Natural Language Processing

Orevaoghene Ahia

Lecture 2: Text Preprocessing 1

Lecture Plan

Lecture 2: Text Preprocessing 1

Goal : Understand how to text preprocessing methods and useful libraries

What did we treat last week ?

Any questions from last week’s class ?

What is text ?

- We can think of text as a sequence of :
  - Characters
  - Words
  - Phrases
  - Sentences
  - Paragraphs

Text Preprocessing Techniques

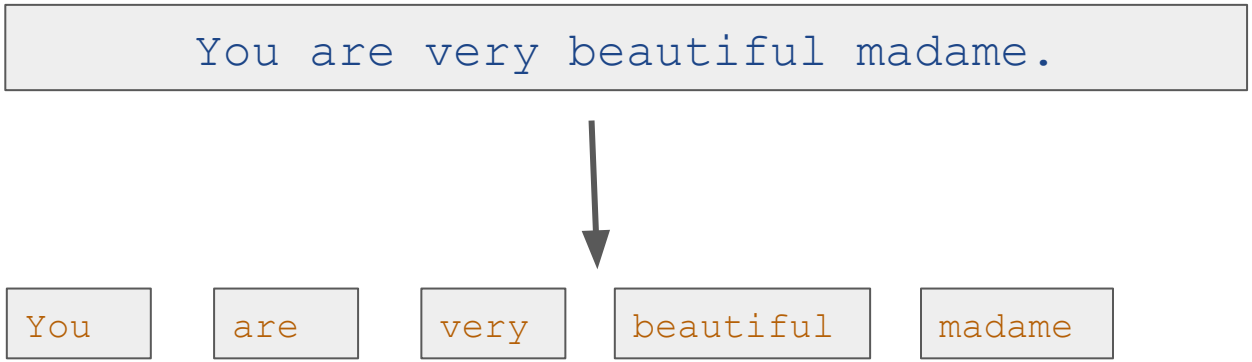
- Converting text to a meaningful format for analysis
- Preprocessing and cleaning text

NLP Toolkits

- NLTK (Natural Language Toolkit)
  - Python library for handling natural language processing (NLP) tasks
- TextBlob
  - Python library for processing textual data, wraps around NLTK and Pattern
- spaCy
  - Built in cython, so it’s fast
- Gensim
  - Great for topic modelling and document similarity

Tokenization

- Tokens are the basic units of text involved in NLP.
- Tokenization is the process of splitting text into smaller pieces called **tokens.**
- Words, numbers, punctuation marks, and others can be considered as tokens



Issues in Tokenization

- Finland’s Capital → Finland Finlands Finland’s ?
- What’re , I’m, isn’t → What are, I am, is not ?
- Helwett-Packard → Helwett Packard ?
- State-of-the-art → State of the art ?
- San Francisco → One token or two ?
- m.p.h , Ph.D → ? ? ?

Tokenization: Language issues

French

- L'ensemble → one token or two?
  - L?L'? Le?
  - Want L'ensemble to match with un ensemble.

German noun compounds are not segmented

- Wohnungsreinigung
- "House cleaning"
- German information retrieval systems need compound splitter.

Chinese and Japanese have no space between words.

- 你很美丽

Case Folding

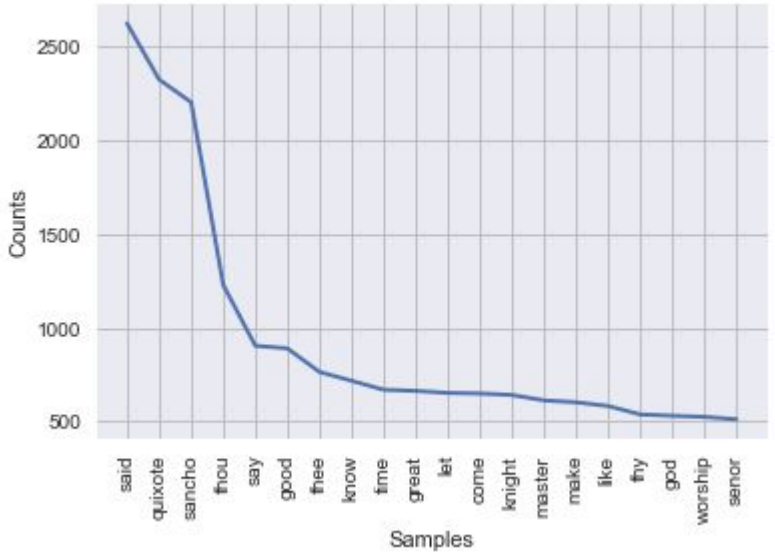
- Applications like Information Retrieval: Reduce all letters to lowercase
- Possible exception: Upper-case in mid-sentences. For Instance:
  - FED vs fed
  - SAIL vs sail
  - General Motors vs general motors
- For Sentiment Analysis, Machine Translation and Information Extraction,
  - Case is helpful. (US is different from us)

