Proposal

The growth of social networks is increasing tremendously and it is one of the best way to keep connected with family and friends be it whether they are overseas or domestically located. Twitter has a great source of rich information from millions of users on the internet. Therefore it is a fitting source for applying data mining.

Mining twitter can help us understand how people use new communication technologies to to develop social connections.

The purpose of this report is to illustrate how to use Twitter API to data mine. It provides step-by-step explanations on how the information is extracted from Twitter. There are 319 million monthly active users on Twitter. There are many topics and reasons to extract data from the users. The report will cover specific script as tweets relating to Wal-Mart. There will be detailed explanation of the script on how the data is extracted. The API is so simple and easy to use that any person with little programming background will be able to analyze the data.

Tools Needed

In order to do the project the following compiler, library and API were needed. We have detailed instructions on how we downloaded each item in the coming pages.

1. Download Python IDLE
2. Download Python twitter library
3. How to get Twitter API

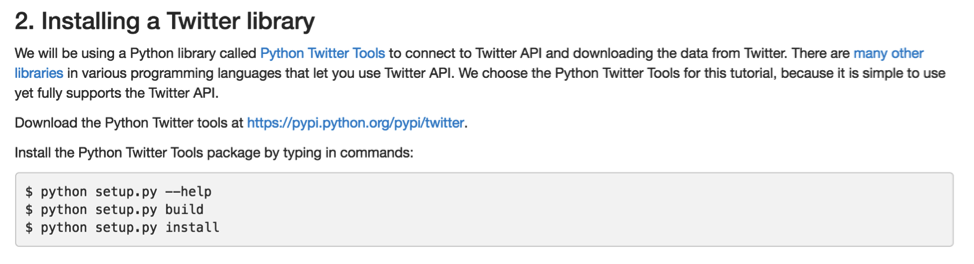
Download Python IDLE

1. Python IDLE. We downloaded the latest version of Python IDLE which is 3.6.

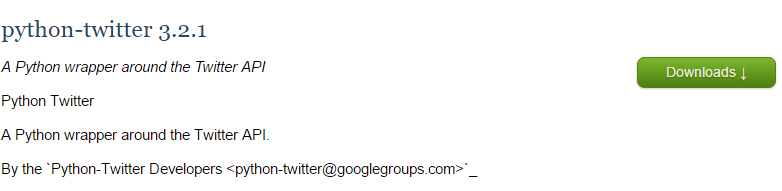


1. Download Python IDLE from <http://www.python.org/download/>

Download Python Twitter library

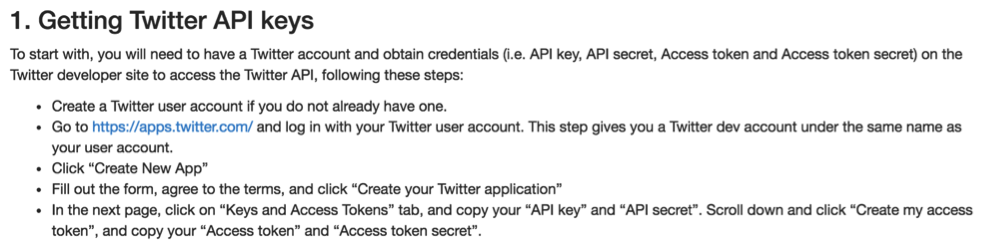


1. We used Python library called Python Twitter Tools.
2. Python Twitter Tools is used to connect to Twitter API which is the used to download data from Twitter.
3. There are other libraries available for other programming languages but we choose Python Twitter Tools because it is simple to use.
4. Python Twitter Tools is downloaded from https://pypi.python.org/pypi/python-twitter/3.2.1
5. Install Python Twitter Tools package by typing the following commands:
   1. $ easy\_install pip
   2. $ python get-pip.py
   3. $ pip install simplejson
   4. $ pip install twitter



