Social Media & Text Analysis

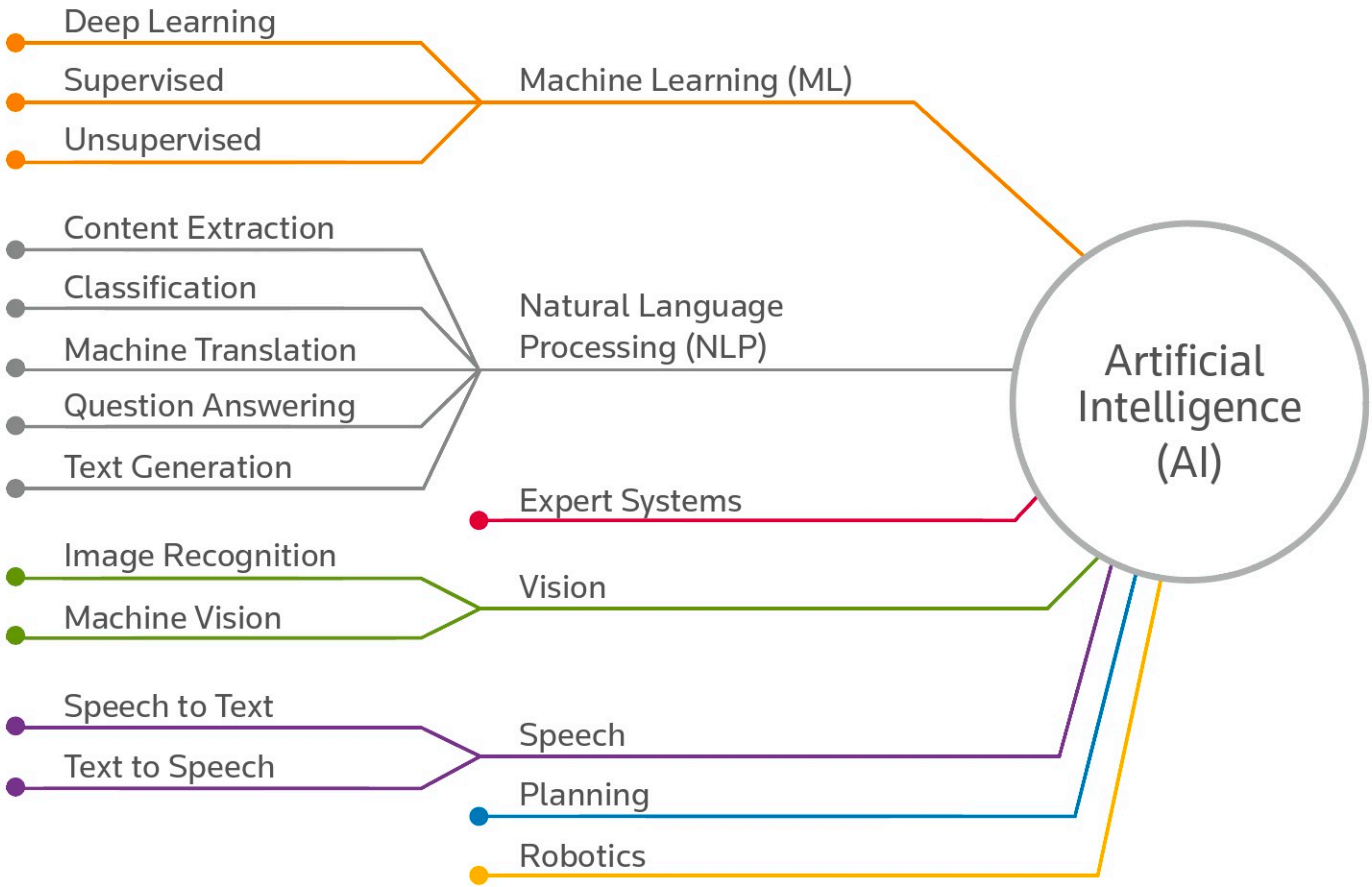
part 2 - Intro to NLP

Follow @cocoweixu

Ohio State University

Instructor: Wei Xu

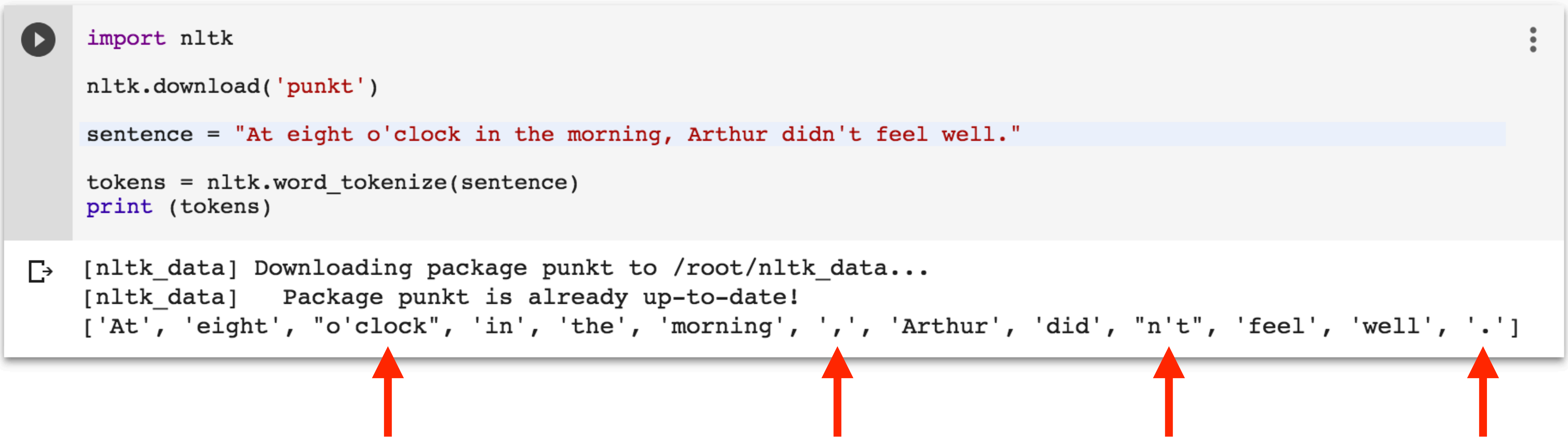
Website: socialmedia-class.org



Source: http://www.fullai.org/short-history-artificial-intelligence/

Basic Text Processing

- Tokenization:



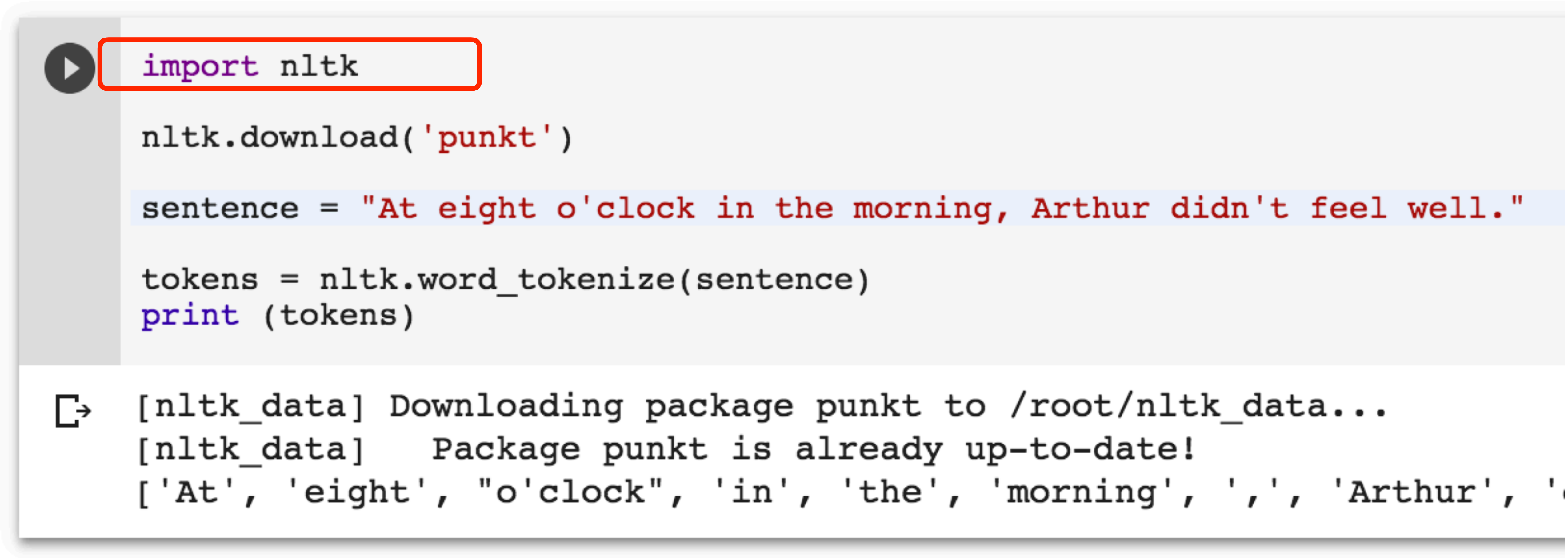
```
import nltk

nltk.download('punkt')

sentence = "At eight o'clock in the morning, Arthur didn't feel well."

tokens = nltk.word_tokenize(sentence)
print (tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
['At', 'eight', "o'clock", 'in', 'the', 'morning', ',', 'Arthur', 'did', "n't", 'feel', 'well', '.']
```

