Tracy Dong

Write-Up

For this task, my goal was to scrape the law firm listings at OpenDI New York here for **name, address, number of ratings, and phone number**. To accomplish this, I had to split this challenge into several steps.

1. Since I had never used Scrapy before, I was not familiar with its syntax and methods. So, for the first few days, I studied and did some practice projects with Scrapy before I tackled this challenge. I learned about building Scrapy spiders as well as using its CSS selectors. I defined the goal of this project as scraping the firm name, address, number of ratings, and phone numbers.
2. When researching the website, I noticed that the HTML addresses for the subsequent pages were similar, and I could write a function to generate links to each page in **start_requests**.
3. Next, I used CSS selectors to locate the firm name, address, number of ratings and coded this in the **parse** function.
4. Since I needed to follow the link in the Details button to get the phone number, I built a helper function, **parse_content**, that took the linked URL and returned the complete firm information.
5. Finally, I used **scrapy crawl angel -o data.jl** to save the scraped information to the data.jl file.

One of the main issues I ran into was obtaining the phone number before I saved the firm information to the data.jl file. Initially, either the parse_content function wouldn't run, or the phone numbers were found after saving the other components. After a lot of research, I learned that I could pass the incomplete version of the firm information down to parse_content, and that function could add the phone number property to the data before saving it to the data.jl file. If I were to do this challenge again, I would find a way of accessing the next page of the website dynamically instead of hardcoding it in the start_requests function. Overall, I really enjoyed programming this and learning about Scrapy! Before completing this challenge, I had used Requests before, and I can say that Scrapy is far more versatile and exciting.