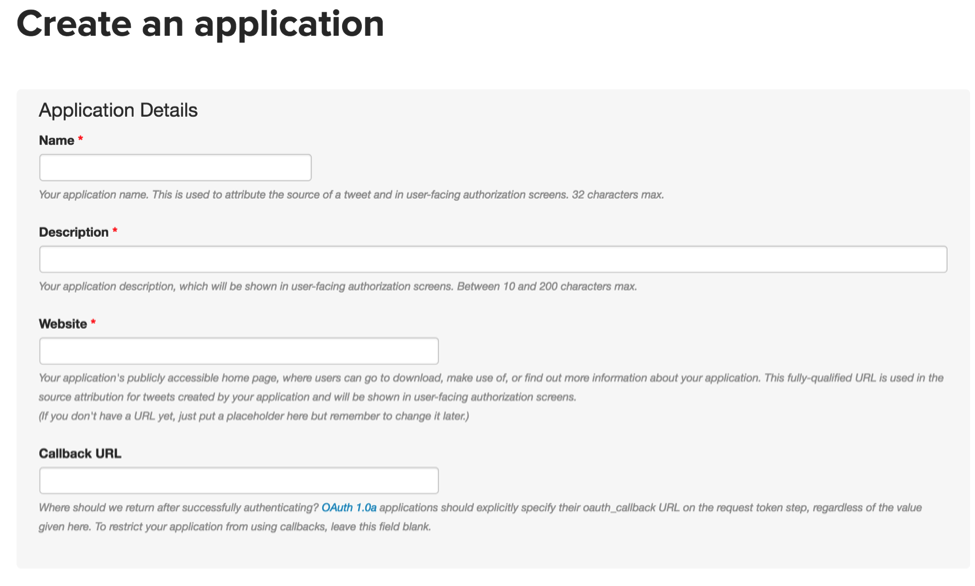
How to get Twitter API

1. Create twitter account unless you have one
2. <https://apps.twitter.com/> go to this link and sign in with your twitter account.
   1. This allows you to create a Twitter developer account with the same user name.
   2. Even if someone does not own a twitter account, one can still create Twitter developer account.
3. Click “Create New App”.



4. Fill out the form, agree to terms and conditions.

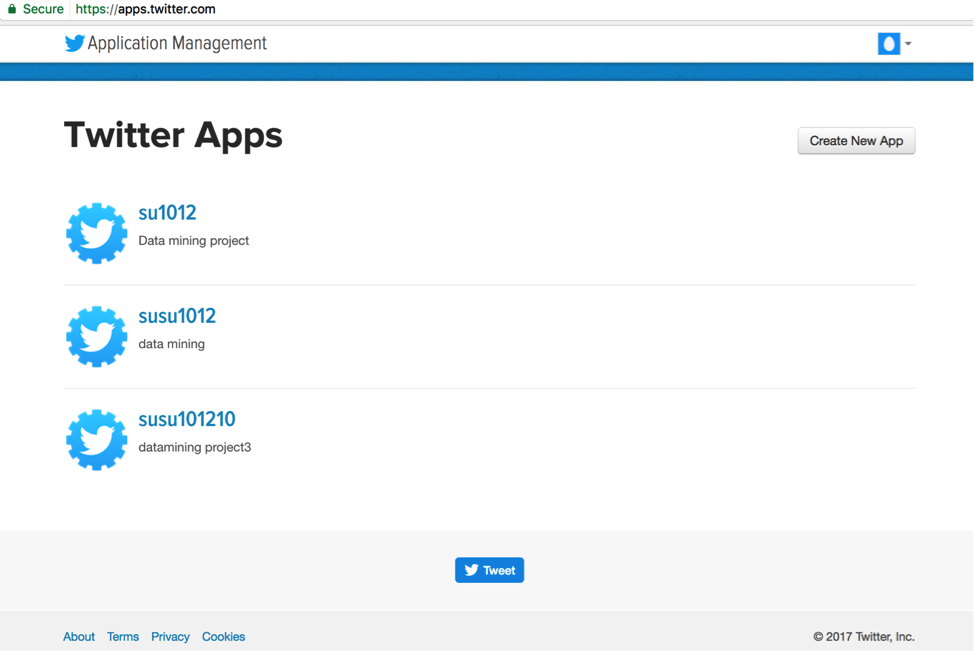
5. Click “Create your Twitter application”.



