LINKEDIN DATA EXTRACTION AND MAIL AUTOMATION

15CSE387 OPEN LAB

PROJECT REPORT

Submitted by

1. VANDANAPU AKHIL [BL.EN.U4CSE17138]
2. KAMAL KANDULA [BL.EN.U4CSE17510]
3. T V PREMCHAND [BL.EN.U4CSE17523]
4. Y V SRAVAN KUMAR REDDY [BL.EN.U4CSE17524]
5. AJITH PAI [BL.EN.U4CSE17541]
6. ESKALA SREKAR [BL.EN.U4CSE17553]

AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE - 560035

JUNE-2020

ABSTRACT

Web scraping is a technique used to extract data from websites. The extracted data can be saved in a table format for further usage. We are implementing web scraping in our application "LinkedIn data extraction and Mail automation" to extract the data from a particular profile and store it in a table (csv) format. Once the data extraction is done and it is stored in a table format, data obtained is shared with an organization or person mail according to the requirement. The mail is sent to the user in an automated process. The data can be used further by the organization to make analysis and it makes the work easier for them.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION……1
1.1 OBJECTIVE……3
1.2 MOTIVATION……4
1.3 PROBLEM STATEMENT……..5
1.4 PROPOSED SOLUTION……6
CHAPTER 2 USE CASE DIAGRAM AND DESCRIPTION……..7
CHAPTER 3 IMPLEMENTATION……..8
CHAPTER 4 RESULTS……..23
CHAPTER 5 CONCLUSION AND FUTURE SCOPE……..24
CHAPTER 6 REFERENCES……...25

1. INTRODUCTION

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting, etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.

A web scraping software will automatically load and extract data from multiple pages of websites based on your requirement. It is either custom-built for a specific website or is one that can be configured to work with any website. With the click of a button, you can easily save the data available on the website to a file on our computer.

The uses of web scraping for business and personal requirements are endless. Each business or individual has his or her own specific need for gathering data. Here we are discussing a few of the common usage scenarios.

1. For Marketing: Lead Generation.
2. For Businesses / eCommerce: Market Analysis, Price Comparison, Competition Monitoring.
3. Gathering data from multiple sources for analysis.
4. For Research.

Beautiful soup (library in python):

Beautiful Soup is a Python library for getting data out of HTML, XML, and other markup languages. Say you've found some webpages that display data relevant to your research, such as date or address information, but that do not provide any way of downloading the data directly. Beautiful Soup helps you pull particular content from a webpage, remove the HTML markup, and save the information. It is a tool for web scraping that helps you clean up and parse the documents you have pulled down from the web.

Selenium:

Selenium is a popular open-source web-based automation tool.

1.1 OBJECTIVE

Develop a software system that is able to scrape the data from a LinkedIn profile page. Profile pages can be extracted by giving a keyword as an input and get the links of their LinkedIn profile pages. Data extracted from the LinkedIn page are Profile name, Position, City, Connection count, Skills, Company Interests, and Profile contact links. This extracted data is sent to an organization or person through the mail id provided automatically.

1.2 MOTIVATION

A lot of companies put the job role out and get thousands of entries! One of the criteria the company uses for the processing of application is LinkedIn. It gets really challenging to pick each entry and go through the profile. So, to effectively resolve the issue we have automated the process of LinkedIn summarization. Giving appropriate data in a structured format just in a single line. Making the Job very simple.

1.3 PROBLEM STATEMENT

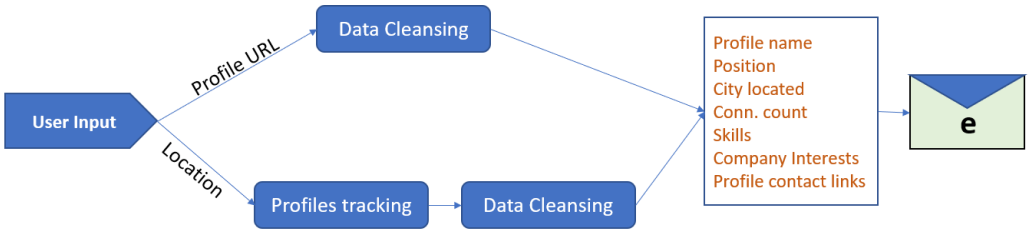
Develop a software program that can automate the data extraction process from LinkedIn profiles and automatically sends the data extracted via email platforms.

1.4 PROPOSED SOLUTION

The profile reading and extracting data from the LinkedIn profile can be automated using some web automation tools and the result obtained, i.e., the LinkedIn profiles data can be sent to the client/user via automated email.

- Web Scraping
- Data cleansing
- Automation

2. USE CASE DIAGRAM AND DESCRIPTION

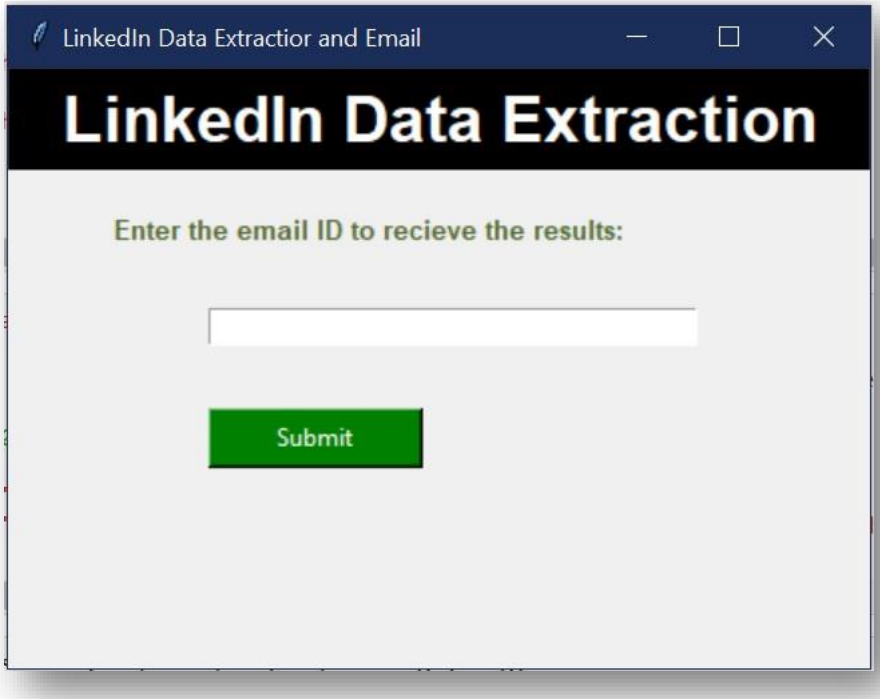


The above-given figure represents the use case diagram.

