Management of Big Data and Tools – DS8003 – Fall 2016

Assignment 04

NAJLIS, BERNARDO - Student Number #500744793

Datasets:

1. **full_text_clean.txt:** (userid, lat, lon, tweet, modified_lat, modified_lon) [[D2L -> Assignment 3 – Pig -> full_text_clean.txt]
2. **cities_clean.txt:** (city_name, lat, lon, modified_lat, modified_lon) [D2L -> Assignment 3 – Pig -> cities_clean.txt]
3. Copy **full_text_clean.txt and cities_clean.txt** to the virtual machine (**Filezilla**)
4. Copy **full_text_clean.txt and cities_clean.txt** from virtual machine into HDFS (**hadoop fs -put**)

Submission:

1. Submit the pig commands (copy it into a file and submit)
2. Submit using Assessment -> Dropbox -> Pig Assignment 3

Assignment (Total of 10)

1) Get a 10% sample from full_text_clean.txt and store results in full_text_small.txt **(2)**

2) Find the top 3 words used in tweets from file full_text_clean.txt **(2)**

3) Calculate total number of records in cities_clean.txt file **(2)**

4) Find closest city for each tweet **(4)**

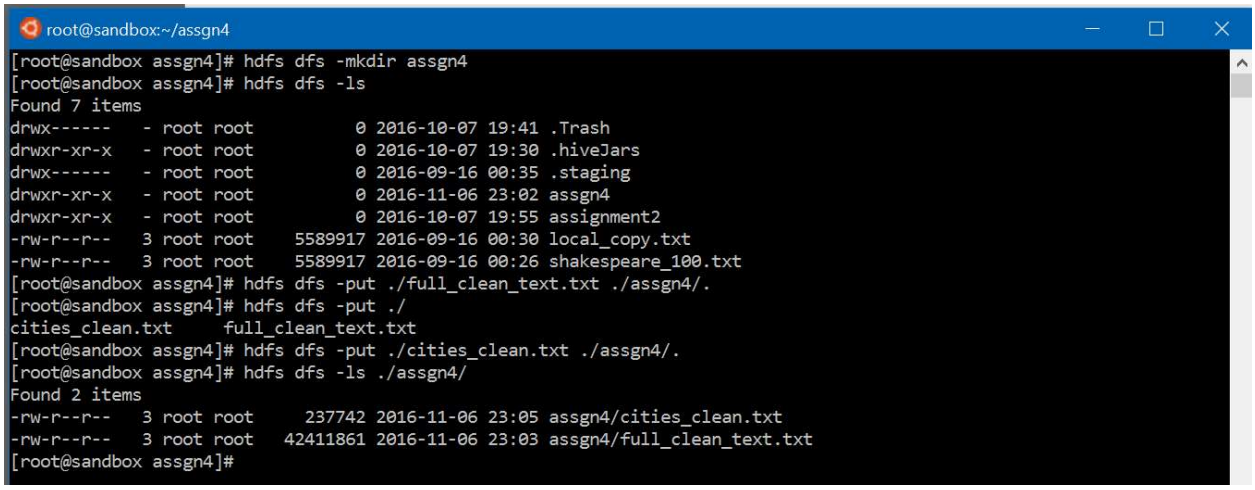
Hint:

- The lat-lon in full_text_clean.txt file doesn't match directly with the lat-lon of the cities in cities_clean.txt file. For that purpose, I have pre-processed both files to include a modified lat and lon column (last two columns of both files). So for each of geo-tagged tweets, you will map to multiple nearby cities using the last two columns of both files. After that, for each geo-tagged tweet, you then calculate the distance using the actual lat-lon values and pick the closest city.
- Calculating Euclidean Distance (pig example)
    - SQRT((lat_1 – lat_2) * (lat_1 – lat_2) + (lon_1 – lon_2) * (lon_1 – lon_2))
    - Lat_1/Lon_1 refer to lat/lon in full_text_clean.txt
    - Lat_2/Lon_2 refer to lat/lon in cities_clean.txt

