PROJECT 1

BÁO CÁO ĐỒ ÁN VỀ CRAWLER DỮ LIỆU

I/ THÔNG TIN

Author	version	Description
20424030 - Đặng Trung Hiếu	1.0	Tham gia viết script crawl của trang Foody
20424025 - Nguyễn Thị Thúy Hằng	1.0	Tham gia viết script crawl của trang Google review
20424051 - Nguyễn Thành Long	1.0	Xây dựng core,viết script crawl của trang ShopeeFood, Login foody và hỗ trợ các plugin khác

II/ KẾ HOẠCH PHÂN CÔNG

Tuần	20424030	20424025	20424051
Tuần 1	Họp về vấn đề chọn chủ đề và các thông tin cần crawl, chọn trang web để crawl. Chọn ngôn ngữ lập trình và thống nhất cấu trúc Project	Họp về vấn đề chọn chủ đề và các thông tin cần crawl, chọn trang web để crawl. Chọn ngôn ngữ lập trình và thống nhất cấu trúc Project	Họp về vấn đề chọn chủ đề và các thông tin cần crawl, chọn trang web để crawl. Chọn ngôn ngữ lập trình và thống nhất cấu trúc Project Tạo source code trên github
Tuần 2	Bắt đầu tiến hành nghiên cứu và code để crawl dữ liệu ở trang Foody	Bắt đầu tiến hành nghiên cứu và code để crawl dữ liệu ở trang Google Review	Bắt đầu tiến hành nghiên cứu và code để crawl dữ liệu ở trang Shopee Food và kiểm tra tiến độ thành viên để support nếu cần
Tuần 3	Họp thống nhất kế hoạch và nêu các khó khăn với nhau để cùng	Họp thống nhất kế hoạch và nêu các khó khăn với nhau để	Họp thống nhất kế hoạch và nêu các khó khăn với nhau

	nhau nghĩ cách giải quyết. Sau khi thống nhất tiếp tục code xử lý những khó khăn gặp phải	cùng nhau nghĩ cách giải quyết. Sau khi thống nhất tiếp tục code xử lý những khó khăn gặp phải	để cùng nhau nghĩ cách giải quyết. Sau khi thống nhất tiếp tục code xử lý những khó khăn gặp phải
Tuần 4	Chạy kiểm kết quả lần cuối và bắt đầu hoàn thành báo cáo	Chạy kiểm kết quả lần cuối và hỗ trợ viết và hoàn thành báo cáo	Tổng hợp 3 bản crawl vào Project, đảm bảo project hoạt động ổn nhất.
Tuần 5	Hoàn thiện báo cáo và đóng gói	Hoàn thiện báo cáo và đóng gói	Thực hiện crawl data demo và hoàn thiện báo cáo và đóng gói.

III/ ĐÁNH GIÁ BÁO CÁO

STT	Tiêu chí	Giải quyết vấn đề	Tỉ lệ hoàn thành
1	Trình bày chủ đề, lý do chọn chủ đề, trang web lấy hoặc trang web là hạt giống.	- Thống nhất chọn chủ đề " Ăn uống " và các trang web dùng để crawl là: Foody, Shopee Food, Google Review. - Sử dụng chủ đề ăn uống vì sau khi nghiên đây là sở thích của các thành viên và là chủ đề được quen thuộc của người dùng, các trang web về ăn uống có rất nhiều và đa dạng, những trang web lớn và có tính ổn định, dữ liệu để crawl về để xử lý là nhiều và đa dạng.	10%
2	Mô tả thuật toán, cấu trúc mã nguồn, các thành phần hệ thống	Cấu trúc mã nguồn chia làm các phần chính: 1. config: chứa các file cấu hình chung như môi trường, database, logger, launcher,... 2. database: chứa dữ liệu crawler 3. utils: chứa các hàm hỗ trợ: db, lock, metric, signal, tracert 4. model: mô tả mô hình lưu trữ `website_id_quan: str`	30%

		website: str url: str district: str rate: float active_time: str full_name: str = None phone: str = None rate_count: float = None favorite: float = None price_from: float = None price_to: float = None other_service: str = None 5. plugin: chứa thuật toán chính của từng website: Foody, GoogleReview và ShopeeFood Các thành phần trong hệ thống: 1. Hệ thống menu lựa chọn 2. Hệ thống logger 3. Hệ thống đo lường Thuật toán chính sẽ trình bày phía dưới.	
3	Các vấn đề xảy ra đối với crawler và phương pháp xử lý	- Vấn đề lấy trang: kích thước frontier tăng lên khá nhanh, nên sẽ giới hạn lại trước khi lấy tiếp; với tốc độ mạng chậm, thì sẽ xử lý bằng cách bỏ qua hoặc chờ 1 khoảng thời gian nếu không tải thành công thì bỏ qua - Vấn đề phân tích trang: Với những trang SPA, hoặc có sử dụng ajax, webpack sử dụng phát sinh mã id,class rất khó, => chưa có phương án xử lý => có thể bỏ qua nếu lỗi. - Trích xuất liên kết: khi sử dụng thư viện python thì đã xử lý các vấn đề url relative - Lưu trữ data: sử dụng 1 field định danh làm khóa chính lưu vào database SQLite - Vấn đề xử lý song song: áp dụng đa luồng trong python để giải quyết - áp dụng kĩ thuật lock để tránh tranh chấp xử lý - Trang Shopee Food phải khai thác api thông qua phân tích - Trang Foody phải Login mới có thể crawl được tất cả dữ liệu. - Trang Google Review phải sử dụng	10%

