TAITOTALO

Tekstianalyysi

Joitakin esimerkkejä

- Puheen tunnistus (speech recognition)
  - Puhutun kielen muuttaminen tekstiksi
- Konekääntäminen
- Luonnollisen kielen prosessointi (Natural Language Processing NLP)
  - esimerkiksi puheohjausjärjestelmät (kännykkä)
  - LLM (Large Language Model) Suuret kielimallit
- Chatbot
  - Joko luonnollisen kielen prosessointi tai ennalta määrätyt vastaukset
- Tunneanalyysi (sentiment analysis)
  - Onko sana negatiivinen, positiivinen vai neutraali
  - Käytetään esimerkiksi mittaamaan kuluttajien reaktiota
- Syväoppiminen
  - RNN (Recurrent Neural Network)

TAITOTALO

Tekstin prosessointi

Miten teksti muutetaan konekielelle luettavaksi ja analysoitavaksi

TAITOTALO

1) Kirjaimen muuttaminen samaan kokoon

- Tekstin muuttaminen niin, että kaikki on kirjoitettu pienellä kirjaimella

```
teksti = "Tämä On Esimerkki Isoista JA Pienistä Kirjaimista."

# Muuta kaikki kirjaimet pieniksi
yhdenmukainen_teksti = teksti.lower()

# Tulosta yhdenmukainen teksti
print("Yhdenmukaistettu teksti:\n", yhdenmukainen_teksti)
```

✓ 0.0s

```
Yhdenmukaistettu teksti:
 tämä on esimerkki isoista ja pienistä kirjaimista.
```

2) Tekstin jakaminen pienempiin merkitysyksiköihin (tokenize)

- Token on yksikkö, jonka luonnollisen kielen käsittely tunnistaa
- Mahdollistaa muun muassa taivutusmuotojen analysoinnin, esiintymistiheyden laskemisen
- Sana, numero, merkki, jne.
- Esimerkiksi "Hän meni kauppaan." voidaan jakaa merkitysyksiköiksi:
  - ’hän’, ’meni’, ’kauppaan’, ’.’

- ['Tämä', 'on', 'esimerkki', 'tekstistä', ',', 'jossa', 'on', 'paljon', 'pysäytysanoja', '.']

3) Pysäytyssanojen (stop words) poistaminen

- Sanoja, jotka eivät olennaisesti lisää informaatiota, esimerkiksi ’ja’, ’on’
- Auttaa keskittymään olennaisiin sanoihin
- Valmiita listoja

```
#löytyymyös suomeksi
stop_words = stopwords.words('english')
print(stop_words)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'wh
ich', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'd
o', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'agains
t', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'agai
n', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such',
'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll
', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'have
n', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't",
'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

```
1  stop_words = set(stopwords.words('finnish'))
2  print(stop_words)
```

✓ 0.0s

