Project #1

by
Dipika Mahashabde
(Cwid:803004274)

CPSC 473-02
Professor: Kenytt Avery

Department of Computer Science

California State University, Fullerton
Spring 2016

A.

1. PROGRAMMING ENVIRONMENT

- IDE: Visual Studio Code
- PLATFORM: Linux
- COMPILER: Python

2. LIBRARIES

- Json: Json library is used because we have our twitter data stored in a json file. That is to parse the json file we import the library also to use the functions of json in python.
- Collections: collection library is used because we have to sort the popular Hash-tags we have got from the data. So by using the functions of collections we have sorted the dictionary of Hash-tags and its score.
- PPrint: It keeps objects on a single line if it can, and breaks them onto multiple lines if they don't fit within the allowed width. Construct "pretty – print" objects explicitly if you need to adjust the width constraint.
- urllib2: The urllib2module defines functions and classes which help in opening URLs (mostly HTTP) in a complex world such as basic and digest authentication, redirections, cookies and more.
- Oauth2: This lib is used for downloading the objects from the twitter. The following kinds of messages may be delivered on a stream like Blank lines, Delete etc.

3. TOOL

- Visual Studio Code.

B.
PROCEDURE

1. Gathering of Data:

The questions we need to ask our self before gathering the data is, what type of data we want? How to access the data?, how much data we want? Many more.

- Step 1: (Understanding)

As Twitter has security we cannot access the data of any one therefore we make use of "Public Streams" which is to take the public data flowing in twitter.

- Step 2: As we have mentioned in the above part that to access the data we need Oauth to access the API, it becomes difficult to access, so for now we take help from third party tool or library to access it.
  - Step a. Create an account of Twitter in order to access the data.
  - Step b. Go to **https://dev.twitter.com/apps** and log in with your twitter login Id and password. Click "Create New App"
  - Step c. Fill out the form and agree to the terms. Put in a dummy website if you don't have one you want to use. Like http://www.google.com.
  - Step d. On the next page, click the "API Keys", Get your access to the tokens
  - Step e. Click the button "Create My Access Token".
  - Step f. Download twitter.py form GitHub. Now copy the Different types of key mentioned in that py file.
  - Step g. run this command line code: Run for 15 mins

    $ python twitter.py > tweet.json

    This will give you a file containing the data in format of a json file.

    After this take a screenshot of the twitter page where you get the most popular trending hashtags.

2. Exploring the Data:

Json file is in format of key, value pair similar to the dictionary in python.
Study the json file how it is formatted and how it is stored by taking one object form the json file

Step 1: First question,

To see how many objects are there in the json file? We access the file, line by line as it stores one object on one line. Thus, we take the count of no_of_line in the json file as no_of_objects.

Step 2: Second question,

To count how many tweets are present? To check if there are tweets I have taken the first parameter of the tweet to check if it is present or no ('created_at') that will give me the no of tweets.

Step 3: Third question,

To count how many tweets have Hash-tags present? The hash-tag details are inside entities so we traverse till the entities-> hashtags and check if it has text in it, the details of the hashtags are stored in an array format so now we check if the text part is empty or no if empty ignore otherwise increment the count of no of tweets having hashtags.

With this itself we calculate how many times a particular hashtags has occurred in our tweets. Thus, keeping the records of all the hashtags no of times it has occurred.

3. Analyzing the Data:

Step 4: Fourth question,

How to check top most popular hashtags? From the above data we have got the no of times the hashtags has occurred, now we sort the data in descending order taking the first 10 hashtags which will give the most popular hashtags.

Step 5: Comparison,

Compare your answer with the screenshot taken check if you have got some trends same. And give the reason for it.

4. Sentiment Analyze:

Step 6: what is sentiment analyze? What is the sentiment of a hashtag?

Sentiment analyze checks if the sentence is positive or negative. So now first we will need is a file containing words with some sentiments to it to map our words to the sentiments. We take a file named AFINN-96 to access the sentiments of words. From this we store the sentiments of each word of a tweet and add it up to sentiment of that tweet and map that sentiment to the Hash-tag present in that tweet. Thus we get many sentiments to one hashtag. We take the average of these sentiments as the sentiment of the hashtag.

Example:

Tweet 1 text: Sachin is the God of Cricket. #Sachin #cricket
Tweet 2 text: Cricket is amazing #cricket.
First we map the sentiment value of each word in the text excluding the hashtag add it up.
Secondly, we assign that sentiment to the hashtag, now we see that #cricket will have two sentiments. We add them up and take the average of it. Thus cricket will get its own sentiment.

5. Conclusion:

Step 7: Most Positive and Negative Hashtags of popular ones

The first 10 popular hashtags will also have sentiments assigned to it we check the most positive sentiment of them and the most negative sentiment of it and display it.

Thus we conclude that the one of the popular hashtag is a positive hashtag or a negative hashtag

6. HOW TO RUN MY CODE

Step 1: Go to the command prompt
Step 2: Take file path where you have saved Python file
Step 3: type this in the command prompt:-
>>>python Project1.py
>>>Enter file name “TYPE THE JSON FILE NAME WHRER YOU HAVE TAKEN THE TWITTER OBJECTS”
>>>just Enter
Step 4: In some time the output will be shown.

C.

