Project Progress Report

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

We have already finished researching the necessary technologies, to identify and crawl faculty directory pages and faculty webpages. After doing quite a bit of research online, we decided to prefer Beautiful Soup over Scrapy Python library. Due to the fact that scrapy cannot handle Javascripts, provide limited resources online, and all the group members are familiar with Python, we decided to use Beautiful Soup instead of Scrapy. Meanwhile, we also finished setting up a consistent environment for each team member using Google Colab instead of Docker because it is better for collaboration.

In the next step, we will start implementing and testing the crawler. As per TA's feedback, we decided to expand the scope of the project by formatting the scraped data into a JSON file and uploading it to GitHub repo so that the scraped data can be used again by future students.

There are a few challenges we are facing: teammates are in different time zones, little knowledge of Python and the Beautiful Soup library, and the need to learn JSON structure.