International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SCRAPING OF SOCIAL MEDIA DATA USING PYTHON-3 AND PERFORMING DATA ANALYTICS USING MICROSOFT POWER BI

Prashant Dutta*[1] & Aushmin Lodh[2]
*[1]MPPKVVCL, Jabalpur
[2]Vivavideo, Gurgaon

DOI: https://doi.org/10.29121/ijesrt.v9.i7.2020.8

ABSTRACT

The manifestation of humanity is driven by fulfillment of desires. These desires are satiated by the society and its resources. But after the advent of social media the societal boundaries have shrunken but desires haven't, hence the desires are now fulfilled through social media. The aforementioned phenomenon was recognized by the business plutocrats very early and have started to satisfy human desires using social media as a tool. But before satisfying the desires, the businesses needs to identify the specific desires of an individual. The identification of specific desires/needs will help the marketing agencies to develop user specific marketing strategies. These desires are explicitly available through the expressions of sentiments in the social media. The sentiment analysis can provide an insight to the desires of an individual. These patterns and insights helps the businesses to market their product to the right person. The sentiments and expressions can be captured using the scraping technique. The aforesaid points highlight's the course of study followed by this paper and it is to perform data analytics of the social media data scraped using python.

KEYWORDS: Python, Facebook, Twitter, Data Mining, Marketing, Power BI, Social Media, Sentiment analysis.

1. INTRODUCTION

As phrased by many economist "the data is the new oil" has real significance in the digital economy of the current era. There was a time in late 1800s when oil wasn't tapped efficiently and its real usage was unrecognized. Same stands right for "data" in the 21st century. The data is still untapped and mostly unrecognized. The world generates 2.5 quintillion bytes of data each day and it's surprising to note that- 'of the total data generated till date, 90 percent was generated in the last 2 years only'.

The social media has seen an exponential boom of data growth. Every day millions of users upload their photos, videos in the various social media platforms. Other platforms like 'youtube' and 'tik-tok' has brought a new trend of mammoth data with each file ranging from the size of 500kb to 128GB. Some insights on the number of users and size of data processed per minute/day/month for some prominent social media platforms are as under:-

S. No	Social Media Platform	Total Number of Users	Data Processed Each minute/day/month
1	Facebook	2.5+ Billion	500+ TB
2	Twitter	350+ Million	700+ million tweets per day
3	Youtube	2+ Billion	450+ hours of new video every minute
4	Instagram	1.0+ Billion	100+ million photos and videos everyday
5	Snapchat	350+ Million	4+ billion snaps each day
6	Tik-Tok	800+ Million	1+ billion video views per day
7	Whatsapp	2+ Billion	65+ billion 'WhatsApp' messages sent in a day, 2+ billion minutes of WhatsApp voice/video calls in a day

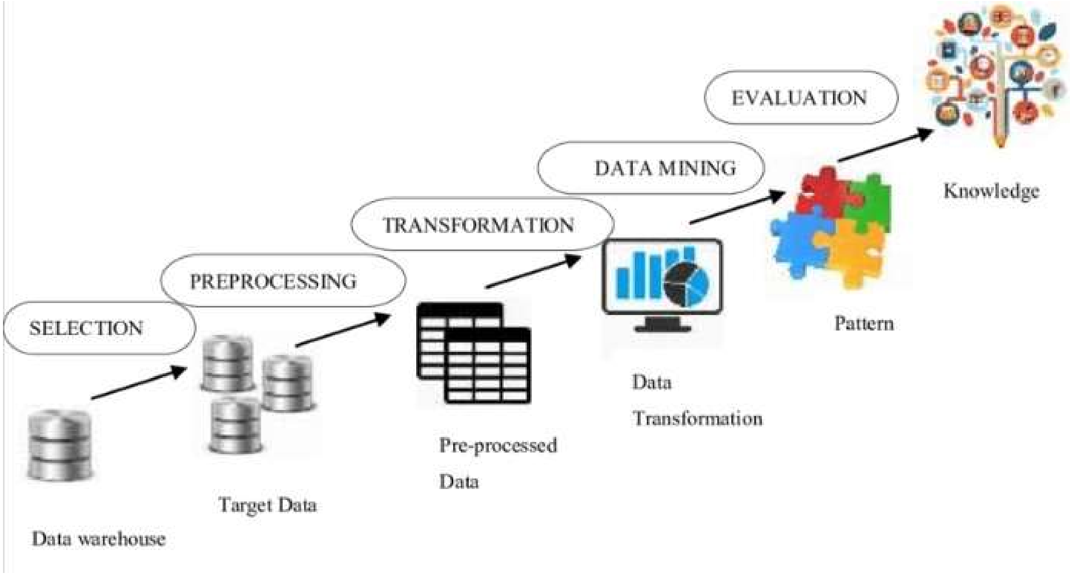
8	WeChat	1.0+ Billion	410+ million audio/video calls in a day, 46TB of data utilized in one minute during peak-hours
9	LinkedIn	690+ Million	280+ billion 'feed-updates' are observed in a year
10	Pinterest	350+ Million	2+ billion searches in a month

Statistics suggests that 250+ million users are joining the social media every year. The No of Tweets has increased to 5 Lakhs tweets per minute from 3 lakhs tweets per minute in 5 years. The nos of video uploaded in youtube has increased 3 times since 2014. On an average an user spends nearly one hour in facebook daily and the number of posts in facebook has increased by 25% in last 2 years. Almost 3.5+ Billion Google-searches are done across the internet every minute i.e 2+ trillion searches in a year over the internet i.e around 40,000+ 'queries' per-second. The above data showcases the fact that "Social-Media-Data" is growing exponentially and will continue to grow in an unprecedented rate. The "Social-Media-Data" has a significant information, which if extracted correctly, it can benefit any company or organization. That information is the user-sentiment/user-opinion/user-emotions. If any organization knows what an user wants, then it can market the right product to the right user. Further, Facebook & whatsapp data also suggest that 'when' the user wants a particular thing. For example if any user is seen online during late-nights, then it can be presumed that he/she needs a coffee to be awake, hence a coffee making company can advertise its products to him/her.

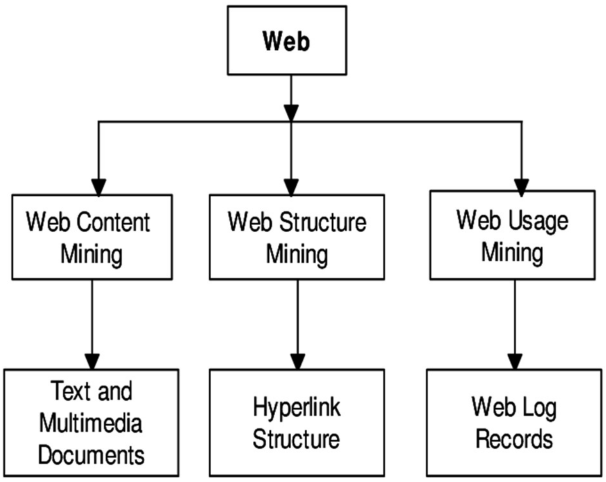
The challenge here is to extract the desired information from a gigantic pool of oceanic data spread across the various nodes of the internet. To address this problem a new technique of Data-Mining has come into picture and is known as social-media-mining.

1.1 Social Media Mining:-

Social Media Mining is a type of Data Mining and taken a different form from its predecessor – "Web Mining".

