

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)

The screenshot shows a Twitter homepage for user Ankit Chachada. The left sidebar displays trending hashtags: #GIFparty, #WorldBookDay, #KanhalyaKumar, #FurnitureASong, No Man's Sky, Echo Dot, Milwaukee, and Freestyle 4. The main feed features a tweet from The Associated Press (@AP) about a European Council president's appeal to economic migrants, accompanied by a photo of a man in a space suit. The right sidebar shows 'Who to follow' suggestions including Medical Vocabulary and TeX tips.

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

The screenshot shows a Twitter homepage for user Ankit Chachada. The left sidebar displays trending hashtags: #FlipkartFashionWeekend, #1mgToughLove, #MebeImatch, #Dream11, #FabulousAsShels, Manmohan Singh, Adrian Levy, Ali Javed, and Ajmal Kasab. The main feed features a tweet from Times of India (@timesofindia) about the Reserve Bank of India's policies, accompanied by a photo of a man in a space suit. The right sidebar shows 'Who to follow' suggestions including Anna Sharman and GoogleCloudPlatform.

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_assignment.rb:

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

sentiment.rb:

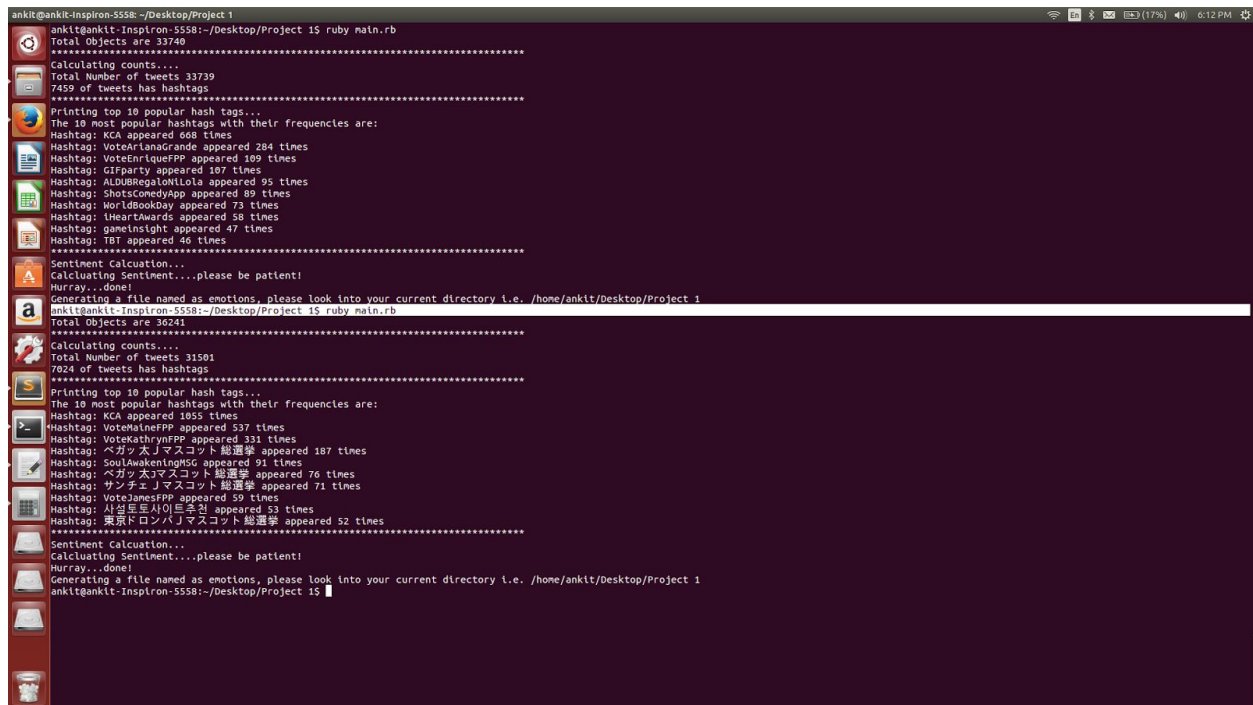
It takes the `AFINN-111.txt` and `tweets_with_top_hashtags` as input and performs sentiment analysis and generates a new file `emotions` which displays the final values of tweets with their sentiment and sentiment value.

