Twitter Data Analysis

Scenario:

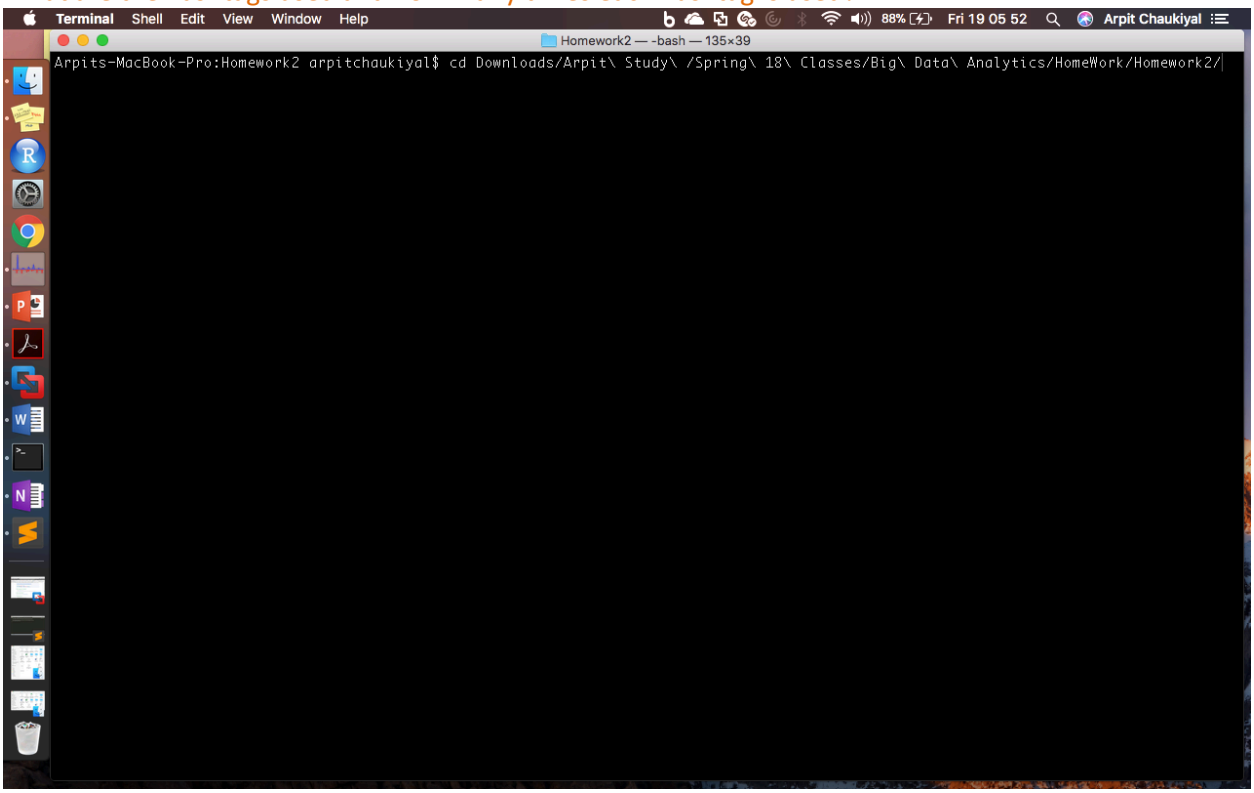
You are provided with two files to perform basic data analysis:

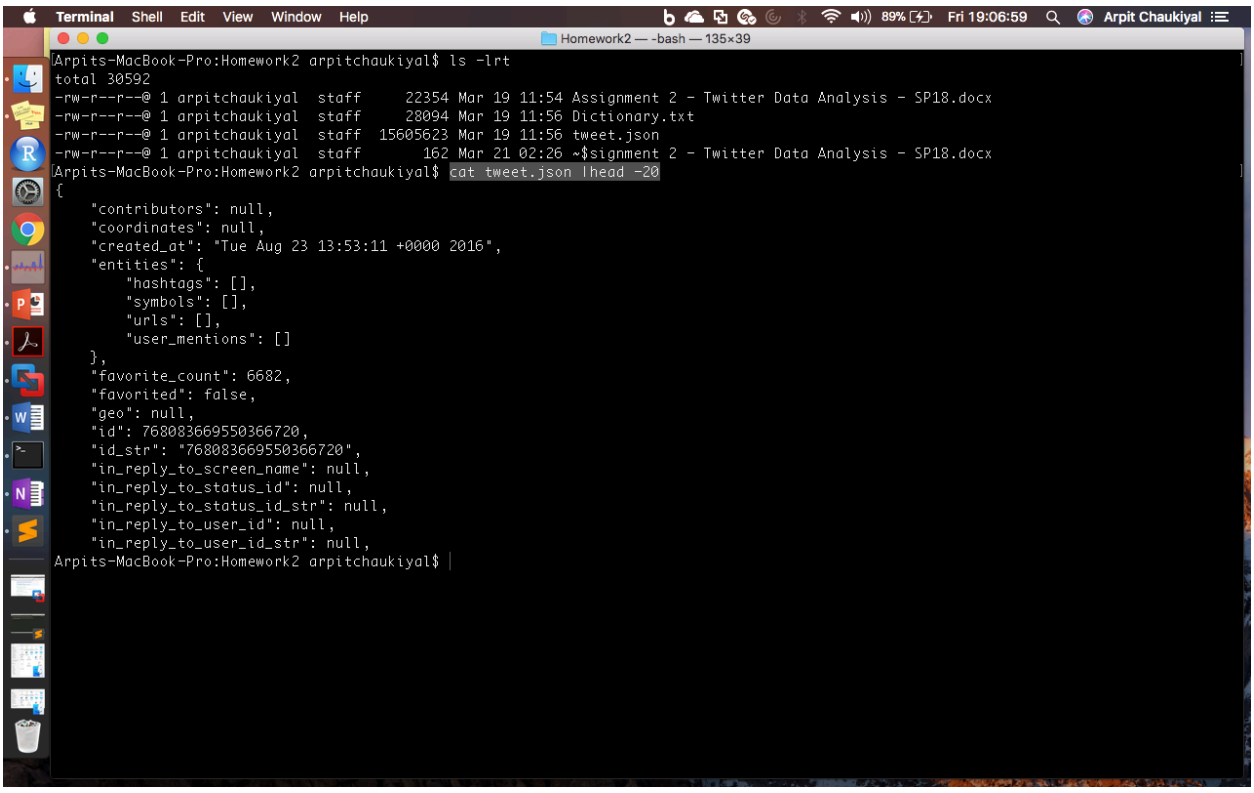
- tweet.json – Twitter feed
- Dictionary.txt – Contains words with a rating for each word

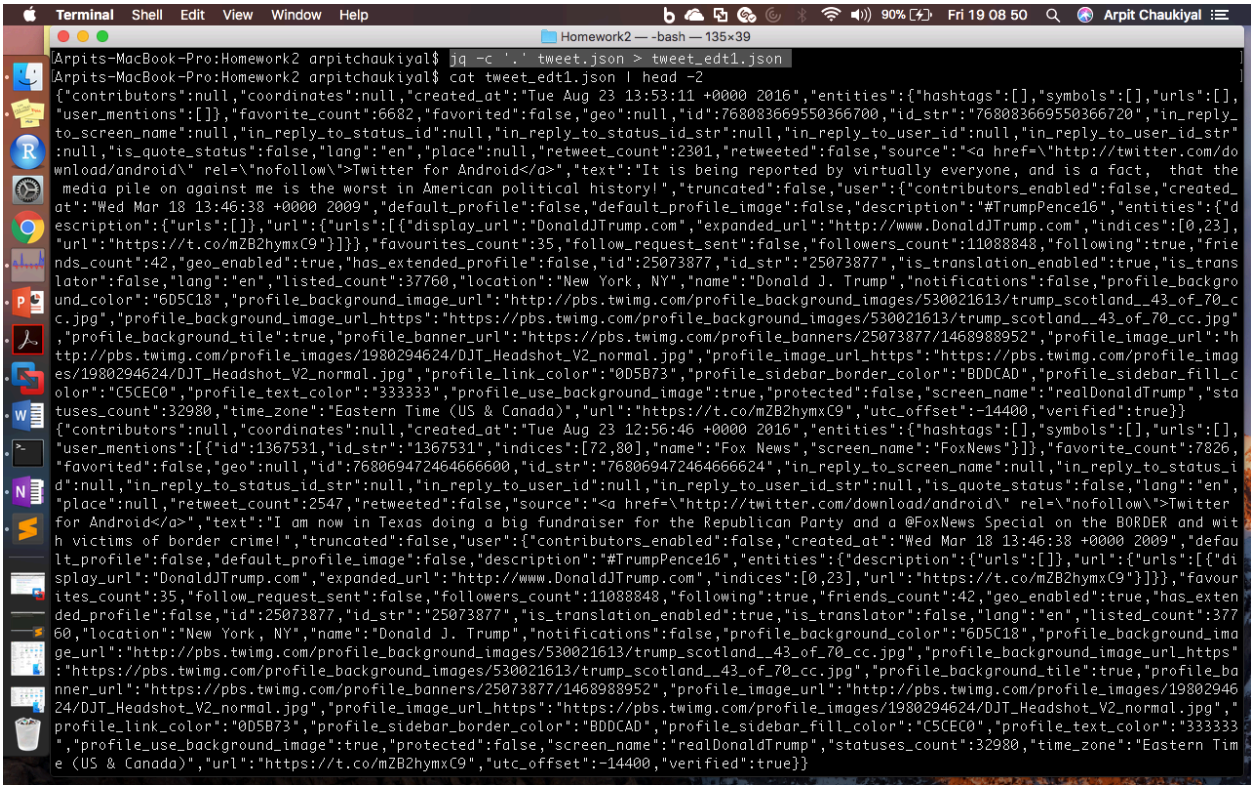
You are required to analyze user comments in the tweet.json file and assign it a positive or negative polarity based on the words available in the Dictionary file. (Please feel free to add more words to the dictionary file if needed). You might have to clean the JSON file before importing it in Hive.

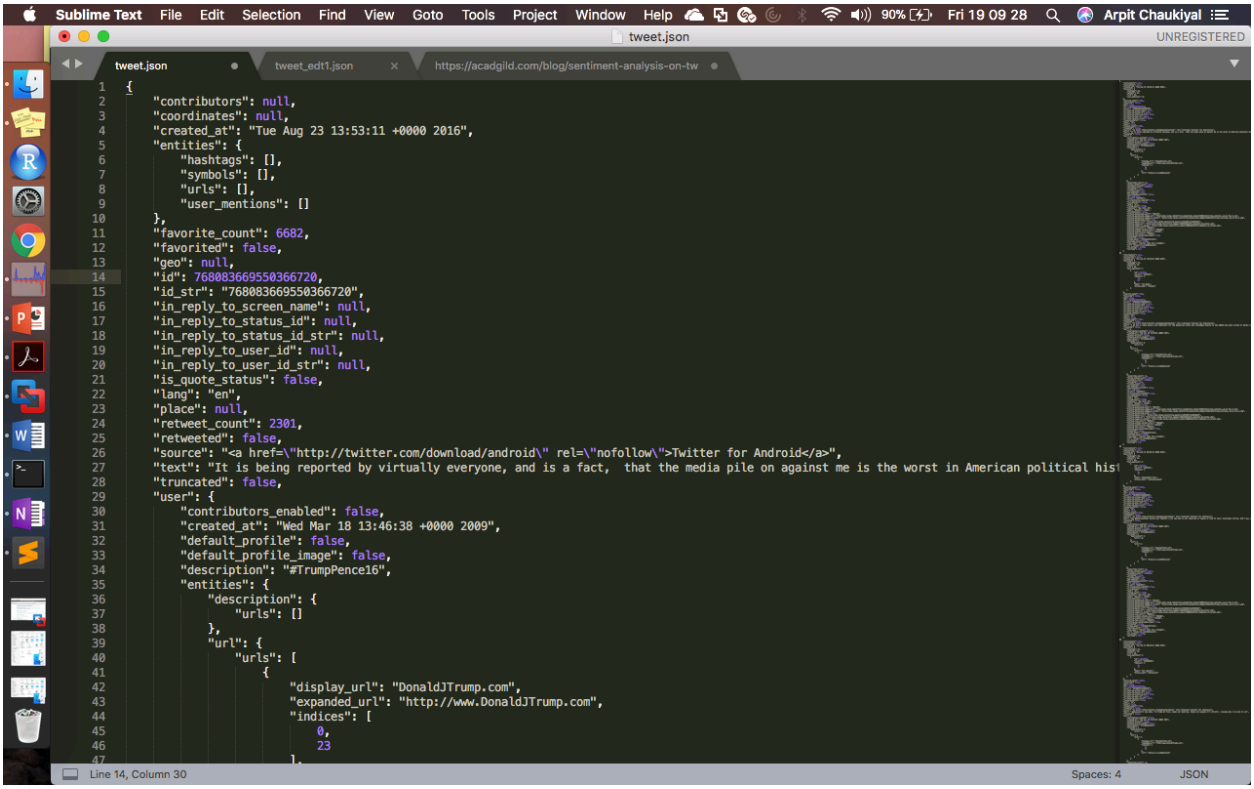
Insights:

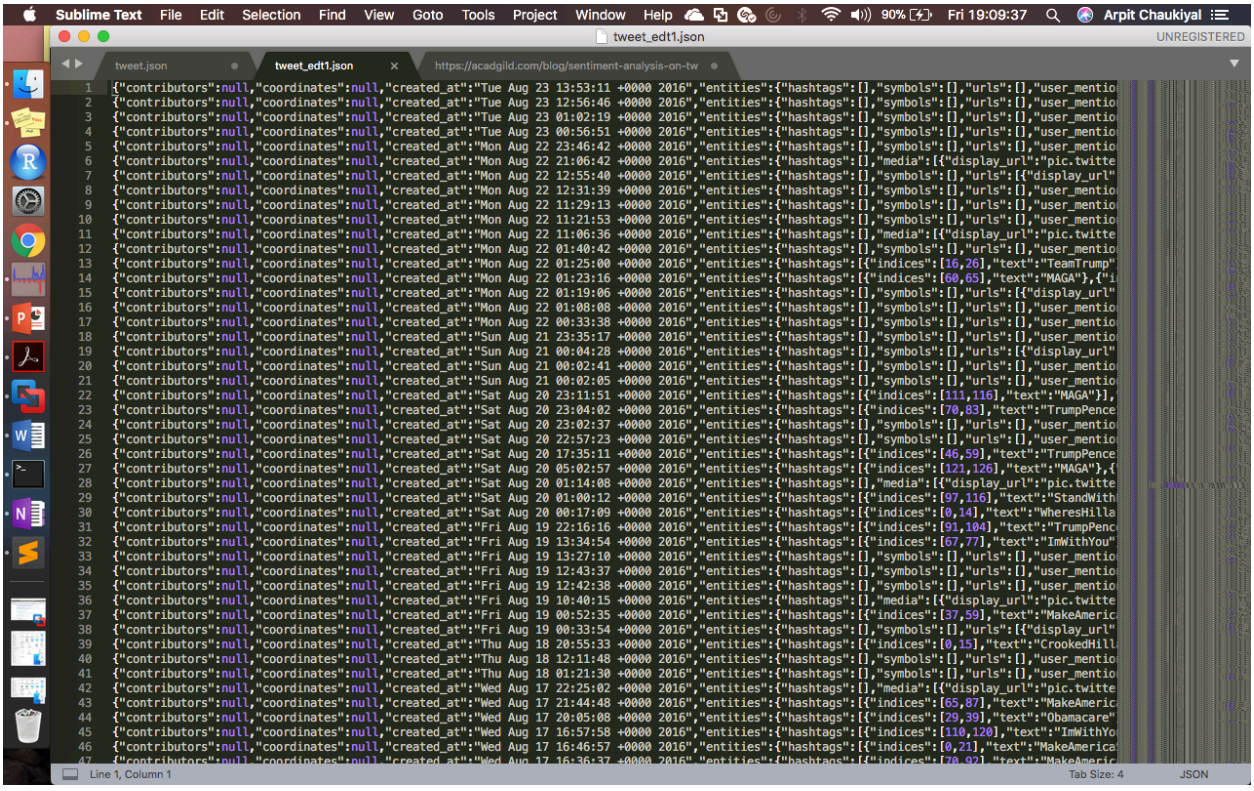
a. What are the hashtags used and how many times each hashtag is used?