6. Click on “Keys and Access tokens” tab and copy the “API key” and “API secret”.

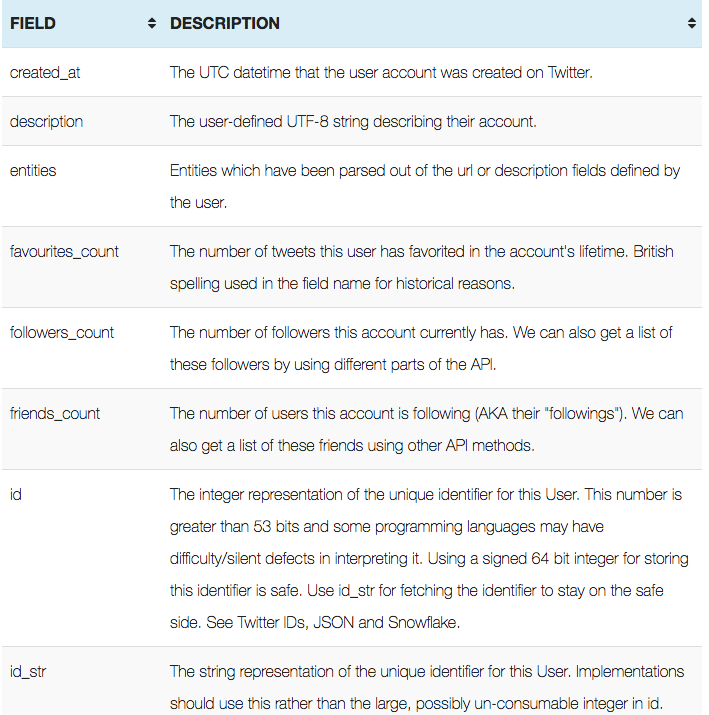
7. Scroll down and click “Create my access token” and copy the “Access token” and “Access token secret”.

This is the account created for the project



How to access Twitter API

There are different API methods for accessing informations such as: tweets, retweets, users, following relationships, etc. Here is a list of the most useful of the variables returned by the API for each user modified descriptions are taken from Twitter website





Type of API used

Twitter provides three kinds of APIs:

1. Search API: lets users to query Twitter content. Users can find tweets for a set of keywords, users or location posted in the past.
2. REST API: allows developers to have access to some of the core basic information of Twitter including timelines, status updates, and user information.
3. Streaming API: enables for large quantities of keywords to be specified and tracked, fetching geotagged tweets from certain region.



System architecture for Streaming API

We used Streaming API for our project. Since this method maintains a persistent connection that reduces network interference similarly when server produces continuous stream of data.

Code

The code above is used to collect data with the word ‘walmart’. On terminal we typed the following code:

$ python collect.py > collect.json

With the code above all data is saved in collect.json

Untitled.png

The image above shows the file size of collect.json

Result of collect.json



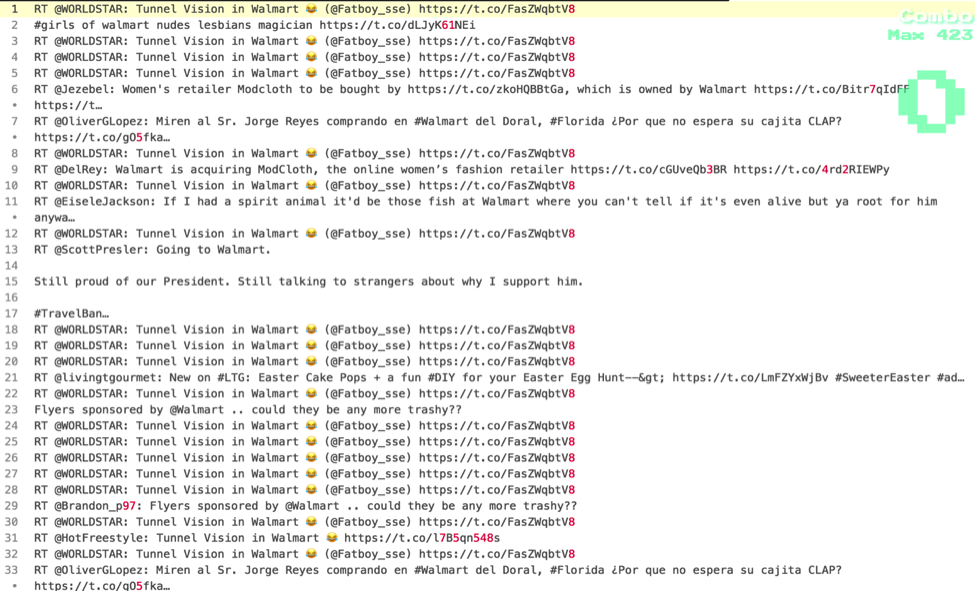
Since streaming API is used collecting data had to be done without any interruption and due to that reason it took 4 hours to gather all the data and the size of the data collected in 55MB.

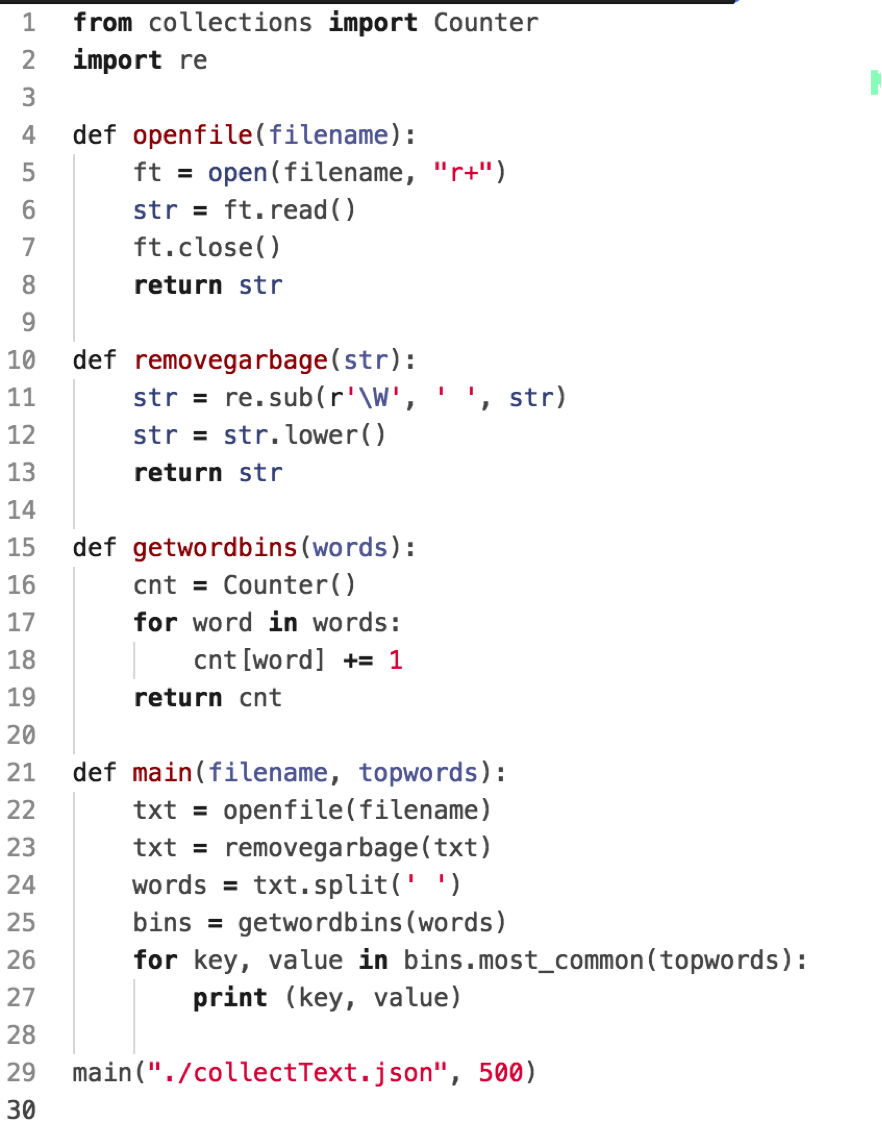


The above code extracts only text area. Typing the following code in terminal.

