Assignment: -

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.

Scraping webpages can be difficult when dealing with CAPTCHAs like hCaptcha. A developer can, however, get around this problem using a number of methods and tactics. Here are some recommendations from a project manager to assist the developer:

• Verify the terms of service: Make sure that scraping the website's pages complies with all applicable terms of service. To prevent legal problems, abide by the website's policies and guidelines.

• Speak with the Website Owner: Inform the website administrator or owner of your purpose to scrape their website. They may occasionally offer you an API or a customised arrangement for data access that allows you to get around CAPTCHAs.

• Delay queries: To simulate human behaviour, add delays between your queries. Rapid scraping might set off anti-bot defences.

• Employ a Headless Browser: Use headless browsers that can run JavaScript and interact with web sites like a human user, such as Selenium or Puppeteer. This can assist in getting around JavaScript-based CAPTCHAs.

• CAPTCHA Solving Services: Take into account employing services like 2Captcha or Anti-Captcha to solve CAPTCHAs. For a price, these services hire people to solve CAPTCHAs on your behalf.

• Using Machine Learning to Solve CAPTCHAs: Create or apply machine learning models to identify and resolve CAPTCHAs. This might be a better long-term answer.

• User-Agent Rotation: Change your user-agent string to represent various devices or browsers. Your requests may appear more authentic as a result of this.

• Session management: To replicate human browsing behaviour, keep sessions and cookies active across requests.

• CAPTCHA Bypass Methods: Some CAPTCHAs can be gotten around by examining their source code and understanding how they operate. This, however, can be tricky and isn't necessarily morally or legally correct.

• Respect Robots.txt: Be sure to follow the directives in the website's robots.txt file, which might determine which pages can and cannot be crawled.

Web scraping might be subject to legal and ethical restrictions, thus it's crucial to act properly and within the limits established by the website owner and the law. The structure of the site, the CAPTCHA system being used, and the project's requirements will all influence the specific method.

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.

Because LinkedIn does not publicly disclose exact income data, estimating the income range of LinkedIn profiles can be difficult. However, you can infer some details and information from LinkedIn profiles to make educated estimates. Here are some ideas about how to go about this task:

• Employ Job Titles and Industries: Information regarding Job Titles and Industries is frequently included in LinkedIn accounts. Specific wage ranges are linked to particular job titles and sectors. You can look up salary ranges for a certain industry and make income projections based on job titles.

• Location: The geographic location has a big impact on income. When estimating income, take into account the location details supplied on LinkedIn profiles and apply local wage statistics or cost of living indices.

• Years of Experience: The number of years of experience is often disclosed in LinkedIn profiles. Professionals with more experience typically have greater salaries. Based on your experience, you can estimate your income using this information.

• Education: The degree of education might also affect how much money is expected. Higher-paid professionals frequently have advanced degrees. Examine the education section of LinkedIn profiles and take that into account when estimating.

• Company investigate: If the individuals' LinkedIn profiles list the organisations they are employed by, you can investigate those organisations to learn the salaries associated with particular positions there. The resources Glassdoor and Payscale are useful for this.

• Establish Contact: Reach out to potential contacts and start conversations to learn more about them. In talks, certain experts might voluntarily divulge their salaries.

• Gather extra Data: Gather extra data from open sources like wage surveys, official labour statistics, or reports from a particular industry. Cross-reference your estimates using this data.

• Machine Learning Models: Create or use machine learning models to forecast income using profile data that is already accessible. For the first training of such a model, you could want labelled data.

• Assume a Wide Range: Since there isn't much information on LinkedIn, it's frequently preferable to give an estimate of your income rather than a precise figure. For each profile, for instance, you may give a low, mid, and high estimate.

• Considerations Regarding Privacy: Be considerate of people's privacy. Despite the fact that LinkedIn profiles are open to the public, it is still crucial to handle data carefully and avoid drawing any conclusions that can be construed as being intrusive.

Remember that extrapolating income ranges from LinkedIn profiles involves some guesswork and potential for mistake. While conducting such analysis, it's crucial to uphold moral and legal norms, respect for people's privacy, and adherence to LinkedIn's platform terms of use.

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

Solution.