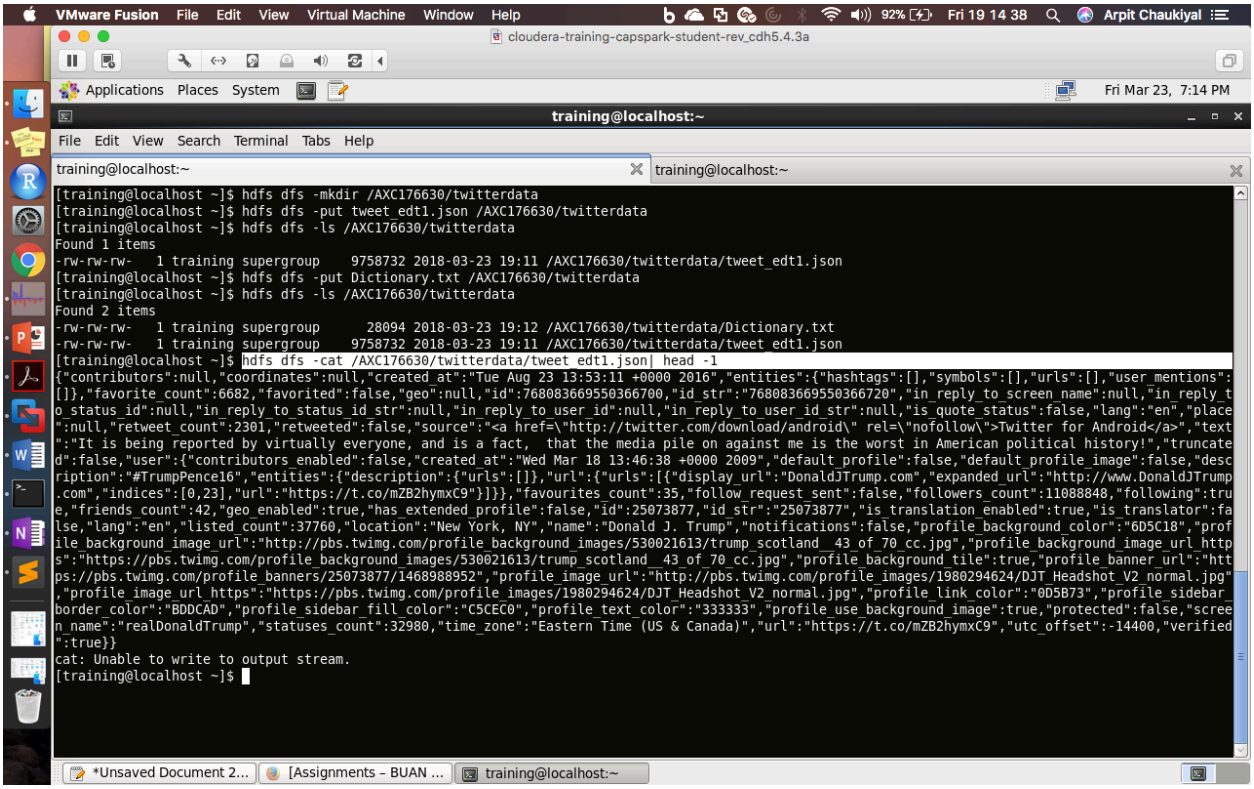


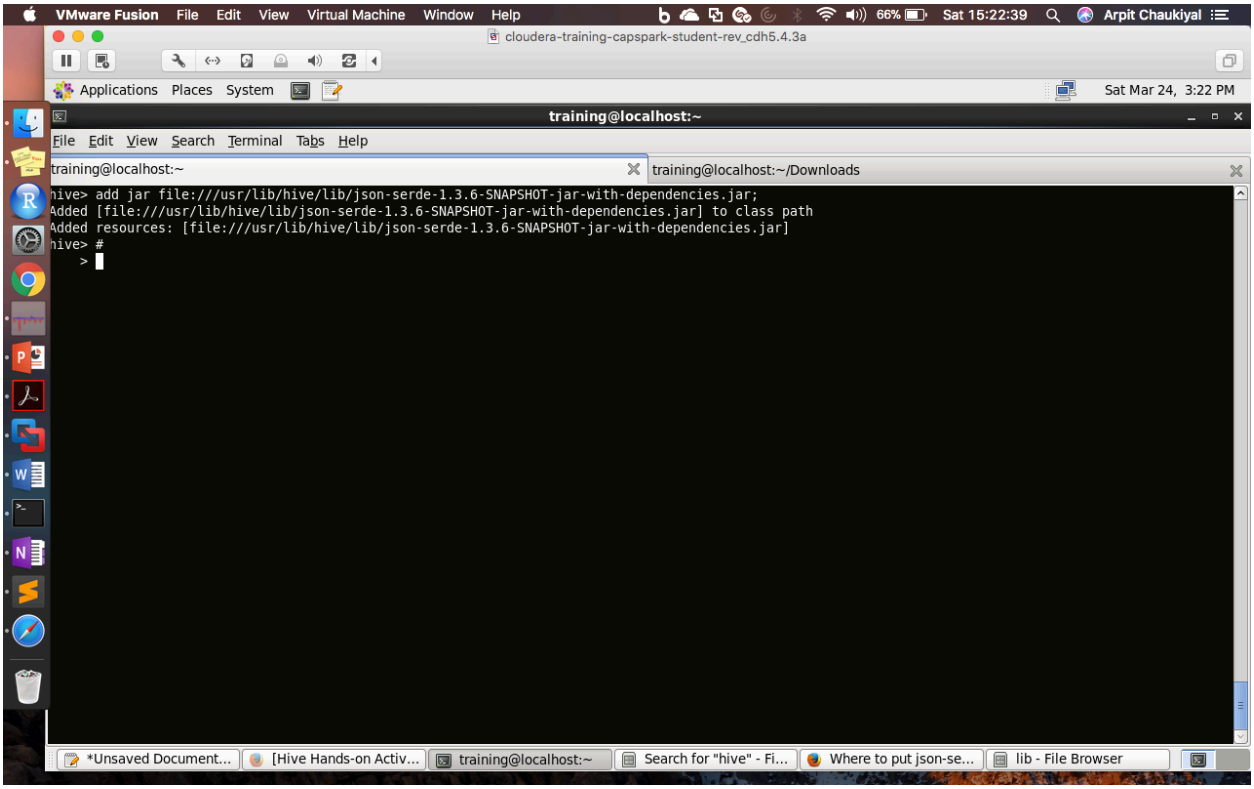

Terminal Shell Edit View Window Help
89% Fri 19:06:59 Arpit Chaukiyal
Homework2 — -bash — 135×39
Arpits-MacBook-Pro:Homework2 arpitchaukiyal$ ls -lrt
total 30592
-rw-r--r--@ 1 arpitchaukiyal staff 22354 Mar 19 11:54 Assignment 2 - Twitter Data Analysis - SP18.docx
-rw-r--r--@ 1 arpitchaukiyal staff 28094 Mar 19 11:56 Dictionary.txt
-rw-r--r--@ 1 arpitchaukiyal staff 15605623 Mar 19 11:56 tweet.json
-rw-r--r--@ 1 arpitchaukiyal staff 162 Mar 21 02:26 ~$signment 2 - Twitter Data Analysis - SP18.docx
Arpits-MacBook-Pro:Homework2 arpitchaukiyal$ cat tweet.json |head -20
{
"contributors": null,
"coordinates": null,
"created_at": "Tue Aug 23 13:53:11 +0000 2016",
"entities": {
"hashtags": [],
"symbols": [],
"urls": [],
"user_mentions": []
},
"favorite_count": 6682,
"favorited": false,
"geo": null,
"id": 768083669550366720,
"id_str": "768083669550366720",
"in_reply_to_screen_name": null,
"in_reply_to_status_id": null,
"in_reply_to_status_id_str": null,
"in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
Arpits-MacBook-Pro:Homework2 arpitchaukiyal$

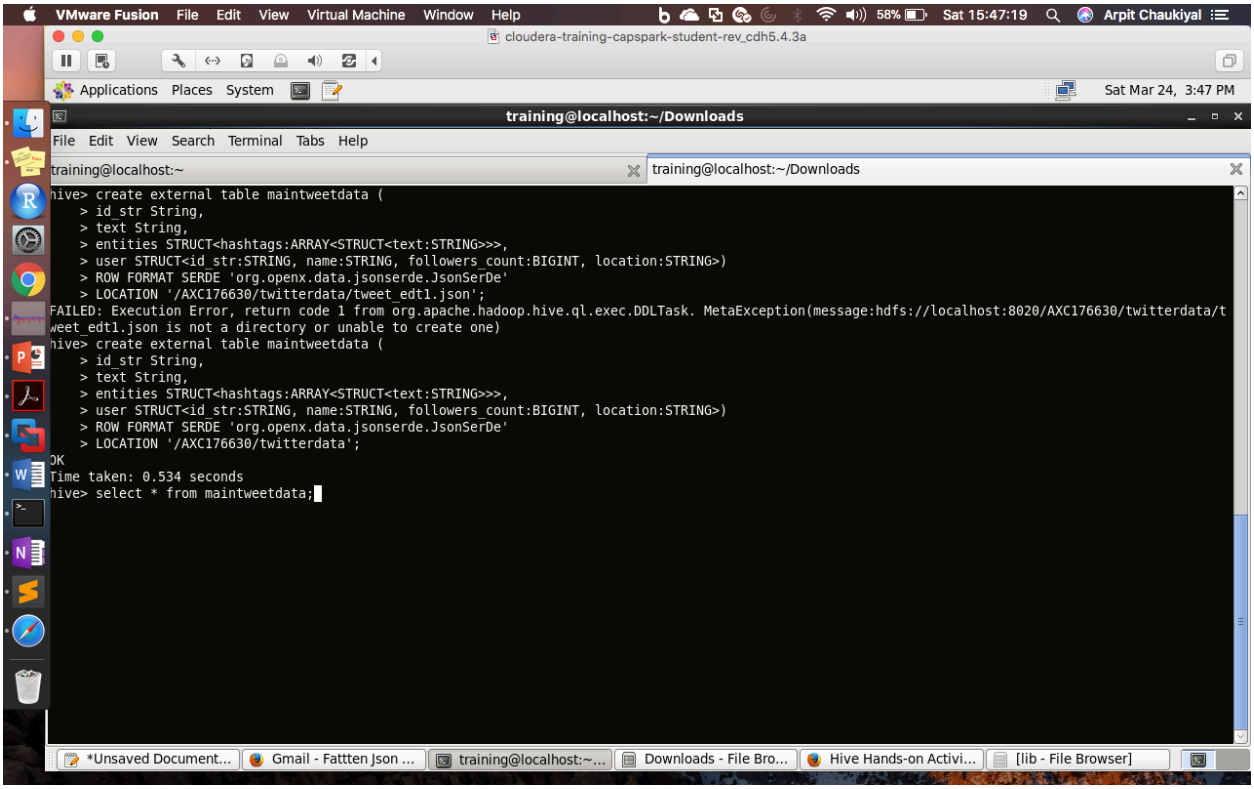

Terminal Shell Edit View Window Help
90% Fri 19 08 50 Arpit Chaukiyal
Homework2 — -bash — 135×39
Arpits-MacBook-Pro:Homework2 arpitchaukiyal$ jq -c '.' tweet.json > tweet_edt1.json
Arpits-MacBook-Pro:Homework2 arpitchaukiyal$ cat tweet_edt1.json | head -2
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 13:53:11 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],
"user_mentions":[]},"favorite_count":6682,"favorited":false,"geo":null,"id":768083669550366720,"id_str":"768083669550366720","in_reply_
to_screen_name":null,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str"
:null,"is_quote_status":false,"lang":"en","place":null,"retweet_count":2301,"retweeted":false,"source":"<a href=\"http://twitter.com/do
wnload/android\" rel=\"nofollow\">Twitter for Android","text":"It is being reported by virtually everyone, and is a fact, that the
media pile on against me is the worst in American political history!","truncated":false,"user":{"contributors_enabled":false,"created_
at":"Wed Mar 18 13:46:38 +0000 2009","default_profile":false,"default_profile_image":false,"description":"#TrumpPence16","entities":{"d
escription":{"urls":[]},"url":{"urls":[{"display_url":"DonaldJTrump.com","expanded_url":"http://www.DonaldJTrump.com","indices":[0,23],
"url":"https://t.co/mZB2hymxC9"}]}},"favourites_count":35,"follow_request_sent":false,"followers_count":11088848,"following":true,"frie
nds_count":42,"geo_enabled":true,"has_extended_profile":false,"id":25073877,"id_str":"25073877","is_translation_enabled":true,"is_trans
lator":false,"lang":"en","listed_count":37760,"location":"New York, NY","name":"Donald J. Trump","notifications":false,"profile_backgro
und_color":"6D5C18","profile_background_image_url":"http://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_c
c.jpg","profile_background_image_url_https":"https://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg"
,"profile_background_tile":true,"profile_banner_url":"https://pbs.twimg.com/profile_banners/25073877/1468988952","profile_image_url":"h
ttp://pbs.twimg.com/profile_images/1980294624/DJT_Headshot_V2_normal.jpg","profile_image_url_https":"https://pbs.twimg.com/profile_imag
es/1980294624/DJT_Headshot_V2_normal.jpg","profile_link_color":"0D5B73","profile_sidebar_border_color":"BDDCAD","profile_sidebar_fill_c
olor":"C5CEC0","profile_text_color":"333333","profile_use_background_image":true,"protected":false,"screen_name":"realDonaldTrump","sta
tuses_count":32980,"time_zone":"Eastern Time (US & Canada)","url":"https://t.co/mZB2hymxC9","utc_offset":-14400,"verified":true}}
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 12:56:46 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],
"user_mentions":[{"id":1367531,"id_str":"1367531","indices":[72,80],"name":"Fox News","screen_name":"FoxNews"}]},"favorite_count":7826,
"favorited":false,"geo":null,"id":768069472464666600,"id_str":"768069472464666624","in_reply_to_screen_name":null,"in_reply_to_status_i
d":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"is_quote_status":false,"lang":"en",
"place":null,"retweet_count":2547,"retweeted":false,"source":"Twitter
for Android","text":"I am now in Texas doing a big fundraiser for the Republican Party and a @FoxNews Special on the BORDER and wit
h victims of border crime!","truncated":false,"user":{"contributors_enabled":false,"created_at":"Wed Mar 18 13:46:38 +0000 2009","defau
lt_profile":false,"default_profile_image":false,"description":"#TrumpPence16","entities":{"description":{"urls":[]},"url":{"urls":[{"di
splay_url":"DonaldJTrump.com","expanded_url":"http://www.DonaldJTrump.com","indices":[0,23],"url":"https://t.co/mZB2hymxC9"}]}},"favour
ites_count":35,"follow_request_sent":false,"followers_count":11088848,"following":true,"friends_count":42,"geo_enabled":true,"has_exten
ded_profile":false,"id":25073877,"id_str":"25073877","is_translation_enabled":true,"is_translator":false,"lang":"en","listed_count":377
60,"location":"New York, NY","name":"Donald J. Trump","notifications":false,"profile_background_color":"6D5C18","profile_background_ima
ge_url":"http://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg","profile_background_image_url_https"
:"https://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg","profile_background_tile":true,"profile_ba
nner_url":"https://pbs.twimg.com/profile_banners/25073877/1468988952","profile_image_url":"http://pbs.twimg.com/profile_images/19802946
24/DJT_Headshot_V2_normal.jpg","profile_image_url_https":"https://pbs.twimg.com/profile_images/1980294624/DJT_Headshot_V2_normal.jpg","
profile_link_color":"0D5B73","profile_sidebar_border_color":"BDDCAD","profile_sidebar_fill_color":"C5CEC0","profile_text_color":"333333
","profile_use_background_image":true,"protected":false,"screen_name":"realDonaldTrump","statuses_count":32980,"time_zone":"Eastern Tim
e (US & Canada)","url":"https://t.co/mZB2hymxC9","utc_offset":-14400,"verified":true}}

