1. **Explain One-Hot Encoding**

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.

1. **Explain Bag of Words**

A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded

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

1. **Explain TF-IDF**

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents.

1. **What is OOV problem?**

Out of vocabulary words are words that are not in the training set, but appear in the test set, real data. The main problem is that the model assigns a probability zero to out of vocabulary words resulting in a zero likelihood. This is a common problem, specially when you have not trained on a smaller data set.

1. **What are word embeddings?**

In natural language processing (NLP), word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

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

The CBOW model architecture tries to predict the current target word (the center word) based on the source context words (surrounding words).

1. **Explain SkipGram**

Skip-gram is one of the unsupervised learning techniques used to find the most related words for a given word. Skip-gram is used to predict the context word for a given target word. It's reverse of CBOW algorithm. Here, target word is input while context words are output

1. **Explain Glove Embeddings.**

GloVe stands for Global Vectors for word representation. It is an unsupervised learning algorithm developed by researchers at Stanford University aiming to generate word embeddings by aggregating global word co-occurrence matrices from a given corpus