**Họ tên SV: Đỗ Minh Vũ**

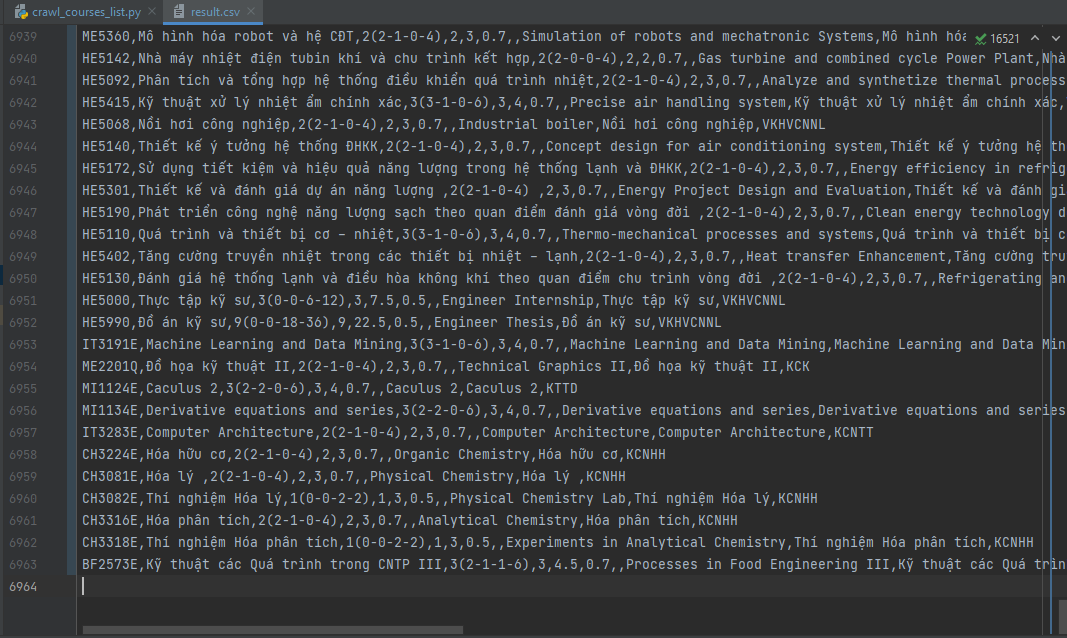
**MSSV: 20173471**

**Báo cáo hằng tuần**

**(Tuần thứ 1: 25/10/2021 – 31/10/2021)**

1. Kết quả làm việc

* Crawl tất cả dữ liệu (6962 môn học) từ 349 page, cùng với phần thông tin bổ sung của tất cả môn học.
* Lưu kết quả dưới dạng file csv như hình:



1. Các bước thực hiện

* Cài đặt thư viện Selenium để mô phỏng thao tác chuyển trang.

***pip install selenium***

* Chỉnh sửa code để có thể lấy toàn bộ dữ liệu trong nhiều trang. Source Code như trong link github:

***https://github.com/dominhvu2605/Project3***

* Xử lý: với mỗi trang, tính số lượng môn học trong danh sách hiển thị. Sử dụng vòng lặp for để lấy dữ liệu từng môn:

+ Mỗi môn sẽ thực hiện giả lập ấn nút mở rộng để xem thông tin môn học, lưu các dữ liệu chi tiết đó vào biến.

+ Cào dữ liệu trong bảng đối với mỗi môn học.

+ Ghi dữ liệu thu thập được của từng môn học vào file csv.

* Để chạy code, gõ dòng lệnh sau tại terminal của IDE (Ghi đè kết quả lên file nếu file đã tồn tại)

***scrapy crawl course -O result.csv***

* Kết quả thu được như mục 1.

1. Các vấn đề và hướng giải quyết

* Các vấn đề
* Tuy đã crawl được tất cả dữ liệu nhưng thời gian chạy tốn nhiều thời gian. Nguyên nhân có thể do driver xử lý các response chậm, do đó cần sử dụng hàm sleep() để chờ nhận được tất cả source code khi thực hiện ấn nút.
* Vẽ ma trận quan hệ của các học phần.
* Hướng giải quyết
* Tìm hiểu thêm các hàm để xử lý ngoại lệ, bắt hết các thành phần khi thực hiện ấn nút mà không phải cài đặt thời gian cố định cho hàm sleep().
* Tìm hiểu về thư viện Graphviz để vẽ ma trận quan hệ của các môn học.