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 Thuấn – MSSV: 1814217

Nguyễn Thế Viễn – MSSV: 1814764 Phạm Đức Duy Anh - MSSV: 1810814 Huỳnh Ngọc Tấn - MSSV: 1813592 Bùi Hữu Đang - MSSV: 1811828



Trường Đại Học Bách Khoa Tp.Hồ Chí Minh Khoa Khoa Học và Kỹ Thuật Máy Tính

Mục lục

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



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 Thuấn:
- Phạm Đức Duy Anh: https://www.chotot.com/
- Bùi Hữu Đang: https://careerbuilder.vn/
- Huỳnh Ngọc Tấn:

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

- Tìm hiểu về scrapy (Thuấn và Viễn) đã push cách cài đặt và sử dụng lên github nhóm
- Tìm hiểu về beautiful
soup (Duy Anh, Tấn và Đang) đã pu
sh cách cài đặt và sử dụng lên github lên nhóm

Nhóm tiến hành crawl dữ liệu, lưu thông tin crawl được vào file json, phần 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 băng command pip install scrapy trên terminal, máy có thể bị lỗi như sau:

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

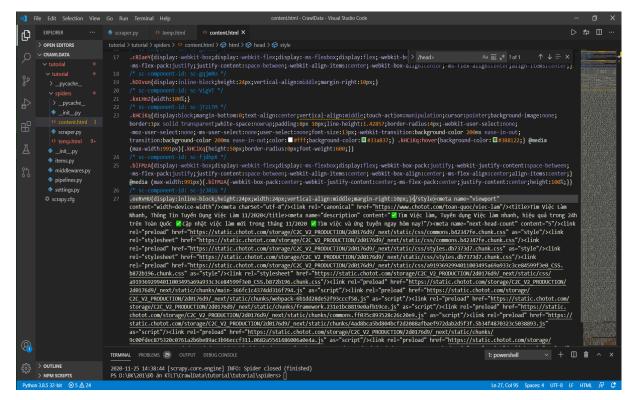
Để khắc phục lỗi này, ta có 2 cách:

- 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 lí trường hợp này, ta đổi nội dung file sang dạng chuỗi và dùng cái method của python để lấy được các trường dữ liêu tương ứng.

2.3 Trang web yêu cầu đăng nhập để hiển thi 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 han chế

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



3 Kết quả đạt được

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

```
21 lines (21 sloc) 20.3 KB
  2 {"job_title": " OFFICE ADMINISTRATOR ($2,200- $2,400)", "salary": " SGD $2,200- $2,400", "location": " Singapore", "position": " ", "job_type": " Permanent Staff", "age": " 21
      {"job title": " Waitress", "salary": " Discuss when interview", "location": " Singapore", "position": " Worker", "job type": " Permanent Staff", "age": " 21 - 30", "quantity"
     {"job_title": "Công nhân sản xuất tại công ty máy dập (SGD $2600)", "salary": "SGD $2600", "location": "Singapore", "position": "Worker", "job_type": "Permanent Staff",
  5 {"job_title": " FRONT DESK OFFICER (FOREIGN BANK/$2500)", "salary": " SGD $2500", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age":
      {"job_title": " JUNIOR SECRETARY (SGD $2200)", "salary": " SGD $2200 -$2300", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 2
     {"job_title": " RECEPTIONIST (Hotel / Front Line / $2350)", "salary": " $2350", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 2
  8 {"job_title": " Automotive technician ($2600 )", "salary": " SGD $2600", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 22-35",
      {"job_title": " BARTENDER (RESTAURANT/ $2200 + tip)", "salary": " $2200", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 24-30",
 10 {"job_title": " Marine Engine Specialist ($2200 - $3000)", "salary": " $2200 - $3000", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "ag
 11 {"job_title": " Waiter / Waitres ( $2200 SGD)", "salary": " ( $2200 SGD)", "location": " Singapore", "position": " Worker", "job_type": " Permanent Staff", "age": " 21-28",
      {"job title": "CHEF (SEAFOOD/VEGETARIAN) - SGD $4000", "salary": "SGD $4000", "location": "Singapore", "position": "Supervisor/Manager", "job type": "Permanent Staff",
 13 {"job_title": " BUTLER - HOTEL SERVICES ( $2200)", "salary": " $2200", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 24-30", "q
 14 {"job_title": "Restaurant Assistant Manager (Restaurant / $2400 - $2600)", "salary": "SGD $2400- $4000", "location": "Singapore", "position": "Supervisor/Manager", "job_ty
      {"job_title": " EXPRESS SERVICE - CALL CENTRE HOTEL SERVICES ($2200)", "salary": " $2200", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff",
 16 {"job_title": "IT Development Specialist ( SGD $4500 - $6000 )", "salary": "SGD $4500 - $6000", "location": "Singapore", "position": "Specialist", "job_type": "Permanent
 17 {"job_title": " CUSTOMER SERVICE OFFICER ($2300 - $2400)", "salary": " SGD $2300- $2400", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff",
      {"job_title": " Saleman (SGD $2300 - $2600)", "salary": " SGD $2300 - $2600", "location": " Singapore", "position": " Specialist", "job_type": " Permanent Staff", "age": " 22
 19 {"job title": " Glass Manufacturing industry- Production Supervisor ", "salary": " SGD $2600 ", "location": " Singapore", "position": " Supervisor/Manager", "job type": " Perm
 20 {"job_title": " FINANCE ANALYST (SGD $2,800)", "salary": " SGD $2,800", "location": " Singapore", "position": " ", "job_type": " Permanent Staff", "age": " 22-30", "quantity"
```

