**Information Retrieval Systems**

**Lab Practical and date** – Practical 2, Wednesday 29nd July 2020

**Name and Roll Number**- Het Shah, 17BIT103

**Practical Objective**- Generate a collection of atleast 10 documents (i.e. corpus) from minimum two domains , and perform preprocessing task.

* Tokenization
* Stop Words Removal
* Stemming
* Case conversion

**Steps Involved**

● The Tokenization task was performed on an input file that consisted of at least 10 documents (i.e. corpus) from a minimum of two domains(1st domain is of movies and the other is banking system).

● After that, the Stop Words Removal in the same file list is been created to remove all the helping verbs, punctuation, etc in the document.

● In the third step, Stemming is being performed on the statement to bring the words back to their root form by removing the certain suffixes like ‘er’, ’ed’, ’ing’, ’tion’ etc.

● Finally, the case conversion methods were applied to sentences to convert from lower case to upper and vice-versa

**Python Package Used**

* In this experiment we didn’t use any packages. We manually performed the whole process starting from tokenization to case conversion.
* We used inbuild python data structures such as Dictionary and arrays and to take the input we used the I/O operations from the files

**Sample Input/Output**

The input is in the form of corpus consisting of 10 text files by the name of main.py

The output is presented in the form of document term matrix in the form of output.pdf. After each step the output the corpus of words are printed as well.

**Conclusion**

In this practical, we manually coded all the steps for text preprocessing and took into consider 10 different documents consisting of 2 domains. In the end, I also printed the document matrix for the same/