**Name: Soham Modi**

# Assignment

**Q1. 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**

There are various ways with which we can overcome this issue:

1. Solve CAPTCHA Automatically:

   This method may not be reliable, but we can access some CAPTCHA solving services or some APIs to automatically solve the CAPTCHA. This solution can be more expensive an may not work effectively in all cases

2. Solve CAPTCH manually:

   We can employ a team of some freelancers to manually solve CAPTCHA when they appear. This method is slower and more labour-intensive but can effective in case of challenging CAPTCHA

3. Machine Learning:

   We can train a Machine Learning model to detect and solve CAPTCHAs. This is the most advanced approach so far and it requires a large labelled training dataset for getting higher accuracy

4. Request permission from the Website Admin:

   If we have any way so that we can contact the Website Admin, we should get in contact with them and request to scrape the site.

5. Delay Between Requests:

   CAPTCHAs are used to detect machine behaviour. We can introduce some delays between the requests so that it will mimic as if a human is accessing the site

**Q2. 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?**

There are various factors that affect income range such as:

- Location of the job
- Job Profile
- Position
- Company
- Education and Experience

We can collect all these data from the profiles with the help of a team of freelancers (to reduce human resource cost and increase the efficiency). On the basis of these attributes, we can develop a predictive model using Machine Learning algorithms (e.g., Multiple Regression)

With this approach we can predict the income range of a person by giving this data to the model.

We can also get in touch with a Data Scientist or Analyst to create a custom model for our objective based on the available data

**Q3. We have a list of 1L company names, need to find LinkedIn company links of these profiles, how to go about this?**

This can be easily achieved using Web Scraping method. We can use various available libraries such as Requests, BeautifulSoup in Python which help us to navigate to the data present on the page. There are various stages to achieve this:

- Data Preparation:

  Assemble all the into a .csv file

  Validate the data if it contains any null values

- Web Scraping

  Create a Python script to using the above-mentioned libraries and extract the website link from their respective profiles by finding links in their profile and store them in format according to the requirement

**Q4. How to identify list of companies whose tech stack is built on Python. Give names of 5 companies, if possible, by your suggested approach**

There are various ways by which we can identify list of companies whose tech stack is built on Python:

1. Their Website or LinkedIn Profile

   We can go through the website and check the technological page or their Blogs. We can also go through their LinkedIn profile to check whether there are any hints of Python

2. Job Requirements

   In the Job Requirements if they are seeking the qualities related to Python then they of course have some stack built on Python.

3. We can also search whether they have any Python related articles or projects available on Internet

There are many companies whose stack is built on Python:

- Google
- Meta
- Spotify
- Amazon
- Microsoft

**Q5. Need to find an API, through which we can send LinkedIn messages to other LinkedIn users**

As per me LinkedIn does not provide API to send messages to other LinkedIn users. But what we can do is, we can use a website known as phantombuster.com to access the messages endpoint of LinkedIn and send customized messages according to the condition given. We can also schedule or send messages repeatedly without doing any manual work