VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY HO CHI MINH UNIVERSITY OF TECHNOLOGY



PROGRAMMING INTERGRATION PROJECT

Instructor: Quản Thành Thơ

Huỳnh Ngọc Thiện

Author:

Name	Student ID	Class
Nguyễn Nam Kha	2052515	CC02
Nguyễn Trần Minh Phước	2052662	CC02
Phan Mai Tấn Lợi	2052158	CC02
Cao Anh Huân	1952712	CC02
Nguyễn Duy Thành	1952456	CC02

Ho Chi Minh city, 12/2022

CONTRIBUTIONS

Name	Student ID	Contribution
	2052515	Define Hook in React. Handle
Nguyễn Nam Kha		interactions between front-end and
		back-end, manipulate data get by calling
		API from server and Swagger
		Define front-end routes using React
Nguyễn Trần Minh Phước	2052662	router, styling page using Sass, drawing
		chart using Chart.js
Nguyễn Duy Thành	1952456	Define model (user, project, grouppage,
		baipost) of the app using MongoDB
		Atlas. Write instruction markdown.
Phan Mai Tấn Lợi	2052158	Add logic, CRUD operation for user,
		project by using Node.js. Draw use case
		diagram, class diagram, MVC .
Cao Anh Huân	1952712	Add logic, CRUD operation for
		grouppage, baipost by using Node.js.
		Draw activity diagram.

Contents

1. Introduction	1
2. Design and technology	2
2.1. Use case diagram	
2.2. Class diagram	
2.3. Activity diagram	
2.4. Architectural approach	
2.5. Technologies	7
3. Project outcome	8
3.1. Log in / Sign up page	8
3.2. Home page	9
3.3. Project page	11
3.4. Group page	14
3.5. Post page	17
4. Summary and further development	18

1. Introduction

Crawl data is the practice of gathering data and information from websites for a variety of uses. As a result, all pages on the website as well as relevant links will be accessed by search engine bots (Search Engine) like Google, Bing, etc. for data statistics.

This is a crucial step for both SEO and search engine optimization because it provides the most accurate evaluation of the website's quality. Search rankings are improved and made more precisely as a result.

Web crawlers in particular are extremely "sensitive" to new websites or updates to well-known websites. These technologies also have strict control over links that are no longer available because they directly affect the caliber of the pertinent website.

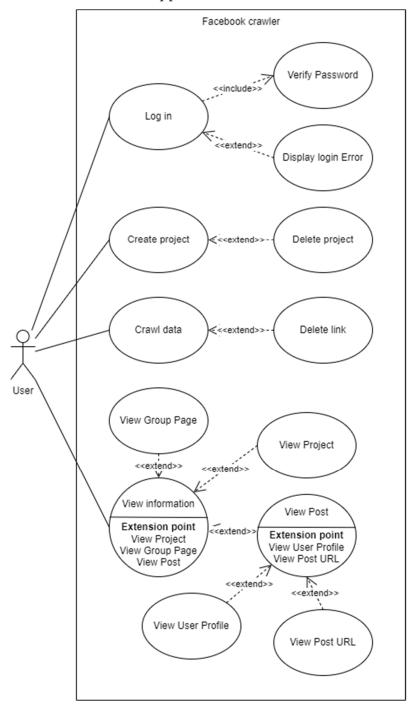
Currently, search engines frequently perform data crawling automatically and independently. In other words, it is challenging for any user or company to obstruct Google, Bing, etc. from collecting data.

In this project, we will try to crawl data from facebook to understand better how big companies we mentioned earlier collect data. With the help of Swagger API, which was created by Mr Huynh Ngoc Thien, we only need to create a database to store data and an app to display this data.

2. Design and technology

2.1. Use case diagram

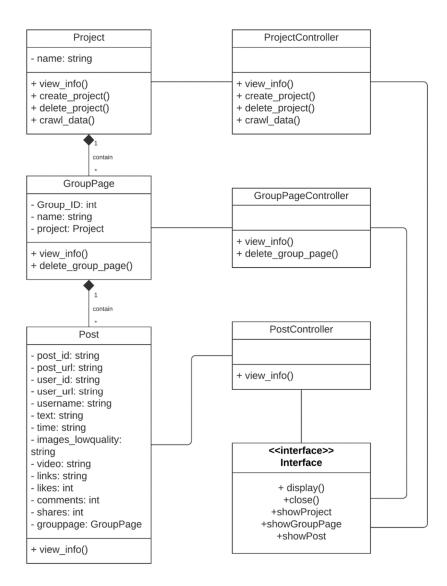
This is all use cases of our application:



Use-case scenario for *Crawl data module*:

Use - case name:	Crawl Data	
Description:	User crawls data from a group page by giving the group page link and Facebook cookie.	
Actor:	User	
Trigger:	Click on "Create" button below the Input Links bar	
Preconditions:	 Input Links must be correct and directed to the available group page (chosen group page must be public). Facebook cookie text file is required. 	
Postconditions:	The user successfully crawls data.	
Normal flow:	 The system shows an Input Links box to receive links and a list of links from which the system has crawled data. The user gives the chosen link in the Input Links bar and cookie text file by clicking Choose file. The user clicks the "Create" button to begin crawling data progress. The loading icon appears and the "Create" button gets disabled. The system crawls data and stores them in the database. The system updates the link list and shows them to the user. 	
Alternative flow:		
Exceptions:		

2.2. Class diagram

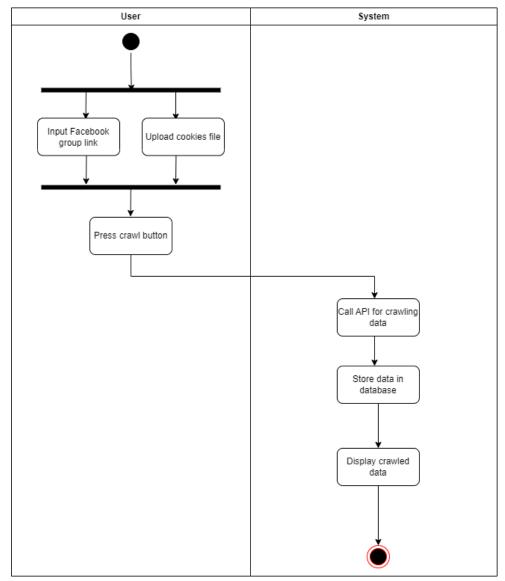


Description:

- The user can view the project, group page or post using view_info().
- The user can create any project using create_project().
- The user cannot create a group page or post. New group pages and posts can only be created via crawling data progress in the project.
- Posts are attached to the group page and group pages are attached to the project, which means deleting a group page also deleting all posts in that group page and deleting a project also deleting all group pages in that project. The user cannot delete posts.

2.3. Activity diagram

This is the activity diagram for Crawling data function:

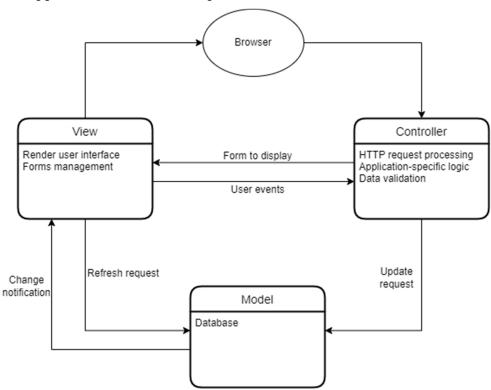


Description:

- The user can start crawling data by giving out a Facebook group link and upload cookies file, then press the Crawl button.
- The crawling progress includes calling the API for crawling data and storing them in the database.
- The system finally displays the result to the user.

2.4. Architectural approach

Our application follows MVC pattern as bellow:



	Description
Model	Model is responsible for managing system data and interacting with
	databases as well as responding to requests from Controller.
	- Database: Functions that interact with the database such as
	create or update data (project, group page, post,).
View	View defines layout, displays application data through UI, processes
	user's interaction (click, scroll,) and sends them to controller.
	- Render user interface: Provide UI to display data. In our
	application, it is responsible for displaying login screens, project
	lists, group page lists, post lists, posts' contents,
	- Forms management: Handle user's input and send to controller.
Controller	Controller manages user's interaction and contains logic that
	updates the Model or View in response to input from the users.
	- Application-specific logic: Mapping user's action to model
	updates or view updates.
	- Data validation: Validate user's input such as username, password,

2.5. Technologies

We use **Node.js**, **Express** for Backend; **React** for Frontend; **MongoDB Atlas** for databases; **Swagger API** for crawling data.



Node.js is an open-source, cross-platform JavaScript runtime environment. With npm(Node Package Manager), we have installed many useful packages that supported us during the development of the app, such as bcrypt, cors, jsonwebtoken, mongoose, axios, chart.js, form-data, sass,...



