TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

Tehnoloogia Instituut

Jürgen Lorenz

Autorsuse tuvastamine sõnavaralise ja märgipõhise tekstianalüüsi meetoditega

Bakalaureusetöö (12 EAP)

Juhendaja: Tõnu Tamme, MSc

Tartu 2015

Autorsuse tuvastamine sõnavaralise ja märgipõhise tekstianalüüsi meetoditega

Interneti populaarsuse kasvu tõttu on verbaalse eneseväljenduse tavad oluliselt muutunud. Järjest sagedamini tekitab probleeme tekstide või dokumentide autori puudumine. Inimesed avaldavad veebis üha julgemini oma mõtteid või arvamusi, kuid jäävad seejuures anonüümseks. Autori nime varjamine põhjustab probleeme ka teadusmaailmas, luuretegevuses, kriminaalõiguses jm. Tekstide autorsuse tuvastamiseks on mitmeid erinevaid võimalusi. Selle bakalaureusetöö käigus üritatakse lahendada autorsuse tõendamise probleeme. Uuritakse, kui suure tõenäosusega on kindel autor analüüsitava dokumendi kirjutanud. Analüüsitakse teksti sõnavaralisi ja märgipõhiseid tunnusjooni ning tuvastatakse konkreetsele autorile omane kirjutamisstiil. Tunnusjoonte määratlemine võimaldab luua nii autori kirjutamisstiilile kui ka tekstile omase mudeli, mida saab võrrelda tundmatu tekstiga ning selle põhjal kindlaks teha, kas uuritav autor on antud teksti kirjutanud. Kahe erineva autori kirjutatud kolme teksti analüüsimisel saadud tulemused on paljutõotavad. Tõenäoliselt on tulevikus võimalik töös katsetatud meetodit rakendada erinevate tekstide autorsuse tuvastamiseks ja kinnitamiseks.

Võtmesõnad: Autorsuse tuvastamine, sõnavaraline analüüs, märgipõhine analüüs, n-grammid, teksti analüüs

Sisukord

Joonised

Tabelid

1 Sissejuhatus

Interneti populaarsuse kasvu tõttu on verbaalse eneseväljenduse tavad oluliselt muutunud. Järjest enam luuakse erinevaid tekste ülemaailmses veebikeskkonnas, kus kirjutiste levitamine on varasemast märksa kiirem ning lihtsam. Interneti kasutamine võimaldab inimestel kommenteerides, blogides või muud teksti luues anonüümseks jääda. Et teadmata autorite tekstide hulk järjest kasvab, on paljude tekstide või dokumentide autorsuse määratlemine oluliseks muutunud. Tekstide autorsuse tuvastamise vajadus on tekkinud paljudel aladel, näiteks luures, kriminaalõiguses, interneti turvalisuse tagamisel jm.

1.1 Probleemi tutvustus

Autorsuse identifitseerimise probleemi on püütud lahendada juba aastakümneid, kuid seni ei ole suudetud toimivat lahendust leida. Eriti oluline on autorsuse kinnitamine selleks, et ei tekiks võimalust omastada võõraid tekste ning esitleda neid kui enda omi. Ainuüksi ühe artikli publitseerimisega võib kaasneda tuntus ning sellega võib seotud olla ka rahalise toetuse määramine. Et ei tekiks vaidlust autorsuse üle või sellest tulenevaid juriidilisi kaasusi, tuleb igale kirjutatud tekstile või dokumendile autori nimi lisada.

Kahjuks pole paljude tekstide juurde autorit märgitud ning tekib vajadus autorsuse tuvastamise järele. Autorsuse tuvastamine on protsess, mille käigus üritatakse kindlaks määrata teksti tõenäolist autorit. Teksti autorsuse tuvastamine tegeleb üldjuhul järgmise kolme probleemiga:

- Profileerimise probleem – on antud dokument, millel puudub kindel autorite kogum. Ülesandeks on koguda autori kohta võimalikult palju demograafilist ja psühholoogilist informatsiooni.

- "Nõel heinakuhjas"-probleem - Tekstil on palju potentsiaalseid autoreid, kelle kirjatööde arv andmebaasides on piiratud, mistõttu on nende autorite profiil äärmiselt pealiskaudne. Ülesanne on teha kindlaks, kes neist autoritest konkreetse teksti kõige tõenäolisemalt kirjutas.

- Tõendamise probleem – Tekstil on üks potentsiaalne autor. Ülesanne on kindlaks teha, kas see autor on kirjutanud vastava teksti või ei. [1].

1.2 Töö eesmärk

Kõige sagedamini kasutab autorsuse tuvastamine tekstianalüüsi meetodeid. Uuritakse mingi määratud hulga autorite tekste ja luuakse iga autori kohta tema kirjutamisstiili mudel. Seejärel võrreldakse neid mudeleid tundmatu tekstiga ning tehakse kindlaks, kas keegi valitud autoritest on uuritava dokumendi kirjutanud. Veelgi tõetruum autorsuse kinnitamise võimalus on analüüsida ühe kindla autori tekstide kogumit ning teha seejärel tekstianalüüsi abil kindlaks, kas vastav isik on küsimuse all oleva dokumendi kirjutanud.

Bakalaureusetöö eesmärk on anda ülevaade autorsuse tuvastamise sõnavaralise ja märgipõhise tekstianalüüsi meetoditest. Tutvutakse nimetatud meetoditega, seejärel realiseeritakse nende meetodite töö programmina ning rakendatakse programmi suvaliselt valitud algandmetel, et selgitada välja valitud meetodite kasutamise praktilisus autorsuse tuvastamisel.

2 Ülevaade probleemist

Teksti analüüsi alused pärinevad 1851. aastast, mil inglise loogik Augustus de Morgan soovitas kirjas sõbrale, et autorsuse küsimused võib lahendada, tehes kindlaks, et kas üks tekst "ei tegele pikemate sõnadega"kui teine. 19. sajandi lõpus sattus de Morgani hüpotees Ameerika füüsiku Thomas Mendenhalli [2] tähelepanu alla. Järgnevalt avaldas ta muljetavaldavate tulemustega akadeemilise töö erineva pikkusega sõnade sagedusest. Ta rakendas oma tehnikat väitele, et kas Francis Bacon on Shakespeare'i tööde tõeline autor. Selleks kasutas Mendenhall Francis Bacon'i poolt kirjutatud töid, kus oli kasutusel umbes 200,000 sõna ning eeldatavalt Shakespeare'i poolt kirjutatud töid, kus sõnade arv ulatusi kuni 400,000-ni.

Järgmise 30 aasta jooksul ei toimund erilisi arenguid autorsuse tuvastamise valdkonnas, vaid inglane G. Udny Yule [3] ja Ameerika keeleteadlane George Zipf [4] töötasid vahelduvate stiili iseärasuste kallal.

1960ndate alguses demonstreerisid Ameerika statistikud Frederic Mosteller [5] ja David Wallace [6] statistilisi meetodeid, et uurida autorsuse tuvastamist. Nende töö oli suureks läbimurdeks teksti analüüsimise ning autorsuse tuvastamine valdkondades.