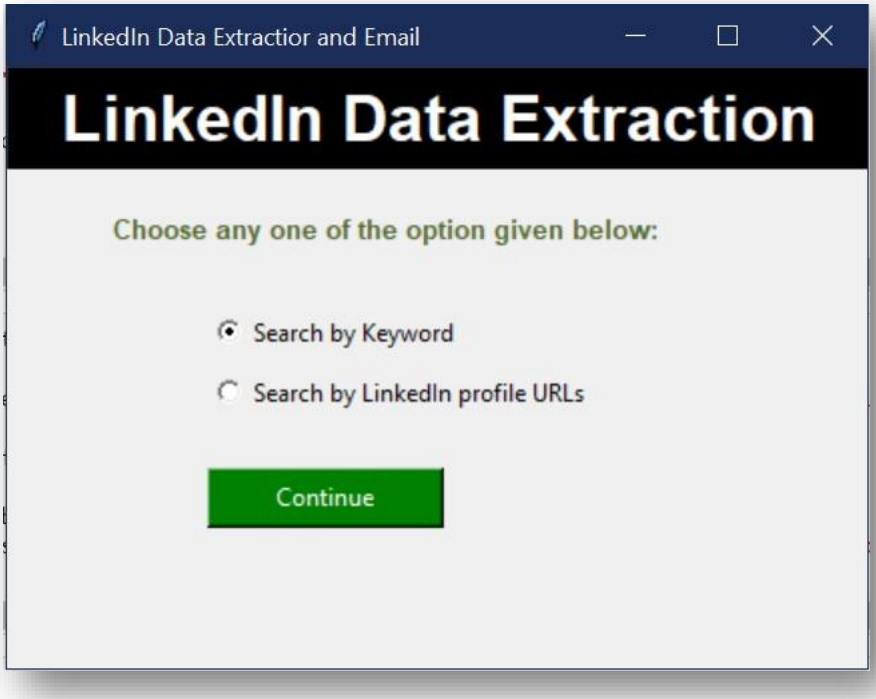
DESCRIPTION:

- Email ID and profile links/locations are taken as inputs from the user.
- Based on the input, the request is processed, and data is extracted in the same way.
- The data extracted is cleansed to remove extra whitespaces and next line characters such as '\n.'
- The final processed data is sent as the output in text format to the email ID given as input by the user.

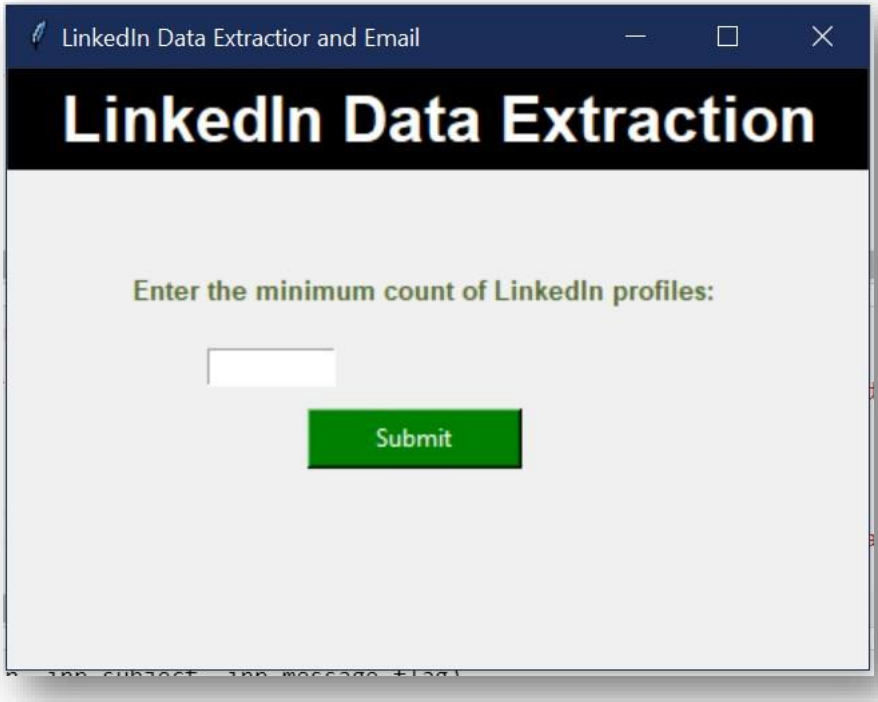
3. IMPLEMENTATION



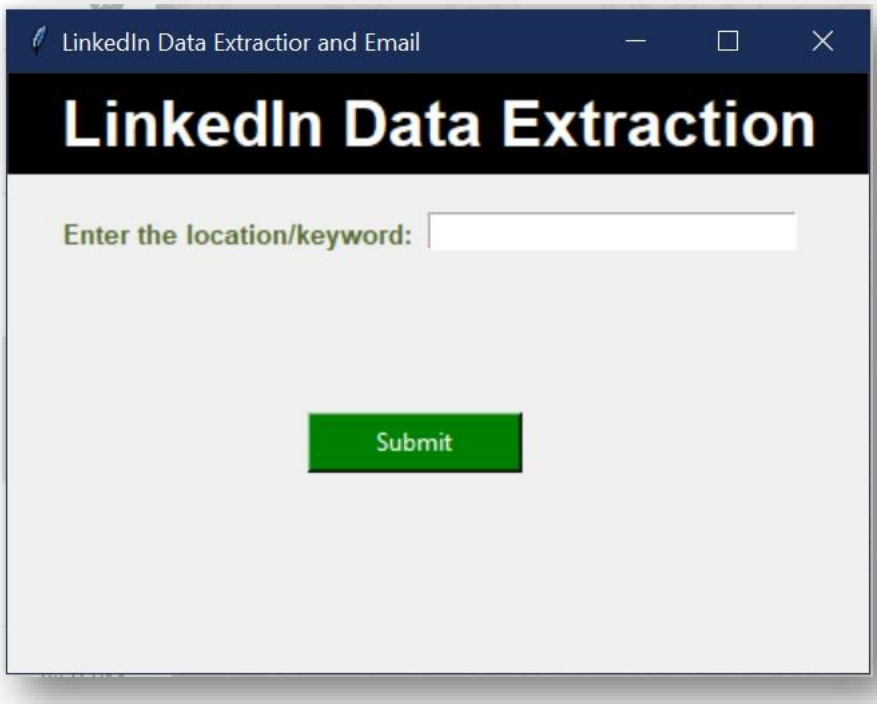
This is the Tkinter module that is used as the front end for our project. This is the first page of our application, and we take the email of the user or customer to whom we must send the mail.



