

Web Scraping

The company data is extracted from “https://www.crunchbase.com/organization/” website by inputting the required company name after the ‘/’ of the url of the given website;

(i.e, google https://www.crunchbase.com/organization/google)

Details’:

1. Many lists are used to store data used further combined to form a dataframe.
2. Multiple Functions are created to perform appropriate data extract operation.
3. The code can be have more than one company name as input, the names are inputting by keeping a space after one and another.

‘No. of companies: apple google microsoft’

4. **Selenium** and **BeautifulSoup** libraries are used for web page traversal and data extraction.
5. **Regex** library is used to extract data on the basis of required feature.

For example:-

```
def NA(text):  
    p1="Acquisitions"  
    if re.search(p1, text):  
        a1='YES'  
    else:  
        a1='NO'  
    return(a1)
```

6. Functions :-

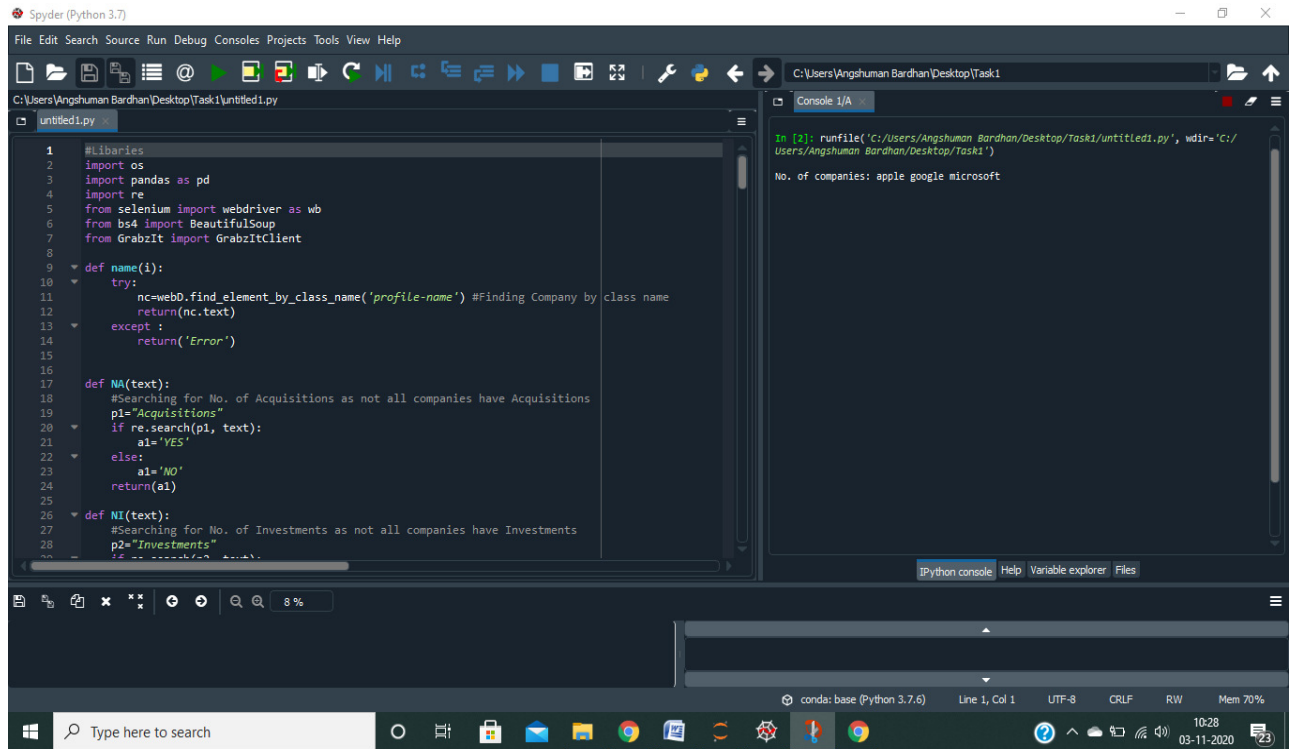
- **name()**-To extract title name of the web page which is the company name.
- **NA() and NI()**-It is used to search for the word ‘Acquisitions’ and ‘Investments’ respectively present in the company web page of the website as not all company data of the website has it.
- **Founder()**-To obtain the names of founder of the company.
- **web()**-To extract website link of company.
- **FLT()**-To extract social media link of company.
- **reduce() and nan()** -The data obtain from web() and FLT() are in a list which contain ‘None’ value and duplicate value so to remove ‘None’ elements nan() is operated and reduce() removes the duplicate values.
- **arr() and arr2()** -These two functions are used to rearrange the list format in a proper way which further used in the operation of dataframe creation.

