Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

Automaatse märksõnastamise KRATT: detailanalüüs

Eesti Rahvusraamatukogu riigihange „Automaatse märksõnastamise KRATT: detailanalüüs, sh prototüübi loomine“ (viitenumber 212433)

Autorid:
Marit Asula
Erik Jürmann
Hele-Andra Kuulmets
Raul Sirel
Silver Traat
Kristiina Vaik

TEXTA OÜ
Tallinn 2019

Sisukord

Sissejuhatus

Käesolev dokument sisaldab teavikute märksõnastaja prototüübi detailanalüüsi. Kirjeldatud on märksõnastaja prototüübi aluseks olevad lähteandmed, kasutatud algoritmid, metoodikad ja tarkvaralahendused, märksõnastamise kvaliteedi analüüsimisel kasutatavad mõõdikud ning rakenduse disain, liidestused, API-d. Lisaks antakse ülevaade arendatud prototüübist ning kasutajate hinnagutest prototüüprakenduse kasutatavusest.

1. Terminoloogia

Käesoleva peatüki eesmärk on anda ülevaade dokumendis kasutatud terminoloogiast ning kirjeldada vähem intuitiivseid graafikute tõlgendamise aspekte.

1.1. Notatsioon ja mõisted

Matemaatilised ja statistilised terminid

Sümbol	**Nimetus**	**Kirjeldus**
μ	Keskväärtus	Iseloomustab kõige suurema tõenäosusega esinevat väärtust.
σ	Standardhälve	Iseloomustab, kui palju erinevad jaotuses olevad väärtused keskväärtusest ehk mida väiksem on standardhälve, seda enam on väärtused koondunud keskväärtuse ümber ja vastupidi - mida suurem on standardhälve, seda suurem on väärtuste hajuvus.
Mo	Mood	Kõige sagedamini esinev väärtus.
Med	Mediaan	Jaotuse andmepunktide keskmine väärtus (ehk väärtus, millest mõlemal pool asub võrdses koguses andmepunkte).
Q1	Esimene kvartiil	Andmepunkt jaotuses, millele eelneb 25% kõigist jaotuse andmepunktidest.
Q2	Teine kvartiil	Andmepunkt jaotuses, millele eelneb 50% kõigist jaotuse andmepunktidest (mediaan).
Q3	Kolmas kvartiil	Andmepunkt jaotuses, millele eelneb 75% kõigist jaotuse andmepunktidest.
Q4	Neljas kvartiil	Andmepunkt jaotuses, millele eelneb 100% kõigist jaotuse andmepunktidest.

IQR	Kvartiilide vahe	Kolmanda ja esimese kvartiili vahe (Q3-Q1).
out	Erind	Andmepunkt, mis on väiksem kui Q1-1.5*IQR või suurem kui Q3+1.5 IQR. Lihtsustatult tähendab see mediaanist tugevalt hälbivat andmepunkti. NB! Antud dokumendis andmete kirjeldamisel on terminit kasutatud valdavalt ülejäänud jaotusest tugevalt hälbivate andmepunktide kohta, mis joonistel märgatavalt ülejäänud jaotusest erinevad.
μ+	Keskväärtus erinditega	Keskväärtus, mille arvutamisesse on kaasatud erindid.
σ+	Standardhälve erinditega	Standardhälve, mille arvutamisesse on kaasatud erindid.
μ-	Keskväärtus erinditeta	Keskväärtus, mille arvutamisesse pole kaasatud erindeid.
σ-	Standardhälve erinditeta	Standardhälve, mille arvutamisesse pole kaasatud erindeid.

Andmeallikad

EMS	Eesti märksõnastik
ERB	Eesti rahvusbibliograafia
DIGAR	RR digitaalarhiiv DIGAR
ESTER	ELNET Konsortsiumi ühine e-kataloog

Tõesusklassid

Sümbol	Termin	Definitsioon	Näide
TP	Tõsipositiiv (*true positive*)	Õigesti klassifitseeritud positiivne märgend.	Dokumendiga on seotud märgend "kalandus" ning märgendaja lisab vastava märgendi.
FP	Väärpositiiv (*false positive*)	Valesti klassifitseeritud positiivne märgend.	Dokumendiga ei ole seotud märgend "ilukirjandus", ent märgendaja lisab selle.
TN	Tõsinegatiiv (*true negative*)	Õigesti klassifitseeritud negatiivne märgend.	Dokumendiga ei ole seotud märgend "ameerika" ning märgendaja ei lisa seda.
FN	Väärnegatiiv (*false negative*)	Valesti klassifitseeritud negatiivne märgend.	Dokumendiga on seotud märgend "keskkond", ent märgendaja ei lisa seda.

Masinõppe mõõdikud

Sümbol	Nimetus	Valem	Kirjeldus
TPR	Tõsipositiivide määr (*true positive rate*)	TP/(TP+FN)	Tõsipositiivide osakaal kõikide positiivsete elementide hulgas.
FPR	Väärpositiivide määr (*false positive rate*)	FP/(TN+FP)	Väärpositiivide osakaal kõikide negatiivsete elementide hulgas.
	Täpsus (*precision*)	TP/(TP+FP)	Mõõdik mudelite kvaliteedi hindamiseks, mis väljendab ennustatud tõeste märgendite suhet kõikide ennustatud märgenditega (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%).
	Saagis[1] (*recall*)	TP/(TP+FN)	Mõõdik mudelite kvaliteedi hindamiseks, mis väljendab ennustatud tõeste märgendite suhet kõikide relevantsete märgenditega (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%).
	F1-skoor	2*(Täpsus*Saagis)/ (Täpsus+Saagis)	Mõõdik mudelite kvaliteedi hindamiseks, mis kombineerib täpsuse ja saagise (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%).

Üldised definitsioonid

- **Juhendatud masinõpe** (*supervised machine learning)* - Mudeldamine märgendatud andmestikul.

- **Juhendamata masinõpe** (*unsupervised machine learning*) - Mudeldamine märgendamata andmestikul.

- **Näitepõhine hindamine** (*example-based evaluation*) - Mudelite hindamine kasutades hindamise aluseks dokumente ehk hinnatakse, kui hästi on mudel keskmiselt võimeline ennustama iga dokumendiga seotud märgendite hulka.

- **Märgendipõhine hindamine** (*label-based evaluation*) - Mudelite hindamine kasutades hindamise aluseks dokumentidega seotud märgendeid ehk hinnatakse, kui hästi on mudel keskmiselt võimeline ennustama iga märgendi sobivust juhuslikule testanmestiku dokumendile.

[1] Saagis on tegelikult sama, mis tõsipositiivide määr, aga kuna need terminid on antud dokumendis kasutusel pisut erinevates kontekstides, on need tabelis lahku löödud.

- **Segment** - Automaatsete vahenditega failist eraldatud lehekülg. Täpsuse huvides ei kasutata terminit “lehekülg” sünonüümina, kuna füüsilise teose lehekülgede arv ei pruugi kattuda automaatselt segmenteeritud dokumendi lehekülgede arvuga.

- **Ülesobitamine** (*overfitting)* - Mudelite liigne sõltuvus treeningmaterjalist, mis tähendab nõrka üldistusvõimet teistele andmetele.

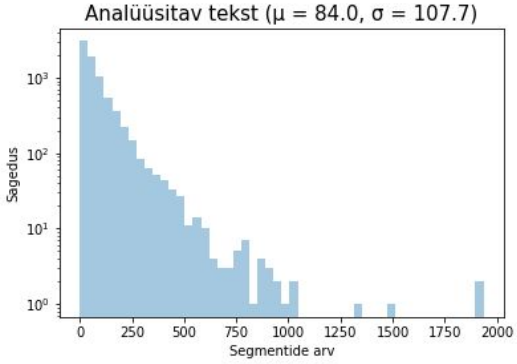
- **Märgend** - kasutatakse antud dokumendi kontekstis “märksõna” sünonüümina.

1.2. Graafikute tõlgendamine

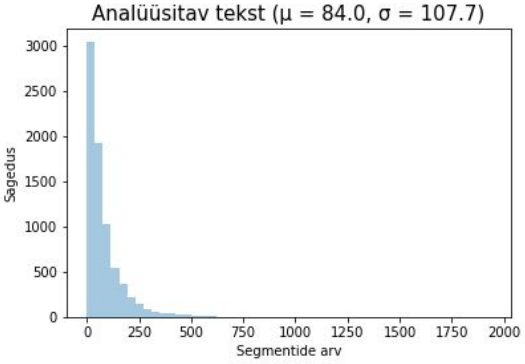
Järgnevas sektsioonis kirjeldatakse, kuidas tõlgendada dokumendis esitatud graafikuid, mis ei pruugi olla intuitiivselt hoomatavad, kuid annavad tihtipeale andmetest hea ülevaate. Näiteks sektordiagrammide puhul saab lugeja tõenäoliselt aru, et eesmärk on väljendada erinevate andmeklasside osakaale, samas võivad aga näiteks kastidiagrammid andmeteadusega igapäevaselt mitte kokkupuutuvate lugejate jaoks ebaselgeks jääda.

Logaritmiline skaala

Tuleb tähele panna, et tihti on andmete jaotuseid kujutavate graafikute y-telg viidud andmetest parema ülevaate saamiseks üle logaritmilisele skaalale, mis tähendab, et visuaalselt võrdse pikkusega intervallide vahed on eksponentsiaalse erinevusega. Näiteks näeme, et joonise 1a y-telje väärtused on 10 astmed, vastavalt $10^0 = 1$, $10^1 = 10$, $10^2 = 100$, $10^3 = 1000$. Logaritmilisele skaalale viimise põhjuseks on ülevaate võimaldamine ka väikese sagedusega andmetest - näiteks näeme samalt jooniselt, et andmestikus eksisteerib üks 1500-leheküljeline tekst, kui aga y-telje andmepunktid oleks esitatud lineaarselt (võrdsetele intervallidele vastaks võrdne arv dokumente), ei oleks antud punkt joonisel nähtav, kuna selle relatiivne sagedus on liiga väike (vt. joonis 1b).



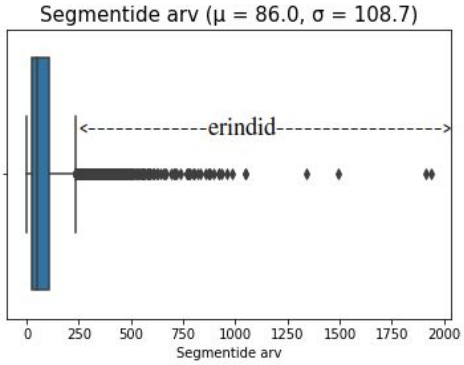
***Joonis 1a**. Andmete jaotus logaritmilisel skaalal.*



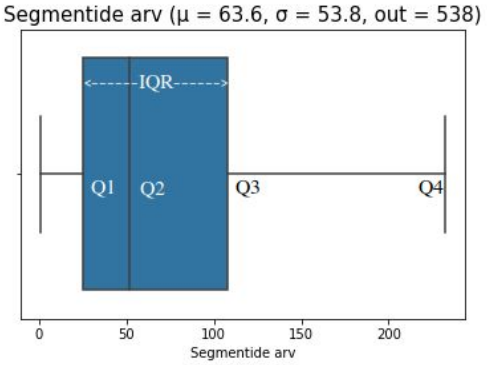
***Joonis 1b**. Andmete jaotus lineaarsel skaalal.*

Kastidiagrammid

Kastidiagrammide (*boxplot)* eesmärk on anda ülevaade andmete jaotusest. Selleks jagatakse (järjestatud) andmepunktid neljaks võrdseks osaks, mille lõpp-punkte kutsutakse kvartiilideks (vastavalt Q1, Q2, Q3, Q4 - esimesele kvartiilile eelneb 25% andmepuntkidest, teisele 50%, kolmandale 75% ning neljandale 100%). Vahemikku esimesest kuni kolmanda kvartiilini (joonisel kujutatud sinise kastina) nimetatakse kvartiilide vaheks (*interquartile range)* ning tähistatakse IQR. Graafiku miinimum- ja maksimumpunkt (joonisel kujutatud nö. vurrudena) arvutatakse esimesest kvartiilist 1.5*IQR lahutamisega ning kolmandale kvartiilile 1.5*IQR liitmisega. Miinimumpunktist väiksemaid ning maksimumpunktist suuremaid väärtuseid kutsutakse erinditeks ning neid tähistakse dokumendis kasutatud joonistel mustade punktidena. Joonistel 2a ja 2b on toodud (ka järgnevas analüüsis kasutusel olevad) kastidiagrammide näited - joonis 2a sisaldab infot koos erinditega ning joonis 2b annab parema ülevaate sama andmestiku andmepunktide jaotusest ilma erinditeta. Vertikaalne joon sinise kasti keskel noteerib mediaani - väärtust, millest mõlemale poole jääb võrdses koguses andmepunkte. Näeme jooniselt 2b, et pooled andmetest on koondunud üpris tihedalt vahemikku 0-50, samas kui teine pool paikneb tunduvalt hajusamalt vahemikus 50-250. Joonis 2a. annab aga ülevaate, et üksikud andmestiku andmepunktide väärtused küündivad 250-st 2000-ni.



***Joonis 2a**. Erinditega kastidiagramm..*



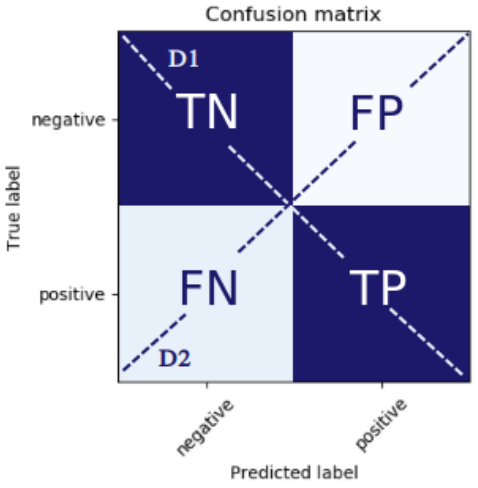
***Joonis 2b**. Erinditeta kastidiagramm.*

Mudelite soorituse visualiseerimine Texta Toolkitis: eksimismaatriks ja ROC-kõver

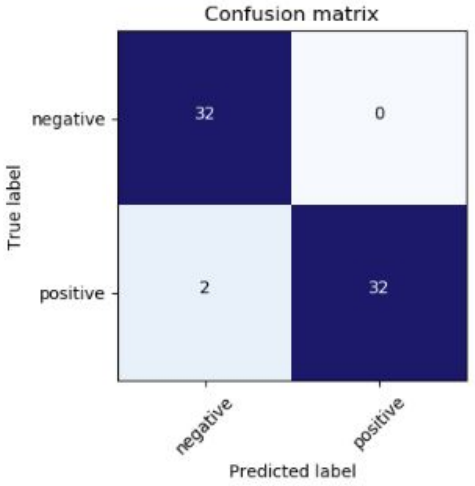
Järgnevalt on toodud seletused Texta Toolkitis mudelite sooritust visualiseerivate graafikute kohta.

Eksimismaatriksi (*confusion matrix*) eesmärk on anda kiire visuaalne ülevaade mudeli ennustamistulemustest jagades valideerimiseks kasutatud märgendid neljaks grupiks: tõsinegatiivsed, väärpositiivsed, väärnegatiivsed ning tõsipositiivsed (vt. joonis 3a ja 3b). Igasse gruppi kuuluvate märgendite arvule vastavalt kujunevad ka joonisel olevad värvid:

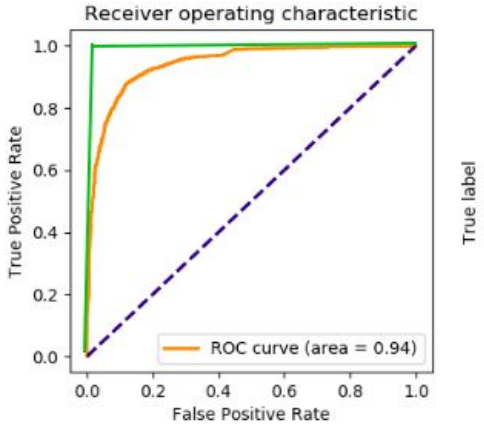
mida suurem on grupi relatiivne sagedus, seda tumedamat tooni sinine on märgendi kast (antud projektis kasutatud joonistel). Seega annab tume toon diagonaalile D1 (vt. joonis 3a) jäävates kastides märku, et märgendaja töötab hästi (väljendab õigesti klassifitseeritud märgendeid) ning tume toon diagonaalile D2 (vt. joonis 3a) jäävates kastides märku, et märgendaja tulemused on viletsad (väljendab valesti klassifitseeritud märgendeid).



Joonis 3a. *Eksimismaatriksi grupid (TN - tõsinegatiivsed, FP - väärpositiivsed, FN - valenegatiivsed, TP - tõsipositiivsed).*



Joonis 3b. *Näide eksimismaatriksist koos igasse gruppi kuuluvate märgendite sagedustega.*



Joonis 3c. *ROC-kõver*

ROC-kõver (*receiver operating characteristic curve*) illustreerib binaarsete klassifitseerijate süsteemi ennustusvõimet näidates, kuidas muutub valepositiivide määr (kujutatud x-teljel) tõsipositiivide määra kasvamisel (kujutatud x-teljel). Sinine katkendlik diagonaalne joon joonisel 3c kujutab hüpoteetilise täiesti juhuslikke ennustusi tegeva (nö. kulli ja kirja viskava) mudeli ROC-kõverat, roheline joon hüpoteetilise perfektse mudeli ROC-kõverat ning oranž joon reaalselt treenitud mudeli ROC-kõverat.

2. Andmeanalüüs

Järgnevas sektsioonis antakse ülevaade hankija edastatud andmetest: kirjeldatakse dokumentide formaati, hinnatakse nende parsitavust ja eraldatud teksti kvaliteeti; kirjeldatakse andmete segmenteerimist ning selle mõju mudelite ehitamisele ning antakse ülevaade dokumentidega seotud märgenditest - nende jagunemist erinevateks tüüpideks, eri tüüpide jaotust dokumentide lõikes, märgendite jagunemist valdkondadeks, hinnatakse treeningmaterjalina kasutatavate märgendite arvu ning selle kasvu segmenteerimisel.

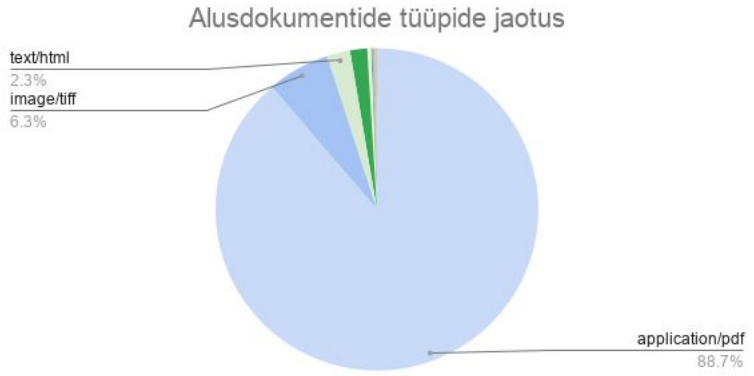
2.1. Dokumentide formaat

Antud peatüki eesmärk on anda ülevaade edastatud andmete formaadist. Järgnevalt kirjeldatakse nii teoste täistekstide, metaandmete kui märksõnastiku väljavõtte formaati.

Projektis kasutatakse ainult Eestis kõigis keeltes ilmunud vabalt kättesaadavaid **raamatuid**. Hankija andis digitaalarhiivist DIGAR pakkujale üle juurdepääsupiiranguta väljaanded alates aastast 1940. Lisaks võeti kasutusele ka e-kataloogis ESTER kirjeldatud veebiressursid, mis ei olnud arhiveeritud DIGARis. Projekti metaandmed (ainult raamatute kohta ja andmed raamatu keelest sõltumata eesti keeles) pärinevad Eesti rahvusbibliograafiast.

2.1.1. Teoste täistekstid

Hankija edastas binaarvoona 11087 teost, millest 9327 olid unikaalsed teosed ning 1760 varasemate dokumentide täiendid, lisad või duplikaadid. Unikaalsetest dokumentidest ei õnnestunud metaandmetega siduda 856 teost, jättes sõelale 8471 dokumenti. Järgneval analüüsil võetakse arvesse ainult unikaalseid dokumente, mille metaandmetega sidumine õnnestus. Edastatud dokumentide formaadiks on valdavalt pdf (7515 dokumenti, 88.7% kõigist dokumentidest), järgnevad tiff (532 dokumenti, 6.3% kõigist dokumentidest) ja html (197 dokumenti, 2.3% kõigist dokumentidest). Täielik dokumendiformaatide jaotus on toodud tabelis 1, visuaalse ülevaate annab joonis 4.



Joonis 4. *Alusdokumentide tüüpide jaotus. Enamuse, 88.7% dokumentidest, moodustavad pdfid; järgnevad tiffid 6.3%-ga ning htmlid 2.3%-ga.*

Sisutüüp	Laiend	Sagedus	Osakaal
application/pdf	.pdf	7515	88.71%
image/tiff	.tif	532	6.28%
text/html	.html	197	2.32%
application/epub+zip	.epub	146	1.72%
application/msword	.wiz	30	0.35%
text/rtf	.rtf	14	0.17%
application/vnd.ms-excel	.xls	8	0.09%
application/zip	.zip	8	0.09%
application/octet-stream	.a	7	0.08%
image/jpeg	.jpg	5	0.06%
text/plain	.txt	4	0.05%
text/xml	.xml	4	0.05%
error	None	1	0.01%

Tabel 1*. Dokumendiformaatide jaotus.*

2.1.2. Metaandmed

Hankija edastas teoste metaandmed MARC XML vormingus ning täistekstidega on metaandmed võimalik siduda MARC ID-ga 907 väljal asuva ERB ID või MARC ID-ga 856 väljal asuvast Digari URList eraldatud ID alusel. 856 teose täisteksti polnud edastatud metaandmete põhjal võimalik andmetega siduda.

Metaandmete parsimise käigus selgus, et andmete kodeering võib kirjeti erineda, mistõttu lüüakse samale terminile vastav märgend vahepeal lahku - eelkõige mõjutab see täpitähti sisaldavaid märksõnu. Näiteks võib isikunime märgenditele keskenduvas peatükis näha, et “Konrad Mägi” on esindatud kaks korda - põhjuseks on, et märgendis esineva “ä”-tähe kodeeringud on antud nimega seotud kirjetes erinevad. Kuna sama termini dubleerimine a) vähendab terminiga seotud näidete hulka, mis omakorda mõjub negatiivselt mudeli sooritusvõimele ning b) tähendab, et lävendi ületamisel vastavaid märgendajaid dubleeritakse, tuleks tulevaste eeltöötlusprotsesside kavandamisel võimalike erisustega kodeeringutes arvestada.

2.1.3. Eesti märksõnastik (EMS)

Eesti märksõnastik (saadaval leheküljel http://ems.elnet.ee) on kõiki ainevaldkondi hõlmav tesauruse struktuuriga märksõnastik raamatute, perioodikaväljaannete, artiklite, nootide, helisalvestiste, kaartide ja muude teavikute eestikeelseks märksõnastamiseks. EMS on raamatukogudes väljaannete märksõnastamiseks ametlikult kasutatav sõnastik, mille kasutus märksõnade allikana on nõutud.

Hankija edastas Eesti märksõnastiku MARC21 formaadis. Eesti märksõnastik sisaldab kokku 61163 terminit, mis jagunevad 39753 märksõnaks ja 21410 äraviiteterminiks, sealjuures on vormimärksõnu 1342, ajamärksõnu 124, kohamärksõnu 6925 ning teemamärksõnu 31380. EMS sisaldab 48 ainevaldkonda ja 51733 inglisekeelset vastet. Projekti raames kasutatavate tekstidega on seotud 386 unikaalset vormimärksõna, 57 ajamärksõna, 482 kohamärksõna ning 8003 teemamärksõna. Seega tuleb tähele panna, et treeningmaterjalis mitte kasutusel olevaid märksõnu juhendatud masinõppe meetoditel baseeruvad mudelid ennustada ei suuda.

2.2. Dokumentide eeltöötlus

See peatükk kirjeldab hankija edastatud dokumentide töötlemise tulemusi - kui paljudest dokumentidest õnnestus eraldada tekst ning kui palju eraldatud tekstidest on omakorda kasutatavad mudelite treeningmaterjalina.

2.2.1. Teksti eraldamine

Teose täistekste sisaldavatest dokumentidest teksti eraldamiseks kasutati Apache Tika tööriistakomplekti.

Parsimisel eraldati tekst edukalt 8127 dokumendist, teksti eraldamine ebaõnnestus 344 dokumendi puhul (vt. joonis 5). Ebaõnnestunuks loetakse dokumente, millest Apache Tika teksti eraldada ei suutnud.

Joonis 5. *Dokumentide parsimise edukus. Tekst õnnestus edukalt eraldada 95.9% dokumentidest, eraldamine ebaõnnestus 4.1%-l.*

Kõige sagedamini ebaõnnestus parsimine pdfide puhul (203 dokumenti), siinkohal tuleb aga silmas pidada, et pdf on ka kõige suurema esindatusega sisutüüp ning suhtarv kõikide pdfide seas on ainult 2.3%, samas kui tif tüüpi failide parsimine ebaõnnestus 126 dokumendi puhul, mis antud sisutüübi lõikes moodustab 24%-se osakaalu. Täieliku ülevaate ebaõnnestunud parsimisega failide sisutüüpidest annab tabel 2.

Sisutüüp	Laiend	Sagedus
application/pdf	.pdf	203
image/tiff	.tiff	126
application/octet-stream	.a	7
application/epub+zip	.epub	4
application/msword	.wiz	1
application/vnd.ms-excel	.xls	1
image/jpeg	.jpg	1
text/xml	.xml	1

Tabel 2*. Ebaõnnestunud parsimise sisutüüpide jaotus*

2.2.2. Teksti kvaliteedi hindamine

Kuna võib juhtuda, et pärast näiliselt edukat teksti eraldamist (Tika lõpetab töö ilma veateateta), pole eraldatud tekst siiski mudeli treeningmaterjalina kasutamiseks kõlbulik, on enne mudelite treenimist kasulik teha tekstidele kvaliteedikontroll, mis valideerib, kas eraldatud tekst on inimloetav. Näiteks dokumentidest, mis sisaldavad viletsa pildikvaliteediga skaneeritud raamatute lehekülgi, ei õnnestu parseril enamjaolt teksti korrektselt tuvastada. Tekstide kvaliteedi hindamiseks treenisime statistilise mudeli, mille abil on võimalik sisendtekstid klassifitseerida treeningmaterjalina kasutatavaks ning mittekasutatavaks.

Statistiline mudel treeniti eesti, inglise ja vene keele tekstide peal, aga töötab ka teiste keelte peal, millel on sarnane häälikusüsteem (vokaalide ja konsonantide vaheldumine). Mudel töötab tõenäosuste ja lävendite alusel. Iga treeningmaterjalis oleva dokumendi põhjal õpib mudel bigrammide[2] tõenäosuslikku jaotumist ehk mida sagedasem on teatud bigramm, seda suurem on ka tõenäosus kohata seda bigrammi loomulikus keeles, nt on suurem tõenäosus, et tähele *r* järgneb täht *a* kui *j*. Lisaks on mudelile etteantud ka hulk juhuslikke häid (korrektselt tuvastatud) ja halbu (halvasti tuvastatud) tekstinäiteid, mille bigrammide alusel saadud keskmiste tõenäosuste järgi genereeritakse lävend, milleks oleme seadnud heade ja halbade keskmise tõenäosuste vahepealse. Uute tekstide klassifitseerimisel genereerib mudel iga teksti bigrammide tõenäosusjaotuse, võrdleb mudeli jaotusega ning arvutab iga teksti keskmise tõenäosuse. Seejärel teeb mudel selle lävendi alusel otsuse, kas tekst klassifitseeritakse kasutatavaks või mittekasutatavaks. Mittekasutatav tekst on näiteks:

flPOI‘PAMMA, METOflM‘IECKI/IE
yKAsAHuH n KOHTPOJIbeIE
SAHAHI/Ifl no ’PYCCKOMY $131)le

8127 parsitud dokumendist tuvastati kokku 270 mittekasutatavat täisteksti.

Sõelale jäänud 7857 dokumendist, mille parsimine õnnestus ja tekst läbis kvaliteedikontrolli, eemaldati lisaks veel allesjäänud pildiformaadis failid (.tif ja .jpg), kuna need sisaldasid üksnes teose kaanepilti, mis märgendaja treeningmaterjalina olulist infot ei sisalda ning võib tulemusi negatiivselt mõjutada. Pärast pildifailide eemaldamist jäi lõplikuks treeningmaterjaliks **7668 teost**.

2.3. Ülevaade andmetest

Selle peatüki eesmärk on anda põgus ülevaade treeningmaterjalina kasutatavast 7668 teosest nendega seotud metaandmetest.

Metaandmetest eraldati järgnevad väljad (sulgudes vastav MARC ID): teose keel (041), teose autor (100), teose pealkiri (245), teose kirjastamise ja ilmumisega seotud informatsioon (260), teose füüsiline kirjeldus nagu mõõtmed maht jne (300), isikunimi märksõnana (600), kollektiivi nimi märksõnana (610), ajutise kollektiivi nimi või sündmus märksõnana (611), teemamärksõna (650), kohamärksõna (651), ajamärksõna (653), vormimärksõna (655), teised teosega seotud isikud nagu tõlkijad, illustraatorid jne (700), digitaalse teose asukoha info (856) ja ERBi ID (907). Kogu info peale valitud märksõnade on kasutusel üksnes mugavamaks navigatsiooniks ja andmetest tervikuna parema ülevaate saamiseks, mis omakorda aitab prognoosida treenitud mudelite üldistusvõimet uutele andmetele. Metaandmetest eraldatud info koos täpsustatud alamväljadega leiab tabelist 3.

[2] Näiteks sõna *bigramm* bigrammid on ‘bi’, ‘ig’, ‘gr’, ‘ra’, ‘am’, ‘mm’.