Autorsuse tuvastamise peamine idee on selles, et mõõtes teatud parameetreid, on võimalik eristada kahe erineva autori kirjutatud tekste [7]. Tänapäeval on interneti tähtsus aina kasvamas ning sellega kaasneb tekstide, artiklite ning muude autorita dokumentide levimine. Võib olla küll mingi aimdus dokumendi autorist, kuid seda ei ole võimalik täielikult kinnitada. Autorsuse tuvastamise probleem on muutunud aktuaalseks ja seda on püütud erinevate meetmetega lahendada. Autorsuse tuvastamist kasutatakse lisaks kirjandusteaduslikule rakendusele (tekstide analüüs) ka mitmetes erinevates valdkondades, näiteks luuretegevuses (sõnumite omistamine, sõnumite sidumine autorsuse alusel), kriminaalõiguses (ahistavate sõnumite saatjate tuvastamine, enesetapukirjade kontrollimine), tsiviilõiguses (autoriõiguste vaidlused) jne. Autorsuse tuvastamiseks on võimalik kasutada tekstianalüüsi meetodeid.

Probleemi tutvustuse peatükis tõin välja kolm põhilist probleemi, millega autorsuse tuvastamine igapäevaselt tegeleb. Järgnevalt käsitlen iga probleemi eraldi.

Profileerimise probleemi [8] puhul puudub kindel autorite kogum, kuhu võiks dokumendi potentsiaalne autor kuuluda. Seetõttu on ülesandeks koguda võimalikult palju

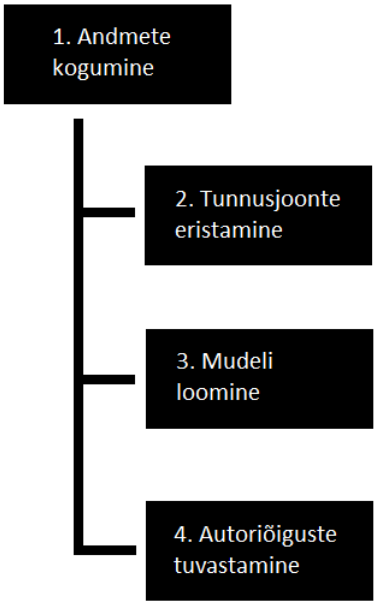
erinevat informatsiooni autori kohta. Profileerimise käigus üritatakse üldjuhul kindlaks teha neli aspekti. Nendeks on autori sugu, vanus, emakeel ning iseloom. Hinnatakse viit iseloomuomadust: avatus, kohusetundlikkus, ekstravertsus (rohkem välismaailmale kui enesele suunatud psüühika), neurootilisus ning vastuvõetavus/meeldivus [9]. Kogutud informatsiooni põhjal on võimalik potentsiaalsete autorite hulka poole võrra vähendada. Näiteks juhul, kui dokument on kirjutatud hispaania keeles ning autoriks on naine, võib välja arvata autorid, kes hispaania keelt ei valda või hispaania keeles kirjutada ei oska ning samuti kõik meessoost autorid. Autori sugu on võimalik tuvastada tekstis esinevate erinevate märksõnade põhjal. Lihtsamaks näiteks võib tuua fraasi "minu poiss-sõber", mis viitab enamikel juhtudel sellele, et autor on naissoost. Soo kindlaks tegemisel uuritakse enam kui ühte märksõna.

"Nõel heinakuhjas"-probleemi puhul on vaatluse all on üks autorita dokument, mida võrreldakse olemasolevate autorite tekstide kogumikuga, mis sisaldab iga autori kohta minimaalsel hulgal informatsiooni. Eesmärgiks on tuvastada, kes vaadeldava dokumendi kõige tõenäolisemalt kirjutas. Selleks tuleb luua iga autori kirjutamisstiili spetsiifiline mudel, mida saab võrrelda küsimärgi all oleva tekstiga. Võrreldes autori mudelit ning tundmatut teksti omavahel, saame tulemuseks tõenäosuse, kas see autor on kirjutanud selle teksti. Leides kõige tõenäolisemad autorid, saab neid omavahel võrrelda ning otsida autorite kohta lisainformatsiooni. Näiteks võimaldab vastavate autorite lisatekstide analüüsimine autorit kirjeldavat mudelit teha veelgi täpsemaks.

Tõendamise probleem põhineb ühe autori kirjatöödel ning ühe autorita dokumendi olemasolul. Eesmärgiks on välja selgitada, kas vaadeldava dokumendi on kirjutanud antud autor. See probleem on vägagi lähedane plagiarismile, mille puhul on vaja tõestada, kas töö autor on teksti ise kirjutanud või on see kusagilt maha kirjutatud või kopeeritud. Näiteks kui on soov ainult kindlaks teha, et vaadeldava teksti kirjutas üks või teine autor, siis tuleks luua selline autorite mudel, mis suudaks neid autoreid teineteisest eristada. Eelnevalt mainitud mudelit tuleks rakendada vaadeldavatel dokumentidel ja hinnata, kumma autori tunnusjooned on lähedasemad tundmatu teksti tunnusjoontele. Teisalt, kui aga on vaja kindlaks teha, kas üks autor kirjutas tundmatu dokumendi, tuleb autori mudelit tekstiga võrrelda ning selle tulemuseks on protsent, mil määral ühtivad autori teiste tekstide ning küsimuse all oleva teksti omadused.

3 Metoodika

Käesolevas bakalaureusetöös käsitletakse autorsuse tõendamist juhul, kui küsimuse all on kaks dokumenti, millest ühe autor on teadmata ning teise autor teada. Ülesandeks on teha kindlaks, kas esimese dokumendi autor on autoriks ka tundmatule tekstile. Kõige lihtsam lähenemine autorsuse tuvastamise probleemi lahendamisele on esitatud joonisel 1



Joonis 1: Autorsuse tuvastamise etapid

1. Esimeseks sammuks on andmete kogumine, et saada võimalikult palju andmeid potentsiaalsete autorite ning tundmatu teksti kohta.

2. Teiseks sammuks on kindlale autorile omaste iseärasuste/tunnusjoonte eraldamine. Tuleb tuua välja kõik tunnusjooned, mis on autorile omased ning teda teistest autoritest eristavad.

3. Kolmandaks sammuks on mudeli genereerimine ehk igale autorile omase mudeli loomine. Saadud mudeli võrdlemisel tundmatu teksti mudeliga on võimalik kindlaks teha, kas tekst on selle autori kirjutatud.

4. Neljandaks sammuks on autorsuse tuvastamine [10].

3.1 Uurimismeetodid

Uurimismeetoditena kasutatakse eritunnusjooni ehk iseärasusi, mida on tekste analüüsides võimalik tuvastada. Tabelis 1 on ära toodud erinevad tunnusjooned ning nende klassid.

