MSiA 490 SEC20 Special Topics: Text Analytics

Wang, Timo

Lab 1 - Tokenization



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.



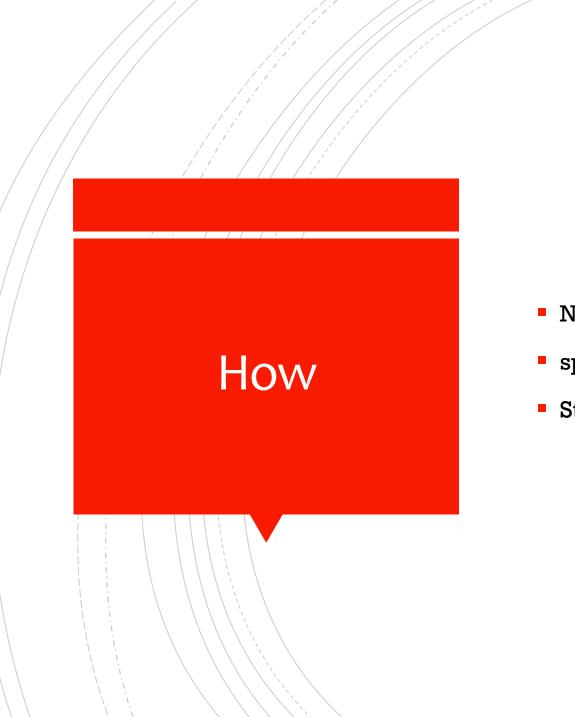
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo erat dolores et ea. Stet clita kasd tempor gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



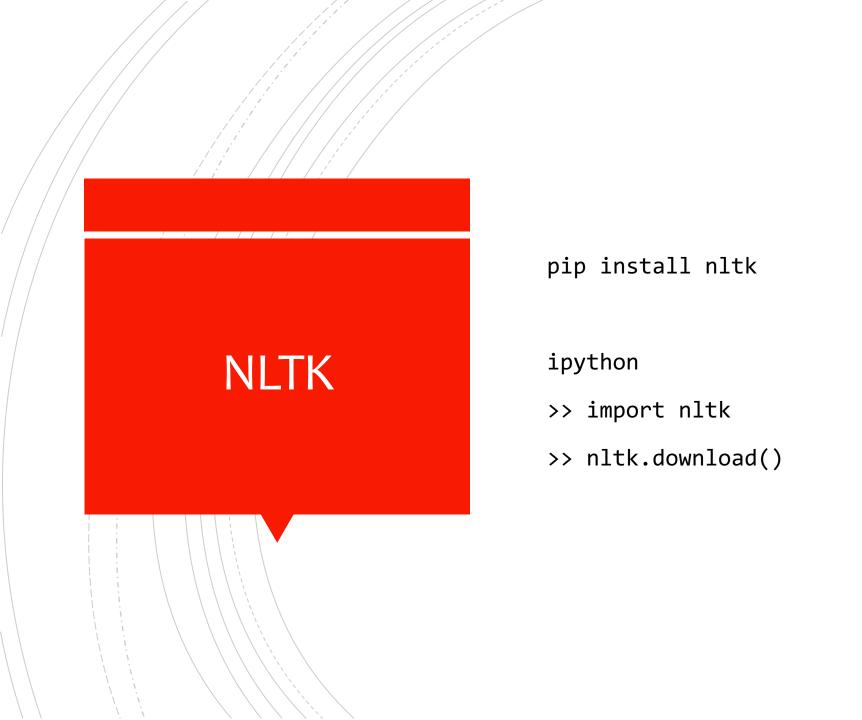
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo erat dolores et ea. Stet clita kasd tempor gubergren, no sea takimata



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



- NLTK
- spaCy
- StanfordNLP



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', '.']
```



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', '.']
```



Log into Cluster/Open Terminal



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



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