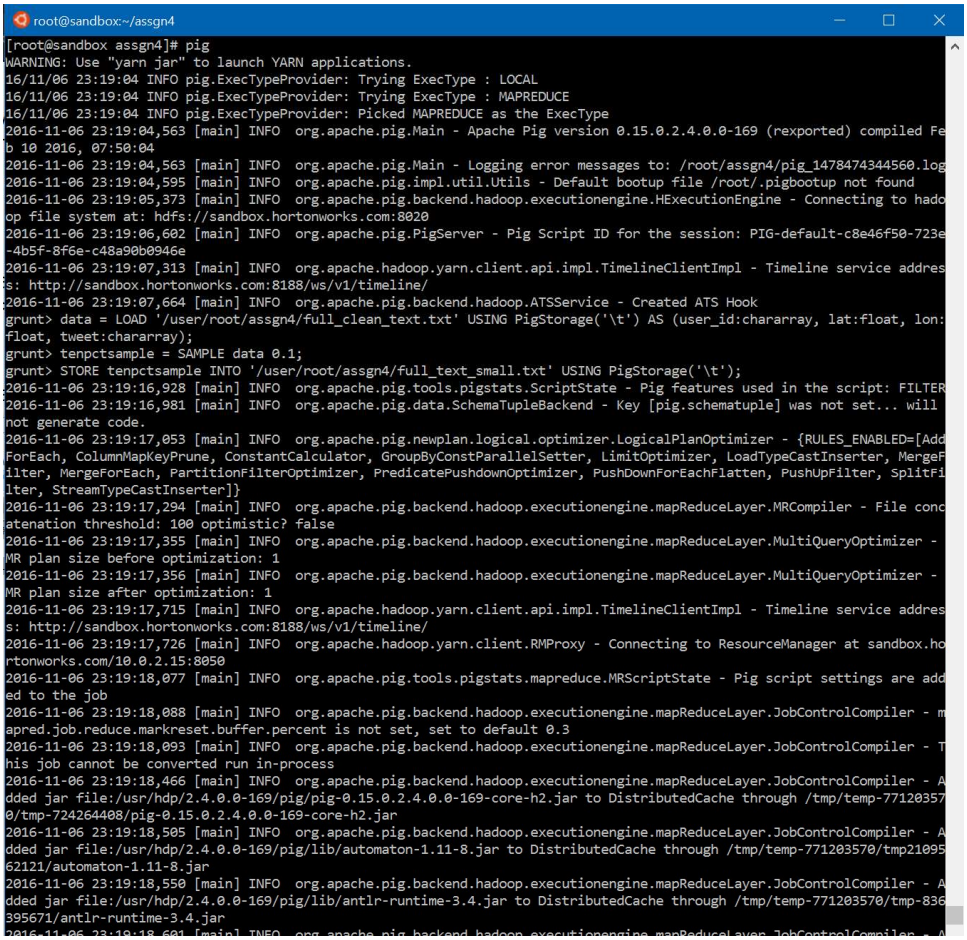
Resolution

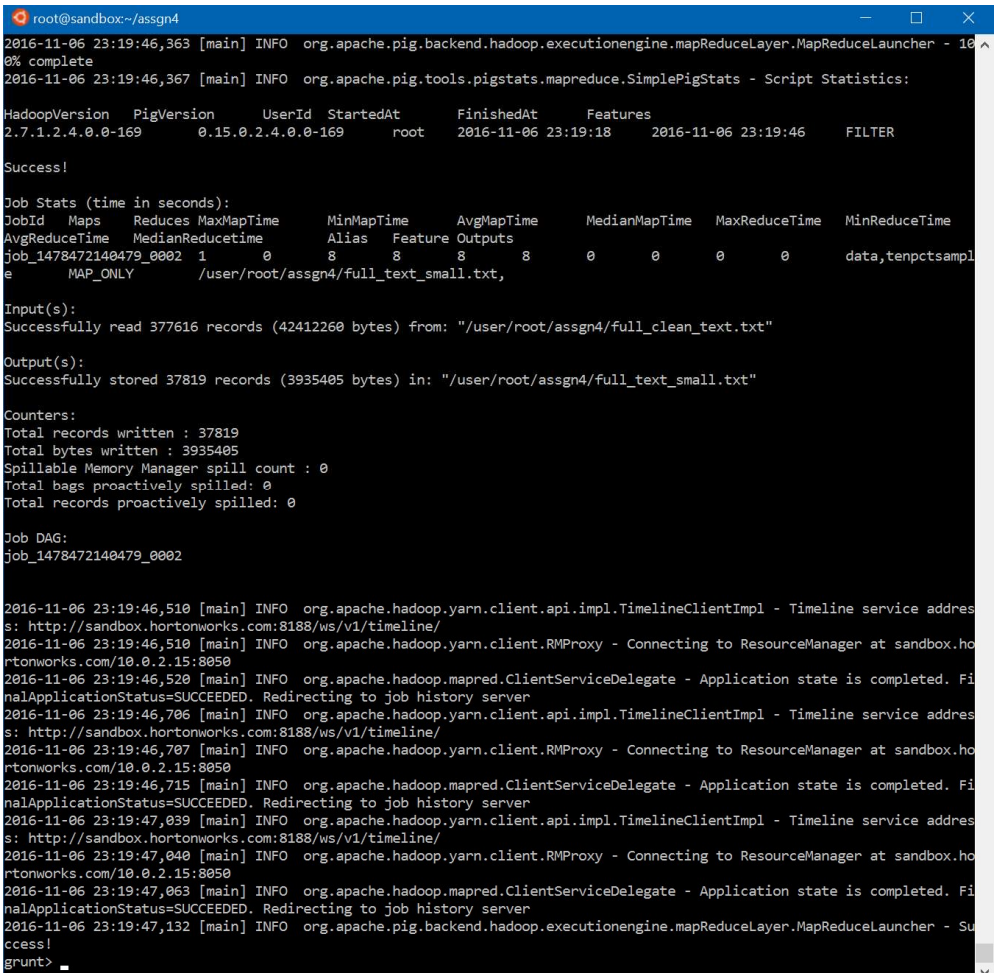
0. Environment setup
    a. Files copied into HDFS and increase Java VM heap size for PIG.

```
hdfs dfs -mkdir assgn4
hdfs dfs -put ./full_clean_text.txt ./assgn4/.
hdfs dfs -put ./cities_clean.txt ./assgn4/.
export PIG_HEAPSIZE=8192
```