		selenium để fetch và extract data	
4	Các tính năng phức tạp của crawler	- Giới hạn lượng thông tin cần láy - Xử lý tốc độ mạng không ổn định - Thao tác qua database, độ ổn định cao - Áp dụng đa luồng để crawl, độc lộc nhau - Cho phép lựa chọn chủ đề **ĐỘNG** - Cho phép lựa chọn website crawl cùng lúc - Sử dụng api đã được phân tích cho kết quả nhanh - Áp dụng kĩ thuật vào selenium để tăng hiệu suất craw data (headless,network,cookie,...) - Có nhiều màn hình để theo dõi từng lường, thống kê - Có chức năng thống kê tốc độ lấy dữ liệu	10%
5	Đánh giá hiệu năng crawler	- Đa luồng nên khả năng khai thác dữ liệu tối đa - Sử dụng api được phân tích trước đó đem lại tốc độ cao - Khả năng chịu lỗi cao không gây crash ứng dụng - Dữ liệu thống kê lấy về có giá trị rất cao - Số lượng thông tin lấy về cao hơn những gì team mong đợi	5%
6	Mô tả và đánh giá dữ liệu thu thập được.	- Tổng quan từ 3 trang Foody, Shopee Food, Google Review - Dữ liệu đề ra để lấy và lưu vào DB: (website_id, website, full_name, district, rate, active_time, price_from, price_to, favorite) - Các dữ liệu sau khi crawl và xử lý về, thu được hợp lệ so với đề ra, do cấu trúc trang web các dữ liệu lấy về đôi khi có vài trang bị lỗi và phải catch xử lý - Dữ liệu rất đa dạng	5%
7	Tiền xử lý dữ liệu thu thập	Format lại dữ liệu quận lowercase Format lại để lấy được url đầy đủ Format lại số Rate của từ 3 trang web theo 1 định dạng để lưu vào DB Format lại số phone của quán theo 1 định dạng thống nhất Format lại giờ hoạt động của quán **Mẫu vd chưa xử lý tiền dữ liệu**: - website: Foody - website_url: /ho-chi-minh/sang-trang/foods/san-fu-lu - name: San Fu Lu - phone: (+84) 824110244 - time: 10h30 - 20h30 - rate: 3.7	10%

		- district: quan-1 - time: 10h30 - 20h00 || chưa cập nhật - price: 500.000vnd - 1.000.000vnd **Mẫu vd sau khi xử lý tiền dữ liệu:** - website: Foody - website_url: https://www.foody.com.vn/ho-chi-minh/sang-trang/foods/san-fu-lu - full_name: San Fu Lu - phone: 0824110244 - time: 10h30 - 20h30 - rate: 7.4 - district: 1 - time: 10h30 - 20h00 || None - price_to: 500.000 - price_from: 1.000.000	
8	Báo cáo rõ ràng các mục đã thực hiện, có thể hiện mức độ hoàn thiện của từng công việc	- Báo cáo đã trình bày đủ các mục đề ra, trong đó có cả giải thích các bước đầy đủ và tỉ lệ hoàn thành	20%
Tổng			100%

IV/ GIẢI THUẬT:

1. Crawl theo chiều sâu: tải các trang từ cùng 1 mức trước khi chuyển sang trang kế tiếp

2. Trong quá trình crawl sẽ có thống kê và thể hiện tốc độ lấy dữ liệu.

3. Foody:
    a. Login: người dùng sẽ sử dụng tài khoản của mình để thực hiện login, từ đây, chúng ta export ra cookie và chia sẻ cho lần chạy sau.
    b. Thực hiện lấy chủ đề lớn: request lên trang template lấy ra data và cho người dùng chọn lựa
    c. Thực hiện crawl data url từ trang số 1 ứng với chủ đề người dùng chọn cho vào frontier
    d. Lặp trên frontier
    e. Sẽ có 2 trường hợp:
        i. Trường hợp là link danh sách chi nhánh: ta thực hiện fetch data url chi nhánh và lấy tất cả link cửa hàng cho vào frontier

ii. Trường hợp là link của hàng, ta thực hiện áp dụng request-html để lấy data và trích xuất lưu vào db

f. Kiểm tra có tồn tại trang thứ 2 không, nếu có ta lặp lại bước C với trang thứ 2. Ngược lại ta kết thúc quá trình crawl và thông báo

