COLLECTING TWEETS USING TWITTER STREAMING API'S

Principles of Big Data Management (Phase 2 of Project)

Instructor

Dr. PRAVEEN RAO, Ph.D.

Abstract:

Phase 2 of this project deals with the following requirements:

1. Writing interesting analytical queries on twitter data that we collected.
2. Developing interesting visualizations.

Title:

Technology in different domains.

Technologies and Tools used:

1. Hadoop
2. Spark
3. Scala
4. Tableau
5. Java and MapReduce programs

Queries and Analysis:

Query-1 :

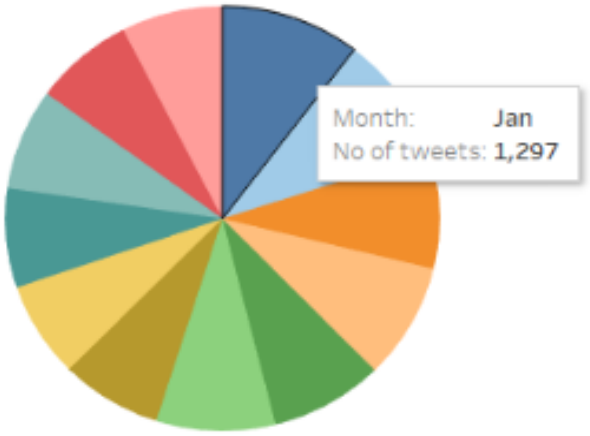
In this query, We found number of users created based on month.

Code :

```
scala> val Query1 = sqlContext.sql("SELECT substring(user.created_at,5,3) as month, count(user.id) from tweetDatatable group by month");
19/05/03 10:08:14 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Query1: org.apache.spark.sql.DataFrame = [month: string, count(user.id AS `id`): bigint]

scala> Query1.show();
+-----+----------------------+
|month|count(user.id AS `id`)|
+-----+----------------------+
|  Oct|                   968|
|  Sep|                   946|
|  Dec|                   946|
|  Aug|                   912|
|  May|                  1054|
|  Jun|                  1110|
|  Feb|                  1207|
|  Nov|                   928|
|  Mar|                  1101|
|  Jan|                  1297|
|  Apr|                  1109|
|  Jul|                   943|
+-----+----------------------+
```

Visualization :



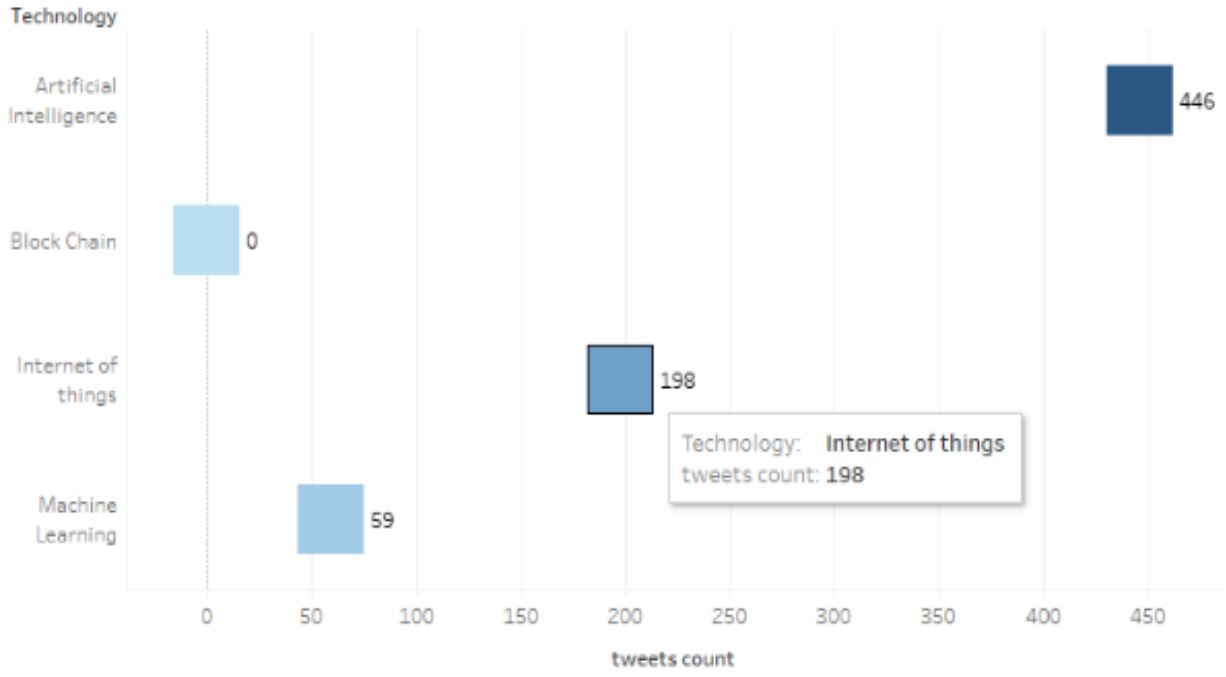
Query-2 :

In this query we found the trend in twitter for Artificial Intelligence, Internet of Things, Machine learning, Block chain etc.,

Code :

```
scala> val Query2 = sqlContext.sql("SELECT COUNT(*) AS NumberOfTweets, 'Artificial Intelligence' as Language FROM Technology where text LIKE '%Artificial Intelligence%' or text like '%AI%' UNION SELECT COUNT(*) AS Number(
Query2:Query2.show();
+-------------+--------------------+
|NumberOfTweets|            Language|
+-------------+--------------------+
|          198|  Internet of things|
|          446|Artificial Intell...|
|            0|         Block Chain|
|           59|    Machine Learning|
+-------------+--------------------+
```

Visualization :



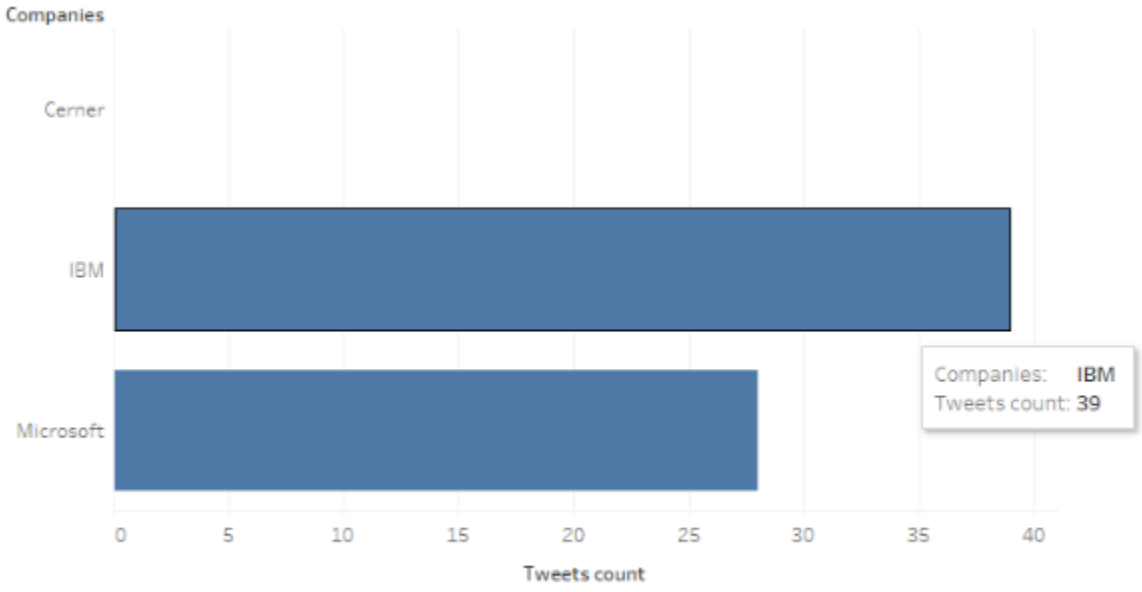
Query-3 :