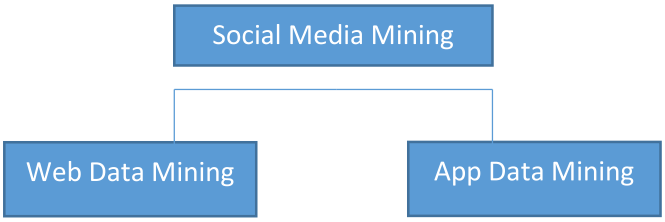


A diagrammatic representation of Data Mining process



A diagrammatic representation of Web Mining approach

Looking to the above approaches of Data Mining and Web Mining the approach towards Social media mining can be presumed as under:-



Social media mining is basically a process of extracting data from the Web and Apps and extrapolating a pattern from the data to provide the insight on the user behavior and sentiments. These user behavior and sentiments can help the advertisers and marketing agencies to target their desired customers. Since the data is ever expanding and ever changing hence Machine learning and A.I plays a major role in analytics of these data.

1.2 Process of Social Media Mining, Analysis and Prediction:-

a) Association:

Association examines patterns in a data base. Association is identified its association rule also known as “if-then” rule.

“If” is the antecedent and “then” is the consequent. Antecedent is an item found in a data set, and a consequent is an item which is found in grouping with the antecedent.
Example: if a man buys an Air-Conditioner, then it’s most likely that he will also buy a voltage stabilizer.

b) Classification:

Classification as the word suggests is used to arrange similar items in different pockets. The motive of classification is to forecast the target pocket for a particular data or item. For example a search for car by a user can be classified under mid-segment cars, high-segment cars, and luxury cars.

c) Tracking patterns:

As the name suggests, the pattern of the data created is observed and valuable insights are drawn from it. For example, the pattern in which mobile phones are sold during festive seasons can help the manufacturers decide their production rate.

d) *Prediction*:
Prediction is the pinnacle of data mining, where in the historical data is analyzed for patterns and future inferences are made. For example, the rainfall pattern of previous years can be helpful for the Power Generating companies to decide their production of electricity.

The aforementioned techniques can be utilized by 'Marketing' agencies to find their prospective customers, however the above techniques needs to be amalgamated with the "Sentiment Analysis" technique. The amalgamation of the above two shall be exemplified as under:-

S.No	Data Mining Techniques	Sentiment Analysis
1	Classification	The Sentiments of user's needs to be segregated and classified under similar groups. They should be categorized under the group names of "likes", "dislikes", "anger", "sadness", "excitement" etc
2	Tracking Pattern	Patterns needs to be observed about the no of "likes" and "dislikes".
3	Prediction	Based on the observed pattern, prediction can be made whether the user will "like" or "dislike" the new product/service.
4	Association	Based on the "likes" and "dislikes" the user can be associated for marketing of other products (which can be procured in complement of the already procured product).

2. METHODS & TECHNIQUES USED IN THIS PAPER

This Paper has a 3 steps approach:

A. Scrapping
B. Data Conversion
C. Analyzing using Power BI

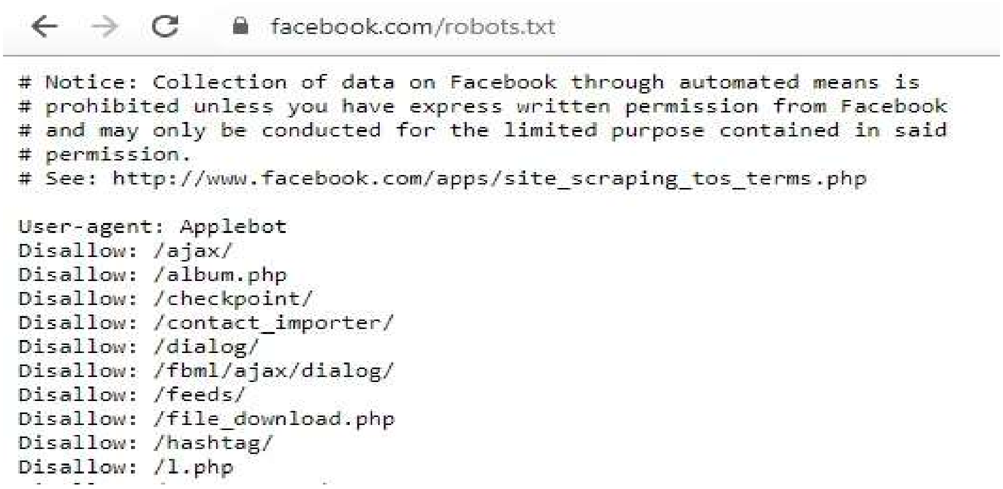
Firstly we will gather data through scraping, then we will convert it into excel sheet, and at last we will perform analytics on the data using power BI

2.1 Data Mining using Python Scraping:

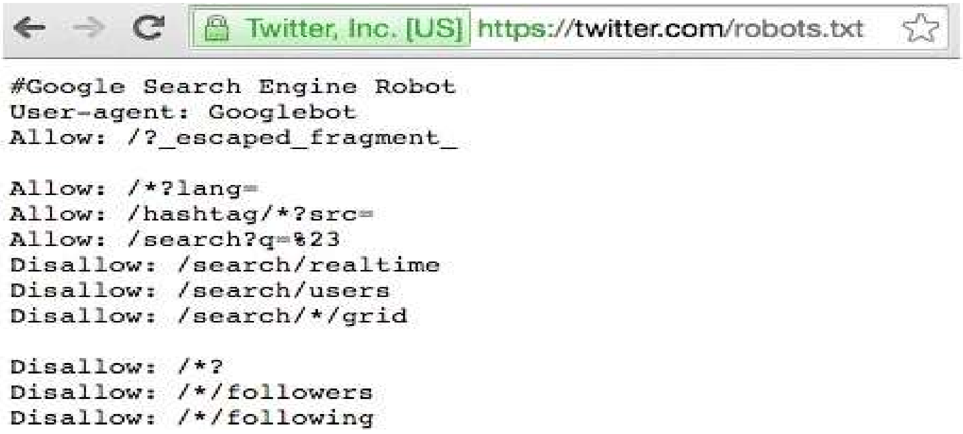
The social media data is basically of 2 forms: Private & Public. Any user who shares his/her opinion or sentiments on social media, shares it in either public domain or in private domain. The public data mining is legitimate but private data mining is unethical and illegal.

Infact, Facebook and Twitters have disabled crawlers using Robots.txt.

If you type https://www.facebook.com/robots.txt or https://www.twitter.com/robots.txt this shall be the resultant page, in which all kinds of 'bots' and crawlers are disallowed.



```
# Notice: Collection of data on Facebook through automated means is
# prohibited unless you have express written permission from Facebook
# and may only be conducted for the limited purpose contained in said
# permission.
# See: http://www.facebook.com/apps/site_scraping_tos_terms.php

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /dialog/
Disallow: /fbml/ajax/dialog/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
```



```
#Google Search Engine Robot
User-agent: Googlebot
Allow: /?_escaped_fragment_

Allow: /*?lang=
Allow: /hashtag/*?src=
Allow: /search?q=%23
Disallow: /search/realtime
Disallow: /search/users
Disallow: /search/*/grid

Disallow: /*?
Disallow: /*/followers
Disallow: /*/following
```