4. ShopeeFood:
    a. Gọi api lấy ra danh sách quận và danh sách chủ đề lớn cho người dùng chọn lựa
    b. Thực hiện lấy tất cả id cửa hàng ứng với chủ đề người dùng đã chọn
    c. Chia nhỏ danh sách các cửa hàng thực hiện thành mỗi cụm 25 cửa hàng
    d. Lặp trên từng id cửa hàng
        i. Truy vấn api lấy data từ id cửa hàng
        ii. Thực hiện chuẩn hóa
    e. Kết thúc quá trình crawl

5. Google Review:
    a. Cho phép người dùng tự lựa chọn chủ đề và nhập vào cho chương trình
    b. Từ keyword người dùng, chương trình sẽ tiến hành thiết lập selenium và khai thác dữ liệu từ trang google review
    c. Kiểm tra trên trang có tồn tại cửa hàng nào không:
        i. Nếu có thì thực hiện lấy data thông qua tương tác click và bóc tách dữ liệu
        ii. Nếu không thì kết thúc quá trình crawl
    d. Nếu chưa kết thúc quá trình, ta kiểm tra xem có nút qua trang hay không?
        i. Nếu có thì ta tương tác click vào và đợi dữ liệu load xong quay lại bước c
        ii. Nếu không thì kết thúc quá trình crawl

V/ MÃ NGUỒN CHƯƠNG TRÌNH:

Link drive: Tai đây

Link github (dự phòng): Tai đây

VI/ BIÊN DỊCH và CHẠY CHƯƠNG TRÌNH

- Project sử dụng python 3, hệ thống lưu trữ là SQLite
- Hướng dẫn chạy chương trình:
    + Đọc hướng dẫn cài đặt file “README.md”

Preview README.md

WebScience-TopicalCrawler

Thu Thập Dữ Liệu Từ Web

- Cài đặt các package:

```
python -m pip install -r requirements.txt
```

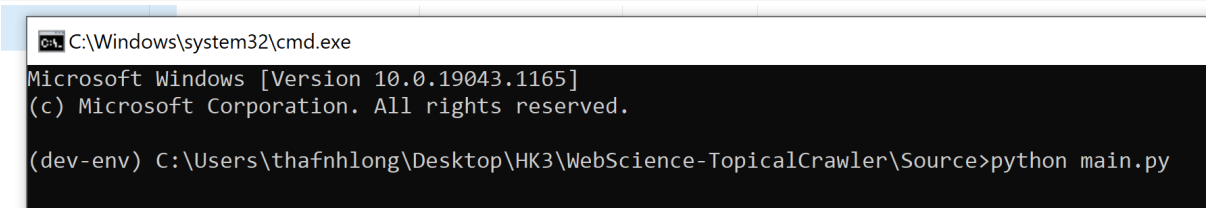
- Tải webdriver giải nén và đưa vào thư mục `webdriver`
  - Chrome: https://sites.google.com/a/chromium.org/chromedriver/downloads
  - Edge: https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/
  - Firefox: https://github.com/mozilla/geckodriver/releases
  - Safari: https://webkit.org/blog/6900/webdriver-support-in-safari-10/
- File cấu hình môi trường:

> /config/env.py

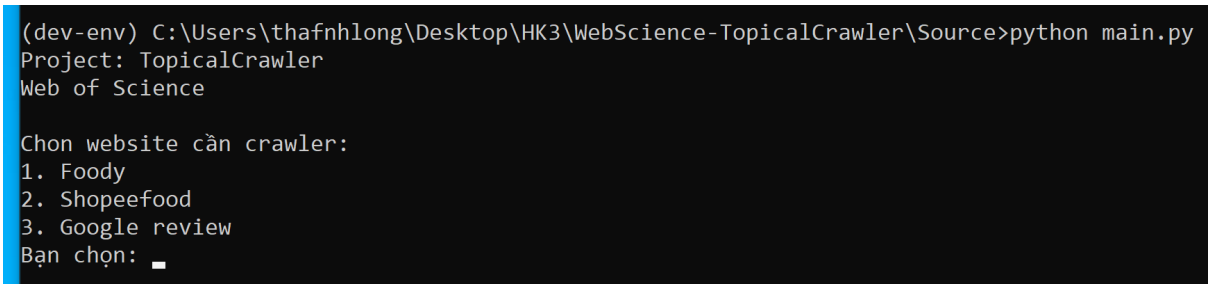
  - python_launcher: cài đặt theo môi trường (python, python3)
  - external_log: quyết định có mở console để theo dõi log hay không
- Chạy chương trình:

```
python main.py
```

+
+ Sử dụng chrome webdriver: https://sites.google.com/a/chromium.org/chromedriver/downloads
+ Tiến hành chạy:
    + Thực hiện mở command line và cd tới thư mục Source