In this query we found how Microsoft, Cerner and IBM associates/ employees are actively tweeting about AI.

```
scala> val Query3 = sqlContext.sql("SELECT 'Microsoft' as Company, count(*) as Count from Technology where text like '%Microsoft%' and (text like '%technology%' or text like '%tech%') UNION SELECT 'IBM' as Company, count(*)
Query3: org.apache.spark.sql.DataFrame = [Company: string, Count: bigint]

scala> Query3.show();
+---------+-----+
|  Company|Count|
+---------+-----+
|Microsoft|   28|
|   Cerner|    0|
|      IBM|   39|
+---------+-----+
```

Visualization :



Query-4 :

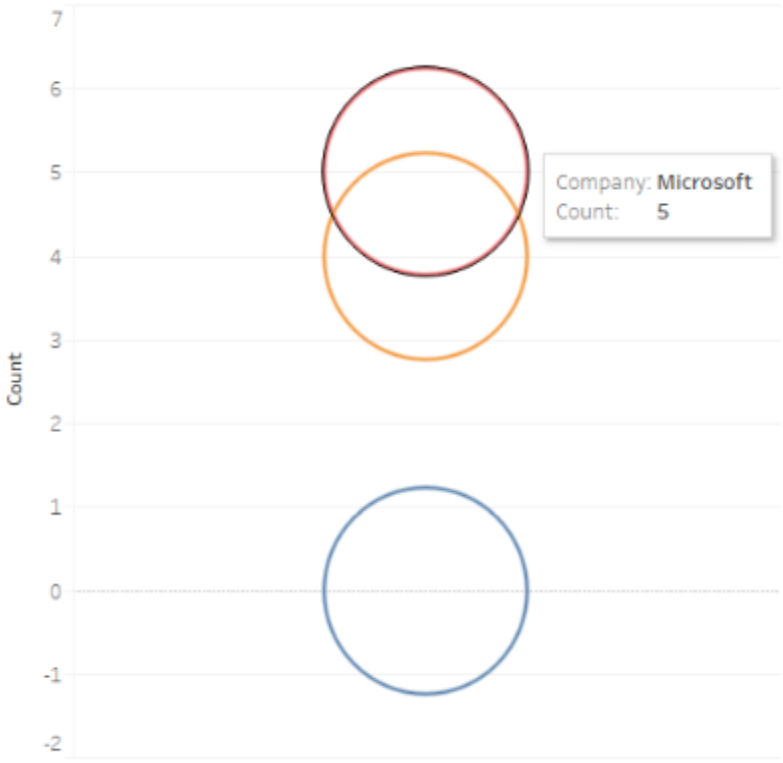
In this query we found number of tweets made by companies like Amazon, Microsoft, IBM on technologies like Artificial Intelligence, Machine Learning, Internet of Things etc.,

Code :

```
scala> val Query4 = sqlContext.sql("SELECT 'Microsoft' as Company, count(*) as Count from Technology where text like '%Microsoft%' and (text like '%AI%' or text like '%IoT%' or text like '%ML%') UNION SELECT 'IBM' as Company
Query4: org.apache.spark.sql.DataFrame = [Company: string, Count: bigint]

scala> Query4.show();
+---------+-----+
|  Company|Count|
+---------+-----+
|   Amazon|    0|
|      IBM|    4|
|Microsoft|    5|
+---------+-----+
```

Visualization :



Query-5 :

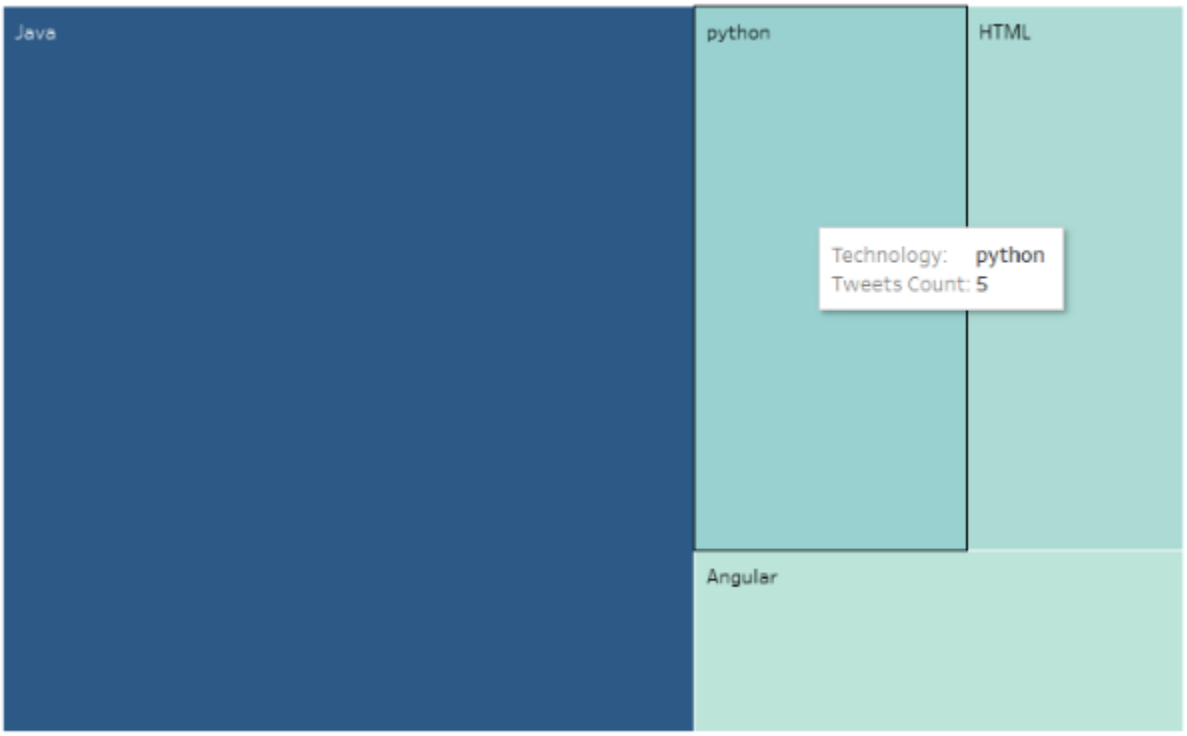
In this query we found number of tweets made on different technologies like Java, Python, HTML, Angular etc.,

