

# **Project 1 Report**

Team Number:9

## **UMKC School of Computing and Engineering**

### Done by:

- Nageswara Rao Nandigam(nrnxh9)
- Revanth Chakilam(rcww4)
- Syed Moin(slrp3)
- Devender Sarda(dspc8)

# **Project Tasks:**

- Main Requirements:
- Collect tweets in JavaScript Object Notation (JSON) format (at least 100K record).
  - > Find the list of top ten used hashtags in your collection.
- Create a directory in HDFS for each hashtag from the top ten hashtag list.
  - Create additional two directories: "Others" and "None"
- Store the tweets on files in HDFS
  - If a tweet contains a hashtag from the top ten list, store the tweet in that hashtag's directory.
  - If a tweet contains one or more hashtags, but none of the hashtags are in the top ten list, store the tweet in the "Others" directory.
  - ➤ If a tweet does not contain a hashtag, store it in the "None" directory.



## Collect tweets in JavaScript Object Notation (JSON) format (at least 100K record)

We created a twitter development account to get consumer key and access keys and using this tokens we developed python script to collect tweets in JSON format. We programed a easy way to get the tweets as may we want by changing single parameter. So we are able to get 1 lakh tweets which stored in local disk file.

### Screen shot of Pyhton program to collect tweets

```
Open ▼
    import json
except ImportError:
     import simplejson as json
# Import the necessary methods from "twitter" library from twitter import Twitter, OAuth, TwitterHTTPError, TwitterStream
# Variables that contains the user credentials to access Twitter API
ACCESS_TOKEN = '434908233-C896jBM6t1Iq7ta0FJ93YQeD3QF2737.122.

ACCESS_SECRET = 'JTPffFFLE6Qk1j4c2YTLDeeF8kkjop1Vtw4ANanVS1ivx'

CONSUMER_KEY = 'bjnYTbDv0d61oShbWWyaD2leP'
CONSUMER_SECRET = 'mCy4iONQfevhJKmfqK1Kt0GjHb1uG0z2f3VUrhN3upmPZMXgsp'
oauth = OAuth(ACCESS_TOKEN, ACCESS_SECRET, CONSUMER_KEY, CONSUMER_SECRET)
# Initiate the connection to Twitter Streaming API
twitter_stream = TwitterStream(auth=oauth)
iterator = twitter_stream.statuses.sample()
tweet_count = 100000
for tweet in iterator:
     tweet_count -= 1
     print json.dumps(tweet)
     if tweet_count <= 0:</pre>
          break
```