$python extractText.py > extractText.json

Result of extractText.json

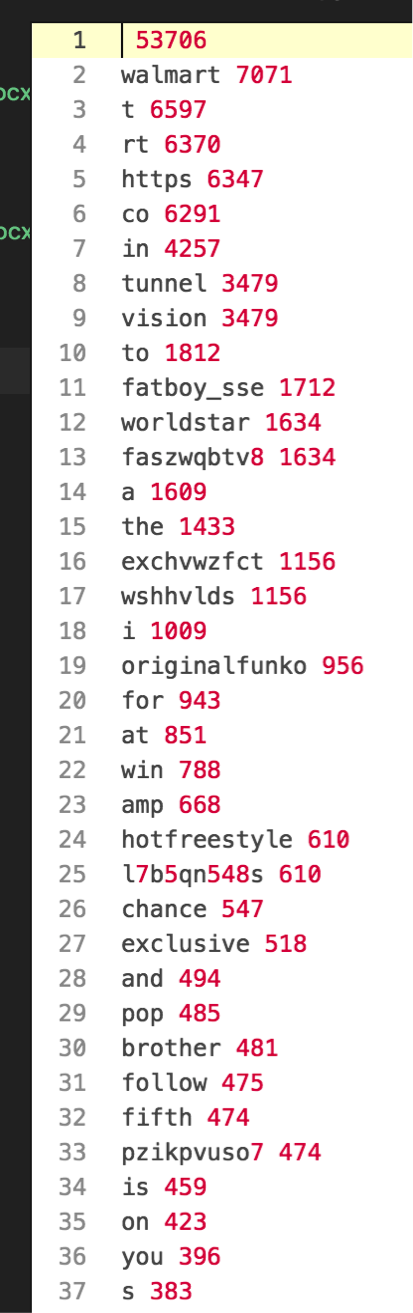




The code above is to find out occurrence of top 500 words. All the data is stored in ranking.json. It can be done by typing the following code in terminal