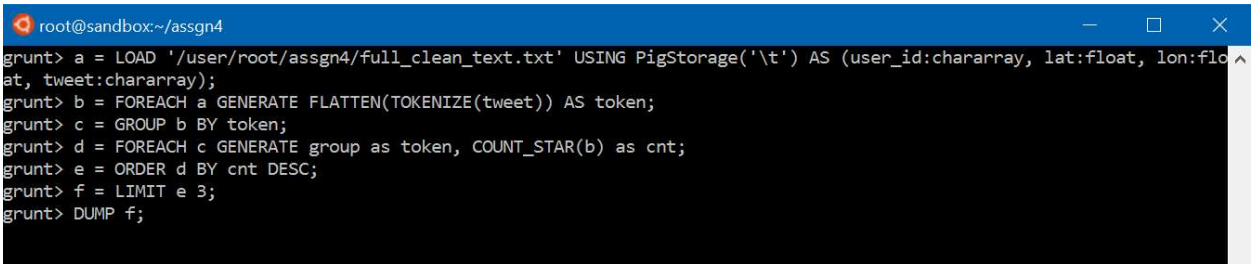


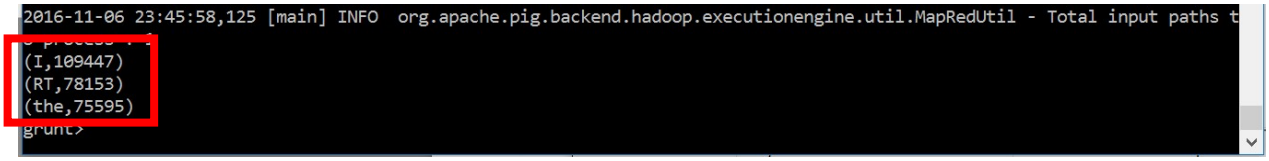
1.
```
data = LOAD '/user/root/assgn4/full_clean_text.txt' USING
PigStorage('\t') AS (user_id:chararray, lat:float, lon:float,
tweet:chararray, mod_lat:int, mod_lon:int);
tenpctsample = SAMPLE data 0.1;
STORE tenpctsample INTO '/user/root/assgn4/full_text_small.txt'
USING PigStorage('\t');
quit

hdfs dfs -get /user/root/assgn4/full_text_clean.txt/part-m-00000
./full_text_clean.txt
```


root@sandbox:~/assgn4
[root@sandbox assgn4]# pig
WARNING: Use "yarn jar" to launch YARN applications.
16/11/06 23:19:04 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/06 23:19:04 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/06 23:19:04 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-06 23:19:04,563 [main] INFO org.apache.pig.Main - Apache Pig version 0.15.0.2.4.0.0-169 (rexported) compiled Fe
b 10 2016, 07:50:04
2016-11-06 23:19:04,563 [main] INFO org.apache.pig.Main - Logging error messages to: /root/assgn4/pig_1478474344560.log
2016-11-06 23:19:04,595 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2016-11-06 23:19:05,373 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://sandbox.hortonworks.com:8020
2016-11-06 23:19:06,602 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-c8e46f50-723e
-4b5f-8f6e-c48a90b0946e
2016-11-06 23:19:07,313 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service addres
s: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
2016-11-06 23:19:07,664 [main] INFO org.apache.pig.backend.hadoop.ATSService - Created ATS Hook
grunt> data = LOAD '/user/root/assgn4/full_clean_text.txt' USING PigStorage('\t') AS (user_id:chararray, lat:float, lon:
float, tweet:chararray);
grunt> tenpctsample = SAMPLE data 0.1;
grunt> STORE tenpctsample INTO '/user/root/assgn4/full_text_small.txt' USING PigStorage('\t');
2016-11-06 23:19:16,928 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2016-11-06 23:19:16,981 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will
not generate code.
2016-11-06 23:19:17,053 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[Add
ForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
ilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFi
lter, StreamTypeCastInserter]}
2016-11-06 23:19:17,294 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File conc
atenation threshold: 100 optimistic? false
2016-11-06 23:19:17,355 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size before optimization: 1
2016-11-06 23:19:17,356 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size after optimization: 1
2016-11-06 23:19:17,715 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service addres
s: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
2016-11-06 23:19:17,726 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.ho
rtonworks.com/10.0.2.15:8050
2016-11-06 23:19:18,077 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are add
ed to the job
2016-11-06 23:19:18,088 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - m
apred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-11-06 23:19:18,093 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - T
his job cannot be converted run in-process
2016-11-06 23:19:18,466 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - A
dded jar file:/usr/hdp/2.4.0.0-169/pig/pig-0.15.0.2.4.0.0-169-core-h2.jar to DistributedCache through /tmp/temp-77120357
0/tmp-724264408/pig-0.15.0.2.4.0.0-169-core-h2.jar
2016-11-06 23:19:18,505 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - A
dded jar file:/usr/hdp/2.4.0.0-169/pig/lib/automaton-1.11-8.jar to DistributedCache through /tmp/temp-771203570/tmp21095
62121/automaton-1.11-8.jar
2016-11-06 23:19:18,550 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - A
dded jar file:/usr/hdp/2.4.0.0-169/pig/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/temp-771203570/tmp-836
395671/antlr-runtime-3.4.jar