Express is a minimal and flexible **Node.js** web application framework that provides a robust set of features for web and mobile applications such as building RESTful APIs



React is a JavaScript library for building user interfaces based on UI components. The most characteristic feature of React is that it supports developing a Single-page application (SPA), which interacts with the user by dynamically rewriting the current web page with new data from the web server, instead of loading entire new pages, thereby providing faster transitions between pages and avoid wasting resources.

mongo DB.

MongoDB is a NoSQL database using JSON-like documents with optional schemas, so it offers a more flexible way in designing the data structures comparing to relational database tables. MongoDB Atlas is a fully-managed cloud database that handles all the complexity of deploying, managing, and healing your deployments on the cloud service provider of your choice (AWS, Azure, and GCP). MongoDB Atlas is the best way to deploy, run, and scale MongoDB in the cloud.



Swagger API is a simplified API developed by Mr. Huynh Ngoc Thien, who is our advisor for this project. Swagger is an open source and professional toolset for users, teams, and enterprises. Mr. Thien's API allows our website to crawl data of a specific Facebook group based on input of the user.

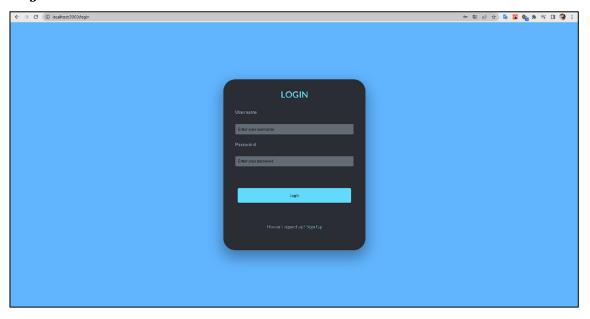
3. Project outcome

All of our code and set up instruction can be found here:

https://github.com/VideySora/fb-crawler

3.1. Log in / Sign up page

Log in



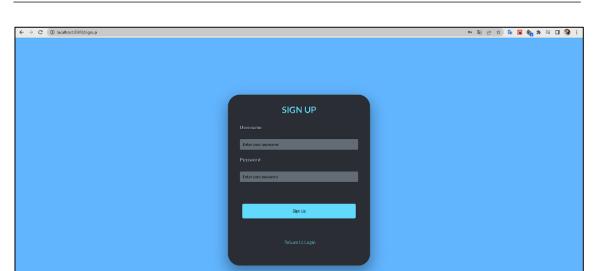
Login form contains a username, a password and a login button which is required to check identity.

To login successfully, an account is required, we can create an account via sign up button

Sign up

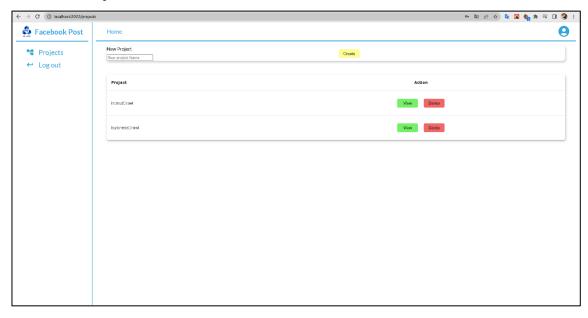
Similar to log in, sign up form requires a username and a password; pressing the Sign up button to send the form to the back end. We can always go back to the login interface by clicking the "Return to Login" button.

After successful login confirmation, the home interface which is also the Create project interface will be displayed.



3.2. Home page

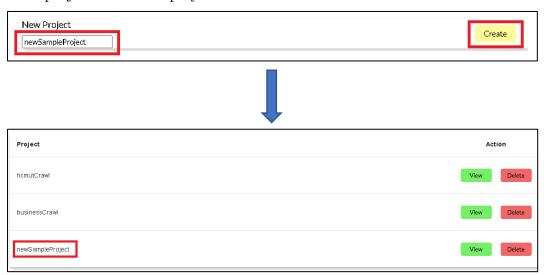
Home interface



This is the home interface, which is also the project list. Here we can create and view our projects. The created projects will be stored and displayed once we create them.

