**Information Retrieval Systems**

**Lab Practical and date** – Practical 6, Wednesday 30th September 2020

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

**Practical Objective**- Perform document clustering.

Input: Document collection (without class information)

Output: Cluster of documents

**Steps Involved**

* We input various documents of the topic blockchain, ai, movies and bank and add them to the corpus
* Various preprocessing steps such as stop word removal, case conversion, stemming, tokenization, punctuation were removed.
* The corpus was splin into training data and testing data in the ratio of 80 percent training and 20 percent testing
* We used the inbuild k-means library of sklearn and we train the model
* We validate the model on the testing set.

**Python Package Used**

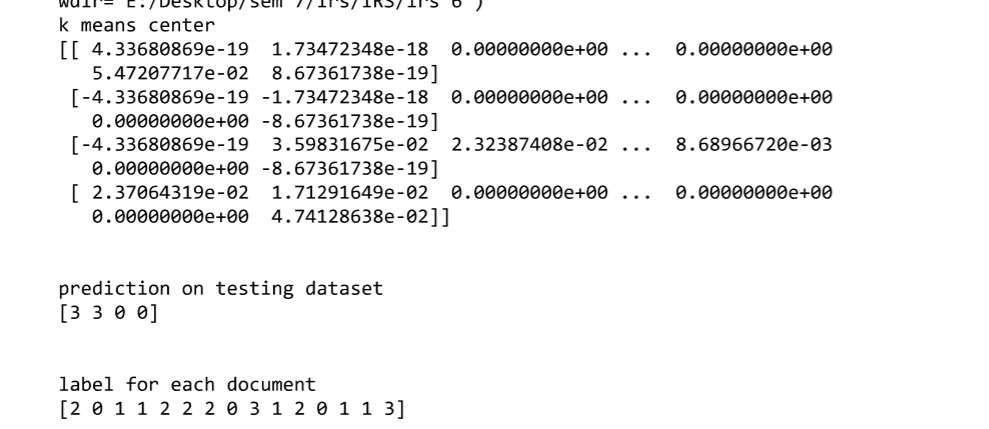
* We used inbuild python data structures such as Dictionary and arrays and to take the input we used the I/O operations from the files
* We used SKLearn to train the k means model
* We used NTLK library to perform the pre-processing steps

**Sample Input/Output**

The input is in the form of corpus consisting of 20 text files located in 4 different directory.

The output is presented in the form of k means cluster centers and the label for each document is also printed

The output can seen in the output.pdf file



**Conclusion**

In this practical, did the text preprocessing and took into consider 20 different documents consisting of 4 domains and performed document clustering on it.