TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Marek Kesküll 192262IABM

Algoritmiline ESG panuse hindamine ettevõtete aruannetes: vektorruumil põhinev semantiline sarnasusskoor

Magistritöö

Juhendaja: Innar Liiv
Doktorikraad

Tallinn 2021

Autorideklaratsioon

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

Autor: Marek Kesküll

10.05.2021

Annotatsioon

Viimaste aastate suured üleilmsed muutused kestliku arengu valdkondades on lisanud suurendatud tähelepanu ESG (*Environmental, Social, Governance*) faktorite arvestamisele investeerimises ja firmade otsuste vastu võtmisel. Ülemaailmne pandeemia on kasvatanud ka teadlikkust ja tuge sotsiaalsetele probleemidele, mis langevad ESG faktorite alla, nagu näiteks töötaja tervis ja turvalisus. ESG hinnangute andmisel kasutatakse tänapäeval manuaalseid juhendeid ja intervjuusi, mille alusel kalkuleeritakse ESG skoor. Antud tööga selgitasime välja, et vektorruumil põhinevat mudelit saab rakendada piisavalt hästi pankade aruannetest ESG andmete välja lugemiseks ja usutava tulemuse andmiseks.

Uuringu läbiviimiseks koguti Eesti pankade ingliskeelsed 2016-2020. aasta aruanded kraapides veebi ning seejärel puhastati andmed, töödeldi andmeid ja ehitati valmis mudel, mida rakendati andmete peal.

Tööga loodi vektorruumil põhinev mudel, mis suudab mõõta ESG valdkondade panust otse Eesti pankade ettevõtete aruannetest ning järjestada ettevõtte aastate kvartaliaruanded vastavalt sarnasusele. Vektorruumil põhinevat koosinussarnasuse mudeli algoritmide abil on võimalik leida kvartaliaruannete vaheline sarnasus, mille abil on võimalik interpreteerida ESG panust võrreldes aruandeid ja ESG termineid ruumis.

Mudeli tulemused kuvati nii tabelite kui ka graafilisel kujul. Mudelist saadud tulemused näitavad selgelt, et Eesti pangad kajastavad oma kvartaliaruannetes ESG seotud faktoreid ebapiisavalt või üldsegi mitte, kuid ESG-ga arvestamine ja raporteerimine on ajas kasvanud.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 65 leheküljel, 5 peatükki, 48 joonist, 2 tabelit.

Abstract

Algorithmic evaluation of the ESG contribution in company reports: a semantic similarity score based on vector space

The major global changes in the field of sustainable development in recent years have added increased attention to the integration of ESG (*Environmental, Social, Governance*) factors into investment and company decision-making. The global pandemic has also increased awareness and support for social issues that fall under ESG factors, such as worker health and safety. Today, manual guides and interviews are used to calculate ESG scores, on the basis of which the ESG score is calculated. With this thesis, we found out that the vector space-based model can be applied well enough to read ESG data from banks' reports and give a plausible result.

To conduct the analysis, the English-language 2016-2020 annual reports of Estonian banks reports were scraped from the web, data was cleaned, data was processed, and a model was built to be applied to the data.

In this thesis, a model was created based on vector space, which can measure the contribution of ESG areas directly from the reports of Estonian bank companies and rank the company's quarterly reports according to similarity. Algorithms for the cosine similarity model based on vector space can be used to find similarities between quarterly reports, which can be used to interpret the ESG contribution by comparing reports and ESG terms in space.

The results of the model were displayed in both tabular and graphical form. The results obtained from the model clearly show that Estonian banks do not adequately or not at all reflect ESG-related factors in their quarterly reports, but the reflection and reporting of ESG has increased over time.

The thesis is in Estonian and contains 65 pages of text, 5 chapters, 48 figures, 2 tables.

Lühendite ja mõistete sõnastik

ESG	Environmental, Social, Governance
EL	Euroopa Liit
CSR	Corporate Social Responsibility
TF-IDF	Term frequency – inverse document frequency
LDA	Latent Dirichlet allocation
LSI	Latent Semantic Indexing
GloVe	Global Vectors for word representation
ISIN	International Securities Identification Number
HTML	HyperText Markup Language
ESMA	European Securities and Markets Authority
UNGC	United Nations Global Compact
ÜRO	Ühinenud Rahvaste Organisatsioon
CO2	Süsihappegaas
DNA	Desoksüribonukleiinhape
Java	Objektorienteeritud programmeerimiskeel
PHP	Skriptimiskeel
RAM	Muutmälu
MIME	Internetiprotokoll
API	Rakendusliides
PDF	Porditav dokumendivorming
.TXT	Tekstifail
MVP	Vähim elujõuline toode
NLTK	Natural Language Toolkit
Gensim	Avatud lähtekoodiga teek keele töötlemiseks

Sisukord

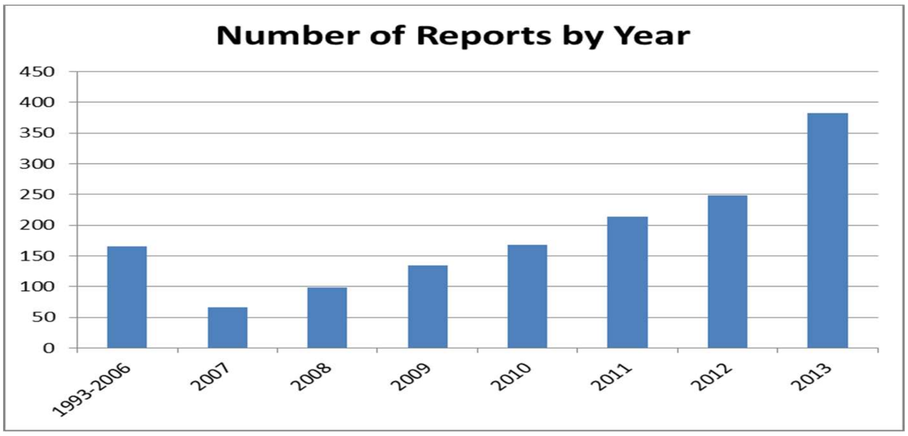
Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Kasvav tähelepanu sotsiaalsele investeerimisele investorite hulgas tekitab küsimuse, kuidas ettevõtted keskkondlikke, sotsiaalsete ja juhtimisega(*ing k. environmental, social and governance,* edaspidi ESG) seotud valdkondadega tegelevad. Estwatchi uuringu järgi enne 2019. aasta suve polnud Eesti pangad, suuremad teadusasutused, ühiskond ega teised osapooled pööranud avalikult tähelepanu kestlikusse arengusse. See tõstatas küsimuse, kas asutused üldse arvestavad ESG riske, mille ignoreerimine ohustab ettevõtte käekäiku, protsesse ja keskkondlikku heaolu. [1]

Suurt rolli mängivad ka finantsasutused, kes finantseerides erinevaid projekte, otseselt ja kaudselt mõjutavad kogu keskkonna käekäiku ning mõju ühiskonnale. Seepärast ESG seotud teguritega arvestamise avalikustamise tugevdamiseks uuendas Euroopa Väärtpaberiturujärelevalve (ESMA) krediidireitingute avalikustamise nõuete suuniseid [2]. Nüüd lasub finantsasutustel eriline roll jälgida vastutustundlikkust projektide finantseerimisel ja klientide varade haldamisel. Ettevõtted, kes varem tähelepanu ei suunanud vastutustundlikkusele, kohandavad enda käitumist üha rohkem ajas. Seda näitab ka allpool olev joonis, kus Russell 3000 USA firmad, kes igal aastal ESG raporteid koostavad, on järsult kasvanud.



Joonis 1 Russell 3000 ettevõtete ESG raporteerimine.

Allikas:Audit Analytics [3]

ESG tähendus on tõusnud viimastel aastatel investorite teadvusesse ning jätkusuutlik rahandus, sellega arvestamine investeerimisotsuste langetamisel on saanud lahutamatuks osaks investoritele. ESG-le pani algpunkti „Pariisi kokkulepe“, millega seati eesmärgiks tegeleda kliimamuutustega ning ühildada rahastamine kestliku arenguga.

Alates 2019. 11.detsembrist, kui Euroopa Liit sõlmis „rohelise kokkuleppe“, alustas ESG valdkond lumepallina veeremist. Maailma suurima varahalduri Blackrocki fondi juht Larry Fink on öelnud, et lähima 5 aasta jooksul hakkavad kõik investorid arvestama ESG faktoreid investeerimisel [4].

Euroopa liidus kehtib ühine klassifitseerimissüsteem või EL taksonoomia, mille eesmärk on soodustada erasektori investeeringuid kestlikku majanduskasvu ja aidata kaasa kliimaneutraalse majanduse saavutamisele. See süsteem annab ettevõtjatele ja investoritele ühise määratluse selle kohta, milliseid majandustegevusalasid saab pidada keskkonnasäästlikuks. [5]

Taksonoomia võimaldab investoritel suunata oma investeeringud kestlikumatesse tehnoloogiatesse ja ettevõtetesse. See süsteem on väga oluline vahend, mis aitab ELil saavutada 2050. aastaks kliimaneutraalsuse ja 2030. aastaks Pariisi kokkuleppe eesmärgid. [5]

ESG kriteeriumite jälgimist mõõdetakse läbi CSR (*ingl. K Corporate Social Responsibility*). Ettevõtte ühiskondlik vastutus ehk vastutustundlik ettevõtlus on kontseptsioon ja juhtimisvahend, mille alusel ettevõtted integreerivad sotsiaalsed ja keskkonnaeesmärgid vabatahtlikult oma tegevusse ning suhetesse huvirühmadega. CSR-i probleem on see, et kõik ettevõtted seda ei jälgi ja CSR-i reitingud on antud kolmandate agentuuride poolt. [6]

Viimastel aastatel on saanud üha populaarsemaks majandusaruannete analüüsimine erinevate uurimisküsimuste lahendamiseks tekstikaeve meetoditega. Mitmeid kordi on uuritud ka ESG valdkonna küsimusi aruannetes[7, 8, 9]. Enamus ESG seotud uuringuid, mis on tehtud kasutades tekstianalüüsi, on lahendatud sõnasageduste leidmisel tekstist ja selle konteksti panemisel. Selles töös mõõdame erinevate objektide sarnasust n-mõõtmelises ruumis ning rakendame vektorruumil põhinevat mudelit, mis kasutab koosinussarnasust (*ingl k. cosine similarity)* majandusaruannete teksti sõnade vektoresituse ja ESG-ga seotud terminite vahel.

1.1 Ülesande püstitus ja eesmärk

Töö põhieesmärgiks on luua vektorruumil põhinev mudel, mis suudab mõõta ESG valdkondade panust otse ettevõtete aruannetest ning järjestada aastate aruanded vastavalt sarnasusele. Tänu sarnassuskoorile ja järjestusele saame ülevaate, kui palju üks või teine ettevõte ESG teemadel informatsiooni kajastab oma aruannetes.

Ettevõtte majandusaruanne annab täpse ülevaate erinevatest tegevustest ning investorite meelitamiseks peaks aruanne sisaldama ka vastutustundlikkuse ja ESG-ga seotud valdkondade termineid.

Mudeli loomine ja tulemused keskenduvad sellele, kui palju firmad raporteerivad erinevaid ESG aspekte. See lubab analüüsida, kui palju on firmad teadlikud ESG valdkondadest ja selle raporteerimisest. Mudeli tulemustele põhinedes saab teha arukamaid investeerimisotsuseid, arvestades ettevõtete panust keskondlikesse, sotsiaalsetesse ja ühingujuhtimise valdkondadesse.

Analüüsimiseks valitakse Eesti pankade (Swedbank, LHV, Luminor ja SEB) inglisekeelsed majandusaasta aruanded alates 2016. aastast ning rakendatakse andmete peal juba olemasolevat vektorruumil põhinevat koosinussarnasuse mudelit.

Töö alameesmärgiks on välja selgitada, kas vektorruumil põhinev mudel oskab piisavalt hästi pankade aruannetest ESG andmeid välja lugeda ja usutava tulemuse anda.

Mudeli tulemusi valideeritakse Estwatchi loodud analüüsi vastu, mis leidis aset 2020. aasta alguses ning analüüsiti Eesti pankade ESG tegevust, mille tulemusena omistati neile hinnanguline skoor.

1.2 Töö struktuur

Antud töö koosneb viiest osast.

Esimeses peatükis antakse töö sissejuhatus, eesmärgid ja ülesandepüstitus.

Teises peatükis antakse ülevaade ärilisest taustast ning tehakse kirjanduse ülevaade. Autor toob välja, mis on ESG ja kuidas seda rakendatakse. Samuti räägitakse varasematest rakendustest.

Kolmandas osas tutvustatakse põhjalikult tööriistu, mida töös kasutati ja mis aitasid tulemuseni jõuda. Samuti tuuakse välja juba olemasolevad kasutusel olevad teegid ja raamistikud.

Neljandas osas kirjeldatakse ka töö tegemise metoodikat ehk protsessi, kuidas eesmärgini jõuti. See kätkeb endas nõudeid, tehnilist arhitektuuri, programmeerimiskeele valikut. Samuti tuuakse välja juba olemasolevad kasutusel olevad teegid ja raamistikud.

Viiendas osas kirjeldatakse peamisi tulemusi ehk tehniline dokumentatsioon tulemist, mida autor saavutas tööga. Tuuakse välja kogu analüüs ja järeldused: nõuete analüüs ja abstraktsioon, tulemuste interpretatsioon, järeldused resultaadist.

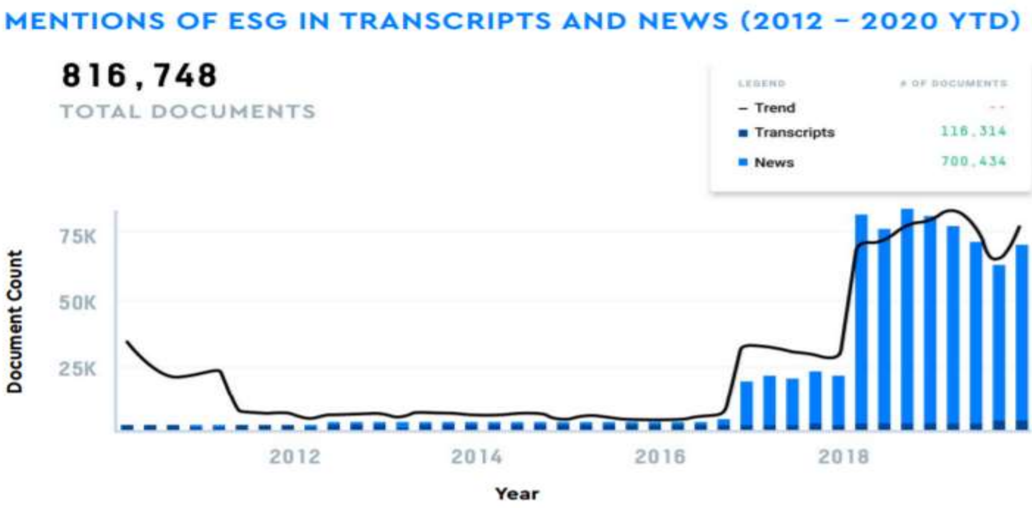
2 Äriline taustinformatsioon ja kirjanduse ülevaade

2.1 Äriline taust

2.1.1 ESG investeerimine

Akronüüm ESG viitab paljudele keskkonna-, sotsiaalsetele ja juhtimiskriteeriumitele, mille alusel ettevõtteid mõõdetakse. See peegeldab tarbijate kasvavat tundlikkust selle suhtes, kuidas ettevõtted ühiskonda panustavad. Ja see pakub üha suuremat huvi investoritele, kes on mures sellepärast, et ettevõtted võtavad kasutusele tavad, mis vähendavad riski ja tagavad nende pikaajalise jätkusuutlikuse. Selle tulemusena, ESG valdkonna teemad üha rohkem kujundavad ettevõtete äritegevust üle maailma. [10]

Alates aastast 2005, mil ESG termin populariseeriti, on investorid üha enam väärtustanud ideed kasutada ESG-tegureid investeerimisotsuste tegemisel. ESG investeerimise idee on suur areng sotsiaalselt vastutustundliku investeerimise suunas, kuid ESG pakub ka laiemat raamistikku sotsiaalse mõju uurimiseks, mitte ainult negatiivsete tulemustega seotud ettevõtete välistamiseks [10]. Viimase 10 aasta jooksul, uuringud indikeerivad, et ESG trend on megatrend, aga see ei ole selline trend, nagu oli 90ndate tehnoloogiamull. Alloleval joonisel on näha, et väljaütlemised ja uudised ESG valdkondade kohta on kasvanud järsult alates 2016. aastast ning jätkab kasvamist. [11]



Joonis 2 ESG populaarsus.

Allikas: AlphaSense

ESG investeerimise trend on olnud viimastel aastatel tõusujoones. Raporti järgi ülemaailmsed jätkusuutlikud investeeringud tõusid ligi 34% kahe aasta jooksul, 2016-2018. Uuring viidi läbi viies suuremas regioonis ning Euroopa ja USA olid suuremad turud investeeringute suhtes, vastavalt 14 triljonit ja 12 triljonit dollarit. [12]

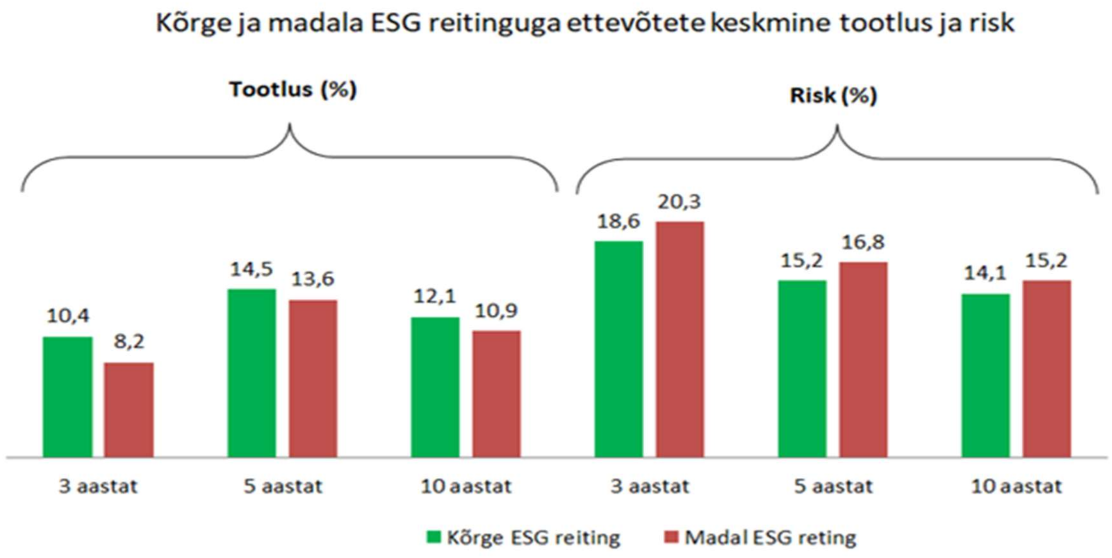
Faktorid ja valdkonnad, mis peegeldavad ESG võivad olla paljudes raamistikes erinevad ja paljud ESG uurivad artiklid viitavad sellele, et ei ole olemas ühtset raamistikku, millega kõiki ettevõtteid üle maailma mõõta. Üldiselt jaotatakse 3 suurde valdkonda ning igal valdkonnal on omad faktorid, mida jälgitakse [13]:

- Keskkond
    - Biokütused, kliimamuutuste strateegia, heitkoguste juhtimine ja aruandlus, juurdepääs maale, bioloogilise mitmekesisuse juhtimine, vesi, keskkonnastandardid, reostuse kontroll, tootevõimalused, tarneahela keskkonnastandardid, jäätmed, taaskasutus;
- Sotsiaalne
    - Ligipääs meditsiinile, HIV/AIDs, tervislikkus, toote ohutusnõuded, privaatsus ja vabadus arvamust avaldada, turvalisus, mitmekesisus, tervis ja turvalisus, tarneahela tööstandardid, altkäemaks ja korruptsioon, poliitiline mõjusus, teadlik turundus, vilepuhujate süsteemid, avalikustamine ja aruandlus, jätkusuutlikkuse probleemide juhtimine, sidusrühmade kaasamine, UNGC(*Ühinenud Rahvaste Organisatsiooni ülemaailmne kokkulepe*);
- Ühingujuhtimine
    - Audit ja kontroll, juhatuse struktuur, töötasu, aktsionäride õigused, läbipaistvus ja toimivus.