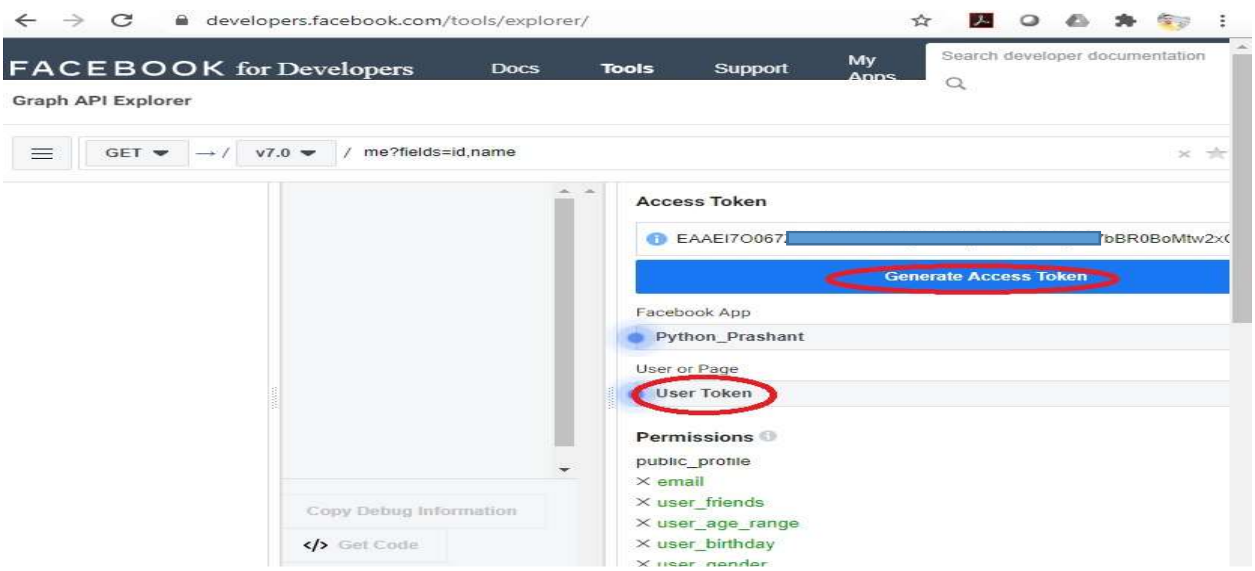
However if any user shares his/her secret token or key then his her data can be drilled down. In this paper we will try to drill down our(1st Authors) Private data using his secret token or key, also we will retrieve the publically available data using python scraping tools.

i. Private Data Mining:

2.1.1.1 Facebook:

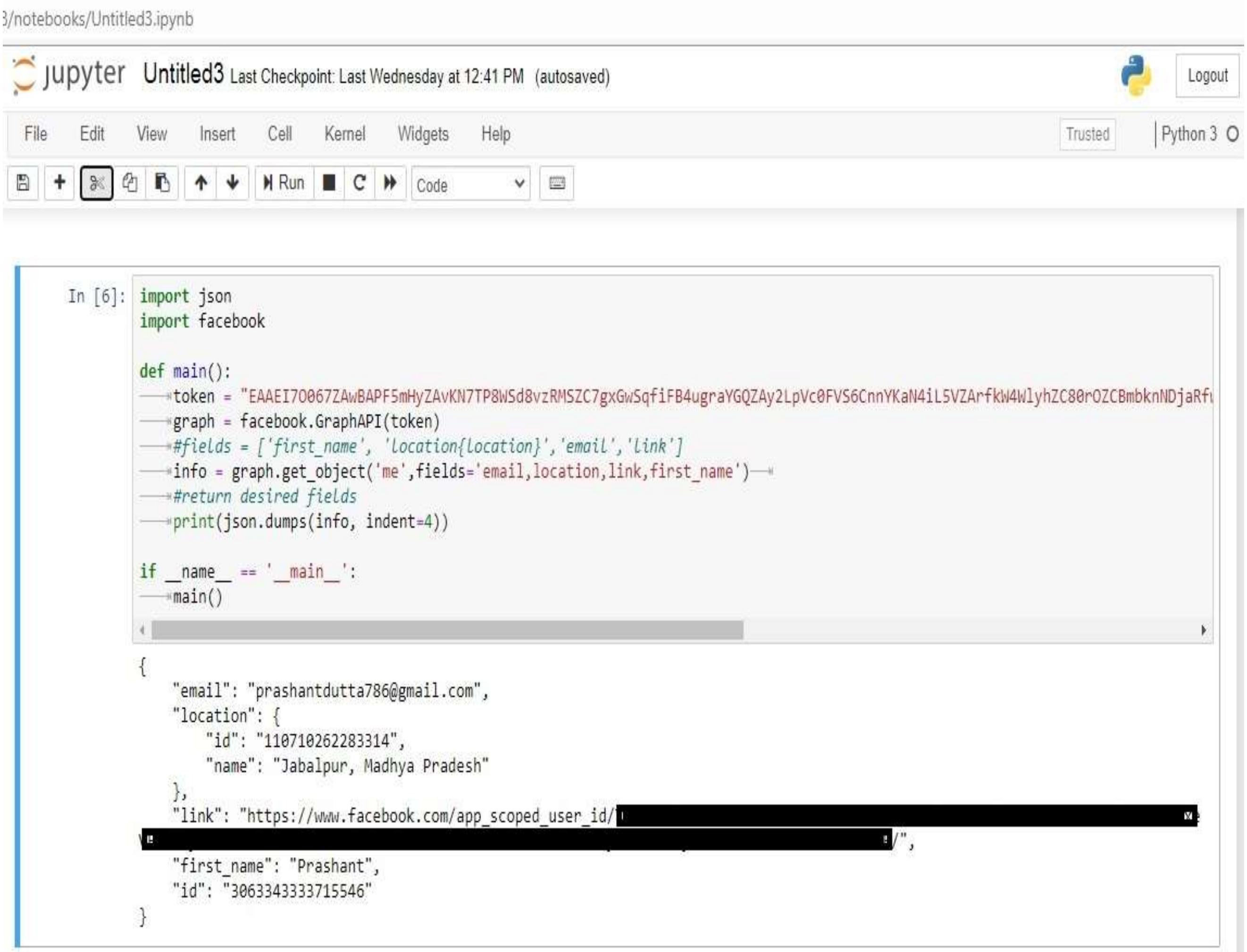
Before the data mining process, one needs to visit https://developers.facebook.com

i. Create an APP there
ii. After the App is created, go to https://developers.facebook.com/tools/explorer/
iii. After this generate the access token and provide the desired permissions



Access Token
EAAEI7O067[redacted]bBR0BoMtw2x(
Generate Access Token
Facebook App
Python_Prashant
User or Page
User Token
Permissions
public_profile
× email
× user_friends
× user_age_range
× user_birthday
× user_gender

In the next step, retrieve the data using Python code


3/notebooks/Untitled3.ipynb
jupyter Untitled3 Last Checkpoint: Last Wednesday at 12:41 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

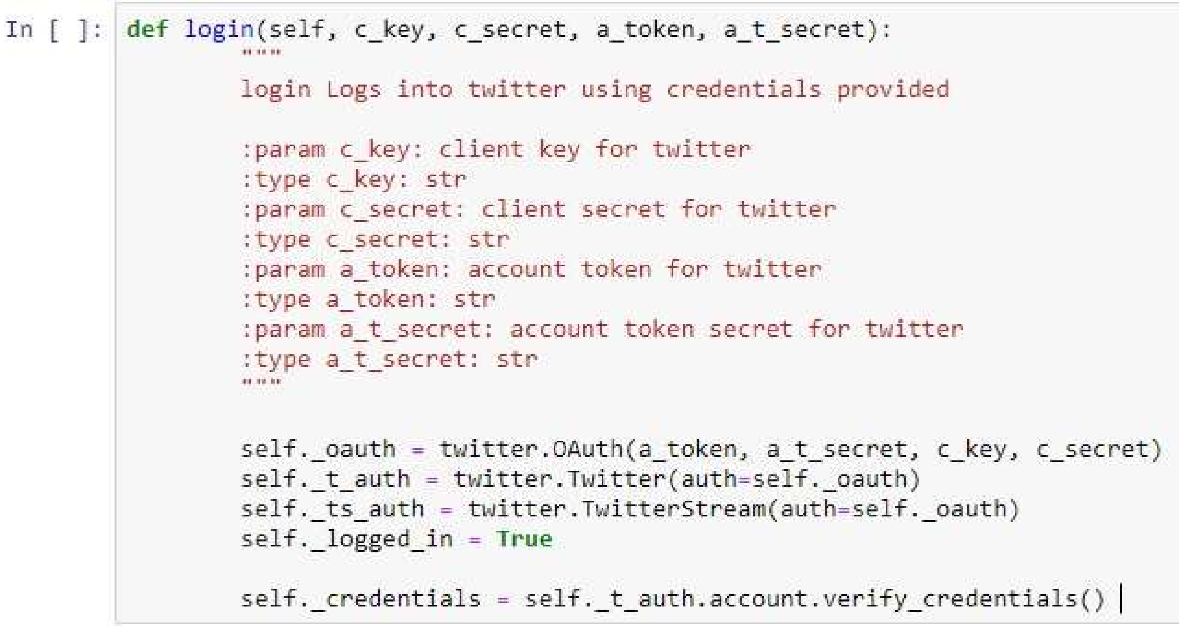
In [6]:

```
import json
import facebook

def main():
    token = "EAAEI7O067ZAwBAPF5mHyZAvKN7TP8WSd8vzRMSZC7gxGwSqfiFB4ugraYGQZAy2LpVc0FVS6CnnYKaN4iL5VZArfkW4WlyhZC80rOZCBmbknNDjaRf
    graph = facebook.GraphAPI(token)
    #fields = ['first_name', 'location{location}','email','link']
    info = graph.get_object('me',fields='email,location,link,first_name')
    #return desired fields
    print(json.dumps(info, indent=4))

if __name__ == '__main__':
    main()
```

