Project 1
Data Mining On Twitter Data

By Ankit Chachada, CWID: 802973818

Programming environment, libraries, and tools:

Programming Language: ruby 2.2.1p85

Operating System: Ubuntu

Text Editor: Sublime Text Editor, Ubuntu Text Editor

Data obtained : twitter.com

(https://stream.twitter.com/1.1/statuses/sample.json)

Libraries used: Standard Ruby libraries used such as JSON,Array, Hash, String

Procedures used to obtain data:

1. Application was created on twitter https://apps.twitter.com/

2. Data was downloaded from

https://dev.twitter.com/streaming/reference/get/statuses/sample using the same

application created in step1.

3. Obtained oauth signature from

https://dev.twitter.com/oauth/tools/signature-generator/<id> and used curl command

provided on that page to download the data. The following is curl command

```
curl --get 'https://stream.twitter.com/1.1/statuses/sample.json' --data
'language=en' --header 'Authorization: OAuth
oauth_consumer_key=<CONSUMER_KEY>, oauth_nonce=<OAUTH>,
oauth_signature=<OAUTH_SIGNATURE>, oauth_signature_method="HMAC-SHA1",
oauth_timestamp=<TIMESTAMP>, oauth_token=<OAUTH_TOKEN>, oauth_version="1.0"'
--verbose -o tweets.json.
```

The data I downloaded twice one without passing language parameter and one with language parameter set to English(en).

Part 1: Data Gathering

How large is the file?

Ans: The file size is 150 MB(english tweets)

The file size is 139 MB(generalize tweets)

Part 2: Data Counting

How many objects are in the file?

Ans: Total Objects are 33740(english tweets)

Total Objects are 36241(generalize tweets)

How many are tweets?

Ans: Total Number of tweets 33739(english tweets)

Total Number of tweets 31501(generalize tweets)

How many of the tweets have [hashtags](#)?

Ans: 7459 of tweets has hashtags (english tweets)

7024 of tweets has hashtags (generalize tweets)

Part 3: Data Analysis

What are the 10 most popular hashtags?

Ans: The 10 most popular hashtags for english tweets are

[["KCA", 668], ["VoteArianaGrande", 284], ["VoteEnriqueFPP", 109], ["GIFparty", 107], ["ALDUBRegaloNiLola", 95], ["ShotsComedyApp", 89], ["WorldBookDay", 73], ["iHeartAwards", 58], ["gameinsight", 47], ["TBT", 46]]
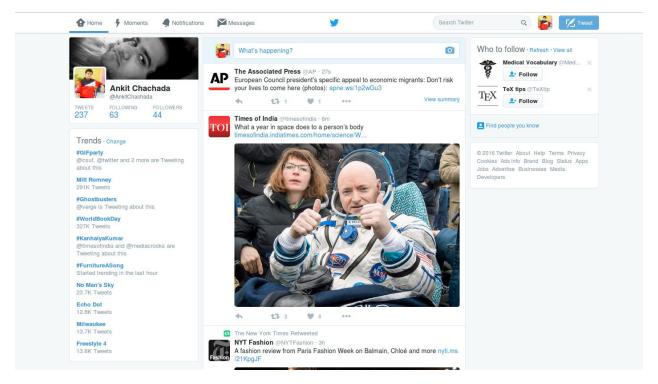
The 10 most popular hashtags for generalize tweets are

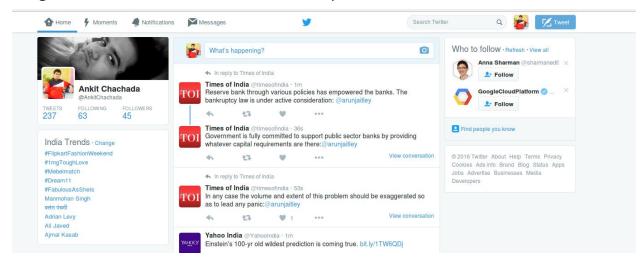[["KCA", 1055], ["VoteMaineFPP", 537], ["VoteKathrynFPP", 331], ["ベガッ太Ｊマスコット総選挙", 187], ["SoulAwakeningMSG", 91], ["ベガッ太Jマスコット総選挙", 76], ["

サンチェＪマスコット総選挙", 71], ["VoteJamesFPP", 59], ["사설토토사이트추천", 53], ["東京ドロンパＪマスコット総選挙", 52]]

How does your list compare with the trending hashtags shown on Twitter? **GIFparty, WorldBookDay** were present in my homepage. (English tweets)

For generalized tweets **none** of them were present.

Information about the code:

The code folder consists of three ruby files i.e. twitter_assignment.rb and sentiment.rb and main.rb and AFINN-111.txt.

main.rb:

This runs the program it calls twitter_assignment.rb and sentiment.rb

twitter_assigment.rb:

This files takes the json file as input and perform counting calculations and display the number of tweets, hashtags and popular hashtags. Also this file is responsible to extract tweets with popular hashtags and generate a file named as tweets_with_top_hashtags, and it is used in deriving sentiments.

sentiment.rb:

It takes the AFINN-111.txt and tweets_with_top_hashtags as input and perform sentiment analysis and generates a new file emotions which displays the final values of tweets with their sentiment and sentiment value.

Running Instructions:

1. Please install ruby 2.2.1p85
2. Open twitter_assignment.rb, change the file name in initialize_file method with your file.
3. Open terminal and go to the folder location and run ruby main.rb.
4. The output of sentiments is generated in emotions file.

Execution Screenshot:

The following screenshot displays the execution of the program. It was run twice one for english tweets and one for generalize tweets. They are separated by white bar.

The following screenshot displays the final output. There are three columns tweet, emotion and sentiment value. For example consider second line, first is tweet, followed by sentiment i.e. POSITIVE and followed by value i.e. 2

Future Scope & Limitations

1. Currently the look and feel of this application is not good. So in future we can implement it.
2. Also the code needs to be refactor. Currently the execution time is very large so in future code will be refactored and optimized.

References

1. https://dev.twitter.com/streaming/public
2. http://ruby-doc.org/core-2.2.2
3. https://apps.twitter.com/
4. https://www.sublimetext.com/
5. http://ruby-doc.org/stdlib-2.0.0/libdoc/json/rdoc/JSON.html
6. http://rubular.com/