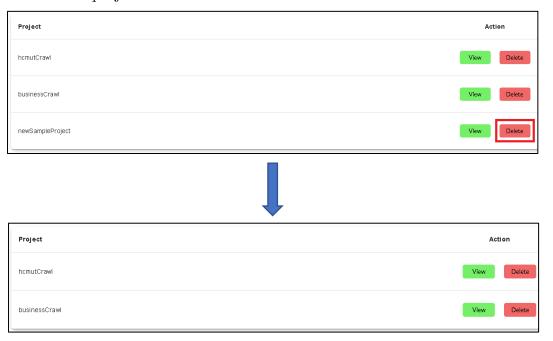
Create project

Below the "New Project" on the top is the New project Name bar which indicates the name of the project we want to create. To create a new project, we enter the name into the mentioned bar and press the Create button, the newly created project will be displayed as below:



Delete project

Delete button in the project section sends a delete request to the controller to remove the project from the database.

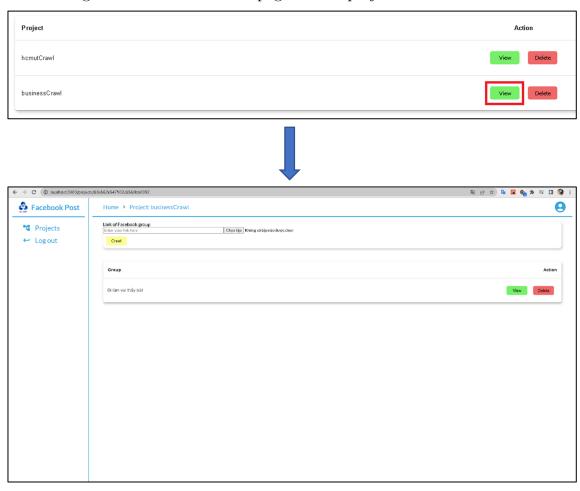


Deleting a project includes deleting all data which were stored in the project, such data will be discussed further down below.

3.3. Project page

View project

From the Home page, we can view a project by pressing the View button, which navigates to the information page of that project



This is the information page of project *businessCrawl*. The page show information of all crawled groups. Currently in this project there is only one group called "Di làm vui thấy bà". Now we will try crawling another group called "Ăn sập Sài Gòn", which has the following url: https://www.facebook.com/groups/219499848640636

Download cookie file

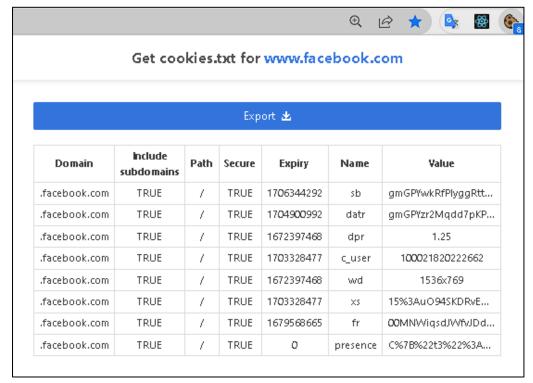
First, we have to download the cookie file. To download cookie, we use this extension from Chrome



Here is the link to install the extension:

https://tinyurl.com/yjmmmxfr

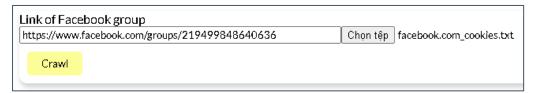
Then open Facebook on your browser, click on the Extension icon, we can see an interface like this:



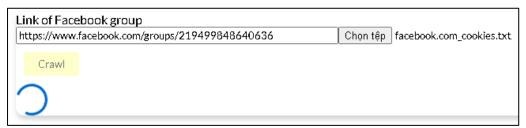
Click the *Export* button, the .txt cookie file will be downloaded to your device.

Crawl data

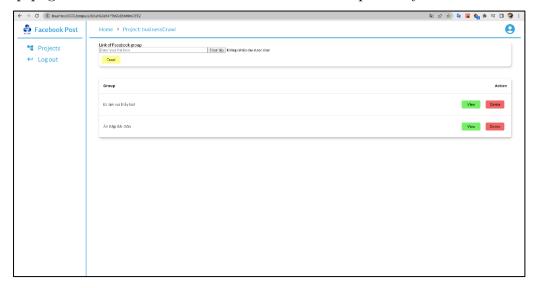
Navigate back to fb-crawler website, upload the .txt cookie file and input link of your desired facebook group to the input field, like below:



After clicking the *Crawl* button, the interface will show a spinning circle indicating that the crawling data is in progress. During this time, the Crawl button will be grayed out so users cannot spam the Crawl button. The progress may take some time, from a few seconds up to a few minutes.

