TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika õppekava

Roman Šumailov

Objektiivne raamatute soovitussüsteem

Bakalaureusetöö (9 EAP)

Juhendaja: Siim Karus

Tartu 2015

Objektiivne raamatute soovitussüsteem

Lühikokkuvõte:

Antud töö eesmärgiks oli objektiivselt mõõdetavatel atribuutidel põhineva raamatute soovitussüsteemi loomine. Loodud süsteem kasutab soovituste andmiseks raamatute hinnangute põhjal õpitud lineaarse regressiooni mudelit. Töös hinnatakse antud lähenemise efektiivsust taolise probleemi lahendamisel. Mudeli põhjal on võimalik arvutada erinevusi ingliskeelsete raamatute vahel saades sisendiks analüüsitud raamatupaarid. Iga raamatu sisu puhul analüüsitakse teda kirjeldavaid atribuute nagu teksti positiivsus või keskmine lausete pikkus. Treeningandmetena kasutati andmestikku lugejate hinnangutest raamatutele ja avatud raamatute kogumiku. Tulemusena saadi mudel, mis suudab korrektselt leida erinevusi raamatute vahel treeningandmete piires, kuid ei suuda korrektselt leida inimlikust vaatepunktist sarnaseid raamatuid treeningandmete vähesuse ja kvaliteedivaesuse tõttu. Mudeli töö korrektsuse põhjal järeldati, et suurema ja parema treeningandmestiku puhul antud töös kasutatud lähenemisel on potentsiaali arendada ka inimliku subjektiivsuse vaatepunktist korrektse raamatusoovitaja.

Võtmesõnad:

Soovitussüsteem, Lineaarne regressioon, Raamatud, Projekt Gutenberg

Objective Book Recommender System

Abstract:

This work's goal was the creation of a book recommender system based on objectively measurable attributes. For recommendations, the created system uses linear regression model trained on the books' ratings. In this work effectiveness of the described approach as a solution for the given problem is measured**.** The model is able to estimate the differences between two books in English. For every book's content, a set of attributes describing the book are calculated, such as text positivity or average sentence length. Books' ratings and open books collection were used as the training dataset. As a result, the model is able to find differences correctly in terms of the training data provided, but is not able to correctly find the nearest books from human point of view due to the lack of a large good quality dataset. Based on the model's work correctness, it was concluded that given a big enough good quality training dataset the approach used in this work has potential to find a model capable of finding the nearest books in human understanding.

Keywords:

Recommender system, Project Gutenberg, Books, Linear regression

Sisukord

1. Sissejuhatus

Raamatute soovitussüsteemide valdkonnas paljud põhinevad ühisel kasutajate-poolsel filtreerimisel (*collaborative filtering*), mis on soovitussüsteemide arendamisel kasutatav lähenemine [1]. Nii nende kui ka teiste lahenduste algoritmid ja kood on kinnised. Samuti pole palju informatsiooni sellest, millised raamatute atribuudid on nende hindamisel lugejatele tähtsad. Käesoleva töö eesmärk on arvutada mudel, mis oleks võimeline leidma raamatutevahelisi erinevusi numbrilises ekvivalendis ehk tinglikke kaugusi. Valminud töö tulemuseks on raamatuid kirjeldavatel atribuutidel baseeruv filtreerimine, mis on ehitatud kasutajate-poolset filtreerimist kasutatavatele andmetele – raamatute hinnangutele. Lisaks hinnatakse selles töös kasutatava lähenemise efektiivsust püstitatud probleemi lahendamisel, potentsiaali ning valminud mudeli konkurentsivõimelisust olemasolevatele raamatute soovitussüsteemidele. Kõrvaltulemusena töö käigus uuritakse mõnede levinud raamatute hinnangute andmestike ja avatud raamatute kogumiku kasutamist, töötlemist ja sellega esinenud probleeme.

Esimeses peatükis kirjeldatakse raamatute soovitussüsteemide tarkvaraturu hetkeseisu, pärast antakse detailsema ülevaate antud töö eesmärgist ja selle uudsusest. Lõpuks, kuna töö koosneb mitmest eri osast, kirjutatakse lahti probleemi lahenduskäik ja lahendamiseks kasutatav meetod. Teises peatükis on ülevaade töö lahendamiseks kasutatavatest tehnoloogiatest, nagu programmeerimiskeel, kolmandate osapoolte teegid jms. Kolmandas peatükis on kirjeldatud töös kasutatavad andmed ja nende töötlusprotsess. Järgmises peatükis räägitakse töö jaoks disainitud andmebaasimudelist, et seletada, kuidas jaotatakse ja kasutatakse neljandas peatükis kirjeldatud andmeid. Viiendas peatükis on juttu mudeli treenimise meetodist, selle rakendusest ja saadud tulemustest. Viimases peatükis testitakse ning analüüsitakse töö tulemusi.

Lisades on toodud MySQL käsud ja arendatud tarkvara vajaminevad meetodid, millega saavutatakse üks või teine töös kirjeldatud tulemus. Samuti on seal toodud testimistulemuste tabelid.

2. Taust

Antud peatükis antakse ülevaade sellest, millised püstitatud ülesande lahendused on olemas ja millised nendest on tarkvaraturul laiemalt levinud. Lisaks sellele kirjeldatakse detailsemalt töö eesmärki ja uudsust. Kuna töö koosneb mitmest eraldi osast, millest igaühte läheb teatud sammul vaja, siis viimases alampeatükis on lühidalt antud töö eesmärgi saavutamiseks sammhaaval teostatavate operatsioonide käik.

2.1 Sarnaste lahenduste turuseis

Maailmas on juba olemas hulk raamatute soovitussüsteemide lahendusi. Suurem osa nendest aga põhineb kasutajate-poolsel filtreerimisel. Nende puhul korjatakse ja analüüsitakse palju kasutajaid puudutavaid andmeid, selliseid nagu hinnangud, sõbraloendid, ühised loetud raamatud. Näiteks, LifeHacker'is[1] mainitud raamatute soovitussüsteemide hulgas on sellised neli viiest: Goodreads, LibraryThing, Reddit's BookSuggestions Subreddit, Olmenta. TheGuardian[2] ajalehe poolt soovitatute hulgast tõenäoliselt neli kaheksast (Goodreads, Amazon.com, My Independent Bookshop, Riffle), DigitalBookWord'i[3] poolt pakututest viis kümnest (Amazon.com, BarnesandNoble.com, Goodreads, Riffle, Sony Reader Store). Summas, kolmeteistkümnest leitud soovitussüsteemist üheksa toetuvad kasutajate-poolsele filtreerimisele ja viis toimivad teistel algoritmidel. Nendest silmapaistvam on Whichbooks[4], kus võib valida mõned parameetrid ja neid kasutades endale raamatu valida, kuid nende puhul on tõenäoliselt tegemist käsitsi hinnatud parameetritega. Antud töö tulemusena saadakse teistsugust samal printsiibil lahendust ingliskeelsetele raamatutele. Raamatute erinevuste leidmine hakkab toimima toetudes raamatut kirjeldavatele atribuutidele ja nende hindamine hakkab toimuma objektiivselt - arvuti poolt, inimsekkumiseta. Samas, mudeli treenimine omakorda toetudes inimeste hinnangutele raamatutele.

2.2 Eesmärk ja innovaatilisus

Kasutajate-poolsele filtreerimisele ehitatud süsteemide üks puudujääkidest on näiteks külma stardi probleem, see on olukord, kus soovitussüsteemil pole piisavalt statistilisi andmeid soovituste tegemiseks, tavapäraselt soovitussüsteemi käivitamise alguses [2]. Lisaks, kõik leitud raamatute soovitussüsteemid on kommertsveebirakendused kinnise koodiga, mida ei saa iseseisvalt kasutada või täiendada. Antud töö kood saab olema avalik ja avatud kas kasutamiseks või edasiseks arendamiseks.

Käesoleva töö idee on tulnud Muusika Genoomprojektilt (*Music Genome Project*)[5], mis on Pandora[6] internet-raadios kasutatav tehnoloogia. Tänu sellele tehnoloogiale on Pandora võimeline mängima sarnaseid laule ja nii vastama kuulaja maitsele. Seal iga laul omab mingi hulga atribuute, mida eksperdid käsitsi hindavad. Hiljem, kasutades oma algoritme, toimub laulude võrdlemine ja soovitus.

Antud töö eesmärgiks on hinnangutele ja raamatu sisu analüüsile ehitada raamatutevahelist erinevust hindav mudel ning hinnata selle lähenemise täpsust ja potentsiaalset konkurentsivõimelisust võrreldes levinumate lähenemistega. Mudeli uudsus seisneb selles, et ta

[1] http://lifehacker.com/five-best-book-recommendation-services-1577706074
[2] http://www.theguardian.com/books/booksblog/2014/may/10/top-book-recommendation-platforms-what-are-your-favourites
[3] http://www.digitalbookworld.com/2013/top-ten-book-recommendation-platforms/
[4] http://www.openingthebook.com/whichbook/
[5] https://www.pandora.com/about/mgp
[6] http://www.pandora.com/about

on ehitatud hinnangutele ja sisupõhise filtreerimise (*content-based filtering*, soovitamine toetudes objektide kirjeldustele ning kasutaja eelistustele [3]) elemendile - objekti kirjeldusele (raamatu atribuutide analüüs), ja mudel on võimeline raamatuid võrdlema saades sisendiks analüüsitud raamatu atribuute. Lisaks, töö tulemus annab ülevaate lugejatele olulisematest raamatu atribuutidest raamatu hindamisel. Töö kõrvaltulemustena antakse ülevaate tööst mõne raamatute hinnangute andmestiku ja avatud raamatute kogumikuga.

