Estnltk — open source tools for Estonian natural language processing

The Why Linguistics Conference May 7-9 2015, Tartu, Estonia

Timo Petmanson
University of Tartu
tpetmanson@gmail.com

Why natural language processing?

computer science + linguistics

- For researches / software developers
  - cleaning text, preparing for indexing & search
  - information extraction, text analysis

- For linguists
  - same reasons as for computer scientists in forms of usable tools
  - needed to answer any questions about large corpora

Why estnltk?

Easy to learn and use

Solves common problems for software developers

- tokenization, stemming, POS-tagging, named entity recognition

Practical tools for linguists

- sentiment analysis, text classification, information extraction

Free and open source under GPLv2 license

Download & install

$ pip install estnltk

warning: the talk is about the development version, which will be released by the end of May 2015

Github: https://github.com/estnltk/estnltk

Docs: http://estnltk.github.io/estnltk/index.html

Fixing spelling mistakes

```
>>> from estnltk import Text
>>> text = Text('Vikastes lausetes on trügivigasid!')
>>> text.fix_spelling()

>>> print (text)
Vigastes lausetes on trükivigasid!
```

Fixing spelling mistakes

More detailed output

```
>>> from estnltk import Text
>>> from pprint import pprint
>>> text = Text('Vikastes lausetes on trügivigasid!')
>>> text.word_texts
['Vikastes', 'lausetes', 'on', 'trügivigasid', '!']
>>> text.spellcheck_words()
[False, True, True, False, True]
>>> text.spellcheck_suggestions()
[['Vigastes', 'Vihastes'], [], [], ['trükivigasid'], []]
```

Fixing spelling mistakes

Even more detailed output

```
>>> from estnltk import Text
>>> from pprint import pprint
>>> text = Text('Vikastes lausetes on trügivigasid!')
>>> pprint(text.spellcheck())
[{'spelling': False,
  'suggestions': ['Vigastes', 'Vihastes'],
  'text': 'Vikastes'},
 {'spelling': True, 'suggestions': [], 'text': 'lausetes'},
 {'spelling': True, 'suggestions': [], 'text': 'on'},
 {'spelling': False, 'suggestions': ['trükivigasid'], 'text': 'trügivigasid'},
 {'spelling': True, 'suggestions': [], 'text': '!'}]
```

Cleaning text from unwanted characters

Filtering and displaying invalid characters

```
>>> from estnltk import TextCleaner
>>> tc = TextCleaner()
>>> text = '''Ilmateade:¶¶
... Esmaspäeval (27.04) liikus madalrõhkkond Põhjalahelt Soome kohale.¶'''
>>> tc.invalid_characters(text)
'¶ō'
>>> text = text.replace('ō', 'õ')
>>> print (tc.clean(text))
Ilmateade:

Esmaspäeval (27.04) liikus madalrõhkkond Põhjalahelt Soome kohale.
```

ō latin small letter o with macron (U+014D)
õ latin small letter o with tilde (U+00F5)

Cleaning text from unwanted characters

Predefined alphabets

```
>>> from estnltk.textcleaner import *
>>> print (EST_ALPHA)
abcdefghijklmnoprsšzžtuvwõäöüxyzABCDEFGHIJKLMNOPRSŠZŽTUVWÕÄÖÜXYZ
>>> print (RUS_ALPHA)
абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ
>>> print (PUNCTUATION)
!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~–
>>> print (DIGITS)
0123456789
>>> WHITESPACE
'\t\n\x0b\x0c\r '
```

Cleaning text from unwanted characters

Checking if text contains only characters present in the alphabet

```
>>> from estnltk.textcleaner import TextCleaner, ESTONIAN, RUSSIAN
>>> tc_et = TextCleaner(ESTONIAN)
>>> tc_ru = TextCleaner(RUSSIAN)

>>> tc_et.is_valid('Segan suhkrut malbelt tassis, kus nii armsalt aurab tee.')
True
>>> tc_ru.is_valid('Segan suhkrut malbelt tassis, kus nii armsalt aurab tee.')
False
>>> tc_et.is_valid('Дождь, звонкой пеленой наполнил небо майский дождь.')
False
>>> tc_ru.is_valid('Дождь, звонкой пеленой наполнил небо майский дождь.')
False
```

