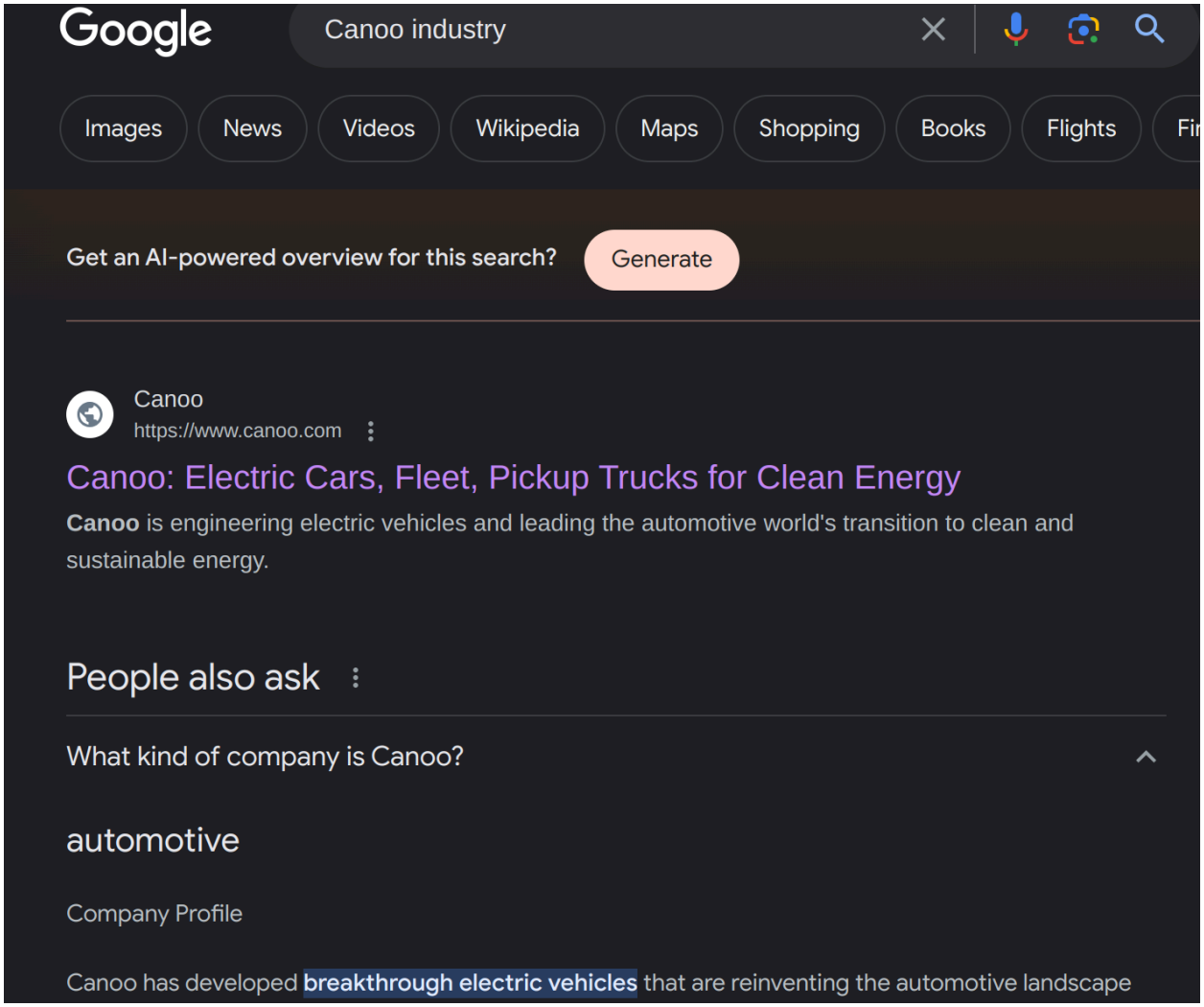
Internship Assignment 2024 Coding Challenge

AI/ML Engineering Intern

Name: Vinayak Rao Dikshit
Mob. +91 8120215502
Github to repo: https://github.com/VinayakRaoDikshit/LizmotorsAssn.git
My Github: https://github.com/VinayakRaoDikshit

Explanation of approach:

I utilized the fact that the "People Also Search for" section on a google search results page mostly contains one liner answers to people's queries. So, I used Selenium to open the instance of a Chrome Browser session and searched for relevant answers in that section. For example, here is a screenshot of the query for the search result of "Canoo industry".