breaking text up into words, phrases, symbols, or other meaningful elements called tokens.

Basic Text Processing

- Tokenization:



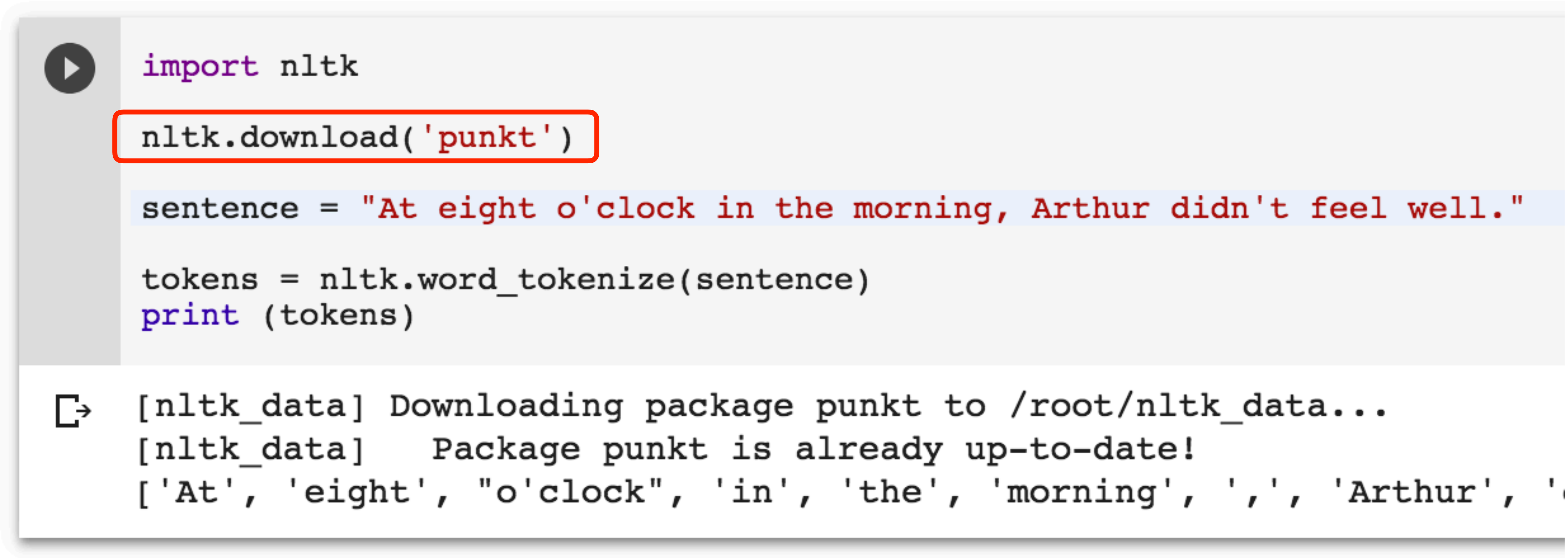
```
import nltk

nltk.download('punkt')

sentence = "At eight o'clock in the morning, Arthur didn't feel well."

tokens = nltk.word_tokenize(sentence)
print (tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
['At', 'eight', "o'clock", 'in', 'the', 'morning', ',', 'Arthur', '
```

To start using Python NLTK (Natural Language Toolkit) library.