Pikki ESG teemalisi raporteid ei koosta firmad mitte ainult uute investorite jaoks, vaid paljude juhid on ka ise tõdenud, et kindla ESG poliitika juurutamine on positiivset mõju avaldanud nende äri tootlikkusele, kohandades oma ärimudelid ümber selliselt, et saadaks kasu jätkusuutlikest ja säästvatest trendidest. [14]

ESG poliitika paikapanemine on küll ettevõttele oluline kulu, kuid mitmed uuringud on näidanud, et pikas perspektiivis muudab see ettevõtte kliendid lojaalsemaks, stabiliseerides seeläbi ka müügitulu. [14]

Allolev joonis näitab, et ESG poliitikat integreerinud USA ettevõtted on nii lühi kui ka pikal perioodil ületanud tulemuslikkusega teisi turuosalisi, kes ESG-ga pole arvestanud.



Joonis 3 ESG tootlus.

Allikas: JPMorgan

LHV analüütiku Raido Tõnissoni sõnul on jätkusuutlike ettevõtete riski mõõtev volatiilsus olnud kõigil ajaperioodidel madalam kui teistel, andes ESG ettevõtetele kõrgema riski ja tulu suhte. Samas mainib Raido, et vettpidavate järelduste tegemiseks on antud ajaperioodid liiga lühikesed, näitavad olemasolevad andmed, et jätkusuutlikesse ettevõtetesse investeerimine toob investorile parema tulemuse. [14]

2.1.2 Jätkusuutlik rahandus ja Euroopa taksonoomia

Euroopa Liidu poolt 5. oktoobril 2016 heaks kiidetud ÜRO (*Ühinenud Rahvaste Organisatsioon*) kliimamuutuste raamkonventsiooni alusel vastu võetud Pariisi kokkuleppe (edaspidi „Pariisi kokkulepe“) eesmärk on tugevdada kliimamuutustele reageerimist, muu hulgas viies investeerimisvood kooskõlla arenguteega, mis on suunatud vähese kasvuhoonegaaside heite ja kliimamuutustele vastupanuvõimelise arengu saavutamisele. [15]

Euroopa komisjon võttis 11. detsembril 2019 vastu „Euroopa roheline kokkulepe“, mille eesmärk on muuta Euroopa Liit rohelise majandusega ühiskonnaks, kus 2050. aastaks ei ole enam kasvuhoonegaaside heiteid ja kus majanduskasv ei olene ressursikasutusest [16]. Euroopa rohelise kokkuleppe rakendamiseks on vaja, et investoritele antaks selgeid ja pikaajalisi märke, et hoida ära varade muutumist kasutamatuks ja panustada vastutustundlikkusse rahastusse. [16]

Euroopa Liit võttis vastu 12. juulil 2020 määruse, millega loodi „jätkusuutlike investeeringute soodustamise raamistik“. See paneb aluse ELi õigusraamistikule, milles on finantssüsteemi keskmesse asetatud ESG-ga seotud kaalutlused, et toetada Euroopa majanduse muutmist keskkonnahoidlikumaks ja vastupidavamaks. [17]

Et investeeringud oleksid vastutustundikumad, tuleks otsuste tegemisel arvestada ESG teguritega.

See peaks tagama, et kõik turu osalised (fondide valitsejad, kindlustusandjad, tööandja kogumispensioni asutused), kindlustuse turustajad ja investeerimisnõustajad, kes saavad oma klientidelt volituse teha nende nimel investeerimisotsuseid, lisavad keskkonna-, sotsiaal- ja juhtimiskaalutlused oma siseprotsessidesse. [17]

ELi eesmärkide saavutamiseks on vaja olulisi ESG-ga teguritega arvestatud investeeringuid. Hinnanguliselt on 2030. aasta kliima- ja energeetikaeesmärkide saavutamiseks vaja ainuüksi lisainvesteeringuid 180 miljardit eurot aastas. [18]

2.1.3 Varasemad rakendused

Palju organisatsioone avalikustavad jätkusuutlikkuse reitinguid, mis mõõdavad ettevõtete panust jätkusuutlikusse. Need organisatsioonid analüüsivad firmasid ESG kriteeriumite alusel. [19]

Reitingute suurimateks probleemideks on see, et reitinguagentuuride skooride kalkulatsioonid on erinevad ning agentuurid kasutavad erinevaid andmeid ja muutujaid. Chelli ja Gendron sõnul põhinevad ESG hinnangute hindamisprotsessid erinevatel hindamismetoodikatel ja ka viisidel. [20]

Reitingukriteeriumid on tuletatud äriorganisatsioonide vastutuse eetilisest vaatenurgast ja jätkusuutlikkuse mõiste varieerub selle rakendamisel ideoloogiliste piirangute järgi [21]. Euroopa komisjon on vastu võtnud määruse, mis määrab ära, kuidas kajastatakse finantsvarade rühmas ESG tegureid. Varade haldurid selgitavad, kasutades väljatöötatud materjale ja vorme, kuidas reflekteeritakse varades kirjeldatud keskkonna-, sotsiaal- ja juhtimistegureid. Varade halduritel lasub kohustus uuendada selgitusi iga kord, kui ESG teguritega toimuvad muutused ja igal juhul ühe korra aastas. [22]

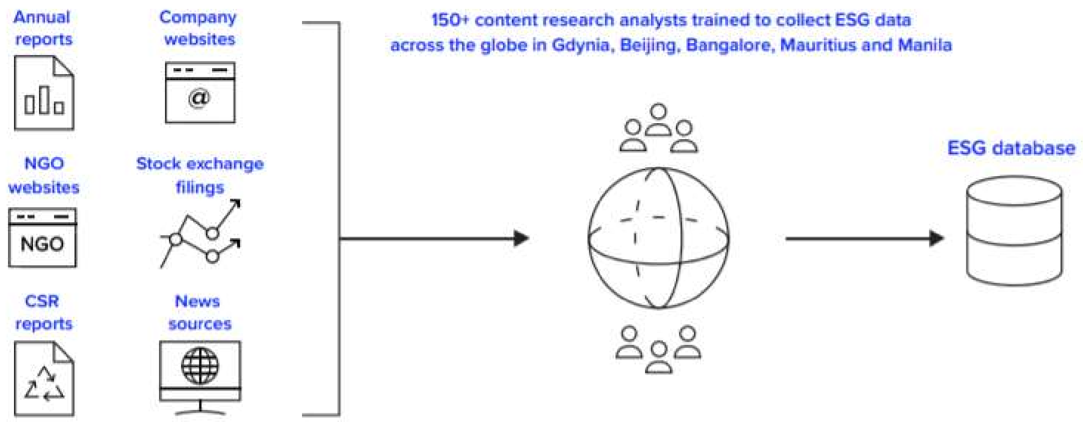
VORM SELLE SELGITAMISEKS, KUIDAS KAJASTATAKSE VÕRDLUSALUSE KIRJELDUSES KESKKONNA-, SOTSIAAL- JA JUHTIMISTEGUREID

SELGITUS SELLE KOHTA, KUIDAS KAJASTATAKSE VÕRDLUSALUSE KIRJELDUSES KESKKONNA-, SOTSIAAL- JA JUHTIMISTEGUREID	
1. JAGU – KESKKONNA-, SOTSIAAL- JA JUHTIMISTEGURITE ARVESSEVÕTMINE	
Punkt 1. Võrdlusaluse halduri nimi.	
Punkt 2. Võrdlusaluse või võrdlusaluste rühma liik. *Valige II lisas esitatud loetelust asjaomane alusvara.*	
Punkt 3. Võrdlusaluse või võrdlusaluste rühma nimi.	
Punkt 4. Kas võrdlusaluste halduri portfellis on ELi kliimaüleminekü võrdlusaluseid, Pariisi kokkulepet järgivaid ELi võrdlusaluseid, keskkonna-, sotsiaal- ja juhtimisalaseid eesmärke järgivaid võrdlusaluseid või keskkonna-, sotsiaal- ja juhtimistegureid arvesse võtvaid võrdlusaluseid?	☐ Jah ☐ Ei
Punkt 5. Kas võrdlusalus või võrdlusaluste rühm järgib keskkonna-, sotsiaal- ja juhtimisalaseid eesmärke?	☐ Jah ☐ Ei
Punkt 6. Kui vastus punktile 5 on positiivne, esitage allpool üksikasjad (punktisumma) seoses II lisas loetletud keskkonna-, sotsiaal- ja juhtimisteguritega iga võrdlusaluste rühma kohta koondtasandil. Keskkonna-, sotsiaal- ja juhtimistegurid avalikustatakse võrdlusaluste rühma tasandil agregeeritud kaalutud keskmise väärtusena.	
a) Kombineeritud keskkonna-, sotsiaal- ja juhtimistegurite loetelu:	Üksikasjad iga teguri kohta:

Joonis 4 Vorm selgitamaks, kuidas kajastatakse ESG tegureid.

Allikas: Euroopa teataja [22]

Refinitiv, mis on tegutsenud turul juba pikka aega, mis kogub infot just ESG kohta erinevatele firmadele üle maailma.



Joonis 5 Refinitiv mudel.

Allikas: Refinitiv [23]

Nende algoritmid on ehitatud järgmiselt: andmeid võetakse mitmest kohast - aruanded, CSR raportid, ettevõtte veebilehed, uudised. Andmeid masintöödeldakse, tehakse sentimendi analüüsi ja siis klassifitseeritakse mudeli tulemuste põhjal välja kategooriatele skoor.

Refinitivi mudeli järgi antakse skoor mitmele erinevale kategooriale:

Joonis 6 Refinitiv ESG skoor.

Allikas: Refinitiv [23]

3 Tehniline taust ja masintöötlusvahendid

Käesolevas töös keskendun dokumentide sarnasuste leidmisele, et kaardistada ja mõõta dokumentide omavahelised sarnasused ning nende põhjal hinnata ettevõtete panust ESG valdkondadesse.

Sarnasuste leidmiseks on tekstitöötluses erinevaid viise: tavaline TF-IDF mudel, eeltreenitud mudelid, LDA, LSI, vektorruumil põhinevad mudelid. Selles peatükis kirjeldan vektorruumil põhinevat mudelit, mis mõõdab koossinussarnasust dokumentide vahel.

3.1 Vektorruum

Lineaaralgebra üheks põhimõisteks on vektorruumi mõiste. Ühtlasi on see üks sagedamini kasutatavaid algebralise struktuuri mõisteid tänapäeva matemaatikas. Näiteks, paljud matemaatilises analüüsis vaadeldavad funktsioonide hulgad on oma algebraliste omaduste poolest vektorruumid. [24]

Hulka V nimetatakse vektorruumiks üle arvukorpuse K, kui temas on antud kaks tehet – liitmine (igale kahele elemendile $\alpha, \beta \in V$ on vastavusse pandud parajasti üks element $\alpha + \beta \in V$), ja skalaariga korrutamine (igale arvule $a \in K$ ja hulga V elemendile α on vastavusse pandud parajasti üks element $a\alpha \in V$), nii et on täidetud järgmised aksioomid [25]:

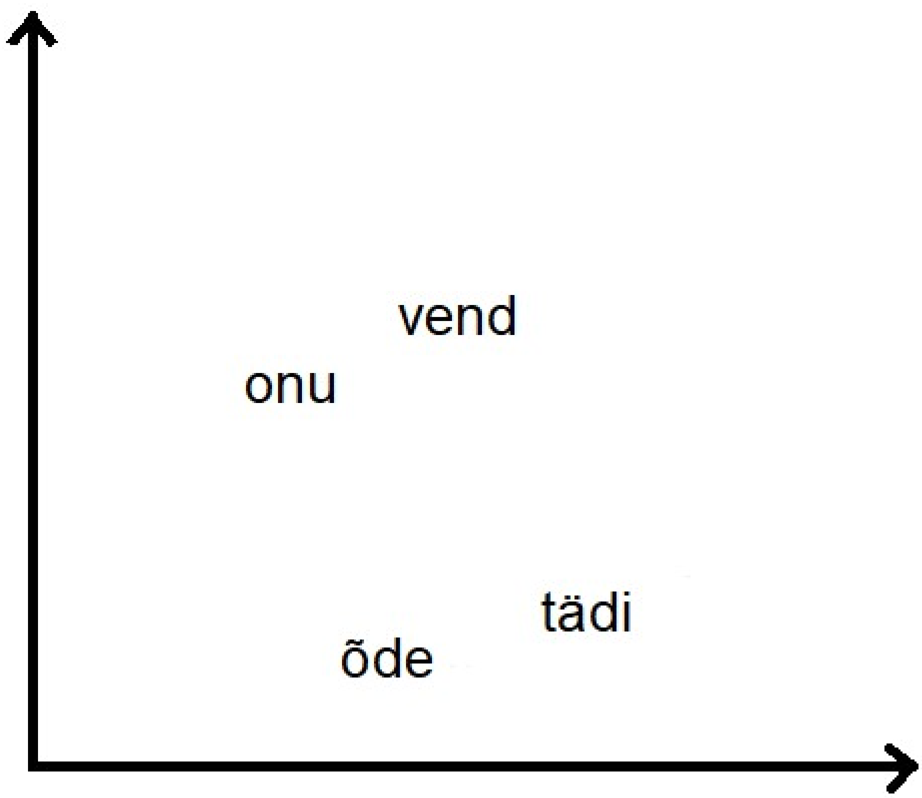
1) $\alpha + \beta = \beta + \alpha$ iga $\alpha, \beta \in V$ korral (liitmise kommutatiivsus);
2) $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ iga $\alpha, \beta, \gamma \in V$ korral (liitmise assotsiatiivsus);
3) $\exists \theta \in V : \theta + \alpha = \alpha$ (nullvektori olemasolu);
4) $\forall\, \alpha \in V$ korral $\exists - \alpha \in V$ nii, et $\alpha + (-\alpha) = \theta$ (vastandvektori olemasolu);
5) $1\alpha = \alpha$ (unitaarsus);
6) $(ab)\, \alpha = a(b\, \alpha)$ iga $a, b \in R$ ja $\alpha \in V$ korral (assotsiatiivsus arvude korrutamise suhtes);

7) $a(\alpha + \beta) = a\alpha + a\beta$ iga $a \in R$ ja $\alpha, \beta \in V$ korral (distributiivsus vektorite liitmise suhtes);

8) $(a + b)\alpha = a\alpha + b\alpha$ iga $a, b \in R$ ja $\alpha \in V$ korral (distributiivsus arvude liitmise suhtes).

Vektorruumil põhinevat sõnadega rikastatud mudelit peetakse sõna tähenduste ruumilisteks esitusteks, mis põhinevad arusaamal, et tähenduslikku sarnasust saab esitada kui sõnade lähedust n-mõõtmelises ruumis, kus n võib tähistada iga täisarvu alates ühest. [26]

Allpool toodud joonisel näeme ühte võimalikku näidet mudelist, kuhu on rikastatatud 4 erinevat sõna ning neil on tähenduslik sarnasus. Sõna *onu* on vektorruumis lähedamal sõnale *vend* kuna sõna *onu* tähenduslik sarnasus on lähedamal sõnale *vend*.



Joonis 7 Kahemõõtmeline sõnaruum.

Allikas: Autor

3.2 Sõnad vektoresituses – GloVe eeltreenitud mudel

GloVe tähendab „Global vectors“ ehk globaalsed vektorid. See mudel kasutab korpuse globaalset statistikat ja lokaalset statistikat, et tuletada sõnavektoreid.

GloVe meetod põhineb ideel, et semantilisi sarnasusi sõnade vahel on võimalik saada kätte samaaegse esinemise maatriksist.

Kui korpuses on *V* arv sõna, samaaegse esinemise maatriks *X* oleks *V x V* maatriks, kus *i'nda* rea ja *j'nda* tulba väärtus X_{ij} indikeerib, kui mitu korda sõna *i* on koos esinenud sõnaga *j*. [27]

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Joonis 8 Koosesinemise maatriks lausele „the cat sat on the mat“.

Allikas: TowardsDataScience matemaatiline seletus [27]

GloVe on sisuliselt bilineaarne mudel, millel kasutatakse lineaarse regressiooni kaalutud väikseimate ruutude meetodit. [28]

Näiteks, mudeli põhjal väljaarvutatud tõenaosused sõnade *ice, gas, water* ja *fashion* kohta.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Joonis 9 Arvutatud tõenäosused 6 miljardist sõnadest koosnevast korpusest.

Allikas: Glove dokumentatsioon [28]

Nagu võib arvata, sõna *ice* esineb koos *solid* sõnaga sagedamini, kui *gas* sõnaga ning sõna *steam* esineb sagedamini koos sõnaga *gas*, kui *solid* sõnaga. Mõlemad sõnad *ice* ja *steam* esinevad sagedamini koos neile sarnase sõnaga *water* ja vähem sagedamini esinevad koos sõnaga *fashion*. Sellisel viisil tõenäosusstatistika suhete teooria muundab lahti väga tähtsat informatsiooni abstraktsete kontseptsioonide kohta. [28]

Antud töös kasutatakse GloVe eeltreenitud mudelit nimega *glove-wiki-gigaword-50,*kus on 400 000 vektorit ning dimensioone on 50. [28]

3.3 Distributiivne semantiline sarnasus

Distributiivne semantika võimaldab mudelites rakendatuna mõõta keeleelementide (nt sõnatähenduste, tekstide) sarnasusi [29]. Selliste mudelite üldine tööpõhimõte on järgmine: distributiivne info (lingvistilise elemendi esinemistingimused teiste elementide suhtes) kogutakse kokku vektoritena (ehk leitakse sõnade vektoresitused) ning seejärel esitatakse elementide tähenduslik sarnasus nende vektorite sarnasuse kaudu. [30]

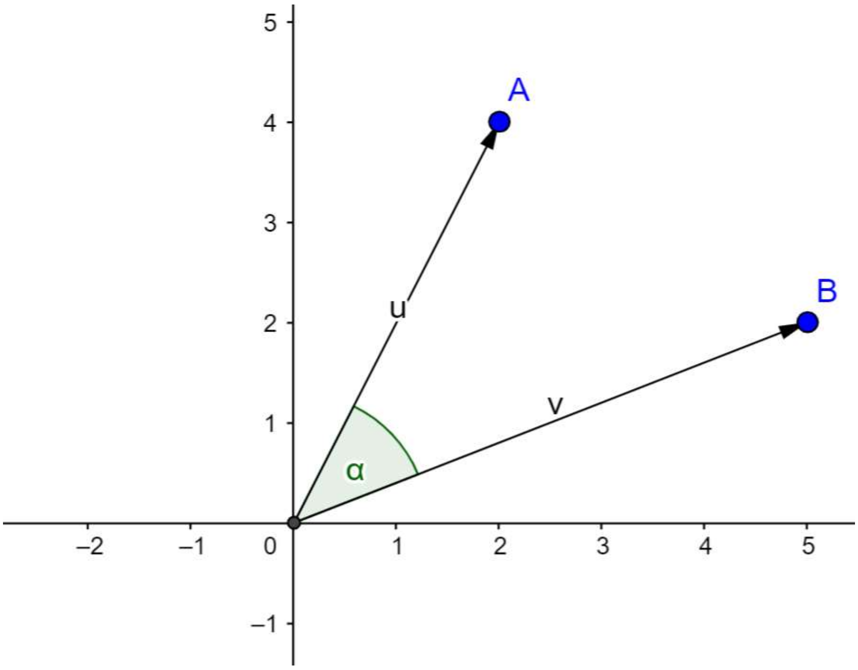
Neid distributiivse semantika mudeleid, mis kasutavad kontekstina keeleelemente nimetatakse vektorruumi ehk semantilise ruumi ehk sõnaruumi mudeliteks[26, 31], mis tuletavad sõnade tähenduse samas kontekstis koosesinemise põhjal.

