1. What are Corpora?

**Ans : A corpus is a collection of authentic text or audio organized into datasets. In natural language processing, a corpus contains text and speech data that can be used to train AI and machine learning systems.**

1. What are Tokens?

**Ans : A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence.**

1. What are Unigrams, Bigrams, Trigrams?

**Ans : A 1-gram (or unigram) is a one-word sequence. A 2-gram (or bigram) is a two-word sequence of words, like “I love”, “love reading”, or “Analytics Vidhya”. And a 3-gram (or trigram) is a three-word sequence of words like “I love reading”, “about data science” or “on Analytics Vidhya”.**

1. How to generate n-grams from text?

**Ans : # Creating a function to generate N-Grams**

**def generate\_ngrams(text, WordsToCombine):**

**words = text.split()**

**output = []**

**for i in range(len(words)- WordsToCombine+1):**

**output.append(words[i:i+WordsToCombine])**

**return output**

**# Calling the function**

**generate\_ngrams(text='this is a very good book to study', WordsToCombine=3)**

1. Explain Lemmatization

**Ans : Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma .**

1. Explain Stemming

**Ans : This is the idea of reducing different forms of a word to a core root. Words that are derived from one another can be mapped to a central word or symbol, especially if they have the same core meaning.**

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

**Ans : POS tagging is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. This task is not straightforward, as a particular word may have a different part of speech based on the context in which the word is used.**

1. Explain Chunking or shallow parsing

**Ans : Chunking is a process of extracting phrases from unstructured text, which means analyzing a sentence to identify the constituents(Noun Groups, Verbs, verb groups, etc.) However, it does not specify their internal structure, nor their role in the main sentence. It works on top of POS tagging.**

1. Explain Noun Phrase (NP) chunking

**Ans : Noun chunks are “base noun phrases” – flat phrases that have a noun as their head. You can think of noun chunks as a noun plus the words describing the noun – for example, “the lavish green grass” or “the world's largest tech fund”. To get the noun chunks in a document, simply iterate over Doc.**

1. Explain Named Entity Recognition

**Ans : In simple words, Named Entity Recognition is the process of detecting the named entities such as person names, location names, company names, etc from the text.**