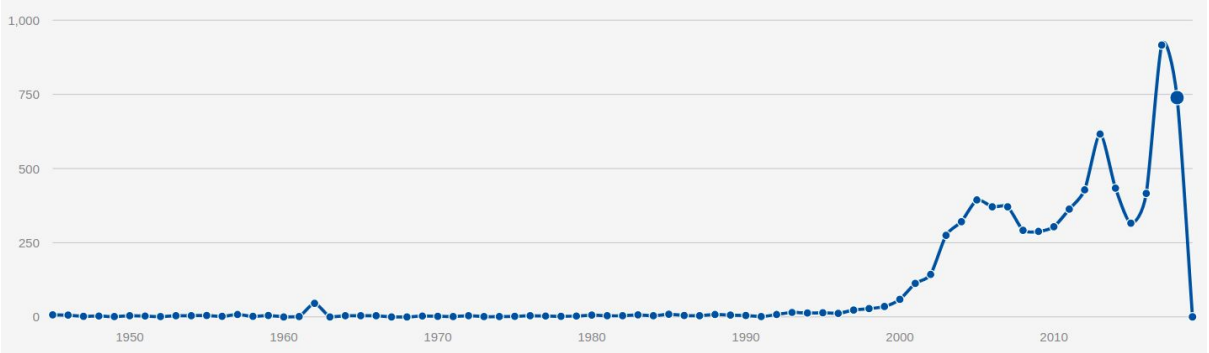
MARC ID	Välja sisu	Alamväli	Alamvälja sisu	Kasutatakse mudeldamisel[3]
001	ERB id	a	ERB id	ei
041	keel	a	originaalkeel(ed)	ei
		b	kokkuvõtte keel(ed)	ei
		h	teose keel(ed)	ei
100	autor	a	teose autori nimi	ei
		d	teose autori sünnikuupäev	ei
		d	teose autori surmakuupäev	ei
245	pealkiri	a ja b	teose pealkiri	ei
260	kirjastamise info	a	kirjastamise koht	ei
		b	kirjastaja	ei
		c	väljalaskekuupäev	ei
		e ⇨ f	trükikoda	ei
300	formaat	a	lehekülgede arv	ei
		c	füüsilise teose mõõdud	ei
600	isiku-märksõna	a	isikumärksõna (nimi)	ei
		c	isikumärksõna (täpsustus)	ei
		d	isikumärksõna (sünnikuupäev)	ei
		d	isikumärksõna (surmakuupäev)	ei
610	kollektiivi nime märksõna	a	kollektiivi nime märksõna (nimi)	ei
		b	kollektiivi nime märksõna (täpsustus)	ei
611	ajutise kollektiivi või sündmuse märksõna	a	ajutise kollektiivi või sündmuse märksõna (nimi)	ei
		c	ajutise kollektiivi või sündmuse märksõna (asukoht)	ei
		d	ajutise kollektiivi või sündmuse märksõna (daatum)	ei

[3] Kasutatavad märksõnade tüübid on täpsemalt kirjeldatud peatükkides 2.5.3.1-2.5.3.7

		e	ajutise kollektiivi või sündmuse märksõna (täpsustus)	ei
		n	ajutise kollektiivi või sündmuse märksõna (järjekorra number)	ei
650	teema-märksõna	a	teemamärksõna	jah
651	koha-märksõna	a	kohamärksõna	jah
653	aja-märksõna	a	ajamärksõna	jah
655	vormi-märksõna	a	vormimärksõna	jah
700	teised isikud	a	teosega seotud isiku nimi (illustraatorit, tõlkijad, koostajad jne)	ei
		d	seotud isiku sünnikuupäev	ei
		d	seotud isiku surmakuupäev	ei
		e	seotud isiku roll (nt. tõlkija)	ei
856	digitaalse teose asukoha info	u	Digar id	ei
907	ERB id	a	ERB id	ei

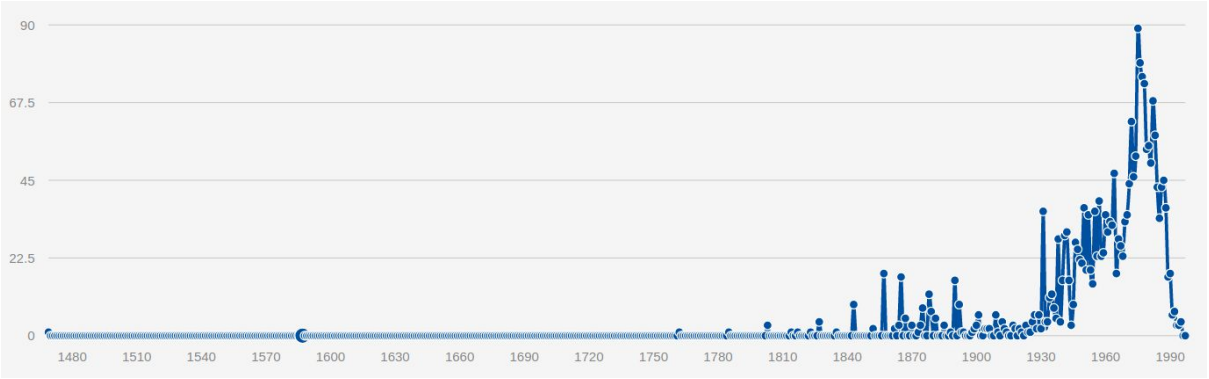
Tabel 3. *Projektis eraldatud ja kasutatud metaandmed.*

Projektis kasutatavad teosed on publitseeritud aastatel vahemikus 1945-2019 (vt. joonis 6), millest valdav osa on ilmunud pärast 2000. aastat. (Ilmumisaasta statistika põhineb MARC väljal 260). Siinkohal tuleb tähele panna, et kõnealune ilmumisaasta väljendab konkreetse teose ja trüki kindla kirjastuse väljalaskekuupäeva, mitte raamatu originaalset ilmumisaastat.



***Joonis 6**. Avaldatud teoste sagedused aastate lõikes.*

Autorite sünniaastad jäävad vahemikku 1469-1997. Kõige varasemad teosed pärinevad autoritelt Niccolò Machiavelli (sünniaastaga 1469), Wilhelm Christian Friebe (sünniaastaga 1762) ning Carl Friedrich von Ledebour (sünniaastaga 1785) (vt. joonis 7).

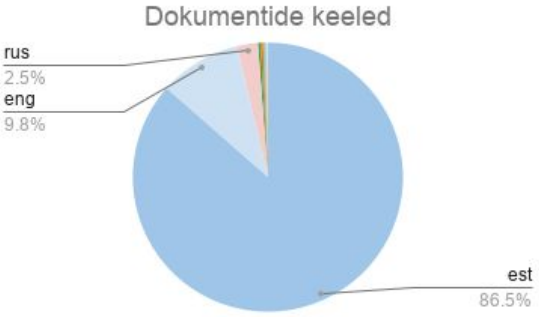


***Joonis 7**. Autorite sünniaastad*

Teoste kontekstist aimduse saamiseks on järgnevalt esitatud 10 juhuslikult valitud teose pealkirjad:

Miku ja Mirjami 6 kummalist kohtumist
Liginullenergiahooned täna ja homme artiklite kogumik
Šotimaa : Glasgow on parem kui Edinburgh
21st International Conference on Types for Proofs and Programs, TYPES 2015 Tallinn, Estonia, 18-21 May 2015 : abstracts
Raasiku vald 25 : 1992-2017
Aktuaalset mahepõllumajanduses.
Fiscal policies and business cycles in an enlarged euro area
Lapse arengu hindamise ja toetamise juhendmaterjal koolieelsetele lasteasutustele
Rahvusvahelistumise ja ekspordi uuring loomemajanduse valdkonnas uuringu aruanne : 31.10.2017
Elukvaliteet Euroopas kokkuvõte

Metainfo andmetel (MARC väli 041 h) on teoste tekstides kasutusel 21 erinevat keelt, neist 86.5% moodustab eesti keel (6632 dokumenti), 9.8% inglise keel (751 dokumenti) ja 2.5% vene keel (194 dokumenti) (vt. joonis 8). Metaandmetes sisalduva info põhjal on iga teosega seotud üks keel. Täieliku ülevaate kasutusel olevatest keeltest annab tabel 4.



***Joonis 8**. 86.5% dokumentidest sisaldavad eestikeelset teksti, 9.8% ingliskeelset ning 2.5% venekeelset.*

	Keel	Dokumentide arv		Keel	Dokumentide arv
1.	eesti	6632	**11.**	jaapani	3
2.	inglise	751	**12.**	kreeka	2
3.	vene	195	**13.**	ungari	2
4.	soome	25	**14.**	portugali	2
5.	saksa	20	**15.**	hispaania	2
6.	prantsuse	10	**16.**	hiina	1
7.	läti	7	**17.**	horvaatia	1
8.	rootsi	5	**18.**	panjabi	1
9.	itaalia	4	**19.**	türgi	1
10.	hollandi	3	**20.**	kõmri	1

***Tabel 4**. Teoste keeled.*

2.4. Segmenteerimine

Selle peatüki eesmärk on anda ülevaade teoste segmenteerimisest: selle alusest, võimalikust mõjust tulemustele ning segmenteeritud andmete jaotusest ja kvaliteedist.

2.4.1. Segmenteerimise metoodika

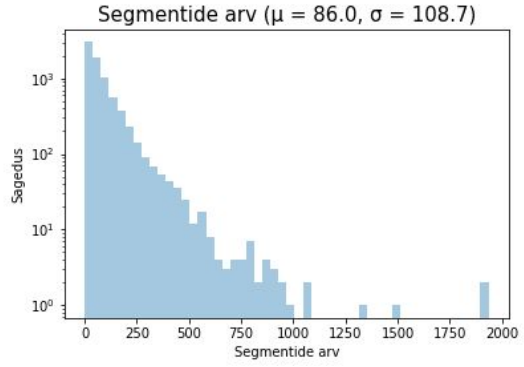
Kuna projekti skoobis olevad teosed võivad olla väga mahukad, on nii andmetes navigeerimise, mudelite efektiivse treenimise kui ka märgendite kasutatavuse tõttu mõistlik algsed dokumendid segmenteerida. Dokumentide segmenteerimiseks on erinevaid

võimalusi: eraldatavateks segmentideks võivad olla näiteks nii peatükid, lõigud kui ka leheküljed. Antud projekti raames kasutame segmenteerimise aluseks lehekülgi, kuna skoopi jäävate dokumentide sisu ja vormi variatiivsus on suur ning automaatne peatükkideks segmenteerimine oleks antud juhul keeruline. Lisaks varieerub teoste lõikes ka peatükkide arv ning pikkus. Viimane kehtib ka lõikude kohta, lisaks on lõikude piirid mõnevõrra sõltuvad ka teksti eraldamise tööriistast. Võimalik oleks dokumentide tekst segmenteerida ka kindlaksmääratud pikkuse alusel - näiteks fikseeritud pikkuse 1000 korral saaksime tulemuseks x segmenti, millest igaüks sisaldaks 1000 tähemärki/sõna. Selle metoodikaga läheb aga kaduma info lehekülgede hõivatusest tekstiga, mis omakorda võib olla oluline tunnus dokumentide tüübi määramisel (romaanide puhul on suurem osa leheküljest tekstiga hõivatud, samas kui luuleraamatu puhul võib lehekülgedel olla palju tühja pinda). Lehekülgede alusel segmenteerimine aitab saada ka kergesti hoomatava ülevaate teoste mahtudest, mis ei pruugi alati olla kooskõlas metaandmetes kirjeldatud lehekülgede arvuga.

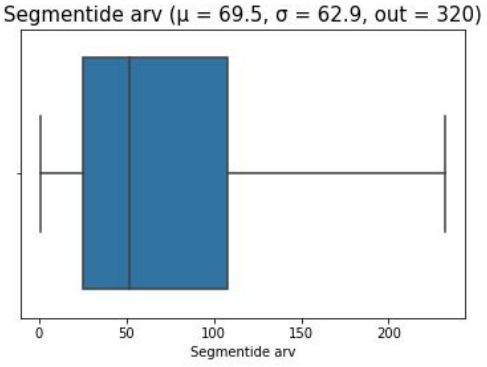
Teostest eraldatud keskmine segmentide arv on 86, standardhälve 108.7 ning mood 1, mis tähendab, et kõige sagedasemad on 1-leheküljelise segmendiga dokumendid. Siinkohal tuleb aga tähele panna segmenteerimise tehnikat: esmalt konverteeritakse kõik alusdokumentide tüübid pdf-failideks, seejärel lõigutakse resultaadid “lehekülgedeks” (segmentideks) ning siis eraldatakse igalt segmendilt tekst. Kui alusdokument on aga näiteks .epub formaadis ning konverteeritud pdfilt teksti eraldamine ebaõnnestub, kasutatakse segmendina lihtsalt tervet algdokumendist eraldatud teksti, mis tähendab, et kõik ühest segmendist koosnevad dokumendid ei pruugi reaalsuses koosneda üksnes ühest leheküljest. Selliseid dokumente, mille segmenteerimine ei õnnestunud ja ühe segmendina on kasutatud tervikteksti, on kokku 617.

2.4.2. Segmentide jaotus

Valdava osa projekti skoobis olevate teoste segmentide arv jääb vahemikku 1-1000, välja arvatud mõned tugevamalt hälbivate segmentide arvuga teosed (vt. joonis 9a). Sealjuures jääb 75% teoste segmentide arv vahemikku 1-100 (vt. joonis 9b), erindite väärtused ulatuvad aga ligi 2000-ni - kõige mahukam projektis olev teos on 1936 “lehekülge” (segmenti) pikk “Eesti-Udmurdi sõnaraamat”.



Joonis 9a. *Segmentide jaotus (erinditega)*

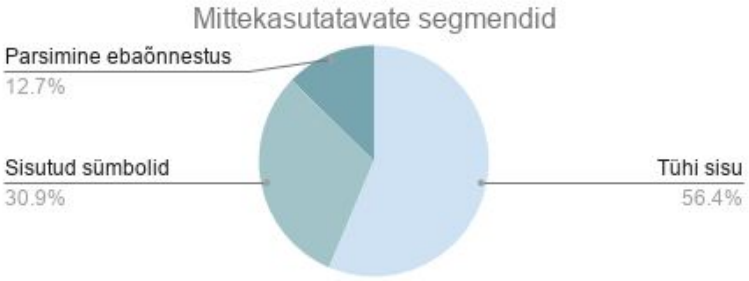


Joonis 9b. *Segmentide jaotus (erinditeta)*

2.4.3. Segmentide kvaliteet

Nagu täistekstide puhul, on ka segmentide juures kasulik kindlaks teha nende sisuline kvaliteet. Seega rakendasime segmentidele sama mudelit, mida ka täistekstidele, et teha kindlaks nende loetavus, sest isegi kui täistekst, mille juurde segment kuulub, kvaliteedikontrolli läbis, ei pruugi see tähendada, et kõik teose leheküljed eraldiseisvalt seda samuti teevad. Kokku eraldati teostest 659576 segmenti, nendest 1949 parsimine ebaõnnestus, 8655-st tuvastati tühi sisu ning 4748-st sisutud sümbolid (vt. joonised 10 ja 11, tabel 5).

***Joonis 10**. Segmentide kasutatavus*



Joonis 11. Mittekasutatavate segmentide jagunevus

Treeningandmestikku lisatakse info kvaliteedi kohta, kuid täiesti välja ebakvaliteetse tekstiga segmente ei visata, kuna ka seosetu sümboljada võib teatud juhtudel tunnusena kasulik olla (näiteks viide, et dokument sisaldab pilte, mis omakorda annab aluse vormimärksõna ennustamiseks).

Kvaliteedi klass	Segmentide arv	Kasutatav
Analüüsitav tekst	644224	JAH
Tühi sisu	8655	EI
Sisutud sümbolid	4748	EI
Parsimine ebaõnnestus	1949	EI

***Tabel 5**. Segmentide kvaliteet*

2.5. Märgendid

Antud peatüki eesmärk on anda põhjalik ülevaade treeningandmetega seotud märgenditest: kasutusel olevatest tüüpidest, eri tüüpide sagedustest ja jaotustest, kasutatavusest treeningmaterjalina ning segmenteerimise mõjust kasutatavusele.

2.5.1. Märgendite tüübid ja jaotus

Vaatleme mudelite arendamisel treeningandmestikuga seotud seitset erinevat märgendi-tüüpi: isikunimi (MARC ID: 600), kollektiivi nimi (MARC ID: 610), ajutine kollektiiv või sündmus (MARC ID: 611), kohamärksõna (MARC ID: 651), ajamärksõna (MARC ID: 653), vormimärksõna (MARC ID: 655) ning teemamärksõna (MARC ID: 650) - põhirõhk antud projektis on just viimase ennustamisel. Kõikide tüüpide peale kokku on projekti kuuluvate teostega seotud 10098 unikaalset märgendit, mis jagunevad järgmiselt: 457 isikunime, 585 kollektiivi nime, 128 ajutist sündmust, 482 kohamärksõna, 57 ajamärksõna, 386 vormimärksõna ning 8003 teemamärksõna. Kuna projektis kasutamiseks planeeritud masinõppe mudelite adekvaatse ennustamise jaoks on minimaalne vajalik märgendite sagedus 50, teeme kindlaks ka seda lävendit ületavate märksõnade arvud iga tüübi lõikes. Pärast lävendi rakendamist ületavad selle 3 kollektiivi nime, 5 ajamärksõna, 8 kohamärksõna, 28 vormimärksõna ning 95 teemamärksõna. Ükski isikunime ega ajutise kollektiivi või sündmuse märksõna lävendit ei ületa. Ülevaade unikaalsete märgendite sagedustest ning lävendi ületanud märgendite arvudest iga tüübi lõikes on koondatud tabelisse 6.

MARC ID	Märgendi tüüp	Märgendite koguarv	Kasutatavate märgendite arv
600	Isikunimi	457	0
610	Kollektiivi nimi	585	3
611	Ajutine kollektiiv või sündmus	128	0
650	Teemamärksõna	8003	95
651	Kohamärksõna	482	8
653	Ajamärksõna	57	5
655	Vormimärksõna	386	28

***Tabel 6**. Kõikide märgendite ning kasutatavate märgendite sagedused tüüpide lõikes.*

Kuna automaatne märgendaja võiks spetsialistide tööd võimalikult täpselt imiteerida ning tagada tulemuseks optimaalse märksõnade arvu, on kasulik analüüsida ka senist märgendite arvu dokumentide lõikes. Keskmiselt on iga dokumendiga seotud 0.1 isikunime märgendit, 0.3 kollektiivi nime märgendit, 0.0 ajutise kollektiivi või sündmuse märgendit, 5.5

teemamärksõna, 0.8 kohamärksõna, 0.1 ajamärksõna, 1.6 vormimärksõna. Kõiki märksõnu kokku on ühel keskmisel dokumendil 8.3 (vt. tabel 7).

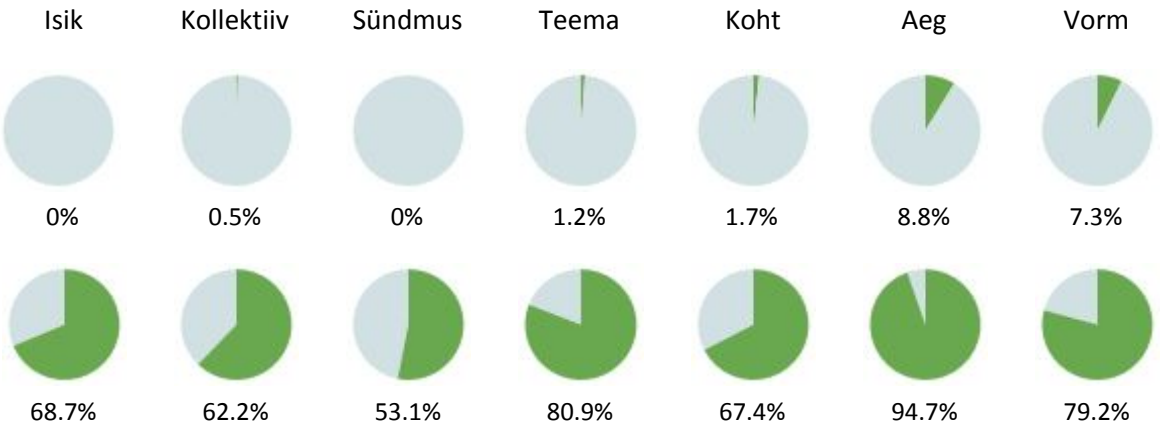
Märgendi tüüp	Keskmine	Mood
Isikunimi	0.1	0
Kollektiivi nimi	0.3	0
Ajutine kollektiiv või sündmus	0.0	0
Teemamärksõna	5.5	5
Kohamärksõna	0.8	1
Ajamärksõna	0.1	0
Vormimärksõna	1.6	2
Kõik märksõnad	8.3	9

***Tabel 7**. Iga dokumendiga seotud märgendite arv tüüpide lõikes.*

2.5.2. Segmenteerimise mõju märgendite kasutatavusele

Tuginedes varasemale kogemusele, on tekstide edukaks märgendamiseks iga ennustatava märgendi jaoks vaja vähemalt 50 näidisdokumenti. See tähendab, et kui soovime tulevikus märgendada kõik kalandusega seotud tekstid märksõnaga “kalandus”, peab treeningandmestikus eksisteerima vähemalt 50 vastava märgendiga dokumenti. Kuna antud projekti treeningandmestik sisaldab aga väga varieeruva pikkusega tekste, annab segmenteeritud andmestiku analüüs siinkohal tõenäoliselt parema ülevaate kui täistekstide oma. Siinkohal tuleb aga silmas pidada, et esitatud tulemused ei ole siiski päris samaväärsed samas mahus unikaalsetest dokumentidest koosnevate andmestikega - näiteks on Ernest Hemingway romaani “Vanamees ja meri” üheks teemamärksõnaks “ameerika” - romaan on 88 lehekülge pikk, mis tähendab, et saaksime teoorias üksnes selle segmenteerimisel kätte vajaliku arvu treeningandmeid märksõna “ameerika” ennustava mudeli treenimiseks. Samas on kaheldav, et tulemuseks saadud mudel suudaks vastavalt märgendada näiteks teose “Muusikaõpik gümnaasiumile. III, 20. sajandi muusika ja Eesti nüüdislooming”, millele ERB-i andmetel samamoodi märgend “ameerika” kuuluma peaks.

Kui segmenteerimata andmestikul oli võimalik ehitada mudelid vaid näiteks 1.2% teemamärksõna ennustamiseks, siis segmenteerimisega tõusis kasutatavus 80.9%-ni. Samamoodi on märgatav kasv ka teistel märksõnatüüpidel, mille kasutatavus segmenteerimata andmestikul varieerus 0%-lt kuni 8.8%-ni, segmenteeritul aga 53.1%-lt 94.7%-ni (vt. joonis 12 ja tabel 8).



***Joonis 12**. Eritüüpi märgendite kasutatavus enne segmenteerimist (ülemine rida) ning pärast segmenteerimist (alumine rida)*

Tulemuste osas tuleb siiski aga peatüki alguses kirjeldatud põhjustel teatav kriitikameel säilitada, kuna segmenteeritud tekstid pärinevad siiski piiratud arvust teostest põhjustades mudelite ülesobitamise ohu.

			Segmenteerimata		**Segmenteeritud**	
MARC ID	**Märgendi tüüp**	**Kokku**	**Kasutatav (sagedus)**	**Kasutatav (%)**	**Kasutatav (sagedus)**	**Kasutatav (%)**
600	Isikunimi	457	0	0. 0%	314	68.7%
610	Kollektiivi nimi	585	3	0.5%	364	62.2%
611	Ajutine kollektiiv või sündmus	128	0	0. 0%	68	53.1%
650	Teemamärksõna	8003	95	1.2%	6477	80.9%
651	Kohamärksõna	482	8	1.7%	325	67.4%
653	Ajamärksõna	57	5	8.8%	54	94.7%
655	Vormimärksõna	386	28	7.3%	306	79.2%

***Tabel 8.** Märgendite kasutatavus enne ja pärast segmenteerimist.*

Tabel 9 annab ülevaate kuidas muutuvad eritüüpi märgendite keskmised sagedused, standardhälbed ja erindite arvud enne ja pärast segmenteerimist. Enne segmenteerimist jääb erindeid arvestav keskmine sagedus vahemikku 1.3-31.4, erindeid mitte kaasav keskmine sagedus aga vahemikku 1-4.6. Pärast segmenteerimist jääb erindeid arvestav keskmine sagedus vahemikku 113.2-2585.1, erindeid mitte kaasav keskmine sagedus aga vahemikku 63.5-866.4. Erindite arvud enne segmenteerimist jäävad vahemikku 9-1056 ning pärast segmenteerimist vahemikku 7-823.

	Enne segmenteerimist					Pärast segmenteerimist				
	μ+	**σ+**	**μ-**	**σ-**	**out**	**μ+**	**σ+**	**μ-**	**σ-**	**out**
Isik	1.3	0.9	1.0	0.0	80	142.4	172.0	121.7	109.3	60
Kollektiiv	3.7	27.6	1.2	0.5	76	288.0	1992.7	97.4	91.3	51
Sündmus	1.3	1.1	1.0	0.0	20	113.2	199.8	63.5	50.0	10
Teema	5.3	14.0	2.3	1.8	1056	447.9	1250.9	216.9	213.3	823
Koht	12.3	155.6	1.3	0.6	97	1072.2	13840.8	136.5	147.9	51
Aeg	13.3	25.1	4.6	4.5	9	1634.2	2331.7	866.4	991.9	7
Vorm	31.4	154.7	4.4	5.0	56	2585.1	12013.1	354.4	457.8	59

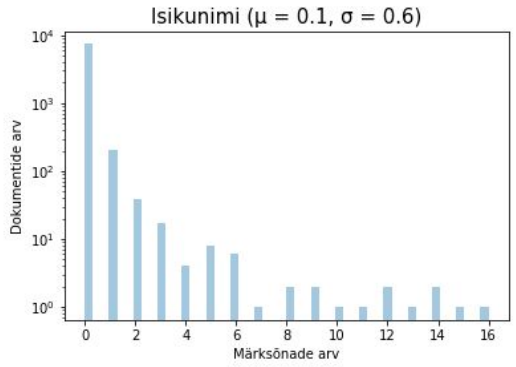
Tabel 9. *Märgendite keskmised sageduses, standardhälbed ja erindite arv enne ja pärast segmenteerimist.*

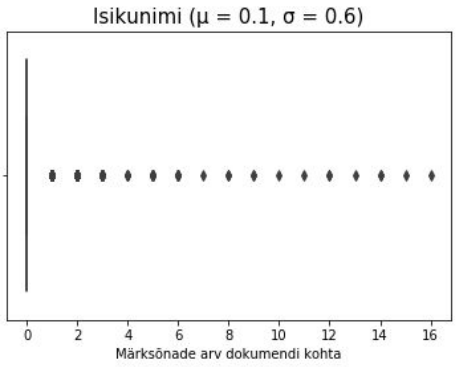
2.5.3. Eritüüpi märksõnade detailne kirjeldus

Järgnevas peatükis antakse detailne ülevaade kõigi seitsme märksõna tüübi kohta (isikunimi, kollektiivi nimi, ajutine kollektiiv või sündmus, teemamärksõna, kohamärksõna, ajamärksõna, vormimärksõna). Sealjuures kirjeldatakse märksõnade jaotust dokumentide lõikes (mitu antud tüüpi märksõna ühel keskmisel dokumendil küljes on), unikaalsete märksõnade sageduste jaotust enne ja pärast segmenteerimist ning lisaks tuuakse välja 10 kõige sagedasemat märksõna enne ja pärast segmenteerimist.

2.5.3.1. Isikunimi märksõnana

Lähteandmestikus on kokku 457 unikaalset isikunime märgendit, sealjuures on iga dokumendiga seotud 0 kuni 16 isikunime märksõna. Keskmine isikunimede märksõnade arv ühe dokumendi kohta on 0.1 standardhälbega 0.6. Jooniselt 13a näeme, et valdav osa treeningandmetest - üle 7000 dokumendi - isikunime märgendit ei sisalda. Kokku sisaldab vähemalt 1 isikunime märgendit 294 dokumenti, millest suurem osa sisaldab täpselt 1 märgendit. 2-3 märgendit sisaldavad kokku umbes 20 dokumenti ning 4-16 märgendit vaid mõned üksikud dokumendid.

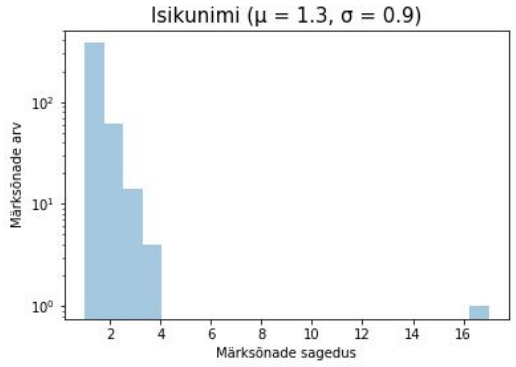


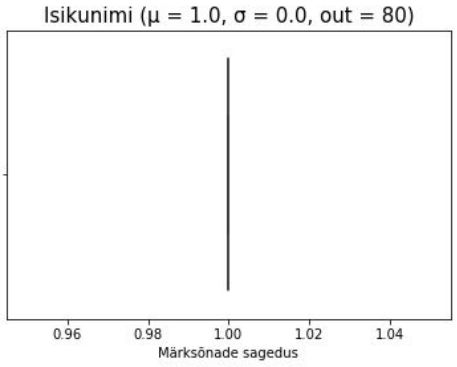


Joonis 13a ***Joonis 13b***

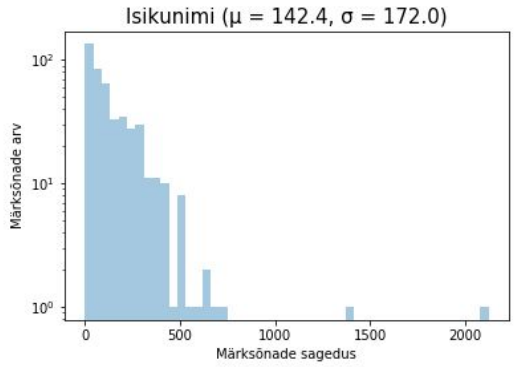
Joonis 13*. Isikunime märksõnade arv dokumentide lõikes*

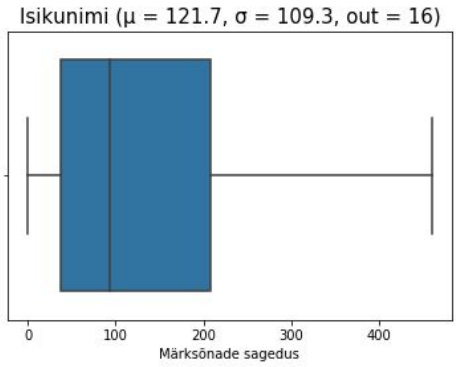
Unikaalsete isikunimede märksõnade sagedus enne segmenteerimist jääb vahemikku 1-4, välja arvatud üks tugevalt hälbiv erind sagedusega 17 (vt. joonis 14a). Keskmine isikunime märksõna sagedus erindeid arvestades on 1.3 standardhälbega 0.9 ning pärast 80 erindi eemaldamist 1 standardhälbega 0. Pärast segmenteerimist jääb isikunime märksõnade sagedus vahemikku 1-718, välja arvatud kaks tugevalt hälbivat erindit sagedustega 1395 ning 2123 (vt. joonis 14c). Sealjuures on 75% isikunime märgendite sagedus vahemikus 1-200 (vt. joonis 14d). Pärast segmenteerimist on keskmine isikunime sagedus erindeid arvestades 142.4 standardhälbega 172 ning pärast 16 erindi eemaldamist 121.7 standardhälbega 109.3. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 12.





Joonis 14a. *Enne segmenteerimist (erinditega)* ***Joonis 14b.*** *Enne segmenteerimist (erinditeta)*





Joonis 14c. *Pärast segmenteerimist (erinditega)*

Joonis 14d. *Pärast segmenteerimist (erinditeta)*

Joonis 14*. Isikunime märksõnade sagedusjaotused enne ja pärast segmenteerimist*

Kõige populaarsem isikunime märksõna on “Jeesus Kristus”, mis on seotud 17 erineva dokumendiga, järgnevad “Konrad Mägi”, “Enn Nõu”, “A. H Tammsaare” ja “Friedebert Tuglas” 4 dokumendiga. Pärast andmete segmenteerimist püsib esikolmiku järjekord muutumatu: kõige populaarsem märksõna on endiselt “Jeesus Kristus”, mis on seotud 2123 segmendiga, järgnevad “Konrad Mägi” 1395 segmendiga ning “Enn Nõu” 718 segmendiga. 10 kõige sagedasemat isikunime märksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 10.

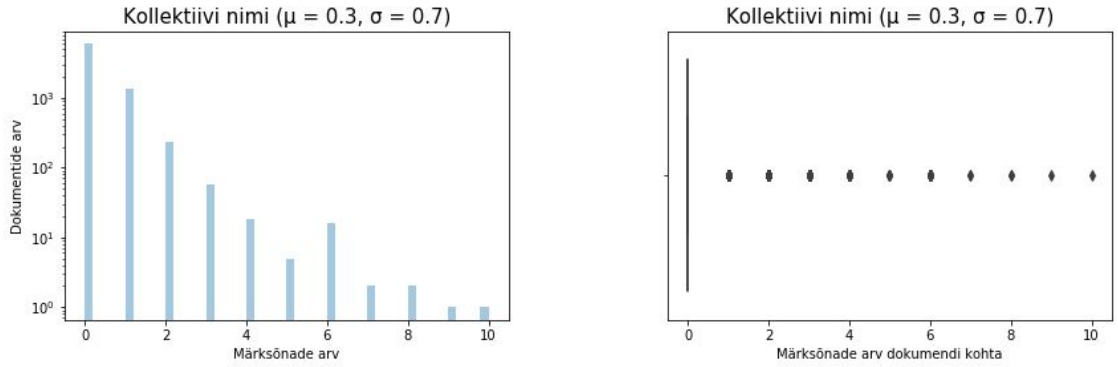
Isikunimed					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	Jeesus Kristus	17		Jeesus Kristus	2123
2.	Konrad Mägi	4		Konrad Mägi	1395
3.	Enn Nõu	4		Enn Nõu	718
4.	A. H Tammsaare	4		Raivo Uukkivi	691
5.	Friedebert Tuglas	4		Konrad Mägi	645
6.	Aino Kallas	3		Helga Nõu	627
7.	Friedrich Reinhold Kreutzwald	3		Marek Vahula	584
8.	Pahlen	3		Istvan Kantor	534
9.	Lev Tolstoi	3		Heino Jõe	528
10.	Raoul Kurvitz	3		Peeter	517

Tabel 10. 10 kõige sagedasemat isikunime märksõna enne ja pärast segmenteerimist.

