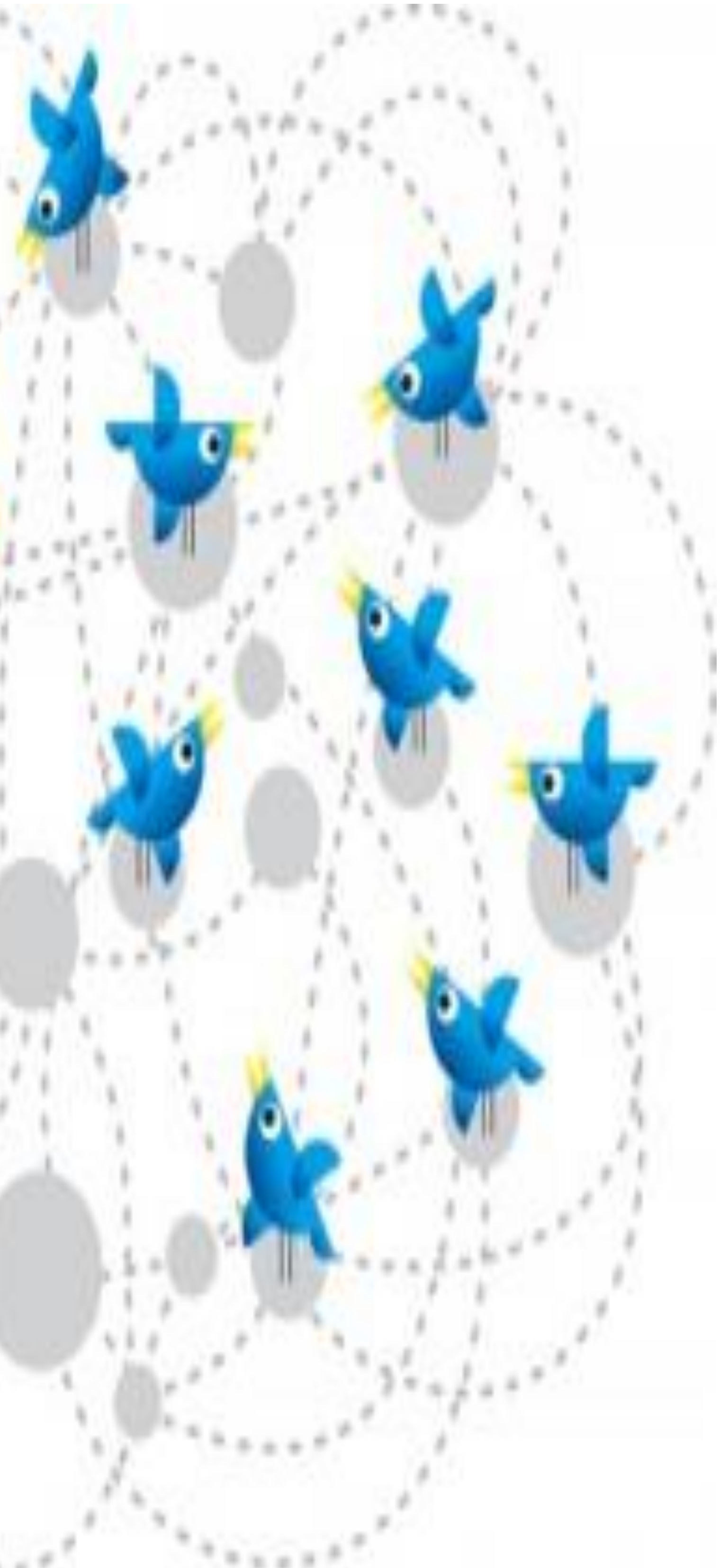


Twitter Data Crawler

MASTER’S 1 FINAL PROJECT

Academic Year 2018/2019

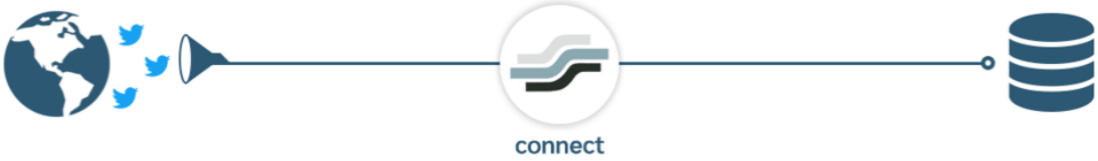
By Maria Afara and Ali Fakih
Supervised by Dr. Ali Choumane

Acknowledgements

This software would not have come into being without the guidance of our supervisor Dr. Ali Choumane. He has been very supportive and his feedback has greatly improved our work.

