

# Estnltk — Pythoni teek eestikeelsete tekstide töötlemiseks

Eesti keeletehnoloogia 2015 konverents

29.-30. oktoober, Ahhaa keskus, Tartu

Timo Petmanson

Tartu Ülikool

[tpetmanson@gmail.com](mailto:tpetmanson@gmail.com)



# Estnltk eesmärgid

Lihtne õppida ja õpetada

Tüüpülesannete lahendamine

sõnestamine, lemmatiseerimine, morfoloogiline analüüs, nimeolemite tuvastus

Tööriistade loomine

teksti klassifitseerimine, struktureeritud info eraldamine tekstidest

Tasuta ning avatud lähtekoodiga GPLv2 alusel



# Paigaldamine ja juhendid

Linux Mint 17

```
sudo apt-get install g++ python3-dev python3-pip python3-wheel python3-numpy swig  
sudo pip3 install estnltk
```

Paigaldusjuhend ka Windowsi jaoks, aga Mac OS X tugi puudub

<http://estnltk.github.io/estnltk/1.3/tutorials/installation.html>

Loomuliku keele töötlemise kursus TÜ-s: <https://courses.cs.ut.ee/2015/pynlp/fall>

Lähtekood: <https://github.com/estnltk/estnltk>

Kasutusjuhend: <http://estnltk.github.io/estnltk/>



# Trükivigade parandamine

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

Vigastes lausetes on trükivigasid!

# Trükivigade parandamine

Üksikasjalikum kirjeldus

```
from estnltk import Text
text = Text('Vikastes lausetes on trügivigasid!')
print (text.word_texts)
['Vikastes', 'lausetes', 'on', 'trügivigasid', '!']
print (text.spelling)
[False, True, True, False, True]
print (text.spelling_suggestions)
[['Vigastes', 'Vihastes'], [], [], ['trükivigasid'], []]
```

# Tükeldamine

Näidistekst

```
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.')
```



# Lausestamine

Statistiline lausestaja treenitud Eesti Päevalehe artiklitel

```
print (text.sentence_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.']
```

# Sõnestamine

Kohandatud NLTK WordPunktTokenizer

```
print (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', '.']
```



# Lemmatiseerimine

**print** (text.lemmas)

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

# Sõnaliikide määramine

```
print (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')]
```

# Mitmesused

**print** (Text('töötatud').analysis)

```
[[{'clitic': '',  
  'ending': 'tud',  
  'form': 'tud',  
  'lemma': 'töötama',  
  'partofspeech': 'V',  
  'root': 'tööta',  
  'root_tokens': ['tööta']}],  
 {'clitic': '',  
  'ending': '0',  
  'form': '',  
  'lemma': 'töötatud',  
  'partofspeech': 'A',  
  'root': 'tööta=tud',  
  'root_tokens': ['töötatud']}],  
 [{'clitic': '',  
  'ending': '0',  
  'form': 'sg n',  
  'lemma': 'töötatud',  
  'partofspeech': 'A',  
  'root': 'tööta=tud',  
  'root_tokens': ['töötatud']}],  
 [{'clitic': 'd',  
  'ending': 'd',  
  'form': 'pl n',  
  'lemma': 'töötatud',  
  'partofspeech': 'A',  
  'root': 'tööta=tud',  
  'root_tokens': ['töötatud']}]]]
```

# Morfoloogilise analüüsi meetodid

text.text, text.word\_texts, text.word\_spans, text.analysis, text.  
roots, text.lemmas, text.endings, text.forms, text.postags, text.  
root\_tokens

```
print (Text('allmaaraudteejaam').root_tokens)
```

```
[['all', 'maa', 'raud', 'tee', 'jaam']]
```

# Morfoloogiline süntees

```
from esnltk import synthesize
```

```
print (synthesize('pood', 'pl p', partofspeech='S'))
```

```
['poode', 'poodisid']
```

```
print (synthesize('palk', 'sg kom'))
```

```
['palgaga', 'palgiga']
```

# Nimeolemite tuvastamine

```
text = Text('Eesti piirneb põhjas üle Soome lahe Soome  
Vabariigiga. Riigikogu on Eesti Vabariigi parlament. 2005. aastal  
sai peaministriks Andrus Ansip.')
```

```
print (zip(text.named_entities, text.named_entity_labels))  
[('Eesti', 'LOC'), ('Soome laht', 'LOC'), ('Soome Vabariik', 'LOC'),  
( 'Riigikogu', 'ORG'), ('Eesti vabariik', 'LOC'), ('Andrus Ansip',  
'PER')]
```

# Nimeolemite tuvastamine

```
print (zip(text.word_texts, text.labels))
```

[('Eesti', 'B-LOC'), ('piirneb', 'O'), ('põhjas', 'O'), ('üle', 'O'),  
('Soome', 'B-LOC'), ('lahe', 'I-LOC'), ('Soome', 'B-LOC'),  
('Vabariigiga', 'I-LOC'), ('.', 'O'), ('Riigikogu', 'B-ORG'), ('on',  
'O'), ('Eesti', 'B-LOC'), ('Vabariigi', 'I-LOC'), ('parlament', 'O'), ('.',  
'O'), ('2005.', 'O'), ('aastal', 'O'), ('sai', 'O'), ('peaministriks', 'O'),  
('Andrus', 'B-PER'), ('Ansip', 'I-PER'), ('.', 'O')]

