**“Statement Classification Tool Based on Frequent Pattern Mining of Twitter Stream”**

Mark Dazey

May 2013

**Abstract**

For this project, I built a web application using Ruby on Rails and Java. I used this project as an opportunity to learn Ruby on Rails, as well as increase my skill set in Java, HTML, and SQL. The project was designed to classify a statement. The classification is based on frequent patterns from the Twitter stream along with the keywords used to classify the statements in the stream.

**Summary**

For this project, I built a web application using Ruby on Rails and Java. I used this project as an opportunity to learn Ruby on Rails, as well as increase my skill set in Java, HTML, and SQL. The project was designed to classify a statement. The classification is based on frequent patterns from the Twitter stream along with the keywords used to classify the statements in the stream.

**Introduction**

There is an enormous ammount of data now passing through Twitter. In 2011, Twitter reported it took them over three years to reach one billion tweets. In that same year the time it took to reach one billion tweets had decreased to one week, they were now averaging 140 million tweets a day2. The next year, Twitter reported that they were now receiving approximately 340 million tweets a day and one billion every three days1. In March of this year Twitter reported over 400 million tweets are being created a day3. At 400 million tweets being created a day there are almost three billion tweets created every week.

The Library of Congress along with Twitter announced in April of 2010 that the entire twitter archive had been donated to the Library of Congress. The Library of Congress announced in January of 2013 they had received 170 billion twitter posts and that the number was growing daily4. The Library of Congress has said they have had approximately 400 request from researchers wanting access to the data4.

There are things to be learned from the Twitter ecosystem. There are many organizations and people who have realized this and are taking steps to start unlocking this data. Twitter provides developers with access to the public stream of tweets. Twitter has two levels of access to the stream of tweets. The first is the sample stream which returns a small random sample of the stream of public tweets. The second is the firehose stream which returns all public tweets5.

**Overview**

For this project I wanted to use new technologies and learn additional skills and techniques. In order to accomplish the task of learning something new I looked for a language I had not previously used for web-development. I decided by using Ruby on Rails I would have the most opportunity to learn new topics. Through Ruby on Rails I was also able to learn the Model View Controller architecture.

The backend part of this project was done using Java, in my current position I am working as a Java developer and took this opportunity to further my skills in the Java programming language. I was able to learn additional skills in the language including multi threaded programming with the use of thread pools. I also was able to develop my skills in processing a stream from a web site.

Due to Twitter's access restrictions on the Firehose stream I was only able to access the sample stream. For the pruposes of this project this stream was sufficent, however better results could have been obtained with access to a larger dataset from a more shorter period of time.

1. <http://blog.twitter.com/2012/03/twitter-turns-six.html>
2. <http://blog.twitter.com/2011/03/numbers.html>
3. <http://blog.twitter.com/2013/03/celebrating-twitter7.html>
4. <http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/>
5. <https://dev.twitter.com/docs/api/1.1>
6. <https://dev.twitter.com/docs/api/1.1/get/statuses/sample>