Chapter 1: Introduction

Twitter is an online social networking/media site that allows users to send and read messages called "tweets" in real time. It is what's happening in the world and what people are talking about right now. Its popularity as a fast information dissemination platform has led to applications in various domains (e.g., business, disaster recovery, intelligent transportation, smart cities, military scenarios, etc.). Twitter data constitutes a rich source that can be used for capturing information about any topic imaginable. This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products, services and users. Users on Twitter are generating about half billion tweets every day. Some of these tweets are available to researchers and developers through Twitter's public APIs (application programming interfaces).



At a high level, APIs are the way computer programs “talk” to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an endpoint: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers). Twitter allows access to parts of its service via APIs to allow people to build software that integrates with Twitter, like in our software.

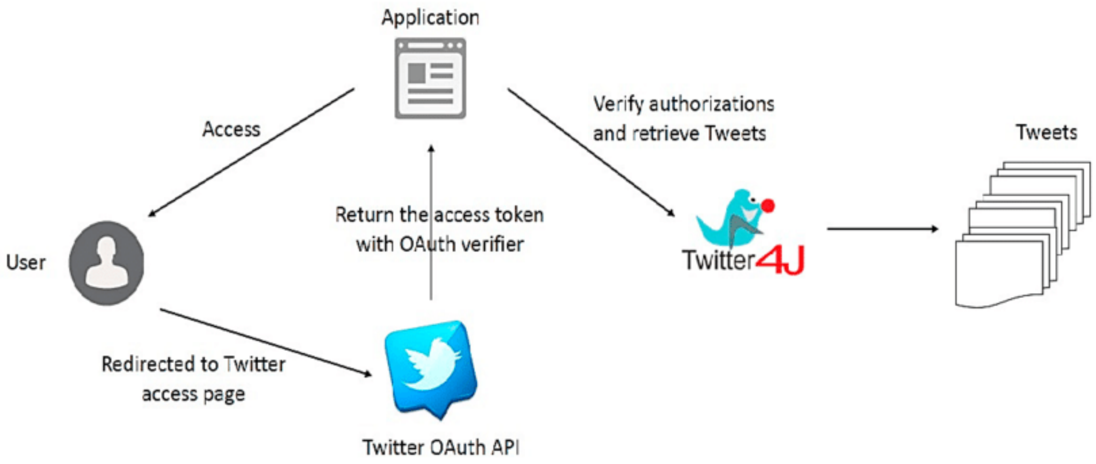
1-Using Twitter streaming Api with twitter4j

Streaming APIs are becoming more popular among application developers because of their low latency. One such emerging API is the Twitter API (version 1.1), which is highly popular among developers due to its low latency access to global stream of Tweet data. Streaming APIs are famous for their real-time streaming nature, which makes it possible for developers to access Tweets and related information as they are created by their respective users.

1-1 Accessing Twitter Data

To access twitter APIs, it is required to register an **application**. It is done through these steps:

1. Create a Twitter app to access Twitter Streaming API (https://apps.twitter.com/)
2. Download Twitter4j
3. generate/re-generate both Consumer key/secret and Access Token/Token secret



This completes App creation and these Access information will be required later when writing the java application.

Chapter 2: Business specification

1-Main Features

- Collect a user's profile information from Twitter given the user's Twitter ID or file of user's Twitter IDs.
- Collects a user's tweets given the user's ID or file of user's Twitter IDs.
- View user's social network informations given a file of tweets for a certain user.
- Collects the similar tweets of a given tweet.

2-Principle of the software

The aim of Twitter data crawler software is as its name states to crawl data which can be either tweets or friends to collect datasets. This is done by entering a specific criteria which differs between each of the 4 pages in the software which are:

- Login page
- Graph page
- User's page
- View tweets page
- Similar tweets page

And during the crawl process the feedbacks can be observed of the crawl that is taking place, and the crawl can be stopped, resumed or even canceled.