2.3 Lähenemine

Töö idee seisneb selles, et leida mudeli kasutades vähimruutude regressiooni, mis kahe antud raamatu atribuutide põhjal oleks võimeline arvutama raamatute erinevust. Käesolevas töös raamatute erinevust mõõdetakse numbrilises väärtuses, mida suurem number, seda suurem erinevus.

Regressioonmudel on *y* muutuja sõltuvus ühest või mitmest *x* regressorist, mis on kujul:

$$y = \sum_{1}^{k} b_k x_k \quad \text{, kus} \qquad [4]$$

b_k - regressiooni parameetrid (koefitsiendid)

x_k - regressorid (mudeli faktorid, atribuudid)

k - mudeli faktorite arv

y - sõltuv muutuja

Siin *y* muutuja tähistab numbrilist erinevust (ka kaugust) kahe raamatu vahel. Regressoriteks on raamatut kirjeldavad mõõdetavad atribuudid, sellised nagu keskmine lausete pikkus või nimisõnade suhteline arv tekstis. Kahe raamatu vahe leidmiseks rakendatava regressioonmudeli parema poole leitakse järgmiselt: iga regressori jaoks (x_k) leitakse mõlema raamatu nende atribuutide absoluutvahe ning saadud arvu pannakse regressori väärtuseks. Näiteks, kahe raamatu erinevuse leidmisel esimese raamatu positiivsuse väärtusest lahutatakse teise raamatu oma, nendest võetakse absoluutväärtus ja see number määratakse valemi parema poole positiivsuse regressoriks ja korrutatakse vastava koefitsiendiga.

Antud töös tulemuse saavutamiseks otsustati kasutada vähimruutude lineaarset regressiooni, sest ta täidab töö tingimusi: ta ennustab sõltuva muutuja väärtuse toetudes ühele või mitmele sõltumatule muutujale. Andmed on selle jaoks sobivad: nad on lineaarsed ja sõltumatud, kõik atribuudid kirjeldavad raamatu. Lisaks sellele, lineaarse regressiooni kasutamine ning sellest arusaamine on lihtne selle ala mitte-spetsialistile.

Raamatu kirjeldamiseks kasutatakse järgmisi atribuute:

1. keskmine lausete pikkus,
2. keskmine sõnade pikkus,
3. tähtede arv (ilma tühikuteta),
4. kuupäevade arv,
5. geograafiliste kohtade arv,
6. erinevate tegelaste arv,
7. raamatu positiivsus,
8. suhteline omadussõnade arv tekstis (omadussõnade protsent tekstis),

9. suhteline nimisõnade arv tekstis (nimisõnade protsent tekstis),
10. suhteline tegusõnade arv tekstis (tegusõnade protsent tekstis),
11. lausete arv,
12. sõnade arv.

Atribuudid olid valitud lähtuvalt kasutatud tehnoloogiate ja teekide võimelisusest, kasutati võimalikult palju mõõdetavaid atribuute, et suurendada otsitava mudeli täpsust ja aktuaalsust.

Üldistatult, tulemuste saavutamiseks hõlmab töö käik järgmisi samme.

1. Laaditakse alla avatud andmestiku inimeste hinnangutega raamatutele.
2. Lugejate hinnangud töödeldakse kasutuskõlblikeks antud töö jaoks.
3. Kasutades hinnanguid leitakse erinevused hinnatud raamatute vahel, nendest saavad sõltuvad *y* muutujad regressioonmudeli treenimiseks. Osa hinnanguid eraldatakse hiljemaks kasutamiseks testandmetena.
4. Laaditakse alla avatud kogumiku raamatutest.
5. Raamatukogumiku loetakse sisse ja töödeldakse sobivaks antud töö jaoks.
6. Pärast töötlemist jäänud raamatud analüüsitakse ja iga raamatu atribuudid väärtustatakse.
7. Raamatupaaride erinevuste ja analüüsitud raamatute abil arvutatakse regressorite koefitsiendid ning salvestatakse andmebaasi.
8. Koefitsiendid kasutatakse erinevuste leidmiseks kõikide raamatute vahel, tulemused salvestatakse andmebaasi.
9. Eelmises punktis saadud erinevused võrreldakse punktis 3. eraldatud testandmetega ning leitakse suurim ja vähim erinevused, dispersioon ning veel mõned punkthinnangud, mis on lähemalt kirjeldatud testimise peatükis.
10. Punktis 8. saadud erinevuste abil leitakse mõned sarnasemad raamatud ja hinnatakse tulemust.
11. Tehakse järeldused.

3. Arhitektuur

Töö realiseerimiseks kasutati programmeerimiskeelena Java SE 1.8, sest Java on üsna levinud ning erinevaid teeke on sellele küllalt. Programmeerimiskeskkonnaks valiti IntelliJ IDEA 14[7], andmebaasi puhul otsustati MySQL 5.5[8] kasuks, sest valitud treeningandmestik on tehtud MySQL andmebaasi jaoks ja MySQL funktsionaalsus laseb soovitud viisil treeningandmed töödelda. Samuti on see levinud tasuta andmebaas, mis tähendab, et kui kellelgi läheb selles töös saadud andmestiku vaja, siis inimene tõenäoliselt oskaks neid kohe kasutada. Lisaks mainitud põhjustele on autor nimetatud töö arhitektuuri elementides kõige pädevam. Töös kasutatavateks tehnoloogiateks on Gradle[9] ja Hibernate[10]. Hibernate - objekt-relatsiooniline kaardistusraamistik (*Object-Relational Mapping*) Java jaoks, mis laseb kaardistada objekt-orienteeritud mudelid objektidele traditsioonilises relatsioonilises andmebaasis. Võetud kasutusele, sest oluliselt kergendab Java ja MySQL vahelist tööd võrreldes puhta JPA-ga. Kaardistamise raamistikuks on valitud konkreetselt Hibernate, sest autor on sellega enim tuttav. Gradle - projekti ehitustööriist, võetud kasutusele kolmandate osapoolte teekide ühendamiseks projekti, Gradle abiga on nende haldamine (lisamine, uuendamine, kustutamine) lihtsam võrreldes puhta käsitsi tegevusega.

Antud töö raames analüüsitakse HTML (*HyperText Markup Language*) vormingus antud raamatu sisu ja määratakse hinnangud raamatuid kirjeldavatele atribuutidele, selleks on vaja keele töötlemistööriistu. Vaadeldi järgmisi teeke ja tehnoloogiaid:

1. The Stanford NLP[11],
2. LingPipe[12],
3. Apache UIMA[13],
4. GATE[14],
5. FrameNet[15],
6. FreeLing[16],
7. Apache OpenNLP[17].

Otsustati Apache OpenNLP teeki kasuks järgmistel põhjustel: ta pakub vajaliku funktsionaalsust, tema dokumentatsioon on selge ja kergesti loetav, tema kasutamine on lihtsam kui teiste variantidega.

Töö sisuliseks eesmärgiks on lineaarse regressioonmudeli leidmine, mida on vaja arvutada. Selle alamülesande lihtsustamiseks kasutatakse Apache Math Commons 3.5[18] teegi võimalusi, sest ta on lihtsasti loetav ja kasutusele võetav.

[7] https://www.jetbrains.com/idea/
[8] https://www.mysql.com/
[9] https://gradle.org/
[10] http://hibernate.org/orm/
[11] http://nlp.stanford.edu/software/
[12] http://alias-i.com/lingpipe/
[13] http://uima.apache.org/
[14] https://gate.ac.uk/
[15] https://framenet.icsi.berkeley.edu/fndrupal/about
[16] http://nlp.lsi.upc.edu/freeling/
[17] https://opennlp.apache.org/
[18] http://commons.apache.org/proper/commons-math/

4. Andmed

Antud peatükis kirjeldatakse töös kasutatavaid andmeid, nende päritolu ja eesmärki. Samuti antakse ülevaadet nende töötlemise protsessist, selle käigus esinevatest probleemidest ja andmestike kasutamisega kaasnevatest ohtudest.

4.1 Sõnade positiivsuse hinnangud

Positiivsuse atribuudi hindamiseks on vaja sõnade positiivsuste hinnangutega andmestiku. Selleks kasutati afektiivsete hinnangute andmestiku[19] (*affective ratings*) - selles on umbes 14000 hinnatud ingliskeelset sõna. Andmestikuks on csv formaadi fail, kus on nimekiri sõnadest. Iga sõna kirjeldavad valents (*valence*) ehk sõna meeldivus, virgutus (*arousal*) ehk sõna poolt põhjustatavate emotsioonide intensiivsus ja dominantsus (*dominance*) ehk sõna kontrolli tase [2]. Antud töös võetakse valentsi sõna positiivsuse hinnanguna. Andmestikufailis on iga sõna nimetavas käändes, aga raamatute tekstides on kindlasti ka teistes. Selleks, et tagada maksimaalse positiivsuse hinnangut omavate sõnade tuvastust tekstides, leitakse andmestikus olevate sõnade algvormid ning samade valentsväärtustega salvestatakse eraldi csv faili, kust neid edaspidi kasutamiseks võetakse. Algvormide leidmiseks kasutatakse *Porter*'i tüvemoodustuse (*stemming*) Java jaoks kirjutatud algoritmi[20] . Nende salvestamiseks vajalikud käivitamiseks meetodid on treening- ja testandmete genereerimise lisas.

