Date __/__/____

Assignment - 4

* **Title:-** Stemming and feature extraction.

* **Problem statement :-** Consider a suitable text data set. Remove stop words apply stemming and feature extraction selection techniques to represent documents as vectors. Classify documents and evaluate precision and recall.

* **Objective:-** ⇒ Implementation of the problem statement using Python.
  ⇒ Remove stop words apply stemming and feature selection.

* **Outcomes :-** Students will be able to :-
  ⇒ Implement the problem statement using python.
  ⇒ Remove stop words apply stemming and feature selection.

* **S/W & H/W Requirements:-**
  ⇒ Fedora 20 | Window 10.
  ⇒ Jupyter Notebook.

* **Theory:-**

Stop words :- In computing, stop words are words which are filtered out before or after processing of text. Through these words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search.

⇒ Stemming :- stemming is the process of reducing inflected for sometimes derived words to their word stem, base or root form – generally a written word form. The stem need not be identical to the morphological root of the word, it is usually sufficient that related words map to the same stem, even if this stem is not itself a valid root.

⇒ Feature Extraction :- In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons :-

1) Simplification of models to make them easier to interpret to use.
2) Shorter training times.
3) To avoid the curse of dimensionality.
4) Enhanced generalization by reducing over fitting (reduction of variance).

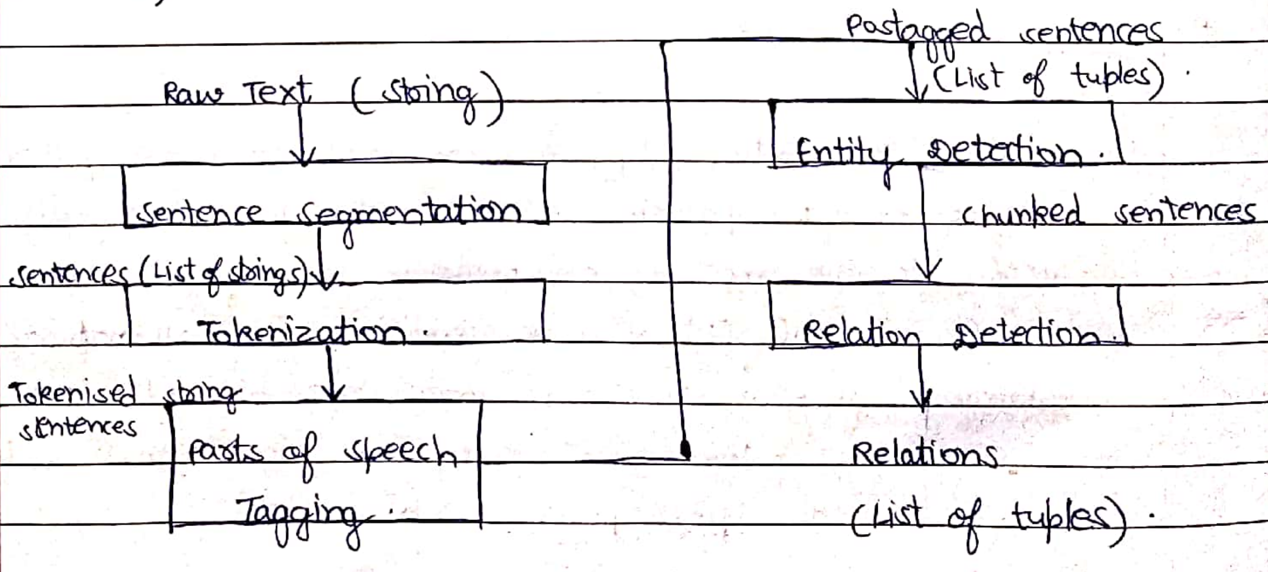

Fig1- Feature Extraction Architecture

* Precision :- Precision mentions the proportions of the positive identifications that was actually correct. It means the percentage of your results that are relevant.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

* Recall :- Recall mentions the proportions of actual positive that were identified correctly. Recall refers to the percentage of total relevant results correctly classified by your algorithm

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

* Conclusion :- We have successfully removed stop words, applied stemming and feature selection techniques to represent documents as vectors and also calculated precision & recall.