Tokenization

```
>>> from estnltk import Text
>>> text = Text('''Keeletehnoloogia on arvutilingvistika
praktiline pool.
... Keeletehnoloogid kasutavad arvutilingvistikas välja töötatud
... teooriaid, et luua rakendusi (nt arvutiprogramme),
... mis võimaldavad inimkeelt arvuti abil töödelda ja
mõista.''')
```

Tokenization

Words

```
>>> text.word_texts
['Keeletehnoloogia', 'on', 'arvutilingvistika', 'praktiline',
'pool', '.', 'Keeletehnoloogid', 'kasutavad',
'arvutilingvistikas', 'välja', 'töötatud', 'teooriaid', ',',
'et', 'luua', 'rakendusi', '(', 'nt', 'arvutiprogramme', '),',
'mis', 'võimaldavad', 'inimkeelt', 'arvuti', 'abil', 'töödelda',
'ja', 'mõista', '.']
```

Tokenization

Sentences

```
>>> text.sentence_texts
['Keeletehnoloogia on arvutilingvistika praktiline pool.',
 'Keeletehnoloogid kasutavad arvutilingvistikas välja
töötatud\n'
 'teooriaid, et luua rakendusi (nt arvutiprogramme),\n'
 'mis võimaldavad inimkeelt arvuti abil töödelda ja mõista.']
```

Morphological analysis

```
>>> text.lemmas
['keeletehnoloogia', 'olema', 'arvutilingvistika', 'praktiline',
'pool', '.', 'keeletehnoloog', 'kasutama', 'arvutilingvistika',
'välja', 'töötama|töötatud', 'teooria', ',', 'et', 'looma',
'rakendus', '(', 'nt', 'arvutiprogramm', '),', 'mis',
'võimaldama', 'inimkeel', 'arvuti', 'abil', 'töötlema', 'ja',
'mõistma', '.']
```

Morphological analysis

```
>>> zip(text.word_texts, text.postags)
[('Keeletehnoloogia', 'S'), ('on', 'V'), ('arvutilingvistika', 'S'),
('praktiline', 'A'), ('pool', 'S'), ('.', 'Z'), ('Keeletehnoloogid', 'S'),
('kasutavad', 'V'), ('arvutilingvistikas', 'S'), ('välja', 'D'), ('töötatud',
'A|V'), ('teooriaid', 'S'), (',', 'Z'), ('et', 'J'), ('luua', 'V'), ('rakendusi',
'S'), ('(', 'Z'), ('nt', 'Y'), ('arvutiprogramme', 'S'), ('),', 'Z'), ('mis',
'P'), ('võimaldavad', 'V'), ('inimkeelt', 'S'), ('arvuti', 'S'), ('abil', 'K'),
('töödelda', 'V'), ('ja', 'J'), ('mõista', 'V'), ('.', 'Z')]
```

Morphological analysis

```
text.text, text.word_texts, text.word_spans, text.analysis,
text.roots, text.lemmas, text.endings, text.forms, text.postags,
text.root_tokens
```

```
>>> Text('allmaaraudteejaam').root_tokens
[['all', 'maa', 'raud', 'tee', 'jaam']]
```

Morphological synthesis

```
>>> from estnltk import synthesize
>>> synthesize('pood', 'pl p', partofspeech='S')
['poode', 'poodisid']
>>> synthesize('palk', 'sg kom')
['palgaga', 'palgiga']
```

Named entity recognition

```
>>> from estnltk import Text
>>> text = Text('Jan Palmér sõitis Tartust Tallinnasse istungile Estonian Airi kontorisse')
>>> zip(text.named_entities, text.named_entity_labels)
[('Jan Palmér', 'PER'), ('Tartu', 'LOC'), ('Tallinn', 'LOC'), ('Estonian Air|Airi', 'ORG')]
```

Named entity recognition

Word level labels

```
>>> zip(text.word_texts, text.labels)
[('Jan', 'B-PER'), ('Palmér', 'I-PER'),
('sõitis', 'O'),
('Tartust', 'B-LOC'), ('Tallinnasse', 'B-LOC'),
('istungile', 'O'),
('Estonian', 'B-ORG'), ('Airi', 'I-ORG'),
('kontorisse', 'O')]
```