# Ajaväljendite tuvastamine

```
text = Text('Potsataja ütles eile, et vaatavad nüüd Genaga viie  
aasta plaanid uuesti üle.')
```

```
print (zip(text.timex_texts, text.timex_values))
```

```
[('eile', '2015-10-29'), ('nüüd', 'PRESENT_REF'), ('viie aasta',  
'P5Y')]
```



# Estnltk-s on veel asju

- ❖ Osalausestaja
- ❖ Verbiahelate tuvastaja
- ❖ Wordneti liides
- ❖ TEI korpuste lugemise liides
- ❖ Wikipedia lugemise liides
- ❖ Estnltk kihtide ilutrükk
- ❖ Grammatikapõhine infoeraldusmoodul
- ❖ Elasticu andmebaasi tugi
- ❖ Korpusepõhine ühestaja

Kasutusjuhend: <http://estnltk.github.io/estnltk/>



# Masinõppetarkvara tekstide klassifitseerimiseks

Töötab Exceli ja CSV failidega

Põhjalik tagasisidesüsteem

täpsus, saagis ja F1-skoor

olulised tunnused

probleemsed tekstid, millega tarkvara hätta jääb

Käsureatööriistad + API

Lähtekood: <https://github.com/estnltk/textclassifier>

Kasutusjuhend: <http://estnltk.github.io/estnltk/1.3/tutorials/textclassifier.html>



# Andmestik kommentaaride ja hinnangutega

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

# Tarkvara õpib hinnanguid ennustama

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	?

# Õpitud mudeli täpsus

## Classification report

### Precision Recall F1-score

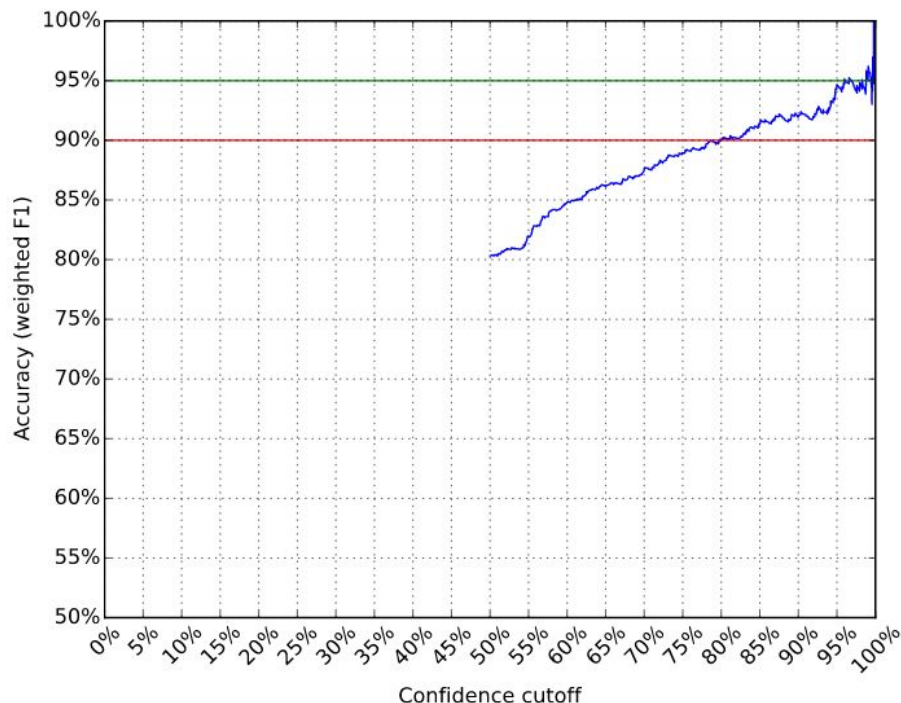
77.3      85.2    81.0

### 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





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, süski, soov, soovima, super, suur, suurepärase, 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, süski, soov, soovima, super, suur, suurepärase, 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 <b>positiivset</b> aga õnneks ka mitte negatiivset.	Positiivne
26	Sülearvutite valik suur ja <b>hinnad</b> on soodsad	Positiivne
31	Probleeme pole olnud, suhtlemine väga personaalsel tasemel. <b>Hinnad</b> 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 <b>hinnad</b> suht kallid.	Positiivne
74	Mõned asjad ostnud. Viimati ostsin HDD, aga temperatuur <b>oli</b> 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 <b>nädala</b> pärast helistada. Helistatigi 2 <b>nädala</b> pärast ja anti uus ketas. Teenindus võiks tõesti olla parem, aga kokkuvõttes täitsa norm pood.	Positiivne

...



**Siim Orasmaa**

temporal expression extractor, clause detector, verb phrase detector, advanced disambiguator

**Aleksander Tkatchenko**

named entity recognition

**Timo Petmanson**

vabamorf Python wrapper, text classification tool, grammar module

**Karl-Oskar Masing**

wordnet module

**Andres Matsin**

wikipedia module

**Annett Saarik**

database module

**Karl Valliste**

prettyprinter module

**Uku Raudvere**

future modules

**Tarmo Vaino**

vabamorf library

**Heiki-Jaan Kaalep**

vabamorf library, language engineering

**Neeme Kahusk**

eurowordnet module

**Sven Laur**

project manager & visionary

