Text Analytics

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

In [26]:

```
import nltk
nltk.download("all")
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
[nltk_data]    | Downloading package abc to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_eng to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_eng is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_ru is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_rus to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_rus is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package basque_grammars to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package basque_grammars is already up-to-date!
[nltk_data]    | Downloading package bcp47 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package bcp47 is already up-to-date!
[nltk_data]    | Downloading package biocreative_ppi to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package biocreative_ppi is already up-to-date!
[nltk_data]    | Downloading package bllip_wsj_no_aux to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package bllip_wsj_no_aux is already up-to-date!
[nltk_data]    | Downloading package book_grammars to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package book_grammars is already up-to-date!
[nltk_data]    | Downloading package brown to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package brown_tei to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown_tei is already up-to-date!
[nltk_data]    | Downloading package cess_cat to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cess_cat is already up-to-date!
[nltk_data]    | Downloading package cess_esp to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cess_esp is already up-to-date!
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package city_database to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package city_database is already up-to-date!
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package comparative_sentences to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package comparative_sentences is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package comtrans to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package comtrans is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package conll2007 to
```

```
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2007 is already up-to-date!
[nltk_data]    | Downloading package crubadan to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package crubadan is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package dolch is already up-to-date!
[nltk_data]    | Downloading package english_wordnet to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package english_wordnet is already up-to-date!
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package europarl_raw is already up-to-date!
[nltk_data]    | Downloading package extended_omw to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package extended_omw is already up-to-date!
[nltk_data]    | Downloading package floresta to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package floresta is already up-to-date!
[nltk_data]    | Downloading package framenet_v15 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package framenet_v15 is already up-to-date!
[nltk_data]    | Downloading package framenet_v17 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package framenet_v17 is already up-to-date!
[nltk_data]    | Downloading package gazetteers to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package ieer to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ieer is already up-to-date!
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package indian to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package indian is already up-to-date!
[nltk_data]    | Downloading package jeita to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package jeita is already up-to-date!
[nltk_data]    | Downloading package kimmo to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package kimmo is already up-to-date!
[nltk_data]    | Downloading package knbc to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package knbc is already up-to-date!
[nltk_data]    | Downloading package large_grammars to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package lin_thesaurus to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package lin_thesaurus is already up-to-date!
[nltk_data]    | Downloading package mac_morpho to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package mac_morpho is already up-to-date!
[nltk_data]    | Downloading package machado to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package machado is already up-to-date!
[nltk_data]    | Downloading package masc_tagged to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package masc_tagged is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker_tab to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_ne_chunker_tab is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_treebank_pos_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger_tab to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_treebank_pos_tagger_tab is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package moses_sample to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package moses_sample is already up-to-date!
```

```
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package mte_teip5 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package mte_teip5 is already up-to-date!
[nltk_data]    | Downloading package mwa_ppdb to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package mwa_ppdb is already up-to-date!
[nltk_data]    | Downloading package names to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package names is already up-to-date!
[nltk_data]    | Downloading package nombank.1.0 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package nombank.1.0 is already up-to-date!
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data]    | Downloading package nps_chat to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package nps_chat is already up-to-date!
[nltk_data]    | Downloading package omw to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package omw-1.4 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package omw-1.4 is already up-to-date!
[nltk_data]    | Downloading package opinion_lexicon to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package opinion_lexicon is already up-to-date!
[nltk_data]    | Downloading package panlex_swadesh to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package panlex_swadesh is already up-to-date!
[nltk_data]    | Downloading package paradigms to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package paradigms is already up-to-date!
[nltk_data]    | Downloading package pe08 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pe08 is already up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package perluniprops is already up-to-date!
[nltk_data]    | Downloading package pil to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pil is already up-to-date!
[nltk_data]    | Downloading package pl196x to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pl196x is already up-to-date!
[nltk_data]    | Downloading package porter_test to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package porter_test is already up-to-date!
[nltk_data]    | Downloading package ppattach to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ppattach is already up-to-date!
[nltk_data]    | Downloading package problem_reports to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package problem_reports is already up-to-date!
[nltk_data]    | Downloading package product_reviews_1 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package product_reviews_1 is already up-to-date!
[nltk_data]    | Downloading package product_reviews_2 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package product_reviews_2 is already up-to-date!
[nltk_data]    | Downloading package propbank to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package propbank is already up-to-date!
[nltk_data]    | Downloading package pros_cons to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pros_cons is already up-to-date!
[nltk_data]    | Downloading package ptb to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ptb is already up-to-date!
[nltk_data]    | Downloading package punkt to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package punkt_tab to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package punkt_tab is already up-to-date!
[nltk_data]    | Downloading package qc to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package qc is already up-to-date!
[nltk_data]    | Downloading package reuters to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package reuters is already up-to-date!
[nltk_data]    | Downloading package rslp to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package rte to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
```

```
[nltk_data]    |   Package rte is already up-to-date!
[nltk_data]    | Downloading package sample_grammars to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sample_grammars is already up-to-date!
[nltk_data]    | Downloading package semcor to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package semcor is already up-to-date!
[nltk_data]    | Downloading package senseval to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package senseval is already up-to-date!
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package sentiwordnet to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sentiwordnet is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package spanish_grammars to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package spanish_grammars is already up-to-date!
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package tagsets is already up-to-date!
[nltk_data]    | Downloading package tagsets_json to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package tagsets_json is already up-to-date!
[nltk_data]    | Downloading package timit to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package timit is already up-to-date!
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package toolbox is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package udhr to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr is already up-to-date!
[nltk_data]    | Downloading package udhr2 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr2 is already up-to-date!
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package unicode_samples is already up-to-date!
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package universal_tagset is already up-to-date!
[nltk_data]    | Downloading package universal_treebanks_v20 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package universal_treebanks_v20 is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package vader_lexicon is already up-to-date!
[nltk_data]    | Downloading package verbnet to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package verbnet is already up-to-date!
[nltk_data]    | Downloading package verbnet3 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package verbnet3 is already up-to-date!
```

```
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package webtext is already up-to-date!
[nltk_data]    | Downloading package wmt15_eval to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wmt15_eval is already up-to-date!
[nltk_data]    | Downloading package word2vec_sample to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package word2vec_sample is already up-to-date!
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet2021 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet2021 is already up-to-date!
[nltk_data]    | Downloading package wordnet2022 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet2022 is already up-to-date!
[nltk_data]    | Downloading package wordnet31 to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet31 is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package words is already up-to-date!
[nltk_data]    | Downloading package ycoe to
[nltk_data]    |     C:\Users\PL\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ycoe is already up-to-date!
[nltk_data]    |
[nltk_data]  Done downloading collection all
```