Hình 3: Data sample crawl được từ trang https://vietsingworks.com/

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

```
1202 lines (1202 sloc) 258 KB
                                                                                                                                                         Blame 🗆 🛭 🗓
          { "Position": "Chuyên Viên Tư Vấn Bán Hàng Online", "Company": "Công Ty TNHH Thương Mại Dịch Vụ Điện Từ Hà Nguyễn", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Min
          { "Position": "Nhân Viên Kinh Doanh Tại Hải Phòng", "Company": "Cty CP Thủy Sản Anh Minh", "Salary": "10 triệu - 12 triệu", "Location": "Hải Phòng" },
          { "Position": "Nhân Viên Sale Nội Thất", "Company": "Công Ty TNHH Xuất Nhập Khẩu Được Phẩm Đức Minh", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh" },
         { "Position": "Nhân Viên Sale Admin (hỗ Trợ Kinh Doanh)", "Company": "Công Ty TNHH Ngô Thanh Thanh", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh" },
          { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Môi Trường Sen Vàng", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh, Bình Dương" },
          { "Position": "Nhân Viên Tư Vấn Visa Lương Trên 15 Triệu", "Company": "Công Ty TNHH Thương Mai Dịch Vụ Quốc Tế Việt Thế Giới", "Salary": "15 triệu - 20 triệu", "Location":
          { "Position": "Nhân Viên Kinh Doanh", "Company": "Silver Kite Apartment", "Salary": "7 triệu - 10 triệu", "Location": "Hà Nội" },
         { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Orai Việt Nam", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh" },
          { "Position": "Nhân Viên Phát Triển Thị Trường (tp Hà Nội)", "Company": "Công Ty Cổ Phần Waytech Việt Nam", "Salary": "6 triệu - 30 triệu", "Location": "Hà Nội" },
         { "Position": "Nhân Viên Hành Chính Văn Phòng (quận Bình Tân)", "Company": "Công Ty Cổ Phần Đầu Tư Và Phát Triển Nam Hưng", "Salary": "7 triệu - 10 triệu", "Location": "Hồ
         { "Position": "Nhân Viên Bán Hàng Thời Trang - Quận 9 Hcm", "Company": "Công Ty TNHH May Vĩnh Phú", "Salary": "5 triệu - 7 triệu", "Location": "Hồ Chí Minh" },
          { "Position": "Bào Hiểm Manulife Tìm Đối Tác Khởi Nghiệp Part Time / Full Time Tai Hà Nội (dự án Khởi Nghiệp Và Phát Triển Lãnh Đạo Fta Kế Cân) Hà Nội", "Company": "Smile
         { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Bảo An Sinh", "Salary": "7 triệu - 10 triệu", "Location": "Bình Định, Đà Nẵng, Quảng Bình" },
         { "Position": "Nhân Viên Sale Online (màng Nha Khoa Thấm Mỹ)", "Company": "Công Ty TNHH Thiên Phúc - Nha Khoa Dora", "Salary": "8 triệu - 10 triệu", "Location": "Hồ Chí Mi
         { "Position": "Nhân Viên Marketing Facebook", "Company": "Công Ty Cổ Phần Thương Mại Và Dịch Vụ Asiatech Việt Nam", "Salary": "6 triệu - 9 triệu", "Location": "Hà Nội" },
          { "Position": "Nhân Viên Tư Văn Chăm Sóc Khách Hàng", "Company": "Công Ty TNHH Trang Thiết Bị ô Tô Việt Nam", "Salary": "8 triệu - 12 triệu", "Location": "Hồ Chí Minh" },
         { "Position": "Chuyên Viên Tư Vấn (thu Nhập Trên 20tr / Tháng)", "Company": "Infinox", "Salary": "20 triệu - 30 triệu", "Location": "Hà Nội, Hồ Chí Minh" },
```

