Math 5366 Homework 24

1. The UCI Machine Learning Repository hosts the Sentiment Labelled Sentences Data Set, containing 3000 sentences, each labeled as 1 (positive) or 0 (negative).

   `http://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences`

   These sentences were obtained from `imdb.com`, `amazon.com`, and `yelp.com` (1000 sentences per site). Split the data into 80% training data and 20% test data, and build models for predicting the sentence labels using each of the following methods. Then calculate the classification accuracy and AUC for each model.

   (a) AFINN score

   (b) Bag of Words (random forest using term frequencies)

   (c) Random forest using tfidf

   (d) Normalized sentiment difference index

   Comments:

   - One of the most difficult parts of this problem will be importing the data. We will discuss this in class. ☺
   - Since there are three data sets, it's possible to build a separate model for each one, but another option would be to build a model for the entire problem, using a factor variable `site`, whose levels are `imdb`, `amazon`, and `yelp`. It would be interesting to compare these two approaches to see which one performs better.

2. Download 2000 tweets for Verizon and 2000 tweets for AT&T, and calculate the AFINN ratings for all of the tweets. Is there a statistically significant difference between the AFINN ratings for these two companies?