**NLP Assignment 1 Solution**

**1. Explain One-Hot Encoding?**

**Ans:** One-hot encoding is a process used to convert categorical data, or data with text labels, into a numerical form that a computer can understand. It does this by creating new columns for each categorical feature and assigning a 1 or 0 (hot or cold, respectively) to each row to indicate the presence or absence of a particular feature. For example, if a dataset had a feature called "Gender" with three options - Male, Female, and Other - one-hot encoding would create three new columns, Male, Female, and Other, and assign a 1 or 0 to each row to indicate which option it was.

**2. Explain Bag of Words?**

**Ans:** Bag of Words (BoW) is a technique used in natural language processing (NLP) for representing text. It is a way of extracting features from text for use in machine learning algorithms. BoW is a model representation used to simplify the often complex task of understanding natural language. It is a process of representing text as numerical feature vectors. It is one of the most common techniques used in NLP for feature extraction and is used to represent text in the form of a bag of words. The bag of words model ignores grammar and word order, but keeps track of the frequency of the words in the text. The text is represented as a numerical vector in which each word is represented by a number indicating the frequency of occurrence in the text.

**3. Explain Bag of N-Grams?**

**Ans:** Bag of N-Grams is a type of feature representation used in Natural Language Processing (NLP). It is a technique used to represent text data as numerical features, where each feature represents a collection (or “bag”) of adjacent words or “N-Grams”. N-Grams are a sequence of N words taken together, and can be a single word (Unigram), two words (Bigram), three words (Trigram), and so on. Bag of N-Grams is used in supervised learning algorithms such as text classification, sentiment analysis, and language modeling. The advantage of Bag of N-Grams is it captures the context of words in a sentence, which is especially useful when the meaning of words changes depending on the context.

**4. Explain TF-IDF?**

**Ans:** TF-IDF (term frequency–inverse document frequency) is a statistical measure used in natural language processing (NLP) to reflect how important a word is to a document in a corpus. It is the product of two statistics, term frequency (TF) and inverse document frequency (IDF). The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. The weighting of TF-IDF is intended to represent the importance of a word in the document.

**5. What is OOV problem?**

**Ans:** OOV (Out-of-Vocabulary) is a problem in natural language processing (NLP) where a system cannot handle words that are not included in its vocabulary. OOV words are words that the system has not seen before and therefore cannot understand or process. This is a common problem when dealing with text data, as unseen words appear frequently in natural language. To address this problem, NLP systems use techniques such as word embeddings, language models, and lexicon expansion.

**6. What are word embeddings?**

**Ans:** Word embeddings are a type of representation for text data, where each word in the corpus is represented as a vector of real numbers. Word embeddings can capture semantic and syntactic similarities between words, allowing models to understand the meaning of words in context and accurately make predictions. They are widely used in Natural Language Processing (NLP) applications such as sentiment analysis, text classification, and machine translation.

**7. Explain Continuous bag of words (CBOW)?**

**Ans:** Continuous Bag-of-Words (CBOW) is a natural language processing model that predicts a target word from its context. The model takes as input a set of context words (also known as the “bag”) and predicts the target word that belongs in the context. CBOW is based on the idea that words that appear in the same context are likely to be related. The model uses the context words to predict the target word by taking into account the context of the words. The model is trained on a large corpus of text and learns to predict the target word from its context.

**8. Explain SkipGram?**

**Ans:** SkipGram is a type of neural network architecture used mainly for natural language processing (NLP). It is used to predict a target word from its surrounding context, meaning it predicts the target word given a set of words that come before and after it in the text. SkipGram works by taking in a word and its surrounding context, and then outputting a probability distribution over all possible target words that could fill in the gap. The probability distribution is then used to determine the most likely target word, given the context. SkipGram is often used in word embedding techniques, where it is used to learn meaningful dense vector representations of words.

**9. Explain Glove Embeddings?**

**Ans:** Glove (Global Vectors for Word Representation) is a popular word embedding technique developed by Stanford researchers. It is a type of word vector representation that uses a neural network to learn a vector representation of words from large datasets. The resulting vectors contain semantic information about the words that is useful for many natural language processing (NLP) tasks. Glove embeddings are pre-trained on a huge corpus of text and can be used to represent any word in the corpus. The semantic information captured in the embeddings can be used to compute similarity between words, detect relationships between words, or even classify text. Glove embeddings are used in many NLP applications such as language modeling, sentiment analysis, and text classification.