Text Mining and NLTK in Python

Text mining is the process of deriving high quality information from text. The overall goal is to turn the text into data for analysis, via application of Natural Language Processing (NLP).

The **NLTK library** is the natural language toolkit for building Python programs to work with human language data and it also provides an easy to use interface.

Tokenization is the first step in NLP. It is the process of breaking strings into tokens which in turn are small structures or units. Tokenization involves three steps which are 1) *breaking a complex sentence into words*, 2) *understanding the importance of each word* with respect to the sentence and 3) *producing a structural description of an input sentence*.

Stemming usually refers to normalizing words into its base form or root form. For example, the words *waited, waiting* and *waits*. The root word is 'wait'.

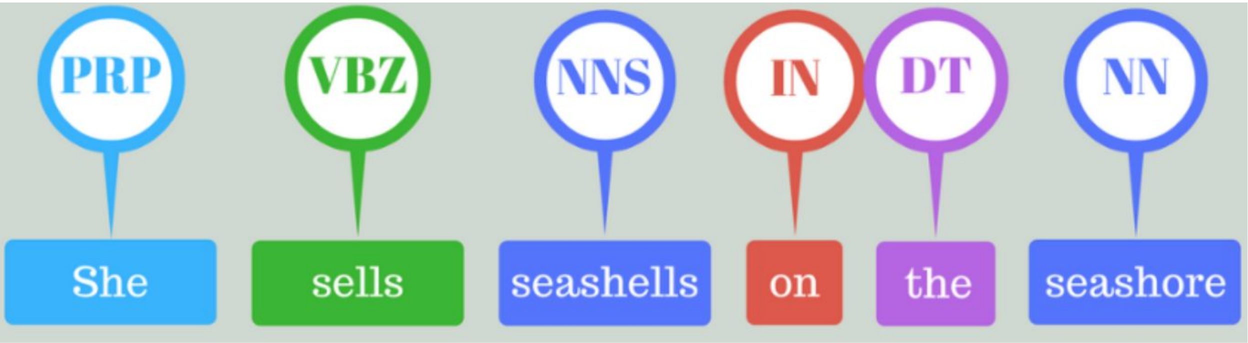
There are **two main methods in Stemming**, Porter Stemming (removes common morphological and inflectional endings from words) and Lancaster Stemming (a more aggressive stemming algorithm).

In simpler terms, **Lemmatization** is the process of converting a word to its *base form.* The difference between stemming (above) and lemmatization is, lemmatization considers the context and converts the word to its meaningful base form, whereas stemming *just removes the last few characters, often leading to incorrect meanings and spelling errors*.

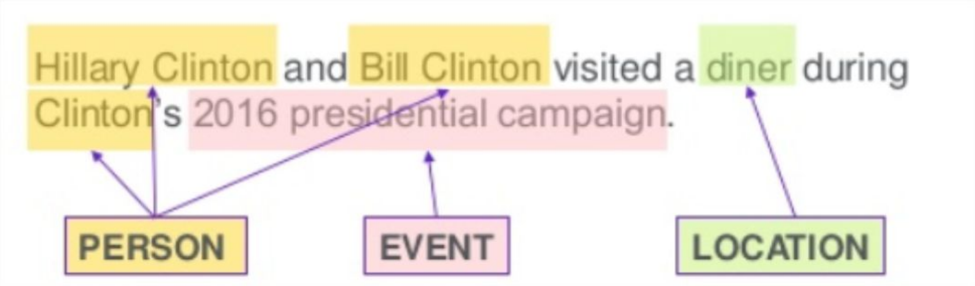
In reality, **stemming** works more often and faster, but **lemmatization** is good if the program knows the root word for what you have input.

Stop words are the most common words in a language ("the", "a", "at", "for", "above", "on", "is", "all"). These words do not provide any meaning and are usually removed from texts. We can remove these stop words using the nltk library, too.

Part-of-speech tagging is used to assign parts of speech to each word of a given text (such as nouns, verbs, pronouns, adverbs, conjunction, adjectives, interjection) based on its definition and its context. There are many tools available for POS taggers and some of the widely used taggers are NLTK, Spacy, TextBlob, Standford CoreNLP, etc.



Named entity recognition (which we do not use) is the process of detecting the named entities such as the person name, the location name, the company name, the quantities and the monetary value.


Hillary Clinton and Bill Clinton visited a diner during Clinton's 2016 presidential campaign.
PERSON
EVENT
LOCATION

Ref: Sujit Pal

Chunking (which we also do not use) means picking up individual pieces of information and grouping them into bigger pieces. In the context of NLP and text mining, chunking means a grouping of words or tokens into chunks.

Sources:

- Text Mining in Python: Steps and Examples
- Part-of-Speech Tagging in NLTK