MSiA490 SEC20/28
Special Topics: Text Analytics
Lab 1 - Tokenization

Timo Wang

Northwestern University

September 17, 2020

What is tokenization?

Raw text

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.

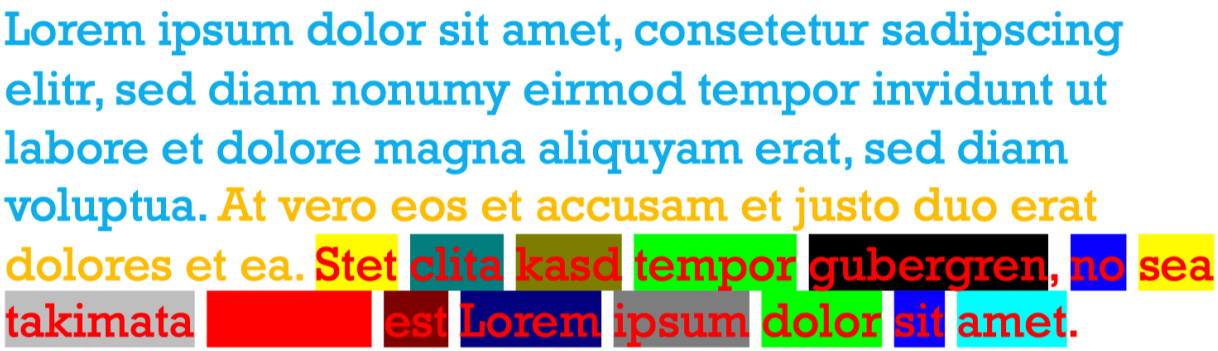
What is tokenization?

Sentence-level 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.

What is tokenization?

Word-level tokenization



How to do tokenization?

Overview

- NLTK
- Spacy
- Stanford Stanza

How to do tokenization?

NLTK – Installation

Shell

```
pip install nltk
```

Python Console

```
import nltk
nltk.download()
```

How to do tokenization?

NLTK – Sentence-level

Python Console

```
from nltk.tokenize import sent_tokenize

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

How to do tokenization?

NLTK – Word-level

Python Console

```
from nltk.tokenize import word_tokenize

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

How to do tokenization?

Spacy – Installation

Shell

```
pip install spacy
python -m spacy download en_core_web_sm
```

How to do tokenization?

Spacy – Sentence-level

Python Console

```
import spacy

nlp = spacy.load('en_core_web_sm')
nlp.add_pipe(nlp.create_pipe('sentencizer'))

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

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

How to do tokenization?

Spacy – Word-level

Python Console

```
import spacy

nlp = spacy.load('en_core_web_sm')
nlp.add_pipe(nlp.create_pipe('sentencizer'))

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

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

How to do tokenization?

Stanford Stanza – Installation

Shell

```
pip install stanza
```

Python Console

```
import stanza
stanza.download('en')
```

How to do tokenization?

Stanford Stanza – Sentence-level

Python Console

```
import stanza

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

nlp = stanza.Pipeline('en')
doc = nlp(text)
sents = [' '.join([token.text for token in sentence.
    tokens]).strip() for sentence in doc.sentences]
```

How to do tokenization?

Stanford Stanza – Word-level

Python Console

```
from functools import reduce

import stanza

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

nlp = stanza.Pipeline('en')
words_by_sentence = [[token.text for token in sentence.
    tokens] for sentence in doc.sentences]
words = reduce(lambda lst1,lst2: lst1 + lst2,
    words_by_sentence)
```

Extra resources

- Miniconda: https://docs.conda.io/en/latest/miniconda.html.
- NLTK: https://www.nltk.org/
- Spacy: https://spacy.io/usage/spacy-101
- Stanford Stanza: https://stanfordnlp.github.io/stanza/


HANDS ON

Quiz

Task 1

Which library do you find easiest to use for tokenization?

A NLTK

B Spacy

C Stanford Stanza

D Other

Quiz

Task 2

Which library runs fastest for POS tagging?

A NLTK

B Spacy

C Stanford Stanza

D Other

Quiz

Task 3

Which library appears most memory efficient on your machine/OS of choice?

A NLTK

B Spacy

C Stanford Stanza

D Other