

DÀI 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

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

Lớp:	L01
GVHD:	Lưu Quan Huân
SV thực hiện:	Nguyễn Thế Viễn 1814764
	Nguyễn Ngọc Thuấn 1814217
	Phạm Đức Duy Anh 1810814
	Huỳnh Ngọc Tấn 1813952
	Bùi Hữu Dang 1811828



## Mục lục

<b>1 Lời mở đầu</b>	<b>3</b>
<b>2 Phân tích yêu cầu đề tài</b>	<b>4</b>
2.1 Phân tích yêu cầu đề tài . . . . .	4
2.2 Các bước thực hiện yêu cầu đề tài . . . . .	4
<b>3 Tìm hiểu về các công cụ để thực hiện đề tài</b>	<b>5</b>
3.1 Ngôn ngữ lập trình sử dụng cho việc crawl dữ liệu . . . . .	5
3.2 Tìm hiểu các thư viện của python phục vụ cho việc crawl dữ liệu . . . . .	5
3.2.1 Scrapy . . . . .	5
3.2.2 BeautifulSoup . . . . .	6
3.2.3 Selenium . . . . .	7
3.3 Tìm hiểu các công cụ phục vụ cho việc lưu trữ dữ liệu . . . . .	8
3.3.1 Định dạng dữ liệu . . . . .	8
3.3.2 JSON . . . . .	10
3.3.3 Ưu điểm của JSON . . . . .	10
3.4 Tìm hiểu các công cụ quản trị Hệ cơ sở dữ liệu . . . . .	11
3.4.1 Cách kết nối Mysql sử dụng python . . . . .	11
3.5 Tìm hiểu các thư viện của python phục vụ cho việc phân tích, thống kê dữ liệu . . . . .	12
3.5.1 Pandas . . . . .	12
3.5.2 Matplotlib . . . . .	12
3.5.3 NumPy . . . . .	13
<b>4 Hiện thực việc crawl dữ liệu từ các trang web</b>	<b>14</b>
4.1 Crawl dữ liệu từ trang http://vietnamworks.com/ . . . . .	14
4.2 Crawl dữ liệu từ trang http://vietsingworks.com/ . . . . .	15
4.3 Crawl dữ liệu từ trang http://careerbuilder.vn/ . . . . .	16
4.4 Crawl dữ liệu từ trang http://mywork.com.vn/ . . . . .	17
4.5 Crawl dữ liệu từ trang http://1001vieclam.com/itviec/ . . . . .	17
4.6 Crawl dữ liệu từ trang http://chotot.vn/ . . . . .	18
4.7 Tổng kết bước tiến hành crawl dữ liệu . . . . .	20
<b>5 Hiện thực làm sạch dữ liệu</b>	<b>21</b>
5.1 Tiến hành làm sạch dữ liệu ở mức unit (cá nhân) . . . . .	21
5.2 Tiến hành làm sạch dữ liệu ở mức program (chương trình) . . . . .	21
5.3 Tiến hành lọc dữ liệu bị thiếu trường dữ liệu - Missing data . . . . .	22
5.4 Tiến hành lọc dữ liệu trùng lặp . . . . .	23
5.5 Tổng kết bước làm sạch dữ liệu thu được . . . . .	25



---

<b>6 Mô hình hóa dữ liệu và xây dựng Hệ cơ sở dữ liệu để lưu trữ</b>	<b>26</b>
6.1 Mô hình hóa dữ liệu . . . . .	26
6.1.1 Sơ đồ quan hệ thực thể (ER Diagram) . . . . .	26
6.1.2 Mô tả . . . . .	27
6.1.3 Ánh xạ lược đồ liên kết thực thể sang lược đồ dữ liệu quan hệ - Mapping . . . . .	27
6.2 Xây dựng hệ cơ sở dữ liệu để lưu trữ . . . . .	27
6.3 Tổng kết bước mô hình hóa dữ liệu và xây dựng Hệ cơ sở dữ liệu . . . . .	30
<b>7 Hiện thực phân tích dữ liệu thu thập được từ các trang</b>	<b>31</b>
7.1 Phân tích thông tin tuyển dụng theo giá trị tiền lương . . . . .	31
7.2 Phân tích thông tin tuyển dụng theo vị trí địa lý . . . . .	32
7.3 Phân tích giá trị tiền lương trung bình theo vị trí địa lý . . . . .	33
7.4 Tổng kết bước phân tích dữ liệu . . . . .	34
<b>8 Đề xuất ứng dụng</b>	<b>35</b>
8.1 Ý tưởng ứng dụng . . . . .	35
8.2 Công cụ hiện thực ứng dụng . . . . .	35
8.3 Web tuyển dụng . . . . .	37
8.3.1 Trang chủ . . . . .	37
8.3.2 Thông tin công việc . . . . .	40
8.3.3 Chi tiết công việc . . . . .	43
8.3.4 Trang đăng tin tuyển dụng . . . . .	44
<b>9 Tổng kết về dự án</b>	<b>47</b>
9.1 Các kết quả đã đạt được . . . . .	47
9.2 Những điểm hạn chế . . . . .	47
<b>10 Link Project</b>	<b>47</b>
<b>Tài liệu</b>	<b>48</b>



