**ADITYA AMIN**

**ASSIGN : NLP-02**

1. What are Corpora?

In natural language processing (NLP), a corpus (plural: corpora) refers to a large and structured collection of text documents or linguistic data. It serves as a fundamental source of linguistic information and is used for various language analysis tasks, including machine learning, computational linguistics, and linguistic research.

1. What are Tokens?

In natural language processing (NLP), a token refers to a unit of textual information that is considered as a single entity. It can be as small as a single character or as large as a word or a phrase, depending on the level of analysis. Tokenization is the process of breaking down a text into individual tokens.

1. What are Unigrams, Bigrams, Trigrams?

Unigrams, bigrams, and trigrams are terms commonly used in natural language processing (NLP) to describe sequences of words within a text. They refer to different levels of n-grams, where an n-gram represents a contiguous sequence of n items (typically words) in a text.

1. How to generate n-grams from text?

Tokenization: Start by tokenizing the text into individual words or other appropriate units, depending on your task and the level of analysis required. This can be done using libraries or functions specifically designed for tokenization, or by using regular expressions to split the text into tokens.

Generate n-grams: Once the text has been tokenized, you can generate n-grams by sliding a window of size n over the tokens. The window moves one token at a time, extracting the contiguous sequence of n tokens within it. This process is repeated until all n-grams have been extracted.

Store or process the n-grams: The generated n-grams can be stored in a suitable data structure, such as a list or a dictionary, depending on your requirements. You can also further process or analyze the n-grams, such as counting their occurrences, calculating statistics, or using them as features for machine learning models.

1. Explain Lemmatization

Lemmatization is a process used in natural language processing (NLP) to reduce words to their base or root form, known as the lemma. The lemma represents the canonical or dictionary form of a word, which can be used to group together different inflected forms of the same word.

1. Explain Stemming

Stemming is a process used in natural language processing (NLP) to reduce words to their root or stem form. It involves removing prefixes, suffixes, and other affixes from words to obtain a common base form.

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

Part-of-speech (POS) tagging is a process in natural language processing (NLP) that assigns grammatical tags or labels to each word in a sentence, indicating its syntactic category or part of speech. POS tagging helps in understanding the role and function of words within a sentence.

1. Explain Chunking or shallow parsing

Chunking, also known as shallow parsing, is a natural language processing (NLP) technique used to group words together into meaningful syntactic units called chunks. It involves identifying and labeling contiguous sequences of words in a sentence based on their grammatical structure and relationship.

1. Explain Noun Phrase (NP) chunking.

Noun Phrase (NP) chunking is a specific type of chunking in natural language processing (NLP) that focuses on identifying and extracting noun phrases from sentences. Noun phrases are grammatical structures that include a noun and words that modify or describe that noun.

1. Explain Named Entity Recognition.

Named Entity Recognition (NER) is a natural language processing (NLP) technique that aims to identify and classify named entities within a text. Named entities are specific terms or phrases that refer to real-world objects, such as persons, organizations, locations, dates, quantities, and more.