**NLP Assignment-1**

1. **Explain One-Hot Encoding**

A one hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

1. **Explain Bag of Words**

A problem with modeling text is that it is messy, and techniques like machine learning algorithms prefer well defined fixed-length inputs and outputs.

Machine learning algorithms cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers. This is called feature extraction or feature encoding.

A popular and simple method of feature extraction with text data is called the bag-of-words model of text. A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms.

1. **Explain Bag of N-Grams**

A bag-of-n-grams model records the number of times that each n-gram appears in each document of a collection. An n-gram is a collection of n successive words. bagOfNgrams does not split text into words. To create an array of tokenized documents, see tokenizedDocument .

1. **Explain TF-IDF**

TF-IDF or ( Term Frequency(TF) — Inverse Dense Frequency(IDF) )is a technique which is used to find meaning of sentences consisting of words and cancels out the incapabilities of Bag of Words technique which is good for text classification or for helping a machine read words in numbers. However, it just blows up in your face when you ask it to understand the meaning of the sentence or the document**.**

1. **What is OOV problem?**

Out-of-vocabulary (OOV) are terms that are not part of the normal lexicon found in a natural language processing environment. In speech recognition, it's the audio signal that contains these terms. Word vectors are the mathematical equivalent of word meaning.

1. **What are word embeddings?**

It is an approach for representing words and documents. Word Embedding or Word Vector is a numeric vector input that represents a word in a lower-dimensional space. It allows words with similar meaning to have a similar representation. They can also approximate meaning. A word vector with 50 values can represent 50 unique features.

Features: Anything that relates words to one another. Eg: Age, Sports, Fitness, Employed etc. Each word vector has values corresponding to these features.

Goal of Word Embeddings

To reduce dimensionality

To use a word to predict the words around it

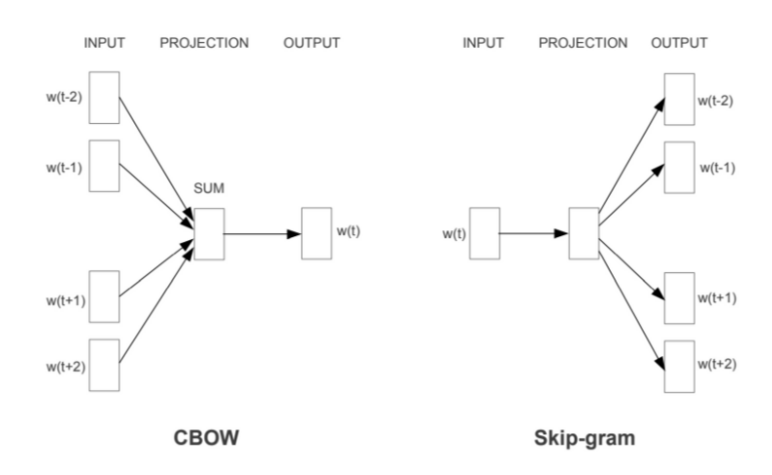
Inter word semantics must be captured

1. **Explain Continuous bag of words (CBOW)**

In the CBOW model, the distributed representations of context (or surrounding words) are combined to predict the word in the middle.

1. **Explain SkipGram**

While in the Skip-gram model, the distributed representation of the input word is used to predict the context.



1. **Explain Glove Embeddings.**

GloVe stands for global vectors for word representation. It is an unsupervised learning algorithm developed by Stanford for generating word embeddings by aggregating global word-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the word in vector space.