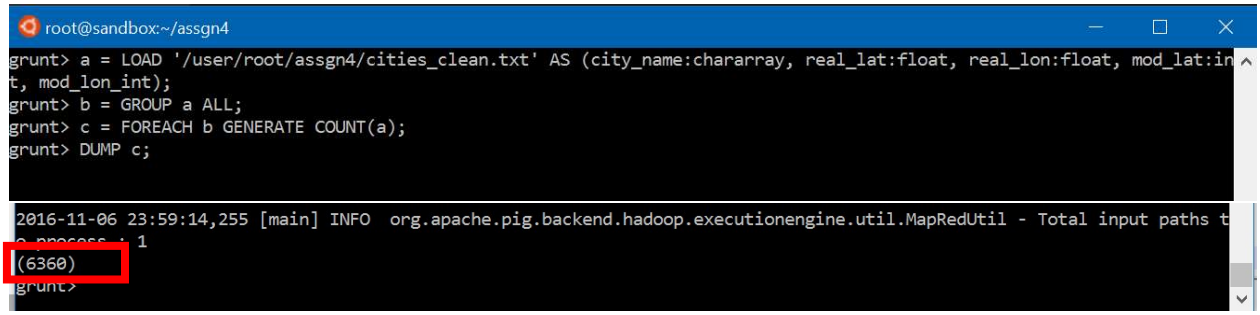


2.
```
a = LOAD '/user/root/assgn4/full_clean_text.txt' USING
PigStorage('\t') AS (user_id:chararray, lat:float, lon:float,
tweet:chararray, mod_lat:int, mod_lon:int);
b = FOREACH a GENERATE FLATTEN(TOKENIZE(tweet)) AS token;
c = GROUP b BY token;
d = FOREACH c GENERATE group as token, COUNT_STAR(b) as cnt;
e = ORDER d BY cnt DESC;
f = LIMIT e 3;
DUMP f;
```



