NLP Section 3

Text Preprocessing

Preprocessing

- Data preprocessing is the first step in any data analysis or machine learning pipeline

- Cleaning, transforming and organizing raw data to ensure that it is ready for modeling

Text Cleaning

- Text cleaning depends on the context, sometimes you remove certain components of the text and sometimes not.
- Typically involves cleaning the following:
  1- Symbols
  2- Numbers
  3- Punctuations
  4- Extra White spaces

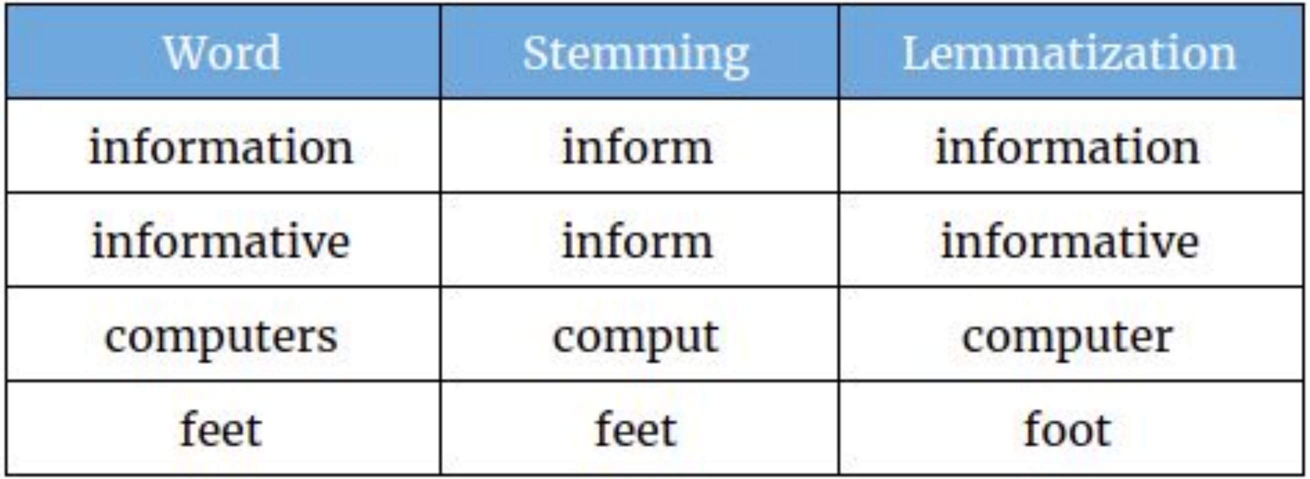
Spell Checking

- Spell checking is the process of correcting miswritten text typically using a dictionary

- Most famous tools:
  1- Textblob
  2- Symspell (superior)

Normalization

- Text normalization involves either stemming or lemmatization
- Lemmatization is almost always better than stemming



Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

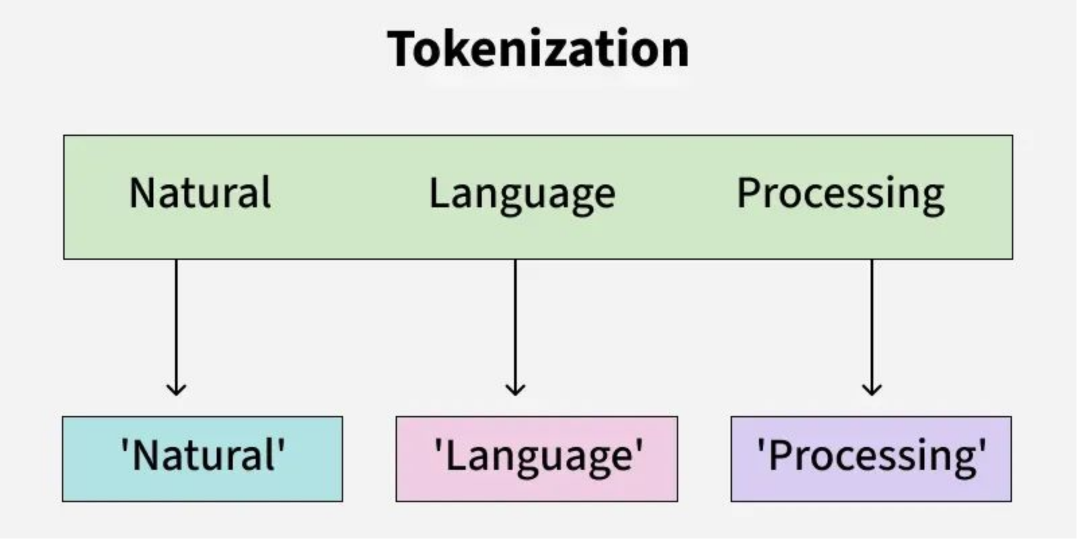
Stopwords

- Stopwords such as:
  - a
  - of
  - on
  - I

  and so on.. are typically not important for analysis purposes.

Tokenization

- The process of splitting text into tokens
- A token can be a subword or a combination of words depending on the technique



Code

Practice and showing the difference between preprocessed text and non-processed text and how it affects the outcome