**Student Name: Nayan Tejani**

**Batch - BDAP PT Batch 01**

**Email - nayan.bdapm10150111@spjain.org**

**NoSQL Assignment (query over twitter records)**

**Pre-requisites:**

- MongoDB is installed on the target machine where this code provided in this assignment solution will be executed.

- path of mongoDB bin folder is configured in system variables>Path. This will enable global access to the mongoDB server executables

- mongod daemon is already running on the machine at the time of executing this code

1. [4 marks] Write and submit code to load all tweets from the JSON file and store them in your local MongoDB server.

**Solution**:

The below solution has been implemented using windows based approach. Similar approach can be taken for unix based import

Provided file name : "twitter.json"

Windows based steps: -

>> On a linux / unix machine, unzip the provided json tar file using the command

'tar – xv zf twitter.json.tar.gz'

>> copy the file from linux back to your windows file system

Assuming the mongoDB server does not contain a database name twitter\_records (at the time of running the below command)

The syntax for import is

mongoimport --db name\_of\_database --collection name\_of\_collection <"absolute path of the json file"

>> Launch a command prompt window and simply run the below command to import the json file to the mongoDB server

mongoimport --db twitter\_records --collection tweets < "D:\Innovations and Excellance\Tech-Mag\Big Data\NoSQL\twitter data\twitter.json"

>> To verify whether database got created and review its collections use below commands: -

show dbs

use twitter\_records

show collections

db.tweets.count()

db.tweets.find().limit(10).toArray()

>> check 1 or few records from the tweets collections to study the data structure of those records

db.tweets.findOne()

db.tweets.find().pretty().limit(1)

db.tweets.find({},{"user":1,"name":1,"\_id":0,"retweet\_count":1,"entities":1}).pretty().limit(5)

**Conclusion / Result**:

**mongoimport --db twitter\_records --collection tweets < "D:\Innovations and Excellance\Tech-Mag\Big Data\NoSQL\twitter data\twitter.json"**

the file size is over 1.2 GB and it will take approximately 2-3 mins to complete the import (depends upon the machine configuration where import is being processed)

Upon completion, it should display last line as "imported 308930 documents"

2. [4 marks] What is the 5th most frequent hashtag in the dataset? Submit along with code. (the answer should be derived using a mongodb query)

Solution:

>> We need to use the aggregation framework for this. Firstly need to find records where hashtags exists

>> The hashtags are case-insensitive. Hence #BigData and #bigData should refer to the same hashtag. Here we need to convert the hasttag.text in Group by to upper / lower case.

>> since we need the 5th most frequent hashtag, we need to skip 4 and also include a clause to limit the result to 1

>> final query would be like as below: -

**db.tweets.aggregate(**

**{$match:{"entities.hashtags":{$exists : true}}},**

**{$project: {"user.screen\_name":1,"user.name":1,"entities.hashtags":1}},**

**{$unwind : "$entities.hashtags"},**

**{$group : {\_id:{$toUpper:"$entities.hashtags.text"},HashTagCount:{$sum:1}}},**

**{$sort : {HashTagCount:-1}},**

**{$skip:4},**

**{$limit:1}**

**)**

**Conclusion / Result:**

5th most frequent hashtag is “**NEWS”**

**{**

**"\_id" : "NEWS",**

**"HashTagCount" : NumberInt(15258)**

**}**

3. [4 marks] What is the 7th most frequent @-mention in the dataset? Submit along with code. (the answer should be derived using a mongodb query)

**Solution**:

>> we will use aggregation framework for this

>> We need to query the records having user\_mentions tag within entities

>> set the projection of columns to be shown

>> unwind the user\_mentions array fields

>> group by user\_mentions.screen\_name and generate the count of records

>> sort the MentionsCount by descending order

>> to get the 7th most frequent @-mention, skip the first 6 and limit the result by 1. This will be the 7th top most @-mention

**db.tweets.aggregate(**

**{$match:{"entities.user\_mentions":{$exists : true}}},**

**{$project:{"user.screen\_name":1,"user.name":1,"entities.user\_mentions.screen\_name":1}},**

**{$unwind : "$entities.user\_mentions"},**

**{$group : {\_id:{$toUpper:"$entities.user\_mentions.screen\_name"},MentionedCount:{$sum:1}}},**

**{$sort : {MentionedCount:-1}},**

**{$skip:6},**

**{$limit:1})**

**Conclusion / Result:**

**User with screen\_name : “Forbes” has the 7th most frequent mention**

**{**

**"\_id" : "FORBES",**

**"MentionedCount" : NumberInt(2711)**

**}**

4. [4 marks] Which twitter user has 3rd most number of followers? Submit along with code. (the answer should be derived using a mongodb query)

**Solution**:

>> we will use aggregation framework for this

>> firstly we need to query the records having user.followers\_count>0

>> the results should be sorted using descending sort on user.followers\_count

>> since we want 3rd most number of followers\_count, hence we need to skip the first 2 using skip(2). if nth most number of followers, then skip n-1 records by using skip(n-1)

>> since we want only 1 record i.e. hence limit the results by 1 using limit(1)

>> restrict the number of fields displayed in the results since we just want few details about the user and other fields may not be required to be displayed

>> final query would like as below: -

**db.tweets.find(**

**{"user.followers\_count":{$ne:null,$gt:0}},**

**{text:1,source:1,created\_at:1,"user.name":1,"user.screen\_name":1,"user.followers\_count":1,"user.friends\_count":1}**

**).sort({"user.followers\_count":-1}).limit(1).skip(2)**

**Conclusion / Result**:

**{**

**"\_id" : ObjectId("577fa0e2bfc80e10014ee20e"),**

**"created\_at" : "Thu Feb 18 14:25:03 +0000 2016",**

**"text" : "Employers are quietly using big data to track employee pregnancies https://t.co/w8m3A0kYKF",**

**"source" : "<a href=\"http://www.socialflow.com\" rel=\"nofollow\">SocialFlow</a>",**

**"user" : {**

**"name" : "TIME.com",**

**"screen\_name" : "TIME",**

**"followers\_count" : NumberInt(9714612),**

**"friends\_count" : NumberInt(843)**

**}**

**}**

To Test the query correctness:

db.tweets.find({"user.followers\_count":{$gt:9714612}}).pretty() --> this will return only 2 records. so that means that there are only 2 users who have the followers\_count > than our current results

Thus proving that the user fetched in our query has the 3rd most number of followers

5. [4 marks] Which twitter user's tweets were retweeted 8th most number of times? Submit along with code. (the answer should be derived using a mongodb query)

**Solution**:

**db.tweets.find(**

**{"retweeted\_status" : {$exists : true}},**

**{"retweeted\_status.user.screen\_name":1,"retweeted\_status.user.name":1,"retweeted\_status.retweet\_count":1}**

**).pretty().sort({"retweeted\_status.retweet\_count":-1}).skip(7).limit(1)**

**Conclusion / Result**:

User name is "Christopher Mims" who tweets have been re-tweeted 8th most number of times

**{**

**"\_id" : ObjectId("577fa0cbbfc80e10014e7005"),**

**"retweeted\_status" : {**

**"user" : {**

**"name" : "Christopher Mims",**

**"screen\_name" : "mims"**

**},**

**"retweet\_count" : NumberInt(991)**

**}**

**}**