BÁO CÁO TUẦN

Thời gian: 30/11 - 06/12/2020

Danh sách nhóm

- Nguyễn Thế Viễn – MSSV: 1814764

- Nguyễn Ngọc Thuấn – MSSV: 1814217

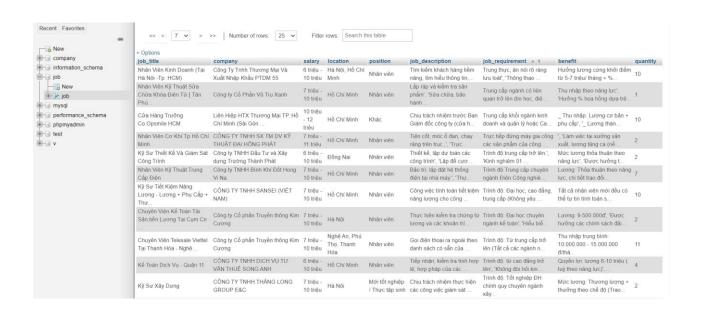
- Phạm Đức Duy Anh - MSSV: 1810814

- Bùi Hữu Đang – MSSV: 1811828

- Huỳnh Ngọc Tấn – MSSV: 1813592

Các công việc đã thực hiện

- Thiết kế database, công cụ đưa dữ liệu từ file JSON lên database MySQL (Viễn).
- Tìm hiểu các định dạng file biểu diễn dataset (JSON) (Viễn).
- Tìm hiểu chuẩn hóa dữ liệu, phân loại dữ liệu (Viễn, Thuấn, Duy Anh, Đang, Tấn).
- Triển khai việc Crawl dữ liệu và clean data 1000 jobs từ trang http://www.vietsingworks.com/ https://www.vietnamworks.com/ (Thuấn, Viễn) đã đã push code lên git.
- Triển khai việc Crawl dữ liệu và clean data 3000 jobs từ trang https://www.chotot.com/ (Duy Anh) đã đã push code lên git.
- Triển khai việc Crawl dữ liệu và clean data 4000 jobs từ trang https://careerbuilder.vn/ (Đang) đã đã push code lên git.
- Triển khai việc Crawl dữ liệu và clean data 2000 jobs từ trang http://1001vieclam.com/ (Tấn) đã đã push code lên git.



Khó khăn gặp phải

- Việc làm sạch dữ liệu còn nhiều khó khăn.

Dự kiến công việc tuần tới:

- Source code làm sạch dữ liệu, source code lưu trữ dữ liệu raw (dataset raw) vào database.
- Source code truy vấn dữ liệu, model data.
- Tiến hành tìm hiểu về một số kỹ thuật phân tích dữ liệu để có thể phân tích bộ dữ liệu mà mình đã thu được.
- Thảo luận đề xuất một ứng dụng cơ bản trên bộ dữ liệu mà nhóm đã crawl trên dữ liệu này (có thể apply một model AI hoặc đơn giản là build một ứng dụng có UI với dataset này...).