Project 2

BIG DATA

Group 3

Smriti Gupta
Swastika Bhat

Table of Contents

Part 1:

Query1:

Command:

Command to run the program to find the movies and the number of reviews

spark-submit --class analyzeimdbdatabase.PopularMovies --master yarn-cluster popularMovies2.jar ./dataset_large/movies/movies_large.csv ./dataset_large/reviews/reviews_large.csv ./output/Project2PopularMovies

You can find the source code at the following location
Group03Project2/Part1/SourceCode/

ScreenShots:

1. bigdata03@linux60818:~/Project2_Output/Project2PopularMovies (ssh)

```
[bigdata03@linux60818 ~]$ spark-submit --class analyzeimdbdatabase.PopularMovies --master yarn-cluster popularMovies2.j
ar ./dataset_large/movies/movies_large.csv ./dataset_large/reviews/reviews_large.csv ./output/Project2PopularMovies
18/06/12 15:41:36 INFO client.RMProxy: Connecting to ResourceManager at name1.hadoop.dc.engr.scu.edu/10.16.128.201:8032
18/06/12 15:41:36 INFO yarn.Client: Requesting a new application from cluster with 24 NodeManagers
18/06/12 15:41:36 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability
 of the cluster (19000 MB per container)
18/06/12 15:41:36 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
18/06/12 15:41:36 INFO yarn.Client: Setting up container launch context for our AM
18/06/12 15:41:36 INFO yarn.Client: Setting up the launch environment for our AM container
18/06/12 15:41:36 INFO yarn.Client: Preparing resources for our AM container
18/06/12 15:41:37 INFO yarn.Client: Uploading resource file:/DCNFS/users/student/bigdata03/popularMovies2.jar -> hdfs:/
/name1.hadoop.dc.engr.scu.edu:8020/user/bigdata03/.sparkStaging/application_1525447797409_22536/popularMovies2.jar
18/06/12 15:41:37 INFO yarn.Client: Uploading resource file:/tmp/spark-73d6c5f2-d46b-460e-8eca-09367cc8507c/__spark_con
f__559107702383029275.zip -> hdfs://name1.hadoop.dc.engr.scu.edu:8020/user/bigdata03/.sparkStaging/application_15254477
97409_22536/__spark_conf__559107702383029275.zip
18/06/12 15:41:37 INFO spark.SecurityManager: Changing view acls to: bigdata03
18/06/12 15:41:37 INFO spark.SecurityManager: Changing modify acls to: bigdata03
18/06/12 15:41:37 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with vi
ew permissions: Set(bigdata03); users with modify permissions: Set(bigdata03)
18/06/12 15:41:37 INFO yarn.Client: Submitting application 22536 to ResourceManager
18/06/12 15:41:37 INFO impl.YarnClientImpl: Submitted application application_1525447797409_22536
18/06/12 15:41:38 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:41:38 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.users.bigdata03
         start time: 1528843297470
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22536/
         user: bigdata03
18/06/12 15:41:39 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:41:40 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:41:41 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:41 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 10.16.128.108
         ApplicationMaster RPC port: 0
         queue: root.users.bigdata03
         start time: 1528843297470
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22536/
```

```
1. bigdata03@linux60818:~/Project2_Output/Project2PopularMovies (ssh)
         user: bigdata03
18/06/12 15:41:42 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:43 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:44 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:45 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:46 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:47 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:48 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:49 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:50 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:51 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:52 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:53 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:54 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:55 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:56 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:57 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:58 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:41:59 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:42:00 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:42:00 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.users.bigdata03
         start time: 1528843297470
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22536/
         user: bigdata03
18/06/12 15:42:01 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:42:02 INFO yarn.Client: Application report for application_1525447797409_22536 (state: ACCEPTED)
18/06/12 15:42:03 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
18/06/12 15:42:03 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 10.16.128.123
         ApplicationMaster RPC port: 0
         queue: root.users.bigdata03
         start time: 1528843297470
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22536/
         user: bigdata03
18/06/12 15:42:04 INFO yarn.Client: Application report for application_1525447797409_22536 (state: RUNNING)
```