4.2 Raamatute hinnangud

Selles alampeatükis on toodud raamatuhinnangute andmestiku valikud, põhjendused ja probleemid. Lisaks kirjeldatakse andmestiku töötluse protsessi, et ta sobiks antud töös kasutamiseks.

Andmestiku valik

Mudeli treenimiseks on vajalikud kasutajate hinnangud raamatutele. Valiti BX (*Book-Crossing Dataset*) andmestiku[21] ja Amazoni arvustuste[22] (*Amazon reviews*) vahel. Otsustati BX andmestiku kasuks, sest Amazoni andmestikus olevad kirjed ei ole kokkuviidavad kasutatava raamatute tekstide andmebaasiga. Amazoni andmestikus on nimekiri objektidest, mis koosnevad pealkirjast, hinnangust, kasutajast ja ASIN-ist (*Amazon Standard Identification Number*), mis on sama, mis ISBN (*International Standard Book Number*) Amazoni raamatute puhul [3]. Raamatu analüüs toetub aga Gutenbergi projekti[23] (*Gutenberg Project*) raamatutele, millel ISBN-i pole, on vaid pealkiri ja autor. Samal ajal, Amazon ei võimalda mõistlikult lihtsat viisi massiliselt raamatute pealkirju ja autoreid saada kasutades ISBN-e. Juhul kui kasutusele oleks võetud Amazoni andmestik, siis ainsaks võimaluseks ühildada hinnangud analüüsitud raamatutega oleks pealkirja põhjal autorit arvestamata. See jätab liiga suure riski, et kaks eri autorit kirjutasid raamatu sama pealkirjaga ning kokkuvõttes võib see töö tulemusi väga tugevalt mõjutada. BX andmestikus aga on olemas eraldi kasutaja, raamatu ISBN, pealkiri, autor ja hinnangud.

[19] http://crr.ugent.be/archives/1003
[20] http://tartarus.org/martin/PorterStemmer/
[21] http://www2.informatik.uni-freiburg.de/~cziegler/BX/
[22] http://snap.stanford.edu/data/web-Amazon-links.html
[23] https://www.gutenberg.org/

BX andmestikust kasutatakse kahte MySQL tabelit:

1. *BX-Books* - raamatute nimekiri, iga objekti kohta on esitatud ISBN, pealkiri, autor, väljastamise aasta, väljaandja, väikse, keskmise ja suure suurusega raamatukaane illustratsioon. Selles on 271379 kirjet.
2. *BX-Book-Ratings* - nimekiri raamatute hinnangutest. Iga objekti kohta on antud kasutaja, ISBN ja hinnang skaalal 0-9. Selles on 1149780 kirjet.

Andmestiku töötlus

Sellisel kujul, nagu andmed imporditud on, nad töösse ei sobi järgmistel põhjustel.

1. Hinnangute seas on ISBN-id, mida raamatute seas ei ole.
2. Kuna Gutenbergi projekti raamatu ainsaks võimalikuks unikaalsuse kirje tagajaks on raamatu pealkiri ja autor, siis seda kasutatakse ka töös. Aga andmestikus võib olla mitu sama pealkirja ja autoriga raamatut (erinevad ISBN-id eri väljaannete tõttu, näiteks), mistõttu korduvad raamatud tuleb eemaldada ja jätta ainult 1 eksemplar. Samas, ei tohi kaotada hinnanguid, mis on tehtud kustutatud raamatutele.
3. Mõned kasutajad on andnud ainult 1 hinnangu, sellistest hinnangutest pole antud töös kasu.
4. Paljud hinnangud võrduvad 0-ga, mis tõenäoliselt tähendab originaalselt hinnangu puudust (näiteks, tootele oli pandud ainult kommentaar).
5. Andmestiku skript on kirjutatud versioonist 5.5 vanema MySQL andmebaasi jaoks.

Selleks, et BX andmed sobiksid töö konteksti ja oleksid analüüsitud raamatutega ühilduvad, on vaja neid enne rakendamist töödelda. Nende töötlemiseks vajalikud sammud on kirjeldatud allpool.

1. Kustutatakse kõik hinnangud, mille ISBN-e pole raamatute seas. Järele jääb 1031175 hinnangut.
2. Kustutatakse kõik hinnangud, mis võrduvad 0-ga. Tõenäoliselt on need hinnangud lihtsalt kommentaarid, aga kuna neid on väga palju hinnangute summa kohta (716109 kirjet ehk ~62% kogu andmestikust), siis võivad nad valesti tulemusi mõjutada. Järele jääb 383852 hinnangut.
3. Kustutada raamatud, mis pole BX ja Gutenbergi raamatute kogumikel ühised. Vastavalt, kustutada kõik hinnangud, millel on kustutatud raamatute ISBN-id. Järele jääb 1116 raamatut ja 1589 hinnangut.
4. Raamatute hulgas võivad olla mõned korduva pealkirja ja autoriga, mis on tavaliselt üks ja sama eri väljaandega raamat, aga raamatu unikaalsuse tagamiseks antud töös on raamatu pealkiri ja autor. Selle parandamiseks kustutatakse hinnangud, kus üks ja sama kasutaja hindas mitu sama pealkirja ja autoriga raamatut jättes järele ühe hinnangu antud raamatu jaoks. Seejärel leitakse kõik raamatud, millel on sama pealkiri ning autor ja nendest jäetakse alles üks. Et väärtuslikud hinnangud ei kaoks koos kustutatud raamatutega, siis kõikidele kustutatud raamatute hinnangutele omastatakse allesjäetud raamatute ISBN-id. Näiteks, kui kustutatud raamatu ISBN-i number on 1112223334 ja allesjäetud oma 5556667778, siis kõikidele esimese raamatu hinnangutele omistatakse ISBN-i väärtuseks 5556667778. Järele jääb 457 raamatut ja 1563 hinnangut.

5. Hinnangutest kustutatakse kõik need, mis on kasutaja ainsad ehk kui teatud kasutajal on ainult üks hinnang, siis see hinnang kustutatakse. Tehakse seetõttu, et raamatutevaheliste erinevuste leidmiseks on vaja võrrelda sama kasutaja hinnanguid eri raamatute jaoks, järelikult, igal kasutajal peab olema vähemalt kaks hinnangut. Järele jääb 588 hinnangut.
6. Lõpuks kustutatakse kõik need raamatud, millel pole rohkem hinnanguid ning pärast seda sammu jääb kasutamiseks 195 raamatut.

BX andmestiku töötlemise protsess koos vajaminevate MySQL koodide ja arendatud tarkvara vajaminevate meetoditega on kirjeldatud BX andmestiku töötlemise lisas. Pärast töötlemist jääb töös kasutamiseks 195 raamatut ja 588 hinnangut.

4.3 Mudelid Apache OpenNLP jaoks

Tabelis 1 on Apache poolt eeltreenitud mudelid[24] vastavate andmete leidmiseks teksti seest. Neid mudeleid kasutatakse erinevate atribuutide määramiseks.

Tabel 1. Kirjeldab töös kasutatud Apache OpenNLP eeltreenitud mudeleid.

Mudeli kirjeldus	**Mudeli faili nimetus**	**Mille leidmiseks kasutatakse**
kuupäeva otsimismudel	en-ner-date.bin	kuupäevade arv
koha otsimismudel	en-ner-location.bin	kohtade arv
isikute otsimismudel	en-ner-person.bin	tegelaste arv
sõnaliikide otsimismudel	en-pos-maxent.bin	suhtelised omadussõnade, nimisõnade ja tegusõnade arvud tekstis
lause otsimismudel	en-sent.bin	laused, lausete arv
lausete lekseemideks jaotaja	en-token.bin	lausete lekseemideks jaotamiseks

4.4 Raamatud

Avatud andmestiku raamatud, mida töös analüüsitakse ja kasutatakse lineaarse regressioonmudeli loomiseks. Valiti järgmiste allikate seast:

1. HathiTrust[25],
2. Digital Public Library of America[26],
3. The European Library[27],
4. Google Books[28],
5. ibiblio[29]
6. Internet Memory Foundation[30],

[24] http://opennlp.sourceforge.net/models-1.5/
[25] http://www.hathitrust.org/
[26] http://dp.la/
[27] www.theeuropeanlibrary.org
[28] https://books.google.com/
[29] http://www.ibiblio.org/
[30] http://internetmemory.org/en/

7. Open Library[31],
8. M Library[32],
9. Pandora[33],
10. Project Gutenberg[34],

Otsustati Gutenbergi projekti kasuks, sest kõikide ülalmainitute seast on see ainus avatud raamatute kogumik, kus raamatuid saab andmestikuna alla laadida ja nendel on ühine formaat (HTML), mille seest võib vähemalt mingil määral eristada pealkirja, autori ja sisu. Allalaaditud raamatud on jaotatud kataloogide järgi ning iga raamatu failid on eraldi zip arhiivis. Raamatu failideks on kas üks või mitu HTML faili, lisaks sellele võib raamat olla jaotatud eraldi kaustadeks ja seal võivad olla ka pildid. Ühist raamatuarhiivi ehituse kirjeldust Gutenbergi projekti leheküljel ei ole, antud töös kasutatav arusaam raamatuarhiivi struktuurist on järeldatud käsitsi raamatufailide uurimisest. Paljud raamatud on salvestatud eri moel, suurem osa nendest on üks HTML fail, mille seest võib leida pealkirja, autori ja sisu. Sellest tulenevalt antud töös keskenduti HTML failide otsimisele ja lugemisele. Kui faili seest on võimalik leida pealkirja, autori ja sisu, siis loetakse raamat leituks, vastasel juhul fail jäetakse vahele.