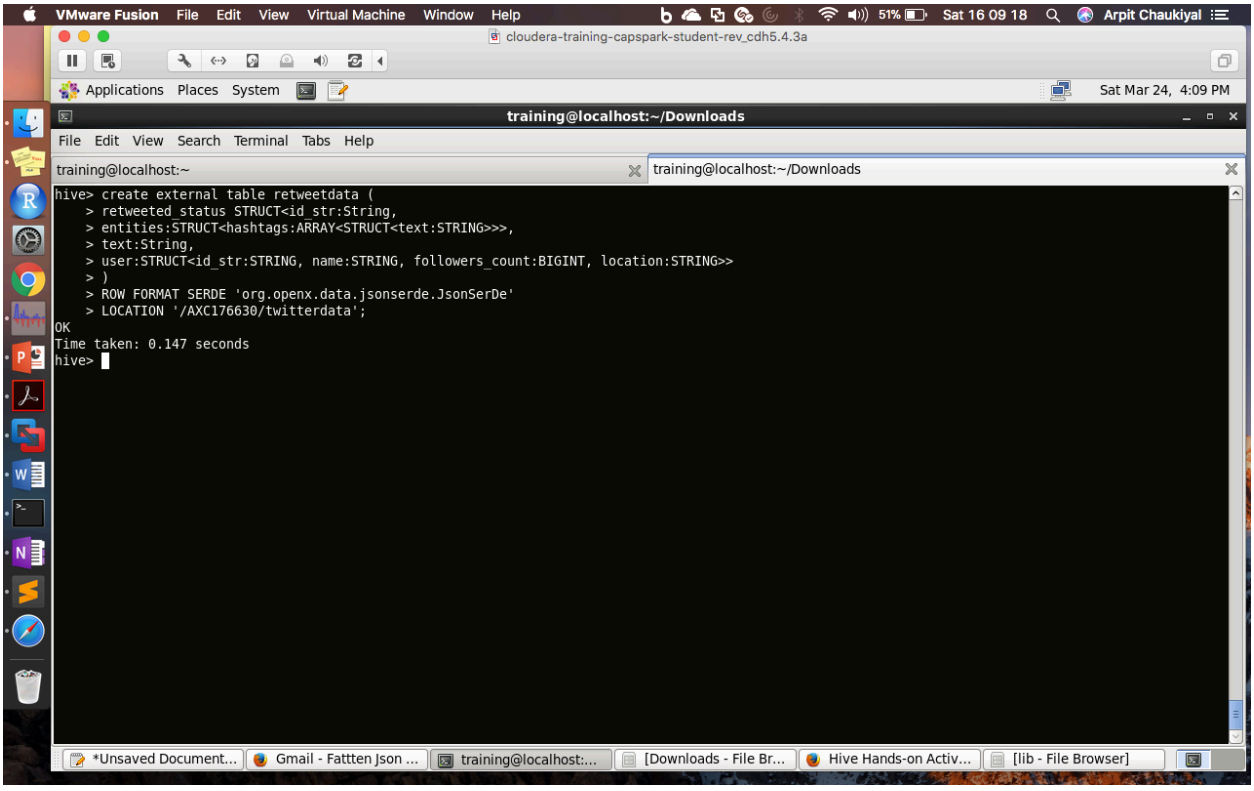

Sublime Text File Edit Selection Find View Goto Tools Project Window Help 90% Fri 19 09 28 Arpit Chaukiyal
tweet.json
UNREGISTERED
tweet.json
tweet_edt1.json
https://acadgild.com/blog/sentiment-analysis-on-tw
{
"contributors": null,
"coordinates": null,
"created_at": "Tue Aug 23 13:53:11 +0000 2016",
"entities": {
"hashtags": [],
"symbols": [],
"urls": [],
"user_mentions": []
},
"favorite_count": 6682,
"favorited": false,
"geo": null,
"id": 768083669550366720,
"id_str": "768083669550366720",
"in_reply_to_screen_name": null,
"in_reply_to_status_id": null,
"in_reply_to_status_id_str": null,
"in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
"is_quote_status": false,
"lang": "en",
"place": null,
"retweet_count": 2301,
"retweeted": false,
"source": "Twitter for Android",
"text": "It is being reported by virtually everyone, and is a fact, that the media pile on against me is the worst in American political hist
"truncated": false,
"user": {
"contributors_enabled": false,
"created_at": "Wed Mar 18 13:46:38 +0000 2009",
"default_profile": false,
"default_profile_image": false,
"description": "#TrumpPence16",
"entities": {
"description": {
"urls": []
},
"url": {
"urls": [
{
"display_url": "DonaldJTrump.com",
"expanded_url": "http://www.DonaldJTrump.com",
"indices": [
0,
23
],
Line 14, Column 30
Spaces: 4
JSON

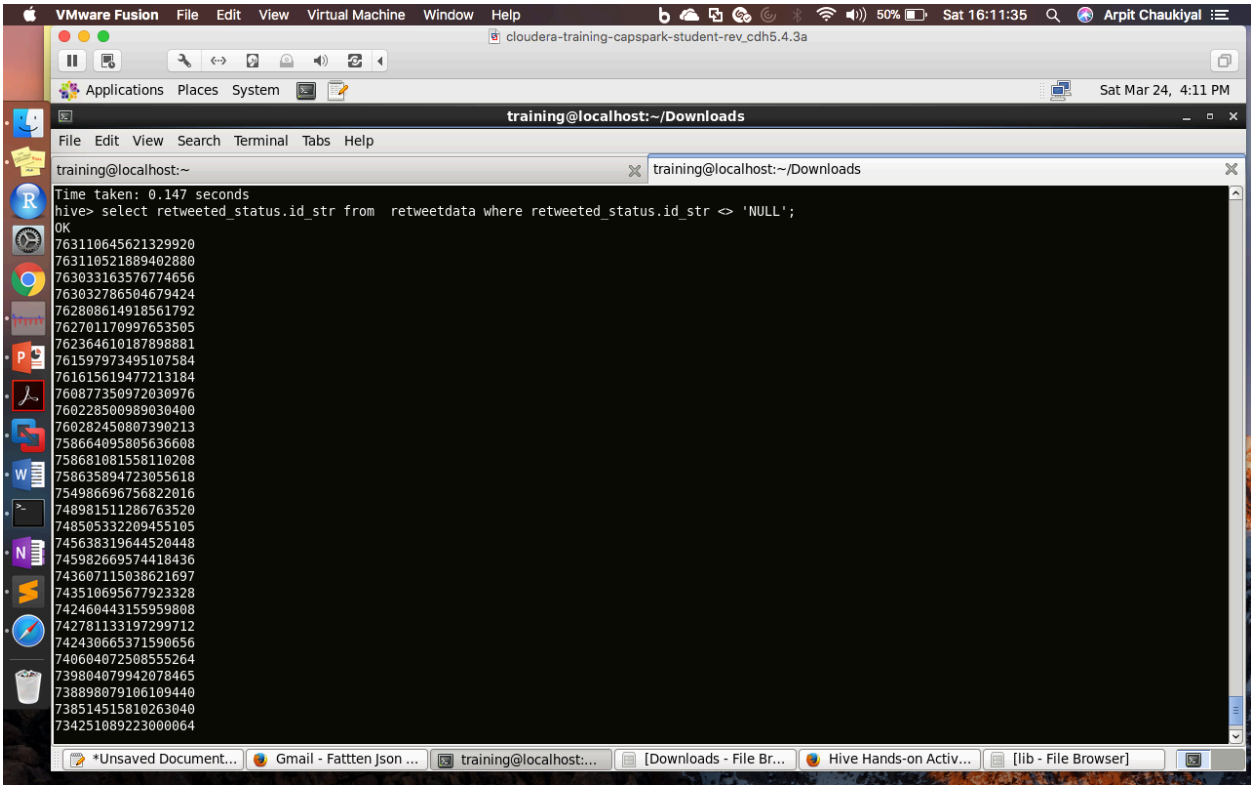

Sublime Text File Edit Selection Find View Goto Tools Project Window Help 90% Fri 19:09:37 Arpit Chaukiyal
tweet_edt1.json
UNREGISTERED
tweet.json
tweet_edt1.json
https://acadgild.com/blog/sentiment-analysis-on-tw
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 13:53:11 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 12:56:46 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 01:02:19 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 00:56:51 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 23:46:42 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 21:06:42 +0000 2016","entities":{"hashtags":[],"media":[{"display_url":"pic.twitte
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 12:55:40 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[{"display_url"
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 12:31:39 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 11:29:13 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 11:21:53 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 11:06:36 +0000 2016","entities":{"hashtags":[],"media":[{"display_url":"pic.twitte
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 01:40:42 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 01:25:00 +0000 2016","entities":{"hashtags":[{"indices":[16,26],"text":"TeamTrump"
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 01:23:16 +0000 2016","entities":{"hashtags":[{"indices":[60,65],"text":"MAGA"},{"i
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 01:19:06 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[{"display_url"
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 01:08:08 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Mon Aug 22 00:33:38 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sun Aug 21 23:35:17 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sun Aug 21 00:04:28 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[{"display_url"
{"contributors":null,"coordinates":null,"created_at":"Sun Aug 21 00:02:41 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sun Aug 21 00:02:05 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 23:11:51 +0000 2016","entities":{"hashtags":[{"indices":[111,116],"text":"MAGA"}],
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 23:04:02 +0000 2016","entities":{"hashtags":[{"indices":[70,83],"text":"TrumpPence
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 23:02:37 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 22:57:23 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 17:35:11 +0000 2016","entities":{"hashtags":[{"indices":[46,59],"text":"TrumpPence
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 05:02:57 +0000 2016","entities":{"hashtags":[{"indices":[121,126],"text":"MAGA"},{
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 01:14:08 +0000 2016","entities":{"hashtags":[],"media":[{"display_url":"pic.twitte
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 01:00:12 +0000 2016","entities":{"hashtags":[{"indices":[97,116],"text":"StandWith
{"contributors":null,"coordinates":null,"created_at":"Sat Aug 20 00:17:09 +0000 2016","entities":{"hashtags":[{"indices":[0,14],"text":"WheresHilla
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 22:16:16 +0000 2016","entities":{"hashtags":[{"indices":[91,104],"text":"TrumpPenc
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 13:34:54 +0000 2016","entities":{"hashtags":[{"indices":[67,77],"text":"ImWithYou"
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 13:27:10 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 12:43:37 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 12:42:38 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 10:40:15 +0000 2016","entities":{"hashtags":[],"media":[{"display_url":"pic.twitte
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 00:52:35 +0000 2016","entities":{"hashtags":[{"indices":[37,59],"text":"MakeAmeric
{"contributors":null,"coordinates":null,"created_at":"Fri Aug 19 00:33:54 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[{"display_url"
{"contributors":null,"coordinates":null,"created_at":"Thu Aug 18 20:55:33 +0000 2016","entities":{"hashtags":[{"indices":[0,15],"text":"CrookedHill
{"contributors":null,"coordinates":null,"created_at":"Thu Aug 18 12:11:48 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Thu Aug 18 01:21:30 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
{"contributors":null,"coordinates":null,"created_at":"Wed Aug 17 22:25:02 +0000 2016","entities":{"hashtags":[],"media":[{"display_url":"pic.twitte
{"contributors":null,"coordinates":null,"created_at":"Wed Aug 17 21:44:48 +0000 2016","entities":{"hashtags":[{"indices":[65,87],"text":"MakeAmeric
{"contributors":null,"coordinates":null,"created_at":"Wed Aug 17 20:05:08 +0000 2016","entities":{"hashtags":[{"indices":[29,39],"text":"Obamacare"
{"contributors":null,"coordinates":null,"created_at":"Wed Aug 17 16:57:58 +0000 2016","entities":{"hashtags":[{"indices":[110,120],"text":"ImWithYo
{"contributors":null,"coordinates":null,"created_at":"Wed Aug 17 16:46:57 +0000 2016","entities":{"hashtags":[{"indices":[0,21],"text":"MakeAmerica
Line 1, Column 1
Tab Size: 4
JSON

