Group 9
Sibelius Seraphini, Harold Austin
Xiaohu Nian, Samresh Jyotishi

Course Project Report - February 24, 2014

For this week, we conducted the first experiment to develop a tweet classifier that determines whether a tweet is related to crime or not. Our experiment is described here:

1. Extract the text from the tweets and transform all text to lowercase
2. Generate a Document Term Sparse Matrix, i.e, a matrix of term counts
3. Transform this matrix using the Tf-Idf weight
4. Train the Linear SVM Classifier using labels that was manually set.
5. Calculate the Precision, Recall, and F1 score using a 5-fold cross validation approach

To train our classifier, we used a collection of 3237 tweets of the user @NYCityAlerts (https://twitter.com/NYCityAlerts). These tweets were manually labeled where 978 were labeled as crime related tweets and 2259 as not related to crime. We used the scikit-learn library (http://scikit-learn.org/stable/) for the feature extraction, classifier and to calculate the metrics. These are the results using a 5-fold cross validation:

Precision: 0.97 (+/- 0.02)
Recall: 0.95 (+/- 0.05)
F1 Score: 0.96 (+/- 0.03)

Even though, our classifier achieved such precision, and recall, we are still collecting and labelling more tweets from different users to check the performance of the classifier in a general context. Moreover, we will use Grid Search to perform hyperparameter optimization to find the best parameter for our features (unigram, bigram, trigram, etc) and for our classifier.

Additionally, we are also starting using Mingle (http://www.thoughtworks.com/products/mingle-agile-project-management) to manage our course project.