Tokenization and Basic Terminologies

Corpus: It just refers to a paragraph.

Documents: It refers to sentences.

Vocabulary: Unique words in a corpus.

Words: words in corpus.

# Tokenization

Tokenization is the process of converting a text (in the form of corpus or documents) into tokens (which can be in form of sentences or words). It is a text preprocessing technique.