A Tkinter module that is used as the front end for our project. This is the second page of our application, where the user gives his input choice of either extracting the data from profile URLs or using a keyword.

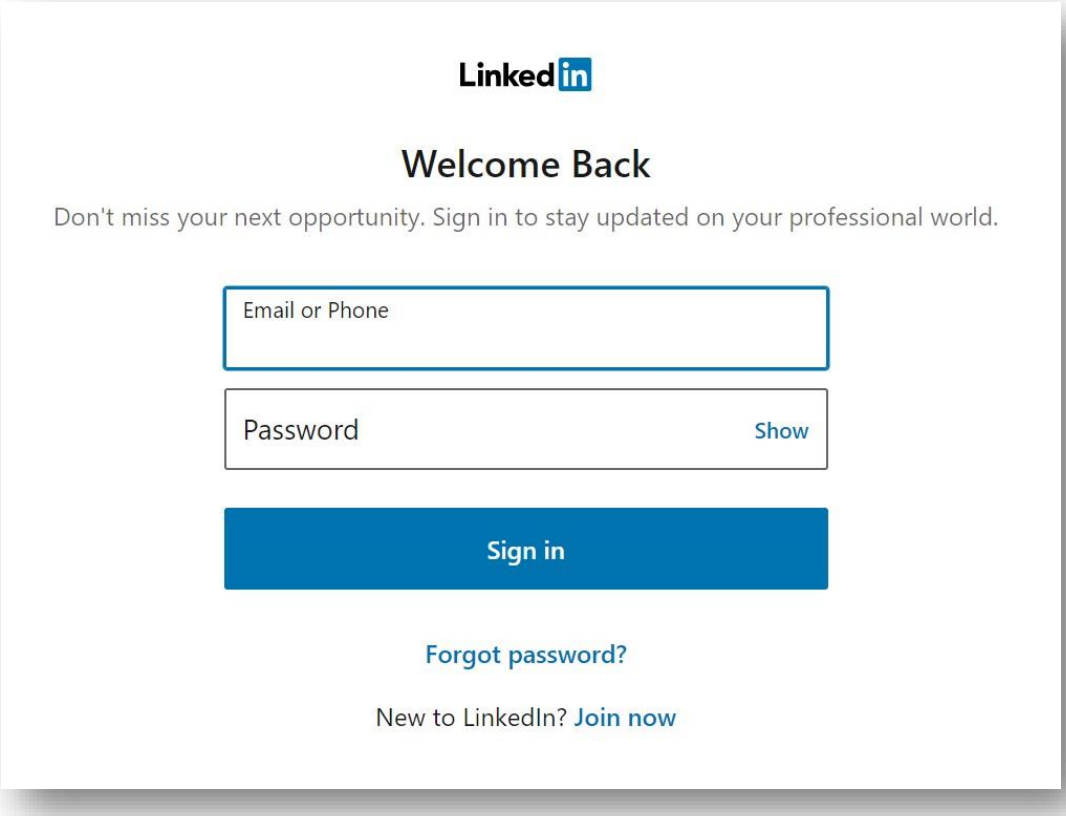


A Tkinter module that is used as the front end for our project. This module appears as the third page of our application if the user gives his input choice as extracting the profile data using a keyword. And here user can enter the minimum number of profiles he wants to scrape utilizing the keyword.