As stated above the criteria that needs to be entered in order to crawl, differs between, the page and what needed to be crawled so we will sate the principle of each page.

2-1 Login Page

In This page the user should add his twitter developer account to access to the application.

He can save his account(s) in a XML file so in this case when the user want to open the application he only need to browse his accounts from the file instead of writing his twitter developer account every time.

The format of the file should be like this:

```
<Apps>
  <App>
    <ConsumerKey>xxxxxxxxx</ConsumerKey>
    <ConsumerSecret> xxxxxxxxx </ConsumerSecret>
    <AccessToken> xxxxxxxxx </AccessToken>
    <AccessTokenSecret> xxxxxxxxx </AccessTokenSecret>
  </App>
  <App>
    <ConsumerKey>xxxxxxxxx</ConsumerKey>
    <ConsumerSecret> xxxxxxxxx </ConsumerSecret>
    <AccessToken> xxxxxxxxx </AccessToken>
    <AccessTokenSecret> xxxxxxxxx </AccessTokenSecret>
  </App>
  …
<Apps>
```

2-2 Graph Page

2-2-1 Page Description

The main idea of this page is to crawl the graph of friends, friends of friends…. for a specific twitter user Account or a file that contains multiple twitter user id.

We use the Breadth first search algorithm to build the graph.

When the input is a single id the graph mean that we will start crawling from the root node (Twitter User Id) and building the graph until we reach the target depth chose by the user.

But when the input is file of IDs then each of those IDs have its specific graph.

At the right position of the GUI there's a window that show the feedback happened while crawling like:

- Number of nodes crawled.
- Current depth of the graph.
- The number of crawled nodes for each input id.
- Any Exception that may happen while crawling like:
    - Account Is Temporarily Locked
    - User has been suspended!
    - Node Doesn't Exist!
    - There's No Internet Connection
    - Account Exceed The Limit Of Queries!
    - And others…

At the top left position there's the number of twitter developer accounts that are active.

2-2-2 User Guide

Number of accounts

At any moment, even while crawling a graph the user can press the plus button (Top of the window) and add another twitter developer account.

PS: the user can benefits from this feature when all his working accounts exceeded the limit of queries, so if he add an account the graph crawling will continue without waiting 15 min (The time needed to reset the number of queries for each twitter developer account).

Input

In the first input the user can write one of this two options:

- Twitter User Id   Ex: 12.
- Screen Name of Twitter user in the following format:
    - https://twitter.com/Cristiano
    - Cristiano
    - @Cristiano

Or He can left this input empty and choose a file of Twitter Users IDs (.txt File).

The format of the file is very simple, the file has no header and each Id is located in a line (Each line contains single ID).

Depth

The user should choose the depth of the graph he want to build and the maximum value if 20.

PS: If the depth is 5 and the average number of friends of each user is 200 then the graph will contain about 320,000,000,000 connection.

Save File

In both input case (User Id or File of IDs) the output will be saved in a single file as follow (Source Destination) per line.

The user should choose where to save the output file (Path) and in which format (.txt or .csv).

Feedback

The user can extract all the feedback and save it in a file at any moment.

Start Crawling Button

Needed to start the graph crawling.

Stop Crawling Button

If the user press this button while crawling the reached result will be save in the file.

Resume Crawling Button

This button will be clickable only after clicking the stop crawling button or exiting the application while crawling. So if the user want to continue his crawl from the stopped point he can press this button and his work will continue normally.

Cancel Current Crawl

This button will be clickable only after clicking the stop crawling button or exiting the application while crawling. So by clicking this button the feedback window will reset and all the inputs fields will be enabled and ready for a new crawl.

2-3User's Page

2-3-1Page Description

The main idea of this page is to crawl the tweets for a specific twitter user Account or a file that contains multiple twitter user id. It crawl at most the latest 3200 tweets (Limited chose by Twitter).

The user can choose the dates range and language of tweets that he wants to crawl.

