**Assignment**

**Qus1. A developer is assigned a task to scrape 1 lakh website pages from a directory site, while scrapping he is facing such captcha, which are placed to stop people from scrapping as a project Coordinator suggest ways to solve this problem?**

When we encounter captcha while web scraping, it can be challenging to bypass, but there are several strategies, we can consider as a project coordinator to address this issue:

* **Contact the Website Owner or Administrator**: Reach out to the website owner or administrator and explain the purpose of our web scraping. In some cases, they may provide us with direct access to the data or offer us an API, which would eliminate the need to bypass the captcha.
* **Use Official APIs or Data Feeds**: Check if the website offers an official API or data feeds for accessing the information we need. Using these APIs can be a legitimate and efficient way to obtain data without dealing with captcha.
* **Proxy Rotation**: Implement a proxy rotation system. Rotate our IP addresses regularly to avoid being detected and blocked by the CAPTCHA. This can help distribute our requests across multiple IP addresses to make it more challenging for the website to block us.
* **CAPTCHA Solvers**: Consider using captcha solving services or tools like 2Captcha, Anti-Captcha. These services solve captcha for a fee, and we can integrate them into our scraping code.
* **Headless Browsing**: Employ headless browsing with tools like Selenium, Puppeteer, or Playwright. These tools can simulate user interaction with the website, making it more challenging for the website to detect automation. However, be cautious about the legality and terms of service of the website we're scraping.

**Qus2. Our client has around 10k LinkedIn people profiles, he wants to know the estimated income range of these profiles. Suggest ways to do this?**

According to me there are five ways to do this

* **Job Titles and Seniority**: Job titles and seniority often correlate with income levels. We can analyze the job titles and seniority of the profiles to estimate income. For instance, executives and senior managers generally earn more than entry-level positions.
* **Company Size**: The size of the companies where these profiles work can also be indicative of income. Larger companies tend to offer higher salaries than smaller ones.
* **Industry**: Different industries offer different salary levels. We can categorize profiles based on the industries they work in and then research average income levels for those industries.
* **Consult with Data Analysts or Data Scientists**: If we have the budget and the data, consider consulting with data analysts or data scientists who specialize in building models to estimate income ranges.
* **Connections**: People with many LinkedIn connections may be more established in their field, which could correlate with higher incomes.

**Qus3. We have a list of 1L company names, need to find LinkedIn company links of these profiles, how to go about this?**

**Manual Search**:

Manually search for each company on LinkedIn by entering the company name in the LinkedIn search bar. This is practical for a small number of companies but not feasible for 100,000 companies.

**Database Management**:

As you collect LinkedIn company profile URLs, manage them in a database or spreadsheet for easy access and organization. Make sure to maintain proper documentation.

**LinkedIn Scraper Tools (Use with Caution)**:

Use web scraping tools like Scrapy or Beautiful Soup in Python to scrape LinkedIn search results. Perform searches with company names and extract the LinkedIn profile URLs.

However, be cautious when scraping LinkedIn, as it goes against LinkedIn's User Agreement, and it may result in account suspension or legal consequences.

**LinkedIn Sales Navigator**:

If we have access to LinkedIn Sales Navigator, we can perform advanced searches to find company profiles based on their names. Export the search results to obtain the LinkedIn URLs.

**Qus4. How to identify list of companies whose tech stack is built on Python. Give names of 5 companies, if possible, by your suggested approach?**

**Google**: Google uses Python extensively for web development, scripting, and data analysis.

**Facebook**: Python is used for various purposes at Facebook, including web development and data analysis.

**Instagram**: Instagram, a subsidiary of Facebook, uses Python for its web backend and data analysis.

**Dropbox**: Dropbox uses Python for server-side applications and various tools.

**Spotify**: Python is used for data analysis and backend services at Spotify.

**Qus5. Need to find an API, through which we can send LinkedIn messages to other LinkedIn users.**

To use the LinkedIn API to send a direct message to a person, we will first need to create an application and obtain an API key. Once we have the API key, we can use it to make authenticated API calls to the LinkedIn API. We can find the documentation for the LinkedIn API, including information on how to create an application and obtain an API key, on the LinkedIn Developer Network website. Once We have the API key, we can use the LinkedIn API to send a direct message to a person by making a POST request to the appropriate endpoint.