Raamatu kindlakstegemist raskendab see asjaolu, et Gutenbergi projekti raamatufailides pole ühist standardit pealkirja, autori ja sisu eristamise jaoks. Ka nende eristamise struktuur antud töös on järeldatud failide käsitsi uurimisest. Pealkirja, autori ja sisu jaoks pole eraldi `<div>` silte. Suuremal osal juhtudest on pealkiri ja autor leitavad `<pre></pre>` siltide ning sisu `<body></body>` siltide vahelt koos teise infoga. Seetõttu antud töös loetakse raamatu sisuks kõik, mis jääb `<body></body>` vahele, kuna täpsemat meetodit selle leidmiseks pole. Sama kehtib pealkirja ja autori puhul, antud töös hoitakse kinni `<pre></pre>` siltidest. Sellegipoolest ei kõlba need raamatud otsekoheselt kasutamiseks antud töös, sest allalaaditud kogumiku raamatud võivad mitte ühtida BX andmestiku hinnatud raamatutega. Samuti Gutenbergi projekti kasutamisel on oht laadides alla ingliskeelsed raamatud, saada mõned raamatud hoopis teistes keeltes. Allalaetud kogumikust, mis pidi koosnema ainult ingliskeelsetest sõnadest, leiti mõned saksakeelsed ja üks hiinakeelne. Antud raamatute kogumikust luuakse olemasolevate raamatute kataloog nimekirjaga raamatutest, kus iga raamatu veergudeks on pealkiri ja autor. Raamatute nimekirja kataloogi loomine koosneb järgmistest sammudest.

1. Läbitakse kõik allalaaditud kogumiku kataloogid, arhiivid võetakse lahti ja nendest otsitakse HTML faile. Kui mõni fail vastab raamatu tingimustele, siis loetakse raamat leituks, salvestatakse andmebaasi ja liigutakse järgmist otsima. Kuna kataloogi loomise eesmärk on näha, millised raamatud on kogumikus olemas, siis iga leitud raamatu puhul andmebaasi salvestatakse ainult pealkirja ja autori.
2. Loodud kataloogist kustutatakse kõik raamatud, mis pole varem leitud BX tabelis olevate raamatutega ühised. Järele jääb 195 raamatut.

Nagu ka hiljem testimise tulemustest näha (Tabel 4 ja Tabel 6) on, esialgse uurimuse läbiviimiseks on 195 raamatut piisav andmestik. Gutenbergi projekti raamatute allalaadimise skript, ka nende sisselugemise ja puhastamise protsess vastavate kirjutatud tarkvara meetodite rakendamisega on Gutenbergi raamatute allalaadimise ja töötlemise lisas.

[31] https://openlibrary.org/developers/dumps
[32] http://www.lib.umich.edu/michigan-digitization-project
[33] http://pandora.nla.gov.au/
[34] https://www.gutenberg.org/

5. Andmebaas

Selles peatükis kirjeldatakse tööks vajaminevad andmebaasi tabelid, nende ülesehitust, eesmärke ja kasutust antud töös. Andmebaasi ER (*entity-relationship*) diagramm on töös kasutatud andmebaasi ER diagrammi lisas.

5.1 Treening- ja testandmed

Selles alampeatükis räägitakse järgmistest tabelist:

1. *BX-Book-Ratings* - hinnangute tabel,
2. *BX-Books* - raamatute tabel,
3. *gutenberg_books_catalogue* - gutenbergi raamatute kataloogitabel,
4. *analyzed_books* - analüüsitud raamatute tabel,
5. *training_euclidean_distances* - treeningerinevuste tabel,
6. *testing_euclidean_distances* - testimiserinevuste tabelis.

Hinnangute tabelis (*BX-Book-Ratings*) on imporditud BX andmestiku raamatute hinnangud. Neid hinnanguid kasutades leitakse raamatupaaride vahelised erinevused. Iga hinnangu objekti kirjeldavad järgmised veerud: kasutaja unikaalne identifitseerimisnumber (*User-ID*), ISBN (*ISBN*), raamatu hinnang (*Book-Rating*).

Raamatute tabelis (*BX-Books*) on imporditud BX andmestiku raamatud. Neid kasutatakse raamatu pealkirja ja autori leidmiseks ISBN järgi hinnangute tabelis. Igal raamatu objektil on järgmised veerud: ISBN (*ISBN*), raamatu pealkiri (*Book-Title*), raamatu autor (*Book-Author*), väljaande aasta (*Year-Of-Publication*), väljaandja (*Publisher*), raamatukaane pilt väike (*Image-URL-S*), raamatukaane pilt keskmine (*Image-URL-M*), raamatukaane pilt suur (*Image-URL-L*).

Gutenbergi raamatute kataloogitabelis (*gutenberg_books_catalogue*) on nimekiri projekti raamatutest, kus veerudeks on pealkiri (*title*) ja autor (*autho*r).

Analüüsitud raamatute tabelis (*analyzed_books*) on analüüsitud Gutenbergi raamatud koos väärtustatud atribuutide ja raamatu identifikaatoriga - pealkirja ja autoriga. Selles tabelis iga objekti veerud on vastavalt: pealkiri (*title*), autor (*author*), sõnade arv (*words*), tähemärkide arv (*characters*), lausete arv (*sentences*), keskmine lausete pikkus (*averSentenceLength*), keskmine sõnade pikkus (*averWordLength*), suhteline nimisõnade arv tekstis (*elNouns*), suhteline omadussõnade arv tekstis (*relAdjectives*), suhteline tegusõnade arv tekstis (*relVerbs*), eri nimede arv tekstis e. eri tegelaste arv (*names*), teksti positiivsustase (*positiveness*), unikaalsete geograafiliste kohtade arv (*locations*), mainitud kuupäevade arv (*dates*).

Treeningerinevuste tabel (*training_euclidean_distances*) on loodud hinnangutest arvutatud raamatutevaheliste kauguste salvestamiseks regressioonmudeli treenimise jaoks. Veerudeks on esimese raamatu ISBN number (*isbn1*), teise raamatu ISBN number (*isbn2*), nendevaheline erinevus numbrilises väärtuses (*distance*).

Testimiserinevuste tabelis (*testing_euclidean_distances*) on testimiseks raamatutevahelised kaugused. Tabel on sama ülesehitusega, nagu treenimiserinevuste tabel. Seda tabelit täidetakse osaga treenimiserinevuste andmetest. Treenimiserinevuste tabeli andmete põhjal treenitakse regressioonmudel, saadud mudeliga arvutatakse raamatute vahed ja neid vahesid võrreldakse testimiserinevuste tabeli andmetega. Veerudeks on esimese raamatu ISBN number (*isbn1*), teise raamatu ISBN number (*isbn2*), nendevaheline erinevus numbrilises väärtuses (*distance*).

5.2 Tulemused

Tulemuste salvestamiseks on disainitud järgmised tabelid:

1. *regression_parameters* - regressorite koefitsientide tabel,
2. *book_difference* - raamatuerinevuste tabelis,
3. *testing_results* - testimistulemuste tabel.

Regressorite koefitsientide tabel (*regression_parameters*) hoiab leidud regressorite parameetreid. Veerudeks on regressori nimetus (*regressor*) ja koefitsiendi väärtus (*parameter*).

Raamatuerinevuste tabelis (*book_difference*) on salvestatud juba leitud regressioonmudelit kasutades arvutatud raamatutevahelised erinevused. Veerudeks esimese raamatu ISBN (*isbn1*), teise raamatu ISBN (*isbn2*), erinevus (*difference*).

Testimistulemuste tabel (*testing_results*) sisaldab testimise tulemusi. See tabel on vajalik, sest teste rakendatakse eri treening- ja testimise raamatutevaheliste erinevuste jaotustega ehk ühel juhul võetakse kõikidest treeningerinevustest näiteks 15% ja kantakse üle testimiserinevustesse ning teisel korral 20%. Veerudeks on testimistulemuste rea identifitseerimisnumber (*id*), minimaalne (*min_difference*), maksimaalne vahe (*max_difference*), keskmine vahe (*aver_difference*), treeningerinevuste protsendimaht (*training_data_percentage*, pärast eraldamist testimiserinevutesse).

6. Õppimine

Antud peatükis kirjeldatakse regressorite koefitsientide otsimist ning selleks kasutatavat meetodit. Sellele eelneb vajalike treeningandmete arvutamine, mille protsessist on ka ülevaade antud. Lõpuks analüüsitakse saadud tulemusi.

6.1 Treeningandmete eeltöötlus

Lineaarse regressioonmudeli treenimiseks on vaja raamatutevahelisi erinevusi, mida saadakse töödeldud kasutajate hinnangutest. Samuti on vaja analüüsitud raamatuid, mida saadakse Gutenbergi projekti raamatukogumikust.

Treenimiseks vajalikud raamatuerinevused on Eukleidese kaugused raamatute vahel, kus mõõtmeteks on kasutajad, erinevused saadakse järgmiselt.

