Project report: Social data analytics - web scraping

Project description:

- For a given list of companies, The program fetches information on some important employees in the companies from Linkedin and stores the information as a csv file, “company_employee_details.csv”.

  Employee details fetched: Name, Position, company, education, location, Linkedin profile url.

- For a given list of companies, The program fetches some important information about various companies from Wikipedia and stores the information as a csv(comma separated values) file, “company_details.csv”.

  Company details fetched: Name, type, industry, founded, founders, Headquarters, area served, products/services, Revenue(US$), number of employees.

- **Current test cases worked on:** [‘Boeing’, ‘Costco’, ‘Starbucks’, ‘Paccar’, ‘Wells Fargo’, ‘Zones LLC’]

User stories:

- As a member of a marketing division, I want to be able to give a company or a list of companies as an input so that the program automatically searches the information relating to the company and top employees in the company, stores them in a csv file and displays them neatly for me to view so that I can analyze the information and know what I’m dealing with and prepare accordingly.

- As an analyst, I want to be able to get as much information as I can about a company and its key employees so that I can come up with some interesting insights that will be useful for various purposes such as in marketing, Job applications etc.

Technologies used:

Language	**Python 3.7**
Frameworks/modules	**csv**: To create csv file **Selenium, parsel**: For data extraction,

	parsing relative HTML pages **Pandas**: To view information efficiently
Misc(IDE)	**Jupyter Notebook**, file type: ".ipynb"
Browser	**Mozilla Firefox**
Operating system	**Linux,** (works on Windows too)

Functional requirements:

- Accessible through ipython notebook file.
- Provide appropriate output messages for each company scanned.
- The system should gather information from Linkedin's website and continue extraction of other information if some are not available.
- The system should write all the information collected into a csv file so it can be accessed later.
- The software should automatically gather all data and store it appropriately, without any user intervention.

Scope of improvement:

- Modification of code with less redundancies and better logic paths.
- Storage of information in databases using SQL, for very large data and more control.
- Retrieve more information from various websites. Better and more fine-tuned retrieval of current data.
- Develop an interface for user feasibility.