1. **What is Corpora?**

A corpus is a collection of authentic text or audio organized into datasets. Authentic here means text written or audio spoken by a native of the language or dialect. A corpus can be made up of everything from newspapers, novels, recipes, radio broadcasts to television shows, movies, and tweets.

1. **What are Tokens?**

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?**

* Single word is called Unigrams.
* Pair of words is called Bigrams.
* Three words bundled together make a Trigram.

1. **How to generate n-grams from text?**

We can use the nltk library to generate n-grams, by removing punctuations and stop words after that using list comprehension to generate the desired n-grams.

1. **Explain Lemmatization.**

Lemmatization takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

1. **Explain Stemming.**

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

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

It is a process of converting a sentence to forms – list of words, list of tuples (where each tuple is having a form (word, tag)). The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

1. **Explain Chunking or shallow parsing.**

Shallow parsing (also chunking or light parsing) is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc).

1. **Explain Noun Phrase (NP) chunking.**

Noun phrase chunking deals with extracting the noun phrases from a sentence. While NP chunking is much simpler than parsing, it is still a challenging task to build an accurate and very efficient NP chunker. The importance of NP chunking derives from the fact that it is used in many applications.

1. **Explain Named Entity Recognition.**

The named entity recognition (NER) is one of the most data preprocessing tasks. It involves the identification of key information in the text and classification into a set of predefined categories. An entity is basically the thing that is consistently talked about or referred to in the text.