GOCOMET ASSIGNMENT

Siddhesh Jadhav
jsiddhesh70@gmail.com
Roll no -18102C2030

You need to create a simple application that takes the product names and parameters as input and then scrape Amazon and Flipkart for the products list.

To run the program which I am submitting it requires python and some of this library to install on your system.
1 - Python
2 - selenium(Used for Scraping)
3 - Web Driver Manager
4 - xlsxwriter (To write result into excel)

→ While doing this problem the biggest challenge was to understand the selenium and understanding how we can use xpath by using logic.

1. At the beginning I learned the concept of Beautiful Soup library but after some research I found that selenium will be good for doing web scraping.
2. I did not know how to get the xpath or generate it in a logical way, so first I learned how to get the xpath to the required element.
3. After the complete I took the problems to solve.
4. I have done the problem in which users provide a product name and I provide them the list of products from amazon and flipkart with attributes like Product name,model number,source,link, etc. and write down the details into the excel file.

The browser scrapping behavior should be as close to user behavior as possible (you need to be creative on this how you will implement it) [it should be like the user is searching for products on Amazon/flipkart]

→ I have researched a lot about how and why amazon blacklists the IPs. We can use the headers, proxy ips, and slow down the scraping speed.

We can do web scraping to get the desire, but as a bot it requests the server many times within less time.This causes the load on the server Hence most of the companies don't allow scraping on their websites.

We can avoid getting blacklisted by:

1. We can do that by using proxy ip server and VPN and rotate the IPs at a particular interval.
2. But using the free proxy IPs amazon as well as google doesn't allow us using those IPs and visit their website. To do that we need to use a paid server for IPs.

Thank you.