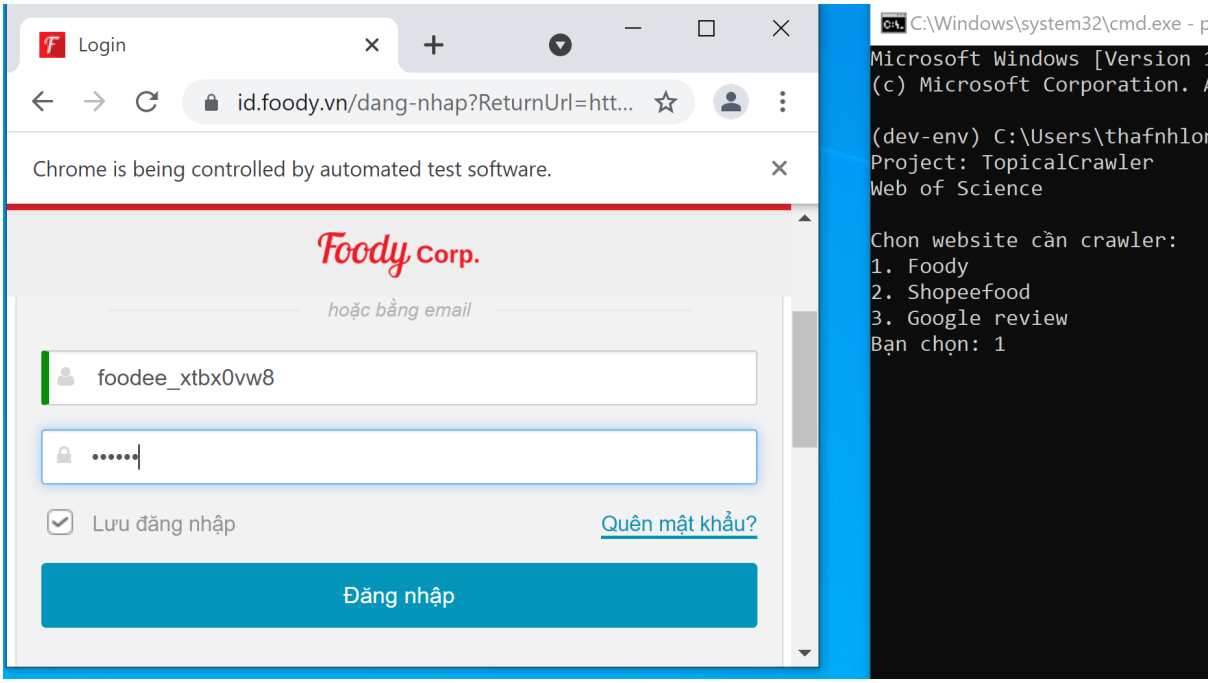


    + Đánh lệnh khởi chạy

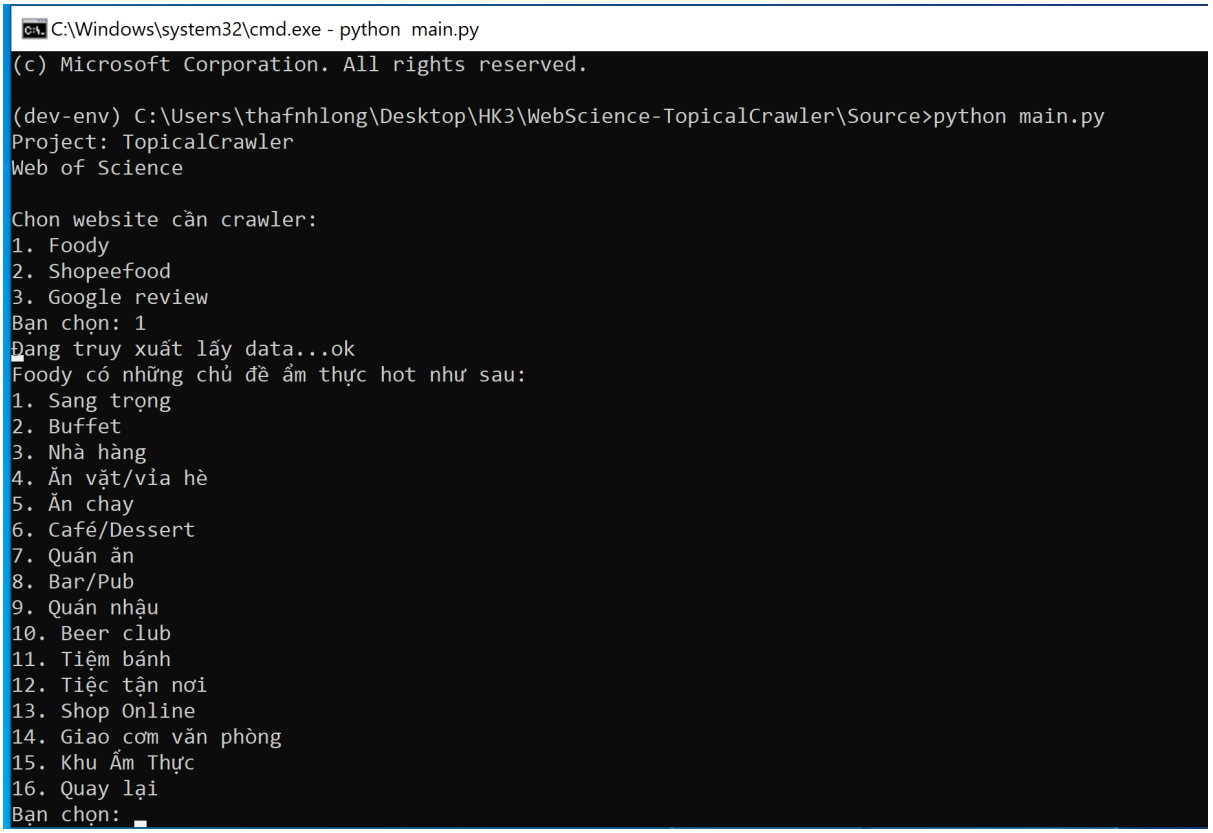


    + Lựa chọn 1 plugin

