1. What are Corpora?

Collection of texts.

1. What are Tokens?

Tokens are small units or building blocks in nlp. They can be words, characters or subwords.

1. What are Unigrams, Bigrams, Trigrams?

Also known as N-grams, they are continuous sequences of words or tokens in a document.

1. How to generate n-grams from text?

Depending on the value of “n”, we generate sequences of length “n” from the document.

If the document is “ Data Science is cool”, a bigram is basically all possible pairs of words which are ‘neighbors’ in the document. E.g [‘Data Science’, ‘Science is’, ‘is cool’].

1. Explain Lemmatization

Lemmatization is a nlp pre-processing technique of grouping together different inflected forms of a word for analysis. E.g ‘Run’, ‘Runs’, ‘Ran’ can all be lemmatized to ‘run’.

1. Explain Stemming

Stemming in nlp is used to reduce inflected forms of words to their stem version. While it sounds similar to stemming, it is a crude way of just chopping off ends of a word and sometimes ending with words which are not actual words.

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

POS tagging is the process of labelling each word in the sentence or paragraph with it’s appropriate Part of Speech, also known as it’s syntactic function in the sentence. Example of POS are nouns, pronouns, adjective, verbs.

1. Explain Chunking or shallow parsing

A process of dividing a text into syntactically related group, or chunks. E.g [NP My Dog] [VP likes] [NP his food].

1. Explain Noun Phrase (NP) chunking

Chunking to only return Noun Phrases (NP) from a sentence and ignore the rest of the chunks or groups.

1. Explain Named Entity Recognition

It is a subtask of information extraction which seeks to locate and classify all the Named Entitles in a text.