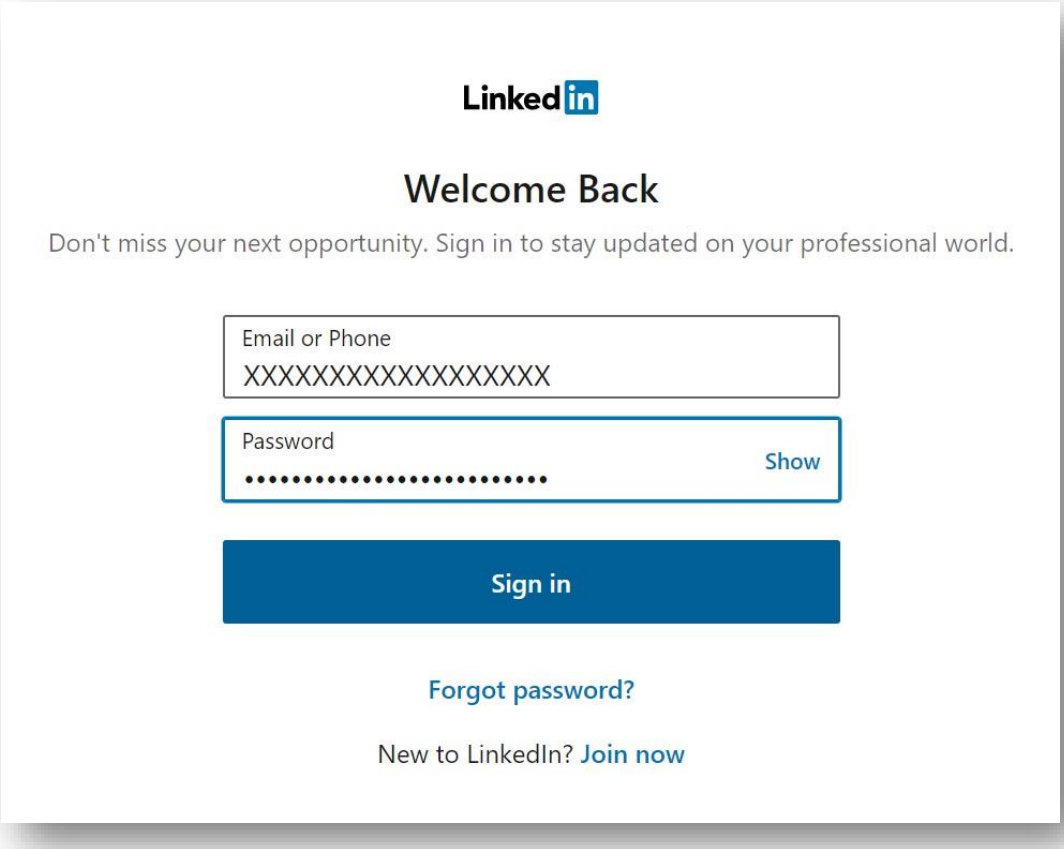


A Tkinter module that is used as the front end for our project. This module appears as the fourth page of our application if the user gives his input choice as extracting the profile data using a keyword. And here, users can enter the keyword to scrape the data from LinkedIn.

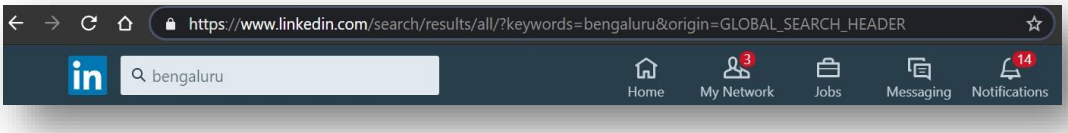
A Tkinter module that is used as the front end for our project. This module appears as the fourth page of our application if the user gives his input choice as extracting the profile data using URLs of the LinkedIn users. Here user can enter all the profile URLs he or she want to scrape the data from LinkedIn.



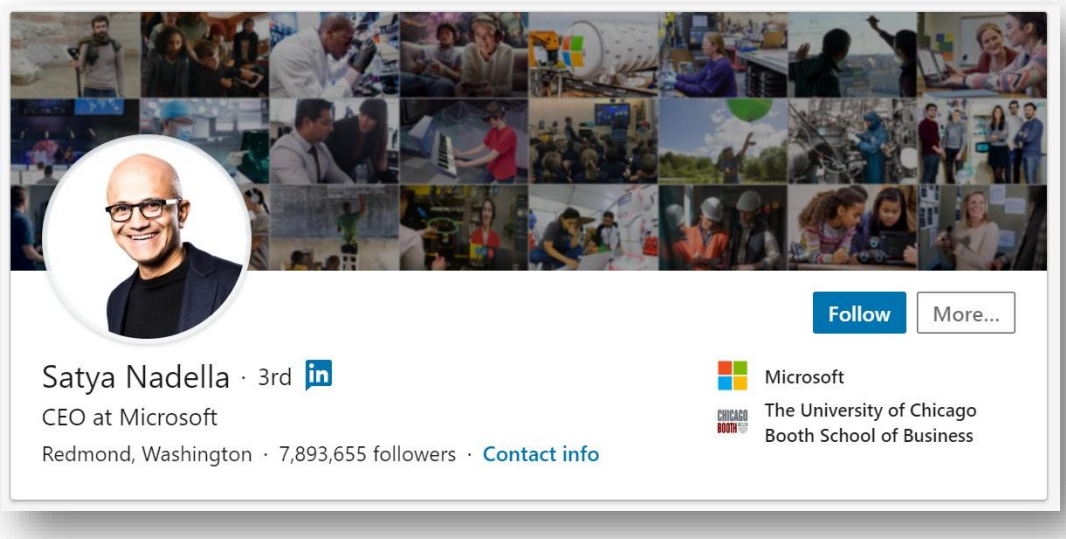
This is the LinkedIn webpage that is used to sign in. The signing in is done automatically by selenium (a Library in python) and web Chrome driver (an automation tool). Using the above two, we will automate the process of signing in



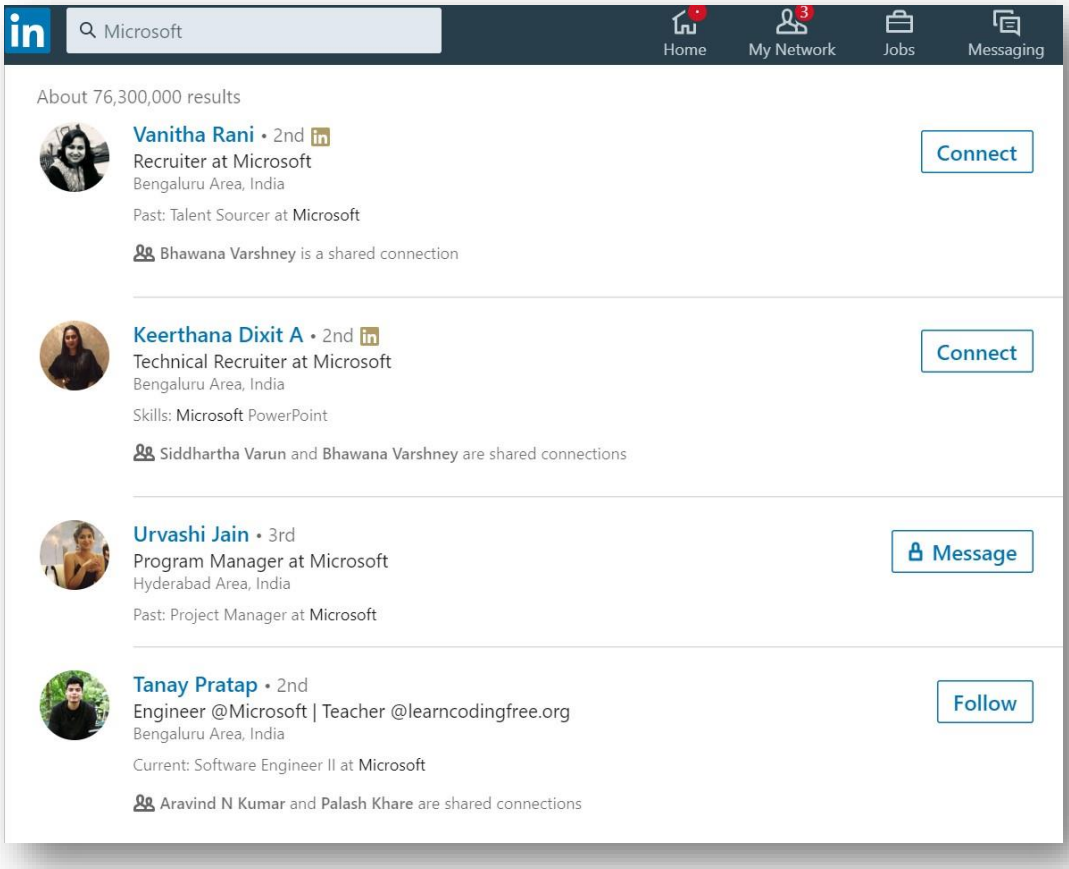
The email is filled in the respective field of email with a time delay of 2 seconds, and the password is hidden and encoded in the backend for security measures. However, in the frontend, the password is already hidden.