Python ▼ Tab Width: 8 ▼ Ln 4, Col 30 ▼ INS



# Principles of Big data and Management 01/26/2017

#### **Output:**

["favorited": false, "contributors": null, "truncated": false, "text": "My son struggles to grasp the intricacies of the... water fountain. https://t.co/NG60iks {"delete": {"status": {"user\_id\_str": "1177832576", "user\_id": 1177832576, "id": 822930086852603904, "id\_str": "822930086852603904"}, "timestamp\_ms": "148796206" {"delete": {"status": {"user id str": "4746158896", "user id": 4746158896, "id": 835198438639751169, "id str": "835198438639751169"}, "timestamp ms": "148796206 {"delete": {"status": {"user\_id\_str": "26093694", "user\_id": 26093694, "id": 800414173788336128, "id\_str": "800414173788336128"}, "timestamp\_ms": "148796206139 {"delete": {"status": {"user id str": "40116700", "user id": 40116700, "id": 504318792852598784, "id str": "504318792852598784"}, "timestamp ms": "1487962061457 ("favorited": false, "contributors": null, "truncated": false, "text": "<\u3054\u3008d\u3073\u3063\u305f\u3093", "is quote status": false, "in reply to status {"favorited": false, "contributors": null, "truncated": false, "text": "RT @SoWhiteGurl: These 25 Hilarious Drunk Text Fails Will Leave You Speechless https://t {"favorited": false, "contributors": null, "truncated": false, "text": "RT @SSOSLA NOTICIAS: Brothers #iHeartAwards #BestMovieSong #GIRLSTALKBOYS -Nat\ud83d\udc {"favorited": false, "contributors": null, "truncated": false, "text": "RT @boob445822: #\u0627\u0644\u0635\u0631\u062e\u064a \u064a\u0642\u062a\u0644\u0639 \u0 {"favorited": false, "contributors": null, "truncated": false, "text": "@Heyitshassan So glad to hear this! \ud83d\ude02\ud83d\ude4c", "is quote status": false {"favorited": false, "contributors": null, "truncated": false, "text": "\*smh auto-fecking-incorrect strikes again", "is quote status": false, "in reply to statu {"favorited": false, "contributors": null, "truncated": false, "text": "(\u0648\u0645\u0627\u0623\u0638\u0646\u0627\u0644\u0633\u0627\u0633\u0627\u0633\u0627\u0633\u0627\u0633\u0627\u0633\u0627\u0633\u06427\u0633\u0644\u0633\u06427\u0633\u06427\u0633\u0644\u06437\u0644\u0644\u06437\u0644 {"favorited": false, "contributors": null, "truncated": false, "text": "\u0633\u0628\u0623\u0624\u0647\u0641\u0641\u0641\u0647\u0648\u0642\u06421\u0648\u0628\u0628\u0628\u0628\u0628\u0628\u0628\u0645\u0645\u0648\u064 ("favorited": false, "contributors": null, "truncated": false, "text": "@dwegawa00 i love you more than you know hanbanan! thanks for being my crazy self's best {"favorited": false, "contributors": null, "truncated": false, "text": "Current status: mining @janeschmidt 's blog for advice!", "is\_quote\_status": false, "in\_ {"favorited": false, "contributors": null, "truncated": false, "text": "RT @Yoraa7: \u0628\u0639\u062f \u0625\u0630\u0646 \u0627\u0644\u0646\u0635\u06637\u0627\u0628\u {"favorited": false, "contributors": null, "truncated": false, "text": "RT @justindbizz22: As long as you love me, we could be starving, we could be homeless and ("favorited": false, "contributors": null, "truncated": false, "text": "RT @chrisviteri: Favor objetividad. Donaciones ya se consumieron y eran 4 gatos y regard {"favorited": false, "contributors": null, "truncated": false, "text": "RT @marilena madrid: RT \u00bfTE APUNTAS A HACER VOLUNTARIADO?. Visitar enfermos mentala {"favorited": false, "contributors": null, "truncated": false, "text": "RT @ FatosDoMundo : Uma das melhores sensa\u00e7\u00e7\u00e7bes da vida \u00e9 ter a certeza qu {"favorited": false, "contributors": null, "truncated": false, "text": "@BAFTA \ud83e\udd1e\ud83d\ude42", "is quote status": false, "in reply to status id": 885 {"favorited": false, "contributors": null, "truncated": false, "text": "Thats the only reason I see y'all loving DBZ...cause it's so many more entertaining/bett {"favorited": false, "contributors": null, "truncated": false, "text": "RT @stevebrown442: More fantastic football action on @LiveSportsFM https://t.co/1HsRN4ht ("favorited": false, "contributors": null, "truncated": false, "text": "RT @peytonnhaenn: i don't really care if someone doesn't like me lmao y'all barely like {"favorited": false, "contributors": null, "truncated": false, "text": "RT @MSNBC: O'Malley on Trump's CPAC speech: These are the things we hear from \"authorit {"favorited": false, "contributors": null, "truncated": false, "text": "RT @Reza41183821: \u0628\u0647\u062a\u0641\u06cc\u0646\u0648\u0627\u06a8\u0648\u0627\u06a8\u0648 {"favorited": false, "contributors": null, "truncated": false, "text": "RT @stelovesrdj: Never let this die https://t.co/020RWpfssz", "possibly\_sensitive": fals {"favorited": false, "contributors": null, "truncated": false, "text": "RT @roysavoy: Williamsburg.\n‡ricohgr2 ‡vsco https://t.co/U2TgbcXWq7 https://t.co/zXsFYk ("favorited": false, "contributors": null, "truncated": false, "text": "RT @RealMaddalena: Hostaria da Pietro \u2764a real taste of #Rome near Campo Marzio #Fri {"delete": {"status": {"user id str": "232251962", "user id": 232251962, "id": 800926545745317888, "id str": "800926545745317888"}, "timestamp ms": "14879620618 {"favorited": false, "contributors": null, "truncated": false, "text": "RT @teenahshai: \*laughs in lgbt\* https://t.co/uuwAD645kc", "possibly sensitive": false, {"favorited": false, "contributors": null, "truncated": false, "text": "RT @washingtonpost: Science societies have long shunned politics. But now they\u2019re r {"delete": {"status": {"user id str": "1695098550", "user id": 1695098550, "id": 835199298451144704, "id str": "835199298451144704"}, "timestamp ms": "148796206 {"favorited": false, "contributors": null, "truncated": false, "text": "RT @NASASunEarth: Our solar observatory watched this dance of solar material over the su {"favorited": false. "contributors": null. "truncated": false. "text": "RT @ashlevguiet: not to be dramatic but i wish you were dead https://t.co/VMlmxtYcTi". "