Hình 4: Data sample crawl được từ trang https://mywork.com/



3.3 Data sample của trang web https://vietsingworks.com/

```
{
    "jobs": [
   "ID": "1",
   "Tlu00ean c'u00f4ng ty": "Cong ty TNHH Brasol",
"\u0110\u1ecba \u0111i\u1ec3a": "S\u1ed1 6, Phan Chu Trinh, P. T\u00e2n Th\u00e0nh, Q. T\u00e2n Ph\u00fa, TP.HCM",
"V\u1ecb tr\u00ed tuy\u1ec3n d\u1ee5ng": "Senior Digital Marketing",
    "Ng\u00e0y c\u1eadp nh\u1eadt": "27/11/2020",
    "Ng\u00e0nh ngh\u1ec1": "Ti\u1ebfp th\u1ecb / Marketing\r\n
                                                                                                                \r\n\t\t\t
                                                                                                                                                                                                                            Ti\u1ebfp th\u1ecb
                                                                    \r\h\t\t\t
tr\u1ef1c tuy\u1ebfn\r\n
                                                                                                                         \r\n
                                                                                                                                                                                Qu\u1ea3ng c\u00e1o / \u0110\u1ed1i ngo\u1ea1i /
Truy\u1ec1n Th\u00f4ng",

"H\u00ecnh th\u1ee9c": "Nh\u00e2n vi\u00ean ch\u00ednh th\u1ee9c",
   "L\u01b0\u01alng": "15 Tr - 20 Tr VND",
"Kinh nghi\u1e7m": "3 - N\u0193m",
"C\u1ea5p b\u1eadc": "Tr\u01b0\u1edfng nh\u00f3m / Gi\u00e1m s\u00e1t",
    "H\u1ebft h\u1ea1n n\u1ed9p": "30/12/2020",
    "Ph\u00fac 1\u1ee3i": [
      'Ch\u1ebf \u0111\u1ed9 b\u1ea3o hi\u1ec3m".
    " Ch\u1ebf \u0111\u1ed9 th\u01b0\u1edfng",
    " T\u0103ng l\u01b0\u01a1ng",
    " Ngh\u1ec9 ph\u00e9p n\u0103m"
   "M\u00f4 t\u1ea3 C\u00f4ng vi\u1ec7c": [
     "- Ph\ules Tr\u00elch v\u12eadn h\u00e0nh, l\u00ean k\u1ebf ho\u1ealch v\u00e0 tri\u10eat nkhai c\u00e1c ch\u01b0\u01alng tr\u00ecnb Digital Marketing (SEO / SEM, Online Advertising, Email Marketing, Mobile
Marketing, Social Media, SMS Marketing, Affiliate Marketing\u2026) c\u1ea7a (\u009fang ty v\u00e0 kh\u00e1ch h\u00e0lng \u0111\u1ea3 \u0111\u1ea3 \u0111\u1ea1t \u0111\u01e3c m\u1ea6 \u011tue1e3 m\u1ea7 \u011tue1e5 ti\u00e0 uth\u01b0\u01a1ng hi\u1ea7
\u000e0 kinh doanh. H\u1ed7 tr\u1ee3 t\u01b0 v\u1ea5n v\u1ec1 Digital Marketing cho kh\u00e1ch h\u00e1ch h\u00e0ng.- \u00e110|u1ec1 xu\u1ea5t v\u00e0 th\u1ef1c hi\u1ec7n c\u00e0t ch\u01b0\u01a1ng tr\u00ecnh khuy\u1ebfn
m\u1eaii, truy\u1ec1n th\u00f4ng, qu\u1ea3ng b\u00e1 v\u00e0 x\u00e1 v\u00e0 x\u00e1 d\u10e1ing th\u01e1ing th\u01e1ing hi\u1ec7u cho c\u00f4ng ty v\u00e1 kh\u00e1h h\u00e1ne0e1 hh\u00e1ng. Thu th\u1eadp, ph\u00e1 v\u00e1 th\u00e1ne0en th\u00e1 th\u00e1 b
th/wlecb tr/w01b0|wleddng, s\w1ea3n ph\w1ea9m v\w000e0 c\w000e1c ch\w01b0|w01a1ng tr/w00ecnh Digital Marketing, Social Media c\w1ee7a c\w00e1c \w0111\w1ed1i th\w1ee7 c\w1ea1nh tranh. \w0110\w1ea1 ranh. \w0110\w1ea5 t
c\u00e1c gi\u1ea3i ph\u00e1p li\u00ean quan.- Theo d\u00f5i, \u0111\u00e1n gi\u00e1 hi\u1ec7u qu\u1ea3 ch\u01b0\u01ang tr\u00ecnh \u011ang di\u1ec5n ra, l\u1eadp b\u00e1 b\u00e1b c\u00e1b \u00e1
S\uleed d\uleesng t\uledit c\u00e1c c\u00e14ng c\u10e5 Digital Marketing, ch\ulecdn l\u1efla lo\u1efla lo\u1eali h\u00ernh n\u00ean s\u1eed d\u1eesng \u0111\u1ec3 ti\u1ebf c\u1ean v\u1edbi \u0111\u00efang \u0111\u1edi
```