Output was saved in a file and downloaded. The part files are present in the path Output was saved in a file and downloaded. The part files are present in the path Group03Project2/Part1/Output/PopularMovies

In total 6 part files were created. The part files contain (7659 + 8146 + 8051 + 5661 + 7736 + 7862) = 45115 records in total sorted in the ascending order of the number of reviews.

```
[bigdata03@linux60818 ~]$ hadoop fs -ls ./output
Found 1 items
drwxrwx---   - bigdata03 supergroup          0 2018-06-12 15:05 output/Project2PopularMovies
[bigdata03@linux60818 ~]$ hadoop fs -ls ./output/Project2PopularMovies
Found 7 items
-rw-rw----   3 bigdata03 supergroup          0 2018-06-12 15:05 output/Project2PopularMovies/_SUCCESS
-rw-rw----   3 bigdata03 supergroup     232412 2018-06-12 15:05 output/Project2PopularMovies/part-00000
-rw-rw----   3 bigdata03 supergroup     249966 2018-06-12 15:05 output/Project2PopularMovies/part-00001
-rw-rw----   3 bigdata03 supergroup     259128 2018-06-12 15:05 output/Project2PopularMovies/part-00002
-rw-rw----   3 bigdata03 supergroup     193713 2018-06-12 15:05 output/Project2PopularMovies/part-00003
-rw-rw----   3 bigdata03 supergroup     267360 2018-06-12 15:05 output/Project2PopularMovies/part-00004
-rw-rw----   3 bigdata03 supergroup     258823 2018-06-12 15:05 output/Project2PopularMovies/part-00005
[bigdata03@linux60818 ~]$ mkdir Project2_Output
[bigdata03@linux60818 ~]$ hadoop fs -copyToLocal ./output/Project2PopularMovies /home/bigdata03/Project2_Output/
[bigdata03@linux60818 ~]$ cd Project2_Output/
[bigdata03@linux60818 Project2_Output]$ ls
Project2PopularMovies
```

Query2

Command:

Command to run the program to find movies with average rating more than 4 and having more than 10 reviews

spark-submit --class analyzeimdbdatabase.TopRatedMovies --master yarn-cluster topRatedMovies2.jar ./dataset_large/movies/movies_large.csv ./dataset_large/reviews/reviews_large.csv ./output/Project2TopRatedMovies

You can find the source code at the following location
Group03Project2/Part1/SourceCode/topRatedMovies

Screenshots:

2. bigdata03@linux60813:~ (ssh)

```
[bigdata03@linux60813 ~]$ hadoop fs -rmr ./output/Project2TopRatedMovies; spark-submit
opRatedMovies --master yarn-cluster topRatedMovies2.jar ./dataset_large/movies/movies_
ews/reviews_large.csv ./output/Project2TopRatedMovies; date +%s%N | cut -b1-13
rmr: DEPRECATED: Please use 'rm -r' instead.
rmr: `./output/Project2TopRatedMovies': No such file or directory
18/06/12 21:41:18 INFO client.RMProxy: Connecting to ResourceManager at name1.hadoop.c
032
18/06/12 21:41:18 INFO yarn.Client: Requesting a new application from cluster with 24
18/06/12 21:41:18 INFO yarn.Client: Verifying our application has not requested more t
ity of the cluster (19000 MB per container)
18/06/12 21:41:18 INFO yarn.Client: Will allocate AM container, with 1408 MB memory ir
18/06/12 21:41:18 INFO yarn.Client: Setting up container launch context for our AM
18/06/12 21:41:18 INFO yarn.Client: Setting up the launch environment for our AM conta
18/06/12 21:41:18 INFO yarn.Client: Preparing resources for our AM container
18/06/12 21:41:18 INFO yarn.Client: Uploading resource file:/DCNFS/users/student/bigda
fs://name1.hadoop.dc.engr.scu.edu:8020/user/bigdata03/.sparkStaging/application_152544
jar
18/06/12 21:41:19 INFO yarn.Client: Uploading resource file:/tmp/spark-ca4576a7-a33c-4
conf__4611593524417313816.zip -> hdfs://name1.hadoop.dc.engr.scu.edu:8020/user/bigdata
525447797409_22561/__spark_conf__4611593524417313816.zip
18/06/12 21:41:19 INFO spark.SecurityManager: Changing view acls to: bigdata03
18/06/12 21:41:19 INFO spark.SecurityManager: Changing modify acls to: bigdata03
18/06/12 21:41:19 INFO spark.SecurityManager: SecurityManager: authentication disabled
 view permissions: Set(bigdata03); users with modify permissions: Set(bigdata03)
18/06/12 21:41:19 INFO yarn.Client: Submitting application 22561 to ResourceManager
18/06/12 21:41:19 INFO impl.YarnClientImpl: Submitted application application_15254477
18/06/12 21:41:20 INFO yarn.Client: Application report for application_1525447797409_2
18/06/12 21:41:20 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.users.bigdata03
         start time: 1528864879620
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525
         user: bigdata03
18/06/12 21:41:21 INFO yarn.Client: Application report for application_1525447797409_2
18/06/12 21:41:22 INFO yarn.Client: Application report for application_1525447797409_2
```