1. Iga raamatupaari puhul leitakse mõlema raamatu hinnangud, mida hindas üks kasutaja.
2. Erinevuse leidmiseks rakendatakse Eukleidese kauguse võrrandit

$$d = \sqrt{\sum_{1}^{n}(q_n - p_n)^2} \quad \text{, kus} \qquad [4]$$

d - otsitav raamatupaari vaheline erinevus

q – esimene raamat

p – teine raamat

q_n - *n* kasutaja hinnang esimesele raamatule

p_n - *n* kasutaja hinnang teisele raamatule

n – kasutajate arv, kes hindasid neid kahte raamatut

Kirjeldatud protseduur korratakse iga raamatupaari jaoks. Lõpuks saadakse treenimiseks ja testimiseks 1124 raamatu erinevust, nende hulgas erinevust 0 väärtusega on 526, mis moodustab peaaegu poole. Selliseid raamatuid, mis üksteise poolest üldse ei erine, tõenäoliselt kohtab väga harva, seetõttu on nulliga erinevuste väärtuslikus küsitav, aga kuna nad on kujunenud tõelistel hinnangutel, siis teostatakse õppimine ja testimine mõlemal juhul: kui on null-erinevused ja kui neid ei ole.

Raamatute analüüside saamiseks on alustatakse Gutenbergi projekti raamatute otsimisest. Iga raamat leitakse samal moel, nagu ka raamatute kataloogi loomisel. Pärast raamatu kindlakstegemist (failis eksisteerivad pealkiri, autor ja sisu), kontrollitakse, kas säärane raamat on kataloogis olemas (pärast selle puhastamist raamatutest, mis ei ühti BX raamatute nimekirjaga). Kui on, siis raamat analüüsitakse, selle atribuudid väärtustatakse, luuakse analüüsitud raamatu kirje ja salvestatakse analüüsitud raamatute tabelisse.

Kirjeldatud moel leitakse raamatupaaride vahelised kaugused (erinevused) ning kõik raamatud, mille vahel on olemas erinevus, analüüsitakse. Treeningandmete leidmiseks vajalikud koodid ja käivitatavad meetodid on antud treening- ja testandmete genereerimise lisas.

6.2 Meetod

Regressioonmudeli arvutamine käib vähimruutude meetodiga, mille mõte seisneb selles, et antakse ette jada sõltuvatest muutujatest ja nendele vastavatest väärtustatud sõltumatute muutujate (regressorite) jadadest, misjärel toimub sobivamate koefitsientide õppimine läbi ruutude summade minimiseerimise punktide vahel. Sisuliselt leitakse paremini sobiv joon $y = ax + b$ etteantud andmestiku jaoks.

$$min = \sum_{0}^{n} (y_n - (b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_k x_{n,k}))^2 \quad \text{, kus} \qquad [8]\ [4]$$

min – Viga, vahe otsitava ja ennustatava vahel, mida tuleb minimeerida

n – treeningjadade arv

k – regressorite arv

y_n – sõltuvad muutujad

b_k - regressiooni parameetrid (koefitsiendid)

$x_{n,k}$ - regressorid (mudeli faktorid), nad on maatriksrepresentatsioonis, seepärast *n* tähendab jada (rea) numbrit ning *k* veeru (regressori) numbrit.

Iga raamatupaari puhul sõltuvaks muutujaks on selle paari vaheline erinevus. Regressoriteks on raamatupaari atribuutidevahelised kaugused, mis leitakse järgmiselt: loetakse sisse mõlemad analüüsitud raamatud ning leitakse iga vastavate atribuutide paari absoluutvahe. Igast atribuudipaari vahest saab vastava regressori väärtus. Näiteks keskmise lause pikkuse regressori määramiseks leitakse absoluutvahe esimese ja teise raamatu keskmise lause pikkuste vahel.

6.3 Rakendus

Atribuutide koefitsientide treenimiseks on pärast andmete töötlemist säilinud 588 raamatu erinevust ja 195 raamatut. Iga raamatupaari saadud erinevused (sõltuvad muutujad) ning vastavatele raamatupaaridele saadud regressorite jadad vastandatakse ning antakse argumentidena üle vähimruutude meetodile, kus toimub regressoritele vastavate koefitsientide leidmine, selleks kasutatakse Apache Math Commons teeki. Saadud koefitsiendid salvestatakse andmebaasi.

Seda protseduuri korratakse 4 korda eri erinevuste jaotustega treening- ning testerinevuste vahel. Proovitakse treenida koefitsiente treening- ja testandmete järgmiste protsentuaalsete jaotustega: 90%/10%, 85%/15%, 80%/20%, 70%/30%. Tulemused esitatakse tabelina, kus kõik numbrilised väärtused on ümardatud kolme komakohani. Õppimiseks vajaminevad tarkvarameetodid on kirjeldatud regressorite parameetrite arvutamise lisas. Juhised selleks, kuidas uuesti luua teisiti jaotatud treeningandmed korduvaks õppimiseks on treening- ja testandmete genereerimise lisas.

6.4 Tulemused

Selles alampeatükis on toodud koefitsientide treenimise tulemused tabelites Tabel 2 ja Tabel 3 koos nende analüüsidega.

Tabel 2. Eri jaotuste tulemused on jaotatud veergude järgi. Esimeses reas on protsentuaalne jaotus, teises hinnangute arvud treenimiseks ja testimiseks, järgmised read on regressorid ja nende parameetrid. Selles on toodud nulle sisaldanud erinevustel arvutatud koefitsiendid.

Treeningandmed/Testandmed	**90%/10%**	**85%/15%**	**80%/20%**	**70%/30%**
Treeningandmed/Testandmed	**1012/112**	**955/169**	**899/255**	**787/337**
Regressor				
konstant	0,736	0,770	0,755	0,688
keskmine lausete pikkus	0,014	0,016	0,028	0,031
keskmine sõnade pikkus	0,319	0,246	0,104	0,020
tähemärkide arv	0,000	0,000	0,000	0,000
kuupäevade arv	0,001	0,001	0,001	0,001
kohtade arv	0,001	0,001	0,000	-0,001
tegelaste arv	-0,002	-0,002	-0,003	-0,003
sõnade arv	0,000	0,000	0,000	0,000
suhteline omadussõnade arv	-0,130	-0,122	-0,113	-0,131
suhteline nimisõnade arv	0,041	0,051	0,053	0,059
suhteline tegusõnade arv	0,021	-0,012	-0,010	-0,015
lausete arv	0,000	0,000	0,000	0,000
positiivsus	0,813	1,047	0,578	0,884

Tulemustest on näha, et kui treenida regressorite koefitsiente erinevuste peal, mis sisaldavad nulle, siis tähtsamateks regressoriteks võib nimetada keskmist sõnade pikkust ja teksti positiivsust. Seejuures negatiivsed koefitsiendid sisuliselt automaatselt vähendavad raamatute erinevust, kuigi nende mõju on peaaegu võrdne nulliga, välja arvatud suhtelise omadussõnade arvu näitaja puhul, mis võrreldes teiste koefitsientidega omab märkimisväärset mõju. Sealhulgas, sõnaliikidest ainult nimisõnade suhteline arv kõikidel juhtudel suurendab erinevust raamatute vahel. Tasub märkida, et mida rohkem andmeid eraldati testandmeteks, seda rohkem kasvasid mõnede regressorite tähtsused ja seda enam vähenesid teised. Näiteks, keskmine lausete pikkuse tähtsus lõpuks suurenes umbes kaks korda võrreldes esialgse näitajaga, samas keskmise sõnade pikkuse tähtsus kahanes drastiliselt, ja teksti positiivsuse mõju varieerus.

Järgmisena proovitakse treenida koefitsiendid samadel erinevustel, kuid ilma nendeta, mis võrduvad nulliga.

Tabel 3. Tabeli ülesehitus on analoogne Tabeliga 2. Selles on toodud nulle mittesisaldanud erinevustel arvutatud koefitsiendid.

Treeningandmed/Testandmed	**90%/10%**	**85%/15%**	**80%/20%**	**70%/30%**
Treeningandmed/Testandmed	**538/60**	**508/90**	**478/120**	**419/179**
Regressor				
konstant	2,101	2,142	2,127	0,789
keskmine lausete pikkus	-0,018	-0,022	-0,022	0,016
keskmine sõnade pikkus	-0,055	-0,061	-0,123	0,227
tähemärkide arv	0,000	0,000	0,000	0,000
kuupäevade arv	0,000	0,000	0,000	0,006
kohtade arv	0,002	0,002	0,002	0,001
tegelaste arv	-0,002	-0,002	-0,003	-0,002
sõnade arv	0,000	0,000	0,000	0,000
suhteline omadussõnade arv	0,001	-0,024	-0,030	-0,121
suhteline nimisõnade arv	0,013	0,011	0,014	0,053
suhteline tegusõnade arv	0,006	0,022	0,017	-0,019
lausete arv	0,000	0,000	0,000	0,000
positiivsus	0,778	0,979	1,345	1,074

Juhul, kui elimineerida treeningandmetest need raamatute erinevused, mis võrduvad nulliga, siis üldiselt kõikide regressorite tähtsused kasvavad. Näiteks, konstandiks saadakse kolmel juhul 2, võrreldes umbkaudse 0,7-ga, mida võib Tabelis 2 näha. Sellised regressorid nagu kuupäevade, kohtade ja tegelaste arv jäid tähtsusetuks. Suhteliste omadussõnade, nimisõnade ja tegusõnade mõju vähenes. Olulisteks regressoriteks siinpuhul jäid positiivsus ja keskmine sõnade pikkus. Vastupidiselt Tabelis 2 toodud tulemustele Tabelis 3 testimiseks eraldatud andmete kasvuga keskmise sõna pikkuse tähtsus aina kasvas. Nulle mittesisaldaval andmestikul treenitud koefitsientide tulemused on usutavamad, sest, et null-erinevusega raamatute olemasolu tõenäosus on väga väike, aga andmestikus oli selliseid umbes pool. Samas, nii Tabelist 2 kui Tabelist 3 võib näha, et suurem osa atribuutidest omavad naeruväärselt väikse mõju raamatute erinevuste määramisel, suurema osa otsustavad positiivsus, suhteline nimisõnade arv, keskmine sõnade pikkus ja konstant.

