Text analysis

- What? Why

encode text data into numeric variables-empirical analysis

Common Sources of Text Data: Corporate Annual Reports, News and Social Media, Government and Legal Documents

A set of documents is called a corpus (plural corpora)

- Text Cleaning and Preprocessing
  - Tokenize: parse documents into smaller units, such as words or phrases

Tokenization

- Split up a document into tokens
- Common tokens
  - Words: e.g., "hello", "blueno", "laptop", etc.
  - Punctuation: e.g., . , " ' ! ?, etc.
- Other tokens
  - Replace very uncommon words with an unknown token: <UNKNOWN>
  - End sentences (or sentence like structures) with a stop token: <STOP>
  - Replace all numbers with a single token: e.g., 100 → <NUM>
  - Replace common words ("a", "the", etc.) with <SWRD>
- Tokenization is a pain (there are lots of edge cases), but luckily, it is a solved problem
- "The dog ran in the park joyously!" →
  c("<SWRD>", "dog", "ran", "<SWRD>", "<SWRD>", "park", "<UNKNOWN>", "!", "<STOP>")

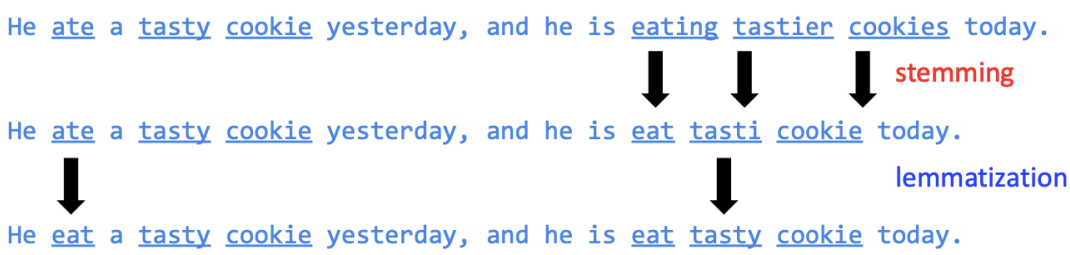
  - Remove stop words (e.g., a, the, and etc.)

Stop words: those that appear frequently in text analysis but contribute less to understanding the meaning of a sentence

  - Stemming & Lemmatization: standardize words with similar meaning

Stemming & Lemmatization

- Goal: standardize words with a similar meaning
- Stemming reduces words to their base, or root, form
- Lemmatization makes words grammatically comparable
  (e.g., am, are, is ⟹ be)



  - Normalize

Normalization

Examples:

- make all words lowercase
- remove any punctuation
- remove unwanted tags

Image Source

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

```
Has Dr. Bob called? He is waiting for the results of Experiment #6.
                    ⬇
has dr bob called he is waiting for the results of experiment 6

<p>Text.</p><!-- Comment --><a href="#breakpoint">More text.</a>'
                    ⬇
text more text
```

- Descriptive Text Analysis
  - Word frequency statistics
  - Word Cloud Analysis

Visualises words in a document with sizes proportional to how frequently the words are used.

- Sentiment Analysis

  Classifies a document as expressing a positive, negative, or neutral opinion

  Especially useful for analyzing reviews or social media posts

Positive vs. Negative Words

Researchers have built lists of words with "positive" and "negative" connotations

A+ Acclaim Accomplish Accurate Achievement Admire ...	Abnormal Abolish Abominable Abominate Abort Abrasive ...

For each chunk of our own text, we can calculate how many words lie in these "positive" or "negative" groupings

```
I love all the delicious free food in the CIT, but working in the Sunlab makes me sad.
```

We can also add common Internet slang to lists of "positive" and "negative" words: e.g, "luv", "yay", "ew", "wtf"

- Topic modeling

Topic models allow us to cluster similar documents in a corpus together. They require no prior information, training set, or labelling of texts before estimation.

  - LDA model

  Latent Dirichlet Allocation (LDA) model

  a probabilistic language model

  Idea:

  The number of topics, K, is selected by the researcher

  Each of the K topics is a probability distribution (a Dirichlet distribution) over a fixed vocabulary of N words

  Each of the D documents is a probability distribution (a Dirichlet distribution) over the K topics

  Each word in each document is drawn from the topic-specific probability distribution (multinomial distribution) over words

  Procedures:

  1. Determine the proportion of topics in the article (Dirichlet distribution).
  2. For each word, select the topic first, then select the word within the topic (multinomial distribution).
  3. Reverse inference, inferring hidden themes from documents.

Reference:

https://cs.brown.edu/courses/cs100/lectures/lecture24.pdf
https://lse-me314.github.io/lecturenotes/ME314_day11.pdf