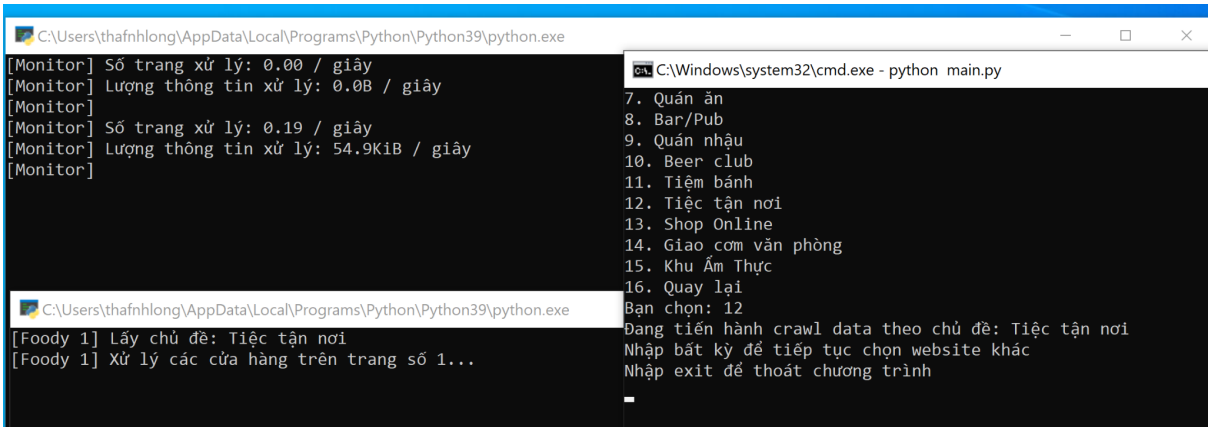
Ví dụ: Foody, thì chương trình sẽ hiện lên cửa sổ, chúng ta sẽ điền tài khoản và mật khẩu