Tunnusjooned/iseärasused	
Sõnavaraline	sümbolite põhjal(sõna pikkus, lause pikkus jne.) sõnavara rikkus sõnade kordus sõnade n-gramm vigade arv
Märgipõhine	märgi tüüp (tähed, numbrid jne.) märkide n-gramm (kindla pikkusega) märkide n-gramm (erineva pikkusega) tihendusmeetodid
Süntaktiline ehk lauseline	lause osad teksti osad lause ja fraasi struktuur ümberkirjutamise reeglite sagedus vigade arv
Semantiline ehk tähenduslik	sünonüümid semantilised sõltuvused
Rakenduspõhine	sisuspetsiifiline keelespetsiifiline struktuuriline

Tabel 1: Erinevad tunnusjooned [7]

Sõnavaralised ja märgipõhised tunnusjooned hõlmavad analüüsitavat teksti vastavalt kas sõnade või märkide jadana. Sõnavaralised tunnusjooned on palju keerulisemad kui märgipõhised [7]. Süntaktilised ja semantilised tunnusjooned vajavad sügavamat keelelist analüüsi ning rakenduspõhised tunnusjooned saab määratleda ainult kindlas keeles või tekstidomeenis. Selle bakalaureusetöö raames kasutatakse tekstide analüüsimiseks sõ-

navaralisi ja märgipõhiseid tunnusjooni, sest süntaktiliste ja semantiliste tunnusjoonte realiseerimine vajab sügavamat keelelist analüüsi, mida on raske programmina realiseerida. Seetõttu toon järgnevates peatükkides detailsemalt välja sõnavaralised ja märgipõhised tunnusjooned, mida töös tekstide analüüsimiseks kasutatakse.

3.1.1 Sõnavaralised tunnusjooned

Lihtne ja elementaarne viis, kuidas analüüsitavat teksti vaadelda, on võtta seda sümbolite jadana, kus igale sümbolile vastab üks sõna, number või kirjavahemärk [11]. Esimesena proovis lihtsate meetmetega autorsust tuvastada Thomas Corwin Mendenhalli 1887. aastal, kui ta üritas autorsust uurida lausete ja sõnade pikkuse määramise abil. Sõna pikkuse ja lause pikkuse meetme eeliseks on see, et neid saab rakendada mistahes tekstile mistahes keeles. Sõnavara rikkus on teine tunnusjoon, mille alusel saab teksti analüüsida. Selle põhjal saab kindlaks teha autori keelelise mitmekesisuse. Tüüpiliseks näiteks on tüübi ja sümboli suhe (type-token ratio) V/N, kus V on unikaalsete sõnade arv ning N on kõikide sõnade koguarv. Korrutades saadud arvu sajaga, saadakse teksti sõnavaraline tihedus [1].

Lihtsaks ja väga efektiivseks autorsuse tuvastamise meetodiks on veel kõige sagedamini esinevate sõnade väljavõte ehk sõnade korduse tunnusjoon. Tuues välja tekstis kõige rohkem esinevad sõnad, on võimalik kaht vaadeldavat teksti võrrelda. Igal autoril on nii-öelda lemmiksõnad, mida ta tahes-tahtmata rohkem kasutab.

Viimase tunnusjoonena võib välja tuua n-grammid. N-gramm on n-objekti järjestus, mis on piiratud n-i asendava arvu suurusega etteantud tekstis [12]. Objektideks võivad olla häälikud, silbid, tähed, sõnad või muu, mis on vastavalt rakendusele vajalik. Kui n-grammi suuruseks on üks, siis kutsutakse seda unigrammiks, kahe puhul bigrammiks, kolme puhul trigrammiks jne. N-grammi mudel on tõenäosuslik keelemudel, mille abil prognoositakse selle järjestuse järgmist kõige tõenäolisemat objekti. Kaks peamist n-grammi eelist on algoritmi lihtsus ning ulatuslik võime – lihtsalt n-i suurendades on võimalik talletada rohkem informatsiooni. Sõnavaraliste n-grammide puhul on üheks ob-

[1]

$$.Teksti\_tihedus = \frac{unikaalne}{kogu} \cdot 100 \quad (1)$$

unikaalne = unikaalsete sõnade arv

kogu = kõikide sõnade arv

jektiks üks sõna. Sõnavaraliste n-grammide näited on välja toodud tabelis 2.

3.1.2 Märgipõhised tunnusjooned

Märgipõhiste tunnusjoonte puhul vaadeldakse teksti pelgalt märkide jadana. Niiviisi on võimalik piiritleda mitmeid erinevaid meetmeid, nagu näiteks tähestikuline märkide loendus, numbrite loendus, suurte ja väikeste tähtede loendus, tähtede sagedus, kirjavahemärkide loendus ja palju muudki. Sellist informatsiooni on kerge tekstidest leida mistahes keeles ja see on tõestatud kasuliku meetodina, millega arvuliselt mõõta ehk kvantifitseerida autori kirjastiili.

Järgmiseks märgipõhiseks tunnusjooneks on märgipõhine n-gramm. Nagu eelnevalt sai mainitud on n-gramm vägagi täpne ning märkide tasandil on väga hea välja tuua n-grammide esinemissagedused. Näiteks tuues välja eelmise lause 4-grammi märkide tasandil: |Nagu|, |agu_|, |gu_e|, |u_ee|, |_eel|, |eeln| [2] ja nii edasi. See meetod suudab välja tuua stiilinüansse, sealhulgas ka sõnavaralist informatsiooni (näiteks: |Nagu|, |_sai|), konteksti informatsiooni, kirjavahemärkide ja kapitaliseerimise kasutamist. Veel on sellise meetodi eeliseks välise müra talumine. Juhtudel, kui vaadeldavad tekstid on mürarikkad, sisaldavad kirjavigu või kirjavahemärkide kummalist kasutamist, ei ole märgipõhise n-grammi kasutamine sellest oluliselt mõjutatud. Näiteks sõnad „kaugemalt“ ja „kaugematl“ annaks mitmeid sarnaseid märkide n-gramme, kuigi üks sõna on ebakorrektne. Sõnade n-grammi kasutades aga loetakse need sõnad erinevateks. Selliseid väikseid vigu võib lugeda ka autori tunnusjooneks. Märgipõhiste n-grammide näited on välja toodud tabelis 2.

[2] Sümbolid „|“ ja „_“ tähistavad vastavalt n-grammide piire ja ühekordseid tühikuid.

n-gramm	Näidis	1-gramm	2-gramm	3-gramm
		unigramm	bigramm	trigramm
Märgipõhine	Mina ja sina	M, i, n, a, _, j, a, s, i, n, a	Mi, in, na, a_, _j, ja, a_, _s, si, in, na	Min, ina, na_, a_j, _ja, ja_, a_s, _sin, sin, ina
Sõnavaraline	Tere õhtust kallis sõber	Tere, õhtust, kallis, sõber	Tere õhtust, õhtust kallis, kallis sõber	Tere õhtust kallis, õhtust kallis sõber

Tabel 2: N-grammide näidised

3.1.3 Hii ruut statistik

Hii-ruut-test on kõige klassikalisem kahemõõtmelise sagedustabeli alusel teostav test, mille abil saab otsustada tabelis esitatud tunnuste seotuse statistilise olulisuse üle [13]. Hii-ruut-test võrdleb andmete alusel konstrueeritud sagedustabelit nii-öelda ideaalse, sõltumatuse juhule vastava sagedustabeliga. Lävendi arvutamiseks kasutati veebist leitavat hii-ruudu kalkulaatorit [14]. Joonisel 2 on välja toodud hii-ruut-statistiku arvutamise valem.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} p_i \left(\frac{O_i/N - p_i}{p_i} \right)^2$$

Joonis 2: Hii-ruut-statistiku valem

3.2 Aparatuur

Programm on kirjutatud *Python*'i keeles ning kasutusel on Python 3.2 versioon. Programm on toodud välja osadena ning selgitustega peatükis Lisad.

