BÁO CÁO TUẦN 2
Thời gian: 30/11 – 6/12/2020

Danh sách nhóm:
- 1920058 - Lê Tất Thiện
- 1927038 - Trần Phương Tĩnh
- 1920068 - Dương Quang Tuấn
- 1927040 - Nguyễn Đức Tuấn

Các công việc đã thực hiện
- Viết prototype function crawlWeb (Thiện), đã update vào file module/tinh.js trong repository DAKTLT-Backend.
- Hoàn thiện function crawlWeb và chạy thử (Tĩnh), đã update vào file module/tinh.js trong repository DAKTLT-Backend.
- Tìm hiểu format chuẩn, cách làm sạch data set (Tuấn Nguyễn).
- Tuần 1 đã crawl 10k record và gửi cho Tĩnh (Tuấn Dương), tuần 2 làm sạch và đưa về format chuẩn mong muốn (Tĩnh, Thiện), đã push file 10kRecords.json lên folder samplerecord trong repository DAKTLT-Backend.
- Tìm hiểu và setup frontend (Tuấn Dương, Tuấn Nguyễn), đã push lên repository DAKTLT-Frontend.
- Kết nối backend với mongoDB, insert 10k record vào mongoDB và tạo các model cho web (Thiện), đã push lên repository DAKTLT-Backend.
- Viết báo cáo tuần 2 (Tĩnh), đã push báo cáo lên folder report/tuan2 trong repository DAKTLT-Backend.

Các khó khăn gặp phải
- Web https://www.muabannhadat.com.vn/ bị sập nên chuyển sang crawl web khác.
- 10k record được crawl về ở tuần 1 chưa đủ thuộc tính, chưa sạch và chưa đúng format chuẩn mong muốn, ở tuần 2 này nhóm đã viết thêm function crawlWeb để crawl trực tiếp từng mẩu tin của trang chính, địa chỉ để crawl lấy từ 10k record cũ. Từ đây crawl được 10k record mới đủ thuộc tính, sạch, và đúng format chuẩn mong muốn, lưu vào file json, sau đó insert vào mongoDB.

Dự kiến công việc tuần tới
- Sử dụng MongoDB để lưu trữ và graphQL để thao tác và truy vấn API.
- Backend: Viết API trả về dữ liệu.
- Frontend: Hiển thị dữ liệu.