## 1 Lời mở đầu

- Trong thời đại công nghiệp hóa hiện đại hóa ngày nay, đặc biệt là trong kỉ nguyên công nghiệp 4.0, các tiến bộ về khoa học kỹ thuật ngày càng được áp dụng vào đời sống để tăng năng suất cũng như đem lại giá trị tinh thần cho con người. Nổi lên đó, trí tuệ nhân tạo (còn được biết đến với tên viết tắt là AI – Artificial Intelligence) đã có những bước phát triển đặc biệt mạnh mẽ và mang lại nhiều lợi ích cho cuộc sống.
- Những đặc điểm cơ bản của trí tuệ nhân tạo AI mà ta có thể kể đến như:
  - Ứng dụng các hệ thống học máy để giúp mô phỏng trí tuệ của con người trong những xử lý mà con người sẽ làm tốt hơn cả máy tính.
  - Trí tuệ nhân tạo giúp cho máy tính có được những trí tuệ của con người như khả năng biết suy nghĩ, biết lập luận để có thể giải quyết được vấn đề, biết giao tiếp do hiểu ngôn ngữ, tiếng nói, biết học và tự thích nghi...
  - Trí thông minh nhân tạo là một trong những ngành trọng yếu của tin học, liên quan đến cách cung cấp, sự học hỏi và khả năng thích ứng thông minh của máy móc mà chúng ta không ngờ.
- Để xây dựng được một hệ thống trí tuệ thông minh nhân tạo, ta cần cung cấp cho nó một tập dữ liệu tập huấn đủ lớn và chất lượng để có thể đạt được kết quả chính xác nhất. Xuất phát từ nhu cầu thực tiễn đó, trong dự án của môn Đồ án Kỹ thuật lập trình này, nhóm chúng em sẽ tiến hành thu thập dữ liệu từ các trang web, tiến hành làm sạch dữ liệu, xây dựng hệ cơ sở dữ liệu cũng như phân tích, đề xuất một ứng dụng với tập dữ liệu đã thu được.



## 2 Phân tích yêu cầu đề tài

### 2.1 Phân tích yêu cầu đề tài

- Trong dự án của môn Đồ án Kỹ thuật lập trình lần này, nhóm chúng em được yêu cầu xây dựng công cụ và thu thập thông tin và phân loại các tin tuyển dụng từ các trang web như:

- <http://vietnamworks.com/>
- <http://vietsingworks.com/>
- <http://careerbuilder.vn/>
- <http://mywork.com.vn/>
- <http://1001vieclam.com/itviec/>
- <http://chotot.vn/>

Từ các dữ liệu đã thu thập được từ các trang web trên, nhóm phải tiến hành quá trình làm sạch dữ liệu, tiến hành mô hình hóa chúng và phân tích, cũng như đề xuất được một ứng dụng sử dụng được tập dữ liệu đã thu thập và làm sạch

### 2.2 Các bước thực hiện yêu cầu đề tài

- Bước 1: Tìm hiểu các công cụ để crawl dữ liệu từ các trang web cũng như các công cụ để lưu trữ, tổ chức và phân tích dữ liệu đã thu được.
- Bước 2: Tiến hành làm sạch dữ liệu ở mức cá nhân, làm sạch dữ liệu theo một định dạng chung, lọc dữ liệu bị trùng lặp (nếu có).
- Bước 3: Xây dựng hệ cơ sở dữ liệu biểu diễn tập dữ liệu đã thu được.
- Bước 4: Phân tích tập dữ liệu đã thu được.
- Bước 5: Đề xuất một ứng dụng để sử dụng tập dữ liệu đã thu được.



