**Yelp Dataset Challenge**

Chiranjib Ghorai

655493675

**ReadMe:**

1. Com.data.setup package has data setup classes.
2. CreateTFIDFTestFile.java is used to create the sparse vector from the csv file that will be used in the model.
3. PorterStemmer.java class is used to stem the review text.
4. Com.mahout.bayes package has the Naïve Bayes implementation in mahout.
5. NaiveBayes.java class implements Naïve Bayes where TRAIN\_DATA\_PATH and TEST\_DATA\_PATH values are to be populated with train and test data respectively in csv format.
6. Com.spark.regression package contains decision tree and random forest implementations.
7. DecisionTreeImpl.java class implements decision tree using following configuration:

**Impurity** = variance

**Maximum depth** = 30

**Maximum bins** = 10000

1. RandomForestImpl.java class implements the random forest algorithm using following configuration:

**Number of classes** = 6

**Number of trees** = 30

**Subset strategy** = Let the algorithm choose

**Impurity** = gini

**Maximum depth** = 30

**Maximum bins** = 10000

**Seeds** = 99999

1. Bigram and unigram folder contains all the data in CSV and in TF-IDF format.