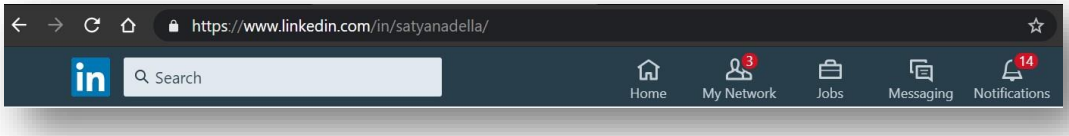
After signing in to email, we must extract the data either by finding the profiles which we have searched by keyword or from the profile links provided by the URL. Here is a sample picture of the URL, which is done by examining the profiles based on keywords.



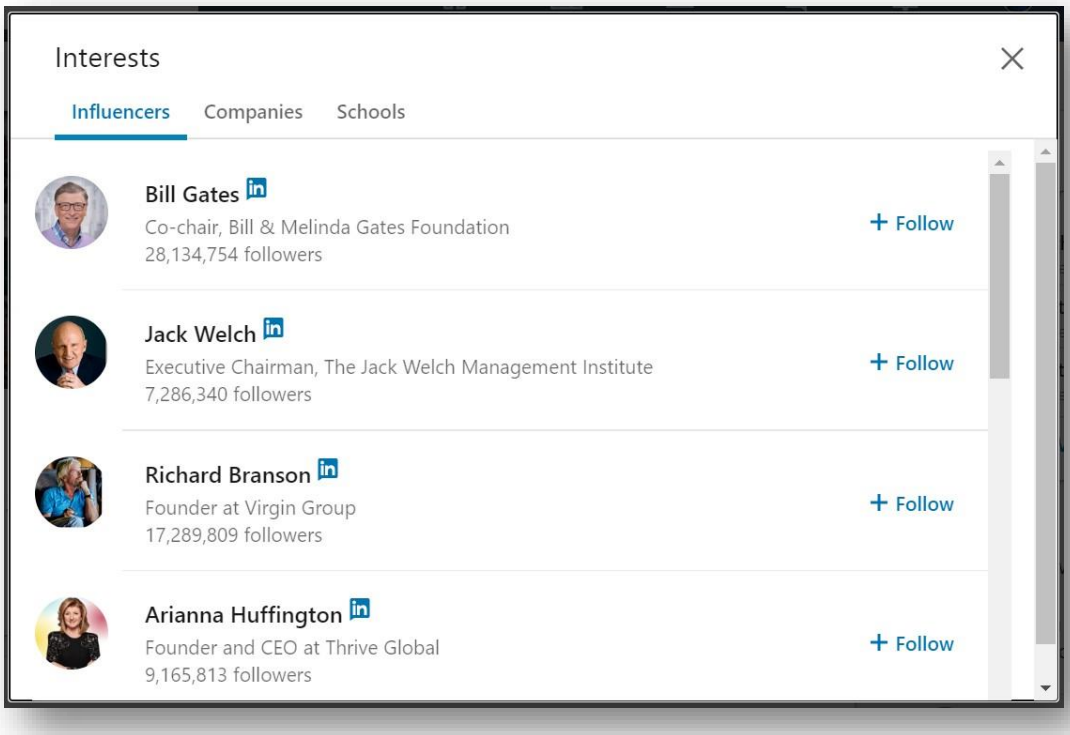
This is the LinkedIn profile account, which we will see on the first page according to the keyword search. Then after we are extracting the details from the web page to structure the data provided.



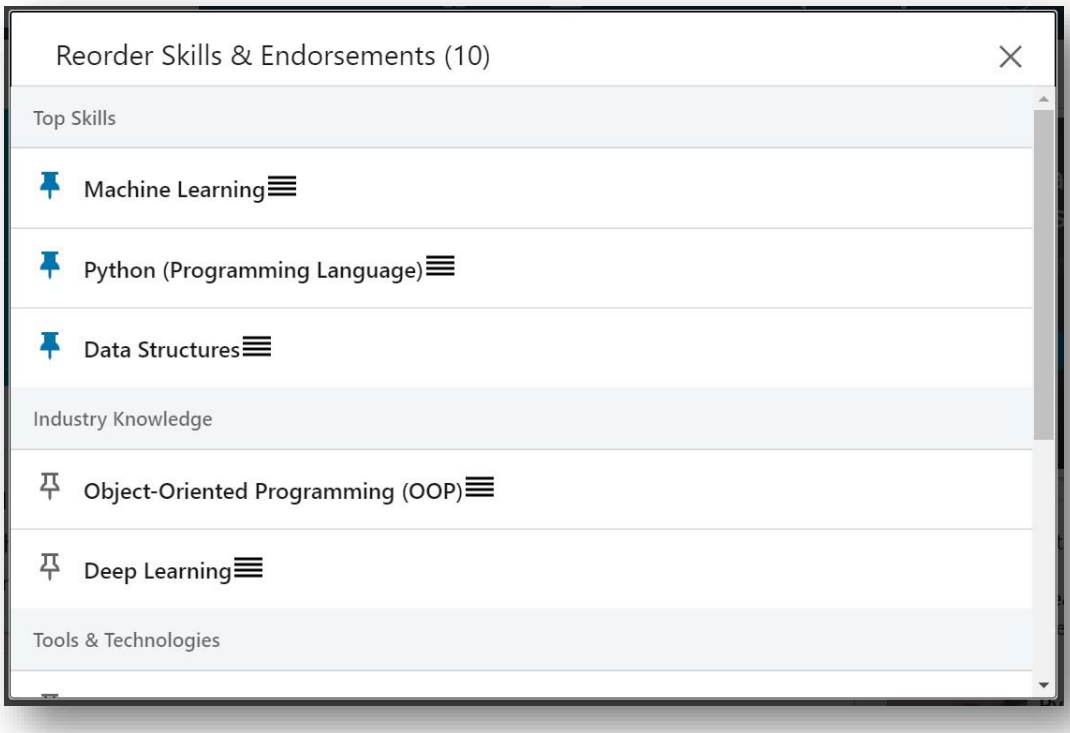
This is an overview of all the profiles we have got (where the exhaustive search part of humanity has been removed). After that, these profile URLs are extracted from the pages which are further processed to extract details