3.3 Uuritav objekt

Uuritavaks probleemiks on kahe erineva teksti kokkulangevus. Kas kahe teksti tunnusjoonte põhjal on võimalik kindlaks teha, kas need tekstid on kirjutanud sama autor? Millist informatsiooni annavad välja arvutatud parameetrid autorite ja tekstide kohta?

Tekstide analüüsimiseks arvutatakse välja:

- keskmine sõna pikkus
- keskmine lause pikkus
- teksti tihedus
- sõnade arv tekstis
- unikaalsete sõnade arv
- sümbolite arv tekstis
- erinevate 2-grammide arv[3]
- erinevate 3-grammide arv
- Hii-ruut-statistiku 5 põhilise karakteristiku põhjal

3.4 Materjalid

Testitavateks tekstideks kasutati kolme ajalehe Postimees veebiportaalis avaldatud artiklit kahelt autorilt. Ühelt autorilt pärineb üks tekst ning teiselt autorilt kaks teksti. Analüüsitud artiklid on järgmised:

1. Vootele Päi kirjutatud arvamusartikkel "Eesti võib teha endast sõjapõgenike vastuvõtmise kontekstis ärahellitatud oiviku" [15].
2. Vootele Päi arvamusartikkel "Mis on «Harta12» autorite tegelik agenda?" [16].
3. Vootele Päi uudisartikkel "Kremli hingus Prantsuse välispoliitika kohal" [17].

[3]N-grammide puhul võrdleme mõlema teksti 200 enim esinevat n-grammi ning võrdleme nende hulkade kokkulangevust

4. Priit Pulleritsu artikkel ”Vahetame riiki” [18].

5. Priit Pulleritsu artikkel ”Arutu kulutus” [19].

6. Priit Pulleritsu blogi postitus ”Kuidas ma sain Jõgeva rattarallil oma parima koha” [20].

4 Tulemused

Artiklite võrdlemiseks tehti tekstid ühepikkusteks. Lähtudes esialgsetest artiklite pikkustest võeti analüüsitava artikli pikkuseks ligikaudu 3000 sümbolit.

4.1 Artiklite karakteristikute tabelid

Peatükis Lisad välja toodud programmi abil leiti kõigi kolme teksti karakteristikud (k.a need karakteristikud, mida lõplikus analüüsis ei kasutata) ning need on välja toodud tabelites 3, 4, 5, 6, 7 ja 8.

Vootele Päi	Sõjapõgenike vastuvõtmine
keskmine sõna pikkus	6,08
keskmine lause pikkus	23,0
teksti tihedus (%)	70,6
tekstis sõnu	484
unikaalseid sõnu	341
sümbolite arv	2999
erinevaid 2-gramme	430
erinevaid 3-gramme	1495

Tabel 3: Vootele Päi kirjutatud arvamusartikli "Eesti võib teha endast sõjapõgenike vastuvõtmise kontekstis ärahellitatud oiviku" analüüsi tulemused

Vootele Päi	Harta 12
keskmine sõna pikkus	6,08
keskmine lause pikkus	20,5
teksti tihedus (%)	71,34
tekstis sõnu	493
unikaalseid sõnu	351
sümbolite arv	3040
erinevaid 2-gramme	424
erinevaid 3-gramme	1442

Tabel 4: Vootele Päi kirjutatud arvamusartikli "Mis on "Harta12" autorite tegelik agenda?" analüüsi tulemused

Vootele Päi	Kremli hingus
keskmine sõna pikkus	6,31
keskmine lause pikkus	18,48
teksti tihedus (%)	72,72
tekstis sõnu	463
unikaalseid sõnu	336
sümbolite arv	2965
erinevaid 2-gramme	438
erinevaid 3-gramme	1459

Tabel 5: Vootele Päi kirjutatud artikli "Kremli hingus Prantsuse välispoliitika kohal" analüüsi tulemused

Priit Pullerits	Vahetame riiki
keskmine sõna pikkus	5,87
keskmine lause pikkus	13,89
teksti tihedus (%)	67,2
tekstis sõnu	501
unikaalseid sõnu	336
sümbolite arv	2982
erinevaid 2-gramme	476
erinevaid 3-gramme	1521

Tabel 6: Priit Pulleritsu kirjutatud artikli "Vahetame riiki" analüüsi tulemused

Priit Pullerits	Arutu kulutus
keskmine sõna pikkus	5,5
keskmine lause pikkus	16,56
teksti tihedus (%)	68,49
tekstis sõnu	531
unikaalseid sõnu	363
sümbolite arv	3002
erinevaid 2-gramme	439
erinevaid 3-gramme	1469

Tabel 7: Priit Pulleritsu kirjutatud artikli "Arutu kulutus" analüüsi tulemused

Priit Pullerits	Jõgeva rattaralli
keskmine sõna pikkus	5,16
keskmine lause pikkus	11,37
teksti tihedus (%)	66,43
tekstis sõnu	558
unikaalseid sõnu	370
sümbolite arv	2940
erinevaid 2-gramme	424
erinevaid 3-gramme	1420

Tabel 8: Priit Pulleritsu kirjutatud blogi postituse "Kuidas ma sain Jõgeva rattarallil oma parima koha?" analüüsi tulemused

4.2 Artiklite omavaheline võrdlus

Seejärel võrreldi mõlema autori tekste samas valdkonnas. Vootele Päi puhul võrreldi kahte arvamusartiklit ning Priit Pulleritsu kahte uudist. Järgnevalt võrreldi sama autori tekste erinevates valdkondades, Pulleritsu kirjutatud uudist ja blogi postitust ning Vootele Päi kirjutatud arvamusartiklit ning uudist. Lõpuks võrreldi erinevate autorite tekste samas valdkonnas, Pulleritsu kahte uudist Vootele Päi uudisega. Tekste võrreldi omavahel peatükis 4.1 välja toodud tulemuste põhjal.

4.2.1 Vootele Päi kahe arvamusartikli võrdlus

Tabelis 9 on Vootele Päi kahe arvamusartikli võrdluse tulemused.

	Sõjapõgenike vastuvõtmine	Harta 12	Vahe (%)
keskmine sõna pikkus	6,08	6,07	0,12
keskmine lause pikkus	23,0	20,5	12,20
teksti tihedus (%)	70,6	71,34	1,04
erinevaid 2-gramme	430	424	1,42
erinevaid 3-gramme	1495	1442	3,68
tekstis sõnu	484	493	1,82
unikaalseid sõnu	341	351	2,84
Hii-ruut	2,79	2,66	
sümbolite arv	2999	3040	

Tabel 9: Vootele Päi arvamusartiklite 'Eesti võib teha endast sõjapõgenike vastuvõtmise kontekstis ärahellitatud oiviku" ja "Mis on «Harta12» autorite tegelik agenda?"

