiselt
elu	aeg	vastus
teadma	välja	andma
püüdma	töö	eelistama
otsus	võima	nägema
tulema	suutma	number
hakkama	ütlema	looma
ütlema	uskuma	aitama
andma	tahtma	just
rohkem	tundma	välja
palju	hästi	tähtis
mõtlema	teadma	tunne
välja	tihti	aeg
suutma	keegi	õige
lahendus	tunne	palju
eesmärk	uus	suutma
leidma	plaan	sõltuma
võtma	jääma	tegelikult
arvama	oskama	otsus
positiivne	oluline	leidma
plaan	elu	vaja
ära	ette	võima
õppima	nägema	puhul
oluline	jõudma	ent
tundma	läbi	mõlema
soovima	rohkem	valima
üritama	võimalik	tegutsemine

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Liset Marleen Pak**,
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Tekstikaeve rakendamine isiksusetestidel personali värbamise eesmärgil,
(*lõputöö pealkiri*)

mille juhendaja on Anna Leontjeva, PhD,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **21.05.2018**