**Text preprocessing approaches**

Text, often, do not have the structure of the data that we normally encounter for data such as tables with relations defined between them. This text can be challenging to process because of some issues such as grammatical errors in the text, incorrect spellings, abbreviations to name a few. Even when there are no mistakes that I just mentioned, there can by synonyms and homographs (words that share same spelling but different meanings eg: number – more numb OR numerical value). Because of these reasons, the text must undergo a significant preprocessing before we can use it as an input for the data mining algorithms. Few of the preprocessing approaches are bag of words, N-gram sequences, Named entity Extraction. Terminology, common to all of these approaches is

Document – Piece of text. For eg: a Facebook post, reddit comment

Corpus – a collection of documents

Token or terms – a documents contain tokens or terms.

**Bag of words:** In this approach, each document is considered as a collection of individual words and treats each word as a potentially important keyword of the document. It ignores the order of words, grammar, and structure of sentences. Each word is considered as a token and every document as represented as a one or zero based on the availability of the token in the document. The next step is to identify the frequency of the token in the document as in some applications importance of a token increases as the frequency increases. Some pre-processing steps such as normalizing the case, stemming (for eg: cooks, cooking, cooked are all reduced to cook) and removal of stop words (the, and, of and on)

**N-gram sequences:**In this approach, the sequence of adjacent words is considered as terms which is called a bi-gram. Bag of n-gram up to 3 means that the document is represented as individual words, two adjacent words together and adjacent word triples. These N-grams are particularly useful when the phrases are of significance instead of the individual words. Eg: exceed\_analyst\_expectation can be more significant in a story that the individual words exceed, analyst, expectation. Main advantage of N-gram sequence is that we do not need to have any complex parsing algorithms or any linguistic knowledge, but it has some disadvantages such that it can drastically increase the size of the feature set and it can quickly go out of hand. Data mining using this approach may need special considerations to storage space.

**Named Entity Extraction:**Phrase extraction sometimes needs to be more sophisticated where we would like to extract common entities in the documents such as United States of America, New York Yankees, Game of Thrones. The above two approaches may not be able to capture the significance of these phrases and some pre-processing needs to be done to understand what sequences contain proper names. Normalization also needs to be done so that phrases like United States of America, U.S.A., U.S. are all linked to United States of America. This process is knowledge intensive and needs training and sometimes hand coding.