Code :

```
scala> val Query5 = sqlContext.sql("SELECT COUNT(*) AS NumberOfTweets, 'HTML' as Language FROM Technology where text LIKE '%HTML%' UNION SELECT COUNT(*) AS NumberOfTweets,'Java' as Language FROM Technology where text LIKE '%
Query5: org.apache.spark.sql.DataFrame = [NumberOfTweets: bigint, Language: string]

scala> Query5.show();
+--------------+--------+
|NumberOfTweets|Language|
+--------------+--------+
|            17|    Java|
|             5|  python|
|             3| Angular|
|             4|    HTML|
+--------------+--------+
```

Visualization :



Query-6 :

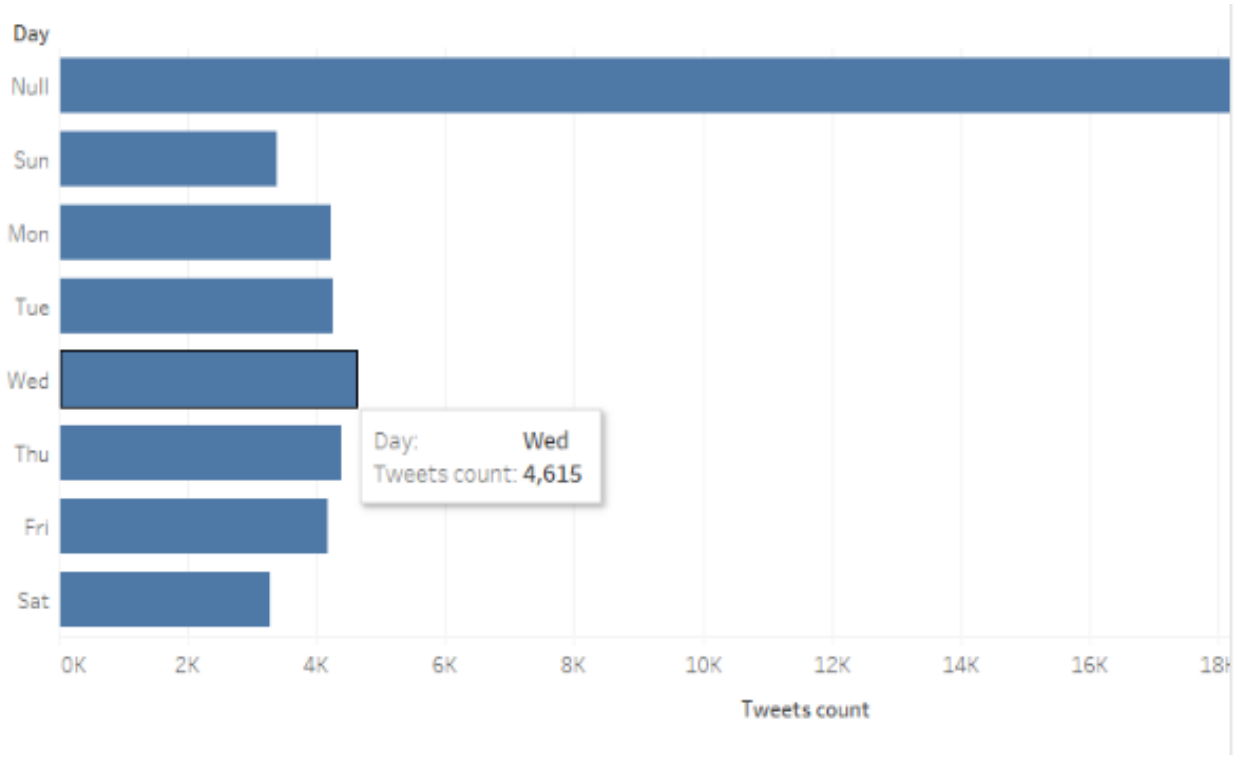
In this query we counted number of tweets made on technologies according to days.

Code :

```
scala> val Query6=sqlContext.sql("SELECT substring(user.created_at,1,3) as day,count(*) as count from technology group by day");
19/05/05 08:36:55 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Query6: org.apache.spark.sql.DataFrame = [day: string, count: bigint]

scala> Query6.show();
+----+-----+
| day|count|
+----+-----+
| Sun| 3399|
|null|22095|
| Mon| 4218|
| Thu| 4397|
| Sat| 3277|
| Wed| 4615|
| Tue| 4250|
| Fri| 4184|
+----+-----+
```

Visualization :



Query-7 :

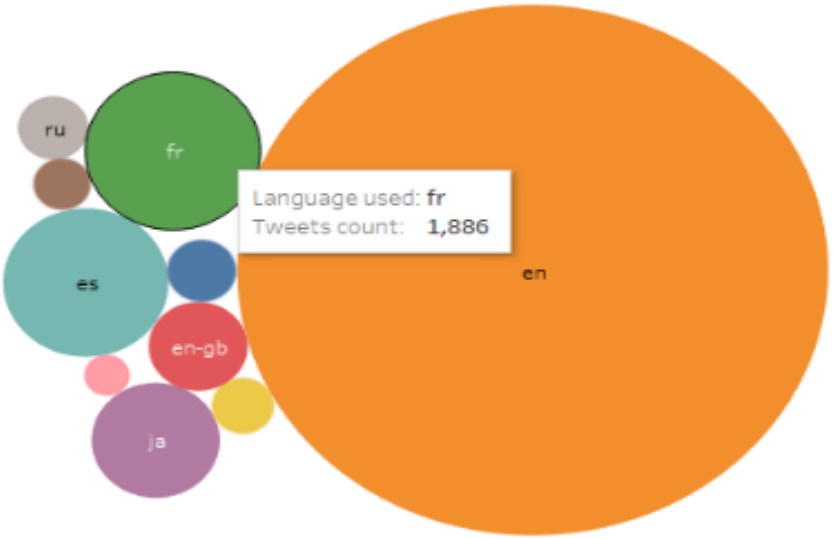
In this query, we found different languages used to tweet about technologies and their count respectively.

