1. **What are Corpora?**

A corpus can be defined as a collection of text documents.

1. **What are Tokens?**

Tokens are the building blocks of Natural Language. Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.

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

* When the machine parses the text one word at a time, then it is a **unigram**.
* When the text is parsed two words at a time, it is a **bigram**.
* The set of words is a **trigram** when the machine parses three words at a time.

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

**import re**

def generate\_ngrams(s, n):

# Convert to lowercases

s = s.lower()

# Replace all none alphanumeric characters with spaces

s = re.sub(r'[^a-zA-Z0-9\s]', ' ', s)

# Break sentence in the token, remove empty tokens

tokens = [token for token in s.split(" ") if token != ""]

# Use the zip function to help us generate n-grams

# Concatentate the tokens into ngrams and return

ngrams = zip(\*[token[i:] for i in range(n)])

return [" ".join(ngram) for ngram in ngrams]

1. **Explain Lemmatization**

The method of mapping all the various forms of a word to its base word (also called “lemma”) is known as Lemmatization. Although this may appear close to the definition of stemming, these are actually different. For instance, the word “better,” after stemming, remains the same. However, upon lemmatization, this should become “good,”. Lemmatization needs greater linguistic knowledge. Modelling and developing efficient lemmatizers still remains an open problem in NLP research.

1. **Explain Stemming**

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers.

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

* Parts of speech tagging better known as [POS tagging](https://www.mygreatlearning.com/blog/pos-tagging/) refers to the process of identifying specific words in a document and group them as part of speech, based on its context. POS tagging is also known as grammatical tagging since it involves understanding grammatical structures and identifying the respective component.
* POS tagging is a complicated process since the same word can be different parts of speech depending on the context. The same generic process used for word mapping is quite ineffective for POS tagging because of the same reason.

1. **Explain Chunking or shallow parsing**

Shallow parsing (also chunking, "light parsing") is an analysis of a sentence which identifies the constituents (noun groups, verbs, verb groups, etc.), but does not specify their internal structure, nor their role in the main sentence

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

A noun phrase is a phrase that contains a noun and operates, as a unit, as a noun.

1. **Explain Named Entity Recognition**

Named Entity Recognition is the process of detecting the named entities such as person names, location names, company names, etc from the text.