Darrell Nelson
IST 652
05/26/2019

HW02: Semi-structured Data

Data

The dataset is the "bball_mm2" text file that consists of tweets about the 2017 NCAA Tournament affectionately known as March Madness. These tweets were given out in the Asynchronous material in section 8.2.2 of this class. Each tweet will be listed as a json structure. The data will be manipulated using pymongo and pandas dataframes.

Data Questions

Based off of the information contained in this dataset there are 5 pertinent data questions that will help extract valuable information from this data-set:

1. What is the average friend count (reach) of each user that posted?
2. What's the difference between a tweet from the most and least followed user?
3. What time zones are these tweets from?
    a. Which time zone is the most popular?
4. Which time zone gets the most retweets?
5. How many twitter fans have their geo location enabled?

Program Description

In order for this program to run the files "db_fn.py", "db_from_file.py", and "bball_mm2.txt" must all be stored in the working directory of the command prompt being used. First, the mongoDB server must be active. Once that is active, running the following line of code will import the tweets from "bball_mm2" into a file. Then, by running this program in the same working directory with the command line "python Twitter_Code.py" will output the necessary data to answer all 5 of the questions above.

Answers/Output:

Question 01: What is the average friend count (reach) of each user that posted?

Code Output:

Question01:
What is the average friend count of each user that posted?
Average friend count of each user that posted: 1974.9615

Answer: On average each tweeter user had approximately 1,975 followers.

Question 02: What's the difference between a tweet from the most and least followed user??

Code Output:

Question02:
What's the difference between a tweet from the most and least followed user?
Max Follower User Name: John Legere
Max Follower Text: It's been a crazy weekend of basketball. Question is... Do you think I still have a chance to win? #MarchMadness https://t.co/7tyH1MArkk
Number of Followers: 3888114

Min Follower User Name: John Hornbostel
Min Follower Text: RT @marchmadness: ICYMI: The end of North Carolina/Kentucky was incredible. #MarchMadness https://t.co/m1KNEPU6L7
Number of Followers: 0

Answer: It's interesting to note that both of their names are John! Also, the twitter user with the most followers provides an original tweet that references his own March Madness bracket. Providing original content for his followers to enjoy. The user with the least amount of followers regurgitated (retweeted) someone else's opinion about a high profile game. The influential (max followers) user is able to provide exclusive content to their followers and add value to his page. Therefore, it appears that a strong way to increase a user's twitter following is by providing new original content.

Question 03: What time zones are these tweets from? Which time zone is the most popular?

Code Output:

Question03: What time zones are the tweets from?
Time Zone Freq. List:

	0
Eastern Time (US & Canada)	1388
Central Time (US & Canada)	924
NaN	3388
Pacific Time (US & Canada)	1032
Mountain Time (US & Canada)	196
Amsterdam	24
Atlantic Time (Canada)	308
MDT	4
Almaty	4
London	64
Solomon Is.	4
Quito	224
America/Chicago	24
Tehran	4
Mexico City	4

Jerusalem 8
Buenos Aires 24
Alaska 40
Santiago 12
Hawaii 44
Indiana (East) 20
Ljubljana 4
Greenland 4
Madrid 16
Arizona 84
America/New_York 20
Pretoria 4
UTC 4
Paris 16
Brasilia 20
Midway Island 8
Bern 4
Taipei 8
Berlin 4
Africa/Brazzaville 4
CST 4
America/Los_Angeles 8
Dublin 8
New Caledonia 4
EST 4
Baghdad 4
Tokyo 8
Central America 4
Sydney 4
America/Detroit 4
Riyadh 4
Rome 4

Answer: The NCAA tournament has an extremely diverse viewership with tweets about the games coming in from Europe, Asia, Africa, Australia, North and South America. The time zone with the most traffic is "Eastern Time (US & Canada)". This probably explains why many sporting events, not just college basketball, refer to prime time as the time slots where viewership is highest on the eastern seaboard (7 – 10pm EST).

Question 04: Which time zone gets the most retweets?

Code Output:

Question04: Which time zone gets the most retweets?
Time zone by retweet count:
time_zone
Africa/Brazzaville 12

Alaska 79244
Almaty 53972
America/Chicago 209824
America/Detroit 38964
America/Los_Angeles 156
America/New_York 39216
Amsterdam 79804
Arizona 281992
Atlantic Time (Canada) 1468048
Baghdad 38956
Berlin 17560
Bern 1772
Brasilia 2400
Buenos Aires 1904
CST 38956
Central America 4
Central Time (US & Canada) 4171800
Dublin 77916
EST 284
Eastern Time (US & Canada) 4744352
Greenland 392
Hawaii 342652
Indiana (East) 9656
Jerusalem 16
Ljubljana 4
London 174244
MDT 38956
Madrid 1228
Mexico City 0
Midway Island 108
Mountain Time (US & Canada) 620528
New Caledonia 16
Pacific Time (US & Canada) 2613532
Paris 39736
Pretoria 0
Quito 1000912
Riyadh 0
Rome 0
Santiago 2396
Solomon Is. 1772
Sydney 38960
Taipei 36
Tehran 24
Tokyo 3544
UTC 8
Name: retweet, dtype: int64

Answer: Unsurprisingly, the time zone where the games play in primetime has the most retweets.