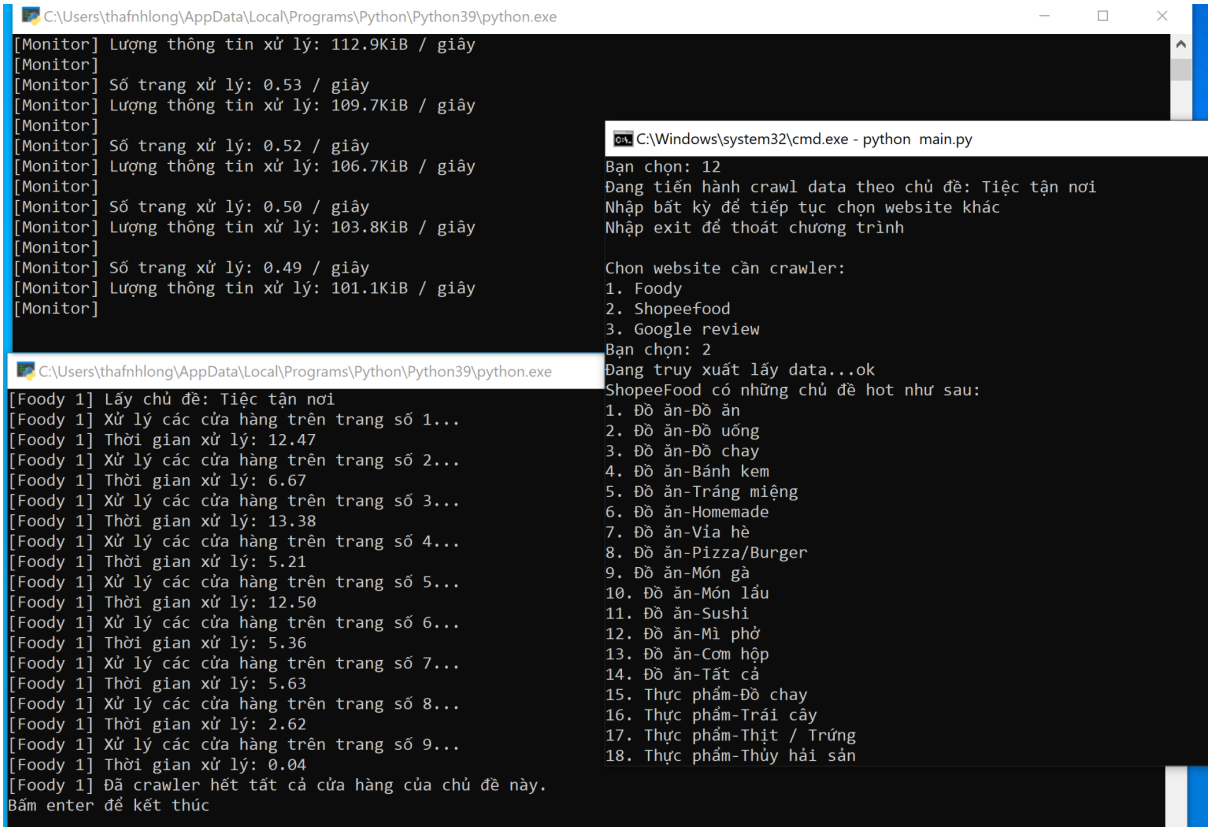
+ Sau khi bấm đăng nhập, chương trình sẽ tự kiểm tra xem đăng nhập thành công hay không và kết quả trả về :



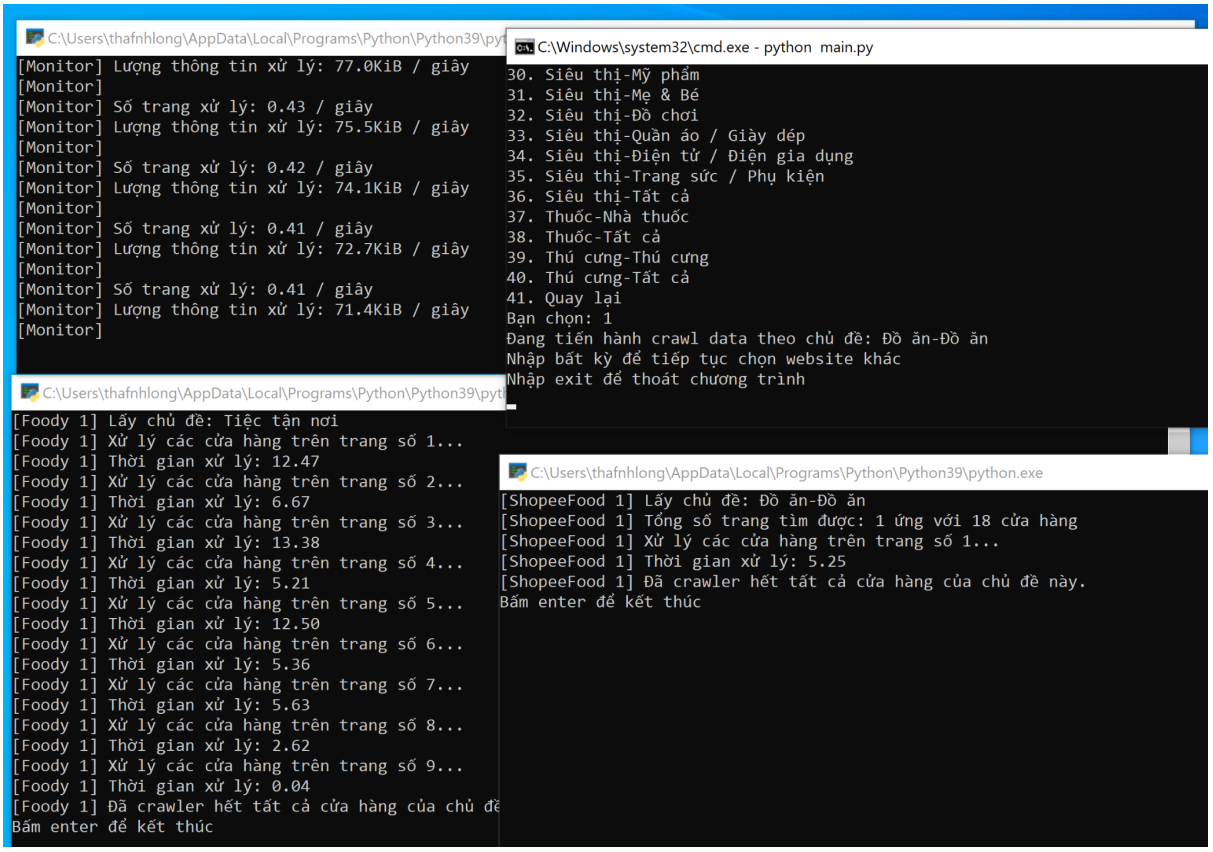
+ Lựa chọn 1 chủ đề lớn: ví dụ 12: tiệc tận nơi