### 3 Tìm hiểu về các công cụ để thực hiện đề tài

#### 3.1 Ngôn ngữ lập trình sử dụng cho việc crawl dữ liệu

- Dự án của môn Đồ án Kỹ thuật lập trình lần này yêu cầu chúng em phải sử dụng một ngôn ngữ lập trình hỗ trợ việc xử lý dữ liệu trên các trang web. Bên cạnh đó, chúng em cũng phải ứng dụng việc tính toán, thống kê, phân tích trên tập dữ liệu thu thập được. Với các yêu cầu đã được phân tích như trên, các ngôn ngữ có thể đáp ứng với nhu cầu này có thể kể đến như: R, Matlab và đặc biệt là Python.
- Python là ngôn ngữ lập trình hướng tới đối tượng bậc cao, dùng để phát triển nhiều ứng dụng khác nhau cũng như được ứng dụng rộng rãi trong ngành Khoa học công nghệ. Python dễ dàng để tìm hiểu, sử dụng cũng như tương thích với các nền tảng khác.
- Bên cạnh đó, Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng.
- Đặc biệt, Python còn cung cấp các thư viện để sử dụng các công cụ được viết sẵn, thông qua các package manager như PyPI (pip), pipenv, poetry, pyflow, conda,... Do đó việc xử lý và thống kê, phân tích dữ liệu cũng trở nên đơn giản và tiện lợi hơn. Với các lý do trên, nhóm chúng em quyết định sử dụng Python làm ngôn ngữ chính trong dự án lần này.

#### 3.2 Tìm hiểu các thư viện của python phục vụ cho việc crawl dữ liệu

Để thuận lợi cho việc crawl dữ liệu từ các trang web, nhóm chúng em sử dụng các thư viện được cung cấp bởi Python như Scrapy, BeautifulSoup và Selenium.

##### 3.2.1 Scrapy

- Scrapy là một khung ứng dụng để thu thập dữ liệu các trang web và trích xuất dữ liệu có cấu trúc có thể được sử dụng cho nhiều ứng dụng hữu ích, như khai thác dữ liệu để giám sát, kiểm tra tự động, xử lý thông tin hoặc lưu trữ lịch sử.
- Scrapy cung cấp rất nhiều tính năng mạnh mẽ để việc crawl trở nên dễ dàng và hiệu quả, chẳng hạn như:



- Hỗ trợ tích hợp để chọn và trích xuất dữ liệu từ các nguồn HTML / XML bằng cách sử dụng bộ chọn CSS mở rộng và biểu thức XPath, với các phương thức trợ giúp để trích xuất bằng cách sử dụng biểu thức chính quy.
  - Một bảng điều khiển trình bao tương tác (IPython nhận thức) để thử các biểu thức CSS và XPath để quét dữ liệu, rất hữu ích khi viết hoặc gỡ lỗi trình thu thập thông tin.
- Hướng dẫn cài đặt Scrapy:  
Cài đặt các gói của python, sau đó gõ lệnh trên terminal:  

```
1 pip install Scrapy
```

### 3.2.2 BeautifulSoup

- BeautifulSoup là một thư viện Python để phân tích dữ liệu có cấu trúc. Cho phép tương tác với HTML theo cách tương tự như cách tương tác với một trang web bằng các công cụ dành cho nhà phát triển. BeautifulSoup cho thấy một số chức năng trực quan mà có thể sử dụng để khám phá HTML mà mình nhận được.
  - BeautifulSoup cung cấp một số phương pháp đơn giản và thành ngữ Pythonic để điều hướng, tìm kiếm và sửa đổi cây phân tích cú pháp như một bộ công cụ để phân tích tài liệu và trích xuất những gì bạn cần. Không cần nhiều mã để viết một ứng dụng.
  - BeautifulSoup tự động chuyển đổi tài liệu đến sang Unicode và tài liệu đi sang UTF-8. Bạn không phải suy nghĩ về các mã hóa, trừ khi tài liệu không chỉ định một mã hóa và BeautifulSoup không thể tự động phát hiện một mã hóa. Sau đó, bạn chỉ cần chỉ định mã hóa ban đầu.
  - BeautifulSoup nằm trên các trình phân tích cú pháp Python phổ biến như lxml và html5lib, cho phép bạn thử các chiến lược phân tích cú pháp khác nhau hoặc tốc độ giao dịch để linh hoạt.
- Hướng dẫn cài đặt BeautifulSoup:  

```
1 pip install requests
2 pip install html5lib
3 pip install bs4
```