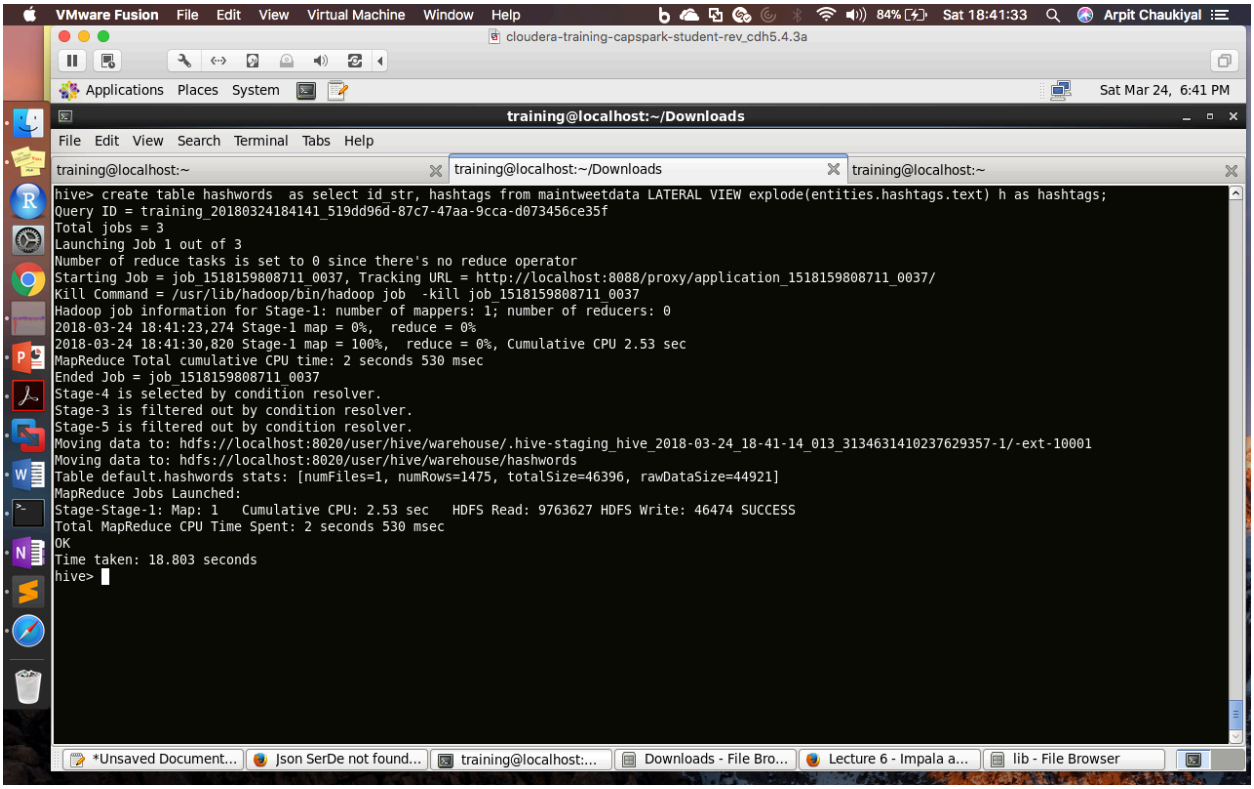

VMware Fusion File Edit View Virtual Machine Window Help
92% Fri 19 14 38 Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Fri Mar 23, 7:14 PM
training@localhost:~
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
[training@localhost ~]$ hdfs dfs -mkdir /AXC176630/twitterdata
[training@localhost ~]$ hdfs dfs -put tweet_edt1.json /AXC176630/twitterdata
[training@localhost ~]$ hdfs dfs -ls /AXC176630/twitterdata
Found 1 items
-rw-rw-rw- 1 training supergroup 9758732 2018-03-23 19:11 /AXC176630/twitterdata/tweet_edt1.json
[training@localhost ~]$ hdfs dfs -put Dictionary.txt /AXC176630/twitterdata
[training@localhost ~]$ hdfs dfs -ls /AXC176630/twitterdata
Found 2 items
-rw-rw-rw- 1 training supergroup 28094 2018-03-23 19:12 /AXC176630/twitterdata/Dictionary.txt
-rw-rw-rw- 1 training supergroup 9758732 2018-03-23 19:11 /AXC176630/twitterdata/tweet_edt1.json
[training@localhost ~]$ hdfs dfs -cat /AXC176630/twitterdata/tweet_edt1.json| head -1
{"contributors":null,"coordinates":null,"created_at":"Tue Aug 23 13:53:11 +0000 2016","entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentions":[]},"favorite_count":6682,"favorited":false,"geo":null,"id":768083669550366700,"id_str":"768083669550366720","in_reply_to_screen_name":null,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"is_quote_status":false,"lang":"en","place":null,"retweet_count":2301,"retweeted":false,"source":"Twitter for Android","text":"It is being reported by virtually everyone, and is a fact, that the media pile on against me is the worst in American political history!","truncated":false,"user":{"contributors_enabled":false,"created_at":"Wed Mar 18 13:46:38 +0000 2009","default_profile":false,"default_profile_image":false,"description":"#TrumpPence16","entities":{"description":{"urls":[]},"url":{"urls":[{"display_url":"DonaldJTrump.com","expanded_url":"http://www.DonaldJTrump.com","indices":[0,23],"url":"https://t.co/mZB2hymxC9"}]}},"favourites_count":35,"follow_request_sent":false,"followers_count":11088848,"following":true,"friends_count":42,"geo_enabled":true,"has_extended_profile":false,"id":25073877,"id_str":"25073877","is_translation_enabled":true,"is_translator":false,"lang":"en","listed_count":37760,"location":"New York, NY","name":"Donald J. Trump","notifications":false,"profile_background_color":"6D5C18","profile_background_image_url":"http://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg","profile_background_image_url_https":"https://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg","profile_background_tile":true,"profile_banner_url":"https://pbs.twimg.com/profile_banners/25073877/1468988952","profile_image_url":"http://pbs.twimg.com/profile_images/1980294624/DJT_Headshot_V2_normal.jpg","profile_image_url_https":"https://pbs.twimg.com/profile_images/1980294624/DJT_Headshot_V2_normal.jpg","profile_link_color":"0D5B73","profile_sidebar_border_color":"BDDCAD","profile_sidebar_fill_color":"C5CEC0","profile_text_color":"333333","profile_use_background_image":true,"protected":false,"screen_name":"realDonaldTrump","statuses_count":32980,"time_zone":"Eastern Time (US & Canada)","url":"https://t.co/mZB2hymxC9","utc_offset":-14400,"verified":true}}
cat: Unable to write to output stream.
[training@localhost ~]$
*Unsaved Document 2...
[Assignments - BUAN ...
training@localhost:~

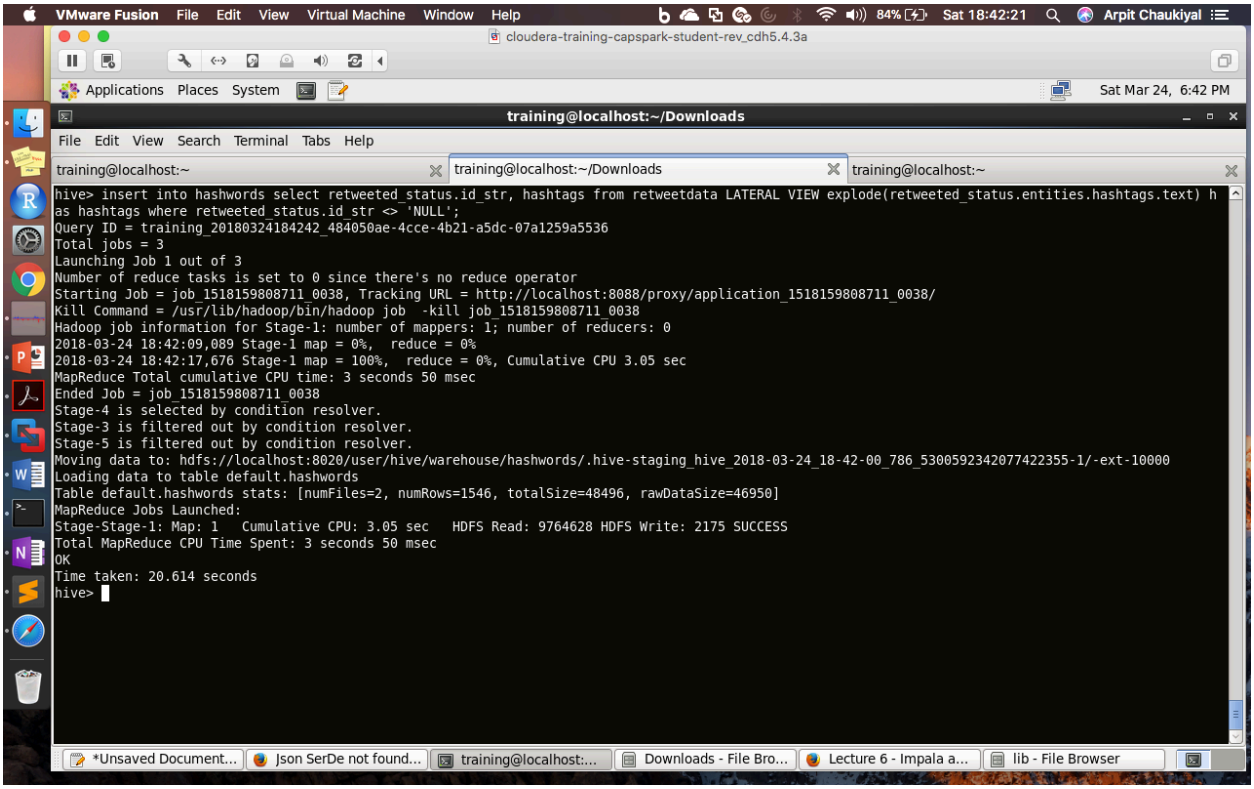

VMware Fusion File Edit View Virtual Machine Window Help
66% Sat 15:22:39 Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 3:22 PM
training@localhost:~
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
hive> add jar file:///usr/lib/hive/lib/json-serde-1.3.6-SNAPSHOT-jar-with-dependencies.jar;
Added [file:///usr/lib/hive/lib/json-serde-1.3.6-SNAPSHOT-jar-with-dependencies.jar] to class path
Added resources: [file:///usr/lib/hive/lib/json-serde-1.3.6-SNAPSHOT-jar-with-dependencies.jar]
hive> #
>
*Unsaved Document...
[Hive Hands-on Activ...
training@localhost:~
Search for "hive" - Fi...
Where to put json-se...
lib - File Browser

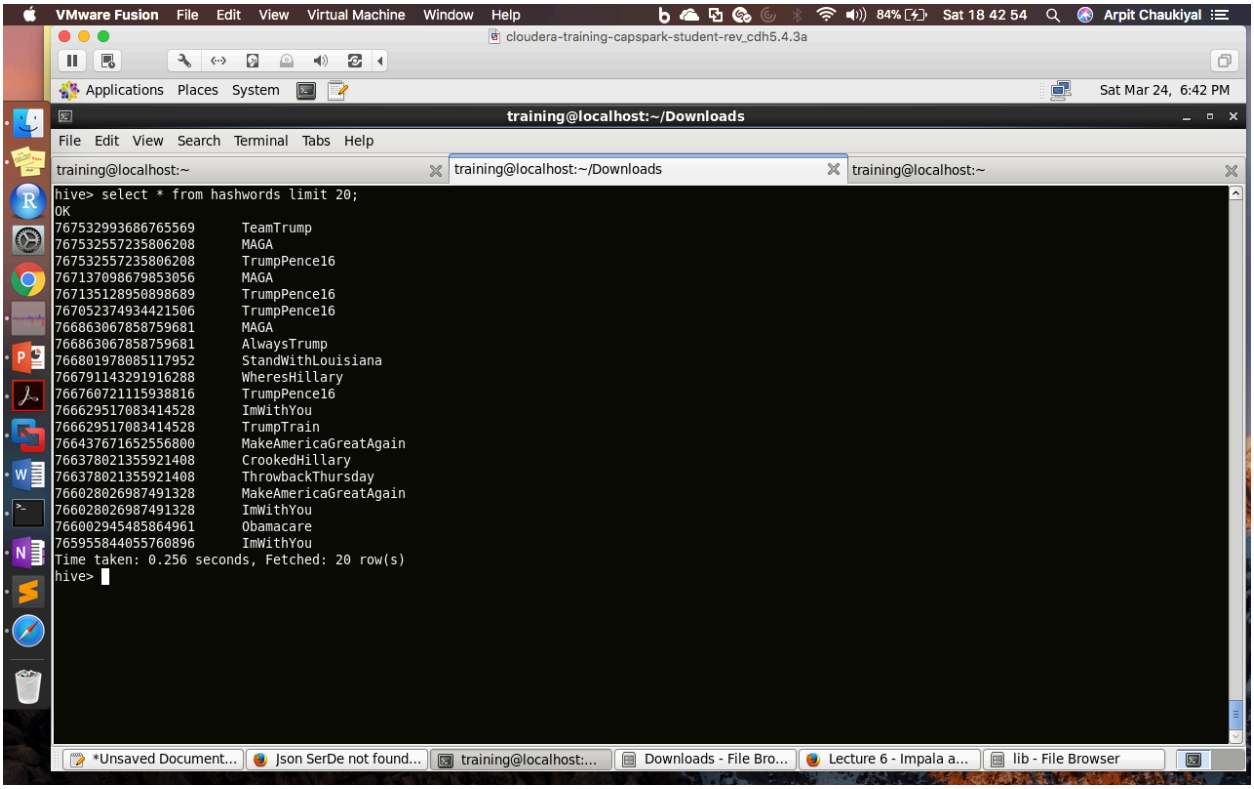

VMware Fusion File Edit View Virtual Machine Window Help
Sat 15:47:19
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 3:47 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
hive> create external table maintweetdata (
> id_str String,
> text String,
> entities STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>,
> user STRUCT<id_str:STRING, name:STRING, followers_count:BIGINT, location:STRING>)
> ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
> LOCATION '/AXC176630/twitterdata/tweet_edt1.json';
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:hdfs://localhost:8020/AXC176630/twitterdata/tweet_edt1.json is not a directory or unable to create one)
hive> create external table maintweetdata (
> id_str String,
> text String,
> entities STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>,
> user STRUCT<id_str:STRING, name:STRING, followers_count:BIGINT, location:STRING>)
> ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
> LOCATION '/AXC176630/twitterdata';
OK
Time taken: 0.534 seconds
hive> select * from maintweetdata;
*Unsaved Document...
Gmail - Fattten Json ...
training@localhost:~...
Downloads - File Bro...
Hive Hands-on Activi...
[lib - File Browser]

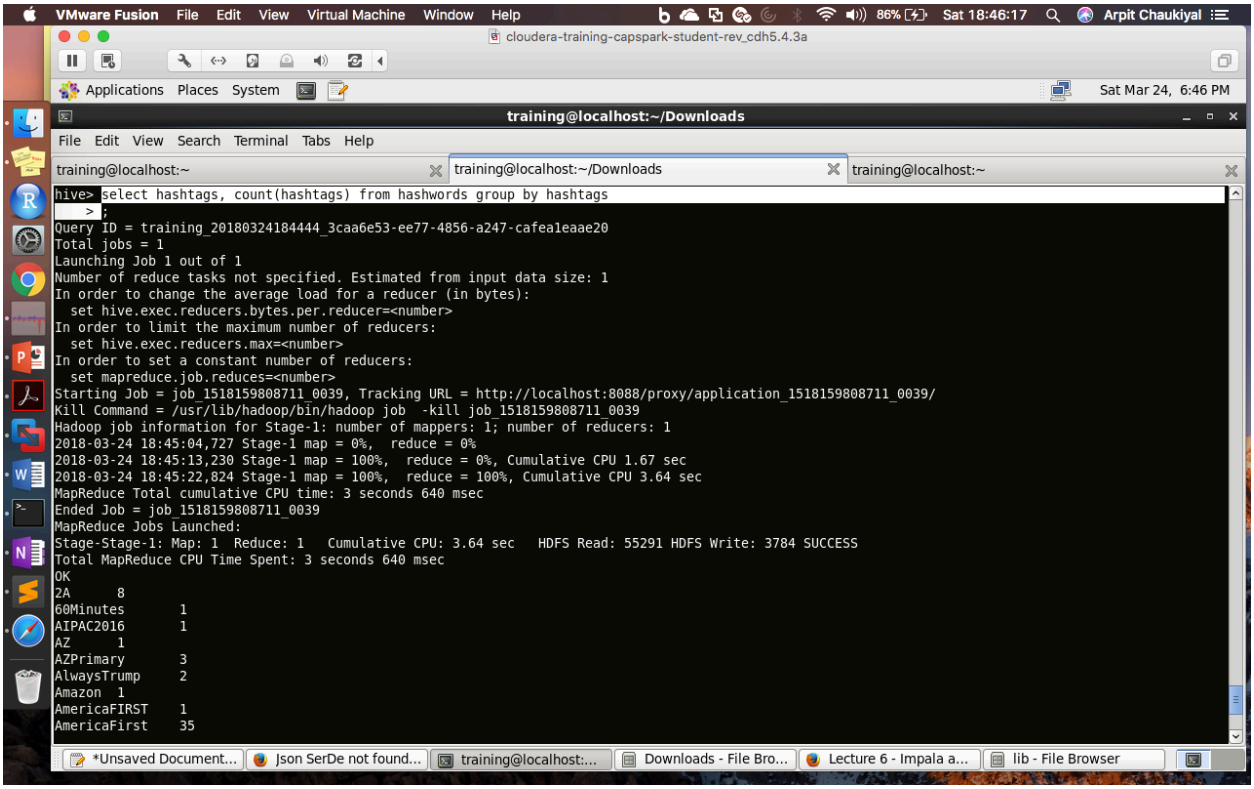

VMware Fusion File Edit View Virtual Machine Window Help
Sat 16 09 18
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 4:09 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
hive> create external table retweetdata (
> retweeted_status STRUCT<id_str:String,
> entities:STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>,
> text:String,
> user:STRUCT<id_str:STRING, name:STRING, followers_count:BIGINT, location:STRING>>
>)
> ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
> LOCATION '/AXC176630/twitterdata';
OK
Time taken: 0.147 seconds
hive>
*Unsaved Document...
Gmail - Fattten Json ...
training@localhost:...
[Downloads - File Br...
Hive Hands-on Activ...
[lib - File Browser]