2.5.3.2. Kollektiivi nimi märksõnana

Lähteandmestikus on kokku 585 unikaalset kollektiivi nime märksõna, sealjuures on iga dokumendiga seotud 0 kuni 10 kollektiivi nime märksõna. Keskmine kollektiivi nimede märksõnade arv ühe dokumendi kohta on 0.3 standardhälbega 0.7. Kokku sisaldab

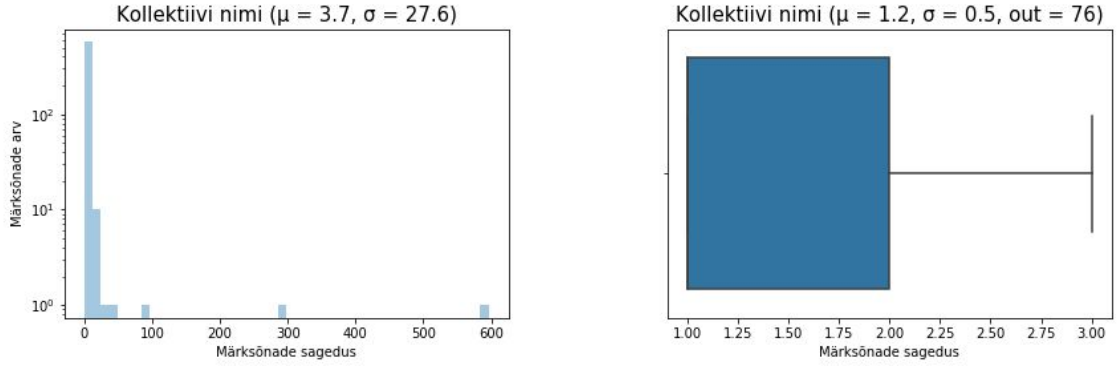
vähemalt ühte kollektiivi nime 1713 dokumenti - üle 1000 dokumendi on seotud täpselt ühe kollektiivi nimega, umbes 400 dokumenti on seotud 2 kollektiivi nimega, 3-4 kollektiivi nime sisaldavad kokku umbes 100 dokumenti, 5-6 kollektiivi nime umbes 30 dokumenti ning 7-10 kollektiivi nime vaid mõned üksikud dokumendid. Kollektiivi nimede märksõnade jaotusest dokumenditi annab ülevaate Joonis 15.



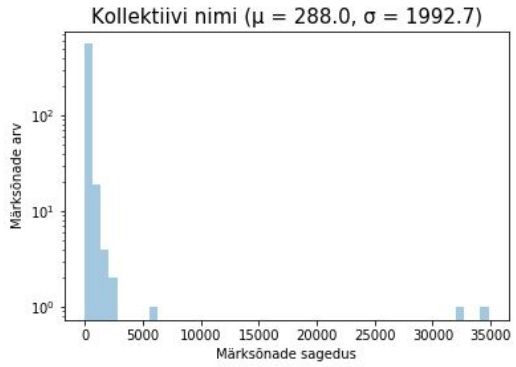
Joonis 15a ***Joonis 15b***

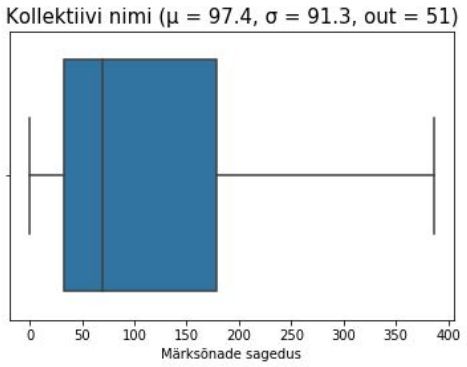
Joonis 15*. Kollektiivi nimede märksõnade arv dokumentide lõikes*

Unikaalsete kollektiivi nime märksõnade sagedus enne segmenteerimist jääb vahemikku 1-39, välja arvatud 3 tugevalt hälbivat erindit sagedustega 94, 287 ja 596 (vt. joonis 16a). Sealjuures jääb 75% kollektiivi nime märksõnade sagedus vahemikku 1-2 (vt. joonis 16b). Keskmine kollektiivi nime märksõna sagedus erindeid arvestades on 3.7 standardhälbega 27.6 ning 76 erindi eemaldamisel 1.2 standardhälbega 0.5. Pärast segmenteerimist jääb kollektiivi nime märksõnade sagedus vahemikku 1-2579 välja arvatud kolme tugevalt hälbiva erindi puhul, mille sagedusteks on 5970, 32587 ja 34819 (vt. joonis 16c). Sealjuures on 75% kollektiivi nime märgendi sagedus vahemikus 1-200 (vt. joonis 16d). Pärast segmenteerimist on keskmine kollektiivi nime sagedus erindeid arvestades 288 standardhälbega 1992.7 ning 51 erindi eemaldamisel 97.4 standardhälbega 91.3. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 14.



Joonis 16a. *Enne segmenteerimist (erinditega)* ***Joonis 16b.*** *Enne segmenteerimist (erinditeta)*





Joonis 16c. *Pärast segmenteerimist (erinditega)*

Joonis 16d. *Pärast segmenteerimist (erinditeta)*

Joonis 16*. Kollektiivi nimede märksõnade sagedusjaotused enne ja pärast segmenteerimist*

Kõige populaarsem kollektiivi nime märksõna on “Euroopa Liit”, mis on seotud 596 dokumendiga, järgneb “Eesti” vastavalt 287 dokumendiga ja “Euroopa Kontrollikoda” 94. dokumendiga. Pärast segmenteerimist tõuseb kõige sagedasemaks märksõnaks “Eesti” 34819 segmendiga, järgneb Euroopa Liit 32587 segmendiga ning Euroopa Kontrollikoda 5970 segmendiga. 10 kõige sagedasemat kollektiivi nime märksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 11.

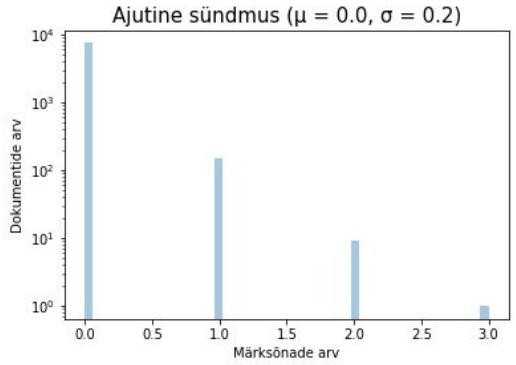
Kollektiivide nimed					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	Euroopa Liit	596		Eesti	34819
2.	Eesti	287		Euroopa Liit	32587
3.	Euroopa Kontrollikoda	94		Euroopa Kontrollikoda	5970
4.	Euroopa Komisjon	39		NSV Liit	2579
5.	Euroopa Parlament	27		Põhja-Atlandi Lepingu Organisatsioon	2204
6.	Tartu Ülikool	23		Euroopa Komisjon	2085
7.	Eesti Pank	20		Tartu Ülikool	2042
8.	Eesti Kunstiakadeemia	18		Euroopa Parlament	1838
9.	Eesti Rahvusraamatukogu	17		Euroopa Kohus	1653
10.	Põhja-Atlandi Lepingu Organisatsioon	16		Eesti NSV Ülemnõukogu	1371

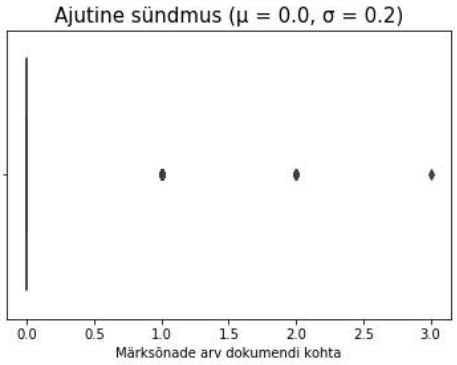
Tabel 11*. 10 kõige sagedasemat kollektiivi nime märksõna enne ja pärast segmenteerimist.*

2.5.3.3. Ajutise kollektiivi nimi või sündmus märksõnana

Lähteandmestikus on kokku 128 ajutise kollektiivi või sündmuse märksõna, sealjuures on iga dokumendiga seotud vastavaid märksõnu 0 kuni 3. Keskmine ajutise kollektiivi või

sündmuse märksõnade arv ühe dokumendi kohta on 0.0 (ümardatud ühe komakohani) standardhälbega 0.2. Kokku sisaldab vähemalt ühte ajutise kollektiivi või sündmuse märksõna 162 dokumenti, millest suurem osa on seotud täpselt ühe ajutise kollektiivi või sündmuse märksõnaga. 2 ajutise kollektiivi või sündmuse märksõna sisaldavad umbes 10 dokumenti ning 3 ajutise kollektiivi või sündmuse märksõna 1 dokument. Ajutise kollektiivi või sündmuste märksõnade jaotusest dokumenditi annab ülevaate joonis 17a.

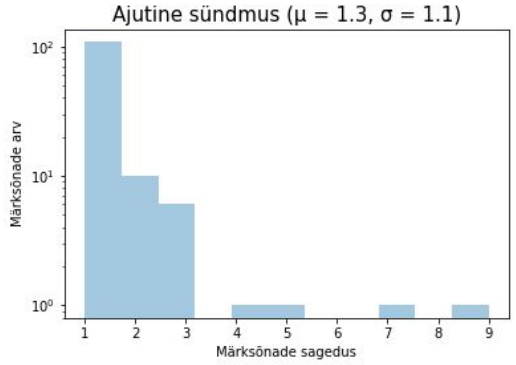


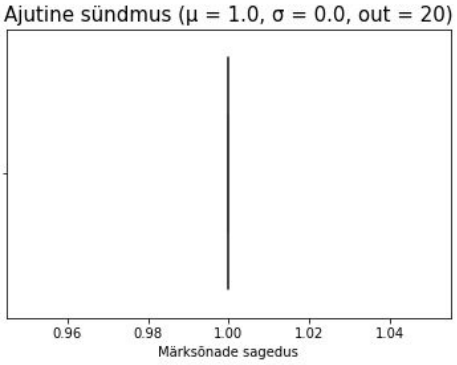


Joonis 17a ***Joonis 17b***

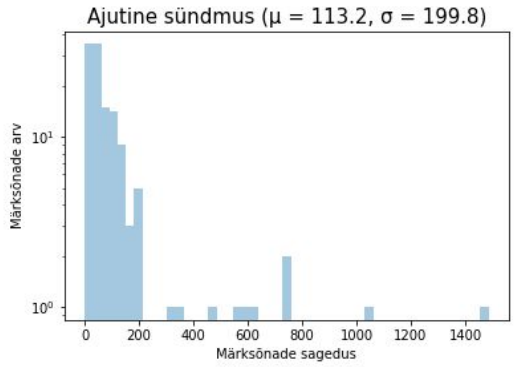
Joonis 17. *ajutise kollektiivi või sündmuse märksõnade arv dokumentide lõikes.*

Unikaalsete ajutise kollektiivi või sündmuse märksõnade sagedus enne segmenteerimist jääb vahemikku 1-3, välja arvatud 4 tugevalt hälbivat erindit sagedustega 4, 5, 7 ja 9 (vt. joonis 18a). Keskmine ajutise kollektiivi või sündmuse märksõna sagedus erindeid arvestades on 1.3 standardhälbega 1.1 ning 20 erindi eemaldamisel 1 standardhälbega 0. Pärast segmenteerimist jääb valdava osa ajutise kollektiivi või sündmuse märksõnade sagedus vahemikku 1-737, välja arvatud 2 tugevalt hälbivat erindit sagedustega 1052 ja 1486 (vt. joonis 18c). Sealjuures jääb 75% ajutise kollektiivi või sündmuse märgenditest vahemikku 1-100 (vt. joonis 18d). Pärast segmenteerimist on keskmine ajutise kollektiivi või sündmuse märksõna sagedus erindeid arvestades 113.2 standardhälbega 199.8 ning pärast 10 erindi eemaldamist 63.5 standardhälbega 50. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 18.

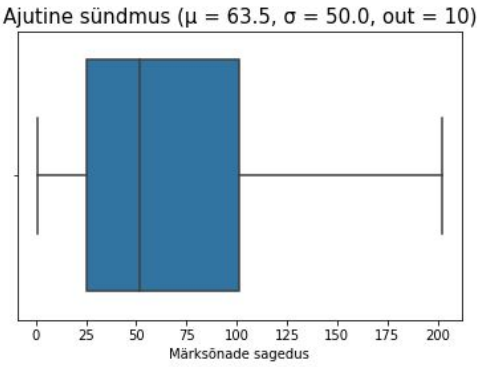




Joonis 18a. *Enne segmenteerimist (erinditega)* ***Joonis 18b.*** *Enne segmenteerimist (erinditeta)*



***Joonis 18c**. Pärast segmenteerimist (erinditega)*



***Joonis 18d**. Pärast segmenteerimist (erinditeta)*

***Joonis 18**. ajutise kollektiivi või sündmuse märksõnade sagedusjaotused enne ja pärast segmenteerimist*

Kõige populaarsed ajutise kollektiivi või sündmuse märksõna on “Eesti matkajuhtide kokkutulek”, mis on seotud 9 dokumendiga, järgnevad “Rahvaloendus Eestis” 7 dokumendiga ning “Taastuvate energiaallikate uurimine ja kasutamine, konverents” 5 dokumendiga. Pärast segmenteerimist tõuseb kõige sagedasemaks märksõnaks “Rahvaloendus Eestis” 1486 segmendiga, järgnevad “Euroopa meedia ja kommunikatsiooni doktorikool” 1052 segmendiga ja “Taastuvate energiaallikate uurimine ja kasutamine, konverents” 737 segmendiga. 10 kõige sagedasemat kollektiivi nime märksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 12.

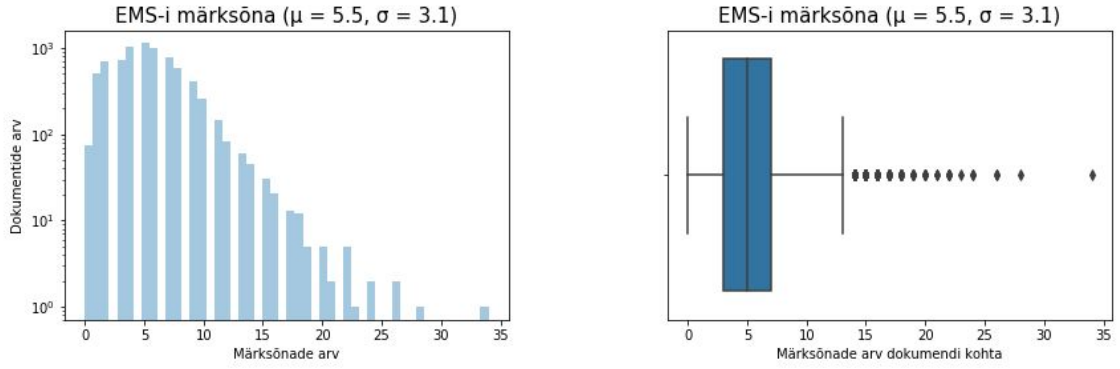
Ajutised kollektiivid või sündmused					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	Eesti matkajuhtide kokkutulek	9		Rahvaloendus Eestis	1486
2.	Rahvaloendus Eestis	7		Euroopa meedia ja kommunikatsiooni doktorikool	1052
3.	Taastuvate energiaallikate uurimine ja kasutamine, konverents	5		Taastuvate energiaallikate uurimine ja kasutamine, konverents	737
4.	Euroopa kultuuripealinn	4		Euroopa kultuuripealinn	735
5.	Noorte etnoloogide ja folkloristide konverents	3		Majanduspoliitika teaduskonverents	624
6.	Matsalu loodusfilmide festival	3		Euroopa Liiduga liitumise mõju Eesti majanduspoliitikale, teadus- ja koolituskonverents	589
7.	Euroopa meedia ja kommunikatsiooni doktorikool	3		Eesti majanduspoliitika teel Euroopa Liitu, teadus- ja koolituskonverents	575
8.	Kreutzwaldi päevad	3		Venezia arhitektuuribiennaal	460
9.	Avatud ühiskonna foorum	3		Eesti matkajuhtide kokkutulek	336

10.	Maailmafilm, visuaalse kultuuri festival	3		Matsalu loodusfilmide festival	311

***Tabel 12**. 10 kõige sagedasemat ajutise kollektiivi või sündmuse märksõna enne ja pärast segmenteerimist.*

2.5.3.4. Teemamärksõna

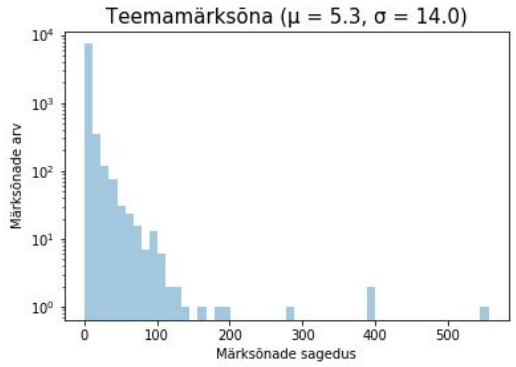
Lähteandmestikus on kokku 8003 unikaalset teemamärksõna, sealjuures on iga dokumendiga seotud 0 kuni 34 teemamärksõna. Keskmine teemamärksõnade arv ühe dokumendi kohta on 5.5 standardhälbega 3.1. Kokku sisaldab vähemalt ühte teemamärksõna 7596 dokumenti, vastav märksõna puudub 53 dokumendil. Joonis 19b annab kiire ülevaate, et pooled dokumentidest sisaldavad 0-5 märksõna, suurem osa ülejäänutest 6-13 märksõna. 15 või rohkem märksõna on seotud vaid mõne üksiku dokumendiga.



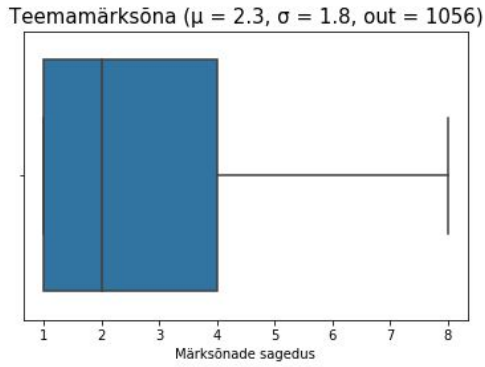
Joonis 19a ***Joonis 19b***

***Joonis 19**. Teemamärksõnade arv dokumentide lõikes.*

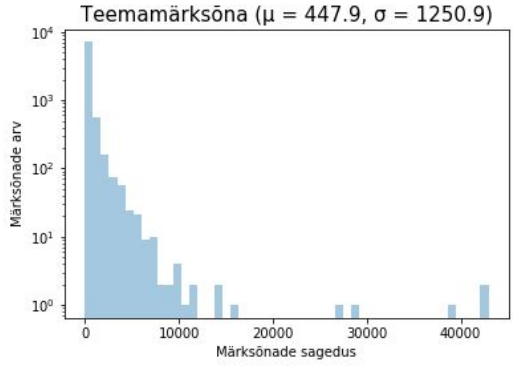
Unikaalsete teemamärksõnade sagedus enne segmenteerimist jääb valdavalt vahemikku 1-200, välja arvatud kolm tugevalt hälbivat erindit sagedustega 393, 400 ning 556 (vt. joonis 20a). Sealjuures jääb 75% teemamärksõna sagedus vahemikku 1-4 (vt. joonis 20b). Keskmine teemamärksõna sagedus erindeid arvestades on 5.3 standardhälbega 14 ning 1056 erindi eemaldamisel 2.3 standardhälbega 1.8. Pärast segmenteerimist jääb valdava osa teemamärksõnade sagedus vahemikku 1-14000, välja arvatud seitse tugevalt hälbivat erindit sagedustega 42954, 42360, 39334, 28533, 27447, 16002 ja 14132 (vt. joonis 20c). Sealjuures jääb 75% teemamärksõna sagedus vahemikku 1-400 (vt. joonis 20d). Pärast segmenteerimist on keskmine teemamärksõnade sagedus erindeid arvestades 447.9 standardhälbega 1250.9 ning 823 erindi eemaldamisel 216.9 standardhälbega 213.3. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 20.



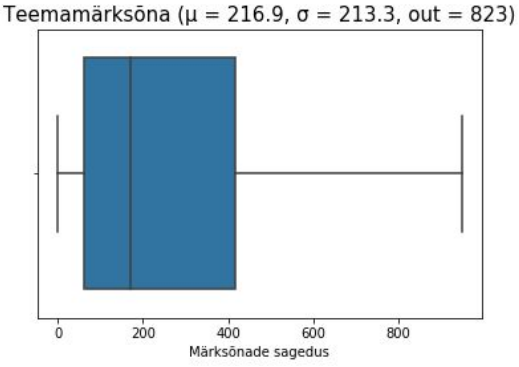
Joonis 20a. *Enne segmenteerimist (erinditega)*



Joonis 20b. *Enne segmenteerimist (erinditeta)*



Joonis 20c*. Pärast segmenteerimist (erinditega)*



Joonis 20d*. Pärast segmenteerimist (erinditeta)*

Joonis 20*. Teemamärksõnade sagedusjaotused enne ja pärast segmenteerimist.*

Kõige populaarsem teemamärksõna on “eesti” olles seotud 556 dokumendiga, järgnevad “auditeerimine” 400 dokumendiga ning “kontroll” 393 dokumendiga. Pärast segmenteerimist on kõige populaarsem märksõna endiselt “eesti” 42 954 segmendiga, järgnevad “parlamendid” 42 360 segmendiga ning “istungjärgud” 39 334 segmendiga. 10 kõige sagedasemat teemamärksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 13.

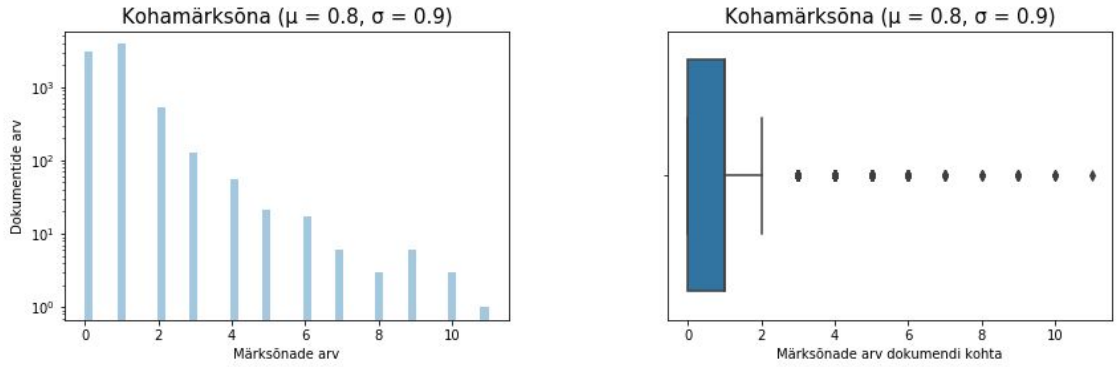
Teemamärksõna					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	eesti	556		eesti	42954
2.	auditeerimine	400		parlamendid	42360
3.	kontroll	393		istungjärgud	39334
4.	eesti keel	279		ajalugu	28533
5.	ajalugu	198		eesti keel	27447
6.	uuringud	185		uuringud	16002
7.	hindamine	157		auditeerimine	14132
8.	raamatupidamine	136		kontroll	13826

9.	parlamendid	133		hindamine	12008
10.	finantseerimine	128		kodulugu	11400

***Tabel 13**. 10 kõige sagedasemat teemamärksõna enne ja pärast segmenteerimist.*

2.5.3.5. Kohamärksõna

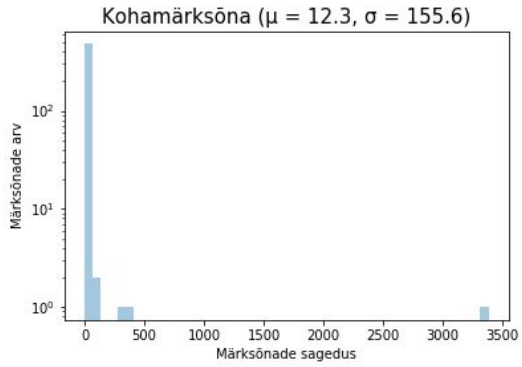
Lähteandmestikus on kokku 482 unikaalset kohamärksõna, sealjuures on iga dokumendiga seotud 0 kuni 11 kohamärksõna. Keskmine kohamärksõnade arv ühe dokumendi kohta on 0.8 standardhälbega 0.9. Kokku sisaldab vähemalt ühte kohamärksõna 4639 dokumenti, millest suurema osa on seotud 1-3 kohamärksõnaga. 4-6 kohamärksõna on kokku umbes 100 dokumendil ning 7-11 märksõna kokku umbes 30 dokumendil. Kohamärksõnade jaotusest dokumenditi annab ülevaate Joonis 21.



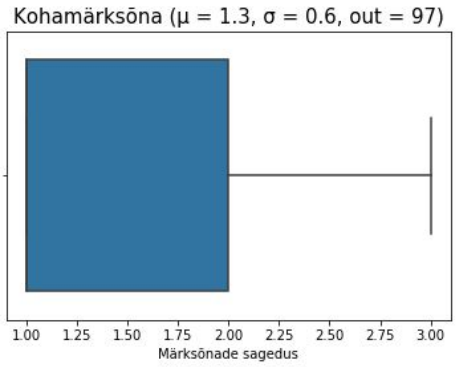
Joonis 21a ***Joonis 21b***

***Joonis 21**. Kohamärksõnade arv dokumentide lõikes.*

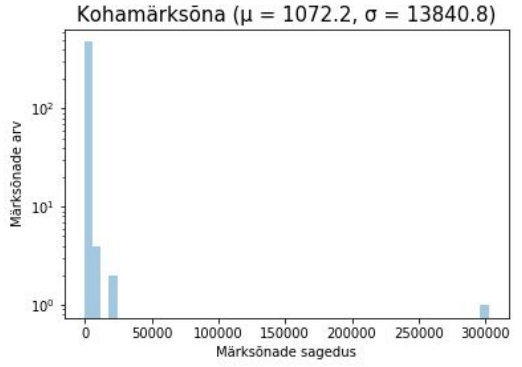
Unikaalsete kohamärksõnade sagedus enne segmenteerimist jääb valdavalt vahemikku 1-394, välja arvatud üks tugevalt hälbiv erind sagedusega 3385 (vt. joonis 22a). Sealjuures jääb 75% kohamärksõna sagedus vahemikku 1-2 (vt. joonis 22b). Keskmine kohamärksõna sagedus erindeid arvestades on on 12.3 standardhälbega 155.6 ning 97 erindi eemaldamisel 1.3 standardhälbega 0.6. Pärast segmenteerimist jääb valdava osa kohamärksõnade sagedus vahemikku 1-23207 välja arvatud üks tugevalt hälbiv erind sagedusega 302339 (vt. joonis 22c). Sealjuures jääb 75% kohamärksõnade sagedus vahemikku 1-300 (vt. joonis 22d). Pärast segmenteerimist on keskmine kohamärksõnade sagedus erindeid arvestades 1072.2 standardhälbega 13840.8 ning 51 erindi eemaldamisel 136.5 standardhälbega 147.9. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 22.



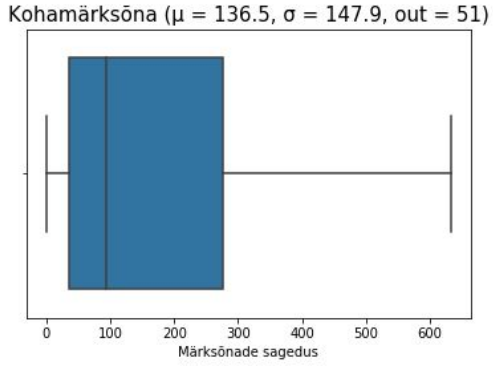
Joonis 22a. *Enne segmenteerimist (erinditega)*



Joonis 22b. *Enne segmenteerimist (erinditeta)*



Joonis 22c*. Pärast segmenteerimist (erinditega)*



Joonis 22d*. Pärast segmenteerimist (erinditeta)*

Joonis 22*. Kohamärksõnade sagedusjaotused enne ja pärast segmenteerimist.*

Kõige populaarsem kohamärksõna on “Eesti” olles seotud 3385 erineva dokumendiga, järgnevad “Euroopa Liidu maad” 394 dokumendiga ning “Euroopa” 295 dokumendiga. Pärast segmenteerimist on kõige populaarsem kohamärksõna endiselt “Eesti” segmentide arvuga 302339, järgnevad “Euroopa” 23207 segmendiga ning “Euroopa Liidu maad” 21494 segmendiga. 10 kõige sagedasemat kohamärksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 14.

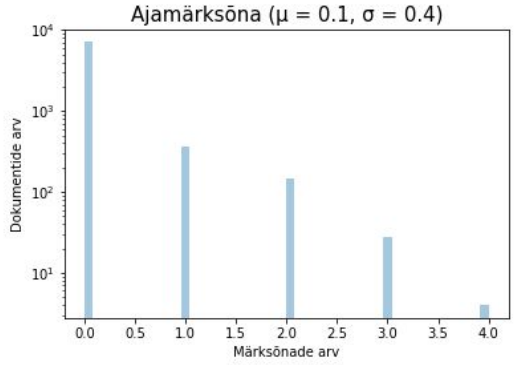
Kohamärksõna					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	Eesti	3385		Eesti	302339
2.	Euroopa Liidu maad	394		Euroopa	23207
3.	Euroopa	295		Euroopa Liidu maad	21494
4.	Tallinn	89		Tallinn	8860
5.	Baltimaad	75		NSV Liit	8554
6.	Venemaa	65		Venemaa	8123
7.	Tartu	55		Baltimaad	7243
8.	Soome	50		Soome	5513

9.	Ida-Virumaa	44		Tartu	4742
10.	NSV Liit	40		Rootsi	3591

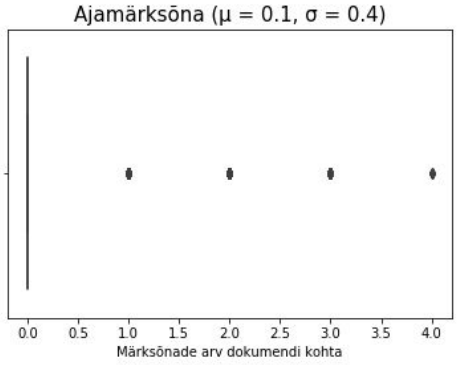
***Tabel 14**. 10 kõige sagedasemat kohamärksõna enne ja pärast segmenteerimist.*

2.5.3.6. Ajamärksõna

Lähteandmestikus on kokku 57 unikaalset ajamärksõna, sealjuures on iga dokumendiga seotud 0 kuni 4 ajamärksõna. Keskmine ajamärksõnade arv ühe dokumendi kohta on 0.1 standardhälbega 0.4. Kokku sisaldab vähemalt ühte ajamärksõna 554 dokumenti, millest suurem osa on seotud 1-2 märksõnaga. 3 ajamärksõna sisaldab umbes 30 dokumenti ning 4 märksõna 1 dokument. Ajamärksõnade jaotusest dokumenditi annab ülevaate Joonis 23.



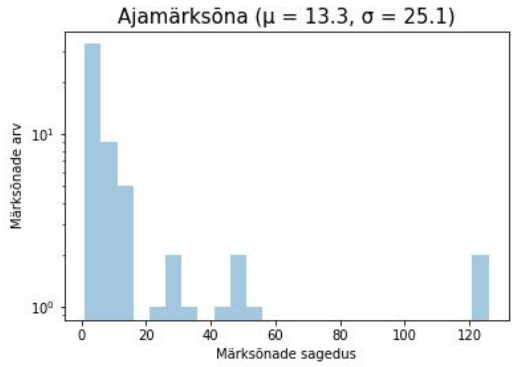
Joonis 23a



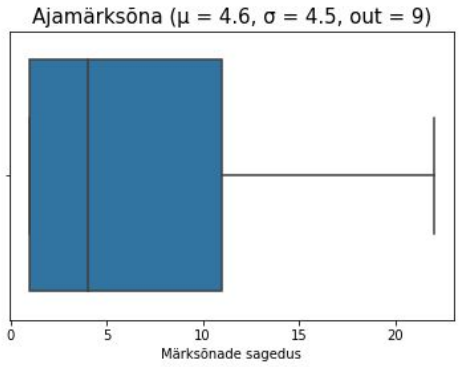
Joonis 23b

***Joonis 23**. Ajamärksõnade esinemissagedused.*

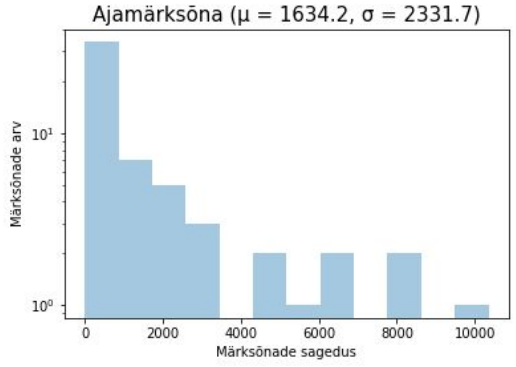
Unikaalsete ajamärksõnade sagedus enne segmenteerimist jääb valdavalt vahemikku 1-60, välja arvatud kaks tugevalt hälbivat erindit sagedustega 125 ja 126 (vt. joonis 24a). Sealjuures jääb 75% ajamärksõna sagedus vahemikku 1-11 (vt. joonis 24b). Keskmine ajamärksõna sagedus erindeid arvestades on on 13.3 standardhälbega 25.1 ning 9 erindi eemaldamisel 4.6 standardhälbega 4.5. Pärast segmenteerimist jääb ajamärksõnade sagedus vahemikku 1-10349 (vt. joonis 24c). Sealjuures jääb 75% ajamärksõnade sagedus vahemikku 1-2000 (vt. joonis 24d). Pärast segmenteerimist on keskmine ajamärksõnade sagedus erindeid arvestades 1634.2 standardhälbega 2331.7 ning 7 erindi eemaldamisel 866.4 standardhälbega 991.9. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 24.