Code :

```
scala> val Query5 = sqlContext.sql("SELECT user.lang, count(*) AS count FROM technology WHERE lang<>'null' GROUP BY user.lang ORDER BY count DESC LIMIT 10");
Query5: org.apache.spark.sql.DataFrame = [lang: string, count: bigint]

scala> Query5.show();
+-----+-----+
| lang|count|
+-----+-----+
|   en|21412|
|   fr| 1886|
|   es| 1677|
|   ja| 1007|
|en-gb|  602|
|   ru|  307|
|   de|  295|
|   it|  239|
|   pt|  203|
|   nl|  128|
+-----+-----+
```

Visualization :



Query-8 :

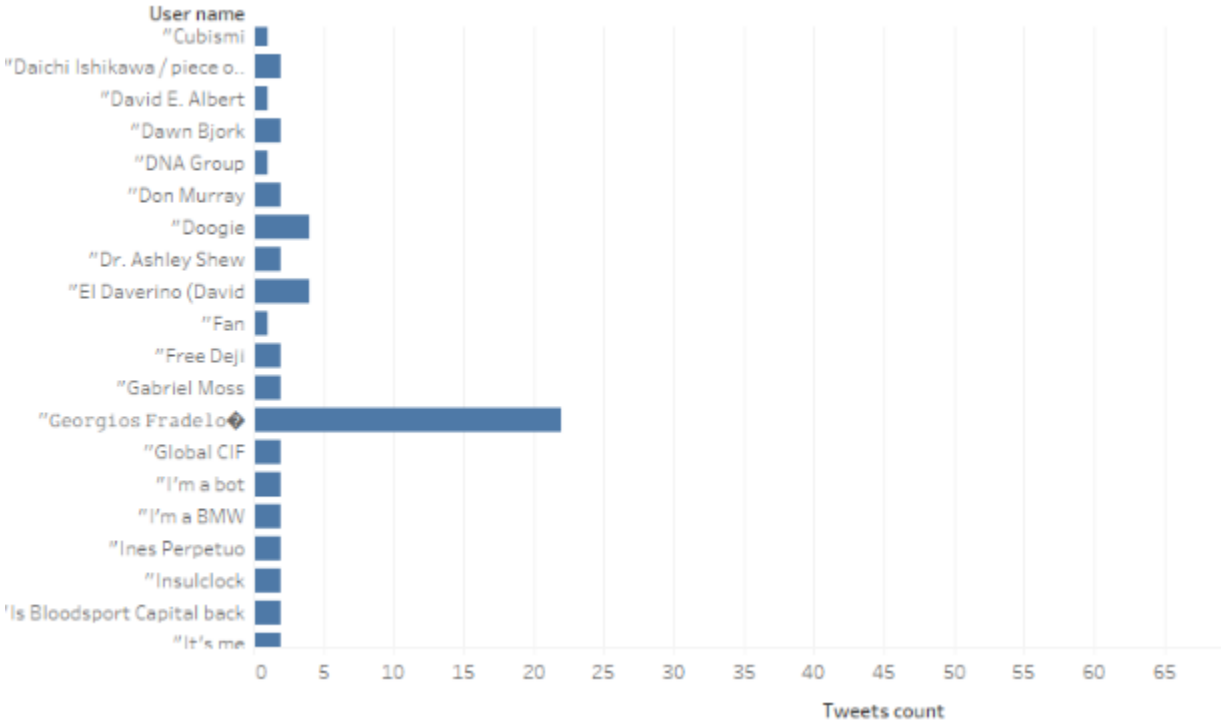
In this query, we found number of tweets made on technologies based on user names.

Code :

```
scala> val Query8 = sqlContext.sql("SELECT count(*) as count, user.name from Technology where user.name is not null group by user.name order by count desc");
Query8: org.apache.spark.sql.DataFrame = [count: bigint, name: string]

scala> Query8.show();
+-----+--------------------------------+
|count|                            name|
+-----+--------------------------------+
|  107|                         Abhishek|
|   85|                  TendenciasTech|
|   82|              SkyPree Marketing|
|   44|             Lauro Espinoza Creel|
|   40|???? ???????????|
|   32|             Money Making Arti...|
|   32|                   startupcrunch|
|   31|            global-tech-news.net|
|   28|             Institute High Tech|
|   28|             ZoeGeop Techonolo...|
|   28|                 QCS Tech Reviews|
|   28|                Podcast Listeners|
|   27|                         technyc|
|   26|                  Business News|
|   25|               The Technology?|
|   22|           ???????? ...|
|   20|                Machine Learning|
|   20|                     Eva Prokop|
|   20|                       Tech Digi|
|   20|                   GoldenHarvest|
+-----+--------------------------------+
only showing top 20 rows
```

Visualization :



Query-9 :

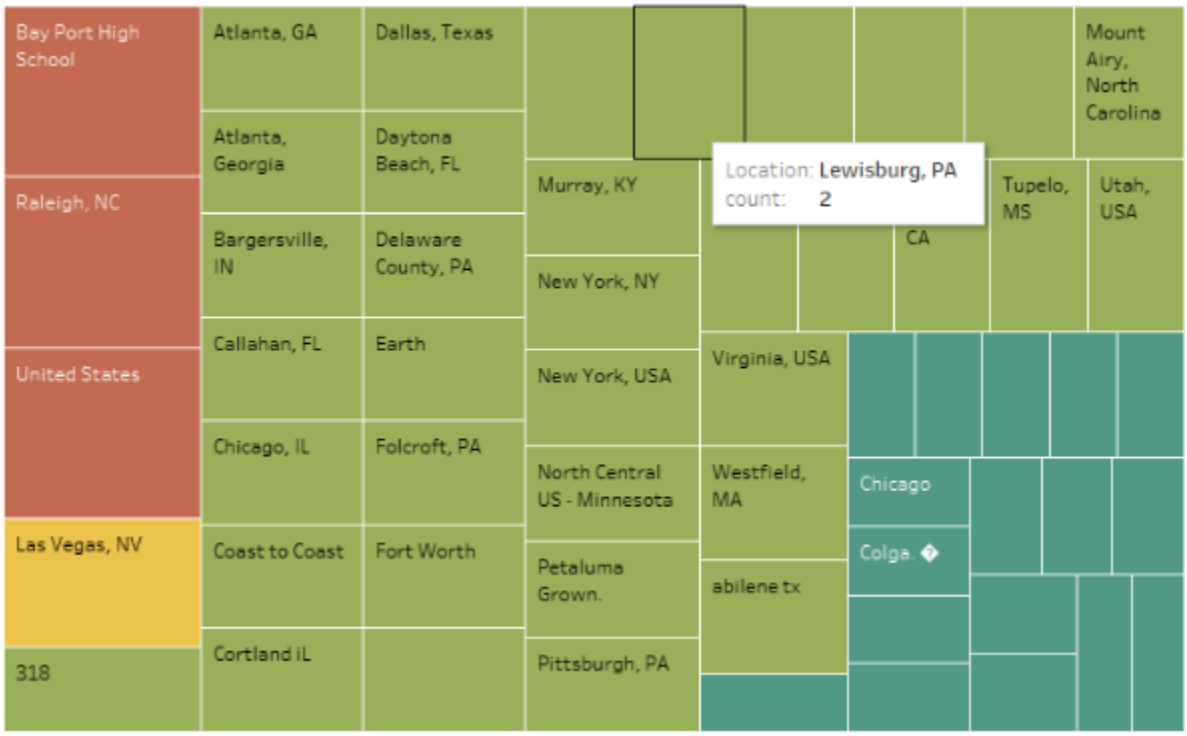
In this query, we found number of tweets made on technologies based on the location.