7. Chrome driver is used for the accessing the web page data.
8. With the help of **BeautifulSoup** the html page is parsed.
9. Path logo and no. of employees are collected with necessary lists.
10. After the lists are formed, they are converted into a dataframe.
11. The dataframe is further used to create a HTML table and csv file.
12. Screenshots of the company links are taken and stored in a folder created by the code with respect to company name.

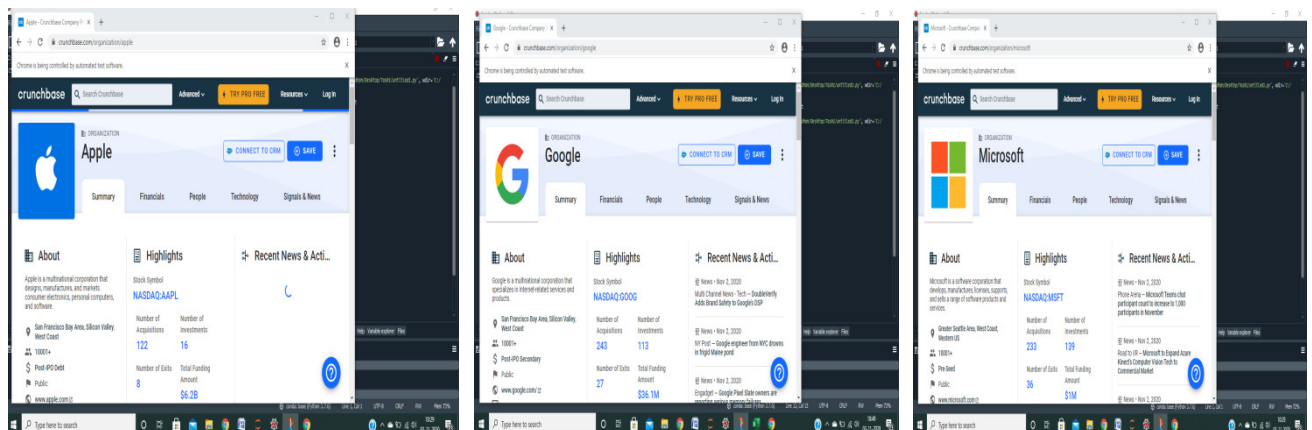
13. **Grabzit** library is used for screenshots. When this library is executed the whole code run time gets effected and some time may show error due to excessive time period.

Screenshots:

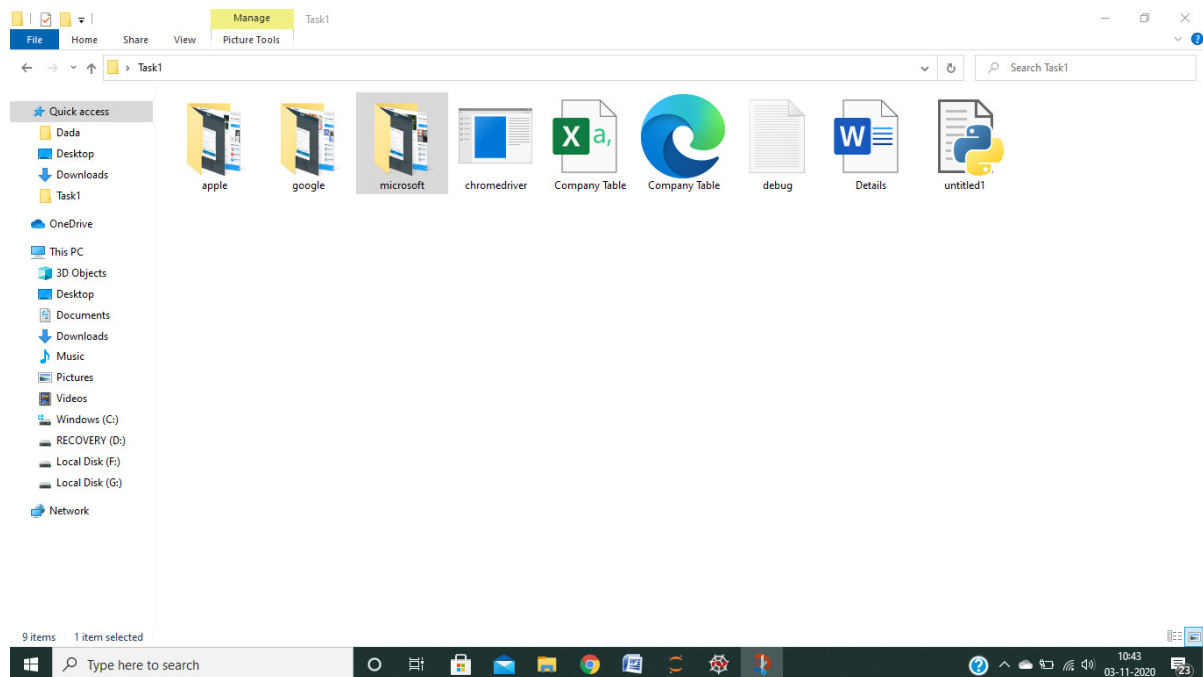
1. Giving apple, google and microsoft as company name inputs



2. Web- page opening of each company



3. Folders of company screenshots are created and screenshots are stored. HTML table and csv file is also created

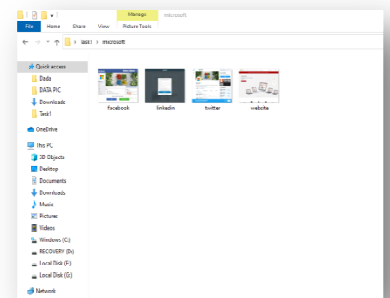
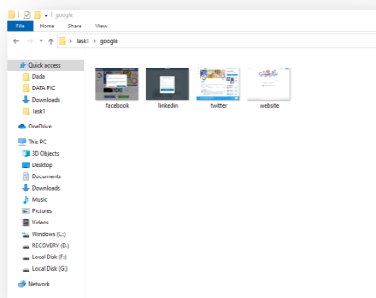
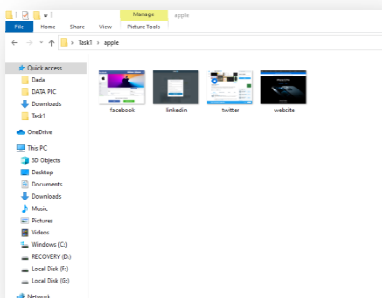