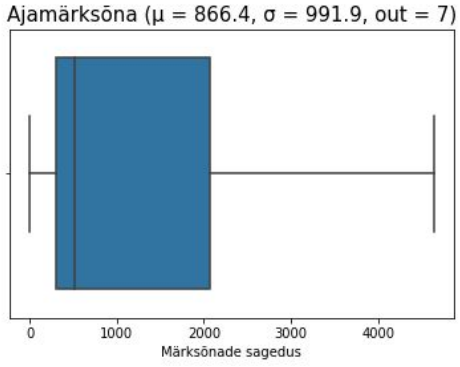
Joonis 24a. *Pärast segmenteerimist (erinditega)*



Joonis 24b. *Enne segmenteerimist (erinditeta)*



Joonis 24c*. Pärast segmenteerimist (erinditega)*



Joonis 24d*. Pärast segmenteerimist (erinditeta)*

Joonis 24*. Ajamärksõnade sagedusjaotused enne ja pärast segmenteerimist.*

Kõige populaarsem ajamärksõna on “2000-ndad” olles seotud 126 erineva dokumendiga, järgnevad “2010-ndad” 125 dokumendiga ning “21. saj. algus” 54 dokumendiga. Pärast segmenteerimist on kõige populaarsem kohamärksõna endiselt “2000-ndad” segmentide arvuga 10 349, järgnevad “2010-ndad” 8205 segmendiga ning “1940-ndad” 8017 segmendiga. 10 kõige sagedasemat ajamärksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 15.

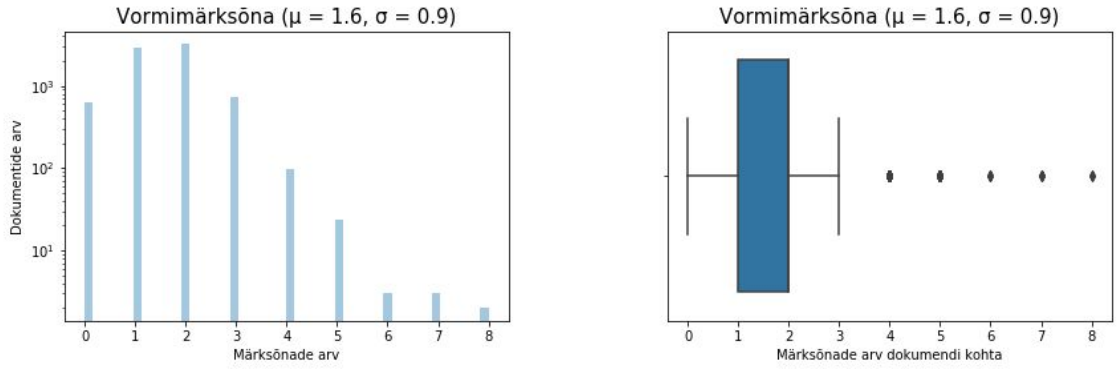
Ajamärksõna					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	2000-ndad	126		2000-ndad	10349
2.	2010-ndad	125		2010-ndad	8205
3.	21. saj. algus	54		1940-ndad	8017
4.	20. saj.	50		20. saj.	6546
5.	1990-ndad	50		21. saj. algus	6262
6.	20. saj. 2. pool	41		1990-ndad	5322
7.	1918-1940	32		20. saj. 2. pool	5130
8.	20. saj. 1. pool	30		1918-1940	4638

9.	1940-ndad	29		19. saj.	3119
10.	19. saj.	22		20. saj. 1. pool	3079

***Tabel 15**. 10 kõige sagedasemat ajamärksõna enne ja pärast segmenteerimist.*

2.5.3.7. Vormimärksõna

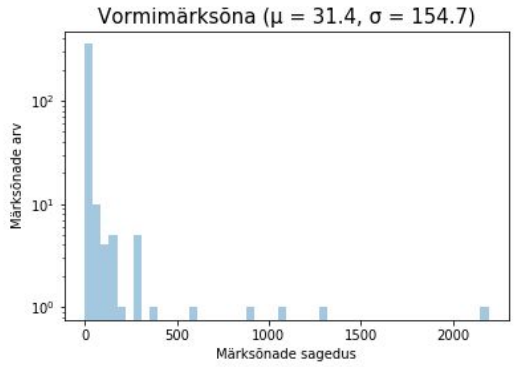
Lähteandmestikus on kokku 386 unikaalset vormimärksõna, sealjuures on iga dokumendiga seotud 0 kuni 8 vormimärksõna. Keskmine vormimärksõnade arv ühe dokumendi kohta on 1.6 standardhälbega 0.9. Kokku sisaldab vähemalt ühte vormimärksõna 7028 dokumenti, millest suurem osa on seotud 1-3 märksõnaga. 4 vormimärksõna sisaldab umbes 100 dokumenti, 5 vormimärksõna 30 dokumenti ning 6-8 vormimärksõnaga on seotud kokku umbes 10 dokumenti. Vormimärksõnade jaotusest dokumenditi annab ülevaate Joonis 25.



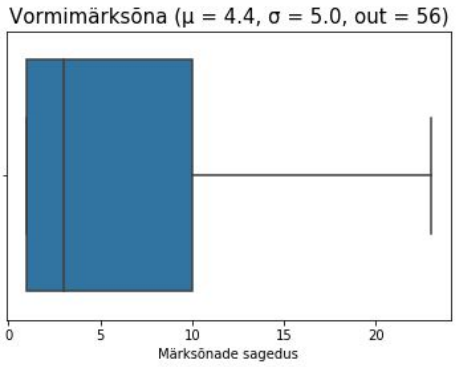
Joonis 25a ***Joonis 25b***

***Joonis 25**. Vormimärksõnade esinemissagedused.*

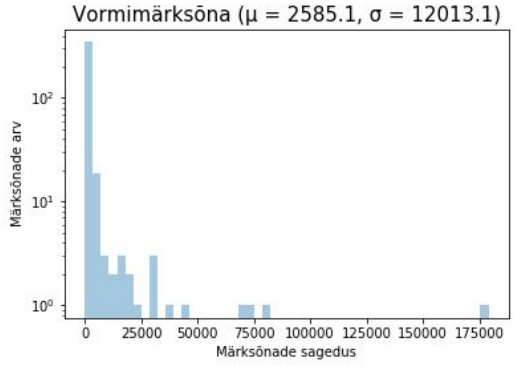
Unikaalsete vormimärksõnade sagedus enne segmenteerimist jääb valdavalt vahemikku 1-360, välja arvatud viis tugevalt hälbivat erindit sagedustega 609, 889, 1071, 1297 ja 2193 (vt. joonis 26a). Sealjuures jääb 75% vormimärksõna sagedus vahemikku 1-10 (vt. joonis 26b). Keskmine vormimärksõna sagedus erindeid arvestades on on 31.4 standardhälbega 154.7 ning 56 erindi eemaldamisel 4.4 standardhälbega 5.0. Pärast segmenteerimist jääb valdava osa vormimärksõnade sagedus vahemikku 1-45000 välja arvatud neli tugevalt hälbivat erindit sagedustega 68635, 74801, 79552 ja 179126 (vt. joonis 26c). Sealjuures jääb 75% vormimärksõnade sagedus vahemikku 1-1000 (vt. joonis 26d). Pärast segmenteerimist on keskmine vormimärksõnade sagedus erindeid arvestades 2585.1 standardhälbega 12013.1 ning 59 erindi eemaldamisel 354.4 standardhälbega 457.8. Sageduste jaotusest enne ja pärast segmenteerimist annab ülevaate joonis 26.



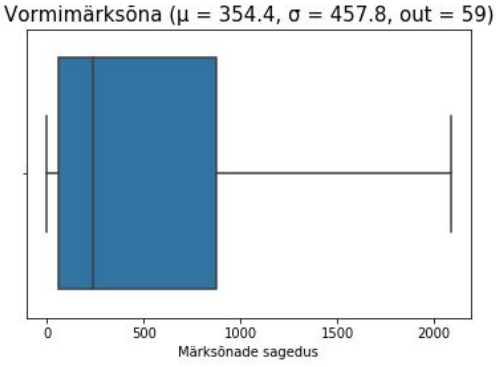
Joonis 26a. *Enne segmenteerimist (erinditega)*



Joonis 26b. *Enne segmenteerimist (erinditega)*



Joonis 26c*. Pärast segmenteerimist (erinditega)*



Joonis 26d*. Pärast segmenteerimist (erinditeta)*

Joonis 26*. Vormimärksõnade sagedusjaotused enne ja pärast segmenteerimist.*

Kõige populaarsem vormimärksõna on “e-raamatud” olles seotud 2193 erineva dokumendiga, järgnevad “võrguväljaanded” 1297 dokumendiga ning “aruanded” 1071 dokumendiga. Pärast segmenteerimist on kõige populaarsem vormimärksõna endiselt “e-raamatud” segmentide arvuga 179 126, järgnevad “dissertatsioonid” 79 552 segmendiga ning “võrguväljaanded” 74801 segmendiga. 10 kõige sagedasemat vormimärksõna koos sagedustega enne ja pärast segmenteerimist on esitatud tabelis 16.

Vormimärksõna					
	Enne segmenteerimist			Pärast segmenteerimist	
1.	e-raamatud	2193		e-raamatud	179126
2.	võrguväljaanded	1297		dissertatsioonid	79552
3.	aruanded	1071		võrguväljaanded	74801
4.	võrguväljaanded	889		aruanded	68635
5.	dissertatsioonid	609		võrguväljaanded	44799
6.	konverentsikogumikud	354		stenogrammid	38741
7.	teatmikud	298		konverentsikogumikud	31878
8.	juhendid	288		artiklikogumikud	30056

9.	käsiraamatud	282		käsiraamatud	28995
10.	ilukirjandus	278		õppematerjalid	24106

***Tabel 16**. 10 kõige sagedasemat vormimärksõna enne ja pärast segmenteerimist.*

2.5.4. Mudeldamisesse kaasatavad märksõnad

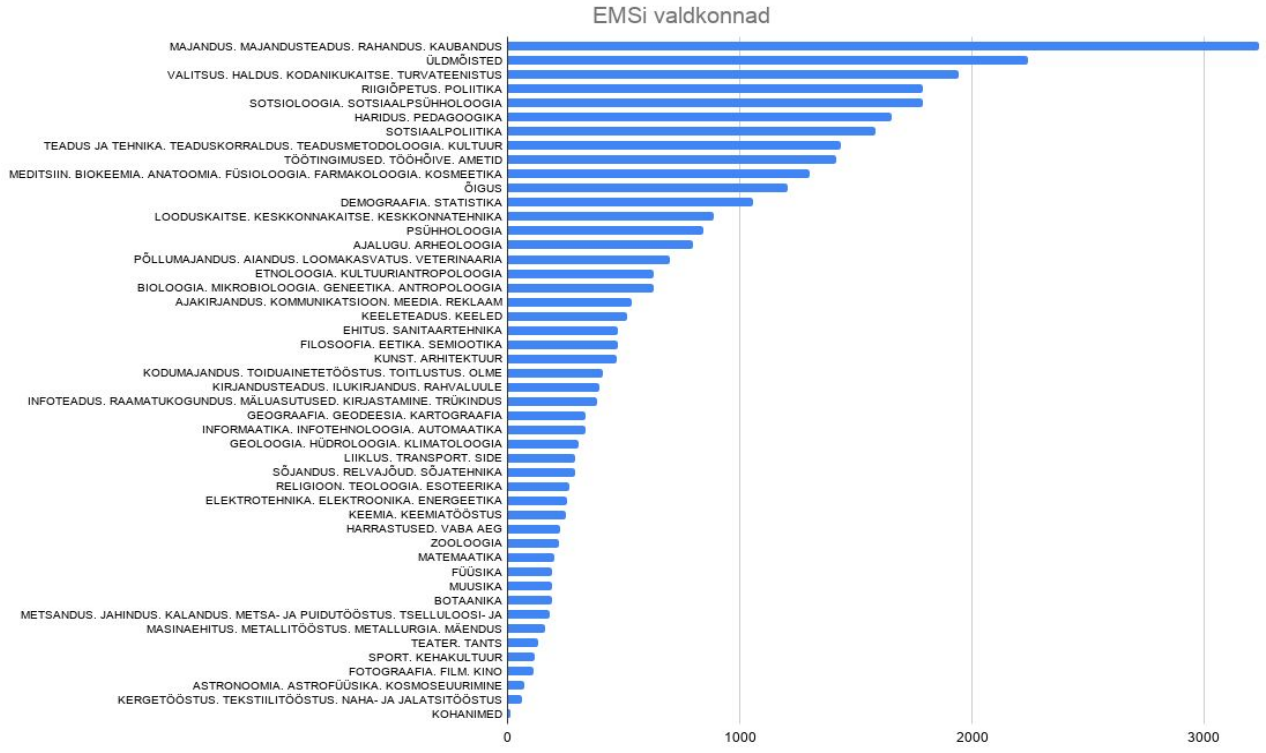
Kuna teatud märksõnade MARC kirjetesse on kaasatud peale märksõna termini ka muud seonduvat infot - näiteks isikumärksõnade puhul muuhulgas ka sünni- ja surmaaastad, siis tuleks kogu see info protsessi automatiseerimiseks koos vastava märksõna põhisisuga ka märgendisse kodeerida, et hiljem oleks võimalik ennustatud märgendi korrektne MARC kirje reprodutseerida. Kuna EMS-ile mittetuginevate märksõnade kirjapiltides võib esineda aga teatud määral variatiivsust, siis tähendaks nende kaasamine täiendavat lisatööd. Näiteks võib isiku sünniaeg olla märgitud nii “$aTamm, Jüri,$d1928-” kui ka “$aTamm, Jüri,$d1928-.” - inimese jaoks siin märgatavat erinevust ei ole ning tõenäoliselt ei teki isegi kahtlust, et viide käib sama isiku kohta. Arvuti silmis on tegemist aga kahe erineva persooniga, kuna kirjete sümboljada pole 100% identne (teise kirje daatumi lõpus on punkt). Seega on mõistlik prototüübi mudelitesse kaasata üksnes sellised märksõnad, mis lisapuhastamist ega ühildamist ei nõua: teemamärksõna (MARC ID-ga 650), kohamärksõna (MARC ID-ga 651), ajamärksõna (MARC ID-ga 653) ning vormimärksõna (MARC ID-ga 655).

2.5.5. EMSi valdkonnad

Kõik EMSi märksõnad on jagatud 48 suuremasse valdkonda. Kuna konkreetsete märksõnade sagedused on tervikteoste lõikes väikesed, võib märksõnade valdkondadega sidumine ning konkreetse märksõna asemel valdkonna ennustamine anda tulevikuks kasulikku infot. Kuna pärast valdkondadesse jagamist ületavad segmenteerimata andmetel vajaliku lävendi peaaegu kõik valdkonnad (va. kohanimed), kaetakse suurem variatsioon erinevaid tekste. Lisaks on kõikidest EMSi märksõnadest märgenditena kasutusel vaid umbes viiendik - et tulevikus oleks võimalik tekste märgendada treeningandmestikus harva või üldse mitte esinevate märgenditega, oleks kasulik uurida ka alternatiivseid (juhendamata masinõppe skoopi kuuluvaid) meetodeid - kui me suudame aga teksti x valdkonna ennustada, siis aitab see võimalike märksõnade hulka piirata ning seega ülesande lahendamisele sammu lähemale astuda.

Treeningmaterjalis olevatel EMSi märksõnadel õnnestus valdkondadega siduda 6613 märksõna ning 1390 märksõna jäid kodeeringuprobleemide tõttu esialgu valdkondadega sidumata. Iga valdkonnaga seotud treeningmaterjali dokumentide arv ning vastav osakaal on toodud tabelis 16 ning kiire visuaalse ülevaate saamiseks on valdkondadesse jagunevust kujutatud joonisel 27. Siinkohal tuleb tähele panna, et ühe dokumendiga võib olla seotud ka mitu erinevat valdkonda. Valdkondade jaotust uurides saame ka infot andmete kallutatusest: tervelt 42.2% dokumenditest kuulub valdkonda “MAJANDUS. MAJANDUSTEADUS.

RAHANDUS. KAUBANDUS", mis tähendab, et tõenäoliselt on sellel andmestikul treenitud mudel võimeline edaspidi antud valdkonda kuuluvaid tekste paremini märgendama kui näiteks 2.9%-se osakaaluga "ZOLOOGIA" valdkonda kuuluvaid teoseid, kuna esimesel juhul on esindatud suurem variatiivsus erinevaid dokumente, mis tähendab, et ülesobitamise oht on väiksem.



***Joonis 27**. Iga EMS-i valdkonnaga seotud dokumentide arv.*

EMSi valdkond	Dokumentide arv	Dokumentide osakaal
MAJANDUS. MAJANDUSTEADUS. RAHANDUS. KAUBANDUS	3237	42.2%
ÜLDMÕISTED	2239	29.2%
VALITSUS. HALDUS. KODANIKUKAITSE. TURVATEENISTUS	1944	25.4%
RIIGIÕPETUS. POLIITIKA	1790	23.3%
SOTSIOLOOGIA. SOTSIAALPSÜHHOLOOGIA	1790	23.3%
HARIDUS. PEDAGOOGIKA	1654	21.6%
SOTSIAALPOLIITIKA	1582	20.6%
TEADUS JA TEHNIKA. TEADUSKORRALDUS. TEADUSMETODOLOOGIA. KULTUUR	1432	18.7%
TÖÖTINGIMUSED. TÖÖHÕIVE. AMETID	1413	18.4%

MEDITSIIN. BIOKEEMIA. ANATOOMIA. FÜSIOLOOGIA. FARMAKOLOOGIA. KOSMEETIKA	1299	16.9%
ÕIGUS	1206	15.7%
DEMOGRAAFIA. STATISTIKA	1056	13.8%
LOODUSKAITSE. KESKKONNAKAITSE. KESKKONNATEHNIKA	886	11.6%
PSÜHHOLOOGIA	844	11.0%
AJALUGU. ARHEOLOOGIA	799	10.4%
PÕLLUMAJANDUS. AIANDUS. LOOMAKASVATUS. VETERINAARIA	699	9.1%
ETNOLOOGIA. KULTUURIANTROPOLOOGIA	630	8.2%
BIOLOOGIA. MIKROBIOLOOGIA. GENEETIKA. ANTROPOLOOGIA	626	8.2%
AJAKIRJANDUS. KOMMUNIKATSIOON. MEEDIA. REKLAAM	535	7.0%
KEELETEADUS. KEELED	511	6.7%
EHITUS. SANITAARTEHNIKA	474	6.2%
FILOSOOFIA. EETIKA. SEMIOOTIKA	471	6.1%
KUNST. ARHITEKTUUR	469	6.1%
KODUMAJANDUS. TOIDUAINETETÖÖSTUS. TOITLUSTUS. OLME	407	5.3%
KIRJANDUSTEADUS. ILUKIRJANDUS. RAHVALUULE	392	5.1%
INFOTEADUS. RAAMATUKOGUNDUS. MÄLUASUTUSED. KIRJASTAMINE. TRÜKINDUS	384	5.0%
GEOGRAAFIA. GEODEESIA. KARTOGRAAFIA	335	4.4%
INFORMAATIKA. INFOTEHNOLOOGIA. AUTOMAATIKA	334	4.4%
GEOLOOGIA. HÜDROLOOGIA. KLIMATOLOOGIA	303	4.0%
LIIKLUS. TRANSPORT. SIDE	289	3.8%
SÕJANDUS. RELVAJÕUD. SÕJATEHNIKA	288	3.8%
RELIGIOON. TEOLOOGIA. ESOTEERIKA	263	3.4%
ELEKTROTEHNIKA. ELEKTROONIKA. ENERGEETIKA	252	3.3%
KEEMIA. KEEMIATÖÖSTUS	250	3.3%
HARRASTUSED. VABA AEG	225	2.9%
ZOOLOOGIA	220	2.9%
MATEMAATIKA	197	2.6%
FÜÜSIKA	191	2.5%
MUUSIKA	190	2.5%
BOTAANIKA	188	2.5%
METSANDUS. JAHINDUS. KALANDUS. METSA- JA PUIDUTÖÖSTUS. TSELLULOOSI- JA PABERITÖÖSTUS	180	2.3%

MASINAEHITUS. METALLITÖÖSTUS. METALLURGIA. MÄENDUS	161	2.1%
TEATER. TANTS	128	1.7%
SPORT. KEHAKULTUUR	113	1.5%
FOTOGRAAFIA. FILM. KINO	111	1.4%
ASTRONOOMIA. ASTROFÜÜSIKA. KOSMOSEUURIMINE	70	0.9%
KERGETÖÖSTUS. TEKSTIILITÖÖSTUS. NAHA- JA JALATSITÖÖSTUS	61	0.8%
KOHANIMED	12	0.2%

***Tabel 17**. Iga EMS-i valdkonnaga seotud dokumentide arv ning osakaal.*

3. Märksõnastamise metoodikate väljatöötamine

Käesolevas peatükis antakse ülevaade erinevatest märksõnastamise viisidest ning põhjendatakse konkreetsete metoodikate valikut.

3.1. Metoodikate jagunemine

Laias laastus saab võimalikud märksõnastamise metoodikad jagada kaheks - juhendatud ja juhendamata masinõppeks. Neist esimesel juhul kasutatakse masinõppemudeli treenimiseks märgendatud andmestikku. Mudeli treenimise käigus õpib mudel etteantud märgendust kasutades ära andmete tunnused, mille järgi on kõige lihtsam eristada eri märgendusega dokumente. Juhendamata masinõppe puhul ei ole mudelile märgendust ette antud. Selle asemel õpib mudel andmeid lihtsalt olemasolevate tunnuste põhjal mingil viisil grupeerima.
Kuna juhendatud masinõpe annab enamasti paremaid tulemusi kui juhendamata masinõpe, kasutatakse märgenduse olemasolul peamiselt just esimest metoodikat, mistõttu on see lähenemine valitud ka antud probleemi lahendamiseks.

3.1.1. Juhendatud masinõpe

Kaks peamist juhendatud masinõppel põhinevat klassifitseerimismudelit on logistiline regressioon (*logistic regression*) ja tugivektormasin (*support vector machine*). Varasemate kogemuste põhjal annavad need kaks mudelit loomuliku keele klassifitseerimisel võrdlemisi sarnaseid tulemusi, kuid logistilise regressiooni tulemused on pisut paremad.

3.1.1.1. Logistiline regressioon

Logistiline regressioon on mudel, mis ennustab positiivse sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt argumenttunnuse väärtuse muutumisest. Veel täpsemalt öeldes kasutatakse antud projektis binaarset logistilist regressiooni, mis eeldab, et uuritaval tunnusel on kaks võimalikku väärtust. See tähendab, et mudel oskab vastust ennustada jah/ei küsimusele. Näiteks saab treenida mudeli, mis ennustab, kas mingi konkreetse väljaandega on seotud märgend “kalandus” või mitte. Selleks, et ennustada iga võimalikku märgendit, treenitakse iga märgendi jaoks, mille puhul on olemas piisavalt treeningandmeid, eraldi mudel. Kõik ühe väljaandega seotud märgendid saadakse seega kõigi treenitud mudelite ennustusi kombineerides.

3.1.1.2. Tugivektormasin

Tugivektormasin on mudel, mis üritab vektorruumi paigutatud positiivsete ja negatiivsete näidete eraldamiseks leida neid võimalikult täpselt eristava eraldaja (joone, tasandi jne, sõltuvalt näidete dimensioonist). Uue andmepunkti klassifitseerimisel leitakse seega kõigepealt antud punkti asukoht vastavas ruumis ning klass määratakse vastavalt sellele, kummale poole eraldajat punkti asukoht jääb. Ehk kui meil on märgend "kalandus", siis on positiivseteks näideteks kõikide antud märgendiga seotud dokumentide vektoresitused ning negatiivseteks näideteks selliste dokumentide vektoresitused, mis antud märgendit ei

sisalda. Dokumentide vektoriteks teisendamine tähendab, et kõigepealt leitakse neile olulised tunnused (tekstide puhul võivad nendeks olla näiteks mingi märgendiga seotud dokumentide iseloomulikud sõnad) ja iga tunnus kandub üle vektorruumi üheks dimensiooniks. Seega on iga tunnuse väärtus ühtlasi vastava dimensiooni koordinaadiks. Kuna sarnaselt binaarsele logistilisele regressioonile on ka siin klassifitseerijad binaarsed, tuleb erinevate märgendite leidmiseks samuti eelnevalt treenitud binaarsed klassifitseerijad kombineerida.

3.1.1.3. Tehisnärvivõrgud

Tehisnärvivõrgud on koondsõna masinõppe mudelitele, mis jäljendavad bioloogilise närvivõrgu arhitektuuri. Kuigi tehisnärvivõrgud on viimasel ajal saavutanud suure populaarsuse, moodustavad need siiski ainult ühe osa kõigist olemasolevatest masinõppemeetoditest. Tehisnärvivõrkude hiljutine suur populaarsus on tingitud nende väga edukast rakendamisest probleemidele, mida klassikaliste masinõppemeetoditega on raske lahendada, peamiselt heli-, pildi- ja videotuvastuses, kuid ka masintõlkes ja muude loomuliku keele töötlusega seotud probleemide puhul.
Üks suuri takistusi tehisnärvivõrkude kasutamisel on piisavalt suure treeningandmestiku puudumine. Hea mudeli treenimiseks on sellele vaja ette anda tuhandeid näiteid ühe märksõna kohta. Minimaalselt võiks treeningandmeid märksõna kohta olla vähemalt mõnisada, kuid seda on siiski kordades rohkem kui enamike märksõnade jaoks on saadaval, mis on ka põhjus, miks antud projektis tehisnärvivõrke ei kasutata.

3.2. Tunnused

Masinõppemudeli ennustamise tulemused sõltuvad ka mudelile etteantud tunnustest, mistõttu tuleb neid valida hoolikalt. Selleks treenitakse mudelid erinevaid tunnuseid või nende kombinatsioone kasutades ning valitakse tulemusi hinnates parim variant. Loomuliku keele töötluse ülesannete puhul võib tunnustena kasutada nii toorteksti, selle derivaate (tekstist tuletatavat morfoloogilist infot) kui ka tekstiga seotud metaandmeid. Mudeldamisel tuleb aga silmas pidada, et kõiki mudeli treenimiseks kasutatud tunnuseid peab olema võimalik kasutada ka mudeli rakendamisel. See tähendab, et kui kasutaksime mudeli ehitamisel lisatunnustena näiteks metaandmetest saadavat teose autorit, “Agatha Christie”, mis aitab mudelil kiiresti aru saada, et tegemist on tõenäoliselt kriminaalromaaniga, siis tuleb sama info märgendajasse sisestada ka mudeli kasutamisel. Kuna teoste märgendamisprotsess toimub aga tihti samal ajal kui teose metaandmete eraldamine, ei saa eeldada, et vastav info märgendamise ajahetkel masinloetaval kujul olemas oleks. Seega on antud projekti raames mõistlikum piirduda teose teksti ning sellest otseselt tuletatava infoga. Projekti raames testitavad tunnused/ tunnuste kombinatsioonid on järgmised: üksustatud tavatekst, lemmad, ning lemmad + sõnaliigid (*part-of-speech tags*). Põgusa ülevaate igast tunnusest ning vastava tunnuse mudeldamisesse kaasamise põhjusest annab tabel 18.

Tunnus	**Kirjeldus**	**Näide**	**Kaasamise põhjendus**
Tavatekst	Tekst töötlemata kujul.	Kui Arno isaga koolimajja jõudis,	Kõige kiirem ja lihtsam viis tulemusi saada, sest võrreldes teiste

		olid tunnid juba alanud.	tunnustega pole andmetele vaja teha eraldi eeltöötlust; annab esmase ettekujutuse, mida üldse tulemustest oodata võiks.
Üksustatud tekst	Tekstis esinevad üksused (sõnad, kirjavahemärgid ja muud sümbolid) on eraldatud tühikuga.	Kui Arno isaga koolimajja jõudis , olid tunnid juba alanud .	Teatud juhtudel võib sõna esinemisvorm sisaldada samuti vajalikku infot, kuid võrreldes töötlemata tavatekstiga on eeliseks kirjavahemärkide sõnadest eristamine.
Lemmad	Tekst on üksustatud ja kõik tekstis esinevad sõnad on viidud algvormi.	kui Arno isa koolimaja jõudma , olema tund juba algama .	Tavateksti kasutamise miinuseks on see, et ühe ja sama sõna erinevaid vorme käsitletakse eraldi tunnustena (näiteks kui tekstis esinevad sõnad “kalandus”, “kalanduse”, “kalandust”, siis neid kõiki käsitletakse eraldi tunnustena, kuigi tegemist on sama sõna eri vormidega). Tulemuseks on väga pikad ja hõredad tunnuste vektorid (eriti morfoloogiliselt rikaste keelte puhul), mille pealt mudel ei suuda piisavalt üldistama õppida. Tekstide lemmatiseerimine lahendab selle probleemi.
Sõnaliigid	Tekstis esinevate sõnade sõnaliigid.	D|J H S S V Z V S D A|V Z (Vaata sümbolite tähendust siit.)	Sõltuvalt teksti žanrist on sõnaliikide esinemissagedustes erinevusi ning see info võib kasuks tulla ka teemade tuvastamisel.

Tabel 18. *Mudeldamisel kasutatavad tunnused.*

4. Märksõnastamise metoodikate valideerimine

Käesolevas peatükis kirjeldatakse, milliseid katseid märksõnastamise metoodikate valideerimiseks tehti ning millistele tulemustele nende käigus jõuti.

4.1 Mudelite valideerimise metoodika

Märksõnastamise mudeleid on võimalik hinnata:

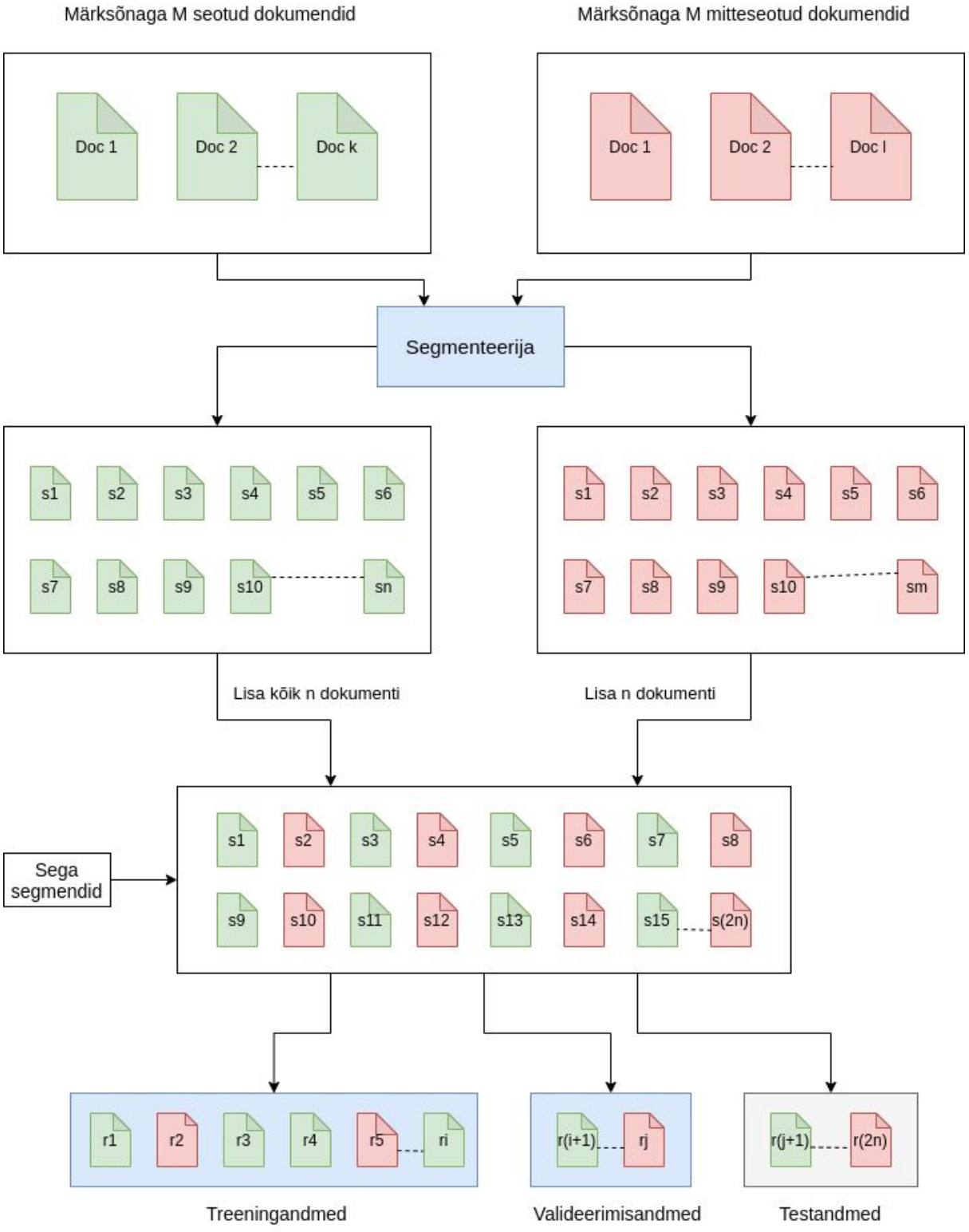
1. **Märgendipõhiselt** - tulemuste valideerimiseks jagatakse iga ennustusmudeli andmestik treening- ja testandmestikuks, treeningandmestiku peal treenitakse vastava märksõna binaarne ennustusmudel ning testandmestiku peal hinnatakse selle sooritust. Lõpptulem saadakse kõikide individuaalsete mudelite skooride keskmise leidmise teel.

