STAT 154
FINAL PROJECT
TEXT MINING

This is the final project for the class. Form teams of 3-4(max) and let the GSI and me know who your team members are by Monday 11/17. The team will receive the same score for the project. Total points: 100 (30% of the course grade). Keep all descriptions and summaries complete but succinct and precise. Provide tables and charts whenever appropriate. Section competition on Wednesday 12/3. **Final write up due by 2pm Monday 12/15 (or earlier). No late project write-ups will be accepted.**

Data: The data will available to you on bCourses. Note these are the actual text messages (not a data matrix).

1. *Description:* Write one paragraph describing the data. (Hint: How many texts are there? Distribution between spam and ham (non-spam). Look at a couple of the texts. What country or countries do they seem to represent?)

2. *Feature Creation:* Define a feature as a unique word in the text. That is a continuous string of alphanumeric characters without white space.

    a. Use your favorite scripting language to parse out the word features from the texts. Set up a dictionary where you list all the unique word features appearing in the texts

    b. Exclude the common word features (known as stop words) listed in http://www.textfixer.com/resources/common-english-words.txt

    c. Derive a word feature matrix where rows= # of texts and cols=# of word features. The number of word features should be around 15,000. In each cell should be the frequency $\in [0,1]$ of the feature in the text. The frequency is the number of times that feature appears in that text divided by the number of word features in the text. Stopwords should be excluded here. (Hint: what should the sum of each row be?) Take out the data of any text messages that are composed of all stopwords (Hint: the sum of the rows of these observations should be 0.)

    d. Improve your word feature matrix. (Hint: make upper case and lower case word features equivalent, for example). Attach the spam/ham label as the final column in your matrix. Report the dimensions of your matrix.

Describe the process you undertook to derive the word feature matrix. How many features did you end up with? Did you encounter programming challenges?

3. *Unsupervised Feature Filtering:* Make a histogram of the # of times word features appear in a text. (Hint: For each column in your data matrix – count the number of times that feature is non-zero). You can filter words that are too rare and too common this way. What minimum and maximum threshold will you use? Apply the filters. How many features do you come up with? Give dimensions. Store this as your word feature matrix.

4. *Power Feature Extraction:* Derive Power features to be used for classification in addition to the Word features. You should create anywhere between 1 and 25 power features. Power features are aspects of data that help discriminate between spam vs. ham – and not included in the word matrix. (Hint: Presence/Absence of Toll Free Area Codes, Length of longest string containing numbers, Presence / Absence of money symbols, etc.). Describe the features and code them. Create a matrix of power features with rows= # of texts and columns = # of power features. Store this as your power feature matrix.

5. *Word and Power Feature Combination:* Create a combined feature matrix. Describe the dimensions of the data. Store this as your combined feature matrix.

6. *Classification on the filtered Word Feature Matrix: Choose* two classification algorithms (SVM and Random Forest) to classify the data. Use V-fold cross validation where V=10. Produce ROC curves for your classification (you should have at least one curve for word (step 6), one curve for power (step 8), and one curve for combined (step 9). Also compute the accuracy of the overall texts, spams and hams separately. Repeat for PPV and NPV. Show in the output the dimension of your feature matrix. Note: Do not use the validation (test set).

7. *VERIFICATION: Submit the word feature matrix to your designated study directory by Tuesday 12/2. This will serve as verification that you have a final output at the end of step 3. In section on Wednesday 12/3, you will be given a validation set of text messages. You will need to produce a word feature matrix from this validation test. Next you will build a classifier on the training set (word matrix from step 3) and predict on the test set (word matrix you just produced. Produce the predicted class labels and submit to GSI before leaving. Total time for processing raw data, training the classifier of your choice, and predicting class label for the test set should take no more than 10 minutes.*

8. Repeat Step 6 on Power Feature Matrix. Write and describe your output. Compare these results (accuracy, PPV, and NPV) with the results from part 7 in which you only used the Word Feature Matrix.

9. Repeat Step 8 on the Combined Feature Matrix. Write and describe your output. Compare these results to the results from steps 7 and 8. Did your results improve? Provide graphs and charts.

10. *Validation set:* Lock the optimal model you chose from steps above and the feature matrix. Develop the model on the feature matrix of your choice. Predict the results on the validation set of text messages given to you in Step 7. Produce a ROC curve, a 2 x 2 table of the classification results; write down the % accuracy and PPV and NPV. Comment on which of these three values would be the best to judge your model (or if you would look at a combination of them). (Hint: what do PPV and NPV mean in terms of classification of emails as spam or ham?) Comment on your final results, as well as your learning process.