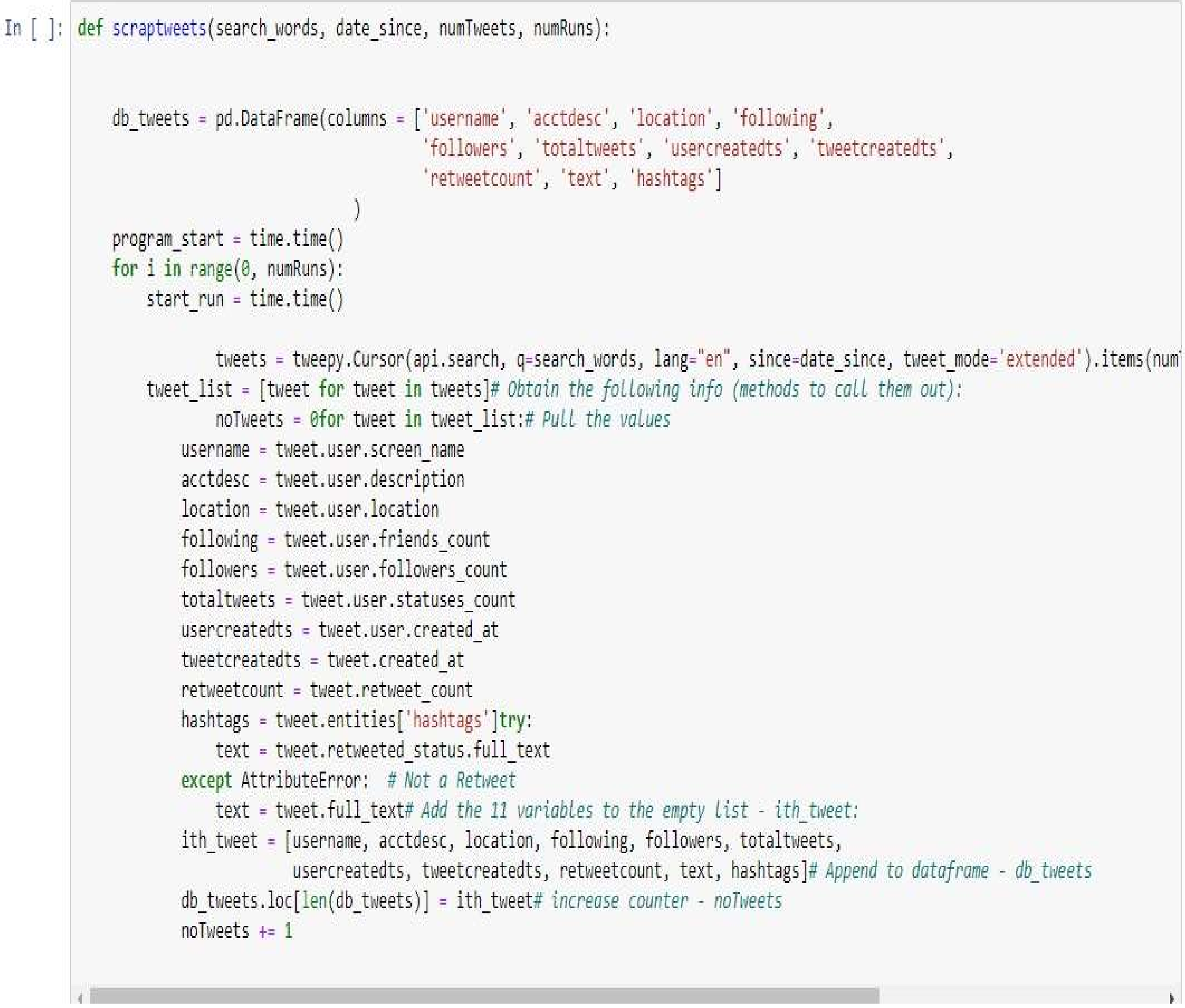
```
{
    "email": "prashantdutta786@gmail.com",
    "location": {
        "id": "110710262283314",
        "name": "Jabalpur, Madhya Pradesh"
    },
    "link": "https://www.facebook.com/app_scoped_user_id/[redacted]/",
    "first_name": "Prashant",
    "id": "3063343333715546"
}
```

2.1.1.2 Twitter:

i. Visit - https://dev.twitter.com/apps
ii. Create New App
iii. Twitter, generally asks for few questions in a questionnaire form and may be via emails also
iv. After twitter is satisfied, the 'Create My Access Token' button gets enabled
v. Now copy the following from the access token- api_key, api_secret, access_token_key, access_token_secret.
vi. After this you can retrieve the tweets using the following code



```
In [ ]: def login(self, c_key, c_secret, a_token, a_t_secret):
                """
                login Logs into twitter using credentials provided

                :param c_key: client key for twitter
                :type c_key: str
                :param c_secret: client secret for twitter
                :type c_secret: str
                :param a_token: account token for twitter
                :type a_token: str
                :param a_t_secret: account token secret for twitter
                :type a_t_secret: str
                """

                self._oauth = twitter.OAuth(a_token, a_t_secret, c_key, c_secret)
                self._t_auth = twitter.Twitter(auth=self._oauth)
                self._ts_auth = twitter.TwitterStream(auth=self._oauth)
                self._logged_in = True

                self._credentials = self._t_auth.account.verify_credentials()
```



```
In [ ]: def scraptweets(search_words, date_since, numTweets, numRuns):


            db_tweets = pd.DataFrame(columns = ['username', 'acctdesc', 'location', 'following',
                                                'followers', 'totaltweets', 'usercreatedts', 'tweetcreatedts',
                                                'retweetcount', 'text', 'hashtags']
                                    )
            program_start = time.time()
            for i in range(0, numRuns):
                start_run = time.time()

                        tweets = tweepy.Cursor(api.search, q=search_words, lang="en", since=date_since, tweet_mode='extended').items(num
                tweet_list = [tweet for tweet in tweets]# Obtain the following info (methods to call them out):
                        noTweets = 0for tweet in tweet_list:# Pull the values
                    username = tweet.user.screen_name
                    acctdesc = tweet.user.description
                    location = tweet.user.location
                    following = tweet.user.friends_count
                    followers = tweet.user.followers_count
                    totaltweets = tweet.user.statuses_count
                    usercreatedts = tweet.user.created_at
                    tweetcreatedts = tweet.created_at
                    retweetcount = tweet.retweet_count
                    hashtags = tweet.entities['hashtags']try:
                        text = tweet.retweeted_status.full_text
                    except AttributeError:  # Not a Retweet
                        text = tweet.full_text# Add the 11 variables to the empty list - ith_tweet:
                    ith_tweet = [username, acctdesc, location, following, followers, totaltweets,
                                 usercreatedts, tweetcreatedts, retweetcount, text, hashtags]# Append to dataframe - db_tweets
                    db_tweets.loc[len(db_tweets)] = ith_tweet# increase counter - noTweets
                    noTweets += 1
```

