TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Vladislav Šikirjavõi 163594IAPM

TEMAATILISTE MUSTRITE KAEVANDAMINE SEADUSETEKSTIDEST

magistritöö

Juhendaja: Ahti Lohk
Doktor

Tallinn 2018

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Vladislav Šikirjavõi

07.05.2018

Annotatsioon

Käesoleva lõputöö eesmärgiks on rakendada teemade modelleerimist eesti seadusetekstidele, et saada teada, milliste “abstraktsete” ehk peidetud teemade kaudu on eesti suurimad seadused omavahel seotud. Teemade modelleerimise algoritmide abil on võimalik leida suurest andmehulgast seoseid ja erinevusi õigusaktide abil, mille inim jõul leidmine oleks märksa keerulisem ja aeganõudvam.

Mainitud eesmärgi saavutamiseks tehakse teoreetiline tutvustus teemasse ja kogutakse kokku algandmed. Järgnevalt luuakse keskkond, kus installitakse uuringuks vajaminevad paketid. Seejärel rakendatakse teemade modelleerimise algoritme eesti õigusaktide andmetele.

Töös käsitletavad probleemid on, et kas teemade modelleerimise rakendamine on eesti õigusaktide jaoks efektiivne ning kas saadud tulemused oleks kasulikud.

Töö põhitulemuseks on eesti seaduste peidetud teemade leidmine kolme algoritmi abil ning saadud mudelite hindamine.

Töö käigus tulid välja parenduskohad, mille kasutamisel saaks teemade modelleerimist eesti seadusetekstidele efektiivsemaks muuta.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 69 leheküljel, 4 peatükki, 9 joonist, 15 tabelit.

Abstract

Extraction of the thematic patterns from the texts of law

The purpose of this thesis is to apply topic modeling to Estonian legal texts in order to find out which “abstract“ or hidden topics the major laws of Estonia are interconnected. Using topic modeling algorithms, it is possible to find the largest amount of data in the relationships and differences in legislation, which would be much more complicated and time-consuming to find for a person.

In order to achieve this goal, a theoretical introduction to the topic and collection of initial data are made. After that, an environment is created, where the needed packages for the study are installed. Then, the topic modeling algorithms are implemented on the data of Estonian legislation.

The problems addressed in the work are that either the implementation of the topic modeling is effective for Estonian legislation and could it be used in some application.

The main result of the thesis is the discovery of hidden topics in Estonian law by using three algorithms and evaluation of the resulting models.

In the course of the work, there were places of improvement, which could make the modeling of the subjects more effective in Estonian legal texts.

The thesis is in Estonian and contains 69 pages of text, 4 chapters, 9 figures, 15 tables.

Lühendite ja mõistete sõnastik

LDA	*Latent Dirichlet Allocation*
LSI	*Latent Semantic Indexing*
HDP	*Hierarchical Dirichlet Process*
NLP	*Natural language processing*
NLTK	*Natural Language ToolKit*
PDF	*Portable Document Format*
LLDA	*Labeled Latent Dirichlet Allocation*
Tf-idf	*Term frequency-inverse document frequency*

Sisukord

Jooniste loetelu

Tabelite loetelu

Sissejuhatus

Suur hulk andmeid kogutakse iga päev. Mida rohkem informatsiooni on saadaval, seda raskem on leida seda, mida otsime. Samuti muutub aastatega Eesti õigusaktide [1] keel aina keerulisemaks ning nii tavainimestele kui ka isegi õigusega igapäevaselt tegelevatele inimestele raskemini arusaadavaks [4]. Seetõttu tuleks õigusaktide tekste organiseerida, analüüsida, et välja otsida ja presenteerida sisu, mis on kasulik. Kuna inim jõul selle lahenduse leidmine oleks märksa keerulisem ja aeganõudvam, siis pakume välja teemade modelleerimist (ingl k *topic modeling*) [12], mida on tehnoloogia tohutu arengu tõttu aina rohkem kasutatud erinevate andmete puhul. Samuti ei ole seda siiani eesti seadusetekstidele rakendatud.

Lühidalt öeldes teemade modelleerimine on üks tehnika tekstide kaevandamise (ingl k *text mining*) valdkonnas. Nagu nimigi ütleb, on see protsess tekstis esinevate teemade (ingl k *topic*) automaatseks tuvastamiseks ja tekstikorpuse peidetud mustrite (ingl k hidden topical patterns) leidmiseks. Teemade modelleerimise algoritmid koosnevad statistilistest meetoditest, mis analüüsivad sõnu antud tekstides ja avastavad peidetud teemad. Tekstide sildistamine (ingl k *labelling*) ei ole vajalik teemade modelleerimise algoritmide jaoks [12].

Teema on aktuaalne, kuna uuringus kasutatakse Gensimi [25], mis on uudne avatud lähtekoodiga Pythonis implementeeritud teemade modelleerimise tööriistakomplekt. Alates 2009.aastast on seda kasutatud ja tsiteeritud üle 800 äri- ja akadeemilises rakenduses, alates meditsiinist kuni kindlustusnõude analüüsi patendiuuringuni. Lisaks sellele on järjest rohkem seadusetekste saadaval digitaalselt kujul, mille puhul tekib võimalusi ja väljakutseid neid lihtsamal moel esitada. Samuti metoodika osas on plaan rakendada tekstikaeves kasutatavaid algoritme *Latent Dirichlet Allocation* (LDA) [13], *Latent Semantic Indexing* (LSI) ja *Hierarchical Dirichlet Process* (HDP) avastamaks seadusetekstide üleseid varjatud temaatilisi mustreid ning kasutada uudset avatud lähtekoodiga Pythoni EstNLTK paketti (alates 08.jaanuar 2016). Puhastatud andmete põhjal luuakse mudelid, mille järgi tehakse edaspidine analüüs. Selle lõputöö eesmärgiks on rakendada teemade modelleerimist eesti seadusetekstidele ning saada teada, milliste

“abstraktsete” ehk peidetud teemade kaudu on eesti suurimad seadused omavahel seotud. Kui uuringu tulemused on head, siis on võimalik, et seda saaks kasutada tulevikus seadustest info otsimisel ühe abivahendina või oleks saadud mudelit võimalik kasutada teiste seadusetekstide peal, mis jagaks need samuti teemade vahel ära.

Töö põhitulemuseks on eesti seaduste peidetud teemade leidmine kolme erineva algoritmi abil ning mudelite hindamine. Samuti saada teada, milline mudel saavutab parima tulemuse.

Töö jaguneb neljaks suuremaks osaks: esmalt alustatakse teema teoreetilisest tutvustusest, mille hulgas räägitakse õigusaktidest, teemade modelleerimisest. Seejärel tutvustatakse uuringut ja selles osalevaid õigusakte. Seaduste osas kasutatakse magistritöö jaoks kakskümmend suurimat õigusakti selle alusel, palju seaduse lõike neis on. Kolmandas osas viiakse läbi uuring. Uuringu osas alustatakse dokumentide kogumise ja integreerimisega ning eeltöötlusega (ingl k *preprocessing*), kus põhitegevuseks on andmete puhastamine (ingl k *data cleaning*), kus eemaldatakse õigusaktide tekstidest kõik ebavajalikud sõnad, kirjavahemärgid ning muu, mis ei ava teemat ning ei anna sisule midagi märkimisväärset juurde. Samuti on plaanis saada kvalitatiivne hinnang isiku käest, kes tunneb seaduseid. Saadud tulemusi on plaanis ka visualiseerida. Kokkuvõttes annab töö autor omapoolsed soovitused tulevikuks.

Magistritöö kirjutati Tallinna Tehnikaülikoolis, infotehnoloogia teaduskonnas, tarkvarateaduse instituudis, 2018.aasta kevadel.

1 Teoreetiline taust ja masintöötlusvahendid

Alljärgnevas peatükis anname ülevaate teoreetilisest taustast, mis on seotud antud lõputööga. Samuti kirjeldame ära, milliste vahendite abil me teemade modelleerimist teeme ning milliseid tehnoloogiad antud lõputöö uuringus kasutame.

1.1 Õigusaktid

Õigusaktid on õigusnormide kirjalik vormistus ja nende süsteemne kogum, mis koondatakse kindlatele nõuetele vastavasse dokumenti. Tänu õigusaktidele annab riik kindlatele käitumisreeglitele üldkohustusliku jõu. Nende abil määratakse kindlaks riigi ja kodanike omavahelised suhted ning korraldatakse kodanike ühiselu [1].

Õigusaktide puhul eristatakse kahte tüüpi. Õigustloovad aktid ehk õiguse üldaktid sisaldavad õigusnorme, mis on pädeva institutsiooni poolt kindlas korras loodud üldise iseloomuga. Samuti on õiguse üksikaktid, mis ei sisalda õigusnormi. Õiguse üksikaktideks on näiteks otsused, korraldused ja käskkirjad [1].

Põhiseaduse § 3 lg 2 sätestab, et täitmiseks saavad kohustuslikud olla üksnes avaldatud seadused. Õigusaktid on saadaval internetis elektroonilise Riigi Teataja leheküljel [2]. Elektrooniline Riigi Teataja on Eesti Vabariigi ametlik võrguväljaanne ja selle kaudu on võimalik otsida erinevaid õigusakte ning näha kõiki hetkel kehtivaid terviktekste [1].

Olenemata sellest, et mitmed Riigi Teataja küsitlused on selle efektiivsust tõestanud, on alati ka vastupidiseid väiteid. 2017.aasta Riigi Teataja küsitluse [3] tulemustes on näha, et vastanutest olid enamasti inimesed, kes tegelevad õigusega igapäevaselt. Seitsmenda punkti juures on näha, et kes kasutab Riigi Teatajat riigiametnikuna on 43%, õppe- ja teadustöös 46% ning isiklikes õigusalastes asjades 44,1%. Ülejäänud protsendid jäävad väikeseks.

Riigikogu põhiseaduskomisjoni esimehe Ken-Marti Vaheri sõnul on Eestis kahjuks viimastel aastatel muutumas õigusaktide keel aina keerulisemaks ja tavainimesele

raskemini arusaadavaks [4]. Samuti märkis Vaher, et eelkõige on õiguskeel muutumas keerulisemaks Brüsselist tulevates regulatsioonides ja püüelda tuleks lihtsama keelekasutuse poole. Probleemiks on ka see, et olemasolevaid seadusi muudetakse liiga tihti, laskmata seadusel korralikult rakenduda.

Hea Õigusloome Tava on Teenusmajanduse Koja poolt koostatud konkurss, mille eesmärk on pöörata avalikkuse ja seadusandja tähelepanu kvaliteetse õigusloome olulisusele [5]. Konkursil osalevad Riigikogu poolt vastuvõetud seaduse, Vabariigi Valitsuse määrused, Vabariigi Valitsuse ministri või kohaliku omavalitsuse määrused. Seadused kandideerivad kahes kategoorias [4]: parim seadus, kus osalevad seadused, mille menetlemise käigus on kõige enam järgitud Head Õigusloome Tava, ning halvim seadus, kus osalevad seadused, mille menetlemise käigus on kõige enam eiratud Head Õigusloome Tava. Seega on parima ja halvima seaduse konkursil mõõdupuuks vastavus Hea Õigusloome Tavale, mitte aga näiteks poliitiline eelistus.

Need õigusaktide kohta käivad küsitlused, uudised ning Hea Õigusloome Tava konkurss näitab, et õigusaktid ei ole sugugi kõigi jaoks arusaadavad. Tegelikult, et seadus toimiks, ei pea ta olema mitte üldiselt, vaid pigem hädavajalikult minimaalselt arusaadav. Teistmoodi mõistes peab seadus olema arusaadav neile, kellele see on täitmiseks. Demokraatlikus riigis ei tohiks seadusi kirjutada nii, et neid mõistaks ainult professionaalsed juristid. Kuid sellegipoolest on seadused oma kogumis paratamatult sellised, et neid mõistavad ainult erialase taustaga inimesed. Seadusandja seisukohalt peab seadus olema kirjutatud nõnda, et võimalikult paljud inimesed sellest võimalikult palju ja hästi aru saaksid [6]. Riigi Teataja kodulehel on seadused küll süstemaatiliselt liigitatud, kuid see ei ole kõikidele lihtsasti arusaadav ning seepärast on plaanis selle lõputöö käigus rakendada teemade modelleerimist nendele seadusetekstidele, et saada teada, milliste “abstraktsete” ehk peidetud teemade kaudu on eesti suurimad seadused omavahel seotud.

Lisaks on veebilehitsejates olemas lisaprogramm nimega xLaw. Selle programmi autoriks on eestlane Egert Nõlv. See rakendus loodi 2015.aasta aprillis ning aasta pärast oli kasutajaid juba üle 600. Programm xLaw on seadustega igapäevaselt kokkupuutuvatele inimestele mõeldud rakendus, mida saab kasutada veebilehitsejatega Chrome ja Firefox. Rakendus omab mitu funktsionaalsust. Tasuta funktsionaalsuste puhul on näiteks xLaw

§'i-põhine info, mis lisab §'de juurde nupud, millele vajutades saab § kohta lisainfot (sh kohtulahendid, artiklid, juhendid, Euroopa Liidu õigus, seletuskirjad) ning xLaw kasutaja privaatse info funktsioon, mis võimaldab koguda ja kasutada oma kommentaare ühtses xLaw märksõnasüsteemis (Riigiteataja.ee, EUR-Lex, Word) [36]. Antud lõputöö käigus on plaanis läbi teha temaatiliste mustrite kaevandamine seadusetekstidest selle tõttu, et see oleks samuti tulevase rakenduse sisendiks, eeltööks või osaks. Kui programm xLaw lihtsustab seadusega igapäevaselt kokkupuutuvate inimeste tööd, siis selle lõputöö käigus saadud tulemus võiks olla mõeldud tulevase rakenduse jaoks, mis oleks pigem suunatud inimestele, kes seadustega igapäevaselt ei tegele, kuid sellegi poolest puutuvad sellega tihti kokku.

1.2 Teemade modelleerimine

Andmekaeve (ingl k *data mining*) on protsess, mis hõlmab meetodeid masinõppe, statistika ja andmebaasi süsteemide ristumisel, ning mille käigus tuvastatakse mustrid suurtes andmekogudes. Andmete kaevandamisel on oluline roll mustrite avastamisel. See on teadusuuringute valdkond, mis on seotud kasulike teadmiste hankimisega suurtest andmehulkadest [7]. Tekstikaeve erinevus võrreldes andmekaevega on see, et tekstikaeve puhul on andmed struktureerimata. Protsessi, mille käigus leitakse mustrid andmetest, nimetatakse mustrite kaevandamiseks. Mustrid on regulaarsused, mis nendes andmetes esinevad, näiteks milliseid kaupu müüakse tavaliselt koos, millised geenid on enamasti aktiivsed teatud haiguse juures või mis tüüpi klient firmale kõige tõenäolisemalt kasumit teenib. Kõik sellised mustrid annavad meile kasuliku ülevaate [8].

Andme- või tekstikaeve mitmesugustes rakendustes seisame silmitsi probleemiga, mille puhul on vaja dokumente liigitada nende teemade järgi või nagu eelmainitud mustrite leidmisega [8]. Tavaliselt tuvastatakse teemad dokumentide iseloomustavate eriliste sõnade leidmisega. Näiteks on pesapalli artiklitel palju lugusid, kus esinevad sellised sõnad nagu “pall”, “kurikas”, “jooksma” ja nii edasi. Kui me oleme teinud kindlaks, et tegemist on pesapalli artiklitega, siis ei ole raske luua seost, miks sellised sõnad esinevad antud dokumentides sageli. Kuid kuni me pole seda kindlaks teinud, ei ole võimalik neid iseloomulikuks pidada [9].

Seega algab klassifitseerimine (ingl k *classification*) sageli dokumentide uurimise ja oluliste sõnade leidmisega. Meie esimene arvamus võib olla selline, et sõnad, mis ilmuvad dokumendis kõige enam, on kõige olulisemad [9]. Kuid see intuitsioon on täpselt vastupidine tõele. Kõige sagedamini ilmuvad tavaliselt sellised sõnad nagu "ja" või "ning", mis aitavad ehitada ideed, kuid tegelikult ise ei oma tähtsust. Selliseid sõnu nimetatakse stoppsõnadeks (ingl k *stop words*), mis eemaldatakse dokumentidest juba enne klassifitseerimist. Stoppsõnu eemaldatakse juba loomuliku keele masintöötluse [10] faasis. Peale stoppsõnade eemaldatakse andmetest kõik kirjavahemärgid, numbrid ja teised sõnad, mis ei oma sisulist tähtsust.

Loomuliku keele masintöötlus ehk NLP (ingl k *Natural language processing*) on arvutiteaduse, tehisintellekti ja arvutilingvistika ühisosa, mis keskendub inimese ja arvuti vahelisele suhtlusele [10]. Peaaegu kõikides loomuliku keele masintöötlusega seotud ülesannetes tuleb läbida mõned eeltöötluse etapid, et teisendada toortekst (ingl k *raw text*) vormi, mis on mudeli ja masina jaoks loetav.