2. bigdata03@linux60813:~ (ssh)

```
[bigdata03@linux60813 ~]$ hadoop fs -rmr ./output/Project2TopRatedMovies; spark-submit --class analyzeimdbdatabase.T
opRatedMovies --master yarn-cluster topRatedMovies2.jar ./dataset_large/movies/movies_large.csv ./dataset_large/revi
ews/reviews_large.csv ./output/Project2TopRatedMovies; date +%s%N | cut -b1-13
rmr: DEPRECATED: Please use 'rm -r' instead.
rmr: `./output/Project2TopRatedMovies': No such file or directory
18/06/12 21:41:18 INFO client.RMProxy: Connecting to ResourceManager at name1.hadoop.dc.engr.scu.edu/10.16.128.201:8
032
18/06/12 21:41:18 INFO yarn.Client: Requesting a new application from cluster with 24 NodeManagers
18/06/12 21:41:18 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capabil
ity of the cluster (19000 MB per container)
18/06/12 21:41:18 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
18/06/12 21:41:18 INFO yarn.Client: Setting up container launch context for our AM
18/06/12 21:41:18 INFO yarn.Client: Setting up the launch environment for our AM container
18/06/12 21:41:18 INFO yarn.Client: Preparing resources for our AM container
18/06/12 21:41:18 INFO yarn.Client: Uploading resource file:/DCNFS/users/student/bigdata03/topRatedMovies2.jar -> hd
fs://name1.hadoop.dc.engr.scu.edu:8020/user/bigdata03/.sparkStaging/application_1525447797409_22561/topRatedMovies2.
jar
18/06/12 21:41:19 INFO yarn.Client: Uploading resource file:/tmp/spark-ca4576a7-a33c-4b60-a419-e75295255f4a/__spark_
conf__4611593524417313816.zip -> hdfs://name1.hadoop.dc.engr.scu.edu:8020/user/bigdata03/.sparkStaging/application_1
525447797409_22561/__spark_conf__4611593524417313816.zip
18/06/12 21:41:19 INFO spark.SecurityManager: Changing view acls to: bigdata03
18/06/12 21:41:19 INFO spark.SecurityManager: Changing modify acls to: bigdata03
18/06/12 21:41:19 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with
 view permissions: Set(bigdata03); users with modify permissions: Set(bigdata03)
18/06/12 21:41:19 INFO yarn.Client: Submitting application 22561 to ResourceManager
18/06/12 21:41:19 INFO impl.YarnClientImpl: Submitted application application_1525447797409_22561
18/06/12 21:41:20 INFO yarn.Client: Application report for application_1525447797409_22561 (state: ACCEPTED)
18/06/12 21:41:20 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.users.bigdata03
         start time: 1528864879620
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22561/
         user: bigdata03
18/06/12 21:41:21 INFO yarn.Client: Application report for application_1525447797409_22561 (state: ACCEPTED)
18/06/12 21:41:22 INFO yarn.Client: Application report for application_1525447797409_22561 (state: ACCEPTED)
18/06/12 21:41:23 INFO yarn.Client: Application report for application_1525447797409_22561 (state: ACCEPTED)
18/06/12 21:41:24 INFO yarn.Client: Application report for application_1525447797409_22561 (state: ACCEPTED)
[bigdata03@linux60813 ~]$ hadoop fs -ls ./output
Found 2 items
drwxrwx---   - bigdata03 supergroup          0 2018-06-12 15:05 output/Project2PopularMovies
drwxrwx---   - bigdata03 supergroup          0 2018-06-12 20:33 output/Project2TopRatedMovies
```

```
2. bigdata03@linux60813:~ (ssh)
18/06/12 21:41:33 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 10.16.128.123
         ApplicationMaster RPC port: 0
         queue: root.users.bigdata03
         start time: 1528864879620
         final status: UNDEFINED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22561/
         user: bigdata03
18/06/12 21:41:34 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:35 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:36 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:37 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:38 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:39 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:40 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:41 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:42 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:43 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:44 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:45 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:46 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:47 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:48 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:49 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:50 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:51 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:52 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:53 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:54 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:55 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:56 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:57 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:58 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:41:59 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:42:00 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:42:01 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:42:02 INFO yarn.Client: Application report for application_1525447797409_22561 (state: RUNNING)
18/06/12 21:42:03 INFO yarn.Client: Application report for application_1525447797409_22561 (state: FINISHED)
18/06/12 21:42:03 INFO yarn.Client:
```

```
18/06/12 21:42:03 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 10.16.128.123
         ApplicationMaster RPC port: 0
         queue: root.users.bigdata03
         start time: 1528864879620
         final status: SUCCEEDED
         tracking URL: http://name1.hadoop.dc.engr.scu.edu:8088/proxy/application_1525447797409_22561/
         user: bigdata03
18/06/12 21:42:03 INFO util.ShutdownHookManager: Shutdown hook called
18/06/12 21:42:03 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-ca4576a7-a33c-4b60-a419-e75295255f4a
1528864924229
```

