STAT 154
FINAL PROJECT
JOB POST MINING

This is the final project for the class. Form teams of 4 and let GSI Austin know who your team members are by 4/11(Mon). The team members will receive the same score for the project as the team provided each member worked on programming for the project. Total points will be 40 (40% of the course grade). Keep all descriptions and summaries complete but succinct and precise. Provide tables and graphs whenever appropriate. Section competition on Monday 5/2. Final write up due by Monday 5/9 (2-5 pm in my office @339 Evans). No late project code or write-up submissions will be accepted.
Data: The training data will be available to you on BOX by 4/20. Note these are actual job posts data (not a data matrix) of two categories or tags: Fraudulent, Not fraudulent. Each entry is a job post from the global job market. You should work strictly with the data set provided to you.

1. [2] Description: Write one paragraph describing the data. (Hint: How many observations are there? Distribution between categories. Look at a couple of the observations in each categories. Is it easy for human to classify these posts as fraudulent or not?)
2. [4] Feature Creation: Some features are given in the training data (binary, categorical and others are complex text). For the complex text features, perform text mining: create several word features from each complex text feature. These word features will be measured as term frequency (TF).
    a. Use R/Python to parse out the word features from the complex text features.
    b. You may need to parse out irrelevant words or common words as needed.
    c. Derive a feature matrix where rows= # of observations and cols=# of features. At the same time create the target vector of having tags "Fraudulent", "Not Fraudulent". Report the dimensions of your feature 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: Are there any unsupervised techniques that can be applied to filter the features? Explain what method you used, if any. Report the dimensions of your feature matrix at this step. How many features do you come up with? Give dimensions. Store this as your word feature matrix.
4. Power Feature Creation: Are there any special features you want to create? We learned a few methods like PCA or you may manually create features after reviewing some job posts that have discriminative power. Describe any power features you may create and include for further analysis. How many power features did you decided to keep?
5. [2] Feature and Power Feature Combination: Create a combined feature matrix. Describe the dimensions of the data. Store this as your combined feature matrix and write how many total features you are using for classification below.

6. [4] Classification on the Feature Matrix: Use the two classification algorithms (SVM and Random Forest) to classify the training data (use final features from step #5). Use V-fold cross validation with V=10 to produce the predictions and to tune your model parameters. Compute the accuracies in each loop. Report the overall accuracy rates and accuracy rates per class. (Note: it might be useful to produce ROC curves for your classification (you should have at least one curve for word, one curve for power, and one curve for combined) for pairwise comparison between classes). Report the dimension of your feature matrix.
7. [10] VERIFICATION: Submit the feature matrix obtained from the training set to your BOX directory by 5/1 (Sun) midnight. This will serve as verification that you have a final output at the end of step 5. Save your final Random Forest model in BOX as well – ready for prediction on an unseen test set.

   In-class competition: In a special session on **Monday 05/02**, you will be given a competition test data set of job posts with no tags. (i) You will need a pre-processing script to produce a feature matrix from this test set. Your feature matrix dimension should be same as step #5. (ii) Next you will need a classification script (that uses a Random Forest classifier) saved on the training set feature matrix and classifies the new observations.

   We will check your saved model in (ii) to make sure it was saved on the training set handed out in the beginning.
   Output the predicted class labels from (ii) on the test set and submit to GSI and upload to BOX before leaving. Total time for (i) processing raw test data into a feature matrix and (ii) using saved model to predict class label for the competition test set should take no more than 5 minutes per team. The accuracy on this test set will be graded. Note: You will need to upload (i) your pre-processing program to deal with raw text files and (ii) saved model -- in BOX ahead of the competition (We'll give you a small practice data set prior to the competition).

8. [5] Perform unsupervised clustering of your final feature matrix. See if the clusters identified track with the classification labels of your training data.
9. [3] Apply methods to improve your features and/or classification results). Provide graphs and charts to describe this process and report the performance of your final classifier and # of final set of features on the training data set.
10. [10] Validation set: Submit best model (SVM, or Random Forest, or any model that you have learned in this class) object built with best features on the training data set to BOX before **Sunday May 8th 6pm**. Along with (i) your saved model, submit your (ii) pre-processing script for converting raw data into the best feature set and (iii) the actual classification program that produced the saved model on training data. Around 7pm – Austin will upload a test set in your team folder. By 8pm, complete (i) pre-processing and (ii) classifying this test set. Save the accuracy metrics overall for each class in a text file in the folder. By 8pm you will no longer have access to your folder. In the write up, comment on your final results, as well as your learning process. Report your final or highest accuracy rate in your write up. Submit the write-up in person on Monday 5/9 2-5 pm at 339 Evans Hall.