Csci 581/543: Data Mining
Project 2
Deadline: April 15, 2018. 11:59 PM

Date: April 03, 2018
Points: 100

Requirements:

- The project must be handed in electronically through Blackboard, in **zip format only**.
- Your submission **must not contain** any information which may reveal your identity.
- Anonymous grading is enabled. If there is any identity-revealing information, **50 marks will be deducted**.

Overview

The objective of this project is to be familiar with mining valuable information from microblog data. We will use Twitter which is the most popular microblog platform in current times.

Task Description

You will be given 1.6 Million tweets from Twitter. Each tweet has the following information-

date, user, text, polarity (1 means "negative", 3 means "neutral" and 5 means "positive")

Your task is to build a supervised learning model that can predict the polarity of a tweet. Note that, you must implement the prediction algorithm by yourself. Third-party implemented prediction algorithms (ex: scikit-learn, weka, etc) are not allowed. However, you can use nltk, numpy and similar packages only for ease of data manipulation purpose.

You have to use your implemented model to predict the polarity of the tweets contained in the evaluation_csci581.csv file.

Your grade on this project will be partially determined by the accuracy of your model. Therefore, you should estimate the performance of your prediction method (for example, using holdout, cross-validation or resampling, etc.) and find ways to improve your method.

Materials Provided

1. twitter_csci581.csv file [It is a large file containing 1.6 Million records. You may want to use software/packages which can handle large text files.]
2. evaluation_csci581.csv file [It is a small file containing 498 records.]

Deliverables [in a single zipped file]

1. All of your source codes.
2. Project Description: Write a document (**pdf file**) about this project, which should include the following information:
    - Describe the method you used to train the model in detail.

- If you have estimated the performance of your method, describe how you did this and what is the estimated performance in terms of accuracy and, optionally, any other measures you applied.
- If you made any improvement according to the estimated performance, highlight what modification you made to your original method and explain why you made the modification. Point out the improvements in terms of accuracy and, optionally, other measures you applied.
- Briefly introduce the most important features of your model.
- The document should be as detailed as possible, preferably with visualizations, but no more than 2 pages.

3. An output file named **output.csv** which contains your predicted polarity for the 498 tweets in the evaluation_csci581.csv file. Note that, **output.csv** should only contain 498 comma-separated numbers and **nothing else**. Don't change the row order of the evaluation_csci581.csv file. Otherwise, it won't be possible for us to determine which polarity in output.csv belong to what record in evaluation_csci581.csv file.

Grading

The full mark of this task is 100.

Basic Points (40 points)

All required materials are submitted as in required format. The code can compile, execute, and make predictions. The prediction should come from a reasonable prediction method (certainly, it cannot be a random guess). Incomplete submission will result in points deduction.

Document Quality Points (30 points)

This part evaluates the quality of your comments in source code and the pdf file in submission. A well written document should be complete, readable and well formatted.

Performance Points (30 points)

This part evaluates the performance of your prediction method. The evaluation is based on the accuracy of the prediction stored in file output.csv of your submission. We will rank all submissions by their accuracy. A submission at rank k gains 30*(n+1-k)/n points, where n is the total number of students. That is, the best submission, which is rank 1, will get 30 points, and the rank 2 submission gets 30*(n-1)/n, and so on.

Bonus Point (20 points)

Put on your data science hat and be imaginative. You are encouraged to use extra information from outside of the provided materials to improve your prediction model. For example, you can try extracting profile information of the users. You are also encouraged to perform exploratory analysis on the data and find interesting patterns.

Explain in the pdf document if you have put any bonus effort and how that improved your prediction model.