4.2.2 Priit Pulleritsu kahe uudisartikli võrdlus

Kolmandaks võrdluseks on Priit Pulleritsu kahe uudisartikli võrdlus, mille tulemused on nähtavad tabelis 10.

	Vahetame riiki	Arutu kulutus	Vahe (%)
keskmine sõna pikkus	5,86	5,5	6,67
keskmine lause pikkus	13,88	16,56	16,14
teksti tihedus (%)	67,2	68,49	1,88
erinevaid 2-gramme	483	439	10,02
erinevaid 3-gramme	1521	1469	3,53
tekstis sõnu	501	531	3,53
unikaalseid sõnu	336	363	7,43
Hii-ruut	9,13	9,13	
sümbolite arv	2982	3002	

Tabel 10: Priit Pulleritsu artiklite "Vahetame riiki" ja "Arutu kulutus" võrdlus

4.2.3 Vootele Päi uudisartikli ja arvamusartikli võrdlus

Tabelis 11 on Vootele Päi uudisartikli "Kremli hingus Prantsuse välispoliitika kohal" ja arvamusartikli "Mis on «Harta12» autorite tegelik agenda?" artikli võrdluse tulemused.

	Kremli hingus	Harta 12	(%)
keskmine sõna pikkus	6,31	6,08	3,92
keskmine lause pikkus	18,48	20,5	9,80
teksti tihedus (%)	72,72	71,34	1,94
erinevaid 2-gramme	433	424	2,12
erinevaid 3-gramme	1459	1442	1,18
tekstis sõnu	463	493	6,08
unikaalseid sõnu	336	351	4,27
Hii-ruut	3,09	3,24	
sümbolite arv	2965	3040	

Tabel 11: Vootele Päi uudisartikli "Kremli hingus Prantsuse välispoliitika kohal" ja arvamusartikli "Mis on «Harta12» autorite tegelik agenda?" võrdlus

4.2.4 Priit Pulleritsu uudisartikli ja blogi postituse võrdlus

Järgmisena võrreldi Priit Pulleritsu "Arutu kulutus" ja blogi postituse "Kuidas ma sain Jõgeva rattarallil oma parima koha" artikleid. Tulemused on toodud välja tabelis 12.

	Arutu kulutus	Jõgeva rattaralli	Vahe (%)
keskmine sõna pikkus	5,5	5,16	6,62
keskmine lause pikkus	16,56	11,38	45,70
teksti tihedus (%)	68,49	66,43	3,10
erinevaid 2-gramme	439	424	3,53
erinevaid 3-gramme	1469	1420	3,45
tekstis sõnu	531	558	4,84
unikaalseid sõnu	363	370	1,89
Hii-ruut	6,09	5,89	
sümbolite arv	3002	2940	

Tabel 12: Priit Pulleritsu uudisartikli "Arutu kulutus" ja blogi postituse "Kuidas ma sain Jõgeva rattarallil oma parima koha" võrdlus

4.2.5 Vootele Päi ja Priit Pulleritsu uudisartiklite võrdlus

Järgmisena võrreldi Vootele Päi "Kremli hingus Prantsuse välispoliitika kohal" ja Priit Pulleritsu "Vahetame riiki" uudisartikleid. Tulemused on toodud välja tabelis 13.

	Kremli hingus	Vahetame riiki	Vahe (%)
keskmine sõna pikkus	6,31	5,87	7,54
keskmine lause pikkus	18,48	13,89	33,05
teksti tihedus (%)	72,72	67,2	8,23
erinevaid 2-gramme	433	483	10,35
erinevaid 3-gramme	1459	1521	4,07
tekstis sõnu	463	501	7,58
unikaalseid sõnu	336	336	0,00
Hii-ruut	12,66	13,11	
sümbolite arv	3002	2940	

Tabel 13: Vootele Päi "Kremli hingus Prantsuse välispoliitika kohal" ja Priit Pulleritsu "Vahetame riiki" uudisartiklite võrdlus

5 Tulemuste arutelu ja analüüs

Peatükis 4.1 toodi välja kõikide tekstide karakteristikud ning seejärel võrreldi neid karakteristikuid omavahel.

5.1 Vootele Päi kahe arvamusartikli võrdluse analüüs

Tabelis 9 on kahe Vootele Päi arvamusartikli võrdluse tulemused. Sellest tabelist on näha, et tunnusjoonte järgi on need kaks artiklit vägagi suure tõenäosusega sama autori poolt kirjutatud, mida need ka on. Ainuke suurem erinevus kahte artikli vahel on keskmine lause pikkus, mis erineb 12,2% võrra. Võttes kokku tabelis esitatud tulemused, on teksti erinevus kumulatiivselt 23,12%, mida tundub paljuvõitu, kuid aritmeetilise keskmise järgi on tekstide erinevus kõigest 3,3%.

2-grammide kokkulangevus kahe Vootele Päi artikli puhul on 80,5% ning 3-grammide kokkulangevus on 51,5%.

Hii-ruut-statistiku järgi, mille väärtusteks on 2,79 ja 2,66, vastavalt Sõjapõgenike vastutvõtmise artikli statistiline sarnasus Harta 12 artiklist ning vastupidi. Mõlemad tulemused on lävendist 12,5915[4] palju madalamad. Hii-ruut-statistiku järgi on Vootele Päi artiklid omavahel statistiliselt sarnased.

5.2 Priit Pulleritsu kahe uudisartikli võrdluse analüüs

Tabelis 10 on kahe Priit Pulleritsu uudisartikli võrdluse tulemused. Tabelist on hästi näha, et tunnusjoonte järgi on ka need kaks artiklit vägagi suure tõenäosusega sama autori poolt kirjutatud. Ainuke suurem erinevus kahte artikli vahel on, nagu ka Vootele Päi arvamusartiklit puhul, keskmine lause pikkus, mis erineb 16,14% võrra. Võttes kokku selle tabeli tulemused, siis kumulatiivselt on teksti erinevus 47,6%, kuid aritmeetilise keskmise järgi on tekstide erinevus kõigest 6,8%.

2-grammide kokkulangevus kahe Priit Pulleritsu artikli puhul on 79% ning 3-grammide kokkulangevus on 41%.

Hii-ruut-statistiku järgi, mille väärtusteks on 9,12 ja 9,17, vastavalt "Vahetame riiki' vastutvõtmise artikli statistiline sarnasus "Arutu kulutus" artiklist ning vastupidi.

[4]Hii-ruut-statistiku lävend 95% tõenäosusega ja vabadusastmete arvuga 6 (kuna meil on 7 karakteristikut, siis vabadusastmete arv on 7-1=6) on 12,5915

Mõlemad tulemusd on lävendist 12.5915 madalamad. Seega hii-ruut-statistiku järgi on Pulleritsu artiklid omavahel statistiliselt sarnased.