#### Find the list of top ten used hashtags in your collection

we developed JAVA program to find out top ten hashtags. We followed coupled of steps to get top 10 hashtags

1) Read the input twitter collections file through bufferedreader reader line by line because each line is related to one tweet and store it in temporary String each time to parse like a JSONObject.

```
----,
46
47
        // Reading the twitter collections file
48
       try
49
           br=new BufferedReader(new FileReader("/home/nag/Desktop/twitter 1L.json"))
50
51
            str=new StringBuilder();
52
           String line=br.readLine();
53
           while(line!=null)
55
               str.append(line);
56
               line=br.readLine();
57
58
          result=str.toString();
59
           str.delete(0, str.length());
60
```

2) We pass the above temp string to JSONObject as a parameter to read like JSON format data and iteratively we read JSON objects until we find hashtag array object, after that we are retrieving hashtag text.

```
. ....
         //Retrieving the JSONobjects from the file and getting hashtag texts
51
52
           JSONObject object=new JSONObject(result);
           if(object.has("retweeted_status"))
55
           JSONObject re=object.getJSONObject("retweeted status");
56
57
           JSONObject e=re.getJSONObject("entities");
           JSONArray hash=e.getJSONArray("hashtags");
58
59
           if(hash.length()!=0){
           JSONObject t=hash.getJSONObject(0);
70
71
72
           for(int i=0;i<1;i++)</pre>
73
74
75
               String s=t.getString("text");
76
```



3) After getting hashtag text, those values stored in hash map as a key-value pair. So, that if same hashtag appears next time it will increment its value by one. So, that we will get to know total repetitions of each hashtag.

```
76
77
                //Storing those hastags in hashmap like key-value pair
78
79
            if(map.containsKey(s.toLowerCase())){
80
                int j=map.get(s.toLowerCase());
81
                map.put(s.toLowerCase(),j+1);
            }
82
83
            else
84
85
86 map.put(s.toLowerCase(),1);
87
88
89
90
       }}}
91
```

4) To find out top ten hashtags from hash map, we took another tree map which has function of sorting the elements by key, so we reverse the key as integer (total count) and hash tag as value. So finally, will get the sorted list of small count to highest count hash tags.

```
92
3
   Set<Entry<String, Integer>> set = map.entrySet();
     Iterator<Entry<String, Integer>> iterator = set.iterator();
)4
95
     while(iterator.hasNext()) {
96
        Map.Entry<String,Integer> mentry = (Map.Entry<String,Integer>)iterator.next()
97
       if(mentry.getValue()>=10)
86
           tr.put(mentry.getValue(),mentry.getKey());
0
   }
)1
```