B - start of the named entity
I - middle or end
O - not named entity

Temporal expressions

```
>>> from estnltk import Text
>>> text = Text('Potsataja ütles eile, et vaatavad nüüd Genaga viie aasta plaanid uuesti üle.')
>>> zip(text.timex_texts, text.timex_values)
[('eile', '2015-05-06'), ('nüüd', 'PRESENT_REF'), ('viie aasta', 'P5Y')]
```

Temporal expressions

Changing document creation date

```
>>> from estnltk import Text
>>> from datetime import datetime
>>> text = Text('Potsataja ütles eile, et vaatavad nüüd Genaga viie
aasta plaanid uuesti üle.', creation_date=datetime(1234, 5, 6))
>>> zip(text.timex_texts, text.timex_values)
[('eile', '1234-05-05'), ('nüüd', 'PRESENT_REF'), ('viie aasta', 'P5Y')]
```

There is more in the API ...

- Clause detector
- Verb chain detection
- Wordnet interface
- TEI corpus reader
- HTML pretty printer

Estonian text classifier tool

- Machine learning software for organizing data into categories.
- Works with Excel and CSV files
- Extensive feedback system
  - precision, recall, accuracy
  - important features
- Command line programs + API

Example training dataset

Kommentaari ID	Kommentaar	Meelsus
8	väga hea firma	Positiivne
10	Viimasel ajal pole midagi halba öelda, aga samas ei konkureeri nad kuidagi Genneti, Ordiga ei hindadelt ega teeninduselt. Toorikute ja tindi ostmiseks samas hea koht ja kuna müüjaid on rohkem valima hakatud, siis võiks 2 ikka ära panna - tuleks 3 kui hinadele ei pandaks kirvest ja toodete saadavus oleks parem.	Negatiivne
11	Fotode kvaliteet väga pro ja “jjk” seal töötamise ajal leiti ikka paljudele asjadele väga meeldivad lahendused. Samas hilisem läpaka ost sujus ka väga meeldivalt - sain esialgse rahas ostusoovi vormistada ümber järelmaksule...äärmiselt asjalik teenindus.	Positiivne
13	Ainult positiivsed kogemused	Positiivne
16	Viimane kord, kui käisin suutis leti taga askeldav ~60 aastane mees tegutseda nii aeglaselt, et minu seal	Negatiivne

hinnavaatlus.ee dataset

Resulting model can predict the sentiment

Kommentaari ID	Kommentaar	Meelsus
8	väga hea firma	?
10	Viimasel ajal pole midagi halba öelda, aga samas ei konkureeri nad kuidagi Genneti, Ordiga ei hindadelt ega teeninduselt. Toorikute ja tindi ostmiseks samas hea koht ja kuna müüjaid on rohkem valima hakatud, siis võiks 2 ikka ära panna - tuleks 3 kui hinadele ei pandaks kirvest ja toodete saadavus oleks parem.	?
11	Fotode kvaliteet väga pro ja “jjk” seal töötamise ajal leiti ikka paljudele asjadele väga meeldivad lahendused. Samas hilisem läpaka ost sujus ka väga meeldivalt - sain esialgse rahas ostusoovi vormistada ümber järelmaksule...äärmiselt asjalik teenindus.	?
13	Ainult positiivsed kogemused	?
16	Viimane kord, kui käisin suutis leti taga askeldav ~60 aastane mees tegutseda nii aeglaselt, et minu seal	?

