Assignment Submission – Varun Vashisht

Assignment: -

1. 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

Ans. As per my understanding of the question, few steps which can be undertaken are:

i. A captcha detection model can be trained on a dataset of multitude of images/text data to make the automation accurate which can be deployed as a browser extension.

ii. One of the main causes of the captcha generation is due to the huge number of requests from a single IP, hence, we can rotate the IP addresses to send requests over the net.

iii. A third-party API like 2Captcha's reCAPTCHA solving API can be employed.

iv. Another approach to this problem is to use the official API for the sites to "officially" use the data provided by them.

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?

Ans. Web-scraping can be deployed here in a sense that:

i. Scrape the profiles for information such as job profile, company size, location and such information.

ii. Use the sites like Glassdoor, Salary.com, Indeed etc. to get the range of the salaries for each profile scraped.

iii. Employing data analytics to calculate and visualise the salary ranges.

iv. In addition to this, an ML model can be trained from the scraped data to create salary range estimates.

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

Ans. Firstly, organise the data of the names in a proper format (csv, spreadsheet) for ease of use.

i. One of the steps that can be used is to use the hardcoded header link format for linkedin and looping over the list of the names to "generate links". (This may not have a very good success rate considering the links may differ from the name of the company).

ii. Another approach to this problem may be to automate the process of iterating over the list, and automating a google search with the "linkedin" suffix (for making scraping easier as we know the top links will definitely be for the linkedin profiles of the company). We may then scrape the google search query data for the links.

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

Ans. One of the basic implementations would be to build a scraper to scrape the data from the google search using a customised query such as "companies using python", "companies with python in their tech stack" etc. and store the results.

My approach got me the results:

i. Dropbox

ii. Instagram

iii. Spotify

iv. Google

v. Netflix

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

Ans. The Messages API provided by Microsoft itself can be used for this purpose. The link to which is: Messages API - LinkedIn | Microsoft Learn. It allows members to create messages to one or more first-degree connections or reply to existing conversations.