01/26/2017

5) As we mentioned above last entries had most frequency hashtag count, so we are iterating from last entry to till 10 entries above to get top list.

```
//Finding out top ten hashtags from the hashmap and displaying in console

tr.lastEntry();
int i=1;
ArrayList<String> a=new ArrayList<String>();
while(i<=10) {
    Map.Entry<Integer,String> mentry = (Map.Entry<Integer,String>)tr.lastEntry(

    System.out.print("key is: "+ mentry.getKey() + " & Value is: ");
    System.out.println(mentry.getValue());
    a.add(mentry.getValue());
    tr.remove(mentry.getKey());
    i++;
```

#### **Output:**

```
<terminated> Hashtag [Java Application] /usr/lib/jvm/java-8-openjdk-amd64/bin/java (Feb 27, 2017, 12:36:39 AM)
```

```
key is: 202 & Value is: الوكر و فولورز نصا بين key is: 117 & Value is: pillowtalk key is: 104 & Value is: workfromhome key is: 93 & Value is: harmonizers key is: 69 & Value is: here key is: 62 & Value is: 사설토토추천사이트 key is: 60 & Value is: videolove key is: 56 & Value is: exsandohs key is: 54 & Value is: kıskanıyorum key is: 48 & Value is: iheartawards
```

Note: As you can see we found some hash tags are in other language too



- Create a directory in HDFS for each hashtag from the top ten hashtag list.
- Create additional two directories: "Others" and "None".

```
ell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project
und 13 items
w-r--r-- 3 nag supergroup 1072811 2017-02-27 00:40 /user/project/
토幸천사이트, videolove, exsandohs, kıskanıyorum, iheartawards]
wxr-xr-x - nag supergroup 0 2017-02-26 21:39 /user/project/Harmonizers
- 0217-02-26 21:39 /user/project/Here
ound 13 items
                                              ب pillowtalk, workfromhome, harmonizers, here, 사 رويطرن زول فو و ركول ا/vser/project/ 00:40 /user/project/ رويبطرن زول في ا
                   nag supergroup
                                                      0 2017-02-26 21:39 /user/project/Kiskaniyorum
                        supergroup
                                                         2017-02-27 00:56 /user/project/Others
                   nag
                                                         2017-02-26 21:39 /user/project/Pillowtalk
2017-02-26 21:39 /user/project/VideoLove
                   nag supergroup
                   nag
                        supergroup
                                                      0 2017-02-26 21:39 /user/project/WorkFromHome
0 2017-02-26 21:39 /user/project/iHeartAwards
WXT-XT-X
                   nag supergroup
                   nag supergroup
                   nag supergroup
                                                                                  /user/project/none
```

- Store the tweets on files in HDFS
- ➤ If a tweet contains a hashtag from the top ten list, store the tweet in that hashtag's directory.
- ➤ If a tweet contains one or more hashtags, but none of the hashtags are in the top ten list, store the tweet in the "Others" directory.
  - ➤ If a tweet does not contain a hashtag, store it in the "None" directory.

We developed a program to store each hashtag tweet information in respective hdfs hash tag directory. To integrate such java functionality to call hdfs filesystem we required to have lot more JAR dependency to be able to copy files in Hadoop. To make it work that way we added all those dependencies. Finally, we able to pass the contents to Huffs file system.

#### Principles of Big data and Management

#### 01/26/2017

```
Quick Access 🗎 🖺
                                                                                                                                                                                                                                                                 🔝 Hashtag.java 🛭
Ð
                                                                                                                                                                                                                                                                   Ð
       168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
191
191
192
193
194
195
197
198
199
200
201
202
203
204
205
206
207
207
208
                                                                                                                                                                                                                                                                  8
                       fw= new FileWriter(f,true);
bw=new BufferedWriter(fw);
bw.write(line);
bw.flush();
bw.flush();
                                                                                                                                                                                                                                                                  ▣
                                                                                                                                                                                                                                                                  bw=null; fw= new FileWriter("/home/nag/Desktop/Pb/Others.txt",true); bw=new BufferedWriter(fw); bw.write(line); bw.flush();
                                                                                                                                                                                                                                                                  <u>=</u>
                             }}
else
                                   bw=null;
fw= new FileWriter("/home/nag/Desktop/Pb/none.txt",true);
bw=new BufferedWriter(fw);
bw.write(line);
bw.flush();
                            }
                       }
                                   bw=null; fw= new FileWriter("/home/nag/Desktop/Pb/none.txt",true); bw=new BufferedWriter(fw); bw.write(line); bw.flush();
                              }
line=brl.readLine();
                       }catch(Exception e)
{
                            e.getMessage();
                      }
                                                                                                                                  Writable
                                                                                                                                                          Smart Insert
                                                                                                                                                                                 161:43
```