Sellised mudelid interpreteerivad sõnadest ruumi, kus sarnasust mõõdetakse sõnadevahelise ehk vektorite vahelise kaugusena selles ruumis.

3.3.1 Koosinussarnasus

Üks levinumaid tekstide ja dokumentide sarnasuse mõõtmise vahend andmeteaduses on koosinussarnasus. Seda rakendatakse mitmetes valdkondades, nagu näiteks leides sarnaseid dokumente, informatsiooni pärimisel, sarnase DNA leidmisel, plagiaadi tuvastamisel. [32]

Koosinussarnasus mõõdab vektorruumis paiknevat kahe nullist erineva vektori vahelise nurga koosinust.



Joonis 10. Kahe vektori vaheline nurk.

Allikas: Autor

Koosinussarnasus arvutatakse järgmiselt:

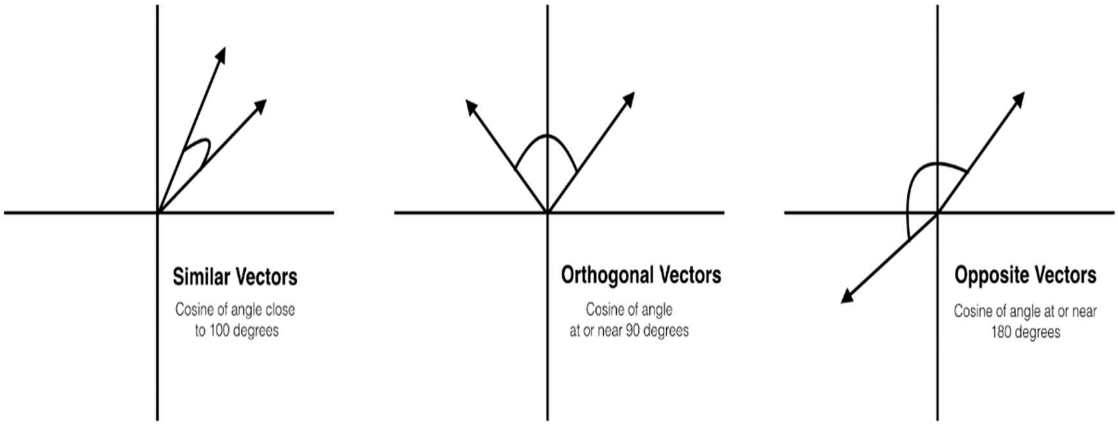
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Joonis 11 Kahe vektori vahelise koosinuse arvutamise valem.

Allikas: Wikipedia

Lühidalt, kaks koosinusvektorit, mis on samas orientatsioonis joondatud, on koosinussarnasus 1. Vastupidi, kui kaks vektorit on ortogonaalsed, siis sarnasus on 0.

Kui kaks vektorit on vastandvektorid, see tähendab, et nad on suunatud täpselt vastassuunas (st tagasi-tagasi), siis on sarnasus -1. Sageli kasutatakse siiski koosinuse sarnasust positiivses ruumis, piiride 0 ja 1 vahel. [33]



Joonis 12 Erinevad vektorid ja nende tähendused

Allikas: Oreilly

3.4 Tööriistad

3.4.1 Python

Python on üldotstarbeline, objektorienteeritud, väike, võimas, lihtne ja lõbus – nagu väidavad keele loojad ja arendajad – vabavaraline programmeerimiskeel, mis on loodud 1991. aastal. Autor on Guido van Rossum (Holland).

Python leiab laialdast kasutamist erinevat liiki tarkvara loomisel, muuhulgas ka veebirakenduste juures ning dokumendipõhistes rakendustes.

Kasutamise ulatuselt on ta võrreldav PHP ja Visual Basicuga. Neist kõrgemal on vaid sellised keeled nagu Java ja C-pere keeled (C, C++, C#), mis on eeskätt süsteemprogrammeerimise keeled. [34]

Antud töös kasutame Pythoni 3.8 versiooni.

3.4.2 Gensim

Gensim on avatud lähtekoodiga Pythonis implementeeritud semantiliste vektoresituses olevate dokumentide töötlemise tööriistakomplekt.

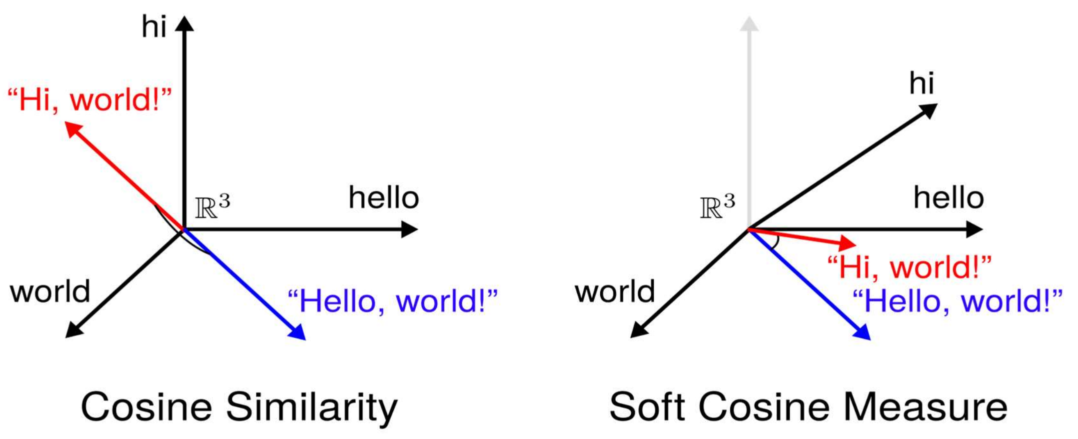
Gensim on spetsiaalselt loodud suurte tekstikogumite käitlemiseks kasutades juhendamata masinõppealgoritme. [35]

Gensimi disaini printsiibid [35]:

- Praktilisus – fokusseeritakse „reaalsetele probleemidele“ ning keskendutatakse inseneeriale, vähem teadusele;
- Mälu kasutus – gensimi paketiga ei ole vaja tervet andmemahtu RAMile paigutada;
- Jõudlus – hästi optimiseeritud vektorruumil põhinevad implementatsioonid.

3.4.3 Soft Cosine

Pehme koosinus (ingl k *SoftCosine*) lubab meil hinnata kahe dokumendi vahelist sarnasust, isegi kui neil ei ole ühtegi sarnast sõna. Allpool toodud pilt illustreerib kahe erineva meetodi kasutust kahe samasuguse lause jaoks. Lausetel ei ole ühtegi ühist sõna, aga kui modelleerida sünonüümia ehk samatähenduslikkus sõnade vahel, *SoftCosine* suudab täpselt mõõta kahe lause vahelise sarnasuse. *SoftCosine* meetod kasutab *bag-of-words* dokumentide vektoresitust. [36]



Joonis 13 Erinevused koosinussarnasuse ja *pehme* koosinussarnasuse vahel.

Allikas: Pehme koosinussarnasuse autori dokumentatsioon [37]

Kontseptsioon meetodi taga on, et kalkuleeritakse standard koosinussarnasus eeldades, et dokumendi vektorid on esitatud mitte ortogonaalsel baasil, kus nurk kahe vektori vahel on tuletatud eeltreenitud mudeli sõnade representatsioonide (ingl embeddings) vahelisest nurgast. Antud töös kasutame Gensimi *SoftCosineSimilarity* meetodit, mis rakendab pehmet koosinust.

3.4.4 TF-IDF

TF-IDF (ingl k *Term Frequency - Inverse Document Frequ*ency) on suurus, millega mõõdetakse sõna "olulisust" korpuses. Arvutamisel võetakse arvesse sõnasagedust (TF) ja dokumendi pööratud esinemissagedus (IDF).

TF on lihtne terminisageduse ja kõigi terminite sageduste summa suhe. IDF näitab, kui oluline sõna korpuses on. Kui sõna esineb ühes dokumendis sageli, võib eeldada, et see on oluline. Kui sama sõna aga esineb sageli ka teistes dokumentides, võib eeldada, et tegemist pole väga informatiivse sõnaga ning sõna olulisust vähendatakse. TF-IDF suhet arvutatakse järgnevalt: TF-IDF = TF*IDF, kus [38]:

$$\mathrm{TF(t)} = \frac{\text{termini(t)sagedus}}{\text{kõigi terminite arv}}$$

ja

$$\mathrm{IDF(t)} = \log\frac{\text{kõigi dokumentide arv}}{\text{dokumentide arv kus esineb termin t}}$$

TF-IDF rakendatakse praktikas paljude tekstikaeveaga seotud ülesannete lahendamiseks näiteks [38]:

- Stopp-sõnade loendi tegemisel. Terminid, mille TF-IDF väärtus on null või nulliähedane võib lisada stopp-sõnade loendisse. Need on terminid, mis esinevad peaaegu kõigis dokumentides;
- Oluliste sõnade tuvastamine. Terminid, millel on kõrge TF-IDFväärtus, on tõenäoliselt kõige olulisemad;
- Dokumentide klassifitseerimine TF-IDF väärtuste põhjal. Neist väärtustest saame luua kaugusmaatriksi, mida võime kasutada mõne klassifitseerimismeetodi sisendina. Analüüsi tulemusena moodustavad sarnasemad dokumendid eraldi rühmad.

3.4.5 Veebist andmete kraapimine ja TIKA teenus

Veebikraapimine on protsess, mis koosneb andmete otsimisest ning eraldamisest veebilehe lähtekoodist, mis üldjuhul on HTML*(Hyper Text Markup Language)* koodis. Kraapimist võib teha manuaalselt, aga enamik juhtudel on see protsess automatiseeritud, et kogu töö oleks efektiivsem ja vähem kulukam.

TIKA teenustepakett on eelkõige mõeldud suurte failide töötlemiseks ja erinevate failiformaatidest teksti eraldamiseks. *TIKA* peamised omadused on [39]:

- Tika server: ressursid on kättesaadavad läbi rakendusteenuse(ingl k. *application programming interface, API*);
- Identifitseerib faili MIME tüübi;
- Identifitseerib faili meta-andmed;
- Identifitseerib dokumendi keele.

3.4.6 Jupyter Notebook

Jupyter Notebook on brauseris töötav rakendus, mis lubab Pythonil suhelda kasutajaliidese vahendusel. Jupyter Notebooki omadused:

- saab interaktiivselt käivitada lühikesi koodijuppe;
- kogu sisend-väljund hoitakse töölehel, kõiki eelnevaid sisestusi saab redigeerida ja vastavaid arvutusi korrata;
- koodijuppide vahele saab lisada erineva kujundusega teksti, valemeid ja jooniseid.

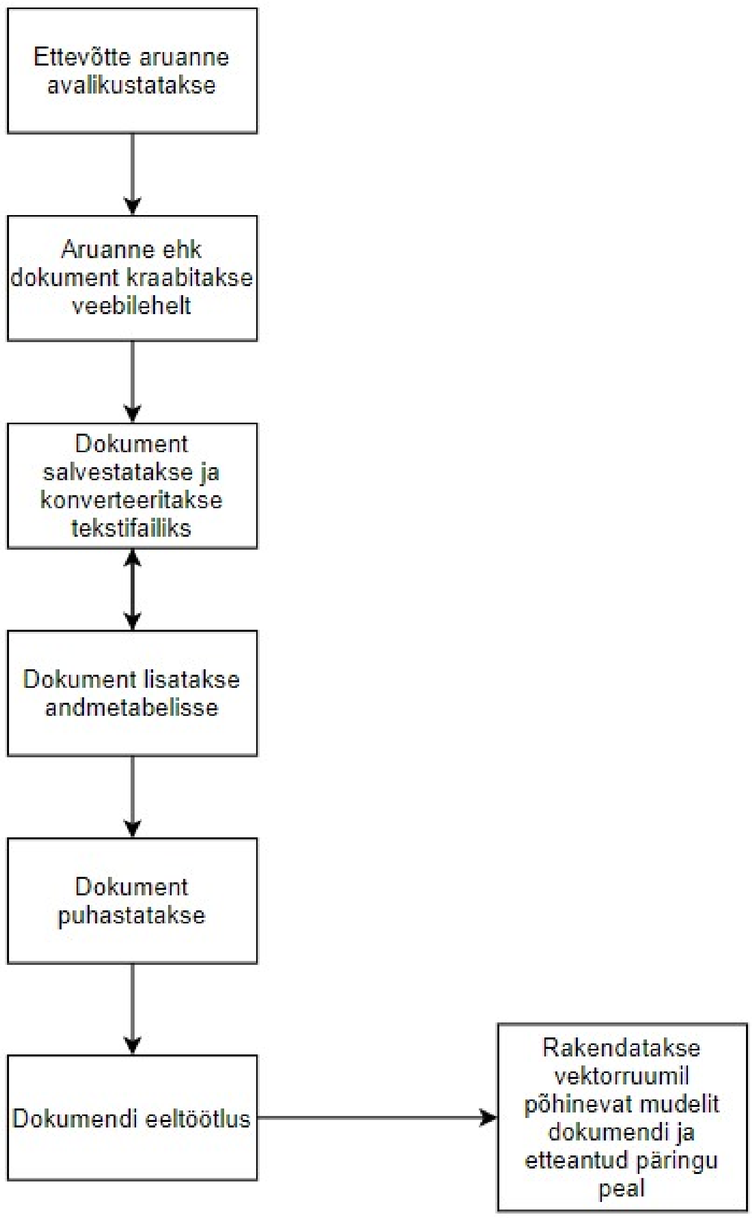
Jupyter notebook on justkui "märkmik", mis sobib arvutuste ja koodijuppide eksperimenteerimiseks ning samal ajal ka dokumentatsiooni tegemiseks.

4 Metoodika

Selles peatükis annab autor ülevaate meetoditest, kuidas veebist saada andmed kätte, töödelda neid, puhastada andmeid ja ehitada valmis vektorruumil põhinev koosinussarnasus mudel.

Metoodika koosneb järgnevasest sammudest:

1. Dokumentide kraapimine ja salvestamine;
2. Dokumentide konverteerimine;
3. Teksti puhastamine;
4. Eeltöötlus tekstile ja päringule;
    a. Stopp-sõnade eemaldamine;
    b. Sõnestamine;
5. Mudeli ehitamine ja käivitamine.



Joonis 14 Metoodika protsessijoonis

Allikas: Autor

4.1 Dokumentide kraapimine ja salvestamine

Andmete saamiseks kasutatakse veebikraapimise metoodikaid. Autor kirjutas valmis meetodi *fetch_pdfs*, mis võtab sisse parameetriteks ettevõtete väärtpaberi unikaalse tunnuskood(ISIN) ja kuupäeva. Meetodis on defineeritud ära aadress, kust infot kraabitakse. Antud töös on defineeritud, et andmete pärimine toimub Nasdaq Baltic kodulehe pealt, kust laetakse alla lähtekood ning eraldatakse kõik ingliskeelsed PDF laiendiga failid etteantud kuupäevaga ja laetakse need lokaalsesse kausta.

Meetod on ehitatud nii, et kui antud meetodi käivitajal ei ole lokaalselt olemas sellise parameetriga(ISIN) kausta, siis see luuakse automaatselt.

Ettevõttete aruanded, mis ei ole Nasdaq Baltic väärtpaberite listis, kraabiti otse nende ettevõtete kodulehtedelt.

4.2 Dokumentide konverteerimine

Protsessi lihtsustamiseks konverteeritakse kõik PDF failid ümber *.txt* failideks, sest PDF failid ei ole olnud kunagi disainitud andmetega töötlemiseks, pigem on PDF formaat disainitud nii, et väljund oleks ilus ja loetav.

Olemuselt koosneb PDF fail voogudest, kuidas joonistada leht valmis. Tekstiandmed ei ole salvestatud paragrahvidena, isegi mitte sõnadena, vaid sümbolitena, mis on joonistatud kindlatele kohtadele lehtedel. Selle tulemusena, suurem osa sisu semantikast kaob ära, kui teksti või *word*'i dokument konverteeritakse PDF failiks – tekstistruktuur muudetakse lehekülgedel hõljuvateks amorfseteks tähtedeks, mis meenutaksid justkui tähesuppi [41].

Kui eelnevas peatükis kirjeldatud meetod on käivitatud ja kõik PDFid on sihtaadressilt alla laetud, siis käiakse *for* tsükliga üle terve allalaetud PDF'ide lokaalne kataloog ning *TIKA* paketi abil tehakse kõik PDF failid lahti, loetakse maha tekst ning salvestatakse faili tekst uue *.txt* failina.

4.3 Teksti esmane puhastamine

Teksti puhastamise protsessiga võib minna peensusteni välja ja see võib tulemusena muutuda väga ajamahukaks, sest igal puhastamise sammul on olemas erandeid. Tänu sellele, autor otsustas läheneda MVP(*minimum viable product*) loogikaga – alustada lihtsalt ja iga iteratsiooniga paremaks minna. Teksti puhastamiseks kasutatakse regulaaravaldisi, mis on olemas Pythoni *re* paketis. Iga dokumendi tekst sisaldab andmeid ja märke, mis ei ole vajalikud edasiseks analüüsiks ega tööks. Antud töös puhastatakse tekst kahes faasis:

1. Esimene puhastamine;
2. Teine puhastamine.

Esimeses puhastamise ringis võetakse ette dokumendi sisu ning kõigepealt tehakse kogu tekst väiketähtedeks, kuna tekstitöötluses ei anna suured tähed meile mitte mingit informatsiooni ja me kohtleme igat sõna samamoodi. Järgmisena eemaldatakse tekstist kogu informatsioon, mis on kandiliste sulgude vahel. Seejärel eemaldatakse kõik punktid ja numbrid tekstist.

Autor vahepeal kontrollis puhastatud teksti ja veendus, et veel on vaja teksti puhastada.

Teises puhastamise ringis eemaldati kõik jutumärgid, tühjad read ja tühimikud ning sümbolid, mis ei ole tähemärgid.

4.4 Teksti ja päringu eeltöötlus

Mudeli ehitamiseks on vaja läbida eeltöötluse sammud. Järgnevad sammud on vajalikud selleks, et mudel oleks võimalikult kasulik ja andmed oleksid sobivad edasise analüüsi jaoks. Allpool kirjeldan samme, mida pean vajalikuks teha tekstiga enne mudeli kasutust.

4.4.1 Stopp-sõnade eemaldamine

Stoppsõnad on tekstikaeves kasutatavad sõnad, mis teksti sisu kohta suurt väärtust ei loo. Stoppsõnade listid on erinevatel projektidel erinevad ning õiget ja lõplikku listi stoppsõnadest on raske määratleda [42]. Kuna meie sihtdokumendid on inglise keeles, siis kasutasime antud töös NLTK(*Natural Language Toolkit*) teegis olevaid stopp-sõnu. Stopp-sõnade hulka kuuluvad inglise keeles vähe informatsiooni andvad sõnad ning NLTK teegis on neid sõnu 179.

Näiteks esimesed listis olevad 10 sõna: ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"].

Stopp-sõnad eemaldasime kõikidest dokumentidest ja samamoodi ka defineeritud päringust, mida võrdleme dokumentidega. Pärast stopp-sõnade eemaldamist olid meie andmed väheinformatiivsetest sõnadest puhastatud ning oli võimalik edasi liikuda mudeli ehitamise järgmise sammuga.

