<https://www.mygreatlearning.com/blog/bag-of-words/>

<https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22>

<https://medium.com/analytics-vidhya/deep-dive-into-word2vec-7fcefa765c17>

For detail mathematics behind Word embedding with CBOW and skip gram model:

<https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>

**NLP**

Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data.

In simple terms, Natural language processing (NLP) is the ability of computers to understand human speech as it is spoken. NLP helps to analyze, understand, and derive meaning from human language in a smart and useful way.

Popular and simple method of feature extraction with text data which are currently used are

* Bag of Words
* TF-IDF
* Word2Vec

**Bag of words**

Bag of words is a Natural Language Processing technique of text modeling. In technical terms, we can say that it is a method of feature extraction with text data. This approach is a simple and flexible way of extracting features from documents.

Examples can be observed in the notebook.

**TF-IDF**

tf–idf or TFIDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.

This concept includes

1. Counts : Count the number of times each word appears in a document.
2. Frequencies : Calculate the frequency that each word appears in a document out of all the words in the document.

**Term Frequency:**  Term frequency (TF) is used in connection with information retrieval and shows how frequently an expression (term, word) occurs in a document. Term frequency indicates the significance of a particular term within the overall document. It is the number of times a word wi occurs in a review rj with respect to the total number of words in review rj.

**Inverse Document Frequency:** The inverse document frequency is a measure of how much information the word provides, i.e., if it’s common or rare across all documents. It is used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score. It is the logarithmically scaled inverse fraction of the documents that contain the word (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient):

**Word2Vec**

The Word2Vec model is used for learning vector representations of words called “word embeddings”. This is typically done as a preprocessing step, after which the learned vectors are fed into a discriminative model (typically an RNN) to generate predictions and perform all sorts of interesting things. It takes the semantic meaning of words.

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

**CBOW (Continuous Bag Of Words) and Skip-Gram are two most popular frames for word embedding**. In CBOW the words occurring in context (surrounding words) of a selected word are used as inputs and middle or selected word as the target. It's the other way round in Skip-Gram, here the middle word tries to predict the words coming before and after it.