hinnavaatlus.ee dataset

Feedback reporting

Classification report

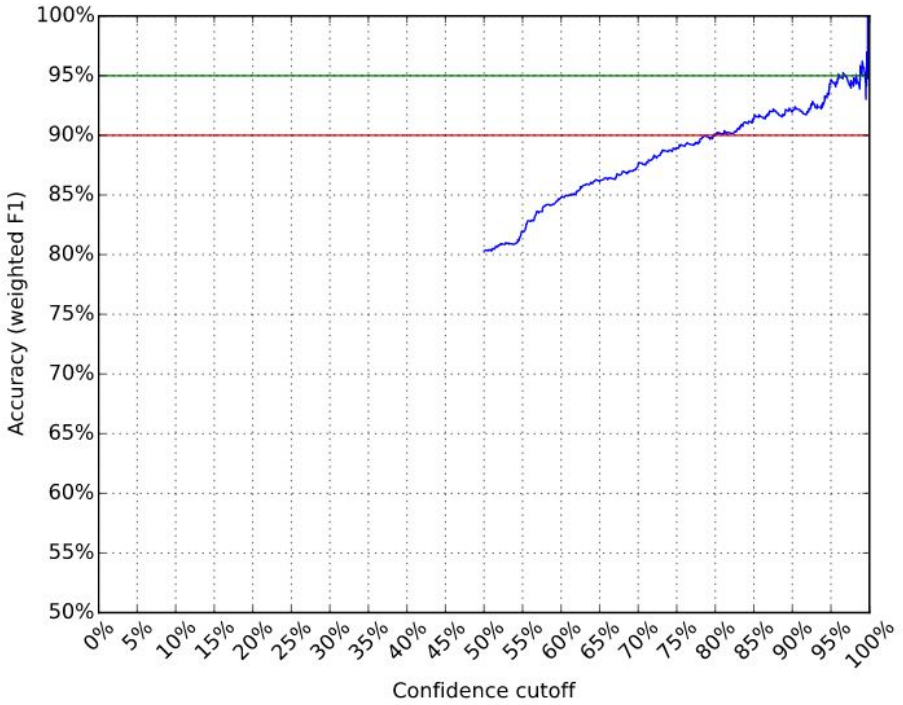
Precision	Recall	F1-score
77.3	85.2	81.0

Ordered by F1

Class	Precision	Recall	F1	Support/Count
Positiivne	77.3	85.2	81.0	492
Negatiivne	73.1	61.7	66.9	321

Ordered by support

Class	Precision	Recall	F1	Support/Count
Positiivne	77.3	85.2	81.0	492
Negatiivne	73.1	61.7	66.9	321



hinnavaatlus.ee dataset

Negatiivne	ainu, algama, **alles**, **arve**, asi, **asima**, eesti, **eima**, **enam**, **enne**, **eriti**, **et**, **garantii**, helistama, hommik, **huvitama**, igati, **juba**, **juhtuma**, julgema, jõudma, kaks, **kallim**, **kee**, kenasti, kiire, **kinni**, **kiri**, **kohale**, koht, **kohtama**, kord, korrektne, kuidas, **kuigi**, käes, kõik, **küsima**, **laos**, leidma, **lisama**, **mail**, **maksma**, meeldiv, **minema**, **minkima**, minut, **mitte**, muidu, **negatiivne**, **nädal**, **omama**, **ootama**, ost, ostma, ostnu, **ots**, **otsama**, **ox**, **pandud**, peale, positiivne, päev, rahu, **rohkem**, **saatma**, sai, **see**, **siin**, **siiski**, **soov**, **soovima**, super, suur, suurepärane, sõbralik, teadma, teenindus, **tegemine**, **teine**, tellima, toimima, tooma, **täiesti**, tänama, **vaatama**, vahetama, **videokaart**, viima, väga, õhtu, õige, ühe, **ei ole**, hea teenindus, hind oli, ja kiire, ja kõik, **on ka**, samal päeval
Positiivne	**ainu**, **algama**, alles, arve, **asi**, asima, **eesti**, eima, enam, enne, eriti, et, garantii, **helistama**, **hommik**, huvitama, **igati**, juba, juhtuma, **julgema**, **jõudma**, **kaks**, kallim, kee, **kenasti**, **kiire**, kinni, kiri, kohale, **koht**, kohtama, **kord**, **korrektne**, **kuidas**, kuigi, **käes**, **kõik**, küsima, laos, **leidma**, lisama, mail, maksma, **meeldiv**, minema, minkima, **minut**, mitte, **muidu**, negatiivne, nädal, omama, ootama, **ost**, **ostma**, **ostnu**, ots, otsama, ox, pandud, **peale**, **positiivne**, **päev**, **rahu**, rohkem, saatma, **sai**, see, siin, siiski, soov, soovima, **super**, **suur**, **suurepärane**, **sõbralik**, **teadma**, **teenindus**, tegemine, teine, **tellima**, **toimima**, **tooma**, **täiesti**, **tänama**, vaatama, **vahetama**, videokaart, **viima**, **väga**, **õhtu**, **õige**, **ühe**, ei ole, **hea teenindus**, **hind oli**, **ja kiire**, **ja kõik**, on ka, **samal päeval**