PART 1: GATHERING OF DATA

Question 1: How large is the file?
Answer 1: My file size is 170.05 MB

PART 2: EXPLORING THE DATA

Question 2: How many objects in the file?
Answer 2: 46365

Question 3: How many tweet?
Answer 3: 40640

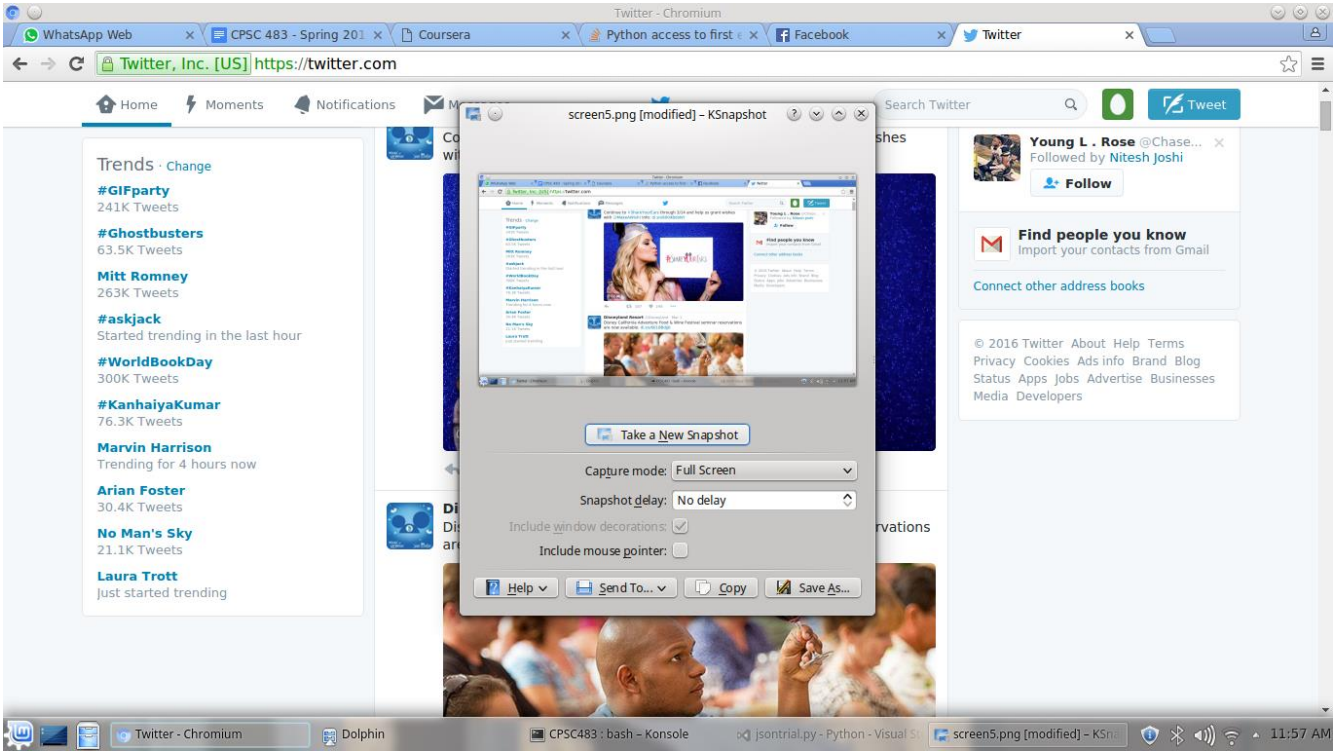
Question 4: How many tweets have Hash-tags?
Answer 4: 8309

PART 3: ANALYZING THE DATA

Question 5: What are the popular Hash-tags ?
Answer 5: Popular Hash-tags are:
KCA 435
GIFparty 105
gameinsight 78
iHeartAwards 77
KanhaiyaKumar 66
androidgames 57
KyState_RYS16 54
BestFanArmy 53
android 52
PorTuSonrisaDamosTodoGravano 52

Question 6: How do your list compare with the trending Hash-tags shown on Twitter?
Answer 6:

1. Screen-Shot of twitter trends:-



2. Trends found from my code:
KCA 435
GIFparty 105
gameinsight 78
iHeartAwards 77
KanhaiyaKumar 66
androidgames 57
KyState_RYS16 54
BestFanArmy 53
android 52
PorTuSonrisaDamosTodoGravano 52

The highlighted ones in the 2nd point are the trends found in the twitter trends also.
As we can see that only two trends could be found in the code done by me the reason for this is that we have taken data of only 15 minutes so we cannot expect that we will get all the popular trends.

PART 4: SENTIMENT ANALYZING

Question 6: What are average sentiments of each Hash-tags?
Answer 6: As I have 8309 Hash-tags it is not feasible for me to show all the average sentiments of all the Hash-tags. So for now we have taken sentiments of popular Hash-tags only, they are :-
{'BestFanArmy': 0.22641509433962265,
'GIFparty': 0.19047619047619047,

'KCA': 0.27586206896551724,
'KanhaiyaKumar': 0.16666666666666666,
'KyState_RYS16': -0.14814814814814814,
'PorTuSonrisaDamosTodoGravano': -0.057692307692307696,
'android': 0.019230769230769232,
'androidgames': -0.10526315789473684,
'gameinsight': -0.05128205128205128,
'iHeartAwards': 0.19480519480519481}

Question 7: What are the most positive and most negative sentiment of popular Hash-tags?
Answer 7: Positive sentiment Hash-tag means the sentiment having the highest value of sentiment is positive Hash-tag:-
KCA is: 0.275862068966

Negative sentiment Hash-tag means the sentiment having the lowest value of sentiment is negative Hash-tag :-
KyState_RYS16 is: -0.148148148148

REFERENCES

1. https://docs.google.com/document/d/1WtMXbJQ_sMz_u9-PVvak9SzJEFUSFuunKYWjee_w86c/edit#
2. https://class.coursera.org/datasci-002/assignment/view?assignment_id=3
3. https://dev.twitter.com/
4. https://github.com/wooorm/afinn-96
5. https://docs.python.org/2/library/pprint.html