1. What are Corpora?

Corpora (singular: corpus) refer to collections or bodies of text that are used for linguistic analysis, language modeling, and various natural language processing (NLP) tasks. Corpora are structured datasets consisting of written, spoken, or recorded language samples from diverse sources and contexts.

Corpora serve as valuable resources for studying language patterns, understanding linguistic phenomena, and training language models. They can be designed for specific purposes, such as studying a particular language, genre, domain, or linguistic phenomenon, or they can be general-purpose corpora covering a wide range of texts.

1. What are Tokens?

Tokens refer to the individual units or elements that make up a text or a sequence of language. In the context of natural language processing (NLP), tokens are typically words, but they can also be other entities like punctuation marks, symbols, or subword units.

Tokenization is the process of splitting a text or a sentence into its constituent tokens. This is an essential preprocessing step in NLP tasks as it helps to organize and structure the text for further analysis. The resulting tokens serve as the atomic units that are processed and analyzed by NLP models and algorithms.

1. What are Unigrams, Bigrams, Trigrams?

Unigrams: Unigrams refer to single tokens or individual words. They are the simplest form of n-grams, where each token is considered independently.

Bigrams: Bigrams are n-grams consisting of two adjacent tokens. They represent pairs of consecutive words in the text.

Trigrams: Trigrams are n-grams composed of three adjacent tokens. They capture triplets of consecutive words in the text.

1. How to generate n-grams from text?

To generate n-grams from text, you can follow these general steps:

Text preprocessing:

Tokenization:

Generate n-grams:

Handle boundary conditions:

1. Explain Lemmatization

Lemmatization is a linguistic and natural language processing (NLP) technique used to reduce words to their base or canonical form, called a "lemma." The lemma represents the dictionary form or the root form of a word, capturing its core meaning.

The process of lemmatization takes into account the context and morphology of the word to determine its lemma. Unlike stemming, which trims words by removing prefixes or suffixes without considering linguistic context, lemmatization aims to produce valid and meaningful words based on language-specific rules and morphological analysis.

1. Explain Stemming

Stemming is a text normalization technique used in natural language processing (NLP) to reduce words to their base or root form, called a "stem." The stem is a truncated version of a word that may not always be a valid word itself but represents the core meaning of the word.

The process of stemming involves removing prefixes, suffixes, and other affixes from words to produce the stem. This simplification aims to map related words to a common form and reduce variations of words that have the same base meaning. Stemming is primarily a rule-based and heuristic-driven approach, and it does not consider the context or semantics of the word.

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

Part-of-speech (POS) tagging, also known as grammatical tagging or word-category disambiguation, is a process in natural language processing (NLP) that assigns grammatical labels or tags to words in a sentence based on their syntactic roles and linguistic properties.

The goal of POS tagging is to determine the appropriate part-of-speech category for each word in a given text. POS tags represent the grammatical function and behavior of words in a sentence, such as nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and more.

1. Explain Chunking or shallow parsing

Chunking, also known as shallow parsing, is a natural language processing (NLP) technique that involves grouping words into meaningful chunks or phrases based on their syntactic structure and relationships. It aims to identify and extract higher-level linguistic units that span multiple words, such as noun phrases, verb phrases, or prepositional phrases.

Chunking typically follows part-of-speech (POS) tagging and is considered a shallow parsing technique because it does not perform a complete parse tree analysis of the sentence. Instead, it focuses on identifying and labeling chunks that are important for capturing the grammatical structure and semantic meaning of the text.

1. Explain Noun Phrase (NP) chunking

Phrase (NP) chunking, also known as noun phrase chunking, is a specific type of chunking or shallow parsing that focuses on identifying and extracting noun phrases from a sentence or text. A noun phrase (NP) is a syntactic structure that consists of a noun or pronoun along with any accompanying modifiers or determiners.

The goal of NP chunking is to group words together and label them as noun phrases, thereby capturing the grammatical structure and extracting meaningful units that represent entities, objects, or concepts in the text.

1. Explain Named Entity Recognition

Named Entity Recognition (NER) is a natural language processing (NLP) task that aims to identify and classify named entities in text, such as names of people, organizations, locations, dates, quantities, and other named entities with specific semantic meanings.

The goal of NER is to automatically identify and extract these named entities from unstructured text, enabling the understanding and analysis of important information within a given context.