CS 410 Project Documentation

Team Member:
Sicheng Jiang - sj62 (group leader)
Zhaokuan Chen - zc56
David Liu - ztliu2

Project Overview:
In this project, we used the python BeautifulSoup package to parse information about UIUC faculty information, then store this information into a JSON file as structured data for future use. Then, we upload the scraped data to MongoDB Atlas. Finally, to help others easily access the scraped data, we built a backend using Django, connected the backend and the MongoDB, and implemented APIs to fetch all the data of the professors or filter the result based on the email of the professors.

Scraping Detail:
Scraping the information of professors is divided into three steps.

In step one, we find URLs to every faculty page from the faculty directory page. In detail, we parse the faculty directory page into a DOM object (a tree). Then, we search for all nodes in the DOM that has the following properties: node has an '<a>' tag with an URL, the URL contains a keyword such as "faculty" or "people", the text on the tag is likely a person's name (determined by the spacy library). Then, we extract all (relative) URLs from the nodes above and append the relative URLs onto the base URL to form absolute URLs. In the end, we return the extracted URLs as a list.

In step two, we extract information such as name, title and email from each faculty page. In detail, we get the faculty page by URL, parse the faculty page into a DOM object, and extract the following information from the DOM: Name, Title, Phone number, Email, Office address, URL, Education.

In step three, we collect all extracted information and output as a JSON object, which is then stored in MongoDB and can be accessed through a web server.
The sample of the JSON file likes like:

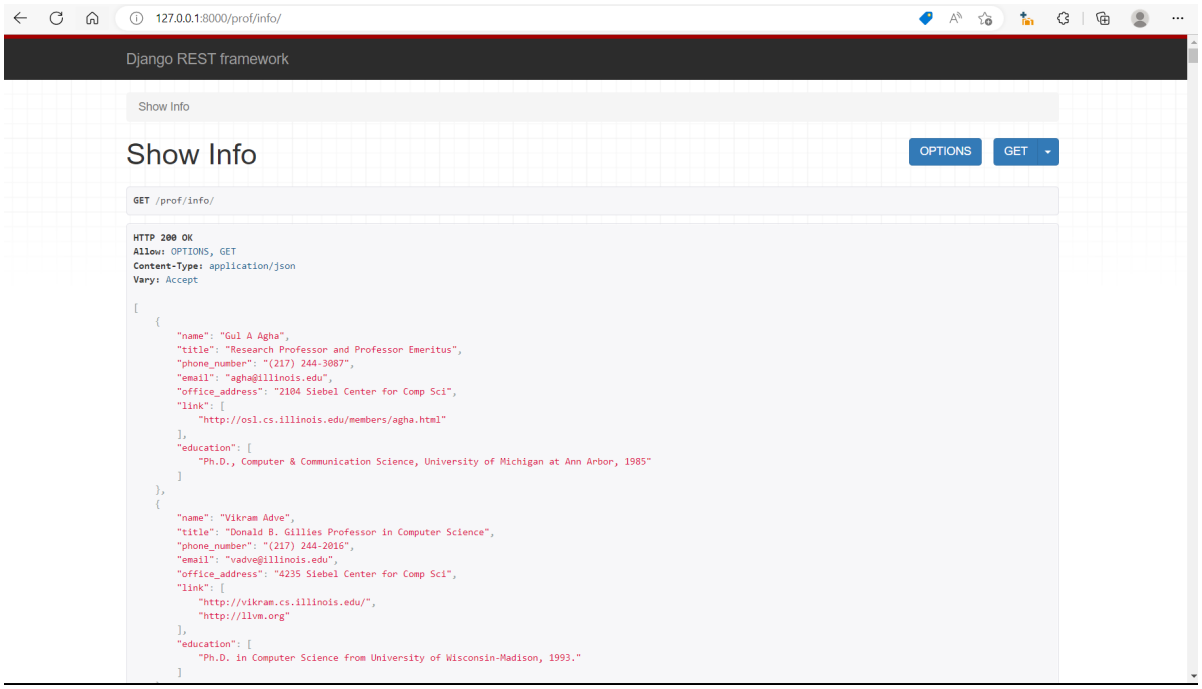

```
[
    {
        "name": "Sarita V Adve",
        "title": "Richard T. Cheng Professor",
        "phone_number": "(217) 333-8461",
        "email": "sadve@illinois.edu",
        "office_address": "4104 Siebel Center for Comp Sci",
```