Frequency Distribution

- Imagine we want to identify words in a text that are most informative about the topic of the text.
- Counting words or sentences appearing in your text.

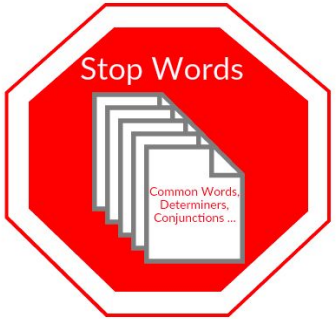
Word Tally

the	𝍸 𝍸 𝍸 ||||
been	𝍸 𝍸 |
message	||||
persevere	|
nation	𝍸 |||



Stop Words Removal

- Stop words can be referred as :
  - Words that have high frequency in a corpus
  - Words empty of true meaning given a context.



With stop-words	Without stop-words
Listening is exhausting	Listening, exhausting
It is a social science	Social, science

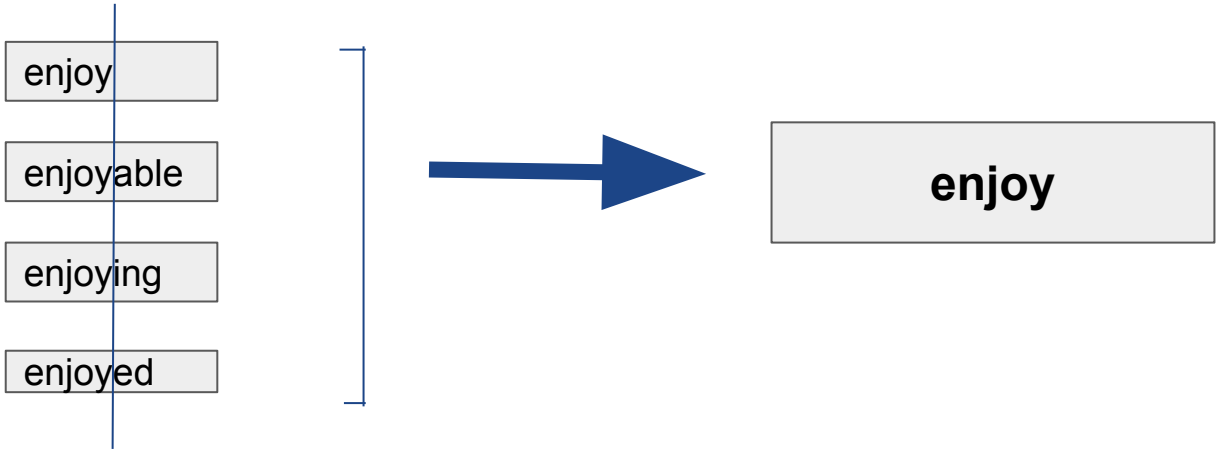
- Words such as articles and some verbs are usually considered stop words because they don’t help us find the context or true meaning of a sentence.
- They can be removed without any negative consequences to your final model.
- You should predefine your own stop words with respect to your corpus.

When should you remove stopwords ?

- If they don’t add any new information to your problem
- Classification problems usually don’t need stop words because you can talk about the general idea of a text even without stop words .
- It is better to keep stop words and do some tests with and without them to see how it affects our model.
- You should never remove stop words without thinking about their impact on the problem you are trying to solve.

Stemming

- Reducing a word to a stem or base form



The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is *Porter's algorithm* (Porter, 1980)

Stemming

- Stemming algorithms are rule based.
- They involve a heuristic process that lops off the end of words.
- A word is looked at and run through a series of conditionals that determine how it is cut down.