### 3.2.3 Selenium

- Selenium là một dự án cho một loạt các công cụ và thư viện kích hoạt và hỗ trợ tự động hóa các trình duyệt web.
- Selenium cung cấp các tiện ích mở rộng để mô phỏng tương tác của người dùng với các trình duyệt, một máy chủ phân phối để mở rộng phân bổ trình duyệt và cơ sở hạ tầng để triển khai đặc tả W3C WebDriver cho phép bạn viết mã có thể hoán đổi cho tất cả các trình duyệt web chính.
- Vì thế, nhóm cũng sử dụng Selenium để thu thập dữ liệu từ web mặc dù nó chủ yếu là một công cụ kiểm tra web tự động, không nhiều về việc thu thập dữ liệu.
  - Hướng dẫn cài đặt Selenium:
    - Đầu tiên, máy tính cần có python (chúng tôi đang sử dụng Python 3) và pip.  
Nếu không có, vui lòng truy cập:  
<https://realpython.com/installing-python/> để cài đặt python  
<https://www.liquidweb.com/kb/install-pip-windows/> để cài đặt pip.
    - Sau đó trong cửa sổ lệnh (hoặc Terminal) gõ:  

```
1 pip install selenium
```
    - Một việc cần làm nữa là tải webdriver (ở đây chúng tôi sử dụng trình duyệt Google Chrome) (Bạn có thể tải webdriver Google Chrome tương thích với hệ điều hành máy tính của bạn từ đây:  
<https://chromedriver.chromium.org/downloads>
    - Sau đó, vui lòng giải nén tệp đã tải xuống và sao chép tệp **chromedriver.exe** giống như tệp này đường dẫn:  
<C:Webdriver//chromedriver.exe>



### 3.3 Tìm hiểu các công cụ phục vụ cho việc lưu trữ dữ liệu

#### 3.3.1 Định dạng dữ liệu

- Data:

- Dữ liệu định tính: Khi dữ liệu được trình bày dưới dạng từ ngữ và mô tả đẽ cập đến văn bản, hình ảnh, video, bản ghi âm, quan sát,...
  - \* Mặc dù bạn có thể quan sát dữ liệu này, nhưng nó chủ quan và do đó, khó phân tích dữ liệu trong nghiên cứu, đặc biệt là để so sánh.
  - \* Ví dụ: Dữ liệu chất lượng đại diện cho mọi thứ mô tả hương vị, kinh nghiệm, kết cấu hoặc ý kiến được coi là dữ liệu chất lượng. Loại dữ liệu này thường được thu thập thông qua các nhóm tập trung, phỏng vấn cá nhân hoặc sử dụng các câu hỏi mở trong các cuộc khảo sát.
- Dữ liệu định lượng: Bất kỳ dữ liệu nào được biểu thị bằng số lượng các số liệu. Loại dữ liệu này có thể được phân biệt thành các loại, được nhóm, đo lường, tính toán hoặc xếp hạng.
  - \* Ví dụ: các câu hỏi như tuổi, thứ hạng, chi phí, chiều dài, cân nặng, điểm số,...

- Phân loại data:

- Quan sát:
  - \* Được chụp trong thời gian thực.
  - \* Không thể tái tạo hoặc lấy lại. Dôi khi được gọi là 'dữ liệu duy nhất'.
  - \* Ví dụ bao gồm chỉ số cảm biến, đo từ xa, kết quả khảo sát, hình ảnh và quan sát của con người.
- Thực nghiệm:
  - \* Dữ liệu từ thiết bị phòng thí nghiệm và trong các điều kiện được kiểm soát.
  - \* Thường có thể tái tạo, nhưng có thể tồn kém để làm như vậy.
  - \* Ví dụ bao gồm trình tự gen, sắc ký đồ, đọc từ trường và quang phổ.
- Mô phỏng:
  - \* Dữ liệu được tạo ra từ các mô hình thử nghiệm nghiên cứu các hệ thống thực tế hoặc lý thuyết.



