1. **What are Corpora?**

A corpus is a collection of texts. We call it a corpus (plural: corpora) when we use it for language research. That makes your class's essays a corpus - a small one.

1. **What are Tokens?**

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.

1. **What are Unigrams, Bigrams, Trigrams?**

A 1-gram (or unigram) is a one-word sequence. For the above sentence, the unigrams would simply be: “I”, “love”, “listening”, “Krisk Naik”, “about”, “data”, “science”, “on”, “Ineuron”, “Youtube”,”channel”

A 2-gram (or bigram) is a two-word sequence of words, like “I love”, “love reading”, or “Ineuron Intelligence”.

And a 3-gram (or trigram) is a three-word sequence of words like “I love reading”, “about data science”

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

Instead of using pure Python functions, we can also get help from some natural language processing libraries such as the [Natural Language Toolkit (NLTK)](https://www.nltk.org/). In particular, nltk has the ngrams function that returns a generator of n-grams given a tokenized sentence.

1. **Explain Lemmatization**

Lemmatization is the grouping together of different forms of the same word.

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. Stemming is important in natural language understanding (NLU) and natural language processing (NLP).

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

Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.

1. **Explain Chunking or shallow parsing**

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**

Dividing sentences into non-overlapping phrases is called text chunking. NP chunking deals with a part of this task: it involves recognizing the chunks that consist of noun phrases (NPs).

1. **Explain Named Entity Recognition**

The named entity recognition (NER) is one of the most data preprocessing task. 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 refer to in the text. NER is the form of NLP