Page No.
Date: / /

ASSIGNMENT 4.

- **Title:** Stop Words, Steming & Feature Selection.

- **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 problem statement using Python. Remove Stop words. apply stemming & feture selection.

- **S/W & H/W requirements:**
- 64 bit OS
- Python 3.8
- Jupyter Notebook.
- 64 bit Processor Machine.
- NLTK library & corpus downloaded.

- **Theory:**

Stop words: In computing stop words are words which are filtered out before or after processing of natural language data. Any group of words can be chosen as stop words for a given purpose. The selection

of stop words also affects the performance of algorithm.

Stemming: Stemming is the process of reducing inflated words to their word stem, base & root form. The stem need not be identical to the morphological root of the word, it is usually sufficient that related words map to the same system.

Feature extraction:
In machine learning & statistics, feature selection is also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of features for use in model construction. Feature selection techniques are used for following reasons:

i) Simplification of models to make it easier for user to be interpreted by user.
ii) Shorter training time
iii) To avoid dimensionality problem.
iv) Enhanced generalization by over fitting.

Libraries used:

i) Pandas
ii) Sklearn
iii) NLTK.

Page No
Date: / /

Conclusion:

Thus we have studied to remove stop words, apply stemming & feature selection techniques to classify documents.