Vabade tekstiväljade puhastamine isikuandmetest

TIS rakenduse administraator

Kalle Olumets

## Muudatused

|  |  |  |  |
| --- | --- | --- | --- |
| **Versioon** | **Aeg** | **Autor** | **Muudatuse kirjeldus** |
| 0.1 | 2012-11-13 | Kalle Olumets | Dokumendi algatus, struktuur |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Sisukord

[Vabade tekstiväljade puhastamine isikuandmetest 1](#_Toc340827434)

[Muudatused 2](#_Toc340827435)

[1 Sissejuhatus 4](#_Toc340827436)

[1.1 Ülevaade 4](#_Toc340827437)

[1.1 Probleemistiku taust 4](#_Toc340827438)

[1.2 Lühendid ja definitsioonid 4](#_Toc340827439)

[2 Vabatekstivälja andmekaeve 4](#_Toc340827440)

[2.1 Isikustatud andmed 4](#_Toc340827441)

[2.2 Vabad tekstikorpused 4](#_Toc340827442)

[2.3 Lause analüüs ja morfoloogiline analüsaator 6](#_Toc340827443)

[2.1 Tehnilised vahendid NLTK 8](#_Toc340827444)

# Sissejuhatus

## Ülevaade

Käesolev dokument kirjeldab isikuandmete eemaldamise probleemistikku TIS’ist. Eelkõige käsitletakse struktureerimata teksti andmekaeve probleemi vabatekstiväljal.

## Probleemistiku taust

Hetkel kasutusel olev võimalus andmetest isikuandmete eemaldamiseks ei ole piisav, (isikukoodid ja Rahvastikuregistrist saadud nimed). Vabatekstis on endiselt infot, mille tulemusena on isikud andmetega kaudselt või otseselt seostatavad ja tuvastatavad. Andmete kasutamine ja kolmandatele osapooltele uurimuslikeks või statistilisteks töödeks edastamine on süvenev vajadus. Koostöös STACC-iga oleks võimalik välja töötada tõhusam tehnoloogilise lahenduse prototüüp, mida ETSA saaks tulevikus kasutada

## Lühendid ja definitsioonid

|  |  |
| --- | --- |
| **Lühend** | **Selgitus** |
| nltk | Natural languange toolkit |
| TIS | Tervishoiu Infosüsteem |
|  |  |
|  |  |
|  |  |

# Vabatekstivälja andmekaeve

## Isikustatud andmed

Kuna isikustatud andmeid leidub ka kõikides teistes tekstides, siis peaks saama kasutada isikustatud andmete leidmiseks ka vabateksti korpuseid, näiteks EKI keeleveebi omi. Kindlasti on neis korpustes, eriti mis puudutab ajakirjandust, pärisnimesid ning aadresse. Üks TIS vabatekstiväljaga seonduv probleem on see, et isiku pärisnimi võib olla tekstis ka leksikaalselt teise funktsioonis, st ta võib olla kas nimisõna, verb, omadussõna vms. Sellise probleemi lahendamiseks ei pea tingimata kasutama TIS vabatekstivälju.

## Vabad tekstikorpused

Vabad eestikeelsed tekstikorpused on saadaval Eesti Keeletehnoloogia veebist:

<http://www.keeletehnoloogia.ee/projektid/koondkorpus>

Näitena võiks vaadata ühe vaba tekstikorpuse kasutatavust:

Näitena võtame sõna *must,* mis esineb ka eestikeelse perekonnanimena.

Building index...

