

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH



ĐỒ ÁN KỸ THUẬT LẬP TRÌNH (CO1031)

Nhóm: BCH_DOAN — Task 2.1: 22/11 - 29/11

Xây dựng công cụ thu thập thông tin và phân loại tin tuyển dụng

GVHD: Lưu Quang Huân
SV thực hiện: Nguyễn Ngọc Thuần – MSSV: 1814217
Nguyễn Thế Viễn – MSSV: 1814764
Phạm Đức Duy Anh - MSSV: 1810814
Huỳnh Ngọc Tấn - MSSV: 1813592
Bùi Hữu Đang - MSSV: 1811828

Tp. Hồ Chí Minh, Tháng 11/2020



Mục lục

1	Các công việc đã thực hiện	2
2	Những khó khăn gặp phải	2
2.1	Cài đặt thư viện scrapy của python	2
2.2	File HTML bị lỗi dẫn đến việc tìm các tag của file khó khăn	3
2.3	Trang web yêu cầu đăng nhập để hiển thị thêm thông tin	3
3	Kết quả đạt được	4
3.1	Data sample của trang web https://vietsingworks.com/	4
3.2	Data sample của trang web https://vietsingworks.com/	4
3.3	Data sample của trang web https://vietsingworks.com/	5
3.4	Data sample của trang web https://chotot.vn/	5
4	Dự kiến công việc cho tuần tới	6

1 Các công việc đã thực hiện

Nhóm đã tiến hành chia việc crawl các trang web cho từng thành viên, cụ thể là:

- Nguyễn Thế Viễn:
- Nguyễn Ngọc Thuần:
- Phạm Đức Duy Anh: <https://www.chotot.com/>
- Bùi Hữu Đăng: <https://careerbuilder.vn/>
- Huỳnh Ngọc Tấn:

Nhóm đã tìm hiểu và sử dụng 2 thư viện là scrapy và beautifulsoup của python để tiến hành crawl dữ liệu, trong đó:

- Tìm hiểu về scrapy (Thuần và Viễn) đã push cách cài đặt và sử dụng lên github nhóm
- Tìm hiểu về beautifulsoup (Duy Anh, Tấn và Đăng) đã push cách cài đặt và sử dụng lên github lên nhóm

Nhóm tiến hành crawl dữ liệu, lưu thông tin crawl được vào file json, phân hiện thực và data sample cho từng trang web đã push lên github

2 Những khó khăn gặp phải

2.1 Cài đặt thư viện scrapy của python

Khi cài đặt thư viện scrapy của python bằng command pip install scrapy trên terminal, máy có thể bị lỗi như sau:

```
building 'twisted.test.raiser' extension
error: Microsoft Visual C++ 14.0 is required. Get it with "Microsoft Visual C++ Build Tools": https://visualstudio.m
icrosoft.com/downloads/
-----
ERROR: Command errored out with exit status 1: 'c:\program files (x86)\python38-32\python.exe' -u -c 'import sys, setup
ools, tokenize; sys.argv[0] = 'C:\Users\Codie Bishwas\AppData\Local\Temp\pip-install-4feur2ly\Twisted\setup.
py'; __file__ = 'C:\Users\Codie Bishwas\AppData\Local\Temp\pip-install-4feur2ly\Twisted\setup.py'; f=getattr(tokenize, 'open', open)(__file__); code=f.read().replace('\r\n', '\n'); f.close(); exec(compile(code, __file__, 'exec'))' install --record 'C:\Users\Codie Bishwas\AppData\Local\Temp\pip-record-3zkhefgk\ins
tall-record.txt' --single-version-externally-managed --compile --install-headers 'c:\program files (x86)\python38-32\Inc
lude\Twisted' Check the logs for full command output.
```

