

# Sentiment Analysis

The platform used to execute the assignment is UBUNTU 18.04 and the language used is python.

Project directory : Sentiment\_Analysis

The Sentiment\_Analysis contains following files :

- **ip.txt** : input file for map reduce.
- **AFINN.txt** : the text file contains tab separated words along with an integer sentiment value based on its meaning ranging from -5 to +5.
- **mapper2.py** : mapper code for the map reduce.
- **reducer2.py** : reducer code for the map reduce.
- **twitterstream.py** : This is the twitter API used to get the live streaming data set from twitter.
- **hadoop-streaming-2.7.3.jar** : this file is used for streaming utility which is needed in running the map reduce program in python.
- **Input and Output snaps**

**Getting the live stream data from twitter using twitter API to create the input file for the program :**

- Install the oauth2 library on the system which is used by the twitterstream.py using command.

***\$ pip install oauth2***

- Create a twitter account.
- Go to the following link and login with twitter account credentials  
<http://dev.twitter.com/apps>
- Go to Create New App and fill the form .
- Go to “Keys and Access Token” section .

- Create your access tokens.
- Copy API key , API secret , Access token , Access token secret from the page which is needed in twitterstream.py
- Copy the values in the twitterstream.py file in the sections given below :

```

1 import oauth2 as oauth
2 import urllib2 as urllib
3
4 # See Assignment 1 instructions or README for how to get these credentials
5 consumer_key = "<API key>"
6 consumer_secret = "<API secret key>"
7 access_token_key = "<Access token>"
8 access_token_secret = "<Access token secret>"
9
10 _debug = 0
11
12 oauth_token = oauth.Token(key=access_token_key, secret=access_token_secret)
13 oauth_consumer = oauth.Consumer(key=consumer_key, secret=consumer_secret)
14
15 signature_method_hmac_sha1 = oauth.SignatureMethod_HMAC_SHA1()
16
17 http_method = "GET"
18
19
20 http_handler = urllib.HTTPHandler(debugLevel=_debug)
21 https_handler = urllib.HTTPSHandler(debugLevel=_debug)

```

- Execute the following command to get the twitter data in ip.txt (Press Ctrl+C when enough data is acquired) :

***\$ py twitterstream.py > ip.txt***

## Running the Program :

- Create a Project directory is hadoop using :

***\$ hadoop fs -mkdir /Sentiment\_Analysis***

- Create a Input directory in project Sentiment\_Analysis

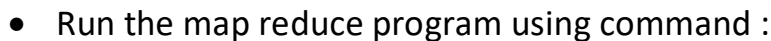
***\$ hadoop fs -mkdir /Sentiment\_Analysis/Input***

- Put your input file (ip.txt) and the cache file (AFINN.txt) in HDFS using :

***\$ hadoop fs -put ip.txt /Sentiment\_Analysis/Input***

***\$ hadoop fs -put AFINN.txt /Sentiment\_Analysis***

- ```
hadoop fs -cat /Sentiment_Analysis/Input/ip.txt
```



```
$ hadoop jar <hadoop-streaming_path> -input <hdfs_input_path> -output  
<hdfs_output_path> -mapper <mapper_code_path> -reducer  
<reducer_code_path> -cacheFile  
<cache file path>#<cache name used in code>
```

**For example :**

```
$ hadoop jar '/home/hadoop/Desktop/Sentiment_Analysis/hadoop-streaming-2.7.3.jar' -input /Sentiment_Analysis/Input/ip.txt -output /Sentiment_Analysis/Output -mapper '/home/hadoop/Desktop/Sentiment_Analysis/mapper2.py' -reducer '/home/hadoop/Desktop/Sentiment_Analysis/reducer2.py' -cacheFile /Sentiment_Analysis/AFINN.txt#cache
```

```
hadoop@prashik-aspire-e5-575: ~/Desktop/Sentiment_Analysis
hadoop@prashik-aspire-e5-575:~/Desktop/Sentiment_Analysis$ hadoop jar '/home/hadoop/Desktop/Mutual_Friends/hadoop-streaming-2.7.3.jar' -input /Sentiment_Analysis/Input/ip.txt -output /Sentiment_Analysis/Output -mapper '/home/hadoop/Desktop/Sentiment_Analysis/mapper2.py' -reducer '/home/hadoop/Desktop/Sentiment_Analysis/reducer2.py' -cacheFile /Sentiment_Analysis/AFINN.txt#cache
20/04/26 21:30:46 WARN streaming.StreamJob: -cacheFile option is deprecated, please use -files instead.
packageJobJar: [/tmp/hadoop-unjar8325799642217469987/] [] /tmp/streamjob2030615487895816911.jar tmpDir=null
20/04/26 21:30:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/26 21:30:48 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/26 21:30:48 INFO mapred.FileInputFormat: Total input files to process : 1
20/04/26 21:30:49 INFO mapreduce.JobSubmitter: number of splits:2
20/04/26 21:30:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1587852006321_0082
20/04/26 21:30:49 INFO impl.YarnClientImpl: Submitted application application_1587852006321_0082
20/04/26 21:30:49 INFO mapreduce.Job: The url to track the job: http://prashik-Aspire-E5-575:8088/proxy/application_1587852006321_0082/
20/04/26 21:30:49 INFO mapreduce.Job: Running job: job_1587852006321_0082
20/04/26 21:30:58 INFO mapreduce.Job: Job job_1587852006321_0082 running in uber mode : false
20/04/26 21:30:58 INFO mapreduce.Job: map 0% reduce 0%
20/04/26 21:31:06 INFO mapreduce.Job: map 50% reduce 0%
20/04/26 21:31:08 INFO mapreduce.Job: map 100% reduce 0%
20/04/26 21:31:18 INFO mapreduce.Job: map 100% reduce 100%
20/04/26 21:31:18 INFO mapreduce.Job: Job job_1587852006321_0082 completed successfully
20/04/26 21:31:18 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=727
  FILE: Number of bytes written=484566
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=181503
  HDFS: Number of bytes written=2857
  HDFS: Number of read operations=0
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=15031
  Total time spent by all reduces in occupied slots (ms)=8945
  Total time spent by all map tasks (ms)=15031
```

- View the map reduce output using command :

***\$ hadoop fs -cat /Sentiment\_Analysis/output/part-00000***

```
hadoop@prashik-aspire-e5-575: ~/Desktop/Sentiment_Analysis
hadoop@prashik-aspire-e5-575:~/Desktop/Sentiment_Analysis$ hadoop fs -cat /Sentiment_Analysis/output/part-00000
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=181293
File Output Format Counters
  Bytes Written=2857
20/04/26 21:31:18 INFO streaming.StreamJob: Output directory: /Sentiment_Analysis/Output
hadoop@prashik-aspire-e5-575:~/Desktop/Sentiment_Analysis$ hadoop fs -cat /Sentiment_Analysis/output/part-00000
Tweet ID : 1254425783960866816
Sentiment Score : 0
Sentiment Analysis : Neutral Tweet
Tweet ID : 1254425783960993792
Sentiment Score : 0
Sentiment Analysis : Neutral Tweet
Tweet ID : 1254425783961096193
Sentiment Score : 2
Sentiment Analysis : Positive Tweet
Tweet ID : 1254425783965061120
Sentiment Score : 0
Sentiment Analysis : Neutral Tweet
Tweet ID : 1254425783965188099
```