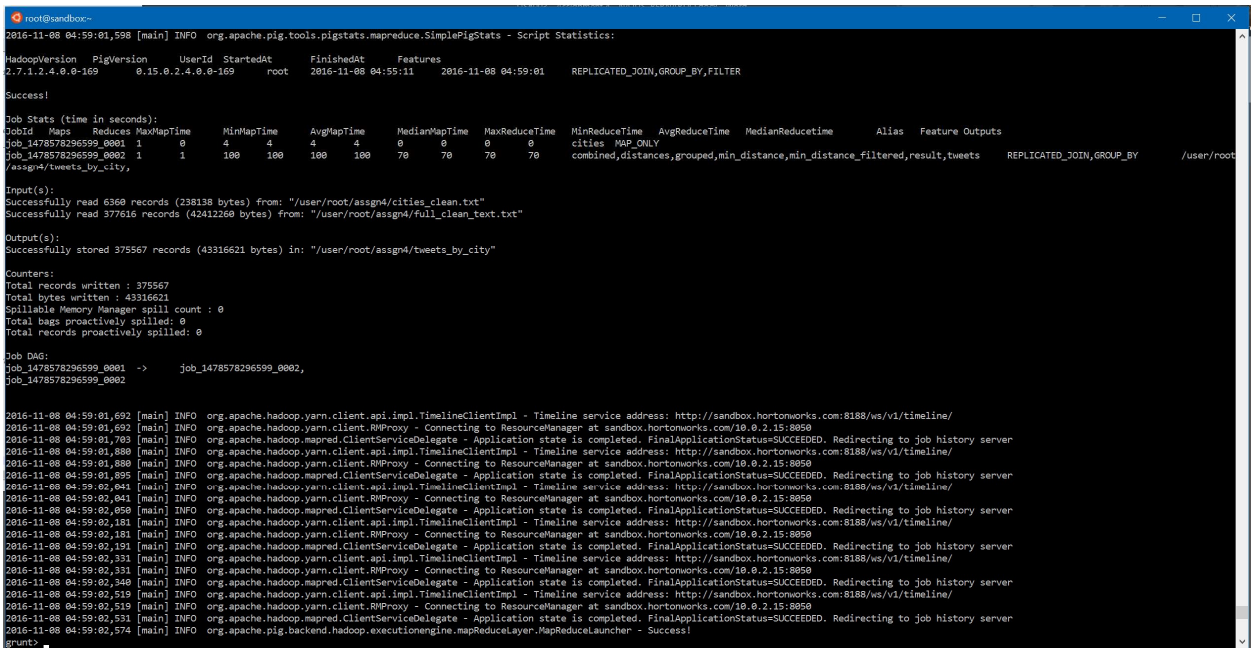


3.
```
a = LOAD '/user/root/assgn4/cities_clean.txt' AS
(city_name:chararray, real_lat:float, real_lon:float,
mod_lat:int, mod_lon_int);
b = GROUP a ALL;
c = FOREACH b GENERATE COUNT(a);
DUMP c;
```