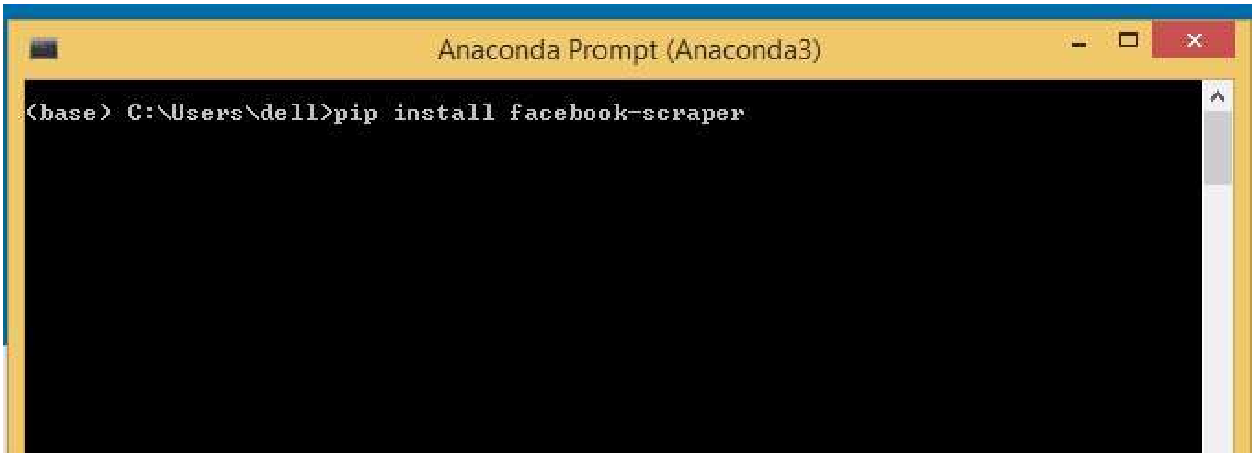
2.1.2 Public Data Mining:

Public data(which is publically available) mining does not require any “key” or “authentication”

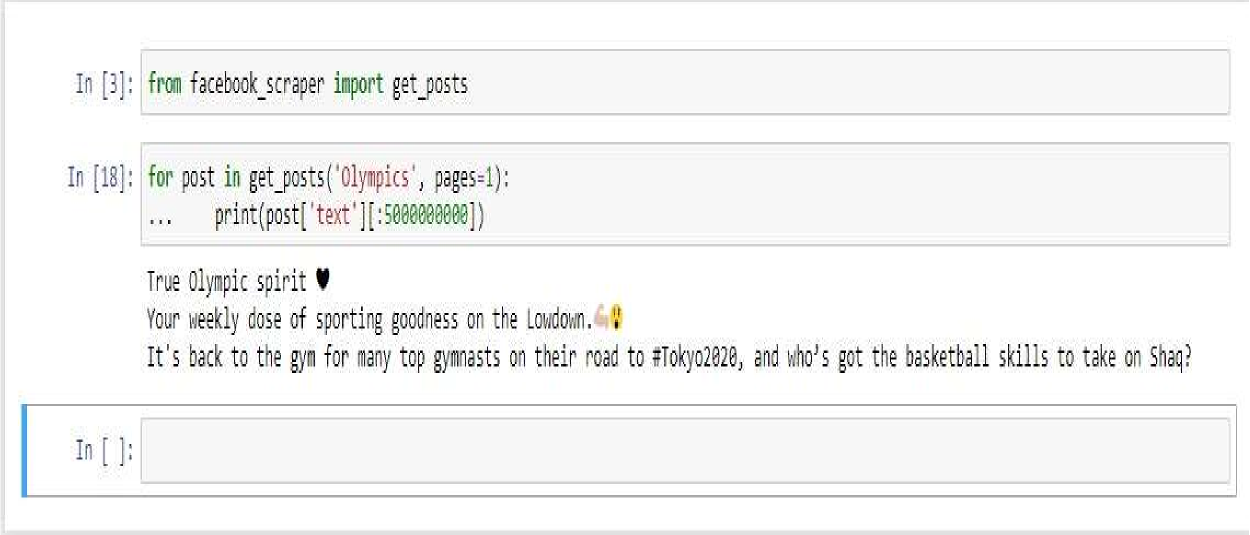
2.1.2.1 Facebook:

The python library provide an API known as Facebook Scraper which helps in scraping public data

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



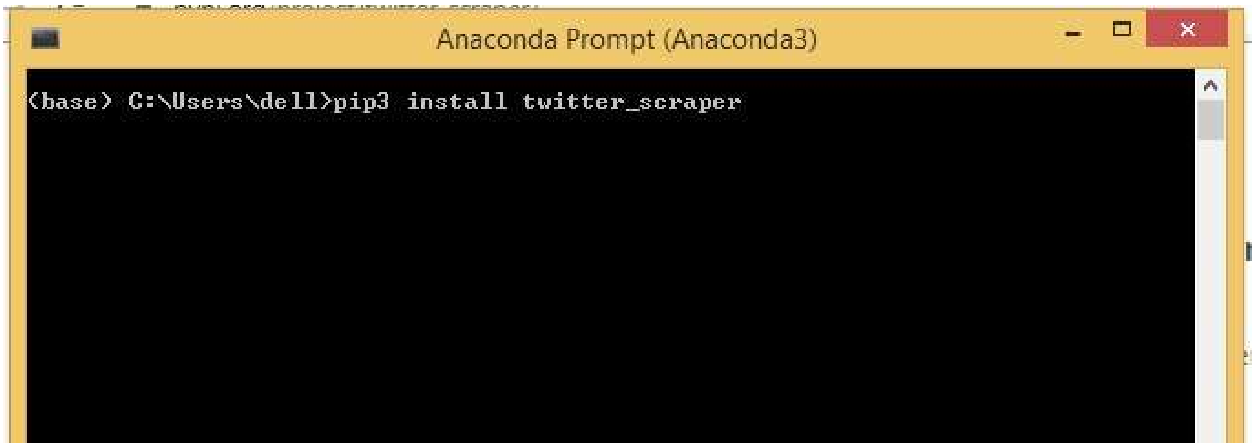
```
(base) C:\Users\dell>pip install facebook-scraper
```