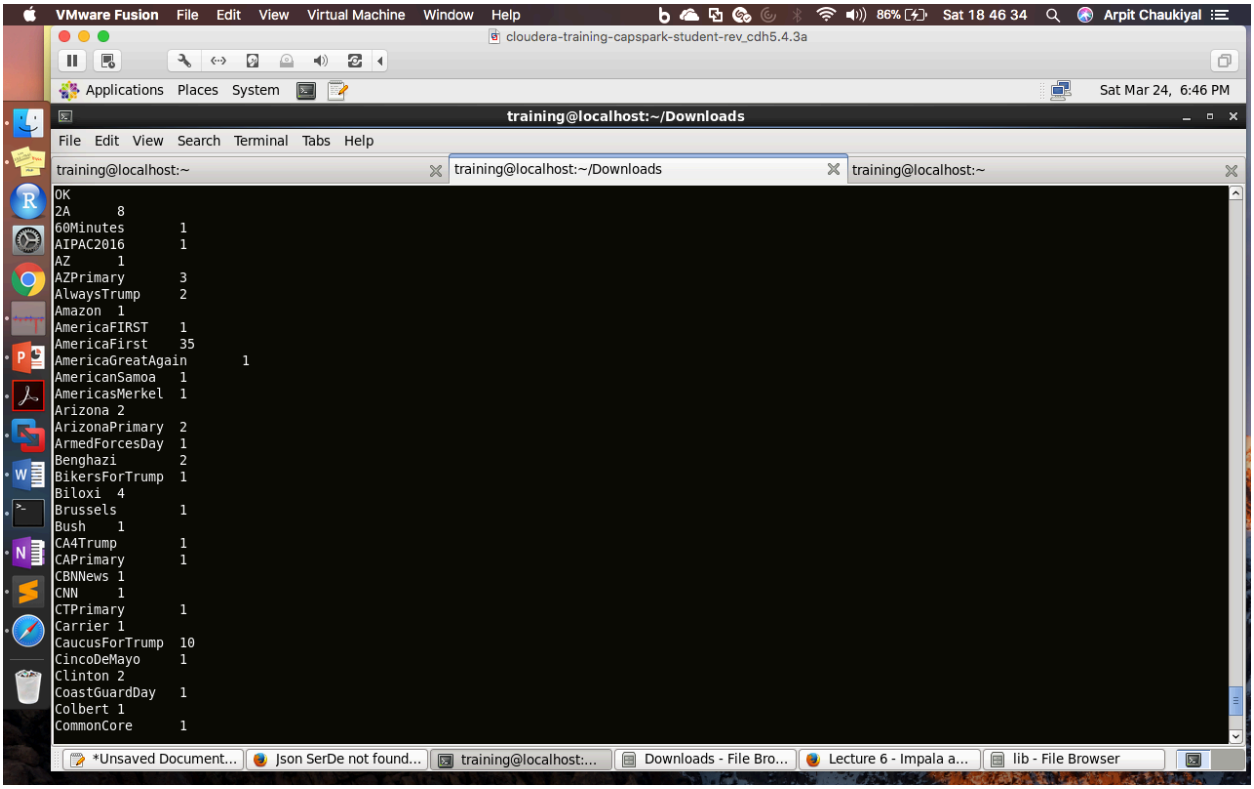

VMware Fusion File Edit View Virtual Machine Window Help
Sat 16:11:35
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 4:11 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
Time taken: 0.147 seconds
hive> select retweeted_status.id_str from retweetdata where retweeted_status.id_str <> 'NULL';
OK
763110645621329920
763110521889402880
763033163576774656
763032786504679424
762808614918561792
762701170997653505
762364610187898881
761597973495107584
761615619477213184
760877350972030976
760228500989030400
760282450807390213
758664095805636608
758681081558110208
758635894723055618
754986696756822016
748981511286763520
748505332209455105
745638319644520448
745982669574418436
743607115038621697
743510695677923328
742460443155959808
742781133197299712
742430665371590656
740604072508555264
739804079942078465
738898079106109440
738514515810263040
73425108922300064
*Unsaved Document...
Gmail - Fattten Json ...
training@localhost:...
[Downloads - File Br...
Hive Hands-on Activ...
[lib - File Browser]


VMware Fusion File Edit View Virtual Machine Window Help
Sat 18:41:33
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 6:41 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> create table hashwords as select id_str, hashtags from maintweetdata LATERAL VIEW explode(entities.hashtags.text) h as hashtags;
Query ID = training_20180324184141_519dd96d-87c7-47aa-9cca-d073456ce35f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0037, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0037/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0037
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-24 18:41:23,274 Stage-1 map = 0%, reduce = 0%
2018-03-24 18:41:30,820 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.53 sec
MapReduce Total cumulative CPU time: 2 seconds 530 msec
Ended Job = job_1518159808711_0037
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/.hive-staging_hive_2018-03-24_18-41-14_013_3134631410237629357-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/hashwords
Table default.hashwords stats: [numFiles=1, numRows=1475, totalSize=46396, rawDataSize=44921]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 2.53 sec HDFS Read: 9763627 HDFS Write: 46474 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 530 msec
OK
Time taken: 18.803 seconds
hive>
*Unsaved Document...
Json SerDe not found...
training@localhost:...
Downloads - File Bro...
Lecture 6 - Impala a...
lib - File Browser

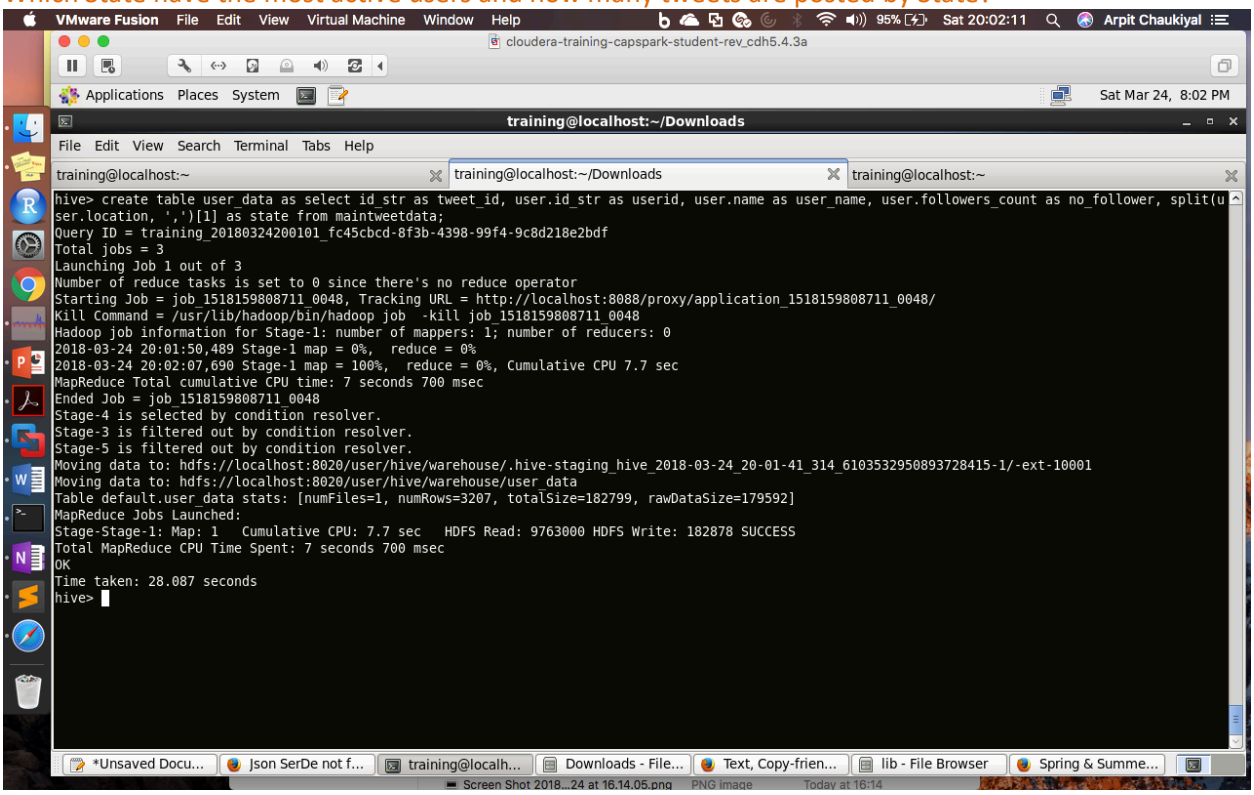

VMware Fusion File Edit View Virtual Machine Window Help
Sat 18:42:21
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 6:42 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> insert into hashwords select retweeted_status.id_str, hashtags from retweetdata LATERAL VIEW explode(retweeted_status.entities.hashtags.text) h
as hashtags where retweeted_status.id_str <> 'NULL';
Query ID = training_20180324184242_484050ae-4cce-4b21-a5dc-07a1259a5536
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0038, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0038/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0038
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-24 18:42:09,089 Stage-1 map = 0%, reduce = 0%
2018-03-24 18:42:17,676 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.05 sec
MapReduce Total cumulative CPU time: 3 seconds 50 msec
Ended Job = job_1518159808711_0038
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/hashwords/.hive-staging_hive_2018-03-24_18-42-00_786_5300592342077422355-1/-ext-10000
Loading data to table default.hashwords
Table default.hashwords stats: [numFiles=2, numRows=1546, totalSize=48496, rawDataSize=46950]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.05 sec HDFS Read: 9764628 HDFS Write: 2175 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 50 msec
OK
Time taken: 20.614 seconds
hive>
*Unsaved Document...
Json SerDe not found...
training@localhost:...
Downloads - File Bro...
Lecture 6 - Impala a...
lib - File Browser

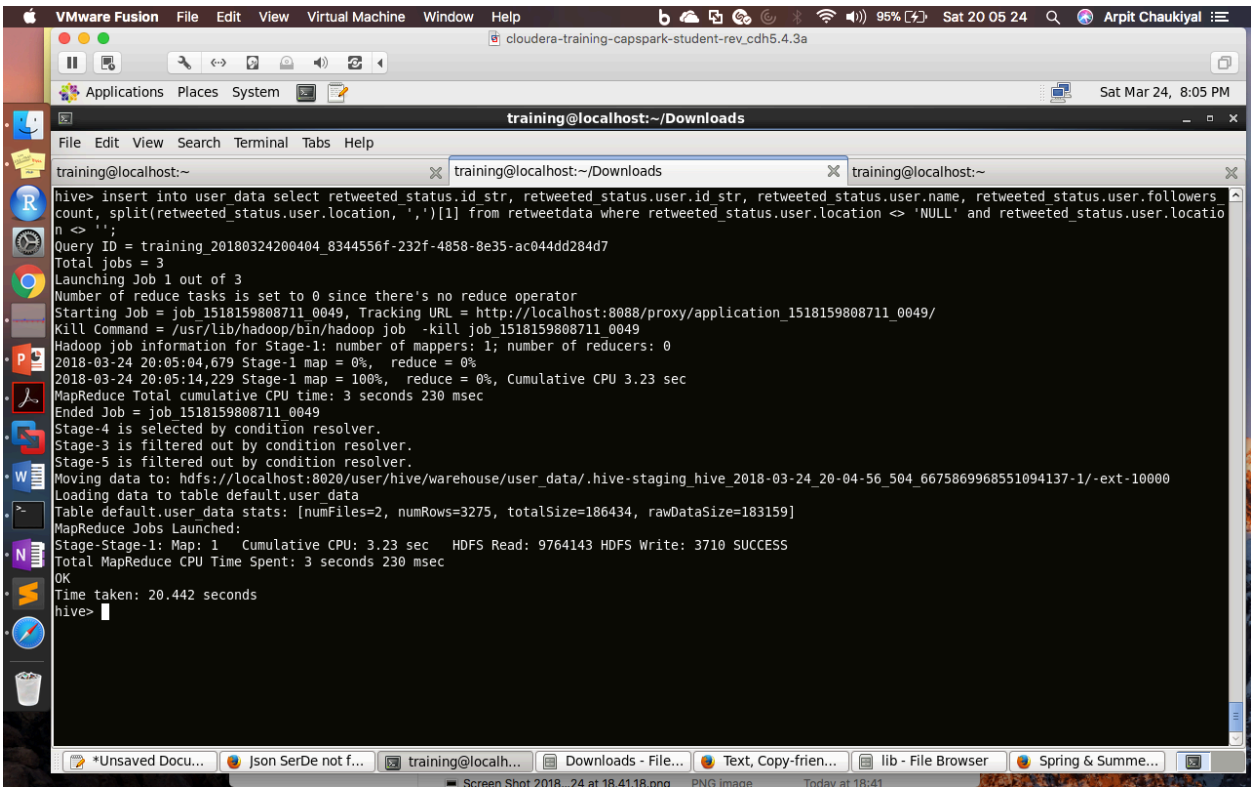

VMware Fusion File Edit View Virtual Machine Window Help
Sat 18 42 54
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 6:42 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> select * from hashwords limit 20;
OK
767532993686765569 TeamTrump
767532557235806208 MAGA
767532557235806208 TrumpPence16
767137098679853056 MAGA
767135128950898689 TrumpPence16
767052374934421506 TrumpPence16
766863067858759681 MAGA
766863067858759681 AlwaysTrump
766801978085117952 StandWithLouisiana
766791143291916288 WheresHillary
766760721115938816 TrumpPence16
766629517083414528 ImWithYou
766629517083414528 TrumpTrain
766437671652556800 MakeAmericaGreatAgain
766378021355921408 CrookedHillary
766378021355921408 ThrowbackThursday
766028026987491328 MakeAmericaGreatAgain
766028026987491328 ImWithYou
766002945485864961 Obamacare
765955844055760896 ImWithYou
Time taken: 0.256 seconds, Fetched: 20 row(s)
hive>
*Unsaved Document...
Json SerDe not found...
training@localhost:...
Downloads - File Bro...
Lecture 6 - Impala a...
lib - File Browser