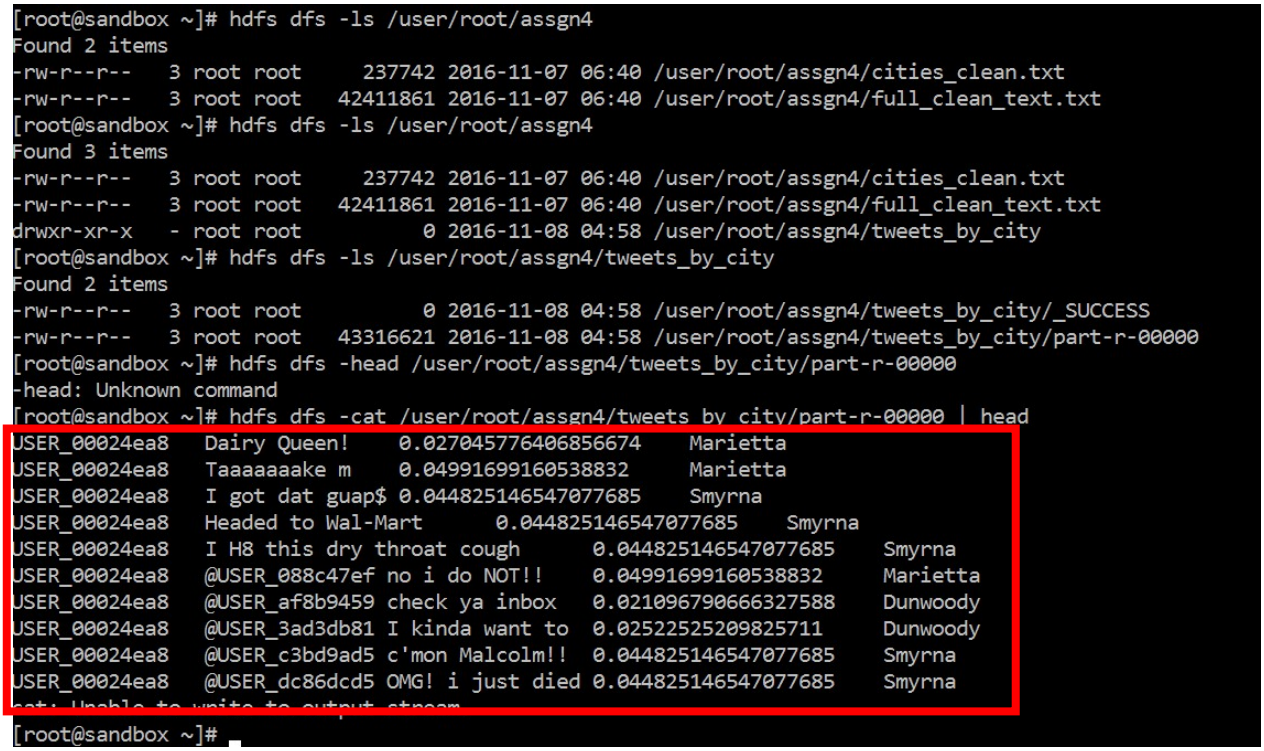


4.
```
tweets = LOAD '/user/root/assgn4/full_clean_text.txt'
      USING PigStorage('\t') AS (
            user_id:chararray,
            real_lat:float,
            real_lon:float,
            tweet:chararray,
            mod_lat:int,
            mod_lon:int
      );

cities = LOAD '/user/root/assgn4/cities_clean.txt' AS (
            city_name:chararray,
            real_lat:float,
            real_lon:float,
            mod_lat:int,
            mod_lon:int
      );

combined = JOIN  tweets BY
      (mod_lat, mod_lon) LEFT OUTER,
      cities BY (mod_lat, mod_lon)
      USING 'replicated';

distances = FOREACH combined GENERATE
            tweets::user_id AS user_id,
            tweets::tweet AS tweet,
            cities::city_name AS city_name,
```

```
          SQRT((tweets::real_lat - cities::real_lat) *
(tweets::real_lat - cities::real_lat) + (tweets::real_lon -
cities::real_lon) * (tweets::real_lon - cities::real_lon)) as
real_dist
      ;

grouped = GROUP distances BY (user_id, tweet);

min_distance = FOREACH grouped
      GENERATE group,
      MIN(distances.real_dist) as distance,
      FLATTEN(distances);

min_distance_filtered = FILTER min_distance BY distance ==
distances::real_dist;

result = FOREACH min_distance_filtered GENERATE
      FLATTEN(group) AS (user_id, tweet),
      distance,
      distances::city_name AS city_name;

STORE result INTO '/user/root/assgn4/tweets_by_city' USING
PigStorage('\t');
```



Result:


[root@sandbox ~]# hdfs dfs -ls /user/root/assgn4
Found 2 items
-rw-r--r--   3 root root   237742 2016-11-07 06:40 /user/root/assgn4/cities_clean.txt
