**ADITYA AMIN**

**ASSIGN : NLP-01**

1. Explain One-Hot Encoding

One-hot encoding is a commonly used technique in machine learning and data processing to represent categorical variables numerically. It is used when we have a categorical feature that cannot be directly used in mathematical models or algorithms that expect numerical input. One-hot encoding converts each category in the feature into a binary vector representation, where only one element is 1 (hot) and the rest are 0 (cold).

1. Explain Bag of Words

Bag of Words (BoW) is a commonly used technique in natural language processing (NLP) to represent text data numerically. It aims to capture the frequency or occurrence of words in a document without considering the order or structure of the text. The name "bag of words" suggests that it treats the text as an unordered collection of words, similar to a bag where the order of the words is irrelevant.

1. Explain Bag of N-Grams

Bag of N-Grams is an extension of the Bag of Words (BoW) approach in natural language processing (NLP) that considers not just individual words but also sequences of N contiguous words, known as "N-grams." An N-gram is a contiguous sequence of N tokens (words or characters) extracted from a given text.

1. Explain TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used in natural language processing (NLP) to evaluate the importance of a term within a document or a corpus of documents. TF-IDF takes into account both the term frequency (TF) and the inverse document frequency (IDF) to calculate a weight for each term.

1. What is OOV problem?

The OOV (Out-of-Vocabulary) problem refers to the challenge of encountering words or tokens in a text that are not present in the vocabulary or training data of a natural language processing (NLP) model. When an unknown word is encountered, it poses difficulties for the model to handle and process that word effectively.

1. What are word embeddings?

Word embeddings are dense vector representations of words in a continuous vector space. They are numerical representations that capture the semantic and syntactic meaning of words based on their contextual usage in a given corpus of text. Word embeddings are widely used in natural language processing (NLP) tasks such as language modeling, sentiment analysis, machine translation, and document clustering.

1. Explain Continuous bag of words (CBOW)

Continuous Bag of Words (CBOW) is a popular algorithm for training word embeddings in natural language processing (NLP). It is a variant of the Word2Vec model, which aims to learn distributed representations of words based on their surrounding context.

1. Explain SkipGram

Skip-gram is a popular algorithm used for training word embeddings in natural language processing (NLP). It is a variant of the Word2Vec model, designed to learn distributed representations of words based on their surrounding context.

1. Explain Glove Embeddings.

GloVe (Global Vectors for Word Representation) is an unsupervised learning algorithm used for generating word embeddings in natural language processing (NLP). It aims to create vector representations for words that capture semantic and syntactic relationships between words based on their co-occurrence patterns in a corpus of text.