4.4.2 Sõnestamine ehk sõnade üksustamine

Sõnade üksustamisel tükeldatakse etteantud dokument väikesteks tükkideks, *tokeniteks*, samal ajal visates välja erinevad tähemärgid, nagu punktuatsioonimärgid [43].

Token on järjestatud tähemärkidest koosnev üksus mingis dokumendis, millel on semantiline väärtus. [43]

Antud töös üksustasime nii kogu dokumentide tekstid ning ka defineeritud päringud, millega võrdlesime dokumente. Andmestiku üksustamiseks kasutasime Python'i meetodit *split(),* mis tagastab listi sõnedest ning vaikimisi üksustab kõik sõned sealt, kus on tühik.

Näiteks *query_E = 'Biofuels, Climate ,Emissions ,land, Biodiversity, Water, Environmental, standards, Pollution, Supply, Waste, recycling'* viidi kujule *['biofuels', 'climate', 'emissions', 'land', 'biodiversity', 'water', 'environmental', 'standards', 'pollution', 'supply', 'waste', 'recycling']*.

4.5 Mudeli ehitamine

Pärast andmete töötlemist ei ole võimalik veel mudelit jooksutada. Mudeli eelduseks on sarnasusmaartiks, mis põhineb GloVe eeltreenitud mudeli vektoresitus sõnadele ja nendevahelisele koosinussarnasusele. Seega genereerisime terminite sarnasusmaatriksi, kasutades Gensimi *WordEmbedddingSimilarityIndex* implementatsiooni GloVe vektorite peal. Tulemuseks on maatriks, mis sisaldab sarnasusi eeltreenitud mudelis olevate ingliskeelsete sõnapaaride vahel.

Sarnasusmaatriksi illustreerimiseks kasutan Python'i *vec2graph* implementatsiooni. Antud joonisel on näha, et sõnega *audit* top 10 lähimat ja sarnasemat sõne eeltreenitud GloVe mudelis on (sulgudes koosinussarnasusskoor):

```
[('auditors', 0.838019609451294),
 ('audits', 0.8341637253761292),
 ('auditing', 0.8266605138778687),
 ('investigations', 0.7749220728874207),
 ('accounting', 0.7711327075958252),
 ('oversight', 0.7665724754333496),
 ('auditor', 0.7644694447517395),
 ('reviewing', 0.7541238069534302),
 ('fec', 0.7462863326072693),
 ('irs', 0.7425697445869446)]
```

Illustreerimaks sõnade seost ja nende sarnasusi ruumis, genereerin top 10 lähimast sõnast graafi.

Joonis 15 Sarnasusmaatriksi vektorite graaf

Allikas: Autor

Algoritmidel põhinevatel mudelitel on tekstitöötlusega kergem tegeleda, kui andmed on viidud sellisele kujule, et need oleks mudelis olevate algoritmide jaoks loetavad. Selleks teisendasime dokumentide ja päringute andmed ühte sõnastikku, kasutades selleks Gensimi *Dictionary* implementatsiooni.

Andmesõnastik koosneb dokumendi sõnadest ja defineeritud päringute sõnadest, vormis 'id' :'sõna':'sagedus', näiteks *'429':'additional':'12'.*

Andmesõnastiku illustreerimiseks kasutati Gensimi meetodit Dictionary.Save ja salvestati terve sõnastiku lokaalselt *.dict* failiks ning siis saadi sõnastikku lugeda.

```
2	accordance	20
3	according	20
4	accounting	20
5	accrued	20
6	activities	20
7	added	20
429	additional	12
8	address	20
9	adequacy	20
10	adjusted	8
11	adjustment	20
12	adjustments	20
441	adoption	4
13	affected	20
14	almost	20
15	also	20
442	amendments	4
16	amortisation	20
17	amount	20
18	amounts	20
19	annual	20
430	application	4
20	applied	20
21	approach	20
22	arise	20
23	arises	20
443	article	9
24	assessment	20
25	assets	20
26	associates	19
27	assumptions	20
28	attributable	20
29	audit	12
30	audited	20
31	auditor	20
32	balance	20
```

Joonis 16 Andmesõnastik

Allikas: Autor

Järgmisena kasutatakse andmesõnastiku peal Gensimi *TF-IDF* implementatsiooni, et mõõta välja sõnade olulisus antud sõnastikus. TF-IDF implementatsioon eemaldab kõik sõned, millel olulisus on vähem kui $1 * 10^{-12}$ kogu andmestikus.

TF-IDF idee seisneb selles, et sõned, millel on kõrgeim skoor dokumendis, on informatiivsed ja meile olulised sõned. Madala skooriga sõned esinevad mitmeid kordi ja on meile vähemolulised.

Et TF-IDF ideed illustreerida, kirjutati valmis tsükkel, mis kasutab *TF-IDF* mudelit ja väljastab antud sõne ning TF-IDF skoori.

```
['capital', 0.05], ['cards', 0.0], ['carrying', 0.0], ['cash', 0.02
], ['central', 0.0], ['certain', 0.0], ['changes', 0.0], ['claims',
0.0], ['closing', 0.0], ['code', 0.0], ['com', 0.0], ['commercial',
0.0], ['commission', 0.0], ['commissions', 0.0], ['common', 0.0], [
'companies', 0.01], ['company', 0.01]
```

Nüüd, kui meil on olemas GloVE eeltreenitud mudeli koosinussarnasused, andmesõnastik ja TF-IDF skoorid, saame implementeerida meie andmestikul põhineva sõnede sarnasuse maatriksi, kasutame selleks Gensimi *SparseTermSimilarityMatrix* implementatsiooni, kuhu anname parameetriteks eeltreenitud mudeli sarnasused, andmesõnastiku ja TF-IDF skoorid. Tulemuseks on meie puhastatud andmete ja eeltreenitud mudeli sõnede hõre maatriks, mis sisaldab sarnasusi sõnapaaride vahel. Hõredaks maatriksiks(*ingl sparse matrix*) nimetatakse maatriksit, mille ridades ja veergudes on üksikuid nullised erinevaid arve [44].

```
matrix([[1.        , 0.        , 0.        , ..., 0.        , 0.        ,
         0.        ],
        [0.        , 1.        , 0.        , ..., 0.        , 0.        ,
         0.        ],
        [0.        , 0.        , 1.        , ..., 0.        , 0.        ,
         0.        ],
        ...,
        [0.        , 0.        , 0.        , ..., 1.        , 0.44220084,
         0.4879511 ],
        [0.        , 0.        , 0.        , ..., 0.44220084, 1.        ,
         0.5966322 ],
        [0.        , 0.        , 0.        , ..., 0.4879511 , 0.5966322 ,
```

Joonis 17 Hõre maatriks

Allikas: Autor

Viimase sammuna ehitame valmis dokumentide maatriksi kasutades Gensimi *SoftCosineSimilarity* implementatsiooni, mis võrreldes tavalise koosinussarnasusega, võrdleb ka sõnade sarnasusi. Meetod tagastab iga dokumendi kohta *SoftCosine* sarnasusskoori vastava defineeritud päringuga.

```
([0.27431798, 0.17771125, 0.28758943, 0.287723  , 0.28371674,
       0.2805964 , 0.2858753 , 0.27897158, 0.29564533, 0.28451794,
       0.2805964 , 0.2858753 , 0.27897155, 0.29564533, 0.28451794,
       0.2858753 , 0.28918618, 0.2918557 , 0.28451797, 0.28451797])
```

Sarnasusskoor väärtusega 0 tähendab, et vektorid on omavahel risti ja need ei ole sarnas ed. Mida lähedamal koosinussarnasus on 1-le, seda väiksem on vektorite vaheline nurk n ing seda suurem on sarnasusskoor.

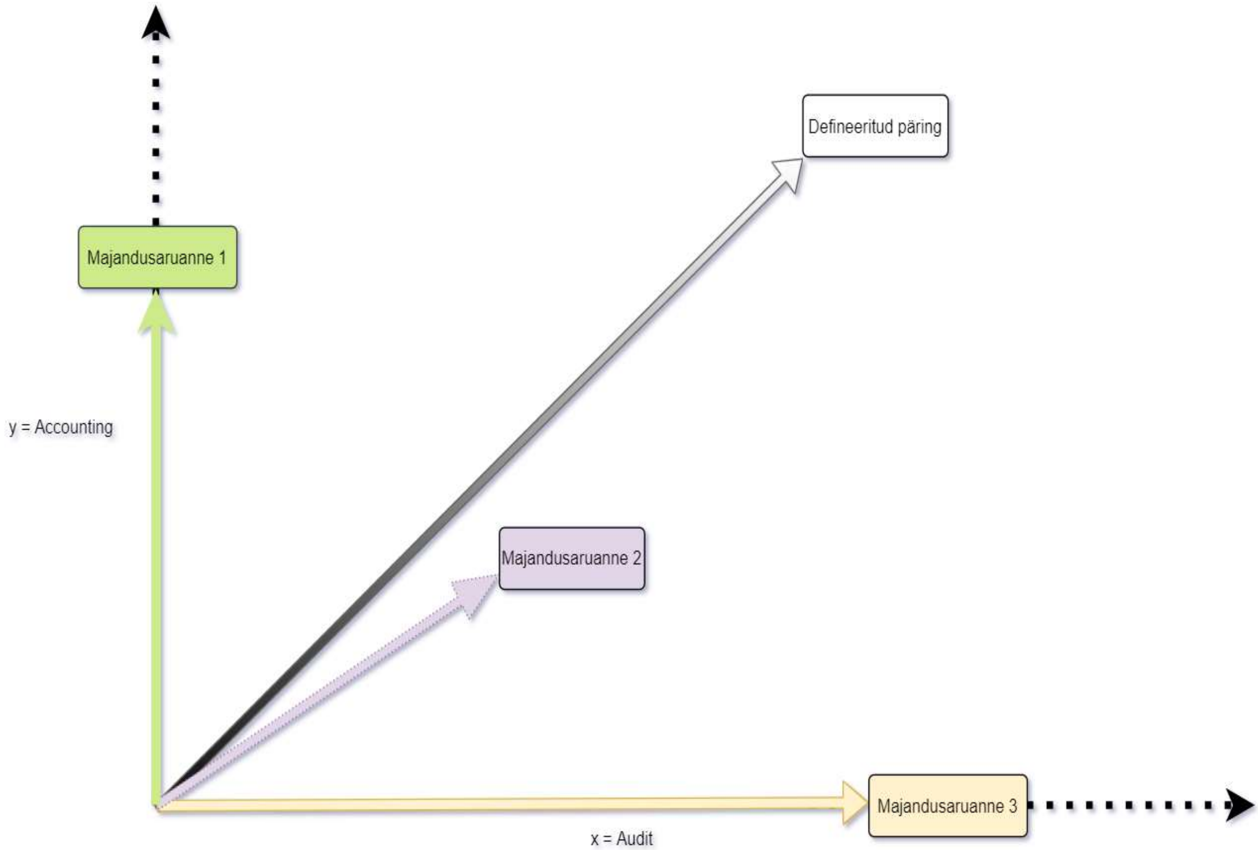
Illustreerimaks tulemusi, panin kõik tulemused allpool väljatoodud andmetabelisse.

Document no	Similarity score with query	Document name
14	0.882082	interim_report_of_q3_2016.txt
9	0.852136	interim_report_of_q1_2016.txt
10	0.842506	interim_report_of_q1_2017.txt
17	0.839587	interim_report_of_q4_2016.txt
6	0.834516	Interim_Report_of_q4_2017.txt
15	0.828033	interim_report_of_q3_2017.txt
12	0.821418	interim_report_of_q2_2016.txt
8	0.685558	Interim_Report_of_q4_2020.txt
3	0.632253	Interim_Report_of_q2_2020.txt
0	0.596540	Interim_Report_of_Q1_2020.txt
5	0.578510	Interim_Report_of_q3_2020.txt
11	0.569276	interim_report_of_q1_2018.txt
13	0.567608	interim_report_of_q2_2018.txt
16	0.567608	interim_report_of_q3_2018.txt
18	0.546087	interim_report_of_q4_2018.txt
1	0.525103	Interim_Report_of_q1_2019.txt
2	0.522308	Interim_Report_of_q2_2019.txt
4	0.512747	Interim_Report_of_q3_2019.txt
7	0.502748	Interim_Report_of_q4_2019.txt

Joonis 18 Sarnasusskoorid dokumenditi

Allikas: Autor

Allpool oleval joonisel on näha, kuidas mudel võrdleb majandusaruandeid etteantud päringuga ning kuidas need võivad asetseda ruumis. Mida lähedamal on majandusaruanne defineeritud päringule, seda kõrgem sarnasusskoor.



Joonis 19 Majandusaruanded vektorruumis koos defineeritud päringuga 2D vaates

Allikas: Autor

4.6 Sarnaseimad sõnad

Kui mudel on arvutanud välja sarnasusskoori dokumentidele, on meil võimalik leida ka hõreda sarnasusmaatriksi abil etteantud terminitele enim sarnaseimad sõnad dokumendis.

Selleks võtame etteantud päringu sõnade id andmesõnastikust ja võrdleme igat dokumendi sõna etteantud päringu sõnaga sarnasusmaatriksis. Kui sarnasusskoor on suurem kui 0, väljastame sarnase sõna. Etteantud päringuna kasutame ühingujuhtimise termineid [45]: *audit and control, board structure, remuneration, shareholder rights, transparency and performance.*

Illustreerimaks antud tulemust väljastan eelmises peatükis leitud sarnasusskooride andmetabelist kõige sarnasema dokumendi kohta 15 kõige sarnasemat sõna:

---

14 0.882 interim_report_of_q3_2016.txt : audit, accounting, auditor, commission, audited, review, disclosure, pricewaterhousecoopers, report, board, concluded, assessment, management, commissions, advisory

---

5 Tulemused

Antud töös uurisime Eesti pankade (Swedbank, LHV, SEB ja Luminori) panust ESG valdkonda otse uuritavate majandusaruannetest, võrreldes majandusaruandeid etteantud ESG päringutega vektorruumis ning kalkuleerides vektoritevahelisi sarnasusi. Sarnasusskooridega saan interpreteerida majandusaruande tähtsust ehk kui majandusaruanne on sarnane etteantud päringuga siis see on indikaator, et antud majandusaruandes tuuakse välja antud valdkonnaga sarnaseid termineid ning arutletakse nende üle.

Analüüs ja uuring viidi läbi uuritavate pankade 2016-2020. aasta inglise keelsete majandusaruannetega, v.a Luminor, kuna Luminoril oli ainult aastast 2018 aruanded saadaval.

Antud analüüsis kasutati väiksemat osa kõikidest turul olevatest ettevõtetest selle tõttu, et võrrelda ainult ühte sektorit korraga. Samuti, et valideerida mudeli tulemusi vastu Estwatchi tehtud uuringut, kus analüüsiti identseid ettevõtteid.

Panuste hindamiseks kasutati defineeritud päringuna termineid [45]:

Keskkondlikud terminid	Sotsiaalsed terminid	Ühingujuhtimise terminid
Biofuels	Access to medicines	Audit and control
Climate	HIV	Board structure
Emissions	AIDs	Remuneration
Land	Nutrition	Shareholder rights
Biodiversity	Product safety	Transparency and Performance
Water	Community relations	
Environmental standards	Privacy and free expression	
Pollution	Security	
Supply	Weak governance zones	
Waste	Diversity	
Recycling	Health and safety	
	ILO core conventions	
	Supply chain labor standards	
	Bribery and corruption	
	Political influence	
	Responsible marketing	
	Whistle-blowing systems	
	disclosure and reporting	
	Governance of sustainability issues	
	Stakeholder engagement	
	UNGC compliance	

Tabel 1 Päringute terminid

5.1 Mudeli tulemused

Eesti pangad arvestavad oma tegevustes ESG faktoreid vähesel määral. Vaadates allpool tulemusi, on ESG faktoritega arvestamisel kõige eeskujulikum LHV pank ning teisel kohal Luminor. SEB ja Swedbank jäävad suurpankadest sarnasusskoori põhjal tahapoole.

Allpool joonisel on näha, et SEB 2016. aasta kvartaliaruannetes on olnud rohkem sarnasusi, kui hilisemastes aruannetes. 2017-2018. aasta kvartaliaruanded, kas kajastasid vähem kui eelnevad aastad või siis antud aastatel ei olnud ESG temaatikaga seotud informatsiooni.

Joonis 20 SEB ESG sarnasusskoorid 2016-2020

Allikas: Autor

Swedbank on ühingujuhtimise teemadel kajastanud informatsiooni stabiilselt sarnasususskooriga 60%, aga sotsiaalsetel ja keskkondlikel teemadel pigem ebapiisavalt, st, et sarnasusskoor on alla poole.

Joonis 21 Swedbank ESG sarnasusskoorid 2016-2020

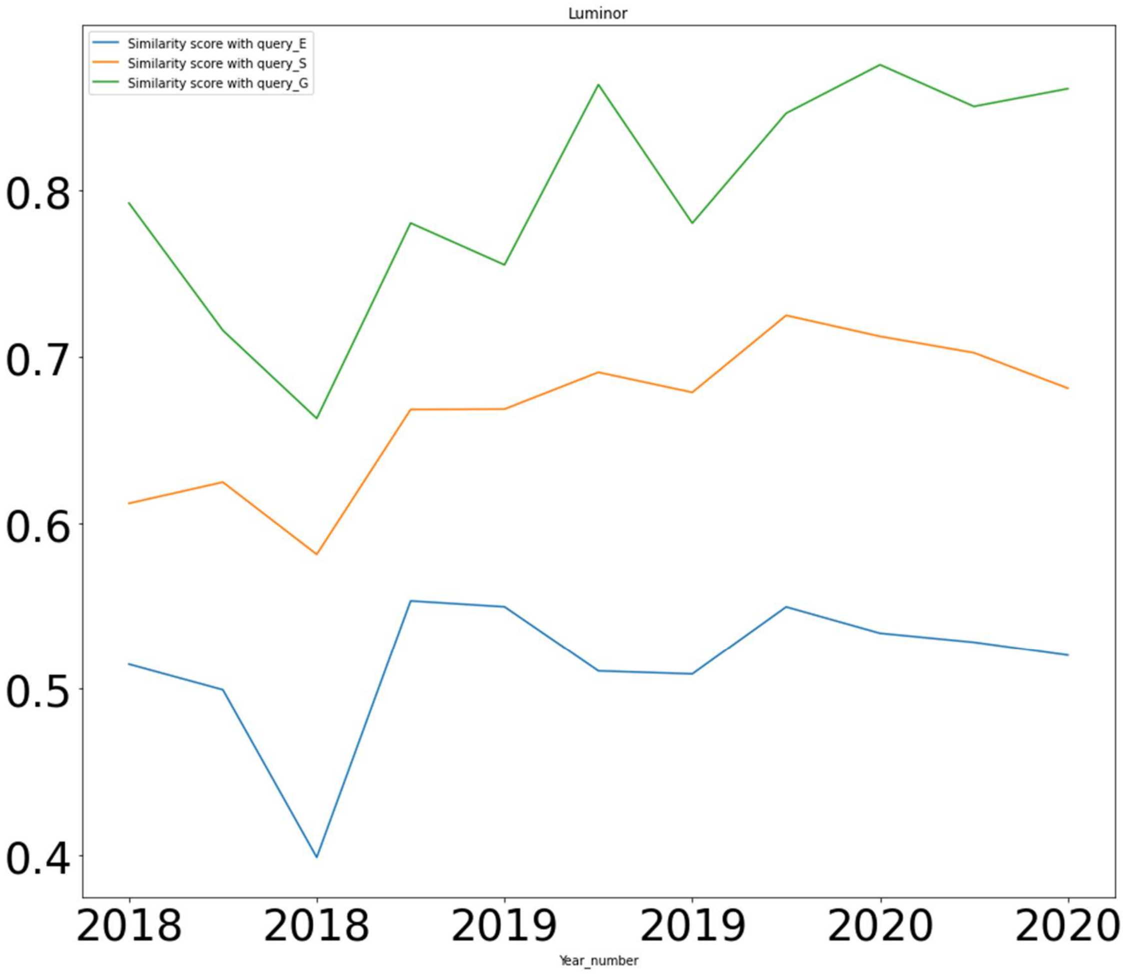
Allikas: Autor

LHV puhul on näha, et ühingujuhtimise ja sotsiaalsetel teemadel kajastavad informatsiooni stabiilselt ja hästi, kuid vajaka jääb keskkonna temaatika. On näha, et alates 2017. aastast on LHV kas ebapiisavalt või üldse mitte rääkinud keskkonna faktoritest.

