

HOME ASSIGNMENT 2

Roll No: 208W1A1299

Name: MOHAMMAD RIZWANULLAH

PROJECT NAME: Twitter data analysis.

HadoopMapReduce-Twitter

Implementing MapReduce algorithms in Hadoop using the Twitter dataset(schema - <https://github.com/episod/twitter-api-fields-as-crowdsourced/wiki>)

Question answered:

1. What hour of the day does @PrezOno's tweet the most on average, using every day we have twitter data? Directory - <https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetsByHour>
2. What day of the week does @PrezOno tweet the most on average? Use the same example as in #1 but for days of the week. Directory - <https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetsByDay>
3. How does @PrezOno's tweet length compare to the average of all others? What is his average length? All others? Directory - <https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetLength>

Instructions:

A sample data file has been included in /data directory to support quick validations through the Hadoop streaming mode. However, the file does not contain tweets from @PrezOno and hence, it would be necessary update the user_name for filtering the tweets.

Sample command: `cat /data/sample-data | ./mapTweetsByHour.py | sort | ./reduceTweetsByHour.py`

To run the map reduce programs in the hadoop cluster, utilize the following command.

`hadoop jar /root/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input /data/twitter -output myoutput -file *.py -mapper mapTweetsByHour.py -reducer reduceTweetsByhour.py`

Execution Screenshots:

```
Applications: Places System cloudera@quickstart:~/Desktop/Worldwide-trade-data
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cd /home/cloudera/Desktop/worldwide-trade-data
[cloudera@quickstart ~]$ ls
2018-2019_export.csv  IndianTradeData.txt  map.txt
2018-2019_import.csv  map.txt
[cloudera@quickstart ~]$ cd /home/cloudera/Desktop/worldwide-trade-data
[cloudera@quickstart ~]$ hadoop jar Worldwidetrade.jar com.sahad.app.CYApp /user/cloudera/IndianTradeData/2018-2019_export.csv /user/cloudera/IndianTradeData/2018-2019_import.csv /user/cloudera/IndianTradeData/output2
Usage: hadoop jar IndianTradeData.jar <com.sahad.app.CYApp /com.sahad.app.CYApp>
p <input-data> <export-data> <output-path>
[cloudera@quickstart ~]$ hadoop jar Worldwidetrade.jar com.sahad.app.CYApp /user/cloudera/IndianTradeData/2018-2019_export.csv /user/cloudera/IndianTradeData/2018-2019_import.csv /user/cloudera/IndianTradeData/output2
22/09/20 09:22:17 INFO Client: Connecting to ResourceManager at /s.0.0.0:8032
22/09/20 09:22:18 INFO InputFileInputFormat: Total input paths to process : 1
22/09/20 09:22:18 INFO InputFileInputFormat: Total input paths to process : 1
22/09/20 09:22:18 INFO MapReduceJobSubmitter: Number of splits:2
22/09/20 09:22:18 INFO MapReduceJobSubmitter: Submitting tokens for job: job_1664462633488_0002
22/09/20 09:22:19 INFO impl: Submitting application application_1664462633488_0002
22/09/20 09:22:19 INFO MapReduceJob: The url to track the job: http://quickstart.cloudera:8080/proxy/application_1664462633488_0002/
22/09/20 09:22:19 INFO MapReduceJob: Running job: job_1664462633488_0002
22/09/20 09:22:26 INFO MapReduceJob: map 0% reduce 0%
22/09/20 09:22:26 INFO MapReduceJob: map 100% reduce 0%
22/09/20 09:22:26 INFO MapReduceJob: map 100% reduce 100%
22/09/20 09:22:40 INFO MapReduceJob: Job job_1664462633488_0002 completed successfully
22/09/20 09:22:40 INFO MapReduceJob: Counters: 58
File System Counters
  FILE: Number of bytes read=251567
  FILE: Number of bytes written=936834
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1812881
  HDFS: Number of bytes written=86561
  HDFS: Number of read operations=0
  HDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=19449
  Total time spent by all reduces in occupied slots (ms)=5582
  Total time spent by all map tasks (ms)=19449
  Total time spent by all reduce tasks (ms)=5582
  Total vcore-milliseconds taken by all map tasks=19449
  Total vcore-milliseconds taken by all reduce tasks=5582
  Total megabyte-milliseconds taken by all map tasks=19915776
  Total megabyte-milliseconds taken by all reduce tasks=5715968
(Worldwide-trade-data) cloudera@quickstart:~$
Killed map tasks=1
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=19449
Total time spent by all reduces in occupied slots (ms)=5582
Total time spent by all map tasks (ms)=19449
Total time spent by all reduce tasks (ms)=5582
Total vcore-milliseconds taken by all map tasks=19449
Total vcore-milliseconds taken by all reduce tasks=5582
Total megabyte-milliseconds taken by all map tasks=19915776
Total megabyte-milliseconds taken by all reduce tasks=5715968
Map-Reduce Framework
  Map input records=213149
  Map output records=9313
  Map output bytes=232875
  Map output materialized bytes=251513
  Input split bytes=578
  Combine input records=0
  Combine output records=0
  Reduce input groups=1677
  Reduce shuffle bytes=251513
  Reduce input records=9313
  Reduce output records=1677
  Spilled Records=18628
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=399
  CPU time spent (ms)=4368
  Physical memory (bytes) snapshot=563458848
  Virtual memory (bytes) snapshot=4319178248
  Total committed heap usage (bytes)=391579008
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=0
File Output Format Counters
  Bytes Written=86561
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/IndianTradeData/output2
and 2 items
w-r--r-- 1 cloudera cloudera 0 2022-09-20 09:22 /user/cloudera/IndianTradeData/output2/ SUCCESS
w-r--r-- 1 cloudera cloudera 86561 2022-09-20 09:22 /user/cloudera/IndianTradeData/output2/part-r-00000
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/IndianTradeData/output2/
```