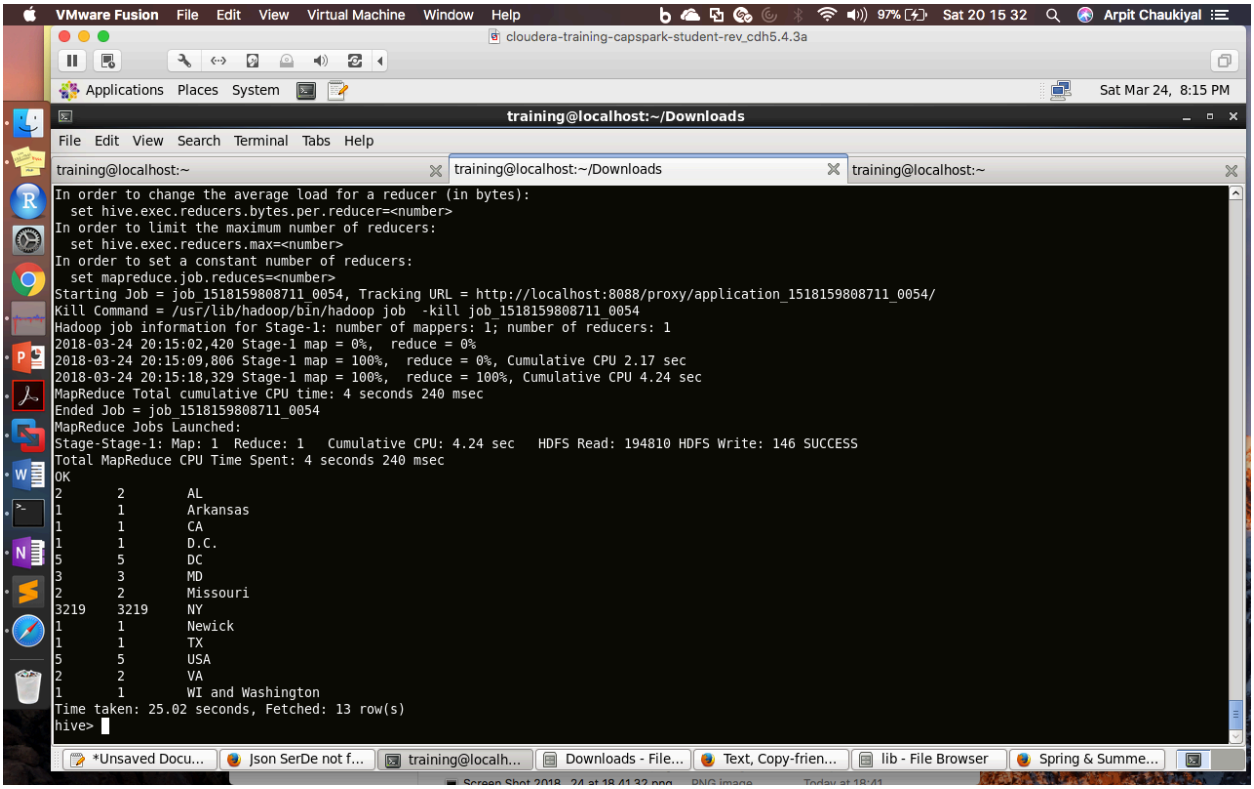

VMware Fusion File Edit View Virtual Machine Window Help
Sat 18:46:17
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 6:46 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> select hashtags, count(hashtags) from hashwords group by hashtags
> ;
Query ID = training_20180324184444_3caa6e53-ee77-4856-a247-cafea1eaae20
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1518159808711_0039, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0039/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0039
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-24 18:45:04,727 Stage-1 map = 0%, reduce = 0%
2018-03-24 18:45:13,230 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.67 sec
2018-03-24 18:45:22,824 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.64 sec
MapReduce Total cumulative CPU time: 3 seconds 640 msec
Ended Job = job_1518159808711_0039
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.64 sec HDFS Read: 55291 HDFS Write: 3784 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 640 msec
OK
2A 8
60Minutes 1
AIPAC2016 1
AZ 1
AZPrimary 3
AlwaysTrump 2
Amazon 1
AmericaFIRST 1
AmericaFirst 35
*Unsaved Document...
Json SerDe not found...
training@localhost:...
Downloads - File Bro...
Lecture 6 - Impala a...
lib - File Browser


VMware Fusion File Edit View Virtual Machine Window Help
Sat 18 46 34
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 6:46 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
OK
2A 8
60Minutes 1
AIPAC2016 1
AZ 1
AZPrimary 3
AlwaysTrump 2
Amazon 1
AmericaFIRST 1
AmericaFirst 35
AmericaGreatAgain 1
AmericanSamoa 1
AmericasMerkel 1
Arizona 2
ArizonaPrimary 2
ArmedForcesDay 1
Benghazi 2
BikersForTrump 1
Biloxi 4
Brussels 1
Bush 1
CA4Trump 1
CAPrimary 1
CBNNews 1
CNN 1
CTPrimary 1
Carrier 1
CaucusForTrump 10
CincoDeMayo 1
Clinton 2
CoastGuardDay 1
Colbert 1
CommonCore 1
*Unsaved Document...
Json SerDe not found...
training@localhost:...
Downloads - File Bro...
Lecture 6 - Impala a...
lib - File Browser

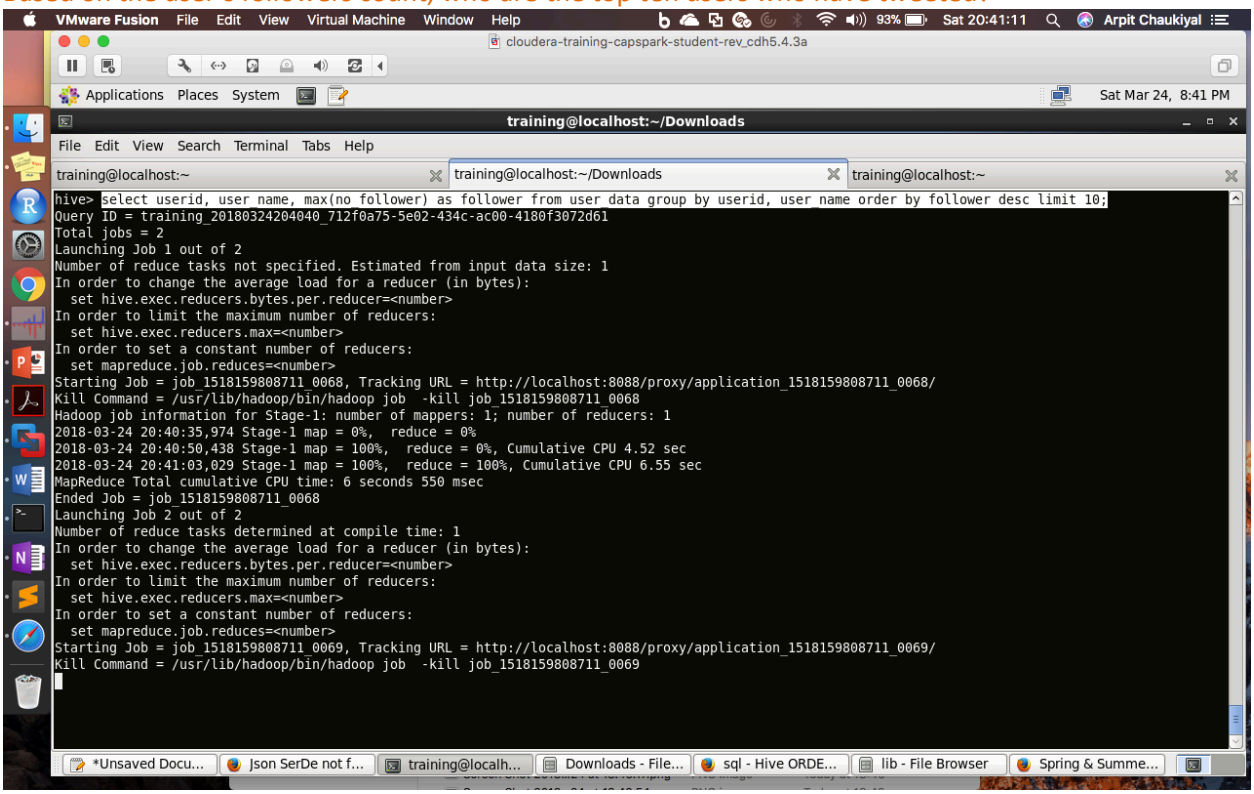
b. Which State have the most active users and how many tweets are posted by State?

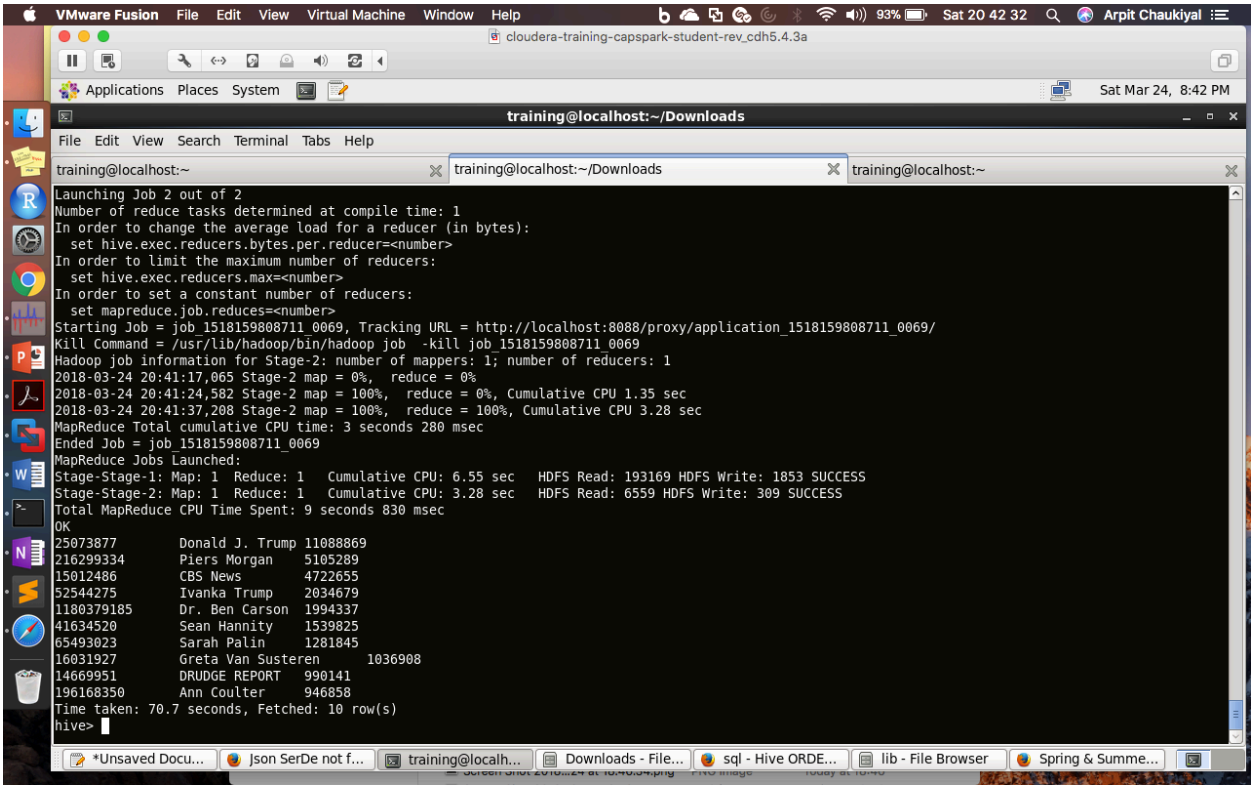




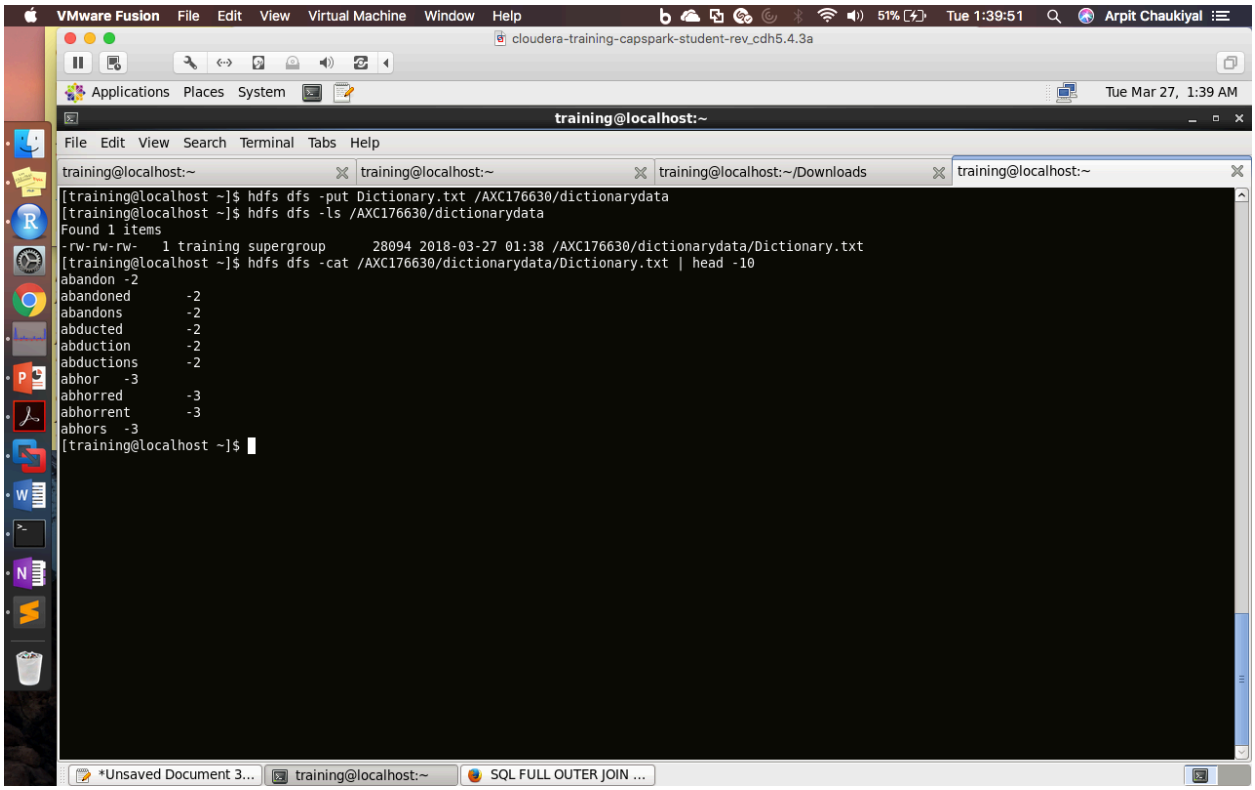


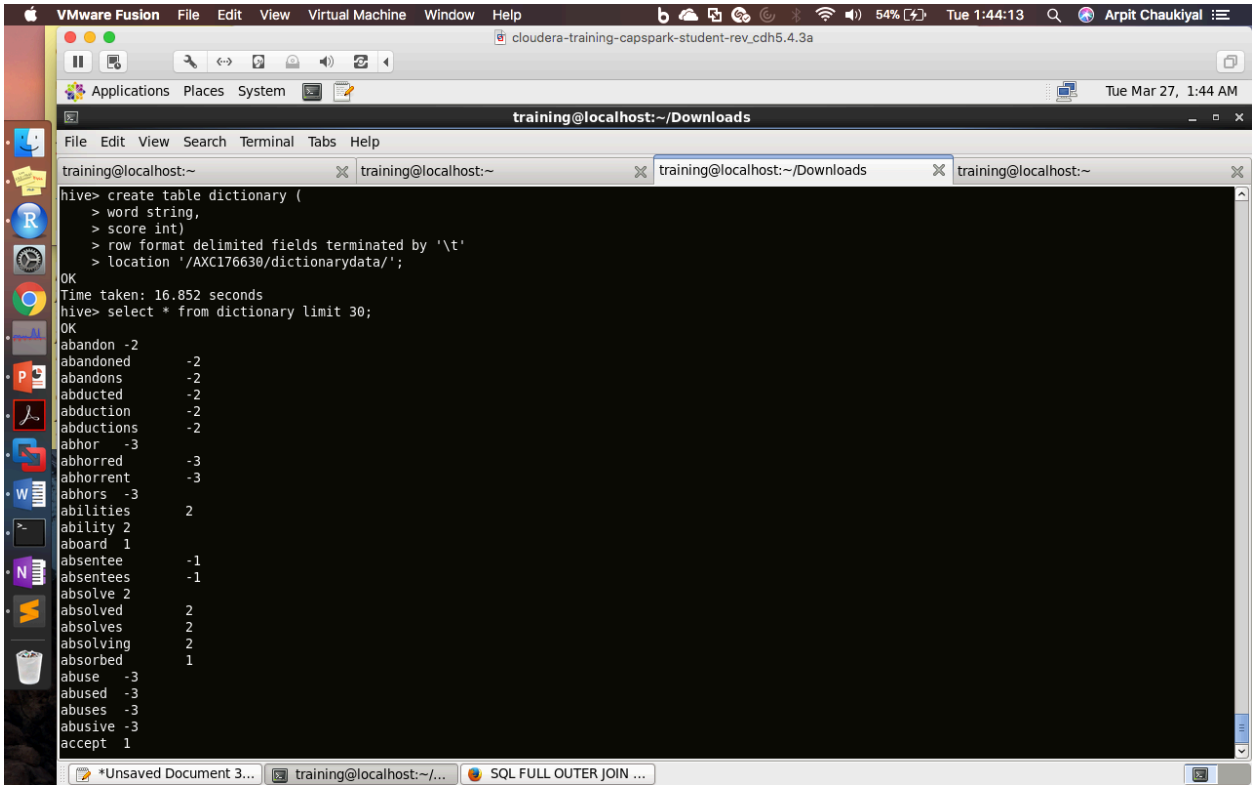
c. Based on the user's followers count, who are the top ten users who have tweeted?