True label: Positiivne
Predicted label: Negatiivne
Count: 66

Kommentaari ID	Kommentaar	Meelsus
23	Olen Tartu esindustest ostnud probleemivabalt igasugu pudi-padi, teenindus on olnud neutraalne, ei midagi mainimistväärselt **positiivset** aga õnneks ka mitte negatiivset.	Positiivne
26	Sülearvutite valik suur ja hinnad on soodsad	Positiivne
31	Probleeme pole olnud, suhtlemine väga personaalsel tasemel. **Hinnad** ja asukoht jätavad aga soovida.	Positiivne
53	Siiani olen kõik asjad aetud saanud. ise tead mida tahad siis ei ole ka probleemi	Positiivne
58	Hehhee, kes siis zorgile ja tema prosekollektsioonile head hinnet ei paneks	Positiivne
62	Suur kaupade valik kohapeal olemas, kuigi **hinnad** suht kallid.	Positiivne
74	Mõned asjad ostnud. Viimati ostsin HDD, aga temperatuur **oli** liiga kõrge, arvuti jooksis kokku. Algul väideti et see normaalne ja tuleb paigaldada jahutus. Aga väikese vaidlemise peale võeti garantiise(aga öeldi et niikuinii saadetakse tagasi), lubati 2 nädala pärast helistada. Helistatigi 2 nädala pärast ja anti uus ketas. Teenindus võiks tõesti olla parem, aga kokkuvõttes täitsa norm pood.	Positiivne

Grammar based information extractor

The problem: blood pressure extraction for medical discharges

Source text	systolic	diastolic
mõõdetud **RR 110/80 mm Hg**	110	80
RR vasakul käel (**mm/Hg**) : **131/103**.	131	103
vererõhu kõikumine **97/52 mmHg** kuni **180/116 mmHg**	97	52
Päevaajal keskmine **RR 142/67 mmHg**	142	67

Defining the symbols of the grammar

Päevaajal keskmine **RR 142/67 mmHg**

```
symbol SystolicValue
regexes
    \d{2,3}
examples
    number >>25<<
    nu,ber >>999<< s

symbol DiastolicValue
regexes
    \d\d\d
    \d\d
```

```
symbol BloodPressureName
words
    v/r
case sensitive words
    RR
    VR
```

```
symbol BloodPressureValue
productions
    SystolicValue '\s?' ('/' '\s?' DiastolicValue)?
```

Defining the symbols of the grammar

```
102  # defineerime lõpuks sümboli <BloodPressure>.
103  # mainime igaks juhuks, et sümbolite järjekord failis pole oluline
104  symbol BloodPressure
105  productions
106      BloodPressureName? '[ -]*' BloodPressureValue ('\s*' BloodPressureUnit)?
107  examples
108      Päevaajal keskmine >>RR 142/67 mmHg<<
109      >>RR 110/80 mm Hg<<
110      >>RR vasakul käel ( mm/Hg ) : 131/103<<.
111      >>VR-157/94<< P-72
112      >>RR 156/85mmHg<<, fr.88x´ .
113      Obj : >>RR 145/80 mmHg<< , fr 77 x regul .
114      Täna >>RR 176/100mmHg<< .
```

Defining the database table layout

```
131  # viimase asjane tutvustame <export> direktiivi, mille abil saab öelda, millised
132  # sümbolid andmebaasi salvestada ja mis tüübiks nad tuleb teisendada.
133  # Näiteks vererõhkude puhul meid ei tegelikult ei huvita midagi peale
134  # süstoolse ja diastoole väärtuse:
135  export BloodPressure
136      SystolicValue   integer
137      DiastolicValue  integer
138
139  # kuigi meil oleks võimalik ka salvestada kirjeldamisel kasutatud nimi ja ühik
140      BloodPressureName string
141      BloodPressureUnit string
```

Contributors & supporters

The Filosoft team: Tarmo Vaino, Heiki-Jaan Kaalep

Timo Petmanson

Siim Orasmaa

Alexander Tkachenko

Karl-Oskar Masing

Raul Sirel

Sven Laur

Eesti keeletehnoloogia riiklik programm (2011-2017)

https://www.keeletehnoloogia.ee/et