CSE508 : Information Retrieval
Assignment 5

Instructions
- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf

Question 1

Download 20_newsgroup dataset from
https://drive.google.com/file/d/1VA4a-wveTVXEy0J_NNv8oZ_YG2smxvPL/view
You need to pick documents of comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space [5 classes] for text classification. You need to use the below as feature vectors
1) Bag of Words Model
2) Word2Vec representation from Google News Pretrained Word2Vec model [you can refer to: http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/]

For both of these features set, implement K-means clustering algorithm [you cannot use any library for k-means] [don't use groundtruth information] and report the error. Draw your inferences

Question 2

You need to pick documents of comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space [5 classes] for text classification from the above dataset.
You need to implement KNN classification (vary k=1,3,5) Perform the above steps on 50:50, 80:20 and 90:10 training and testing split and analyze the accuracy scores. Plot ROC curve and show confusion matrix.
Compare and Analyse this method with previously implemented Naive Bayes Algorithm