Code :

```
scala> val Query9=sqlContext.sql("SELECT user.location,count(text) as count FROM technology WHERE place.country='United States' AND user.location is not null GROUP BY user.location ORDER BY coun
Query9: org.apache.spark.sql.DataFrame = [location: string, count: bigint]

scala> Query9.show();
+--------------------+-----+
|            location|count|
+--------------------+-----+
|Bay Port High School|    4|
|       United States|    4|
|         Raleigh, NC|    4|
|       Las Vegas, NV|    3|
| Delaware County, PA|    2|
|Mount Airy, North...|    2|
|        Knoxville, TN|    2|
|North Central US ...|    2|
|     Atlanta, Georgia|    2|
|          Atlanta, GA|    2|
|          Cortland iL|    2|
|    Rancho Mirage, CA|    2|
|               Earth|    2|
|        Callahan, FL|    2|
|           Queens,NY|    2|
|       HARRISONBURG,|    2|
|      Louisiana, USA|    2|
|           Utah, USA|    2|
|    Bargersville, IN|    2|
|       Westfield, MA|    2|
+--------------------+-----+
only showing top 20 rows
```

Visualization :



Query-10 :

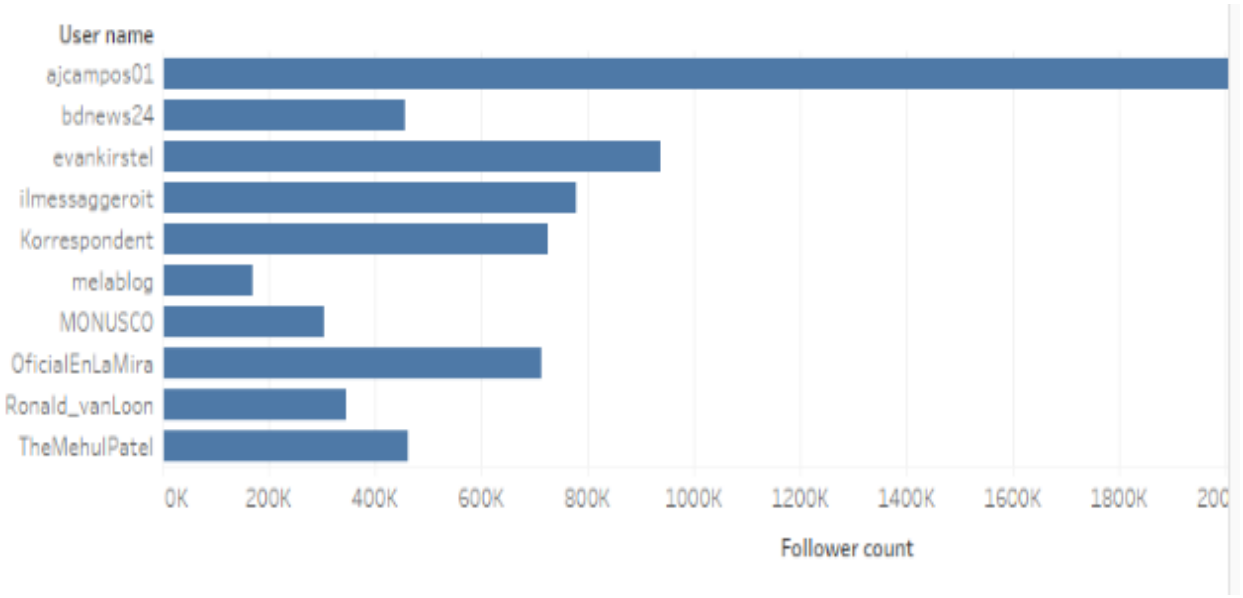
In this query, we found followers count for the non-verified accounts.

Code :

```
scala> val Query10 = sqlContext.sql("SELECT user.screen_name,user.followers_count FROM Technology WHERE user.verified = false ORDER BY user.followers_count DESC LIMIT 2
Query10: org.apache.spark.sql.DataFrame = [screen_name: string, followers_count: bigint]

scala> Query10.show();
+---------------+---------------+
|    screen_name|followers_count|
+---------------+---------------+
|      ajcampos01|        1227793|
|      ajcampos01|        1227793|
|  ilmessaggeroit|         389138|
|  ilmessaggeroit|         389138|
|   Korrespondent|         363049|
|   Korrespondent|         363049|
|OficialEnLaMira |         357310|
|OficialEnLaMira |         357310|
|         MONUSCO|         303123|
|     evankirstel|         234363|
|     evankirstel|         234362|
|     evankirstel|         234362|
|     evankirstel|         234360|
|   TheMehulPatel|         231373|
|   TheMehulPatel|         231373|
|        bdnews24|         227957|
|        bdnews24|         227957|
|  Ronald_vanLoon|         172560|
|  Ronald_vanLoon|         172560|
|         melablog|         171501|
+---------------+---------------+
```

Visualization :



Query-11 :

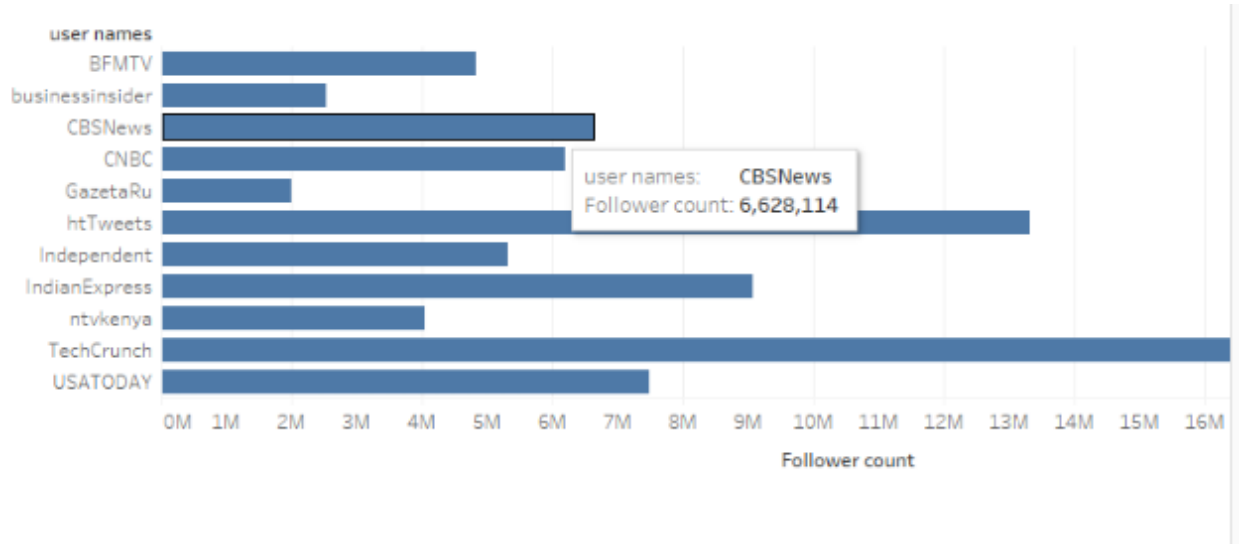
In this query, we found the followers count for verified accounts.

