

MSiA 490 SEC20

Special Topics: Text Analytics

Wang, Timo

The background features a series of concentric circles in light gray, some solid and some dashed, creating a ripple effect. A large red speech bubble is centered on the page, containing the text 'Lab 1 - Tokenization'.

Lab 1 - Tokenization

A red speech bubble graphic with a white outline, pointing downwards. The word "What" is written in white inside the bubble.

What

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo erat dolores et ea. Stet clita kasd tempor gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

A red speech bubble graphic with a white outline, pointing downwards. The word "What" is written inside in white.

What

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo erat dolores et ea. Stet clita kasd tempor gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



What

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo erat dolores et ea. Stet clita kasd tempor gubergren, no sea takimata est Lorem ipsum dolor sit amet.

A large red speech bubble graphic with a white outline, pointing downwards. The word "Why" is written in white inside the bubble.

Why

- Identification of key information
 - Medical terms within a electronic prescription
- Embedding of content
 - Word-level/sentence-level embeddings



How

- NLTK
- spaCy
- StanfordNLP

The NLTK logo is a red speech bubble with the text "NLTK" in white. It is positioned on the left side of the slide. The background features decorative curved lines in the top-left and bottom-right corners.

NLTK

```
pip install nltk
```

```
ipython
```

```
>> import nltk
```

```
>> nltk.download()
```


Sentence Tokenizer

```
from nltk.tokenize import sent_tokenize
```

```
text = 'I am happy. I am sleepy. I am dreamy.'
```

```
sent_tokenize(text)
```

```
['I am happy.', 'I am sleepy.', 'I am dreamy.']
```

Word Tokenizer

```
from nltk.tokenize import word_tokenize
```

```
text = 'I am happy. I am sleepy. I am dreamy.'  
word_tokenize(text)
```

```
['I', 'am', 'happy', '.', 'I', 'am', 'sleepy',  
'.', 'I', 'am', 'dreamy', '.']
```

The SpaCy logo consists of a red speech bubble shape with a small tail pointing downwards. The word "spaCy" is written in white inside the bubble, with "spa" in lowercase and "Cy" in uppercase.

spaCy

```
pip install spacy
```

Sentence Tokenizer

```
from spacy.lang.en import English
```

```
text = 'I am happy. I am sleepy. I am dreamy.'
```

```
nlp = English()
```

```
nlp.add_pipe(nlp.create_pipe('sentencizer'))
```

```
doc = nlp(text)
```

```
[sent.string.strip() for sent in doc.sents]
```

```
['I am happy.', 'I am sleepy.', 'I am dreamy.']
```

Word Tokenizer

```
from spacy.lang.en import English
```

```
text = 'I am happy. I am sleepy. I am dreamy.'
```

```
nlp = English()
```

```
doc = nlp(text)
```

```
[token.text for token in doc]
```

```
['I', 'am', 'happy', '.', 'I', 'am', 'sleepy',  
'.', 'I', 'am', 'dreamy', '.']
```

StanfordNLP

```
pip install stanfordnlp
```

```
ipython
```

```
>> import stanfordnlp
```

```
>> stanfordnlp.download(,en')
```

Sentence Tokenizer

```
import stanfordnlp
```

```
text = 'I am happy. I am sleepy. I am dreamy.'
```

```
nlp = stanfordnlp.Pipeline()
```

```
doc = nlp(text)
```

```
[' '.join([token.text for token in  
sentence.tokens]).strip() for sentence in  
doc.sentences]
```

```
['I am happy .', 'I am sleepy .', 'I am dreamy  
.']]
```

Word Tokenizer

```
import stanfordnlp
from functools import reduce

text = 'I am happy. I am sleepy. I am dreamy.'
nlp = stanfordnlp.Pipeline()
doc = nlp(text)

words_by_sentence = [[token.text for token in
sentence.tokens] for sentence in doc.sentences]

reduce(lambda lst1, lst2: lst1 + lst2,
words_by_sentence)

['I', 'am', 'happy', '.', 'I', 'am', 'sleepy',
'.', 'I', 'am', 'dreamy', '.']
```


The background features a series of concentric circles in light gray, some solid and some dashed, creating a ripple effect. A large, solid red speech bubble is centered on the page, pointing downwards. The text "Setting Things up" is written in white, sans-serif font inside the bubble.

Setting Things up

A red speech bubble with a white outline, containing the text 'Log into Cluster/Open Terminal'. The bubble has a tail pointing towards the bottom left. The background is white with faint, curved, dashed lines in the top left and bottom right corners.

Log into
Cluster/Open
Terminal

Install Conda

<https://docs.conda.io/en/latest/miniconda.html>

Create an Environment

```
conda create -n py36 python=3.6
```

```
conda create -n py27 python=2.7
```



Install Libraries

Get the Data

- <https://www.kaggle.com/crawford/20-newsgroups#sci.med.txt>



Have Fun!