Checking for output file.

HBaseImpalaSparkSolrOozieCloudera ManagerGetting Started

HadoopOverviewDatanodesSnapshotStartup ProgressUtilities

Browse Directory

/user/cloudera/indiantradedata

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxrwxrwx	root	cloudera	11.3 MB	Thu Sep 29 08:22:39 -0700 2022	1	128 MB	2018-2010_export.csv
-rwxrwxrwx	root	cloudera	6.26 MB	Thu Sep 29 08:23:18 -0700 2022	1	128 MB	2018-2010_import.csv
drwxr-xr-x	root	cloudera	0 B	Thu Sep 29 08:32:24 -0700 2022	0	0 B	output1
drwxr-xr-x	cloudera	cloudera	0 B	Thu Sep 29 09:22:44 -0700 2022	0	0 B	output2

Hadoop, 2017.

Output:

part-r-00000(1)			
File Edit View Search Tools Documents Help			
Open Save Undo			
part-r-00000(1)			
country=AFGHANISTAN TIS, year=2010	import 17.8	export 0.0	
country=AFGHANISTAN TIS, year=2011	import 9.8	export 0.5	
country=AFGHANISTAN TIS, year=2012	import 126.0	export 0.0	
country=AFGHANISTAN TIS, year=2013	import 5.3	export 0.0	
country=AFGHANISTAN TIS, year=2014	import 0.3	export 1.4	
country=AFGHANISTAN TIS, year=2015	import 8.9	export 0.0	
country=AFGHANISTAN TIS, year=2016	import 9.4	export 0.1	
country=AFGHANISTAN TIS, year=2017	import 3.8	export 8.9	
country=ALBANIA, year=2010	import 2.8	export 0.0	
country=ALBANIA, year=2011	import 3.1	export 0.0	
country=ALBANIA, year=2012	import 0.1	export 0.0	
country=ALBANIA, year=2013	import 0.8	export 0.0	
country=ALBANIA, year=2014	import 4.3	export 0.1	
country=ALBANIA, year=2015	import 2.1	export 0.0	
country=ALBANIA, year=2016	import 1.2	export 0.0	
country=ALBANIA, year=2017	import 1.8	export 0.0	
country=ALGERIA, year=2010	import 0.5	export 0.0	
country=ALGERIA, year=2011	import 0.1	export 0.0	
country=ALGERIA, year=2012	import 13.2	export 0.0	
country=ALGERIA, year=2013	import 2.7	export 0.7	
country=ALGERIA, year=2014	import 11.5	export 2.2	
country=ALGERIA, year=2015	import 16.7	export 0.0	
country=ALGERIA, year=2016	import 116.3	export 9.9	
country=ALGERIA, year=2017	import 6.2	export 7.7	
country=ALGERIA, year=2018	import 1.9	export 0.0	
country=AMERI SAMOA, year=2011	import 1.8	export 0.0	
country=ANDORRA, year=2010	import 0.2	export 0.0	
country=ANDORRA, year=2017	import 4.5	export 0.0	
country=ANGOLA, year=2010	import 3.5	export 2.7	
country=ANGOLA, year=2011	import 2.1	export 1.3	
country=ANGOLA, year=2012	import 4.3	export 0.0	
country=ANGOLA, year=2013	import 8.4	export 0.0	
country=ANGOLA, year=2014	import 3.5	export 0.5	
country=ANGOLA, year=2015	import 0.2	export 0.0	
country=ANGOLA, year=2016	import 10.9	export 0.0	
country=ANGOLA, year=2017	import 19.5	export 0.0	
country=ANGOLA, year=2018	import 11.4	export 0.0	
country=ANGUILLA, year=2010	import 0.1	export 0.0	
country=ANGUILLA, year=2011	import 0.0	export 0.1	
country=ANGUILLA, year=2012	import 0.1	export 0.0	
country=ANTARTICA, year=2013	import 0.0	export 0.4	
country=ANTARTICA, year=2014	import 0.0	export 4.8	
country=ANTARTICA, year=2015	import 0.0	export 2.2	
country=ANTARTICA, year=2017	import 0.1	export 0.0	

Output: We are able to get the most accurate times where celebrities tweet in which hour and in which day of the week