CSE508 : Information Retrieval
Assignment 05
Deadline: April 8, 2020

Total: 100 marks

Instructions

- The assignment is to be attempted individually.
- Language allowed: Python
- For Plagiarism, institute policy will be followed.
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
- Mention your outputs and analysis done in Analysis.pdf.
- Deliverables: Code files, ReadMe, and analysis.pdf
- Submit code, readme and analysis files in ZIP format with the following name: A5_<roll_no>.zip
- Save all your precomputed indexes and tables which may take time to compute.
- Note: Due to the COVID-19 outbreak and lockdown, it may so happen that the assignment demos cannot be taken, hence you are advised to prepare a well documented .py and .ipynb file and report/analysis with all the justifications that may be necessary.

Download the 20_newsgroup dataset. You need to pick documents of **comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space** [5 classes] for text classification.

Implement the following algorithms for text classification:

1. Naive Bayes
2. kNN (vary k=1,3,5)

Feature selection techniques to be used with both algorithms:

- Tf-IDF

- Mutual Information

Implementation Points:

- Perform the data pre-processing steps.
- Split your dataset randomly into train: test ratio. You need to select the documents randomly for splitting. You are **not** supposed to split documents in sequential order, for instance, choosing the first 800 documents in the train set and last 200 in the test set for the train: test ratio of 80:20.
- Implement the TF-IDF scoring technique and mutual information technique for efficient feature selection.
- For each class - train your Naive Bayes Classifier and kNN on the training data.
- Test your classifiers on testing data and **report the confusion matrix and overall accuracy**.
- Perform the above steps on 50:50, 70:30, and 80:20 training and testing split ratios.
- Compare and analyze the performance of the above-mentioned two classification algorithms for both the feature selection techniques across different train: test ratios. **Use graphs to report the performance comparison**. **Also, mention your inferences from the graphs**. Example of a graph you can report - a graph showing the performance of kNN for different values of k.