We can see that output is already highlighted in the “People also Search for Section”.

Steps taken and Summary:

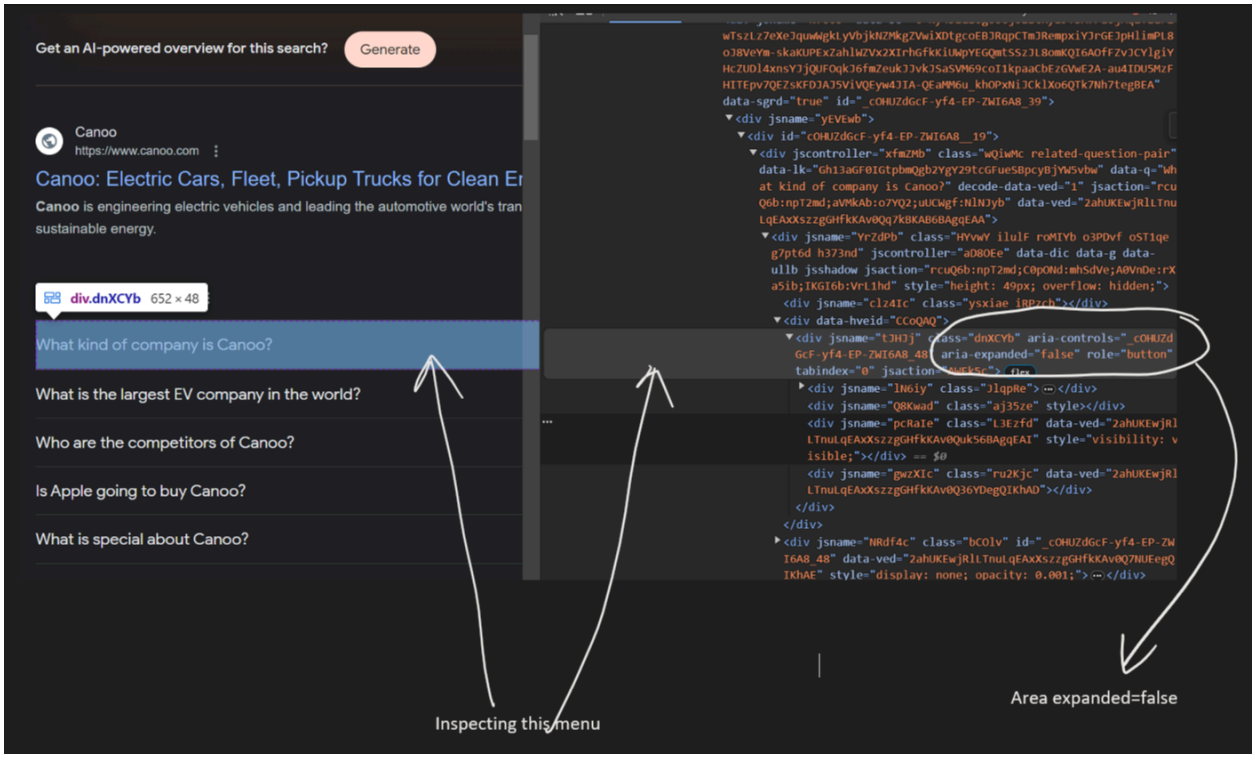
My initial goal was to utilize the ‘beautifulsoup’ library for web scraping. However, that led to a lot of problems (mentioned later) and finally I settled on using Selenium for automation and web scraping. Explanation of the code:

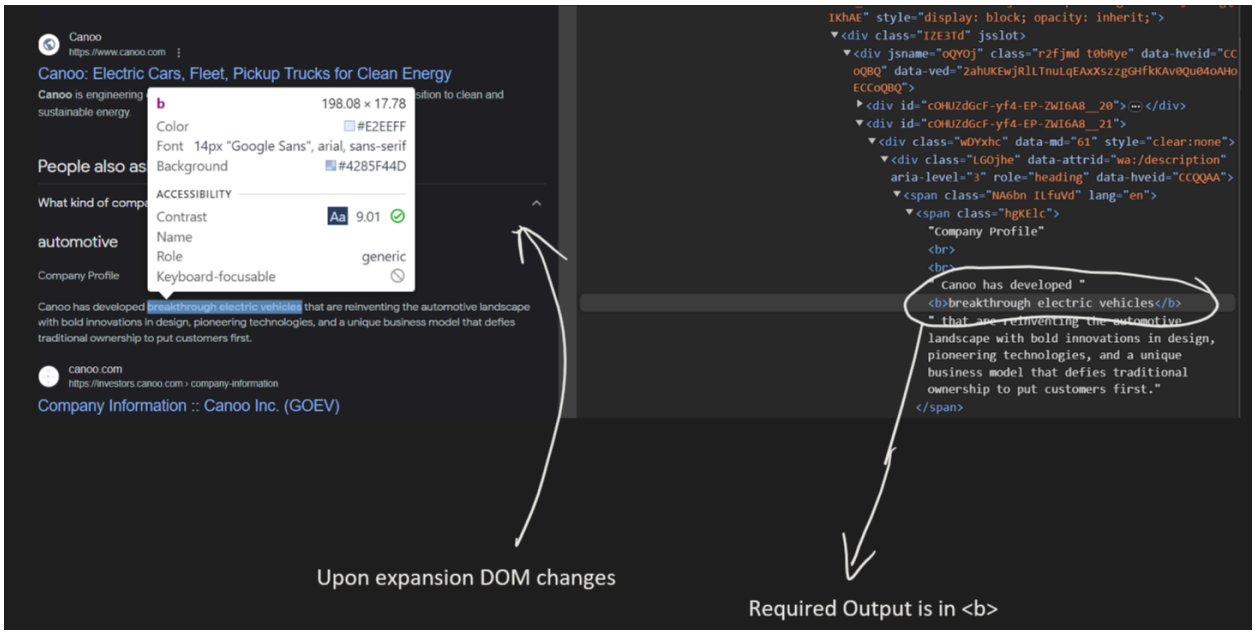
1. Driver Function
   Execution begins from the driver function driverFunc(), which takes queries, one at a time from the user. The user can press '.' to stop the execution of the code.

