Preprocessing Methods

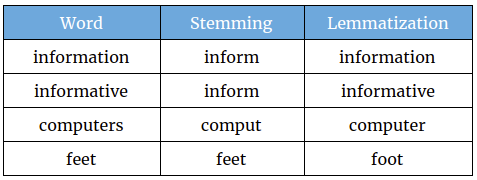
Data preprocessing required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

**Methods of Preprocessing Data:**

* Converting characters to lower cases
* Tokenizing using NLTK’s (Natural Language ToolKit) Tokenizer
* Removing any punctuation (“!” ”?” “,” “.”) or URL
* Correcting any misspelling
* Removing stop words (Common words -> “the”, “is”, “and”)
* Expanding contractions. Such as -> “don’t”, “won’t” etc…
* Removing non-English words with SpaCy. Such as -> “Déjà vu”, ”Eureka”, etc… (Might be removed)
* Lemmatization ( Stemming can be also used as an alternative)

Lemmatization: Returns to the word’s original root

Stemming: Removes or stems the last few characters of a word



**lemmatization provides better results** by performing an analysis that depends on the word's part-of-speech and producing real, dictionary words. As a result, lemmatization is harder to implement and slower compared to stemming.

* Removing emojis
* Limiting each typing length to approximately 50 characters