Homework 4

Sivakarthik Gade

INSTRUCTIONS TO COMPILE AND RUN THE CODE

1. Implemented the assignment in Java. Need Oracle/Sun JDK setup on the target machine to compile and run the code. (Used JDK 1.7.0_45 during development)
2. Extract all java files into some location.
3. Compile instruction through command prompt: 'javac *.java'
4. Run instruction through command prompt:
    a. java TextClassification <full path to data folder> {true/false}
    b. Ex: java TextClassification E:\txt_sentoken true
5. The 2nd argument {true/false} is optional. It can be used to set whether to ignore/not ignore stop words. If chosen to ignore stop words then the Stop_Words.txt file included in the zip file should be placed in the same folder as TextClassification.class file.

README

1. Have implemented the standard Naïve Bayes classifier algorithm. Have two classes pos & neg.
2. Performed 10 fold cross validation as instructed.
3. Generated 10 models for each of the fold training data, generated respective learning model (gets written to lm.txt).
4. Used the learning model generated in each fold to test the testing samples set aside during that fold.
5. The program execution prints the accuracy generated on each fold, and the final accuracy at the end.

REPORT

1. The final accuracy ignoring stop words is 76%.
2. The final accuracy without ignoring stop words is 74.5%.