Sapternab Chatterjee
Assignment Solutions

Question-1

A developer is assigned a task to scrape 1 lakh website pages from a directory site, while scrapping he is facing such hcaptcha, which are placed to stop people from scrapping. As a Project Coordinator, suggest ways to solve this problem.

Solution: -

Dealing with CAPTCHAs, such as hCaptcha, while scraping websites is a common challenge which are designed to deter automated scraping activities. As a project coordinator, here are some strategies we can try to solve this problem by absolutely ensuring that our scraping activities comply with local and international laws and respect terms of service associated with a website:

- **Public APIs:** Firstly, we should be checking if the website provides a public API for accessing their data which makes it the most straightforward and legal way to obtain the data.
- **Rotating User Agents and IP Addresses:** Maybe we could be using a variety of user agents and IP addresses to make our scraping activities appear more like those of a legitimate user. Rotating IP addresses will help us avoid IP bans. We can use a pool of proxy servers to make our requests for rotating IP addresses
- **Delay Requests:** If we keep inserting random time delays between our requests. This kind of delay helps mimic human behavior and reduces the likelihood of being detected as a scraper.
- **Session Management:** Maybe we can maintain sessions and cookies like a regular web user. This can help with persistence and ensure that our scraping bot appears more like a real user.
- **Headless Browsing:** Also we can use headless browsers like Selenium to interact with web pages. This type of approach can make it harder for websites to detect automated scraping.
- **CAPTCHA Solving Services:** We could also be implementing CAPTCHA-solving services, such as 2Captcha, Anti-Captcha, or DeathByCaptcha, which can solve automatically solve CAPTCHAs, though this might incur additional costs for their premium versions.
- **Scraping at Off-Peak Hours:** We can try scraping during times when the website experiences lower traffic to minimize the chances of encountering CAPTCHAs.
- **Adaptive Scraping:** We can try implementing a system that can with time adapt to changes in the website's behavior, including CAPTCHA challenges.
- **CAPTCHA Solving AI:** Nowadays there are thousands of handy AI tools so we can explore using AI-powered CAPTCHA solving solutions, which can analyze and solve CAPTCHAs more intelligently and diligently.
- **Alternative Data Sources:** Lastly, we should try investigating whether we can obtain similar data from alternative sources that don't have CAPTCHA protection.

Question-2

Our client has around 10k LinkedIn people profiles, he wants to know the estimated income range of these profiles. Suggest ways on how to do this?

Solution: -

Here are my suggestions based on my understanding to estimate income range of the 10,000 LinkedIn people profiles that our client just gave us:

- **LinkedIn Salary Insights:** This feature is available to LinkedIn Premium subscribers and provides estimated salary ranges for different job titles, companies, and locations.
- **Third-party salary estimation tool:** Nowadays, there are a number of third-party tools available, such as PayScale and Salary.com, that will provide estimated salary ranges based on different criteria like job titles, companies, and locations.
- **Combination of LinkedIn Salary Insights and a third-party salary estimation tool:** To make more accurate estimates, we could combine the use of both LinkedIn Salary Insights and a third-party salary estimation tool where Salary Insights will get us a preliminary estimate of the person's salary range and after that we can use the third-party salary estimation tool to refine the estimate.
- **Data scientist**: Data scientists can estimate salaries for LinkedIn profiles without training data by using a variety of techniques, including:
    - **Machine learning models trained on other similar data-** This is the most common approach, and it involves using a machine learning model that has been trained on salaries of people with similar job titles, experience levels, and other relevant factors. The data scientist would then use this model to predict the salaries of the LinkedIn profiles.
    - **Natural language processing (NLP) to extract features from the LinkedIn profiles-** NLP models can be used to extract various text features from the LinkedIn profiles, such as the person's job title, company, skills, education, and work experience which can be then used to train a machine learning model to predict salary.
    - **Hybrid approaches-** Data scientists may also use hybrid approaches that combine multiple techniques. For example, a data scientist might train a machine learning model on other similar data and then use NLP to extract features from the LinkedIn profiles to improve the model's predictions.

Question-3

We have a list of 1L company names, need to find LinkedIn company links of these profiles, how to go about this?