Basic Text Processing

- Tokenization:



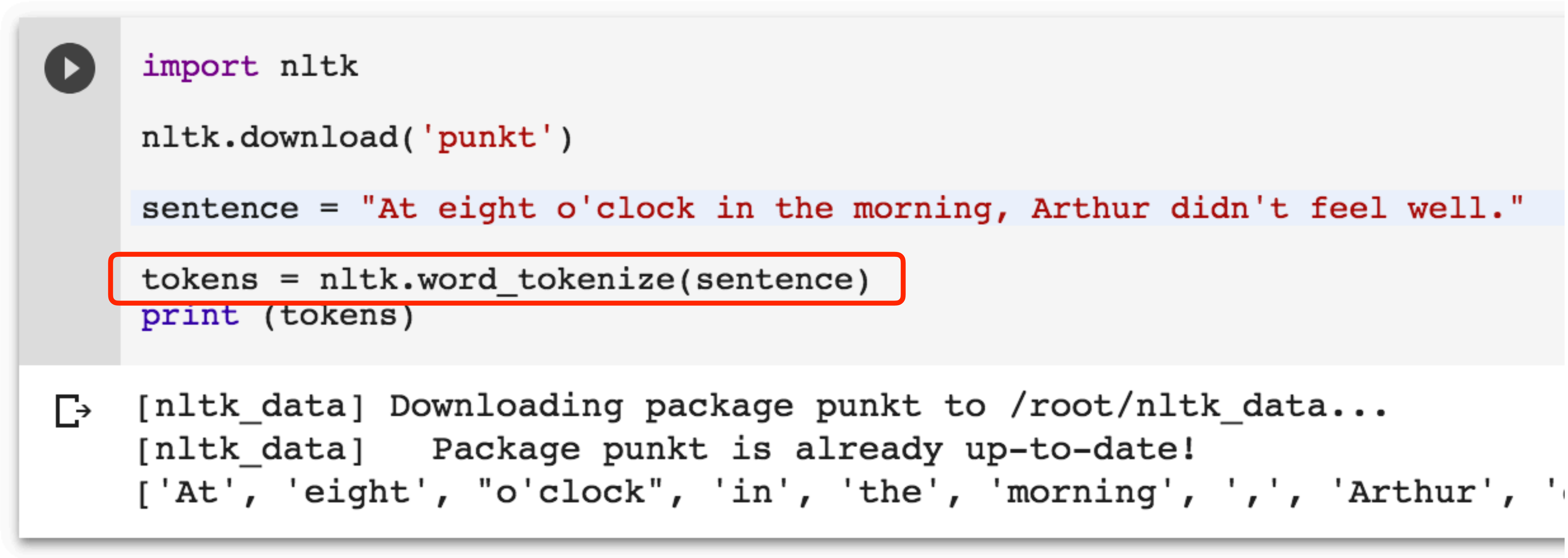
```
import nltk
nltk.download('punkt')
sentence = "At eight o'clock in the morning, Arthur didn't feel well."

tokens = nltk.word_tokenize(sentence)
print (tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
['At', 'eight', "o'clock", 'in', 'the', 'morning', ',', 'Arthur', '
```

Load in a pre-trained tokenizer for English named “Punkt”.

Basic Text Processing

- Tokenization:



```
import nltk

nltk.download('punkt')

sentence = "At eight o'clock in the morning, Arthur didn't feel well."

tokens = nltk.word_tokenize(sentence)
print (tokens)
```