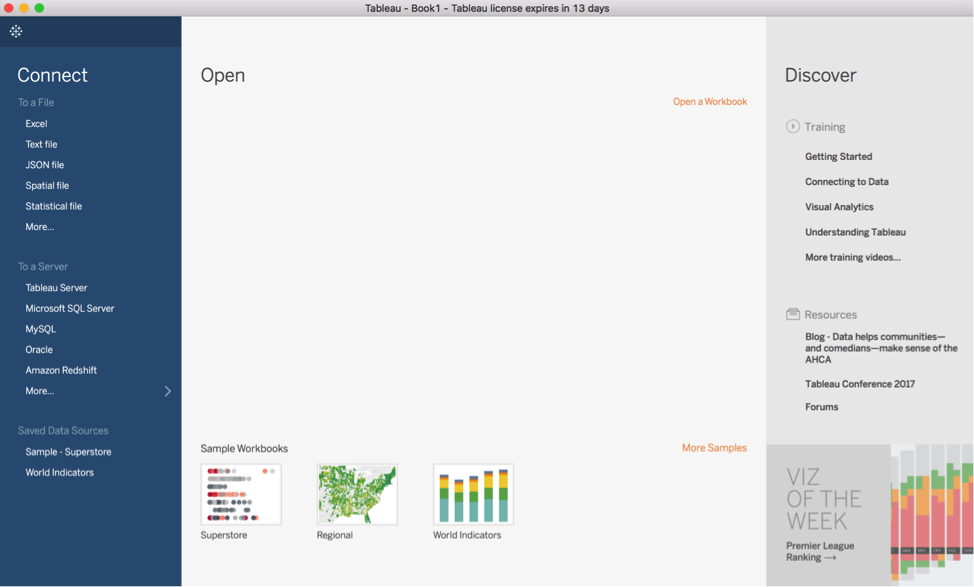
$ python ranking.py > ranking.json

Result of ranking.json



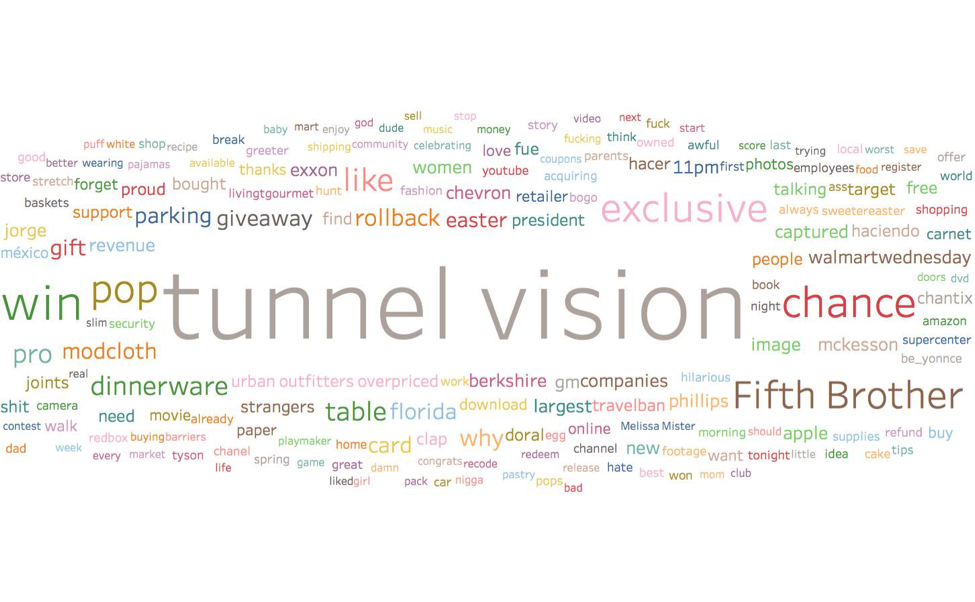
All unnecessary words has been eliminated from ranking.json. Unnecessary words like articles, conjunctions and etc were extracted manually and saved in the excel file.

Tableau is installed from<https://www.tableau.com/> in order to create word cloud image file. Tableau software tool is used with excel file in order to create the word cloud. (Tableau software is not a free software tool but there is a 15 days free trial available).



Tableau

Data visualization



Word cloud

Conclusion

As social media is one of the most popular resource for people today to communicate with each other, it is a great source for analyst to perform data mining in them using their API tools provided which are open source and cost free. Being students clearly these resources that are used for this project was a great knowledge to know how to use and manipulate the data that are available in the social networks. Twitter being one of the growing and most popular social media helped to see how people use different words when expressing their views and thoughts to the world these days.

The fact that the data available in twitter is extremely active and continuously changing if someone decides to keep track of a specific keyword over a specific period of time then we can see how the diction is twitter has changed over that specific period of time. Someone can also see how a company’s reputation has changed over a certain period of time. For example United Airlines before the sensational event that occurred recently, United Airlines did not have such a terrible rating but now due to the event if someone were to use Twitter to data mine United Airlines their reputation is not going to be that great. So in the end one can say that Twitter is one of the most powerful resource to data mine and analyze the data available in it.

Citations:

1. <https://dev.twitter.com/streaming/overview>
2. <https://dev.twitter.com/rest/public/search>
3. <http://socialmedia-class.org/twittertutorial.html>
4. <https://www.tableau.com/>
5. http://www.rdatamining.com/docs/twitter-analysis-with-r