Assignment: DMW 4

Title:- Stemming and feature extraction

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

Objectives:
- Implementation of the problem statement using python.
- Remove stop words apply ~~ing~~ stemming & feature selection.

Outcomes:
Students will be able to
- implement the problem statement using python
- remove stop words, apply stemming & feature selection.

Software and Hardware Requirements:
- Fedora 20 / windows 10
- Jupyter Notebook / Google Colab.

Theory:
Stop words:
1) To computing, stop words are words which are filtered out before or after processing of natural language data (text).

Teacher Signature _____

2) Though "stop words" usually refers 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 the stop words to support phrase search
Any group of words can be chosen as stop words for a given purpose some search engines these are some of most common, ~~short~~ short function words such as the, is, at, which, and on. In this case stop words can cause problems when searching for phrases that include them, particularly in name such as "The who" "The The" or "Take that" other search engines remove some.
nltk tool is used to remove stop words in python.

2) Stemming:
1. Stemming is process of reducing inflected words to their word stem base or root from generally a written word form
2. The stem need not be identical to the morphological root of the word it is usually sufficient the related words map the same stem even if its stem is not itself a valid root.
3. Algorithm for stemming have been studied in computer science since 1960s.
4. Many search engines treat words with the same stem as synonyms as a kind of query expansion a process called conflation

Teacher Signature _____

5) A suffix-stripping algorithm is famous for stemming

3) 'Suffix-stripping' algorithms.

1.) suffix stripping algorithms do not rely on the lookup table. that consists of inflected forms and root form relations.

2) Instead, a typically smaller list of "rules" is stored which provides a path for the algorithms given an input word form to find its root form Some example of the rules include

Ex:- if the word include "ed" remove "ed"
if the word include "ly" remove "ly"

4) Feature Extraction

In ML and statistics, feature selection also known as variable selection attribute selection or variable subset selection is the process of selecting a subset of relevant feature for use in model construction.

1) Feature selection techniques are used for four reasons.

2) Simplification of models to make them easier to Interpreate by reserchers / users.

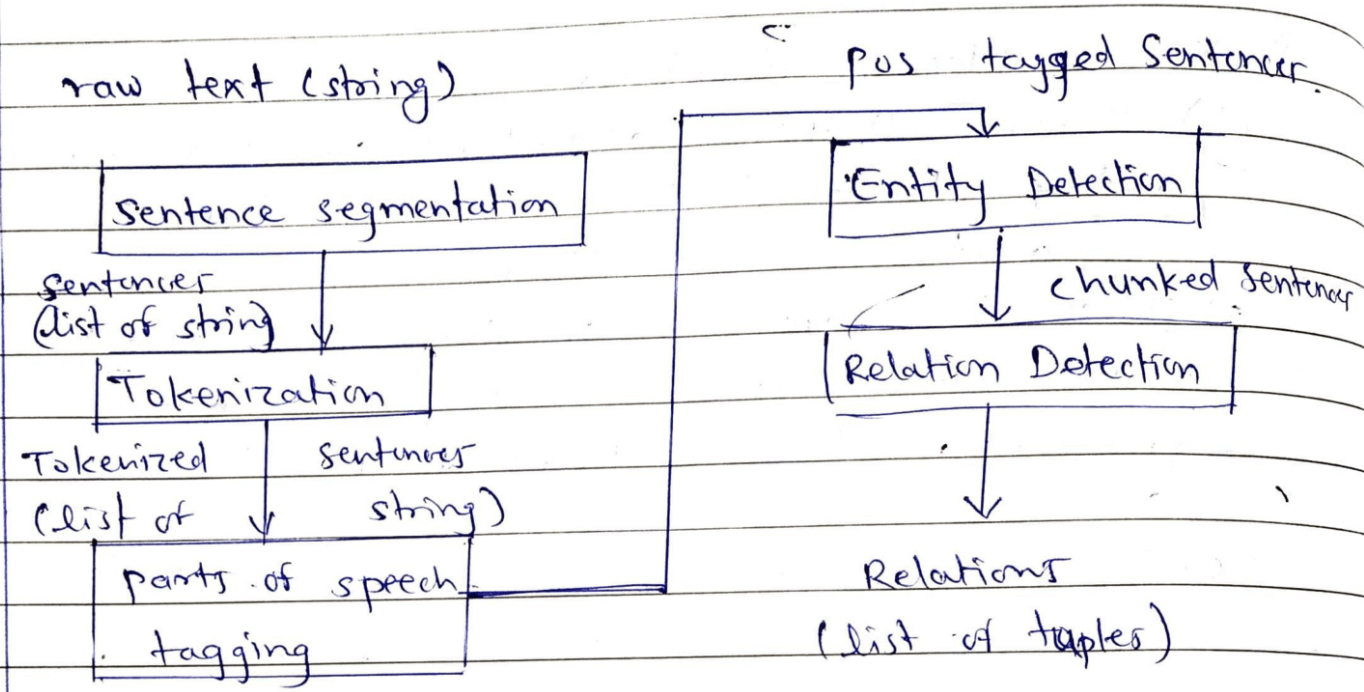
3) Shorter training items.

4) To avoid the curse of dimensionality

5) Enhanced generalization by reducing over fitting

Teacher Signature

Feature extraction architecture



Conclusion:- Thus, we have ~~stt~~ studied to remove stop words apply stemming and feature extraction techniques to represent documents as vectors

Teacher Signature