**Preprocessing of text data**

In any machine learning task, cleaning or preprocessing the data is as important as model building if not more. And when it comes to **unstructured** data like text, this process is even more important.

Some of the common text preprocessing / cleaning steps are:

* Lower casing
* Removal of Punctuations
* Removal of Stopwords
* Removal of Frequent words
* Removal of Rare words
* Stemming
* Lemmatization
* Removal of emojis
* Removal of emoticons
* Conversion of emoticons to words
* Conversion of emojis to words
* Removal of URLs
* Removal of HTML tags
* Chat words conversion
* Spelling correction

These are the different types of text preprocessing steps. But we do not need to do all of these all the times. We need to carefully choose the preprocessing steps based on our use case.

For example, in sentiment analysis, we do not remove the emojis or emoticons as it will convey some important information about the sentiment. Similarly we need to decide based on our use cases.

**Preprocessing with spacy:**

**python -m spacy download en\_core\_web\_sm**

First, you need to download the default model for the English language: en\_core\_web\_sm

**import spacy**

**nlp = spacy.load("en\_core\_web\_sm")**

If these lines run without any errors, then it means that spaCy was installed and that the models and data were successfully downloaded.

**Lemmatization**

A lemma is the base form of a token. The lemma of walking, walks, walked is walk. Lemmatization is the process of reducing the words to their base form or lemmas.

**Stemming**

Stemming refers to reducing a word to its root form. The stem does not have to be a valid word at all. Stemming algorithms remove affixes (suffixes and prefixes). For example, the stem of “university ”is “univers”.

**Homework:**

Write a function that performs the following preprocessing steps on a chosen text entry:

* Lower casing
* Removal of Punctuations
* Removal of Stopwords
* Removal of Frequent words
* Removal of Rare words
* Stemming
* Lemmatization

The text you use must be composed of several sentences.