At the right position of the GUI there's a window that show the feedback happened while crawling like:

- Number of nodes crawled.
- Current depth of the graph.
- The number of crawled nodes for each input id.
- Any Exception that may happen while crawling like:
  - Account Is Temporarily Locked
  - User has been suspended!
  - Node Doesn't Exist!
  - There's No Internet Connection
  - Account Exceed The Limit Of Queries!
  - And others…

At the top left position there's the number of twitter developer accounts that are active.

2-3-2User Guide

Number of accounts

At any moment, even while crawling a graph the user can press the plus button (Top of the window) and add another twitter developer account.

PS: the user can benefits from this feature when all his working accounts exceeded the limit of queries, so if he add an account the graph crawling will continue without waiting any minutes (The time needed to reset the number of queries for each twitter developer account).

Input

In the first input the user can write one of this two options:

- Twitter User Id Ex: 12.
- Screen Name of Twitter user in the following format:
  - https://twitter.com/Cristiano

- Cristiano
- @Cristiano

Or He can left this input empty and choose a file of Twitter Users IDs (.txt File).

The format of the file is very simple, the file has no header and each Id is located in a line (Each line contains single ID).

Date

The user should choose minimum and maximum dates of tweets that he wants to crawl.

Language

The user should choose the languages of tweets that he wants to crawl.

He can check the checkbox 'All', Therefore the application will crawl all tweets regardless its language.

Or he can specify the languages (from the checkComboBox) of tweets that he want to crawl.

Save File

The user should specify the directory that will contains the crawled tweets.

For each input User Id will be a JSON file that contains all its tweets.

Feedback

The user can extract all the feedback and save it in a file at any moment.

Start Crawling Button

Needed to start the tweets crawling.

Stop Crawling Button

If the user press this button while crawling the reached result will be save in the files.

Resume Crawling Button

This button will be clickable only after clicking the stop crawling button or exiting the application while crawling. So if the user want to continue his crawl from the stopped point he can press this button and his work will continue normally.

Cancel Current Crawl

This button will be clickable only after clicking the stop crawling button or exiting the application while crawling. So by clicking this button the feedback window will reset and all the inputs fields will be enabled and ready for a new crawl.

2-4View Tweets Page

2-4-1 Page Description

The main idea of this page is to view the tweets for a specific person from a file which is written in json format.

All the user have to do is to select the file he wishes to view, then the basic information of the person will be viewed along with list of all of his tweets.

For each tweet the date and time it is tweeted are listed along with the tweet text itself and the number of likes and retweets the tweet have and of course a link that direct you to the real tweet post on the default browser.

2-4-2 User Guide

With a click of a bottom which is placed on the top center we can select a user file.

Once selected the file is being loaded and then the informations will be seen sequentially, the user's profile (profile image, screen name, name and description) first then the tweets where they can be scrolled down and for each tweet we can right click on it to copy the tweet text.

2-5 Similar Tweets Page

2-5-1 Page Description

The main idea of this page is to crawl the similar tweets for a specific tweet id or url .

Only the tweets which are similar to this inputted tweet will be crawled and according to specified dates chosen.

At the right position of the GUI there's a window that show the feedback happened while crawling like:

- Number of tweets crawled.
- Current tweet being looking for it similar tweets.
- Any Exception that may happen while crawling like:
    - Account Is Temporarily Locked
    - User has been suspended!
    - There's No Internet Connection
    - Account Exceed The Limit Of Queries!
    - Limit Exceeded
    - Sorry, that page does not exist.
    - When a Tweet cannot be viewed by the authenticating user, usually due to the Tweet’s author having protected their Tweets.
    - The status text is too long.

At the top left position there’s the number of twitter developer accounts that are active.

2-5-2 User Guide

Number of accounts

At any moment, even while crawling for a similar tweet the user can press the plus button (Top of the window) and add another twitter developer account.

PS: the user can benefits from this feature when all his working accounts exceeded the limit of queries, so if he adds an account the graph crawling will continue without waiting 15 min (The time needed to reset the number of queries for each twitter developer account).

Input

In the first input the user can write one of this two options:

- Tweet Id
- Tweet url in the following format:
    - https://twitter.com/username/status/tweetid

Or choose a file that contains tweets ids separated by enter .Each line contains single ID. (.txt File).