Question 05: How many twitter fans have their geo location enabled?

Code Output:

Question05: How many twitter users have their geo location on?
There are 4176 users out of 8000 that have their location enabled
52.2 Percent of people want their location tracked.

Answer: As the age of social media reigns supreme. There has been increased concern on the amount of privacy users are unknowingly forfeiting to these tech-giants. Geo-tracking is one feature that many users are unaware that they are providing when making a tweet. Many times this information is sold to 3rd parties with the actual creators of the data (twitter users) unaware that they consented to such just by signing up. The 52.2% of twitter users that have geo location enabled reflects a growing concern over this issue.

Appendix

Raw Output

```
(base) C:\Users\Darrell\Desktop\Syracuse\Spring_19\IST_652_Scripting\HW02>python Twitter_Code.py
Twitter_Code.py:14: DeprecationWarning: database_names is deprecated. Use list_database_names instead.
  client.database_names()
Twitter_Code.py:17: DeprecationWarning: collection_names is deprecated. Use list_collection_names instead.
  db.collection_names()
Number of tweets imported:  8000

Question01:
What is the average friend count of each user that posted?
Average friend count of each user that posted:  1974.9615

Question02:
What's the difference between a tweet from the most and least followed user?
Max Follower User Name: John Legere
Max Follower Text: It's been a crazy weekend of basketball. Question is... Do you think I still have a chance to win? #MarchMadness https://t.co/7tyH1MArkk
Number of Followers:  3888114

Min Follower User Name: John Hornbostel
Min Follower Text: RT @marchmadness: ICYMI: The end of North Carolina/Kentucky was incredible. #MarchMadness https://t.co/m1KNEPU6L7
Number of Followers:  0

Question03: What time zones are the tweets from?
Time Zone Freq. List:
                              0
Eastern Time (US & Canada)   1388
Central Time (US & Canada)    924
NaN                          3388
Pacific Time (US & Canada)   1032
Mountain Time (US & Canada)   196
Amsterdam                      24
Atlantic Time (Canada)        308
MDT                             4
Almaty                          4
London                         64
Solomon Is.                     4
Quito                         224
America/Chicago                24
Tehran                          4
Mexico City                     4
```

Jerusalem 8
Buenos Aires 24
Alaska 40
Santiago 12
Hawaii 44
Indiana (East) 20
Ljubljana 4
Greenland 4
Madrid 16
Arizona 84
America/New_York 20
Pretoria 4
UTC 4
Paris 16
Brasilia 20
Midway Island 8
Bern 4
Taipei 8
Berlin 4
Africa/Brazzaville 4
CST 4
America/Los_Angeles 8
Dublin 8
New Caledonia 4
EST 4
Baghdad 4
Tokyo 8
Central America 4
Sydney 4
America/Detroit 4
Riyadh 4
Rome 4

Question04: Which time zone gets the most retweets?
Time zone by retweet count:
time_zone
Africa/Brazzaville 12
Alaska 79244
Almaty 53972
America/Chicago 209824
America/Detroit 38964
America/Los_Angeles 156
America/New_York 39216
Amsterdam 79804
Arizona 281992
Atlantic Time (Canada) 1468048
Baghdad 38956
Berlin 17560

Bern 1772
Brasilia 2400
Buenos Aires 1904
CST 38956
Central America 4
Central Time (US & Canada) 4171800
Dublin 77916
EST 284
Eastern Time (US & Canada) 4744352
Greenland 392
Hawaii 342652
Indiana (East) 9656
Jerusalem 16
Ljubljana 4
London 174244
MDT 38956
Madrid 1228
Mexico City 0
Midway Island 108
Mountain Time (US & Canada) 620528
New Caledonia 16
Pacific Time (US & Canada) 2613532
Paris 39736
Pretoria 0
Quito 1000912
Riyadh 0
Rome 0
Santiago 2396
Solomon Is. 1772
Sydney 38960
Taipei 36
Tehran 24
Tokyo 3544
UTC 8
Name: retweet, dtype: int64

Question05: How many twitter users have their geo location on?
There are 4176 users out of 8000 that have their location enabled
52.2 Percent of people want their location tracked.

(base) C:\Users\Darrell\Desktop\Syracuse\Spring_19\IST_652_Scripting\HW02>