1. What are Corpora?

Ans: Corpus or Corpora is a collection of text or paragraphs which can be used to train a model or extract words out of it and convert them into vectors. Often the data is collected in the form of corpus where we have a large amount of text.

1. What are Tokens?

Ans: Tokens are words in the corpus. Tokenization is performed to extract the tokens/words so that further text pre processing steps can be performed such as Stopwords, Stemming, Lemmatization or word2vec.

1. What are Unigrams, Bigrams, Trigrams?

Ans: these are types of n-grams which are a sequence of continuous words. Unigram has one word. Bigrams is a sequence of 2 words and trigrams is a sequence of 3 words.

1. How to generate n-grams from text?
2. Explain Lemmatization

Ans: Lemmatization is used to reduce words into their base words which are meaningful as they are dictionary words. That reduces the number of duplicate words that need to be converted into vectors.

1. Explain Stemming

Ans: Stemming is also used to reduce words into their root/base words but it is done by removing suffixes. Therefore, the words may not be always meaningful and that is what makes it different from Lemmatization.

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

Ans: it is used to tag verbs, adverb, adjective in a sentence.

1. Explain Chunking or shallow parsing

Ans: it is used to group words into chunks which are meaningful. For example, the sentence is a cat sat on the mat can be chunked into [a cat] [sat] [on] [the mat]

1. Explain Noun Phrase (NP) chunking

Ans: Noun phrases are composed of nouns and its modifiers that describe or qualify nouns. For example: The big brown dog ran down the street. The noun is dog and its modifier is big brown. This can be used in NLP tasks such as sentiment analysis or information retrieval.

1. Explain Named Entity Recognition:

Ans: NER is used to identify the names of organization, people or other types of entities from a large corpus of data. Before we apply any ML or DL algorithm to train a model, we have to perform text pre processing which involves tokenization, stopwords, lemmatization/stemming and converting words into vectors by employing any method for conversion such as BOW, TF-IDF, Embedding layer.