**Text Analytics Group Assignment #2**

**New Due Date: 9/19 by 11:59 p.m. on Canvas – Note that the due date for this assignment is postponed to 9/19 because we will cover the PMI approach to sentiment analysis on 9/12**

**The data for this assignment (Yelp Restaurant Review Data) is posted on Canvas.**

This Yelp dataset has information on restaurants (e.g., type of food, price range, etc.) as well as reviews written by patrons. The output variable is the star rating (1-5). It will be best to convert this rating to high (say, ratings of 4 & 5) and low (1, 2, 3).

**Task A.** Ignore the text (reviews) and run a classification model with the numeric data (you can use standard methods like logistic regression, k-nearest neighbors or anything else). What is the best accuracy of your model with numeric data?

**Task B.** Perform a supervised classification on a subset of the corpus using the reviews only. You can write your code in Python or R. What accuracy do you get from this text mining exercise?

**Task C.** Combine the numeric data and the text classification model (in task B) to create a “hybrid” model. It is your task to figure out how to do this. Now run this hybrid classification model and compare the results with those in A and B. Does the numeric data add to the predictive power relative to text?

**Task D.** Use unsupervised sentiment analysis on the reviews (with SentiStrength or any other tool) and use the sentiment scores to predict high/low rating. Compare and contrast the results of tasks B and D. What can you conclude from your analysis?

**Task E.** Implement the PMI approach to sentiment analysis (in either Python or R), and run the classification model with the sentiment scores. How do your results compare with those in Task D?

**Task F.** What are the top 5 “attributes” of a restaurant that are associated with (i) high and (ii) low ratings? That is, when people rate a restaurant high or low, are they more likely to mention service, ambiance, etc.?