ĐẠI HỌC BÁCH KHOA THÀNH PHỐ HỒ CHÍ MINH
KHOA KHOA HỌC & KĨ THUẬT MÁY TÍNH

ĐỒ ÁN KỸ THUẬT LẬP TRÌNH

BÁO CÁO TUẦN

Task 2.1: Tìm hiểu và hiện thực các công cụ Crawl dữ liệu theo yêu cầu

---

GVHD: Lưu Quang Huân

SV thực hiện: Lộc Quốc Huy -1812369
Nguyễn Việt Hưng -1812499
Trần Quang Huy - 1812427

Mục lục

1. **Các công việc đã thực hiện……**

2. **Các khó khăn đã gặp phải……**

3. **Dự kiến công việc tuần tới……**

4. **Link Github……**

1. Các công việc đã thực hiện.

- Nhóm sử dụng ngôn ngữ lập trình Python cùng với tham khảo các tài liệu để crawl dữ liệu về giá cả thông tin đồ điện tử, điện lạnh, điện thoại từ các trang website thương mại điện tử **tiki, shopee, lazada, nguyenkim, dienmaycholon** và làm báo cáo (đã commit lên Github).
- Cụ thể bạn Lộc Quốc Huy: crawl dữ liệu từ **shopee** và làm báo cáo.
- Bạn Nguyễn Việt Hưng crawl dữ liệu từ **nguyenkim, dienmaycholon.**
- Bạn Trần Quang Huy crawl dữ liệu từ **tiki, lazada** và mọi người cùng nhau tìm và chia sẻ các tài liệu tham khảo.

2.Các khó khăn gặp phải.

Trong quá trình thực hiện task nhóm có gặp phải một số khó khăn như sau:

- Không thể lấy dữ liệu từ các trang có sử dụng load dữ liệu động. Cách khắc phục nhóm đã sử dụng Selenium là bộ kiểm thử tự động miễn phí (mã nguồn mở) dành cho các ứng dụng web trên các trình duyệt và nền tảng khác nhau.
- Cài đặt Selenium cũng như import một số thư viện của Python. Cách khắc phục các thành viên đã teamview hướng dẫn cài và giúp đỡ nhau.
- Chưa xác định đúng được đường dẫn tới các thẻ chứa thuộc tính của HTML lấy được từ website. Cách khắc phục nhóm đã tham khảo các tài liệu và hướng dẫn nhau hoàn thành nhiệm vụ.

3.Dự kiến công việc tuần tới.

- Cả nhóm dự kiến tuần tới sẽ làm sạch, chuẩn hóa và phân loại các dữ liệu.
- Thiết kế cơ sở dữ liệu, các định dạng file biểu diễn dataset.
- Xây dựng bộ dataset(~10000), cải tiến công cụ (source code đã xây dựng).

4.Link Github

https://github.com/tqhuy-bk/201CO10313_DA_KTLT.git