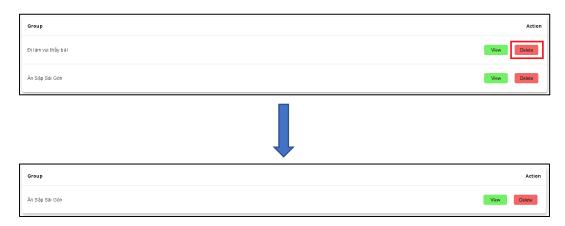


Once finished crawling, the group list will be updated immediately and show group page names and actions that the user can do optionally.



Delete group

Similar to projects, we can View or Delete Group pages whenever we want.

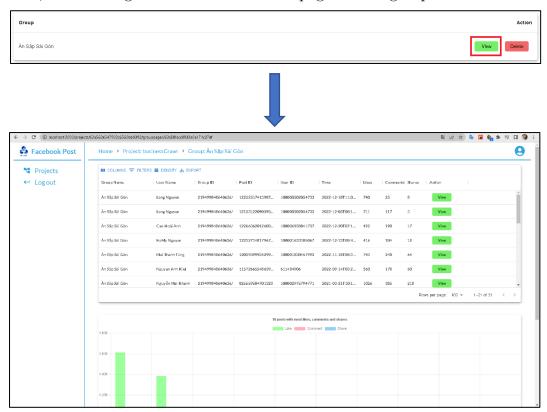


The system will remove the link and its data in the link list from the database and update the shown link list to the user.

3.4. Group page

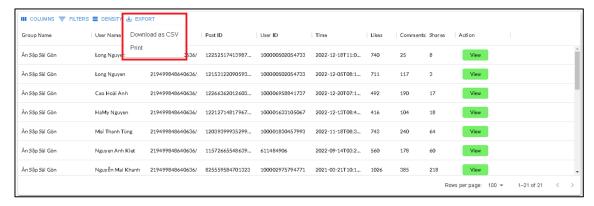
View group

From the Project page, we can view a specific group by pressing the View button, which navigates to the information page of that group.



The Group page section contains a table of posts that we were trying to get the data from. The list shows the information of each post, which includes Post ID, User name, Time, number of Likes, Comments and Shares.

Export CSV

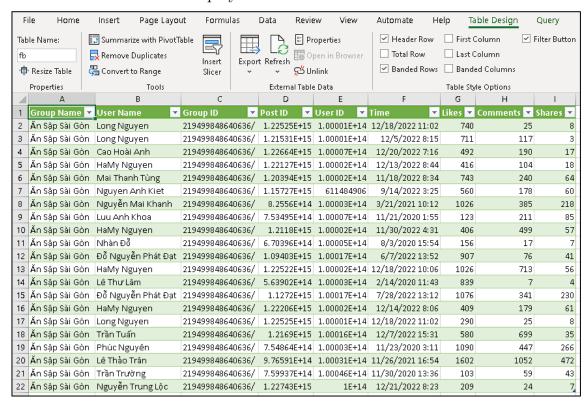


The user can export the list as a CSV file or print it, which will contain all the information shown in the table.