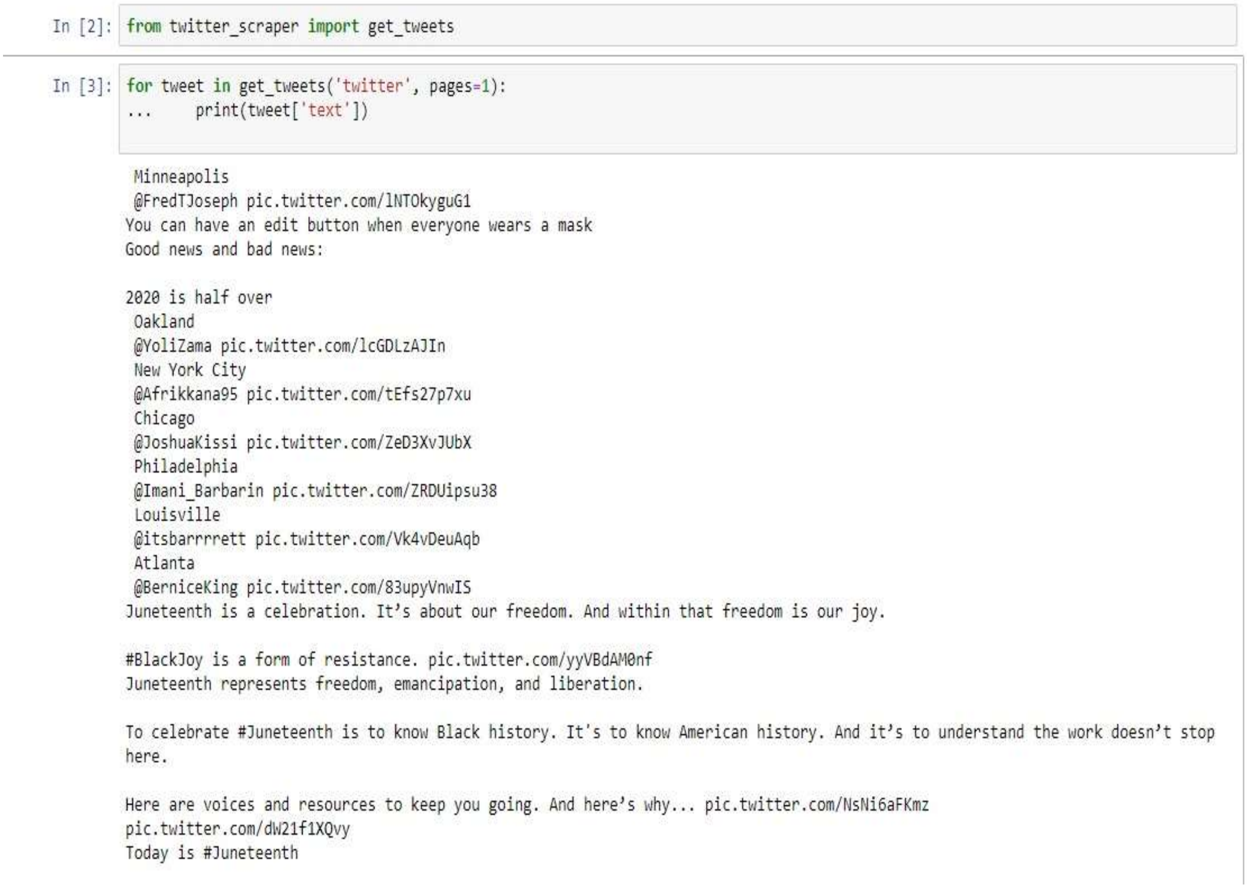
```
In [3]: from facebook_scraper import get_posts

In [18]: for post in get_posts('Olympics', pages=1):
         ...     print(post['text'][:5000000000])

         True Olympic spirit ♥
         Your weekly dose of sporting goodness on the Lowdown.💪😱
         It's back to the gym for many top gymnasts on their road to #Tokyo2020, and who's got the basketball skills to take on Shaq?

In [ ]:
```

2.1.2.2 Twitter:

The python library provide an API known as Twitter Scraper which helps in scraping public data



```
(base) C:\Users\dell>pip3 install twitter_scraper
```

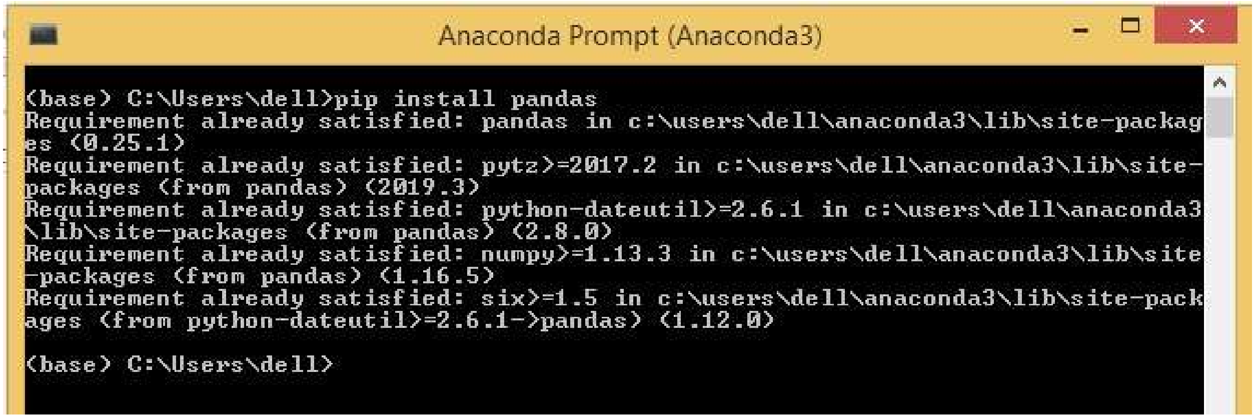


```
In [2]: from twitter_scraper import get_tweets

In [3]: for tweet in get_tweets('twitter', pages=1):
        ...     print(tweet['text'])

 Minneapolis
 @FredTJoseph pic.twitter.com/lNTOkyguG1
You can have an edit button when everyone wears a mask
Good news and bad news:

2020 is half over
 Oakland
 @YoliZama pic.twitter.com/lcGDLzAJIn
 New York City
 @Afrikkana95 pic.twitter.com/tEfs27p7xu
 Chicago
 @JoshuaKissi pic.twitter.com/ZeD3XvJUbX
 Philadelphia
 @Imani_Barbarin pic.twitter.com/ZRDUipsu38
 Louisville
 @itsbarrrrett pic.twitter.com/Vk4vDeuAqb
 Atlanta
 @BerniceKing pic.twitter.com/83upyVnwIS
Juneteenth is a celebration. It’s about our freedom. And within that freedom is our joy.

#BlackJoy is a form of resistance. pic.twitter.com/yyVBdAM0nf
Juneteenth represents freedom, emancipation, and liberation.

To celebrate #Juneteenth is to know Black history. It's to know American history. And it’s to understand the work doesn’t stop
here.

Here are voices and resources to keep you going. And here’s why… pic.twitter.com/NsNi6aFKmz
pic.twitter.com/dW21f1XQvy
Today is #Juneteenth
```

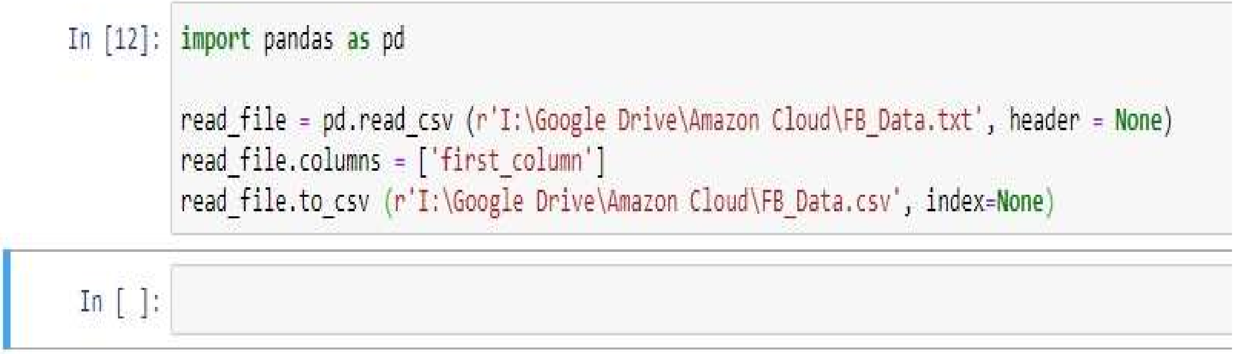
2.2 Data Conversion:

The next step after data mining is converting the raw data into a usable format.

