Assignment:

1. A developer is assigned a task to scrape 1 lack 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.

Dealing with hcaptcha when scraping websites can be a challenge, as these are designed to prevent automated access. Here are several ways to handle CAPTCHAs when scraping web data:

**Obtain Permission:**  The best approach is always to obtain permission from the website owner or administrator before scraping their site. Many websites have terms of service that explicitly prohibit scraping, while others might provide APIs for data access. If you can get permission or access to an API, this is the most ethical and straightforward approach.

**Use a CAPTCHA Solving Service:**There are third-party CAPTCHA solving services that can help bypass CAPTCHAs. These services often use real humans to solve CAPTCHAs. Some popular services include 2Captcha and Anti-Captcha. However, using these services may have associated costs and ethical considerations.

**Delay and Retry:** Implement a delay and retry mechanism. If you encounter a CAPTCHA page, delay your scraping for a while (a few minutes or longer) before attempting to access the page again. Websites often implement CAPTCHAs to deter rapid, automated scraping, so slowing down your scraper can sometimes bypass them.

**Check for APIs:**Investigate whether the website offers an API for accessing its data. APIs are a more structured and authorized way to obtain data, without the need to scrape HTML content.

**Handle CAPTCHA Challenges Programmatically:** Some websites may present CAPTCHA challenges that can be solved programmatically. For example, simple text-based CAPTCHAs may be solvable using Optical Character Recognition (OCR) tools.

\*\*Remember it's crucial to respect a website's terms of service and privacy policies. If a website explicitly prohibits scraping, it's best to seek permission or explore alternative methods to access the data you need. Additionally, always follow ethical guidelines when scraping data from websites\*\*.

1. 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?

Estimating the income range of LinkedIn profiles can be a challenging task because LinkedIn does not typically provide income information. However, we can make educated guesses based on certain factors:-

**Job Titles and Industries:**

Analyze the job titles and industries listed in the profiles. Certain job titles and industries tend to have higher average incomes.

**Company Size:**

Consider the size of the companies listed on the profiles. Larger companies often pay higher salaries than smaller ones.

**Education:**

Examine the educational backgrounds of the profiles. People with advanced degrees tend to earn more than those with only a bachelor's degree or less.

**Years of Experience:**

Look at the years of experience listed in the profiles. Generally, individuals with more years of experience tend to earn higher incomes. We can estimate income ranges based on the average salaries for different experience levels in specific industries.

**Data Analysis Tools:**

Use data analysis tools and software to process the profiles and extract relevant information. Machine learning models can be trained to predict income based on available profile data, although this would require a large dataset of profiles with known income ranges for training.

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

To find LinkedIn company pages for a list of 1L company names, we can use a combination of manual and automated methods.

**Manual Search:** Start with manual searches for the most well-known and easily identifiable companies. For these, simply type the company name in the LinkedIn search bar and look for the official LinkedIn company page. This will be the most accurate method.

**Use LinkedIn Search:** For companies with common names, use LinkedIn's search functionality. Enter the company name and narrow down the results by using filters such as location or industry.

**Web Scraping:** We can use web scraping techniques to automate the process. Tools like Python's Beautiful Soup or Scrapy can help you scrape LinkedIn search results for company pages.

**APIs:** LinkedIn used to provide a Company Profile API that allowed us to retrieve company information. However, access to this API is limited, and we may need to apply for developer access.

If we have access to the API, you can use it to fetch LinkedIn company pages programmatically. This is a more structured and reliable way to gather this information.

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

Identifying a list of companies whose tech stack is built on Python can be challenging because many companies use a combination of programming languages and technologies. However, one approach to discover such companies is to analyze job postings, company websites, and technology-related publications. Here's a simplified approach:

Use Job Search Engines:

Explore Company Websites:

Analyze LinkedIn Profiles:

Use Online Communities:

Here are five well-known companies that have publicly mentioned their use of Python in their tech stack as of my last knowledge update in January 2022:

Google

Facebook

Dropbox

Instgram

Spotify

1. Need to find an API, through which we can send linkedin meaasges to other linkedin users

As updated in January 2022, LinkedIn does not provide a public API that allows sending unsolicited messages to other LinkedIn users. The LinkedIn API has limited functionality and is primarily designed for integration with LinkedIn's official partners for specific purposes such as recruitment, marketing, and analytics.

To send messages on LinkedIn, we should follow these general guidelines:

**Connect with the Recipient**

**Use LinkedIn's Messaging Interface**

**Respect LinkedIn's Rules**

**Do Not Automate Unsolicited Messaging**

**Leverage Premium Features:**

If we have a LinkedIn Premium subscription, you may have access to additional InMail messages, which can be used to contact users who are not in your network.

---------------------------------------------------------------------------------------------------------

Mudabir Afroz Jan

+917006282791

Date:28/10/2023