**Báo Cáo Tuần( Task 2.2)**

Thời gian từ ngày 7/12/2020 – 13/12/2020

**1. Danh sách nhóm DA\_KTLT:**

* Lộc Quốc Huy – 1812369
* Nguyễn Việt Hưng – 1812499
* Trần Quang Huy – 1812427

**2. Các công việc đã thực hiện:**

* Dataset chưa đủ lớn nên nhóm đã crawl thêm dữ liệu để mỗi trang web có khoảng 10.000 item. Phân chia crawl data như sau: Quang Huy crawl Tiki Lazada, Hưng crawl Nguyenkim Dienmaycholon, Quốc Huy crawl Shopee.
* Source code làm sạch dataset của mỗi trang thu được phía trên, phân công làm sạch data tương ứng với phân công công việc phía trên: Loại bỏ các kí tự đặc biệt, loại bỏ các icon, format giá sản phẩm, đồng thời xóa đi các item trùng nhau. Kết quả thu được là dataset đã làm sạch file data.csv
* Source code cập nhập giá thành khi giá thành sản phẩm thay đổi(Quang Huy): So sánh dataset mới crawl với dataset cũ để tự động cập nhập giá thành sản phẩm.
* Tìm hiểu về cách phân tích dữ liệu và cách truy vấn được dữ liệu (Quốc Huy)

**3. Các khó khăn gặp phải:**

Trong quá trình làm task nhóm gặp phải các khó khăn sau:

* Khi làm sạch dữ liệu thì chưa hoàn toàn triệt để, vẫn còn một vài icon đặc biệt chưa xóa được
* Source code cập nhập giá thành còn khá đơn giản và độ phức tạp của giải thuật là O(n2) nên khi với dataset lớn thì chạy khá chạy nên cần cải thiện thêm.

**4. Dự kiến công việc tuần tới:**

* Cải thiện source code cập nhập giá thành sản phẩm hiệu quả hơn.
* Phân tích dữ liệu thu được.
* Xây dựng web/app với IU để dễ dàng truy vấn dữ liệu.