This project is about extract tweets from twitter api and get the count of urls and hashtags from them.

Execution:

1. Got twitter developer account access for collecting tweets.

2. Wrote python code to extract tweets from api and did the same to extract urls and hashtags from the collected tweets.

Got two files and named them as follows :

1. hashtagsintweets.txt
2. ii. urlsfromtweets.txt

3. Created a directory in hadoop with the name input using the below command :

hadoop fs -mkdir input

4. Moved both the files into input directory with the below commands:

hadoop fs -put 'home/cloudera/Downloads/hashtagsintweets.txt' input/

hadoop fs -put 'home/cloudera/Downloads/urlsfromtweets.txt' input/

5. Written scala programs for counting hashtags and urls and executed using below commands:

spark-shell -i wordCount.scala

spark-shell -i hashtagsCount.scala