Assignment A4 41603

- Problem statement: consider a situation test dataset. Remove stop words, apply stemming & feature selection techniques to represent documents as vectors, classify documents & evaluate precision & recall

- Objectives: - Implementation of the problem statement using python
  - Remove stopwords, apply stemming & feature selection
- Outcomes: students will be able to
  - implement the P.S. using python
  - remove stop words, apply stemming & feature selection.
- S/w & H/w requirements: - Fedora 20/ windows 10
  - Jupyter environment
- 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 the language there is an single universal list. Some tools even specifically avoid removing these stop words to support phrase search.

41403

Stemming: Stemming is the process of reducing inflected words to their word stem base or root form generally a written word form. The stem need not be identical to the morphological root to the word it is usually sufficient that related words map to some stem even if this stem is not in itself a valid root.

- Feature selection: In machine learning & statistics feature selection also known as variable selection is the process of selecting a subset of relevant features for use in model construction. Feature selection techniques are used for 4 reasons
  - simplifications of models to make them easier to interpret by users.
  - shorter training trees.
  - to avoid the use of dimentionality
  - enhanced generalization by reducing overfitting.

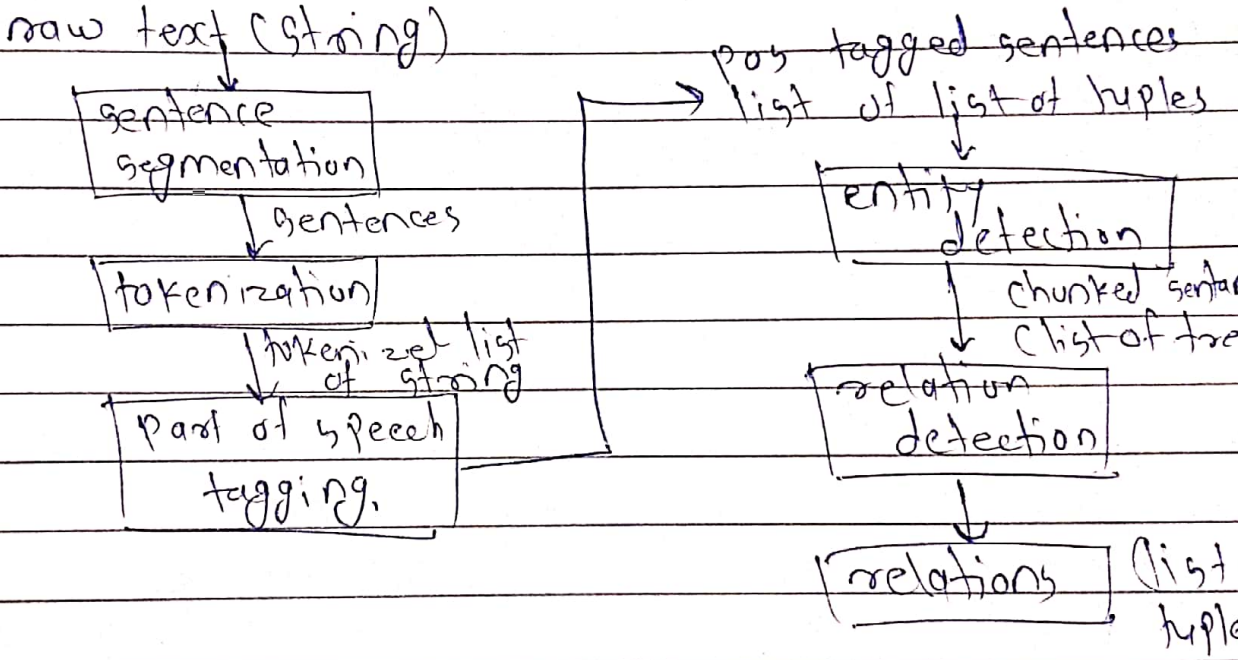
- Precision: It means the proportion of the positive identifications that was actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

41403

- Recall: Recall mentions the proportions of actual positives that were identical correctly.

$$\text{Recall} = \frac{TP}{TP + FN}$$



- conclusion: We have successfully reversed stop words applied stemming & featureselection techniques to represent document of vectors and also calculated precision & recall.