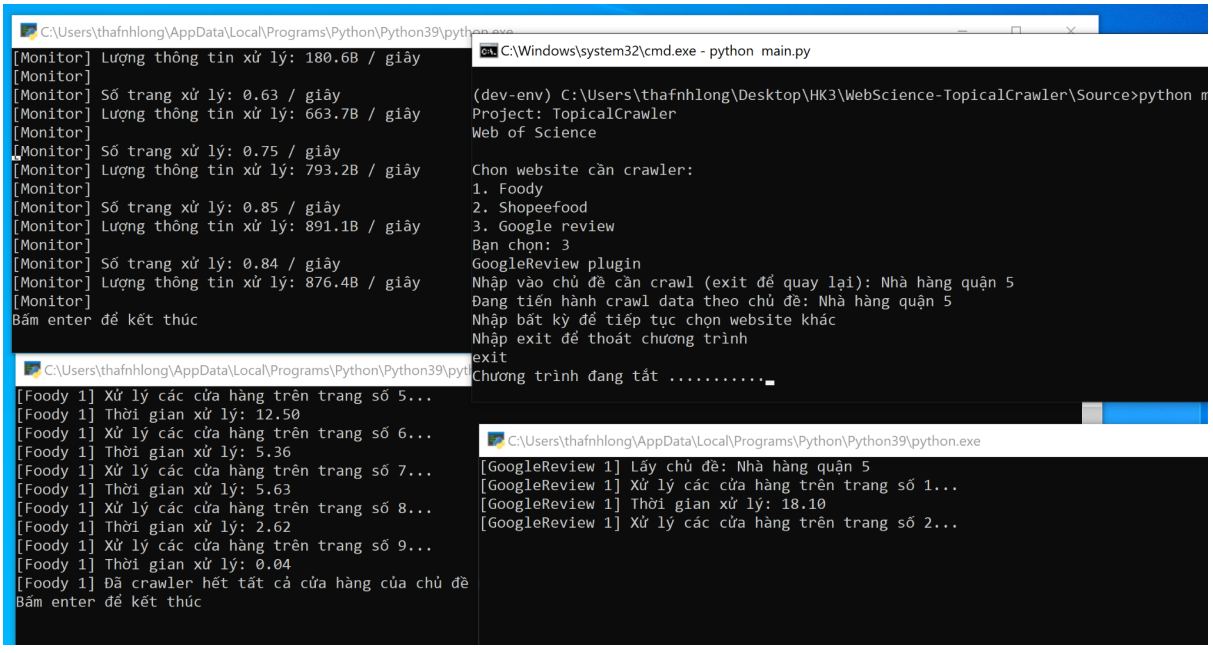
+ Sẽ có 2 cửa sổ hiện lên, 1 là monitor thống kê trên / giây; cửa sổ còn lại là thông tin của luồng xử lý crawl Foody (# Foody 1)
+ Trong lúc đó ta cũng có thể chọn lựa 1 chủ đề khác hoặc kể cả 1 website(plugin) khác
+ Bấm enter để tiếp, hoặc exit để thoát