Joonis 22 LHV ESG sarnasusskoorid 2016-2020

Allikas: Autor

Luminor panga kvartaliaruannetes leidub kajastusi palju ühingujuhtimise temaatikatel, aga pigem vähe on näha keskkondlikel teemadel informatsiooni. Keskkonna valdkonna sarnasusskoor Luminor panga kvartaliaruannete dokumentidega on olnud 50% ringis, mis on pigem ebapiisav.



Joonis 23 Luminor ESG sarnasusskoorid 2018-2020

Allikas: Autor

5.1.1 Keskkondlik faktor

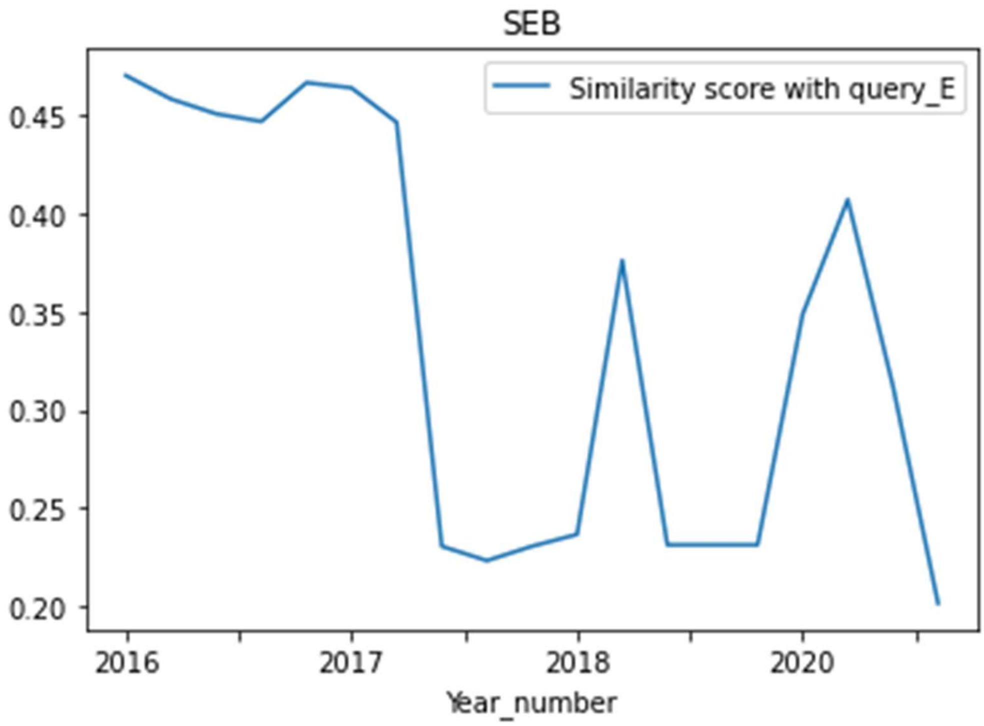
5.1.1.1 SEB

SEB panga keskkondliku faktori sarnasus on kõige tugevam 2016. ja 2017. kvartaliaruannetes ja on olemas sarnaseid sõnu meie defineeritud päringuga. Tabelis on ka näha, et mõnel aruandel ei ole leitud sarnaseid sõnu, aga sarnasus on suurem kui nendel, millel on toodud välja sarnased sõnad. See on tingitud pehme koosinuse loogikast, kus sarnasus leitakse eeltreenitud mudeli sõnade vahel ja kalkuleeritakse sarnasused kokku ning suuremat rolli mängib sõnade arvukus.

Similarity score with query	Document name	Most similar words
0.470402	interim_report_of_q4_2016.txt	sustainable, energy, plants
0.466789	interim_report_of_q1_2017.txt	energy, warming, climate, plants
0.464153	Interim_Report_of_q4_2017.txt	sustainable, energy, plants
0.458451	interim_report_of_q3_2016.txt	fuel, consumption, energy, plants
0.450913	interim_report_of_q2_2016.txt	consumption, energy, plants, efficient
0.447018	interim_report_of_q1_2016.txt	sustainable, energy, alternatives, plants
0.446620	interim_report_of_q3_2017.txt	fuel, consumption, energy, plants
0.407284	Interim_Report_of_q4_2020.txt	sustainable, environmentally
0.376341	Interim_Report_of_q4_2019.txt	
0.348884	Interim_Report_of_Q1_2020.txt	
0.312281	Interim_Report_of_q2_2020.txt	consumption, efficient
0.236731	interim_report_of_q1_2018.txt	
0.231394	Interim_Report_of_q3_2019.txt	
0.231394	Interim_Report_of_q2_2019.txt	
0.231394	Interim_Report_of_q1_2019.txt	
0.230690	interim_report_of_q2_2018.txt	
0.230690	interim_report_of_q3_2018.txt	
0.223377	interim_report_of_q4_2018.txt	
0.201661	Interim_Report_of_q3_2020.txt	

Joonis 24 SEB Environment päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 25 SEB Environment sarnasusskoori graafik 2016-2020

Allikas: Autor

Nagu antud graafikult näha, on SEB puhul 2016. aastal olnud kajastamine antud teemadel parem, kui lähiminevikus, aastatel 2018-2020.

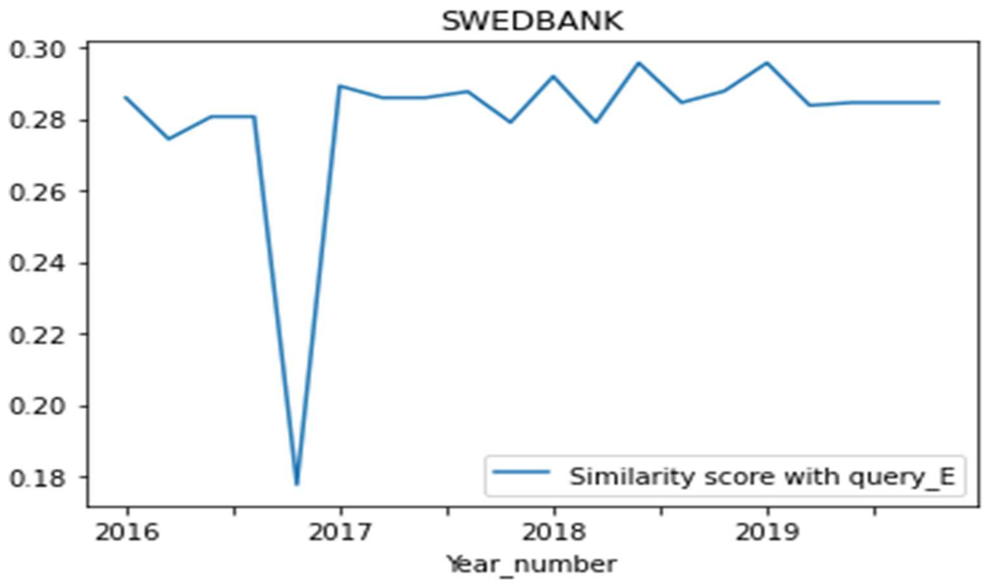
5.1.1.2 Swedbank

Swedbanki kvartaliaruannete sarnasusskoore keskkondlikel faktoritel uurides näeme, et pigem Swedbank raporteerib vähe aruannetes antud teemadel ning sarnasusskoor on alla 30%, mis on kehv. Aruannetes ei kohta ka ühtegi sarnast sõna, mida mudel välja toob, mis tähendab seda, et hõredas sarnasusmaatriksis ei ole ühtegi suurem kui 0 sarnasusskooriga sõna meie etteantud päringuga.

Similarity score with query	Document name	Most similar words
0.295645	Q2_19_eng.txt	
0.295645	Q3_19_eng.txt	
0.291856	Q4_18_eng.txt	
0.289186	Q4_17_eng.txt	
0.287723	Q1_19_eng.txt	
0.287589	Q1_18_eng.txt	
0.285875	Q4_16_eng.txt	
0.285875	Q3_17_eng.txt	
0.285875	Q2_17_eng.txt	
0.284518	Q4_20_eng.txt	
0.284518	Q4_19_eng.txt	
0.284518	Q2_20_eng.txt	
0.284518	Q3_20_eng.txt	
0.283717	Q1_20_eng.txt	
0.280596	Q3_16_eng.txt	
0.280596	Q2_16_eng.txt	
0.278972	Q2_18_eng.txt	
0.278972	Q3_18_eng.txt	
0.274318	Q1_16_eng.txt	
0.177711	Q1_17_eng.txt	

Joonis 26 Swedbank Environment päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 27 Swedbank Environment sarnasusskoori graafik 2016-2020

Allikas: Autor

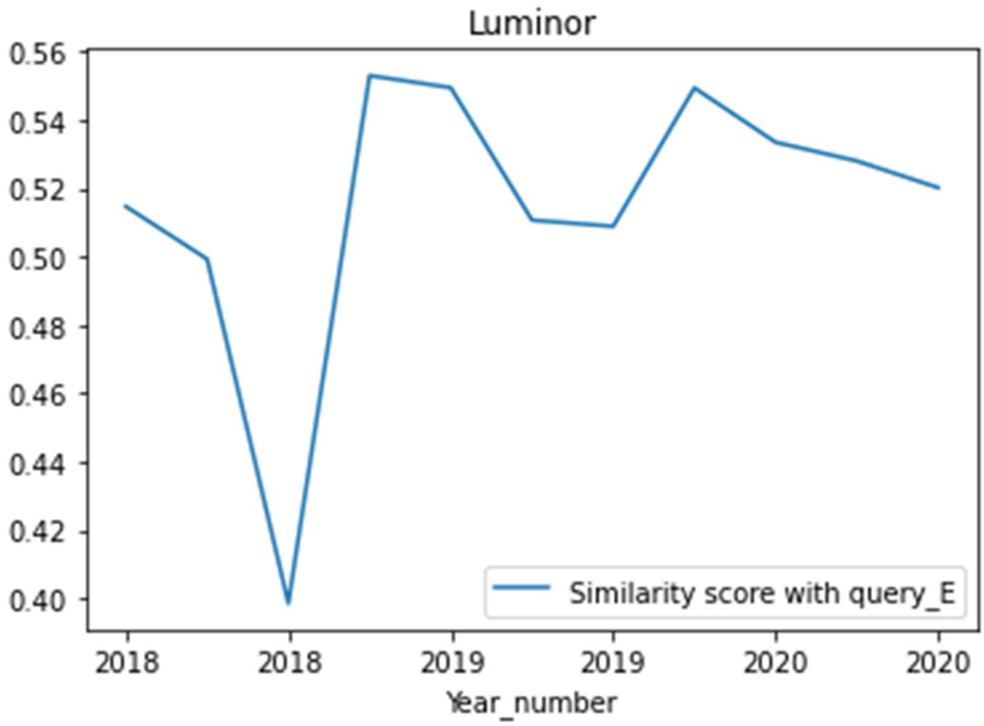
5.1.1.3 Luminor

Luminori kvartaliaruannete põhjal on leidnud ligi 55% sarnasusega 2019. aasta kvartaliaruanded kajastust keskkondlikel teemadel. On olemas ka kõige sarnasemad sõnad, mis sarnananevad töös etteantud päringule.

Similarity score with query_E	Year	Most similar words
0.552983	luminor_q4_2019_interim_report_en.txt	consumption, sustainable, produce, efficiency,...
0.549435	luminor_q3_2019_interim_report_en.txt	renewable, consumption, sustainable, energy, c...
0.549378	luminor_bank_as_interim_report_q4_2020.txt	sustainable, utilization, climate, efficiency,...
0.533469	luminor_bank_as_interim_report_q1_2020.txt	consumption, energy, utilization, efficiency, ...
0.527996	luminor_bank_as_interim_report_q3_2020.txt	energy, utilization, efficient
0.520192	luminor_bank_as_interim_report_q2_2020.txt	consumption, utilization, efficient
0.514686	luminor-ee-4q-report-2018-en.txt	energy, efficiency, efficient
0.510690	luminor_q2_2019_interim_report_eng.txt	energy, produce, efficiency, efficient
0.508860	luminor_q1_2019_interim_report_en.txt	consumption, produce, efficiency, efficient
0.499289	luminor-ee-3q-report-2018-en.txt	
0.398585	luminor-ee-2q-report-2018-en.txt	

Joonis 28 Luminor Environment päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 29 Luminor Environment sarnassuskoori graafik 2018-2020

Allikas: Autor

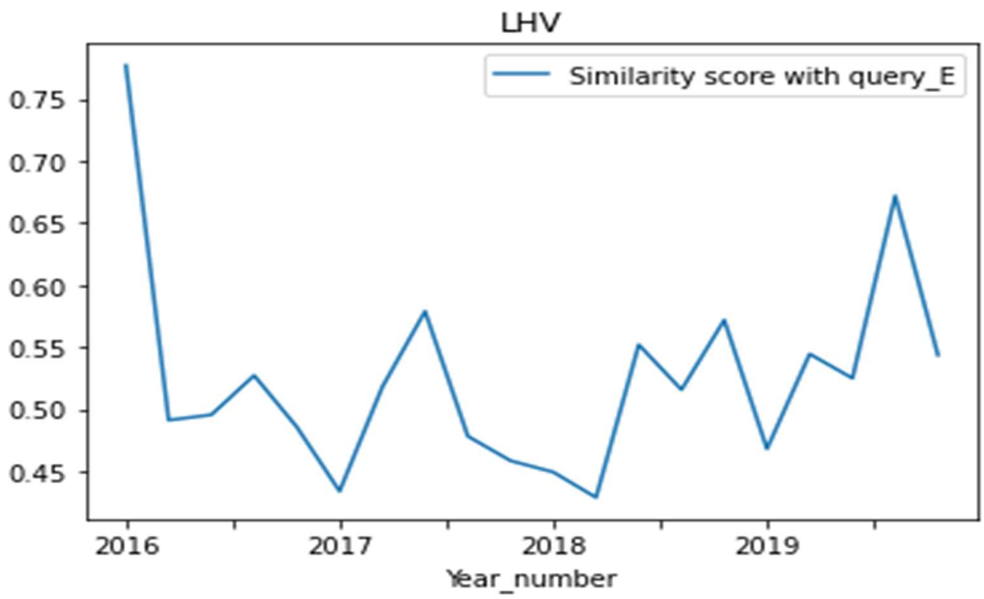
5.1.1.4 LHV

LHV panga puhul on näha kõige suuremat keskkondliku kajastust. Kõige rohkem sarnaneb 2016. aasta 4.kvartali aruanne keskkondlikele terminitele ja järgnevad 2019-2020. aastate aruanded. On näha, et LHV on suhteliselt stabiilne ja tugev kajastaja keskkondlike teemadel aruannetes.

Similarity score with query	Document name	Most similar words
0.776778	2016_q4_en_eur_con_00.txt	consumption, energy, pollution, organic, effic...
0.671864	2020_q4_en_eur_con_00.txt	fuels, fuel, consumption, sustainable, energy,...
0.578893	2017_q4_en_eur_con_00.txt	consumption, sustainable, efficiency
0.571841	2019_q2_en_eur_con_00.txt	fuels, fuel, consumption, sustainable, energy,...
0.551971	2019_q1_en_eur_con_00.txt	fuels, consumption, sustainable, energy, expor...
0.544509	2020_q2_en_eur_con_00.txt	fuel, consumption, sustainable, energy, climat...
0.543637	2020_q1_en_eur_con_00.txt	consumption, sustainable, energy, efficiency
0.527233	2016_q3_en_eur_con_00.txt	consumption, energy, efficiency
0.524971	2020_q3_en_eur_con_00.txt	fuels, consumption, sustainable, energy, expor...
0.518005	2017_q2_en_eur_con_00.txt	consumption, sustainable, energy, efficiency, ...
0.515736	2019_q3_en_eur_con_00.txt	sustainable, energy, produce, efficient
0.495547	2016_q1_en_eur_con_00.txt	consumption, energy
0.491215	2016_q2_en_eur_con_00.txt	consumption, energy
0.485669	2017_q3_en_eur_con_00.txt	consumption, sustainable, produce, efficiency,...
0.478287	2018_q4_en_eur_con_00.txt	consumption, sustainable, energy, climate, org...
0.468106	2019_q4_en_eur_con_00.txt	consumption, sustainable, energy, produce
0.458542	2018_q2_en_eur_con_00.txt	consumption, sustainable, energy, dependence, ...
0.449400	2018_q3_en_eur_con_00.txt	consumption, sustainable, energy, exporting, c...
0.434006	2017_q1_en_eur_con_00.txt	consumption, energy
0.428953	2018_q1_en_eur_con_00.txt	consumption, sustainable

Joonis 30 LHV Environment päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 31 LHV Environment sarnasusskoori graafik 2016-2020

Allikas: Autor

5.1.2 Sotsiaalne faktor

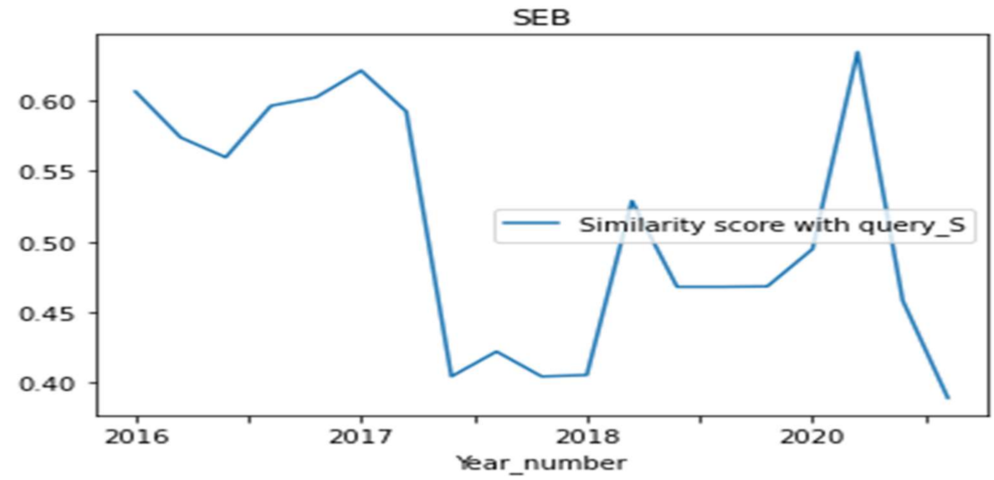
5.1.2.1 SEB

SEB panga sotsiaalse faktori sarnasus on kõige tugevam 2020. aasta 4. kvartaliaruandes ning seda on näha ka allpool väljatoodud tabelis, kus on välja genereeritud ka antud dokumendi kõige sarnasemad sõnad. On näha, et 2016. ja 2017. aasta kvartaliaruannetes on sotsiaalsetel teemadel kõige rohkem räägitud ja on olemas sarnaseid ning täpselt vastavaid sõnu meie defineeritud päringuga.

