**Experiment No: 07**

**Aim: Text Analytics**

**Problem Statement:**

1. Extract Sample document and apply following document preprocessing methods:

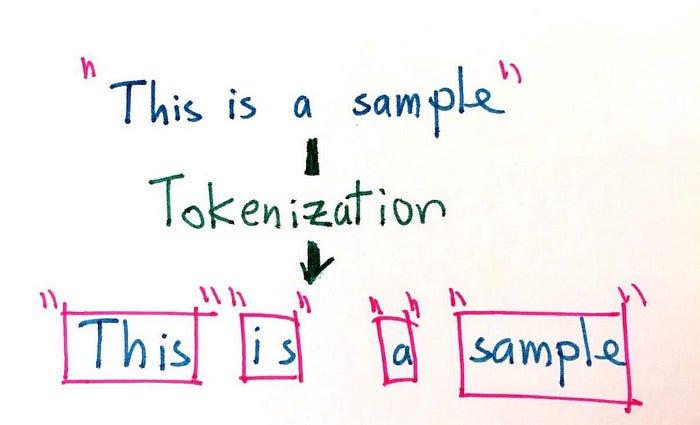
Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

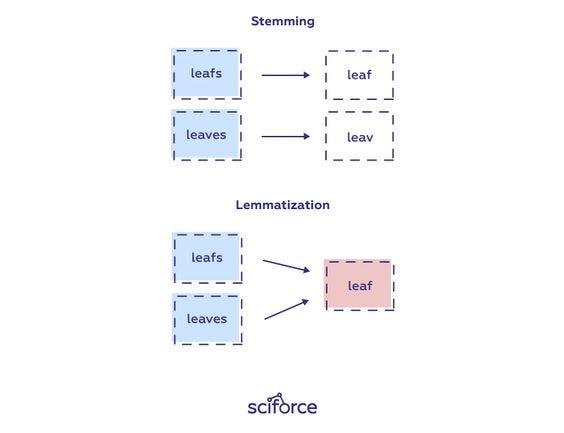
2. Create representation of documents by calculating Term Frequency and Inverse

DocumentFrequency.

**Theory:**

Tokenization is the process of breaking down the given text in natural language processing into the smallest unit in a sentence called a token. Punctuation marks, words, and numbers can be considered tokens.

****



Stemming is definitely the simpler of the two approaches. With stemming, words are reduced to their word stems. A word stem need not be the same root as a dictionary-based morphological root, it just is an equal to or smaller form of the word.

The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead, it uses lexical knowledge bases to get the correct base forms of words.

**Conclusion:** Hence we have extracted sample documents and applied document preprocessing methods like Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.