Output was saved in a file and downloaded. The part files are present in the path Group03Project2/Part1/Output/TopRatedMovies

In total 6 part files were created. The part files contain (61+ 62+ 64+ 65+ 64+65) = 381records in total sorted in the descending order of the average rating.

Spark and MapReduce Comparison

The following table illustrates the time taken by a Java Spark job and a Java map-reduce job run on the cluster

	MapReduce(milliseconds)	Spark (milliseconds)
Query 1	52,800	22,470
Query 2	61,286	44,609

Calculating Spark Job Execution time:

Using the command “date +%s%N | cut -b1-13” we get the number of milliseconds since the epoch. i.e. the end time. Subtracting the start time as given by the Spark console from the end time we get the total execution time in milliseconds. Thus, time taken = start time – end time (in milliseconds).

Using the command (date +%s%N | cut -b1-13; hadoop jar popularmovie-0.0.1.jar popularmovie.PopularMovie ./dataset_large/reviews/reviews_large.csv ./dataset_large/movies/movies_large.csv ./output_large/; date +%s%N | cut -b1-13) and the same logic, we calculated the time taken by the map reduce job.

Observation:

From the table we see that Spark takes less time as compared to MapReduce.

Conclusion:

As per our Analysis, Spark is faster because,

- Spark makes use of lazy evaluation, thus optimizing the DAG and in turn the job performance.
- Spark stores the intermediate RDD’s in memory unlike Mapreduce where the intermediate results are written to file, then read and further tasks are performed. (In mapreduce, both queries have two tasks)