2. **Näitepõhiselt** - tulemuste valideerimiseks jagatakse kõik unikaalsed dokumendid treening- ja testandmestikuks, treeningandmestiku peal treenitakse märgendite ennustusmudelid, mudeli sooritust hinnatakse kõigi testandmestiku dokumentide peal ning lõpptulemiks on mudeli keskmine ennustustulemus kõikidel testandmestiku dokumenditel.

Kuna antud juhul tähendaks näitepõhise valideerimise rakendamine ennustusmudelite näidete olulist vähenemist, mis omakorda tingiks nõrgema üldistusvõimega halvemad mudelid, on valideerimiseks mõistlik kasutada märgendipõhist valideerimismetoodikat.

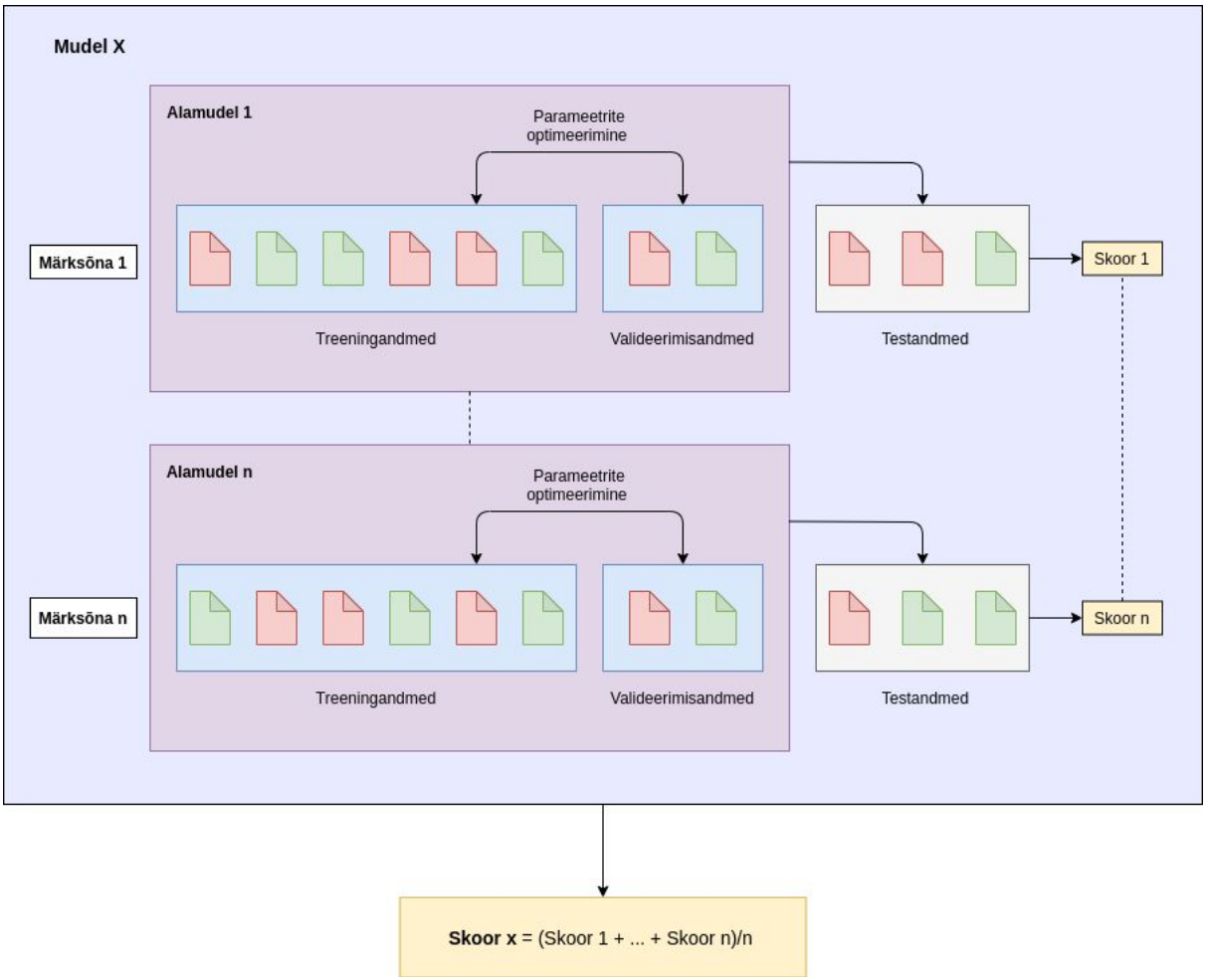
Märgendipõhise valideerimismetoodika töövoog on järgnev: olgu *k* kõikide märgendiga *M* seotud unikaalsete täistekstiliste dokumentide arv ning *l* kõikide unikaalsete täistekstiliste dokumentide arv, mis märgendiga *M* seotud ei ole. Kõik andmestiku dokumendid segmenteeritakse, mille järel tekib märgendiga *M* seotud *k-st* dokumendist *n* segmenti ning märgendiga *M* mitteseotud *l*-st dokumendist *m* segmenti. Kuna antud juhul on *m* alati suurem kui *n* (sest iga märgendiga seotud dokumentide arv on väiksem kui märgendiga mitteseotud dokumentide arv), siis valitakse *m* negatiivse segmendi hulgast juhuslik alamhulk suurusega *n.* Nii oleme saanud märksõna *M* mudeli treenimise andmete hulga suurusega *2n,* millest omakorda valitakse juhuslikud alamhulgad treenimiseks, valideerimiseks ja testimiseks, vastavalt suurustega 1.2n (60%), 0.4n (20%) ja 0.4n (20%). Treeningandmete peal treenitakse ennustusmudel, valideerimisandmeid kasutatakse mudelite parameetritele optimaalsete väärtuste leidmiseks ning testandmete peal hinnatakse märksõna *M* ennustusmudeli tulemusi (vt. joonis 28).

Lõplik tulemus leitakse kõikide märksõnade individuaalsete binaarsete mudelite tulemuste keskmise arvutamise teel (vt. joonis 29).



$$i = 0.6*2n$$
$$j = 0.8*2n$$

Joonis 28. *Märksõna M mudeli treening-, valideerimis- ning testandmestiku koostamise töövoog.*



Joonis 29. *Kõiki märksõnu kaasava mudeli hindamine.*

4.2. Mudelid

Järgnevas sektsioonis analüüsitakse eelnevas peatükis kirjeldatud tunnuseid või nende kombinatsioonide kasutades treenitud mudeleid. NB! Siin ja edaspidi kasutatakse sõna "mudel" samade tunnuste ning parameetrite alusel treenitud binaarsete klassifitseerijate kombinatsiooni kohta. See tähendab, et iga nö. mudel sisaldab omakorda *n* (*n* = märksõnade arv) binaarset alamudelit (vt. ka peatükk 3.1.1.1. Logistiline regressioon). Kuna erinevaid märksõnu on võimalik tüübi alusel kategoriseerida, on võimalik ka iga suurema mudeli alla koondatud märksõnade klassifitseerijad vastavatesse gruppidesse kombineerida ning tulemusi ka kitsamates kategooriates hinnata. Kõikidel treenitud mudelitel on klassifitseerimiseks kasutatud logistilist regressiooni. Tabel 19 annab ülevaate iga mudeli jaoks kasutatud tunnustest, ennustatavate märksõnade tüüpidest ning mudeli tulemustest klassikaliste masinõppe mudelite hindamise mõõdikute - täpsuse, saagise ja F1-skoori põhjal. Minimaalne näidete arv väljendab seatud piirmäära iga ennustatava märgendi jaoks vajaliku näidete arvu kohta ja märgendajate arv antud mudeliga ennustatavate märksõnade arvu (seotud minimaalse näidete arvuga - mida madalam lävend seada, seda rohkem märksõnu on võimalik ennustada, kuid samas langeb jälle mudeli keskmine sooritustulemus). Lisaks üldistele mudeli ennustustulemustele (toodud tabelis 19), on iga

mudeli analüüsis kirjeldatud ka tehniliste mõõdikute tulemused erinevate märgenditüüpide lõikes eraldiseisvalt. Kuigi otsustasime arendatavast prototüübist välja jätta isiku-, kollektiivi-, ning ajutise kollektiivi või sündmuse märgendid (vt. peatükk 2.5.4. Mudeldamisesse kaasatavad märksõnad), on kirjelduses välja toodud siiski ka nende märksõnade tulemused.

	Märgendid	Tunnused	Min näidete arv	Märgendajate arv	Täpsus	Saagis	F1-skoor
1.	Kõik	Tavatekst	50	7915	0.98	0.83	0.90
2.	Kõik	Lemmad	50	7915	0.98	0.85	0.91
3.	Kõik	Lemmad; sõnaliigid	50	7915	0.98	0.86	0.92
4.	EMSi kategooriad	Lemmad, sõnaliigid	100	48	0.83	0.79	0.82

Tabel 19. *Treenitud mudelite parameetrid ja tulemused*

Lisaks märksõnu ennustavatele mudelitele, treenisime mudeli ka EMSi kategooriate ennustamiseks, kuid kuna mudeli tulemused on märksõnu ennustavate mudelite omadest nõrgemad (eriti täpsuse osas, vt. tabel 19) ning mudeli kasuteguri hindamiseks oleks vaja täiendavat analüüsi ja selle põhjal sooritatud eksperimente*, mis ei mahu enam projekti skoopi, siis loetakse see analüüs lõpetatuks.

* Näiteks tekstist sagedasemate iseloomulike sõnade leidmine ning nende tekstiga seotud EMSi kategooriate alla kuuluvate märksõnadega võrdlemine.

Mudel 1

Mudel 1 on treenitud **üksustatud tavateksti** peal ennustamaks kõiki märksõnu minimaalse näidete arvuga 50. Kokku ületas nõutud lävendi 7915 märksõna ning mudeli keskmine täpsus on 0.98, saagis 0.83 ning F1-skoor 0.90. Eritüüpi märksõnu ennustavate mudelite tulemused leiab tabelist 20.

Kõige täpsemini ennustab mudel teemamärksõnu (täpsus 0.98), järgnevad kohamärksõna, kollektiivi nime märksõna ja ajutise kollektiivi või sündmuse nime märksõna täpsusega 0.97 Ajamärksõna täpsus on 0.96 ning vormi- ja isikunime märksõna oma 0.95.

Kõige parem saagise skoor on isikunime märksõnu ennustaval mudelil (0.86), järgnevad kollektiivi nime märksõnade mudel saagisega 0.85, aja- ja vormimärksõnu ennustav mudel saagisega 0.84, teema- ja kohamärksõnade mudelid saagisega 0.83 ning ajutise kollektiivi või sündmuse nime märksõna mudel saagisega 0.81.

Parim F1-skoor on kollektiivi nime märksõnana ennustaval mudelil (0.91), järgnevad aja-, teema- ja isikunime märksõnade mudelid skooriga 0.90, koha- ja vormimärksõnade mudelid skooriga 0.89 ning ajutise kollektiivi või sündmuse nime märksõnade mudel skooriga 0.88.

	Marc ID	Märgend	Märgendajate arv	Täpsus	Saagis	F1-skoor
1.	600	Isik	300	0.95	0.86	0.90
2.	610	Kollektiiv	368	0.97	0.85	0.91
3.	611	Ajutine kollektiiv või sündmus	68	0.97	0.81	0.88
4.	650	Teema	6470	0.98	0.83	0.90
5.	651	Koht	325	0.97	0.83	0.89
6.	653	Aeg	54	0.96	0.84	0.90
7.	655	Vorm	310	0.95	0.84	0.89

Tabel 20. *Tavateksti pealt treenitud mudeli eritüüpi märksõnade mudelite keskmised tulemused.*

Mudel 2

Mudel 2 on treenitud **lemmatiseeritud teksti** peal ennustamaks kõiki märksõnu minimaalse näidete arvuga 50. Kokku ületas nõutud lävendi 7915 märksõna ning mudeli keskmine täpsus on 0.98, saagis 0.85 ning F1-skoor 0.91. Eritüüpi märksõnu ennustavate mudelite tulemused leiab tabelist 21.

Kõige täpsemini ennustab mudel teema-, koha-, aja- ja kollektiivi nime märksõnu (täpsus 0.98), järgnevad isikunime- ja ajutise kollektiivi või sündmuse nime märksõna täpsusega 0.97 ning vormimärksõna täpsusega 0.96.

Kõige parem saagise skoor on ajamärksõnu ennustaval mudelil (0.87), järgnevad teema-, vormi- ja kollektiivi nime märksõnade mudelid saagisega 0.85, koha- ja isikunime märksõnade mudelid saagisega 0.84 ning ajutise kollektiivi või sündmuse nime märksõnade mudel saagisega 0.82.

Parim F1-skoor on ajamärksõnu ennustaval mudelil (0.92), järgevad teema- ja kollektiivi nime märksõnade mudelid skooriga 0.91, vormi-, isikunime- ja kohamärksõnade mudelid skooriga 0.90 ning ajutise kollektiivi või sündmuse nime märksõna mudel skooriga 0.89.

	Marc ID	Märgend	Märgendajate arv	Täpsus	Saagis	F1-skoor
1.	600	Isik	300	0.97	0.84	0.90
2.	610	Kollektiiv	368	0.98	0.85	0.91
3.	611	Ajutine kollektiiv või sündmus	68	0.97	0.82	0.89
4.	650	Teema	6470	0.98	0.85	0.91
5.	651	Koht	325	0.98	0.84	0.90
6.	653	Aeg	54	0.98	0.87	0.92
7.	655	Vorm	310	0.96	0.85	0.90

Tabel 21 *Lemmatiseeritud teksti pealt treenitud mudeli eritüüpi märksõnade mudelite keskmised tulemused.*

Mudel 3

Mudel 3 on treenitud **lemmatiseeritud teksti ning sõnaliikide** peal ennustamaks kõiki märksõnu minimaalse näidete arvuga 50. Kokku ületas nõutud lävendi 7915 märksõna ning mudeli keskmine täpsus on 0.98, saagis 0.86 ning F1-skoor 0.92. Eritüüpi märksõnu ennustavate mudelite tulemused leiab tabelist 22.

Kõige täpsemini ennustab mudel teema-, koha- ja kollektiivi nime märksõnu (täpsus 0.98), järgnevad isikunime-, aja- ja ajutise kollektiivi või sündmuse nime märksõna täpsusega 0.97 ning vormimärksõna täpsusega 0.96.

Kõige parem saagise skoor on ajamärksõnu ennustaval mudelil (0.87), järgnevad teema-, koha-, vormi- ja kollektiivi nime märksõnade mudelid saagisega 0.86, isikunime märksõna mudel saagisega 0.84 ning ajutise kollektiivi või sündmuse nime märksõna mudel saagisega 0.83.

Parim F1-skoor on teema-, koha-, aja ning kollektiivi märksõnu ennustaval mudelil (0.92), järgevad vormimärksõna mudel skooriga 0.91, isikumärksõna mudel skooriga 0.90 ja ajutise kollektiivi või sündmuse märksõna mudel skooriga 0.89.

	Marc ID	Märgend	Märgendajate arv	Täpsus	Saagis	F1-skoor
1.	600	Isik	300	0.97	0.84	0.90
2.	610	Kollektiiv	368	0.98	0.86	0.92
3.	611	Ajutine	68	0.97	0.83	0.89

		kollektiiv või sündmus				
4.	650	Teema	6470	0.98	0.86	0.92
5.	651	Koht	325	0.98	0.86	0.92
6.	653	Aeg	54	0.97	0.87	0.92
7.	655	Vorm	310	0.96	0.86	0.91

Tabel 22. *Lemmade ja sõnaliikide pealt treenitud mudeli eritüüpi märksõnade mudelite keskmised tulemused.*

4.1.1. Prototüübis kasutatav mudel

Näeme, et mudelite tulemused ei erine kuigi palju: nii üksustatud tavateksti, lemmade kui ka lemmade ja sõnaliikide kombinatsioone kasutavate mudelite täpsus on 0.98, kõige madalama saagisega (0.83) on üksustatud tavateksti pealt treenitud mudel, sellele järgneb lemmade pealt treenitud mudel (saagisega 0.85) ning parimaid tulemusi annab lemmade ja sõnaliikide kombinatsioone kasutav mudel (saagisega 0.86). Kuna F1-skoor tuleneb otseselt täpsuse ja saagise kombinatsioonist ning kõikide treenitud mudelite täpsus on võrdne, piisab siinkohal saagise analüüsist.

Lisaks üldtabelis (tabel 19) toodud üldskooridele on mudeli valimisel kasulik vaadata tulemusi ka relevantsete märksõnade alamudelite lõikes (teema-, koha-, aja-, ning vormimärgendid). Ka siin ei erine tulemused palju: tavatekstil treenitud mudeli vastavate märgendite täpsused jäävad vahemikku 0.95-0.98, saagised vahemikku 0.83-0.84. Lemmade pealt treenitud mudelite täpsused on vahemikus 0.96-0.98 ja saagised vahemikus 0.84-0.87 ning lemmade ja sõnaliikide kombinatsioonide peal treenitud mudelite täpsused vahemikus 0.96-0.98 ning saagised vahemikus 0.86-0.87. Kuigi tabeleid 21 ja 22 võrreldes näeme, et lemmade pealt treenitud mudeli relevantsete märgendite ennustamise täpsus on isegi natuke parem kui lemmasid sõnaliikidega kombineeriva oma, on viimane paremate tulemustega saagise osas. Kuna antud ülesande juures on oluline maksimeerida pigem saagist (kasutajale pakutakse rohkem märgendeid), siis saadud tulemuste põhjal on soovitatav eelistada lemmade ning sõnaliikide kombinatsioone kasutades treenitud mudelit (**Mudel 3**).

Siinkohal tuleks aga silmas pidada, et kasutegur morfoloogilise info mudelisse kaasamisest on sõltuv keelest - kuna valdav osa prototüübi mudeli jaoks kasutatud treeningandmestikust oli eestikeelne ning eesti keel on morfoloogiliselt küllalt rikas, siis aitab tekstide lemmatiseerimine tõsta ka mudeli tulemusi. Morfoloogiliselt vaeste keelte puhul (nt. inglise) morfoloogilise info mudeldamissesse kaasamine aga tulemusi märgatavalt tõsta ei pruugi. Kuna iga teksti märgendamisel tõstab tekstist morfoloogilise info eraldamine teatud määral

ka märgendamiseks kuluvat aega (vt. peatükk 5.2. Organisatsioonilised mõõdikud), siis tasub seda tulevikus uute mudelite treenimisel arvesse võtta.

4.2. Optimaalsed vaikeväärtused

Käesoleva peatüki eesmärgiks on treenitud mudeli tulemuste analüüsimisel välja selgitada märgendamisel kasutatavate parameetrite vaikeväärtused.

4.2.1. Segmentide arv

Selle peatüki eesmärgiks on välja selgitada optimaalne segmentide arv teoste märgendamiseks ning teha kindlaks, kas see jääb teose mahust sõltumata konstantseks või muutub dünaamiliselt koos teose mahu suurenemisega.

Kui märgendada tuleb 500-leheküljeline raamat, ei ole terve teose märgendajale edastamine tõenäoliselt vajalik, kuna a) mida mahukam teos (suurem segmentide arv) märgendajale edastada, seda aeglasem on kogu märgendamisprotsess ja b) teatud mahust alates märgendaja tulemused stabiliseeruvad või muutuvad koguni halvemaks (pakutakse rohkem irrelevantseid märgendeid). Et kindlaks teha, kui suur maht (mitu segmenti) teosest märgendajale saata tuleks, vaatleme, kuidas muutuvad mudelite tehniliste mõõdikute skoorid segmentide arvu kasvamise ja mitme segmendi kaasamisel jõuavad need maksimumväärtusteni. Kuna erinevate teoste segmentide arv varieerub palju (ühest pea kahe tuhandeni), siis on täpsema ülevaate saamiseks mõistlik katseseeriasse kaasatud teosed eelnevalt segmentide arvu alusel grupeerida.

Jagame teosed mahu alusel 5 gruppi piiridega 10-49 lehekülge (Grupp 1), 50-99 lehekülge (Grupp 2), 100-299 lehekülge (Grupp 3), 300-400 lehekülge (Grupp 4), ning 500 ja rohkem lehekülge (Grupp 5). Seejärel valime iga grupi jaoks välja 10 juhuslikku teost ning analüüsime iga teose n_{min} * segmenti.

* n_{min} - gruppi kuuluva kõige lühema teose segmentide arv, vastavalt 15, 50, 100, 300 ja 500.

Et kindlaks teha, millisele mõõdikule optimaalse segmentide arvu välja selgitamisel enim keskenduma peaks, uurime kõigepealt, kui palju märgendeid erinevate segmentide arvude puhul keskmiselt ennustatakse. Kui ennustatud märgendite arvud kasvavad kiiresti, tuleks suuremat tähelepanu pöörata täpsusele, kui märgendeid on aga pigem vähe, siis saagisele. Kuna mitme segmendi analüüsimisel on võimalik märgendid esinemissageduste alusel järjestada*, uurime lisaks, kuidas muutub ennustatud märgendite hulk erinevate sageduslävendite puhul.

* Kui märgend *m1* on 100-st analüüsitud segmendist seotud ainult ühe dokumendiga, on tema õigsuse tõenäosus tunduvalt väiksem kui näiteks 47 segmendiga seotud märgendil *m2*.

Tabelist 23 näeme, et 10 segmendi puhul ennustatakse keskmiselt 19 märgendit, 20 segmendi puhul 28 märgendit ning iga järgmise 10 segmendi lisamisel kasvab ennustatud

märgendite hulk 2-6 võrra. Näeme, et ainult 20% lävendi puhul (märgend peab esinema 20%-s analüüsitud segmendis) kahaneb märgendite hulk 10 segmendi puhul poole võrra ning suuremate segmentide arvude puhul veel rohkem. 100%-se lävendi puhul (märgend peab esinema kõikides ennustatud segmentides) ületab selle 10 segmendi analüüsil keskmiselt 1 märgend ning kõrgema segmentide arvu puhul mitte ükski.

	Keskmiselt ennustatud märgendeid					
Segmentide arv	**Lävend puudub**	**Lävend 20%**	**Lävend 40%**	**Lävend 60%**	**Lävend 80%**	**Lävend 100%**
10	19	10	7	5	3	1
20	28	9	7	4	2	0
30	34	8	6	4	2	0
40	39	8	6	4	2	0
50	43	8	6	4	2	0
60	47	7	5	4	2	0
70	50	7	5	3	2	0
80	53	7	5	3	2	0
90	56	7	5	3	1	0
100	58	7	5	3	1	0

Tabel 23: *Keskmiselt ennustatud märgendite arv sõltuvalt analüüsi kaasatud segmentide arvust ja seatud lävendist.*

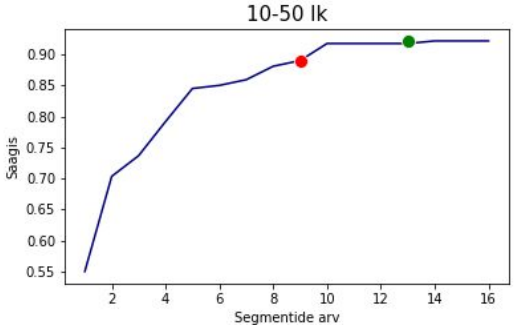
Kuna näeme, et lävendite seadmisega on märgendite arvu võimalik tõhusalt vähendada, seame eesmärgiks võimalikult kõrge saagise saavutamise, kuna saagis väljendab, kui palju tõeseid märgendeid on ennustatud märgendite hulka jõudnud. See tähendab, et iga uue ennustatud märgendi lisamisel saab saagis jääda samaks või tõusta, aga mitte langeda. Seega kui teoses on kokku näiteks 394 lehekülge, aga saagis saavutab maksimumi pärast 134 segmendi läbi töötamist, ei ole märgendajal mõtet enam ülejäänud 260 lehekülge analüüsida, kuna stabiliseerunud maksimaalse saagise puhul saab teine oluline mõõdik - täpsus - jääda üksnes samaks või muutuda halvemaks. See tähendab, et kui saagis jõuab mõistliku arvu segmentide kaasamise järel piisavalt hea tulemuseni, ei ole märgendamisele kaasatavate segmentide arvu leidmisel enam täpsuse analüüs oluline, kuna maksimaalse saagise korral leidub ennustatud märgendite seas alamhulk ka maksimaalse täpsusega.

Vaatleme, mitme segmendi lisamisel jõuab saagis erinevate gruppide lõikes maksimumini ja stabiliseerub (vt. tabel 24 ja joonised 30a-30e). Esimeses grupis jõuab saagis maksimumini 14 segmendi analüüsimise järel, teises grupis 30 segmendi analüüsimise järel, kolmandas grupis 39 segmendi analüüsimise järel, neljandas grupis 282 segmendi analüüsimise järel ning viiendas grupis 45 segmendi analüüsimise järel. Näeme, et neljandas grupis kulub saagise maksimeerimiseni teiste gruppidega võrreldes tunduvalt rohkem aega, aga siinkohal tuleb tähele panna, et eksperimenti kaasatud teoste puhul pole täiendavat jaotusanalüüsi

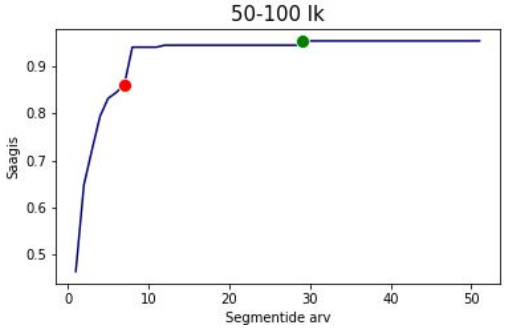
tehtud, mistõttu on tõenäoliselt tegu mõnest erindist tuleneva kallutatusega ning tegelik keskmine jääb ilmselt kolmanda ja viienda grupi vahele (väärtusega u. 42 segmenti).

Teose maht	Analüüsitud segmentide arv	90%-ni jõudmiseks kuluv segmentide arv	Maksimumini jõudmiseks kuluv segmentide arv	Maksimumi väärtus	Maksimumi ja 90%-ni jõudmiseks kuluvate segmentide vahe
10-49 lk	15	10	14	92%	4
50-99 lk	50	8	30	96%	22
100-299 lk	100	4	39	97%	35
300-499 lk	300	5	282	99%	277
500-... lk	500	6	45	94%	39

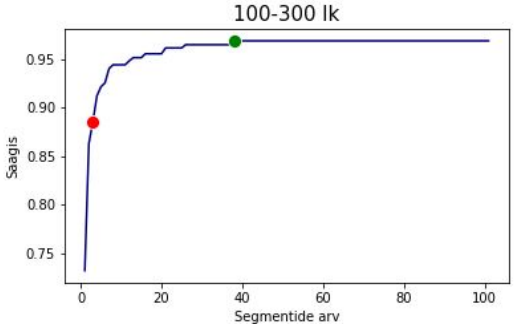
Tabel 24. *Saagise 90%-ni ja maksimumini jõudmiseks kuluv segmentide arv ja omavaheline võrdlus.*



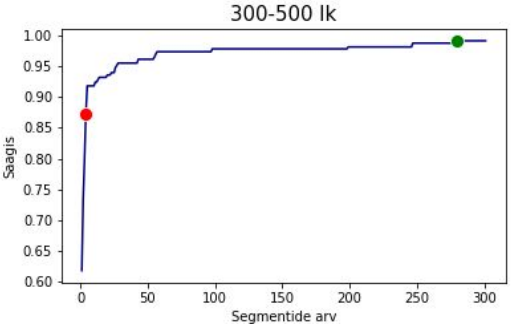
Joonis 30a. *Grupp 1*



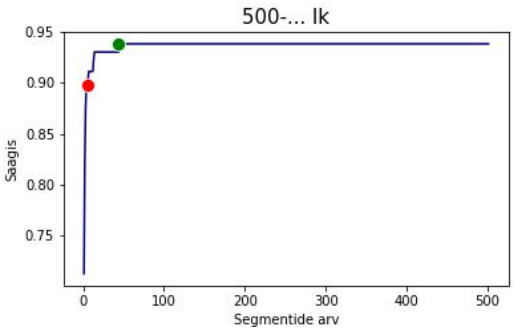
Joonis 30b. *Grupp 2*



Joonis 30c. *Grupp 3*



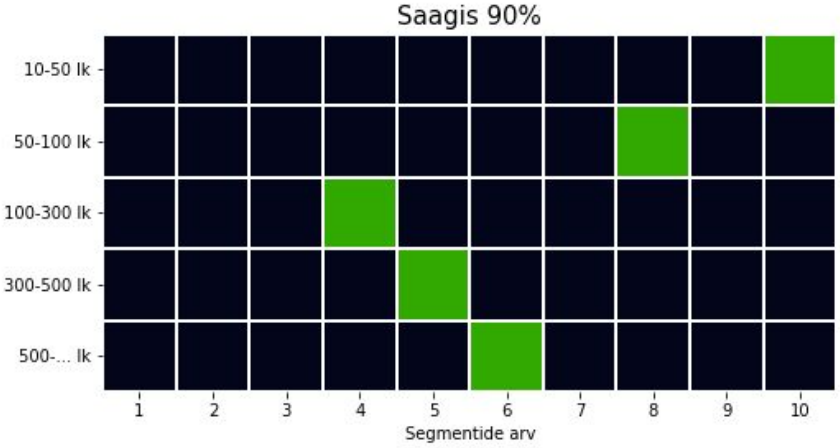
Joonis 30d. *Grupp 4*



Joonis 30e. *Grupp 5*

Joonis 30. *Saagise muutumine segmentide arvu kasvamisel erinevate teosemahtude lõikes. Punane (esimene) täpp tähistab 90%-ni jõudmist (joonisel natuke allapoole nihkes) ning roheline täpp maksimumini jõudmist.*

Kuna ühe segmendi lisamisega kaasnev töötlusaeg (teksti eraldamine + teksti kvaliteedi valideerimine + morfoloogilise info eraldamine + märgendamine) on 6 sekundit, mis 40 segmendi analüüsimisel tähendab 4 minutit, siis oleks kasulik uurida ka, millal muutub saagise skoor üldjuhul piisavalt heaks, kuna mõne saagise protsendi ohverdamine võib anda märgatava ajalise võidu. Uurime, mitme segmendi lisamisel ületab saagise skoor 90% piiri (vt. joonised 30a-30e ja joonis 31). Esimese grupi saagis jõuab 90%-ni pärast 10 segmendi lisamist, teise grupi oma pärast 8 segmendi lisamist, kolmanda grupi oma pärast 4 segmendi lisamist, neljanda grupi oma pärast 5 segmendi lisamist ning viienda grupi oma pärast 6 segmendi lisamist. Kuna näeme, et kõige kauem kulub 90% piirini jõudmiseks esimesel grupil, kuhu kuuluvate teoste maht on kõige väiksem, võib arvata, et põhjus on väikesest valimist tuleneval kallutatusel. Küll aga annab analüüs infot, et kõigi gruppide saagis jõuab vähemalt 90%-ni ainult 10 segmendi kaasamisel.