7. Testimine

Kõigepealt õppimise käigus arvutatud koefitsientide abil leitakse erinevused kataloogi jäänud raamatute vahel. Töö tulemuste testimiseks kasutatakse kahte meetodit. Esiteks, leidud regressioonmudeli abil arvutatud raamatuerinevused võrreldakse testandmete omadega ja arvutatakse minimaalset erinevust testandmetega, maksimaalset erinevust, keskmist erinevust, dispersiooni ja mediaani. Võrdlused toimuvad samade treening- ja testandmete jaotustega, nagu ka regressioonmudeli treenimises. Teiseks, sooritatakse valikulised subjektiivsed testid: kahe raamatu jaoks võetakse neli lähimat raamatut ja hinnatakse sarnasust. Säärane võrdlemine on aga raskendatud seetõttu, et paljud raamatud, mis võiksid olla tõesti sarnased, puuduvad selles töös kasutatud raamatute kogumikust. Veel üheks variandiks oleks võrrelda soovituste tulemusi mõne interneti soovitussüsteemiga, kuid seda teeb väga keerukaks asjaolu, et paljud raamatud, mis internetisüsteemis olemas on, antud töös kasutatavast kogumikust puuduvad. Kõik tabelites toodud numbrid on ümardatud kolme komakohani. Testimiseks kasutatavad tarkvarameetodid ja MySQL käsud on kirjeldatud regressorite parameetrite arvutamise lisas.

Tabel 4. Tabeli ülesehitus on analoogne Tabeli 3 omaga, ainult regressorite asemel on testimistulemuste punkthinnangud. Selles on toodud nende regressorite testimiste tulemused, mis olid treenitud nulle sisaldavatel erinevustel.

Treeningandmed/Testandmed	**90%/10%**	**85%/15%**	**80%/20%**	**70%/30%**
Treeningandmed/Testandmed	**1012/112**	**955/169**	**899/255**	**787/337**
Punkthinnang (võrreldes testandmetega)				
minimaalne vahe	0,006	0,001	0,007	0,019
maksimaalne vahe	4,781	4,879	4,563	4,904
keskmine vahe	1,081	1,071	1,102	1,167
dispersioon	0,832	0,765	0,728	0,890
mediaan	0,925	0,901	0,901	1,005

Nagu näha, nulle sisaldava andmestiku peal treenitud mudeli tulemuste keskmine vahe on umbes 1, dispersioon umbes 0,8 ja mediaan 0,9. Võttes keskmist, mudel eksis umbes 11% võrra hinnangust, mis sellist andmemahtu arvestades on üsna hea tulemus. Seejuures minimaalne leitud on nullilähedane ning maksimaalne üsna suur - umbes 4,5. Lõpuks, treeningandmete eraldusega testandmeteks näitajad ei kasva ega kahane stabiilselt, vaid varieeruvad.

Selleks, et näha, milliseid raamatuid mudel soovitaks, siis edukaima jaotuse jaoks võetakse kaks raamatut: Charles Dickens'i "Oliver Twist" ning Jack Londoni "Martin Eden". Nendele leitakse 4 lähimat raamatut ning nende erinevus. Soovitusi ülejäänud jaotuste jaoks võib leida testimise soovitatavate raamatute tabelite lisast.

Tabel 5. Tabeli 4 85% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“The Varieties of Religious Experience”, William James	-0,021
	“The Education of Henry Adams”, Henry Adams	0,264
	“The Iliad of Homer”, Homer	0,562
	“The Dead Secret”, Wilkie Collins	0,570
“Martin Eden”, Jack London	“The Varieties of Religious Experience”, William James	-0,041
	“The Education of Henry Adams”, Henry Adams	0,598
	“Howards End”, E. M. Forster	0,652
	“Middlemarch”, George Eliot	0,668

Tabel 6. Tabeli ülesehitus on analoogne Tabeli 4 omaga. Antud tabelis on toodud nende regressorite testimiste tulemused, mis olid treenitud nulle mittesisaldavatel erinevustel.

Treeningandmed/Testandmed	**90%/10%**	**85%/15%**	**80%/20%**	**70%/30%**
Treeningandmed/Testandmed	**538/60**	**508/90**	**478/120**	**419/179**
Punkthinnang (võrreldes testandmetega)				
minimaalne vahe	0,014	0,006	0,012	0,000
maksimaalne vahe	2,964	2,932	3,242	4,892
keskmine vahe	0,883	0,932	0,965	1,061
dispersioon	0,385	0,378	0,465	0,740
mediaan	0,949	0,977	0,986	0,919

Mudel, mille treenimisel jäeti välja nulle sisaldavad erinevused, annab märkimisväärselt täpsemaid tulemusi. Selle mudeli maksimaalne vahe, keskmine vahe, dispersioon ja mediaan kõik annavad paremaid tulemusi kui Tabelis 4. Ainus näitaja, mis pisut kasvas, on minimaalne vahe, kuid võrreldes kasvanud täpsusega üle kõigi andmete, selle üksiku erinevuse näitaja olulisus on väike. Tasub mainida, et võrreldes Tabeliga 4, selles korrelatsioon täpsuse ning treeningandmete vähenemise suhtes on otsene. Treeningandmete kogu-se kahanedes ning testandmete kasvus mudeli üldine täpsus langeb, mida näitab kõikide täpsusnäitajate vastav muutus (v.a minimaalne vahe). Nullideta erinevustel treenitud mudel suurema treeningandmete kogusega andis keskmise vahe testtulemustega umbes 0,9, mis on umbes 10% võimalikust vahest, see on 1% vähem kui eelmises tabelis. Dispersioon on koguni kaks korda väiksem, mida illustreerib ka maksimaalse vahe langus. See viitab

sellele, et kvaliteetsete treeningandmete mahu kasvudes hakkab kindlasti kasvama ka mudeli täpsus. Kui kasutajate hinnangute vahed korreleerivad erinevustega raamatute vahel, siis võib väita, et piisava kvaliteetse andmemahu juhul mudel oleks suuteline raamatuid adekvaatselt soovitama.

Tabel 7. Tabeli 6 90% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	**Soovitatavad**	**Erinevus**
"Oliver Twist", Charles Dickens	"Memoirs of Fanny Hill", John Cleland	1,479
	"The Education of Henry Adams", Henry Adams	1,556
	"War and Peace", Leo Tolstoy	1,567
	"Anne of Avonlea", Lucy Maud Montgomery	1,659
"Martin Eden", Jack London	"Memoirs of Fanny Hill", John Cleland	1,312
	"The Education of Henry Adams", Henry Adams	1,517
	"War and Peace", Leo Tolstoy	1,563
	"Middlemarch", George Eliot	1,579

Nagu Tabelitest 5 ja 7 näha, hetkel mõlemal juhul sisu poolest soovitatavad raamatud ei ole kõige sobivamad, näiteks "Oliver Twist'ile" soovitatakse "The Iliad of Homer". Samas, testimistulemused näitavad, et mudeli kõrvalekalle väga suur ei ole ning antud treeningandmete piires töötab mudel korrektselt. See tähendab, et säärased tulemused on tingitud treeningandmete mahust ja kvaliteedist. Hetkel oli treenimiseks nullita hinnanguid 538 hinnangut ja kõigest 195 raamatut, mis teeb keskmiselt 2,76 hinnangut raamatu kohta, mis on väga vähe. Küll aga võib öelda, et antud töös kasutatud atribuutide hulgast raamatute erinevuste hindamisel tõenäoliselt mängivad suuremat rolli keskmine sõnade pikkus ning teksti positiivsus. Üllatav on aga see, et tähemärkide arv, teisisõnu raamatu maht, rolli ei mänginud.

Kui hinnanguid ning raamatuid oleks rohkem, ja nende andmetega poleks probleeme, mis esinesid antud töös kasutatud andmestikega, võib väita, et siis sisulised raamatutesoovitused oleksid märksa tulemuslikumad. See tähendab, et antud töös kasutatud lähenemisel on potentsiaali ja seda võiks kindlasti proovida suuremate andmekogudega. Samuti, hetkeseisuga andmestiku vähesuse tõttu ei saa kinnitada või ümber lükata väidet, et kasutajate hinnangute vahed kirjeldavad mingil määral raamatute erinevust.

8. Kokkuvõte

Selle töö eesmärgiks oli kirjutada raamatu analüsaator ja leida mudel, mis töötades analüsaatoriga oleks võimeline hindama raamatute erinevust ja nende kaudu leidma sarnasemaid toetudes raamatuid kirjeldavate atribuutide analüüsile. Töö tulemusena valminud mudeli täpsus testandmete suhtes on üsna suur (ennustatud erinevuste kõrvalekalle tõelistest keskmiselt 10%) ning etteantud treeningandmete piires on ta võimeline leidma korrektseid raamatute erinevusi. Seejuures töö võimalused olid paljuski piiratud ajapiiri ning treening- ja testandmete kättesaadavusega. Raamatuid kirjeldavaid atribuute on piiratud arv, ka treeningandmeid jäi väheseks, seetõttu töö käigus arvutatud mudel ei ole võimeline tegema inimvaatepunktist korrektseid soovitusi. Kõrvaltulemustena saadi ülevaate tööst BX andmestiku ja Gutenbergi projekti raamatutega, nendega esinenud probleemide ja töötlusvõimalustega. Samuti leiti, et etteantud atribuutide hulgast on inimestele olulisemad raamatu hindamisel teksti positiivsus ning keskmine sõnade pikkus, samas kui raamatu maht erilist rolli ei mängi.