5.3 Vootele Päi uudis- ja arvamusartikli võrdluse analüüs

Tabelis 11 on Vootele Päi uudis- ja arvamusartikli võrdlus. Taaskord võib tabelist välja lugeda, et teksti tihedus on enam-vähem samas suurusjärgus. Keskmine sõnade pikkuse vahe on suurenenud võrreldes Vootele Päi kahe arvamusartikli võrdlusega, kuid keskmine lause pikkus on vähenenud mõne protsendi võrra.

2-grammid ühtivad 79% protsendi ulatuses ning 3-grammid alla pooltel juhtudel, täpsemalt 43,5% juhtudest. Kumulatiivselt on kahe teksti erinevus 29,31%, aritmeetilise keskmise järgi on kahe teksti erinevus 4,19%. Erinevus on küll suurenenud võrreldes arvamusartiklite omavahelise võrdlusega, kuid üldjoontes on Vootele Päi kirjutamisstiil nii arvamus- kui ka uudisartiklite puhul sarnane.

Hii-ruut-statistiku järgi on artiklid omavahel statistiliselt sarnased. Uudisartikli statistiline sarnasus Harta 12 teemat puudutava arvamusartikliga on 3,36, mis on peaaegu neli korda väiksem hii-ruut-statistiku lävendist. Vastupidisel variandil on statistiline sarnasus 3,50, mis on lävendist palju madalam. Hii-ruut-statistiku järgi on Vootele Päi arvamus- ja uudisartikkel statistiliselt sarnased.

5.4 Priit Pulleritsu uudisartikli ja blogi postituse võrdluse analüüs

Eelviimasena võrreldi Priit Pulleritsu uudisartiklit ning blogi postitust, mille tulemused on tabelis 12. Sellest tabelist on väga hästi näha, et tunnusjoonte järgi on need kaks artiklit vägagi suure tõenäosusega sama autori poolt kirjutatud, kuid suuresti erineb ainult keskmise lause pikkuse karakteristik, 45,70% protsendilise erinevusega. See näitab tema kirjutamisstiili erinevust teaduslike artiklite kirjutamisel ning reeglitevabama blogi kirjutamisel. Võttes kokku selle tabeli tulemused, siis kumulatiivselt on teksti erinevus kõigest 69,13% ning aritmeetilise keskmise järgi on tekstide erinevus 9,88%, kuid jättes välja keskmise lause pikkuse karakteristiku, siis tuleb kumulatiivselt erinevuseks 23,43% ning aritmeetiliselt kõigest 3,35%.

2-grammide kokkulangevus kahe Pulleritsu artikli puhul on 75,5% ning 3-grammide

kokkulangevus on 47%.

Hii-ruut-statistiku järgi, mille väärtusteks uudisartikli puhul on 6,09 ning blogi postituse puhul 5,89, mis on lävendist 12,5915 poole madalam. Hii-ruut-statistiku järgi on Priit Pulleritsu "Arutu kulutus" ja "Kuidas ma sain Jõgeva rattarallil oma parima koha?" artiklid statistiliselt sarnased.

Nende tunnusjoonte tulemuste järgi on näha, mis on Priit Pulleritsu artiklitele omane. Keskmine sõna pikkus on umbes 5,65, keskmine lause pikkus on 15,20 sõna ning teksti tihedus on 66,8% juures. Nende parameetritega saab luua Priit Pulleritsu kirjutamisstiili mudeli, mida saab võrrelda teiste küsitavate tekstidega. Kindlasti oleks parem luua autori tekstide mudelit, kui analüüsida olulisel suuremat hulka tema tekste. Sellisel moel loodud mudelit saaks kasutada tundmatu teksti analüüsimisel, et tuvastada kui suure tõenäosusega kirjutas selle teksti just Priit Pullerits.

5.5 Vootele Päi ja Priit Pulleritsu uudisartiklite võrdluse analüüs

Tabelis 13 on Vootele Päi "Kremli hingus Prantsuse välispoliitika kallal" ja Priit Pulleritsu "Vahetame riiki" uudisartiktlite võrdlus. Tabelist võib näha, et need kaks uudisartiklit on üldjoontes erinevad. Enamik tuunusjoontest erinevad üksteisest vähemalt 7% võrra. Unikaalsete sõnade arv on küll juhuse tahtel täpselt sama, kuid teksti sõnade arv erineb peaaegu neljakümne sõna võrra. Vootele Päi kasutab peaaegu viit sõna lauses rohkem, mis on 33,05% rohkem, kui seda teeb Priit Pullerits.

2-grammide kokkulangevus on 77,5% protsenti ning 3-grammide kokkulangevus kõigest 37,5%. Kumulatiivselt on kahe teksti erinevus 70,82%, aritmeetilise keskmise järgi 10,12%. Erinevus on märgatav nii aritmeetiliselt kui ka tabelist üksikud karakteristikuid vaadates. Seega ei saa kinnitada tõenäolist autorit, kuid ei saa ka täielikult seda ümber lükata. Vajalikud on spetsiifilisemad testimised tekstidega ning lisainformatsiooni kogumine autorite kirjutamisstiili kohta uudisartiklite valdkonnas.

Hii-ruut-statistiku järgi on artiklid omavahel statistiliselt erinevad. Vootele Päi uudisartikli statistiline sarnasus Priit Pulleritsu uudisartikliga on 12,66, vastupidiselt on statistiline sarnasus 13,11. Mõlemal puhul on hii-ruut-statistiku väärtus lävendist kõrgem, mis näitab kahe uudisartikli statistilist erinevust.

5.6 Autoritele omased tunnusjooned

Tekstide analüüsimise käigus tulid välja mitmed tunnusjooned, mis on autoritele omased.

Vootele Päile on omane keskmiselt üle kuue sõna kasutamine ühes lauses ning lausete pikkus (keskmiselt üle kahekümne sõna lauses) võrreldes Priit Pulleritsuga. Kuna tema sõnad ja laused on pikemad, siis tal on ka tekstis üldjoontes vähem sõnu. Vootele Päi kolme teksti juures tuli välja, et kõikidest sõnadest on tal üle 70% unikaalseid sõnu.

Priitu Pulleritsule on omane lühemad sõnad ning laused. Ta kasutab rohkem sõnu, kuid võrreldes Vootele Päiga on tal kasutusel vähem unikaalseid sõnu. Umbes 67-68% on tema tekstides unikaalseid sõnu kogu sõnade arvust. Tabelis 12 võrreldi uudisartiklit ja blogi postitust ning tuli välja, et Priit Pullerits kirjutab blogis veelgi lühemate sõnade ning lausetega. Sellest võib järeldada, et ta kirjutab blogis vabamalt, kui uudis- või arvamusartikleid kirjutades.

Autorsuse tuvastamine sõnavaraliste ja märgipõhiste tekstianalüüsi meetoditega

Jürgen Lorenz

Kokkuvõte

Inteneti lai levik tänapäeval on olulise probleemina tõstatanud tekstide autorsuse küsimuse. Mõningatel juhtudel on tingimata vaja kirjutaja isik kindlaks teha. Autorsuse tuvastamise meetodid annavad võimaluse võrdlemisi täpselt määrata kahtluse all oleva teksti autori.

