ASSIGNMENT - 1

1. Explain One-Hot Encoding

Ans: One-hot encoding is a technique used to represent categorical variables as binary vectors. Each category in the variable is assigned a vector with all zeros except for one position. The position with the '1' indicates the category the data point belongs to.

For example, imagine a dataset with a categorical variable "Color" with values "Red", "Green", and "Blue". Using one-hot encoding, "Red" would be represented as [1, 0, 0], "Green" as [0, 1, 0], and "Blue" as [0, 0, 1].

This allows machine learning models to understand categorical data by converting it into a numerical format they can work with.

2. Explain Bag of Words

Ans: The Bag of Words (BoW) model is a simplified way to represent a document. It completely ignores the order of words and focuses only on their presence. Each word in a vocabulary is treated as a unique feature.

For each document, a histogram is created. The value at each position in the histogram represents the number of times the corresponding word appears in the document.

While BoW is simple and easy to implement, it loses information about word order and relationships between words.

3. Explain Bag of N-Grams

Ans: An extension of BoW, Bag of N-Grams considers sequences of N words (N-grams) as features instead of individual words. This captures some context from word order.

For example, a bigram model (N=2) would consider "red car" and "car red" as different features, acknowledging the order of the words.

4. Explain TF-IDF

Ans: Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting scheme used for words in a document to highlight their importance.

Term Frequency (TF): Represents how often a word appears in a document.

Inverse Document Frequency (IDF): Measures how common a word is across all documents in the corpus. Words that appear frequently across all documents have a lower IDF weight.

The TF-IDF score for a word considers both its frequency within a document and its rarity across the corpus. This helps prioritize words that are important for a specific document but not overly common.

5. What is OOV problem?

Ans: The OOV problem arises when a model encounters a word during testing that it has not seen during training. This can happen with new or rare words.

One-hot encoding, BoW, and similar techniques rely on a predefined vocabulary. If a word is not present in the training vocabulary, the model won't be able to represent it.

6. What are word embeddings?

Ans: Word embeddings are a dense vector representation of words. Unlike one-hot encoding, which is sparse, word embeddings capture semantic relationships between words. Words with similar meanings tend to have closer vectors in the embedding space.

This allows models to understand the relationships between words and perform tasks like text similarity, analogy detection, and sentiment analysis more effectively.

7. Explain Continuous bag of words (CBOW)

Ans: CBOW is a neural network architecture used to learn word embeddings. It predicts a target word based on its surrounding context words.

Imagine a sentence "The quick brown fox jumps over the lazy dog". The CBOW model, given "quick" and "brown", would try to predict the most likely word in the center, which is "fox". By adjusting the weights in the network based on the prediction accuracy, CBOW learns to represent words in a way that captures their semantic relationships.

8. Explain SkipGram

Ans: SkipGram is another neural network architecture for learning word embeddings. It works in the opposite way of CBOW. Given a word, it predicts the surrounding context words.

Using the same sentence, SkipGram would take "fox" as input and try to predict words like "quick" and "brown" that appear around it. This approach also helps the model learn semantic relationships between words.

9. Explain Glove Embeddings.

Ans: GloVe (Global Vectors) is a popular word embedding technique that combines statistical information from a large corpus of text data. It leverages the co-occurrence of words to learn their representations.

GloVe considers not just the immediate neighbors of a word but also words that appear in similar contexts, providing a richer representation of word meaning.