BÁO CÁO TUẦN

Thời gian: 22/11 – 29/11/2000

Danh sách nhóm

- Nguyễn Thế Viễn – MSSV: 1814764
- Nguyễn Ngọc Thuấn – MSSV: 1814217
- Phạm Đức Duy Anh – MSSV: 1810814
- Bùi Hữu Đang – MSSV: 1811828
- Huỳnh Ngọc Tấn – MSSV: 1813592

Các công việc đã thực hiện

- Tìm hiểu về scrapy (Thuấn và Viễn) đã push lên git
- Tìm hiểu về beautiful soup (Duy Anh, Tấn, Đang) đã push lên git
- Tìm hiểu về selenium (Viễn và Thuấn) đã push lên git
- Hiện thực crawler từ trang https://careerbuilder.vn/ (Đang) đã push code lên git.
- Hiện thực crawler từ trang https://www.chotot.com/ (Duy Anh) đã push code lên git.
- Hiện thực crawler từ trang https://www.vietnamworks.com/ và https://mywork.com.vn/ (Thuấn và Viễn) đã push code lên git.
- Hiện thực crawler từ trang http://www.vietsingworks.com/ (Thuấn) đã push code lên git.
- Hiện thực crawler từ trang http://1001vieclam.com/ (Tấn) đã push code lên git.

Khó khăn gặp phải

- Cài đặt thư viện scrapy của python bị lỗi, đã khắc phục bằng cách cài Twisted tương thích với máy
- Cấu trúc file HTML của chotot.vn bị lỗi, đã khắc phục bằng cách đưa về chuỗi và dùng các method của python để xử lý
- Trang vietsingwork.com, mục BENEFIT và OTHER REQUIREMENTS của mỗi công việc có format khác nhau, nên chưa crawl được, đang tìm cách xử lý
- Trong quá trình crawl link chưa xác định được lỗi gây ra việc chỉ crawl được 9 công việc đầu ở trong một trang .Trang vietnamwork.com không thể lấy hết các đường link, vì tất cả các trang chưa các job đều được gộp thành 1 đường link.

Dự kiến công việc tuần tới:

- Triển khai công cụ crawl dữ liệu để có được dataset khoảng 10.000 job
- Cải tiến công cụ crawl dữ liệu để tạo thành một công cụ hoàn chỉnh (dùng cho web không public API, yêu cầu đăng nhập, dùng javascript load động).
- Tìm hiểu các hàm làm sạch, chuẩn hóa và phân loại dữ liệu
- Thiết kế database, công cụ đưa dữ liệu từ file lên database