Käesolevas bakalaureusetöös uuriti autorsuse tuvastamist sõnavaraliste ja märgipõhiste tekstianalüüsi meetoditega.

Tehtud uurimuse ning eksperimendi tulemuste põhjal võib järeldada, et sõnavarapõhise ja märgipõhise tekstianalüüsi meetoditega on edukalt võimalik tundmatu teksti autorit tuvastada. Seda näitas väga hästi tabelis 9 välja toodud Vootele Päi kahe arvamusartikli võrdlus ning vastava tabeli analüüs peatükis 5.1, mille tulemuseks oli kahe teksti aritmeetilise keskmise erinevus ligikaudu 3% ning hii-ruut-statistiku väärtus oli üle viie korra väiksem 95% tõenäosuse juures ja 6 vabadusastmete arvuga, mis näitab kui sarnased need kaks artiklit omavahel olid.

Samamoodi näitasid eksperimendi tulemused ka vastupidist ehk kahe teksti erinevust. Tabelis 13 ning peatükis 5.5 esitatud tabelite analüüsi põhjal on näha nii aritmeetiliselt kui ka kumulatiivselt kahe teksti erinevust ning ka hii-ruut-statistik kinnitab seda. Hii-ruut-statistiku väärtus oli mõlema võrdluse puhul lävendist kõrgem. Selline erinevus viitab järjekordselt sellele, et kasutusel olev mudel suudab kahte autorit omavahel eristada.

Tekstide erinevuse hindamiseks oli kasutusel kolm erinevat hindamisevalemit. Kõige paremini sobis hii-ruut statistik, sest see näitas kõige täpsemalt kahe teksti statistilist erinevust seitsme parameetri põhjal. Veel oli kasutusel aritmeetiline keskmine ja kumulatiivne hindamine, millele vastavalt leiti kas seitsme parameetri aritmeetiline keskmise või summa.

Sõnavaralised ja märgipõhised tekstianalüüsi meetodid on vägagi efektiivsed.

Need meetodid näitavad, et tõesti on igal autoril oma nii-öelda käekiri, mis on talle tahes-tahtmata omane. Esmane uurimus autorite tuvastamisel, kasutades kindlaid mudeleid, oli äärmiselt edukas, kuid siiski on vajalikud edaspidised laialdasemad uuringud. Järgmise etapina olekski plaanis teha ulatuslik uuring, mis tooks välja selle programmi võimalikud piirangud, ning vajadusel kohandada mudelit ka tulevikus kasutatavate suuremate andmestike jaoks.

Authorship identification using lexical and sign-based text analysing methods

Jürgen Lorenz

Summary

The wide spreading of internet in today's society has turned the ownership of texts into an important issue. In some cases, it is crucial to find out the persona of the writer. Methods of identifying the author give an opportunity to determine the author of the text in question with reasonable accuracy.

The BA thesis at hand studied identifying authors through methods of lexical and sign-based text analysis.

Based on the results of the study and experiment, it can be concluded that with lexical and sign-based text analysis methods, it is possible to successfully determine the author of an anonymous text. This was shown very well in table 9 where the comparison of Vootele Päi's two opinion articles has been pointed out and the analysis of the table on chapter 5.1, the result of which was the arithmetic difference of the two texts by at least 3% and the chi-square-statistic value was more than five times smaller at 95% probability and the number of 6 DOF's which shows how similar the two articles were.

Likewise, the results of the experiment also proved the opposite which is the difference of two texts. Based on the analysis of tables in table 13 and chapter 5.5, the difference of the two texts can be seen both arithmetically and cumulatively and the chi-square-statistic also confirms it. The value of the chi-square-statistic was higher than the threshold in both comparisons. A difference like this once again indicates that the model in use can distinguish one author from the other.

Three different evaluation formulas were used to evaluate the difference between the texts. The chi-square statistic worked best because it showed it was the most accurate in showing the statistic difference between the two texts based on seven parameters. Arithmetic average and cumulative evaluation were also used through which the arithmetic average or sum of the seven parameters were calculated.

Lexical and sign-based text analysis methods are very effective. These methods

show that every author does indeed have their own so-called script that is unavoidably characteristic to them. The initial study of identifying authors with the help of certain modules was extremely successful but still, further and more large-scale research is needed. The next step would be to carry out an extensive study which would bring out the possible restrictions of this programme and if necessary, the model could be adjusted for larger data systems that will be used in the future.

Viited

[1] M. Koppel, J. Schler, and S. Argamon. "Computational Methods in Authorship Attribution". In: *Journal of the American Society for Information Science and Technology* 60.1 (2009), pp. 9–26.

[2] Wikipedia. *Thomas Corwin Mendenhall*. URL: http://en.wikipedia.org/wiki/Thomas_Corwin_Mendenhall.

[3] Wikipedia. *George Udny Yule*. URL: http://en.wikipedia.org/wiki/Udny_Yule.

[4] Wikipedia. *George Kingsley Zipf*. URL: http://en.wikipedia.org/wiki/George_Kingsley_Zipf.

[5] Wikipedia. *Freferick Mosteller*. URL: http://en.wikipedia.org/wiki/Frederick_Mosteller.

[6] David I. Holmes and Judit Kardos. "Who was the author? An introduction to stylometry". In: *Chance* 16.2 (2003).

[7] E. Stamatatos. "A Survey of Modern Authorship Attribution Methods". In: *Journal of the American Society for Information Science and Technologyl* 60.3 (2009), pp. 538–556.

[8] Shlomo Argamon et al. "Automatically Profiling the Author of an Anonymous Text". In: *Commun. ACM* 52.2 (Feb. 2009), pp. 119–123. ISSN: 0001-0782. DOI: 10.1145/1461928.1461959. URL: http://doi.acm.org/10.1145/1461928.1461959.

[9] Wikipedia. *Big Five personality traits*. URL: http://en.wikipedia.org/wiki/Big_Five_personality_traits.

[10] S. Nirkhi and Dr. R.V. Dharaskar. "Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis". In: *International Journal of Advanced Computer Science and Applications* 4.5 (2013), pp. 32–35.

[11] Wikipedia. *Lexical Analysis*. URL: http://en.wikipedia.org/wiki/Lexical_analysis.

[12] Wikipedia. *n-gram*. URL: http://en.wikipedia.org/wiki/N-gram.

[13] Wikipedia. *Chi-squared test.* URL: http://en.wikipedia.org/wiki/Chi-squared_test.

[14] John Walker. *Chi-Square Calculator.* URL: https://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html.

[15] Vootele Päi. *Eesti võib teha endast sõjapõgenike vastuvõtmise kontekstis ärahellitatud oiviku.* URL: http://arvamus.postimees.ee/3193893/vootele-pai-eesti-voib-teha-endast-sojapogenike-vastuvotmise-kontekstis-arahellitatud-oivikul.

[16] Vootele Päi. *Mis on «Harta12» autorite tegelik agenda?* URL: http://arvamus.postimees.ee/1048450/vootele-pai-mis-on-harta12-autorite-tegelik-agenda.