Using automated web scraping methods and data enrichment providers, you can retrieve LinkedIn corporate profiles for a list of 100,000 company names. A general strategy for doing this is as follows:

•Data Preparation: Make sure your list of business names is organised and spotless. If you want your search to be more accurate, you might try to standardise the names.To keep the LinkedIn company URLs and any further information you may gather, create a spreadsheet or database.

• Automated Web Scraping: Create a script or use a service that can automatically look up LinkedIn corporate profiles based on company names.The public pages on LinkedIn are dynamic and might not be simple to scrape. It's important to read LinkedIn's terms of service and scrape legally.

• LinkedIn Search: Manually look up each firm name using LinkedIn's search function, then note the URL for their LinkedIn page. Even though it takes time, this method can work for shorter lists.

• LinkedIn API: Access to some corporate data is available through LinkedIn's limited API. However, access to this API is frequently restricted, and therefore might not be appropriate for extracting large amounts of data. Applying for access and abiding by their API usage guidelines are required.

• Use the LinkedIn Search Engine: LinkedIn has a search engine of its own. You can manually search for each firm using LinkedIn's search bar and collect the URLs if the list is not too long. Make sure you restrict the results to only showing businesses.

• Manual Verification: Regardless of the technique you select, you might need to manually check the LinkedIn profiles to make sure the firms they belong to are listed. Although it can take some time, this step is crucial for precision.

• Export and Store: Export the LinkedIn company URLs and other relevant information to your database or spreadsheet for further usage or research.

It's important to abide by LinkedIn's terms of use and operate properly because web scraping LinkedIn may be subject to legal and ethical restrictions. Additionally, the structure of LinkedIn's website could change, necessitating periodic modification of your scraping process to account for such updates.

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.

` Because this data isn't always readily accessible to the public, it might be difficult to pinpoint businesses that utilise Python as a component of their IT stack. To make educated assumptions, though, you can use a variety of techniques and resources. Here is a method to find such businesses:

• Social media and LinkedIn profiles:Programming languages like Python are frequently listed in a person's talents and endorsements on LinkedIn. You may look up employee profiles at different businesses to see if Python is listed as a skill. It's a good sign if several employees from a certain organisation list Python in their skill set.

• GitHub Repositories: Open-source projects are hosted on GitHub by a lot of businesses. To find out if Python is the main language used, look for company repositories and examine the code. A tech stack that uses Python can be identified by a sizable number of Python repositories.

• Job Postings: Check out the listings for open positions on employer websites or job boards. In job descriptions, businesses frequently list the programming languages and technologies that are necessary. A Python tech stack is suggested by the frequent mentions of Python in job postings.

• Technology Articles and Blogs: Businesses occasionally write blogs or articles regarding their tech stack. To see if Python is mentioned, you might search for stories regarding a company's technological decisions.

• Tech News and Awards: Businesses that have utilised Python in their tech stack and have received awards or news coverage may be worthy of consideration.

Let me now give you the names of five reputable businesses that are known to use Python as a component of their IT stack:

1. Google: Google heavily utilises Python for machine learning, web development, and many internal tools.

2. Facebook: For backend programming and data analysis, Python is one of the most popular languages utilised there.

3. Instagram: Instagram, a division of Facebook, employs Python for backend services and web development.

4. Dropbox: Python is frequently used across Dropbox's server-side technology.

5. Netflix: Netflix makes use of Python for a number of tasks, including as data analysis, content suggestions, and their content distribution system.

Please be aware that even though these businesses are known to employ Python, Python is probably just one component of a much more varied technology stack. A company's tech stack can change based on the particular projects and services they provide.

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

Solution.

In order to prohibit the automated delivery of messages by third-party applications or APIs, LinkedIn has put in place stringent standards and security safeguards. The ability to send direct messages to other LinkedIn members is not currently available through the LinkedIn API, as of my most recent knowledge update in September 2021. In general, access to the LinkedIn API is restricted to certain use cases including job posting, recruiting, and integrations with approved partner platforms.

The use of an unauthorised API or automated methods to send unsolicited or automated messages may result in account suspension or other consequences for violating LinkedIn's terms of service.

Using the LinkedIn messaging interface directly through their website or mobile app is the advised and legal method of communicating on LinkedIn. This permits private conversation while abiding by LinkedIn's regulations and users' permission.

Check LinkedIn's terms of service and API documentation for any upgrades or modifications that might have taken place after my previous update. Always be sure that you are adhering to LinkedIn's regulations and respecting other users' privacy and consent when using their site.