2. Scraper Function (scraper):

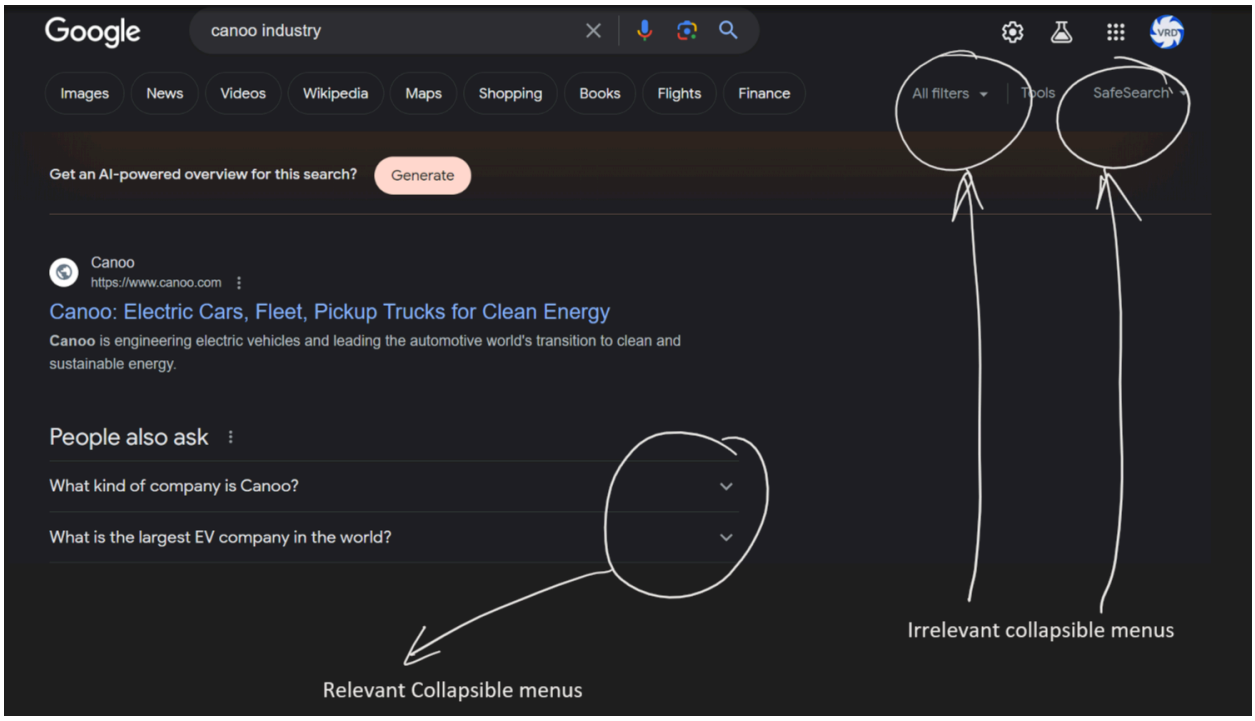
   - This function retrieves relevant results from the web as per the query q.
   - It first navigates to the Google search results page for the given query.
   - It waits for the page to load using time.sleep(2).
   - It finds all the elements associated with collapsible menus that are currently not expanded.
   - It expands all the menus by clicking on them.
   - It searches for all the <b> elements on the page (typically representing search results).
   - It iterates over the list of <b> elements, extracts the text from each element, and appends it to a list (li).
   - It calls the clean function with the list of extracted text and the query.

   The reason we need to "automate" clicking on collapsible menus is because the DOM of the page doesn't get updated to show our required element up until its parent menu is expanded.


Get an AI-powered overview for this search?
Generate
Canoo
https://www.canoo.com
Canoo: Electric Cars, Fleet, Pickup Trucks for Clean Er
Canoo is engineering electric vehicles and leading the automotive world's tran
sustainable energy.
div.dnXCYb 652 × 48
What kind of company is Canoo?
What is the largest EV company in the world?
Who are the competitors of Canoo?
Is Apple going to buy Canoo?
What is special about Canoo?
Inspecting this menu
Area expanded=false


Canoo
https://www.canoo.com
Canoo: Electric Cars, Fleet, Pickup Trucks for Clean Energy
Canoo is engineering
sustainable energy
People also as
What kind of comp
automotive
Company Profile
b 198.08 × 17.78
Color #E2EEFF
Font 14px "Google Sans", arial, sans-serif
Background #4285F44D
ACCESSIBILITY
Contrast Aa 9.01
Name
Role generic
Keyboard-focusable
Canoo has developed breakthrough electric vehicles that are reinventing the automotive landscape with bold innovations in design, pioneering technologies, and a unique business model that defies traditional ownership to put customers first.
canoo.com
https://investors.canoo.com › company-information
Company Information :: Canoo Inc. (GOEV)
Upon expansion DOM changes
Required Output is in <b>

The reason to start the for loop (inside the scraper function) from 5th element, is to skip the below mentioned expandable menus.



3. Clean Function (clean):

- This function removes blank spaces (" and ' ') from the list obtained by scraping.
- It accepts two arguments: l (the list to be cleaned) and q (the query).
- It iterates over the list and removes elements that are either empty strings or contain only spaces.
- It then calls the writeCSV function with the cleaned list and the query.