Loomuliku keele masintöötlus omab mitu eelist. Ettevõtted kasutavad loomuliku keele masintöötlust, et parandada dokumenteerimisprotsesside tõhusust ja täpsust ning tuvastada suurtest andmebaasidest kõige asjakohasem teave. Näiteks võib haigla kasutada loomuliku keele masintöötlust selleks, et arsti struktureerimata märkmetest paika panna diagnoos. Samuti aitab NLP näiteks tööandjatel resümee tekstide abil meelitada erinevaid kandidaate ja palgata rohkem osavamaid töötajaid [11].

Loomuliku keele masintöötlus on lihtsalt üks osa teemade modelleerimisest [12]. Teemade modelleerimine on protsess tekstides esinevate teemade automaatseks tuvastamiseks ja tekstikorpuse peidetud mustrite leidmiseks. Teemade modelleerimise algoritmid koosnevad statistilistest meetoditest, mis analüüsivad sõnu antud tekstides ja avastavad peidetud teemad. Nagu ka eelnevalt mainisime, siis ei ole tekstide sildistamine vajalik teemade modelleerimise algoritmide jaoks [12].

Olemas on mitu lähenemist, mis suudavad peidetud mustreid leida. Lõputöö uuringus kasutame *Latent Dirichlet Allocation* (LDA) [13], mis on laialdaselt kasutatav teemade modelleerimise tehnika. LDA mudelis iga dokument koosneb mitmest erinevast teemast, mis esinevad kogu korpuse ulatuses. Mudeli põhjal saab öelda, et iga sõna on seotud

kindla teemaga [14]. LDA mudel avastab erinevaid teemasid, mida dokumendid esindavad ja kui palju esineb igat teemat antud dokumentides [15].

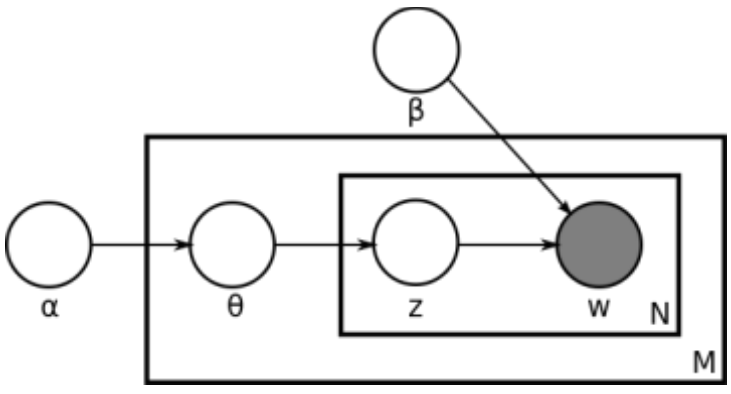
Samuti kasutame uuringus *Latent Semantic Indexing* (LSI) meetodit, mis analüüsib dokumendikogu tervikuna, et näha, millised dokumendid sisaldavad samu sõnu. LSI arvab, et dokumendid, millel on palju ühiseid sõnu, on semantiliselt lähedased. Sama loogika järgi need dokumendid, kus on vähe ühiseid sõnu, on semantiliselt kauged. See lihtne meetod on üllatavalt sarnane sellega, kuidas inimene ise klassifitseeriks dokumendikorpust [16].

Veel lisaks kasutame uuringus *Hierarchical Dirichlet Process* (HDP), mille erinevus LDA-ga on see, et HDP leiab optimaalse teemade arvu tekstikorpusest ise ja sellele algoritmile ei ole seda arvu vaja ette anda [17]. Põhimõtteliselt HDP on LDA laiendus, mille eesmärk on käsitleda sellist olukorda, kus teemade arv pole ette teada.

1.2.1 LDA

Esimene teemade modelleerimise mudel, mida vaatleme, on LDA. LDA on korpuse generatiivne tõenäosuslik mudel, mille põhieesmärgiks on jagada dokumendid juhuslike teemade vahel, kus iga teema omab talle iseloomulikke sõnu. Teemade kogumid on selleks, et võimalikult hästi kirjeldada dokumentides sisaldavat teksti [13]. LDA eelduseks on see, et iga dokument koosneb piiratud teemade kogumikust ja igale sõnale dokumendist on võimalik määrata ühest nendest teemadest.

Joonisel 1 on kujutatud LDA diagrammi:



Joonis 1 LDA diagramm

Antud diagrammi järgi joonisel 1 on näha, et sõna w on sõltuvuses teemaga z, mis on omakorda sõltuvuses dokumendiga θ, mis koosneb mitmest erinevast teemast. Samuti α on parameeter, mis näitab dokumentide teemade jaotust. Parameeter β näitab sõnade jaotust iga teema kohta.

Selleks, et eelneva joonise sisu paremini edasi anda, selgitame allpool lihtsamalt, kuidas LDA leiab teemad dokumentide vahel.

Artikli [18] selgituse abil kujutame ette, et meil on kolm dokumenti A, B ja C, ning me tahame tuvastada, milliseid abstraktseid teemasid need dokumendid sisaldavad.

A: Ma söön kala ja juurvilju.

B: Kalad on koduloomad.

C: Mu kass sööb kala.

Antud lausete põhjal võib LDA algoritm klassifitseerida järgnevalt, et sõnad "söön", "kala", "juurvilju", "sööb" esimeses ja kolmandas lauses kuuluvad ühe teema alla, mida nimetame näiteks "toit". Samuti ülejäänud sõnad "kalad", "koduloomad", "kass" teises ja kolmandas lauses kuuluvad teise teema alla, mida nimetame "koduloomad".

LDA saavutab antud tulemused 3 käiguga.

1. Algoritmile peab ette andma teemade arvu (K). Teemade optimaalse arvu võib eelnevalt leida uuringu käigus või kasutada katse ja eksituse meetodit. Erinevate variantide proovimisel on võimalik leida see üks teemade arv, mille abil on teemade tõlgendatavus kõige parem.

2. Algoritm määrab igale sõnale ajutise teema K teemade hulgast. Teemade määramine on ajutine, kuna see uuendatakse järgmises punktis. Ajutised teemad määratakse Dirichlet' jaotuse järgi, mis tähendab samuti seda, et isegi kui sõna esineb mitu korda, siis see võib olla määratud erinevate teemade külge.

3. Kolmandas punktis käib algoritm kõik sõnad läbi ja uuendab nendele kuuluvad teemad. Tegu on iteratiivse protsessiga. Iga sõna teema uuendamise puhul vaadeldakse kahte kriteeriumit:

- kui valdava osa moodustab sõna sellest teemast?
- kui valdava osa moodustavad teemad dokumendis?

Nende kahe kriteeriumi selgitamiseks toome näite, kus me tahame teada, mis teemaga on seotud sõna “Kala“ dokumendis B.

Tabel 1 LDA teema määramine

Teema	**Dokument A**	**Teema**	**Dokument B**
F	Kala	?	Kala
F	Kala	F	Kala
F	Sööma	F	Piim
F	Sööma	P	Kass
F	Juurviljad	P	Kass

Nüüd võtame ette esimese kriteeriumi: kui valdava osa moodustab sõna sellest teemast? Kuna sõna “Kala” hõlmab kahe dokumendi peale peaaegu pool teemast F ning 0% teemast P, siis sõna “Kala” kuuluks pigem teema F alla.

Tabel 2 LDA teema määramine (1)

Teema	**Dokument A**	**Teema**	**Dokument B**
F	**Kala**	?	Kala
F	**Kala**	**F**	**Kala**
F	Sööma	F	Piim
F	Sööma	P	Kass
F	Juurviljad	P	Kass

Nüüd võtame ette teise kriteeriumi: kui valdava osa moodustavad teemad dokumendis? Kuna sõnad dokumendis B on jaotatud teema F ja teema P vahel 50-50 suhtega, siis sõna “Kala” võib kuuluda nii ühe kui ka teise teema alla.

Tabel 3 LDA teema määramine (2)

Teema	**Dokument A**	**Teema**	**Dokument B**
F	Kala	?	Kala
F	Kala	**F**	**Kala**
F	Sööma	**F**	**Piim**
F	Sööma	**P**	**Kass**
F	Juurviljad	**P**	**Kass**

Arvestades mõlema kriteeriumi tulemusi, saame öelda, et sõna “Kala” kuulub teema F alla. Tulemusena saadakse järgnev jaotus:

Tabel 4 LDA teema määramine (3)

Teema	**Dokument A**	**Teema**	**Dokument B**
F	Kala	F	Kala
F	Kala	F	Kala
F	Sööma	F	Piim
F	Sööma	P	Kass
F	Juurviljad	P	Kass

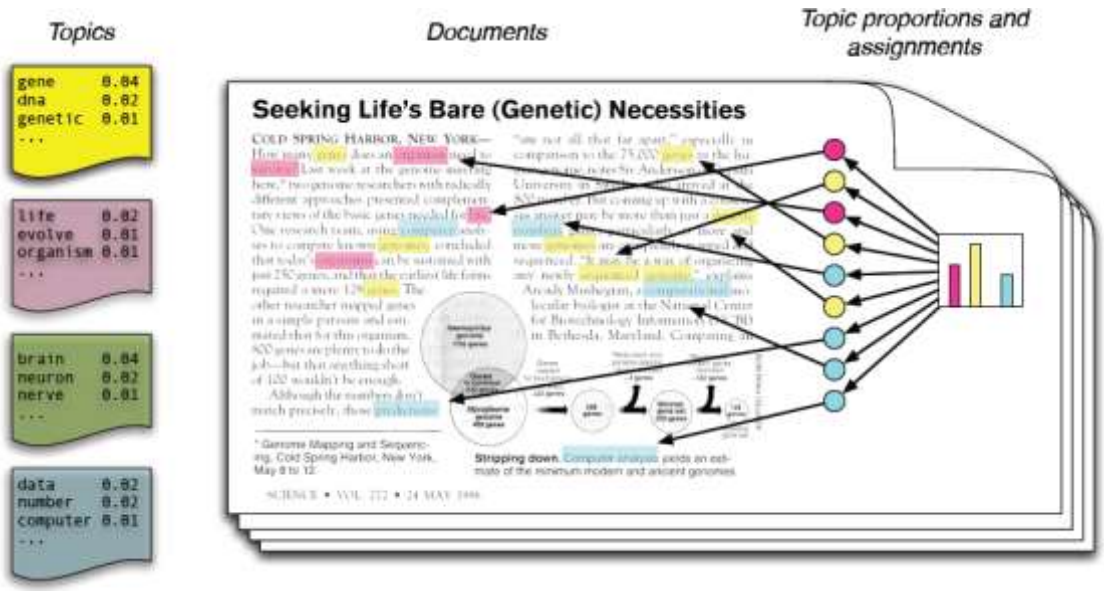
Samuti saab tabeli tulemusi esitada protsentuaalsel kujul:

A: 100% teema F

B: 60% teema F ja 40% teema P

Selle järgi saame öelda, et dokument A räägib toidu kohta ja dokument B millega kasse toita.

Teemade määramise protsessi korratakse iga sõna jaoks kõikide dokumentide vahel, mis liigub läbi kogu dokumentide kogumi mitu korda. See iteratiivne teemade uuendamise funktsioon on LDA võtmefunktsioon, mis loob sidusate teemadega lõpliku lahenduse. See iteratiivne ajakohastamine on LDA võtmefunktsioon, mis loob sidusate teemadega lõpliku lahenduse.



Joonis 2 LDA illustratsioon

Joonisel 2 [14] on kujutatud veel üks viis, kuidas ette kujutada teemade modelleerimise protsessi. Artiklit lugedes kasutame erinevaid värve võtmesõnade teemade paigutamise jaoks. Kui me oleme selle protsessiga lõpetanud, siis saame ühe värviga sõnad paigutada ühe sõnade loendi alla. See sõnade loend on teema ja iga värv kujutab endast erinevat teemat [14].

1.2.2 LSI

Teine tehnika, mida kasutame temaatiliste mustrite kaevandamiseks seadusetekstidest, on *Latent semantic indexing* (LSI). See on indekseerimis- ja otsimismeetod, mis kasutab ainsuse väärtus lagunemise meetodit (ingl k *singular value decomposition*), et tuvastada tekstikogumites sisalduvate mõistete ja nende vahelisi suhteid [19].

LSI tugineb põhimõttel, et ühes ja samas kontekstis kasutatavatel sõnadel on tavaliselt sarnased tähendused. LSI peamiseks tunnuseks on see, et ta suudab tekstielemendi kontseptuaalset sisu välja võtta, luues seoseid nende terminite vahel, mis esinevad sarnases kontekstis [20]. LSI jaoks dokumendid, millel on palju ühiseid sõnu, on semantiliselt lähedased, ning vastupidi, millel on vähe ühiseid sõnu, on semantiliselt kauged. Seepärast klassifitseerib see meetod üllatavalt sarnaselt inimesele ning kuigi LSI algoritm ei tea, mida sõnad antud dokumendikorpuses tähendavad, võib jääda mulje, et see oleks üllatavalt arukas [16].

Mustrite kaevandamiseks kasutab LSI mõningaid lihtsustusi:

1. Dokumendid on kujutatud sõnade kogudena, kus sõnade järjekord ei ole oluline. Tähtsaks peetakse seda, kui tihti esineb üks või teine sõna antud dokumendis.

2. Teemades esinevad sõnad ilmuvad ka dokumentides enamasti koos. Näiteks sõnad “kiirus”, “mootor” ja “gaas” võivad esineda dokumentides, kus räägitakse autodest.

3. Eeldatakse, et sõnadel on ainult üks tähendus. Mõnede näidete puhul ei ole see selge (näiteks palk võib tähendada nii puupalki kui ka rahalist palka), kuid see muudab probleemi juhitavaks.

1.2.3 HDP

HDP on mudel, kus teemad koosnevad erinevatest sõnadest, samuti nagu LDA puhul. Kui LDA puhul fikseeritakse, mitu teemat tahetakse dokumendikorpusest leida, siis HDP puhul tekitab teemade arvu Dirichlet’ protsess [21], mille tulemusena on teemade arv juhuslik muutuja. Nime "hierarhiline" osa viitab sellele, et genereeritavale mudelile lisandub teine tase (teemade arvu loomise protsess) [22].

1.3 Python, Matplotlib, Gensim

Python on lihtne ning võimas programmeerimiskeel [23], mida kasutatakse laialdaselt andmeteaduses. Python on populaarne kolme põhjuse pärast:

- Lihtne tõlgendada ja õppida
- Käitleb edukalt erinevaid andmestruktuure
- Omab võimsaid andmete visualiseerimise tööriistu

Antud töös kasutame Python versiooni 3.5.4.

Matplotlib on Pythoni peamine teaduslik joonestamisteek. See pakub kasutajale väga kvaliteetseid visualiseerimise võimalusi, nagu joondiagramme, histogramme, hajuvusdiagramme jne. Andmete visualiseerimine võib anda väärtuslikke teadmisi ning nende teadmiste põhjal on parem analüüsi läbi viia [24].

Antud töös kasutame Matplotlib versiooni 2.1.2.

Gensim on avatud lähtekoodiga Pythonis implementeeritud teema modelleerimise tööriistakomplekt. Gensim on spetsiaalselt loodud suurte tekstikogumite käitlemiseks kasutades voogesitust ja algoritme. See eristab Gensimi enamikest teistest teaduslikest tarkvarapakettidest, sest teised on mõeldud ainult partii ja mälu töötlemiseks [25].

Antud töös kasutame Gensimi versiooni 3.2.0.

1.4 EstNLTK

EstNLTK (NLTK ehk Natural Language ToolKit) on peamiselt Pythonis kirjutatud kogumik teeke eestikeelsete tekstide töötluseks. EstNLTK eesmärkideks on olemasolevate keeletehnoloogia tööriistade omavaheline liidestamine ja kättesaadavaks muutmine ning uute loomine [26].

EstNLTK kui projekti tuumaks on Pythoni teek, milles sisaldub: eesti keele sõnestamine ehk sõnapiiride tuvastamine ehk üksustamine (ingl k *tokenization*), eesti keele lausestamine ehk lausepiiride tuvastamine, eesti keele osalausestamine ehk osalausepiiride tuvastamine, eesti keele lemmatiseerimine ehk sõnade algvormide (lemmade) määramine ning morfoloogiline analüüs, sõnaliikide määramine ja palju muud.

Antud töös kasutame EstNLTK versiooni 1.4.1.1.

1.5 Jupyter Notebook

Jupyter Notebook, mis kandis kunagi nime IPython Notebook, on selline veebilehitsejas toimiv rakendus, mis lubab Pythoni (või mõne muu interaktiivse keelega) suhelda samalaadse kasutajaliidese vahendusel.

Toome välja mõned Jupyter Notebooki põhifunktsionaalsused:

- saab interaktiivselt käivitada lühikesi koodijuppe

- kogu sisend-väljund säilitatakse töölehel, kõiki eelnevaid sisestusi saab redigeerida ja vastavaid arvutusi korrata
- arvutuste vahele saab lisada rikkaliku kujundusega teksti, valemeid ja jooniseid

Jupyter Notebooki puhul on tegemist interaktiivse märkmeraamatuga, kus saab teha matemaatilisi eksperimente ja selle arvutuskäigu dokumenteerimist samaaegselt. Jupyter Notebooki on võimalik veebis katsetada ka nii, et midagi ei ole enne seda arvutisse installeeritud. See tähendab seda, et serverprogramm ja arvutusmootor ei pruugi asuda tingimata selles samas arvutis, kus programmi kasutatakse (läbi veebibrauseri) [27].

2 Uuringu tutvustus

Õiguslikke tekste pole teemade modelleerimisega väga palju uuritud, kuid sellegipoolest leiab selle valdkonnaga seotud kirjandust ja uuringuid. Autor toob välja mõned näited, kus on kasutatud teemade modelleerimist temaatiliste mustrite kaevandamiseks seadusetekstidest.

2.1 Varasemad rakendused

Täielikult automatiseeritud lähenemist temaatiliste mustrite kaevandamiseks seadusetekstidest pakuti välja artiklis [28]. Selle eesmärgiks oli määruste muutuste valdkonna jaoks ehitada automatiseeritud lahendus, kuna see valdkond oli viimastel aastatel saanud suure tähelepanu ohvriks oma keerukuse tõttu. Selle ülesande lahendamiseks kasutasid autorid Suurbritannia õiguslike tekstide korpust , mis koosnes 41518 dokumendist aastast 2000 kuni 2016, ning nendele andmetele ehitati erinevad mudelid. Antud uuringu käigus andsid häid tulemusi Saffron [29] ja standardne LDA mudel, siis kui põhiterminid valitakse seoses teemade spetsiifilisusega. See uuring on näide sellest, kuidas seaduse alal tegutsejad saaksid leida tähtsamad ja selged teemad kasutades automatiseeritud temaatiliste mustrite kaevandamise vahendit.

Teine LDA mudeliga seotud lähenemine toodi välja [30], kus põhiideeks oli automaatselt temaatiliste mustrite tuvastamine seadusetekstidest, mis on PDF-formaadis, ning samuti põhikonteksti kokkuvõtmine. Autor kasutas temaatiliste mustrite tuvastamise näiteks viieleheküljelist dokumenti, kus alustas tekstide välja võtmisega PDF-failist. Edasi tegeles autor andmete puhastamisega ning seejärel liikuda teemade modelleerimise juurde. Samuti andis autor visuaalse ülevaate antud uuringust. Antud lähenemine andis head tulemused ning saadud teemade ja sõnapilve (ingl k *wordcloud*) abil sai autor koheselt järeldusele jõuda, mis on antud dokumendi sisu.

Selle uuringu autorite [31] eesmärgiks oli organiseerida seaduslikud dokumendid erinevatesse klastritesse (ingl k *cluster*) kasutades koosinuse sarnasust dokumentide ja teemade vahel, et parandada infootsingu jõudlust. Infootsingu osas on tähtis otsida antud dokumenti talle ettemääratud klastri abil, mitte terve korpuse osas. Saadud tulemuste

põhjal jõudis autor järeldusele, et seaduslike dokumentide klasterdamist saab saavutada üksnes koosinuse sarnasuse leidmise abil dokumentide ja teemade vahel.

Neljas uuring, mis on seotud temaatiliste mustrite kaevandamisega, on kirjeldatud artiklis [32]. Artiklis kirjeldatud põhieesmärgiks oli Hiina seaduslike tekstide sarnasuse hindamine. Eksperimentide käigus hinnati LDA, LLDA (ingl k *Labeled Latent Dirichlet Allocation*) ja tf-idf (ingl k term *frequency-inverse document freqeuncy*) kaudu saadud mudeleid. LDA sai eksperimentide käigus paremaid tulemusi kui ülejäänud mudelid. Kuid antud uuringu käigus leiti, et tegu ei olnud täielikult automatiseeritud lähenemisega. Lähenemine vajas manuaalset sekkumist eeltöötluse ja mudeli parameetrite sätestamise osas. Samuti omab mudel lihtsustatud eeldusi, mis vajavad täiustamist.

2.2 Uuringusse kuuluvad õigusaktid

Õigusaktide [1] tutvustused pärinevad Riigiteataja [2] kodulehelt:

1. Investeerimisfondide seadus - Investeerimisfond on juriidiline isik või varakogum, millesse kaasatakse mitme investori kapital eesmärgiga seda vastavalt kindlaksmääratud investeerimispoliitikale kõnealuste investorite kasuks ja ühistes huvides investeerida. Investeerimisfondide moodustamist, asutamist ja valitsemist ning nende osakute, osade, aktsiate ja teiste sarnaste osalust väljendavate õiguste osalust reguleerib investeerimisfondide seadus ehk lühendina IFS, mis võeti vastu 14.12.2016. Investeerimisfondide seadus jõustus 10.01.2017.

   Süstemaatiline liigitus: Haldusõigus - Maksuõigus

2. Võlaõigusseadus - Võlaõigus ehk lepinguõigus moodustab ühe olulisema osa tsiviilõigusest ehk eraõigusest. Võlaõiguse valdkonda reguleeriv põhiseadus on võlaõigusseadus ehk lühendina VÕS, mis võeti vastu september 26.09.2001. Võlaõigusseadus jõustus 01.07.2002. Kuni selle momendini reguleeris lepingulisi suhteid nõukogude ajast pärit Tsiviilkoodeks.

   Süstemaatiline liigitus: Eraõigus - Võlaõigus

3. Tsiviilkohtumenetluse seadustik - Tsiviilkohtumenetluses vaadatakse läbi tsiviilasi, kui seaduses ei ole sätestatud teisiti. Tsiviilasi on eraõigussuhtest tulenev kohtuasi. Tsiviilvaidluste ring on suur ning sinna alla kuuluvad erinevatest lepingutest ja võlasuhetest tulenevad vaidlused, perekonna- ja pärimisasjad, asjaõigust puudutavad vaidlused, äri- ja mittetulundusühingute tegevuse ja juhtimise küsimused, pankrotiasjad ning tööõiguse küsimused. Tsiviilasjade menetlustoimingud tehakse toimingu tegemise ajal kehtiva seaduse järgi. Tsiviilasjade menetlemist reguleerib tsiviilkohtumenetluse seadustik ehk lühendina TsMS, mis võeti vastu 20.04.2005. Tsiviilkohtumenetluse seadustik jõustus 01.01.2006.

   Süstemaatiline liigitus: Kohtumenetlusõigus - Tsiviilkohtumenetlus

4. Kriminaalmenetluse seadustik - Kuritegude kohtueelse menetluse ja kohtumenetluse kord ning kriminaalasjas tehtud lahendi täitmisele pööramise kord reguleerib kriminaalmenetluse seadustik ehk lühendina KrMS, mis võeti vastu 12.02.2003. Kriminaalmenetluse seadustik jõustus 01.07.2004.

   Süstemaatiline liigitus: Kohtumenetlusõigus - Kriminaalmenetlus

5. Kindlustegevuse seadus - Kindlustustegevus on kindlustuslepingu alusel kindlustusvõtja või kindlustatu kindlustusriskide ülevõtmine ja kindlustusjuhtumi toimumise korral kahju hüvitamine, kokkulepitud rahasumma maksmine või lepingu täitmine muul kokkulepitud viisil. Kindlustustegevust ja kindlustusvahendust ning nende järelevalvet reguleerib kindlustustegevuse seadus ehk lühendina KindlTS, mis võeti vastu 10.06.2015. Kindlustegevuse seadus jõustus 01.01.2016.

   Süstemaatiline liigitus: Haldusõigus – Kindlustus

   Süstemaatiline liigitus: Karistusõigus - Väärteod

6. Äriseadustik - Äriseadustik ehk lühendina ÄS on Eesti Vabariigi seadus, mis korraldab ettevõtjate tegevust. Ettevõtja käesoleva seaduse tähenduses on füüsiline isik, kes pakub oma nimel tasu eest kaupu või teenuseid ning kellele kaupade müük või teenuste osutamine on püsiv tegevus, ning käesolevas

seaduses sätestatud äriühing. Äriseadustik võeti vastu 15.02.1995 ja jõustus 01.09.1995.

Süstemaatiline liigitus: Haldusõigus - Teabeõigus, andmekogud ja statistika

Süstemaatiline liigitus: Eraõigus - Äriühingud

7. Liiklusseadus - Liikluskorralduse Eesti teedel, liiklusreeglid, liiklusohutuse tagamise alused ja põhinõuded, tee omaniku kohustused ja teede rahastamise, teekasutustasu tasumise tingimused ja määrad, mootorsõidukite, trammide ja nende haagiste ning maastikusõidukite registreerimise ja neile esitatavad nõuded, juhtimisõiguse andmise, mootorsõidukijuhi töö- ja puhkeaja ning liiklusregistri korraldamise ja pidamise nõuded ning vastutuse liiklusreeglite rikkumise eest reguleerib liiklusseadus ehk lühendina LS, mis võeti vastu 17.06.2010. Liiklusseadus jõustus 01.07.2011.

   Süstemaatiline liigitus: Haldusõigus - Liiklus ja transport

   Süstemaatiline liigitus: Karistusõigus - Väärteod

8. Riigihangete seadus - Riigihanke korraldamise reeglid, riigihankega seotud isikute õigused ja kohustused, riikliku järelevalve ja haldusjärelevalve tegemise, vaidlustuste lahendamise korra ning vastutuse käesoleva seaduse rikkumise eest reguleerib riigihangete seadus ehk lühendina RHS, mis võeti vastu 14.06.2017.

   Süstemaatiline liigitus: Haldusõigus - Riigihanked

9. Halduskohtumenetluse seadustik - Halduskohtumenetlus on menetlus haldusasja lahendamiseks. Haldusasi on halduskohtus lahendatav kohtuasi. Halduskohtu pädevus ning halduskohtusse pöördumise ja halduskohtumenetluse kord niivõrd, kuivõrd see ei ole reguleeritud teiste seaduste ning otsekohaldatavate välislepingute ja Euroopa Liidu õiguse normidega reguleerib halduskohtumenetluse seadustik ehk lühendina HKMS, mis võeti vastu 27.01.2011. Halduskohtumenetluse seadustik jõustus 01.01.2012.

Süstemaatiline liigitus: Kohtumenetlusõigus - Halduskohtumenetlus

10. Karistusseadustik - Karistusseadustik ehk lühendina KarS on süütegude temaatikat käsitlev Eesti seadus. Seadustik võeti vastu 06.06.2001 ja jõustus 01.09.2002. Enne karistusseadustikku kehtis Eestis kriminaalkoodeks ja karistusseadustik põhineb suuresti kriminaalkoodeksil.

    Süstemaatiline liigitus: Karistusõigus - Karistusõiguse üldregulatsioon

11. Riigilõivuseadus - Riigilõiv on seaduses sätestatud juhul ja käesolevas seaduses sätestatud määras tasutav summa lõivustatud toimingu tegemise eest. Riigilõivu kehtestamise, tasumise, tasumise kontrollimise ja tagastamise korra aluseid ning sätestab riigilõivuvabastused, riigilõivumäärad ja tehinguväärtuse määramise korra reguleerib riigilõivuseadus ehk lühendina RLS, mis võeti vastu 10.12.2014. Riigilõivuseadus jõustus 01.01.2015 ja osaliselt 01.01.2018.

    Süstemaatiline liigitus: Haldusõigus - Maksuõigus

12. Kaitseväeteenistuse seadus - Kaitseväekohustus on Eesti kodaniku kohustus osaleda riigikaitses ja käesolevas seaduses sätestatud toimingute tegemises. Isik, kellel on kaitseväekohustus, on kaitseväekohustuslane. Kaitseväeteenistus on kaitseväekohustuslase teenimine sõjaväelise auastmega ametikohal. Kaitseväeteenistuses olev isik on kaitseväelane. Kaitseväekohustuse, kaitseväeteenistuse ja asendusteenistuse subjektid, nende õigusliku seisundi, kaitseväekohustuse täitmise korralduse ning kaitseväelaste suhtes rakendatavad ergutused, distsiplinaarvastutuse alused, distsiplinaarkaristused ja distsiplinaarmenetluse korra reguleerib kaitseväeteenistuse seadus ehk lühendina KVTS, mis võeti vastu 13.06.2012. Kaitseväeteenistuse seadus jõustus 01.04.2013.

    Süstemaatiline liigitus: Haldusõigus - Avaliku teenistuse eriregulatsioonid

    Süstemaatiline liigitus: Haldusõigus - Pensionide eriregulatsioonid

    Süstemaatiline liigitus: Haldusõigus - Riigikaitse

13. Väärtpaberituru seadus - Väärtpaber käesoleva seaduse tähenduses, ka selle kohta dokumenti väljastamata, on järgmine vähemalt ühepoolse tahteavalduse alusel üleantav varaline õigus või kohustus või leping. Väärtpaberite avalikku pakkumist ja nende reguleeritud väärtpaberiturul kauplemisele võtmist, investeerimisühingute tegevust, investeerimisteenuste osutamist, aruandlusteenuse osutamist, reguleeritud väärtpaberituru ja väärtpaberiarveldussüsteemi toimimist, järelevalve teostamist väärtpaberituru ja selle osaliste üle ning nende vastutust reguleerib väärtpaberituru seadus ehk lühendina VPTS, mis võeti vastu 17.10.2001. Väärtpaberituru seadus jõustus 01.01.2002 ja osaliselt 01.05.2004.

    Süstemaatiline liigitus: Haldusõigus – Rahandus

    Süstemaatiline liigitus: Karistusõigus - Väärteod

14. Audiitortegevuse seadus - Vandeaudiitor on isik, kes on sooritanud arvestusala eksperdi kutseeksami raamatupidamise ja vandeaudiitori eriosa, kellele on antud vandeaudiitori kutse ning kes on andnud vande. Majandusarvestusala audiitortegevuse õiguslikud alused, vandeaudiitorile ja vandeaudiitorite ühingule esitatavad nõuded, vandeaudiitori ja vandeaudiitorite ühingu tegevuse õiguslikud alused ja vastutus, atesteeritud siseaudiitorile ja avaliku sektori siseaudiitorile esitatavad nõuded, siseaudiitorite tegevuse õiguslikud alused avaliku sektori ja avaliku huvi üksustes, auditi ja ülevaatuse kohustus, auditikomitee tegevuse alused ning siseaudiitorite õigus kutsetegevuseks, Audiitorkogu õiguslik seisund, pädevus ja vastutus, samuti tegevuse ja töökorralduse ning rahastamise alused, siseaudiitorite, vandeaudiitorite, vandeaudiitorite ühingute ja Audiitorkogu üle järelevalve teostamise alused, audiitortegevuse registri tegevuse alused reguleerib audiitortegevuse seaduse ehk lühendina AudS, mis võeti vastu 27.01.2010. Audiitortegevuse seaduse jõustus 08.03.2010.

    Süstemaatiline liigitus: Haldusõigus - Majandustegevuse üldregulatsioon

    Süstemaatiline liigitus: Haldusõigus – Rahandus

    Süstemaatiline liigitus: Haldusõigus - Vabad elukutsed

15. Välismaalaste seadus - Välismaalaste Eestisse saabumise, Eestis ajutise viibimise, elamise ja töötamise aluseid ning vastutust käesolevas seaduses sätestatud kohustuste rikkumise eest reguleerib välismaalaste seadus ehk lühendina VMS, mis võeti vastu 09.12.2009. Välismaalaste seadus jõustus 01.10.2010.

    Süstemaatiline liigitus: Haldusõigus - Rahvastikuõigus

16. Väärteomenetluse seadustik - Väärtegu kujutab endast pisemate rikkumiste menetlemist. Väärtegude kohtuvälise menetluse ja kohtumenetluse kord ning väärteo eest kohaldatud karistuse täitmisele pööramise kord reguleerib väärteomenetluse seadustik ehk lühendina VTMS, mis võeti vastu 22.05.2002. Väärteomenetluse seadustik jõustus 01.09.2002.

    Süstemaatiline liigitus: Kohtumenetlusõigus - Väärteomenetlus

17. Atmosfääri kaitse seadus - Atmosfääriõhk koosneb troposfääri, stratosfääri ja mesosfääri õhukihist, mis ulatub kuni 100 kilomeetri kõrguseni maapinnast. Välisõhu keemilise ja füüsikalise mõjutamise kohta esitatavad nõuded, meetmed välisõhu kvaliteedi säilitamiseks ja parandamiseks, osoonikihi kaitsmise nõuded, meetmed kliimamuutuste leevendamiseks ja kasvuhoonegaaside heitkoguste vähendamiseks, riikliku järelevalve korralduse käesolevas seaduses sätestatud nõuete täitmise üle, vastutuse käesolevas seaduses sätestatud nõuete täitmata jätmise eest reguleerib atmosfääriõhu kaitse seadus ehk lühendina AÕKS, mis võeti vastu 15.06.2016. Atmosfääriõhu kaitse seadus jõustus 01.01.2017.

    Süstemaatiline liigitus: Keskkonnaõigus – Keskkonnakaitse

    Süstemaatiline liigitus: Karistusõigus - Väärteod

18. Finantskriisi ennetamise ja lahendamise seadus - Finantskriis on olukord, kus raha pakkumine ületab raha nõudlust. Sisuliselt mõistetakse selle all olukorda, kus likviidsus ehk maksevalmidus muutub olematuks, kuna vaba raha suunatakse pankadest välja. See omakorda sunnib pankasid müüma teisi investeeringuid, et korvata puudujääke, sest muidu kukuvad kriisi sattunud

pangad kokku. Krediidiasutuste ja investeerimisühingute suhtes kriisiennetusmeetmete ning kriisilahendusmeetmete ja -õiguste rakendamist, kui on oht, et nende finantsolukord võib kiiresti halveneda, või on tõenäoline, et nad on maksejõuetud või võivad muutuda tulevikus maksejõuetuks reguleerib finantskriisi ennetamise ja lahendamise seadus ehk lühendina FELS, mis võeti vastu 18.02.2015. Finantskriisi ennetamise ja lahendamise seadus jõustus 29.03.2015.

Süstemaatiline liigitus: Haldusõigus – Rahandus

Süstemaatiline liigitus: Karistusõigus - Väärteod

19. Elektroonilise side seadus - Nõuded üldkasutatavatele elektroonilise side võrkudele ja teenustele, elektrooniliste kontaktandmete otseturustuseks kasutamisele, raadioside pidamisele, raadiosageduste ja numeratsiooni haldamisele ja raadioseadmele ning riiklik järelevalve nende nõuete täitmise üle ja vastutus nende nõuete rikkumise eest reguleerib elektroonilise side seadus ehk lühendina ESS, mis võeti vastu 08.12.2004. Selle seaduse eesmärk on luua elektroonilise side arenguks vajalikud tingimused, et soodustada elektroonilise side võrkude ja teenuste arengut konkreetseid tehnoloogiaid eelistamata ning tagada elektroonilise side teenuse kasutajate huvide kaitse vaba konkurentsi soodustamise teel ja raadiosageduste ning numeratsiooni otstarbekas ja õiglane planeerimine, eraldamine ning kasutamine. Elektroonilise side seadus jõustus 01.01.2005.

    Süstemaatiline liigitus: Haldusõigus - Post ja side

    Süstemaatiline liigitus: Karistusõigus - Väärteod

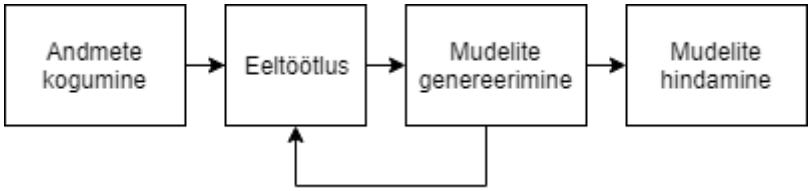
20. Krediitasutuste seadus - Krediidiasutus võib tegutseda aktsiaseltsi või tulundusühistuna ja temale kohaldatakse aktsiaseltsi või hoiu-laenuühistu kohta sätestatut, kui käesolevast seadusest ei tulene teisiti. Krediidiasutuse asutamist, tegevust, lõpetamist, vastutust ning järelevalve teostamist krediidiasutuse üle reguleerib krediitasutuste seadus ehk lühendina KAS, mis võeti vastu 09.02.1992.

Süstemaatiline liigitus: Haldusõigus – Rahandus

Süstemaatiline liigitus: Karistusõigus – Väärteod

3 Uuring

Antud uuringus osales kakskümmend suuremat õigusakti, mille andmed saadi Riigiteataja [2] kodulehelt. Need kakskümmend suuremat õigusakti moodustasid umbes 30% (täpsemalt 28.7%) ülejäänutest seadustest tekstimahu poolest. Seaduste tekstimahu suurused leidis lõputöö juhendaja. Antud eksperimendi käigus kasutati kakskümmend kõikidest õigusaktidest selle tõttu, et kasutada iga õigusakti süstemaatilist liigitust analüüsi tegemise jaoks ning jääda töö piiridesse, kuna seadused on mahukad ning keerulised.



Joonis 3 Tekstikaeve protsess

Antud lõputöös kasutatakse tekstikaeve standardprotsessi, mis on kirjeldatud antud raamatus [37], kuid meie lõputöö jaoks viime sisse antud uuringu jaoks vajalikke muudatusi nagu näha ülemisel joonisel. Esimeseks osaks on dokumentide kogumine, kus kogutakse kõik uuringusse kuuluvad dokumendid kokku. Teises sammus toimub andmete eeltöötlus. Selles faasis puhastakse dokumentide andmed nii, et me saaks algoritme rakendada. Andmete puhastamise osasse kuuluvad stoppsõnade, erinevate märkide ning numbrite eemaldamine. Selles samas sammus töötleme me dokumente nii, et muuta seda veelgi arusaadavamaks. Selles faasis tehakse lemmatiseerimist, segmenteerimist. Selle lõputöö puhul teeme me lemmatiseerimise ning sõnade üksustamise faasi juba sel hetkel, kui me andmed programmi sisse loeme. Järgmise osana eeltöötluse sammus tehakse dokumentide ümberkujundamine selleks, et see tekst oleks algoritmide jaoks loetavas vormingus. Me kõik teame, et masinloetavad algoritmid ei suuda teksti tõlgendada. Nad saavad vaid hinnata reaalarvude vormingut. Selles faasis saame me kasutada erinevaid lähenemisviise, nagu näiteks sõna vektorid, uni-grammid ja nii edasi. Eelviimane osa kujutab endast modelleerimist, kus genereerime kolme erineva algoritmi abil mudeleid. Viimasena toimub mudelite hindamine. Selles faasis me tuvastame ka algoritmi, mida edaspidi kasutame (meeles peab pidama seda, et algoritm ja selle tulemus oleks

tõlgendatav). Mudelite hindamist saab teostada erinevate tehnikate abil: ristvalideerimine, koherentsuse arvutamine ja nii edasi. Samuti on võimalik kasutada ka erinevaid visualiseerimise tööriistu [37].

3.1 Eeltöötlus

Pärast esimest sammu ehk andmete kogumist Riigiteataja [2] kodulehelt ja integreerimist, on vaja mudeli ehitamise jaoks läbida eeltöötluse sammud. Andmete puhastamine on tähtis kasuliku mudeli genereerimiseks. Need sammud on vajalik läbida selle jaoks, et andmed, millega me edasi tegeleme, oleks sobivad analüüsi osa jaoks. Allpool kirjeldame samme, mida peame vajalikuks meie tekstide korpuse jaoks. Esimese sammuna kirjeldame sõnade üksustamist ehk tokeniseerimist (ingl k *tokenization*), mis hõlmab endast teksti jagamist lauseteks, paragrahvideks või sõnadeks. Me jagasime teksti sõnadeks. Samuti on vaja viia sõnad lemmatiseeritud kujule, kuid need mõlemad osad tegime me ära andmete integreerimise ajal. Järgmisena tuleb eemaldada kõik ebavajalikud sõnad, stoppsõnad ja numbrid, mis ei anna mingit lisainformatsiooni juurde. Samuti tuleb andmeid viia masinloetavale kujule, kuid see teostatakse eraldi etapina. Pärast seda on võimalik mudeleid genereerida.

3.1.1 Sõnade üksustamine

Sõnade üksustamine on viis jagada sõnu nende üksuste järgi, mida ei peeta kõige raskemaks väljakutseks keeletöötluse osas. Lihtsamalt öeldes on vaja dokumentide sisust üles leida sõnad [15]. Kuid samuti need üksused võivad olla ka laused või paragrahvid. Tavaliselt on see esimene osa keeletöötlusest ning samuti omab see märkimisväärset mõju järgnevate keeletöötluse etappide jaoks.

Keeleteaduses on morfoloogia teatud keele morfeemide ja teiste keeleliste üksuste struktuuri tuvastamine, analüüsimine ja kirjeldamine, näiteks sõnajuured, lemmad, sufiksid, kõneosad jne [33]. EstNLTK omab võimalust morfoloogilise analüüsi ja sünteesi jaoks tänu Vabamorf morfoloogilisele analüsaatorile.

Samuti enne sõnade üksustamist transformeerisime kõik sõnad nende algkujule, et andmed, millega me edasi toimetasime, oleks kasulikumad. Selleks me kasutasime

EstNLTK morfoloogilise analüüsi funktsiooni *lemmas* [26]. Enne seda kasutame funktsiooni *text*, mis teeb meie jaoks ära sõnade üksustamise.

3.1.2 Stoppsõnade eemaldamine

Kuigi stoppsõnade kohta öeldakse, et nad on kõige sagedamini esinevad sõnad keeles, siis tegelikkuses ei ole ühtegi kindlat universaalset listi nende kohta [34]. Stoppsõnad eemaldatakse andmete hulgast enne või pärast keeletöötlust.

Alguses kustutasime sõnu, mis esinevad eesti keeles väga sagedalt, nagu näiteks "aga", "ei", "et", "ja", "jah" ja nii edasi. Selleks me kasutasime [35], kus olid mitmed sõnad juba olemas. Sellegipoolest polnud see list piisav meie andmete jaoks. Seetõttu pidime me lisama korduvaid sõnu, mis ei andnud juurde mingit lisainformatsiooni (need sõnad olid pigem seadustele spetsiifilised): "leping", "seadus", "lõik", "käesolev", "isik" ja nii edasi. Pärast seda oli meie andmed stoppsõnadest puhastatud ning oli võimalik liikuda järgmise sammu juurde.

3.1.3 Muud eemaldamised

Õigusaktides esineb palju numbreid. Nende eemaldamiseks kasutasime Pythoni [23] *isnumeric()* meetodit, kuid siinkohal tuleb mainida, et see eemaldas numbrid, mis ei ole osa sõnast, vaid on eraldiseisvad. Seetõttu kasutasime teist meetodit *isalnum()*, mille abil kõik kuupäevad, viited ja ülejäänud numbrid saime andmetest eemaldatud, mis ei ole mudeli genereerimisel sugugi kasulikud.

Pärast andmete uurimist otsustasime, et tuleb eemaldada ka lühemad sõnad, mis ei anna sisule midagi kasulikku juurde. Me eemaldasime andmetest sõnad, mis olid lühemad kui kolm tähemärki.

Samuti eemaldasime sõnad, mis on esinevad harva, ning ka sõnad, mis esinevad tihti. Selleks me kasutasime sellist sagedust, mille puhul me vaatame, kui tihti esinevad antud sõnad erinevates dokumentides ehk õigusaktides. Me otsustasime eemaldada sõnu, mis esinevad rohkem kui 7 dokumendis. Selleks me kasutasime Gensimi [25] funktsiooni *filter.extremes()*. Tänu sellele me filtreerisime välja väga sagedased sõnad, mis ei olnud meil stoppsõnade listis, nagu näiteks "kehtivus", "jagu", "vastuvõtmine".

3.2 Mudelite genereerimine

Pärast eeltöötluse samme ei ole veel võimalik mudeleid genereerida. Enne tuleb viia andmed sellisele kujule, et need oleks algoritmide jaoks loetavad. Selleks me teisendasime dokumentide sõnastiku *bag-of-words* formaati. Selleks me kasutasime Gensimi [25] doc2bow implementatsiooni. Nüüd oli võimalik järgmises etapis mudeleid genereerida. Selle osa jaoks kasutasime Gensimi [25] *LdaModel*, *LsiModel* ja *HdpModeli* implementatsioone. Mudelite genereerimise osas on oluline otsustada parameetrite suhtes.

Esimene asjana on vaja otsustada, et mitu teemat soovime me tuvastada. Kahjuks ei ole sellele küsimusele õiget vastust ja heade tulemuste saavutamine tuleb proovimise teel. Me katsetasime 4 kuni 20 teema avastamise vahel. HDP mudeli genereerimise puhul pole teemade arvu vaja ette anda. LDA ja LSI puhul oli märgata, et kui teemade arv ületas arvu 12, siis olid tekkinud teemad juba liiga väikese osakaaluga ning selle tõttu otsustasime edaspidi teha analüüsi kuni 12 teemani. Arvatavasti üle 12 teema avastamine ei toonud antud dokumendikorpuse suhtes häid tulemusi sellepärast, et meie dokumendikorpuses olid vaid 20 õigusakti ning kahekümnele dokumendile ei ole mõistlik väga palju erinevaid teemasid avastada, kuna selle käigus ei leia me enam tegelikult varjatud mustreid. Mudeleid ehitasime kogu dokumendikorpusele ehk kahekümnele õigusaktile.

3.3 Mudelite hindamine

Pärast mudelite genereerimist liikusime me edasi mudelite hindamise osa juurde. Kuigi teemade modelleerimine on võimas protsess tekstides esinevate teemade automaatseks tuvastamiseks ja tekstikorpuse peidetud mustrite leidmiseks, siis võib selle mudeli hindamise ja tõlgendamise osas tulla ette raskusi. Mudelite hindamiseks kasutati koherentsust, mis vaatleb koostatud teemade sõnade kogumit ja hindab nende teemade tõlgendatavust. Samuti kasutati kvalitatiivset hinnangut, et saada arvamus mudelite kohta ka seadustega igapäevaselt tegeleva inimese käest.

3.3.1 Koherentsus

Teema koherentsuseks peetakse kvantitatiivset lahendust, mille eesmärgiks on hinnata mudeleid nende inim mõistetava tõlgendatavuse järgi. Teema koherentsus vaatleb koostatud teemade sõnade kogumit ja hindab nende teemade tõlgendatavust [39].

Artiklis [38] on kirjeldatud teema koherentsust niimoodi, et kujutame ette, et me saame vett paljudest erinevatest kohtadest. Viis, kuidas me seda vett testime, põhineb sellel, et pakume seda vett paljudele inimestele ning küsime neilt hiljem tagasisidet selle kohta. Kui paljud inimesed ütlevad, et vesi on hea, siis me arvame, et vesi on hea. Kui aga paljud inimesed ütlevad, et vesi on halb, siis me arvame vastupidist. Sellel juhul meie hinnang sõltub tagasiside abil, mida me saame inimeste käest. Kuid sellel juhul ei oska me vee kvaliteeti täpselt hinnata ning seepärast saab seda nimetada kvalitatiivseks hinnanguks.

Selle parandamise jaoks paigaldame me neli erinevat toru algallika külge, kus igale anname numbrilise väärtuse selle kohta, mis kvaliteediga vesi on. Kui nüüd on see paigaldatud, siis ei pea me enam tuhandetelt inimeselt eraldi küsima vee kvaliteeti, vaid me saame selle väärtuse täpselt seoses nende väärtusega tänu neljale paigaldatud torule.

Vesi on selles olukorras teemad, mida me saame teemade modelleerimise algoritmide käest. Väärtus, mis veega kaasas käib, ning torud moodustavad teema koherentsuse. Neli toru mis me selle jaoks kasutasime on:

- Segmentatsioon (ingl k *segmentation*): kus vesi paigutatakse erinevatesse klaasidesse ning iga klaasi vee kvaliteet erineb üksteisest
- Tõenäosuse arvutus (ingl k *probability estimation*): kus iga klaasi vee hulk on mõõdetud
- Kinnitusmeede (ingl k *confirmation measure)*: kus vee kvaliteet igas klaasis on mõõdetud ja numbriline väärtus seostatakse iga klaasiga
- Liitmine (ingl k *aggregation*): kus vee kvaliteedi numbrilised väärtused liidetakse kindla valemi järgi, et saada vastuseks üks kindel väärtus

See oli lihtsam selgitus teema koherentsuse saamise kohta. Mudelite koherentsuse arvutamisel kasutasime ka sama artikli autori abi koodi osas [38]. Üldiselt on teema

koherentsus viis teemade modelleerimise mudelite hindamiseks. Mudelite hindamiseks kasutasime koherentsust selle tõttu, et see võtab arvesse sõnade konteksti ning semantilisi seoseid. Koherentsuse arvutamiseks on mitmeid meetmeid, kuid *Cv* on neist kõige paremini kooskõlas inimese tõlgendatavusega [39]. Gensim [25] kasutab *Cv* vaikimisi.

3.3.2 Kvalitatiivne hinnang

Käesolevas magistritöös kasutame ka kvalitatiivset hinnangut, et uuringu käigus saada inimese arvamust antud uuringu kohta. Selleks kasutasime palju abi antud allikast [41]. Samuti oli küsimuste loomise eeskuju võetud selle lõputöö järgi, kuid me muutsime küsimusi artikli [31] abil, et see oleks meie lõputöö teemaga piisavalt hästi seotud. Küsimuste abil tahtsime saada vastuseid enamasti küsimustele, et kas sellest uuringust oleks kasu kellegi jaoks ning kas saadud tulemused võiksid kasulikud olla. Kvalitatiivsete hinnangute kasutamine on kasulik sel juhul, kui täpse uuringu kohta on informatsiooni vähe või see teave, mis selle käigus saadakse, aitab lõputööle palju kaasa. Kvalitatiivsete hinnangute saamisel kasutatakse intervjuust saadud andmeid kohe edaspidise analüüsi jaoks. Kuna saadud informatsioon võib varieeruda, siis ei ole võimalik saadud andmeid ette planeerida. See on selle tõttu, et intervjuust saadud informatsioon võib sisaldada selliseid detaile, mida töö autor ei osanud ette planeerida [40]. Nii kui saadakse andmeid intervjuu käigus, siis üritatakse neid kohe analüüsida. See võimaldab täpsustavate küsimuste esitamist, mis aitaks lõputööle kaasa.

Eestis on võimalik õppida õigusteadust mitmes õppeasutuses: Tallinna Tehnikaülikool, Tartu Ülikool, Tallinna Ülikool. Kõikides õppeasutustes on nominaalne õppeaeg 3 aastat ning on olemas nii päeva- kui ka kaugõppe võimalus. Samuti pakub Tartu Ülikool avatud ülikooli võimalust. Igas asutuses saab spetsialiseeruda Eesti õigusele. Seminarides omandatakse õigusaktide tõlgendamise ja praktilise rakendamise oskust. Lisaks auditoorsele tööle teevad üliõpilased iseseisvat tööd seminarideks valmistumisel ning kirjalike üliõpilastööde koostamisel.

Antud lõputöös on kvalitatiivne hinnang vajalik mudelite hindamisel, et leida sobivam mudel nii arvuliste väärtuste kui eksperdi koosmõjul. Kuna antud lõputööga seoses olid olemas sarnased uurimistööd, siis kasutati uurimise läbiviimiseks poolstruktureeritud intervjuud.

Intervjueeritavaks osutus Tartu Ülikooli 3.kursuse üliõpilane. Intervjueeritav omab õigusaktide tõlgendamise ja praktilise rakendamise oskus, mis on ka oluline selle lõputöö jaoks. Intervjueeritav oli nõus intervjuus vabatahtlikult osalema. Intervjuu läbiviimise aja leppisid lõputöö autor ja intervjueeritav omavahel kokku.

4 Tulemused

Lõputöö uuringut alustasime 20 õigusaktiga ning nende andmete integreerimise ja puhastamisega. Esimesed kakskümmend õigusakti moodustasid tekstimahu poolest umbes 30% (täpsemalt 28.7%) kõikidest (361) õigusaktidest. Antud eksperimendi käigus kasutati väiksemat osa kõikidest õigusaktidest selle tõttu, et kasutada iga õigusakti süstemaatilist liigitust analüüsi tegemise jaoks. Samuti, et püsida töö piirides, kuna seadused on väga mahukad ning keerukad. Enne mudelite hindamist viisime läbi mitmeid eksperimente, kus avastasime mudelite puhul erinevaid arvu teemasid. Samuti jälgisime, kas teemad on tõlgendatavad inimese jaoks, ning kui ei olnud, siis proovisime andmete puhastamisel teha täiustusi. Eesmärgiks oli saavutada head tulemused ning selle tõttu käisime pidevalt kahe sammu vahel. Allpool on toodud näitena koodi kolm esimest osa:

```
import os
import re
import operator
import matplotlib.pyplot as plt
import warnings
import gensim
import estnltk
import numpy as np
warnings.filterwarnings('ignore')
from estnltk import Text
from gensim.models import CoherenceModel, LdaModel, LsiModel, HdpModel
from gensim.models.wrappers import LdaMallet
from gensim.corpora import Dictionary
from pprint import pprint
```

```
#https://markroxor.github.io/gensim/static/notebooks/lda_training_tips.html
main = 'Õigusaktid/'

# kaustad koos Eesti õigusaktidega ning nende sisselugemine
folders = ['Atmosfääriõhu', 'Audiitor', 'Elektroonilise', 'Finantskriisi', 'Halduskohtu', 'Investeerimisfondide', 'Kaitseväe', 'K
directories = ['/' + folder for folder in folders]

# Read all texts into a list.
documents = []
for x in directories:
    docs = os.listdir(main + x)
    for doc in docs:
        with open(main + x + '/' + doc, errors='ignore', encoding="UTF-8") as docid:
            txt = docid.read()
            text = Text(txt)
            new = text.lemmas
        documents.append(new)
```

```
# http://www.tekstikaeve.ee/blog/2018-04-18-eestikeelsete-stoppsonade-loend/
# stoppsõnade sisselugemine ja eemaldamine dokumendikorpusest

files = open(path, errors='ignore', encoding="UTF-8")

file_list =  files.readlines()

final_list = []
for i in file_list:
    final_list.append(i.strip())

documents = [[token for token in document if not token in final_list] for document in documents]

other = ["osaline", "seadus", "riigikogu", "liik", "riigikogu", "nõukogu", "paragrahv", "jaotis", "kategooria", "käesolev", "akt"

documents = [[token for token in document if not token in other] for document in documents]

# numbritest koosnevate sõnade eemaldamine
documents = [[token for token in document if token.isalnum()] for document in documents]

documents = [[token for token in document if not token.isnumeric()] for document in documents]

# sõnade eemaldamine, mis on pikkuselt lühemad kui 2 tähemärki
documents = [[token for token in document if len(token) > 2] for document in documents]

# Gensimi dictionary genereerimine
dictionary = Dictionary(documents)

# sõnade väljafiltreerimine, mis esinevad üle 30% dokumentidest
dictionary.filter_extremes(no_below=1, no_above=0.3)
print(dictionary)
```

Joonis 4 Näide teemade modelleerimise koodist

Kogu kood koosneb kuuest erinevast osast. Joonisel on kujutatud kolm esimest osa. Esimeses plokis imporditakse vajalikud paketid. Teises osas loeb kood andmeid ning normaliseerib need. Normaliseerimine parandab andmete loetavust. Seejärel toimub andmete puhastamine, kus eemaldatakse erinevad stoppsõnad ning ülejäänud sõnad, mis ei anna mudeli ehitamisel midagi kasulikku juurde. Neljandas plokis ehitatakse mudelid. Esimene mudel on LDA, teine LSI ja kolmas HDP. Esimese ja teise mudeli puhul katsetasime 4 kuni 20 teema avastamise vahel. Samuti printisime välja antud avastatud

teemade võtmesõnad (ingl k *keywords*). Kolmanda mudeli puhul prindime nii palju teemasid, kui algoritm ise leidis. Pärast seda leiame kõigi kolme mudeli koherentsuse väärtuse, mille järgi hindame mudelite teostusvõimet. Sellele järgneb viies osa, kus leiame dokumentide ja teemade vahelisi seoseid. Viimases osas katsetame mudelit uute nägemata dokumentide peal.

4.1 Võtmesõnad

Pärast dokumendikorpuse sisu puhastamist ja normaliseerimist oli kokku 12467 unikaalset lemmatiseeritud sõna. Esialgu oli meil kõikide dokumentide peale 20377 unikaalset sõna, millest 281 olid stoppsõnad, 5664 olid numbritega sõnad ning 65 olid lühemad sõnad. Samuti filtreerisime välja sõnad, mis esinesid rohkem kui 7 dokumendis dokumentidest.

Me genereerisime LDA mudeli meie dokumentide korpusele. Me proovisime 4 kuni 20 teema avastamise vahel, kuid näitena toome tabelis välja kümme võtmesõna teemadest, kus teemade arv on 6. Võtmesõnade ees olev arv näitab sõnade tõenäosust antud teema all. Teemasid tõlgendused mõtlesime me ise välja, kuid igaüks saab neid enda moodi tõlgendada.

Tabel 5 LDA teemade võtmesõnad ja tõlgendused

Teema	**Võtmesõnad**	**Teema tõlgendus**
1	0.025*"makseteenus" + 0.024*"vangistus" + 0.015*"veksel" + 0.014*"kindlustusvõtja" + 0.013*"vedaja" + 0.011*"veos" + 0.010*"tarbijakrediidileping" + 0.009*"üürnik" + 0.008*"tšekk" + 0.008*"maksja	Kindlustuse ja üüriga seotud vangistus
2	0.051*"elamisluba" + 0.016*"viisa" + 0.015*"sideettevõtja" + 0.013*"järelevalvenõukogu" + 0.013*"audiitorkogu" + 0.011*"siseaudiitor" + 0.011*"sideteenus" + 0.008*"viibimisaeg" + 0.007*"audiitorettevõtja" + 0.007*"finantskonglomeraat	Elamisluba, viisa ning viibimisaeg

Teema	Võtmesõnad	Teema tõlgendus
3	0.040*"hankija" + 0.026*"hankeleping" + 0.023*"riigihange" + 0.021*"pakkumus" + 0.019*"osaühing" + 0.014*"kohtuistung" + 0.014*"ringkonnakohus" + 0.011*"kostja" + 0.010*"hankemenetlus" + 0.010*"apellatsioonkaebus	Riigihanked ja pakkumine
4	0.021*"inspektsioon" + 0.018*"süüdistatav" + 0.013*"emitent" + 0.013*"kaitsja" + 0.012*"heitkogus" + 0.012*"kohtuistung" + 0.011*"kahtlustatav" + 0.011*"õhukvaliteet" + 0.010*"ringkonnakohus" + 0.010*"saasteaine	Õhukvaliteet ja saasteaine, ning sellega seonduv inspektsioon ja süüdistus
5	0.021*"tegevväelane" + 0.014*"haagis" + 0.013*"juhiluba" + 0.012*"ülem" + 0.010*"tegevteenistus" + 0.009*"kaitseväelane" + 0.009*"sõidutee" + 0.009*"mootorsõidukijuht" + 0.008*"maanteeamet" + 0.008*"tehnonõue	Kaitseväeteenistus ja sõidukid
6	0.087*"fondivalitseja" + 0.037*"pensionifond" + 0.033*"osak" + 0.028*"finantsjärelevalve" + 0.022*"valitseja" + 0.019*"depositoorium" + 0.019*"fond" + 0.018*"aktsiaseltsifond" + 0.018*"eurofond" + 0.017*"kindlustusgrupp	Fondi aktsiate või osakute korraldamine

Nagu näha tabelist, siis see näitab võimalikke avastatud teemasid, kui teemade arvuks oli kuus. Teemadega käivad kaasas nende võtmesõnad ning tabelis on olemas 10 põhi võtmesõna iga teema kohta. LDA puhul on kohe näha, et avastatud teemad on kohe tõlgendatavad ning võib teha järeldusi kohe sellest, millest võib juttu tulla.

Nii nagu LDA puhul, siis genereerimise LSI mudeli samuti meie dokumentide korpusele. Me proovisime 4 kuni 20 teema avastamise vahel, kuid näitena toome tabelis välja kümme võtmesõna teemadest, kus teemade arv on 6.

Tabel 6 LSI teemade võtmesõnad ja tõlgendused

Teema	Võtmesõnad	Teema tõlgendus
1	0.063*"fondivalitseja" + 0.038*"hankija" + 0.027*"pensionifond" + 0.025*"osak" +	Fondid ja riigihanked

Teema	Võtmesõnad	Teema tõlgendus
	0.025*"hankeleping" + 0.022*"riigihange" + 0.020*"pakkumus" + 0.016*"valitseja" + 0.015*"depositoorium" + 0.014*"fond	
2	0.020*"makseteenus" + 0.012*"veksel" + 0.012*"kindlustusvõtja" + 0.011*"veos" + 0.011*"vedaja" + 0.009*"haagis" + 0.009*"juhiluba" + 0.008*"tarbijakrediidileping" + 0.007*"üürnik" + 0.007*"tšekk	Kindlustus ja üür
3	0.027*"inspektsioon" + 0.025*"vangistus" + 0.018*"emitent" + 0.009*"investeerimisteenus" + 0.009*"prospekt" + 0.008*"finantsjärelevalve" + 0.007*"finantskonglomeraat" + 0.007*"väärtpaberiturujärelevalve" + 0.006*"ülevõtmispakkumine" + 0.006*"kauplemissüsteem	Finants ning sellega seonduv vangistus
4	0.030*"elamisluba" + 0.021*"kohtuistung" + 0.019*"ringkonnakohus" + 0.016*"süüdistatav" + 0.011*"kaitsja" + 0.010*"kahtlustatav" + 0.009*"viisa" + 0.008*"kostja" + 0.008*"apellatsioonkaebus" + 0.008*"menetlusabi	Elamisluba ja viisa ning sellega seonduv karistus
5	0.031*"kindlustusgrupp" + 0.027*"sideettevõtja" + 0.021*"finantsjärelevalve" + 0.019*"sideteenus" + 0.017*"kindlustustegevus" + 0.012*"solventsuskapitalinõue" + 0.012*"omavahend" + 0.012*"kindlustusmaakler" + 0.011*"raadiosagedus" + 0.010*"lõppkasutaja	Kindlustus, side
6	0.035*"osaühing" + 0.026*"tegevväelane" + 0.018*"järelevalvenõukogu" + 0.017*"audiitorkogu" + 0.015*"sissemakse" + 0.015*"osakapital" + 0.015*"ülem" + 0.014*"siseaudiitor" + 0.013*"tegevteenistus" + 0.012*"kaitseväelane	Kaitseväeteenistus ja kapital

Nagu näha tabelist, siis see näitab võimalikke avastatud teemasid, kui teemade arvuks oli kuus. Teemadega käivad kaasas nende võtmesõnad ning tabelis on olemas 10 põhi võtmesõna iga teema kohta. Mõned teemad on avastanud pigem laia teemat, nagu näiteks “fondid ja riigihanked” või “kaitseväeteenistus ja kapital”, kus ei ole võimalik täpselt

paika panna, millest võib juttu täpsemalt olla. Samas teema “elamisluba ja viisa ning sellega seonduv karistus” abil saab kohe teha järeldusi, et mis sealt võib leida.

Samuti genereerisime HDP mudeli meie dokumentide korpusele. Näitena toome tabelis välja kümme võtmesõna teemadest, kus teemade arvuks on esimesed 6, kuna HDP leiab nii palju teemasid, kui palju ta ise leiab.

Tabel 7 HDP teemade võtmesõnad ja tõlgendused

Teema	**Võtmesõnad**	**Teema tõlgendus**
1	0.021*makseteenus + 0.019*süüdistatav + 0.013*veksel + 0.012*kindlustusvõtja + 0.012*kahtlustatav + 0.011*kaitsja + 0.011*vedaja + 0.010*veos + 0.008*tarbijakrediidileping + 0.008*kohtuistung	Makseteenustega seotud süüdistatavad
2	0.107*fondivalitseja + 0.046*pensionifond + 0.042*osak + 0.028*valitseja + 0.025*depositoorium + 0.023*fond + 0.023*aktsiaseltsifond + 0.022*eurofond + 0.020*finantsjärelevalve + 0.019*osakuomanik	Fondid
3	0.044*inspektsioon + 0.028*emitent + 0.025*heitkogus + 0.022*õhukvaliteet + 0.020*saasteaine + 0.016*kasvuhoonegaas + 0.014*prospekt + 0.013*investeerimisteenus + 0.013*kauplemissüsteem + 0.012*ühik	Õhukvaliteet ja saasteaine ning sellega seonduv inspektsioon
4	0.021*haagis + 0.021*juhiluba + 0.014*sõidutee + 0.013*mootorsõidukijuht + 0.012*maanteeamet + 0.012*tehnonõue + 0.012*liiklus + 0.011*robotliikur + 0.010*jalakäija + 0.010*sõitma	Liiklus ja sõidukid
5	0.032*kohtuistung + 0.031*kostja + 0.028*ringkonnakohus + 0.026*hageja + 0.025*vahekohus + 0.018*menetlusabi + 0.016*apellatsioonkaebus + 0.015*maksekäsk + 0.013*kassatsioonkaebus + 0.012*tsiviilasi	Kohus
6	0.125*hankija + 0.081*hankeleping + 0.071*riigihange + 0.065*pakkumus + 0.032*hankemenetlus + 0.024*alusdokument + 0.021*vaidlustuskomisjon + 0.019*ehitustöö + 0.016*vaidlustus + 0.015*kontsessioonileping	Riigihanked ja pakkumine

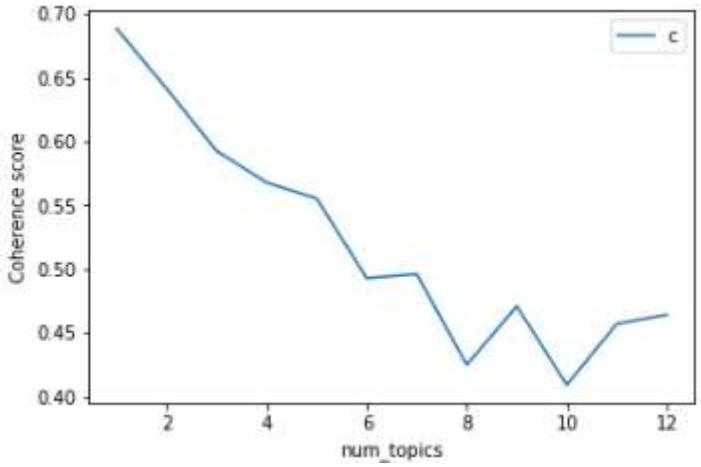
Nagu näha tabelist, siis see näitab võimalikke avastatud teemasid, kui teemade arvuks oli kuus. Kuna HDP avastas 20 teemat, siis tabelis esitame 6 esimest teemat. Teemadega käivad kaasas nende võtmesõnad ning tabelis on olemas 10 põhi võtmesõna iga teema

kohta. HDP kõikide 20 teema nägemisel oli võimalik kohe tähele panna, et mõned teemad hakkasid korduma ning seetõttu oli keeruline teemasid eristada üksteisest.

4.2 Mudelite koherentsus

Meie mudelite hindamiseks kasutasime me teemade koherentsust. Mudelite koherentsuse saamiseks kasutasime Gensimi *models.coherencemodel* moodulit, mis ongi mõeldud mudelite koherentsuse arvutamiseks. See implementatsioon on neljaastmeline mudelite koherentsuse arvutamise viis, mida mainisime punktis 3.3.1. Joonisel avastatud teemasid on kuni 12, kuna eelmise punkti eksperimentide käigus jõuti järeldusele, et peale 12 teemade avastamist läksid tulemused inimese tõlgendatavuse jaoks halvaks. Nii LDA kui ka LSI puhul tegime 10 katset ning võtsime arvesse parimad tulemused analüüsi jaoks.

1. Tulemused LDA näitel:



Joonis 5 LDA koherentsuse tulemused graafiliselt

Joonisel 2 on näha LDA koherentsuse tulemusi graafiliselt. Kuid selleks, et teha konkreetseid järeldusi, viidi antud joonise andmed tabelikujule.

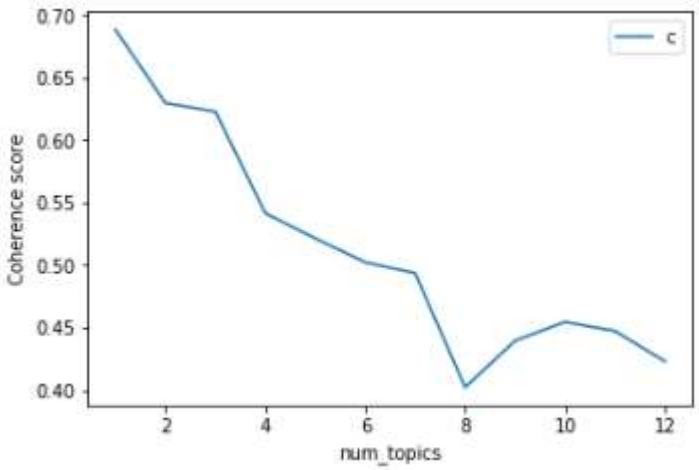
Tabel 8 LDA koherentsuse tulemused

LDA teemade arv	**Koherentsuse tulemus**
5	0.5552
6	0.4928

LDA teemade arv	**Koherentsuse tulemus**
7	0.4959
8	0.4248
9	0.4707
10	0.4092
11	0.4569
12	0.4640

Mida nullile lähemal koherentsuse tulemus on, seda vähem kasulikum on mudel. Samuti põhjus, miks tabeli kujul ei pandud kirja koherentsuse väärtusi kuni 4 teema avastamiseni, oli see, et nii LDA kui LSI tulemused olid sinnamaani peaaegu, et ühtlased.

2. Tulemused LSI näitel:



Joonis 6 LSI koherentsuse tulemused graafiliselt

Joonisel 3 on näha LSI koherentsuse tulemusi graafiliselt.. Kuid selleks, et teha konkreetseid järeldusi, viidi antud joonise andmed samuti tabelikujule.

Tabel 9 LSI koherentsuse tulemused

LSI teemade arv	**Koherentsuse tulemus**
5	0.5215
6	0.5022
7	0.4936
8	0.4022
9	0.4395
10	0.4545
11	0.4473
12	0.4229

3. Tulemused HDP näitel:

Tabel 10 HDP koherentsuse tulemused

HDP teemade arv	**Koherentsuse tulemus**
20	0.4045

Antud tabelite põhjal on näha meie poolt genereeritud mudelite koherentsuse väärtused. Et saada parem ülevaade keskmisest tulemusest, siis tekitame uue tabeli, kus paneme kõikide mudelite keskmise väärtuse kirja. HDP puhul on väärtus üks, kuid LDA ja LSI puhul arvutame keskmise väärtuse 5-12 teemade arvu vahel.

Tabel 11 Mudelite tulemused

Mudel	**Koherentsuse keskmine tulemus**
LDA	0.4712
LSI	0.4605
HDP	0.4045

Nagu tabelist näha, siis LDA mudel sai kõige parema koherentsuse väärtuse tulemuse. See väärtus on seetõttu parem, et LDA leidis dokumendikorpuses paremad teemad, mis on inimeste jaoks paremini tõlgendatavad. Muidugi ei saa alati usaldada ainult arvulisi väärtusi ning seetõttu soovisime saada ka kvalitatiivset hinnangut mudelitele.

4.3 Kvalitatiivne hinnang

Intervjuu viidi läbi 23.aprillil 2018. Intervjuu kestis 30 minutit. Intervjuu käigus kasutati poolstruktureeritud küsimustikku, mis koosnes kaheksast avatud küsimusest (Lisa 2). Küsimuste järjekord oli paigas, kuid intervjuu käigus esitati ka mitmeid täpsustavaid küsimusi. Samuti andis intervjueeritav palju lisainformatsiooni intervjuu käigus. Esimene, kolmas ja viies intervjuuküsimus olid sõnastatud nii, et intervjueeritav annaks jah/ei vastust. Saadud vastusest sõltusid vastavalt teise, neljanda ja kuuenda küsimuse sõnastus.

Intervjuu käigus saime teada, et intervjueeritav arvab, et seadusega tegelevatel inimestel on uurimistöödest igapäevatöös kasu. Ta tõi kohe näitena xLaw tarkvara [36], mis on spetsialiseerunud õigustehnoloogia valdkonnale ning mida me mainisime punktis 1.1. Intervjueeritava jaoks xLaw aitab tööprotsesse lihtsustada ning muuta üleüldist tööd efektiivsemaks. Intervjueeritava arvamus temaatiliste mustrite kaevandamisest seadusetekstidest oli selline, et see muudaks inimeste elu kiiremaks ning säästaks palju aega. Kuid samas mainis juuratudeng, et see ei pakuks juristile eriti lisaväärtust, vaid pigem oleks see suunatud tudengitele ning tavainimestele, kes seadustega igapäevaselt kokku ei puutu. Näitena tõi ta seda, et 10-15-aastase kogemusega prokuröril on nii karistusseadustik kui ka teised õigusaktid tõenäoliselt selged.

Intervjueeritav andis veel lisainformatsiooni. Nimelt arvas ta, et tavainimestele on seadused ja sellega seonduv väga keeruline, isegi välistades kindlaid õigusakte. Ta tõi näitena, et kui on mingi praktiline olukord, mis ei ole lepinguga reguleeritud ning oleks vaja teada saada, millised on tagajärjed, siis võiks programm 1000 paragrahvi seast leida potentsiaalselt õiged, sest vastasel juhul ei ole tavainimestel midagi teha soorituskondiktsiooni või regressikondiktsiooni juhtumitega. See vastus näitab seda, et kui antud lõputööst teha edasiarendus, siis oleks võimalik teha rakendus, mis oleks

tavainimestele väga kasulik seaduste vahel informatsiooni leidmisel või nende klasterdamisel.

Mudelite hindamisel pani intervjueeritav tähele, et avastatud teemade kümne põhisõna seas oli ka neid, mis väga sinna ei sobi. Näiteks LDA seitsme teema avastamise puhul pani ta tähele, et sõna “vangistus” ei sobi põhisõna hulka, kuna see oli pigem äriseadustikule viitav teema ning see käib eraõiguse kohta. Samuti arvas, et ei sobi sõna “kohtuistung”, kuid hetk hiljem arvas, et see on oluline lüli, kuna see võib abistada inimest, kes otsiks kuidas apellatsiooni esitada.

HDP puhul arvas intervjueeritav, et 20 õigusakti kohta on 20 avastatud teemat liiga palju ning pigem iga teema on seotud kindlalt õigusaktiga. Selle tõttu välistasime ka HDP analüüsimist edaspidi.

Me eksperimenteerisime mitme erineva teemade arvu avastamise vahel LDA ja LSI puhul ning vaatasime tulemusi koos intervjueeritavaga. Ta oli üht meelt, et LDA tulemused olid tõlgendatavamad ning sobivad edaspidise analüüsi jaoks paremini kui LSI. LSI puhul tõi ta mitmeid kordi rohkem märkusi, kus üks või teine sõna ei sobi antud teemasse.

Kuna intervjueeritav pidas LDA mudeli tulemusi kõige paremaks ning koherentsuse tulemuste poolest oli LDA kõige parem, siis tegime edaspidise analüüsi LDA mudeli kohta. Teemade arvuks meie dokumendikorpuse jaoks valisime 8. Kuigi koherentsuse järgi oli teemade arvu 9 puhul parem tulemus, siis intervjueeritava arust oli 8 teema puhul saadud tulemus parem. Seetõttu otsustasime usaldada intervjueeritava intuitsiooni. Tulemused on nähtavad allpool olevas tabelis. Lõputöös kasutatava LDA mudeli salvestasime eraldi failina.

Tabel 12 LDA teemade võtmesõnad ja tõlgendused

Teema	**Võtmesõnad**	**Teema tõlgendus**
1	0.022*"vangistus" + 0.020*"osaühing" + 0.014*"kohtuistung" + 0.013*"ringkonnakohus" + 0.011*"kostja" + 0.010*"apellatsioonkaebus" + 0.010*"menetlusabi" + 0.009*"hageja" + 0.009*"sissemakse" + 0.009*"osakapital	Kohtuistung ja vangistus
2	0.090*"fondivalitseja" + 0.038*"pensionifond" + 0.035*"osak" + 0.026*"finantsjärelevalve" + 0.023*"valitseja" + 0.020*"depos	Fondid

Teema	Võtmesõnad	Teema tõlgendus
	itoorium" + 0.020*"fond" + 0.019*"aktsiaseltsifond" + 0.018*"eurofond" + 0.017*"kindlustusgrupp	
3	0.020*"haagis" + 0.019*"juhiluba" + 0.013*"sõidutee" + 0.012*"mootorsõidukijuht" + 0.012*"maanteeamet" + 0.011*"tehnonõue" + 0.011*"liiklus" + 0.010*"robotliikur" + 0.010*"jalakäija" + 0.010*"sõitma	Sõidukid
4	0.045*"elamisluba" + 0.027*"makseteenus" + 0.016*"veksel" + 0.015*"kindlustusvõtja" + 0.014*"vedaja" + 0.014*"viisa" + 0.012*"veos" + 0.010*"tarbijakrediidileping" + 0.009*"üürnik" + 0.009*"tšekk	Elamisluba
5	0.067*"hankija" + 0.043*"hankeleping" + 0.038*"riigihange" + 0.038*"inspektsioon" + 0.035*"pakkumus" + 0.025*"emitent" + 0.017*"hankemenetlus" + 0.013*"alusdokument" + 0.012*"prospekt" + 0.011*"vaidlustuskomisjon	Riigihanked
6	0.021*"heitkogus" + 0.020*"õhukvaliteet" + 0.019*"sideettevõtja" + 0.018*"saasteaine" + 0.014*"kasvuhoonegaas" + 0.014*"sideteenus" + 0.010*"ühik" + 0.010*"välisõhk" + 0.010*"heiteallikas" + 0.010*"käitaja	Atmosfäär
7	0.022*"süüdistatav" + 0.016*"tegevväelane" + 0.016*"kaitsja" + 0.015*"kohtuistung" + 0.014*"kahtlustatav" + 0.012*"ringkonnakohus" + 0.011*"järelevalvenõukogu" + 0.011*"audiitorkogu" + 0.009*"ülem" + 0.009*"siseaudiitor	Süüdistus ja kohus
8	0.045*"kriisilahendusmenetlus" + 0.031*"kriisilahendusasutus" + 0.024*"teisendamine" + 0.020*"kriisilahendusmeede" + 0.018*"finantstoetus" + 0.015*"erihaldur" + 0.014*"ühisotsus" + 0.014*"finantsjärelevalve" + 0.014*"kriisilahenduskava" + 0.014*"pangandusjärelevalve	Kriisiolukord

4.4 Õigusaktide ja teemade vaheline sarnasus

Me arvutasime välja iga dokumendi ja teema vahelise sarnasuse kogu dokumendikorpuse peale Gensimi *inference()* meetodi abil. Allpool järjestasime antud tulemused nende teemade järjekorra järgi. Kõige suurema kaaluga tulemused tegime tumedaks.

Õigusaktide ja teemade vahelise sarnasuse tulemused viisil *Teema – Sarnasuse väärtus* (lühendite selgitused on saadaval punktis 2.2):

1. IFS: (1 - 0.022, **2 - 11799.999**, 3 - 0.007, 4 - 0.013, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)
2. VÕS: (1 - 0.022, 2 - 0.017, 3 - 0.006, **4 - 15690.987**, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)
3. TsMS: (**1 - 7798.0093**, 2 - 0.017, 3 - 0.006, 4 - 0.013, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)
4. KrMS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.013, 6 - 0.018, **7 - 10795.005**, 8 - 0.005)
5. KindlTS: (1 - 0.022, **2 - 5250.996**, 3 - 0.007, 4 - 0.013, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)
6. ÄS: (**1 - 4428.0049**, 2 - 0.0169, 3 - 0.0065, 4 - 0.0127, 5 - 0.0133, 6 - 0.0182, 7 - 0.0215, 8 - 0.0049)
7. LS: (1 - 0.0217, 2 - 0.0169, **3 - 10986.984**, 4 - 0.0127, 5 - 0.0133, 6 - 0.0182, 7 - 0.0215, 8 - 0.0049)
8. RHS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, **5 - 7378.992**, 6 - 0.018, 7 - 0.022, 8 - 0.005)
9. HKMS: (**1 - 1680.859**, 2 - 0.017, 3 - 0.006, 4 - 0.013, **5 - 47.835**, 6 - 0.018, **7 - 1211.362**, 8 - 0.005)
10. KarS: (**1 - 5528.016**, 2 - 0.017, 3 - 0.006, 4 - 0.013, 5 - 0.013, 6 - 0.018, 7 - 0.0215, 8 - 0.005)
11. RLS: (**1 - 4210.138**, 2 - 0.017, **3 - 48.893**, 4 - 0.013, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)
12. KVTS: (1 - 0.022, 2 - 0.017, 3 - 0.0065, 4 - 0.013, 5 - 0.013, 6 - 0.018, **7 - 5188.005**, 8 - 0.005)
13. VPTS: (1 - 0.022, **2 - 83.161**, 3 - 0.0076, 4 - 0.013, **5 - 6494.869**, 6 - 0.018, 7 - 0.022, 8 - 0.005)

14. AudS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.013, 6 - 0.018, **7 - 2894.013**, 8 - 0.005)

15. VMS: (1 - 0.022, 2 - 0.017, 3 - 0.007, **4 - 4162.002**, 5 - 0.013, 6 - 0.018, 7 - 0.022, 8 - 0.005)

16. VTMS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.013, **6 - 726.278**, **7 - 1501.759**, 8 - 0.005)

17. AÕKS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.013, **6 - 6058.995**, 7 - 0.022, 8 - 0.005)

18. FELS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.0136 - 0.018, 7 - 0.022, **8 - 2817.997**)

19. ESS: (1 - 0.022, 2 - 0.017, 3 - 0.007, 4 - 0.013, 5 - 0.013, **6 - 3570.003**, 7 - 0.022, 8 - 0.005)

20. KAS: (1 - 0.022, **2 - 759.215**, 3 - 0.007, 4 - 0.013, 5 - 0.013, **6 - 2734.804**, 7 - 0.022, 8 - 0.005)

Üleval pool olevad tulemused on esitatud kaalu väärtuse järgi. Protsentuaalsete väärtuste jaoks on vaja iga dokumendi kõik arvud kokku liita ning siis kindla teema väärtuse jagada selle väärtusega. Näitena 9.dokumendi puhul saame protsentuaalseteks väärtusteks alustades esimesest: 1 - 57,2%, 5 – 1,6%, 7 - 41,2%. Siit saab järeldada, et dokument number 9 on kõige rohkem seotud teemaga number 1 (ehk 57,2% sõnadest olid määratud 1 teema alla) ja samuti on suur seos teemaga number 7. Dokumentide ja teemade seoseid on võimalik leida sama loogika järgi ülejäänute dokumentide puhul.

Teises tabelis näitame, kuidas õigusakte liigitati Riigi Teataja kodulehel [2] nende süstemaatilise liigituse järgi:

Tabel 13 Riigi Teataja süstemaatiline liigitus

Õigusakt	**Haldusõigus**	**Eraõigus**	**Kohtu-menetlus-õigus**	**Karistus-õigus**	**Keskkonna-õigus**
IFS	x				

Õigusakt	**Haldusõigus**	**Eraõigus**	**Kohtu-menetlus-õigus**	**Karistus-õigus**	**Keskkonna-õigus**
VÕS		x			
TsMS			x		
KrMS			x		
KindlTS	x			x	
ÄS	x	x			
LS	x			x	
RHS	x				
HKMS			x		
KarS				x	
RLS	x				
KVTS	x				
VPTS	x			x	
AudS	x				
VMS	x				
VTMS			x		
AÕKS				x	x
FELS	x			x	
ESS	x			x	
KAS	x			x	

Kui vaatleme samaaegselt dokumentide ja teemade sarnasuse tulemusi ning Riigi Teataja süstemaatilist liigitust õigusaktide poolt, siis märkame, et on olemas mitmeid sarnasusi kui ka erinevusi. IFS ja KindlTS on Riigi Teataja kodulehel seotud süstemaatilises liigituses Haldusõiguse järgi ning ka teemade osas on nad seotud teemaga number 2. ESS ja KAS on süstemaatilises liigituses seotud nii Haldusõiguse kui ka Karistusõiguse järgi ning ka teemades on nad seotud omavahel teemaga number 6. Kohtumenetlusõiguse koha pealt on ka teemade puhul sarnasust VTMS, KrMS ning HKMS, kus teema numbriks oli 7. Neid sarnasusi võib leida mitmeid veel.

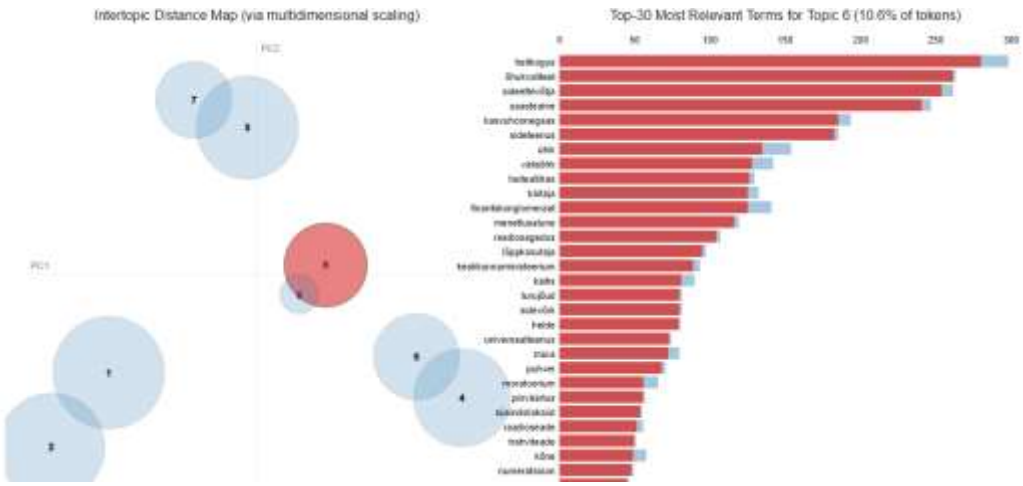
KindlTS, LS ja KarS on kõik Riigi Teataja süstemaatilise liigituse järgi Karistusõiguse all, kuid kõikide LDA poolt avastatud teemad erinevad üksteisest. See räägib sellest, et LDA mudel leidis nende õigusaktide erinevustega liigituse võrreldes Riigi Teataja omadega. See ei ole sugugi halb, sest kindlasti saab tekste tegelikult erinevat moodi liigitada, kuna tegelikult avastas LDA antud dokumendikorpuses varjatud mustrid, mis on ka meie jaoks kergesti tõlgendatavad, kes seadusega igapäevaselt ei tegele.

Kõige suuremad erinevused õigusaktide ja teemade vahel tulevad siis, kui vaadata süstemaatilise liigituse osas Haldusõigust. See näitab seda, et see liigitus on pealiskaudne ning tegelikult need seadused, mis kuuluvad selle liigituse alla, erinevad üksteisest päris palju. Karistusõiguse puhul on sarnasused ja erinevused pooleks. Kuid näiteks veel kui vaadata Kohtumenetlusõiguse süstemaatilist liigitust, kus on TsMS, KrMS, HKMS ja VTMS, siis vastavalt nende põhi avastatud teemad on TsMS puhul 1, KrMS puhul 7, HKMS puhul 1 ja 7 ning VTMS puhul 6 ja 7. HKMS avastatud teemad järgi saab teha järeldust, et tal on seos TsMS-ga teema number 1 järgi ja KrMS-ga teema number 7 järgi.

4.5 LDA tulemuste visualiseerimine ja kontrollimine

LDA teemade visualiseerimiseks kasutasime Pythoni teeki PyLDAvis [42]. PyLDAvis on teemade modelleerimisest saadud teemade visualiseerimise tööriist. See pakett eraldab infot LDA mudelist, et antud tulemusi veebis interaktiivselt visualiseerida.

LDA tulemused Pythoni teeki PyLDAvis näitel:

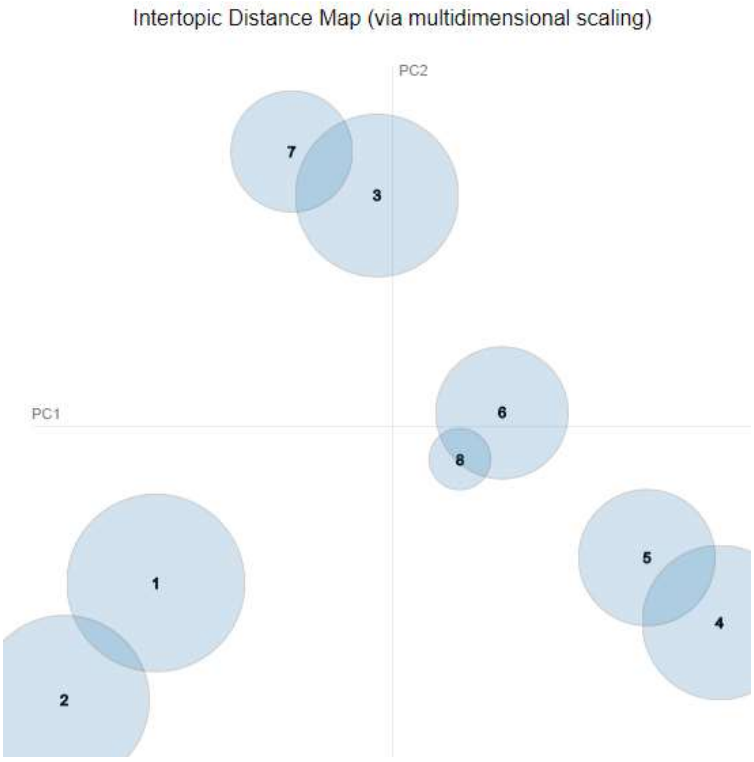


Joonis 7 LDA tulemused PyLDAvise abil

Vaatame jooniselt 7 LDA mudeli teemade visualiseerimist. Teemad on näidatud vasakul pool ning sõnad paremal. Joonise lugemise puhul tuleb arvestada antud asju:

1. Suuremalt kujutatud teemad on dokumendikorpuses sagedasemad ehk meie näite puhul on nendeks teemadeks 1 ja 2
2. Läheduses olevad teemad on rohkem sarnased, lahutuses olevad aga vähem (näiteks teemad 6 ja 8 on omavahel rohkem sarnased)
3. Teema valimise puhul näeme paremal pool kohe selle teemaga seotud põhi 30 võtmesõna.
4. Kui paremalt pool hiirega sõna peale minna, siis muutuvad teemade ringide suurused selle järgi, kui palju seda sõna esineb ühes või teises teemas

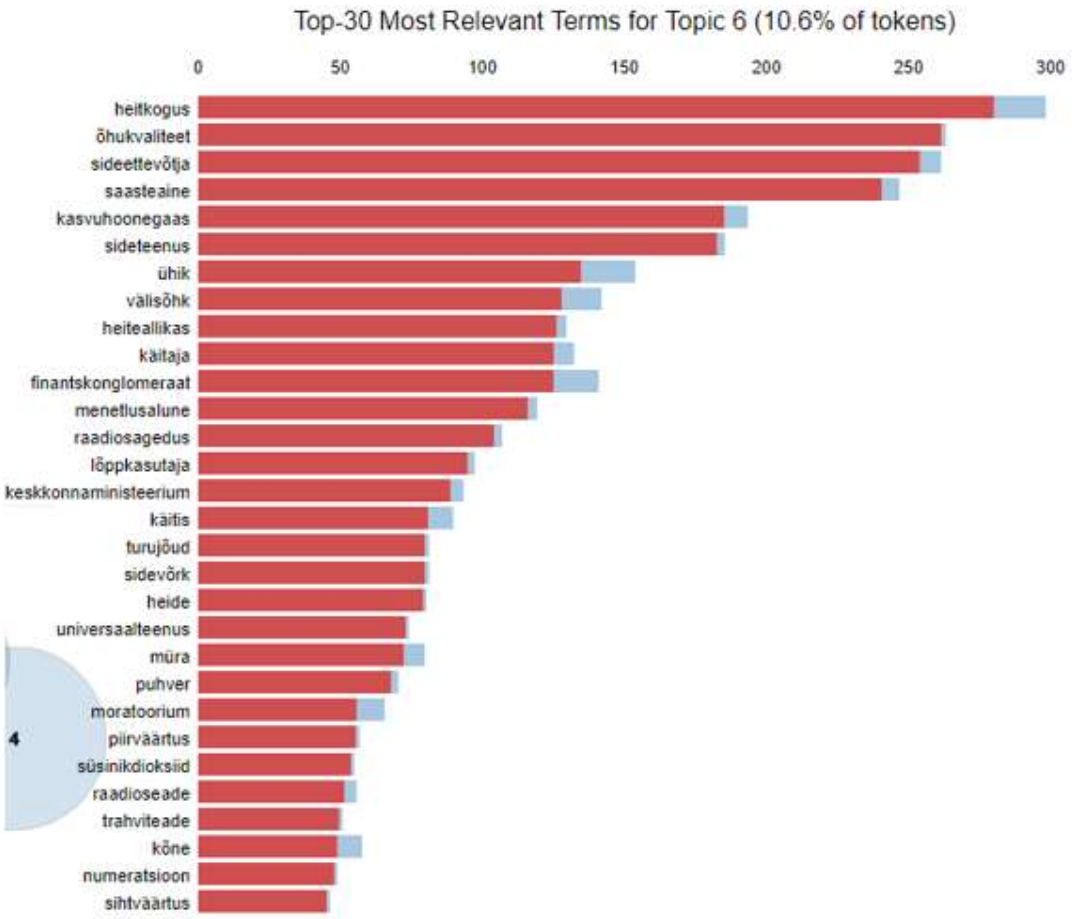
Selleks, et saada parem ülevaade, vaatame PyLDAvise vasakut poolt eraldi antud jooniselt:



Joonis 8 LDA teemad PyLDAvise abil

Joonisel on näha, kui seotud on teemad 6 ja 8 omavahel. Samuti on 5 ja 4 ning 7 ja 3 omavahel seotud. Teemad 1 ja 2 on seotud, kuid mitte nii palju nagu eelnevad teemad omavahel. Kuuenda ja kaheksanda teemade puhul saab teha järelduse, et põhjus, miks nad omavahel seoses on see, et kriisilaheduse teema on osake sellest, mis on seotud õhukvaliteedi ja saasteainetega, et kindlas atmosfääriga seotud kriisiolukorras on seadused sellised. Ka teisi seoseid on võimalik luua selle loogika järgi.

Veelgi parema arusaama jaoks vaatame PyLDAvise paremat poolt eraldi antud jooniselt:



Joonis 9 LDA teema sõnad PyLDAvise abil

LDA-d saab kasutada automaatse sildistamise jaoks. Me võime läbi käia kõik teemad ning panna neile silt külge selle järgi, millest antud teemas juttu on. Joonisel 9 on näha, et teema number 6 on suuremas osas seotud õhukvaliteediga, kuid on ka mitu seost sidega. Sellest järeldub see, et LDA pani need mõlemad valdkonnad ühe teema alla arvatavasti sellepärast, kuna me avastasime 8 teemat. Samas kui me oleks valinud teema number 5,

siis käiks see riigihangete kohta. Mida paremini algandmeid puhastaks, seda paremad oleks ka avastatud teemad. Tulemusi saab mõjutada ka sellega, et mitu teemat me soovime avastada. Veelgi kui kasutada rohkem dokumente ehk näiteks kõiki õigusakte, siis oleks ka teemad arvatavasti paremini tõlgendatavad, kuid see jääb tuleviku töö jaoks [43].

Selleks, et näha, et leitud teemad on tõesti kasulikud, siis proovime programmist läbi lasta uued dokumendid. Näitena kasutame rakendusaktide tekste, mida saame Riigi Teataja kodulehelt [2]. Seal on võimalik eraldi õigusakti lehelt leida ka seotud rakendusaktid.

Näitena toome välja atmosfääri kaitse seaduse all olevat rakendusakti "Lasteaiahoonetes energiatõhususe ja taastuvenergia ja kasutuse edendamise toetuse kasutamise tingimused ja kord". Kui esmapilgul dokumenti läbi vaadata, siis on võimalik jõuda järeldusele, et see on suuremas osas seotud atmosfääri kaitse seadusega, kuid samas on olemas osa, mis on seotud riigihangete seadusega. Laseme antud dokumendi läbi meie LDA mudeli ning vaatame, millised on saadud tulemused.

Tabel 14 Uue dokumendi tulemused

LDA teema	**Tõenäosus**
1	0.160
2	0.022
3	0.045
4	0.071
5	0.196
6	0.464
7	0.001
8	0.041

Tabelis number 14 näeme uue dokumendi saadud tulemusi. Nagu kohe silma jääb, siis kõige suurem seos on teemaga number 6. Kui vaadata lõputöö punkti 4.3, siis näeme, et see teema on seotud enamasti atmosfääriga. Samuti on suur osakaal teemal 1 ning teemal

5. Pärast dokumendi analüüsimist jõudsime järeldusele, et teema 1 on seotud dokumendiga selle tõttu, et seal on palju kirjutatud menetlusest. Teema 5 puhul on tugev seos riigihangetega, kuna antud dokumendis on olemas osa, kus räägitakse projekti rahastamisest ning hangete läbiviimisest.

Teise näitena toome välja audiitortegevuse seaduse all olevat rakendusakti "Riigi äriühingu ja sellise äriühingu, kus riigil on vähemat otsustusõigus, ning riigi asutatud sihtasutuse auditikomitee moodustamise ja tasustamise ning töökorra põhimõtted".

Tabel 15 Teise uue dokumendi tulemused

LDA teema	**Tõenäosus**
1	0.040
2	0.001
3	0.001
4	0.001
5	0.001
6	0.453
7	0.497
8	0.001

Tabelis number 15 näeme teise uue dokumendi saadud tulemusi. Kohe jääb silma, et dokument on tugevalt seotud kahe teemaga. Arvestades seda, et AudS on tugevalt seotud teemaga number 7, siis on ka see tulemus kõige suurem, kuid see on peaaegu võrdväärne teemaga number 6 saadud tulemusega. Analüüsides dokumenti jõuame järeldusele, et mudeli puhul on võimalik teha veel mitmeid parenduskohti, kuna antud dokument on tegelikult suhteliselt vähe seotud teemaga number 6.

Üldiselt tahtsime kasutada iga õigusakti rakendusakti, kuid Riigi Teataja kodulehel polnud finantskriisi ennetamise ja lahendamise seaduse, karistusseadustiku ning riigilõivuseaduse kohta rakendusakte. Kui arvesse võtta kõiki tulemusi, siis mudel ei toiminud hästi krediidiasutuste seaduse, tsiviilkohtumenetluse seadustiku ning võlaõigus

seaduse rakendusaktide puhul. Seal ei olnud avastatud teemad kooskõlas õigusaktidega. Kuid kui näiteks vaadata tsiviilkohtumenetluse seadustiku rakendusakti, siis oli see tegelikult seotud suures osas kriminaalmenetluse seadustikuga, ning ka mudel avastas teema, mis on tugevalt seotud just kriminaalmenetluse seadustikuga. Kõikide teiste õigusaktide rakendusaktid (14 rakendusakti 17-st) puhul olid tulemused head ning mudel suutis avastada põhiteemad, mis on samad ka õigusaktide puhul.

4.6 Järeldused ja ettepanekud

Antud lõputöö uuringus saavutasime me parimad tulemused nii koherentsuse kui ka kvalitatiivse hinnangu abil just LDA mudeli puhul. Kuigi praeguste tulemuste abil oli võimalik saadud teemasid tõlgendada ning õigusaktidega seotud rakendusakte jagada nende samade teemade vahel ära, siis kindlasti saab seda mudelit paremaks muuta.

Parema tulemuse saamiseks peaks käsitlema kõiki õigusakte, mis on saadaval Riigi Teataja kodulehel. Selle puhul tuleks kindlasti tõlgendatavaid teemasid ning uusi seoseid seaduste vahel juurde. Samuti andmete puhastamise etapp mõjutab oluliselt analüüsitulemusi. Mida paremad andmed olid kasutatud analüüsi jaoks, seda paremad on ka saadud tulemused. Selles etapis saab näiteks juurde lisada mitmeid stoppsõnu, mis ei anna sisule mingit lisainformatsiooni juurde.

Praegu keskendub andmete puhastamise faas ainult eesti keelele, kuigi Riigi Teataja kodulehel on õigusaktid saadaval ka inglise keeles. Tulevikutöö jaoks oleks võimalik eesti õigusaktide teemade modelleerimist rakendada inglise keelsetele tekstidele ning selle tegemisel oleks kindlasti rohkem abi lõputöös mainitud varasematest rakendustest.

Failide teksti sisu integreerimist saab veelgi parandada lisades tuge teiste failitüüpide jaoks. Riigi Teataja kodulehel on õigusaktid saadaval ka XML ja PDF formaadis.

Kokkuvõte

Käesoleva lõputöö eesmärk oli rakendada teemade modelleerimist eesti seadusetekstidele, mida ei oldud siiani veel rakendatud. Ülesanne, mida lahendati oli järgmine: saada teada, milliste "abstraktsete" ehk peidetud teemade kaudu on eesti suurimad seadused omavahel seotud. Teemade modelleerimise algoritmide abil on võimalik leida suurest andmehulgast seoseid ja erinevusi õigusaktide vahel, mille inim jõul leidmine oleks märksa keerulisem ja aeganõudvam. Uuringu läbiviimiseks koguti algandmeid (õigusaktide teksti failid Riigi Teataja kodulehelt), installeeriti vajalikud paketid nagu EstNLTK ja PyLDAvis. Tulemused kuvati nii tabelite kujul kui ka graafiliselt.

Töö oluliseks tulemuseks oli eesti seaduste peidetud teemade leidmine kolme algoritmi abil ning saadud mudelite hindamine. Mudelite hindamiseks kasutati teemade modelleerimise tööriistakomplekti nimega Gensim. Mudeleid hinnati koherentsuse ning kvalitatiivse hinnangu abil. Tänu sellele saadi aru, milline mudel toimis kõige paremini antud dokumendikorpuse peal.

Autor ootas enne uuringu läbiviimist, et leiab milline mudel on kõige sobilikum eesti seadusetekstide jaoks. LDA mudel sai parima tulemuse nii koherentsuse järgi kui ka kvalitatiivse hinnangu abil, kus intervjueeritav pidas selle mudeli tulemusi kõige paremini tõlgendatavaks. Samuti hindasime LDA saadud tulemusi võrreldes Riigi Teataja süstemaatilise liigitusega, kus oli näha mitmeid sarnasusi kui ka erinevusi. Näiteks Kohtumenetlusõiguse süstemaatilise liigituse puhul olid tulemused pigem sarnased, kuid Haldusõiguse puhul eristusid väga palju. See, et tulemused erinesid süstemaatilisest liigitusest, ei ole sugugi halb, vaid pigem näitab see seda, et seadusi on võimalik erinevat moodi liigitada. Samuti uute dokumentide ehk õigusaktidega seotud rakendusaktide testimisel olid tulemused paljulubavad. 14 rakendusakti puhul avastas mudel õige põhiteema, mis oli võrdväärne sellega seotud õigusaktiga. Kuid kolmel juhul ei suutnud mudel avastada õiget põhiteemat ning see näitab seda, et arenguruumi veel on.

Praegused tulemused on saadud kahekümne suurima õigusakti näitel ning saadud tulemused ei ole sugugi halvad, leides milliste peidetud teemade kaudu on eesti suurimad

seadused omavahel seotud. Kuid üks lähenemise puudustest on kindlasti see, et antud lõputöö raames ei käsitletud kõiki seadusi, kuna taheti jääda töö piiridesse, eelkõige sellepärast, et seadused on väga mahukad. Kvalitatiivse hinnangu käigus arvas intervjueeritav, et saadud tulemused ei pakuks juristile eriti lisaväärtust, vaid pigem oleks see suunatud tudengitele ning tavainimestele, kes seadustega igapäevaselt kokku ei puutu. Siit saab järeldada seda, et kui antud lõputööst teha edasiarendus, siis oleks võimalik teha rakendus, mis oleks tavainimestele väga kasulik seaduste vahel informatsiooni leidmisel või seadustega seotud tekstide liigitamisel, ning selline edasiarendus võiks jääda tuleviku uurimisteemadeks.

Summary

The purpose of this thesis was to apply topic modeling to Estonian law texts, which has not been implemented until now. The task that was solved was to find out which of "abstract" or hidden topics in Estonian law texts are interconnected. Using topic modeling algorithms, it is possible to find the largest amount of data in the relationships and differences in legislation, which would be much more complicated and time-consuming to find for a person. Initial data was collected for the thesis from the Riigi Teataja website. Also, all necessary packages were installed such as EstNLTK and PyLDAvis. Results were displayed both in tabular and in graphic form.

The main outcome of this thesis was the discovery of hidden topics in Estonian law texts using three algorithms and evaluation of the resulting models. The topic modeling toolkit called Gensim was used to evaluate the models. Models were evaluated using coherence and qualitative assessment. This made it possible to understand which model worked best on the initial data.

Before the study, we wanted to find out which model is most suitable for Estonian legal texts. The LDA model received best results both in terms of coherence and in terms of qualitative assessment, which results of this model had the strongest interpretation. We also evaluated the results of the LDA compared to the systematic classification of Riigi Teataja, which saw a number of similarities and differences. For example, in the case of systematic classification of court proceedings, the results were pretty similar, but in the case of Administrative Law, the results were different. The fact that the results are differentiated is not bad at all but rather indicates that laws can be classified differently. The results of the testing of implementing new documents were promising. In the case of 14 implementing acts, the model identified the correct topic, which was equivalent to the related legal act. However, in three cases, the model was unable to detect the correct topic, and this shows that there is still room for improvement.

The current results are of the twenty largest acts and the results were not bad at all, considering it was possible to find out by which hidden themes were the Estonian laws interconnected. However, one of the drawbacks of the approach is certainly that all the laws were not covered in the framework of this thesis, because it was important to stay within the bounds of work, especially because the laws are very large and complex. In

the qualitative assessment, the interviewee said that the results obtained would not add much value to the lawyer, but would rather be aimed for students and ordinary people who are not in contact with the law on a daily basis. From this, it can be concluded that if this graduation work progressed, then it would be possible to make an application that would be very useful for ordinary people when seeking information or classifying texts related to the law, and could be the topic for future studies.

Kasutatud kirjandus

1. Õigusaktid [WWW]
https://www.eesti.ee/et/oigusabi/oigussuesteem/oigusaktid/

2. Riigi Teataja [WWW]
https://et.wikipedia.org/wiki/Riigi_Teataja

3. Riigi Teataja küsitlus (2017) [WWW]
https://www.riigiteataja.ee/failid/eRT_kysitlus_2017.pdf

4. Postimees (2017). Vaher: Eesti õigusaktide keel muutub kahjuks aina keerulisemaks (2017) [WWW]
https://www.postimees.ee/4000375/vaher-eesti-oigusaktide-keel-muutub-kahjuks-aina-keerulisemaks

5. Hea Õigusloome Tava [WWW]
http://teenusmajandus.ee/kojast/hea-oigusloome-tava/

6. Maret Maripuu (2016). Seaduse arusaadavuse arusaadavus [WWW]
https://rito.riigikogu.ee/wordpress/wp-content/uploads/2016/04/Seaduse-arusaadavuse-arusaadavus.pdf

7. Wikipedia (2017). Andmekaeve [WWW]
https://et.wikipedia.org/wiki/Andmekaeve

8. Jilles Vreeken (2009). Making Pattern Mining Useful [WWW]
http://www.cs.uu.nl/groups/ADA/pubs/2009/making_pattern_mining_useful-vreeken.pdf

9. Jure Leskovic, Anand Rajaraman, Jeffrey D. Ullman (2014). Mining of Massive Datasets [WWW]
http://infolab.stanford.edu/~ullman/mmds/book.pdf

10. Scott Sims (2015). Everything You Need To Know About Natural Language Processing [WWW]
https://www.kdnuggets.com/2015/12/natural-language-processing-101.html

11. Terena Ball (2018). What is Natural Language Processing? The Business Benefits of NLP Explained [WWW]
https://www.cio.com/article/3258837/artificial-intelligence/what-is-natural-language-processing-the-business-benefits-of-nlp-explained.html

12. Shivam Bansal (2016). Beginners Guide to Topic Modelling in Python [WWW]
https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

13. David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation [WWW]
http://jmlr.csail.mit.edu/papers/v3/blei03a.html

14. David M. Blei (2012). Probabilistic topic models [WWW]
https://dl.acm.org/citation.cfm?doid=2133806.2133826

15. Jordan Barber. Latent Dirichlet Allocation (LDA) with Python [WWW]
https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html

16. Latent Semantic Indexing [WWW]
http://www.seobook.com/lsi/lsa_definition.htm

17. Yee W. Teh, Michael I. Jordan, Matthew J. Beal, David M. Blei (2005). Hierarchical Dirichlet Processes [WWW]
https://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf

18. A. Ng, K. Soo (2015). Topic Modelling with LDA Introduction [WWW]
https://algobeans.com/2015/06/21/laymans-explanation-of-topic-modeling-with-lda-2/

19. Latent Semantic Analysis (2018) [WWW]
https://en.wikipedia.org/wiki/Latent_semantic_analysis

20. Barbara Rosario (2000). Latent Semantic Indexing: An Overview [WWW]
https://www.cse.msu.edu/~cse960/Papers/LSI/LSI.pdf

21. Wikipedia (2018). Dirichlet Process [WWW]
https://en.wikipedia.org/wiki/Dirichlet_process

22. Ana Paula de Oliveira Sales (2011). Clustering Multiple Related Datasets with a Hierarchical Dirichlet Process [WWW]
https://en.wikipedia.org/wiki/Dirichlet_process

23. What is Python? Executive Summary [WWW]
https://www.python.org/doc/essays/blurb/

24. Müller, C., Guido, S. (2016). Introduction to Machine Learning with Python [WWW]
http://proquestcombo.safaribooksonline.com/book/programming/machine-learning/9781449369880/1dot-introduction/scikit_learn_html?query=((scikit+learn))#X2ludGVybmFsX0h0bWxWaWV3P3htbGlkPTk3ODE0NDkzNjk4ODAlMkZtYXRwbG90bGliX2h0bWwmcXVlcnk9KChzY2lraXQlMjBsZWFybikp

25. Radim Rehurek (2018). Introduction [WWW]
https://radimrehurek.com/gensim/intro.html

26. EstNLTK [WWW]
https://estnltk.github.io/

27. Valter Kiisk (2018). Python ja Jupyter Notebook [WWW]
http://kodu.ut.ee/~kiisk/python.html

28. James O'Neill, Leona O'Brien, Cecile Robin, Paul Buitelaar (2017). An analysis of topic modelling for legislative texts [WWW]
http://www.grctc.com/wp-content/uploads/2017/06/ASAIL_paper.pdf

29. Saffron (2017). Knowledge extraction framework [WWW]
http://saffron.insight-centre.org/

30. Oguejiofor Chibueze (2018). NLP for topic modelling and summarization of legal documents [WWW]
https://towardsdatascience.com/nlp-for-topic-modeling-summarization-of-legal-documents-8c89393b1534

31. Ravi kumar V, K. Raghuveer (2012). Legal documents clustering using latent Dirichlet allocation [WWW]
http://research.ijais.org/volume2/number6/ijais12-450384.pdf

32. Yue Wang, Jidong Ge, Yemao Zhou, Yi Feng, Chuanyi Li, Zhongijn Li, Xiaoyu Zhou, Bin Luo (2017). Topic Model Based Text Similarity Measure for Chinese Judgement Document [WWW]
https://link.springer.com/chapter/10.1007/978-981-10-6388-6_4

33. EstNLTK(2014). Morphological Analysis [WWW]
https://estnltk.github.io/estnltk/1.1/tutorials/morf_analysis.html

34. Wikipedia (2018). Stop words [WWW]
https://en.wikipedia.org/wiki/Stop_words

35. Kristel Uiboaeg (2018). Eestikeelsete stoppsõnade loeng [WWW]
http://www.tekstikaeve.ee/blog/2018-04-18-eestikeelsete-stoppsonade-loend/

36. ExtendLaw [WWW]
http://extendlaw.com/et

37. Charu C. Aggarwal (2015). Data Mining: The Textbook [WWW]
http://www.charuaggarwal.net/Data-Mining.htm

38. Devashish Deshpande (2016). What is Topic Coherence? [WWW]
https://rare-technologies.com/what-is-topic-coherence/

39. Michael Röder, Andreas Both, Alexander Hinneburg (2015). Exploring the Space of Topic Coherence Measures [WWW]
http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf

40. Marilyn D. White, Emily Marsh (2006). Content Analysis: A Flexible Methodology [WWW]
https://www.researchgate.net/publication/32957977_Content_Analysis_A_Flexible_Methodology

41. Karin Eilmann (2011). Ämmaemanda õppekava üliõpilaste ja lõpetanute arvamused [WWW]
https://tervis.ut.ee/sites/default/files/ode/karineilmann_2011.pdf

42. Ben Mabey (2015). PyLDAvis documentation [WWW]
https://media.readthedocs.org/pdf/pyldavis/latest/pyldavis.pdf

43. Carson Sievert, Kenneth E. Shirley (2014). LDAvis: A method for visualizing and interpreting topics [WWW]
https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

Lisa 1 – Keskkond

Antud peatükis annab autor kirjelduse vajamineva keskkonna loomise kohta. Algselt installisime arvutisse PyCharm versiooni 2018.1. Seejärel loodi projekt ning projekti sätete all lisati programmeerimiskeeleks Python 3.5.4. Samuti lisati projekti interpretaatori alt mitmeid projekti jaoks vajaminevaid pakette:

- Gensim versioon 3.2.0
- EstNLTK versioon 1.4.1.1
- Numpy 1.14.0
- Jupyter 1.0.0
- Matplotlib 2.1.2
- PyLDAvis 2.1.1

Seejärel kirjutati PyCharmi terminalis käsk: *jupyter notebook*. Siis avanes keskkond veebis, kust valiti *Main.ipynb* ning jooksutati antud faili. Samuti on olemas *Eksperiment.ipynb* fail, mis erineb *Main.ipynb* selle tõttu, et ta oli rohkem eksperimentide tegemise jaoks esialgu ning samuti on seal rohkem koodi ning väljundeid.

Lisa 2 – Intervjuu

Uurimisküsimustest lähtuvad intervjuu küsimused:

1. Kas teie arvates on seadusega tegelevatel inimestel uurimistöödest igapäevatöös kasu?

2. Kui uuritav vastab 1. küsimusele jah, siis: a) Kuidas on teie arvates seadusega tegelevatel inimestel uurimistöödest igapäevatöös kasu?

   Kui uuritav vastab 1. küsimusele ei, siis: b) Miks teie arvates ei ole seadusega tegelevatel inimestel uurimistöödest igapäevatöös kasu?

3. Kas teie arvates on seadusega tegelevatel inimestel sellest kasu, kuid nende tööprotsesse üritatakse optimeerida ning muuta töö efektiivsemaks?

4. Kui uuritav vastab 3. küsimusele jah, siis: a) Kas te olete ise kasutanud mõnda tarkvara, mis on muutnud teie töö efektiivsemaks õiguse vallas?

   Kui uuritav vastab 3. küsimusele ei, siis: b) Miks teie arvates ei ole seadusega tegelevatel inimestel sellest kasu, et nende tööprotsesse üritatakse optimeerida ning töö muuta efektiivsemaks?

5. Kas teie arvates on võimalik antud mudelite tulemuste abil seaduste vahel informatsiooni lihtsamini otsida või grupeerida?

6. Kui uuritav vastab 5. küsimusele jah, siis: a) Miks teie arvates abistaks need tulemused seaduste vahel informatsiooni lihtsamini otsida või grupeerida?

   Kui uuritav vastab 5. küsimusele jah, siis: a) Miks teie arvates ei abistaks need tulemused seaduste vahel informatsiooni lihtsamini otsida või grupeerida?

7. Millise mudeli tulemused tundusid teie jaoks kõige arusaadavamad?

8. Miks just selle mudeli tulemused tundusid teie jaoks kõige arusaadavamad?