Solution: -

There may be a few ways to find LinkedIn company links of 1 Lakh company names:

- **Manually search for each company on LinkedIn:** This is the most time-consuming method, but it is the most reliable. To do this, simply navigate to the LinkedIn homepage and type the company name into the search bar. Clicking on the company's name in the results list then allows us to view its LinkedIn profile.
- **Use a LinkedIn company finder tool:** Nowadays there are a number of LinkedIn company finder tools available, such as PhantomBuster and Botster which allows us to search for LinkedIn company links in bulk. To use a LinkedIn company finder tool, simply enter the list of company names and the tool will return a list of LinkedIn company links.
- **Use a web scraping library:** Web scraping libraries, such as Scrapy and Beautiful Soup, can be used to scrape LinkedIn company links from the LinkedIn website. Now, this method may be more technical than the others, but it surely is the most efficient way to scrape LinkedIn company links in bulk.

Question-4

How to identify list of companies whose tech stack is built on Python. Give names of 5 companies if possible, by your suggested approach?

Solution: -

Based on my understanding, here are some ideas on how we can prepare an approach to know which companies' tech stack is built on Python:

- **Job Listings:** Searching on job boards and company career pages for job listings mentioning Python as a required skill. Companies which are actively seeking Python developers are more likely to have Python in their tech stack.
- **LinkedIn and Professional Networks:** We can also examine LinkedIn profiles of employees at tech companies. Many professionals list their skills and technologies on their profiles where we can look for companies with a significant number of employees listing Python as a skill.
- **GitHub Repositories:** Next we can search on GitHub for organizations or repositories that have a substantial number of projects written in Python by using GitHub's advanced search to filter repositories by primary language.
- **Tech Blogs and Forums:** We may also conduct research on tech blogs, forums, and discussion boards where companies may discuss their tech stacks where some of them might share insights into their use of Python through blog posts.
- **Open Source Contributions:** We can start to look for companies that actively contribute to or maintain open source Python projects are likely to have Python as a core component of their tech stack.
- **Technology Reports and Surveys:** Lastly, we can explore industry surveys, reports, and studies, like the Stack Overflow Developer Survey and Redmonk Programming Language Rankings, which can provide information on the popularity of Python among companies.

Now based on this approach, here are five companies that use Python to build up their tech stack:

- **Google:** Google is known for using Python extensively, particularly in projects like YouTube, Google Cloud, and various internal tools.
- **Facebook:** Python is used in several Facebook projects, including Django for web development and PyTorch for machine learning.
- **Dropbox:** Dropbox's server-side code is primarily written in Python, which helps power its file-sharing and cloud storage services.
- **Instagram:** Instagram, owned by Facebook, utilizes Python for backend development and data analysis.
- **Pinterest:** Pinterest relies on Python for a variety of tasks, including web development and data processing.

Question-5

Need to find an API, through which we can send LinkedIn messages to other LinkedIn users.

Solution: -

There is no official LinkedIn API for sending messages to other LinkedIn users. However, there are a number of unofficial APIs available that can be used to send LinkedIn messages.

One such API is the LinkedIn Messaging API, which is provided by Beeper. This API allows us to send and receive LinkedIn messages from our own application, but it is a paid API and we will need to create an account and obtain an API key in order to use it.

Another option is to use a third-party LinkedIn automation tool such as Phantombuster, Zapier, or LinkedIn Automation API. These tools offer a variety of features, such as the ability to send bulk messages, personalize messages, and schedule messages. However, all of these tools are paid tools, and we will need to subscribe to a plan in order to use them.

If one is looking for a free option, we can try using a LinkedIn extension such as Linked Helper. This extension allows us to send messages to our LinkedIn connections directly from browser.

We must keep in mind that using an unofficial LinkedIn API or tool to send messages is against LinkedIn's terms of service which if LinkedIn detects that we are using an unofficial API or tool to send messages, your account may be suspended or banned. Nevertheless, here is a summary of the different options available:

API/Tool	Paid or Free	Features
LinkedIn Messaging API	Paid	Send and receive LinkedIn messages from your own application
Phantombuster	Paid	Send bulk messages, personalize messages, schedule messages, and more
Zapier	Paid	Integrate with other apps to automate LinkedIn tasks, such as sending messages
LinkedIn Automation API	Paid	Send bulk messages, personalize messages, and more
Linked Helper	Free	Send messages to LinkedIn connections directly from your browser