Assignment by Thanishma

1. **A developer is assigned a task to scrape 1 lakh website pages from a directory site, while scraping 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 CAPTCHA challenges when scraping websites is a common issue, as many websites implement them to prevent automated scraping.

There are different ways to bypass hcaptcha while scraping:

- Review the Website's Terms of Service: Before proceeding, ensure that you are not violating the website's terms of service by scraping their data. Some websites explicitly prohibit scraping, and ignoring these terms could lead to legal issues.
- Use APIs (if available): Check if the website offers an API that allows access to the data you need. APIs are structured data access points that are more web scraping-friendly and can be used without the need to solve CAPTCHAs.
- Rotate IP Addresses and User Agents: Employ a pool of rotating IP addresses and user agents to avoid IP bans and detection by the website's security mechanisms. This can help distribute requests and make it harder for the website to identify your scraper.
- Browser Automation Tools: Utilize browser automation tools like Selenium or Puppeteer to automate the interaction with the website. These tools can simulate human behavior by opening a web browser and interacting with the website, including solving CAPTCHAs manually if necessary.
- Avoid Hitting the Same Page Repeatedly: Instead of scraping the same page over and over, try to navigate the website more like a human user by visiting different pages, following links, and interacting with the site in a natural way.
- Break the Task into Smaller Parts: Instead of scraping 1 lakh pages in one go, break the task into smaller chunks and scrape a reasonable number of pages at a time. This may help avoid triggering CAPTCHAs.

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?**

   LinkedIn profiles do not usually show the exact salary of the members, but there are some ways to get an approximation based on other factors, such as job title, location, industry, education, experience, skills, etc.

   Some of the possible methods are:

- Use LinkedIn Salary: This is a tool that LinkedIn provides to show the salary range for different jobs and locations, based on data from LinkedIn members. You can search for the job title and location of the profiles you have and see the median and percentile salaries. However, this method may not work for all profiles, as not all members share their salary data, and the data may be outdated or skewed by outliers.
- Use Talent Insights: This is a product that LinkedIn offers to help recruiters and employers find and analyze talent pools. You can create a talent pool report based on the profiles you have and see the compensation information for that pool. This data is inferred from LinkedIn Salary data and estimated based on similar roles, companies, and regions. However, this method requires a subscription to Talent Insights, which may be costly, and the data may not be very precise or representative.
- Use Statistical Modeling: This is a method that LinkedIn uses internally to power its salary product. It involves using machine learning algorithms to model the relationship between salary and various factors, such as job title, location, industry, education, experience, skills, etc. You can try to replicate this method by using your own data or public data sources, such as salary.com or payscale.com. However, this method requires a lot of technical expertise and computational resources, and the results may not be very accurate or generalizable.

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

LinkedIn company links are the URLs that lead to the official pages of organizations on LinkedIn. They usually have the format of https://www.linkedin.com/company/<company-name>. However, not all companies have the same name on LinkedIn as they do in other sources, so finding the exact links can be challenging.

There are different ways to find LinkedIn company links from a list of company names, but none of them are perfect. Here are some possible methods:

- Use a web scraping tool: You can use a tool that automates the process of searching for company names on LinkedIn and extracting the links from the results. For example, you can use Derrick, a Google Sheets add-on that lets you find LinkedIn company links from company names in a few clicks. You can also use PhantomBuster, an online platform that offers a similar service. However, these tools may not work for all company names, and they may require a subscription or a fee.
- Use a manual search: You can manually type and enter the company name into the search bar at the top of your LinkedIn homepage, and then click on Companies at the top of the search results page. You can also use filters such as Locations, Industry, or Company size to narrow down the results. Then, you can select the correct organization name in the results list and copy the link from the address bar. However, this method may be time-consuming and tedious, and it may not yield accurate results for some company names.
- Use a statistical model: You can use a machine learning algorithm that predicts the LinkedIn company link from the company name, based on various features such as spelling, similarity, popularity, etc. You can train your own model using your own data or public data sources, or you can use an existing model that someone else has built. However, this method requires a lot of technical expertise and computational resources, and the results may not be very reliable or generalizable.

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

   Python is a popular and versatile programming language that is used by many tech companies for web development, data analysis, machine learning, and more. There are different ways to find out which companies use Python in their tech stack, but none of them are very accurate or comprehensive. Here are some possible methods:

- Use a web search tool: You can use a tool like Bing to search for keywords like "companies that use Python" or "Python web development". You can also use filters like date, region, or language to narrow down the results. However, this method may not show all the relevant companies, and the information may be outdated or incomplete.
- Use a web scraping tool: You can use a tool that automates the process of extracting data from websites, such as BeautifulSoup or Scrapy. You can write a script that crawls through various websites that list or rank tech companies, such as StackShare or BuiltWith, and scrapes the information about their tech stack. However, this method may require some technical skills and may violate some websites' terms of service.
- Use a statistical model: You can use a machine learning algorithm that predicts the probability of a company using Python in their tech stack, based on various features such as company size, industry, location, etc. You can train your own model using your own data or public data sources, such as [Crunchbase] or [LinkedIn], or you can use an existing model that someone else has built. However, this method requires a lot of technical expertise and computational resources, and the results may not be very reliable or generalizable.

   Using web search tool some of the popular companies that use python are:

   Google, Netflix, Dropbox, stripe, Instagram