```
{'minä', 'minussa', 'ketä', 'jolta', 'keille', 'heidät', 'noita', 'teidät', 'joissa', 'meillä', 'näillä', 'hänet', 'tuona', 'heihin', 'nuo', 'mistä', 'sen', 'hänelle', 'j
```

IOT ALT
TAITOTALO

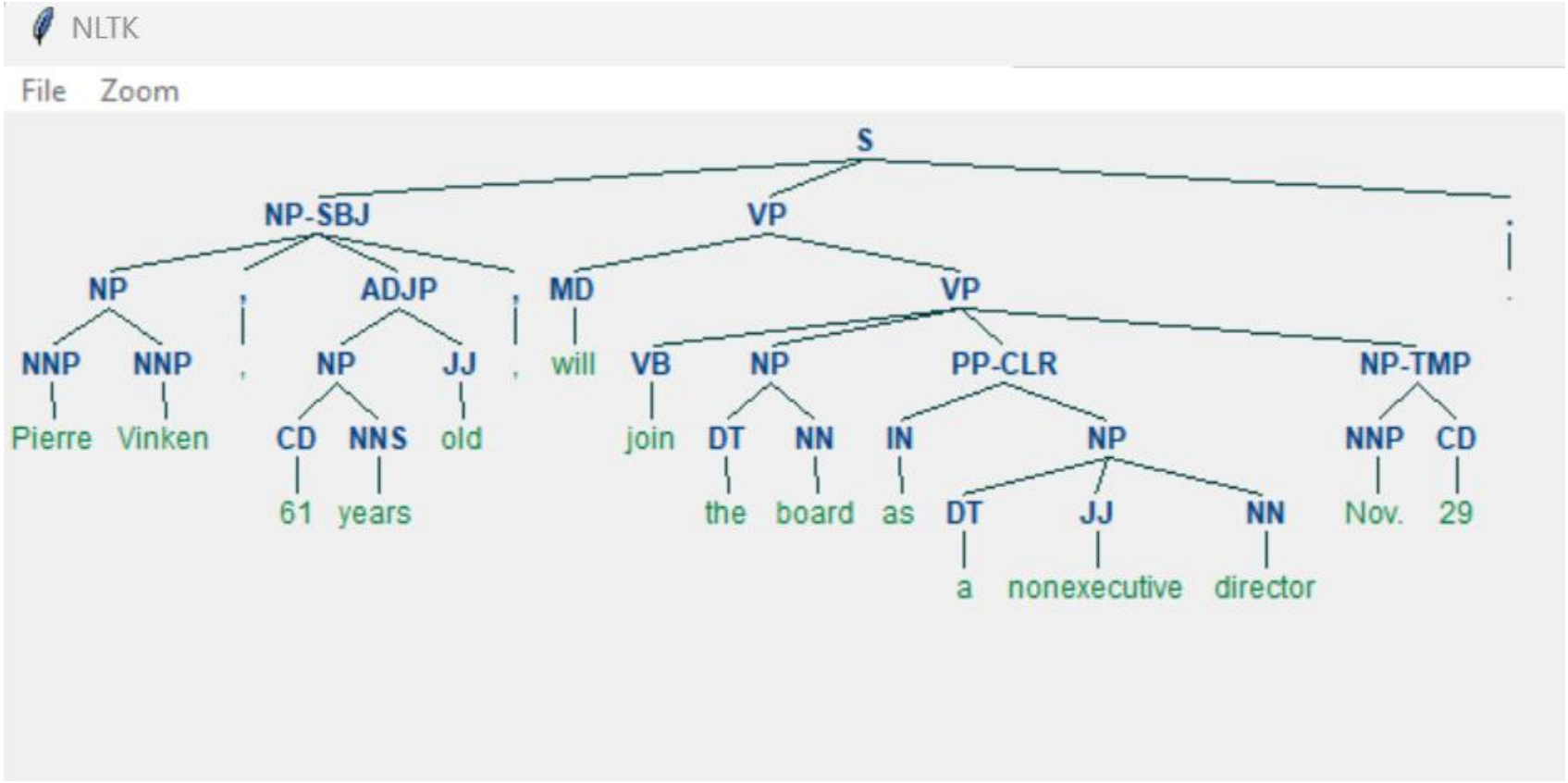
Sanan vartalon hakeminen (stemming)

- Haetaan sanojen kanta
- Suomenkielessä haastavampaa kuin englanninkielessä
- Esim. juokse-
  - Juoksen, juoksemme, juokseminen, mutta juosta tai juoksu
  - Vrt. run, running

Sanan perusmuodon palauttaminen (lemmatization)

- Sanan perusmuoto ilman taivutusta
- Vastaa ns. sanakirjamuotoa
- Esimerkki:
  - "Juoksen, juokset, juoksemme, juoksette, juoksevat”, perusmuoto on ’juosta’

Sanaluokkien tunnistus ja jaottelu



Natural Language Toolkit

- NLTK :: Natural Language Toolkit
- Natural Language Toolkit Tutorial
- https://www.nltk.org/howto/sentiment.html

spaCy-kirjasto

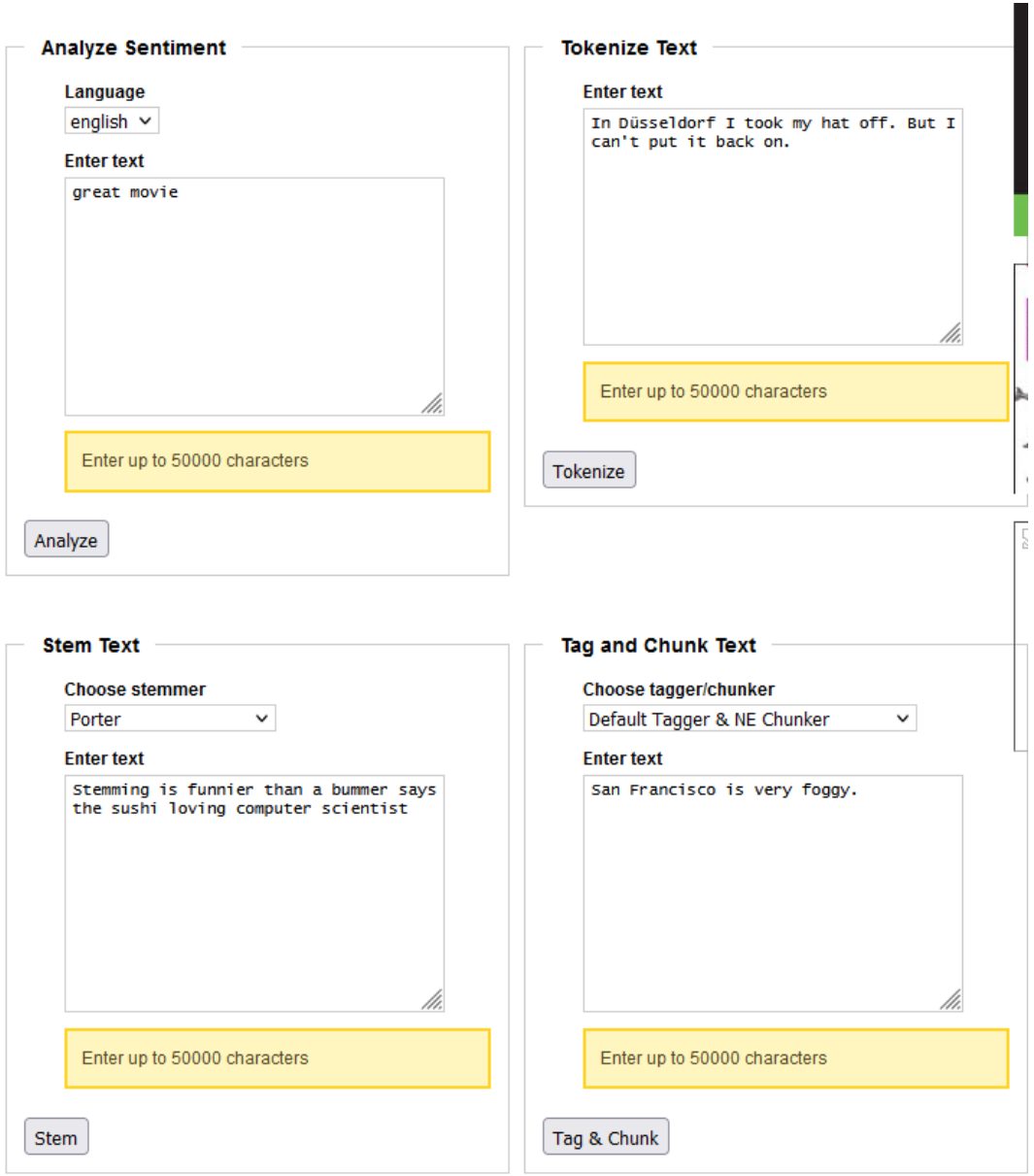
- https://spacy.io/
- https://spacy.io/models/fi
- https://applied-language-technology.mooc.fi/html/index.html
- https://applied-language-technology.mooc.fi/html/notebooks/part_ii/05_evaluating_nlp.html?highlight=sentiment
- Natural Language Processing With spaCy in Python – Real Python

Tehtävä 1

1. Tutustu demon avulla eri tekstikäsittelytyökaluihin:
   http://text-processing.com/demo/

2. Etsi teksti netistä ja kopioi se
3. Avaa valikot ja kokeile eri työkaluja ja kieliversioita
4. Tallenna kuvakaappaus kokeilustasi.
5. Mitä opit tehtävästä?

(http://snowball.tartarus.org/algorithms/finnish/stemmer.html)



TAITOTALO

Sentiment-analyysi

Positiiviset ja negatiiviset sanat

- Sanalistoja käytetään sentiment-analyysin pohjana
- Verrataan tekstistä poimittuja merkityksellisiä sanoja tunnesanalistaan
- Lasketaan tunnesanojen osuus koko tekstistä

