Python for Data Science

Natural Language Processing: nltk

Dr. Ilkay Altintas

Describe what natural language processing (NLP) means

List a few examples of NLP in everyday life

Explain what NLTK is

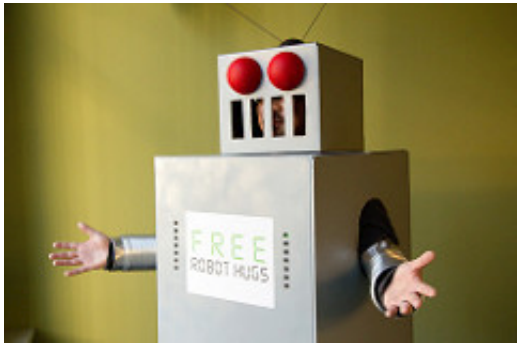
Intro to Natural Language Processing (NLP)

Algorithms to analyze, understand and derive meaning from human language

Hard computational problem because human language is **ambiguous**, needs **context** and ability to **link concepts**

Applications: summarize text, generate keywords, identify sentiment of text

Real Life examples of NLP



Speech recognition engines like Siri, Google Now or Alexa

Automatic translation like Google Translate or Facebook automatic translation of statuses

Chat bots that can answer question via Facebook Messenger, for example provided by Techcrunch, Disney or Whole Foods

nltk

Natural Language Toolkit in Python

Work with human language data
Includes over 50 datasets

- Complete library of easy to use algorithms for processing text
- Available for free under open source license

http://nltk.org

Natural Language Processing: nltk corpora

- Describe what corpus means
- List some datasets in the nltk corpora
- Recite the basic features of the movie reviews corpus in nltk

nltk corpora

corpus (plural corpora) is a collection of text in digital form, assembled for text processing

nltk provides a **download interface** to pre-processed text datasets.

```
nltk.download()
```

```
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader> l

Packages:
  [ ] abc..................... Australian Broadcasting Commission 2006
  [ ] alpino.................. Alpino Dutch Treebank
  [ ] averaged_perceptron_tagger Averaged Perceptron Tagger
  [ ] averaged_perceptron_tagger_ru Averaged Perceptron Tagger (Russian)
  [ ] basque_grammars..... Grammars for Basque
  [ ] biocreative_ppi..... BioCreAtIvE (Critical Assessment of Information
                           Extraction Systems in Biology)
  [ ] bllip_wsj_no_aux.... BLLIP Parser: WSJ Model
  [ ] book_grammars....... Grammars from NLTK Book
  [ ] brown................... Brown Corpus
  [ ] brown_tei........... Brown Corpus (TEI XML Version)
  [ ] cess_cat............ CESS-CAT Treebank
  [ ] cess_esp............ CESS-ESP Treebank
  [ ] chat80.............. Chat-80 Data Files
  [ ] city_database....... City Database
  [ ] cmudict............. The Carnegie Mellon Pronouncing Dictionary (0.6)
  [ ] comparative_sentences Comparative Sentence Dataset
  [ ] comtrans............ ComTrans Corpus Sample
  [ ] conll2000........... CONLL 2000 Chunking Corpus
  [ ] conll2002........... CONLL 2002 Named Entity Recognition Corpus

Hit Enter to continue:
```

nltk movie reviews corpus

nltk.download("movie_reviews")

```
~ altintas$ ls nltk_data/corpora/movie_reviews

    README neg pos
```

2000 files:

1000 positive reviews in the pos/ folder
1000 negative reviews in the neg/ folder

nltk movie reviews corpus

```
nltk.download("movie_reviews")
```

2000 files:

1000 positive reviews in the pos/ folder
1000 negative reviews in the neg/ folder
average 800 words per review

Natural language processing: tokenize

Explain what Tokenization means
Use nltk word tokenizer

Tokenization

The first step in analyzing text is to split it into words: Tokenization

Corner cases:

- punctuation
- contractions
- hyphenated words

Example: "New York-based"

First Attempt without nltk

Naively just split on whitespace

See **Tokenize text in words**

Tokenize with nltk

nltk.word_tokenize

Sophisticated tokenizer specific to English,
it requires the *punkt* corpus.

It correctly identifies also punctuation.

Natural language processing: build a bag-of-words model

Explain what bag-of-words mean

Understand how you can build machine learning features from words

Give examples of stopwords

Bag-of-words Model

Bag-of-words = text as unordered collection of words

simple model
discards sentence structure
useful to identify topic or sentiment

Building Features with Words

	outstanding	movie	family	worse	uninvolving	interesting
Review 1	True	True	False	False	False	False
Review 2	False	True	False	True	True	False
Review 3	True	True	True	False	False	False

Filter out Stopwords and Punctuation

The movie_reviews tokenized words also include punctuation and stopwords.

Stopwords are very common words that have no intrinsic meaning like "the", "is", "which".

Natural Language Processing: Plotting Frequency of Words

Count how many times an item appears in a list
Plot word frequency in logarithmic axes
Plot word counts histograms

Number of Words in Movie Reviews Corpus

~1.6 million words

just 710 thousand after filtering punctuation and stopwords

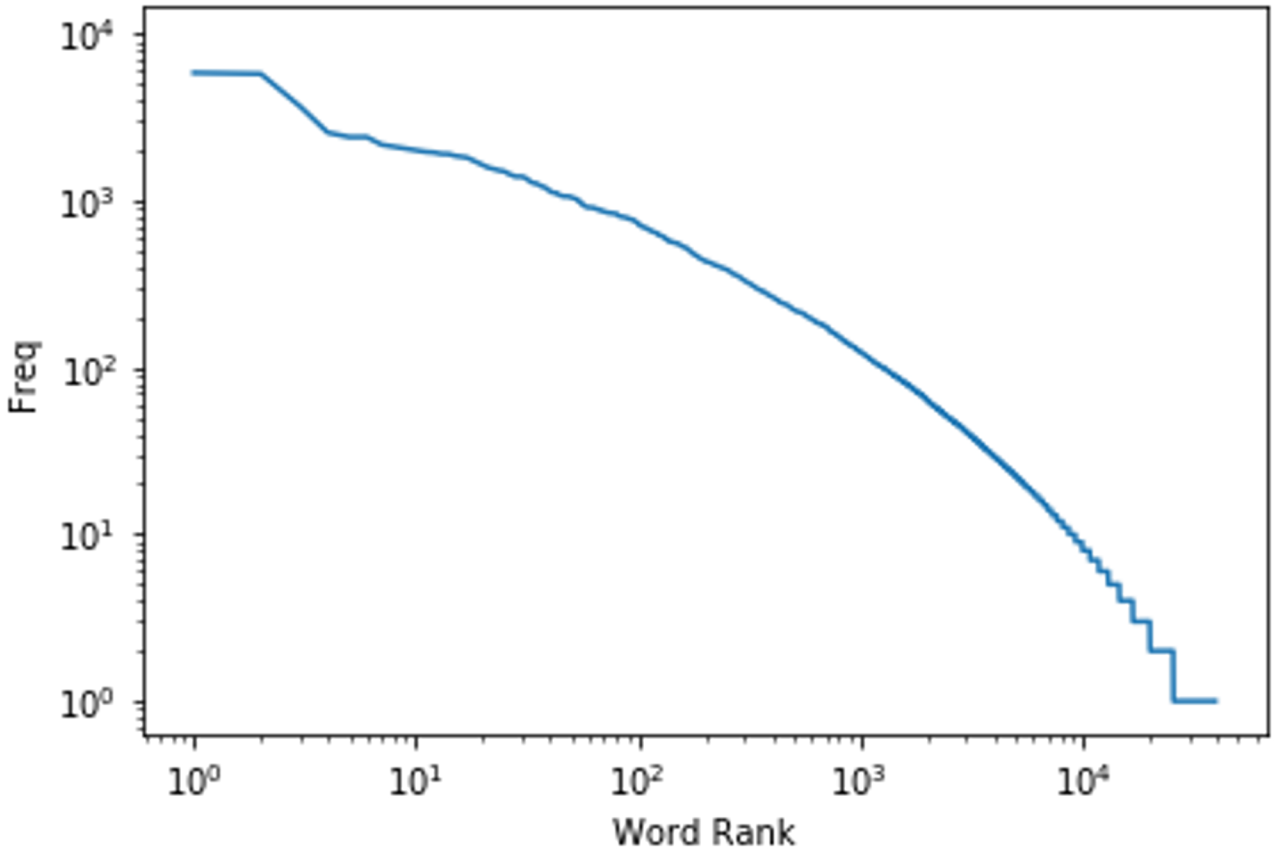
Using Counter

Part of the collections package in the Python Standard Library

Counts how many time an item is repeated

```
counter = Counter(filtered_words)
counter["movie"]
5771
```

Plotting Word Frequency



Histogram of Word Counts

Use hist from matplotlib to create a histogram

Choose bin number and optionally log axes

Python for Data Science

Natural Language Processing: Sentiment Analysis

Explain what is Sentiment Analysis
Train a Sentiment Analysis classifier with nltk
Check accuracy on training and test data

What is Sentiment Analysis

Identify attitude or emotion encoded in a text

Can be implemented as a Machine Learning Classifier

Example: prediction on the appearance of words in a review

Build features/label pairs

The function implemented previously creates a set of features.

Create a pair of feature and positive/negative label for each review.

Naive Bayes Classifier

Naive Bayes Classifier is a simple classifier based on Conditional Probabilities.

In the training phase, it detects the probability that each feature (word) appears in a category (positive or negative).

Once trained, it collects the "votes" for all words in the new review and finds the most probable label.