1. Explain One-Hot Encoding

One-hot encoding is a technique used to represent categorical data in a numerical format. It is commonly employed in machine learning and data analysis tasks to handle categorical variables, which are variables that can take on a limited number of distinct values or categories.

In one-hot encoding, each unique category is represented by a binary vector, where all elements are zero except for the position corresponding to the category, which is set to one. This binary vector is often referred to as a "dummy variable."

1. Explain Bag of Words

Bag of Words (BoW) is a simple and widely used technique in natural language processing (NLP) to represent text data as numerical feature vectors. It treats a document as an unordered collection or "bag" of words, disregarding grammar and word order, and focuses solely on the presence or absence of words in the document.

1. Explain Bag of N-Grams

Bag of N-Grams is an extension of the Bag of Words (BoW) model in natural language processing (NLP) that takes into account not just individual words but also contiguous sequences of words known as "n-grams." An n-gram is a sequence of n consecutive words in a document.

1. Explain TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical representation technique used in natural language processing (NLP) and information retrieval to assess the importance of terms in a document within a collection of documents.

TF-IDF takes into account two factors: term frequency (TF) and inverse document frequency (IDF)..

Term Frequency (TF):

Inverse Document Frequency (IDF):

1. What is OOV problem?

The OOV problem, or Out-of-Vocabulary problem, refers to the challenge encountered when a word or token that is present in the data being processed is not recognized or found in the vocabulary or lexicon of a language model or NLP system.

When a language model or NLP system encounters an out-of-vocabulary word, it typically treats it as an unknown token or assigns a special placeholder to represent it. This can happen for several reasons.

1. What are word embeddings?

Word embeddings are distributed representations of words in a numerical form that capture the semantic and syntactic relationships between words. They are dense vector representations that encode semantic information based on the words' context and usage patterns within a given corpus or text dataset.

1. Explain Continuous bag of words (CBOW)

Continuous Bag of Words (CBOW) is a model architecture used in natural language processing (NLP) and neural network-based language models for word representation and prediction. CBOW aims to predict a target word based on its context, where the context is defined by the surrounding words.

1. Explain SkipGram

Skip-gram is a model architecture used in natural language processing (NLP) and neural network-based language models for learning word embeddings. It is an alternative to the Continuous Bag of Words (CBOW) model. While CBOW predicts a target word given its context, Skip-gram predicts the surrounding context words given a target word.

1. Explain Glove Embeddings.

GloVe (Global Vectors for Word Representation) is a word embedding model that aims to capture the semantic meaning of words by considering their co-occurrence patterns in a corpus of text. It was introduced as an alternative approach to learning word embeddings, complementary to models like Word2Vec.

The GloVe model leverages the statistics of word co-occurrence across the entire corpus to learn word embeddings. It combines the advantages of global and local word co-occurrence information to generate high-quality embeddings. The underlying intuition is that words that frequently appear together and have similar contexts are likely to have similar meanings.