Facebook Profile Sociality Analyzer

TECHNICAL DOCUMENT

AUTHORS:

Table of Contents

[**1.** **INTRODUCTION** 3](#_Toc502008179)

[**2.** **FACEBOOK DATA STRUCTURE AND RETREIVING MECHANISM** 4](#_Toc502008180)

[**2.1.** **Data retrieving mechanism** 4](#_Toc502008181)

[**2.2.** **Data structure** 5](#_Toc502008182)

[**2.3.** **Data storing and encryption** 7](#_Toc502008183)

[**3.** **PROGRAMING LANGUAGE** 9](#_Toc502008184)

[**3.1.** **Language and working environment** 9](#_Toc502008185)

[**3.2.** **Dependencies** 9](#_Toc502008186)

[**4.** **MODULES AND CLASSES** 10](#_Toc502008187)

[**4.1.** **Built-in modules** 10](#_Toc502008188)

[**4.2.** **Third-party modules** 10](#_Toc502008189)

[**4.3.** **Self-built modules** 10](#_Toc502008190)

Table of Figures

[Figure 1. Example of a return of JSON (fields=posts) 5](#_Toc502008291)

[Figure 2. The maximum deep of data level could be supported by one retrieving time from 6](#_Toc502008292)

[Figure 3. Facebook Data structure created by the application 7](#_Toc502008293)

[Figure 4. The process of file i/o based on an encoded filename (user-name, id) 7](#_Toc502008294)

[Figure 5. Data storing on file system 7](#_Toc502008295)

[Figure 6. An encrypted file opened in Notepad 8](#_Toc502008296)

[Figure 7. A screenshot as an example of auto-limite-retries by the app 11](#_Toc502008297)

[Figure 8. Paging keyword showed in returned JSON in case of any sub-page exists 11](#_Toc502008298)

[Figure 9. Process of retrieving Posts 11](#_Toc502008299)

[Figure 10. Retrieving finish messages. Above: finished the posts request alone and saved. Below: finished all data and saved 12](#_Toc502008300)

1. **INTRODUCTION**

This application aims to collect and analyze the reactions of any user’s friends. We are motivated by a situation, in which most of the end-users of Facebook Social Network (FB) are not going to know their real connections. For example, many users have many friends in their list that without any clear knowledge about how the real interactions of those friends to them as comments and reactions. Those kinds of interaction are very important because those are what they really want to have when joining a social network, where people interactions were the main idea. It also come from a practical issue that many FB users took much of time to go on their friends list and tried to figure out who did not interact anything with them and gave those people un-friend actions.

Therefore, we write this app to help those people to have a clearly look at how the interaction between them and their friends, how the frequencies of comments or reactions they received from their friends or whose friends mostly care about them. We wrote this application using Python as the programming language and Git-hub as a sub-version control. In this application, we figured out the possible ways of data retrieving and also the ways to control the received data and the receiving processes. After that, we figured out how to deal with the data extraction and plotting graphs of interactions or clustering. Finally, we deploy into a web-style framework that makes the app could be used online as an FB application.

1. **FACEBOOK DATA STRUCTURE AND RETREIVING MECHANISM**
   1. **Data retrieving mechanism**

In this application, we use Graph API (version 2.11) supported by Facebook to retrieve data. Facebook (FB) allows applications to retrieve a personal profile or a page with limited data fields. However, applications first must get the user token from the user profile. We obtained this by design a login webpage, registered as FB App. By this way, every Facebook user could be invited to login through this webpage and accept our requested data fields. After all, we would have that user token and their permissions to retrieve necessary data, in our case, are posts, comments, and reactions.

We will request directly to FB Graph API (https://developers.facebook.com/tools/explorer/) with necessary fields to retrieve the data. We could retrieve as this URL: [https://graph.facebook.com/v2.11/me?fields=id,name,posts{comments{reactions,comments{reactions}},reactions}](https://graph.facebook.com/v2.11/me?fields=id,name,posts%7bcomments%7breactions,comments%7breactions%7d%7d,reactions%7d). We figured out there were two two ways of data retrieving:

* + 1. All-in-one retrieving: (1) phase only
       1. All the posts, comments, and reactions are retrieved at once

*Pros:* reducing requests from the FB server with more retrieved data at once

*Cons:* received data is in more levels and can be risk at data lost that result one or many retries. Codes are more complex to understand

* + 1. Sequential retrieving: (2) phases
       1. All the posts will be retrieved and saved to a file.
       2. Using the retrieve posts data to extract all post-id and use that post-id to retrieve other fields (comments and reactions). Fields that are retrieved in this phase will then append together and to each of the post.

*Pros:* codes are cleaner and better managed. Less amount of data per retrieve time that result a less risk of data loss, disconnection or retry.