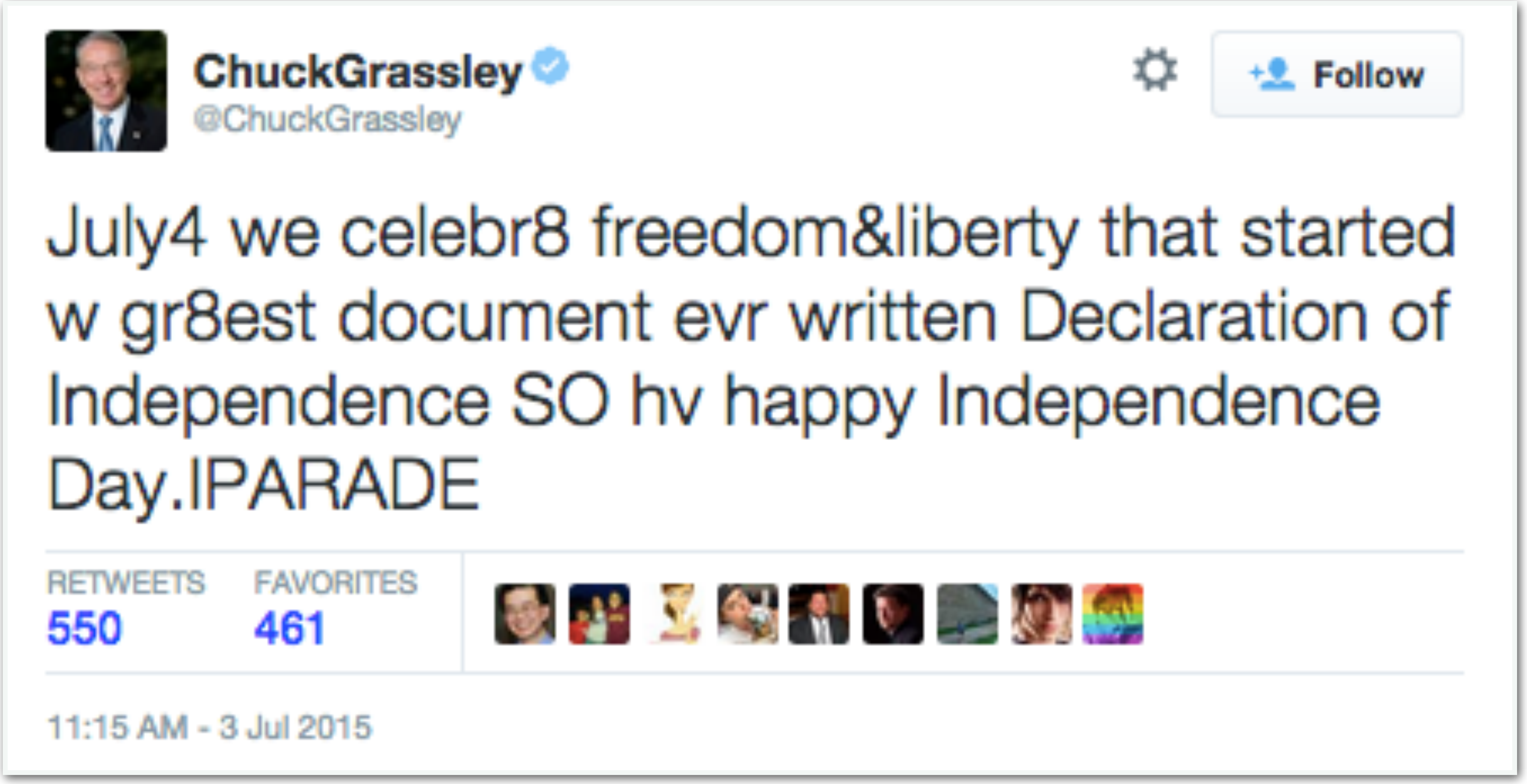
```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
['At', 'eight', "o'clock", 'in', 'the', 'morning', ',', 'Arthur', '
```

Calling the **word_tokenize()** function in the **nltk** module.

Try it Out!

- We will use **Google's Colab** programming environment:

What if we try to tokenize some tweets?



ChuckGrassley
@ChuckGrassley

July4 we celebr8 freedom&liberty that started w gr8est document evr written Declaration of Independence SO hv happy Independence Day.IPARADE

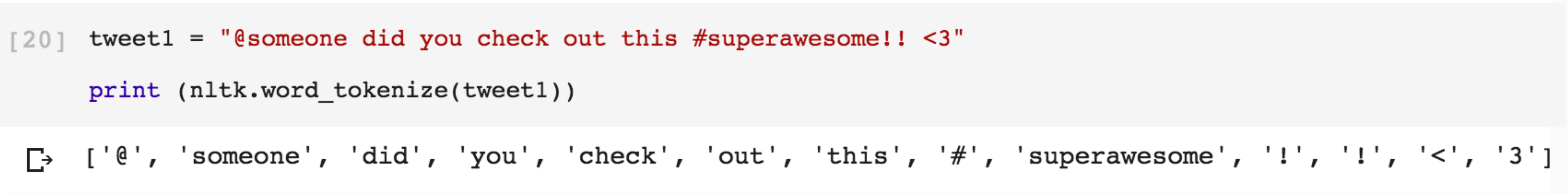
RETWEETS 550
FAVORITES 461

11:15 AM - 3 Jul 2015

Try it Out!