4. Link based screenshots

Apple

Google

Microsoft



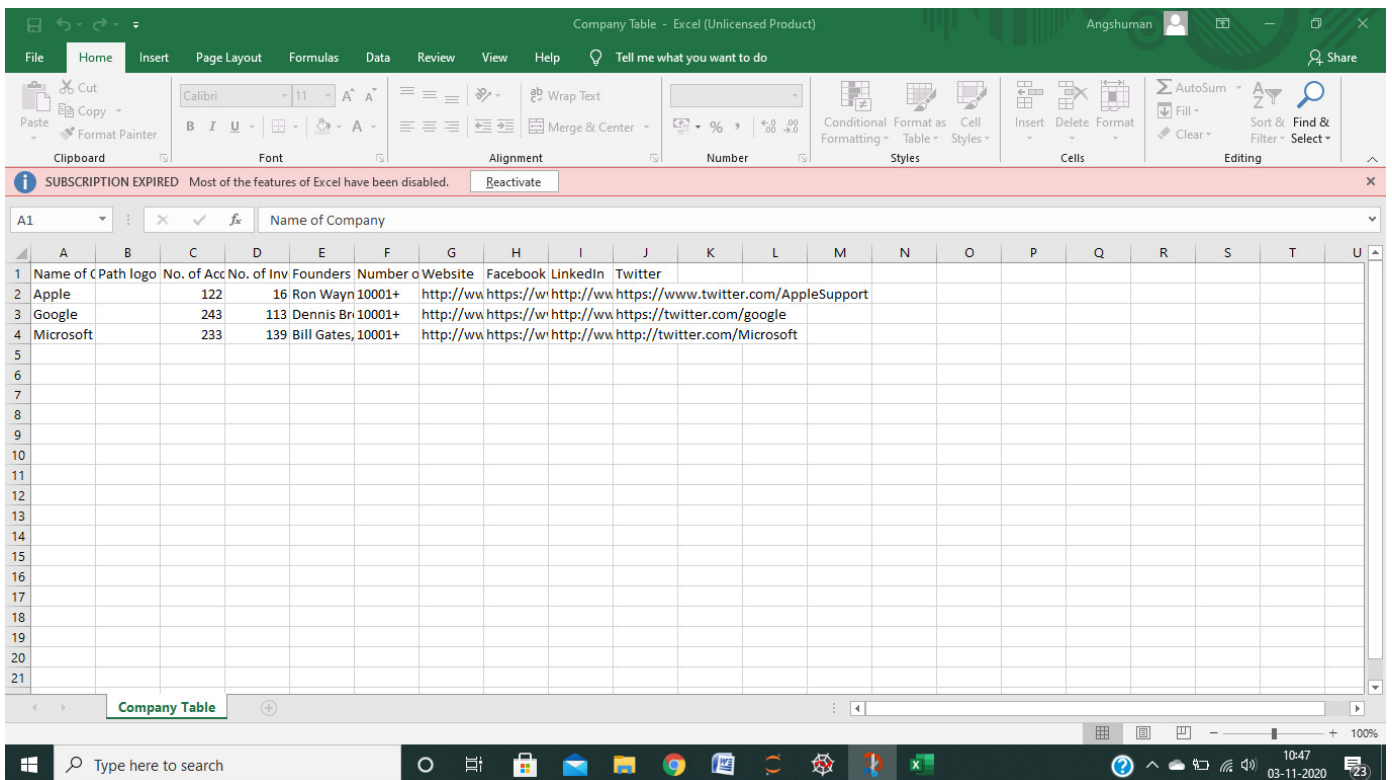
5. HTML TABLE



The screenshot shows a web browser window with the address bar displaying the file path: C:/Users/Angshuman%20Bardhan/Desktop/Task1/Company%20Table.html. The browser window contains an HTML table with 10 columns and 4 rows of data.

	Name of Company	Path logo	No. of Acquisitions	No. of Investments	Founders	Number of employees	Website	Facebook	LinkedIn	Twitter
0	Apple		122	16	Ron Wayne, Steve Jobs, Steve Wozniak	10001+	http://www.apple.com	https://www.facebook.com/apple/	http://www.linkedin.com/company/apple	https://www.twitter.com/AppleSupport
1	Google		243	113	Dennis Brown, Larry Page, Sergey Brin, Wesley Chan	10001+	http://www.google.com/	https://www.facebook.com/Google	http://www.linkedin.com/company/google	https://twitter.com/google
2	Microsoft		233	139	Bill Gates, Paul Allen	10001+	http://www.microsoft.com	https://www.facebook.com/Microsoft	http://www.linkedin.com/company/microsoft	http://twitter.com/Microsoft

6. CSV FILE



The screenshot shows the Microsoft Excel interface with the 'Company Table' file open. The data is displayed in a grid with columns A through U and rows 1 through 21. The data is as follows:

	Name of Company	Path logo	No. of Acquisitions	No. of Investments	Founders	Number of employees	Website	Facebook	LinkedIn	Twitter
1	Apple		122	16	Ron Wayne, Steve Jobs, Steve Wozniak	10001+	http://www.apple.com	https://www.facebook.com/apple/	http://www.linkedin.com/company/apple	https://www.twitter.com/AppleSupport
2	Google		243	113	Dennis Brown, Larry Page, Sergey Brin, Wesley Chan	10001+	http://www.google.com/	https://www.facebook.com/Google	http://www.linkedin.com/company/google	https://twitter.com/google
3	Microsoft		233	139	Bill Gates, Paul Allen	10001+	http://www.microsoft.com	https://www.facebook.com/Microsoft	http://www.linkedin.com/company/microsoft	http://twitter.com/Microsoft