**Information Retrieval (CS317)**
Programming Assignment No. 3
Spring 2020

**Submission Date: May 15, 2020**

## Assignment Objective

This assignment focuses on the tasks of classification and clustering.

## Classification

First task is related to document classification. We have discussed three methods for the task of text/document classification. These are Rocchio's, Naïve Bayesian and KNN. In this assignment you need to work on KNN using VSM model to hold documents and Euclidian distance measure to estimate closeness of the instances in neighborhood. There is no training required for the KNN algorithm, all you need to divide your data into suitable splits of train and test. Using the training data, you need to check the labels of test instances using k=3. From the 5 classes in the dataset you need to create you train and test sets from 737 instances. Using your idea of a good split set you are free to create the train and test set. The evaluation of classification will be performed on Accuracy of classification for the test split. For a detail description of KNN read the chapter 14 of the textbook.

## Clustering

For the task of clustering you need to implement a K-means clustering algorithm, assuming all documents are represented by VSM of suitable feature space. There are 4613 features in the dataset. The feature selection is what you need to perform as per your understanding. The evaluation of clustering will be performed by measure purity of the dataset. For a detail description of K-means read the chapter 16 of the textbook.

## Datasets

For this assignment the dataset we have selected is from BBC Sport website corresponding to sports news articles in five topical areas from 2004-2005. There are 737 Documents, and 4613 Terms. There are 5 classes from sports these are athletics, cricket, football, rugby, and tennis.

## Evaluation/ Grading Criteria

The grading will be done as per the scheme of implementations, evaluation measures, clarity of code, split set criteria for classification, feature selection for clustering, and efficiency of the code. Extra marks for intuitive GUI for the assignment.

**<The End>**