- We will use **Google's Colab** programming environment:

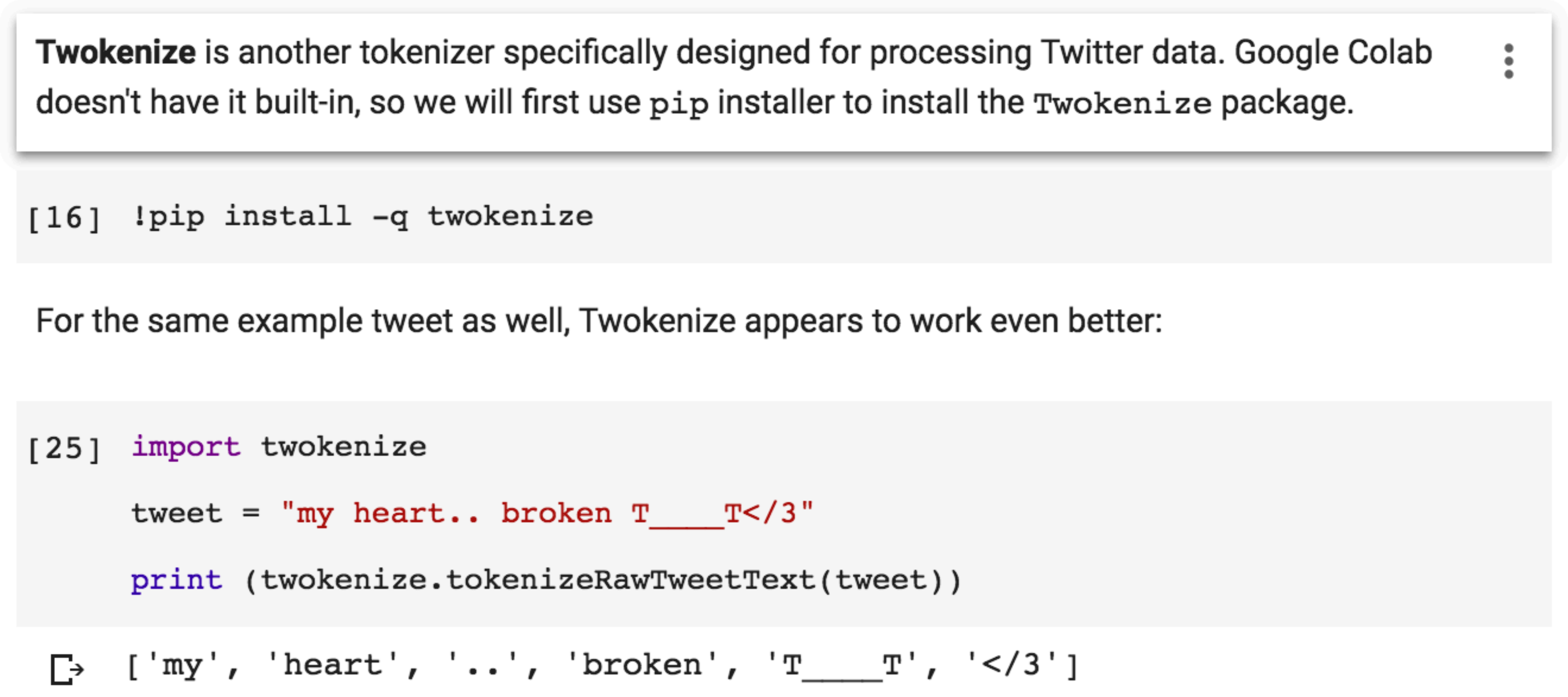
What if we try to tokenize some tweets?



```
tweet1 = "@someone did you check out this #superawesome!! <3"
print (nltk.word_tokenize(tweet1))
```

```
['@', 'someone', 'did', 'you', 'check', 'out', 'this', '#', 'superawesome', '!', '!', '<', '3']
```

Twitter-specific Tokenizer



Twokenize is another tokenizer specifically designed for processing Twitter data. Google Colab doesn't have it built-in, so we will first use `pip` installer to install the `Twokenize` package.

