1. 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.
2. 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.
3. A 1-gram (or unigram) is a one-word sequence. For the above sentence, the unigrams would simply be: “I”, “love”, “reading”, “blogs”, “about”, “data”, “science”, “on”, “Analytics”, “Vidhya”. A 2-gram (or bigram) is a two-word sequence of words, like “I love”, “love reading”. And a 3-gram (or trigram) is a three-word sequence of words like “I love reading”, “about data science”.
4. 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. 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 .
2. 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).
3. The process of assigning one of the parts of speech to the given word. It is generally called POS tagging. In simple words, we can say that POS tagging is a task of labelling each word in a sentence with its appropriate part of speech. We already know that parts of speech include nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories.
4. 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.).
5. 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 a accurate and very efficient NP chunker. The importance of NP chunking derives from the fact that it is used in many applications.
6. 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**.