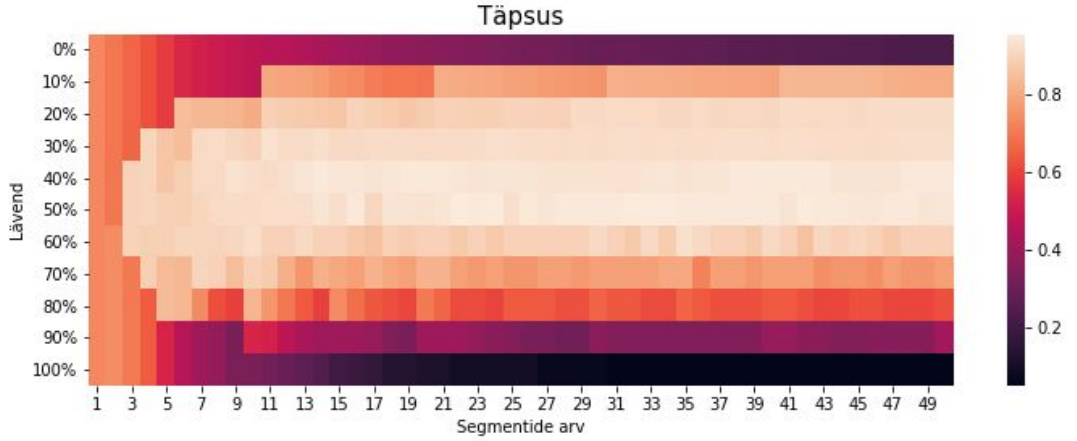


Joonis 31. *Mitme segmendi lisamisel jõuab saagis 90%-ni erineva mahuga teoste lõikes.*

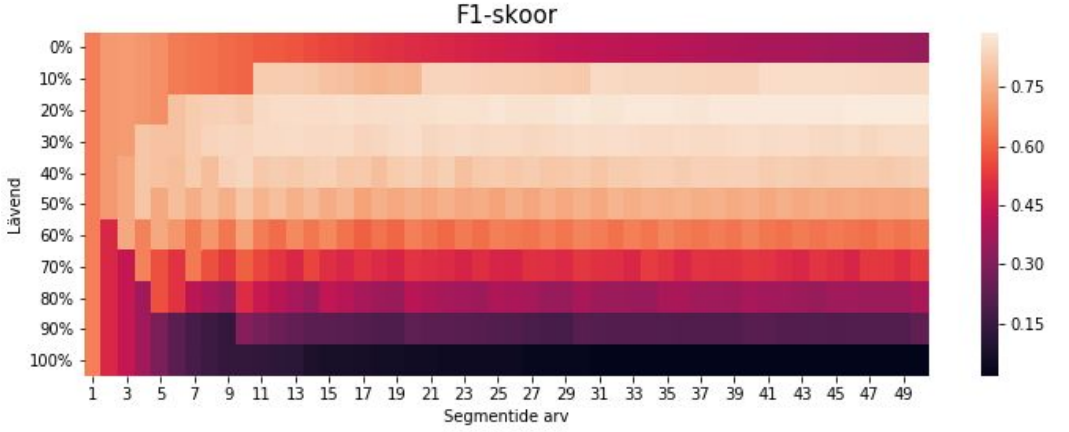
Seega **soovitame kaasatavate segmentide vaikeväärtuseks seada 10**, kuid anda kasutajale võimaluse seda soovi korral muuta (siin peab kasutaja lihtsalt oskama arvestada iga segmendi lisamisel kasvava märgendamise ajakuluga). Kuna kõigi gruppide saagised ületavad selle väärtuse puhul 90% piiri, ei ole mahu alusel eristamine vajalik.

4.2.2. Märgendite lävend

Selle peatüki eesmärgiks on analüüsida, milline märksõnade lävend annab kindlaks määratud segmentide arvu juures kõige paremaid tulemusi. Selleks analüüsime, kuidas muutuvad täpsus ja nii saagist kui täpsust kaasav F1-skoor erinevatel lävenditel 1-50 kaasatud segmendi ulatuses. Joonisel 32 on kujutatud täpsuse sõltuvus lävendist ja kaasatud segmentide hulgast (mida heledam toon, seda parem skoor) ning joonisel 33 sama info F1-skoori kohta. Näeme, et keskmiselt on täpsus kõige kõrgem lävendite vahemikus 30%-60% ning muutub küllalt stabiilseks juba 10 segmendi lisamisel. F1-skoori puhul jäävad parimad tulemused lävendivahemikku 20%-40%. Antud tulemuste põhjal **soovitame vaikelävendiks määrata 40%**, mis aitab tagada võimalikult suure täpsuse ohverdamata sealjuures oluliselt saagist. See tähendab, et 10 segmendi kaasamisel on keskmiseks tagastatavate märgendite arvuks 7 (vt. ülal tabel 23). Siiski võiks kasutajale jääda ka lävendi muutmise võimalus (kõrgema lävendi puhul tagastatakse vähem märgendeid ning madalama puhul rohkem).



Joonis 32: *Märgendaja täpsus sõltuvalt kaasatud segementide arvust ja seatud lävendist.*



Joonis 33: *Märgendaja F1-skoor sõltuvalt kaasatud segementide arvust ja seatud lävendist.*

Kokkuvõtlikult saame optimaalseteks vaikeväärtusteks:

Segmentide arv	10
Märksõnade lävend	40%

4.2.3. Näited

Järgnevalt on toodud mõned juhuslikult valitud näited illustreerimaks mudeli täpsuse muutumist lävendi suurendamisel. Kõik näidetes olevad märgendid on leitud 10 segmendi analüüsimisel.

Ennustatud märgendite hulk = TP + FP
(õigesti ennustatud märgendid ja valesti ennustatud märgendid) *

Tegelike märgendite hulk = TP + FN
(õigesti ennustatud märgendid ja ennustamata jäetud õiged märgendid) *

* vt. ka peatükki 5. Mõõdikute süsteem

Näide 1

Autor	**Teos**	**Maht**
Enn Pirrus	“Eestimaa suured kivid: suurte rändrahnude lugu”	172 lk

	Lävend = 0	**Lävend = 0.4**
TP	650 rändrahnud 650 mineraloogia 650 looduskivid 650 loodusmälestised 650 looduskaitse 655 võrguväljaanded 651 Eesti	650 rändrahnud 650 mineraloogia 650 looduskivid 650 loodusmälestised 650 looduskaitse 655 võrguväljaanded
FP	650 kodulugu 650 matkamine 650 kivimid	
FN		651 Eesti
TN		650 kodulugu 650 matkamine 650 kivimid
Täpsus	70%	100%
Saagis	100%	86%
F1-skoor	82%	92%

Näeme, et lävendi puudumisel on küll ennustatud märgendite hulgas kõik tegelikult teosega seotud märgendid (saagis=100%), kuid lisaks ka 3 märgendit, mis algselt teosega seotud pole: “kodulugu”, “matkamine” ja “kivimid”. Lävendi tõstmisel 40%-ni kaovad kõik need märgendid ennustatute hulgast, kuid lisaks eemaldatakse ka korrektne märgend “Eesti”.

Näide 2

Autor	**Teos**	**Maht**
Roosmarii Kurvits	“Eesti ajalehtede välimus 1806-2005”	423 lk

	Lävend = 0	**Lävend = 0.4**
TP	650 ajakirjandusajalugu 650 ajalehed 650 ajalehekujundus 650 graafiline disain 650 sisuanalüüs 650 toimetamine 650 trükiajakirjandus 651 Eesti 653 19. saj. 653 20. saj. 655 dissertatsioonid	650 ajakirjandusajalugu 650 ajalehed 650 ajalehekujundus 650 graafiline disain 650 sisuanalüüs 650 toimetamine 650 trükiajakirjandus 651 Eesti 653 19. saj.
FP	650 ajakirjandus 650 ajakirjandusuuringud 650 ajalugu 650 faktid 650 sisuloome 650 tekstianalüüs 650 tekstid 655 võrguväljaanded	
FN	653 2000-ndad	653 20. saj. 653 2000-ndad 655 dissertatsioonid
TN		650 ajakirjandus 650 ajakirjandusuuringud 650 ajalugu 650 faktid 650 sisuloome 650 tekstianalüüs 650 tekstid 655 võrguväljaanded
Täpsus	58%	100%
Saagis	92%	75%

F1-skoor	71%	86%

Näeme taaskord, et ilma lävendita on saagis kõrge ning tegelikest teosega seotud märgenditest on ennustamata jäänud ainult üks (ajamärksõna “2000-ndad”). Küll aga on ligi pooled ennustatud märgenditest (originaalmärgenduse põhjal) valed. Tõstes lävendi 40%-ni, kaovad ennustatud märgendite seast taaskord valepositiivsed, kuid koos nendega ka tõesed märgendid “20. saj” ja “dissertatsioonid”.

Näide 3

Autor	**Teos**	**Maht**
Vello Kala	“Hüdrograafia alused”	315 lk

	Lävend = 0	**Lävend = 0.4**
TP	650 hüdrograafia 655 kõrgkooliõpikud	650 hüdrograafia 655 kõrgkooliõpikud
FP	650 füüsika 650 satelliitnavigatsioonisüsteemid 650 sisehindamine 650 teadus- ja arendustegevus 650 tehnikateadused 650 tehnohooldus 650 tööstusrobotid 655 artiklikogumikud 655 e-raamatud 655 juhendid 655 kooliõpikud 655 kutsekooliõpikud 655 ülesanded	655 e-raamatud
FN		
TN		
Täpsus	13%	67%
Saagis	100%	100%
F1-skoor	23%	80%

Näeme taaskord eelmiste näidetega sarnast mustrit: lävendi puudumisel on saagis küll maksimaalne, kuid seekord on ennustatud 15 märgendist valepositiivseid kokku tervelt 13. Lävendi 40%-ni tõstmisel vabanetakse aga peaaegu kõigist neist peale ühe (“e-raamatud”).

4.3. Märksõnade sageduse mõju mudelite ennustamise edukusele

Et valdav osa mudelitest on treenitud täistekstide mõttes väga väikese treeningandmestiku peal, on mõistlik uurida, kuidas treeningandmete arv mudelite ennustamise edukust mõjutab. Tabelis 25 on toodud juhuslikult valitud mudelite keskmised näitajad erinevate märksõnadega seotud dokumentide arvu kohta. Dokumentide arvud veergudesse on võetud sagedusjaotuse kvartiilide järgi ning iga arvu kohta on valitud sada juhuslikku mudelit.
Tabelit vaadates on selge, et ainult ühest väljaandest koosneva treeningandmestiku peal treenitud mudelil puudub üldistusvõime ning see ülesobitub tugevalt nähtud segmentidele, mis väljendub madalas saagises. Samas juba väikese väljaannete arvu suurenemisega paraneb saagis oluliselt ning ka täpsus muutub paremaks.

Märksõnaga seotud dokumentide arv	1	2	3 kuni 4	rohkem kui 4
Keskmine täpsus	0.899	0.959	0.978	0.965
Keskmine saagis	0.537	0.646	0.747	0.835
Keskmine f1	0.672	0.772	0.847	0.895

Tabel 25. *Märksõnaga seotud dokumentide arvu mõju ennustamise edukusele (lemmad + sõnaliigid)*

5. Mõõdikute süsteem

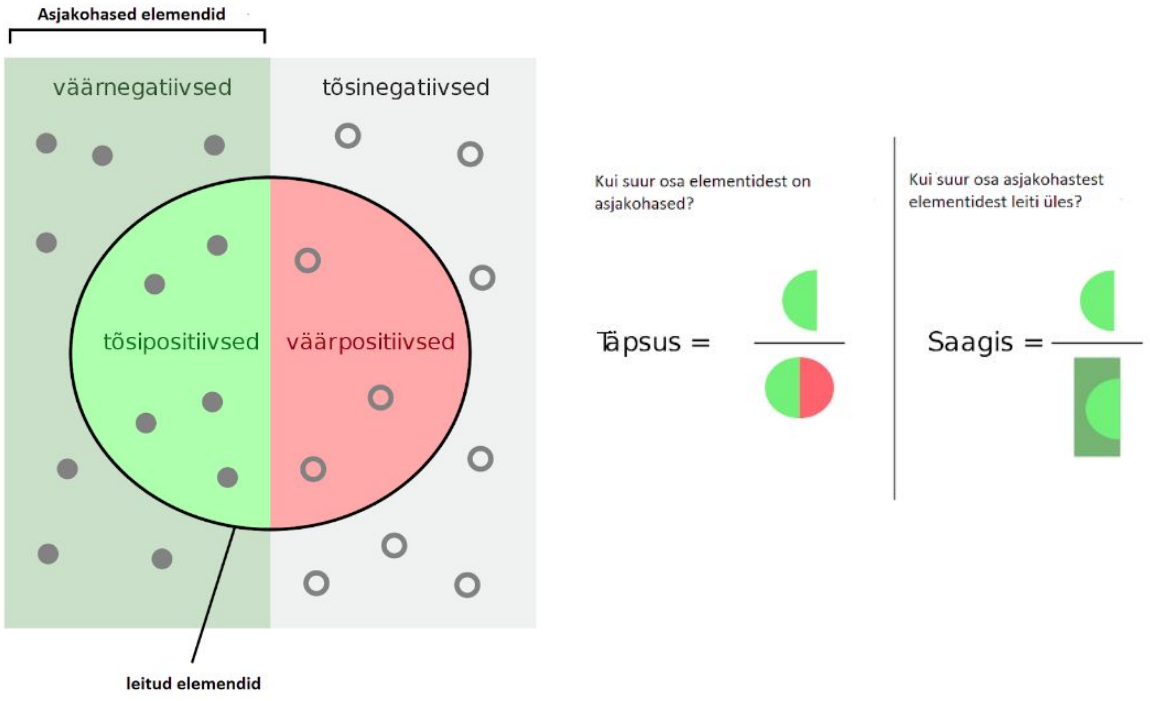
Käesolevas peatükis kirjeldatakse projekti tehnilisi ja organisatsioonilisi parameetreid ja eesmärke, millega hinnatakse projekti edukust.

5.1. Tehnilised mõõdikud

Antud projekti tehnilisi tulemusi hinnatakse masinõppe seisukohast klassikaliste mõõdikute alusel, milleks on saagis (*recall*), täpsus (*precision*) ja F1-skoor. Hindamine toimub automaatselt ning selleks jagatakse andmestik kaheks, kus 80% andmeid kasutatakse treenimiseks ning 20% valideerimiseks. Eelnimetatud tehniliste mõõdikute definitsioonid ning eesmärgiks seatud tulemused leiab tabelist 26 (vt. ka joonis 34b).

Mõõdik	**Eesmärk**	**Selgitus**
Täpsus	0.7	Täpsus näitab, kui suure osa moodustavad õigesti klassifitseeritud (tõsipositiivsed) märgendid kõikide ennustatud märgendite hulgast.
Saagis	0.8	Saagis näitab, kui suure osa moodustavad õigesti klassifitseeritud (tõsipositiivsed) märgendid kõikide asjakohaste märgendite hulgast.
F1-skoor	0.8	F1 skoor on saagise ja täpsuse keskmine, mis arvutatakse automaatselt.

Tabel 26. *Tehnilised mõõdikud*



Joonis 34a. *Väärnegatiivsed, tõsinegatiivsed, tõsipositiivsed ja väärpositiivsed elemendid.*

Joonis 34b. *Täpsus ja saagis.*

Tehniliste mõõdikute probleemid

Vaatleme, milliseid märgendeid ennustavad kaks erinevat mudelit teosest “Põhja-Aasia - Vaikse Ookeani Divisjon” pärinevale segmendile (vt. tabel 27). Teosega treeningandmestikus seotud märgendid on “adventistid” ja “misjon”. Mõlemad mudelid ennustavad need märksõnad korrektselt, kusjuures esimene mudel väljastabki ainult need märksõnad tagades seega parima võimaliku täpsuse, saagise ning F1-skoori. Mudel 2 ennustab aga lisaks ka märksõnad “usuõpetus”, “eesti” ning “adventism”, mistõttu on mudeli täpsus ning F1-skoor esimese mudeliga võrreldes märgatavalt madalamad (vastavalt 0.4 ja 0.57). Ennustatud “valesid” märgendeid analüüsides näeme aga, et semantiliselt ei erine need palju teosega tegelikult seotud märgenditest - “adventism” jagab korrektse märgendiga “adventistid” koguni sama tüve. See tähendab, et mudeli automaatselt arvutatavate tehniliste mõõdikute skoorid ei pruugi alati anda lõpliku ülevaadet mudeli kvaliteedist.

	Mudel 1	**Mudel 2**
adventistid	TP	TP
misjon	TP	TP
usuõpetus	TN	FP
eesti	TN	FP
adventism	TN	FP
Täpsus	1.0	0.4
Saagis	1.0	1.0
F1-skoor	1.0	0.57

Tabel 27. *Kahe erineva mudeli ennustustulemused*

5.2. Organisatsioonilised mõõdikud

See peatükk keskendub organisatsioonilistele mõõdikutele, mille hindamine aitab kindlaks teha projekti edukuse.

Projekti organisatsioonilised mõõdikud koos vastavate väärtustega leiab tabelist 28.

Mõõdik	Manuaalselt	Masinõpe
Kiirus: kui kaua võtab aega ühe teose märksõnastamine käsitsi vs masinõppe toega.	14 minutit	1-10 minutit sõltuvalt teose mahust ning kaasatud segmentide arvust
Automaatse märgendaja tööriista paigaldamise lihtsus: kui kaua võtab paigaldamine aega.	N/A	1-5 minutit (+ 1h koos virtuaalmasina paigaldusega)
Automaatse märgendaja tööriista uuendamise lihtsus: kui kaua võtab API uuendamine aega.	N/A	1-5 minutit
Kui kaua võtab uue mudeli treenimine aega.	N/A	1 märksõna ennustava binaarse mudeli treenimine kuni 15 min (sõltuvalt treeningandmete mahust);

Tabel 28. *Organisatsioonilised tulemid.*

Märgendamise ajakulu

Kuna dokumendi automaatse märgendamise ajakulu on lisaks otsesele märgendamis-protsessile (dokumendi märgendamismudeliga analüüsimisele) seotud ka erinevate eeltöötlusprotsessidega ning on lisaks sõltuv dokumendi mahust, on mõistlik uurida erinevates etappides kuluvat aega ning konstantse ajahinnangu asemel kaasata muutuvatest parameetritest tekkiv varieeruvus. Dokumentide märgendamise põhietapid on järgmised:

1) Dokumentide konverteerimine PDFiks (vajalik segmenteerimiseks);
2) Konverteeritud PDFi segmentideks tükeldamine;
3) $k_{seg} + v_{seg}$ PDFi segmendist teksti eraldamine (k_{seg} = eelnevalt kindlaks tehtud optimaalne segmentide arv, v_{seg} = varusse kaasatavate segmentide arv[4]);
4) Eraldatud tekstisegmentide automaatne kvaliteedianalüüs;
5) Kvaliteedianalüüsi läbinud tekstisegmentidest morfoloogilise informatsiooni eraldamine (mida kasutatakse mudeli tunnustena);
6) Segmentide märgendamine.

Märgendamise ajakulu hindamiseks viisime läbi 20 katset iga märgendamise alamprotsessi jaoks ning leidsime iga katseseeria keskmise. Iga alamprotsessi keskmine tulemus on toodud tabelis 29.

[4] Varusegmentide lisamine on vajalik, kuna erinevate eeltöötlus- ja valideerimisprotsesside käigus võib osa segmentidest kasutuskõlbmatuks osutuda. See tähendab, et kui kasutaja valib märgendemiseks näiteks 10 segmenti, aga 4 neist ei läbi kvaliteedikontrolli, saadetakse märgendajale ainult 6 segmenti, mis omakorda võib märgendustulemusi negatiivselt mõjutada. Varu kaasamine aitab tagada, et märgendajale saadetakse lõpuks sama arv segmente nagu kasutaja soovis.

Sümbol	Kirjeldus	Väärtus	Ühik
t_{conv}	Dokumendi PDFiks konverteerimisele kuluv aeg (ühe MB kohta)	78	s/MB
t_{seg}	PDFi segmenteerimisele kuluv aeg (ühe MB kohta)	1.4	s/MB
t_{text}	Ühest segmendist teksti eraldamisele kuluv aeg	0.1	s
t_{val}	Eraldatud tekstisegmendi kasutatavuse valideerimiseks kuluv aeg	0.002	s
t_{morph}	Valideeritud tekstisegmendist morfoloogilise info eraldamisele kuluv aeg	3.32	s
t_{tag}	Ühe segmendi märgendamiseks kuluv aeg	2.5	s
s_{doc}	Dokumendi suurus	**x**	MB
k_{seg}	Märgendamiseks kasutatavate segmentide arv	**y** = 10	seg
v_{seg}	Varusse kaasatud segmentide arv	**z** = 5	seg
b_{doc}	Dokumenditüübi konverteerimiskordaja (dokument on PDF)	0	
	Dokumenditüübi konverteerimiskordaja (dokument pole PDF)	1	

Tabel 29. *Märgendamise ajakulu parameetrid ja konstandid*

Erinevaid alamprotsesse arvestades saame ühe dokumendi märgendamiseks kuluva aja valemi:

$$b_{doc} \cdot s_{doc} \cdot t_{conv} + s_{doc} \cdot t_{seg} + (k_{seg} + v_{seg}) \cdot t_{text} + (k_{seg} + v_{seg}) \cdot t_{val} + k_{seg} \cdot t_{morph} + k_{seg} \cdot t_{tag}$$

Pärast leitud konstantide asendamist jääb valemisse kolm parameetrit: dokumendi suurus (s_{doc}), kaasatavate segmentide arv (k_{seg}) ning varusse kaasatavate segmentide arv (v_{seg}), millest kasutaja jaoks olulisim on kaasatavate segmentide arv (dokumendi suuruse ning varusse kaasatavate segmentide osas puudub kasutajal kontroll). Pärast lihtsustamist taandub valem järgnevale kujule:

1) Kui dokument **on** PDF (konverteerida pole vaja):

$$t = 1.4 \cdot s_{doc} + 5.9 \cdot k_{seg} + 0.1 \cdot v_{seg}$$

2) Kui dokument **ei ole** PDF (lisandub konverteerimiseks kuluv aeg):

$$t = 79.4 \cdot s_{doc} + 5.9 \cdot k_{seg} + 0.1 \cdot v_{seg}$$

See tähendab, et ühe segmendi lisamisel tõuseb dokumendi märgendamiseks kuluv aeg umbes **6 sekundit**. Mõned näited märgendamiseks kuluvast ajast sõltuvalt dokumendi suurusest, dokumendi tüübist, kaasatud põhi- ning varusegmentide arvust leiab tabelist 30.

Dokumendi suurus	2 MB	1.7 MB	7 MB
Dokumendi tüüp	.pdf	.epub	.pdf
Kaasatud segmentide arv	10	10	20
Varusegmentide arv	5	5	5
Märgendamiseks kuluv aeg	62.5 s	194.7 s	128.5 s

Tabel 30: *Märgendamise ajakulu sõltuvalt dokumendi suurusest, tüübist, kaasatavate põhi- ning varusegmentide arvust.*

6. Funktsionaalsed ja mittefunktsionaalsed nõuded

Käesolevas peatükis on kirjeldatud automaatse märksõnastaja süsteemi (edaspidi KRATT) funktsionaalsed ja mittefunktsionaalsed nõuded.

6.1. Funktsionaalsed nõuded

Järgnevalt on toodud süsteemile esitatud funktsionaalsed nõuded koos neile vastavate peatükkide viidetega.

	Nõue	Viide
3.1.1.1.	KRATT peab automaatselt märksõnastama täisteksti	7.1. Süsteemi arhitektuuri kirjeldus; 7.2. Märgendajate haldamine; 7.3. Märgendajate rakendamine
3.1.1.2.	KRATT peab omama veebipõhist liidest	7.2. Märgendajate haldamine; 7.3. Märgendajate rakendamine
3.1.1.3.	KRATT peab suutma tuvastada väljaande keele	7.1.2. Krati mooduli kirjeldus
3.1.1.3.	KRATT peab suutma automaatselt märksõnastada eestikeelseid väljaandeid	7.1. Süsteemi arhitektuuri kirjeldus; 7.1.2. Krati mooduli kirjeldus; 7.3. Märgendajate rakendamine
3.1.1.4.	KRATT peab olema nii programmeeritud, et oleks võimalus liidestada teisi keeli (inglise, vene)	7.1. Süsteemi arhitektuuri kirjeldus; 7.1.1. TEXTA Toolkit-i mooduli kirjeldus
3.1.1.5.	KRATT peab eristama teema-, aja-, koha-, vormi- ja žanrimärksõnu	2.5.4. Mudeldamisesse kaasatavad märksõnad
3.1.1.6.	KRATT peab eristatama isikute ja organisatsioonide nimesid ning kasutama neid vajadusel märksõnadena	2.5.4. Mudeldamisesse kaasatavad märksõnad
3.1.1.7.	KRATT peab kasutama märksõnastamiseks ainult Eesti märksõnastikus (EMS) olevaid märksõnu (märksõnastik on pidevas muutumises)	2.5.4. Mudeldamisesse kaasatavad märksõnad
3.1.1.8.	KRATT peab kasutama sisendiks täisteksti .epub, .pdf, .html või .txt formaadis, mis on kättesaadav läbi URL'i	8.2. Toetatud failiformaadid ja sisendi töötlemine
3.1.1.9.	KRATT peab kasutama sisendiks täisteksti .epub, .pdf, .html või .txt formaadis, mis on üleslaetav kohalikust arvutist veebilehitseja vahendusel	8.2. Toetatud failiformaadid ja sisendi töötlemine
3.1.1.10.	KRATT peab väljastama genereeritud asjakohased märksõnad, mis iseloomustavad sisendiks olnud teksti	7.3. Märgendajate rakendamine
3.1.1.11.	KRATT peab logima kõik pöördumised	8.4. Logimine ja lõppväljund

3.1.1.12.	KRATT peab võimaldama seadistada, kui mitu märksõna väljaande kohta väljastatakse, millise täpsusega märksõna määratakse ning milliseid märksõnu tuleb tüüpjuhtudel vältida ja mida tuleb kindlasti kasutada	4.2. Optimaalsed vaikeväärtused; 7.3. Märgendajate rakendamine
3.1.1.13.	KRATT peab saatma loodud märksõnad e-kataloogi ESTER bibliokirjetesse (kirjete uuendamine toimub kas läbi Sierra REST API või MARC21 kirjete ülelaadimise teel.)	7.3. Märgendajate rakendamine
3.1.1.14.	KRATT peab omama APIt, mis on liidestatav muude mäluasutuste süsteemidega (nt AIS, MUIS, RIKS)	8. API ja integratsioonide analüüsid
3.1.1.14.1.	Sisendiks URL, millelt päritakse täisteksti	8.2. Toetatud failiformaadid ja sisendi töötlemine
3.1.1.14.2.	Väljundiks märksõnad JSON vormingus	8.4. Logimine ja lõppväljund
3.1.1.15.	KRATT peab teadmusbaasi loomiseks ja/või õpetamiseks omama avalikku veebipõhist liidest	7.2. Märgendajate haldamine; 7.3. Märgendajate rakendamine
3.1.1.16.	KRATT peab tuvastama kasutaja TARA teenuse kaudu (https://e-gov.github.io/TARA-Doku/)	8.1. Kasutajate autentimine
3.1.1.17.	KRATT peab võimaldada eristada vähemalt kahte rolli: RR töötaja ja treenija (*crowdsourcing*'u teostaja)	8.1. Kasutajate autentimine
3.1.1.18.	KRATT peab võimaldama RR töötajal töö tulemust valideerida (rakendus pakub välja märksõnad, töötaja kinnitab tulemuse)	6.3. Märgendajate rakendamine
3.1.1.19.	KRATT peab pakkuma märksõnu vastavalt väljatöötatud metoodikale, kasutaja valib välja sobivaimad märksõnad - toimub rakenduse õpetamine;	8.3. Märgendite pärimine ja tagasiside
3.1.1.20.	KRATT peab logima kõik märksõnadega seotud tegevused (millistele teostele millised märksõnad välja pakuti ja millised omistati)	8.4. Logimine ja lõppväljund

6.2. Mittefunktsionaalsed nõuded

Järgnevalt on toodud süsteemile esitatud mittefunktsionaalsed nõuded koos neile vastavate peatükkide viidetega.

	Nõue	Viide
3.1.2.1.	Soovituslik on järgida RMIT üldiseid mittefunktsionaalseid ja tehnilisi nõudeid, kõrvalekalded on lubatud keeletehnoloogiliste vahendite kasutamisel.	Nõue kinnitati 21.11.2019 koosoleku protokolliga.
3.1.2.2.	Pakkuja peab esitama nõuded serveri spetsifikatsioonile prognoosiga viieks (5) aastaks.	7.4 Nõudeid kasutatavale riist -ja tarkvarale

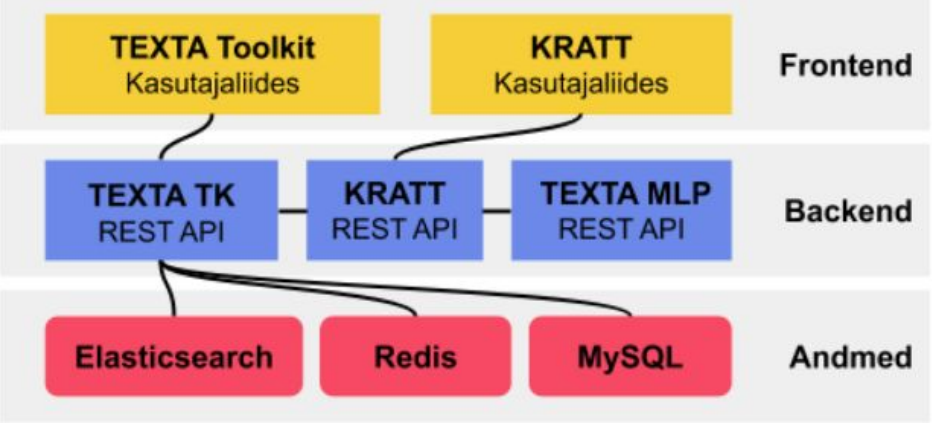
	Nõuded tuleb esitada detailanalüüsi etapis või selle lõpetamisel.	
3.1.2.3.	Loodav lahendus peab olema lihtsalt skaleeritav ega tohi nõuda arhitektuuri muutmist.	7.6 Rakenduse skaleeritavus
3.1.2.4.	Tulevase ja olemasolevate infosüsteemide platvormid (rakendusserver, andmebaas, kolmanda osapoole komponendid) ja topoloogia peab olema enne reaalse arenduse algust Hankijaga kooskõlastatud.	7.4. Nõudeid kasutatavale riist- ja tarkvarale
3.1.2.5.	Kõik kasutajaliidese disainiotsused peavad olema kooskõlastatud Hankijaga.	7.3. Märgendajate rakendamine

7. Disain

Käesoleva peatüki eesmärk on anda ülevaade prototüübi ülesehitusest, selles kasutatavate statistiliste märgendajate treenimise ja rakendamise põhimõtetest, prototüübi tehnilistest nõuetest, rakenduse skaleeritavusest ning Kratt API ja TTK API toimimisest ja omavahelisest integratsioonist.

7.1. Süsteemi arhitektuuri kirjeldus

Loodav tarkvaralahendus on modulaarne ning koosneb mitmest eraldiseisvast teenusest ja abiteenusest, mis on kujutatud joonisel 35.



Joonis 35. *Loodava lahenduse frontend, backend ja andmekihid ning nende omavahelise interaktsiooni skeem.*

Märgendajate treenimine, valideerimine ja ületreenimine toimub läbi TEXTA Toolkit-i graafilise kasutajaliidese[5], mis on programmeeritud Angularis ning pakendatud eraldiseisva Dockeri konteinerina. Märgendajate treenimise eest *backend*'is vastutab TEXTA Toolkit-i RESTful lahendus[6], mille kaudu on võimalik käivitada märgendajate treenimiseks vajalikke asünkroonseid töövoogusid. TEXTA Toolkit-i *backend* on vastutav ka märgendajate serveerimise eest Krati REST *backend*'ile üle API, mis omakorda omab API-t ning teenindab päringuid teavikute märgendamisel ja seal toimub kasutajate autentimine läbi TARA teenuse. Krati REST *backend* liidestub ühtlasi TEXTA MLP (*Multilingual processor*) teenuse külge, mis on ette nähtud tekstide lemmatiseerimiseks ja sõnaliikide määramiseks. Krati kasutajaliidese, mis on programmeeritud Angularis, kaudu on võimalik lõppkasutajal sisendtekste üles laadida, millele leitakse *backend*'i päringu peale märgendid, mis omakorda kasutajale kuvatakse.

