Assignment 4

Title : Stemming and feature selection techniques using vectors

Problem statement : Consider a suitable text. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors, classify documents as vectors precision and recall.

Objective : 1) Implementation of problem using python.
2) Remove stop words, applying stemming and feature selection.

Outcome : 1) Understanding the stemming and feature selection process.
2) Learn about precision and recall.

Theory :

1) stop words :
In computing stop words are words which are filtered out before or after processing of natural language data (text). Though stop words are usually refer to the most common values or words in a language, there is no universal list of stop words used by all

natural processing tools.

Example for search engines :

the , is , at , which, on, etc.

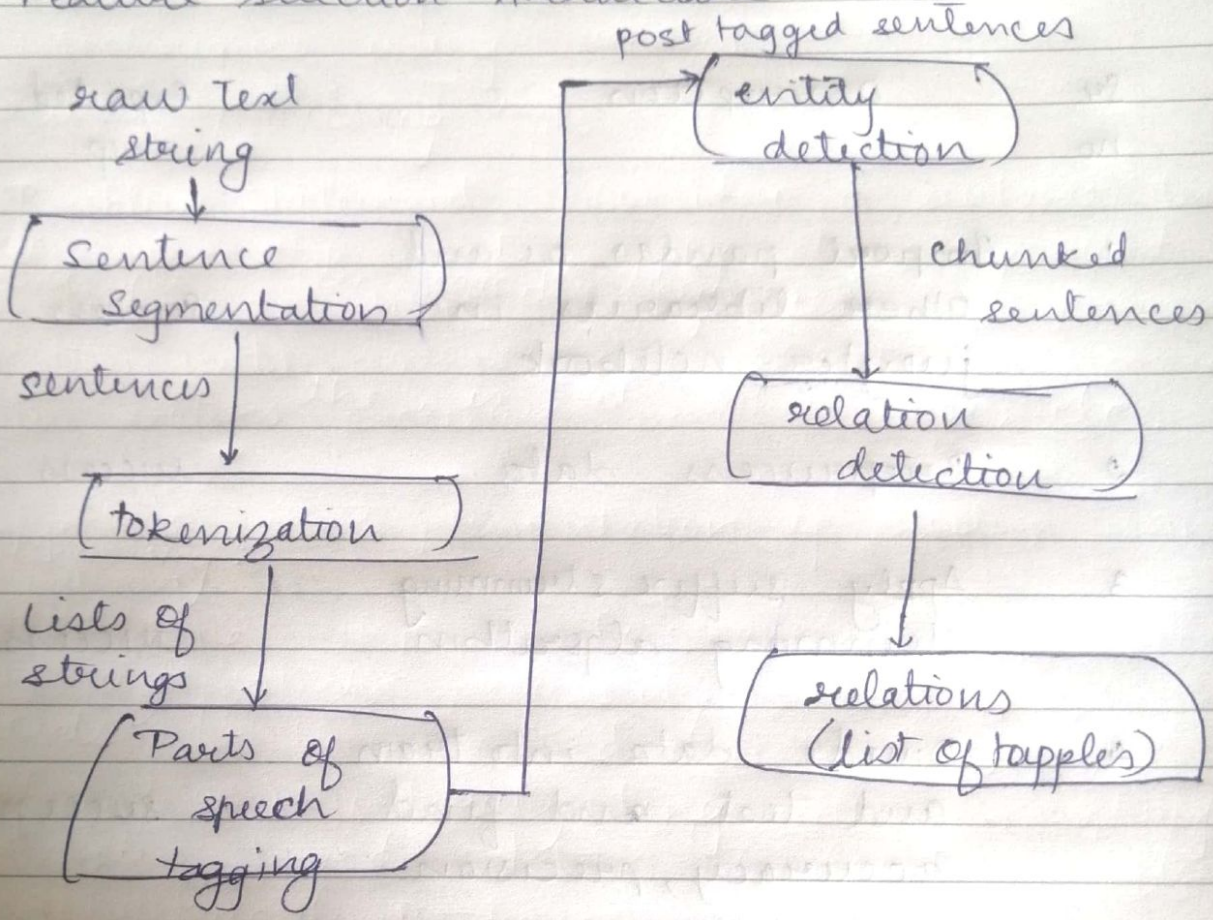
2) Stemming :

Stemming is the process of reducing inflected words to their word stem , base or root from generally a written word form. The stem need not be identical to the morphological root of the word.

Many search engines treat words with same stems as synonyms as a kind of query expansion , a process called conflation

3) Feature Selection :
In machine learning and statistics feature selection also known as variable subset selection , attribute or variable subset selection is a process of selecting a subset of relevant features (variables, predictions) for use in model construction.

Feature Selection Architecture



Reasons for use of feature selection technique.

1) Simplification of models to make easy to interpret

2) Lesser training time

3) To avoid curse of dimensionality

4) Enhanced generalization by reducing overfitting

· Test Cases :

Sr no	Description	Expected o/p	Actual o/p
1	Import pandas, os and other libraries in jupyter notebook	Success	Success
2	Preprocess data	success	success
3	Apply suffix strimming stemming algorithm	success	success
4	Divide data into train and test and find accuracy, precision	success	success

Conclusion : Thus we have studied to remove stop words, apply stemming and feature selection techniques to represent documents as vectors.