[17] Vootele Päi. *Kremli hingus Prantsuse välispoliitika kohal.* URL: http://pluss.postimees.ee/3090623/kremli-hingus-prantsuse-valispoliitika-kohal.

[18] Priit Pullerits. *Vahetame riiki.* URL: http://tartu.postimees.ee/2749500/priit-pullerits-vahetame-riiki.

[19] Priit Pullerits. *Arutu kulutus.* URL: http://tartu.postimees.ee/2712864/priit-pullerits-arutu-kulutus.

[20] Priit Pullerits. *Kuidas ma sain Jõgeva rattarallil oma parima koha?* URL: http://suusk.blogspot.com/2015/05/pullerits-kuidas-ma-sain-jogeva.html.

Lisad

Programmi funktsioonid

Teksti analüüsimiseks on kõik järgnevad funktsioonid välja kutsutud eraldi teksti parameetrite saamiseks ning seejärel tehtud vajalikud arvutused võrdlemaks teise teksti parameetritega.

Tähtede loendur

```
def tahecounter(sona, n):
    count = Counter(sona)
    if n == 1:
      sobivad=''abcdefghijklmnopqrstuväöõüxzyw1234567890''
    else:
      sobivad= ''abcdefghijklmnopqrstuväöõüxzyw1234567890*!()=)"'?/][@£''
    summa = sum(count[letter] for letter in sobivad)
    return summa
```

See funktsioon loeb kokku sõnades esineva tähtede ja vajadusel ka sümbolite arvu. Parameetriteks võtab ta ühe sõna ning n-i väärtuse. Selleks kasutab ta *Python*'i teegis *collections* asuvat sissehitatud funktsiooni *Counter*. Tahecounter funktsioonis on kaks erinevat võimalust, mis on määratud n-i väärtusega:

1. n==1 korral ta loeb ainult tähed, mis esinevad tekstis
2. muu n-i väärtuse korral loeb ta kõik tähed ning sümbolid, mis esinevad tekstis

Lõpuks tagastab tähtede või tähtede ja sümbolite arvu, mis esines antud sõnas.

Sõna pikkuse loendur

```
def sonapikkus(fail):
    sonadearv = []
    tahecount = []
    laused = fail.split('.')
    for lause in laused:
        sonad = lause.split(' ')
        sonadearv.append(len(sonad))
        for i in sonad:
            i = i.lower()
            b = count_letters(i)
            if b == 0:
                continue
            else:
                tahecount.append(b)
    keskmine_tahecount= sum(tahecount)/sum(sonadearv)
```

See funktsioon leiab keskmise sõna pikkuse. Sisendparamateeriks võtab ka kogu teksti. Peale teksti sisse lugemist, poolitab laused '.' märgitud kohast. Iga lause läbimisel lisab ta sõnade listi iga sõna, et oleks võimalik pärast sõnade arvu kokku lugeda. Seejärel võtab iga lause eraldi ning poolitab lause sõnadeks tühiku kohast. Järgmisena töötleb sõna väiketähtede kujule ning seega kasutab 5.6 välja toodud esimese funktsiooni esimest punkti, et leida tähtede arv sõnas. Nii käib programm läbi kõik sõnad ning laused ja lõpuks tagastab keskmise tähtede arvu sõnas. Selleks liidab kokku kõikide sõnade tähtede arvu ning jagab selle kõikide sõnade arvuga.

Lause pikkus sõnades

```
def lausepikkus (fail):
    sonadearv = []
    laused = fail.split('.')
    for lause in laused:
        sonad = lause.split(' ')
        sonadearv.append(len(sonad))
    keskmine_sonadearv = (sum(sonadearv) 1)/(len(sonadearv) 1)
    return keskmine_sonadearv
```

Lause pikkus sõnades funktsioon võtab parameetriks kogu teksti. Taaskord poolitab ta teksti lauseteks ning seejärel sõnadeks. Funktsioon tagastab lõpuks keskmise sõnadearvu lausetes, mis on kogu sõnade arv jagatud lausete arvuga.

Unikaalsete sõnade arv

```
def unikaalne (fail):
    erisonad = []
    for punct in string.punctuation:
        fail = fail.replace(punct,"")
        fail = fail.lower()
    laused = fail.split(" ")
    for lause in laused:
        if lause == "":
            continue
        else:
            sonad = lause.split(' ')
            lause = lause.lower()
            if lause not in erisonad:
                erisonad.append(lause)
    return len(erisonad)
```

Funktsioon võtab kogu teksti ning alguses asendab iga kirjavahemärgistuse tühja sõnega ning viib teksti väiketähtede kujule. Tühiku kohast erladab sõnad üksteisest ning hakkab sõna-sõna haaval kontrollima, kas erisõnade listis on olemas see sõna, kui ei ole siis lisab selle sinna listi. Lõpuks väljastab unikaalsete arvu.

n-grammid

```
def ngrammid(word_list, n):
    ngrams = dict()
    for i in range(len(word_list)    n + 1):

        gram = tuple(word_list[i:i+n])

        if gram in ngrams:
            ngrams[gram] += 1
        else:
            ngrams[gram] = 1
    ngrams = sorted(ngrams, key=ngrams.get, reverse=True)
    return ngrams
```

See funktsioon on n-grammide väljastamiseks. Parameetriteks võtab kogu teksti ning n-i väärtuse, mis n-gramme vaja väljastada on. Ta käib teksti läbi vastavalt n-tähe haaval ning lisab n-grammide sõnastikku vastava n-grammi ning ka tema esinemiste arvu tekstis. Seejärel toimub sõnastiku sorteerimine suuremast väiksemani ning sõnastiku tagastamine.

Teksti pikkus

```
def tekstipikkus (fail):
    sonadearv = []
    tahecount = []
    laused = fail.split('.')
    for lause in laused:
        sonad = lause.split(' ')
        sonadearv.append(len(sonad))
        for i in sonad:
            i = i.lower()
            b = tahecounter(i,2)
            if b == 0:
                continue
            else:
                tahecount.append(b)
    return sum(tahecount), sum(sonadearv)
```

Võttes parameetriks kogu teksti, väljastab see programm sümbolite arvu tekstis, kus kasutab 5.6 välja toodud esimese funktsiooni teist punkti. Lisaks väljastab see ka sõnade arvu tekstis.

Sõnavaraline tihedus

```
def sonavaralinetihedus (fail):
    sonadearv = []
    laused = fail.split('.')
    for lause in laused:
        sonad = lause.split(' ')
        sonadearv.append(len(sonad))
    b = unikaalne (fail)
    tihedus = b/(sum(sonadearv) 1)*100
    return tihedus
```

Funktsioon arvutab välja sõnavaralise tiheduse ehk protsentuaalselt unikaalsete sõnade esinemise arvu kõikide sõnade seas. See funktsioon kasutab unikaalsete sõnade arvu leidmiseks Lisade peatükis 5.6 välja toodud funktsiooni.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Jürgen Lorenz (sünnikuupäev: 11 mai 1993),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose:
Autorsuse tuvastamine sõnavaraliste ja märgipõhiste tekstianalüüsi meetoditega,

mille juhendaja on Tõnu Tamme,

1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartu, 27.05.2015