Rakenduse andmekihis on kolm abiteenust/moodulit:

- Elasticsearch märgendajate treenimiseks ja vajalike andmete hoiustamiseks;
- Redis, asünkroonsete töövoogude sõnumivahendaja (*message broker*);

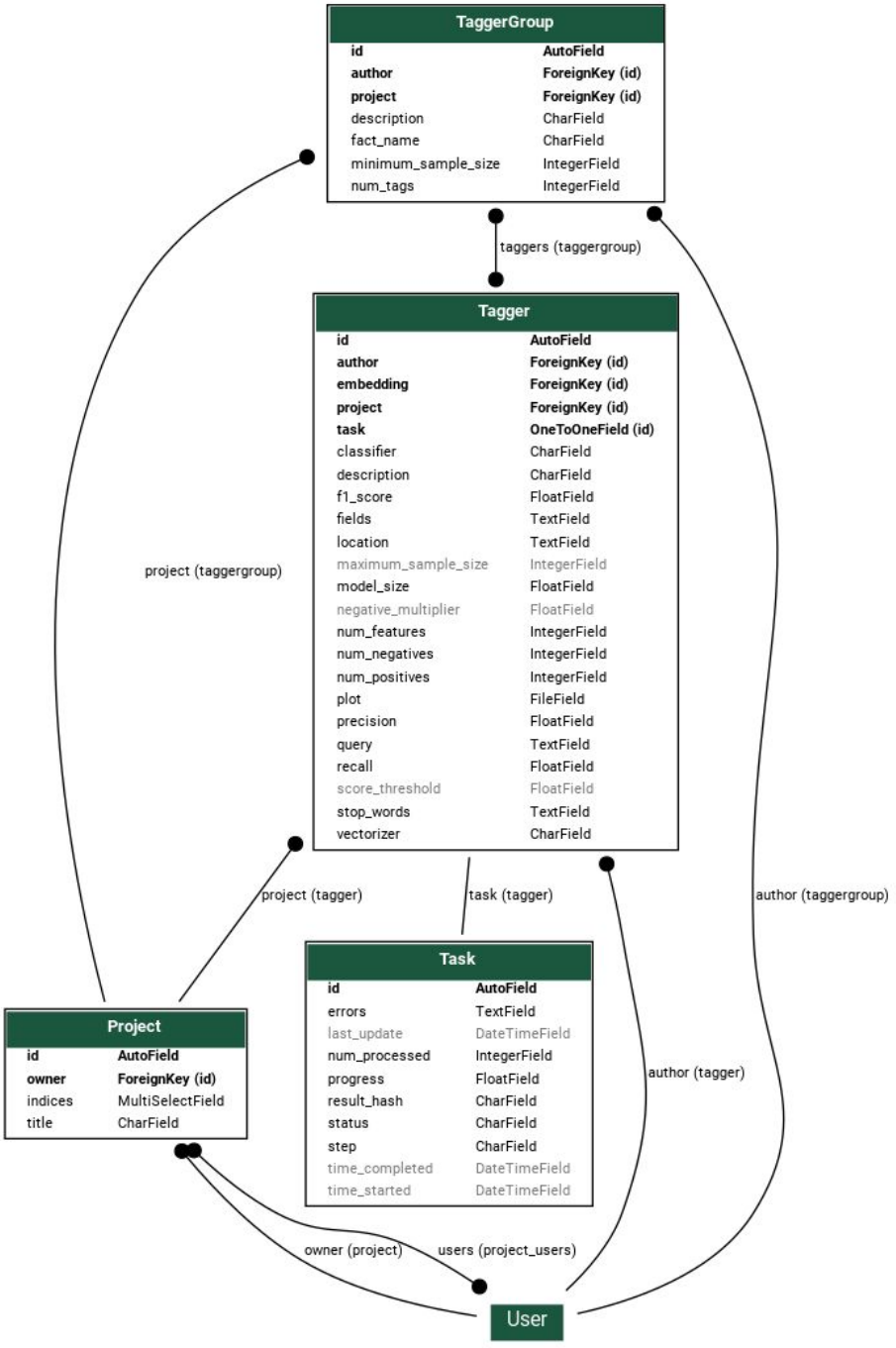
5 https://git.texta.ee/texta/texta-rest-front

6 https://git.texta.ee/texta/texta-rest

- MySQL TEXTA Toolkit-i andmemudeli hoiustamiseks.

7.1.1. TEXTA Toolkit-i mooduli kirjeldus

TEXTA Toolkit on vabavaraline rakendus, mis võimaldab API kaudu treenida erinevaid masinõppemudeleid, sh projekti raames kasutatavaid logistilisel regressioonil ja tugivektormasinal põhinevaid tekstimärgendajaid. TEXTA Toolkit on programmeeritud Django REST raamistikus ning on varustatud andmemudeliga (vt. joonis 36), mida hoitakse MySQL andmebaasis (MySQL on vaikevalik, ent on asendatav ka Postgres andmebaasiga).



Joonis 36. *TEXTA Toolkit-i andmemudeli alamosa, mida rakendatakse tarkvaralahenduse loomisel.*

Andmemudelis (vt. joonis 36) on salvestatud:

- kasutajad, nende õigused ja seotus märgendusprojektidega;
- märgendajate andmed ja statistika;
- märgendajagruppide andmed ja statistika;
- märgendajate treenimise progress ja tekkinud veateated.

Järgnevalt on kirjeldatud TEXTA Toolkit-i komponendid, mida käesoleva probleemi lahendamisel kasutatakse. Kuivõrd TEXTA Toolkit-i andmemudel on ingliskeelne, on esitatud komponendi nimetus koos selle kaudse eestikeelse vaste ning kirjeldusega.

Project. Projektid on objektid, mille eesmärk on koondada kokku samade lähteandmete alusel loodud ressursid (sh märgendusmudelid). Iga projektiga on võimalik siduda 1 või enam kasutajat, kellel on ligipääs projekti algandmetele ning loodud ressurssidele.

Tagger ehk märgendaja on objekt, milles määratakse märgendaja treenimise aluseks olevad parameetrid (kirjeldatud märgendajate haldamist puudutavas peatükis). Pärast treeningprotsessi lõppu lisatakse sinna mudeli kvaliteedinäitajaid (saagis, täpsus, F1-skoor).

Tagger Group ehk märgendajate grupp on kogum märgendusmudeleid, mis on treenitud sama algandmestiku ja parameetristiku alusel. Objekti luues on võimalik treenida ja rakendada tuhandeid märgendusmudeleid.

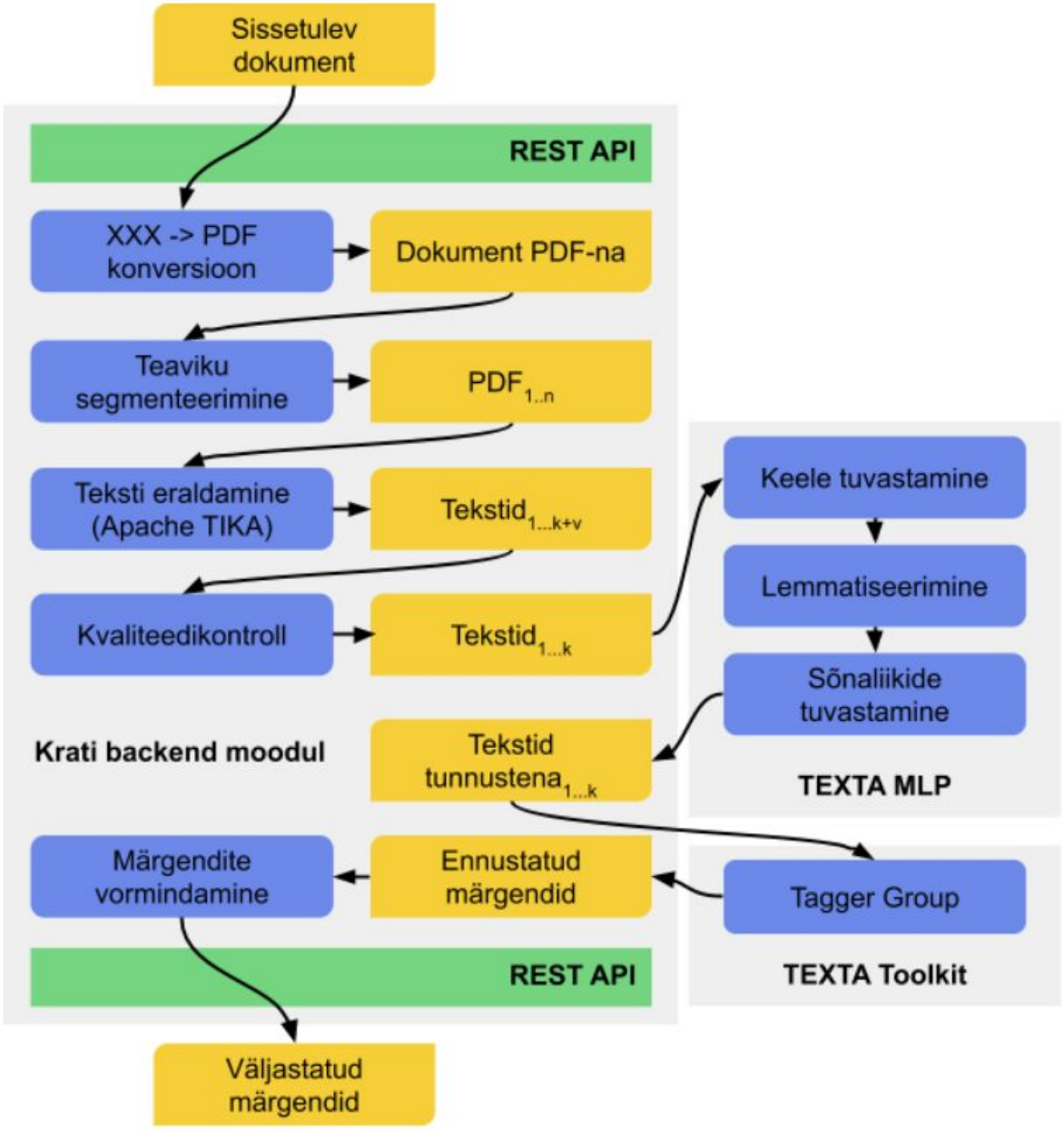
Task ehk ülesanne on objekt asünkroonsete treeningülesannete haldamiseks. Objekti kasutatakse treeningprotsessi staatuse ja veateadete monitoorimisel. Ülesandeobjekt on seotud iga treenitava ressursiga (sh märgendajaga).

7.1.2. Krati mooduli kirjeldus

Krati *backend* moodul on programmeeritud kasutades programmeerimiskeelt Python ning Django REST raamistikku. Mooduli eesmärk on eeltöödelda ja märgendada sissetulev teavik, kasutades selleks käesolevas dokumendis kirjeldatud meetodeid. Mooduli aluseks olev töövoog on kirjeldatud joonisel 37.

Moodul on kasutatav läbi REST API (mille spetsifikatsioon on kirjeldatud peatükis API ja integratsioonide analüüsid), kuhu saadetakse fail kasutaja arvutist või veebiressursi puhul aadress faili asukohaga. Sissetulev dokument konverteeritakse PDF formaati, et võimaldada dokumendist lehekülgede eraldamist ehk segmenteerimist (.txt formaadis sisendfailide puhul jääb see etapp koos järgmisega vahele). PDF formaadis dokumendi segmenteerimise tulemusel tekib *n* PDF dokumenti, millest igaüks sisaldab algse dokumendi ühte lehekülge. Eraldatud lehekülgede PDFide seast valitakse *k+v* juhuslikku segmenti, kus *k* on märgendamisele kuuluvate segmentide arv ning *v* varusse kaasatavate segmentide arv. Valitud segmendid suunatakse seejäral Apache TIKA teenusesse, mis eraldab igalt leheküljelt teksti. Eraldatud tekstid suunatakse kvaliteedikontrolli, mille käigus valitakse kontrolli läbinud segmentide hulgast *k* juhuslikku üksust. Valitud tekstid saadetakse TEXTA MLP teenusesse, mis tuvastab tekstide keelsuse, lemmatiseerib nendes olevad sõnad ning

määrab sõnaliigid. Saadud tulemusi kasutatakse tunnustena tekstidele märgendite ennustamisel, mis toimub TEXTA Toolkit-i rakenduses, kasutades selleks Tagger Group funktsionaalsust. Seejärel ennustatud märgendid puhastatakse, viiakse lõppkasutajale sobivasse formaati ja väljastatakse vastusena algsele API päringule.



Joonis 37. *Krati backend moodulis kasutatava töövoo kirjeldus.*

7.1.3. Lisamoodulite ja -funktsionaalsuste kirjeldus

Käesolevas peatükis antakse põgus ülevaade moodulitest ja funktsionaalsustest, mida prototüübi raames ei realiseerita, kuid mis on vajalikud märksõnastamise töövoo täielikuks automatiseerimiseks. Iga kirjeldatud moodul vajab hankijaga kooskõlastatud eraldiseisvat analüüsi.

7.1.3.1. Moodul: Treeningandmete parser

Mooduli eesmärk on andmete ettevalmistamine uute mudelite treenimiseks ja/või olemasolevate mudelite uuendamiseks. Moodul peab võimaldama:

1) Sisse lugeda teaviku täisteksti sisaldava faili;
2) Sisse lugeda teosega seotud märksõnade MARC kirje;
3) Vajadusel teaviku täisteksti faili ühildama märksõnade MARC failiga (kui erinevate teavikute märksõnade info on salvestatud ühte faili);
4) Sisseloetud andmed eelnevalt kooskõlastatud formaadis Elasticsearchi andmebaasi üles laadima.

7.1.3.2. Moodul: e-kataloogi ESTER bibliokirjete automaatne uuendaja

Mooduli eesmärk on Krati tuvastatud märksõnad automaatselt saata e-kataloogi ESTER bibliokirjetesse. Täiendavad nõuded selgitatakse välja koostöös hankijaga.

7.1.3.3. Moodul: EMS-i uuenduste kontrollija

Mooduli eesmärk on suhelda EMS-iga ning koguda infot märksõnade uuenduste kohta. Juhul kui märksõna x kasutamine asendatakse märksõna y kasutamisega, peaks see muutus kajastuma ka märksõna x mudelis (märksõna nimi asenduma y-ga). Lisaks tuleks automaatselt asendada ka kõik märksõnaga x märksõnastatud teosed märksõnaga y. Täiendavad nõuded selgitatakse välja koostööd hankijaga.

7.1.3.4. Täiendus: aktiivõppe võimaldamine

Lisaarendus Krati moodulile, mis peab võimaldama mudelite ületreenimist ning uute märksõnade lisamist.

7.1.3.5. Täiendus: nõuded märksõnade formaadile

Prototüüpi on kaasatud üksnes märksõnade nimetus, tulevikus peaks mudelite treenimisel ja rakendamisel arvestama aga kõigi märksõna alamväljade sisuga ning korrektsete indikaatoritega kaasamisega. Märksõnade täpsustused ning neis sisalduvad kirjavahemärgid peavad muutumatul kujul säilima.

7.2. Märgendajate haldamine

Märgendajate treenimine ja kvaliteedi valideerimine toimub TEXTA Toolkit-i graafilise kasutajaliidese kaudu. Järgnevalt on kirjeldatud mudeli treenimise protsess koos kuvatõmmistega.

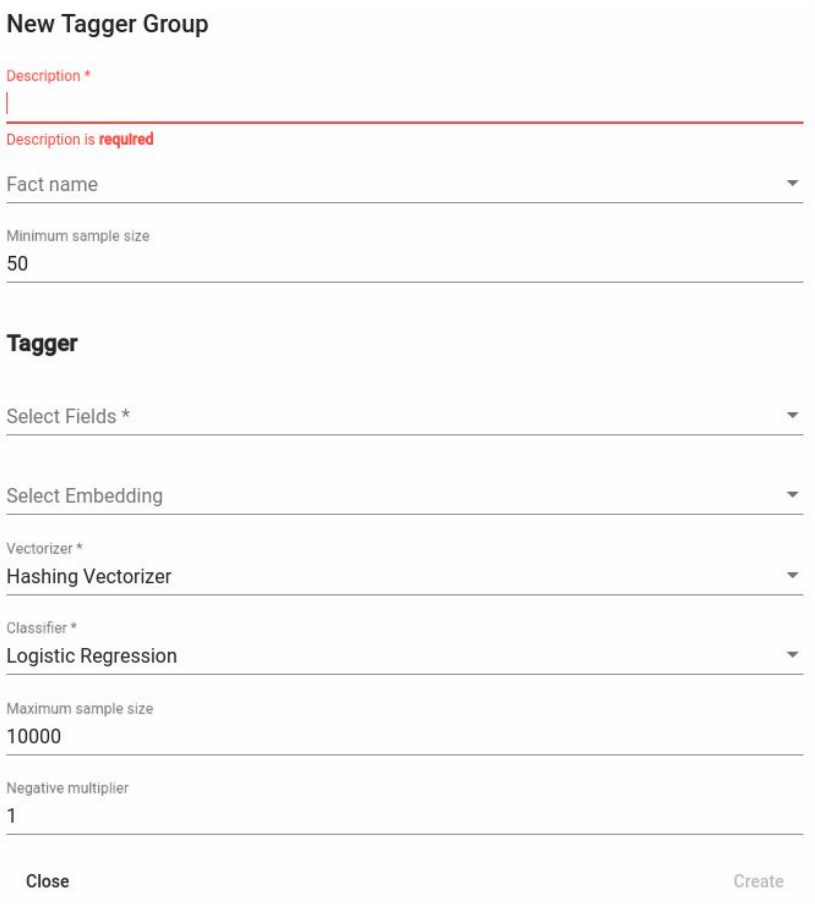
Kuivõrd tegemist on mitmemärgendilise probleemiga (*multi-label classification*) ehk tekstile tuleb anda rohkem kui üks märgend (ja sellest lähtuvalt treenida ka rohkem kui üks mudel), kasutatakse TEXTA Toolkit-i Tagger Group funktsionaalsust, mis võimaldab treenida samade parameetrite alusel suure hulga märgendajaid (TEXTA Toolkit-is Tagger), mida hiljem tekstile rakendada.

Tagger Group-i treenimisel on võimalik seadistada järgnevad parameetrid (tabel 31 ja joonis 38):

Välja nimetus	**Kirjeldus**	**Kohustuslik**	**Vaikeväärtus**
Description	Märgendajate gruppi kirjeldav tekst	jah	puudub
Fact name	Elasticsearchi dokumentides sisalduv märgendinimi, mille alusel leitakse iga märgendaja treenimiseks vajalikud sisendandmed. Võimalikud valikud esitatakse soovitusena.	jah	puudub
Minimum sample size	Minimaalne märgendi esinemise arv. Kui märgend esineb vähem kui x korda, siis märgendajat ei treenita.	ei	50
Fields	Dokumendiväljad, mille alusel märgendajad treenitakse. Võimalikud valikud esitatakse rippmenüüs.	jah	puudub
Embedding	Toolkitis treenitud keelemudel, mida kasutatakse dokumentide sõnavara grupeerimiseks ja dimensionaalsuse vähendamiseks. Antud projektis märgendajate treenimisel keelemudeleid ei kasutata.	ei	puudub
Vectorizer	Tekstide vektoriseerimise metoodika. Võimalikud valikud: Hashing	ei	Hashing Vectorizer

	Vectorizer, Count Vectorizer, Tfidf Vectorizer.		
Classifier	Märgendamisel kasutatav klassifitseerimisalgoritm. Võimalikud väärtused: Logistic Regression, LinearSVC.	ei	Logistic Regression
Maximum sample size	Maksimaalne näidete hulk märgendamise treenimisel. Kasutatakse mäluplahvatuse ärahoidmiseks.	ei	10000
Negative multiplier	Negatiivsete näidete kordaja. Kui tavaolukorras võetakse treeningandmeteks x dokumenti, mis sisaldavad antud märgendit ning samas mahus dokumente, mis ei sisalda antud märgendit (negatiivsed näited), siis antud parameetrit muutes on võimalik negatiivsete näidete määra suurendada või vähendada.	ei	1

Tabel 31. *Märgendajate treenimisel kasutatavad parameetrid ja nende vaikeväärtused.*



Joonis 38. *Märgendajate grupi (Tagger Group) treenimisel määratavad parameetrid graafilises kasutajaliideses vaadatuna.*

Treenitud ja treenimisel olevad märgendajate grupid on nähtavad koos progressi puudutava informatsiooniga eraldi vaates (joonis 39). Kasutajale on kuvatud treenimise tulemusel tekkinud märgendajate arv, keskmine F1-skoor ning saagis ja täpsus. Märgendajate grupiga seotud märgendusmudeleid on võimalik taastreenida valides “Edit” -> “Models Retrain” vastava grupi realt kasutajaliideses.

Id	Author	Description	Fact Name	Min Sample Size	Total Tags	Avg. f1_score	Avg. precision	Avg. recall	Progress	Edit
26	admin	NLIB ALL KW postags + lemmas	ALL_KW	50	7915	0.91	0.94	0.64	7913 / 7915	⋮

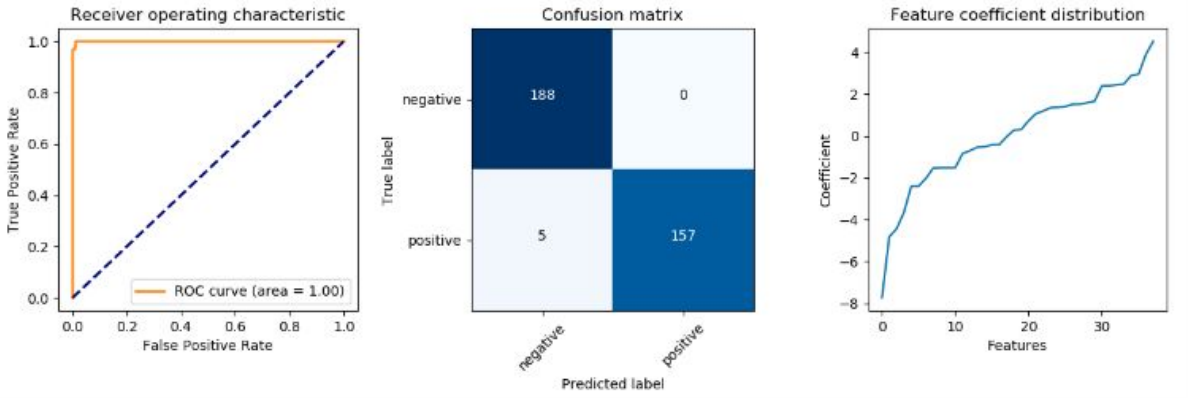
Joonis 39. *Kuvatõmmis treenitud märgendajate grupist (Tagger Group).*

Treenitud märgendajate gruppi kuuluvaid märgendajaid on võimalik hallata eraldi vaates, kus on kasutajale kuvatud märgendi nimi, treenimise aluseks olnud andmeväljad, treeningprotsessi algus ja lõpp ning tekkinud märgendaja kvaliteedinäitajad (joonis 40).

Id	Author	Description	Fields	Time started	Time completed	F1 score	Precision	Recall	Task	Edit
88576	admin	655 lepingud	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:50:30 PM	Nov 25, 5:32:43 PM	0.99	1.00	0.97	completed	⋮
89344	admin	650 saasteseire	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:50:57 PM	Nov 25, 6:41:42 PM	0.93	1.00	0.85	completed	⋮
89856	admin	650 surm	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:51:14 PM	Nov 25, 7:20:22 PM	0.88	1.00	0.75	completed	⋮
90368	admin	650 prohvetiraamatud (Vana Testament)	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:51:31 PM	Nov 25, 7:57:41 PM	0.98	1.00	0.93	completed	⋮
90624	admin	650 pillimeistrid	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:51:39 PM	Nov 25, 8:17:08 PM	0.94	1.00	0.80	completed	⋮
90880	admin	650 näitusekataloogid	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:51:47 PM	Nov 25, 8:36:00 PM	0.97	1.00	0.92	completed	⋮
91136	admin	650 spionaaž	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:51:56 PM	Nov 25, 8:54:22 PM	0.86	1.00	0.62	completed	⋮
91648	admin	650 konfliktid	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:52:13 PM	Nov 25, 9:29:30 PM	0.86	1.00	0.57	completed	⋮
91904	admin	650 mõisted	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:52:21 PM	Nov 25, 9:47:38 PM	0.95	1.00	0.84	completed	⋮
92160	admin	611 Terve loom ja tervislik toit 2014, konverents	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:52:30 PM	Nov 25, 10:05:33 PM	0.78	1.00	0.30	completed	⋮
92416	admin	650 fosforiit	text_mlp.lemmas,text_mlp.postags	Nov 25, 12:52:38 PM	Nov 25, 10:23:32 PM	1.00	1.00	1.00	completed	⋮

Joonis 40. *Kuvatõmmis treenitud märgendajate nimekirjast järjestatuna täpsuse (precision) alusel.*

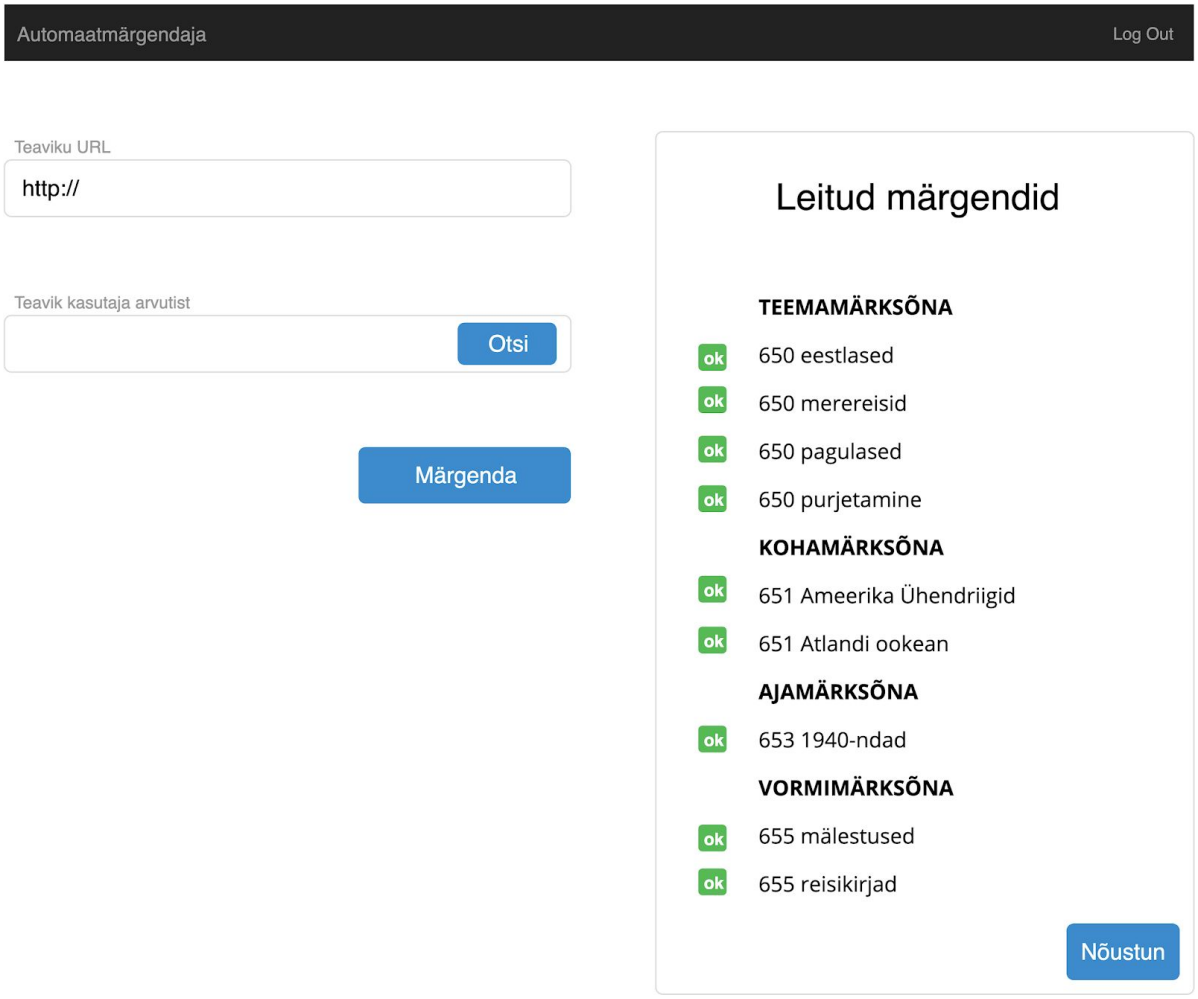
Märgendajale klikkides on kasutajal võimalik näha ka automaatvalideerimise aluseks olevat informatsiooni, kus on kirjeldatud tunnuste jaotumine, eksimismaatriks (*confusion matrix*) ning ROC-kõver (joonis 41).



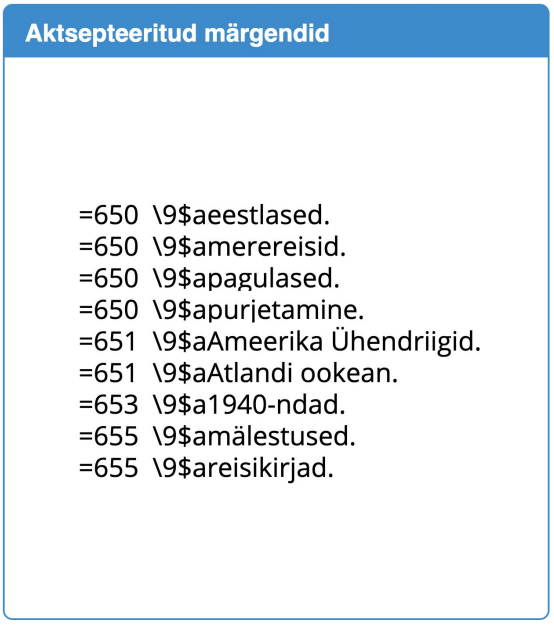
Joonis 41. *Märgendajate valideerimiseks kasutatavad mõõdikud.*

7.3. Märgendajate rakendamine

Lõppkasutajal on võimalik treenitud märgendajaid rakendada kasutades graafilist kasutajaliidest või API-t (spetsifikatsioon on kirjeldatud peatükis 8), mida on omakorda võimalik liidestada teiste süsteemidega (REST API kaudu e-kataloogi ESTER bibliokirje muutmiseks). Graafilise kasutajaliidese kaudu saab kasutaja üles laadida teose faili enda arvutist või esitada teose faili aadressi veebis (vt. joonis 42). “Märgenda” nupule vajutades tehakse päring Krati *backend* moodulile, mis tagastab ennustatud märgendid, mis omakorda kuvatakse lõppkasutajale “Leitud märgendid” sektsioonis. Kasutajal on võimalik valideerida sobilikud märgendid klikkides “ok” nupule. Vajutades nupule “Nõustun”, väljastatakse valitud märgendid osalise MARC kirjena, mis sisaldab valitud märgendeid (vt. joonis 43). Kasutajale kuvatud osaline MARC kirje on kopeeritav ning liidetav olemasolevale MARC kirjele.



Joonis 42. *Lõppkasutajale mõeldud graafilise kasutajaliidese imitatsioon.*



Joonis 43. *Lõppkasutajale kuvatav osalise MARC kirje imitatsioon aktsepteeritud märgenditega.*

Antud punktis kirjeldatud lahendus kinnitati 13.11.2019 toimunud koosoleku protokolliga.

7.4. Nõudeid kasutatavale riist- ja tarkvarale

Loodav lahendus on pakendatud kasutades Dockerit, mis tagab rakenduse kasutatavuse nii Windowsi kui ka Linuxi operatsioonisüsteemiga serverites. Rakendust serveerivas masinas peab olema installeeritud Docker ja docker-compose. Järgnevad riistvaralised nõuded serverile on prognoositud kehtima viieks aastaks.

Rakenduse paigaldus ja edukas kasutamine eeldab serverit või virtuaalmasinat, millel on minimaalselt:

- 24 tuuma,
- 32GB vahemälu,
- 1TB SSD.

Rakendus on lõppkasutajale kasutatav läbi kaasaegse veebibrauseri, nt Chrome, Firefox, Safari jne (Internet Explorer ei ole toetatud!).

7.5. Kasutatavad tarkvarad ja tarkvara moodulid

Kasutatavad tarkvarad ja tarkvara moodulid kinnitatakse eraldi koosoleku protokollina.

7.6. Rakenduse skaleeritavus

Rakenduse skaleerimiseks on kaks peamist meetodit. Esimese meetodi puhul on võimalik käivitada rohkem Dockeri konteinereid ning sissetulev võrgukoormus jaotada kõikide töötavate konteinerite vahel läbi pöördproksi. Teise meetodi puhul on võimalik luua rohkem Celery protsesse ja ühendada neid rakenduse sees oleva Redisega. Viimasel juhul loevad kõik Redise külge haagitud Celery protsessid Redise sees olevad sõnumid, täidavad seal olevaid käske ja salvestavad lõpptulemuse tagasi Redisesse. Kumbki meetod ei vaja arhitektuurseid muutuseid.

8. API ja integratsioonide analüüsid

Käesoleva peatüki eesmärgiks on anda üleüldine kirjeldus Kratt API toimimisest, toetatud sisenditest ja väljunditest ning integratsioonist eraldiseisvate komponentidega nagu TARA ja TTK API.

