1. Explain One-Hot Encoding

One-hot encoding is a binary representation of categorical variables as a vector of 0s and 1s. Each category is assigned a unique index, and for a given input category, only one element in the vector is set to 1 (hot), while the others are set to 0 (cold). It is used to represent categorical data in a format that can be easily processed by machine learning algorithms.

1. Explain Bag of Words

Bag of Words is a text representation technique where a document is represented as an unordered collection of words, ignoring grammar and word order. It creates a vector where each dimension represents the frequency of a word in the document. BoW is commonly used in natural language processing for tasks like text classification and sentiment analysis.

1. Explain Bag of N-Grams

Bag of N-Grams extends the Bag of Words model by including sequences of N contiguous words (n-grams) as features. It captures local word order and dependencies within a certain window of words. This approach can improve the representation of text by considering phrases or multi-word expressions.

1. Explain TF-IDF

TF-IDF is a numerical statistic used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It combines the term frequency (how often a term appears in a document) with the inverse document frequency (how unique a term is across the corpus). High TF-IDF values indicate that a word is important to a specific document but relatively rare in the corpus.

1. What is OOV problem?

The OOV problem occurs when a word that is not in the vocabulary of a language model or system is encountered. OOV words cannot be processed or understood by the model, leading to potential errors. Techniques like handling OOV words with unknown tokens or subword embeddings are used to mitigate this problem.

1. What are word embeddings?

Word embeddings are dense, continuous-valued vector representations of words in a high-dimensional space. They capture semantic relationships between words and are learned from large text corpora. Word embeddings, such as Word2Vec and GloVe, are used to improve the performance of natural language processing tasks by providing distributed representations of words.

1. Explain Continuous bag of words (CBOW)

CBOW is a type of word embedding model that learns to predict a target word based on the context words surrounding it. It uses a sliding context window to generate a word vector for the target word. CBOW is trained to minimize the prediction error between the target word and its context.

1. Explain SkipGram

SkipGram is another type of word embedding model that learns to predict context words based on a target word. It aims to predict the context words surrounding a target word in a sentence. SkipGram is used to capture the relationships between a word and its context words.

1. Explain Glove Embeddings.

GloVe (Global Vectors for Word Representation) is an unsupervised word embedding model that combines the strengths of global matrix factorization and local context window methods. It creates word vectors by capturing word co-occurrence statistics across a large text corpus. GloVe embeddings are known for their ability to represent semantic relationships between words effectively.