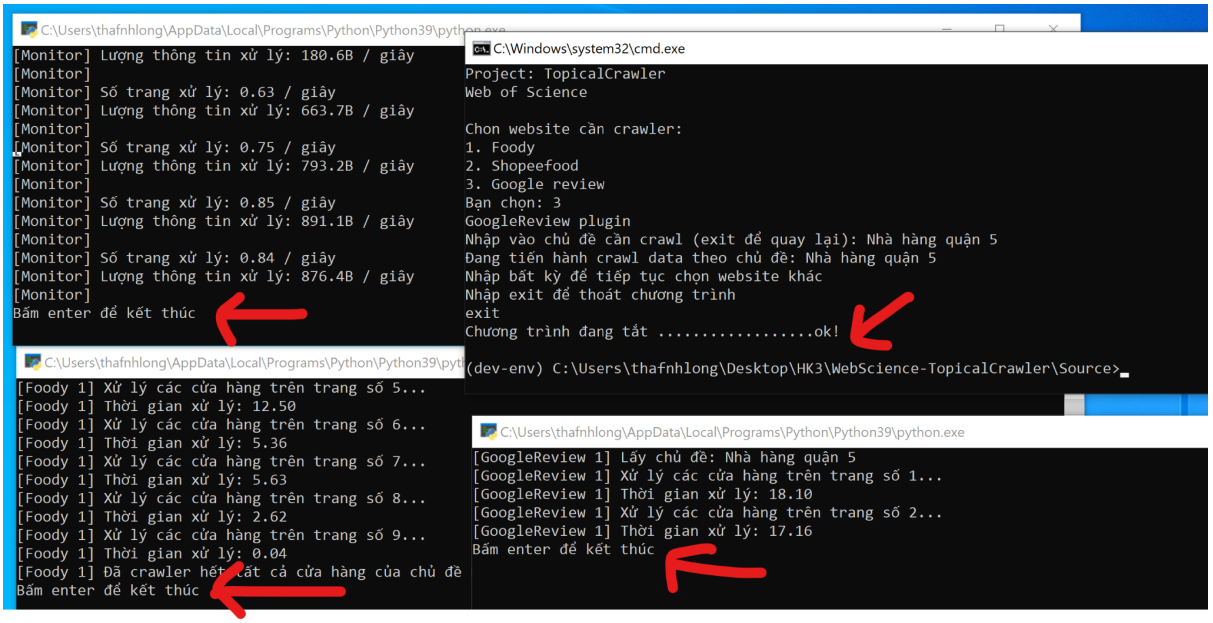
+ Giả sử lựa chọn tiếp Plugin ShopeeFood, chọn 1 chủ đề



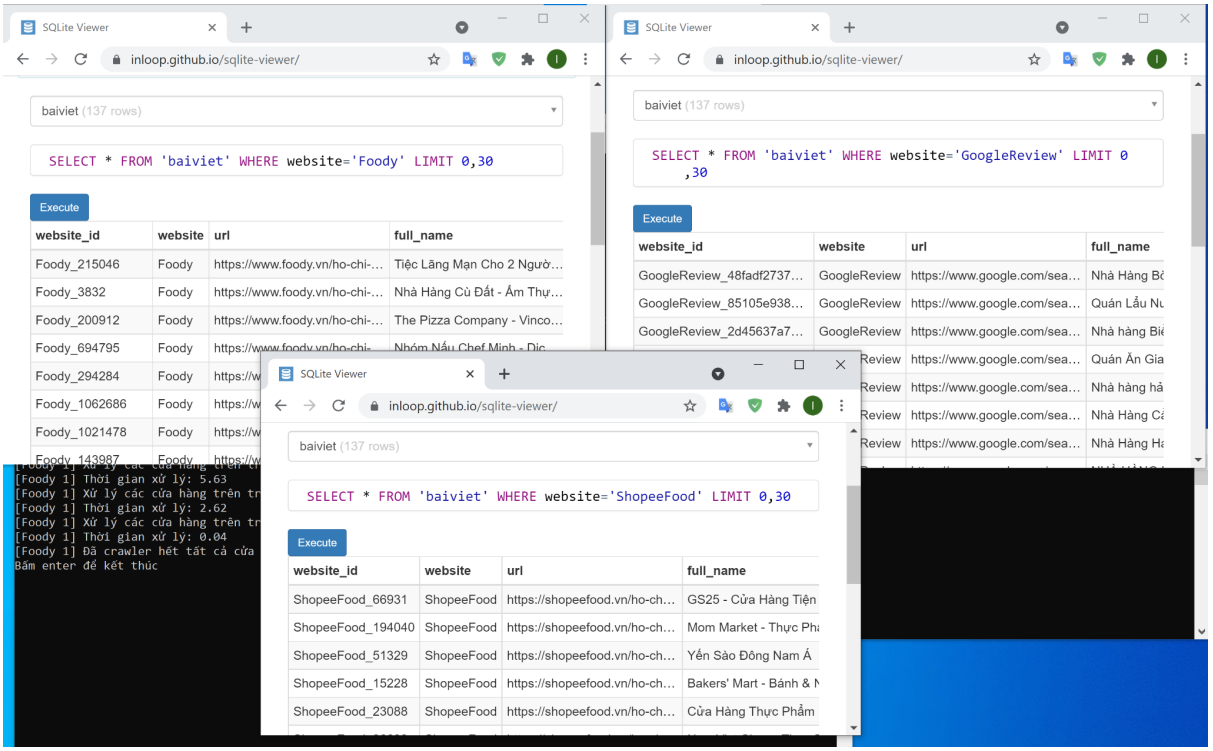
+ Chúng ta có thể ngưng crawl bằng cách nhập exit ở console quản lý



+ Luồng chính sẽ quản lý xóa bộ nhớ thừa và tự tắt khi hoàn thành



+ Kiểm tra data đã crawl



VII/ TÀI LIỆU THAM KHẢO

https://selenium-python.readthedocs.io/
https://docs.python-requests.org/projects/requests-html/en/latest/
https://stackoverflow.com/
https://codelearn.io
https://simple-deeplearing.medium.com/