Depth

The user should choose the depth of the graph he want to build and the maximum

Between 2 Dates

The user can specify between what dates does he wants the crawled tweets to be in.

Save File

The user should choose in what format does he wants to save the results, either in json or text format (put in radio bottoms).

In each case the user also should first select the folder he wants to save the result in, then for a single tweet id a folder will be generated holding its name and each crawled similar tweet will be saved in a single file either .txt or .json according to what has been chosen.

Feedback

The user can extract all the feedback and save it in a file at any moment.

Start Crawling Button

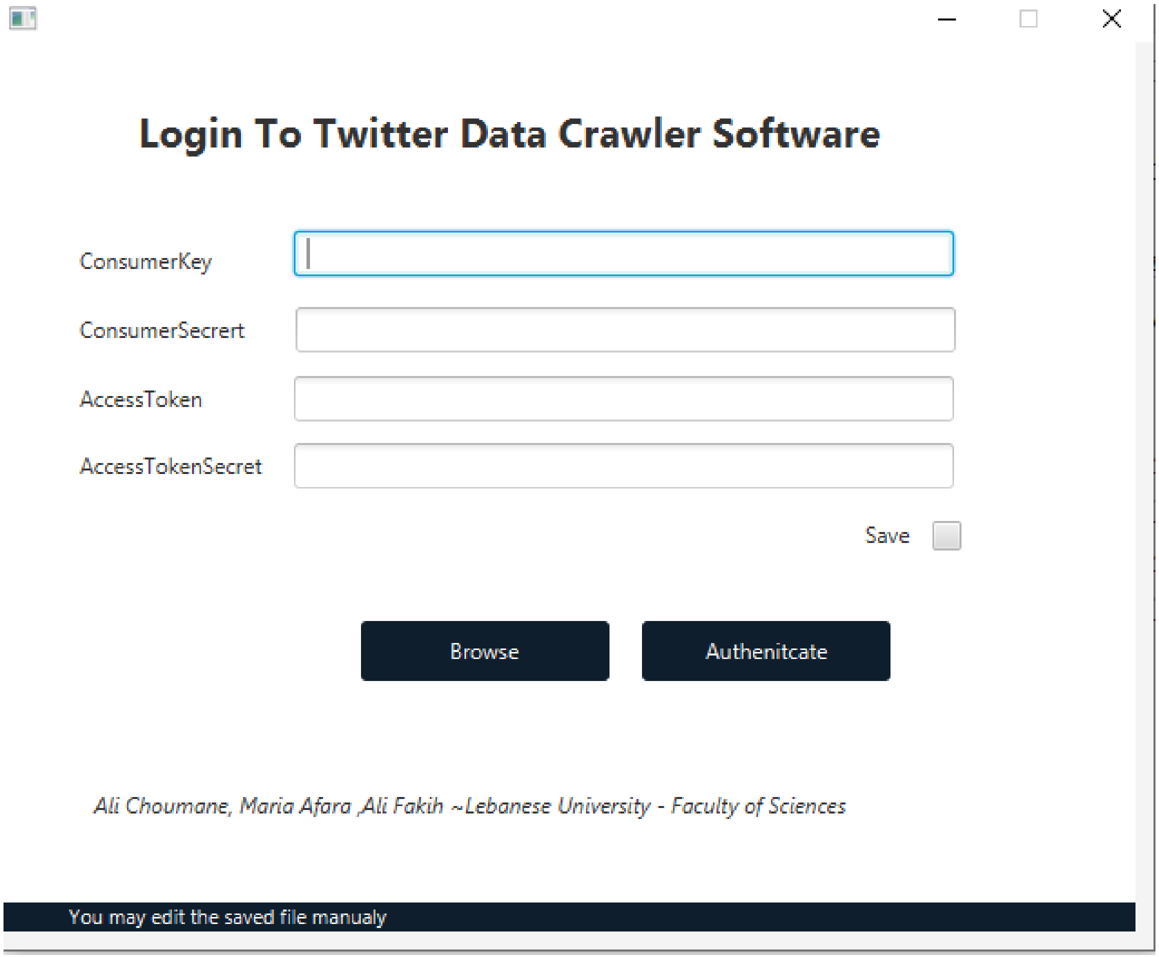
Needed to start the crawling.