Overstemming

- Overstemming comes from when too much of a word is cut off.
- This can result in nonsensical stems, where all the meaning of the word is lost or muddled.
- Can result in words being resolved to the same stems, even though they probably should not be.
- For instance :
  - University, universal, universities, universe
- A stemming algorithm will resolve this four words to the stem “univers”
- A better resolution is to have “university and universities” stemmed together and

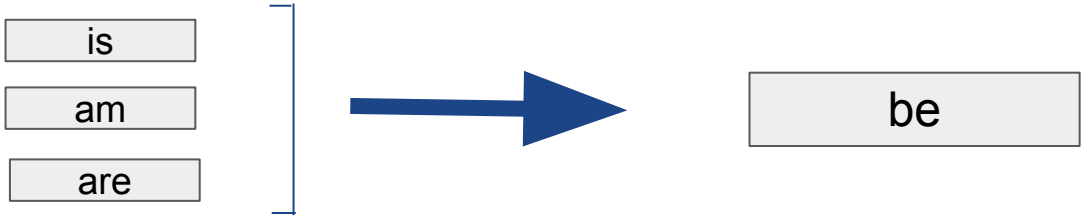
  “Universal and universe” stemmed together.

Understemming

- The opposite of overstemming.
- As a result of having several words that are forms of one another.
- Would be nice for them all to the same stem, but unfortunately, they do not.
- For instance:
  - data will be stemmed to “dat” and datum will be stemmed to “datu”
  - What about date ?

Lemmatization

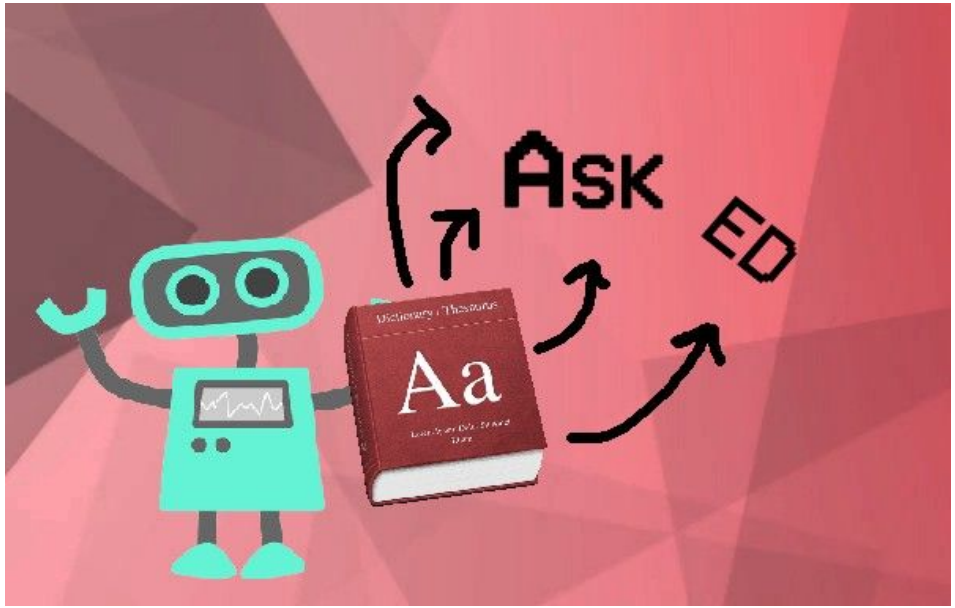
- Reducing words to their root word and transforms the root word using vocabulary and morphological analysis.
- More sophisticated than stemming



- While **stemming** works on an individual word without knowledge of the context, **lemmatization** requires a dictionary look-up.

Lemmatization

- Its harder to create a lemmatizer in a new language than a stemmer.
- Because lemmatizers require a lot more knowledge about the language structure .



When not to use ...

- Lemmatization and stemming are useful when we have limited data and a simple model.
- When we use complex models like neural networks, its not advisable to use lemmatization nor stemming
- Simple models cannot really learn complex things so they require some special data preprocessing before hand.

Summary

- Text Data is usually messy .
- Preprocessing should be done before analysis.
- There are many other preprocessing Techniques

Next Week

- Assignment 1; explore other preprocessing techniques befor class.
- Mini-Project 1
- Other preprocessing techniques.
- We might introduce Text Representation.