Code :

```
scala> val Query10 = sqlContext.sql("SELECT user.screen_name,user.followers_count FROM Technology WHERE user.verified = true ORDER BY user.followers_count DESC LIMIT 20
Query10: org.apache.spark.sql.DataFrame = [screen_name: string, followers_count: bigint]

scala> Query10.show();
+---------------+---------------+
|    screen_name|followers_count|
+---------------+---------------+
|      TechCrunch|       10031055|
|      TechCrunch|       10031055|
|        htTweets|        6654072|
|        htTweets|        6654072|
|         CBSNews|        6628114|
|        USATODAY|        3741062|
|        USATODAY|        3741062|
|           CNBC|        3093696|
|           CNBC|        3093696|
|  IndianExpress|        3027055|
|  IndianExpress|        3026993|
|  IndianExpress|        3026993|
|    Independent|        2660160|
|    Independent|        2660142|
|businessinsider|        2547083|
|          BFMTV|        2415473|
|          BFMTV|        2415465|
|       ntvkenya|        2027313|
|       ntvkenya|        2027313|
|       GazetaRu|        1995501|
+---------------+---------------+
```

Visualization :



Query-12 :

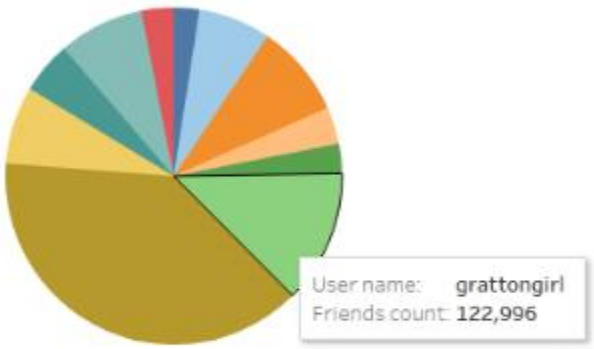
In this query, we found the friends count for verified account

Code :

```
scala> val Query10 = sqlContext.sql("SELECT user.screen_name,user.friends_count FROM Technology WHERE user.verified = true ORDER BY user.friends_count DESC LIMIT 20");
19/05/06 08:12:50 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Query10: org.apache.spark.sql.DataFrame = [screen_name: string, friends_count: bigint]

scala> Query10.show();
+-------------+-------------+
|  screen_name|friends_count|
+-------------+-------------+
|  HerbertRSim|       185836|
|  HerbertRSim|       185836|
|  grattongirl|        41002|
|  grattongirl|        40997|
|  grattongirl|        40997|
|nikkeibpITpro|        39535|
|nikkeibpITpro|        39535|
|Kevin_Jackson|        35872|
|Kevin_Jackson|        35872|
|    frenchweb|        34061|
|    DavidPapp|        33418|
|    DavidPapp|        33418|
|       ZAGrrl|        29795|
|derStandardat|        28705|
|derStandardat|        28705|
|derStandardat|        28705|
|  grattonboy|        27341|
|    Lenovodc|        24062|
|    Lenovodc|        24062|
|    cfarivar|        23557|
+-------------+-------------+
```

Visualization :



Query-13 :

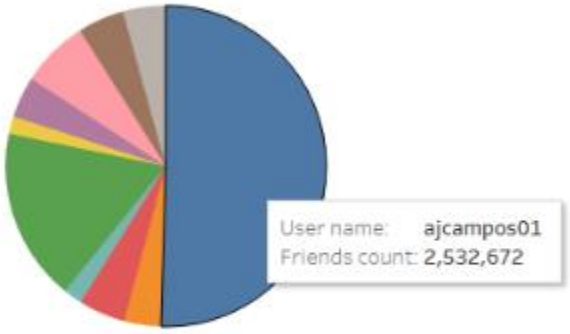
In this query, we found the friends count for the non-verified accounts.

Code :

```
scala> val Query10 = sqlContext.sql("SELECT user.screen_name,user.friends_count FROM Technology WHERE user.verified = false ORDER BY user.friends_count DESC LIMIT 20");
Query10: org.apache.spark.sql.DataFrame = [screen_name: string, friends_count: bigint]

scala> Query10.show();
+--------------+-------------+
|   screen_name|friends_count|
+--------------+-------------+
|     ajcampos01|      1266336|
|     ajcampos01|      1266336|
|    evankirstel|       218615|
|    evankirstel|       218615|
|    evankirstel|       218613|
|    evankirstel|       218613|
|Ronald_vanLoon|       171635|
|Ronald_vanLoon|       171635|
|      BSkylstad|       118228|
|      BSkylstad|       118228|
|stojkovic_alex|       117484|
|stojkovic_alex|       117484|
| The_News_DIVA|       108282|
| The_News_DIVA|       108282|
|    ipfconline1|       103759|
|    ipfconline1|       103759|
|       chrisifg|        96121|
|       bdnews24|        93101|
|       bdnews24|        93101|
| GotStockTips|        88750|
+--------------+-------------+
```

Visualization :



Query-14 :

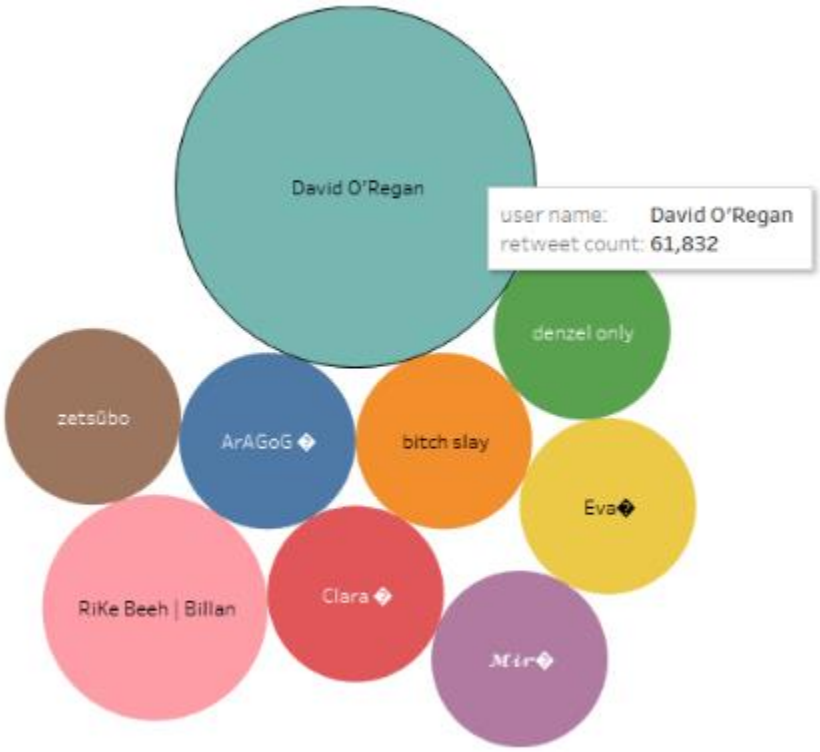
In this query, we the found number of retweets on a particular tweet for a technology based on user.