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.



```
ankit@ankit-Inspiron-5558: ~/Desktop/Project 1
ankit@ankit-Inspiron-5558:~/Desktop/Project 1$ ruby main.rb
Total Objects are 33740
Calculating counts...
Total Number of tweets 33739
7459 of tweets has hashtags
Printing top 10 popular hash tags...
The 10 most popular hashtags with their frequencies are:
Hashtag: KCA appeared 668 times
Hashtag: VoteArianaGrande appeared 284 times
Hashtag: VoteEnriquePP appeared 109 times
Hashtag: GIFparty appeared 107 times
Hashtag: ALDUBRegaloNiLola appeared 95 times
Hashtag: ShotsComedyApp appeared 89 times
Hashtag: WorldBookDay appeared 73 times
Hashtag: Heartwards appeared 58 times
Hashtag: gameinsight appeared 47 times
Hashtag: TBT appeared 46 times
Sentiment Calculation...
Calculating Sentiment....please be patient!
Hurray...done!
Generating a file named as emotions, please look into your current directory i.e. /home/ankit/Desktop/Project 1
ankit@ankit-Inspiron-5558:~/Desktop/Project 1$ ruby main.rb
Total Objects are 36241
Calculating counts...
Total Number of tweets 31501
7024 of tweets has hashtags
Printing top 10 popular hash tags...
The 10 most popular hashtags with their frequencies are:
Hashtag: KCA appeared 1055 times
Hashtag: VoteKainePP appeared 537 times
Hashtag: VoteKathrynPP appeared 331 times
Hashtag: ベガッスJマススコット総選挙 appeared 187 times
Hashtag: SoulAwakeningMSG appeared 91 times
Hashtag: ベガッスJマススコット総選挙 appeared 76 times
Hashtag: サンチェJマススコット総選挙 appeared 71 times
Hashtag: VoteJamesPP appeared 59 times
Hashtag: 사설토도사이드주전 appeared 53 times
Hashtag: 東京トロンハJマススコット総選挙 appeared 52 times
Sentiment Calculation...
Calculating Sentiment....please be patient!
Hurray...done!
Generating a file named as emotions, please look into your current directory i.e. /home/ankit/Desktop/Project 1
ankit@ankit-Inspiron-5558:~/Desktop/Project 1$
```

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

```
Tweet:Emotion:SentimentValue
RT @TwitterBE: Thanks for joining us today at #SportsYear2016 event. #Score! #GIFparty https://t.co/W10pEr3Ut https://t.co/33MAf5Bz8f | POSITIVE | 2
RT @MIAngelEsLal!: Queen #KCA #VotaLaliEsposito https://t.co/wZE0AGgl9f | NEUTRAL | 0
RT @JHonlineTV: ★ LIVE! STREAM ★
https://t.co/zAmCSFX9g0
https://t.co/uqQkieFUYP
https://t.co/KZ0oYUBXjo
#ALDUBRegaloNiLola https://t.co/... | NEUTRAL | 0
RT @MaichardOFC2016: Dean's snapchat update #ALDUBRegaloNiLola https://t.co/Vy4sRXjV0e | NEUTRAL | 0
I can't stop! #GIFparty https://t.co/MUOzu8dbm0 | NEGATIVE | -1
RT @JaDineLand: @TeamkabebeBaby hi team kabebe! Were the second acc of JaDine Land, can you give us a shout out? Thank you! 🙏
#VoteJamesPP | POSITIVE | 2
RT @rics11600: GOOD MORNING!! #ALDUBRegaloNiLola
@huggable_me @izaoreal @MarianEBrown77 @Ysa_2kD @zshyrel21 @lhenmerencillo @JurenJoren @Ch... | POSITIVE | 3
RT @MFCATA 34.13
```

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/>