Displaying 15 of 156 matches:

suurte mõõtmetega võrreldes odavad . must äri keerleb ümber telerite , muusika

rebima . Asemele kästi panna sini - must - valge mütsimärk . Et soomlased võt

t olen inimestele lähedane - olgu ta must , valge , värviline . Kui sind aland

” ja ahvi vahele paigutati alul kogu must rass , mis hiljem astmestati omakord

bi vaatepildi , kuidas tema maja ees must auto peatus ja sellest välja tormanu

vimuusikute klubi Muusik 1992 - 1993 must - valge muusikaleht . Väljaandjaks M

. Bäng Kaks korda kuus ilmunud õhuke must - valge üllitis , mis lõpetas kooper

A Tahab surra targemana kui sündis . must preester Paavo Kangur “ Minu Ameerik

t , ” ütleb Pajula . Hardo Pajula on must preester , ütleb kolleeg Sulev Vedle

, millele järgneb pärissõda , räägib must preester . Palju on inimkond õppinud

**Savisaar ( 49 . ) . Ajaloolane Aadu must ( 21 . ) kandideerib koos tütre Kadr**

, et ükskõik , kuhu ma lähen , jääb must maha märg rada . Vahel lasen Kuul ta

ideestik on kokkuvõetav järgmiselt - must noorsugu Ameerikas ei suuda oma unis

elt oma olemuselt üdini negatiivne , must idee . Sest kui vaadata seda tolle l

pop “ Suitsetav Zeig ” , reggaelik “ must / valge ” või uimastavalt kütkestav

Saadud lihtne tulem ilmestab pärisnimede ja temaga foneetiliselt samade sõnade esinemist vabatekstikorpuses. Tumedamas kirjas on väljatoodud lõik, kus sõna *must* esines kui pärisnimi. Antud juhul kasutati ühte osa Eesti Ekspressi tekstikorpusest:

<http://www.cl.ut.ee/korpused/segakorpus/ekspress/index.php?lang=et>

Eesti Keeletehnoloogia veebis on ka ajakirja „Eesti Arst“ 2002 – 2004 väljaannetest moodustatud korpus. <http://www.cl.ut.ee/korpused/segakorpus/eestiarst/>

Vaatleme sellest sõna *must* esinemist.

concordance:

Building index...

Displaying 5 of 5 matches:

kivitunne kõhus , aeg - ajalt esinev must väljaheide ja söögiisu langus ) alga

*ib olla püsiv . Blum M , Harris SS , must A , Phillips SM , Rand WM , Dawson -*

Joonis 2 . Tüüpiline pulsisageduse ( must ) ja vere laktaadisisalduse ( punane

“ suur katk ” , “ tohutu katk ” , “ must surm ” , “ tõsine idamaa katk ” . Üh

ivad obsessiivsed vastandamised nagu must / valge , hea / halb , parem / vasak

Võib öelda et Eesti Arsti selles tekstikorpuses esineb sõna must kui omadussõna va. teine rida, milles selle sõnaliik on kahtluse all.

Seega saaks Eesti Ekspressi(või mõnda muud ajakirjandus vms. korpust) ja Eesti Arsti korpust kasutades saada pärisnimede ja nendega foneetiliselt samaste kuid teises funktsioonis olevate sõnade probleemi lahendamiseks vajaliku tekstikorpuse.

## Lause analüüs ja morfoloogiline analüsaator

Pärisnime ja temaga foneetiliselt sama kuid muu liiki sõna erineb lauses oma funktsiooni poolest. Seega saab kasutada morfoloogilist analüsaatorit, et määrata sõna liik. Kahjuks mulle tuntud eesti keele morfoloogiline analüsaator <http://www.filosoft.ee/html_morf_et> määrab ära küll pärisnimesid, kuid kui see on samane foneetiliselt mingi muu sõnaga millel on tähendus, siis ta seda pärisnimeks ei pea. Näiteks lauses

Aadu  
    Aadu+0 //\_H\_ sg g, sg n, //

Must  
    mina+st //\_P\_ sg el, //  
    must+0 //\_A\_ sg n, //  
    must+0 //\_S\_ sg n, //

sõi  
    söö+i //\_V\_ s, //

musta  
    must+0 //\_A\_ adt, sg p, //  
    must+0 //\_A\_ sg g, //  
    must+0 //\_S\_ adt, sg p, //  
    must+0 //\_S\_ sg g, //  
    musta+0 //\_V\_ o, //

leiba  
    leib+0 //\_S\_ adt, sg p, //

Seega tuleks leida muud iseloomulikud kirjeldav omadused, mis eristaks pärisnime muudest sõnadest. Ühe hüpoteesina võiks selleks olla lähedased sõnad. Võib oletada näiteks, et perekonna nimed käivad tavaliselt koos eesnimega. Et neid hüpoteese kontrollida, uleb luua treeningu nimekiri ja testimise nimekiri sõnadest ja eelnenud sõnast/sõnadest ning leida treeningu käigus enim kirjeldavaid omadusi.