Kokkuvõtvalt võib öelda, et testimistulemused näitasid, et antud lähenemisel on potentsiaali suurema ja kvaliteetsema treeningandmestiku puhul. Samuti võib sisse tuua uusi atribuute, mis võiksid suurendada mudeli paindlikust. Paremini treenitud mudel võiks olla lahenduseks, näiteks, külma stardi probleemile uutes soovitussüsteemides või olla osaks hübriidsest soovitussüsteemist. Lisaks, võib antud töös kasutatud lähenemisega uurida täpsemalt eri inimeste gruppide jaoks olulisemaid atribuute raamatute hindamisel ja läbi selle, näiteks, koostada teismelistele sobivama kohustusliku kirjanduse nimekirja.

9. Tsiteeritud teosed

[1] B. S. a. J. Y. Greg Linden, „Amazon.com Recommendations Item-to-Item Collaborative Filtering,“ February 2003. [Võrgumaterjal]. Available: http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf. [Kasutatud 14 05 2015].

[2] S. Sahebi ja W. W. Cohen, „Community-Based Recommendations: a Solution to the,“ [Võrgumaterjal]. Available: http://www.dcs.warwick.ac.uk/~ssanand/RSWeb11/7Sahebi.pdf. [Kasutatud 14 05 2015].

[3] R. v. Meteren ja M. v. Someren, „Using Content-Based Filtering for Recommendation,“ [Võrgumaterjal]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.5743&rep=rep1&type=pdf. [Kasutatud 14 05 2015].

[4] „Multiple Linear Regression,“ 1997-1998. [Võrgumaterjal]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm. [Kasutatud 14 05 2014].

[5] A. Warriner ja V. B. M. Kuperman, „Norms of valence, arousal, and dominance,“ 2013. [Võrgumaterjal]. Available: http://crr.ugent.be/papers/Warriner_et_al_affective_ratings.pdf. [Kasutatud 14 05 2015].

[6] „About ISBN and ASIN,“ [Võrgumaterjal]. Available: http://www.amazon.co.uk/gp/help/customer/display.html?nodeId=898182. [Kasutatud 14 05 2015].

[7] E. W. Weisstein, „Distance.,“ [Võrgumaterjal]. Available: http://mathworld.wolfram.com/Distance.html. [Kasutatud 14 05 2015].

[8] „Linear Regression,“ 1998. [Võrgumaterjal]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm. [Kasutatud 14 05 2015].

Lisad

I. BX andmestiku töötlemine

Siin antakse juhised, kuidas kasutades arendatud tarkvara, töödelda BX andmestiku selleks, et ta oleks kõlblik kasutamaks antud töö eesmärkide saavutamiseks.

Enne tabelite importimist tuleb muuta nende tabelite loomisskripti 21. rida, asendada seal sõna "TYPE" sõnaga "ENGINE", sest "TYPE" on mõeldud vanemate MySQL versioonide jaoks ja 5.5-ga tööle ei hakka.

Algul on tabelis *BX-Books* 271379 raamatut ja tabelis *BX-Book-Ratings* 1149780 hinnangut. Andmete töötlemiseks on vajalikud järgnevad sammud.

1. Kustutada tabelist *BX-Book-Ratings* hinnangud nende ISBN-dega, kus on ISBN-d, mis pole *BX-Books*'is. Selleks MySQL jaoks käivitada käsu:

```
DELETE FROM `BX-Book-Ratings`
WHERE `ISBN` NOT IN (
SELECT `ISBN` FROM `BX-Books`);
```

   Kustutatud 118605 kirjet, jäänud 1031175 hinnangut.

2. Kustutada tabelist *BX-Book-Ratings* hinnangud, mis võrduvad nulliga, käsuga:

```
DELETE FROM `BX-Book-Ratings`
WHERE `Book-Rating` = 0;
```

   Kustutatud 647323 hinnangut, järele jääb 383852.

3. Kustutada tabelist *BX-Books* kõik raamatud, mis pole ühised *gutenberg_books_catalogue* tabeli kirjetega. Järgneva käsu käivitamiseks peavad enne olema sisse loetud Gutenbergi projekti raamatud.

```
DELETE FROM `BX-Books` WHERE (`Book-Title`,`Book-Author`) NOT IN
(SELECT title,author FROM gutenberg_books_catalogue);
```

   Vastavalt, kustutada hinnangud, mis jäid oma raamatutest ilma.

```
DELETE FROM `BX-Book-Ratings`
WHERE ISBN NOT IN (
SELECT ISBN FROM
`BX-Books`);
```

   Kustutatud raamatuid - 270263, järele jäänud 1116. Kustutatud hinnanguid 382263, järele jäänud 1589.

4. Nüüd tuleb elimineerida hinnangud korduvatele raamatutele. Selleks tuleb kävitada klassi `BXRatingModel` meetod `clearDuplicateRatings`.

   Kustutatud 26 hinnangut. Järele jäänud 1563.

5. Nüüd tuleb kustutada korduvad raamatud ning korduvate raamatute hinnangud omastada jäänud raamatutele. Selleks rakendada klassi `BXRatingModel` meetod `reassignDuplicateBooksRatingsAndDeleteBooks`.

   Pärast käsu käivitamist hinnangute arv ei muutunud (1563), millest võib järeldada, et kõik töötas õigesti. Raamatuid aga oli kustutatud 659 ning järel on 457.

6. Nüüd tuleb kustutada need hinnangud, mille kasutajal on need hinnangud ainsad. Selle tarvis käivitada MySQL käsk:

```
DELETE FROM `BX-Book-Ratings` WHERE `User-ID` IN
(SELECT `User-ID` FROM
(SELECT * FROM `BX-Book-Ratings`
GROUP BY `User-ID`
HAVING COUNT(`User-ID`) < 2) as tempTable);
```

Kustutatud 975 hinnangut, järel 588.

7. Lõpuks kustutada kõik hinnanguteta jäänud raamatud MySQL käsuga:

```
DELETE FROM `BX-Books`
WHERE ISBN NOT IN
(SELECT ISBN FROM `BX-Book-Ratings`);
```

Kustutatud 262 raamatut, järel 195.

II. Gutenbergi raamatute allalaadimine ja töötlemine

Gutenbergi raamatukogumiku saab alla laadida kasutades tarkvara GNU Wget[35]. Skript Gutenbergi inglisekeelsete raamatute allalaadimiseks html formaadis[36] [37]:

```
wget -H -w 2 -m
http://www.gutenberg.org/robot/harvest?filetypes[]=html&langs[]=en \
--referer="http://www.google.com" \
--user-agent="Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.6)
Gecko/20070725 Firefox/2.0.0.6" \
--header="Accept:
text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,
image/png, /*;q=0.5" \
--header="Accept-Language: en-us,en;q=0.5" \
--header="Accept-Encoding: gzip,deflate" \
--header="Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7" \
--header="Keep-Alive: 300"
```

Sellega laaditakse alla umbes 65GB raamatuid arhiveeritud näol.

Allalaaditud kogumiku töötlemine töös kasutamiseks hõlmab järgmisi samme (selle töötamiseks peab olema raamatukogumiku kaust projekti juurkaustas):

1. Käivitada klassi `GutenbergBookModel` meetod `startTraverseAndCatalogueCreation`.

   Pärast operatsiooni lõppu on kataloogis 32721 raamatut.

2. Kustutada raamatud, mis pole Gutenbergi kataloogil BX andmestikuga ühised (koodi korrektseks tööks peab olema BX kogumiku töötlemine valmis), selleks käivitada järgmine MySQL kood:

```
DELETE FROM gutenberg_books_catalogue WHERE (`title`,author) NOT IN
(SELECT `Book-Title`,`Book-Author` FROM `BX-Books`);
```

   Kustutatud 32526 raamatut. Järel 195 raamatut. See käsk võtab umbes 3500 sekundit tööks Intel Core i5 protsessori ja 8GB RAM mäluga.

[35] http://www.gnu.org/software/wget/

[36] http://www.gutenberg.org/wiki/Gutenberg%3aInformation_About_Robot_Access_to_our_Pages#Getting_All_EBook_Files

[37] http://webapps.stackexchange.com/questions/12311/how-to-download-all-english-books-from-gutenberg

III. Treening- ja testandmete genereerimine

Selles lisas on juhised, kuidas tekitada treening- ja testandmed. Seda tuleb teha pärast seda kui imporditud andmed (BX kogumik) on puhastatud ja töödeldud.

Kõigepealt tuleb leida raamatupaaride erinevused ja salvestada need *training_euclidean_distances* tabelisse BX andmestiku kasutades. Selleks käivitada klassi `TrainingEuclideanDistanceModel` meetod `findAndInsertBXEuclideanDistances`. Sellega oli loodud 1124 kaugust. Järgmise sammuna tuleb osa erinevuste kirjesid eraldada testandmeteks. Selleks MySQL käsud:

1. *n*-i asendada soovitud arv kirjeid, mida eraldada.

```
INSERT INTO testing_euclidean_distances(
SELECT isbn1, isbn2, distance
FROM book_recommender_data.training_euclidean_distances
LIMIT n);
```

2. Nüüd kustutada treeningvahede tabelist kõik eraldatud kirjed.

```
DELETE FROM training_euclidean_distances
WHERE (isbn1,isbn2) IN (
SELECT isbn1,isbn2
FROM testing_euclidean_distances);
```

Et teostada koefitsientide treenimist ja testimist juhul, kus 0-ga raamatuerinevusi ei ole, tuleb nendest lahti saada järgmise käsuga:

```
DELETE
FROM training_euclidean_distances
WHERE distance = 0;
```

Kui seda rakendada, siis kustutatakse 526 hinnangut ning järele jääb 598.

