**Assignment - 1**

1. Explain One-Hot Encoding

Ans: One-Hot Encoding is a technique used in machine learning and natural language processing to represent categorical data numerically. In this encoding scheme, each category is represented as a binary vector, where each element of the vector corresponds to a category, and only one element is 1 (hot) while the rest are 0 (cold). For example, if we have three categories A, B, and C, the one-hot encoding representation of category B would be [0, 1, 0].

1. Explain Bag of Words

Ans: Bag of Words (BoW) is a simple and commonly used technique for text representation in natural language processing. It involves creating a vocabulary of unique words present in a corpus and representing each document as a vector where each dimension corresponds to the frequency of a word in the document. The order of words is disregarded, and only the occurrence of words matters.

1. Explain Bag of N-Grams

Ans: Bag of N-Grams is an extension of the Bag of Words model that considers sequences of words (N-grams) instead of individual words. Instead of representing documents using single words, Bag of N-Grams represents documents using all possible contiguous sequences of N words. This approach captures some of the contextual information present in the text.v

1. Explain TF-IDF

Ans: TF-IDF (Term Frequency-Inverse Document Frequency) 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 frequency of a term (word) in a document (Term Frequency) with its inverse frequency across all documents in the corpus (Inverse Document Frequency). Words that occur frequently in a document but rarely in the corpus are assigned higher TF-IDF scores, indicating their importance in that document.

1. What is OOV problem?

Ans: OOV (Out of Vocabulary) problem refers to the issue encountered when a word that is present in the input data is not found in the vocabulary of a model or system. This problem often arises in natural language processing tasks when dealing with unseen or rare words that were not encountered during training.

1. What are word embeddings?

Ans: Word embeddings are dense, low-dimensional vector representations of words in a continuous vector space. These representations are learned from large corpora of text using techniques like Word2Vec, GloVe, or FastText. Word embeddings capture semantic and syntactic similarities between words, allowing machine learning models to understand the relationships between words and generalize better to unseen data.

1. Explain Continuous bag of words (CBOW)

Ans: Continuous Bag of Words (CBOW) is a type of word embedding model used to learn distributed representations of words in a continuous vector space. CBOW predicts a target word based on its context words within a fixed window size. It aims to learn to predict a target word given the surrounding context words.

1. Explain SkipGram

Ans: SkipGram is another type of word embedding model used to learn distributed representations of words. Unlike CBOW, SkipGram predicts context words given a target word. It aims to learn the representation of a target word that is useful for predicting the words that co-occur with it in a sentence.

1. Explain Glove Embeddings.

Ans: GloVe (Global Vectors for Word Representation) is a word embedding technique that learns vector representations of words by factorizing the logarithm of the word co-occurrence matrix. GloVe embeddings capture global word-word co-occurrence statistics across the entire corpus, resulting in embeddings that encode both semantic and syntactic information about words.