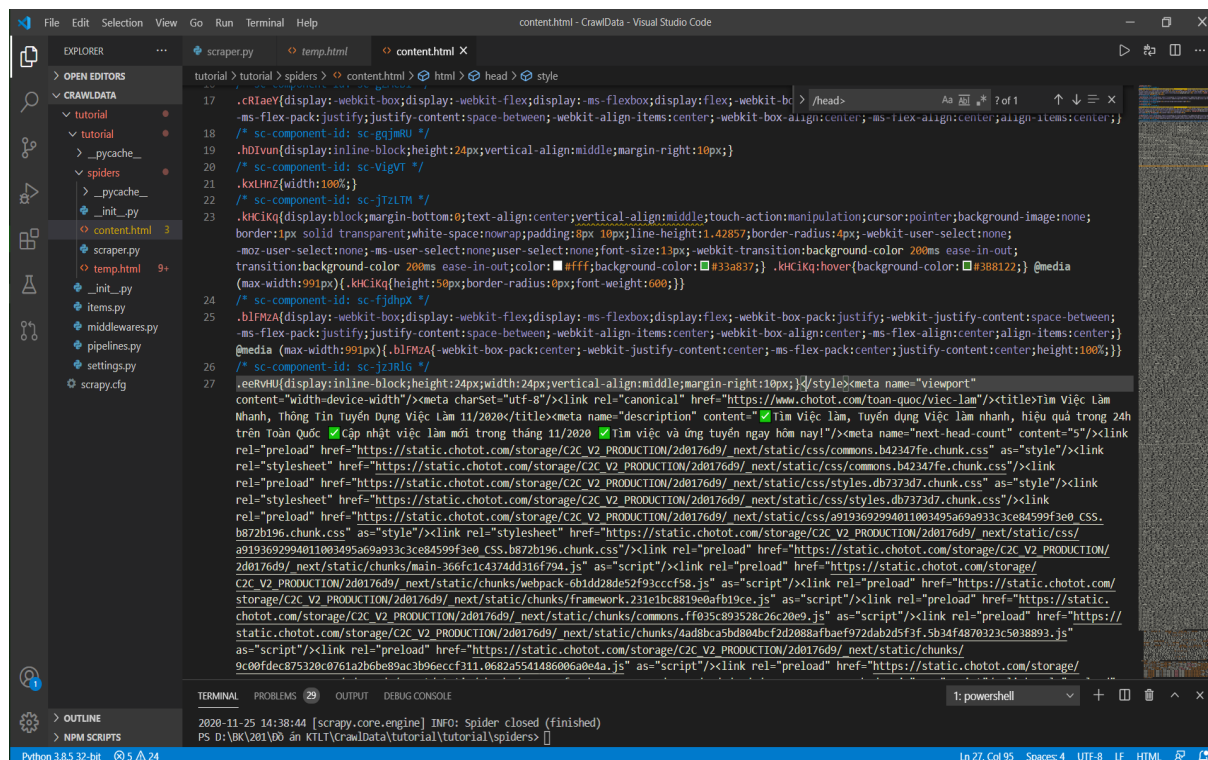
Hình 1: Lỗi khi cài đặt thư viện scrapy của python

Để khắc phục lỗi này, ta có 2 cách:

- Cách 1: Mở terminal, cài đặt Twisted bằng **command pip install Twisted**. Tuy nhiên trong vài trường hợp, cách này có thể bị lỗi.
- Cách 2: Tải Twisted wheel (file .whl) trên trang web <https://www.lfd.uci.edu/~gohlke/pythonlibs/#twisted>, chọn version tương thích với máy. Tiếp theo ta bật terminal, chuyển đến folder chứa file .whl đã download và cài đặt bằng command **pip install Twisted_file_name.whl**

2.2 File HTML bị lỗi dẫn đến việc tìm các tag của file khó khăn

Cụ thể ở trang web <https://chotot.vn/>, khi ta gọi request trả về nội dung của file HTML, ta thu được như hình:



Hình 2: Nội dung file HTML trả về khi gửi request đến trang web

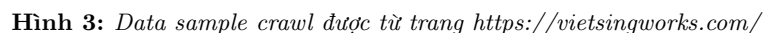
Để xử lý trường hợp này, ta đổi nội dung file sang dạng chuỗi và dùng cái method của python để lấy được các trường dữ liệu tương ứng.

2.3 Trang web yêu cầu đăng nhập để hiển thị thêm thông tin

Cụ thể ở trang web <https://vietnamworks.com/>, khi ta gọi request trả về nội dung của file HTML, web chỉ trả về một số lượng trường dữ liệu nhất định và số lượng link trả về cũng bị hạn chế

Hiện tại nhóm vẫn đang tìm hiểu về selenium, ... để tìm hiểu thêm cách giải quyết vấn đề này

3.1 Data sample của trang web <https://vietsingworks.com/>

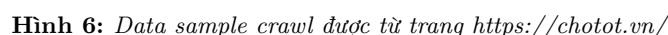


3.2 Data sample của trang web <https://vietsingworks.com/>



[illegible]

3.4 Data sample của trang web <https://chotot.vn/>





4 Dự kiến công việc cho tuần tới

- Triển khai việc crawl dữ liệu để có được dataset khoảng 10000 items
- Cải tiến công cụ crawl dữ liệu (tìm hiểu về selenium, ...) để tạo thành công cụ hoàn chỉnh cho việc crawl dữ liệu (dùng cho web không public API, yêu cầu đăng nhập, dùng javascript load động, ...)
- Tìm hiểu các làm sạch, chuẩn hóa và phân loại dữ liệu
- Thiết kế database, công cụ đẩy data lên database