Code :

```
scala> val Query10 = sqlContext.sql("SELECT user.name ,retweeted_status.retweet_count AS Retweet_Count FROM technology WHERE retweeted_status.retweet_count IS NOT NULL
Query10: org.apache.spark.sql.DataFrame = [name: string, Retweet_Count: bigint]

scala> Query10.show();
+------------------+-------------+
|              name|Retweet_Count|
+------------------+-------------+
|      David O'Regan|        30916|
|      David O'Regan|        30916|
|Rike Beeh | Billan|        24216|
|              Eva?|        14820|
|          ArAGoG ?|        14818|
|            zets?bo|        14817|
|         bitch slay|        14816|
|           Clara ?|        14815|
|        denzel only|        14814|
|              ????|        14813|
+------------------+-------------+
```

Visualization :



Query-15:

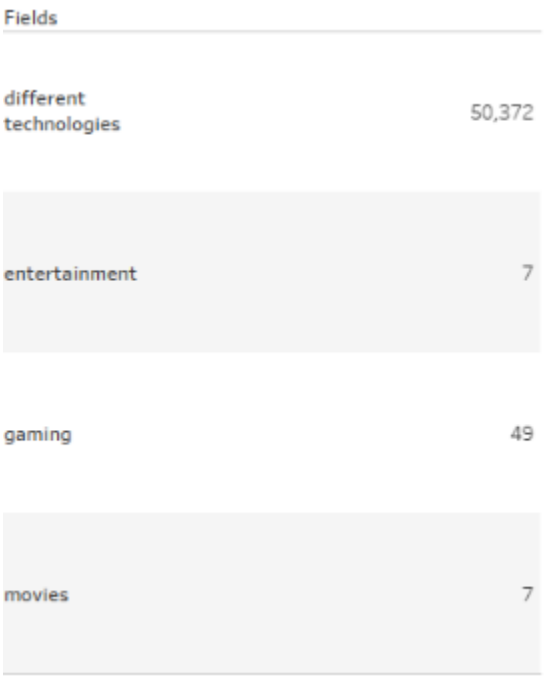
In this query, we found technologies in different fields like gaming, entertainment, movies ets.,

Code :

```
scala> val Query9=sqlContext.sql("select count(*) as count,q.text from (select case when text like '%gaming%' then 'gaming' when text like '%entertainment%' then 'enter
Query9: org.apache.spark.sql.DataFrame = [count: bigint, text: string]

scala> Query9.show();
+-----+--------------------+
|count|                text|
+-----+--------------------+
|    7|              movies|
|    7|       entertainment|
|50372|different technol...|
|   49|              gaming|
+-----+--------------------+
```

Visualization :



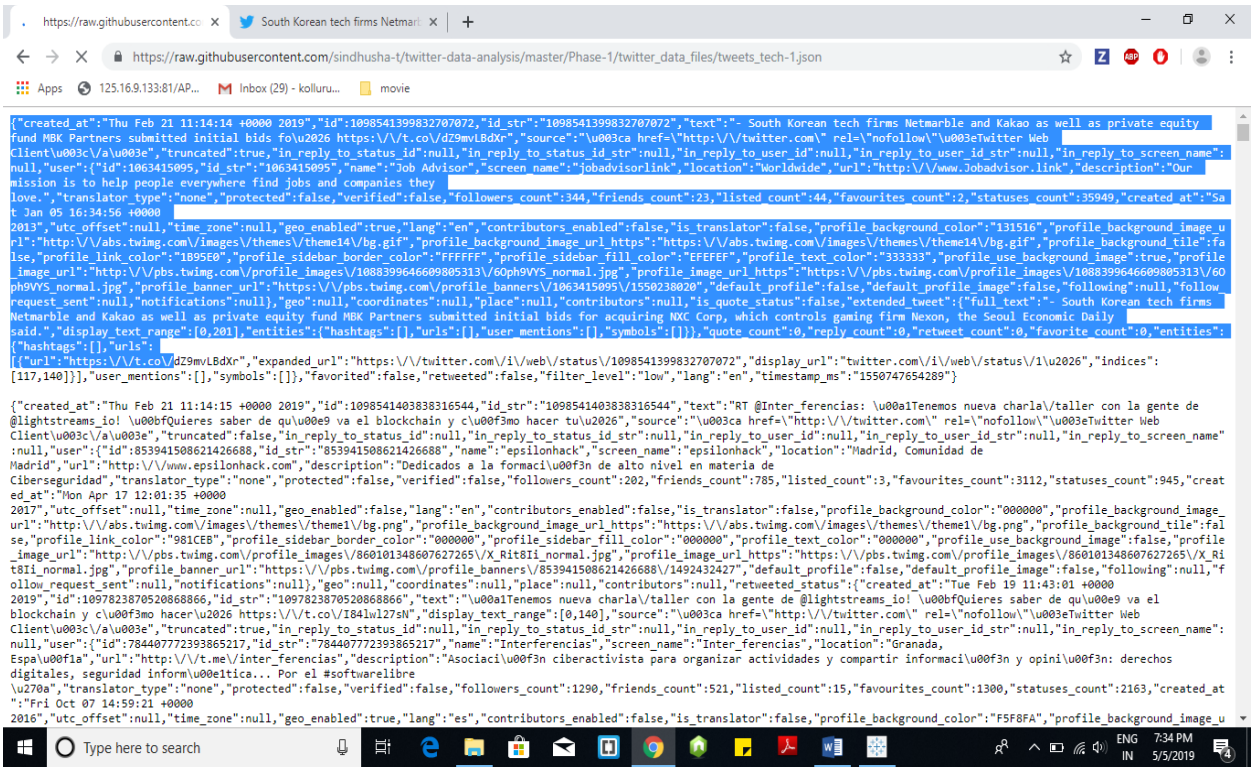
Fields	
different technologies	50,372
entertainment	7
gaming	49
movies	7