4. Write to CSV Function (writeCSV):

- This function appends the query and its result to a CSV file named canoo.csv.
- It accepts two arguments: query (the user's query) and result (a list containing the output of the web-scraped data).
- It initializes a DictWriter object to write data to the CSV file.
- If the CSV file is empty, it writes the header row.
- It joins the elements of the result list into a single string separated by commas, and then appends the query and result to the CSV file.

To summarize, the code:

- Opens a new Chrome browser window.
- Takes queries as input from the user.
- Scrapes relevant search results from Google for each query.
- Cleans and formats the results.
- Appends the query and its corresponding results to a CSV file named canoo.csv

Problems that I faced:

Issue#1

First, I tried to simply use Python's google-search library to make google searches for the given queries automatically. Then use the requests library to get a requests.Response object of the first 2-3 links.

Finally, I tried to use beautiful soup to scrape information from that object.

The problem with this approach, however, is that there is no guarantee that the first link or even the top 3 URLs have the correct information.

For Example: The google search result for the query- *"What is the industry of Canoo ?"* does not give a phrasal answer like "Electric vehicles". There is no way of finding out this answer simply by web scraping the top 3 sites. ***Regular Expressions*** and other pattern matching techniques would also not work here.

To my knowledge, one would have to use *Natural Language Processing* or *Advanced search algorithms* for that purpose.

Solution 1 "People also search for" section

I observed that the first one or two results of these sections generally contained answers to questions in highlighted form.

So, if I could get/extract this highlighted text, the problem would get solved.

Issue#2 (This issue still persists in the code)

To extract the text from the **"People also search for"** section, I utilized the fact that they are enclosed in <b> tags.

However, those specific <b> tags could not be differentiated from other <b> tags of the page.

I tried to search by background-colour CSS property, as the highlighted text has a different rgba value from the other text, **rgba(66, 133, 244, 0.3)** to be specific. But I wasn't able to isolate this in the given time.

I also tried using **Class selectors**, but the class ids change as soon as we make any interaction with the page.

So, I ended up using the text from all <b> on the page.

This made the solution work, but also introduced useless data in the CSV result. I am still working on finding a walkaround for this.

Advantages of my approach:

- It is highly generic in nature, no *if-then-else* used. You can search for almost any query and get results in the CSV file.

Disadvantages of my approach:

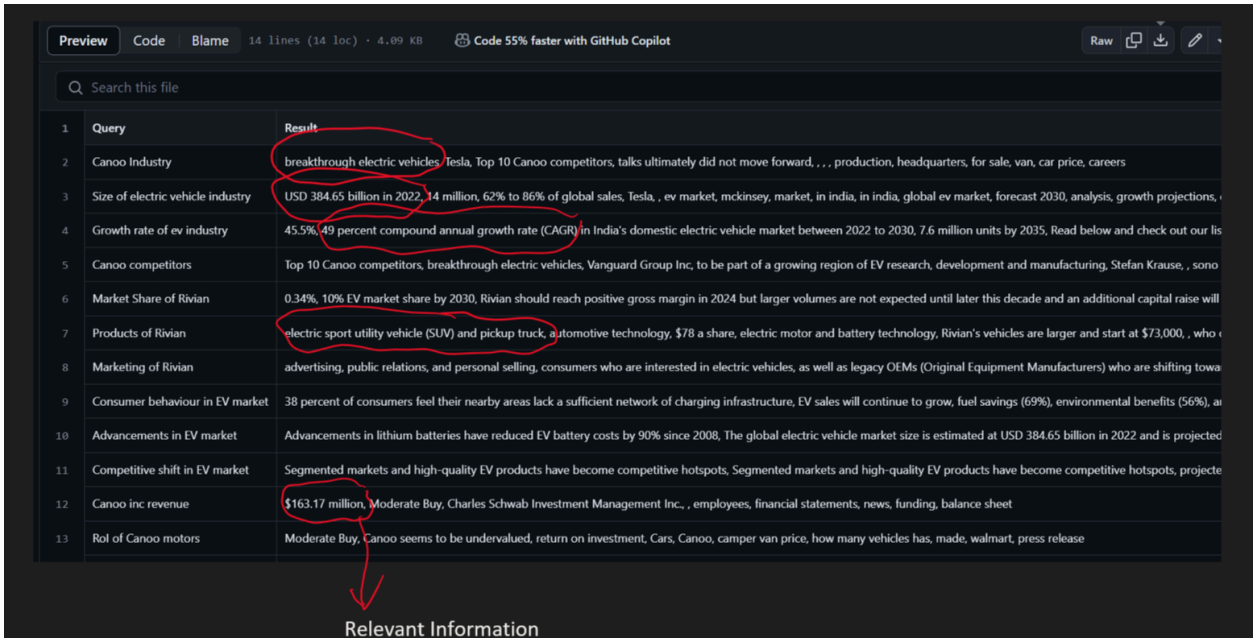
- The desired result is surrounded by irrelevant information which I am unable to remove as of now.