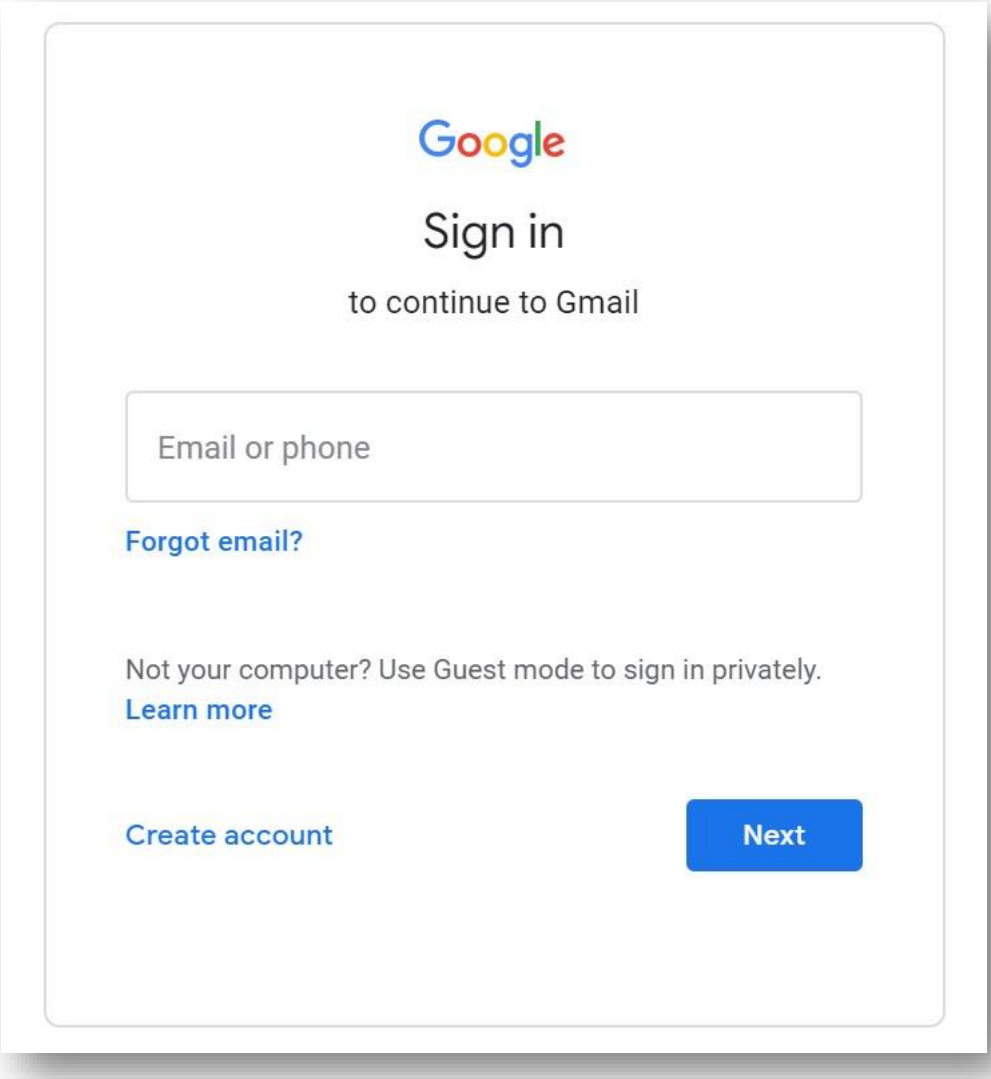
In LinkedIn, as mentioned before, we must extract the data either by finding the profiles which we have searched by keyword or from the profile links provided by the URL. Here is a sample picture of the profile URL. This URL is automatically searched by the program, which is further used in data extraction.



