**NLP Assignment-1**

**Q1. Explain One-Hot Encoding.**

**Ans:** One-Hot Encoding is a technique used in NLP and machine learning to represent categorical variables as binary vectors. It converts each category or word into a binary vector, where each position in the vector corresponds to a unique category. The value of the corresponding position is set to 1, indicating the presence of a category, while the values of all other positions are set to 0. This encoding allows categorical data to be used in mathematical models and helps extract meaningful patterns from text data.

**Q2. Explain Bag of Words.**

**Ans:** Bag of Words is a text representation technique that treats a document as a collection of individual words, disregarding grammar and word order. It creates a vocabulary of unique words and represents each document as a numerical vector, where each position corresponds to a word and the value represents its frequency. Bag of Words is used in NLP tasks like text classification but doesn't consider word meaning or context.

**Q3. Explain Bag of N-Grams.**

**Ans:** Bag of N-Grams is an extension of Bag of Words (BoW) that considers sequences of words called n-grams. It creates a vocabulary of unique n-grams, generates numerical vectors for each document, counts occurrences of words and n-grams, and represents the dataset as a matrix. By capturing word relationships, Bag of N-Grams provides more contextual information than BoW. However, it still has limitations like high dimensionality. Techniques like TF-IDF, feature selection, and advanced language models are used to address these limitations and enhance performance.

**Q4. Explain TF-IDF.**

**Ans:** TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used in NLP to evaluate the importance of a word in a document. It combines the frequency of a word within a document (TF) and its rarity across the document collection (IDF). TF measures word frequency, while IDF measures word uniqueness. The TF-IDF score is obtained by multiplying these values. It helps identify important and distinctive words in a document collection, aiding in tasks like document similarity and keyword extraction.

**Q5. What is OOV problem?**

**Ans:** The OOV (Out-of-Vocabulary) problem in NLP refers to encountering words in text that are not present in the model's vocabulary or training data. OOV words pose a challenge as the model lacks information about them. Strategies to handle OOV words include using special tokens, subword tokenization, morphological analysis, or dictionary-based approaches. Effectively addressing the OOV problem is essential for better performance and generalization of NLP models.

**Q6. What are word embeddings?**

**Ans:** Word embeddings are dense vector representations of words in NLP. They capture semantic and syntactic information and are learned from large amounts of text data. Unlike sparse one-hot encodings, word embeddings encode relationships between words. Similar words have vectors closer together. Word embeddings are widely used in NLP tasks, enhancing performance and understanding of language semantics.

**Q7. Explain Continuous bag of words (CBOW).**

**Ans:** CBOW (Continuous Bag of Words) is a word embedding model in NLP. It predicts a target word by using its surrounding context words. CBOW creates a context vector by averaging the word embeddings of the context words and using it to train a neural network. The model learns meaningful word representations that capture semantic relationships. CBOW is computationally efficient and commonly used in language modeling and sentiment analysis.

**8. Explain SkipGram.**

**Ans:** Skip-gram is a word embedding model in NLP that predicts context words given a target word. It learns representations of words by training a neural network to capture the relationships between target and context words. Skip-gram is effective for capturing semantic similarities and works well with large and diverse datasets. It has applications in language modeling, machine translation, and information retrieval.

**Q9. Explain Glove Embeddings.**

**Ans:** GloVe (Global Vectors for Word Representation) is a word embedding model in NLP. It learns word representations by analyzing the co-occurrence of words in a large text corpus. GloVe combines global and local context information to capture semantic relationships between words. The resulting embeddings are dense vectors that represent words in a high-dimensional space. GloVe embeddings are widely used in NLP tasks to improve understanding and processing of text data.