Assignment No : DMW-4

* Title : Stemming and Feature Selection Techniques.

* Problem statement :

Consider a suitable text dataset. Remove stop words. Apply stemming and feature selection techniques to represent documents as vectors. classify documents and evaluate precission and recall

* Objectives : To,
1) Implemen~~ted~~ the given classification problem using Python.
2) Remove stem words, apply stemming and feature selection.

* Outcomes : Students will be able to,
1) Apply stemming features selection and stopwards elimination.
2) classify text for given model

* S/W and H/W requirements :

Google colab, Python, Google chrome, 8GB RAM, 500 GB HDD.

* Concepts related to theory:

Stopsword : are the words which are filled out before or after processing of data. It refers to the most common words in a language. There is no single list of stopwords used by all NLP Tools.

Any group of word can be chosen as stopwards for a given purpose eg: the, is, at, etc.

Stemming -

It is the process of reducing inflated words from their word stern, base or root form. Generally a written word form. The stem need not be identical to the morphological root of the word, it is actually sufficient that related words map to the same stem even if this stem is not itself a valid root. The suffix stripping algorithm is famous for stemming.

eg: nominate, nomination → nominal
walking, walked → walk.

Suffix stripping algorithm do not rely on lookup tables. Instead, a typically smaller list of 'rules' is stored which provides a path for algorithms, given input word form.

Feature Extraction:

In Machine Learning, feature selection and extraction also known as variable selection, attribute selection or variable subset selection is the process of selecting a subset of relevant features for use in model iteration.

Feature selection is used for following reasons:-

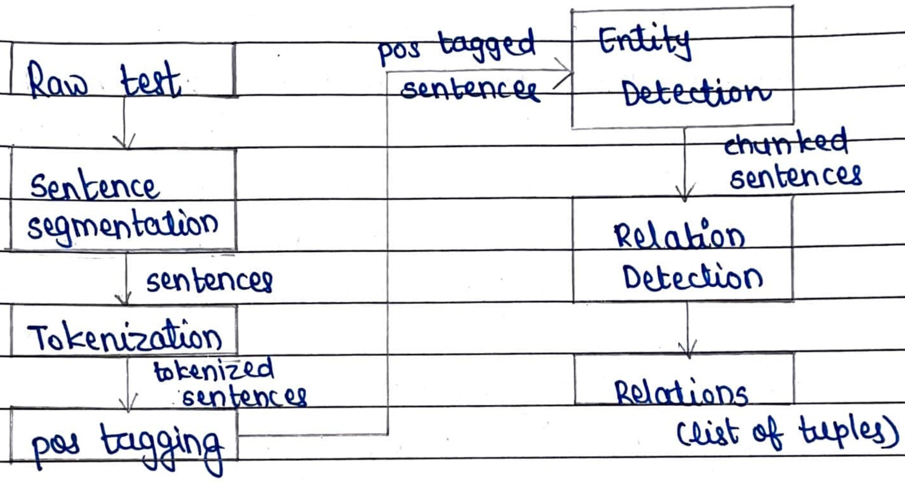
1> Simplification of models to make them easier to interpret.
2> Shorter training times.
3> To avoid dimensionality.
4> Enhancing generalization by reducing overfitting.

* Algorithms-

1> Read, train and test data from excel into dataframes.
2> Remove stopwords from data.
3> Remove special characters from data.
4> Apply stemming.
5> Normalize data.

6> Apply feature selection.
7> Use of classification model to classify documents.
8> Calculate precision and recall.

Feature Selection Architecture.



* conclusion:

Thus, we have successfully implemented text classification by applying removal of top words stemming and feature selection on imDB movie review dataset in python.