Enne kui saab analüüsida raamatute positiivsust, peab afektiivsete hinnangutega sõnadest tekitama algvormid ja salvestama nad eraldi faili, mida arendatud raamatu analüsaator kasutama hakkab. Selleks on vaja kävitada `AnalyzedBookModel` klassi meetod `stemValencedWords`.

Nüüd on võimalik hakata analüüsima Gutenbergi kataloogi jäänud raamatud ning tulemused salvestama tabelisse *analyzed_book*. Protsessi teostamiseks on vaja käivitada klassi `AnalyzedBookModel` meetod `startTraverseAndAnalysis`. Analüüsi lõpuks on analüüsitud 195 raamatut, mis ühtib kataloogi raamatute arvuga.

Kui on vaja luua raamatute treening- ja testerinevused otsast peale, siis tuleb korrata vastavad antud lisa juhised, aga enne seda puhastada erinevuste tabelid:

```
DELETE FROM training_euclidean_distances;
DELETE FROM testing_euclidean_distances;
```

IV. Regressorite parameetrite arvutamine

Olemasolevatel treeningandmetel atribuutide koefitsientide arvutamiseks ja salvestamiseks tabelisse *regression_parameters* käivitada klassi `RegressorParameterModel` meetodi `calculateAndSaveRegressorParameters`.

Et saadud regressorite parameetreid testida (võrrelda *test_euclidean_distances* tabeli kirjetega), tuleb kõigepealt kasutades saadud koefitsiente arvutada raamatuvahed, mida võrdlema hakatakse. Selleks on klassi `BookDifferenceModel` meetod `calculateAndSaveBookDifferences`, see meetod arvutab erinevused ja salvestab nad andmebaasi tabelisse *book_difference*, saadakse 37830 erinevust. Vahetult testimiseks on vaja käivitada klassi `TestingResultsModel` meetodi `compareResultsWithTestData`, seejuures meetodis real 173 tuleb manuaalselt panna kui suur protsent eraldati treeningandmetest:

```
testResult.setTrainingDataPercentage(soovitudProtsent);
```

Tulemused salvestatakse tabelisse *testing_results*.

Konkreetsele raamatule saab lähimaid leitud raamatuid leida järgmiste käsudega. *isbn1* ja *isbn2* võib vahetada omavahel kohtadega, sest kui ei leidu ühtepidi erinevust, võib leiduda teistpidi.

1. Kasutades ISBN-i

```
SELECT * FROM book_difference
JOIN `BX-Books`
ON book_difference.isbn1 = `BX-Books`.ISBN
WHERE isbn2 = isbn
ORDER BY difference ASC;
```

2. Kasutades pealkirja ja autorit

```
SELECT * FROM book_difference
JOIN `BX-Books`
ON book_difference.isbn1 = `BX-Books`.ISBN
WHERE isbn2 = (
SELECT isbn
FROM `BX-Books`
WHERE `Book-Title` = pealkiri
AND `Book-Author` = autor)
ORDER BY difference ASC;
```

V. Testimise soovitatavate raamatute tabelid

Tabel 8. Tabeli 4 90% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“Babbitt”, Sinclair Lewis	0,564
	“Child of Storm”, H. Rider Haggard	0,689
	“Greenmantle”, John Buchan	0,697
	“Anne of Avonlea”, Lucy Maud Montgomery	0,742
“Martin Eden”, Jack London	“The Varieties of Religious Experience”, William James	0,036
	“The Education of Henry Adams”, Henry Adams	0,617
	“Middlemarch”, George Eliot	0,643
	“The Rise of Silas Lapham”, William Dean Howells	0,643

Tabel 9. Tabeli 4 80% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“The Varieties of Religious Experience”, William James	-0,079
	“The Education of Henry Adams”, Henry Adams	0,249
	“Babbitt”, Sinclair Lewis	0,442
	“The Dead Secret”, Wilkie Collins	0,509
“Martin Eden”, Jack London	“The Varieties of Religious Experience”, William James	-0,009
	“L'Assommoir”, Emile Zola	0,622
	“Middlemarch”, George Eliot	0,632
	“Howards End”, E. M. Forster	0,657

Tabel 10. Tabeli 4 70% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“The Varieties of Religious Experience”, William James	-0,167
	“Babbitt”, Sinclair Lewis	0,222
	“The Education of Henry Adams”, Henry Adams	0,310
	“The Dead Secret”, Wilkie Collins	0,371
“Martin Eden”, Jack London	“The Varieties of Religious Experience”, William James	-0,123
	“Middlemarch”, George Eliot	0,477
	“The Golden Bowl”, Henry James	0,522
	“L'Assommoir”, Emile Zola	0,532

Tabel 11. Tabeli 6 85% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“Memoirs of Fanny Hill”, John Cleland	1,401
	“The Education of Henry Adams”, Henry Adams	1,432
	“War and Peace”, Leo Tolstoy	1,483
	“Anne of Avonlea”, Lucy Maud Montgomery	1,683
“Martin Eden”, Jack London	“Memoirs of Fanny Hill”, John Cleland	1,155
	“The Education of Henry Adams”, Henry Adams	1,381
	“War and Peace”, Leo Tolstoy	1,521
	“Swann's Way”, Swann's Way	1,539

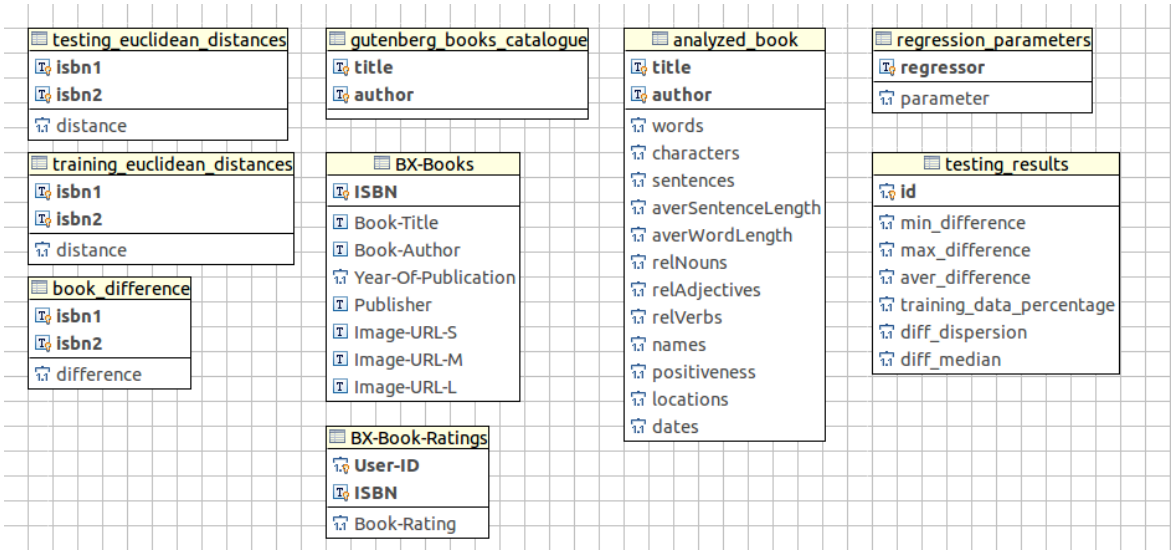
Tabel 12. Tabeli 6 80% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“Memoirs of Fanny Hill”, John Cleland	1,403
	“The Education of Henry Adams”, Henry Adams	1,432
	“War and Peace”, Leo Tolstoy	1,483
	“Anne of Avonlea”, Lucy Maud Montgomery	1,683
“Martin Eden”, Jack London	“Memoirs of Fanny Hill”, John Cleland	1,141
	“The Education of Henry Adams”, Henry Adams	1,292
	“War and Peace”, Leo Tolstoy	1,493
	“Middlemarch”, George Eliot	1,517

Tabel 13. Tabeli 6 70% treeningandmete jaotust sisaldavate tulemustega tehtud nelja sarnasema raamatu esiletoomine.

Raamat	Soovitatavad	Erinevus
“Oliver Twist”, Charles Dickens	“The Varieties of Religious Experience”, William James	0,235
	“The Education of Henry Adams”, Henry Adams	0,272
	“The Iliad of Homer”, Homer	0,558
	“The Dead Secret”, Wilkie Collins	0,573
“Martin Eden”, Jack London	“The Varieties of Religious Experience”, John Cleland	-0,053
	“The Education of Henry Adams”, Henry Adams	0,608
	“Howards End”, E. M. Forster	0,662
	“Middlemarch”, George Eliot	0,669

VI. Töös kasutatud andmebaasi ER diagramm



Joonis 1. Töös kasutatud andmebaasi ER diagramm

Töös kasutatud andmebaasi loomise skript on projekti juurkaustas failina *book_recommender_mysqldump.sql* .

VII. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina **Roman Šumailov** (sünnikuupäev: 19.07.1993)
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Objektiivne raamatute soovitussüsteem,
(*lõputöö pealkiri*)

mille juhendaja on Siim Karus,
(*juhendaja nimi*)

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

Tartus, **14.05.2015**