```
#https://gist.github.com/mkulakowski2/4289437
pos = open('positive-words.txt','r').read().split()
print(pos)
```

```
['a+', 'abound', 'abounds', 'abundance', 'abundant', 'a
e', 'accomodative', 'accomplish', 'accomplished', 'acc
'achievible', 'acumen', 'adaptable', 'adaptive', 'adequ
y', 'adorable', 'adore', 'adored', 'adorer', 'adoring',
ntageous', 'advantageously', 'advantages', 'adventures
ion', 'affection', 'affectionate', 'affinity', 'affirm
le', 'agile', 'agilely', 'agility', 'agreeable', 'agree
e', 'amazed', 'amazement', 'amazes', 'amazing', 'amazi
le', 'amicability', 'amicable', 'amicably', 'amity', 'a
g', 'applaud', 'appreciable', 'appreciate', 'appreciate
```

```
#https://gist.github.com/mkulakowski2/4289441
neg = open('negative-words.txt','r').read().split()
print(neg)
```

```
['2-faced', '2-faces', 'abnormal', 'abolish', 'abominable', '
upt', 'abruptly', 'abscond', 'absence', 'absent-minded', 'abs
'abysmal', 'abysmally', 'abyss', 'accidental', 'accost', 'acc
'acerbic', 'acerbically', 'ache', 'ached', 'aches', 'achey',
'adamantly', 'addict', 'addicted', 'addicting', 'addicts', 'a
d', 'adulteration', 'adulterier', 'adversarial', 'adversary',
'aggravating', 'aggravation', 'aggression', 'aggressive', 'ag
e', 'agonizing', 'agonizingly', 'agony', 'aground', 'ail', 'a
d', 'alienation', 'allegation', 'allegations', 'allege', 'all
'ambivalent', 'ambush', 'amiss', 'amputate', 'anarchism', 'an
```

Tehtävä 2

- Hae joku teksti netistä esimerkiksi käyttämällä BeautifulSoup-kirjastoa(web scraping). Vaihtoehtoisesti voit ladata tiedot csv-tiedostoon ja lukea sen.
- Esimerkiksi Free eBooks | Project Gutenberg
- Prosessoi teksti analysoitavaan muotoon
- Tee sanoille tunneanalyysi
- Pohdi ryhmässä mitä johtopäätöksiä analyysista voi tehdä ja mitä opit tehtävästä

(https://jss367.github.io/getting-text-from-project-gutenberg.html)