Part 2: Emerging Topic Detection & Sentiment Analysis

Command :

spark-submit --class emergingtopicdetector.EmergingTopicDetector --master local EmergingTopicDetector-jar-with-dependencies.jar <consumerKey> <consumerSecret> <accessToken> <accessTokenSecret>

```
1. bash
piyush@Piyushs-MacBook-Air:~/Desktop$ spark-submit --class emergingtopicdetector.EmergingTopicDetector --master local
EmergingTopicDetector-jar-with-dependencies.jar 0LdEBYiS7BHQHmWuf2EhP6LeT yijtynjCefki823wYlSg7Be2ruQcx0yFacl0fuKRhcfW
hTUGNq 850196196601716740-IifmurkoXbZDHMhhAXCncdP4WQHz5A2 FqgKIclEeyBwCYjrSYQBpYFh8f8oA7bDJEwqB1ywwcRd6
```

Output:

A part of the result of sentiment analysis for emerging topics is shown below. You can check the output file *'outputFile.txt'* generated in the folder where the '*EmergingTopicDetector-jar-with-dependencies.jar'* file is located.

```
piyush@Piyushs-MacBook-Air:~/Desktop$ spark-submit --class emergingtopicdetector.EmergingTopicDetector --master local EmergingTop
icDetector-jar-with-dependencies.jar 0LdEBYiS7BHQHmWuf2EhP6LeT yijtynjCefki823wYlSg7Be2ruQcx0yFacl0fuKRhcfWhTUGNq 850196196601716
740-IifmurkoXbZDHMhhAXCncdP4WQHz5A2 FqgKIclEeyBwCYjrSYQBpYFh8f8oA7bDJEwqB1ywwcRd6
2018-06-12 18:50:22 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
2018-06-12 18:50:23 INFO  StanfordCoreNLP:88 - Adding annotator tokenize
2018-06-12 18:50:23 INFO  TokenizerAnnotator:88 - No tokenizer type provided. Defaulting to PTBTokenizer.
2018-06-12 18:50:23 INFO  StanfordCoreNLP:88 - Adding annotator ssplit
2018-06-12 18:50:23 INFO  StanfordCoreNLP:88 - Adding annotator parse
2018-06-12 18:50:24 INFO  ParserGrammar:88 - Loading parser from serialized file edu/stanford/nlp/models/lexparser/englishPCFG.se
r.gz ... done [1.2 sec].
2018-06-12 18:50:24 INFO  StanfordCoreNLP:88 - Adding annotator sentiment

2018-06-12 18:50:35 INFO  TwitterStreamImpl:62 - Establishing connection.
2018-06-12 18:50:36 INFO  TwitterStreamImpl:62 - Connection established.
2018-06-12 18:50:36 INFO  TwitterStreamImpl:62 - Receiving status stream.
```

```
2018-06-12 18:50:35 INFO  TwitterStreamImpl:62 - Establishing connection.
2018-06-12 18:50:36 INFO  TwitterStreamImpl:62 - Connection established.
2018-06-12 18:50:36 INFO  TwitterStreamImpl:62 - Receiving status stream.

.....................................................................................................................
                                                        Window :1       Time: 1528854680155 ms
.....................................................................................................................

Topic ->5thFlowerPathWithBTS  ,  Sentiment -> NEUTRAL  ,        Content -> RT @snowberrytae: Q: what's your wish?
BTS: "the 7 of us always together."

#5thFlowerPathWithBTS @BTS_twt https://t.co/vse6tPK4Ve
Topic ->5thFlowerPathWithBTS  ,  Sentiment -> NEGATIVE  ,        Content -> RT @taesinqularity: #5thFlowerPathWithBTS
"No one knew, not even them." https://t.co/1lR25jgJ7b
Topic ->5thFlowerPathWithBTS  ,  Sentiment -> NEGATIVE  ,        Content -> RT @furerumatsu: ARMY's BIG Happiness List!! Let's walk
the #5thFlowerPathWithBTS in 2018🔥🔥🔥🔥🔥
#BTSFesta2018
#5thAnniversaryBTS

Feel fr…
Topic ->5thFlowerPathWithBTS  ,  Sentiment -> NEUTRAL  ,        Content -> RT @HallyuSG: [🎉] Happy 5th Anniversary to BTS! (@BTS_tw
t @bts_bighit) #5thFlowerPathWithBTS https://t.co/QvupmcnwAv

.....................................................................................................................
                                                        Window :2       Time: 1528854740239 ms
.....................................................................................................................

Topic ->BTS  ,   Sentiment -> NEGATIVE  ,        Content -> RT @kpoppingcom: #BTS 5th anniversary party by Naver x Dispatch

#5thFlowerPathWithBTS

🔗https://t.co/b3R5efM463 https://t.co/n7XbQne84S
Topic ->BTS  ,   Sentiment -> NEUTRAL  ,        Content -> RT @radiodisney: Congrats to #BTS on their 5th anniversary! Watch them t
alking about #FakeLove and #RDMA nominations while at the Radio Dis…
Topic ->BTS  ,   Sentiment -> NEGATIVE  ,        Content -> RT @1thepollskpop: [REQ]

BEST VOCALIST OF ALL TIME 🎤
(dont be biased!!)

#EXO #CHEN #EXOL #BTS #BTSARMY #JIMIN #BIGBANG #TAEYANG #SUPERJU…
Topic ->BTS  ,   Sentiment -> POSITIVE  ,        Content -> RT @abc7george: MUSIC NEWS - @billboard staff picks the top 50 #BTS song
s.  Wonder if #BTSARMY agrees.  Dare I ask? @BTS_twt
https://t.co/…
Topic ->BTS  ,   Sentiment -> NEGATIVE  ,        Content -> RT @soompi: UNICEF Thanks #BTS And ARMY For Raising Over $1 Million Thro
ugh "Love Myself" Campaign #5thFlowerPathWithBTS https://t.co/twsun…
Topic ->BTS  ,   Sentiment -> NEGATIVE  ,        Content -> RT @GetOnSwag: ⚡"180612 BTS Dinner Party" #BTS  #2018BTSFESTA

https://t.co/GgZo6OfXUB


.....................................................................................................................
                                                        Window :3       Time: 1528854800441 ms
.....................................................................................................................
```