Hình 5: Data sample crawl được từ trang https://careerbuilder.vn/

t\u01b0\u1ea3ng kh\u00e1ch h\u00e0ng v\u00e0 \u0111\u1ea1t m\u1ea5c ti\u00eau th\u01b0\u01a1ng hi\u1ec7u & kinh doanh.- Ph\u1ed1i h\u1ea3p v\u1ed1i \u01b1\u1ea3p v\u1ed1i \u01b1\u1ea7u di \u00e7m x\u00e7m x\u00

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

```
14606 lines (14606 sloc) 599 KB
                                                                                                                                                        Raw Blame 🖫 🖉 🗓
        "iobs": [
          "Tiêu đề": "Bao Ăn Ở, Bảo Vệ Nhà Máy Ở KCN Đại Đăng,Bình Dương",
          "Hình thức trả lương": "Theo tháng",
          "Loại công việc": "Toàn thời gian"
          "Ngành nghề": "Bảo vê/An ninh/Vê sỹ",
          "Kinh nghiệm": "< 1 năm",
          "Tên công ty": "Công ty Cổ Phần Bảo vệ Trung Nam",
         "Số lượng tuyển dụng": "6",
"Các quyền lợi khác": "Lễ Tết X2,3",
          "Học vấn tối thiểu": "Cấp 2",
          "Chứng chỉ / kỹ năng": "Không yêu cầu",
          "Tuổi tối thiểu": "18",
          "Tuổi tối đa": "60"
          "Địa chỉ": "Xã Vĩnh Tân, Thị xã Tân Uyên, Bình Dương"
          "Tiêu đề": "Tuyển 1000 Nam Nữ Thời Vụ- Lắp Ráp Điện Tử",
          "Hình thức trả lương": "Theo tháng",
          "Loại công việc": "Thời vụ",
          "Ngành nghề": "Thơ điện/Điện tử/Điện lanh".
```

Hình 6: Data sample crawl được từ trang 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 chính cho việc crawl dữ liệu (dùng cho web không public API, yêu cầu đăng nhập, dùng javascript load động, ...)
- $\bullet\,$ Tìm hiểu các làm sạch, chuẩn hóa và phân loại dữ liệu
- Thiết kế database, công cụ đẩy data lên database