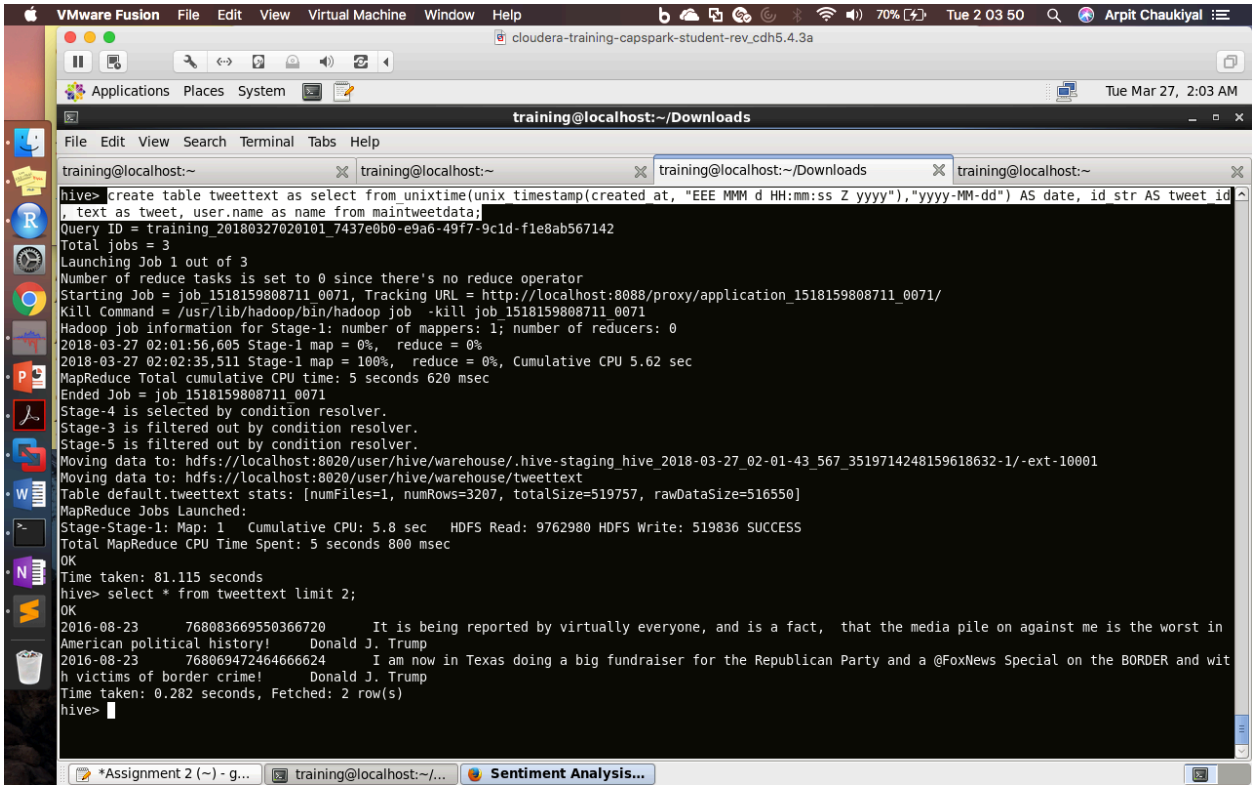


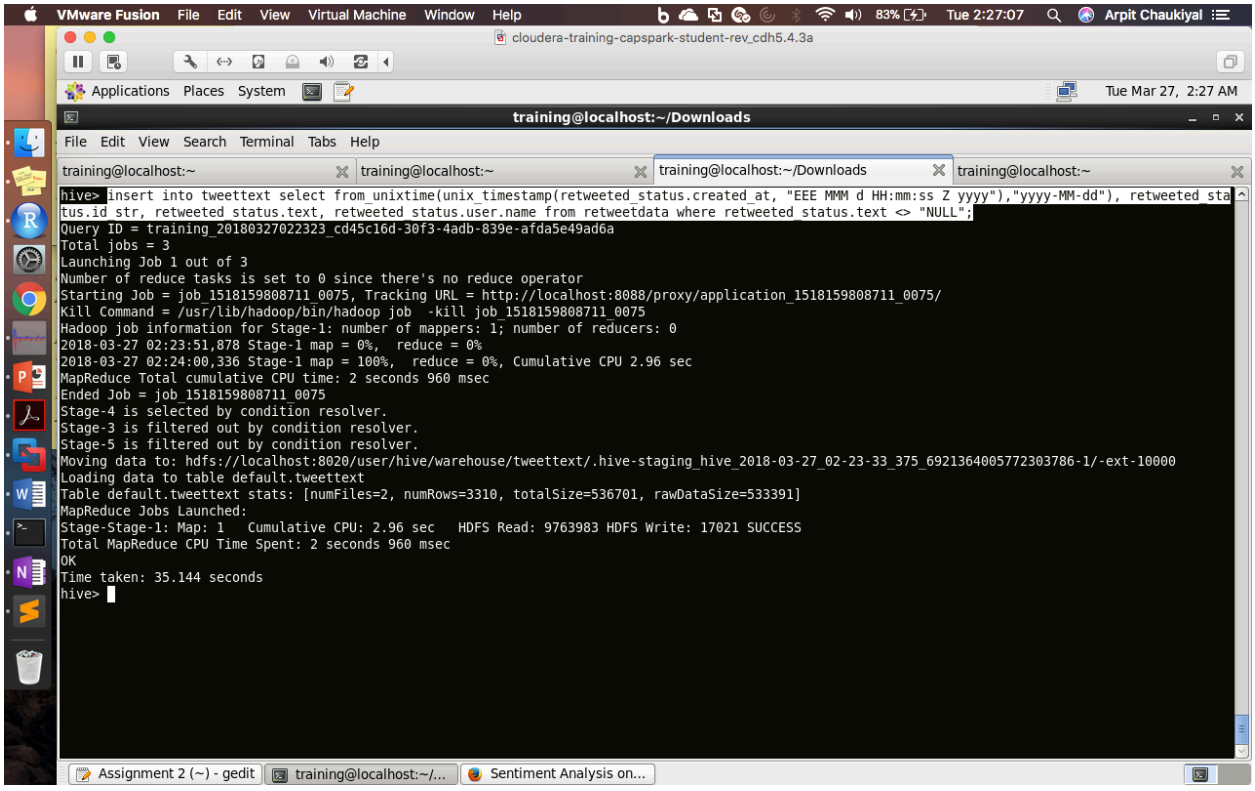

VMware Fusion File Edit View Virtual Machine Window Help
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Sat Mar 24, 8:42 PM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1518159808711_0069, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0069/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0069
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-24 20:41:17,065 Stage-2 map = 0%, reduce = 0%
2018-03-24 20:41:24,582 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.35 sec
2018-03-24 20:41:37,208 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.28 sec
MapReduce Total cumulative CPU time: 3 seconds 280 msec
Ended Job = job_1518159808711_0069
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.55 sec HDFS Read: 193169 HDFS Write: 1853 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.28 sec HDFS Read: 6559 HDFS Write: 309 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 830 msec
OK
25073877 Donald J. Trump 11088869
216299334 Piers Morgan 5105289
15012486 CBS News 4722655
52544275 Ivanka Trump 2034679
1180379185 Dr. Ben Carson 1994337
41634520 Sean Hannity 1539825
65493023 Sarah Palin 1281845
16031927 Greta Van Susteren 1036908
14669951 DRUDGE REPORT 990141
196168350 Ann Coulter 946858
Time taken: 70.7 seconds, Fetched: 10 row(s)
hive>
*Unsaved Docu...
Json SerDe not f...
training@localh...
Downloads - File...
sql - Hive ORDE...
lib - File Browser
Spring & Summe...

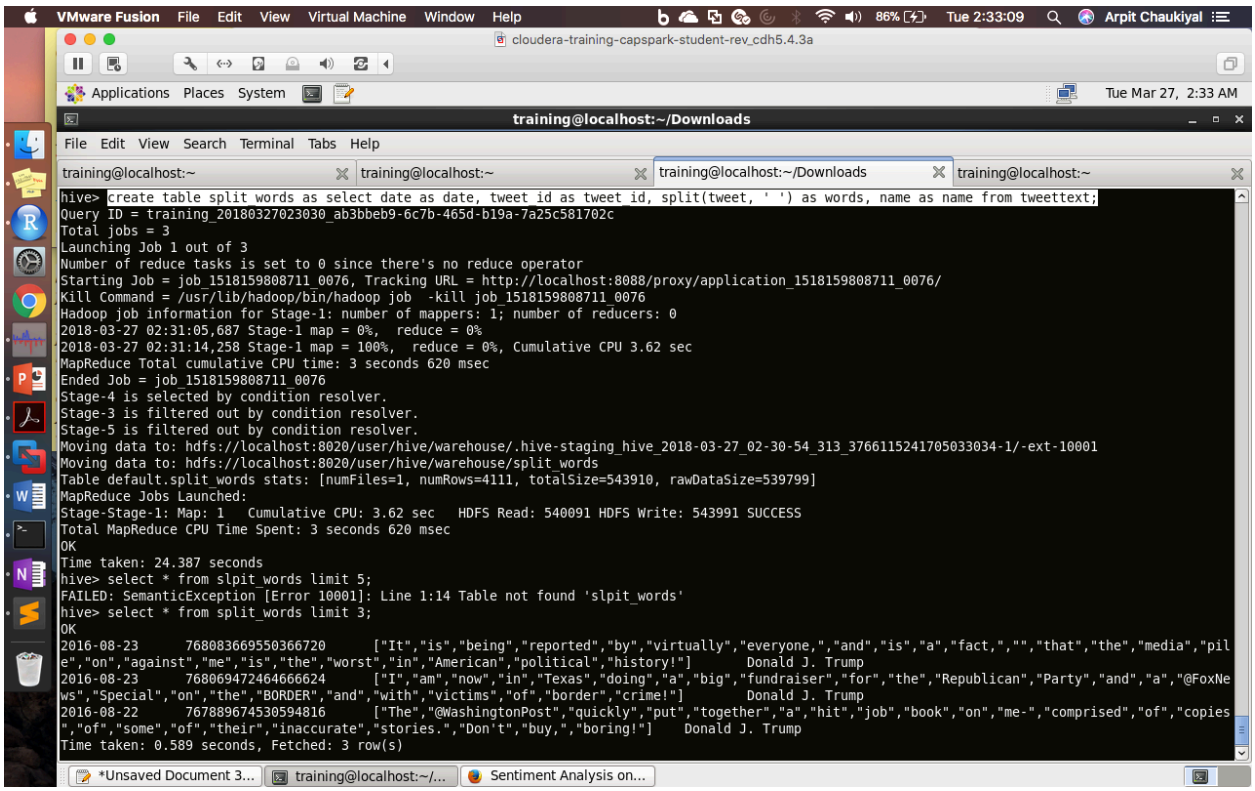
d. What is the polarity score for each tweet that was posted? Does the tweet have a positive or negative sentiment?
Use the dictionary.txt for the score
Note: Include the date in the format 'yyyy-mm-dd', with tweet id, user name and the score.