Here is the .csv file displayed in VS Code, every data entry is separated by commas:

```
sers > APC > Desktop > 🗟 fb.csv
 Group Name, User Name, Group ID, Post ID, User ID, Time, Likes, Comments, Shares
 Ån Sập Sài Gòn,Long Nguyen,219499848640636/,1225251741398770,100000502054733,2022-12-18T11:02:10,740,25,8
Än Sập Sài Gòn, Long Nguyen, 219499848640636/,1215312209059390,100000502054733,2022-12-05T08:15:41,711,117,3
Än Sập Sài Gòn, Cao Hoài Anh, 219499848640636/,1226636201260324,100006958841737,2022-12-20T07:16:28,492,190,17
 Ăn Sập Sài Gòn,HaMy Nguyen,219499848640636/,1221271481796796,100001633105067,2022-12-13T08:44:27,416,104,18
 Ăn Sập Sài Gòn, Mai Thanh Tùng, 219499848640636/, 1203939993529945, 100001830457993, 2022-11-18T08:34:31, 743, 240, 64
 Ån Sập Sài Gòn,Nguyễn Mai Khanh,219499848640636/,825559584701323,100002975794771,2021-03-21T10:12:52,1026,385,218
 Ån Sập Sài Gòn,HaMy Nguyen,219499848640636/,1211801369410474,100001633105067,2022-11-30T04:31:28,406,499,57
 Ăn Sập Sài Gòn,Nhàn Đỗ,219499848640636/,670395580217725,100005025181615,2020-08-03T15:54:29,156,17,7
        Sài Gòn,Đỗ Nguyễn Phát Đạt,219499848640636/,1094027187854560,100016543657583,2022-06-07T13:52:33,907,76,41
    Sập Sài Gòn, HaMy Nguyen, 219499848640636/, 1225223724734905, 100001633105067, 2022-12-18T10:06:32, 1026, 713, 56
    Sập Sài Gòn,Lê Thư Lâm,219499848640636/,563901594200458,100002650015153,2020-02-14T11:43:10,839,7,4
 Ăn Sập Sải Gòn,Đỗ Nguyễn Phát Đạt,219499848640636/,1127197404537538,100016543657583,2022-07-28T13:12:58,1076,341,230
 Ån Sập Sài Gòn,HaMy Nguyen,219499848640636/,1222063771717567,100001633105067,2022-12-14T08:06:25,409,179,61
 Ån Sập Sài Gòn,Long Nguyen,219499848640636/,1225251741398770,100000502054733,2022-12-18T11:02:10,290,25,8
 Ăn Sập Sài Gòn,Trần Tuấn,219499848640636/,1216897235567554,100015507585201,2022-12-07T15:31:15,580,699,35
 Ăn Sập Sài Gòn,Phúc Nguyên,219499848640636/,754864378437511,100002789971500,2020-11-23T03:11:26,1090,447,266
 Ån Sập Sài Gòn,Trần Trường,219499848640636/,759936864596929,100045860692115,2020-11-30T13:36:48,103,59,43
 Ăn Sập Sài Gòn,Nguyễn Trung Lộc,219499848640636/,1227432274514050,100000193157368,2022-12-21T08:23:39,209,24,7
```

Here is the .csv file displayed in Excel:



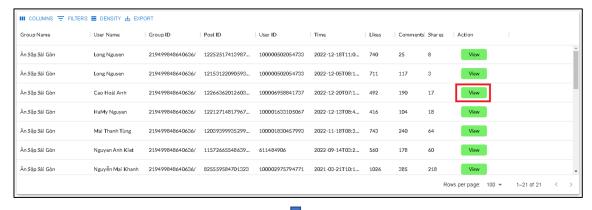
Bar chart

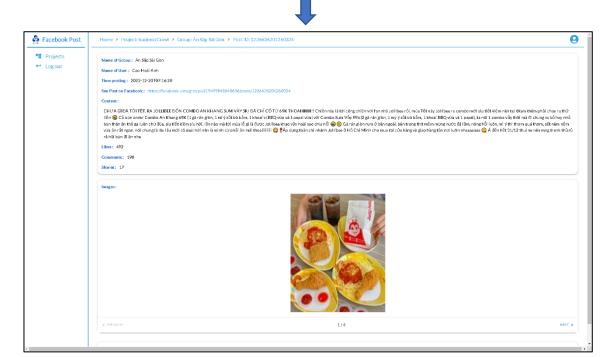
We draw a bar chart to display 10 posts with most likes, comments and shares.



3.5. Post page

From the Group page, we can view a specific post by pressing the View button, which navigates to the information page of that post.





Post interface displays the Name of group and user, Time posting, Post profile url, the content and number of Like, Comments and Shares. It also shows the images of that post, displayed in an Image Slider.

4. Summary and further development

Favor:

- Swagger API supports crawling data.
- Advisors support with care.
- Have a good view of expected results from the beginning.

Difficulty:

• Study many programming languages in a short time.

To crawl data more efficiently, our application would need to be able to crawl data from other websites than Facebook. This method extends the research range so that our application can crawl more data. However, we still need to manage a large amount of data, especially when people upload more and more data everyday. To resolve this, our group suggests using another method in which the bot starts crawling data from the most popular or recent data and links to other pages.

Our group discovered a lot of important information from this subject. The first is specialized knowledge of the software development process, including understanding of UML diagrams, github, and a recent application called Swagger API. Not only that, the difficulties encountered during the working process also gave the participants a chance to hone their self-study techniques, teamwork coordination, and time management abilities. A helpful experience for future employment is also having first-hand knowledge of the procedures involved in running a restaurant or how to create a stunning user interface.