# Precision Recall and ROC assignment

In this assignment, you will be given python code that has been previously executed in an iPython notebook and you can use this code to answer questions about the output of the code. Please try and run this code on your own! Add your own features. Please bring your answers to the next class for discussion.

The code reads tweets from a twitter sentiment analysis data set available here (http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/).

For more on sentiment analysis see here https://en.wikipedia.org/wiki/Sentiment_analysis.

The tweets have been previously loaded as a thorn file available here (http://www.cs.ou.edu/~cgrant/teaching/cs5970sp16/sad/sad.thorn).

The thorn file contains tweet text and labeled sentiment classifications (0 – negative, 1 – positive). The python file reads the tweet text and trains three different classifiers based on the training labels. It then splits the files into a test set and a training set. Each algorithm is trained on the training set and evaluated on the test set.

Use the source code and the logged output to answer the following questions. You are free to take the code and rerun it to get a better understanding of what the code does. Only 10000 tweets are used but you are free to change the parameters to help you evaluate the code. All the questions will refer to the listed run and specified parameters.

**The link to the iPython code is here:**
http://www.cs.ou.edu/~cgrant/teaching/cs5970sp16/sad/sad.html

Note: To understand some of the functions you will have to search for details on the web. The code used is from the NLTK and also sklearn http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics as well as a graphing library called ggplot. We will be using these modules going forward so you can use this as a chance to get familiar with these tools. Also, "sad" stands for sentiment analysis data set — pun intended.

1. How many positive tweets were extracted from the thorn file?
2. How many negative tweets were extracted from the thorn file?
3. What are the names of the three classification algorithms used?
4. What is the name of the function used to compute the precision, recall, and f1-score?
5. Which algorithm has the best average recall?
6. Which algorithm has the best average precision?
7. Which algorithm has the best average f1 score?
8. Which algorithm has the highest area under the curve?

9. Which algorithm took the longest to train?
10. What is the name of the package was used to plot the ROC curve?