LP2- Assignment -A4

Title:- Text analysis

Problem statement:- Consider a suitable text analysis dataset. Remove stop words, apply stemming & feature selection technique to represent docs as vectors, classify docs & evaluate precision, recall.

Objective:-
1) Implement problem statement using python.
2) Perform text classification after preprocessing.

Outcome:- Students will learn to perform text preprocessing before performing classification

S/W & H/W:-
1) Python 3
2) 64 bit OS

Theory:-

1) Stop words:-
- These words refer to the most commonly words in the long sentence.
- Some of the most common stop words are as, the, at, is, and, on, of, etc.
- Stopwords can cause problems when searching for phrases when that include them particularly in phrases such as "The who" "The The", or "take the".

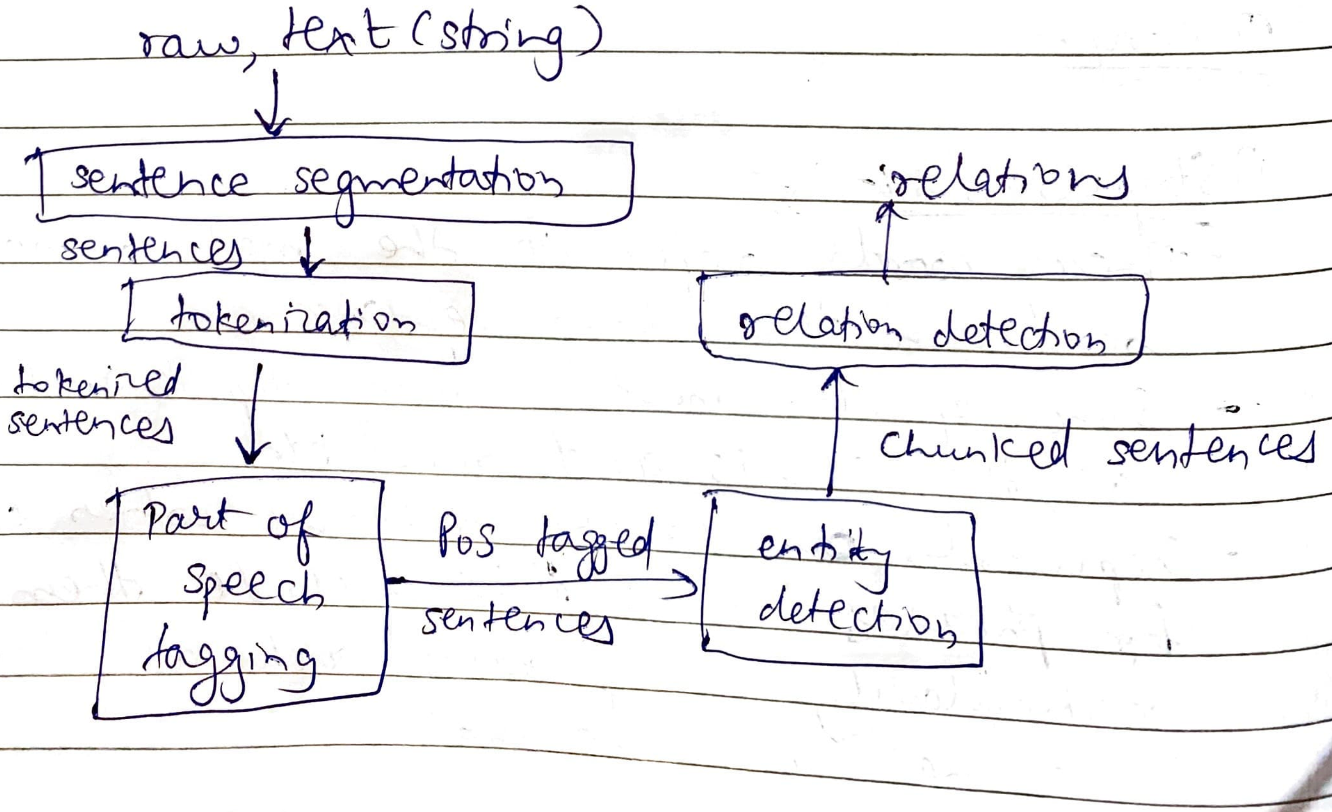
2) Stemming:-

- It is the process of reducing inflected words to their word stem, base of root form - generally.
- The stem need not be identical to the morphological root ~~of~~ the word, it is usually suffivent that releted words map to the same stem, even if this stem is not in itself a valid root.

3) Feature extraction:-

- It is the process of selecting a subset of relevant features for use in model construction. It is used for:-

a) Simplification of models.
b) Shorter training time
c) To avoid curse of dimensionality.
d) Enhanced generalization.



Precision & Recall:-

True positive (TP):- The case when model predicts positive label & actual label is positive.

True negative (TN):- The case when model predicts negative label & the actual label is negative.

False positive (FP):- The case when model predict positive label & actual label is not positive.

False negative (FN):- The case when model predict negative label & actual label is not negative.

Precision:- Ratio of TP & all positives.

$$P = \frac{TP}{TP + FP}$$

Recall:- Ratio of TP and all (TP + FN).

$$R = \frac{TP}{TP + FN}$$

Conclusion:- We have successfully performed text preprocessing steps & performed classification of text documents.