- \* Mô hình và siêu dữ liệu trong đó dữ liệu đầu vào quan trọng hơn dữ liệu đầu ra.
- \* Ví dụ bao gồm các mô hình khí hậu, mô hình kinh tế và kỹ thuật hệ thống.
- Bắt nguồn hoặc biên dịch:
  - \* Kết quả phân tích dữ liệu hoặc tổng hợp từ nhiều nguồn.
  - \* Có thể tái tạo (nhưng rất đắt).
  - \* Ví dụ bao gồm khai thác văn bản và dữ liệu, cơ sở dữ liệu được biên dịch và mô hình 3D.
- Tham chiếu hoặc chuẩn:
  - \* Tập dữ liệu thu thập cố định hoặc không phải trả tiền, thường được đánh giá ngang hàng và thường được xuất bản và sắp xếp.
  - \* Ví dụ bao gồm cơ sở dữ liệu trình tự gen, dữ liệu điều tra dân số, cấu trúc hóa học.
- Định dạng tệp:
  - \* Định dạng tệp nên được chọn để đảm bảo chia sẻ, truy cập lâu dài và bảo quản dữ liệu của bạn.
  - \* Chọn các tiêu chuẩn và định dạng mở dễ sử dụng lại.
  - \* Nếu bạn đang sử dụng một định dạng khác trong giai đoạn thu thập và phân tích nghiên cứu của mình, hãy đảm bảo bao gồm thông tin trong tài liệu của bạn về các tính năng có thể bị mất khi tệp được chuyển sang định dạng bảo quản, cũng như bất kỳ phần mềm cụ thể nào sẽ cần thiết để xem hoặc làm việc với dữ liệu.



### 3.3.2 JSON

- JavaScript Object Notation (JSON) là một giải pháp thay thế đang thu hút rất nhiều sự chú ý. Điều đầu tiên khi nhắc đến JSON với các nhà phát triển là được thiết kế nhẹ nhàng để có thể dễ dàng đọc, trao đổi dữ liệu và thực thi. Tuy nhiên, đó không phải là lý do duy nhất bạn nên sử dụng JSON cho tích hợp API Restful.
- Các API RESTful phụ thuộc vào việc trao đổi dữ liệu dễ dàng, đáng tin cậy và nhanh chóng. JSON phù hợp với hóa đơn cho từng thuộc tính này. Đây là lý do tại sao sử dụng ngôn ngữ này nhiều sẽ giúp bạn tạo ra các dịch vụ API thú vị.
- Mặc dù CSV là định dạng tệp phổ biến nhất cho dữ liệu “phẳng”, JSON là định dạng tệp phổ biến nhất cho dữ liệu “dạng cây” có khả năng có nhiều lớp, như các nhánh trên cây:

```
1      {
2      [
3          {
4              "id":0,
5              "type":"banana",
6              "amount":12
7          },
8          {
9              "id":1,
10             "type":"apple",
11             "amount":7
12         }
13     ]
14 }
```

- Đối với tệp JSON, bản xem trước tab Dữ liệu sẽ hiển thị một cây tương tác với các nút trong tệp JSON được đính kèm. Bạn có thể nhấp vào các phím riêng lẻ để mở và thu gọn các phần của cây, khám phá cấu trúc của tập dữ liệu khi bạn tiếp tục. Các tệp JSON không hỗ trợ mô tả hoặc chỉ số cột.

### 3.3.3 Ưu điểm của JSON

- JSON nhanh hơn:

Cú pháp JSON rất dễ sử dụng. Chúng ta chỉ phải sử dụng như một cú pháp giúp chúng ta dễ dàng phân tích dữ liệu và thực thi dữ liệu nhanh hơn. Vì cú pháp của nó rất nhỏ và trọng lượng nhẹ, đó là lý do mà nó thực thi phản hồi theo cách nhanh hơn. Cách tiếp cận nhẹ của JSON có thể tạo ra những



cải tiến đáng kể trong các API RESTful hoạt động với các hệ thống phức tạp.

- Phân tích cú pháp máy chủ:

Phân tích cú pháp phía máy chủ là phần quan trọng mà các nhà phát triển muốn nếu quá trình phân tích cú pháp sẽ nhanh ở phía máy chủ thì chