TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

**Sirle Orav-Hinno**

# Euroopa Liidu Kohtu otsustest fraasidele sarnaste lõikude otsingu analüüs CountVectorizer ja Word2Vec baasil

**Magistritöö (15 EAP)**

Juhendaja: Dage Särg
MA
Kaasjuhendaja: Risto Hinno
MSc

Tartu 2021

## Euroopa Liidu Kohtu otsustest fraasidele sarnaste lõikude otsingu analüüs CountVectorizer ja Word2Vec baasil

**Lühikokkuvõte:**

Magistritöös analüüsitakse, kas CountVectorizer või Word2Vec abil on võimalik luua nutikam märksõna otsing, mis annaks etteantud fraasile sarnaseid Euroopa Kohtu otsuste lõike. Hetkel olemasolevad InfoCuria ja EUR-Lex otsingusüsteemid ei võimalda Euroopa Liidu Kohtu otsust kuvada selliselt, et selle lugemist saaks alati alustada kohtu analüüsist. Samuti ei kuva need sisult sarnaste sõnadega tulemusi. Eeltoodust tulenevalt on Euroopa Kohtu otsustest fraasidele vastava informatsiooni leidmine aeganõudev. Magistritöö käigus loodi kolm Euroopa Kohtu otsuste andmetabelit (kohtu hinnangu ja resolutsiooni tekstid, ainult kohtu hinnangu tekstid ning ainult resolutsiooni tekstid), kus iga Euroopa Kohtu lahendi kohta on kohtu hinnangu ja resolutsiooni osad lõikudena eraldi ridadel. Pärast seda rakendati nende andmestike peal CountVectorizerit ja Word2Veci, et saada kätte lõikude vektorid, mida testimiseks kasutatava fraasidega võrrelda. Testimiseks kasutati kümmet maksuõiguse fraasi. Töö tulemusena leiti, et CountVectorizer või Word2Vec abil on võimalik luua nutikam märksõna otsing (tulemustena kuvataks kasutajale kohtuotsuse lõigud, mitte tervikteksid), kuid see toimib kohtu hinnangu osast juristide ja kohtunike jaoks kasulike lõikude leidmiseks. Kasulike kohtuotsuste resolutsioonide leidmiseks toimivad jätkuvalt paremini InfoCuria ja EUR-Lex otsingusüsteemid.

**Võtmesõnad:**

Loomuliku keele töötlus, sarnaste tekstide leidmine, Euroopa Kohtu otsused, CountVectorizer, TfidfVectorizer, Word2Vec

**CERCS:** P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

## Analysis of searching for similar phrases in sections of judgements of the Court of Justice of the European Union based on CountVectorizer and Word2Vec

**Abstract:**

The master's thesis analyzes whether CountVectorizer or Word2Vec can be used to create a smarter keyword search that would yield paragraphs of the Court's judgments similar to the given phrase. The current InfoCuria and EUR-Lex search systems do not display the judgment of the Court of Justice of the European Union in such a way that you can start reading from the Court's analysis part, nor do they display results with similar words. So it

is time-consuming to find information corresponding to the phrases in the judgments of the Court. In the master's thesis, the author created three data tables of court decisions (court assessment and resolution texts, court assessment texts only and resolution texts only), where court assessment and resolution sections are divided into separate lines. After that author applied CountVectorizer and Word2Vec on these datasets and received vectors. These vectors were used to compare with tax law phrases for testing. The result was that CountVectorizer or Word2Vec could be used to create a smarter keyword search (the results would show the paragraphs of the judgment, not the full texts), but it works only when you want to find useful sections of the court assessment part. The InfoCuria and EUR-Lex search engines continue to work better when you want to find practical court resolutions.

**Keywords:**
Natural language processing, finding similar texts, judgements of the Court of Justice of the European Union, CountVectorizer, TfidfVectorizer, Word2Vec

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

## Sisukord

# 1. Sissejuhatus

Juristid, advokaadid ja kohtunikud kasutavad õigusküsimuste analüüsis mitmesuguseid õigusallikaid. Muuhulgas on allikaks kohtupraktika, mis on eriti oluline kohtuvaidlustes. Kuigi Eestis ei kehti pretsedendiõigus, viitavad juristid, advokaadid ja vandeadvokaadid (edaspidi ühiselt nimetaud: juristid) kohtuotsustes toodule õigusanalüüsides, kohtule esitatavates dokumentides ning kohtunikud omakorda kohtuotsustes. Sellised viited on vajalikud selleks, et kohtupraktika oleks ühtne ning teisalt toetab varasem praktika juriidilist argumentatsiooni. Samuti on kohtupraktika teadmine vajalik, sest teatud juhtudel on kõrgema astme kohtu otsus alama astme kohtule siduv.

Eesti kuulub 2004. aastast Euroopa Liitu. Eesti seadused peavad olema kooskõlas Euroopa Liidu õigusega ning õiguse tõlgendus peab olema liiduüleselt ühtne. Seetõttu on Eesti juristide ja kohtunike jaoks lisaks Eesti kohtupraktikale üheks oluliseks õigusallikaks Euroopa Liidu Kohtu (edaspidi ka: Euroopa Kohus) otsustes esitatud põhjendused. Euroopa Kohtu otsuseid on võimalik otsida InfoCuria andmebaasist [1] või EUR-Lex [2] portaalist. Euroopa Kohtu otsus koosneb enamasti kohtu koja nimetusest, kuupäevast, kohtuasja numbrist, poolte nimedest, menetlusosaliste kirjeldusest, asjakohaste õigusaktide kirjeldusest, asjaolude kirjeldusest, poolte ja menetlusosaliste seisukohtadest, Euroopa Kohtu analüüsist, kohtukulude jaotusest ning resolutsioonist (vaata näidiskohtuotsust Lisast I). Need kohtuotsuse osad ei pruugi alati olla samas järjestuses nagu nimetatud ning kohtuotsus võib olla täiendavalt liigendatud (lisatud täiendavaid alapealkirju).

Juristid ja kohtunikud alustavad kohtuotsuse lugemist enamasti resolutsioonist ja liiguvad siis kohtu analüüsi juurde ning vajadusel vaatavad pärast nimetatute lugemist või lugemise ajal ülejäänud kohtu otsuse osasid. Enamasti on analüüs ja resolutsioon esitatud ligikaudu poolest kohtuotsusest alates või paikneb lõpus. Igapäevases töös tähendab selline kohtuotsuse struktuur juristi või kohtuniku jaoks[1], et vajaliku põhjenduse leidmiseks tuleb esmalt otsida kohtulahendit InfoCuria andmebaasist või EUR-Lex portaalist. Seejärel tuleb kerida lahend resolutsioonini või kohtu analüüsini (mõnikord eelneb kerimisele ctrl+f klahvikombinatsiooniga märksõna otsing) ning olenevalt kohtuotsusest töötada keskmiselt läbi 10-30 lõiku ning liikuda seejärel järgmise otsingu tulemuse juurde, korrates eelnevat protseduuri.

[1] Magistritöö autor on töötanud pikemat aega juristina, mistõttu siin ja järgnevalt tugineb magistritöö autor oma töökogemusele.

Võimalus on otsingut teha ka Google'i otsingu kaudu, kuid otsingu tulemused viivad siiski eelnimetatud andmebaasi või portaali tulemuste juurde.

Kuigi EUR-Lex võimaldab kuvada otsingu tulemusi selliselt, et näidataks ainult kokkuvõtvat infot ja resolutsiooni, võib oluline informatsioon sisalduda kohtu analüüsis olevates põhjendustes. InfoCuria ja EUR-Lex ei võimalda Euroopa Liidu Kohtu otsust kuvada selliselt, et selle lugemist saaks alati alustada kohtu analüüsist. Samuti ei pruugi tulemust anda Ctrl+f märksõna otsing, kuivõrd sõna võib olla asendatud sisult sarnase sõnaga näiteks sõna „ettevõtlus" asemel võib olla kasutatud sõna „majandustegevus", mis tähendab, et otsingut tuleb teha kummagi märksõna kohta eraldi. Võib olla ka olukordi, kus fraasi (näiteks „käibemaksukohustuslaseks registreerimine") otsingul kuvatakse lahendeid, mis puudutavad sõna „registreerimine" kuigi olulisem võib olla nende kahe sõna kogum. Seega eeltoodust tulenevalt on Euroopa Kohtu otsustest fraasidele vastava informatsiooni leidmine aeganõudev.

Eesti kohtutele esitatavate hagide, kaebuste, vastuste või seisukohtade tähtaeg on piiratud. Näiteks kaebusele vastuse koostamiseks annab kohus enamasti juristidele aega 21-30 kalendripäeva. Samuti on üldine põhimõte, et juristidel tuleb kohtule esitada seisukohad nii varakult kui võimalik, et kohus saaks nendega otsuse tegemisel arvestada. Seega on juristide jaoks oluline, et vajalik informatsioon nende koostamiseks oleks kättesaadav võimalikult lihtsalt ja kiirelt. See tähendab, et info otsinguks ja selle töötlemiseks kuluks võimalikult vähe aega. Samuti on vajaliku info lihtne ja kiire leidmine oluline kohtunike jaoks. Kohtunikel võib olla käes 40-60 kohtuasja. Selleks, et need mõistliku aja jooksul lahendatud saaks ning kohtumenetlus ei veniks, ongi oluline, et info oleks kergelt kättesaadav. Eeltoodust tulenevalt, kui info otsing (sealhulgas asjakohase varasema kohtupraktika leidmine) kulgeb kiiresti ja relevantne informatsioon on esitatud esmajärjekorras, on see ajavõit nii juristi kui kohtuniku jaoks.

Eelnevalt kirjeldatud probleemi lahendus võib olla kuvada otsijale kohtuotsuse tervikosa asemel kohtuotsuse lõike, mis on otsingufraasile sarnased. Selleks, et sarnaseid lõike leida on vajalik keeletöötlus. Keeletöötluseks ja sarnaste tekstide leidmiseks on olemas erinevaid meetodeid. Osad neist loevad tekstis sõnad lihtsalt kokku, teised suudavad ka semantilist analüüsi (täpsemalt on neid käsitletud töö kolmandas peatükis). Samas olenevalt ülesandest võib lihtsam meetod anda piisava täpsusega vajaliku vastuse, mistõttu kõrgemataseme-lisema meetodi kasutamine ei pruugi olla põhjendatud. Selleks, et selgitada välja, millist meetodit võiks rakendada fraasile sarnaste kohtuotsuste lõikude otsingule, on magistritöö jaoks valitud kaks erineva tasemega meetodit – CountVectorizer ja Word2Vec. Kuigi töö autor

katsetas töö kirjutamise käigus ka EstBERTi ja Doc2Veci, siis selgus, et need ei sobi töös toodud eksperimentide tegemiseks, sest vajavad konteksti (ümbritsevaid sõnu) rohkem kui otsinguks kasutatud paarisõnaline fraas võimaldas. Seetõttu piirduti töös, ning ka selleks, et magistritöö maht ei läheks liiga suureks, kahe meetodi kasutamisega. Selgitamaks välja, kas nimetatud meetodite abil otsing toimiks paremini kui senised süsteemid, on võrdlusena kasutatud InfoCuria ja EUR-Lex otsinguid.

Seega on käesoleva töö eesmärk selgitada välja, kas CountVectorizer või Word2Vec abil on võimalik luua nutikam märksõna otsing, mis annaks etteantud fraasile sarnased Euroopa Kohtu otsuste lõigud.

Töö eesmärgi saavutamiseks seadis töö autor järgnevad uurimisülesanded:

1. Anda ülevaade loomuliku keele töötlusest ja masinõppest.
2. Anda ülevaade CountVectorizer ja Word2Vec ning sarnasuse hindamise meetoditest ja nende rakendamise võimalustest.
3. Anda ülevaade Euroopa Kohtu otsustest kui töö aluseks olevatest andmetest ning nende eeltöötlemiseks tehtud toimingutest.
4. Analüüsida CountVectorizer ja Word2Vec rakendamisest saadud tulemusi ning selgitada välja, kas nende abil on võimalik luua nutikam märksõna otsing.

Töö on aktuaalne, sest õigustehnoloogia (*legal tech*) tööriistade kasutusele võtmine tähendab juristide ja advokaatide töös efektiivsuse tõusu [3] ning viimaste aastate jooksul on õigustehnoloogia tööriistade kasutusele võtmine kasvutrendis [4]. Kui varem otsisid juristid vajalikku informatsiooni raamatutest, siis praeguseks tehakse enamik õigusanalüüsist elektrooniliselt või elektroonilistele allikatele tuginedes [4]. Kui nutikam märksõnaotsing töötab, siis selle baasil on võimalik hiljem välja arendada juriste ja kohtunikke abistav otsingurakendus efektiivsemaks Euroopa Kohtu otsuste lõikude kui õigusallikate leidmiseks.

Magistritöö koosneb seitsmest osast (sh sissejuhatus, kokkuvõte ja viidatud kirjandus). Töö teises peatükis on ülevaade loomuliku keele töötlusest ja masinõppest ning varasematest valdkonna uuringutest. Kolmandas peatükis on ülevaade vektorite moodustamise ja sarnasuse hindamise meetoditest (CountVectorizer ja Word2Vec). Neljandas peatükis selgitatakse Euroopa Kohtu otsuste struktuuri ja nende eeltöötlust. Viiendas peatükis analüüsitakse CountVectorizeri ja Word2Veci rakendamisest saadud tulemusi ning selgitakse, milline on neist sarnaste kohtuotsuste lõikude leidmiseks parim. Tööl on üks lisa ning neljanda ja

viienda peatüki aluseks olev kood ning tulemusena saadud andmetabelid on tehtud kättesaadavaks GitHubi[2] ja Google Drive[3] kaudu.

[2] https://github.com/sirleor/Mag.-tooks/tree/master/Mag.too%20puhas%20kood
[3] https://drive.google.com/drive/folders/1aE_WQzBxnzmGJwQeZAnRl6LqJoQEyYcg?usp=sharing

## 2. Loomuliku keele töötlusest ja masinõppest

Magistritöös töödeldakse Euroopa Kohtu otsuseid, mis on loomuliku keele andmed. Loomuliku keele töötlus tähendab lihtsalt öeldes arvuti võimet töödelda loomuliku keele andmeid selliselt, et arvutid saaks efektiivselt suhelda inimestega [5]. Suhtlus tähendab inimese abistamist erinevates ülesannetes. Üheks võimalikuks viisiks suhtluse korraldamisel on masinõpe [6]. Masinõpe tähendab, et programm leiab näidisandmetest mustrid ja teeb ennustusi uute andmete kohta [6]. Näiteks on masinõppe võimalik abil kommentaaridest, artiklitest, postitustest jms leida kajastatud teemasid, tekste teemade järgi klassifitseerida või eraldada muud väärtuslikku informatsiooni, mis lihtsustab inimese jaoks suurema hulga informatsiooni töötlemist [7]. Käesoleval juhul on masinõppe ülesandeks leida kohtuotsustest juristide ja kohtunike jaoks olulisi lõike ehk eraldada neist väärtuslikku informatsiooni.

Masinõpe loomuliku keele töötlusena tähendab teksti andmetest mudelite loomist ehk tekstide matemaatilist esitust [6]. Masinõppe mudel on andmete peal treenimise tulemus [6]. Mudel õpib sellest, mida ta on varem näinud ehk andmetest ja ülesandest, mis talle ette antakse [6]. See tähendab, et kui treeningandmed on kehvad (sisaldades näiteks ennustuseks ebavajalikku informatsiooni), siis ei saa ka ennustuse tulemus olla hea. Masinõpe jaotub juhendatud ja juhendamata õppeks [6]. Nende vahe on selles, et juhendatud õppe korral on osa dokumentidest märgendatud siltidega, mida masin peaks leidma [6]. Juhendamata õppe korral neid andmeid ei ole ehk masin peab ise sildid leidma [6].

Käesolevas töös kasutatakse juhendamata masinõppe meetodeid. Kohtuotsused sisaldavad märksõnade osa. Samas analoogiliste kohtuotsuste osade leidmiseks võib märksõna otsing jätta välja need lahendid, kus märksõna puudub. Selleks, et saada laiem tulemus, mis juriste ja kohtunikke võiks aidata ning testimaks, kas see võiks toimida olemasolevatest (InfoCuria ja EUR-Lex) süsteemidest paremini, said töös valitud juhendamata masinõppe meetodid. Neid katsetati ilma märksõnadeta kohtulahendite osade peal.

Info tekstist eraldamise ja klassifitseerimismudelite treenimise tehnikaid on mitmeid [5]. Üks lihtsamatest meetoditest põhineb sõnade hulga kokku lugemisel (*Bag-of-Words*), sellest veidi keerulisem on sõnade seoste leidmine (*Word Embedding*) ja veel komplitseeritum konteksti seostel (*Contextualized Embeddings*) põhinev meetod [5]. Käesolevas töös on käsitletud ja kasutatud sõnade hulga meetodil baseeruvat CountVectorizerit ja sõnade seostel baseeruvat Word2Vec meetodit. Põhjus nende kasutamiseks on selles, et eksperimenteerimiseks kasutati fraase, mis koosnevad kahest sõnast ning seetõttu ei sisalda need väga palju

konteksti informatsiooni (eelnev või järgnev sõna, mis konteksti rohkem avaks, kahesõnalise fraasi korral puudub).

Varasemalt on teadustööde raames CountVectorizerit kasutatud näiteks majandusuudiste kokkuvõtete seostamiseks aktsiaturu kõikumistega [8] ja turismiobjektide koondamiseks pealkirjade ja koordinaatide alusel [9]. Word2Veci on kasutatud näiteks teksti lihtsustamise programmis [10] ja masintõlkeks [11]. Eestikeelsete kohtulahendite analüüsimiseks autorile teadaolevalt selliselt neid meetodeid kasutatud ei ole.

Eesti kohtulahendite masintöötlemise kohta on magistritöö tasemel uurimuse teinud Katrin Valdson [12]. Valdson uuris mustripõhise informatsiooni eraldamise võimalusi Eesti kohtulahenditest [12]. Selleks kasutas Valdson *Python*i EstNLTK teeki ning lõi töö tulemusena andmebaasi, kus iga kohtuotsuse jaoks on välja toodud selle segmenteeritud tekst [12]. Selle andmebaasi alusel on võimalik arendada tarkvara, mis võimaldab juristidel mugavamalt ja efektiivsemalt otsida enda jaoks relevantseid kohtulahendeid [12]. Samuti on varasemalt tehtud äri- ja süsteemianalüüs kohtulahendite avalikustamise protsessi osas [13] ning püütud ennustada Euroopa Inimõiguste Kohtu või spetsiifiliselt teatud riikide kohtute otsuseid [14] [15] [16].

Eeltoodud töödes ei ole siiski uuritud võimalust luua CountVectorizer või Word2Vec baasil fraasidele sarnaste Euroopa Kohtu otsuste lõikude kuvamise süsteemi ning seda eesti keele alusel. Seega annab magistritöö ühtlasi panuse Eesti õigustehnoloogia arengusse. Eelnevalt mainiti, et Euroopa Kohtu otsuseid analüüsitakse kahe meetodi abil. Selleks, et aru saada, milliseid tulemusi neist peaks teoreetiliselt saama ja kuidas need toimivad, on järgmises peatükis neid täpsemalt käsitletud ning selgitatud nende kasutamise võimalusi.

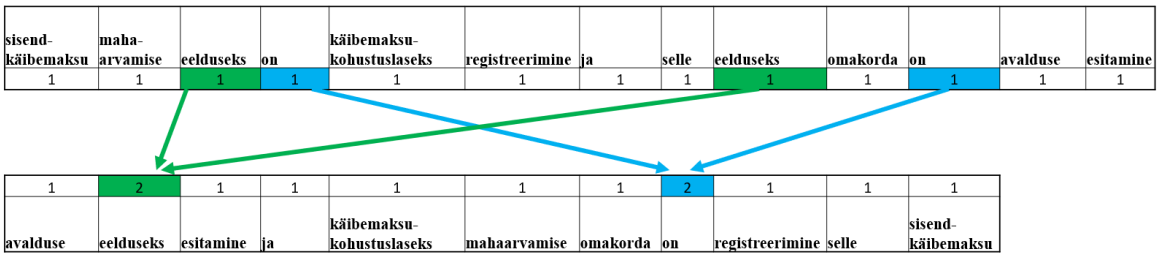
# 3. Vektorite moodustamise ning sarnasuse hindamise mudelid

Selleks, et keel oleks arvuti jaoks töödeldav, tuleb see viia arvude kujule [5]. Selleks, et sõnu arvude kujul näidata, kasutatakse vektoreid [5]. Vektorid paiknevad semantilises ruumis ning on arvutite jaoks mõistetavad [5]. Inimesel on sellist ruumi keeruline ette kujutada, sest selles ruumis on väga palju dimensioone [5]. Semantilises ruumis vektorite esitlemise eesmärk on enamasti leida seoseid sõnade vahel (sarnaseid sõnu) ning selleks mõõdetakse vektorite vahelist kaugust ruumis [5]. Meetodeid, mis kasutavad vektoreid, on palju, kuid alljärgnevalt sai valitud töösse neist kaks: CountVectorizer (koos TfidfVectorizeriga) ja Word2Vec. Alljärgnevalt selgitatakse nende toimimist lähemalt.

## 3.1 CountVectorizer ja TfidfVectorizer meetodid

Üks lihtsamaid viise sõnade numbriliseks vektoresitluseks, on kasutada sõnade kokku lugemise baasil tehnikat [5]. See tähendab, et meetod loeb kokku sõna esinemise ja sõnade koosesinemise sageduse ehk esitab teksti sõnade hulgana (*bag-of-words*) [5]. Selline tehnika on efektiivne teksti klassifitseerimise ülesanneteks [5]. Sõnade hulgal baseeruv tehnika põhineb vektori loomisel, millel on nii palju dimensioone, kui korpuses on unikaalseid sõnu või väljendeid [5]. Kui sõna on ruumis olemas, siis selle esitlus on 1 ja kui pole, siis on see tähistatud 0ga [5]. Iga kord kui sõna esineb, siis suurendatakse vastavat arvu ühe võrra (on ka versioon, kus sõna esinemise korral on esitlus 1 ja selle mitteesinemise korral 0) [5]. Seega eeldab sõnade hulgal põhinev meetod suurt salvestusmahtu ning selle arvutusvõimsus on limiteeritud [5]. Eelnevat selgitab järgmine näide.

Võtame näiteks lause „sisendkäibemaksu mahaarvamise eelduseks on käibemaksukohustuslaseks registreerimine ja selle eelduseks omakorda on avalduse esitamine“. Selle lause unikaalsed sõnad on tähestiku järjekorras järgmised [„avalduse“, „eelduseks“, „esitamine“, „ja“, „käibemaksukohustuslaseks“, „mahaarvamise“, „omakorda“, „on“, „registreerimine“, „selle“, „sisendkäibemaksu“]. Seega tulemuseks saadud vektor on 11 elemendi pikkune (vastab unikaalsete sõnade arvule) [5]. Järgmisena tuleb vaadata, mitu korda need sõnad lauses esinevad, nt sõna „avalduse“ esineb lauses üks kord, aga sõna „eelduseks“ kaks korda [5]. Vastavalt esinemise arvule märgitakse vektorisse sõnade asemel numbrid. Konkreetse lause vektoresitlus on seega [[1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1]] [5]. Joonisel 1 on näide piltlikustatud. Lausete sarnasust hinnatakse selle abil, et vektorid, millel on sarnased numbrite mustrid, on üksteisele sarnasemad [5].




**Joonis 1. Sõnade hulgal põhineva meetodiga vektori moodustamine.** Ülemine tabel on lause tervikuna, rohelise ja sinisega on näidatud kordused. Alumises tabelis on sõnade esinemiste arv kokku loetud.

Teksti esitamiseks sõnade hulga kujul on olemas Pythoni teegis Scikit-learn meetod CountVectorizer(), mida on töös kasutatud [17]. Sõnade hulgal põhinev meetod jagab teksti väiksemateks tükkideks (tookeniteks, inglise keeles *token* ehk näiteks lause sõnadeks või sõnad tähemärkideks) ning loob vastavad vektorid [18]. Samas sõnade hulgal põhinev meetod ei tuvasta sõnade vahelisi seoseid ega semantilist informatsiooni. Samuti ei arvesta sõnade hulgal põhinev meetod sõna tähtsust lausetes [5]. Näiteks sõna „käibemaksukohustuslaseks“ on lauses olulise tähtsusega indikeerimaks, millisest valdkonnast lause on. Samas sõna „on“ esineb lauses kaks korda, kuid ei anna lause sisu kohta olulist informatsiooni eristamaks konkreetset lauset sisu poolest teistest lausetest. Eelduslikult võiks seega enne sõnade hulgal põhinev meetodi kasutamist sellised sõnad (neid nimetatakse stoppsõnadeks ning need ei lisa tekstile olulist väärtust [19]) nagu „on“, „ei“, „ja“ eemaldada, sest muidu on oht, et vektorite võrdlusel võib sõna „on“ saada määravaks kriteeriumiks (selle esinemise rohkuse tõttu), kuigi olulisem on sõna „käibemaksukohustuslaseks“.

Siiski on olemas meetod, mis võimaldab seada sõnadele kaalud, millega sõnade hulgal põhineva meetodi tulemust parandada. Selleks meetodiks on TF-IDF [20]. TF-IDF loogika on vähendada terminite tähtsust, mis on sagedased mitmetes tekstides, kuid mis ei anna teksti kohta selle eristamist võimaldavat informatsiooni [21]. TF-IDF kasutamiseks on Python teegis Scikit-learn meetod TfidfVectorizer, mida on töös kasutatud [17]. TF (*Term Frequency*) tähistab sõnasagedust [21]. Selle arvutamiseks tuleb leida konkreetse sõna esinemiste arv dokumendis ja jagada see kogu sõnade arvuga dokumendis [22]. Dokumendiks on hulk sõnu [23]. Näiteks võib dokumendiks olla lause, lõik või muu pikem tekst [23]. TF abil toimub tulemuste normaliseerimine, mis tähendab, et teksti pikkus ei mõjuta seda, kumb dokument on tähtsam [23]. Ilma sellist normaliseerimist tegemata võiks pikemas dokumendis olla sõna kordusi rohkem ning seetõttu peaks meetod seda olulisemaks [23].

IDF (*Inverse Document Frequency*) tähendab pöördväärtust dokumendi sagedusest, mis mõõdab konkreetse sõna informatiivsust [24]. Selle rakendamise vajadus on tingitud sellest, et hoolimata TF kasutamisest, on stoppsõnade arv tekstides suur ja see omakorda tähendab, et see põhjustab suuri sõnale vastavaid väärtusi ka pärast normaliseerimist [24]. Suured sõna väärtused tähendavad, et nende tähtsus on samuti suur, kuigi tegelikkuses annavad väärtust (võimaldavad sarnasust paremini hinnata) pigem need sõnad, mis ei ole kõigis dokumentides ühtlaselt levinud [24]. Selleks, et hindamise tulemuse täpsust parandada, ongi vajalik IDF täiendav kasutamine [24]. IDF arvutamiseks tuleb esmalt leida sõna esinemise arv dokumentides ja normaliseerimiseks jagada kogu dokumentide arv leitud sõna esinemiste arvuga (kogu dokumentide arv/sõna esinemiste arv dokumentides) [25].

Siiski suuremate dokumentide hulkade korral IDFi selliselt arvutades oleks ühe sõna väärtus endiselt väga suur [26]. Seetõttu võetakse sellest logaritm [26]. Kuivõrd mõnel juhul võib tekkida nulliga jagamine on selle vältimiseks vajalik liita number 1, mistõttu lõplik IDF arvutus näeb välja järgmine: log(kogu dokumentide arv/(sõna esinemiste arv dokumentides+1)). Seega kogu TF-IDF saamiseks tehtav arvutus on näidatud valemis (1) [26].

$$\frac{sõna\ esinemiste\ arv\ dokumendis}{kogu\ sõnade\ arv\ dokumendis} \times \log\left(\frac{kogu\ dokumentide\ arv}{sõna\ esinemiste\ arv\ dokumentides+1}\right) \quad (1)$$

CountVectorizer ja TfidfVectorizer kasutamine võimaldab seega luua vektorid, mis põhinevad sõnade esinemiste arvu kokku lugemisel. Siiski on nende kasutamise puuduseks see, et sõnade vahelisi seoseid ja semantilist infot ei võimalda need tuvastada. Samas on olemas meetodid, mis arvestavad sõnade vaheliste seostega. Järgnevalt käsitletakse Word2Vec meetodit.

## 3.2 Word2Vec meetod

Word2Vec on laialt kasutatav algoritm, mis põhineb närvivõrkudel [5]. Word2Vec ennustusi saab kasutada, et luua sõna seosed teiste sõnadega või klasterdada või klassifitseerida neid teema järgi [27]. Word2Vec on kasulik näiteks automaatse teksti märgendamise, soovitussüsteemide ja masinõppe osas [28]. Samuti saab Word2Veci klastreid kasutada masintõlkeks, info eraldamiseks ja küsimustele vastamise süsteemides, otsinguks, sisu analüüsis näiteks teaduslikes otsingutes, õiguslikes analüüsides, samuti e-kaubanduse ja kliendisuhte halduses [29] [27] [21]. Word2Vec toimib hästi analoogia ja sarnasuse suhetes [21]. Seega on Word2Vec kasutusalad laialdased.

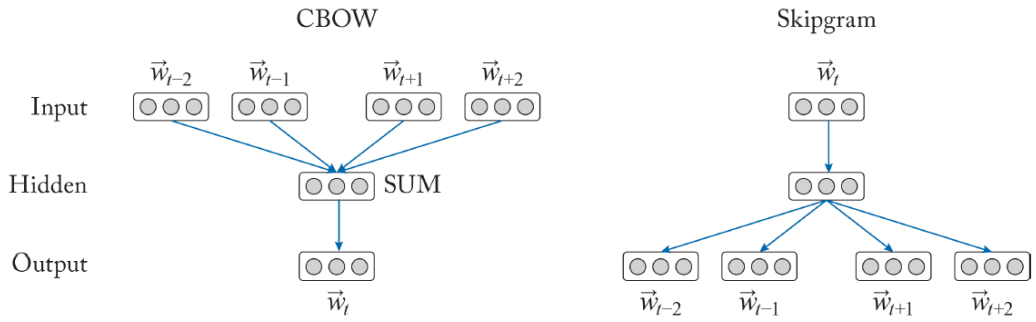
Sarnaselt sõnade hulgal põhinevale meetodile muudab ka Word2Vec sisendiks antud sõnad vektoriteks ja ei vaja selleks märgendatud teksti, kuid erinevus seisneb selles, et Word2Vec arvestab sõnade järjestusega [28]. See tähendab, et Word2Vec kasutab treenimiseks konkreetse sõna naabruses olevaid sõnu (kontekstsõnu) [30]. Word2Vec võib sisendiks antud sõna abil ennustada väljundi konteksti või kasutada sõna konteksti, et ennustada väljundsõna [30].

Word2Vec eelduseks on, et sarnastel sõnadel on enamasti sarnased vektorid (sarnaste väärtustega vektorid) [31]. Word2Vec eesmärgiks on grupeerida ühes vektorruumis kokku sarnaste sõnade vektorid [27]. Sarnasused tuvastatakse seejuures automaatselt [27]. Nimelt loob Word2Vec vektorid, mis on sõnade kontekstide numbrilise esinemise jaotused [27]. Kogu väljundiks on sõnastik, milles iga elemendiga on seotud vektor, mida saab anda ette sügavale närvivõrgule või kasutada lihtsalt sõnade vaheliste seoste tuvastamiseks [27]. See tähendab, et Word2Vec suudab ennustada täpsemalt kui sõnade hulgal põhinev meetod, mida sõna tähendab [27]. Selline ennustus sõltub sõna varasemast ilmnemisest tekstides ning sarnasused tuvastatakse matemaatiliselt [27].

Selleks, et sarnasusi tuvastada kasutab Word2Vec kas Skip-gram või CBOW versiooni (*Continuous bag of words* ehk katkematu sõnade kogum) [5]. Skip-gram ennustab sõnade konteksti (vt Joonis 2) [29]. Täpsemalt öeldes annab Skip-gram tõenäosusjaotuse sisendiks antud sõnale lähedaste sõnade osas. Skip-gram võtab näiteks sõnade paarid - sõna1 ja sõna2 (need saadakse, kui ette antud suurusega nii öelda aken liigub üle kogu teksti andmete) ja treenib nende põhjal 1-peidetud-kihi närvivõrgu [28]. Treenimine põhineb sisendiks antud sõna põhjal [28]. Näiteks kui lauseks on „sisendkäibemaksu mahaarvamise eelduseks on käibemaksukohustuslaseks registreerimine“, siis võimalikud treenimise näited võivad olla sõna „sisendkäibemaksu“ korral (sisendkäibemaksu, mahaarvamise), (sisendkäibemaksu, eelduseks) [5]. Seega proovib Skip-gram meetod ennustada sõna „sisendkäibemaks“ abil naabersõnu („mahaarvamise“ ja „eelduseks“), mis tähendabki konteksti [5]. Ilmselt „sisendkäibemaks“ ja „mahaarvamine“ võiksid olla ligilähedasemad, kui näiteks „sisendkäibemaks“ ja „avaldus“, kuid see sõltub tekstidest, mida treenimiseks kasutada [5].

CBOW (vt Joonis 2) meetod kasutab sõnade konteksti, et ennustada olemasolevat sõna [29]. Selleks vaatab meetod keskmist sõna ümbritsevaid sõnu (mõni sõna enne ja pärast keskmist sõna) [32]. Sõnade järjekord ei ole seejuures oluline [32]. See tähendab, et kui võtta näiteks eelnevalt kasutatud eelnev sisendkäibemaksu mahaarvamise kohta käiv lause, siis CBOW püüab sõnade „sisendkäibemaksu“ ja „eelduseks“ abil ennustada sõna „mahaarvamise“.

Konteksti järgi sarnaseks hinnatud sõnavektorid nihutatakse üksteisele lähemale, kohandades vektoris olevaid kaale [27]. Kui sõnavektorid on hästi treenitud, siis asetavad need sarnased sõnad ruumis üksteisele lähemale [27].




**Joonis 2. CBOW ja Skip-gram mudelite tööpõhimõtted** [5]**.** *Input* – sisend, *Hidden* – varjatud, *Output* – väljund, *SUM* – summeerimine $w_t$ – sõna, $w_{t-2}$, $w_{t-1}$, $w_{t+1}$, $w_{t+2}$ – $w_t$-le eelnevad ja järgnevad sõnad [5].

Üheks Word2Veci miinuseks on see, et Word2Vec arvutab ühe staatilise esitluse iga sõna jaoks ja seda sõltumata kontekstist, milles see sõna esineb. Seega ei suuda Word2Vec interpreteerida sõnu selliselt nagu inimene seda teeks [5]. Word2Vec miinuseks on ka see, et treenimine võib võtta kaua aega, kui laused on ette antud ühe pika jadana [27]. Põhjus on selles, et Word2Vec jaoks on tekstide piirid olulised, sest koosesinemise statistika kogutakse ette antud tekstide järgi [27]. Samuti ei ole meetod võimeline looma vektoreid tundmatute sõnade jaoks [28].

Selleks, et treenimine läheks kiiremini, on võimalik kasutada Gensimi teeki, mis võimaldab laadida kasutamiseks eeltreenitud mudeli, eraldada sõnade vektorid, treenida mudelit algusest ja eeltreenitud mudelit täiustada [33]. Seejuures tuleb treenimiseks kasutada sobivat keeleversiooni, kuivõrd erinevates keeltes on sõnade koosesinemised erinevad [34]. Käesolevas töös on kasutatud Eesti keele koondkorpusel treenitud Word2Vec mudelit [34]. Korpuses on kollektsioon ajalehtede, ilukirjanduse, teaduse ja seaduste tekstidest [34].

Eeltoodust tulenevalt peaks teoreetiliselt Word2Vec mudel toimima paremini, kui sõnade hulgal põhinev meetod, sest hindab lisaks sõnade arvu kokku lugemisele ka naabruses olevaid sõnu ja nende konteksti. Samas Word2Veci täpsus oleneb treenimiseks kasutatud tekstide hulgast ja sisust. Tegemist pole sügava närvivõrguga, mis võib seada piirid ennustuse täpsusele.

## 3.3 Sarnasuse hindamine

Teksti sarnasuse hindamine on üks olulisi tehnikaid, mida loomuliku keele töötluses kasutatakse, et leida kahe teksti tähenduses lähedus [35]. Tekstide sarnasuse hindamine võimaldab infot eraldada, automaatselt vastata küsimustele, masintõlget ja dokumentide võrdlust [35]. Töös otsitakse CountVectorizer ja Word2Vec abil otsingufraasile sarnanevaid kohtuotsuste lõike. Eelnevalt kirjeldatud CountVectorizer ja Word2Vec annavad tulemusteks sõnavektorid, kuid fraas koosneb kahest ning lõik enam kui kahest sõnast. Selleks, et fraasi ja lõigu sarnasust hinnata on töös summeeritud sõnavektorid üheks ning kasutatud sarnasuse hindamiseks koosinussarnasust. Järgnevalt selgitatakse koosinussarnasust ja selle töös kasutamise põhjuseid lähemalt.

Tekstide semantilise sarnasuse hindamiseks tuleb võrrelda mudelite poolt loodud vektoreid [35]. Kuivõrd vektorid on esitatud matemaatiliselt, siis saab võrdluseks kasutada näiteks Eukleidese kaugust (*Euclidean Distance*), koosinuskaugust (*Cosine Distance*), Manhattani kaugust (*Manhattan Distance*) või mõnda muud kauguse mõõtmise meetodit [35]. Praktikas on aga koosinussarnasus on enim kasutatud mõõtevahend [21].

Koosinussarnasus mõõdab kahe nullist erineva väärtusega vektori vahelise nurga koosinust ruumis [36]. Kui nurk on 0-kraadine, siis on tegemist lähedaste vektoritega ja seega sarnaste tekstidega [36] 0-kraadise nurga koosinus on 1 [36]. Koosinussarnasust arvutatakse selliselt nagu valemis (2) on näidatud [37].

$$k(x, y) = \frac{xy^T}{\|x\|\|y\|} \quad (2)$$

Valemis tähistavad x ja y vektoreid, T on transponeerimine (ehk vektor keeratakse ümber), ||x|| ja ||y|| tähistavad vektorite pikkuseid [38] [36].

Koosinussarnasust kasutatakse enamasti juhul, kui vektorite suurus ei ole oluline [39]. Samas on võimalik seda sarnasust kasutada ka siis, kui mõned omadused muutuvad nii, et kaalud on suuremad [39]. See tähendab, et dokumendid on ebaühtlase pikkusega [39]. Tekstiandmed on enamasti kõige tüüpilisemad, mille peal seda mõõtevahendit kasutada [39].

Kuivõrd ka Euroopa Kohtu otsused on tekstiandmed, otsustati töös kasutada samuti koosinussarnasust. Samuti sai meetod valitud põhjusel, et tekstide pikkus ei hakkaks võrdlustulemusi mõjutama, sest lõikude pikkus võib varieeruda paarist sõnast mitme realiste laseteni. Järgnevas peatükis käsitletakse täpsemalt andmeid, mida analüüsis kasutati, andmete puhastamist ning analüüsitakse saadud tulemusi.

# 4. Euroopa Kohtu otsused kui analüüsi alusandmed ja nende eeltöötlemine

Käesolevas peatükis kirjeldatakse andmete töötluseks tehtud ettevalmistavaid tegevusi. Esmalt kirjeldatakse andmeid, mis said valitud analüüsimiseks ning seejärel selgitatakse andmete eeltöötlemise ehk puhastamise toiminguid. Andmete kirjeldus on vajalik, et oleks arusaadav, milliste andmete pinnalt analüüs tehti. Puhastamine oli vajalik selleks, et saada võimalikult täpseid vasteid. Andmete töötluseks kasutati Jupyter Notebookis Pythonit 3.8.5 ja Colaboratorys Pythonit. Peamiselt kasutati teeke pandas, numpy ja BeautifulSoup, kuid teegid on täpsemalt nimetatud järgnevalt ja nähtavad koodist.

## 4.1 Kohtuotsuste struktuur ja andmete kraapimine

Analüüsitavateks andmeteks said valitud Euroopa Kohtu otsused seisuga 20.02.2021. Kuigi Euroopa Kohtu otsuseid tõlgitakse kõikidesse Euroopa Liidu liikmesriikide keeltesse, siis valiti neist välja ainult eesti keelsed kohtuotsused, sest selliste õigustekstide analüüse on praktikas tehtud üksikuid. Lisaks on eesti keel magistritöö autori emakeel ja seetõttu oli tulemusi lihtsam sisuliselt hinnata.

Sissejuhatuses mainiti, et Euroopa Kohtu otsus koosneb enamasti kohtu koja nimetusest, kuupäevast, kohtuasja numbrist, poolte nimedest, menetlusosaliste kirjeldusest, asjakohaste õigusaktide kirjeldusest, asjaolude kirjeldusest, poolte ja menetlusosaliste seisukohtadest, Euroopa Kohtu analüüsist, kohtukulude jaotusest ning resolutsioonist (vaata näidiskohtuotsust Lisast I). Samas ei pruugi need andmed olla lahendis esitatud täpselt selles järjekorras. See tähendab, et mõnel juhul võivad näiteks poolte seisukohad ja kohtu poolne analüüs olla esitatud vaheldumisi igas õigusküsimuse aspektis. Vahel on kohus otsustanud lisada kohtuotsuse liigendusse täiendavaid alapeatükke. Näiteks leidus analüüsitud andmetes pealkirju nimetusega „Kiirmenetlus“, „Vastuvõetavus“ ja „Esialgsed märkused“. Seega kohtuotsuste struktuur võib varieeruda.

Kuivõrd Euroopa Kohtu otsuste kohta puuduvad avaandmed, siis oli esimene samm kohtuotsuste lingid kraapida avalikust InfoCuria andmebaasi otsingu süsteemist. InfoCuria sai valitud põhjusel, et sellest oli võimalik linke struktureeritult ja lihtsalt kätte saada. See tähendab, et kraapimine ei ole InfoCurias tehniliselt kaitstud. Samuti välditi InfoCuria ülekoormamist seeläbi, et iga kohtuotsuse kraapimisele seati ajaline nihe.

Enne kraapimist täpsustati otsingut täpsustati järgmiselt:

- Dokumendid = Kohtulahendite kogumikus avaldatud dokumendid: Kohtuotsused;
- Kohtuasjade menetlusstaadium = Lõpetatud kohtuasjad.

Eelnevalt kirjeldatud täpsustused olid vajalikud, sest selliselt oli võimalik kätte saada lingid terviklahenditele ning välistada dokumendid, mis kohtu seisukohta ei sisalda (näiteks kohtuvaidluse poolte seisukohad, kohtujuristi arvamused jm). Kuigi ka kohtumäärused võivad Euroopa Kohtu seisukohta sisaldada, jäeti need valimist välja seetõttu, et need võivad sisaldada analüüsi ainult Euroopa Kohtu pädevusest ja mitte sisuliste küsimuste vastuseid. Nende välja jätmine ei mõjuta ka oluliselt analüüsi tulemusi, kuivõrd kohtumääruseid kokku oli vähem kui 3000 (võrdlusena kohtuotsuste linke oli 23 804).

Kohtulahendite kättesaamiseks kraabiti esmalt nende lingid Python Selenium paketi abil. Selenium on enamasti mõeldud süsteemide funktsionaalsuste testimiseks [40], kuid sobis hästi linkide kraapimiseks pärast koodi mõningast muutmist. Kohandatud kood linkide kraapimiseks saadi magistritöö ühelt juhendajalt. Kraapimise tulemusena saadi kätte 23 804 linki, mis kirjutati csv-faili, lihtsustamaks nende hilisemat töötlemist[4].

Pärast linkide saamist oli vajalik nende abil kätte saada HTML-vormingus (*HyperText Markup Language* ehk hüperteksti märgistuskeel [41]) kohtuotsuste andmed ning kirjutada need txt-failidesse. See samm oli vajalik, et iga järgmise töötlusega ei peaks InfoCuria andmebaasi päringuid tegema. Seeläbi tagati väiksem koormus vastavale andmebaasile ning teisalt kiirendati hilisemat andmete töötlemise võimalusi. HTML-vormingus andmete kättesaamiseks kasutati Pythoni teeke BeautifulSoup [42] ja pandas [43]. Linkidest saadi 23 804 erineva suurusega faili[5].

## 4.2 Andmete puhastamine

Uurides  saadud faile lähemalt, selgus, et osad neist siiski otsuse sisu ei sisalda. Selle põhjuseks oli asjaolu, et andmed kraabiti alla asendades lingis keel eesti keele märgiga „ET“, kuid ei kontrollitud, kas vastav keeleversioon üldse eksisteerib. Osad kohtulahendid ei pruugi olla eesti keelde tõlgitud. Kohustuslik on need tõlkida alates Eesti liitumisest Euroopa Liiduga ning vahel läheb tõlgete tegemisega ka aega, mistõttu ei pruukinud tõlked olla kraapimisel kättesaadavad.

[4] Vt koodi osa GitHubist – „1. Linkide_linkide_saamine_selenium.ipynb“ ja Google Drivest fail nimetusega „ELlingid.csv“
[5] Vt koodi osa GitHubist – „2. Lingid_txtks.ipynb“, txt-failid on autori valduses ning neid pole tööle lisatud nende suure mahu tõttu

Selleks, et sisutühjad failid ei hakkaks hilisemat töötlust segama, eemaldati kohtuotsused, mis sisaldasid järgmisi inglise keelseid lauseid: *„The document is not available in that language“*, *„This document cannot be found“*, *„The document is not available in that language.“* ja *„This document cannot be found.“*. Pärast nimetatud failide eemaldamist jäi järele 8049 eesti keelset kohtuotsust[6]. Kohtuotsused salvestati eraldi kausta.

HTML-vorminguga failid tähendasid, et need sisaldasid silte (*tag*) ja muid koodiosi. Seega oli raskendatud kohtuotsuste sisu lugemine ning teisalt analüüsi jaoks sisaldasid need andmeid, mis takistasid täpsete tulemuste kuvamist (nt poolte seisukohad, kohtu koosseis jm). Tavapäraselt andmete kraapimisel võivadki andmed sisaldada osi, mis info otsijale on väärtusetud ja kodeeritust [44]. Selleks, et andmed oleksid meetodite jaoks sobivad, tuleb neid esmalt eeltöödelda [45]. Eeltöötlus võib tähendada, et eemaldada on vaja spetsiaalsed tähemärgid, sildid, stopp-sõnad jne [45].

Enne siltide eemaldamist tuli analüüsida, millisele kohtulahendi osale silt viitab. Selle abil lootis töö autor kätte saada kohtulahendi olulised osad (hinnang ja resolutsioon). HTML-vorminguga failidest nähtus, et sildid olid olemas p-klasside all. Beautiful Soup teeki kasutades oli võimalik kätte saada 123 erinevat silti. Seejärel vaatas töö autor iga silti eraldi ning hindas, kas sildi taga olev tekst tähistab kohtu koda, kohtuasja numbrit, kohtumenetluse keelt, märkust, märksõnu, märkuseid, pealkirja, resolutsiooni või teksti osa. Sellise analüüsi tulemusena oli võimalik siltide määratluse abil failidest moodustada üks csv-fail järgmiste veergude pealkirjadega (vt Tabel 1)[7]:

- „Tag“, mis tähistab lõigu silti;
- „Lõigu sisu“, kus on kohtuotsuse sisu tekstid;
- „Pealkiri1“, mille sisu võib olla:
  - „Koda“ – kohtuotsuse teinud kohtu koda;
  - „Kuupäev“ – kohtuotsuse kuupäev;
  - „Viited“ – märksõnad kohtuotsuse sees, mis viitavad, millise sisuga on kohtuotsus;
  - „Poolte_nimed“ – asjaosaliste nimed;
  - „Keel“ – viide sellele, mis keeles originaalselt kohtumenetlust peeti;

[6] Vt koodi osa GitHubist – „3. Txt_sorteerimine(vahendamine).ipynb“, txt-failid on autori valduses ning neid pole tööle lisatud nende suure mahu tõttu

[7] Vt koodi osa GitHubist – „4. Siltide_saamine.ipynb“ ja „5. Andmed_csv-ks.ipynb“ ning Google Drivest fail „PohitabelET_oige.csv“.

    - „Resolutsioon“ – kohtuotsuse resolutsioonosa ning
    - tühi – tähendab, et tegemist on kohtuotsuse tekstiosa.
- „Kohtuasja_nr“, mis tähistab konkreetse kohtuasja numbrit nii nagu see on InfoCuria andmebaasis;
- „Kohtuasja_ID, mis tähistab tinglikult pandud kohtuasja numbrit (lõigud, mis on sama kohtulahendi osad, on sama Kohtuasja_ID-ga. See oli vajalik, sest kohtuasja number võib olla otsinguks liiga pikk ja sisaldada nii numbreid kui ka tähemärke).

**Tabel 1. Näide csv-faili tabeli päisest, mis saadi HTML-vorminguga failidest siltide abil**

| | Tag | Lõigu sisu | Pealkiri1 | Kohtuasja_nr | Kohtuasja_ID |
|---|---|---|---|---|---|
| 0 | C19Centre | EUROOPA KOHTU OTSUS (neljas koda) | Koda | C-416/20 PPU | 0 |
| 1 | C19Centre | 17. detsember 2020 | Kuupaev | C-416/20 PPU | 0 |
| 2 | C71Indicateur | Eelotsusetaotlus – Eelotsuse kiirmenetlus – Po... | Viited | C-416/20 PPU | 0 |
| 3 | C02AlineaAltA | TR | Poolte_nimed | C-416/20 PPU | 0 |
| 4 | C02AlineaAltA | Generalstaatsanwaltschaft Hamburg, | Poolte_nimed | C-416/20 PPU | 0 |

Algselt oli plaan koheselt csv-faili luues siltide abil välja filtreerida kohtu hinnangu osa, kuid probleemiks osutus see, et kohtulahendite struktuur varieerus (mõnel juhul võis kohtulahendis olla 5 erineva tasandi pealkirja, teisel 3). See omakorda tähendas, et kui proovida leida kahe pealkirja vahelist teksti ainult siltide abil, siis see osade kohtulahendite peal toimis, teiste peal aga andis ebatäpseid tulemusi. Näiteks luges programm osadel juhtudel kohtu hinnanguks hoopis kiirmenetluse või seisukoha küsimise vastuvõetavuse analüüsi, mis aga sisuliseks hinnanguks ehk õigusaktide tõlgenduses abistavaks ei ole. Seega otsustati andmed eelnevalt kirjeldatud viisil csv-faili lugeda ning seejärel edasi töödelda. Kuigi juba alguses saanuks jätta kõrvale koja, kuupäeva, poolte nimed, viited osa, siis on need algses failis säilitatud, et võimaldada magistritöö väliselt hilisemaid täiendavaid töötlusi.

Pärast ühtse csv-faili loomist oli vajalik välja filtreerida kohad, kus kohtu hinnang ja resolutsioon sisaldub. Selleks tuli leida pealkirjad, mis viitasid kohtu hinnangule ning kohtukuludele, sest nagu märgitud, siis sildid andnuks ebatäpse tulemuse. Filtreerimiseks leiti esmalt enim levinud hinnangu osade pealkirjad nt „Euroopa Kohtu hinnang“, „Kohtu hinnang“, „Euroopa Kohtu vastus“. Seejärel nende leiti kohtuotsuste Kohtuasja_ID-d, milles vastavad sõnad esinesid ning lahutati neist nende kohtuotsuste Kohtuasja_ID-d, milles neid ei esinenud. Saadud hulgast vaadeldi iga kohtu lahendit üks haaval, et leida täiendavaid kohtu

hinnangu pealkirju. Sama metoodikat kasutati kohtukulusid väljendavate pealkirjade leidmiseks. Kokkuvõtteks leiti kohtu hinnangute ja kohtukulude pealkirjade kogum.

Kohtu hinnangu pealkirjade ning resolutsioonide lõigud märgendati ära kui „algus“ ning kohtukulude lõigud kui „lõpp“. Täiendavalt märgendati ära „Koda“ (see tähistab Kohtu koja nimetust nt Euroopa Kohtu neljas koda) kui „vahetus“. Põhjus on selles, et osades kohtuotsuses puudus kohtukulude osa, seega oli vajalik leida koht, kus uus kohtulahend võiks alata ning selleks sobis kohtu koda hästi (vt Tabel 2). Seejärel filtreeriti välja algusega lahendid ning täideti ära alguse ja lõpu vahepeale kuuluv osa, mis omakorda välja filtreeriti. Kokkuvõttes jäid järele 6438 kohtuasja andmed, milles on kohtu hinnangu ja resolutsiooni osad lõikudena. Lihtsustamaks töötlemist kirjutati ka need andmed eraldi csv-faili[8].

**Tabel 2. Näide andmete filtreerimiseks tehtud töötlusest** (alguse ja lõpu ning vahetuse osa on siin tabelis veel täitmata)

| | Tag | Lõigu sisu | Pealkiri1 | Kohtuasja_nr | Kohtuasja_ID | algus_lopp |
|---|---|---|---|---|---|---|
| **0** | C19Centre | EUROOPA KOHTU OTSUS (neljas koda) | Koda | C-416/20 PPU | 0 | vahetus |
| **1** | C19Centre | 17. detsember 2020 | Kuupaev | C-416/20 PPU | 0 | None |
| **2** | C71Indicateur | Eelotsusetaotlus – Eelotsuse kiirmenetlus – Po... | Viited | C-416/20 PPU | 0 | None |
| **3** | C02AlineaAltA | TR | Poolte_nimed | C-416/20 PPU | 0 | None |
| **4** | C02AlineaAltA | Generalstaatsanwaltschaft Hamburg, | Poolte_nimed | C-416/20 PPU | 0 | None |
| **...** | ... | ... | ... | ... | ... | ... |
| **1031875** | C01PointnumeroteAltN | 95 Kõnesoleva hüvitise hindamisel tuleks ... | NaN | F-1/05 | 23801 | None |
| **1031876** | C48DispositifIntroduction | Esitatud põhjendustest lähtudes | Resolutsioon | F-1/05 | 23801 | algus |
| **1031877** | C08Dispositif | 1. Tühistada Euroopa Koolitusfondi 25. ju... | Resolutsioon | F-1/05 | 23801 | algus |
| **1031878** | C08Dispositif | 2. Pooled teatavad Avaliku Teenistuse Koh... | Resolutsioon | F-1/05 | 23801 | algus |
| **1031879** | C08Dispositif | 3. Otsus kohtukulude kohta tehakse hiljem. | Resolutsioon | F-1/05 | 23801 | algus |

Eespool mainiti, et üks võimalus andmeid eeltöödelda on eemaldada stopp-sõnad [45]. Samuti otsustati eemaldada kirjavahemärgid, kuna autori hinnangul ei kanna need õigustekstides lisatähendusi. Eestikeelsed stopp-sõnad saadi Kristel Uiboaed poolt koostatud andmestikust [46]. Eestikeelseteks stopp-sõnadeks on näiteks sõnad „on“, „ei“, „tere“, „sa“, „teie“, „oleksid“ [46]. Kirjavahemärkide leidmiseks kasutati Python tekstitöötlus teenuse String konstanti string.punctuation [47]. Kirjavahemärkideks on näiteks koma, hüüumärk, küsimärk jne [44]. Stopp-sõnad ja kirjavahemärgid eemaldati andmestikust. Samuti muudeti

[8] Vt koodi osa GitHubist – „6. Filtreerimine_kohtulahendid_hinnang.ipynb“ ja Google Drivest fail „Pohitabel_filt_algusega.csv“

suurtähed väiketähtedeks, et sõnad oleksid ühtsed (ei eristuks näiteks sõna „Kaks“ sõnast „kaks“. Tulemus liideti andmestiku veergu ning kirjutati eraldi csv-faili[9].

Kokkuvõttes saadi andmete puhastamise tulemusena kolm faili:

- fail, mis sisaldab nii kohtu hinnangu kui resolutsiooni osi (317 521 rida)[10],
- fail, mis sisaldab ainult hinnangu osa (295 245 rida)[11] ja
- fail, mis sisaldab ainult resolutsiooni osa (22 276 rida) [12].

Enne Word2Vecile nende failide ette andmist lemmatiseeriti vastavad andmed kasutades teegi EstNLTK klass Text isendit morph_analysis.lemma[13]. Lemmatiseerimine võimaldas kätte saada sõnade algvorm ja sõnaliik. Eeltöötlus kiirendas Word2Vec mudeli kasutamist. CountVectorizer puhul kasutati lemmatiseerimist meetodi sees ehk koodi osana. Kuigi ilmselt saanuks andmeid puhastada veelgi, siis arvestades töö suunda (võrrelda erinevate meetoditega töötlusest saadud tulemusi), otsustati eelnevalt kirjeldatud puhastamise toimingutega piirduda. Järgnevas peatükis esitatakse analüüsi metoodika ja tulemused.

[9] Vt koodi osa GitHubist – „7. Eemalda_stop_sõnad.ipynb“ ja Google Drivest failid „Pohitabel_stopid_eemald_lower.csv“, „Pohitabel_ainult_reso.csv“, „Pohitabel_ilma_resota.csv“
[10] Vt Google Drivest faili „Pohitabel_stopid_eemald_lower.csv“
[11] Vt Google Drivest faili „Pohitabel_ilma_resota.csv“
[12] Vt Google Drivest faili „Pohitabel_ainult_reso.csv“
[13] Vt koodi osa GitHubist – „9. Lemmad.ipynb“ ja Google Drivest faile „Koondtabel_lemmadega.csv“, „Koondtabel_lemmadega_ainult_reso.csv“ ja „Koondtabel_lemmadega_ilma_resota.csv“

# 5. Euroopa Kohtu otsuste analüüs ja tulemused

Eelnevalt kirjeldatud eeltöötluse tulemusena saadi Euroopa Kohtu otsuste andmete töötlusest kohtuotsuste lõikudega failid, mida on võimalik pöörata Excel failideks. Seeläbi on võimalik failidest otsida lihtsa vaevaga märksõnu. Samas ei võimalda Excel otsing leida sõnale või fraasile sarnaseid sõnu või lõike ning üks haaval märksõnade abil tabeli läbi käimine võib olla samuti aeganõudev protsess. Selleks, et otsingut lihtsustada, kasutati CountVectorizerit (koos TfidfVectorizeriga) ja Word2Veci. Järgnevalt on toodud analüüsi metoodika täpsemalt ning selgitatud saadud tulemusi.

## 5.1 Analüüsi metoodika

Andmete analüüsimiseks kasutati CountVectorizerit (koos TfidfVectorizeriga) ja Word2Veci, et saada kätte lõikude vektorid, mida testimiseks kasutatava fraasidega võrrelda. CountVectorizerit ja Word2Veci kasutati andmestiku kolme eri versiooni peal (kogu andmestik ehk kohtu hinnangu ja resolutsiooni lõigud, andmestik ilma resolutsioonita ja andmestik ainult resolutsiooniga). See oli vajalik, et näha, kuidas meetodid toimivad erinevate andmemahtudega ning samuti spetsiifilisemate tekstidega.

Meetodite headuse hindamiseks kasutati kümmet maksuõigusest pärinevat fraasi:

- „sisendkäibemaksu mahaarvamine“;
- „käibemaksukohustuslaseks registreerimine“;
- „topeltmaksustamise vältimine“;
- „õiguse kuritarvitamine“;
- „maksupettuses osalemine“;
- „maksueelise saamine“;
- „litsentsitasude maksustamine“;
- „ettevõtlusega tegelemine“;
- „dividendide maksustamine“;
- „tagatise määramine“.

Fraasid valis töö autor seetõttu, et oli suuteline vastavas valdkonnas varasema kogemuse pinnalt tulemusi sisuliselt (juriidilisest vaatest) hindama. Samuti vähemalt osade neist (nt „sisendkäibemaksu mahaarvamine“, „dividendide maksustamine“) kohta leidub hulgaliselt kohtupraktikat, mistõttu sai eeldada, et saadakse andmed, mida hinnata. CountVectorizeri

ja Word2Veci kasutamise tulemusena saadud sõnavektorid summeeriti kokku ning võrreldi fraaside summeeritud vektoritega kasutades koosinussarnasuse hindamist.

Koosinussarnasuse hindamise tulemused järjestati ning neist võeti välja 200 kõige sarnasemat teksti[14]. Tulemused kirjutati csv-failidesse ning csv-failid konverteeriti omakorda Exceli failideks, võimaldamaks nende käsitsi hindamist.[15] 200-st tulemusest filtreeriti välja esimesed 20, sest neid vaadataks esmajärjekorras ning nutikama märksõnaotsingu mõte oli, et tulemustest peaks olema võimalik kiiresti vastuseid leida, mis tähendab, et enim asjakohaseid lõike peaks olema esimeste tulemuste seas. Lisaks arvestades, et meetodite headust testiti kümne fraasiga, millest iga kohta saadi kolm tabelit ning 20 rida andmeid (iga rida tähendab ühte lõiku), siis tähendas see kahe meetodi kohta 10x3x20x2=1200 rida andmeid, mis tuli töö autori poolt käsitsi üle vaadata ja märgendada. Rohkemate ridade märgendamise või rohkemate fraasidega testimise korral oleks töö maht läinud väga suureks. Seetõttu piirduti kümne fraasiga ja iga fraasi osas 20 esimese tulemuse hindamisega.

Iga tabeli 20 rea osas märgendati ära tulemused, milles esinesid järgmised osalised sõnad (vt näitena Tabel 3 neljas veerg):

- „sisendkäibe“, „mahaarva“, “maha arva“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „sisendkäibemaksu mahaarvamine“);
- „maksukohus“, „registr“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „käibemaksukohustuslaseks registreerimine“);
- „topelt“, „välti“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „topeltmaksustamise vältimine“);
- „õigus“, „kuri“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „õiguse kuritarvitamine“);
- „pettus“, „osal“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „maksupettuses osalemine“);
- „eelis“, „saam“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „maksueelise saamine“);

[14]Vt koodi osa GitHubist – „8. Oma CountVectorizer1.ipynb“ ja „10. Word2vec.ipynb“

[15] Vt Google Drivest kaustades „dividend“, „ettevõtlusega tegelemine“, „KMKR“, „kuritarvitamine“, „litsentsitasud“, „maksueelis“, „maksupettus“, „sisendkäibemaks“, „tagatis“, „topeltmaks“ olevaid faile. Failid, mille pealkiri sisaldab „Count_vec“ tähistavad CountVectorizeriga saadud tulemusi; „Word2Vec“ tähistab Word2Veciga saadud tulemusi; „ilma_resota“ tähistab ilma resolutsioonita andmetega saadud tulemusi; „ainult_reso“ tähistab ainult resolutsiooniga andmetega saadud tulemusi; „hindamine“ failidest on näha käsitsi läbi viidud hindamine. Kaustade pealkirjade selgitused on Google Drive failis „Analüüside_koondtabel.xlsx“ lehel „Faililaiendid“

- „litsents“, „maks“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „litsentsitasude maksustamine“);
- „ettevõt“, „tegel“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „ettevõtlusega tegelemine“);
- „divid“, „maks“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „dividendide maksustamine“);
- „tagatis“, „määr“ (tabelid, mis saadi sarnasuse hindamise tulemusena fraasiga „tagatise määramine“).

Sellise märgendamise abil oli võimalik näha, kuivõrd mudel hindas sarnasust fraasis sisalduvatest sõnadest lähtuvalt.

Täiendavalt lisati igale lõigule kohtulahendi sisu selgitav märksõna (vt Tabel 3 viies veerg, näiteks sisendkäibemaksu mahaarvamine on käibemaksu valdkonda kuuluv fraas). See on abistavaks informatsiooniks, et näha, millistest valdkondadest on saadud lõigud. See märgendamine on tehtud käsitsi lähtudes sõnade sisaldusest lõigus (näiteks, kui lõigus sisaldus sõna käibemaks, siis tegemist oli käibemaksu valdkonda kuuluva lahendiga ja lisati märksõna „Käibemaks“) ning juhul, kui sõna lõigus ei sisaldunud, käidi läbi lahendid käsitsi InfoCuria andmebaasis, et hinnata, milline märksõna võiks kohtuotsust iseloomustada.

Viimaseks hinnati, kas kohtulahendi lõigust võib olla juristile või kohtunikule kasu. Täpsemalt vaadati, kas lõik on arusaadav eraldiseisvana, kas tegemist võiks olla kohtu hinnanguga ning kas lõigu sisu annab informatsiooni otsitava fraasi kohta või võiks olla vastava otsingufraasi osas abistav (vt Tabel 3 kuues veerg). Kasulikud ei olnud näiteks lõigud, mis on jäänud poolikuks. Näites 1 toodud lõik on üks neist, mis on poolik.

**Tabel 3. Väljavõte ühest hindamise tabelist[16] Word2Vec meetodiga, kõik andmed, „topeltmaksustamise vältimine“**

| Koos_sarnasus | Jarjekord | Tag | Hinnang sisaldab topelt (x) või välti (v) | Kohtulahendi sisu selgitav märksõna | Kasulik? | Lõigu sisu | Pealkiri1 | Kohtuasja_nr |
|---|---|---|---|---|---|---|---|---|
| 0.6890180110931396 | 1 | C01PointnumeroteAltN | v | Tulumaks | Ei | 57 | Järelikult ei saa nõustuda maksupettuste ja n | C-484/19 |
| 0.6745551824569702 | 2 | C01PointnumeroteAltN | xv | Tulumaks | Ei | 53 | Seoses topeltmaksustamise vältimise leping | C-562/07 |
| 0.6745449304580688 | 3 | C01PointnumeroteAltN | v | Tulumaks | Ei | 99 | Seetõttu ei saa asutamisvabaduse piirangut a | C-504/16 ja C-613/16 |
| 0.6452925205230713 | 4 | C01PointnumeroteAltN | v | Tulumaks | Ei | 95 | Saksamaa Liitvabariik leiab piirangu põhjend | C-504/16 ja C-613/16 |
| 0.6440877914428711 | 5 | C01PointnumeroteAltN | v | Tööõigus | Ei | 54 | Odvolací finanční ředitelství lisab, et need õ | C-53/13 ja C-80/13 |
| 0.6414346098899841 | 6 | C01PointnumeroteAltN | xv | Tulumaks | Ei | 34 | Kõigepealt tuleb rõhutada, et Saksamaa Liitv | C-345/04 |
| 0.6400985717773438 | 7 | C01PointnumeroteAltN | v | Tulumaks | Ei | 63 | Seega ei ole põhikohtuasjas kõnealused mak | C-322/11 |
| 0.6349479556083679 | 8 | C01PointnumeroteAltN | xv | Tulumaks | Ei | 32 | Käesolevas asjas nähtub eelotsusetaotlusest, | C-157/10 |
| 0.6328774094581604 | 9 | C01PointnumeroteAltN | xv | Tulumaks | Jah | 42 | Käesoleva otsuse punktist 39 nähtub siiski, | C-416/17 |
| 0.6271044015884399 | 10 | C01PointnumeroteAltN | xv | Tulumaks | Ei | 21 | Tuleb tõdeda, et Põhjamaade Nõukogu liikm | C-319/02 |
| 0.6262299418449402 | 11 | C01PointnumeroteAltN | v | Tulumaks | Ei | 32 | Portugali Vabariik tugineb viimasena põhjen | C-503/14 |
| 0.6198920607566833 | 12 | C01PointnumeroteAltN | xv | Tulumaks | Jah | 49 | Selles osas on selge, et Portugali ja Tuneesi | C-464/14 |
| 0.6128898859024048 | 13 | C01PointnumeroteAltN | xv | Tulumaks | Ei | 65 | Hispaania Kuningriik on lisaks märkinud, et | C-487/08 |
| 0.6128876209259033 | 14 | C01PointnumeroteAltN | v | Tulumaks | Ei | 61 | Mis puudutab piirangu põhjendatust ja propo | C-6/16 |
| 0.6114983558654785 | 15 | C01PointnumeroteAltN | xv | Tulumaks | Jah | 69 | Nimelt ei saa liikmesriik asutamislepingust | C-303/07 |
| 0.6090998649597168 | 16 | C01PointnumeroteAltN | xv | Tulumaks | Jah | 40 | Osaluste puhul, millele direktiiv 90/435 ei l | C-487/08 |
| 0.6088991761207581 | 17 | C01PointnumeroteAltN | v | Tulumaks | Ei | 50 | Lisaks tuleb nentida, et see tasaarveldus võib | C-388/14 |
| 0.6068427562713623 | 18 | C01PointnumeroteAltN | v | Tulumaks | Ei | 61 | Selleks et maksudest kõrvalehoidumise ja m | C-322/11 |
| 0.6046936511993408 | 19 | C01PointnumeroteAltN | v | Tulumaks | Ei | 50 | Portugali Vabariik tugineb põhjendustele, m | C-503/14 |
| 0.6045740842819214 | 20 | C01PointnumeroteAltN | v | Tulumaks | Ei | 77 | Järelikult ei saa vastuvõetavaks pidada põhje | C-484/19 |
| 0.5315229892730713 | 128 | C01PointnumeroteAltN | | Tulumaks | Jah | 68 | Dividendide maksustamine ja mitteresidendi | C-480/16 |
| 0.5299570560455322 | 132 | C08Dispositif | | Tulumaks | Ei | 2. | Liikmesriigi õigusnormid, mis näevad ette te | Resolutsioon C-164/12 |
| 0.52896648645401 | 137 | C01PointnumeroteAltN | | Tulumaks | Ei | 72 | Olgugi et kõnesoleva kapitalikasumi maksus | C-591/13 |
| 0.5279883742332458 | 139 | C01PointnumeroteAltN | | Tulumaks | Ei | 99 | Kui aga liikmesriik on otsustanud residentis | C-190/12 |
| 0.5263089537620544 | 145 | C01PointnumeroteAltN | | Tulumaks | Ei | 55 | Neil asjaoludel, nagu märgib sisuliselt ka ko | C-322/11 |

**Näide 1. Poolik kohtuotsuse lõik** (sulgudes on viide kohtulahendi numbrile)

„– lubas reisibüroodel teatavatel juhtudel märkida arvele käibemaksu kogusumma, millel puudub seos kliendi tegelikult makstava käibemaksuga, ja lubas kliendil siis, kui ta on maksukohustuslane, arvata see kogusumma tasumisele kuuluvast käibemaksust maha, ning“ (C-189/11)

Poolikuks jäänud lõigud võisid olla tingitud sellest, et kohati kasutati kohtuotsustes loetelusid ning sel juhul andmete csv-faili kirjutamisel arvestas programm iga loetelu elementi kui eraldi lõiku. Kasulikud ei olnud ka lõigud, kus on kirjeldatud ainult vaidlusalust küsimust (vt Näide 2). Sellised lõigud ei ole abiks õiguse tõlgendamise juures. Kasulikud olid lõigud, mis sisaldasid kohtu hinnangut (vt Näide 3). Näites 3 toodud lõigust on näha, et kohus selgitab, mida sisendkäibemaksuna maha võib arvata ning sealjuures on viide varsemale praktikale. Seega sellised lõigud märgendati kui kasulikud.

**Näide 2. Ainult vaidlusalust küsimust kirjeldav lõik** (sulgudes on viide kohtulahendi numbrile)

„28 Nagu nähtub eelotsusetaotluse esitanud kohtu esitatud andmetest, on esimese küsimuse eesmärk kindlaks teha, kas liikmesriigi maksuhalduril on õigus teha maksukohustuslase suhtes maksuotsust, mis on tingitud algul tehtud mahaarvamise korrigeerimisest vastavalt riigisisestelt kaubatarnetelt hiljem saadud hinnaalandustele.“(C-684/18)

[16] Vt terviktabelit Google Drive kaustast „topeltmaks“ pealkirjaga „Word2vec_hindamine_topeltmaks.xlsx“ ning selle lehte „Word2vec_tulemus_lemmatud_topel“

**Näide 3. Kasulik lõik** (sulgudes on viide kohtulahendi numbrile)

“ 21 Nimelt võib ühises käibemaksusüsteemis maha arvata üksnes maksu, millega maksustati eelmises etapis maksukohustuslase maksustatavate tehingute jaoks kasutatud kaubad või teenused. Sisendkäibemaksu mahaarvamine on seotud maksu kogumisega vahetult järgnevas etapis. Kui maksukohustuslase poolt omandatud kaupu või teenuseid kasutatakse tehingutes, mis on maksust vabastatud või ei kuulu käibemaksu kohaldamisalasse, ei või käibemaksu arvestada ega sisendkäibemaksu maha arvata (vt selle kohta 11. aprilli 2018. aasta kohtuotsus SEB bankas, C-532/16, EU:C:2018:228, punkt 38).“ (C-374/19)

Pärast märgendamist loeti kokku, mitmes tabeli reas esines osalisi otsingusõnu kokku[17]. Samuti loeti kokku, kui palju esines lõikudes olulisi otsingusõnu[18]. Oluliste otsingusõnade all on mõeldud osalisi otsingusõnu, mis määratlevad fraasi peamise sisu. Nendeks on „sisendkäibe“, „maksukohus“, „topelt“, „kuri“, „pettus“, „eelis“, „litsents“, „ettevõt“, „divid“, „tagatis“. Samuti loeti kokku, kuivõrd palju oli kasulikke kohtuotsuste lõike[19].

Täiendavalt filtreeriti igast tabelist välja maksimaalselt 5 rida, mis otsingusõnu ei sisaldanud ning vaadati, kas need võiksid olla kasulikud (vt Tabel 3 oranžiga märgendatud lahendid). See oli vajalik hindamaks, kas ka ilma otsingusõnata kohtuotsuse lõigud võiksid olla lugejale kasulikud[20]. Samas tuleb märkida, et iga fraasi tulemuste osas neid ei leidnud.

Saadud tulemuste osas arvutati täpsus protsentides, võimaldamaks nende paremat võrdlust[21]. Täiendavalt otsiti iga fraasi InfoCuriast ja EUR-Lexist ning võeti välja esimene lõik (mis vastas olulisele otsingusõnale) ja resolutsioon kuni 20-st esimesest tulemusest alates kohtu hinnangust (iga fraasi kohta need otsingud 20 kohtuotsust ei andnud)[22]. Esimese lõigu väljavõtmine oli vajalik selleks, et imiteerida juristi otsingut. See tähendab, et olemasolevate otsingusüsteemidega vaataks jurist või kohtunik esmalt resolutsiooni ning seejärel püüaks leida esimese lõigu, kus otsingusõna sisaldub. Kuigi praktikas liiguks lugeja järgmiste kohtuotsuste osade juurde, kui vajalikku infot esimene lõik ei sisalda, siis see lõik võimaldab hinnata, kas CountVectorizeri ja Word2Veci tulemused võiksid olla paremad Ctrl+f sõnade

[17] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „Kogu“, „Ilma resolutsioonita“ ja „Ainult resolutsioon“
[18] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „Kogu“, „Ilma resolutsioonita“ ja „Ainult resolutsioon“
[19] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „Kogu“, „Ilma resolutsioonita“ ja „Ainult resolutsioon“
[20] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehte „Otsingusõnata tulemused“.
[21] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „%Kogu“, „% Ilma resolutsioonita“ ja „% Ainult resolutsioon“
[22] Vt Google Drivest faile „InfoCuria andmed.xlsx“ ja „EUR-Lex andmed.xlsx“

otsingust. Lisaks tehti täiendav otsing InfoCuriast ja EUR-Lexist lemmatiseeritud fraasidega[23]. See oli vajalik nägemaks, kas lemmatiseeritud tulemus võiks anda teistsuguseid vasteid. Saadud tulemused salvestati Exceli failidesse, neile lisati juurde iga kohtuotsust iseloomustav märksõna ja hinnati nende kasulikkust[24]. Tulemuste osas arvutati nende täpsus samuti protsentides võimaldamaks võrdlust CountVectorizeri ja Word2Veciga[25].

Kuivõrd CountVectorizer ja Word2Vec võisid anda tulemuseks sama lahendi mitmeid lõike, siis selleks, et võrdlused oleksid InfoCuria ja EUR-Lexiga võrreldavad, loeti kokku kui palju on CountVectorizeri ja Word2Vec 20 esimese tulemuse seas erinevaid lahendeid ja kui palju neist oleksid lugejale kasulikud[26]. Seeläbi on võrdlus InfoCuria ja EUR-Lexi andmetega täpsem.

Analüüsi koondtulemused on salvestatud ülevaatlikkuse huvides eraldi Exceli faili[27]. Järgnevalt on esitatud analüüside tulemused ning nendest tehtud järeldused.

## 5.2 Analüüsi tulemused

Analüüsi tulemused näitasid, et nii Word2Veci kui ja CountVectorizerit saab kasutada kasulike kohtulahendite lõikude leidmiseks. Siiski kohtuotsuste resolutsioonide leidmiseks tuleb rohkem andmeid puhastada, kui seda magistritöö raames teha jõuti. Alljärgnevalt on selgitud täpsemalt nende tulemuste põhjuseid.

Kui vaadata tulemusi, mis on saadud kõigi andmete peal treenimise tulemusena (vt Tabel 4), siis nähtub, et enim otsingusõnu leidus Word2Vec tulemustes. Samuti sisaldasid Word2Vec tulemused ka enim olulisi otsingusõnu, mistõttu oli kasulikke lahendeid rohkem kui CountVectorizeri tulemustes. Kui võrrelda neid tulemusi ainult kohtu hinnangu andmete peal treeningu tulemustega (vt Tabel 5), siis nähtub, et CountVectorizeri abil on leitud võrreldes kogu andmetel treenimisega rohkem otsingusõnu. Oluline vahe nähtub sellest, kui vaadata olulise otsingusõna ja kasulike otsuste lõikude tulemusi. On näha, et CountVectorizer on ainult kohtu hinnangu andmete (ilma resolutsioonita andmed) peal treenituna mõlema osas toiminud paremini, kui Word2Vec. See tähendab, et kui andmed on

[23] Vt Google Drivest faile „InfoCuria andmed.xlsx“ ja „EUR-Lex andmed.xlsx“ igal lehel osa „Lemmatiseeritud“
[24] Vt Google Drivest faile „InfoCuria andmed.xlsx“ ja „EUR-Lex andmed.xlsx“
[25] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „InfoCuria tulemused“ ja „EUR-Lex tulemused“.
[26] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“ lehti „Vektorid ilma kordusteta“ ja „Vektorid % ilma kordusteta“.
[27] Vt Google Drivest faili „Analüüside_koondtabel.xlsx“

spetsiifilisemad (ja neid on vähem), siis toimib CountVectorizer Word2Vec-st paremini, andes rohkem kasulikke lõike.

Samas tuleb mainida, et Word2Vec toimib kiiremini CountVectorizeri. See tähendab, et Word2Veciga saadi tulemused umbes 15-30 minutiga, samas kui CountVectorizeriga tulemuste saamine võis võtta aega paar tundi (kasutati Colaboratoryt, mille mälumaht oli 13305360 kB ja kasutas 2 tuumalist protsessorit). Seetõttu võib kasutatavuse osas ning arvestades kasulike kohtuotsuste lõikude arvu väikeseid erinevusi, olla eelistatud siiski Word2Vec.

**Tabel 4**[28]**. Kõikide andmete (kohtu hinnang + resolutsioon) koondtulemused protsentides** (välja arvatud viimane tulp, mis näitab tükke kokku). Suurema väärtusega tulemused on tumedama värviga tähistatud.

| | sisendkäibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topelt-maksustamise vältimine | õiguse kuri-tarvitamine | maksupettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine | Kokku tk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CountVectorizer otsingusõnu kokku | 20 | 45 | 10 | 50 | 35 | 10 | 55 | 15 | 60 | 35 | 67 |
| Word2Vec otsingusõnu kokku | 100 | 100 | 100 | 100 | 100 | 95 | 100 | 90 | 100 | 100 | 197 |
| CountVectorizer oluline otsingsõna (allajoonitud) | 10 | 45 | 5 | 10 | 20 | 10 | 40 | 5 | 35 | 15 | 39 |
| Word2Vec oluline otsingsõna (allajoonitud) | 55 | 20 | 45 | 20 | 0 | 0 | 0 | 30 | 100 | 40 | 62 |
| CountVectorizer kasulik | 15 | 10 | 5 | 0 | 25 | 5 | 15 | 5 | 10 | 15 | 21 |
| Word2Vec kasulik | 65 | 0 | 20 | 15 | 25 | 0 | 0 | 50 | 75 | 30 | 56 |
| CountVectorizer maksu-;tolliõiguse valdkondi | 65 | 13 | 15 | 1 | 17 | 7 | 12 | 5 | 15 | 4 | 104 |
| Word2Vec maksu-;tolliõiguse valdkondi | 100 | 30 | 95 | 25 | 0 | 5 | 90 | 30 | 95 | 40 | 102 |

**Tabel 5**[29]**. Ilma resolutsioonita andmete (kohtu hinnangu osa) koondtulemused protsentides** (välja arvatud viimane tulp, mis näitab tükke kokku). Suurema väärtusega tulemused on tumedama värviga tähistatud.

| | sisendkäibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topeltmaksustamise vältimine | õiguse kuri-tarvitamine | maksupettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine | Kokku tk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CountVectorizer otsingusõnu kokku | 90 | 100 | 90 | 65 | 80 | 95 | 90 | 70 | 95 | 85 | 172 |
| Word2Vec otsingusõnu kokku | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 80 | 100 | 95 | 195 |
| CountVectorizer oluline otsingsõna (allajoonitud) | 90 | 100 | 90 | 50 | 80 | 95 | 90 | 55 | 85 | 45 | 156 |
| Word2Vec oluline otsingsõna (allajoonitud) | 55 | 15 | 40 | 15 | 0 | 0 | 0 | 20 | 100 | 45 | 58 |
| CountVectorizer kasulik | 50 | 20 | 35 | 20 | 40 | 70 | 40 | 10 | 60 | 25 | 74 |
| Word2Vec kasulik | 70 | 0 | 15 | 10 | 45 | 0 | 0 | 55 | 75 | 30 | 60 |
| CountVectorizer maksu-;tolliõiguse valdkondi | 95 | 100 | 90 | 15 | 95 | 45 | 55 | 40 | 95 | 15 | 129 |
| Word2Vec maksu-;tolliõiguse valdkondi | 100 | 25 | 95 | 20 | 0 | 5 | 100 | 30 | 100 | 25 | 100 |

Kehvemini on Word2Vec ja CountVectorizer toiminud aga resolutsioonandmete osas (vt Tabel 6). Selgituseks, et resolutsioon on kohtulahendi kokkuvõttev lühike osa (1-5 lõiku) ehk osa, kus kohus kirjutab, mida vaidlusküsimusele vastata ja seda loevad juristid ning kohtunikud esmajärjekorras. Samas kohtu analüüs või hinnang ehk põhjendus, kuidas

[28] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „%Kogu“
[29] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „% Ilma resolutsioonita“

resolutsioonini jõuti, võib olla 10-50 lõiku pikk. Seetõttu oli resolutsiooni andmeid analüüsimiseks vähem kui teisi andmeid (üle 20 000 rea, samal ajal kui kohtu hinnangu andmeid oli peaaegu 300 000 rida), millest võivad tingitud olla ka kehvemad tulemused. Samuti tuleb resolutsioonandmete osas arvestada, et need sisaldavadki üldiselt kordustena vähem spetsiifilisemaid üksikuid sõnu (näiteks „sisendkäibemaks“, „käibemaksukohustuslane“) kui neid on hinnangu osas. Põhjus on selles, resolutsioonid on konkreetsed ja enamasti erinevad, samas kui kohtu hinnangu osas pigem mainitakse spetsiifilisemaid sõnu korduvalt, sest resolutsiooni ehk järelduseni jõudmist tuleb eraldi põhjendada ning põhjendusel võidakse kasutada varasemaid kohtuotsuseid.

Samuti vaadates Tabelis 6 toodud tulemusi, on kehvemate tulemuste põhjus ilmselt ka andmete puhastamises, sest sisse on jäänud sissejuhatavad lõigud (näiteks „Esitatud põhjendustest lähtudes Euroopa Kohus (esimene koda) otsustab:“[30]), mis on juristi ja kohtuniku jaoks väärtusetud. See tähendab, et andmete puhastamine ei õnnestunud resolutsioonandmete osas hästi, sest analüüsimiseks jäi sisse ebavajalikku informatsiooni. Samas on Tabelist 6 näha, et hoolimata andmete kehvemast kvaliteedist, on Word2Vec suutnud siiski leida kasulikke lahendeid näiteks dividendide maksustamise osas 65%. See võib olla tingitud asjaolust, et neid lahendeid ja dividendi nimetamist resolutsioonides oli rohkelt.

**Tabel 6[31]. Ainult resolutsioonandmete koondtulemused protsentides** (välja arvatud viimane tulp, mis näitab tükke kokku)

| | sisendkäibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topeltmaksustamise vältimine | õiguse kuritarvitamine | maksupettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine | Kokku tk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CountVectorizer otsingusõnu kokku | 5 | 10 | 5 | 20 | 10 | 5 | 15 | 5 | 15 | 10 | 20 |
| Word2Vec otsingusõnu kokku | 55 | 30 | 5 | 100 | 0 | 5 | 80 | 15 | 100 | 25 | 83 |
| CountVectorizer oluline otsingsõna (allajoonitud) | 5 | 10 | 5 | 0 | 0 | 0 | 10 | 5 | 5 | 0 | 8 |
| Word2Vec oluline otsingsõna (allajoonitud) | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 10 | 55 | 0 | 14 |
| CountVectorizer kasulik | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 5 | 0 | 4 |
| Word2Vec kasulik | 30 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 65 | 0 | 20 |
| CountVectorizer maksu-;tolliõiguse valdkondi | 45 | 30 | 25 | 15 | 20 | 20 | 45 | 5 | 20 | 30 | 51 |
| Word2Vec maksu-;tolliõiguse valdkondi | 95 | 10 | 20 | 10 | 10 | 5 | 80 | 25 | 85 | 20 | 72 |

Kui vaadata tulemusi ilma otsingusõnata kohtuotsuste lõikude osas (vaata Tabel 7), siis on näha, et need head tulemust pole andnud. Maksimaalselt 32-st 6-s lahendis oli kasulikke lõike. Samas ei olnud piisavalt andmeid iga fraasi kohta, et võtta välja 20 tulemust. Seetõttu nende andmete põhjal ei saa teha laiapõhjalisi järeldusi, kuid tundub, et pigem on

[30] Vt nt kaustast topeltmaks faili „Count_vec_hindamine_topeltmaks.xlsx“ lehte „Count_vec_tulemus_ainult_reso_t“

[31] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „% Ainult resolutsioon“

Word2Vecist ja CountVectorizerist abi selleks, et leida lahendeid, kus fraasi üks või mõlemad sõnad sisalduvad.

**Tabel 7[32]. Ilma otsingusõnata tulemused tükkidena**

| | sisend-käibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topelt-maksustamise vältimine | õiguse kuri-tarvitamine | maksu-pettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine | Kokku |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Kogu** | | | | | | | | | | | |
| CountVectorizer (otsingusõnata) | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | **50** |
| Word2Vec (otsingusõnata) | 5 | 5 | 5 | 0 | 2 | 5 | 0 | 5 | 0 | 5 | **32** |
| CountVectorizer (otsingusõnata) kasulik | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | **1** |
| Word2Vec (otsingusõnata) kasulik | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | **2** |
| **Ilma resolutsioonita** | | | | | | | | | | | |
| CountVectorizer (otsingusõnata) | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | **50** |
| Word2Vec (otsingusõnata) | 5 | 5 | 5 | 0 | 1 | 5 | 1 | 5 | 0 | 5 | **32** |
| CountVectorizer (otsingusõnata) kasulikke | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** |
| Word2Vec (otsingusõnata) kasulik | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 4 | 0 | 0 | **6** |

Selleks, et öelda, et Word2Vec ja CountVectorizer toimivad kohtulahendite lõikude osas hästi, tuleb nende tulemusi (kasulikkuse osas) võrrelda olemasolevate süsteemidega ehk InfoCuria ja EUR-Lex andmetega. Tabelist 8 nähtub, et parimaid tulemusi on andnud CountVectorizer ja Word2Vec, kui on kasutatud ilma resolutsioonita andmeid. InfoCuria ja EUR-Lex toimivad hästi juhul, kui otsingufraas sisaldub kohtulahendis märksõnana („sisendkäibemaksu mahaarvamine“, „topelmaksustamise vältimine“). Samas kui nende abil püütakse leida lahendit, milles otsingufraasi pole („maksueelise saamine“, „maksupettuses osalemine“), siis annavad InfoCuria ja EUR-Lex tulemuseks lahendite lõikusid, millest juristil või kohtunikul pole kasu.

**Tabel 8[33]. CountVectorizer, Word2Vec kasulike lahendite võrdlus InfoCuria ja EUR-Lex tulemustega protsentides** (välja arvatud viimane tulp, mis on väljendatud tükkidena)

| | sisend-käibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topelt-maksustamise vältimine | õiguse kuri-tarvitamine | maksupettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine | Kokku tk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Kogu kordusteta** | | | | | | | | | | | |
| CountVectorizer kogu kasulik | 17,6 | 21,4 | 7,1 | 0 | 23,5 | 10 | 16,7 | 7,7 | 16,7 | 15,8 | 19 |
| Word2Vec kogu kasulik | 73,3 | 0 | 0 | 16,7 | 30,8 | 0 | 0 | 40 | 83,3 | 31,3 | 39 |
| **Ilma resolutsioonita kordusteta** | | | | | | | | | | | |
| CountVectorizer ilma resolutsioonita kasulik | 52,9 | 14,3 | 46,2 | 21,1 | 50 | 88,9 | 80 | 14,3 | 66,7 | 27,8 | 56 |
| Word2Vec ilma resolutsioonita kasulik | 72,2 | 27,8 | 20 | 10 | 57,1 | 0 | 0 | 43,8 | 83,3 | 35,3 | 54 |
| **InfoCuria** | | | | | | | | | | | |
| Ilma lemmatiseerimata kasulik I lõik | 60 | 0 | 65 | 86,7 | 0 | 0 | 0 | 0 | 25 | 15 | 46 |
| Lemmatiseeritud kasulik I lõik | 100 | 0 | 90 | 10 | 0 | 0 | 66,7 | 0 | 25 | 5 | 20 |
| **EUR-Lex** | | | | | | | | | | | |
| Ilma lemmatiseerimata kasulik I lõik | 60 | 5 | 35 | 20 | 0 | 0 | 10 | 5 | 15 | 5 | 31 |
| Lemmatiseeritud kasulik I lõik | 40 | 0 | 80 | 10 | 0 | 0 | 40 | 0 | 15,8 | 0 | 19 |

Kui võrrelda andmeid resolutsiooni andmete osas (vaata Tabel 9), siis nähtub, et paremaid tulemusi annavad endiselt InfoCuria ja EUR-Lex. Põhjus on selles, et andmeid analüüsiks oli CountVectorizer ja Word2Vec jaoks vähe ning andmete puhastamine ei olnud hästi

[32] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „Otsingusõnata tulemused“

[33] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „Võdlus Curia, EUR-Lex, vektorid“ (esimene tabel)

õnnestunud. Seetõttu saab järeldada, et resolutsiooni leidmiseks tasub kasutada endiselt InfoCuria või EUR-Lex süsteeme. Kuid kui on vajalik leida kohtu hinnangu osast kasulikku informatsiooni, tasub proovida luua eraldi CountVectorizeril või Word2Vecil põhinev programm.

**Tabel 9[34]. CountVectorizer, Word2Vec kasulike lahendite võrdlus InfoCuria ja EUR-Lex tulemustega protsentides ainult resolutsiooni andmed** (välja arvatud viimane tulp, mis on väljendatud tükkidena)

| | sisend-käibemaksu mahaarvamine | käibemaksu-kohustuslaseks registreerimine | topelt-maksustamise vältimine | õiguse kuri-tarvitamine | maksupettuses osalemine | maksueelise saamine | litsentsitasude maksustamine | ettevõtlusega tegelemine | dividendide maksustamine | tagatise määramine |
|---|---|---|---|---|---|---|---|---|---|---|
| **Resolutsioon** | | | | | | | | | | |
| CountVectorizer kasulik | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 5 | 0 |
| Word2Vec kasulik | 30 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 65 | 0 |
| **Resolutsioon kordusteta** | | | | | | | | | | |
| CountVectorizer ainult resolutsioon kasulik | 5,3 | 0 | 5 | 0 | 0 | 0 | 5,9 | 0 | 6,7 | 0 |
| Word2Vecainult resolutsioon kasulik | 33,3 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 61,5 | 0 |
| **InfoCuria** | | | | | | | | | | |
| Ilma lemmatiseerimata kasulik resolutsioon | 95 | 100 | 95 | 53,3 | 0 | 0 | 0 | 0 | 75 | 0 |
| Lemmatiseeritud kasulik resolutsio | 100 | 0 | 90 | 20 | 100 | 0 | 0 | 0 | 50 | 0 |
| **EUR-Lex** | | | | | | | | | | |
| Ilma lemmatiseerimata kasulik resolutsioon | 95 | 20 | 85 | 40 | 0 | 5 | 5 | 10 | 85 | 0 |
| Lemmatiseeritud kasulik resolutsio | 90 | 15 | 60 | 5 | 100 | 0 | 20 | 0 | 47,4 | 0 |

Kokkuvõttes saab öelda, et kasulike kohtulahendite lõikude leidmiseks on võimalik kasutada Word2Veci ja CountVectorizerit. Need toimivad paremini, kui neid treenida ilma resolutsioonita andmete peal. Seejuures CountVectorizer võib toimida veidi paremini kui Word2Vec, kuid põhjalikemate järelduste tegemiseks oleks vajalik testida neid rohkemate andmete peal. Siiski on näha, et Word2Veci ja CountVectorizerit toimivad paremini (töös toodud eksperimentide osas), kui senised InfoCuria ja EUR-Lex otsingud seda võimaldaksid (välja arvatud resolutsioonandmete leidmisel).

[34] Sama tabel on suuremalt olemas ka Google Drives faili „Analüüside_koondtabel.xlsx“ lehel „Võdlus Curia, EUR-Lex, vektorid“ (teine tabel)

# 6. Kokkuvõte

Käesoleva töö eesmärk oli selgitata välja, kas CountVectorizer või Word2Vec abil on võimalik luua nutikam märksõna otsing, mis annaks etteantud fraasile sarnased Euroopa Kohtu otsuste lõigud. Töö eesmärk saavutati. See tähendab, et töö tulemusena leiti, et CountVectorizer või Word2Vec abil on võimalik luua nutikam märksõna otsing, kuid seda kohtu hinnangu osast juristile või kohtunikule kasulike lõikude leidmiseks. Kasulike resolutsioonide leidmiseks toimivad jätkuvalt paremini InfoCuria ja EUR-Lex otsingusüsteemid.

Töö eesmärgi saavutamiseks anti ülevaade loomuliku keele töötlusest, masinõppest, CountVectorizer ja Word2Vec ning sarnasuse hindamise meetoditest ja nende rakendamise võimalustest. Need olid vajalikud loomaks tööle teoreetiline taust. Samuti anti ülevaade Euroopa Kohtu otsustest kui töö aluseks olnud andmetest ning nende eeltöötlemiseks tehtud toimingutest. Kohtuotsuste eeltöötluse tulemusena saadi kolm andmetabelit, kus iga kohtu hinnangu ja resolutsiooni lõik oli eraldi real ühes tabelis ning eraldi tabelid saadi ka ainult resolutsiooni ja ainult kohtu hinnanguga lõikude kohta.

CountVectorizerit ja Word2Veci treeniti eelnevalt puhastatud andmete peal ning saadud tulemusi võrreldi ette antud fraasidega. Tulemusi võrreldi omakorda üksteisega ja InfoCuria ja EUR-Lex tulemustega. Analüüsist nähtus, et CountVectorizer ja Word2Vec toimisid paremini andmete osas, kus treenimiseks kasutati ainult kohtu hinnangu osa. Seega on võimalik nende abil luua nutikam märksõna otsing ehk süsteem, mis leiaks fraasile vastavad kohtu hinnangute lõigud.

Tööd on võimalik edasi arendada luues programmi, mida juristid või kohtunikud saaksid ise katsetada. See tähendab, et programm võimaldaks tulevikus juristidel või kohtunikel sisestada otsingufraase ja saada vastuseks kohtuotsuste lõigud. Samuti on võimalik tööd edasi arendada seeläbi, et puhastada andmeid täiendavalt, saavutamaks ka resolutsioonide osas paremad soovitused sarnaste kohtulahendite lõikude osas. Samuti on võimalik testida Word2Vec ja CountVectorizeri headust täiendavate testandmetega.

# 7. Viidatud kirjandus


[1] InfoCuria Kohtupraktika andmebaas. https://curia.europa.eu/juris/recherche.jsf?language=et

[2] EUR-Lex portaal. https://eur-lex.europa.eu/homepage.html?locale=et

[3] Palmer J. ANALYSIS: Legal Tech Is Helping Lawyers, But Where's the Love? *Bloomberg Law*, 2020. https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-legal-tech-is-helping-lawyers-but-wheres-the-love (23.07.2021)

[4] Burk B. New Technology and Its Impact on the Practice of Law. *Expert Institute*, 2021. https://www.expertinstitute.com/resources/insights/new-technology-and-its-impact-on-the-practice-of-law/ (23.07.2021)

[5] Pilehvar M. T., Camacho-Collados J. Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. Toronto: Morgan & Claypool Publishers, 2021.

[6] Nadkarni P. M., Ohno-Machado L., Chapman W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 2011, nr 5, lk 544–551. https://doi.org/10.1136/amiajnl-2011-000464 (23.07.2021)

[7] Di Pietro M. Text Classification with NLP: Tf-Idf vs Word2Vec vs BERT. *towards data science*, 2020. https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794 (23.07.2021)

[8] Kaar M. Majandusuudiste kokkuvõtete seostamine aktsiaturu kõikumistega kasutades loomuliku keele masintöötlust. Tallinn: Tallinna Tehnikaülikool, 2019. https://digikogu.taltech.ee/et/Item/f7abc462-6cdb-4469-bf45-fbeb6713c422 (23.07.2021)

[9] Pindis J. Piltide pealkirjade ja koordinaatide alusel turismiobjektide koondamine ja tüübi kategooriate määramine ning neid toetav paindlik raamistik. Tallinn: Tallinna Tehnikaülikool, 2020. https://digikogu.taltech.ee/et/Item/8463ef71-5ed4-4972-9280-da2c983b95f4 (23.07.2021)

[10] Peedosk M. Eesti keele digitaalsete ressursside ja tehnoloogiate rakendamine teksti lihtsustamise programmis. Tartu: Tartu Ülikool, 2017. https://dspace.ut.ee/handle/10062/65676 (23.07.2021)

[11] Tättar A. Juhendamata masintõlge kasutades keeltevahelisi fraaside vektoresitusi. Tartu: Tartu Ülikool, 2018. http://dspace.ut.ee/handle/10062/66186 (23.07.2021)

[12] Valdson K. Mustripõhine informatsiooni eraldamine Eesti kohtulahenditest. Tartu: Tartu Ülikool, 2016. http://hdl.handle.net/10062/56143 (23.07.2021)

[13] Kullarand S. Kohtulahendite avalikustamise protsessi äri- ja süsteemianalüüs. Tallinn: Tallinna Tehnikaülikool, 2021. https://digikogu.taltech.ee/et/Item/7cd875eb-2af6-4445-b276-1c9723ed0fde (23.07.2021)

[14] Medvedeva M., Vols M., Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law* 2019. https://doi.org/10.1007/s10506-019-09255-y (23.07.2021)

[15] Chandna N. AI in legal industry — A case study on predicting judgements through deep learning. *Medium*, 2020. https://medium.com/analytics-vidhya/ai-in-legal-industry-a-case-study-on-predicting-judgements-through-deep-learning-4ca13f4cf8e1 (23.07.2021)

[16] Liu Z., Chen H. A predictive performance comparison of machine learning models for judicial cases. *2017 IEE Symposium Series on Computational Intelligence (SSCI)*, 2017, lk 1-6. https://ieeexplore.ieee.org/document/8285436 (23.07.2021)

[17] sklearn.feature_extraction.text.CountVectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (23.07.2021)

[18] Sarlis S., Maglogiannis I. On the Reusability of Sentiment Analysis Datasets in Applications with Dissimilar Contexts. IFIP Advances in Information and Communication Technology, 2020, nr 583. https://doi.org/10.1007/978-3-030-49161-1_34 (23.07.20121)

[19] Wilbur W. J., Sirotkin K. The automatic identification of stop words, *Journal of Information Science* , 1992, nr 18(1), lk 45-55. https://www.researchgate.net/publication/247786801_The_automatic_identification_of_stop_words (23.07.2021)

[20] Tarımer İ., Çoban A., Kocaman A. E. Sentiment Analysis on IMDB Movie Comments and Twitter Data. *CoRR*, 2019, lk 1-8. http://arxiv.org/abs/1903.11983 (23.07.2021)

[21] Shahmirzadi O., Lugowski A., Younge K. Text Similarity in Vector Space Models: A Comparative Study. *CoRR*, 2018. http://arxiv.org/abs/1810.00664 (23.07.2021)

[22] Salton G., Wong A., Yang C.-S. A vector space model for automatic indexing. *Communications of the ACM*, 1975, nr 18(11), lk 613-620. https://doi.org/10.1145/361219.361220 (23.07.2021)

[23] Turney P. D., Pantel P. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence*, 2010, lk 141-188. http://arxiv.org/abs/1003.1141(23.07.2021)

[24] Soucy P., Mineau G.M. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, 2005, lk 1130-1135. https://dl.acm.org/doi/10.5555/1642293.1642474 (23.07.2021)

[25] Robertson S. Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*, 2004, nr 60(5), lk 503-520. [Online]. https://doi.org/10.1108/00220410410560582 (23.07.2021)

[26] 6.2. Feature extraction. https://scikit-learn.org/stable/modules/feature_extraction.html (23.07.2021)

[27] Word2Vec. Neural word empeddings for NLP in DL4J. https://deeplearning4j.konduit.ai/language-processing/word2vec (23.07.2021)

[28] Word2Vec Model. https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html#sphx-glr-download-auto-examples-tutorials-run-word2vec-py (23.07.2021)

[29] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space, *CoRR*, 2013. https://arxiv.org/pdf/1301.3781.pdf (23.07.2021)

[30] Meng Y., Huang J., Wang G., Zhang C., Han J. Unsupervised Word Embedding Learning by Invorporating Local and Global Contexts, *Front Big Data*, 2020 nr 3(9). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931948/ (23.07.2021)

[31] Goldberg Y., Levy O. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method," *CoRR*, 2014. http://arxiv.org/abs/1402.3722 (23.07.2021)

[32] Word2Vec. https://www.tensorflow.org/tutorials/text/word2vec (23.07.2021)

[33] Word2vec embeddings. https://radimrehurek.com/gensim/models/word2vec.html (23.07.2021)

[34] Orasmaa S., Tkachenko A. estnltk / word2vec-models. https://github.com/estnltk/word2vec-models (23.07.2021)

[35] Wang J., Dong Y. Measurement of Text Similarity: A Survey. *Information (Switzerland), 2020, nr* 11(9), lk 421. https://www.researchgate.net/publication/344010599_Measurement_of_Text_Similarity_A_Survey (23.07.2021)

[36] Ristanti P. Y., Wibawa A., Pujianto U. Cosine Similarity for Title and Abstract of Economic Journal Classification. *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019. https://www.researchgate.net/publication/339174965_Cosine_Similarity_for_Title_and_Abstract_of_Economic_Journal_Classification (23.07.2021)

[37] Affinities and Kernels 6.8. Pairwise metrics. https://scikit-learn.org/stable/modules/metrics.html#cosine-similarity (23.07.2021)

[38] Han J., Kamber M., Pei P. *Data Mining: Concepts and Techniques*.: Elsevier Inc. , 2012. https://doi.org/10.1016/C2009-0-61819-5 (23.07.2021)

[39] Emmery C. Euclidean vs. Cosine Distance. https://cmry.github.io/notes/euclidean-v-cosine (23.07.2021)

[40] Selenium Python bindings for Selenium. https://selenium-python.readthedocs.io/installation.html#introduction (23.07.2021)

[41] Hanson V., Tavast A. Arvutikasutaja sõnastik. http://www.keeleveeb.ee/dict/speciality/aks/dict.cgi?word=html&lang=en (23.07.2021)

[42] Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (23.07.2021)

[43] Pandas documentation. https://pandas.pydata.org/docs/ (23.07.2021)

[44] Comparison of different Word Embeddings on Text Similarity — A use case in NLP. *Medium*, 2019. https://intellica-ai.medium.com/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c (23.07.2021)

[45] Mhatre M., Phondekar D., Kadam P., Chawathe A., Ghag K. Dimensionality reduction for sentiment analysis using pre-processing techniques. *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 2017. https://ieeexplore.ieee.org/document/8282676 (23.07.2021)

[46] Uiboaed K. Eesti keele stoppsõnad / Estonian stop words. DataDOI, 2018. https://datadoi.ee/handle/33/78 (23.07.2021)

[47] Python string — Common string operations. https://docs.python.org/3/library/string.html#string-constants (23.07.2021)

[48] Dubey P. Understand Text Summarization and create your own summarizer in python. towads data science, 2018. https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70 (23.07.2021)

# Lisad

## I. Kohtuotsuse näidis[35]

Kohtu koda: EUROOPA KOHTU OTSUS (teine koda)

Kuupäev: 14. märts 2013(*)

Märksõnad: Käibemaks – Direktiiv 2006/112/EÜ – Artiklid 213, 214 ja 273 – Käibemaksukohustuslaste registreerimine – Käibemaksukohustuslasena registreerimise numbri andmisest keeldumine põhjendusel, et maksukohustuslasel puuduvad deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid – Õiguspärasus – Maksupettuste vastane võitlus – Proportsionaalsuse põhimõte

Kohtuasja number: Kohtuasjas C-527/11,

mille ese on ELTL artikli 267 alusel Augstākās tiesas Senāts (Läti) 12. oktoobri 2011. aasta otsusega esitatud eelotsusetaotlus, mis saabus Euroopa Kohtusse 18. oktoobril 2011, menetluses

Pooled:

**Valsts ieņēmumu dienests**

versus

**Ablessio SIA,**

Kohtu koda: EUROOPA KOHUS (teine koda),

koosseisus: A. Rosas teise koja presidendi ülesannetes, kohtunikud U. Lõhmus (ettekandja), A. Ó Caoimh, A. Arabadjiev ja C. G. Fernlund,

kohtujurist: E. Sharpston,

kohtusekretär: A. Calot Escobar,

arvestades kirjalikku menetlust,

---

35 https://curia.europa.eu/juris/document/document.jsf?text=&docid=135026&pageIndex=0&doclang=ET&mode=lst&dir=&occ=first&part=1&cid=1101378. Kollasega on lisatud autori poolt juurde kohtuotsuse osasid indikeerivad märksõnad.

arvestades seisukohti, mille esitasid:

Menetlusosalised:

– Valsts ieņēmumu dienests, esindaja: T. Kravalis,

– Läti valitsus, esindajad: I. Kalniņš ja I. Ņesterova,

– Eesti valitsus, esindaja: M. Linntam,

– Euroopa Komisjon, esindajad: C. Soulay ja E. Kalniņš,

arvestades pärast kohtujuristi ärakuulamist tehtud otsust lahendada kohtuasi ilma kohtujuristi ettepanekuta,

on teinud järgmise

**otsuse**

1 Eelotsusetaotlus puudutab nõukogu 28. novembri 2006. aasta direktiivi 2006/112/EÜ, mis käsitleb ühist käibemaksusüsteemi (ELT L 347, lk 1), artiklite 213, 214 ja 273 tõlgendamist.

2 Taotlus on esitatud Valsts ieņēmumu dienestsi (Läti maksuamet, edaspidi „maksuamet") ja Ablessio SIA (edaspidi „Ablessio") vahelises kohtuvaidluses Ablessio käibemaksukohustuslasena registreerimisest keeldumise üle.

Asjakohaste õigusaktide kirjeldus:

**Õiguslik raamistik**

Direktiiv 2006/112

3 Mõisted „maksukohustuslane" ja „majandustegevus" on direktiivi 2006/112 artikli 9 lõikes 1 määratletud järgmiselt:

„„Maksukohustuslane" on iga isik, kes mis tahes paigas teostab iseseisvalt mis tahes majandustegevust, olenemata nimetatud tegevuse eesmärgist või tulemustest.

„Majandustegevus” on tootja, ettevõtja ja teenuseid osutava isiku mis tahes tegevus, sealhulgas kaevandamis- ja põllumajandusalane tegevus ning kutsealane tegevus. Majandustegevusena käsitatakse eelkõige materiaalse või immateriaalse vara kasutamist kestva tulu saamise eesmärgil.”

4 Sama direktiivi artikli 213 lõige 1 sätestab:

„Iga maksukohustuslane teatab, millal tema tegevus maksukohustuslasena algab, muutub või lõpeb.

Enda kehtestatavatel tingimustel lubavad liikmesriigid selle teate esitada elektrooniliselt. Nad võivad elektroonilist esitamist ka nõuda.”

5 Direktiivi artikkel 214 on sõnastatud järgmiselt:

„1. Liikmesriigid võtavad vajalikud meetmed, et järgmised isikud registreerida, andes neile individuaalse numbri:

a) iga maksukohustuslane, välja arvatud artikli 9 lõikes 2 nimetatud maksukohustuslased, kes teeb oma riigi territooriumil mahaarvamisõigust andvaid kaubatarneid või osutab teenuseid, välja arvatud sellised kaubatarned ja teenuseosutamised, mille puhul peab käibemaksu artiklite 194–197 ja artikli 199 kohaselt tasuma ainult kaupade soetaja või teenuste saaja või isik, kelle tarbeks kaubad või teenused on mõeldud;

b) iga maksukohustuslane või mittemaksukohustuslasest juriidiline isik, kes ühendusesiseselt soetab artikli 2 lõike 1 punkti b kohaselt käibemaksuga maksustatavaid kaupu või kes kasutab artikli 3 lõikes 3 ettenähtud valikuõigust oma kaupade ühendusesisene soetamine käibemaksuga maksustada;

[...]

2. Liikmesriigid ei pea registreerima teatavaid maksukohustuslasi, kes juhuti teevad [...] tehinguid.”

6 Direktiivi 2006/112 artikli 273 esimene lõik näeb ette:

„Liikmesriigid võivad maksukohustuslaste riigisiseste ja liikmesriikidevaheliste tehingute võrdse kohtlemise põhimõtet järgides kehtestada käibemaksu nõuetekohaseks kogumiseks ning maksudest kõrvalehoidumise ärahoidmiseks vajalikuks peetavaid muid kohustusi, tingimusel et sellised kohustused ei too liikmesriikidevahelises kaubanduses kaasa formaalsusi piiriületamisel.”

Läti õigus

7 Käibemaksuseaduse (Likums Par pievienotās vērtības nodokli, Latvijas Vēstnesis, 1995, nr 49, edaspidi „käibemaksuseadus") artikli 3 lõike 11 teine lõik sätestab põhikohtuasja asjaolude suhtes kohaldatavas redaktsioonis:

„Maksuametil on õigus keelduda isiku käibemaksukohustuslasena registreerimisest, kui see isik:

1) ei ole kättesaadav tema registrijärgsel aadressil või taotlusel märgitud elukohas või

2) ei anna maksuameti taotletud teavet või esitab valeandmeid oma majanduslike, tehniliste ja rahaliste võimaluste kohta deklareeritud majandustegevust ellu viia."

8 Nimetatud seaduse artikli 3 lõike 5 kohaselt:

„Kui füüsilise või juriidilise isiku ja nende isikute grupi või nende lepingulise või kokkuleppelise esindaja poolt eelneva kaheteistkümne kuu jooksul teostatud maksustatavate kaubatarnete ja teenuste osutamise koguväärtus on alla 10 000 Läti lati või ei ületa seda summat, siis on isikutel, nende gruppidel või esindajatel õigus mitte olla maksuameti peetavas registris käibemaksukohustuslasena registreeritud. Seda sätet kohaldatakse ka riigieelarvelistele asutustele. Isik, kes kasutab käesolevas lõikes ettenähtud õigust, on kohustatud end selles registris käibemaksukohustuslasena registreerima kolmekümne päeva jooksul alates hetkest, kui nimetatud väärtus moodustab või ületab kõnealuse summa."

Asjaolude kirjeldus, poolte ja menetlusosaliste seisukohad:

**Põhikohtuasi ja eelotsuse küsimused**

9 Läti piiratud vastutusega äriühing Ablessio esitas maksuametile avalduse enda käibemaksukohustuslasena registreerimiseks. Maksuamet jättis registreerimisavalduse rahuldamata 15. novembri 2007. aasta otsusega, mis pärast vaide esitamist jäeti 27. novembri 2007. aasta otsusega muutmata, kuna ta leidis, et äriühingul puuduvad deklareeritud majandustegevuseks, st ehitusteenuste pakkumiseks vajalikud majanduslikud, tehnilised ja rahalised vahendid.

10 Eelotsusetaotlusest nähtub, et maksuamet põhjendas registreerimata jätmise otsust sellega, et esiteks on tuvastatud, et Ablessiol puudub kinnisvara ega ole sõlmitud ühtki lepingut sellise vara üürimiseks. Pealegi on äriruumide üürileping sõlmitud vaid 4 m2 suuruse mitteeluruumi kohta. Äriühing ei ole kantud ehitusregistrisse ega ole asutamisest saadik läbi viinud tegelikku majandustegevust, äriühingu ainus töötaja on juhatuse esimees, kellele ilmselt palka ei maksta.

11 Ablessio esitas käibemaksukohustuslasena registreerimata jätmise otsuste peale tühistamiskaebuse Administratīvā rajona tiesale (esimese astme halduskohus), kes rahuldas kaebuse 20. oktoobri 2009. aasta otsusega ja kohustas maksuametit äriühingu registreerima. Nimetatud kohus leidis, et Ablessio oli esitanud maksuametile andmed oma deklareeritud majandustegevuseks vajalike majanduslike, tehniliste ja rahaliste vahendite kohta ning et andmete õigsust ei ole vaidlustatud. Seega olid kohtu hinnangul täitmata seaduses sätestatud tingimused, mis lubavad maksuametil jätta ettevõtja registreerimata.

12 Administratīvā apgabaltiesa (piirkondlik halduskohus), kes lahendas maksuameti apellatsioonkaebust, jättis 13. detsembri 2010. aasta otsusega esimese astme kohtu otsuse muutmata, leides samuti, et käibemaksuseadus ei luba maksuametil anda hinnangut, kas isik, kes soovib end käibemaksukohustuslasena registreerida, suudab majandustegevust läbi viia. Selles osas ei oma tähtsust, et see isik on juba esitanud avalduse ja registreerinud mitu äriühingut, mille osalus on kohe pärast registreerimist võõrandatud teistele isikutele, kelle sissetulekute tase on nii madal, et sellest ei piisa osakapitali sissemaksete tegemiseks, kuna nimetatud seadus ei näe ette, et sellised asjaolud on isiku vastavas registris registreerimata jätmise aluseks. Maksukohustuslase kõigi võimalike ebaseaduslike tegude ärahoidmiseks käibemaksu tasumise raames on ette nähtud maksuameti kohustus maksukohustuslast seadusega sätestatud korras kontrollida ja kui ta tuvastab liikmesriigi õigusnormide rikkumise, on ta kohustatud arvutama täiendava maksu ja sanktsioonid.

13 Maksuamet esitas eelotsusetaotluse esitanud kohtule kassatsioonkaebuse Administratīvā apgabaltiesa kohtuotsuse peale, väites, et viimasena nimetatud kohus on käibemaksuseaduse artikli 3 lõike 11 teise lõigu tõlgendamisel õigusnorme rikkunud. Maksuamet leiab, et see säte paneb talle kohustuse kontrollida, kas isik suudab deklareeritud majandustegevust läbi viia.

14 Eelotsusetaotluse esitanud kohus, viidates 21. oktoobri 2010. aasta otsusele kohtuasjas C-385/09: Nidera Handelscompagnie (EKL 2010, lk I-10385), avaldab kahtlusi, kuidas tõlgendada eeskätt direktiivi 2006/112 artikleid 213, 214 ja 273.

