**Information Retrieval Systems**

**Lab Practical and date** – Practical 3, Wednesday 12th August 2020

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

**Practical Objective**- Apply preprocessing techniques ( which you had done in previous experiments)

* Represent term frequencies
* Visualize the data
  + - Plot word frequencies using histogram
    - Show word cloud

**Steps Involved**

1. The Tokenization task was performed on an input file that consisted of a real world data set
2. 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.
3. 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.
4. Finally, the case conversion methods were applied to sentences to convert from lower case to upper and vice-versa
5. After the preprocessing steps were completed, we counted the frequency of each word and then plotted the word cloud and the word frequencies using histrogram
6. Each frequency of the word was also printed.

**Python Package Used**

* Pandas-pandas is a software library written for the Python programming language for data manipulation and analysis
* Mathplotlib- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.
* WordCloud-A tag cloud is a novelty visual representation of text data, typically used to depict keyword metadata on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.
* NLTK- The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

**Sample Input/Output**

The input is in the form of corpus of a file on which the text preprocessing was done. The code in by the name of main.py

The output is presented in the form frequency of each word along with the word count histogram and word cloud representation. There is a separate pdf file for the console output, word-cloud diagram and then word-histogram as well.

**Conclusion**

In this practical, we used the NTLK library for the pre-processing step and used matplotlib and word cloud library to plot the representation.