Similarity score with query	Document name	Most similar words
0.634497	Interim_Report_of_q4_2020.txt	providing, provides, services, internet, provi...
0.621346	Interim_Report_of_q4_2017.txt	access, allows, provides, users, enable, servi...
0.606426	interim_report_of_q4_2016.txt	services, internet, information, available, se...
0.602320	interim_report_of_q1_2017.txt	allows, providers, users, enabling, services, ...
0.596348	interim_report_of_q1_2016.txt	provide, providing, services, internet, inform...
0.592273	interim_report_of_q3_2017.txt	allow, services, internet, information, limite...
0.573902	interim_report_of_q3_2016.txt	enable, services, internet, information, enabl...
0.559815	interim_report_of_q2_2016.txt	existing, services, internet, information, lim...
0.528935	Interim_Report_of_q4_2019.txt	provides, services, internet, information, ser...
0.494767	Interim_Report_of_Q1_2020.txt	provide, providing, provides, services, intern...
0.468358	Interim_Report_of_q1_2019.txt	provides, services, internet, information, ava...
0.468045	Interim_Report_of_q3_2019.txt	provides, services, internet, information, ava...
0.468045	Interim_Report_of_q2_2019.txt	provides, services, internet, information, ava...
0.458341	Interim_Report_of_q2_2020.txt	providing, provides, services, internet, infor...
0.422210	interim_report_of_q4_2018.txt	provides, services, internet, information, ava...
0.405543	interim_report_of_q1_2018.txt	provides, services, internet, information, ava...
0.404504	interim_report_of_q2_2018.txt	provides, services, internet, information, ava...
0.404504	interim_report_of_q3_2018.txt	provides, services, internet, information, ava...
0.389319	Interim_Report_of_q3_2020.txt	providing, provides, services, internet, provi...

Joonis 32 SEB Social päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 33 SEB Social sarnasusskoori graafik 2016-2020

Allikas: Autor

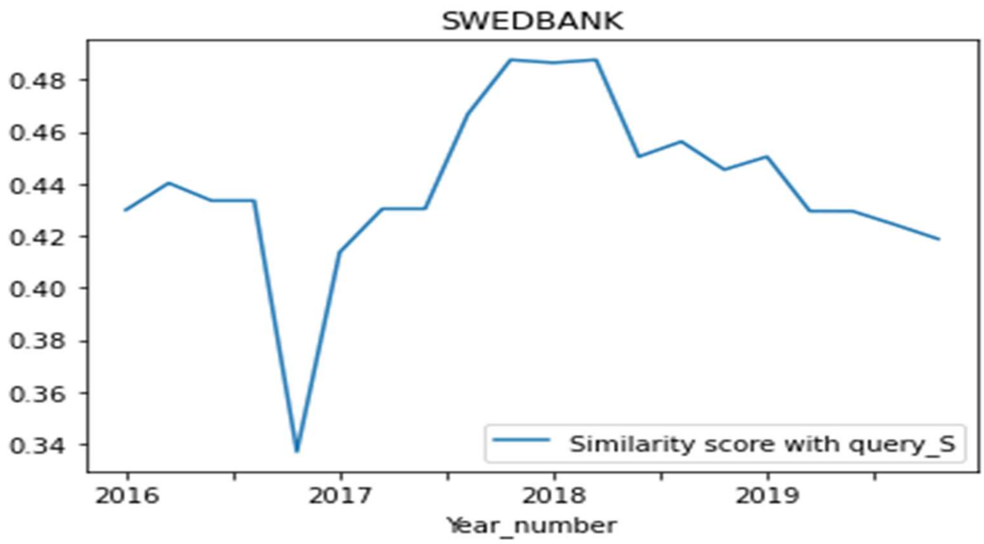
5.1.2.2 Swedbank

Swedbank üleüldine meelestatus sotsiaalsetel teemadel kvartaliaruannetes on pigem kehvapoolne. See tähendab, et nende kvartaliaruannete sarnassuskoor sotsiaalsete terminitega on alla poole. Kõige parem sarnasus esineb 2016. aasta kvartaliaruannetes, kus sarnassus on ligi 49%.

Similarity score with query	Document name	Most similar words
0.487676	Q3_18_eng.txt	services, direct, information, application, re...
0.487676	Q2_18_eng.txt	services, direct, information, application, re...
0.486540	Q4_18_eng.txt	services, direct, information, application, re...
0.466762	Q1_18_eng.txt	services, direct, information, maintains, appl...
0.456246	Q4_19_eng.txt	services, direct, information, required, comme...
0.450428	Q3_19_eng.txt	services, direct, information, required, comme...
0.450428	Q2_19_eng.txt	services, direct, information, required, comme...
0.445427	Q1_19_eng.txt	services, direct, information, required, comme...
0.440285	Q1_16_eng.txt	services, direct, information, maintains, requ...
0.433521	Q2_16_eng.txt	services, direct, information, maintains, requ...
0.433521	Q3_16_eng.txt	services, direct, information, maintains, requ...
0.430377	Q2_17_eng.txt	services, direct, information, maintains, requ...
0.430377	Q3_17_eng.txt	services, direct, information, maintains, requ...
0.429954	Q4_16_eng.txt	services, direct, information, maintains, requ...
0.429499	Q1_20_eng.txt	services, direct, information, required, comme...
0.429465	Q2_20_eng.txt	services, direct, information, required, comme...
0.424254	Q3_20_eng.txt	services, direct, information, required, comme...
0.418778	Q4_20_eng.txt	services, direct, information, required, comme...
0.413624	Q4_17_eng.txt	services, direct, information, maintains, requ...
0.336972	Q1_17_eng.txt	services, direct, information, maintains, requ...

Joonis 34 Swedbank Social päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 35 Swedbank Social sarnasusskoori graafik 2016-2020

Allikas: Autor

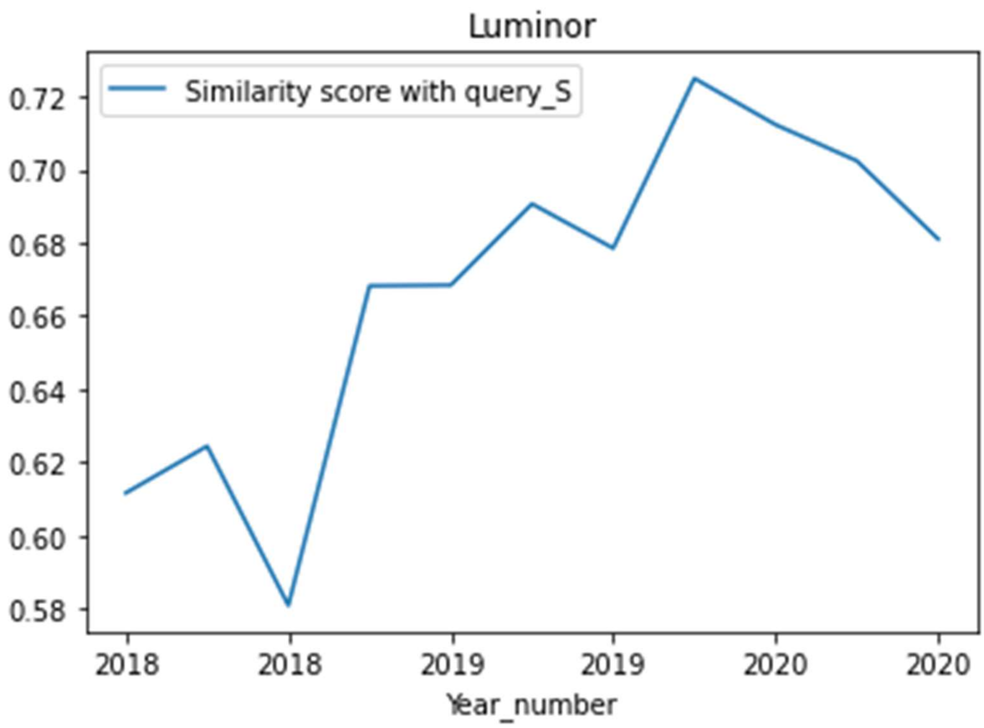
5.1.2.3 Luminor

Luminori sotsiaalne sarnasus kvartaliaruannetes on kõvasti üle poole, st, et eelmise aaasta aruannete põhjal genereeriti sarnasusskoor ligi 72%. Luminori puhul on selgelt näha, et mida aasta edasi, seda rohkem on hakatud sotsiaalsetele teemadele rõhku panema ja nendest rohkem rääkima oma kvartaliaruannetes.

Similarity score with query_S	Year	Most similar words
0.724868	luminor_bank_as_interim_report_q4_2020.txt	access, allows, provides, providers, existing,...
0.712235	luminor_bank_as_interim_report_q1_2020.txt	access, provides, providers, existing, enable,...
0.702376	luminor_bank_as_interim_report_q3_2020.txt	access, provides, providers, existing, users, ...
0.690629	luminor_q2_2019_interim_report_eng.txt	access, provides, providers, existing, secure,...
0.681038	luminor_bank_as_interim_report_q2_2020.txt	access, allows, provides, providers, existing,...
0.678521	luminor_q1_2019_interim_report_en.txt	allows, provides, providers, existing, secure,...
0.668452	luminor_q3_2019_interim_report_en.txt	access, provides, providers, existing, users, ...
0.668242	luminor_q4_2019_interim_report_en.txt	access, provides, providers, existing, allowin...
0.624491	luminor-ee-3q-report-2018-en.txt	provides, services, direct, information, enabl...
0.611732	luminor-ee-4q-report-2018-en.txt	provides, services, information, enables, limi...
0.581026	luminor-ee-2q-report-2018-en.txt	services, direct, information, limited, availa...

Joonis 36 Luminor Social päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 37 Luminor Social sarnasusskoori graafik 2018-2020

Allikas: Autor

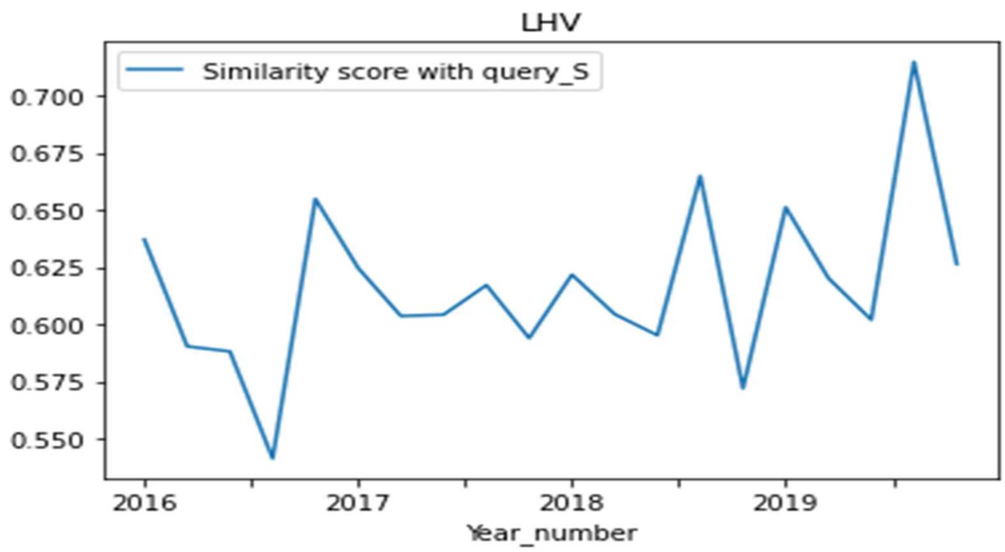
5.1.2.4 LHV

LHV panga puhul on näha, et kvartaliaruannete sotsiaalne sarnasusskoor on stabiilselt tõusnud aastatega. Kõige enam sarnaneb sotsiaalsetele terminitele 2020. aasta 4. kvartali aruanne, 71% sarnasusega.

Similarity score with query	Document name	Most similar words
0.714735	2020_q4_en_eur_con_00.txt	access, allows, provides, existing, services, ...
0.664909	2019_q3_en_eur_con_00.txt	access, allows, provides, providers, existing,...
0.654845	2017_q3_en_eur_con_00.txt	access, provides, providers, existing, link, a...
0.651190	2019_q4_en_eur_con_00.txt	access, allows, provides, existing, link, allo...
0.637056	2016_q4_en_eur_con_00.txt	access, allows, provides, existing, allowing, ...
0.626436	2020_q1_en_eur_con_00.txt	access, allows, provides, existing, secure, al...
0.624654	2017_q1_en_eur_con_00.txt	access, provides, providers, existing, service...
0.621746	2018_q3_en_eur_con_00.txt	access, provides, providers, existing, service...
0.620183	2020_q2_en_eur_con_00.txt	access, allows, provides, existing, secure, al...
0.617179	2018_q4_en_eur_con_00.txt	access, provides, existing, allowing, services...
0.604478	2018_q1_en_eur_con_00.txt	access, allows, provides, existing, allowing, ...
0.604275	2017_q4_en_eur_con_00.txt	access, allows, provides, existing, secure, se...
0.603678	2017_q2_en_eur_con_00.txt	access, provides, existing, services, internet...
0.602055	2020_q3_en_eur_con_00.txt	access, allows, provides, existing, users, ser...
0.595246	2019_q1_en_eur_con_00.txt	access, provides, services, information, enabl...
0.593995	2018_q2_en_eur_con_00.txt	access, provides, existing, services, informat...
0.590424	2016_q2_en_eur_con_00.txt	access, allows, provides, providers, services,...
0.588200	2016_q1_en_eur_con_00.txt	access, allows, provides, providers, services,...
0.572083	2019_q2_en_eur_con_00.txt	access, provides, existing, enable, services, ...
0.541450	2016_q3_en_eur_con_00.txt	access, provides, existing, allowing, services...

Joonis 38 LHV Social päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 39 LHV Social sarnasusskoori graafik 2016-2020

Allikas: Autor

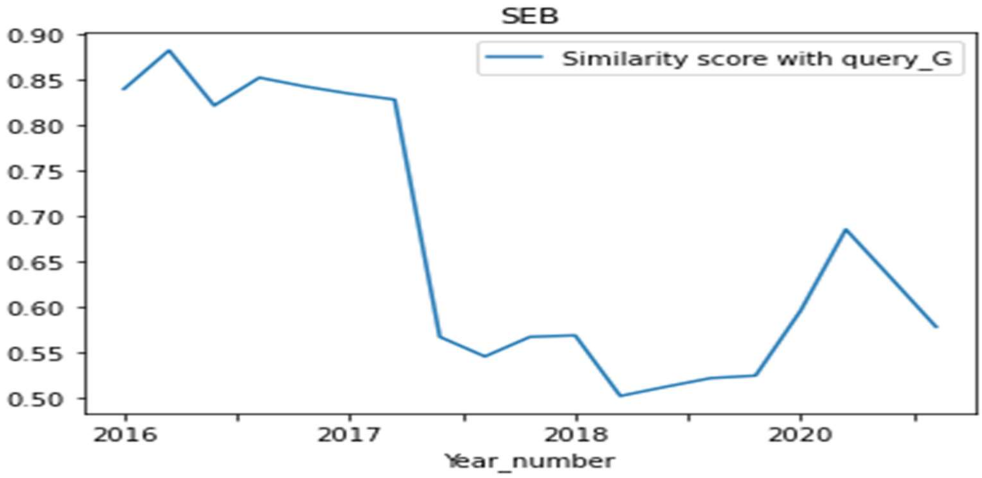
5.1.3 Ühingujuhtimise faktor

5.1.3.1 SEB

SEB panga ühingujuhtimise faktori sarnasus on kõige tugevam 2016. aasta kvartaliaruannetes ning seda on näha ka allpool väljatoodud tabelis, kus on välja genereeritud ka antud dokumendi kõige sarnasemad sõnad. On näha, et 2016. ja 2017. aasta kvartaliaruannetes on ühingujuhtimiste teemadel kõige rohkem räägitud ja on olemas sarnaseid ja täpselt vastavaid sõnu meie defineeritud päringuga. Kõige sarnasemaks dokumendiks sai mudeli järgi 2016. 3. kvartali aruanne, mis sarnanes 89% etteantud päringule.

Similarity score with query	Document name	Most similar words
0.882082	interim_report_of_q3_2016.txt	audit, accounting, auditor, commission, audite...
0.852136	interim_report_of_q1_2016.txt	audit, accounting, auditor, commission, audite...
0.842506	interim_report_of_q1_2017.txt	audit, accounting, auditor, commission, audite...
0.839587	interim_report_of_q4_2016.txt	audit, accounting, auditor, commission, audite...
0.834516	Interim_Report_of_q4_2017.txt	audit, accounting, auditor, commission, audite...
0.828033	interim_report_of_q3_2017.txt	audit, accounting, auditor, commission, audite...
0.821418	interim_report_of_q2_2016.txt	audit, accounting, auditor, commission, audite...
0.685558	Interim_Report_of_q4_2020.txt	audit, accounting, auditor, commission, audite...
0.632253	Interim_Report_of_q2_2020.txt	audit, accounting, auditor, commission, audite...
0.596540	Interim_Report_of_Q1_2020.txt	audit, accounting, auditor, commission, audite...
0.578510	Interim_Report_of_q3_2020.txt	audit, accounting, auditor, commission, audite...
0.569276	interim_report_of_q1_2018.txt	audit, accounting, auditor, commission, audite...
0.567608	interim_report_of_q2_2018.txt	audit, accounting, auditor, commission, audite...
0.567608	interim_report_of_q3_2018.txt	audit, accounting, auditor, commission, audite...
0.546087	interim_report_of_q4_2018.txt	audit, accounting, auditor, commission, audite...
0.525103	Interim_Report_of_q1_2019.txt	audit, accounting, auditor, commission, audite...
0.522308	Interim_Report_of_q2_2019.txt	audit, accounting, auditor, commission, audite...
0.512747	Interim_Report_of_q3_2019.txt	audit, accounting, auditor, commission, audite...
0.502748	Interim_Report_of_q4_2019.txt	audit, accounting, auditor, commission, audite...

Joonis 40 SEB Governance päringu sarnasusskoorid ja sarnaseimad sõnad
Allikas: Autor



Joonis 41 SEB Governance sarnasusskoori graafik 2016-2020

Allikas: Autor

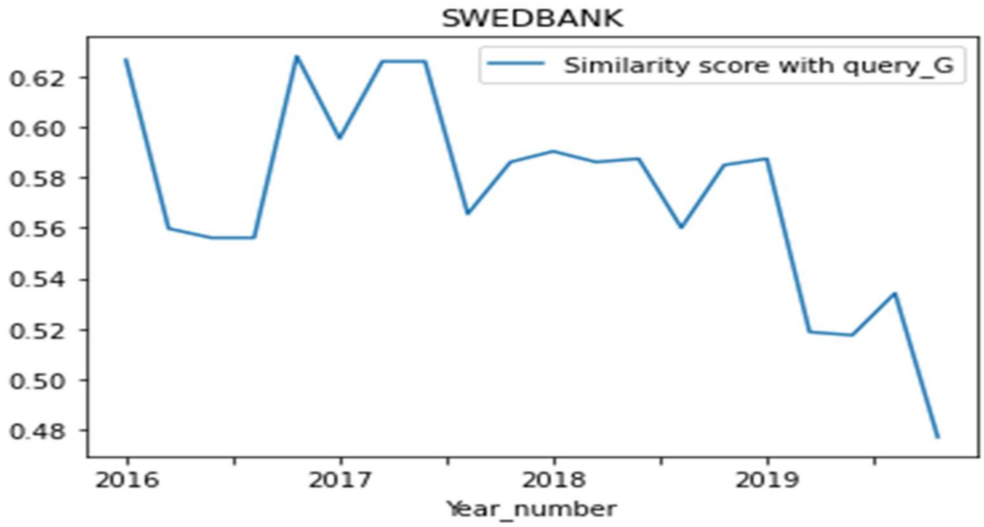
5.1.3.2 Swedbank

Swedbanki ühingujuhtimise faktori sarnasus on kõige tugevam 2017. aasta 1.kvartali aruandes ning seda on näha ka allpool väljatoodud tabelis, kus on välja genereeritud ka antud dokumendi kõige sarnasemad sõnad. Kõige sarnasemaks dokumendiks sai mudeli järgi 2017. aasta 1. Kvartali aruanne, mis sarnanes 62% etteantud päringule.

