**Text Analytics**

**Assignment 3 Classification**

This assignment will give you hands-on experience in building text classification models, using the application of email spam filtering. The target variable represents whether an email is either spam (1) or non-spam (0). Follow the directions and answer following questions.

**Question 1**

Explore different ways to improve the classification performance (accuracy or expected cost). You can consider the following:

1. Feature representation: Compare 3 feature representations; binary vs. frequency vs. tf-idf
2. Classifier: compare 3 classifiers of your choice such as decision trees, neural nets, etc.
3. OPTIONAL: Feature selection: different feature/attribute selection methods or parameters (extra credit)

***Report the evaluation results of your model using split training and testing***. Report the following:

1. Precision and Recall by Class
2. Confusion Matrix.

**Question 2**

Calculate the total cost and expected cost (per email) based on the confusion matrix you obtained in question. Assume the cost for each mis-classified email from Spam to Non-spam is 5, and from Non-spam to Spam is 100.

[Hint: be careful with the dimensions of the confusion matrix: which are the “actuals” and which are the “predictions”?]

Based on your observation, please analyze which combination of feature and classifier is the best.

**Question 3 (Extra credit)**

Run 10-fold cross-validation instead of split sample. Does your conclusion still hold? If the observation is different, could you analyze the cause?