

1. Explain One-Hot Encoding
2. Explain Bag of Words
3. Explain Bag of N-Grams
4. Explain TF-IDF
5. What is OOV problem?
6. What are word embeddings?
7. Explain Continuous bag of words (CBOW)
8. Explain SkipGram
9. Explain Glove Embeddings.

Ans:

1. **One-Hot Encoding:** One-Hot Encoding is a technique used to convert categorical data into a numerical format. In this technique, each categorical value is converted into a binary vector of zeros and ones. For example, suppose there is a categorical feature "fruit" with three categories: apple, banana, and orange. One-hot encoding will create three binary vectors with values [1, 0, 0], [0, 1, 0], and [0, 0, 1] for apple, banana, and orange, respectively.
2. **Bag of Words:** Bag of Words is a technique used for text classification and natural language processing. In this technique, a text is represented as a bag of its words, disregarding grammar and word order but keeping track of the frequency of each word. The Bag of Words technique creates a vocabulary of words from the corpus and represents each document as a vector of word frequencies.
3. **Bag of N-Grams:** Bag of N-Grams is an extension of the Bag of Words technique, where instead of representing each word as a feature, it represents each n-gram (a sequence of n words) as a feature. For example, if n=2, then each feature represents a combination of two adjacent words. Bag of N-Grams can capture more context than Bag of Words.
4. **TF-IDF:** TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used for text classification and information retrieval. It calculates the importance of each word in a document by weighing the frequency of the word in the document and inversely weighing the frequency of the word in the corpus. This technique aims to identify words that are important to a specific document but not frequent in the corpus.
5. **OOV problem:** OOV (Out of Vocabulary) problem refers to the problem of encountering words that are not present in the vocabulary of a language model or a machine learning model. OOV words can negatively impact the performance of a model since it cannot process them, resulting in errors or incorrect predictions.
6. **Word embeddings:** Word embeddings are a type of vector representation of words in natural language processing. It aims to map each word to a high-dimensional vector space in a way that preserves the semantic relationships between words. Word embeddings are trained using neural networks on large amounts of text data and can capture the context and meaning of words.
7. **Continuous bag of words (CBOW):** Continuous bag of words (CBOW) is a type of neural network architecture used for training word embeddings. CBOW aims to predict a target word from its surrounding context words. In other words, given a set of context words, CBOW predicts the probability of the target word.

8. SkipGram: SkipGram is another type of neural network architecture used for training word embeddings. In SkipGram, the model predicts the surrounding context words from a given target word. SkipGram is useful for capturing the meaning of rare words or words with multiple meanings.
9. Glove Embeddings: Glove (Global Vectors for Word Representation) Embeddings is a type of word embedding technique that combines the advantages of both Bag of Words and word co-occurrence matrices. Glove Embeddings uses a co-occurrence matrix to capture the relationships between words and then applies a factorization technique to produce dense vector representations for each word. Glove embeddings are known to perform well in various natural language processing tasks.