Similarity score with query	Document name	Most similar words
0.628034	Q1_17_eng.txt	audit, accounting, auditor, commission, audite...
0.626585	Q4_16_eng.txt	audit, accounting, auditor, commission, audite...
0.625939	Q2_17_eng.txt	audit, accounting, auditor, commission, audite...
0.625939	Q3_17_eng.txt	audit, accounting, auditor, commission, audite...
0.595405	Q4_17_eng.txt	audit, accounting, auditor, commission, audite...
0.590344	Q4_18_eng.txt	audit, accounting, auditor, commission, audite...
0.587409	Q3_19_eng.txt	accounting, auditor, commission, audited, pric...
0.587409	Q2_19_eng.txt	accounting, auditor, commission, audited, pric...
0.586102	Q3_18_eng.txt	audit, accounting, auditor, commission, audite...
0.586102	Q2_18_eng.txt	audit, accounting, auditor, commission, audite...
0.584930	Q1_19_eng.txt	accounting, auditor, commission, audited, pric...
0.565545	Q1_18_eng.txt	audit, accounting, auditor, commission, audite...
0.560061	Q4_19_eng.txt	accounting, auditor, commission, audited, pric...
0.559760	Q1_16_eng.txt	audit, accounting, auditor, commission, audite...
0.556093	Q2_16_eng.txt	audit, accounting, auditor, commission, audite...
0.556093	Q3_16_eng.txt	audit, accounting, auditor, commission, audite...
0.534182	Q3_20_eng.txt	accounting, auditor, commission, audited, pric...
0.518810	Q1_20_eng.txt	accounting, auditor, commission, audited, pric...
0.517537	Q2_20_eng.txt	accounting, auditor, commission, audited, pric...
0.477185	Q4_20_eng.txt	accounting, auditor, commission, audited, pric...

Joonis 42 Swedbank Governance päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 43 Swedbank Governance sarnasusskoori graafik 2016-2020

Allikas: Autor

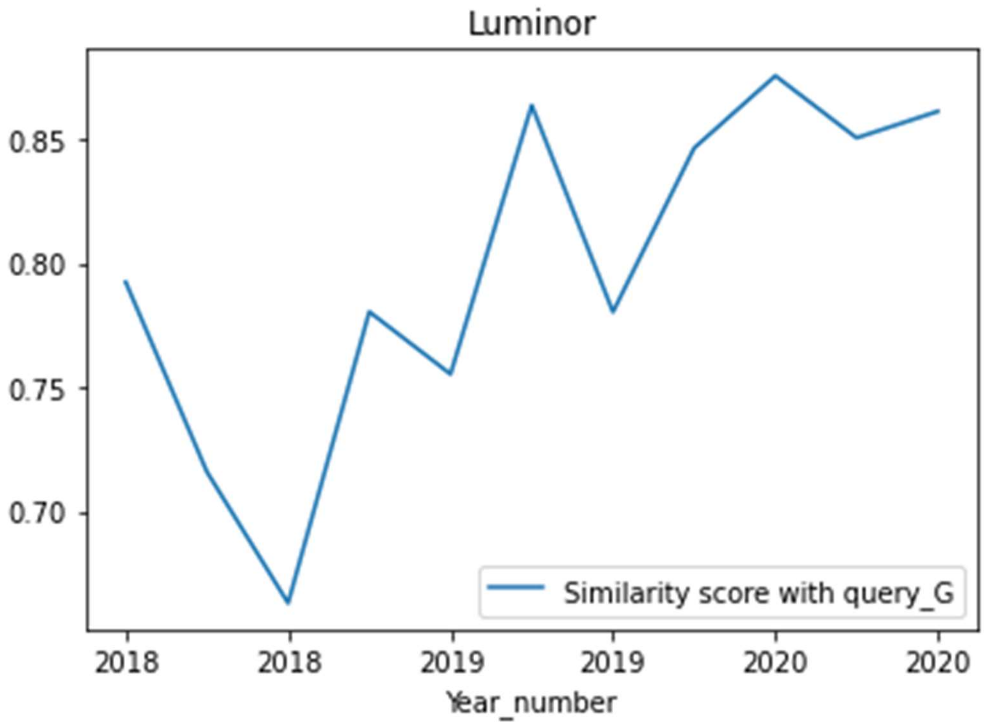
5.1.3.3 Luminor

Luminori ühingujuhtimise faktori sarnasus on kõige tugevam 2020. aasta kvartaliaruannetes ning seda on näha ka allpool väljatoodud tabelis, kus on välja genereeritud ka antud dokumendi kõige sarnasemad sõnad. Kõige sarnasemaks dokumendiks sai mudeli järgi 2020. aasta 1. Kvartali aruanne, mis sarnanes 87% etteantud päringule.

Similarity score with query_G	Year	Most similar words
0.875716	luminor_bank_as_interim_report_q1_2020.txt	audit, auditing, accounting, oversight, audito...
0.863801	luminor_q2_2019_interim_report_eng.txt	audit, auditors, auditing, accounting, auditor...
0.861340	luminor_bank_as_interim_report_q2_2020.txt	audit, auditors, auditing, accounting, oversig...
0.850645	luminor_bank_as_interim_report_q3_2020.txt	audit, accounting, oversight, commission, audi...
0.846555	luminor_bank_as_interim_report_q4_2020.txt	audit, accounting, oversight, commission, audi...
0.792441	luminor-ee-4q-report-2018-en.txt	audit, accounting, commission, disclosure, reg...
0.780445	luminor_q4_2019_interim_report_en.txt	accounting, reviewing, commission, review, dis...
0.780367	luminor_q1_2019_interim_report_en.txt	accounting, commission, review, disclosure, re...
0.755285	luminor_q3_2019_interim_report_en.txt	accounting, reviewing, commission, review, dis...
0.715843	luminor-ee-3q-report-2018-en.txt	accounting, commission, audited, disclosure, r...
0.662836	luminor-ee-2q-report-2018-en.txt	accounting, commission, disclosure, regulatory...

Joonis 44 Luminor Governance päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 45 Luminor Governance sarnasusskoori graafik 2018-2020

Allikas: Autor

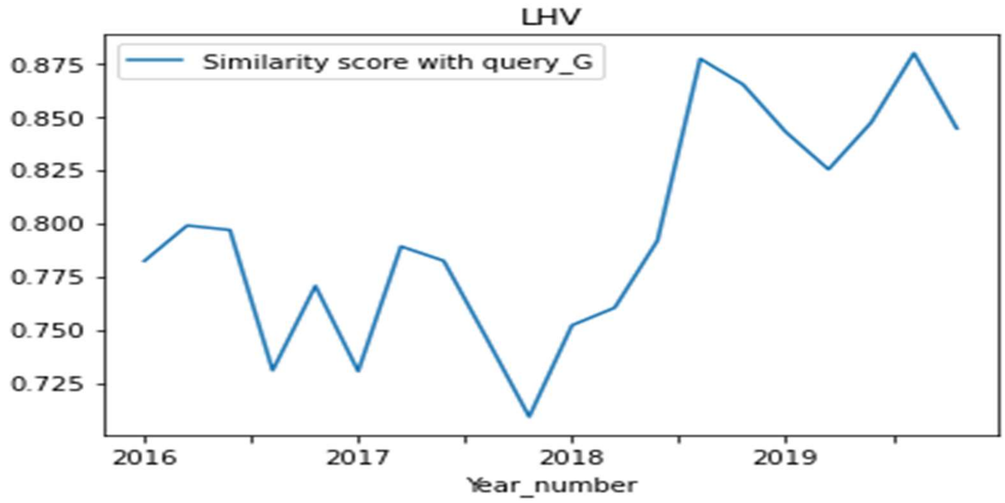
5.1.3.4 LHV

LHV ühingujuhtimise faktori sarnasus on kõikidest pankadest kõige tugevam viimastel aastatel. Kui vaadata antud tabelit, siis LHV 2019-2020. aasta aruanded sarnanesid 87% etteantud ühingujuhtimise päringule. Kõige kõrgema sarnasusskoori sai 2020. aasta 4.kvartali aruanne.

Similarity score with query	Document name	Most similar words
0.879742	2020_q4_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.877110	2019_q3_en_eur_con_00.txt	accounting, commission, audited, disclosure, r...
0.865055	2019_q2_en_eur_con_00.txt	audit, auditors, auditing, accounting, auditor...
0.847203	2020_q3_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.844373	2020_q1_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.842757	2019_q4_en_eur_con_00.txt	accounting, commission, audited, disclosure, r...
0.825171	2020_q2_en_eur_con_00.txt	audit, auditing, accounting, reviewing, commis...
0.798839	2016_q2_en_eur_con_00.txt	audit, accounting, commission, audited, regula...
0.796750	2016_q1_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.791773	2019_q1_en_eur_con_00.txt	accounting, commission, audited, regulatory, r...
0.789043	2017_q2_en_eur_con_00.txt	accounting, reviewing, commission, audited, co...
0.782281	2017_q4_en_eur_con_00.txt	accounting, commission, audited, committee, re...
0.782196	2016_q4_en_eur_con_00.txt	accounting, commission, audited, review, commi...
0.770492	2017_q3_en_eur_con_00.txt	accounting, commission, audited, committee, re...
0.760251	2018_q1_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.752011	2018_q3_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...
0.746069	2018_q4_en_eur_con_00.txt	audit, accounting, commission, audited, review...
0.730839	2016_q3_en_eur_con_00.txt	accounting, commission, audited, review, commi...
0.730447	2017_q1_en_eur_con_00.txt	accounting, reviewing, commission, audited, co...
0.709087	2018_q2_en_eur_con_00.txt	accounting, commission, audited, reviewed, com...

Joonis 46 LHV Governance päringu sarnasusskoorid ja sarnaseimad sõnad

Allikas: Autor



Joonis 47 LHV Governance sarnasusskoori graafik 2016-2020

Allikas: Autor

5.2 Järeldused ja analüüs

Töös ehitatud mudeli ja etteantud päringute tulemused näitavad selgelt, et Eesti pangad kajastavad oma kvartaliaruannetes ESG seotud faktoreid ebapiisavalt. Vaatamata kehvapoolsetele tulemustele on ESG faktorite kajastamises kvartaliaruannetes sarnasusskoori järgi esirindlik Luminor pank ning teisel kohal LHV. Järjestuses kolmandaks tuli SEB ning tema järel Swedbank. Luminori hea positsioon keskmises punktiskooris võib olla ka tingitud sellest, et Luminori aruandeid oli 2 aasta jagu vähem, kui teistel.

Kõigil neljal pangal on olemas oma tugevad faktorid, mida nad raporteerivad sarnasuse järgi rohkem ning peaaegu, et kõigis valdkondades oli näha, et mida aeg edasi, seda paremini ja läbipaistvamalt on ESG temaatikaid aruannetes kajastatud. Alphasense raporti järgi ESG väljaütlemised ja uudised ESG valdkondade kohta on kasvanud järsult alates 2016 aastast.

	Swedbank	SEB	Luminor	LHV
Average similarity score in environment topic	0,27991	0,34034	0,51051	0,52576
Average similarity score in social topic	0,43909	0,51050	0,66760	0,61644
Average similarity score in governance topic	0,57347	0,66877	0,79866	0,79607
Average	1,29	1,52	1,98	1,94

Tabel 2 Eesti pankade ESG keskmised 2016-2020

Näiteks Swedbank ja SEB on asetanud rõhu rohkem rääkida sotsiaalsetest ja ühingujuhtimise teemadest, LHV ja Luminor aga rohkem keskkondlikest ja ühingujuhtimise teemadest.

Kui valideerida mudeli tulemusi vastu inimtehtud uuringut, siis näeme, et mudeli ja manuaalsel hinnangul on mõningad kooskõlad ja samuti ka erisusi.

Estwatchi uuringu eesmärk oli hinnata ja võrrelda Eesti pankasid, kui palju nad ESG temaatikatesse panustavad. Uuring keskendus ka neljale suurimale Eesti pangale ning hindamiseks kasutati säästva panganduse raamistikku (SUSBA), mis kohandati Eesti konteksti.

Uuringu tulemusena selgus, et Eesti pangad ei panusta ESG temaatikatele piisavalt või üldsegi mitte. Kõige suurema punktiskoori sai SEB pank, 34% kogu punktidest ning temale järgnes 21% punktiskooriga Swedbank.

Antud uuringu järgi LHV ja Luminor eiravad uuringukriteeriumite kohaselt oma finantsotsustes ESG temaatikaid.

ESG valdkonnad ja indikaatorid		LHV	Luminor	SEB	Swedbank
Eesmärk	Vastutustundlikkuse strateegia ja osapoolte kaasamine	0.00	0.00	0.83	0.33
	Osalemine vastutustundliku finantssektori edendamises	0.00	0.10	0.40	0.30
Eeskirjad	Üldised vastutustundliku panganduse eeskirjad	0.00	0.00	0.63	0.50
	Valdkondade eeskirjad (metsandus, kliimamuutus, inimõigused, jm)	0.00	0.00	0.44	0.28
Protsessid	ESG riskide hindamine ja haldamine	0.00	0.00	0.13	0.31
	Investeeritavate jälgimine ja mõjutamine vastutustundlikkuse edendamiseks	0.00	0.00	0.57	0.43
Personal	Vastutustundlikkuse protsesside juhtimine ja vastutusalad	0.00	0.00	0.33	0.00
	Personali ESG-alane koolitamine ja tulemuste hindamine	0.00	0.00	0.29	0.00
Tooted	Vastutustundlikud finantstooted ja -teenused	0.13	0.00	0.25	0.00
Portfell	ESG riskide hindamine ja vähendamine portfelli tasandil	0.00	0.00	0.00	0.14
	ESG riskidega seotud ohtude, tegevuste ja eesmärkide avalikustamine	0.00	0.00	0.00	0.00
	Kogu keskmine	0.01	0.01	0.34	0.21

Joonis 48 ESG hinnangud Eesti pankadele.

Allikas: Estwatch

Magistritöö kooskõla antud uuringuga on kindlasti see, et SEB on väga tubli ESG temaatikatega tegeleja, mudeli kriteeriumite põhjal sai SEB pank tugeva keskmise skoori, mis näitab, et SEB on Eesti pankadest tugev eeskuju.

Vastuolu antud uuringuga on see, et Luminor ja LHV on uuringu järgi kõige kehvema keskmise skooriga, kuigi meie mudel andis neile kõige tugevamad keskmised skoorid.

Mudeli tulemuste vastuolu antud uuringuga on kindlasti tingitud ka sellest, et antud töös ehitatud mudel põhineb semantilisel sarnasusel, mis mõõdab, kui sarnane on dokument *Environmental*, *Social* või *Governance* terminitega, mis on masinaga mõõdetud läbi matemaatiliste tehete.

Estwatchi uuringu eripära on see, et antud uuring on koostatud manuaalselt, st, et inimene on käinud ettevõtte esindajatega kohapeal rääkimas ja küsimas küsimusi, mis antud skoori ettevõtetele annavad. Kindlasti on Estwatchi uuringus omajagu subjektiivsust, nagu ka meie mudelis. Mudeli tulemusi ja Estwatchi uuringu tulemusi saab võrrelda ainult lõpptulemustega, st, et kes on kõige parem ESG temaatikatega tegeleja ja kes jääb järjestuses tahapoole.

Antud töö eesmärk oli ehitada valmis mudel, mis mõõdab ESG panust otse ettevõtete aruannetest ja järjestab ettevõtted baseerudes panusele.

Autori arvates on tulemused väga usaldusväärsed, kuna töös kasutatud mudel kasutab aruannetest saadud teksti ja võrdleb seda etteantud päringuga ning kui nende kahe vahel on semantiline sarnasus ja see sarnasus on kõrge, siis saab julgelt väita, et antud aruandes on kajastatud informatsiooni vastava ESG valdkonna kohta.

Töös kasutatud metoodika ei võimaldanud teada saada antud ettevõtete päris ESG skoori, autor kasutas aritmeetilist keskmist ettevõtete keskmiste skooride leidmiseks, et panna firmad pingeritta.

Autor usub, et antud mudelist oleks palju abi just terviku ESG skoori leidmisel, kuna ESG hindamisprotsessid põhinevad erinevatel mudelitel ja kui kombineerida mitu mudelit, saaks täpsema ülevaate. Chelli ja Gendron sõnul põhinevad ESG hinnangute hindamisprotsessid erinevatel hindamismetoodikatel ja ka viisidel [20].

5.3 Võimalikud edasiarendused

Töö käigus välja töötatud lahendust saaks kindlasti optimiseerida paremaks, lisades juurde erinevaid tekstitöötlusega seotud mudeleid ning mudel täiesti ära automatiseerida, st, et mudel töötaks pidevalt ja ilma inimese abita.

Antud töö raames leidsin lahenduse, kuidas mõõta ESG panust otse ettevõtete majandusaruannetest, kuid kindlasti järgmine samm on panna mudelisse juurde rohkem andmeid, mis pärineksid ettevõtte kohta käivatest uudistest, väljaütlemistest, investorkohtumiste transkriptidest.

Tulevikus oleks kindlasti mõistlik kasutada semantilist sarnasust ühe osana tervikust, st, et leitud sarnasusskoor peaks andma osa tervest valdkonna skoorist.

Kokkuvõte

Käesoleva lõputöö eesmärk oli ehitada koosinussarnasusele baseeruvat mudel ja rakendada mudelit Eesti pankade inglisekeelsetele kvartaliaruannetele, mida siiani veel rakendatud ei ole. Ülesanne, mida lahendati oli järgmine: saada teada, milliste pankade kvartaliaruanded on kõige enam sarnasemad etteantud ESG terminitega.

Töö tähtsaimaks tulemuseks mudel, mis arvutab välja ESG faktorite sarnasusskoorid otse ettevõtete aruannete tekstist. Mudelit kasutati Eesti pankade ESG panuse mõõtmiseks otse nende ettevõtete aruannete tekstist. Analüüsiks kasutati Eesti pankade 2016-2020. aasta ingliskeelsed kvartaliaruandeid.

Mudelist saadud tulemused näitavad selgelt, et Eesti pangad kajastavad oma kvartaliaruannetes ESG seotud faktoreid ebapiisavalt või üldsegi mitte, kuid ESG-ga arvestamine ja raporteerimine on ajas kasvanud.

Tööga suudeti tõestada ka, et on võimalik mõõta ESG panust otse ettevõtete aruannete tekstist luues ise koosinussarnasusel põhineva tekstitöötlusmudeli, mis väljastab sarnasusi dokumentide ja etteantud päringu vahel. See annab lootust, et lisades tulevikus juurde rohkem andmeid ja muutujaid, saab suure tõenäosusega praeguse mudeli tulemusi oluliselt paremaks.

Kasutatud kirjandus

[1] U. Lilleväli, „Vastutustundlikkus Eesti panganduses," Jaanuar 2020. [Võrgumaterjal]. Available: https://www.estwatch.ee/wp-content/uploads/2020/02/ Vastutustundlikkus-Eesti-panganduses-Estwatch.pdf.

[2] ESMA, „Guidelines on Disclosure Requirements Applicable to Credit Ratings," [Võrgumaterjal]. Available: https://www.esma.europa.eu/sites/default/files/library/esma33-9-320_final_report_guidelines_on_disclosure_requirements_applicable_to_credit_rating_agencies.pdf.

[3] Audit Analytics, [Võrgumaterjal]. Available: https://blog.auditanalytics.com/environmental-social-and-governance-reporting-by-the-russell-3000/.

[4] Business Insider, [Võrgumaterjal]. Available: https://www.businessinsider.com/blackrock-larry-fink-investors-esg-metrics-2018-11.

[5] Euroopa komisjon, „EU taxonomy for sustainable activities," [Võrgumaterjal]. Available: https://ec.europa.eu/info/business-economy-euro/banking-and-finance/sustainable-finance/eu-taxonomy-sustainable-activities_en.