How to detect emerging topic?:

We decided to detect emerging topic based on the increase in the number of tweets a particular topic gets between two consecutive windows. An emerging topic would be the one that has the maximum increase in the number of tweets from the number of tweets in the previous window.

We have made the following assumptions :

- Topic = hashtag.
- Number of tweets for a topic = number of tweets in which the topic(hashtag) appears

For eg: Consider the below scenario:
Batch size: 10 secs
Window size: 60 secs
Slide duration : 60 secs

Window 1 (6 batches) has topics ‘BigData’ and ‘Trump’.
‘BigData’ has 10 tweets.
‘Trump’ has 20 tweets.

Window 2 (6 batches) has topics ‘BigData’, ‘Trump’ and ‘Hadoop’.
‘BigData’ has 50 tweets.
‘Trump’ has 20 tweets.
‘Hadoop’ has 20 tweets.

Window 3 (6 batches) has topics ‘BigData’, ‘DistributedSystems’ and ‘Hadoop’.
‘BigData’ has 40 tweets.
‘DistributedSystems’ has 20 tweets.
‘Hadoop’ has 50 tweets.

So after Window 2, the emerging topic would be ‘BigData’ since its number of tweets increased by 40 while ‘Trump’ has no increase in tweets and ‘Hadoop’ has an increase of 20 tweets. Max increase in tweets is that of ‘BigData’.

Similarly, after Window 3, the emerging topic is ‘Hadoop’ since its number of tweets has increased by 30 from Window 2.

In other words, for each window (window size of 60 secs) we do the following:
For each tweet in all batches(batch size=10 secs) in a window size of 60 secs, we create a Hashmap of (topic, List<Status>). Status is the tweet object received from Twitter stream.

We, then compare the difference in the number of tweets for a topic using the hashmap created for current window (new hashmap) and hashmap created for previous window (old hashmap). If a topic does not appear in old hashmap then its count in the old hash map is considered to be zero.

We pick the topic (hashtag) with the maximum increase, in number of tweets, as the emerging topic. If two or more topics have maximum increase in number of tweets then all these topics are emerging topics.

After the end of a window size,

- The tweets of the emerging topics are then sent for sentiment analysis.
- Old hash map=new hashmap
- New hashmap is created again for the current window.