**NLP Assignment 1**

1. Explain One-Hot Encoding

Ans-) One-Hot Encoding is a technique used to represent categorical data as numerical data. In this technique, each category is represented as a vector of binary values, where only one value is 1 and the rest are 0s. For example, if we have three categories: cat, dog, and bird, then cat can be represented as [1, 0, 0], dog as [0, 1, 0], and bird as [0, 0, 1].

2. Explain Bag of Words

Ans-) Bag of Words is a technique used to represent text data as numerical data. In this technique, a document is represented as a bag of words, where the frequency of occurrence of each word in the document is counted and stored as a vector. The order of the words is ignored, and only the frequency matters.

3. Explain Bag of N-Grams

Ans-) Bag of N-Grams is a variation of the Bag of Words technique where instead of considering single words, we consider groups of N consecutive words, known as N-grams. This technique helps to capture the context of words better than the Bag of Words technique.

4. Explain TF-IDF

Ans-) TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used to represent text data as numerical data. It measures the importance of a word in a document by weighing its frequency in the document and inversely proportional to the frequency of the word in the corpus. Words that occur frequently in a document but rarely in the corpus are considered important and given a high score, while words that occur frequently in the corpus are considered less important and given a low score.

5. What is OOV problem?

Ans-) OOV (Out-of-vocabulary) problem is a problem that occurs in natural language processing when a word that is not present in the vocabulary of the model is encountered. This problem can occur when the model is trained on a limited vocabulary or when it encounters new or rare words that are not present in the training data.

6. What are word embeddings?

Ans-) Word embeddings are a technique used to represent words as numerical vectors in a high-dimensional space. Each word is represented as a dense vector of real numbers, where words with similar meanings are represented by vectors that are closer together in the space. Word embeddings are commonly used in natural language processing tasks such as language translation and text classification.

7. Explain Continuous bag of words (CBOW)

Ans-) Continuous Bag of Words (CBOW) is a neural network architecture used for word embeddings. In this architecture, the model predicts the current word based on the context of surrounding words. The context words are averaged and fed into a neural network, which predicts the current word.

8. Explain SkipGram

Ans-) SkipGram is another neural network architecture used for word embeddings. In this architecture, the model predicts the surrounding words based on the current word. The current word is fed into a neural network, which predicts the context words.

9. Explain Glove Embeddings.

Ans-) Glove (Global Vectors for Word Representation) Embeddings is a technique used to generate word embeddings. It is based on matrix factorization of the word co-occurrence matrix. The resulting word vectors capture the meaning of words and their relationships with other words in the corpus. Glove embeddings are often used in natural language processing tasks such as sentiment analysis and language translation.