**Stemming**

Stemming, in Natural Language Processing (NLP), refers to the process of reducing a word to its word stem that affixes to suffixes and prefixes or the roots

**Word Embedding**

Numerical representation (specifically Vector) representation of words. More similar words have more similar vectors, thus similar words will be more close to each other than those which are not so similar.

**Techniques for Word Embedding:**

**One Hot Encoding** [used for representing each unique word through their position in the document] [Insight: frequency of words in a document]

If there are N number of total words and K number of unique words in a corpus or document, then each of the unique words will be represented by a vector of N length. In this N, in the places where a unique word appears those positions will be marked by an indicator value like 1.

**Word2Vec** [Insights: can group similar words: can capture words meaning based on the occurrence in the corpus]

As machines can’t work with words like humans do, every word in a context or document is encoded into a vector [as a vector is represented by numbers and machines are good with numbers]

**Bag of Words (BOW):** [used for representing each document through the presence of unique words in them]

*Like* ***Shingles****.*

Gather all the unique words presented in all the documents which is called vocabulary. Then create a vector for unique identification of each document. Total number of entries of the document is equal to the length of the vocabulary. Each entry in a vector that represents a word of the vocabulary is present in the document associated with the vector.

*Ignore frequent words like article or punctuation or prepositions like a, of etc.*

*Reduce the word like playing reduced to play*

*Fixed the misspelled word*

**Grams**

Each word of vocabulary or each token [entry] of a vector is called a gram. Two-word pair of tokens is called bigram. In this way N-gram can be mapped. Only the n-grams appearing in the corpus are modeled. [for example, only the pair which appears in a sentence will be counted or measured]

*tri-gram*

**Continuous Bag of Words (CBOW)** [finds a missing word in a context]

Given a context, if a word is missing from the context, the Neural Network (NN) will try to find the word by analyzing the words surrounding the missing word.

**Window Size**

The number of words given as input in the NN to predict the missing word. [Generally the words are adjacent to the missing words]. If m is window size and c is the position of the missing word, then words from c-m to c+m except the missing word at position c will be given as input into the NN to predict the missing word at position c.

**Skip-gram** [finds the context given a word]

In skip-gram if a word and the window size is given as input then the context is predicted. In skip gram the number of predicted words will be two times of window size.

**TF-IDF** [Term frequency inverse Document frequency] [Measures how relevant a word in a series or corpus is to a text]

**TF**: (Number of times a word appears in the document) / (Total number of words in the document)

**IDF**: log(total number of documents / number of documents contain the term)

**TF-IDF** = TF\*IDF

**Multinomial naive bayes** [Can be used for determining the count or frequency of a word in the text] [Discrete measure] [combined/multiple probability]

[What is the revised probability of an event occurring after taking new information into consideration?]

It predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

**Naive Bayes**

Simple rule of belonging to or not. If an item belongs to one group or collection, it will not belong to any other group. [Feature being classified is not related to any other feature. The presence or absence of one feature does not affect the presence or absence of the other feature]

**Bayes theorem**

The probability of an event occurring based on the prior knowledge of conditions related to an event.

P(A|B) = P(A) \* P(B|A)/P(B)

Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often i.e. frequency. On the other hand, Bernoulli is a binary algorithm used when the feature is present or not.