Sellise lähenemise puhul võib näha, et isegi 910 kirjega treeningnimekirja puhul saab üsna suure tõenäosusega leida enim kirjeldavaid omadusi, mis eristavad pärisnime mitte pärisnimest lauses.

Most Informative Features

eel\_sona\_liik = 'H' p : n = 20.2 : 1.0

eel\_sona\_liik = 'A' n : p = 13.2 : 1.0

eel\_sona\_liik = 'P' n : p = 7.9 : 1.0

sonade\_arv = 20 n : p = 5.1 : 1.0

eel\_sona\_liik = 'D' n : p = 4.3 : 1.0

Accuracy: 0.76

Sellest väljatrükist on näha, et pärisnime enimkirjeldavaks omaduseks on eelnenud sõna pärisnimena esinemine (H – pärisnime lühend, A – omadussõna, P – asesõna, D – määrsõna jne.[[1]](#footnote-2) ). Väikese tähega lühendid tähendavad eeltöödeldud lausetes esinenud sõna pärisnimelisust: p – pärisnimi ja n – mitte pärisnimi.

Kasutades seda treeningnimekirja ja testnimekirja saame luua ka klassifikaatori ja testida seda suvaliste lausetega.

# oodatud tulemus: p

cf.classify(ft('Minu nimi on mall karu, kuidas on siis, tere','karu'))

tulemus: p

# oodatud tulemus: n

cf.classify(ft(' orda enam massi paberitootmiseks kui mets , lisaks on kanepi koristamine ja tr','mets'))

tulemus: n

# oodatud tulemus: n

cf.classify(ft(' Isegi sellise linnuliigi puhul nagu must karkjalg , keda on looduses','must'))

tulemus: n

# oodatud tulemus: p

cf.classify(ft('Keegi ivar valge kurtis elu','valge'))

tulemus: p

# oodatud tulemus: n

cf.classify(ft('selle kala kast oli must','kast'))

# oodatud tulemus: p

cf.classify(ft('direktor kalle kask seisab kindlalt','kask'))

tulemus: n

Viimane väär tulemus tuleb sellest, et morfoloogiline analüsaator käsitleb eesnime *kalle* mitte kui pärisnimena vaid kui nimisõnana.

|  |
| --- |
| direktor     direktor+0 //\_S\_ sg n, // |
| kalle     kalle+0 //\_S\_ sg n, // |
| kask     kask+0 //\_S\_ sg n, // |
| seisab     seis+b //\_V\_ b, // |
| kindlalt     kindel+lt //\_A\_ sg abl, //     kindlalt+0 //\_D\_ // |

Selle jaoks saaks edaspidi leida paremini kirjeldavaid omadusi ... (näiteks üle-eelmise või ülejärgmise sõna liik vms.)

## Tehnilised vahendid NLTK

## Tehniliste vahenditena on tekstikaeveks hea kasutada päriskeele töövahendeid ehk *natural languange toolkit’te.* Kasutasin antud näidete puhul pythoni nltk’d <http://nltk.org/>

# Kokkuvõte ja järeldused

Vabateksti e. struktureerimata teksti andmekaeve on võimalik kui kasutada erinevaid päriskeele töövahendeid (nltk) ja eeltöödeldud tekstikorpusi. Samuti on võimalik lahendada pärisnime ja temaga foneetiliselt samase kuid mitte pärisnime funktsioonis oleva sõna eristamise probleemi lauses, kasutades selleks eeltreenitud vahendit ja seda üsna suure tõenäosusega.

1. <http://www.filosoft.ee/html_morf_et/morfoutinfo.html> [↑](#footnote-ref-2)