-------------EE219 PROJECT 1--------------

Team member:

Yucong Wang 305036163

Shizhong Hao 605035020

-------------CLASSIFICATION ANALYSIS ON TEXTUAL DATA-------------

The following README contains the requirements and steps that are required to

execute codes in this Project.

-------------Folder Structure -------------

- PROJ1\_CODE\_FINAL.ipynb

- Readme.docx

- Proj1\_Report.pdf

------------- Implementation Dependencies -------------

Please make sure these packages are correctly pre-installed before running

the codes

a. nltk

b. numpy

c. matplotlib

d. sklearn

Environment:

This project is coded using Jupyter notebook.

-------------Instruction -------------

Before running the code for each question, you should first run the block at the very beginning to import all the packages we will use.

Each question can be run separately, but you’d better not to run the code with random order(like first run the first block of question i2 then the second block of question h) for it might change the variables we will use, and cause the wrong output.

Graphs and output of each question will be displayed in output automatically.

-------------Basic introduction for each part -------------

QUESTION(a). It plots the histograms of the documents per topic and counts the number of documents in the two groups.

QUESTION(b). It turns the documents into numerical feature vectors and creates the TFxIDF representation.

QUESTION(c). It finds the 10 most significant terms in the required classes.

QUESTION(d). It applies LSI/NMFtechnique and maps the documents to a 50-dimensional vector.

QUESTION(e). It uses hard margin SVM and soft margin SVM to classify the dataset.

QUESTION(f). It uses hard/soft margin SVM with different gamma value to classify the dataset and finds the optimal gamma.

QUESTION(g). It uses multinomial Naive Bayes Classifier to classify the dataset.

QUESTION(h). It uses Logistic Regression Classifier to classify the dataset.

QUESTION(i1). It adds a regularization term to the Logistic Regression Classifier and uses l1 and l2 norm regularization to do the same task as in QUESTION(h).

QUESTION(i2). It solves MULTICLASS CLASSIFICATION. It performs Naïve Bayes classification and multiclass SVM classification with both OneVsOne and OneVsRest strategy.