**Using Hadoop and Spark to Analyze Tweets**

**CS 5540 Group 19 Members:**

**Alex Acsenvil**

**Duy Hoang Ho**

**Tonderai Kambarami**

**Introduction**

For phase 1 of this project we were required to collect tweets using Twitters streaming API, extract all the tweets hashtags and URLs, then use Apache Hadoop and Spark to compute the word count. This was completed in a 3 step process. We first had to sign up for a twitter developer account and create an application through their API, after which we received our API keys. Once we received our keys we created a python script and used that to collect the tweets in JSON format. We used the open sourced Tweepy library to access the twitter API. And finally ran the wordcount program on the collected tweets in both Hadoop and Spark and collected the log file and results.

**Collecting the Tweets**

Like mentioned before, we used the open source library Tweepy to connect to the Twitter streaming API to collect our tweets. The library allowed us to open a stream and collect the tweets in a live environment, filtering based on a key word (in our case we chose ‘America’ as our key word). The script we used was a modified version of the example found within the Tweepy repo. We modified the script to allow us to store the tweets in JSON format, and to parse out and store the Hashtags and URLs in separate files while scanning the live tweets. This made feeding the data into Hadoop and Spark easier, and as a result easier to run the Wordcount program. You can find the modified script in the GitHub repo.

**Hadoop and Spark**

We moved the URLs and Hashtag files to the HDFS and ran the wordcount through Hadoop and Spark from there. The Hadoop files produced an output part-00000 on the HDFS, while as in Spark we added “&>output.txt” to make the output file in our current directory. Attached to the first increment of the project you will find 3 Hadoop files (output and log files) and 2 Spark files (log files).

**References**

[www.tweepy.org](http://www.tweepy.org)

<https://github.com/tktoday/CSEE5540_Project1>