Page No.
Date

Assignment No: 4

Aim:- Consider a suitable text dataset. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors. Classify documents and evaluate Precision, recall.

Objective

1. Implementation of the Problem Statement using Python.
2. Remove stop words, apply stemming and feature selection.

Theory :-

1 Stop words

1. In Computing, stop words are words which are filtered out before or after processing of natural language data (text).

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 these stop words to support phrase search.

Any group of words can be chosen as the stop words for a given purpose for some search engines, these are some of the most common, 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 names such as "The Who", "The The" or "Take That". other search engines remove some of the most common word- including lexical words, such as "want" - from a query in order to improve performance.

Remove stop words with nltk tool in Python module

```
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
data = " All work and no play makes jack dull boy. All work and no play makes jack a dull boy."
    stopWords = set (stopwords.. words ('english') )
    words = word_tokenize (data )
    wordsFiltered = [ ]
    for W in words :
    if W not in stopWords :
    wordsFiltered . append (W )
    Print ( wordsFiltered )
```

Page No.
Date

2] Stemming

1 Stemming is the process of reducing inflected words to their word stem, base or root form - generally a written word form.

2 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 in itself a valid root.

3 Algorithms for stemming have been studied in computer science since the 1960s.

4 Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

5 A suffix-stripping algorithm is famous for stemming

Code Stemming with the nltk tool in Python module

```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
ps = PorterStemmer()
example_words = ["Python", "Pythoner", "Pythoning", "Pythoned", "Pythonly"]
```

```
for w in example_words:
    Print (PS.Stem(w))
```

3] Suffix Stripping algorithms

1 Suffix Stripping algorithms do not rely on a 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:

if the word ends in 'ed', remove the 'ed'.
if the word ends in 'ing', remove the 'ing'
if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Suffix stripping algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations.

The solutions produced by Suffix Stripping algorithms are limited to those lexical categories which have well known suffixes with few exceptions.

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:

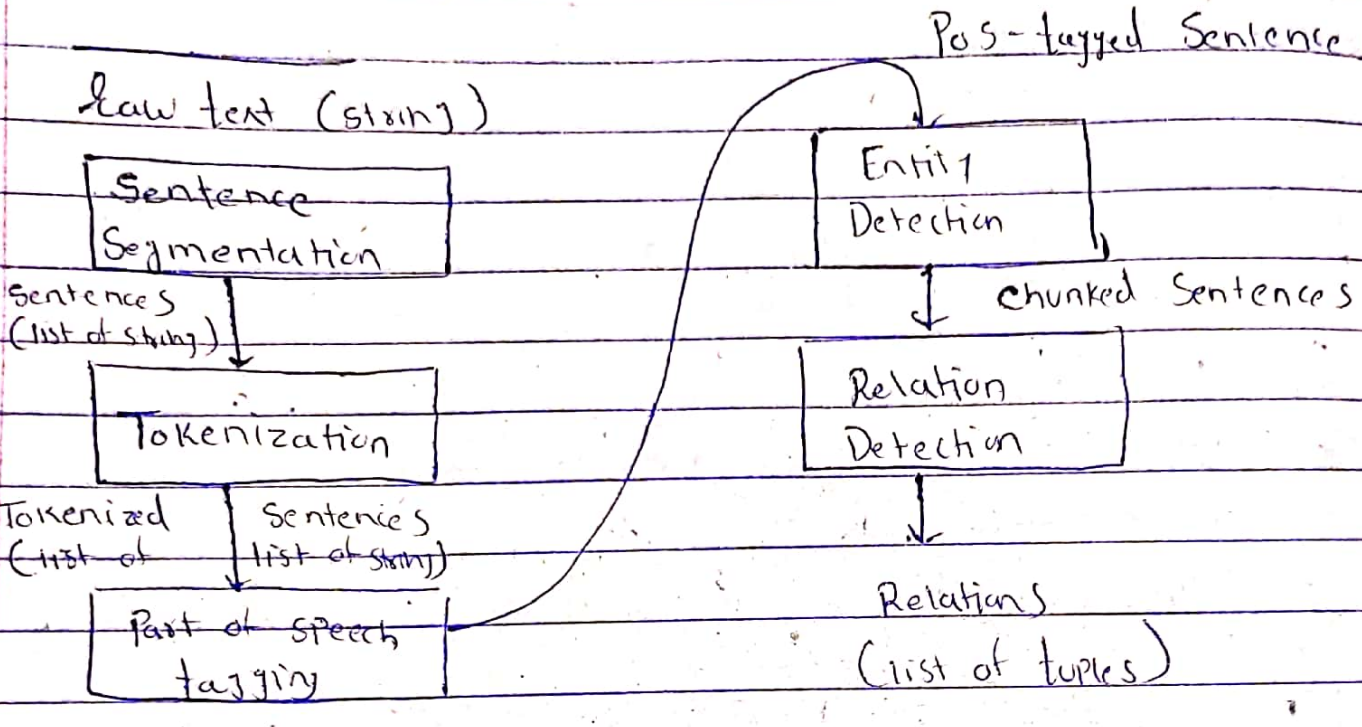
2 Simplification of models to make them easier to interpret by researchers/users

3 Shorter training times

4 To avoid the curse of dimensionality

5 Enhanced generalization by reducing over fitting

- Feature Extraction architecture



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