Data conversion is a process of converting one for of data to an another form. In our context we will be converting the text data collected into an Excel sheet using Panda library of Python.



```
(base) C:\Users\dell>pip install pandas
Requirement already satisfied: pandas in c:\users\dell\anaconda3\lib\site-packag
es (0.25.1)
Requirement already satisfied: pytz>=2017.2 in c:\users\dell\anaconda3\lib\site-
packages (from pandas) (2019.3)
Requirement already satisfied: python-dateutil>=2.6.1 in c:\users\dell\anaconda3
\lib\site-packages (from pandas) (2.8.0)
Requirement already satisfied: numpy>=1.13.3 in c:\users\dell\anaconda3\lib\site
-packages (from pandas) (1.16.5)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-pack
ages (from python-dateutil>=2.6.1->pandas) (1.12.0)

(base) C:\Users\dell>
```



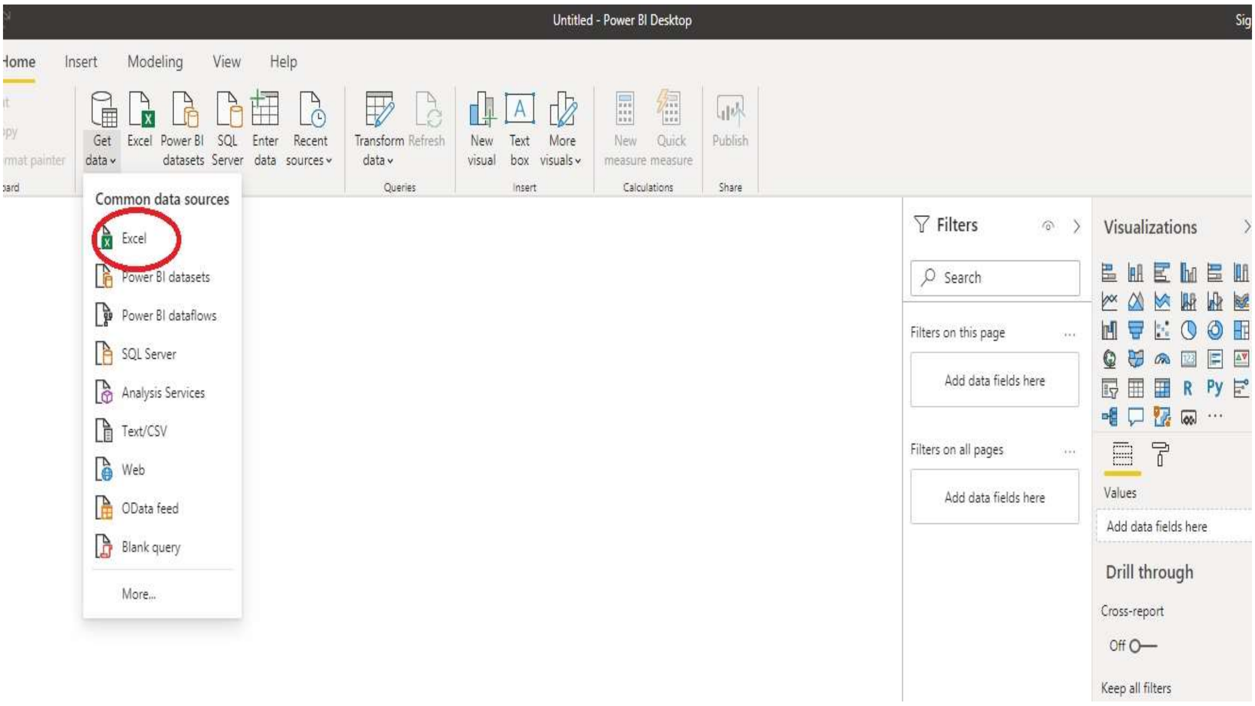
```
In [12]: import pandas as pd

         read_file = pd.read_csv (r'I:\Google Drive\Amazon Cloud\FB_Data.txt', header = None)
         read_file.columns = ['first_column']
         read_file.to_csv (r'I:\Google Drive\Amazon Cloud\FB_Data.csv', index=None)

In [ ]:
```

3. RESULTS AND DISCUSSION

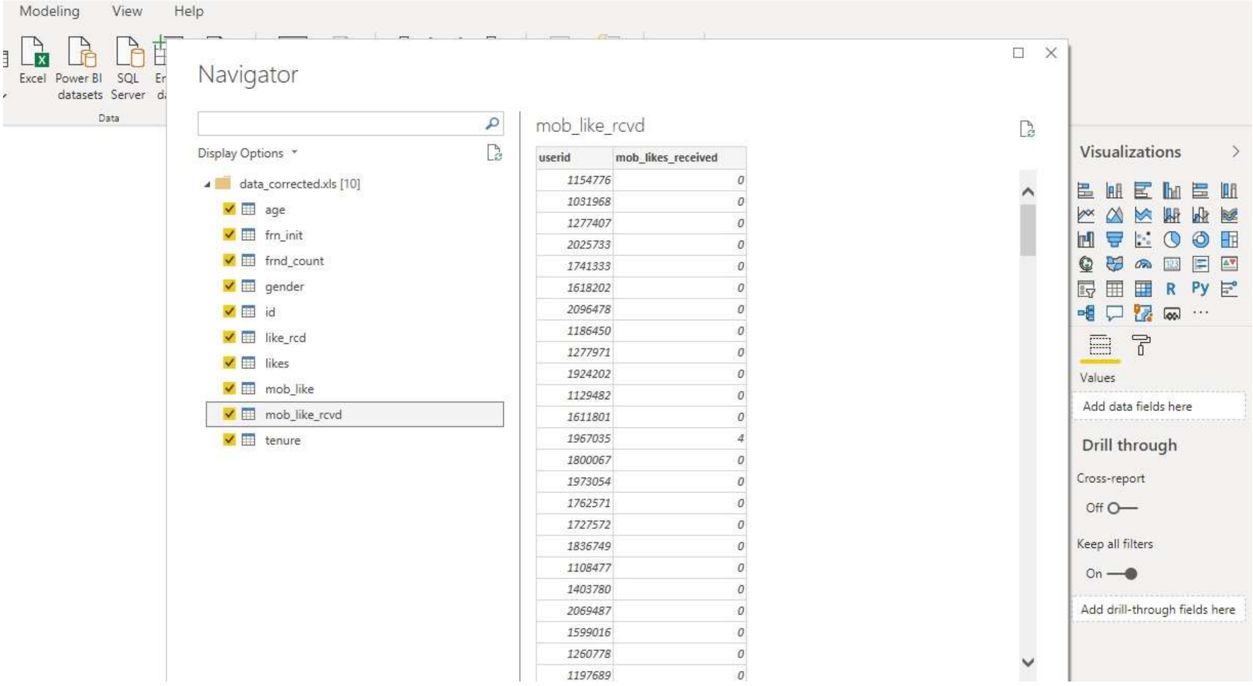
After the data is converted to the desired format(CSV in our case) it is now subjected to analysis
We will be using Microsoft Power BI for Data Analysis.