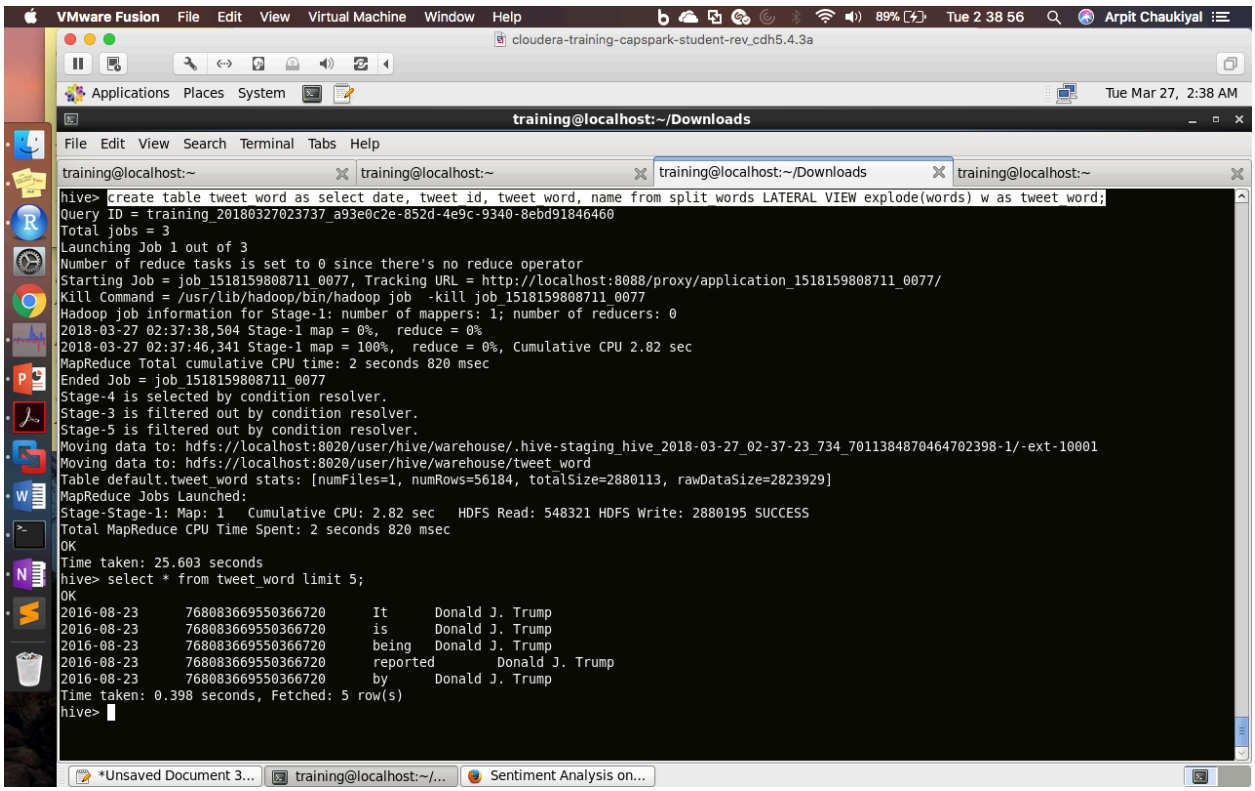


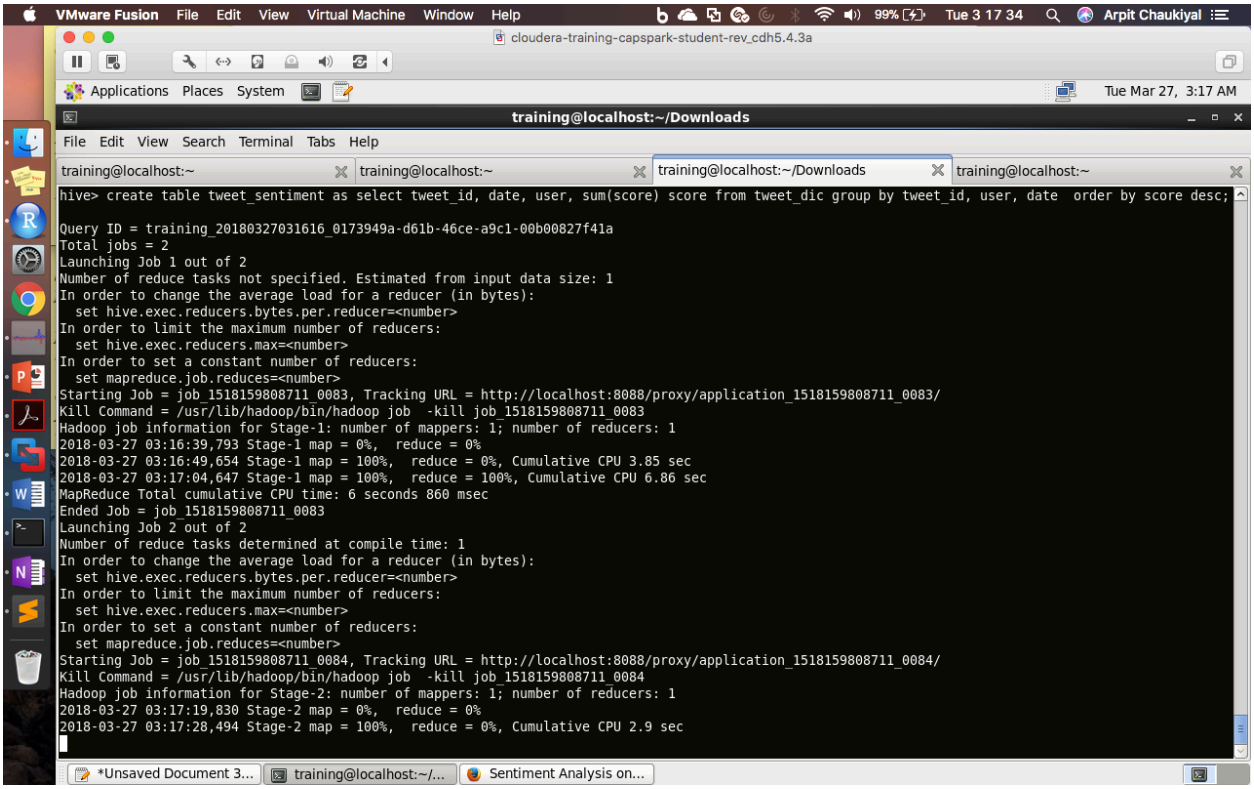


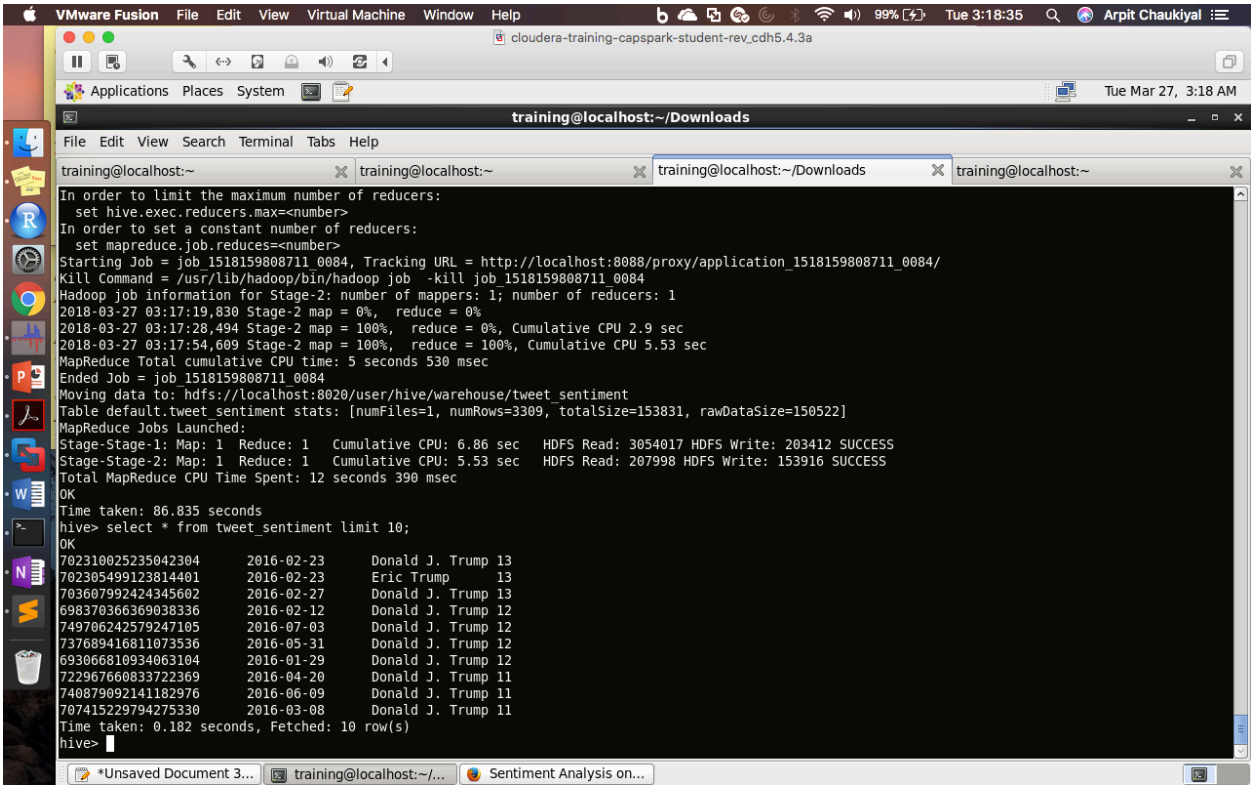

VMware Fusion File Edit View Virtual Machine Window Help
Tue 2 03 50
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 2:03 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> create table tweettext as select from_unixtime(unix_timestamp(created_at, "EEE MMM d HH:mm:ss Z yyyy"),"yyyy-MM-dd") AS date, id_str AS tweet_id
, text as tweet, user.name as name from maintweetdata;
Query ID = training_20180327020101_7437e0b0-e9a6-49f7-9c1d-f1e8ab567142
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0071, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0071/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0071
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 02:01:56,605 Stage-1 map = 0%, reduce = 0%
2018-03-27 02:02:35,511 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.62 sec
MapReduce Total cumulative CPU time: 5 seconds 620 msec
Ended Job = job_1518159808711_0071
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/.hive-staging_hive_2018-03-27_02-01-43_567_3519714248159618632-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweettext
Table default.tweettext stats: [numFiles=1, numRows=3207, totalSize=519757, rawDataSize=516550]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 5.8 sec HDFS Read: 9762980 HDFS Write: 519836 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 800 msec
OK
Time taken: 81.115 seconds
hive> select * from tweettext limit 2;
OK
2016-08-23 768083669550366720 It is being reported by virtually everyone, and is a fact, that the media pile on against me is the worst in
American political history! Donald J. Trump
2016-08-23 768069472464666624 I am now in Texas doing a big fundraiser for the Republican Party and a @FoxNews Special on the BORDER and wit
h victims of border crime! Donald J. Trump
Time taken: 0.282 seconds, Fetched: 2 row(s)
hive>
*Assignment 2 (~) - g...
training@localhost:~/...
Sentiment Analysis...


VMware Fusion File Edit View Virtual Machine Window Help
Tue 2:27:07
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 2:27 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> insert into tweettext select from_unixtime(unix_timestamp(retweeted_status.created_at, "EEE MMM d HH:mm:ss Z yyyy"),"yyyy-MM-dd"), retweeted_sta
tus.id_str, retweeted_status.text, retweeted_status.user.name from retweetdata where retweeted_status.text <> "NULL";
Query ID = training_20180327022323_cd45c16d-30f3-4adb-839e-afda5e49ad6a
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0075, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0075/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0075
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 02:23:51,878 Stage-1 map = 0%, reduce = 0%
2018-03-27 02:24:00,336 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.96 sec
MapReduce Total cumulative CPU time: 2 seconds 960 msec
Ended Job = job_1518159808711_0075
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweettext/.hive-staging_hive_2018-03-27_02-23-33_375_6921364005772303786-1/-ext-10000
Loading data to table default.tweettext
Table default.tweettext stats: [numFiles=2, numRows=3310, totalSize=536701, rawDataSize=533391]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 2.96 sec HDFS Read: 9763983 HDFS Write: 17021 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 960 msec
OK
Time taken: 35.144 seconds
hive>
Assignment 2 (~) - gedit
training@localhost:~/...
Sentiment Analysis on...


VMware Fusion File Edit View Virtual Machine Window Help
86% Tue 2:33:09 Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 2:33 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> create table split words as select date as date, tweet id as tweet id, split(tweet, ' ') as words, name as name from tweettext;
Query ID = training_20180327023030_ab3bbeb9-6c7b-465d-b19a-7a25c581702c
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0076, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0076/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0076
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 02:31:05,687 Stage-1 map = 0%, reduce = 0%
2018-03-27 02:31:14,258 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.62 sec
MapReduce Total cumulative CPU time: 3 seconds 620 msec
Ended Job = job_1518159808711_0076
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/.hive-staging_hive_2018-03-27_02-30-54_313_3766115241705033034-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/split_words
Table default.split_words stats: [numFiles=1, numRows=4111, totalSize=543910, rawDataSize=539799]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.62 sec HDFS Read: 540091 HDFS Write: 543991 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 620 msec
OK
Time taken: 24.387 seconds
hive> select * from slpit_words limit 5;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'slpit_words'
hive> select * from split_words limit 3;
OK
2016-08-23 768083669550366720 ["It","is","being","reported","by","virtually","everyone,","and","is","a","fact,","","that","the","media","pile","on","against","me","is","the","worst","in","American","political","history!"] Donald J. Trump
2016-08-23 768069472464666624 ["I","am","now","in","Texas","doing","a","big","fundraiser","for","the","Republican","Party","and","a","@FoxNews","Special","on","the","BORDER","and","with","victims","of","border","crime!"] Donald J. Trump
2016-08-22 767889674530594816 ["The","@WashingtonPost","quickly","put","together","a","hit","job","book","on","me-","comprised","of","copies","of","some","of","their","inaccurate","stories.","Don't","buy,","boring!"] Donald J. Trump
Time taken: 0.589 seconds, Fetched: 3 row(s)
*Unsaved Document 3...
training@localhost:~/...
Sentiment Analysis on...


VMware Fusion File Edit View Virtual Machine Window Help
89% Tue 2 38 56 Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 2:38 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> create table tweet word as select date, tweet id, tweet word, name from split words LATERAL VIEW explode(words) w as tweet word;
Query ID = training_20180327023737_a93e0c2e-852d-4e9c-9340-8ebd91846460
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1518159808711_0077, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0077/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0077
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-03-27 02:37:38,504 Stage-1 map = 0%, reduce = 0%
2018-03-27 02:37:46,341 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.82 sec
MapReduce Total cumulative CPU time: 2 seconds 820 msec
Ended Job = job_1518159808711_0077
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/.hive-staging_hive_2018-03-27_02-37-23_734_7011384870464702398-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweet_word
Table default.tweet_word stats: [numFiles=1, numRows=56184, totalSize=2880113, rawDataSize=2823929]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 2.82 sec HDFS Read: 548321 HDFS Write: 2880195 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 820 msec
OK
Time taken: 25.603 seconds
hive> select * from tweet_word limit 5;
OK
2016-08-23 768083669550366720 It Donald J. Trump
2016-08-23 768083669550366720 is Donald J. Trump
2016-08-23 768083669550366720 being Donald J. Trump
2016-08-23 768083669550366720 reported Donald J. Trump
2016-08-23 768083669550366720 by Donald J. Trump
Time taken: 0.398 seconds, Fetched: 5 row(s)
hive>
*Unsaved Document 3...
training@localhost:~/...
Sentiment Analysis on...


VMware Fusion File Edit View Virtual Machine Window Help
Tue 3 17 34
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 3:17 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
hive> create table tweet_sentiment as select tweet_id, date, user, sum(score) score from tweet_dic group by tweet_id, user, date order by score desc;
Query ID = training_20180327031616_0173949a-d61b-46ce-a9c1-00b00827f41a
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1518159808711_0083, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0083/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0083
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-03-27 03:16:39,793 Stage-1 map = 0%, reduce = 0%
2018-03-27 03:16:49,654 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.85 sec
2018-03-27 03:17:04,647 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.86 sec
MapReduce Total cumulative CPU time: 6 seconds 860 msec
Ended Job = job_1518159808711_0083
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1518159808711_0084, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0084/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0084
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-27 03:17:19,830 Stage-2 map = 0%, reduce = 0%
2018-03-27 03:17:28,494 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.9 sec
*Unsaved Document 3...
training@localhost:~/...
Sentiment Analysis on...


VMware Fusion File Edit View Virtual Machine Window Help
Tue 3:18:35
Arpit Chaukiyal
cloudera-training-capspark-student-rev_cdh5.4.3a
Applications Places System
Tue Mar 27, 3:18 AM
training@localhost:~/Downloads
File Edit View Search Terminal Tabs Help
training@localhost:~
training@localhost:~
training@localhost:~/Downloads
training@localhost:~
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1518159808711_0084, Tracking URL = http://localhost:8088/proxy/application_1518159808711_0084/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1518159808711_0084
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-03-27 03:17:19,830 Stage-2 map = 0%, reduce = 0%
2018-03-27 03:17:28,494 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.9 sec
2018-03-27 03:17:54,609 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 5.53 sec
MapReduce Total cumulative CPU time: 5 seconds 530 msec
Ended Job = job_1518159808711_0084
Moving data to: hdfs://localhost:8020/user/hive/warehouse/tweet_sentiment
Table default.tweet_sentiment stats: [numFiles=1, numRows=3309, totalSize=153831, rawDataSize=150522]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.86 sec HDFS Read: 3054017 HDFS Write: 203412 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 5.53 sec HDFS Read: 207998 HDFS Write: 153916 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 390 msec
OK
Time taken: 86.835 seconds
hive> select * from tweet_sentiment limit 10;
OK
702310025235042304 2016-02-23 Donald J. Trump 13
702305499123814401 2016-02-23 Eric Trump 13
703607992424345602 2016-02-27 Donald J. Trump 13
698370366369038336 2016-02-12 Donald J. Trump 12
749706242579247105 2016-07-03 Donald J. Trump 12
737689416811073536 2016-05-31 Donald J. Trump 12
693066810934063104 2016-01-29 Donald J. Trump 12
722967660833722369 2016-04-20 Donald J. Trump 11
740879092141182976 2016-06-09 Donald J. Trump 11
707415229794275330 2016-03-08 Donald J. Trump 11
Time taken: 0.182 seconds, Fetched: 10 row(s)
hive>
*Unsaved Document 3...
training@localhost:~/...
Sentiment Analysis on...

e. Do you find any problem in the way sentiment analysis was performed in the previous question? If so, how will you improve it?

Ans. Below are a few problems about the way of sentimental analysis we performed:

- When we break the tweet text into individual words the contextual meaning of those words is lost.
- There is no way we can get the sarcastic sense of the words.
- The numerical value (magnitude) which we had assigned to the words don't have the absolute method of assigning.
- The regular updation of the dictionary is required.

Ways to improve the analysis:

- Techniques like n-grams can be used to increase the accuracy of the analysis.
- Naive Bayes ML algos can be implemented for estimate more correct sentiment.