NLP assingments-1

1. One-hot encoding is a way to represent categorical variables as numerical data. It involves creating a new binary column for each unique category in a categorical variable. Each row is then encoded as a one or a zero in each of the new columns, depending on which category the row belongs to.
2. Bag of words is a way to represent text data when working with natural language processing (NLP) tasks. It involves creating a vocabulary of all the unique words in a corpus of texts, and then creating a matrix where each row represents a document and each column represents a word in the vocabulary. The value in each cell of the matrix is the frequency with which the word appears in the document.
3. Bag of n-grams is similar to bag of words, but instead of considering only single words, it also considers combinations of adjacent words (n-grams). For example, if n is 2, the bag of n-grams representation would consider pairs of adjacent words (bigrams), and if n is 3, it would consider triples of adjacent words (trigrams).
4. TF-IDF (term frequency-inverse document frequency) is a weighting scheme used in NLP to assign importance to words in a document based on how frequently they appear in the document and how rare they are across all documents in a corpus. It helps to identify the most important words in a document by down-weighting words that are common across all documents and up-weighting words that are more specific to a particular document.
5. The OOV (out-of-vocabulary) problem refers to the difficulty of representing words that are not present in the vocabulary of a model. When working with NLP tasks, it is common to build a vocabulary of all the unique words in a corpus of texts, and then to represent each document as a bag of words using this vocabulary. However, if a document contains words that are not in the vocabulary, they will be treated as OOV words and may not be properly represented in the bag of words representation.
6. Word embeddings are numerical representations of words in a low-dimensional vector space. They are used in NLP to capture the meaning of words and the relationships between them in a way that can be understood by a machine learning model. Word embeddings are typically learned from large amounts of text data and are used as the input to NLP models.
7. Continuous bag of words (CBOW) is a method for learning word embeddings. It involves predicting a target word based on the context of the surrounding words. For example, given the sentence "The cat sat on the mat," CBOW would try to predict the word "cat" based on the context of the words "The," "sat," "on," "the," and "mat."
8. SkipGram is another method for learning word embeddings. It is similar to CBOW, but instead of predicting a target word based on the context of the surrounding words, it predicts the surrounding words based on the target word. Using the same example sentence, SkipGram would try to predict the words "The," "sat," "on," "the," and "mat" based on the target word "cat."
9. GloVe (global vectors) is a method for learning word embeddings developed by Stanford University researchers. It involves training a model on a large corpus of text data to learn the co-occurrence statistics of words, and then using these statistics to learn the word embeddings. GloVe embeddings are widely used in N