Joe, Vino, Louis

Intro to Intelligent Security Systems

Mooman

1. Introduction

Our goals for this project was to gather data regarding the tweets during the arab spring, provide an analysis using machine learning techniques, and ultimately make predictions based off of our findings. The idea behind this project is to correlate social media and security concepts. Just as an intrusion detection system was meant to find an anomaly in a large dataset, we hoped to do the same with “human” data and attain the ability to predict human security issues such as civil unrest and revolts.

1. Background

The arab spring was a series of protests, riots, and rebellions in the Arab world that resulted in many political changes throughout the region. This was a unique event in that much of it was organized via social media. Sites such as Facebook, Twitter, and Youtube played a huge role in helping the protestors achieve their goals. In social science this is a very interesting scenario as this is the first time we have massive amounts of data to analyze a revolution with. With this data and proper machine learning techniques, it should be possible to extract sentiments and make predictions about the future.

1. Each of our Findings

We used R as a means to get a time frequency analysis and the twitteR package to get a realtime view on our findings. From the historical data we gathered using the tweet ids found in the public dataset, we were able to resolve them to full json responses using the Twitter API’s. From there we began doing some analysis such as time frequency and sentiment analysis. While our frequency analysis was much more successful than the sentiment analysis, it was cool to learn about all of the things the R programming language could help accomplish.

We also used the python package PySci to apply machine learning algorithms to our data. We were originally going to use weka but found that since we had already begun writing custom code for our project, that this might work out well. As it turns out it did as it had a means to perform a basic sentiment analysis of our data which was one of our ultimate goals.

We used Mongo & Python to develop a pair of programs, the first simply scrapes data from Twitter into the DB. While we did not have time to complete the second, the idea is for it to clean the data. It can either be used to clean the data before feeding it into Weka, or, depending on requirements, to process the data itself in order to update in real time. The primary trade-off is that Weka has built-in algorithms researched by P.H.D students, but even if their knowledge and choices involve better algorithms, writing the code ourselves allows for a larger range of functionality.

1. Conclusion

In the end we were unable to finish everything we had hoped to but we were able to make huge strides towards the overall goal of making predictions about social security. We got data on the frequency of tweets during an anomaly event, and were able to gather good historical data, clean it, and store it in a logical and easy to access way. This project was a great way to learn about various machine learning techniques and different toolds such as R, Weka, Python, and Mongo to try and achieve our goal.