15 Neil asjaoludel otsustas Augstākās tiesas Senāts menetluse peatada ja esitada Euroopa Kohtule järgmised eelotsuse küsimused:

„1. Kas […] direktiivi 2006/112 […] tuleb tõlgendada nii, et see ei luba keelduda maksukohustuslase registreerimisest põhjendusel, et maksukohustuslase osanik on eelnevalt registreerinud mitmel korral individuaalse numbri all muud äriühingud, mis ei ole kunagi tegelikult majandustegevust ellu viinud, ja mille osad võõrandas nende omanik vahetult pärast individuaalse numbri all registreerimist teistele isikutele?

2. Kas direktiivi [2006/112] artiklit 214 koostoimes artikliga 273 tuleb tõlgendada nii, et see lubab maksuametil enne maksukohustuslase registreerimist kontrollida tema võimet

maksustataval tegevusalal tegutseda, juhul kui selle kontrolliga tahetakse tagada käibemaksu nõuetekohast kogumist ning maksupettuste ärahoidmist?”

Euroopa Kohtu analüüs:

**Eelotsuse küsimuste analüüs**

16 Nende küsimustega, mida tuleb analüüsida koos, soovib eelotsusetaotluse esitanud kohus sisuliselt selgitust, kas direktiivi 2006/112 artikleid 213, 214 ja 273 tuleb tõlgendada nii, et need ei luba liikmesriigi maksuametil käibemaksu nõuetekohase kogumise tagamiseks ja maksupettuste ärahoidmiseks keelduda äriühingule käibemaksukohustuslasena registreerimise numbrit andmast vaid sel põhjusel, et ameti arvates puuduvad tal deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid ning et äriühingu osanik on selle numbri juba mitu korda saanud äriühingute jaoks, mis ei ole kunagi teostanud reaalset majandustegevust ja mille osalus on võõrandatud varsti pärast numbri saamist.

17 Olgu meenutatud, et direktiivi 2006/112 artikli 213 lõike 1 esimese lõigu kohaselt on iga maksukohustuslane kohustatud teatama, millal tema tegevus maksukohustuslasena algab, muutub või lõpeb. Sama direktiivi artikli 214 lõige 1 kohustab liikmesriike võtma vajalikke meetmeid, et maksukohustuslased registreerida, andes neile individuaalse numbri.

18 Direktiivi 2006/112 artiklis 214 ette nähtud maksukohustuslaste registreerimise peamine eesmärk on tagada käibemaksusüsteemi nõuetekohane toimimine (vt selle kohta 22. detsembri 2010. aasta otsus kohtuasjas C-438/09: Dankowski, EKL 2010, lk I-14009, punkt 33).

19 Selles osas on Euroopa Kohus juba selgitanud, et käibemaksukohustuslasena registreerimise numbri andmine tõendab maksukohustuslase staatust käibemaksuga maksustamisel ja lihtsustab kontrolli maksukohustuslaste üle eesmärgiga tagada maksu nõuetekohane kogumine. Euroopa Liidu sisese kaubanduse maksustamise üleminekukorra raames on käibemaksukohustuslaste registreerimise eesmärk lihtsustada selle liikmesriigi väljaselgitamist, kus leiab aset tarnitud kauba lõpptarbimine (vt selle kohta 6. septembri 2012. aasta otsus kohtuasjas C-273/11: Mecsek-Gabona, punktid 57 ja 60, ning 27. septembri 2012. aasta otsus kohtuasjas C-587/10: VSTR, punkt 51).

20 Peale selle on käibemaksukohustuslasena registreerimise numbril oluline tähtsus tehtud tehingute tõendamisel. Nimelt on direktiivis 2006/112 rida sätteid, mis puudutavad eeskätt arve esitamist, deklaratsioone ja koondaruandeid ning muu hulgas nõuavad, et neile dokumentidele peab olema märgitud maksukohustuslase, kauba soetaja või teenuste saaja käibemaksukohustuslasena registreerimise number.

21 Eelotsusetaotluse esitanud kohtu küsimustele vastamisel tuleb arvestada neid kaalutlusi.

22 Tuleb nentida, et kuigi direktiivi 2006/112 artiklis 214 on loetletud isikute kategooriad, kes peavad olema individuaalse numbriga registreeritud, ei näe see säte ette, milliseid tingimusi võib käibemaksukohustuslasena registreerimise numbri andmisele seada. Selle artikli ja sama direktiivi artikli 213 sõnastusest nähtub, et liikmesriikidel on tegelikult käibemaksukohustuslaste registreerimise tagamise meetmete võtmisel teatav kaalutlusruum.

23 See kaalutlusruum ei saa mingil juhul olla piiramatu. Nimelt, kui liikmesriigil on lubatud maksukohustuslasele individuaalse numbri andmisest keelduda, ei saa seda võimalust kasutada ilma õiguspärase põhjuseta.

24 Lisaks ilmneb direktiivi 2006/112 artikli 9 lõikes 1 määratletud mõistest „maksukohustuslane", et selle mõiste alla kuulub iga isik, kes mis tahes paigas teostab iseseisvalt mis tahes majandustegevust, olenemata nimetatud tegevuse eesmärgist või tulemustest.

25 Kooskõlas Euroopa Kohtu praktikaga tuleb seda mõistet tõlgendada laialt. Igaüht, kellel on objektiivselt tõendatud kavatsus alustada iseseisvalt majandustegevust ja kes teeb selleks vajalikke esmaseid investeeringuid, tuleb pidada maksukohustuslaseks (vt selle kohta 8. juuni 2000. aasta otsus kohtuasjas C-400/98: Breitsohl, EKL 2000, lk I-4321, punkt 34, ja 1. märtsi 2012. aasta otsus kohtuasjas C-280/10: Polski Trawertyn, punkt 30).

26 Nii kohtupraktikast kui ka direktiivi 2006/112 artikli 213 lõike 1 sõnastusest tuleneb, et maksukohustuslasteks, kes võivad taotleda käibemaksukohustuslasena registreerimise numbrit, peetakse mitte ainult isikuid, kes teostavad juba majandustegevust, vaid ka neid, kes kavatsevad seda alustada ja kes teevad selleks vajalikke esmaseid investeeringuid. Neil isikutel ei tarvitse olla võimalik oma majandustegevuse algstaadiumis tõendada, et neil on juba selleks tegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid.

27 Seega ei luba direktiiv 2006/112 – eelkõige selle artiklid 213 ja 214 – liikmesriigi maksuametil keelduda taotlejale käibemaksukohustuslasena registreerimise numbrit andmast vaid sel põhjusel, et ta ei suuda tõendada, et tal on käibemaksukohustuslasena registreerimise avalduse esitamise ajal olemas deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid.

28 Sellegipoolest on Euroopa Kohtu praktika kohaselt liikmesriikidel õigustatud huvi võtta oma finantshuvide kaitseks vajalikke meetmeid ning maksupettuste ja maksustamise vältimise ning muude võimalike kuritarvituste vastane võitlus on direktiiviga 2006/112 tunnustatud ja toetatud eesmärk (vt eelkõige 21. veebruari 2006. aasta otsus kohtuasjas C-255/02: Halifax jt, EKL 2006, lk I-1609, punkt 71; 7. detsembri 2010. aasta otsus kohtuasjas C-285/09: R., EKL 2010, lk I-12605, punkt 36, samuti 18. oktoobri 2012. aasta otsus kohtuasjas C-525/11: Mednis, punkt 31).

29 Ühtlasi on liikmesriikidel kohustus tagada, et kanded maksukohustuslaste registris oleksid õiged, et kindlustada käibemaksusüsteemi nõuetekohane toimimine. Seega lasub pädeval siseriiklikul ametiasutusel kohustus kontrollida, kas taotlejal on maksukohustuslase

staatus, enne kui ta annab talle käibemaksukohustuslasena registreerimise numbri (vt eespool viidatud kohtuotsus Mecsek-Gabona, punkt 63).

30 Niisiis võivad liikmesriigid kooskõlas direktiivi 2006/112 artikli 273 esimese lõiguga õiguspäraselt võtta meetmeid, millega saab takistada registreerimisnumbri kuritarvitamist eriti nende ettevõtjate poolt, mille tegevus – ja järelikult ka maksukohustuslase staatus – on puhtalt fiktiivne. Sellegipoolest ei tohi need meetmed minna kaugemale sellest, mis on vajalik maksu nõuetekohase kogumise kindlustamiseks ja pettuste ärahoidmiseks, ning nendega ei tohi seada süstemaatiliselt kahtluse alla käibemaksu mahaarvamise õigust ja sellest tulenevalt ka maksustamise neutraalsust (vt selle kohta 27. septembri 2007. aasta otsus kohtuasjas, C-146/05: Collée, EKL 2007, lk I-7861, punkt 26; eespool viidatud kohtuotsus Nidera Handelscompagnie, punkt 49; eespool viidatud kohtuotsus Dankowski, punkt 37, ja eespool viidatud kohtuotsus VSTR, punkt 44).

31 Selles osas tuleb nentida, et sellised kontrollimeetmed, nagu on kehtestatud käibemaksuseadusega, ei saa piirata maksukohustuslaste õigust arvata maha kavandatavate ja mahaarvamisõigust andvate tehingute tarbeks tehtud investeeringutelt tasumisele kuuluvat või tasutud käibemaksu.

32 Nimelt tuleb meelde tuletada, et Euroopa Kohtu väljakujunenud praktika kohaselt on direktiivi 2006/112 artiklis 214 ette nähtud registreerimise ja artiklis 213 sätestatud kohustuste puhul tegemist kontrolliotstarbeliste vorminõuetega, mis ei saa seada kahtluse alla eelkõige õigust mahaarvamisele või käibemaksuvabastusele seoses ühendusesisese tarnega, kui nende õiguste tekkimise sisulised tingimused on täidetud (vt selle kohta eespool viidatud kohtuotsus Nidera Handelscompagnie, punkt 50; 19. juuli 2012. aasta otsus kohtuasjas C-263/11: Rēdlihs, punkt 48, ja eespool viidatud kohtuotsus Mecsek-Gabona, punkt 60).

33 Sellest kohtupraktikast järeldub, et käibemaksukohustuslasena registreerimine on selline vorminõue, mis ei võimalda maksukohustuslast takistada teostamast talle kuuluvat mahaarvamisõigust põhjusel, et ta ei olnud enne soetatud kaupade oma maksustatavas majandustegevuses kasutamist käibemaksukohustuslasena registreeritud (vt selle kohta eespool viidatud kohtuotsused Nidera Handelscompagnie, punkt 51, ja Dankowski, punktid 33, 34 ja 36). Siit tulenevalt ei saa käibemaksukohustuslasena registreerimise numbri andmisest keeldumine põhimõtteliselt avaldada mingit mõju maksukohustuslase õigusele arvata maha sisendkäibemaks, kui selle õiguse tekkimise sisulised tingimused on täidetud.

34 Selleks et individuaalse numbri andmisest keeldumist saaks pidada proportsionaalseks maksupettuste ärahoidmise eesmärgi suhtes, peab see põhinema arvestataval teabel, mis lubab objektiivselt järeldada, et kui sellele maksukohustuslasele antakse käibemaksukohustuslasena registreerimise number, hakkab ta seda tõenäoliselt kuritarvitama. Otsus peab põhinema hinnangul, mis antakse kogumis juhtumi kõigile asjaoludele ja tõenditele, mis on ettevõtja esitatud andmete kontrollimise käigus kogutud.

35 Eelotsusetaotluse esitanud kohtul, kes ainsana on pädev tõlgendama siseriiklikku õigust ja hindama põhikohtuasja asjaolusid, eeskätt seda, kuidas maksuamet on seda õigust

rakendanud (vt eespool viidatud kohtuotsus Mednis, punkt 33 ja seal viidatud kohtupraktika), tuleb hinnata, kas riiklikud meetmed on kooskõlas liidu õigusega ja eeskätt proportsionaalsuse põhimõttega. Euroopa Kohtu pädevuses on üksnes anda liikmesriigi kohtule liidu õiguse tõlgendamiseks juhiseid, mis võimaldavad tal hinnata kooskõla küsimust (vt selle kohta 30. novembri 1995. aasta otsus kohtuasjas C-55/94: Gebhard, EKL 1995, lk I-4165, punkt 19, ja 29. juuli 2010. aasta otsus kohtuasjas C-188/09: Profaktor Kulesza, Frankowski, Jóźwiak, Orłowski, EKL 2010, lk I-7639, punkt 30).

36 Põhikohtuasja asjaolude kohta olgu märgitud, et ainuüksi sellest, et maksukohustuslasel puuduvad deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid, ei piisa tõendamaks, et ta tõenäoliselt paneb toime maksupettuse. Sellegipoolest ei saa välistada, et sedalaadi asjaolud – mida kinnitavad muud objektiivsed seigad, mis tekitavad kahtluse, et maksukohustuslane kavandab pettust – võivad olla teabeks, mida tuleb pettuseohu üldisel hindamisel arvesse võtta.

37 Samuti ei näe direktiiv 2006/112 ette ühtki piirangut selle kohta, mitu käibemaksukohustuslasena registreerimise avaldust võib sama isik esitada, tegutsedes erinevate juriidiliste isikute nimel. Direktiiv ei luba ka järeldada, et juriidilise isiku üle kontrolliõiguse loovutamine pärast tema käibemaksukohustuslasena registreerimist on õigusvastane tegevus. Sellegipoolest võib ka neid asjaolusid pettuseohu üldise hindamise raames arvesse võtta.

38 Eelotsusetaotluse esitanud kohtu ülesanne on käesoleva juhtumi kõiki asjaolusid arvestades kontrollida, kas riigi maksuamet tegi õiguslikult piisavalt kindlaks arvestatava teabe olemasolu, mis lubab järeldada, et käibemaksukohustuslasena registreerimise avaldus, mille esitas Ablessio, võib kaasa tuua individuaalse numbri kuritarvitamise või muu käibemaksupettuse ohu.

39 Kõiki eelnevaid kaalutlusi arvestades tuleb esitatud küsimustele vastata, et direktiivi 2006/112 artikleid 213, 214 ja 273 tuleb tõlgendada nii, et need ei luba liikmesriigi maksuametil keelduda äriühingule käibemaksukohustuslasena registreerimise numbrit andmast vaid sel põhjusel, et ameti arvates puuduvad tal deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid ning et äriühingu osanik on selle numbri juba mitu korda saanud äriühingute jaoks, mis ei ole kunagi teostanud reaalset majandustegevust ja mille osalus on võõrandatud varsti pärast numbri saamist, ilma et maksuamet oleks objektiivseid asjaolusid arvestades kindlaks teinud arvestatava teabe olemasolu, mis võimaldab kahtlustada, et antud käibemaksukohustuslasena registreerimise numbrit hakatakse kuritarvitama. Eelotsusetaotluse esitanud kohtu ülesanne on hinnata, kas maksuamet esitas arvestatavat teavet pettuseohu kohta põhikohtuasjas.

Kohtukulude jaotus:

**Kohtukulud**

40 Kuna põhikohtuasja poolte jaoks on käesolev menetlus eelotsusetaotluse esitanud kohtus poolelioleva asja üks staadium, otsustab kohtukulude jaotuse siseriiklik kohus. Euroopa Kohtule seisukohtade esitamisega seotud kulusid, välja arvatud poolte kohtukulud, ei hüvitata.

Resolutsioon:

Esitatud põhjendustest lähtudes Euroopa Kohus (teine koda) otsustab:

**Nõukogu 28. novembri 2006. aasta direktiivi 2006/112/EÜ, mis käsitleb ühist käibemaksusüsteemi, artikleid 213, 214 ja 273 tuleb tõlgendada nii, et need ei luba liikmesriigi maksuametil keelduda äriühingule käibemaksukohustuslasena registreerimise numbrit andmast vaid sel põhjusel, et ameti arvates puuduvad tal deklareeritud majandustegevuseks vajalikud majanduslikud, tehnilised ja rahalised vahendid ning et äriühingu osanik on selle numbri juba mitu korda saanud äriühingute jaoks, mis ei ole kunagi teostanud reaalset majandustegevust ja mille osalus on võõrandatud varsti pärast numbri saamist, ilma et maksuamet oleks objektiivseid asjaolusid arvestades kindlaks teinud arvestatava teabe olemasolu, mis võimaldab kahtlustada, et antud käibemaksukohustuslasena registreerimise numbrit hakatakse kuritarvitama. Eelotsusetaotluse esitanud kohtu ülesanne on hinnata, kas maksuamet esitas arvestatavat teavet pettuseohu kohta põhikohtuasjas.**

Allkirjad

* Kohtumenetluse keel: läti.

## II. Litsents

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Sirle Orav-Hinno,

*(autori nimi)*

**1.** annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose **Euroopa Liidu Kohtu otsustest fraasidele sarnaste lõikude otsingu analüüs CountVectorizer ja Word2Vec baasil,**

*(lõputöö pealkiri)*

mille juhendaja on Dage Särg ning kaasjuhendaja Risto Hinno,

*(juhendaja nimi)*

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

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

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

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

Sirle Orav-Hinno

***01.08.2021***