Steps:

i. Pull data from the “Get Data” tab in Power Bi, select source as Excel.



ii. Transform and Load Data, Apply and Close



Navigator

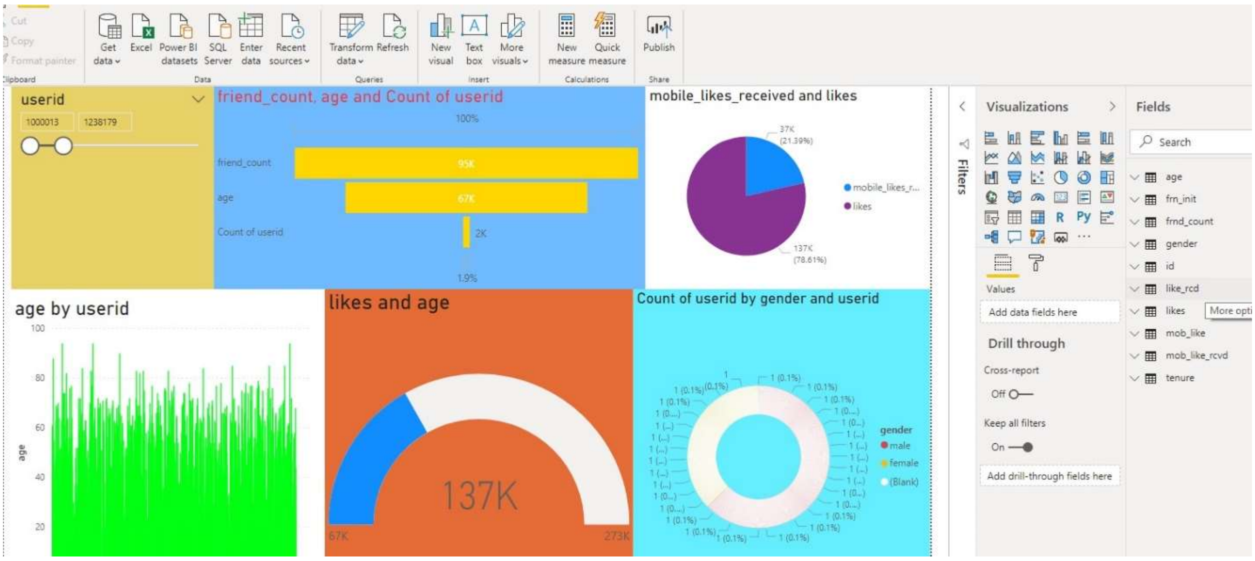
Display Options

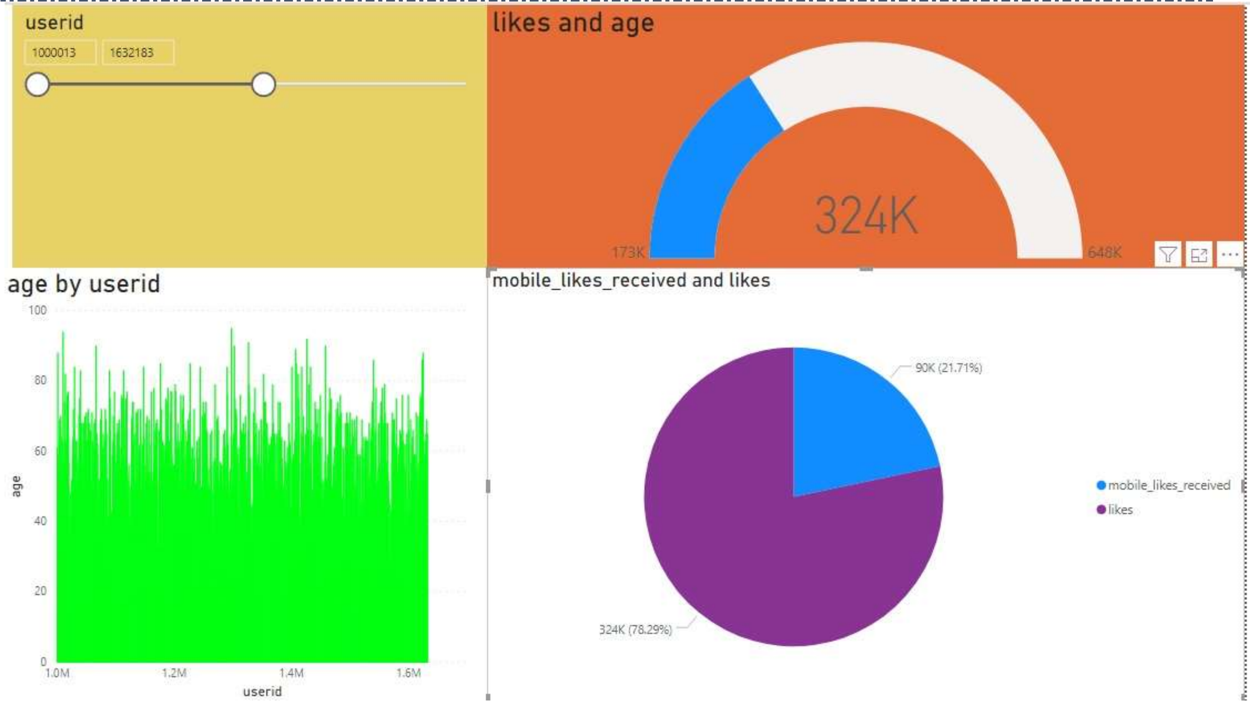
data_corrected.xls [10]
- ☑ age
- ☑ frn_init
- ☑ frnd_count
- ☑ gender
- ☑ id
- ☑ like_rcd
- ☑ likes
- ☑ mob_like
- ☑ mob_like_rcvd
- ☑ tenure

mob_like_rcvd

userid	mob_likes_received
1154776	0
1031968	0
1277407	0
2025733	0
1741333	0
1618202	0
2096478	0
1186450	0
1277971	0
1924202	0
1129482	0
1611801	0
1967035	4
1800067	0
1973054	0
1762571	0
1727572	0
1836749	0
1108477	0
1403780	0
2069487	0
1599016	0
1260778	0
1197689	0

iii. Perform Analysis





The above analysis is a description of :-

i. Users Age
ii. Likes corresponding to the age of user
iii. Likes received by mobile app and web portal
iv. Friends count corresponding to the age of user

4. CONCLUSION

The aforementioned data analytics output can be priceless for any marketing agencies. Marketing agencies uses many tools and means to drill down these social media data. The "sentiment" of an individual is the new gold for marketing companies. The sentiments if captured precisely can be also turning-point for other fields like Politics (Eg. Facebook–Cambridge Analytica data scandal). The analytics done above were just our way of drilling down the insights, every other organization will have its own way of drilling down. However as previously told in this paper that Social media data is increasing in an unprecedented rate hence storing all the scraped data and further performing analysis on it is quite impossible for a single processing unit or server, hence the cloud platform needs to be engaged here. Cloud platforms like AWS and Azure not only provides mammoth storage but also provides you various tools for data analytics. AWS has S3 and Data Lakes for storage and Kinesis for analytics similarly Azure has Azure Analysis services.

REFERENCES

[1] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#657457e260ba
[2] https://datareportal.com/social-media-users
[3] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
[4] https://zephoria.com/top-10-valuable-snapchat-statistics/
[5] https://pypi.org/project/facebook-scraper/
[6] https://pypi.org/project/twitter-scraper/
[7] https://developers.facebook.com/tools
[8] Weissbock, J. and Inkpen, D., 2014, in: Combining Textual Pre-Game Reports and Statis- tical Data for Predicting Success in the National Hockey League, Advances in Artificial Intelligence, Springer International Publishing, pp. 251–262 .

[9] Javed, B.S., 2018, Hybrid semantic clustering of hashtags, Online Social Networks and Media 5 (2018) 23–36 .

[10] Vicient , A. M., 2014, Unsupervised semantic clustering of Twitter hashtags, Proceedings of the 21st European Conference on Artificial Intelligence, pp. 1119–1120 .

[11] Javed , B.S., 2016, Sense-level semantic clustering of hashtags in social me- dia, in: Proceedings of the 3rd Annual International Symposium on Informa- tion Management and Big Data, pp. 140–149 .

[12] https://sproutsocial.com/insights/social-media-sentiment-analysis/.