[6] P. Baier, M. Berninger ja F. Kiesel, „ResearchGate," Jaanuar 2018. [Võrgumaterjal]. Available: https://www.researchgate.net/publication/326703254_Environmental_Social_and_Governance_Reporting_in_Annual_Reports_A_Textual_Analysis.

[7] O. Giles ja D. Murphy, „SLAPPed: the relationship between SLAPP suits and changed ESG reporting by firms," [Võrgumaterjal]. Available: https://doi.org/10.1108/SAMPJ-12-2014-0084.

[8] C. Lokuwaduge ja K. Heenetigala, „Integrating Environmental, Social and Governance (ESG) Disclosure for a Sustainable Development: An Australian Study," [Võrgumaterjal]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bse.1927.

[9] T. Loughran, B. McDonald ja H. Yun, „A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports," [Võrgumaterjal]. Available: https://link.springer.com/article/10.1007/s10551-008-9910-1.

[10] S&P Global, „S&P Global," [Võrgumaterjal]. Available: https://www.spglobal.com/en/research-insights/featured/esg-going-beyond-the-balance-sheet.

[11] AlphaSense, [Võrgumaterjal]. Available: https://go.alpha-sense.com/rs/741-IHO-525/images/AlphaSense_WP_ESG-Sustainable-Success.pdf.

[12] D. Harty, „SPGLOBAL," [Võrgumaterjal]. Available: https://www.spglobal.com/en/research-insights/articles/esg-becoming-mature-market-as-sustainable-assets-hit-30-7-trillion-in-2018.

[13] E. Dimson, O. Karakas ja X. Li. [Võrgumaterjal]. Available: https://watermark.silverchair.com/hhv044.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAApwwggKYBgkqhkiG9w0BBwaggKJMIIChQIBADCCAn4GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQ

Mq9h14J_N9dQprWnqAgEQgIICTw0XjXlmc-5xv1chxzqUMKKzuQCbdJaoO66z73GxXQIWH3Ti.

[14] R. Tõnisson, „LHV,“ 25 March 2021. [Võrgumaterjal]. Available: https://fp.lhv.ee/news/newsView?locale=et&newsId=5576443&fbclid=IwAR1coqyk-s4dYVKB2_gPqOZctLrh5M4RbzJgSBqvbPzPsu8MSLX-pY_qa2o.

[15] EUR-lex, [Võrgumaterjal]. Available: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=CELEX:22016A1019(01).

[16] Euroopa Komisjon, [Võrgumaterjal]. Available: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_et.

[17] EUR-lex, [Võrgumaterjal]. Available: https://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:52018PC0353&from=ET.

[18] EUR-lex, [Võrgumaterjal]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1483696687107&uri=CELEX:52016SC0405.

[19] E. Escrig-Olmedo, M. J. Muñoz-Torres, M. Á. Fernández-Izquierdo ja J. M. Rivera-Lirio, „Lights and shadows on sustainability rating scoring,“ [Võrgumaterjal]. Available: https://www.researchgate.net/publication/266294803_Lights_and_shadows_on_sustainability_rating_scoring.

[20] Y. Gendron ja M. Chelli, „Expertise in Corporate Socio-environmental Assessment: Legitimation and Trials of Strength,“ [Võrgumaterjal]. Available: https://www.cairn-int.info/article-E_CCA_212_0063--expertise-in-corporate-socio.htm#.

[21] S. Timperley. [Võrgumaterjal]. Available: https://www.researchgate.net/publication/33052434_Corporate_Social_Responsibility_Indexes_Measure_for_Measure.

[22] Eur-lex, [Võrgumaterjal]. Available: https://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:32020R1816&qid=1617436119425&from=ET.

[23] Refinitiv, [Võrgumaterjal]. Available: https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf.

[24] I. Tammeraid, J. Majak, S. Pohjolainen ja T. Luodeslampi, „Vektorruumid,“ [Võrgumaterjal]. Available: http://linas.org/mirrors/www.cs.ut.ee/2004.01.04/toomas_l/linalg/a11e/node3.html.

[25] T. Lepikult, „Vektorruum,“ [Võrgumaterjal]. Available: https://enos.itcollege.ee/~lepikult/lineaaralgebra/Vektorruum.pdf.

[26] M. Sahlgren, „The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces,“ Researchgate, 2006.

[27] T. Ganegedara, „TowardsDataScience,“ [Võrgumaterjal]. Available: https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010.

[28] J. Pennington, R. Socher ja C. D. Manning, „GloVe: Global Vectors for Word Representation,“ [Võrgumaterjal]. Available: https://nlp.stanford.edu/projects/glove/.

[29] M. Sahlgren, „The distributional hypothesis," Italian Journal of Disability Studies, 2008.

[30] B. B. Rieger, „On Distributed Representations in Word Semantics," 1992.

[31] H. Schütze, „Word Space," CiteSeer, 1992.

[32] Varun. [Võrgumaterjal]. Available: https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db.

[33] DeepAI, [Võrgumaterjal]. Available: https://deepai.org/machine-learning-glossary-and-terms/cosine-similarity.

[34] Taltech informaatikainstituut, „Tutvumine Pythoniga," [Võrgumaterjal]. Available: http://scratch.ttu.ee/failid/Python_sisse.pdf.

[35] R. Rehurek, „Gensim," [Võrgumaterjal]. Available: https://radimrehurek.com/gensim/intro.html.

[36] G. Sidorov, A. Gelbukh, H. Gomez-Adorno ja D. Pinto, „Soft Similarity and Soft Cosine Measure: Similarity of featurers in vector space model," [Võrgumaterjal]. Available: http://www.scielo.org.mx/pdf/cys/v18n3/v18n3a7.pdf.

[37] V. Novotny. [Võrgumaterjal]. Available: https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/soft_cosine_tutorial.ipynb.

[38] K. Uiboaed, „ Tekstikaeve terminid, Sissejuhatus tekstikaevesse," [Võrgumaterjal]. Available: https://kristel.gitbooks.io/sissejuhatus-tekstikaevesse/content/tekstikaeve-terminid.html.

[39] M. Reis. [Võrgumaterjal]. Available: https://medium.com/@masreis/text-extraction-and-ocr-with-apache-tika-302464895e5f.

[40] V. Kiisk. [Võrgumaterjal]. Available: http://kodu.ut.ee/~kiisk/python.html.

[41] FilingDB, „What's so hard about PDF text extraction? ," [Võrgumaterjal]. Available: https://filingdb.com/b/pdf-text-extraction.

[42] K. Uiboaed, „Eestikeelsete stoppsõnade loend," [Võrgumaterjal]. Available: http://www.tekstikaeve.ee/blog/2018-04-18-eestikeelsete-stoppsonade-loend/.

[43] Stanford, [Võrgumaterjal]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

[44] A. Lahe. [Võrgumaterjal]. Available: https://eopearhiiv.edu.ee/e-kursused/ehitusmehaanika/node254.html.

[45] E. Dimson, „Active ownership," [Võrgumaterjal]. Available: https://academic.oup.com/rfs/article/28/12/3225/1573572?login=true.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Marek Kesküll

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Algoritmiline ESG panuse hindamine ettevõtete aruannetes: vektorruumil põhinev semantiline sarnasusskoor", mille juhendaja on Innar Liiv.
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

10.05.2021

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

Lisa 2 Andmete kogumise ja puhastamise kood

Kogu kood on avalik ka Github'is: https://github.com/makesk/NLP

```
#!/usr/bin/env python
# coding: utf-8

#Kraabime veebilehelt PDFid ja slaidiesitlused maha

import os
import requests
from urllib.parse import urljoin
from bs4 import BeautifulSoup
import tika
from tika import parser
#Tegin siia meetodi, mis võtab nasdaqbalticu lehelt siis ISIN koodi ja kuupäeva järgi kõik .pdf formaadis failid,
#laeb need lokaalselt mulle alla.
def fetch_pdfs(ISIN,date):
    url = f"https://nasdaqbaltic.com/statistics/en/instrument/{ISIN}/reports?date={date}"
    #If there is no such folder, the script will create one automatically
    global folder_location
    folder_location = ISIN
    basedir = r"C:\Users\marek.keskull\Documents\webscraping"
    fold = os.path.join(basedir, folder_location)
    print("The folder location that i am operating in: " + fold)
    if not os.path.exists(fold):os.mkdir(fold)

    response = requests.get(url)
    soup= BeautifulSoup(response.text, "html.parser")
    table = soup.find("tbody")
    #table_trs = table.find_all('tr')
    links = table.select("a[href*='con']")


    for link in links:
        filename = os.path.join(fold,link['href'].split('/')[-1])
        with open(filename, 'wb') as f:
            f.write(requests.get(urljoin(url,link['href'])).content)
 #kasutan tika libraryt, selle peab panema lokaalselt käima eraldi, et kasutada tika rest teenust.
 #Võtan kõik allalaetud PDF'id ning muudan need .txt failideks, et mul jääks alles ainult plain text
    for root, dirs, files in os.walk(fold):
        for file in files:
            path_to_pdf = os.path.join(root, file)
            [stem, ext] = os.path.splitext(path_to_pdf)
            if ext == '.pdf':
                print("Processing " + path_to_pdf)
                pdf_contents = parser.from_file(path_to_pdf)

                path_to_txt = stem + '.txt'
                with open(path_to_txt, 'w',encoding="utf-8") as txt_file:
                    print("Writing contents to " + path_to_txt)
```

```
                    txt_file.write(pdf_contents['content'])
fetch_pdfs("EE3100073644","2021-03-05")

quarters = []
basedir = r"C:\Users\marek.keskull\Documents\webscraping"
path = os.path.join(basedir, folder_location)
files = os.listdir(path)
for file in files:
    path_to_pdf = os.path.join(path, file)
    [stem, ext] = os.path.splitext(path_to_pdf)
    if ext == '.txt':
        quarters.append(file)




data = []
for c in quarters:
    with open(path+ "\\" + c, "rb") as file:
        text = file.read()
        asd = text.strip()
        decoded=str(asd,'utf-8')
        data.append(decoded)

#teen listidest dictionary
ddata = dict(zip(quarters, data))
ddata

def combine_text(list_of_text):
    combined_text = ''.join(list_of_text)
    return combined_text

data_combined = {key: [combine_text(value)] for (key, value) in
ddata.items()}

#pandas dataframe data
import pandas as pd
pd.set_option('max_colwidth',150)

data_df = pd.DataFrame.from_dict(data_combined).transpose()
data_df.columns = ['text']
data_df = data_df.sort_index()
data_df


# # Andmete puhastamine

#
# - Tekst kõik väikeste tähtedega
# - Eemdaldame ebavajaliku teksti, mis on kandiliste sulgude vahel
# - Eemaldame punktid
# - Eemaldame numbrid

import re
import string

def clean_text_round1(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), ' ', text)
    text = re.sub('\w*\d\w*', '', text)
```

```
    return text

round1 = lambda x: clean_text_round1(x)

#vaatame puhastatud teksti
data_clean2 = pd.DataFrame(data_df.text.apply(round1))
data_clean2

# 2. round puhastamisele
def clean_text_round2(text):
    text = re.sub('[‘’“”…]', '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\r', ' ', text)
    text = re.sub(r'\W+', ' ', text)
    return text

round2 = lambda x: clean_text_round2(x)


#vaatame puhastatud teksti
data_clean2 = pd.DataFrame(data_clean2.text.apply(round2))
pd.options.display.max_colwidth = 500
data_clean2

import pickle
data_clean2.to_pickle("pickles/lhv2016t2020.pkl")


# top 30 sõna sõnastikus
top_dict = {}
for c in data.columns:
    top = data[c].sort_values(ascending=False).head(30)
    top_dict[c]= list(zip(top.index, top.values))

top_dict


import pickle
data_dtm.to_pickle("pickles/historicallhvdtm.pkl")
```

Lisa 3 – Mudeli kood

```
#!/usr/bin/env python
# coding: utf-8

# # Sõnade arvutamine
#
# Võtan tekstidest kõik sõnad ja võrdlen neid etteantud sõnadega, mis
on seotud ESG-ga. Defineeritud sõnapaketi võtsin internetist.

import pandas as pd
import nltk
datacorp = pd.read_pickle('pickles/Seb2016t2020.pkl')
datacorp['document'] = datacorp.index

query_G = 'Audit and control, Board structure, Remuneration,
Shareholder rights, Transparency and Performance'
query_S = 'Access to medicines, HIV, AIDs, Nutrition, Product safety,
Community relations, Privacy and free expression, Security, Weak,
governance zones, Diversity, Health and safety, ILO core conventions,
Supply chain labor standards, Bribery and corruption, Political
influence, Responsible marketing, Whistle-blowing systems, disclosure
and reporting, Governance of sustainability issues, Stakeholder
engagement, UNGC compliance'
query_E = 'Biofuels, Climate ,Emissions ,land, Biodiversity, Water,
Environmental, standards, Pollution, Supply, Waste, recycling'
datacorp



import json
import logging
from re import sub
from multiprocessing import cpu_count

import numpy as np

import gensim.downloader as api
from gensim.utils import simple_preprocess
from gensim.corpora import Dictionary
from gensim.models import TfidfModel
from gensim.models import WordEmbeddingSimilarityIndex
from gensim.similarities import SparseTermSimilarityMatrix
from gensim.similarities import SoftCosineSimilarity


import logging

# logimine
logging.basicConfig(format='%(asctime)s : %(levelname)s :
%(message)s', level=logging.WARNING)  # DEBUG # INFO


import nltk

# stoppsõnad
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('english')
```

```
#andmetest võtan andmepealkirjad ja tekstid nimetan dokumentideks

titles = [item for item in datacorp['document']]
documents = [item for item in datacorp['text']]
print(f'{len(documents)} documents')

def preprocess(doc):
    logging.info( 'tokenizing' )
    doc = doc.lower().split()
    doc = [w for w in doc if w not in stopwords]
    return doc

import re
import string
def preprocess_query(doc):
    logging.info( 'tokenizing' )
    doc = doc.lower().split()
    doc = [remove_punc(i) for i in doc]
    doc = [w for w in doc if w not in stopwords]
    return doc


def remove_punc(string):
    punc = '''!()-[]{};:'"\, <>./?@#$%^&*_~'''
    for ele in string:
        if ele in punc:
            string = string.replace(ele, "")
    return string


corpus = [preprocess(document) for document in documents]

query = preprocess_query(query_G)
print(query)


# glove vektoripakk, siin on 400000 vektorit sees

glove = api.load("glove-wiki-gigaword-50")



#vecs = similarity_index.keyedvectors


#from vec2graph import visualize

#visualize(r'C:\Users\marek.keskull\Documents\GitHub\NLP\Vizualization
', vecs, 'audit')



#glove vektorite koosinussarnasus indeksi näide
#most_similar =
similarity_index.keyedvectors.most_similar(positive=['water'],
topn=10)
#most_similar
```

```
#ehitame TF-idf mudeli

#kõigepealt ehitame valmis andmesõnastiku, kus on sees kõik dokumendi
sõnad ja otsingupäringu sõnad vormis: 'võti':'sõna'
logging.info( 'building dictionary' )
dictionary = Dictionary(corpus+[query])

#This module implements functionality related to the
#Term Frequency - Inverse Document Frequency vector space bag-of-words
models.
tfidf = TfidfModel(dictionary=dictionary)


#Builds a sparse term similarity matrix using a term similarity index.
similarity_matrix = SparseTermSimilarityMatrix(similarity_index,
dictionary, tfidf, nonzero_limit=100)


#dictionary = Dictionary(corpus+[query])
#print(dictionary[429])
#basedir =r"C:\Users\marek.keskull\Documents\GitHub\NLP\Vizualization"
#logging.info( 'saving dictionary' )
#dictFile = basedir + '.dict'
#dictionary.save_as_text(dictFile, sort_by_word=True)



#sparse maatriksi kasutamise näide
#similarity_matrix.inner_product(dictionary.doc2bow(query),dictionary.
doc2bow(corpus[7]))
#similarity_matrix.matrix.todense()


similarity_matrix.matrix.nnz


len(dictionary)**2


#Compute soft cosine similarity against a corpus of documents by
storing the index matrix in memory.
index = SoftCosineSimilarity(tfidf[[dictionary.doc2bow(document) for
document in corpus]],similarity_matrix)

doc_similarity_scores =
index.get_similarities(dictionary.doc2bow(query))
doc_similarity_scores




sorted_indexes = np.argsort(doc_similarity_scores)[::-1]
d = []
for idx in sorted_indexes:
    d.append(
        {
            'Document no': idx,
            'Similarity score with query': doc_similarity_scores[idx],
            'Document name': titles[idx]
        }
    )
```

```
final = pd.DataFrame(d)


doc_similar_terms = []
max_results_per_doc = 30
#query = ['audit', 'control', 'board', 'structure', 'remuneration',
'shareholder', 'rights', 'transparency', 'performance']
for term in query:
    #dictionary = Dictionary(corpus+[query])
    #dictionary is query + my corpus(which has 25 documents)

    idx1 = dictionary.token2id[term]
    for document in corpus:
        #print(document.name)
        results_this_doc = []
        for word in set(document):
            idx2 = dictionary.token2id[word]
            score = similarity_matrix.matrix[idx1, idx2]
            if score > 0.0:
                results_this_doc.append((word, score))

        results_this_doc = sorted(results_this_doc, reverse=True,
key=lambda x: x[1])

        results_this_doc =
results_this_doc[:min(len(results_this_doc), max_results_per_doc)]
        #print(results_this_doc)
        doc_similar_terms.append(results_this_doc)



results = []
for idx in sorted_indexes[:30]:
    similar_terms_string = ', '.join([result[0] for result in
doc_similar_terms[idx]])
    results.append(
        {
            'Document no': idx,
            'Similarity score with query': doc_similarity_scores[idx],
            'Document name': titles[idx],
            "Most similar words":similar_terms_string
        }
    )

similar_words = pd.DataFrame(results)
similar_words

import pickle
similar_words.to_pickle('Governanceresults/SEB.pkl')
```

Lisa 4 – Graafikute kood

```
#!/usr/bin/env python
# coding: utf-8

import pandas as pd
e_swed = pd.read_pickle('Environmentresults/Swedbank.pkl')
s_swed = pd.read_pickle('Socialresults/Swedbank.pkl')
g_swed = pd.read_pickle('Governanceresults/Swedbank.pkl')
g_swed


e_swed = e_swed.rename(columns={"Similarity score with 
query":"Similarity score with query_E","Document name":"Year"})
#e_swed = e_swed.drop(columns=["Document no","Most similar words"])
s_swed = s_swed.rename(columns={"Similarity score with 
query":"Similarity score with query_S","Document name":"Year"})
#s_swed = s_swed.drop(columns=["Document no","Most similar words"])
g_swed = g_swed.rename(columns={"Similarity score with 
query":"Similarity score with query_G","Document name":"Year"})
#g_swed = g_swed.drop(columns=["Document no","Most similar words"])
s_swed


dfs = [e_swed, s_swed, g_swed]
dfs = [df.set_index('Year') for df in dfs]
s = dfs[0].join(dfs[1:])
s.index.names = ['Quarter document']
s.mean()


s = s.assign(Year_number = "20" + s.index.str[3:5])
s = s.sort_values(by='Year_number')
s


s.plot(x ='Year_number', y='Similarity score with query_E', kind = 
'line', title = "SWEDBANK")



s.plot(x ='Year_number', y='Similarity score with query_S', kind = 
'line', title = "SWEDBANK")



s.plot(x ='Year_number', y='Similarity score with query_G', kind = 
'line', title = "SWEDBANK")


import pickle
ax3 = s.plot(x='Year_number',mark_right=False,figsize=(15,13), title = 
"SWEDBANK",fontsize = 35)
pickle.dump(ax3, open("swedplot.pickle", "wb"))
```