**Assignment - 2**

1. What are Corpora?

Ans: Corpora (singular: corpus) are large collections of text or linguistic data, often used for linguistic analysis, natural language processing (NLP), or machine learning tasks. Corpora can consist of various types of text, including written documents, transcripts, speeches, social media posts, and more, and they serve as the basis for training and evaluating NLP models and algorithms.

1. What are Tokens?

Ans: Tokens are the individual units of text that result from tokenization, the process of breaking down a text into smaller components, such as words, punctuation marks, numbers, or other meaningful elements. Tokens are the building blocks used in various NLP tasks, including parsing, part-of-speech tagging, and text analysis.

1. What are Unigrams, Bigrams, Trigrams?

Ans: Unigrams, bigrams, and trigrams refer to sequences of one, two, and three adjacent words or tokens, respectively, in a text or corpus. They are used in NLP for tasks such as language modeling, feature extraction, and sequence analysis. For example, in the sentence "The quick brown fox," the unigrams are ["The", "quick", "brown", "fox"], the bigrams are ["The quick", "quick brown", "brown fox"], and the trigrams are ["The quick brown", "quick brown fox"].

1. How to generate n-grams from text?

Ans: N-grams can be generated from text by sliding a window of size n across the text and extracting sequences of n adjacent tokens. This process captures the local context and dependencies between words in the text. N-grams are commonly used in language modeling, machine translation, and other NLP tasks.

1. Explain Lemmatization

Ans: Lemmatization is the process of reducing words to their base or canonical form, known as the lemma, by removing inflections and variations such as tense, case, gender, or number. Lemmatization aims to normalize words so that different forms of the same word are treated as identical, facilitating tasks like text analysis, information retrieval, and language understanding

1. Explain Stemming

Ans: Stemming is the process of reducing words to their root or stem form by removing suffixes or prefixes. Stemming is a more heuristic approach compared to lemmatization and may produce stems that are not actual words. Despite its simplicity, stemming is widely used in information retrieval, text mining, and search engines to group together words with similar meanings.

1. Explain Part-of-speech (POS) tagging

Ans: Part-of-speech (POS) tagging is the process of assigning grammatical categories, or parts of speech, to each word or token in a text based on its context and role in the sentence. POS tagging is essential for various NLP tasks, including syntactic parsing, named entity recognition, and machine translation.

1. Explain Chunking or shallow parsing

Ans: Chunking, also known as shallow parsing, is the process of identifying and grouping together syntactically related words or tokens into larger units or phrases, such as noun phrases (NP), verb phrases (VP), or prepositional phrases (PP). Chunking helps extract structured information from text and is often used as a preprocessing step for deeper syntactic analysis.

1. Explain Noun Phrase (NP) chunking

Ans: Noun Phrase (NP) chunking is a specific type of chunking that focuses on identifying and extracting noun phrases from text. Noun phrases typically consist of a noun and any associated modifiers, determiners, or adjectives. NP chunking is useful for extracting meaningful information from text, such as named entities, key terms, or subject-verb-object relationships.

1. Explain Named Entity Recognition

Ans: Named Entity Recognition (NER) is the process of identifying and categorizing named entities, such as persons, organizations, locations, dates, or numerical expressions, within a text. NER is an important task in NLP for extracting structured information from unstructured text data and is used in various applications, including information extraction, question answering, and document classification.