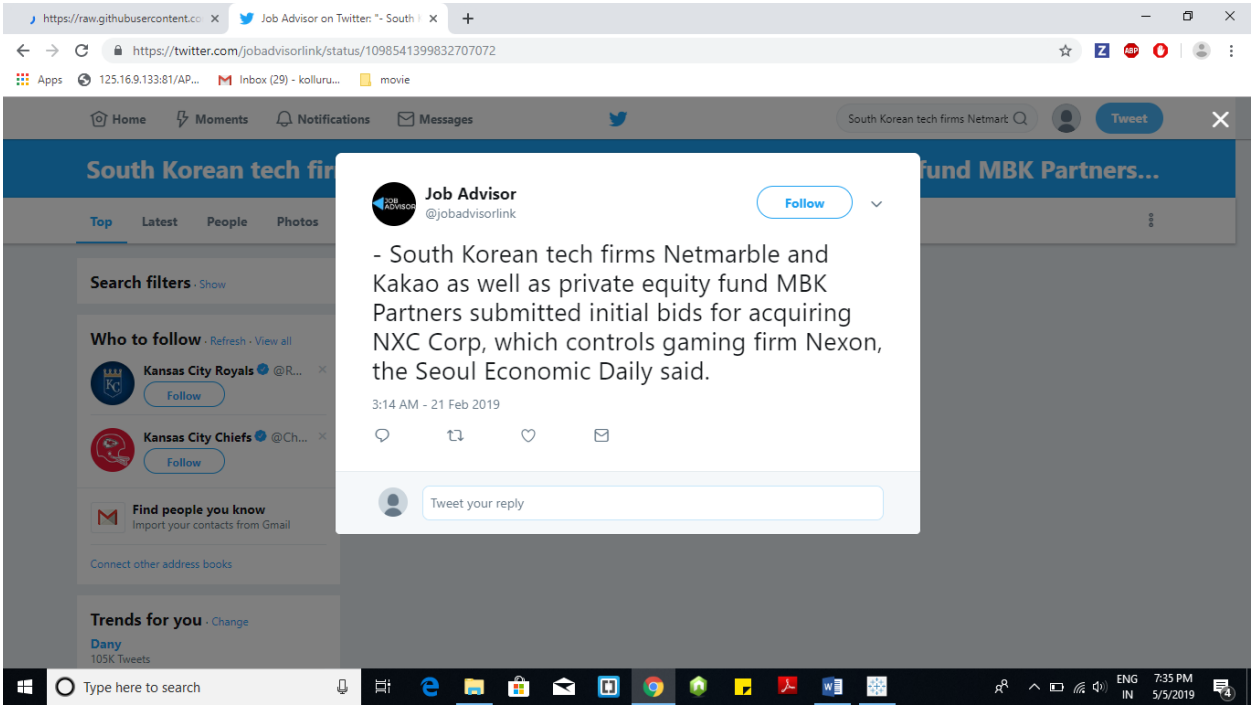
Testing

Manual testing :

We tried to test the results manually by taking the text from the collected tweets and finding the tweets using twitter search engine.

For instance consider the text "*South Korean tech firms Netmarble and Kakao as well as private equity fund MBK Partners submitted initial bids*" which is taken from the collected tweets and searched it manually in twitter to find the tweet.


https://raw.githubusercontent.com/sindhusha-t/twitter-data-analysis/master/Phase-1/twitter_data_files/tweets_tech-1.json
{"created_at":"Thu Feb 21 11:14:14 +0000 2019","id":1098541399832707072,"id_str":"1098541399832707072","text":"- South Korean tech firms Netmarble and Kakao as well as private equity fund MBK Partners submitted initial bids fo\u2026 https:\/\/t.co\/dZ9mvLBdXr","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","truncated":true,"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,"in_reply_to_screen_name":null,"user":{"id":1063415095,"id_str":"1063415095","name":"Job Advisor","screen_name":"jobadvisorlink","location":"Worldwide","url":"http:\/\/www.Jobadvisor.link","description":"Our mission is to help people everywhere find jobs and companies they love.","translator_type":"none","protected":false,"verified":false,"followers_count":344,"friends_count":23,"listed_count":44,"favourites_count":2,"statuses_count":35949,"created_at":"Sat Jan 05 16:34:56 +0000 2013","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"131516","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme14\/bg.gif","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme14\/bg.gif","profile_background_tile":false,"profile_link_color":"1B95E0","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"EFEFEF","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/1088399646609805313\/6Oph9VYS_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/1088399646609805313\/6Oph9VYS_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1063415095\/1550238020","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"extended_tweet":{"full_text":"- South Korean tech firms Netmarble and Kakao as well as private equity fund MBK Partners submitted initial bids for acquiring NXC Corp, which controls gaming firm Nexon, the Seoul Economic Daily said.","display_text_range":[0,201],"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]}},"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/dZ9mvLBdXr","expanded_url":"https:\/\/twitter.com\/i\/web\/status\/1098541399832707072","display_url":"twitter.com\/i\/web\/status\/1\u2026","indices":[117,140]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1550747654289"}
{"created_at":"Thu Feb 21 11:14:15 +0000 2019","id":1098541403838316544,"id_str":"1098541403838316544","text":"RT @Inter_ferencias: \u00a1Tenemos nueva charla\/taller con la gente de @lightstreams_io! \u00bfQuieres saber de qu\u00e9 va el blockchain y c\u00f3mo hacer tu\u2026","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","truncated":false,"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,"in_reply_to_screen_name":null,"user":{"id":853941508621426688,"id_str":"853941508621426688","name":"epsilonhack","screen_name":"epsilonhack","location":"Madrid, Comunidad de Madrid","url":"http:\/\/www.epsilonhack.com","description":"Dedicados a la formaci\u00f3n de alto nivel en materia de Ciberseguridad","translator_type":"none","protected":false,"verified":false,"followers_count":202,"friends_count":785,"listed_count":3,"favourites_count":3112,"statuses_count":945,"created_at":"Mon Apr 17 12:01:35 +0000 2017","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"000000","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"981CEB","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/860101348607627265\/X_Rit8Ii_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/860101348607627265\/X_Rit8Ii_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/853941508621426688\/1492432427","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Tue Feb 19 11:43:01 +0000 2019","id":1097823870520868866,"id_str":"1097823870520868866","text":"\u00a1Tenemos nueva charla\/taller con la gente de @lightstreams_io! \u00bfQuieres saber de qu\u00e9 va el blockchain y c\u00f3mo hacer\u2026 https:\/\/t.co\/I84lwl27sN","display_text_range":[0,140],"source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","truncated":true,"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,"in_reply_to_screen_name":null,"user":{"id":784407772393865217,"id_str":"784407772393865217","name":"Interferencias","screen_name":"Inter_ferencias","location":"Granada, Espa\u00f1a","url":"http:\/\/t.me\/inter_ferencias","description":"Asociaci\u00f3n ciberactivista para organizar actividades y compartir informaci\u00f3n y opini\u00f3n: derechos digitales, seguridad inform\u00e1tica... Por el #softwarelibre \u270a","translator_type":"none","protected":false,"verified":false,"followers_count":1290,"friends_count":521,"listed_count":15,"favourites_count":1300,"statuses_count":2163,"created_at":"Fri Oct 07 14:59:21 +0000 2016","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8FA","profile_background_image_u
Type here to search
ENG IN
7:34 PM 5/5/2019


https://twitter.com/jobadvisorlink/status/1098541399832707072
Home
Moments
Notifications
Messages
South Korean tech firms Netmarb
Tweet
Top
Latest
People
Photos
Search filters · Show
Who to follow · Refresh · View all
Kansas City Royals @R...
Follow
Kansas City Chiefs @Ch...
Follow
Find people you know
Import your contacts from Gmail
Connect other address books
Trends for you · Change
Dany
105K Tweets
Job Advisor
@jobadvisorlink
Follow
- South Korean tech firms Netmarble and Kakao as well as private equity fund MBK Partners submitted initial bids for acquiring NXC Corp, which controls gaming firm Nexon, the Seoul Economic Daily said.
3:14 AM - 21 Feb 2019
Tweet your reply
Type here to search
ENG IN
7:35 PM 5/5/2019