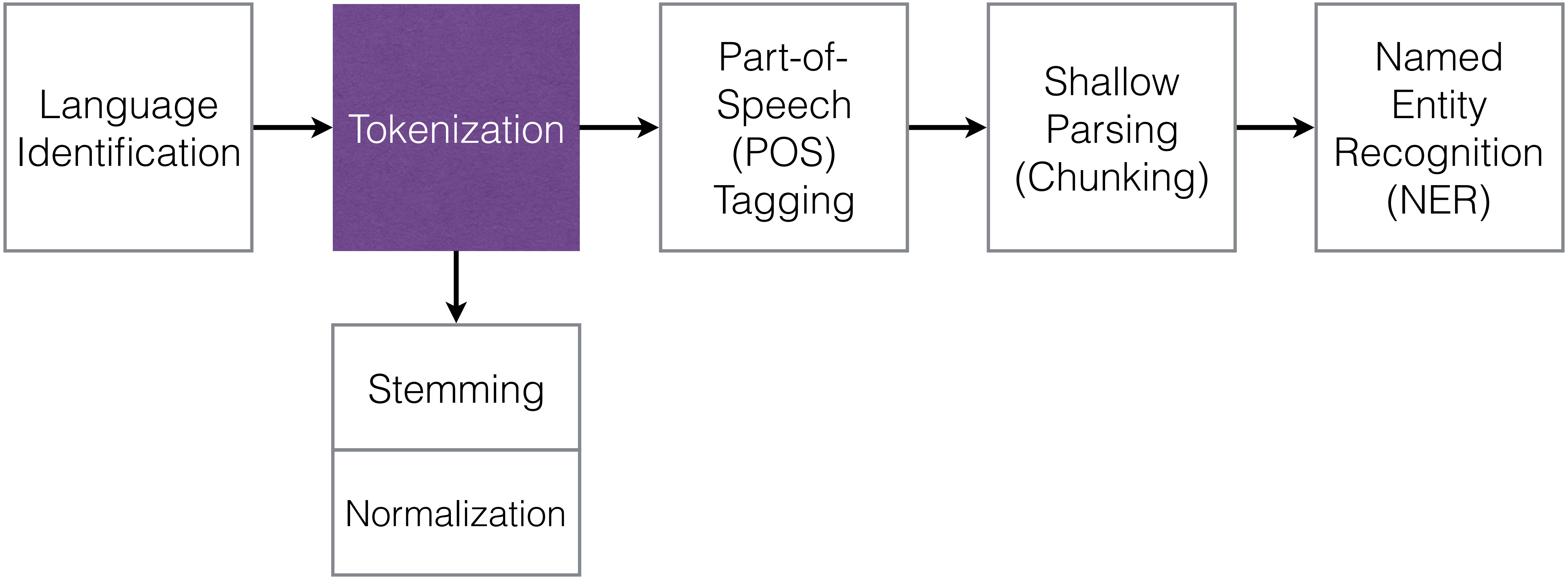
```
[16] !pip install -q twokenize
```

For the same example tweet as well, Twokenize appears to work even better:

```
[25] import twokenize
     tweet = "my heart.. broken T____T</3"
     print (twokenize.tokenizeRawTweetText(tweet))
```

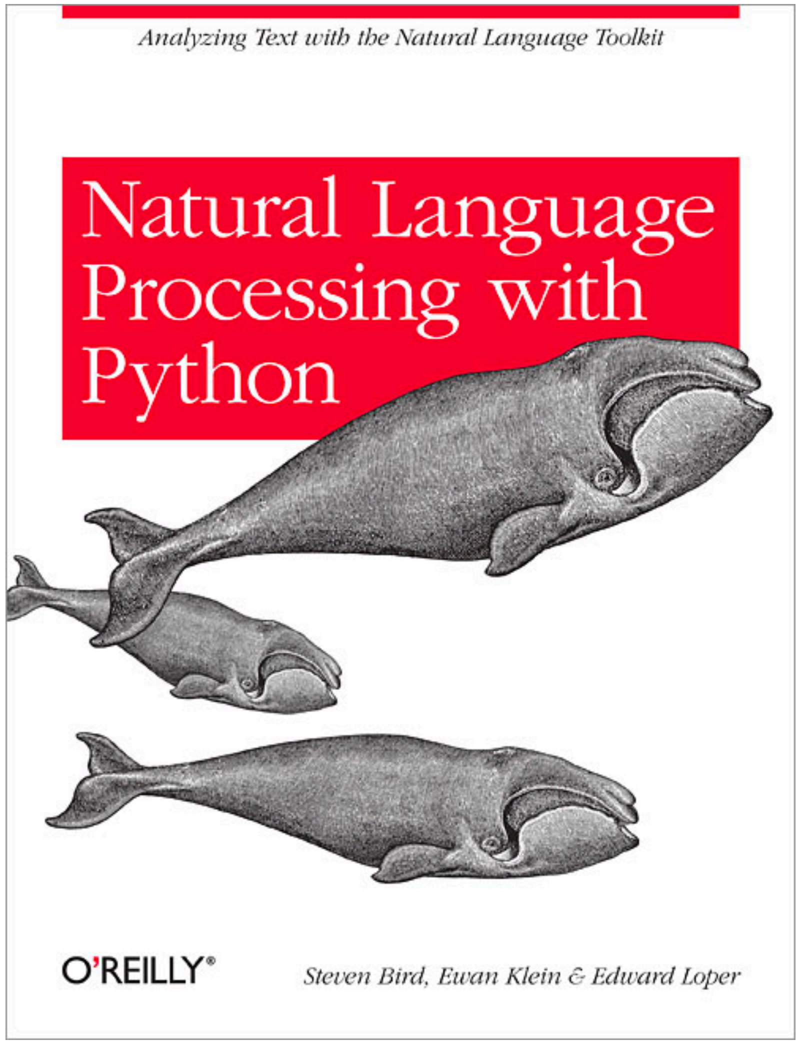
```
['my', 'heart', '..', 'broken', 'T____T', '</3']
```