Stop Crawling Button

If the user press this button while crawling the crawl is stopped and only those which are already been crawled will be saved.

Chapter 3: Screen Shots

1-Login Page

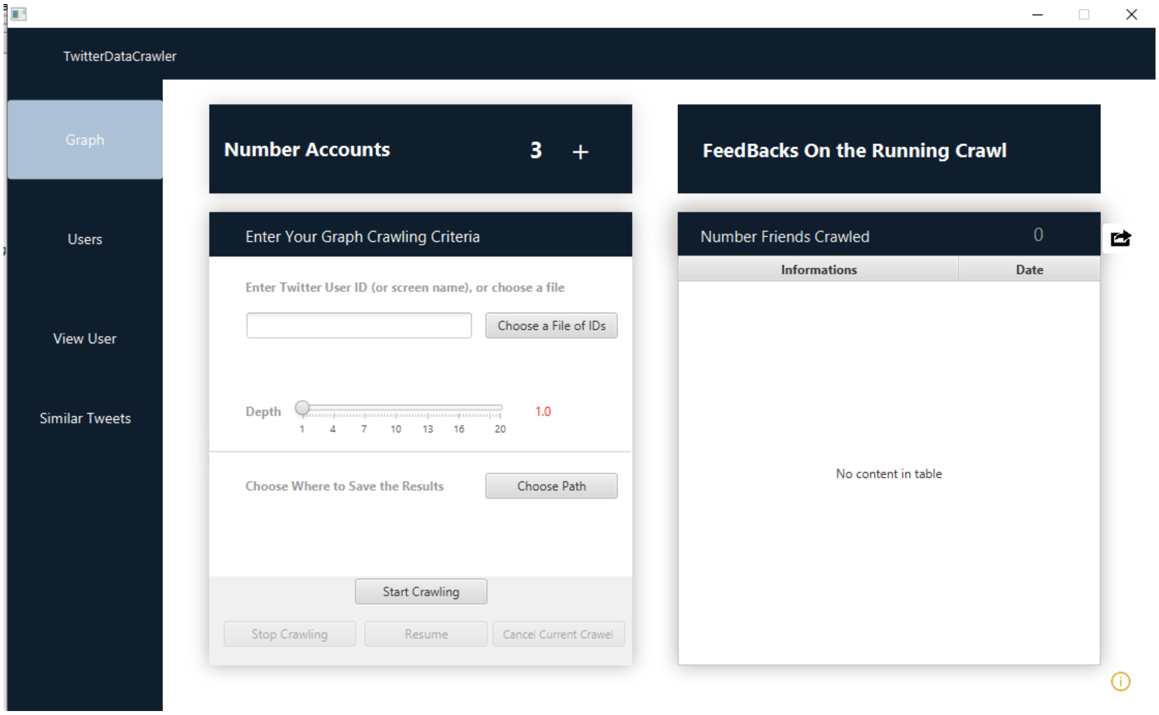


Login To Twitter Data Crawler Software

Ali Choumane, Maria Afara ,Ali Fakih ~Lebanese University - Faculty of Sciences

You may edit the saved file manualy

2- Graph Page



TwitterDataCrawler

Number Accounts 3