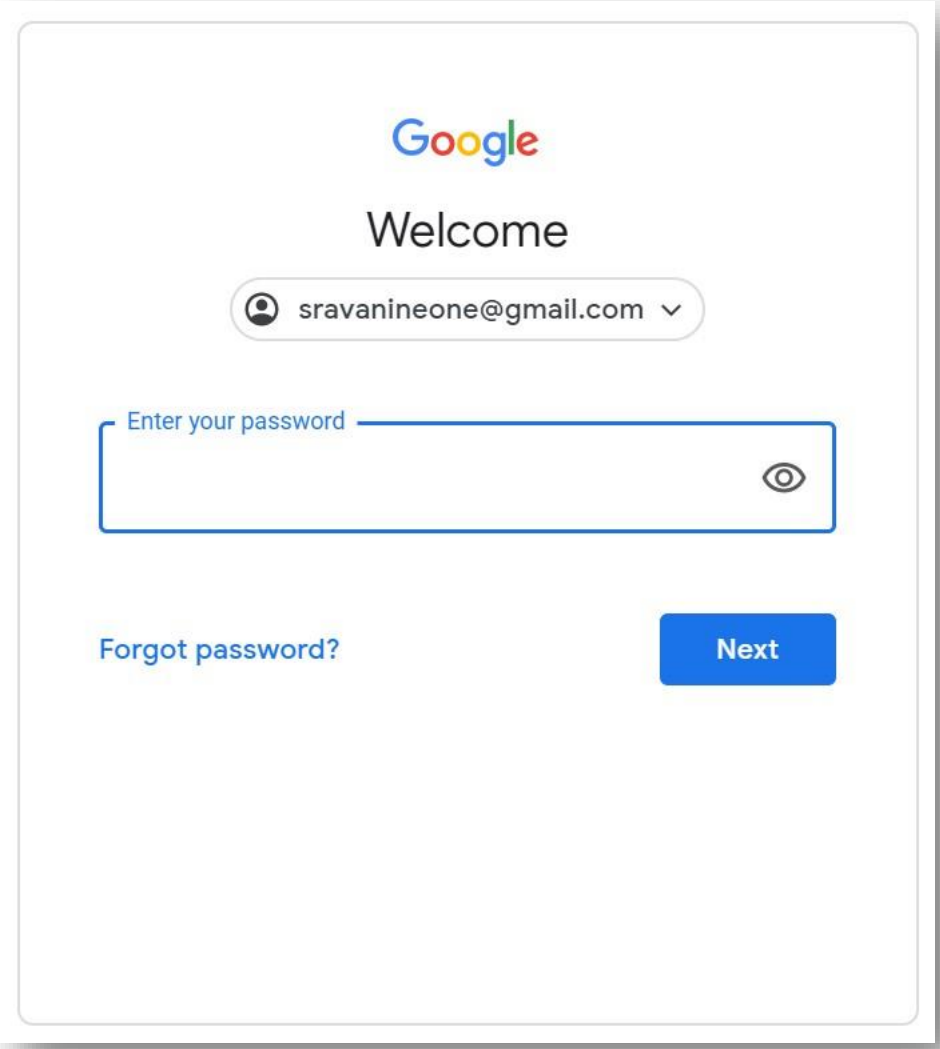
Whenever we extract the data from the LinkedIn, we are also extracting the Interests of that person. When we do this, a window related to Interests will be popped up, and the window is the same one mentioned above in the picture.



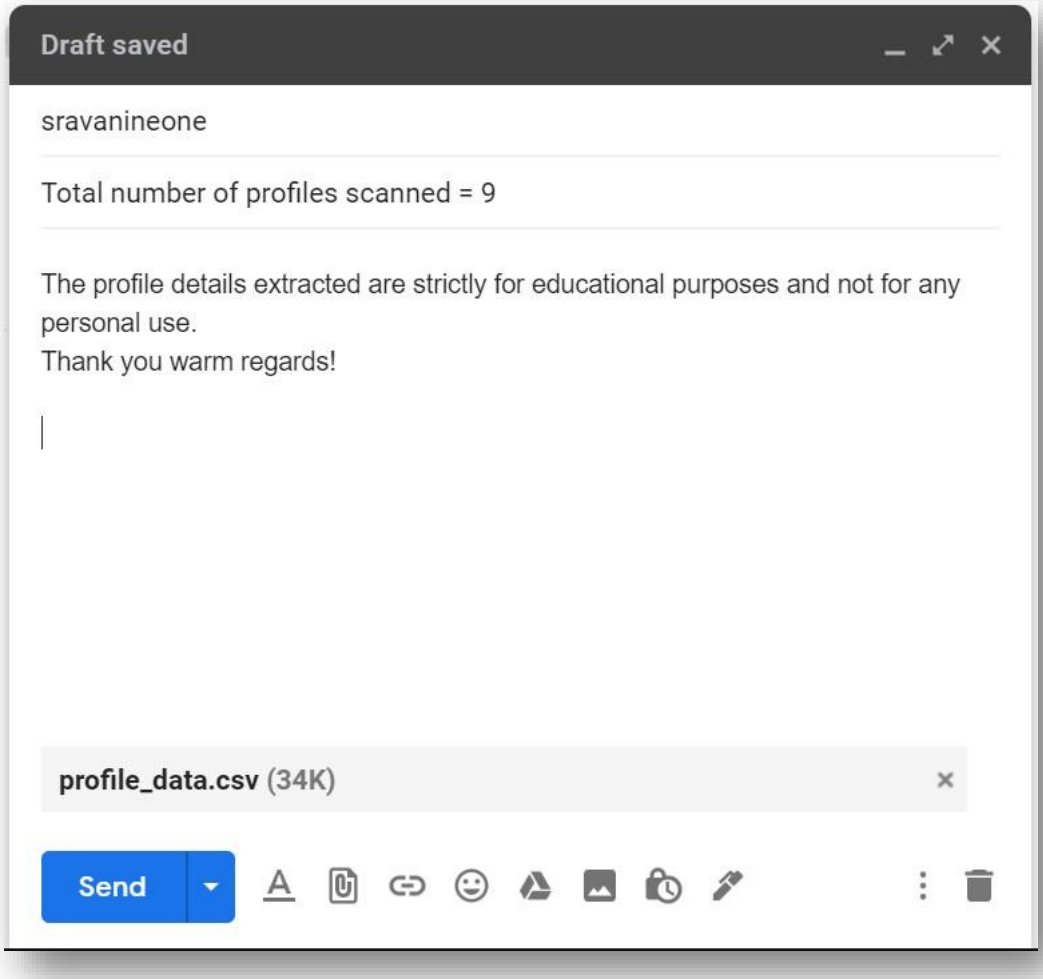
Extracting the data from LinkedIn also contains extracting the skills of that person. When we do this, a window related to skills will be popped up, and the window is the same one mentioned above in the picture.