Sample of Output:

Raw output:

Query,Result

Canoo Industry ,"breakthrough electric vehicles, Tesla, Top 10 Canoo competitors, talks ultimately did not move forward, , , , production, headquarters, for sale, van, car price, careers"
Size of electric vehicle industry ,"USD 384.65 billion in 2022, 14 million, 62% to 86% of global sales, Tesla, , ev market, mckinsey, market, in india, in india, global ev market, forecast 2030, analysis, growth projections, ev market growth"
Growth rate of ev industry,"45.5%, 49 percent compound annual growth rate (CAGR) in India's domestic electric vehicle market between 2022 to 2030, 7.6 million units by 2035, Read below and check out our list of the top EV companies worldwide., 7.6%, , , , , , electric vehicle, projections, market size in india, market size mckinsey, electric vehicle market analysis pdf, market share by company, global, market size"



	Query	Result
2	Canoo Industry	breakthrough electric vehicles, Tesla, Top 10 Canoo competitors, talks ultimately did not move forward, , , , production, headquarters, for sale, van, car price, careers
3	Size of electric vehicle industry	USD 384.65 billion in 2022, 14 million, 62% to 86% of global sales, Tesla, , ev market, mckinsey, market, in india, in india, global ev market, forecast 2030, analysis, growth projections,
4	Growth rate of ev industry	45.5%, 49 percent compound annual growth rate (CAGR) in India's domestic electric vehicle market between 2022 to 2030, 7.6 million units by 2035, Read below and check out our lis
5	Canoo competitors	Top 10 Canoo competitors, breakthrough electric vehicles, Vanguard Group Inc, to be part of a growing region of EV research, development and manufacturing, Stefan Krause, , sono
6	Market Share of Rivian	0.34%, 10% EV market share by 2030, Rivian should reach positive gross margin in 2024 but larger volumes are not expected until later this decade and an additional capital raise will
7	Products of Rivian	electric sport utility vehicle (SUV) and pickup truck, automotive technology, $78 a share, electric motor and battery technology, Rivian's vehicles are larger and start at $73,000, , who
8	Marketing of Rivian	advertising, public relations, and personal selling, consumers who are interested in electric vehicles, as well as legacy OEMs (Original Equipment Manufacturers) who are shifting towa
9	Consumer behaviour in EV market	38 percent of consumers feel their nearby areas lack a sufficient network of charging infrastructure, EV sales will continue to grow, fuel savings (69%), environmental benefits (56%), a
10	Advancements in EV market	Advancements in lithium batteries have reduced EV battery costs by 90% since 2008, The global electric vehicle market size is estimated at USD 384.65 billion in 2022 and is projected
11	Competitive shift in EV market	Segmented markets and high-quality EV products have become competitive hotspots, Segmented markets and high-quality EV products have become competitive hotspots, projecte
12	Canoo inc revenue	$163.17 million, Moderate Buy, Charles Schwab Investment Management Inc., , employees, financial statements, news, funding, balance sheet
13	RoI of Canoo motors	Moderate Buy, Canoo seems to be undervalued, return on investment, Cars, Canoo, camper van price, how many vehicles has, made, walmart, press release

Output in tabular format on github