FeedBacks On the Running Crawl

Enter Your Graph Crawling Criteria

Enter Twitter User ID (or screen name), or choose a file

Depth 1.0

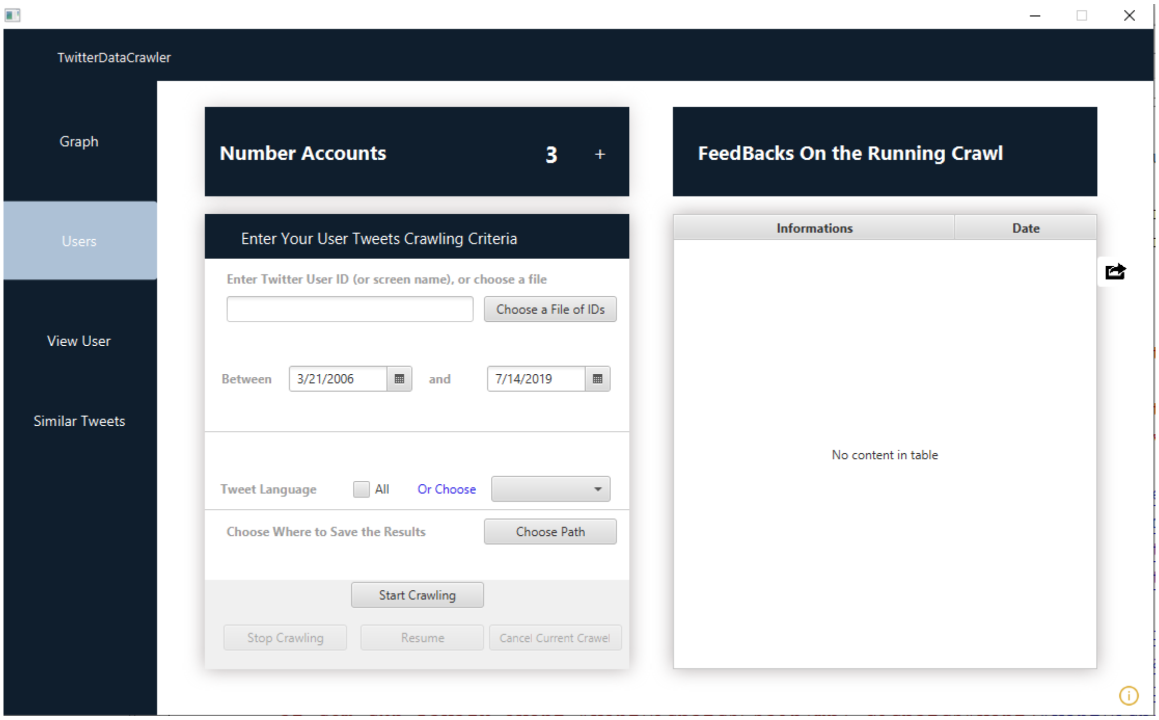
Choose Where to Save the Results

Number Friends Crawled 0

Informations	Date

No content in table

3-User’s Page



Number Accounts 3

FeedBacks On the Running Crawl

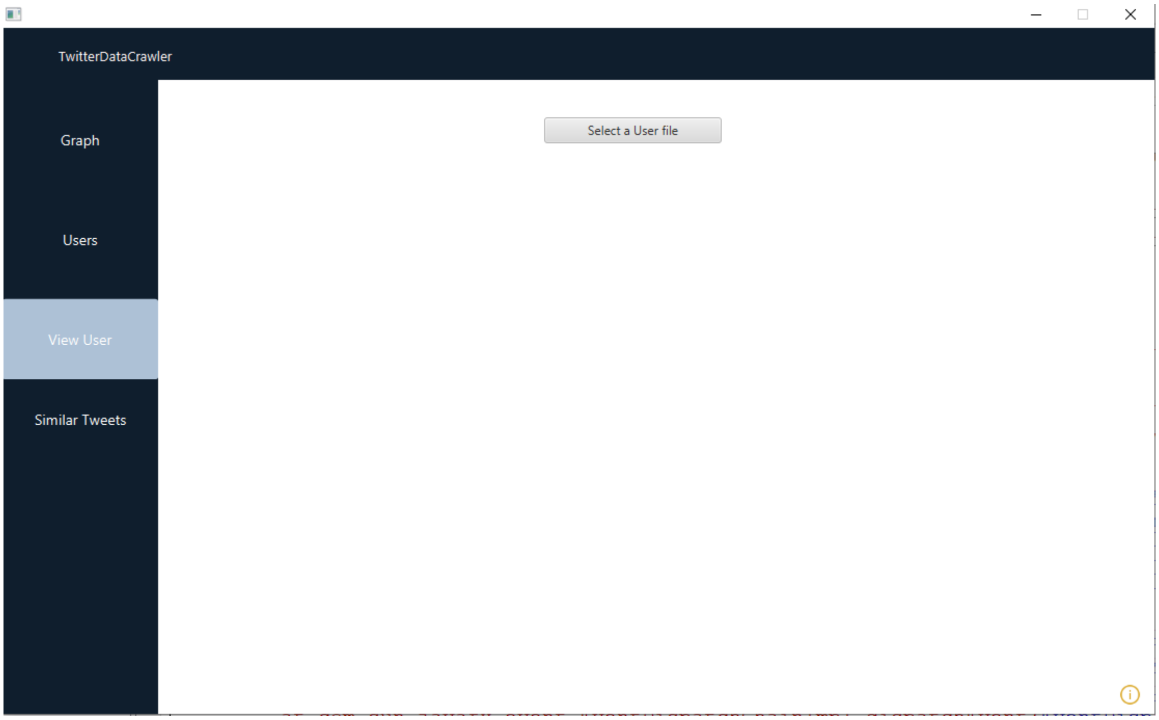
Enter Your User Tweets Crawling Criteria