# Principles of Big data and Management 01/26/2017

```
}
//dumping files to hadoop system in each respective hashtag directory

void hdfs(ArrayList<String> s) throws IOException, URISyntaxException
{
    URI url=new URI("hdfs://localhost:9000"); //-----> (url where hdfs located)
    Configuration conf = new Configuration();

    FileSystem file1= FileSystem.get(url, conf);
    Path a = new Path("/home/nag/Desktop/Pb/none.txt");
    Path b = new Path("/user/project/none");
    file1.copyFromLocalFile(a,b);
    Path ab = new Path("/home/nag/Desktop/Pb/Others.txt");
    Path bc = new Path("/user/project/Others");
    file1.copyFromLocalFile(ab,bc);

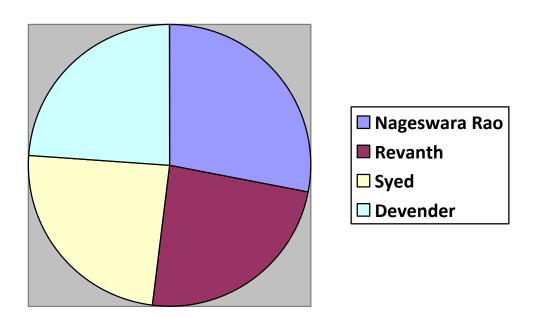
for(int i=0;i<s.size();i++)
{
        Path x = new Path("/home/nag/Desktop/Pb/"+s.get(i)+".txt");
        Path y = new Path("/user/project/"+s.get(i));
        file1.copyFromLocalFile(x,y);
    }
}
</pre>
```

#### **Output:**

```
ن صبطين زرول في وركول ا/nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project
Found 1 items
-rw-r--r-- 3 nag supergroup 1072811 2017-02-27 00:56 /user/project/ل عيطن زرول أو و رك ول الزيح طن زرول أو و رك ول المناطقة (txt
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/Others
                3 nag supergroup 32314221 2017-02-27 00:56 /user/project/Others/Others.txt
------
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/none
Found 1 items
-rw-r--r-- 3 nag supergroup 161409867 2017-02-27 00:56 /user/project/none/none.txt
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/WorkFromHome
Found 1 items
-rw-r--r-- 3 nag supergroup 820832 2017-02-26 21:39 /user/project/WorkFromHome/WorkFromHome.txt
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/ExsAndohs
ls: '/user/project/ExsAndohs': No such file or directory
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/ExsAndOhs
Found 1 items
                3 nag supergroup
                                               432848 2017-02-26 21:39 /user/project/ExsAndOhs/ExsAndOhs.txt
- FW- F-- F--
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls_/user/project/VideoLove
Found 1 items
-rw-r--r-- 3 nag supergroup 469606 2017-02-26 21:39 /user/project/VideoLove/VideoLove.txt
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$ bin/hadoop fs -ls /user/project/Pillowtalk
Found 1 items
-rw-r--r-- 3 nag supergroup 798982 2017-02-26 21:39
nag@nag-Dell-System-XPS-L502X:~/Downloads/hadoop-2.7.3$
                                               798982 2017-02-26 21:39 /user/project/Pillowtalk/Pillowtalk.txt
```



## **Project Contribution:**



## References:

- 1) <a href="https://dev.twitter.com/">https://dev.twitter.com/</a>
- 2) <a href="http://stackoverflow.com/">http://stackoverflow.com/</a>
- 3) <a href="http://hadoop.apache.org/">http://hadoop.apache.org/</a>