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

hCaptcha asks straightforward questions that are simple for people to answer and challenging for machines to answer, helping your favourite online services ward off spam, bots, and abuse. A service you use has decided to add hCaptcha to improve your online experience and lessen fraudulent traffic.

a. Use an API if the target website provides that. It's the most legal way

b. Rotation of Proxy:

To get around IP restrictions, use a pool of dynamic proxies. In order to create the illusion that requests are coming from various areas, these proxies should have different IP addresses.

Make sure the proxy service you use is trustworthy because some proxies may be untrustworthy or blacklisted.

c. Limiting Rate and Delay:

To make your queries behave more like those of a human, introduce delays. To make your scraping activity seem less suspicious, vary the duration of these delays.

Put rate limitation into practice to limit how many requests you send in a minute or an hour. You can prevent security measures from being activated by adhering to rate limits.

d. Employ a Service to Solve CAPTCHAs:

Third-party services and APIs are available that provide automated solutions for captchas. These services solve captchas using machine learning and optical character recognition (OCR).

These services include DeathByCaptcha, Anti-Captcha, and 2Captcha.

These services are frequently not free, and their efficacy varies. Think about the ethical and legal ramifications of using these services as well, as they can be against the terms of service of the website you are scraping.

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?

1.Machine Learning Models: Using the data accessible on LinkedIn profiles, create or employ pre-trained machine learning models to forecast revenue. To provide forecasts, these models might take into account factors like years of experience, education, and job title.

2.Take a look at the job title and industry on your LinkedIn profile. Some job titles and industries have certain income ranges attached to them. For instance, if you're a software engineer in the tech industry, you might make a lot more money than a marketing manager in the same field. You can use salary data from different industries or government statistics to figure out the income ranges for those job titles and industries.

3.You can do surveys and market research to get income data for similar jobs and industries. This won't give you exact numbers, but it'll give you a general idea. You can also use your professional network and industry connections to get an idea of what typical income ranges are for certain roles and industries. People who work in the same field might have a better idea of what the income norms are

4. If the profile is of an employee of a publicly traded company, you can check out their annual reports or financial news websites to see what their financials are like and what their executive salaries are like.

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

To break the business problem of chancing LinkedIn company profiles for a list of 100, 000 company names, I would follow a methodical approach that clarifies Objectives First, I would meet with the stakeholders or clients to easily understand the objects behind collecting LinkedIn biographies. Are they seeking to establish business connections, conduct request exploration, or for some other purpose? The specific pretensions will guide the approach. Data Sources and Quality Assessment I would estimate the quality and applicability of the company names in the list. It is important to ensure that the list is accurate and over- to- date. This step might involve data cleansing and confirmation.

Using the LinkedIn Profile API: The Linkedin Profiles can be accessed using the Linkedin API which can be used to gather the information and do market research where companies can export LinkedIn profiles of industry professionals to analyse their experience, education, skills, and work history. This can help them identify market trends, assess competitors and make informed business decisions.

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

Data Sources

Identify the sources of information you will use. These can include job postings, company websites, LinkedIn, GitHub, and industry-specific platforms or forums.

Keyword Search:

Perform keyword searches on job portals and career pages. Look for job postings that mention Python as a required skill. This can give you insights into companies that use Python.

LinkedIn Company Pages:

Visit the LinkedIn Company Pages of the companies you are interested in. Look for information about their tech stack in the "About" or "Overview" sections.

GitHub Repositories:

Search GitHub for repositories linked to companies. Companies often open-source code, and the tech stack may be mentioned in project descriptions.

Tech News and Blogs:

Look for tech news articles and company blogs that mention the use of Python in a company's tech stack.

Social Media and Forums:

Explore social media platforms, tech forums, and developer communities for discussions or posts related to companies using Python.

Data Enrichment Services:

Consider using data enrichment services that may provide information about a company's tech stack. These services can offer aggregated data from various sources.

Industry Reports:

Look for industry-specific research reports or surveys that may provide insights into the tech stacks commonly used by companies in a particular sector.

Data Validation:

Verify the information you collect from different sources to ensure its accuracy and relevance to your objective

5 companies that use Python:

Google: Google uses Python for various purposes, including web development, machine learning, and data analysis. They have open-sourced Python libraries like TensorFlow.

Facebook: Python is used at Facebook for web development and data analysis. Facebook developed and open-sourced the Python library PyTorch for deep learning.

Dropbox: Dropbox's server-side code is primarily written in Python. They've also open-sourced some Python libraries and tools.

Instagram: Instagram, a subsidiary of Facebook, is known to use Python for its backend services and infrastructure.

Spotify: Spotify uses Python in various aspects of its technology stack, including web development and data analysis.

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

Sending a direct message to a person using the LinkedIn API requires obtaining the appropriate API permissions and making an API request to the LinkedIn API endpoint. The LinkedIn API provides the `/v2/messages` endpoint for sending messages.
Here is an example of how you can send a direct message to a person using the LinkedIn API with the `/v2/messages` endpoint:

1. Obtain API permissions: To use the LinkedIn API, you'll need to obtain the appropriate API permissions. This can be done by creating a LinkedIn Developer Account and registering your application.
2. Authenticate the API request: To authenticate the API request, you'll need to use an access token. The access token can be obtained by following the authentication process outlined in the LinkedIn API documentation.

3. Send the API request: Once you have obtained the access token, you can send a POST request to the `/v2/messages` endpoint with the recipient's LinkedIn ID and the message text in the request body.