1. What are Corpora?

Ans **Corpora** (plural of **corpus**) refer to large, structured collections of texts or linguistic data used for various purposes in fields like linguistics, natural language processing (NLP), and machine learning.

1. What are Tokens?

**Tokens** are individual units or pieces of text that are extracted from a larger body of text during the process called **tokenization**

1. What are Unigrams, Bigrams, Trigrams?

Ans **Unigrams**, **Bigrams**, and **Trigrams** are terms used to refer to the number of words (or tokens) considered together in a sequence

1. How to generate n-grams from text?

To generate **n-grams** from text, you can break down the text into a sequence of tokens (typically words) and then group them into contiguous sequences of **n** tokens.

1. Explain Lemmatization

**Lemmatization** is a text processing technique used in **Natural Language Processing (NLP)** that reduces words to their **base or root form** (called a **lemma**), taking into account the word's meaning and context.

1. Explain Stemming

Ans **Stemming** is a text processing technique used in **Natural Language Processing (NLP)** to reduce words to their **root form** or **stem**

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

Ans **Part-of-speech (POS) tagging** is the process of assigning a part of speech to each word in a sentence.

1. Explain Chunking or shallow parsing

Ans **Chunking**, also known as **shallow parsing**, is a technique in **Natural Language Processing (NLP)** that involves grouping words into **chunks** based on their **syntactic roles**

1. Explain Noun Phrase (NP) chunking

**Noun Phrase (NP) Chunking** is a process in **Natural Language Processing (NLP)** that involves identifying and extracting noun phrases from a sentence

1. Explain Named Entity Recognition

**Named Entity Recognition (NER)** is a fundamental task in **Natural Language Processing (N NLP)** that involves identifying and classifying named entities in text into predefined categories such as **names of persons, organizations, locations, dates, monetary values**, and more.