Enter Twitter User ID (or screen name), or choose a file

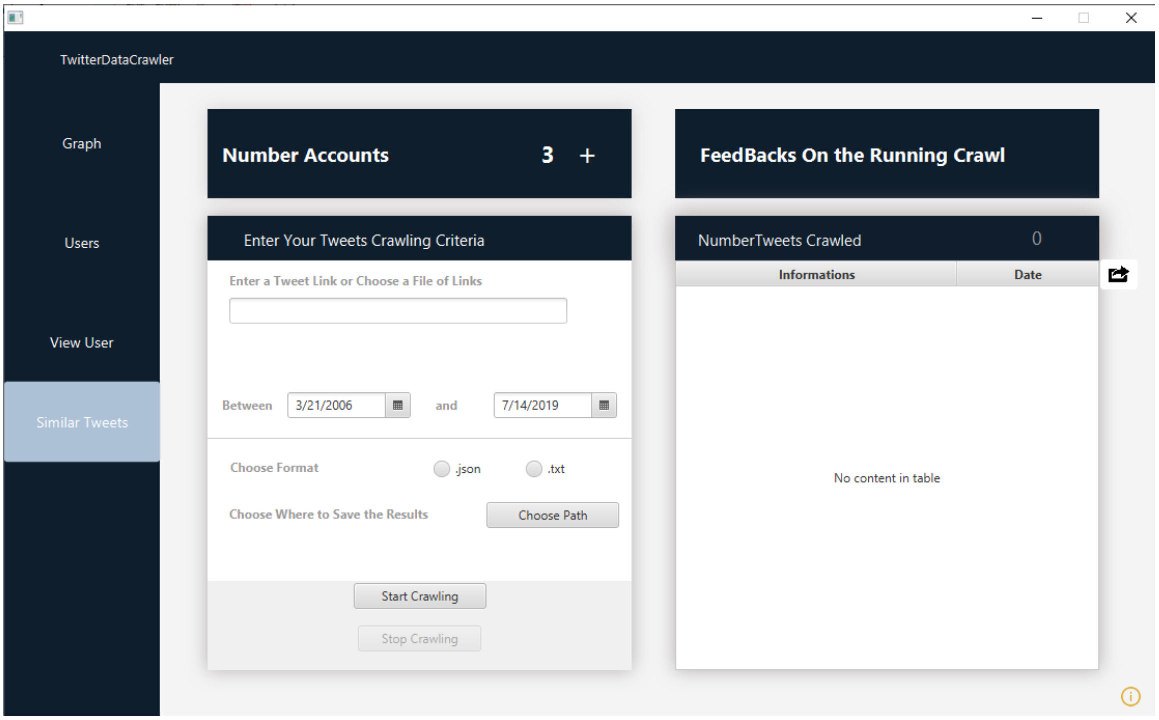
Between 3/21/2006 and 7/14/2019

No content in table

4-View Tweets Page



5- Similar tweet Page



Number Accounts 3 +

FeedBacks On the Running Crawl

Enter Your Tweets Crawling Criteria

Enter a Tweet Link or Choose a File of Links

Between 3/21/2006 and 7/14/2019

Choose Format ○ .json ○ .txt

Choose Where to Save the Results

NumberTweets Crawled 0

Informations	Date

No content in table