Out[26]:

```
True
```

Tokenization

In [27]:

```python
from nltk import word_tokenize, sent_tokenize
```

In [28]:

```python
corpus = "Sachin was the GOAT of the previous generation. Virat is the GOAT of this generation. Shubham will be
```

In [29]:

```python
print(word_tokenize(corpus))
print(sent_tokenize(corpus))
```

```
['Sachin', 'was', 'the', 'GOAT', 'of', 'the', 'previous', 'generation', '.', 'Virat', 'is', 'the', 'GOAT', 'of'
, 'this', 'generation', '.', 'Shubham', 'will', 'be', 'the', 'GOAT', 'of', 'the', 'next', 'generation']
['Sachin was the GOAT of the previous generation.', 'Virat is the GOAT of this generation.', 'Shubham will be t
he GOAT of the next generation']
```

POS tagging

In [30]:

```python
from nltk import pos_tag
```

In [31]:

```python
tokens = word_tokenize(corpus)
print(pos_tag(tokens))
```

```
[('Sachin', 'NNP'), ('was', 'VBD'), ('the', 'DT'), ('GOAT', 'NNP'), ('of', 'IN'), ('the', 'DT'), ('previous', '
JJ'), ('generation', 'NN'), ('.', '.'), ('Virat', 'NNP'), ('is', 'VBZ'), ('the', 'DT'), ('GOAT', 'NNP'), ('of',
'IN'), ('this', 'DT'), ('generation', 'NN'), ('.', '.'), ('Shubham', 'NNP'), ('will', 'MD'), ('be', 'VB'), ('th
e', 'DT'), ('GOAT', 'NNP'), ('of', 'IN'), ('the', 'DT'), ('next', 'JJ'), ('generation', 'NN')]
```

Stop word removal

In [32]:

```python
from nltk.corpus import stopwords
stop_words = set(stopwords.words("english"))
stop_words
```

