**Project Phase 1**

**Objective**

* Collect Tweets using Twitter’s Streaming APIs (e.g., 100K Tweets)
* Extract all the hashtags and URLs in the tweets
* Run the WordCount example in Apache Hadoop and Apache Spark on the extracted hashtags/URLs and collect the output and log files from Hadoop. Add a README file.

**Team**

* **Ruthvic Punyamurtula - rpmgz**
* **Navya Ramya Sirisha - np7yd**
* **Bhanu Sudheer – bvhkf**

**Twitter tweets**

The “Twitter” folder has the contents for

* Around 100k tweets are collected using “tweet.py”
* Source code for hash tag extraction using “hashtag.py”
* Extracted hashtags and url from the collected 100k tweets are stored in “extractedTweets.txt”

Now using this “extractedTweets.txt” as input for WordCount program in Hadoop and Spark

**Hadoop**

* logs.tar.gz – have the total logs collected for the word count (compressed)
* part-r-00000 – this is the output file of the word count run in Hadoop
* WordCount.java is the java file used for compiling and generating JAR to execute word count program

**Spark**

* part-00000 – this is the output file of word count run in spark
* siri\_spark.JPG – this is the image showing the wordcount run in Spark environment

**Images**

* hadoop\_cmd\_1.JPG – this is the image for running word count in Hadoop
* Hadoop\_cmd\_2.JPG – this is the image showing successful map reduce job for

Word count in Hadoop

* Hadoop\_output.JPG – this is the screenshot of the output file