NLP Pipeline


Language Identification
Tokenization
Part-of-Speech (POS) Tagging
Shallow Parsing (Chunking)
Named Entity Recognition (NER)
Stemming
Normalization

More Resources

- NLTK (Natural Language Toolkit):
  - NLTK Book: http://www.nltk.org/book — free online!

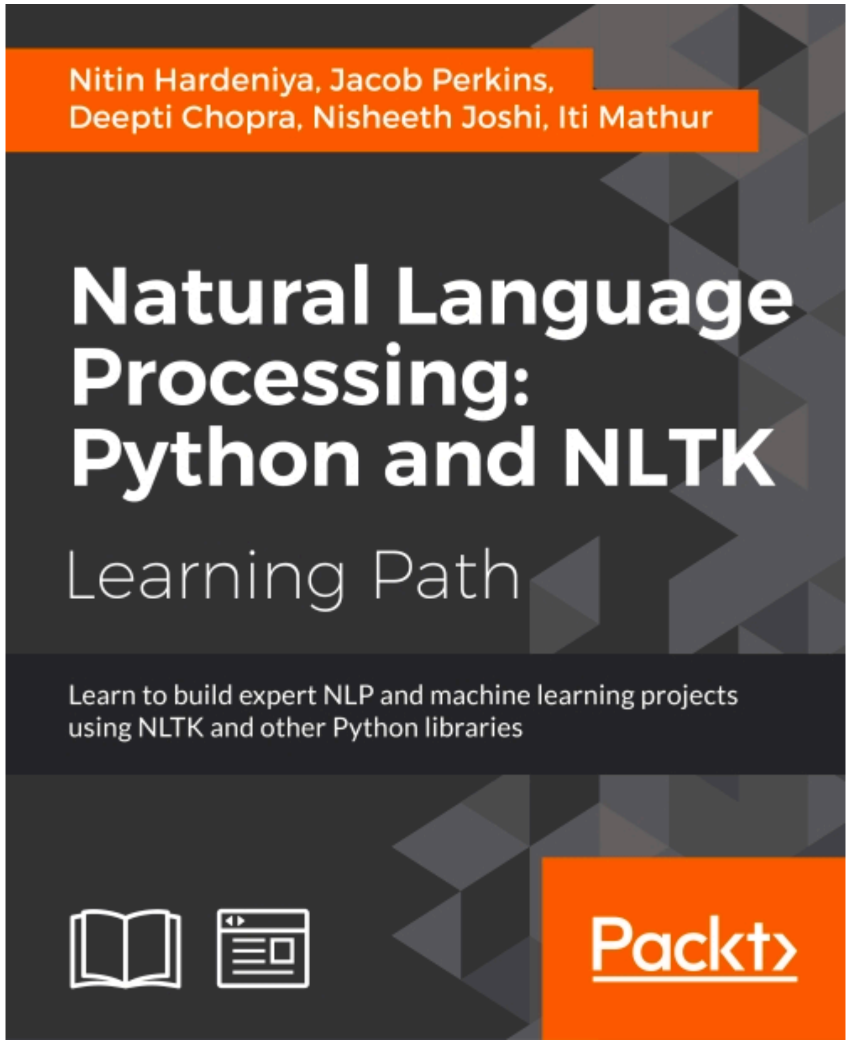


Analyzing Text with the Natural Language Toolkit

Natural Language Processing with Python

O'REILLY®

Steven Bird, Ewan Klein & Edward Loper



Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, Iti Mathur

Natural Language Processing: Python and NLTK

Learning Path

Learn to build expert NLP and machine learning projects using NLTK and other Python libraries

Packt›