*Cons:* conducting more request to server

* 1. **Data structure**

The user token obtained from the user login and acceptance would be transferred to our Main class to retrieve data. The would-be-retrieved FB data is in JSON format (Fig. 1). There are 2 levels of data, we only take care of the 2nd level that is “data”. The Fig. 1 shows the JSON data retrieve from a posts field.

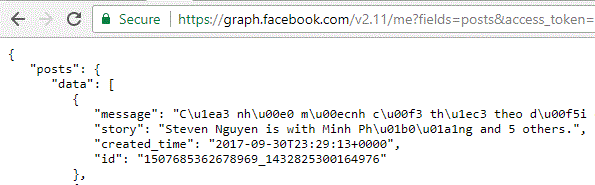


Figure . Example of a return of JSON (fields=posts)

As mentioned above, in phase 2, those data would be extracted to have the id and get other fields. Other fields would be appended continue to the “id” in this figure. In short, we will have a list of contents [list] (fig. 2), with each content have a list of message, story, created\_time, id, comments [list], reactions [list].

Figure . The maximum deep of data level could be supported by one retrieving time from

(Dark Red: Level 0; Red: Level 1; Orange: Level 2; Green: Level 3)

Figure . Facebook Data structure created by the application

* 1. **Data storing and encryption**

In the first phase, data would be retrieved from FB Graph API under HTTPS connection. Then, data will be stored under a file with the name as format {filename}.{extension}. The filename is a string obtained from a hash of SHA-256 followed by a Base64 URL-safe-encoding. The hash function uses FB user-id as password and FB user-name as salt. Therefore, this will help to secure the file information that even the developer could not know the end user information. This is a one-way filename encoding which allows an end-user must login into the system with their permission so that the app could get the user-name and user-id to hash (fig. 3)

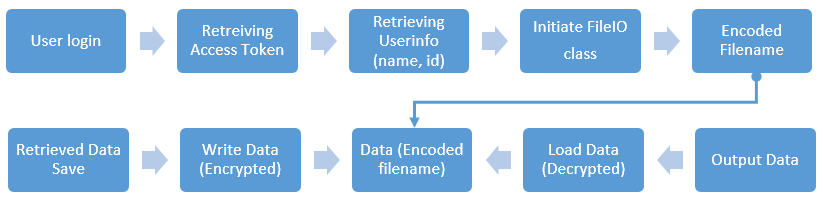


Figure . The process of file i/o based on an encoded filename (user-name, id)

For every phase, data will be saved to different file under different extensions (fig. 4). The scope of this is to store separately the posts file and the posts-comments-reactions file, which helps the app to save and read easier, avoid data losing and also for further development (scalable). The extensions are set as “.posts” for posts only file and “.postsNreactz” for posts-comments-reactions file. Those extensions could be set whatever as we want.



Figure . Data storing on file system

For privacy and security, we also implemented an encryption and decryption (fig. 5) third-party module and applied inside the functions that take the responsibilities on reading and writing the data. Therefore, data would be encrypted and saved to ensure no one except the end-user could load it. Besides, we also implemented a delete method, which will delete even the encrypted save data on the file system whenever the end-users finish their application uses sessions.

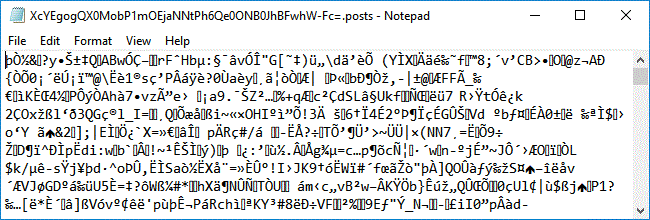


Figure . An encrypted file opened in Notepad

The module that is used for this is PyCryptodome (<http://pycryptodome.readthedocs.io>). File would be encrypted and decrypted based on AES 128bit key. We provided two options to have the secret key.

(1) The key is a 16-bytes key randomized by the app. Hence, this key would be changed from instances of user to user.

(2) The key is the first 16 characters of a combination of FB user-id and user-name. This 16 characters form the max 16-bytes key used in encryption.

By applying two different filename and data encryption, even us could not know which one to decrypted without knowing the user information or if the key is randomized, it could be very secured by the time writing this document. Adding to the security is a file deletion method that would be triggered after finishing data analysis and output to user.

1. **PROGRAMING LANGUAGE**
   1. **Language and working environment**

* Python was used to program this. This application was tested under Python version 3.6.x
* Python flask was used to program the web connection to Facebook.
* We used JetBrains PyCharm and Spyder (Anaconda) to program the application.
* Git-hub was used as a version control for our team works.
  1. **Dependencies**
* To run the code we programmed, please install those dependencies as in one command line using pip or conda
  + pip install flask flask\_oauthlib matplotlib numpy pandas urllib json hashlib base64 bokeh==0.12.2 sklearn
  + conda install flask flask\_oauthlib matplotlib numpy pandas urllib json hashlib base64 bokeh==0.12.2 sklearn

1. **MODULES AND CLASSES**
   1. **Built-in modules**
   2. **Third-party modules**
      1. PyCryptodome: <http://pycryptodome.readthedocs.io>

This module is used to encrypt and decrypt the data content, which were retrieved from the FB, before writing to or reading from file system.

