1. What are Corpora?

Corpora (singular: corpus) are large and structured collections of text, speech, or other linguistic data used for language analysis and natural language processing tasks. Corpora serve as datasets for various purposes, such as training language models, conducting linguistic research, and developing NLP applications. They can consist of text from books, articles, conversations, websites, or any source of written or spoken language.

1. What are Tokens?

Tokens are the individual units or words that text is divided into for processing. In natural language processing, tokenization is the process of splitting text into tokens, which could be words, phrases, sentences, or other meaningful linguistic units. Tokens are the basic building blocks used in various language processing tasks.

1. What are Unigrams, Bigrams, Trigrams?

* Unigrams: Unigrams are single words or tokens in a text. They represent the simplest form of tokenization.
* Bigrams: Bigrams consist of pairs of adjacent words in a text. They capture word relationships between adjacent words.
* Trigrams: Trigrams are sequences of three adjacent words in a text. They provide a more context-rich representation of language and capture relationships between three consecutive words.

1. How to generate n-grams from text?

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

* Tokenize the text into words or phrases.
* Create n-grams by sliding a window of size n over the tokens. For unigrams, n = 1; for bigrams, n = 2; and so on.
* Record the n-grams as they appear in the text. For example, "natural language processing" would generate bigrams like ("natural", "language") and ("language", "processing").

1. Explain Lemmatization

Lemmatization is the process of reducing words to their base or dictionary form, known as lemmas. It involves removing inflections and variations to bring words to their canonical or root form. For example, lemmatizing the word "running" would result in "run."

1. Explain Stemming

Stemming is the process of reducing words to their root or stem by removing suffixes or prefixes. Stemming methods are more aggressive than lemmatization and may result in non-words. For example, stemming "jumps" would yield "jump."

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

Part-of-speech tagging is the process of assigning grammatical categories or tags (such as nouns, verbs, adjectives, etc.) to words in a text. It helps identify the syntactic roles of words and is crucial for various language processing tasks.

1. Explain Chunking or shallow parsing

Chunking, also known as shallow parsing, is the process of identifying and grouping adjacent words or tokens into syntactic chunks or phrases, such as noun phrases (NPs) and verb phrases (VPs). It helps extract structured information from text.

1. Explain Noun Phrase (NP) chunking

Noun phrase (NP) chunking is a specific type of chunking that focuses on identifying and grouping words in a text that constitute noun phrases. Noun phrases typically consist of a noun and any associated modifiers, like adjectives or determiners.

1. Explain Named Entity Recognition

Named Entity Recognition is a natural language processing task that involves identifying and classifying named entities in text, such as names of people, organizations, locations, dates, and more. NER is used to extract structured information from unstructured text data and is essential for applications like information retrieval and question answering.