Out[32]:

```
{'a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'ain',
 'all',
 'am',
 'an',
 'and',
 'any',
 'are',
 'aren',
 "aren't",
        'as',
        'at',
        'be',
        'because',
        'been',
        'before',
        'being',
        'below',
        'between',
        'both',
        'but',
        'by',
        'can',
        'couldn',
        "couldn't",
        'd',
        'did',
        'didn',
        "didn't",
        'do',
        'does',
        'doesn',
        "doesn't",
        'doing',
        'don',
        "don't",
        'down',
        'during',
        'each',
        'few',
        'for',
        'from',
        'further',
        'had',
        'hadn',
        "hadn't",
        'has',
        'hasn',
        "hasn't",
        'have',
        'haven',
        "haven't",
        'having',
        'he',
        "he'd",
        "he'll",
        "he's",
        'her',
        'here',
        'hers',
        'herself',
        'him',
        'himself',
        'his',
        'how',
        'i',
        "i'd",
        "i'll",
        "i'm",
        "i've",
        'if',
        'in',
        'into',
        'is',
        'isn',
        "isn't",
        'it',
        "it'd",
        "it'll",
        "it's",
        'its',
        'itself',
        'just',
        'll',
        'm',
        'ma',
        'me',
        'mightn',
        "mightn't",
        'more',
        'most',
        'mustn',
        "mustn't",
        'my',
        'myself',
        'needn',
        "needn't",
        'no',
        'nor',
        'not',
        'now',
        'o',
        'of',
        'off',
        'on',
        'once',
        'only',
        'or',
        'other',
        'our',
        'ours',
        'ourselves',
        'out',
        'over',
        'own',
        're',
        's',
        'same',
        'shan',
        "shan't",
        'she',
        "she'd",
        "she'll",
        "she's",
        'should',
        "should've",
        'shouldn',
        "shouldn't",
        'so',
        'some',
        'such',
        't',
        'than',
        'that',
        "that'll",
        'the',
        'their',
        'theirs',
        'them',
        'themselves',
        'then',
        'there',
        'these',
        'they',
        "they'd",
        "they'll",
        "they're",
        "they've",
        'this',
        'those',
        'through',
        'to',
        'too',
        'under',
        'until',
        'up',
        've',
        'very',
        'was',
        'wasn',
        "wasn't",
        'we',
        "we'd",
        "we'll",
        "we're",
        "we've",
        'were',
        'weren',
        "weren't",
        'what',
        'when',
        'where',
        'which',
        'while',
        'who',
        'whom',
        'why',
        'will',
        'with',
        'won',
        "won't",
        'wouldn',
        "wouldn't",
        'y',
        'you',
        "you'd",
        "you'll",
        "you're",
         "you've",
         'your',
         'yours',
         'yourself',
         'yourselves'}
```

In [33]:

```
tokens = word_tokenize(corpus)
cleaned_tokens = []
for token in tokens:
  if (token not in stop_words):
    cleaned_tokens.append(token)
print(cleaned_tokens)
```

```
['Sachin', 'GOAT', 'previous', 'generation', '.', 'Virat', 'GOAT', 'generation', '.', 'Shubham', 'GOAT', 'next'
, 'generation']
```

Stemming

In [34]:

```
from nltk.stem import PorterStemmer
```

In [35]:

```
stemmer = PorterStemmer()
```

In [36]:

```
stemmed_tokens = []
for token in cleaned_tokens:
  stemmed = stemmer.stem(token)
  stemmed_tokens.append(stemmed)
print(stemmed_tokens)
```

```
['sachin', 'goat', 'previou', 'gener', '.', 'virat', 'goat', 'gener', '.', 'shubham', 'goat', 'next', 'gener']
```

Lemmatization

In [37]:

```
from nltk.stem import WordNetLemmatizer
```

In [38]:

```
lemmatizer = WordNetLemmatizer()
```

In [39]:

```
lemmatized_tokens = []
for token in cleaned_tokens:
  lemmatized = lemmatizer.lemmatize(token)
  lemmatized_tokens.append(lemmatized)
print(lemmatized_tokens)
```

```
['Sachin', 'GOAT', 'previous', 'generation', '.', 'Virat', 'GOAT', 'generation', '.', 'Shubham', 'GOAT', 'next'
, 'generation']
```

TF-IDF

In [40]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.probability import FreqDist
```

In [41]:

```
corpus = [
    "Sachin was the GOAT of the previous generation",
    "Virat is the GOAT of the this generation",
    "Shubham will be the GOAT of the next generation"]
```

In [42]:

```
vectorizer = TfidfVectorizer()
```

In [43]:

```
vectorizer.fit(tokens)
fdist=FreqDist(tokens)
for word,frequency in fdist.items():
    print(f"{word}:{frequency}")
```

```
Sachin:1
was:1
the:5
GOAT:3
of:3
previous:1
generation:3
.:2
Virat:1
is:1
this:1
Shubham:1
will:1
be:1
next:1
```

In [44]:

```
tfidf_matrix = vectorizer.transform(tokens)
print(tfidf_matrix)
```

```
  (0, 7)	1.0
  (1, 12)	1.0
  (2, 9)	1.0
  (3, 2)	1.0
  (4, 5)	1.0
  (5, 9)	1.0
  (6, 6)	1.0
  (7, 1)	1.0
  (9, 11)	1.0
  (10, 3)	1.0
  (11, 9)	1.0
  (12, 2)	1.0
  (13, 5)	1.0
  (14, 10)	1.0
  (15, 1)	1.0
  (17, 8)	1.0
  (18, 13)	1.0
  (19, 0)	1.0
  (20, 9)	1.0
  (21, 2)	1.0
  (22, 5)	1.0
  (23, 9)	1.0
  (24, 4)	1.0
  (25, 1)	1.0
```

In [45]:

```
print(vectorizer.get_feature_names_out())
```

```
['be' 'generation' 'goat' 'is' 'next' 'of' 'previous' 'sachin' 'shubham'
 'the' 'this' 'virat' 'was' 'will']
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js