-rw-r--r--   3 root root   42411861 2016-11-07 06:40 /user/root/assgn4/full_clean_text.txt
[root@sandbox ~]# hdfs dfs -ls /user/root/assgn4
Found 3 items
-rw-r--r--   3 root root   237742 2016-11-07 06:40 /user/root/assgn4/cities_clean.txt
-rw-r--r--   3 root root   42411861 2016-11-07 06:40 /user/root/assgn4/full_clean_text.txt
drwxr-xr-x   - root root   0 2016-11-08 04:58 /user/root/assgn4/tweets_by_city
[root@sandbox ~]# hdfs dfs -ls /user/root/assgn4/tweets_by_city
Found 2 items
-rw-r--r--   3 root root   0 2016-11-08 04:58 /user/root/assgn4/tweets_by_city/_SUCCESS
-rw-r--r--   3 root root   43316621 2016-11-08 04:58 /user/root/assgn4/tweets_by_city/part-r-00000
[root@sandbox ~]# hdfs dfs -head /user/root/assgn4/tweets_by_city/part-r-00000
-head: Unknown command
[root@sandbox ~]# hdfs dfs -cat /user/root/assgn4/tweets_by_city/part-r-00000 | head
USER_00024ea8   Dairy Queen!   0.027045776406856674   Marietta
USER_00024ea8   Taaaaaaake m   0.04991699160538832   Marietta
USER_00024ea8   I got dat guap$ 0.044825146547077685   Smyrna
USER_00024ea8   Headed to Wal-Mart   0.044825146547077685   Smyrna
USER_00024ea8   I H8 this dry throat cough   0.044825146547077685   Smyrna
USER_00024ea8   @USER_088c47ef no i do NOT!!   0.04991699160538832   Marietta
USER_00024ea8   @USER_af8b9459 check ya inbox   0.021096790666327588   Dunwoody
USER_00024ea8   @USER_3ad3db81 I kinda want to 0.02522525209825711   Dunwoody
USER_00024ea8   @USER_c3bd9ad5 c'mon Malcolm!! 0.044825146547077685   Smyrna
USER_00024ea8   @USER_dc86dcd5 OMG! i just died 0.044825146547077685   Smyrna
[root@sandbox ~]#