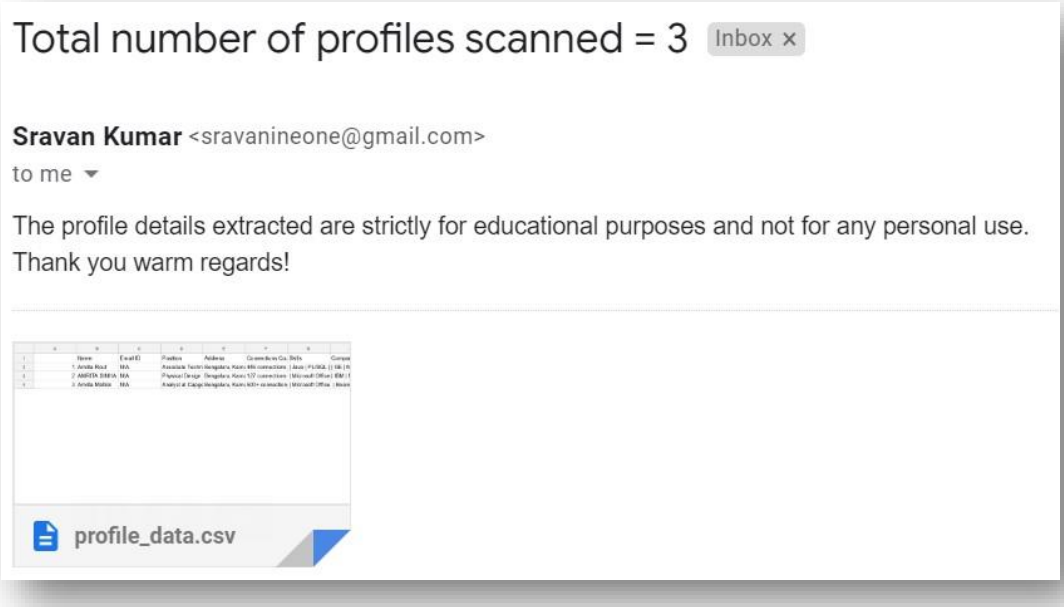
This is the official web page for entering the Gmail of our company so that we can automatically send the mail to the required person. Being Gmail a public key, we will not encrypt it.



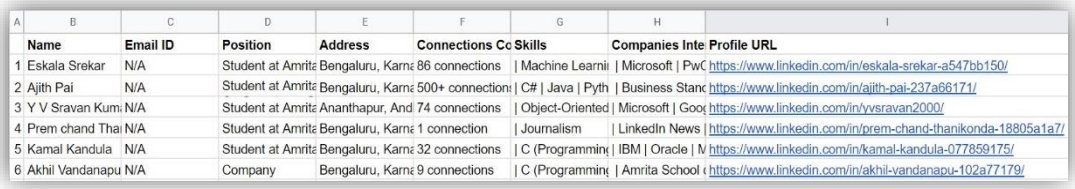
The password which we must enter the field for signing in to the Gmail is filled in the respective place. Here the password is encrypted at the backend, and in the front end, it is hidden as usual.



After signing in to the mail, the next and the foremost step is to send the data to the user, and this is again done with automation (like all the signing in the process). After composing the mail, it will be sent to the customer in a structured format.



This is a mail sent to the user with the structured data. Whenever the profiles are scanned and sent to the user. The subject being the number of patterns being examined and the body is the greetings from the provider.



A	B	C	D	E	F	G	H	I
	Name	Email ID	Position	Address	Connections Co	Skills	Companies Inte	Profile URL
1	Eskala Srekar	N/A	Student at Amrita	Bengaluru, Karna	86 connections	| Machine Learni	| Microsoft | PwC	https://www.linkedin.com/in/eskala-srekar-a547bb150/
2	Ajith Pai	N/A	Student at Amrita	Bengaluru, Karna	500+ connection	| C# | Java | Pyth	| Business Stan	https://www.linkedin.com/in/ajith-pai-237a66171/
3	Y V Sravan Kum	N/A	Student at Amrita	Ananthapur, And	74 connections	| Object-Oriented	| Microsoft | Goo	https://www.linkedin.com/in/yvsravan2000/
4	Prem chand Thai	N/A	Student at Amrita	Bengaluru, Karna	1 connection	| Journalism	| LinkedIn News	https://www.linkedin.com/in/prem-chand-thanikonda-18805a1a7/
5	Kamal Kandula	N/A	Student at Amrita	Bengaluru, Karna	32 connections	| C (Programmin	| IBM | Oracle | N	https://www.linkedin.com/in/kamal-kandula-077859175/
6	Akhil Vandanapu	N/A	Company	Bengaluru, Karna	9 connections	| C (Programmin	| Amrita School	https://www.linkedin.com/in/akhil-vandanapu-102a77179/

The details scanned from the front-end web pages are structured into a profile_data.csv file, as shown in the above figure. The details scanned are seen in the snipped shot at the 1st line of the page.

4. RESULTS

- The profile reading and extracting data from the LinkedIn profile can be automated using some web automation tools.
- The details of a person (like Name, Position, Address, Connection count, Skills, Companies Interested, Profile URL) using his LinkedIn profile link (and store the details in a CSV file).
- Even based on a location also, we can extract the details of the people present in that location.
- The CSV file obtained, i.e., the LinkedIn profiles data, can be sent to the client/user via automated email.

5. CONCLUSION AND FUTURE SCOPE

- We created Tkinter modules/windows for email input, location input, profile URLs input, and related modules as a part of front-end development.
- Data cleansing for data extracted from LinkedIn pages and profiles URLs.
- Data validation and verification did on scraping data from LinkedIn.
- Gmail login automation, email ID parsing, email composer automation, and output file import from local storage.
- LinkedIn keyword analyzing, profile URL's scraping automation, and LinkedIn profile data scraping automation.
- Future Scope: The LinkedIn Automation programs shall be bind together with all other source codes.

6. REFERENCES

i. FORBES, the information on the legality of scrapping data from LinkedIn. https://www.forbes.com/sites/emmawoollacott/2019/09/10/linkedin-data-scraping-ruled-legal/#49f0a9611b54

ii. LinkedIn, to check and verify all the terms and conditions to meet the usage laws. https://about.linkedin.com/

iii. Udemy, to learn about coding in python and using libraries appropriately. https://www.udemy.com/course/pythonforbeginnersintro/

iv. Udemy, to understand the process of LinkedIn data scrapping and implement it. https://www.udemy.com/course/web-scraping-python-tutorial/