* 1. **Self-built modules**
     1. Core

This module is used to retrieve the JSON data from Facebook, handling errors and retries. Public functions could be triggered as:

+ getUser() *return* a <list> userinfo (name, id)

+ getPosts() *return* a <list> data of all posts from the user

+ getComments() *return* a <list> data of all comments from the user

+ getReactions() *return* a <list> data of all reactions from the user

The fig. 6 demonstrate how all posts would be retrieved from FB.

1. The getPosts() will call the \_getData().
2. This \_getData()
   1. Call the \_prepareURL() function to check which kind of data (posts, comments, reactions) would be retrieved and this function will prepare also a string-type URL to retrieve the JSON data from FB Graph API.
   2. Call the \_getFields() to get the string-type field of which would be retrieved. The JSON would be retrieved as dictionary data type. Therefore, this string is used to know which key to extract the data from that dictionary and append to the final data (<list>)
   3. Call the \_getJSON() after having all necessary data (URL and fields)
      1. There is a while loop inside this function to limited-retry whenever it fails to reach the server. The number of retries and the time (in seconds) between every retry could be set in \_\_init\_\_(). This is very useful because the server might down at any time and the whole data retrieving process could be lost and have to be re-done from start. There is another error handler inside to catch whenever an error or wrong JSON, lost connection, expired access token… happens

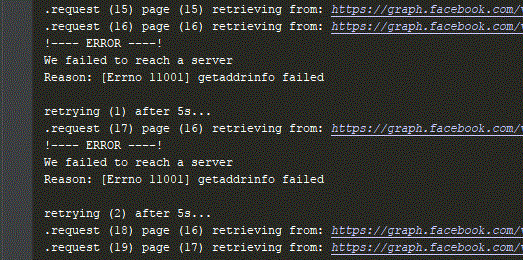


Figure 7. A screenshot as an example of auto-limite-retries by the app

* 1. Call the \_getNextURL() to get the ‘next’ keyword in the retrieved dict-type JSON from facebook if there is any other next page because FB will not return all data at once.

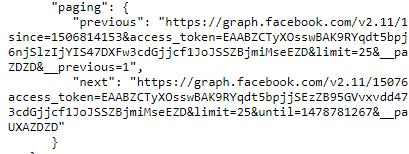


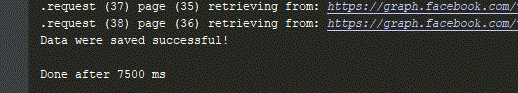
Figure 8. Paging keyword showed in returned JSON in case of any sub-page exists

* 1. Starting a loop with the \_getNextURL() condition, each time, data is retrieved and append to the final data
  2. Call \_checkError() to printout error (if any)
  3. In case the data is None, return an empty list [], otherwise, return data

Figure . Process of retrieving Posts

After the posts are retrieved, as mentioned, post id inside the posts data would be extracted to retrieved other data (comments, reactions). This process is almost the same as the process above, except that, the getComments or getReactions will call \_getSubData(). This \_getSubData() holds a <for loop> that run the \_getData() in the loop.

This happens because for any comments or reactions field of a user posts, there could be one or another sub-dictionary containing more than one reactions or comments. Therefore, for every post id, this will make a loop of \_getData(), in which this \_getData() will get all comments or reactions. Everytime the \_getData() returns result, \_getSubData() would append into the parent data (as in fig. 2). A separate loop like this helps to control things easier instead of integrating everything into \_getData(). After finish, there would be messages printed out. In case of comments and/or reactions are retrieved, there would be a so-called Batch number, which increase every time a comment/reaction is traverse in a post.



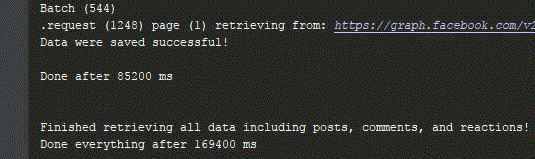


Figure 10. Retrieving finish messages. Above: finished the posts request alone and saved. Below: finished all data and saved

* + 1. FileIO

This module is used to handle file reading and writing. This file also ensures that data would be saved and open in UTF-8 encoding. Filename and data content would be encrypted and decrypted by this class. Public methods/functions could be triggered:

+ writeToFile() encrypting the file and writing the file into the file system

+ readFromFile() *return* a <list> decrypted data content loading from file system

+ deleteFile() auto delete all files relating to a user

+ getFilePath() *return* a <str> path of the file without extension

* + 1. Main

This module implements other modules (Core, FileIO) to retrieve, save and load data from the internet or from the saved data files.