Ülevaatlikult on Kratt API ülesanded järgmised:

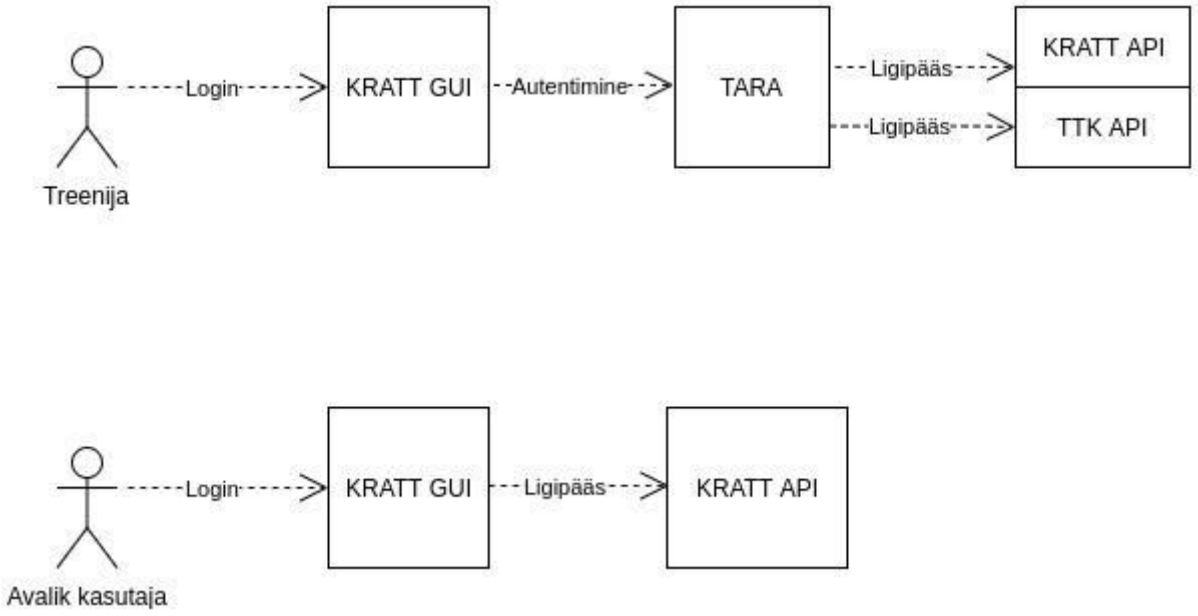
- kasutajate autentimine
- sisendi töötlemine
- märgendite pärimine TTK-st
- märgendite väljastamine JSON formaadis
- tagasiside saatmine TTK-le
- logimine

Järgnevates alapeatükkides on nimetatud funktsionaalsused täpsemalt lahti selgitatud.

8.1. Kasutajate autentimine

Kratt API kasutajad jaotuvad kahte rolli: treenija ja avalik kasutaja. Treenijate isikute tuvastamine toimub läbi riigi autentimisteenuse TARA. Avalikul kasutajal on vaba ligipääs Kratt API-le ning Krati graafilisele kasutajaliidesele. Treenija pääseb ligi Kratt API-le, Krati graafilisele kasutajaliidesele ning lisaks on tal ka administratiivne ligipääs Kratt API tarbeks treenitavatele mudelitele. Mudeleid treenitakse ja hallatakse läbi TEXTA toolkit API (edaspidi TTK API), mille kasutamine on lähemalt selgitatud alapeatükis TEXTA Toolkit-i mooduli kirjeldus.

Kui treenija autendib ennast Kratt API-sse sisse logides esmakordselt, siis tekitatakse talle TTK API administratiivsete õigustega kasutaja, mis tagab kõik vajalikud õigused mudelite treenimiseks ja haldamiseks. Avalik kasutaja ei pea ennast Kratt API kasutamiseks läbi TARA autentima ning seega talle TTK API-sse kasutajat ei tekitata (vt. joonis 44). Küll aga tekitatakse nii tavakasutajale kui ka treenijale esmakordse autentimise korral Kratt API kasutaja, mis tagab ligipääsu Kratt API-le ning Krati graafilisele kasutajaliidesele.



Joonis 44. *Treenija ja avaliku kasutaja autentimise töövoog.*

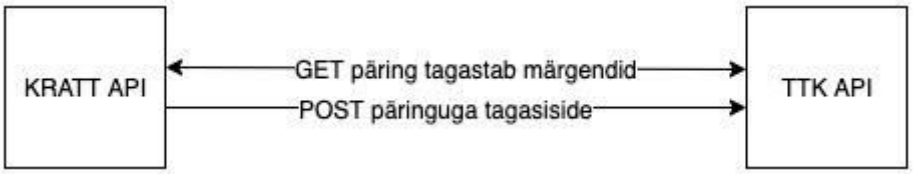
8.2. Toetatud failiformaadid ja sisendi töötlemine

Kui kasutaja on autenditud, siis on võimalik saata Kratt API-ile sisend kas .epub, .pdf, .html või .txt failide näol. Krati kasutajaliideses saab faile üles laadida või URLi kaudu sisestada, misjärel Kratt API töötleb sisestatud faili ning edastab selle TTK API-le, kus toimub failis sisalduva täisteksti automaatmärgendamine (vt. joonis 45). Sisendi töötlemise protsess ja Kratt API edasine käitumine on täpsemalt kirjeldatud alapeatükis Krati mooduli kirjeldus.

Joonis 45. *Failisisendi saatmine Kratt API-sse koos parameetritega.*

8.3. Märgendite pärimine ja tagasiside

Nii avalikul kasutajal kui treenijal on võimalik läbi Kratt API saata päringuid TTK-API-le, mis tagastab leitud märgendite nimekirja. Ainult treenijal on võimalik saata läbi Kratt API tagasisidet TTK API-le märgendite õigsuse kohta (vt. joonis 46). Tagasiside andmise protsess on kujutatud joonisel 42, alapeatükis 7.3. Märgendajate rakendamine.



Joonis 46. *Märgendajate pärimine ja tagasiside saatmine.*

8.4. Logimine ja lõppväljund

Kratt API logib JSON formaadis järgneva:

- sissetulevad veebipäringud ning talletab sealjuures päringuga seonduva kuupäeva, kellaaja, IP-aadressi, otspunkti (*endpoint*), päringu sisu ja logitaseme (*INFO, DEBUG, WARNING, CRITICAL, ERROR*)
- Kratt API koodisiseste veateatede kuupäeva, kellaaja, veateate sisu ja logitaseme
- välja pakutud ja välja valitud märksõnade nimekirjad koos vastava MARC ID, loomise kuupäeva, kellaajaga ja logitasemega.

Logid salvestatakse Kratt API koodihoidlas logide talletamiseks mõeldud kataloogi vastavalt logitüübile, kas veebipäringute logide faili või serverisiseste logide faili. Logide roteerimine lähtub logifailide mahupiirangust, mis on konfigureeritav Kratt API koodihoidlas leitavas üldsättete failis (*settings.py*). Kui logifailide suurus ületab mahupiirangu, siis arhiveeritakse täisfailist koopia ja logifailid tühjendatakse. Nii tagatakse logifailide asjakohasus, inimloetav pikkus ning kõikide logifailide kättesaadavus arhiveerimise kaudu.

Olenevalt päringu sooritamise edukusest tagastab Kratt API kolme tüüpi staatuskoode (xx tähistab kolmekohaliste staatuskoodide esimese ja teise koha variatsioone, mis tähistavad erinevaid vastuse tüüpe API kliendi päringutele) :

- API klient käitus vääralt (kliendi viga, 4xx staatuskood)
- API käitus vääralt (serveri viga, 5xx staatuskood)
- Klient ja API käitusid õigesti (edukas päring, 2xx staatuskood)

Eduka päringu korral tagastab Kratt API, lisaks staatuskoodile, ka märgendid JSON formaadis. Näide tagastatavast JSON väljundist:

```
[{"tag": "merereisid",       "marc_id": "650",  "probability": 0.70},
 {"tag": "Vaikne ookean", "marc_id": "651",  "probability": 0.85},
 {"tag": "reisikirjad",        "marc_id": "655",  "probability": 0.75}]
```

API peale ehitatud graafiline kasutajaliides koondab need märksõnad omakorda kasutaja jaoks mugavalt loetavale kujule (vt. joonis 42 peatükis 7.3. Märgendajate rakendamine), mille seast saab kasutaja selekteerida relevantsed märksõnad. Valitud märksõnadest genereeritakse MARC kirje, mille kasutaja saab hõlpsasti bibliokirjesse kopeerida (vt. joonis 43 peatükis 7.3. Märgendajate rakendamine).

9. Prototüüp

Käesoleva peatüki eesmärk on kirjeldada prototüübi API-t: selle arenduskeskonna loomist ja käivitamist, kasutatud tarkvara, otspunkte koos nende funktsioonidega ning kasutajate tagasisidet koos funktsionaalsete ja sisuliste puudustega.

9.1. Prototüübi käivitamise ning töövalmiduse kontrollimise juhend

9.1.1. Prototüübi arenduskeskkonna loomine ja käivitamine

Prototüübi arenduskeskonna käivitamiseks on nõutud Linux operatsioonisüsteem ja pakihalduse süsteem Conda. Esmalt tuleb avada terminal, liikuda prototüübi kataloogi ning luua Conda virtuaalkeskkond:

```
conda env create -f environment.yaml
```

Seejärel tuleb loodud conda virtuaalkeskkond käivitada ning paigaldada ülejäänud vajalikud tööriistad:

```
conda activate nlib-api
sudo apt-get install libreoffice calibre poppler-utils -y
```

Järgmise sammuna tuleb avada uus terminali aken, aktiveerida varasemalt loodud ja täiendatud conda virtuaalkeskkond, liikuda prototüübi kataloogi ning käivitada Redis server ja Celery:

```
sudo service redis-server start
celery -A nlib_api.taskman worker
```

Seejärel võib eelmises terminali aknas serveri käivitada:

```
python manage.py runserver
```

Rakenduse töövalmiduse kontrollimiseks on võimalik jooksutada automaatteste:

```
python manage.py test
```

9.1.2. Prototüübi käivitamine Dockeris

Prototüübi Dockeris käivitamiseks piisab registrist Dockeri pildi jooksutamisest:

```
docker run -p 80:80 docker.texta.ee/texta/nlib-api:latest
```

Rakendus serveeritakse otspunktis:

```
localhost/api/v1/
```

9.1.3. Prototüübis kasutatavad tarkvarad ja tarkvara moodulid

Kõik kolmanda osapoole tarkvarad ja tarkvara moodulid, mida on prototüübi arenduseks kasutatud ning mida on prototüübi toimimiseks vaja on vabavaralise litsentsiga. Täielik nimekiri kasutatud kasutatud tarkvaradest:

```
_libgcc_mutex=0.1=conda_forge
_openmp_mutex=4.5=0_gnu
amqp=2.5.2=pypi_0
billiard=3.5.0.5=pypi_0
ca-certificates=2019.11.28=hecc5488_0
celery=4.2.2=pypi_0
celery-progress=0.0.9=pypi_0
certifi=2019.11.28=py36_0
cffi=1.13.2=py36h8022711_0
chardet=3.0.4=py36_1003
cld2-cffi=0.1.4=py36he1b5a44_1000
coreapi=2.3.3=pypi_0
coreschema=0.0.4=pypi_0
cryptography=2.8=py36h72c5cf5_1
django=2.1.15=pypi_0
django-cors-headers=3.2.1=pypi_0
django-extensions=2.2.8=pypi_0
djangorestframework=3.11.0=pypi_0
drf-yasg=1.17.0=pypi_0
fontconfig=2.13.1=h86ecdb6_1001
freetype=2.10.0=he983fc9_1
gettext=0.19.8.1=hc5be6a0_1002
ghostscript=9.22=hf484d3e_1001
icu=64.2=he1b5a44_1
idna=2.8=py36_1000
inflection=0.3.1=pypi_0
```

itypes=1.1.0=pypi_0
jansson=2.11=h516909a_1001
jinja2=2.10.3=pypi_0
kombu=4.3.0=pypi_0
langdetect=1.0.7=pypi_0
libffi=3.2.1=he1b5a44_1006
libgcc-ng=9.2.0=h24d8f2e_2
libgomp=9.2.0=h24d8f2e_2
libiconv=1.15=h516909a_1005
libpng=1.6.37=hed695b0_0
libstdcxx-ng=9.2.0=hdf63c60_2
libuuid=2.32.1=h14c3975_1000
libxcb=1.13=h14c3975_1002
libxml2=2.9.10=hee79883_0
markupsafe=1.1.1=pypi_0
meld3=2.0.0=py_0
ncurses=6.1=hf484d3e_1002
openssl=1.1.1d=h516909a_0
packaging=20.1=pypi_0
pcre=8.43=he1b5a44_0
pip=20.0.2=py_2
psutil=5.6.7=py36h516909a_0
pthread-stubs=0.4=h14c3975_1001
pycparser=2.19=py36_1
pydot=1.4.1=pypi_0
pyicu=2.4.2=py36h8412b87_0
pyopenssl=19.1.0=py36_0
pyparsing=2.4.6=pypi_0
pysocks=1.7.1=py36_0
python=3.6.7=h357f687_1006
python-magic=0.4.15=pypi_0
pytz=2019.3=pypi_0
readline=8.0=hf8c457e_0
redis=3.3.11=pypi_0
requests=2.22.0=py36_1
ruamel-yaml=0.16.6=pypi_0
ruamel-yaml-clib=0.2.0=pypi_0
setuptools=45.1.0=py36_0
six=1.14.0=py36_0
sqlite=3.30.1=hcee41ef_0
supervisor=4.1.0=py36_0
termcolor=1.1.0=pypi_0
tika=1.23.1=py_0
tk=8.6.10=hed695b0_0
uritemplate=3.0.1=pypi_0

urllib3=1.25.7=py36_0
uwsgi=2.0.18=py36h79bd928_3
vine=1.3.0=pypi_0
wheel=0.33.6=py36_0
wkhtmltopdf=0.12.4=1
xorg-kbproto=1.0.7=h14c3975_1002
xorg-libx11=1.6.9=h516909a_0
xorg-libxau=1.0.9=h14c3975_0
xorg-libxdmcp=1.1.3=h516909a_0
xorg-libxext=1.3.4=h516909a_0
xorg-libxrender=0.9.10=h516909a_1002
xorg-renderproto=0.11.1=h14c3975_1002
xorg-xextproto=7.3.0=h14c3975_1002
xorg-xproto=7.0.31=h14c3975_1007
xz=5.2.4=h14c3975_1001
yaml=0.1.7=h14c3975_1001
zlib=1.2.11=h516909a_1006

9.2. API otspunktid

Käesolevas alapeatükis anname ülevaate automaatmärksõnastaja prototüübi API otspunktidest.

- /api/v1/ - rakendus on versioneeritud kasutades otspunktide sisest versioonile viitavat nimeruumi. S.t, kõik otspunktid sisaldavad lisaks juurele ka vastava versiooni nimeruumi. Kui on soov juba serveeritavat API-d uuendada, saab luua uue nimeruumi all uuendatud API versiooni ning jätkata sama otspunkti alt vana versiooni serveerimist klientidele, kes seda tarbida soovivad. Ilma versiooni nimeruumita juurotspunktile päringu sooritaja suunab rakendus automaatselt /api/v1/ otspunkti. Nimetatud otspunkt on ühtlasi nimekiri JSON objektidest, mis sisaldavad endas sisestatud faili objekti, leitud märksõnade nimekirja ja keele tuvastuse tulemust. Vastu käesolevat otspunkti käib ka failide POST päring milles sisaldub sisendfaili URL või kettalt üles laetud fail ning soovitud segmentide arv.


```
{
        "id": 1,
        "file_url": "",
        "segment_size": 5,
        "response": [
            {
                "tag": "eesti",
                "marc_id": "650",
                "probability": 1.0
            },
            {
```

```
                "tag": "ilukirjandus",
                "marc_id": "655",
                "probability": 0.8
            },
            {
                "tag": "novellid",
                "marc_id": "655",
                "probability": 0.4
            },
            {
                "tag": "ungari",
                "marc_id": "650",
                "probability": 0.2
            },
            {
                "tag": "antoloogiad",
                "marc_id": "655",
                "probability": 0.2
            },
        ],
        "language_response": [
            {
                "et": 5
            }
        ],
        "task": {
            "id": 12,
            "url": "http://localhost:8000/api/v1/tasks/12",
            "file_model_url": "http://localhost:8000/api/v1/1/",
            "status": "completed",
            "progress": 100.0,
            "step": "Mudelid valmis",
            "errors": {},
            "time_started": "2020-02-19T12:46:28.289194+02:00",
            "last_update": "2020-02-19T12:46:53.055625+02:00",
            "time_completed": "2020-02-19T12:46:53.055653+02:00"
        }
    },
```


Joonis 47. *Näide lõppenud töövoo tulemusel tekitatud JSON objektist. Objektis sisalduvad leitud märksõnad, keeletuvastuse tulemus ning tööülesande objekt.*

- Pärast päringu sooritamist tagastab rakendus JSON objekti, milles sisalduv URL viitab otspunktile `/api/v1/tasks/{task.id}`. Selles otspunktis kajastub iga individuaalse tööülesande progress (*"step"*) lõpptulemus (*"status"*) ning läbikukkunud tööülesande korral ka veateated (*"errors"*). Tuleb olla teadlik, et `/api/v1/` otspunktis leitavate faili mudeli objektide *"id"*-d ei pruugi ühtida nende sisse kätketud tööülesande objektide *"id"*-dega, sest läbi kukkunud tööülesande korral faili mudelit ei looda. Kui tööülesanne on edukalt lõpetatud, siis luuakse uus `/api/v1/` otspunktist kättesaadav faili objekt milles on kätketud keele tuvastuse tulemus, tööülesande objekt, leitud märksõnad ning POST päringus sisalduv asjakohane info (fail või URL, valitud lehekülgede arv).

```
{
    "id": 2,
    "url": "http://localhost:8000/api/v1/tasks/30/",
    "file_model_url": "http://localhost:8000/api/v1/6/",
    "status": "completed",
    "progress": 100.0,
    "step": "Mudelid valmis",
    "errors": {},
    "time_started": "2020-02-19T16:04:53.058215+02:00",
    "last_update": "2020-02-19T16:05:06.709173+02:00",
    "time_completed": "2020-02-19T16:05:06.709183+02:00"
}
```


Joonis 48. *Edukalt lõppenud tööülesanne.*


```
{
    "id": 3,
    "url": "http://localhost:8000/api/v1/tasks/31/",
    "file_model_url": "",
    "status": "failed",
    "progress": 0.0,
    "step": "Töövoog ebaõnnestus",
    "errors": {
        "messages": [
            "Faili formaat ei ole toetatud."
        ],
        "step": "Töövoog käivitatud",
        "stack_trace": [
            "Error at Töövoog käivitatud: [ErrorDetail(string='Faili formaat ei ole
toetatud.', code='invalid')]"
        ]
    },
    "time_started": "2020-02-19T16:31:33.026886+02:00",
    "last_update": "2020-02-19T16:31:45.385806+02:00",
    "time_completed": null
}
```


Joonis 49. *Ebaõnnestunud tööülesanne.*

- /api/v1/convert_to_marc/ - võtab sisendiks valitud märksõnad ja tagastab MARC kirje.
- /api/v1/health/ - tagastab asjakohast infot host arvuti ning vajalike jooksvate teenuste kohta.

```
{
    "mlp": {
        "url": "http://mlp-dev.texta.ee:5000/",
        "alive": true,
        "status": {
            "loaded entities": [
                "/opt/texta-mlp/entity_mapper/data/addresses.json",
                "/opt/texta-mlp/entity_mapper/data/atc.json",
                "/opt/texta-mlp/entity_mapper/data/companies.json",
                "/opt/texta-mlp/entity_mapper/data/substances.json",
                "/opt/texta-mlp/entity_mapper/data/drugs.json"
            ],
            "service": "TEXTA MLP",
            "version": "2.2.1"
        }
    },
    "version": "1.0.0",
    "disk": {
        "free": 327.51826095581055,
        "total": 456.9557571411133,
        "used": 106.15888595581055,
        "unit": "GB"
    },
    "memory": {
        "free": 2.361003875732422,
        "total": 7.7030029296875,
        "used": 4.76885986328125,
        "unit": "GB"
    },
    "cpu": {
        "percent": 20.0
    }
}
```


Joonis 50. */health* vaate tüüpiline tulemus

- api/v1/docs/ - tagastab infot API otspunktide kohta (sisendi ja väljundi formaadid, võimalikud HTTP päringud.

9.3 Kasutajate tagasiside

Käesoleva peatüki eesmärgiks on kirjeldada kasutajate tagasisidet pärast põgusat prototüübi testimist: mis on põhilised funktsionaalsed ja sisulised puudused, millega lisaarenduste tegemisel arvestama peaks.

Iga kasutaja rakendas loodud tööriista 20 faili/URLi peal Google Chrome'i veebilehitsejas. Märksõnastajale edastatavaks lehekülgede arvuks valisid kõik kasutajad teaviku algpikkusest sõltumata 7.

9.3.1. Funktsionaalsed puudused

- Kasutajaliideses võiks kuvada märksõnastamisprotsessi aega.

- Umbes 10% teavikutest avanes alles mitmendal katsel - automaatmärksõnastaja töökindlust tuleks täiustada.

- Kasutajaliideses peaks kuvama terve teaviku keelelise jaotuse (mitte üksnes märksõnastamiseks edastatud lehekülgede oma).

- Märksõnade lävendi liugur on kasutaja jaoks ebaintuitiivne - selle võiks asendada liuguriga, mille eesmärgiks on lihtsalt märksõnade arvu (ükshaaval) suurendamine või vähendamine. NB! Siinjuures tuleb siiski arvestada, et märksõnadel võib olla sama kaal ning seega võivad osad uvatavate märksõnadega võrdväärselt olulised märksõnad kaduma minna.

- Kasutajal võiks olla võimalus ise puuduvaid märksõnu lisada ning rakendada aktiivõpet.

9.3.2. Sisulised puudused

- Juhuslik lehekülgede valik tuleks asendada vaikimisi pooljuhusliku valikuga: soovituslik oleks märksõnastamise protsessi kaasata alati teose esimesed 5 lehekülge; ülejäänud osa võib enamjaolt valida juhuslikult, kuid kasutajal võiks eksisteerida võimalus ka ise lehekülgi valida. NB! See tähendab, et mudelite treenimisel tuleks samuti segmentide valikuga arvestama (raamatute esimesed leheküljed peaksid olema pigem treening- kui testandmestikus).

- Märksõnade vaikelävend võiks olla sõltuv märksõnastatavate lehekülgede arvust - kui märksõnastajale edastatakse vähe lehekülgi, siis peaks lävend paremate tulemuste saavutamiseks automaatselt tavapärasest (0.4) natuke madalam olema.

- Vormimärksõnu pakuti vähe, mis tuleneb suuresti mudeli treeningandmestiku piiratusest. Siiski võiks vormimärksõnade ennustamisel arvesse võtta ka faili tüüpi (nt. epub formaadis teavikud viitavad vormimärksõnale “e-teavik”) ning faili üleslaadimise viisi (URL-ilt laadimine on samuti indikaator vormimärksõnale “e-teavik”).

- Märksõnade mudelid peaksid arvestama EMS-i loogikat kitsamate ja laiemate valdkondade pakkumisel.

9.3.3. Üldhinnang rakenduse kasutatavusele

Kasutajate hinnangul on märksõnastaja ebastabiilne, kuna vaikimisi pakutavate märksõnade arv on liiga varieeruv. Lisaprobleemina toodi välja, et sama teaviku korduval märksõnastamisel pakuti täiesti erinevaid märksõnu. Siinkohal tuleb aga silmas pidada, et märksõnastajale edastatavate lehekülgede arv oli testimise käigus üksnes 7 - kuna märksõnastajale edastatavad leheküljed valitakse juhuslikult, tähendab väike lehekülgede arv, et mahukamate teoste mitmekordes märksõnastamises tekibki madala lävendi juures suurem variatsioon. Mida rohkem lehekülgi märksõnastajale edastatakse, seda väiksemaks muutub ka tuvastatud märksõnade variatsioon teaviku mitmekordsel märksõnastamisel. Kui märksõnastajale edastada kõik teaviku leheküljed, on tulemus teaviku igal märksõnastamisel alati sama.

Kasutajate hinnagul ei tõsta automaatmärksõnastaja prototüüp tööviljakust abivahendina, kuna automaatmärksõnastajaga eelnevalt märksõnastatud teavikute kontrollimisel teistsuguste märksõnade saamine tähendab, et kasutaja peab teavikusse uuesti süvenema, et tulemusi valideerida, mis omakorda tähendab, et protsess kujuneb kokkuvõttes ajakulukamaks kui manuaalne märksõnastamine. Kasutaja hinnangul oleks automaatmärksõnastaja rakendamine mõttekas ainult juhul, kui kasutajapoolset kontrolli ei läbita ning automaatselt tuvastatud märksõnadele pannakse juurde vastav kirje (“märksõnastatud automaatselt”).

Kokkuvõte

Automaatmärksõnastaja Kratt on mõeldud **raamatute** automaatseks märksõnastamiseks.
Prototüübi mudeli treenimiseks kasutati alates aastast 1940 Eestis ilmunud vabalt kättesaadava 7668 raamatu täisteksti, millest 6632 olid eestikeelsed, 751 ingliskeelsed, 195 venekeelsed ning 90 muukeelsed. Raamatud pärinevad valdavalt digitaalarhiivist DIGAR, lisaks võeti kasutusele ka e-kataloogis ESTER kirjeldatud veebiressursid, mis DIGARis arhiveeritud ei ole. Raamatutega seotud märksõnad pärinevad Eesti rahvusbibliograafiast ning põhinevad “Eesti Märksõnastikul” (EMS).
Automaatmärksõnastaja toetatud failiformaadid on .epub, .pdf, .txt ja .html.

Kratt on võimeline ennustama teema-, koha-, aja- ning vormimärksõnu. Isiku-, kollektiivi- ning ajutise kollektiivi või sündmuse märksõnu prototüübi mudelisse ei kaasatud, kuna märksõnad ei põhine EMS-il ning nõuaks täiendavat eeltöötlemist. Ennustatavate märksõnade koguarvuks on 7914, mis valdkondade vahel jagunevad järgnevalt: 6469 teemamärksõna, 325 kohamärksõna, 54 ajamärksõna ning 310 vormimärksõna.

Mudeli treeningandmestiku tehislikuks suurendamiseks kasutati tekstide segmenteerimist: iga raamat jagati lehekülgedeks ning iga leheküljega seoti samad märksõnad, mis olid seotud ka vastava täistekstiga. Segmenteerimise käigus suurenes treeningandmestik 7668 teaviku täistekstilt 659 576 segmendile.

Märksõnade ennustusmudel treeniti dokumendisegmentide lemmade ja sõnaliikide pealt kasutades ennustamiseks logistilist regressiooni. Mudel on treenitud Texta Toolkitis. Mudeli valimiseks ning hindamiseks kasutati automaatseid masinõppemõõdikuid *täpsus, saagis* ja *f1-skoor*. Mudeli hindamiseks kasutati märgendipõhist metoodikat: mõõdeti iga märgendi ennustamise tulemusi ning leiti kõigi märgendite tulemuste aritmeetiline keskmine. Valitud mudeli täpsus on 0.98, saagis 0.86 ning f1-skoor 0.92. Kuigi mudeli skoorid on näiliselt head, tuleb siiski tähele panna, et treeningmaterjali hulk oli küllalt piiratud ning soodustas seega ülesobitamist avalikus kasutuses olevate teoste tüpaažile. See tähendab aga, et teist tüüpi sisuga (nt. ilukirjanduslike) teoste märksõnastamisel ei anna Kratt suure tõenäosusega kuigi häid tulemusi. Parim lahendus tulevikus Krati ennustusvõime parandamiseks on treeningandmestiku laiendamine.

Prototüübi arenduse käigus valmis veebirakendus, mis võimaldab kasutajal üles laadida maksimaalselt 1GB suuruse teaviku ja selle automaatselt märksõnastada. Kasutajal on võimalik automaatselt leitud märksõnu eemaldada ning seejärel kopeerida sobivate märksõnade MARC kirje. Prototüübi raames jäid realiseerimata järgnevad moodulid ja funktsionaalsused: kasutajate autentimine, bibliokirjete automaatne uuendamine, uute treeningandmete parser, EMS-i uuenduste kontrollija ning mudelite aktiivõpe.

Prototüübis realiseeritud töövoog on järgnev: kasutaja valib märksõnastamiseks teaviku ning teaviku märksõnastatavate juhuslikult valitud lehekülgede arvu *n*; dokument konverteeritakse

Krati moodulis PDF-iks (juhul kui see juba vastavas formaadis ei ole); PDF tükeldatakse lehekülgedest; lehekülgede hulgast valitakse *n* juhuslikku; valitud *n* leheküljelt eraldatakse tekst; eraldatud *n* teksti läbivad kvaliteedikontrolli; kvaliteedikontrolli läbinud leheküljed saadetaksmoodulisse, kus 1) tuvastatakse nende keel ja 2) eraldatakse lemmad ja sõnaliigid; eraldatud lemmad ja sõnaliigid saadetakse Texta Toolkiti märgendamismoodulisse, kust tagastatakse vastavalt iga leheküljega seotud märksõnad; märksõnad järjestatakse Krati moodulis esinemissageduse alusel ning väljastatakse APIs koos *n* leheküljelt tuvastatud keelega. Krati graafiline kasutusliides võimaldab omakorda märksõnu esinemissageduse alusel filtreerida ning valitud märksõnade MARC-kirje genereerida.

Kasutajate hinnagul ei tõsta automaatmärksõnastaja prototüüp tööviljakust abivahendina, kuna automaatmärksõnastajaga eelnevalt märksõnastatud teavikute kontrollimisel teistsuguste märksõnade saamine tähendab, et kasutaja peab teavikusse uuesti süvenema, et tulemusi valideerida, mis omakorda tähendab, et protsess kujuneb kokkuvõttes ajakulukamaks kui manuaalne märksõnastamine. Kasutaja hinnangul oleks automaatmärksõnastaja rakendamine mõttekas ainult juhul, kui kasutajapoolset kontrolli ei läbita ning automaatselt tuvastatud märksõnadele pannakse juurde vastav kirje (“märksõnastatud automaatselt”).

Prototüübi eesmärgiks oli muus osas välja selgitada, kas masinõppe metoodikatel tuginev märksõnastamine on tänasega võrreldes asjakohasem ja objektiivsem. Mõlemad on tugevalt sõltuvad treeningandmete hulgast, mis prototüübi arendusel oli küllalt piiratud. Seega ei ole tulemused nii teoreetiliselt kui kasutajate põgusal hinnangul kuigi asjakohased. See tuleneb prototüübi automaatmärksõnastaja nõrgast üldistusvõimest, mis omakorda on tingitud andmete vähesusest. Piisavalt suure ja mitmekesise treeningandmete hulga puhul on võimalik aga treenida paremini üldistuvad, asjakohasemad ja objektiivsemad mudelid.

Kui erinevad kasutajad märksõnastavad erinevatel aegadel sama teose erinevaid väljaandeid, võib tulemuseks saadud märksõnade hulk olla erinev. Prototüübi eesmärk oli selles osas tagada suurem ühtlus - et sama sisuga teos märksõnastataks alati samamoodi. Kuigi märksõnastamise protsess on iseenesest deterministilik - sama sisendiga kaasnevad alati samad märksõnad -, tuleb siiski arvestada võimalikke variatsioone lehekülgede alamhulga valimisest tingitud juhuslikkusest. See tähendab, et kui Kratile ette anda teose A trükkide A1 ja A2 kõik (või samad) segmendid, siis oleks nende kahe teose leitud märksõnade hulk idente. Kuna ajalise efektiivsuse tagamiseks edastatakse märksõnastajale vaikimisi aga juhuslikult valitud 10 segmenti, võib ka märksõnade hulgas seetõttu erinevusi esineda. Ühtlasemaks märksõnastamiseks oleks vajalik märksõnastamise protsessi optimeerimine, et märksõnastajale saaks edastada võimalikult suure osa dokumendist.

Lisad

Lisa 1: Prototüübi kasutusjuhend

Lisana 1 on kaasatud fail “Automaatmärksõnastaja Kratt prototüübi kasutusjuhend.pdf”. Prototüübi kasutusjuhendi eesmärk on anda ülevaade prototüübi veebirakenduse sisendist, väljundist ning selle võimaldatud funktsionaalsustest.

Lisa 2: Prototüübi testlood

Lisana 2 on kaasatud fail “KRATT_testlood.zip”, mis sisaldab kuue kasutaja läbiviidud prototüübi testimise tulemusi .xlsx failides.