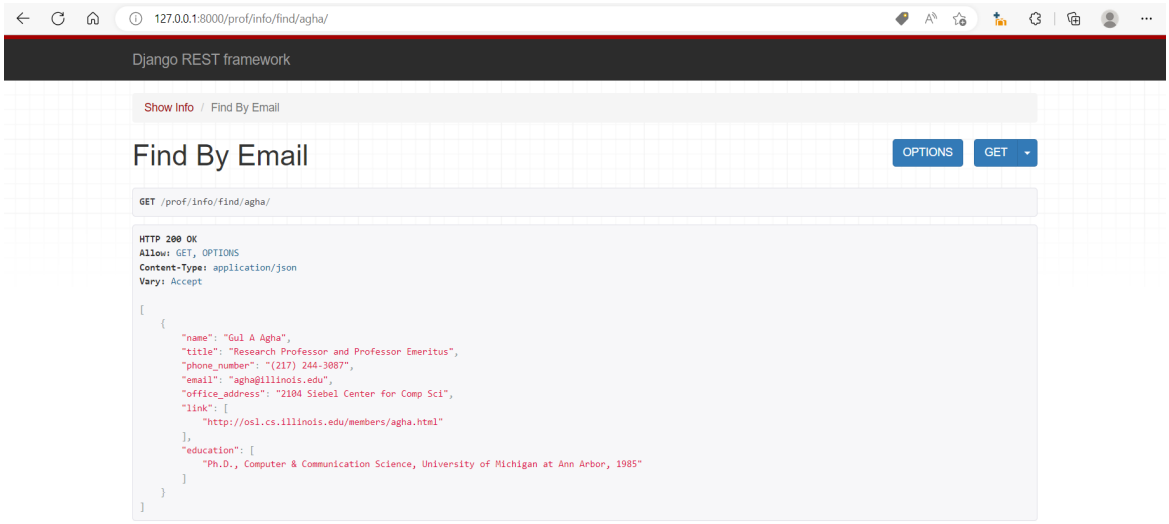
```
            "link": [
                "http://www.cs.uiuc.edu/~sadve"
            ],
            "education": [
                "Computer Science, Ph.D., University of Wisconsin-Madison, 1993"
            ]
        }
]
```


MongoDB Detail:
We uploaded the JSON into MongoDB using the following command:
mongoimport --uri mongodb+srv://CS410FinalProject:YCoH1NhOE3pI4XRU@cluster0.fhrqfoj.mongodb.net/proj --collection professor --type json --jsonArray --file "C:\Users\Zhaokuan Chen\Documents\GitHub\CS410CourseProject\faculty_info.json"

Backend Detail:
We connected the backend and the database using PyMongo. Then we created two API. The first API is “GET prof/info/” that returns all the scraped data to the User in JSON format. The second API is “GET prof/info/find/<str:email>/” that returns the professor information based on the email in the URL.

Backend Presentation:





Evaluation:

The project not only works as we described in the project proposal but also provides APIs for users to gain easy access. In the project proposal, we only planned to write a scraper. In the final build, we not only built a scraper but also uploaded the data scraped to MongoBD, implemented a backend to establish connection to the database, and provided APIs for users to gain easy access. We would say that we worked better than expected.

Contribution of Each Member:
Each of the team members contributed around 30 hours for this project. And everyone contributed equally to the project.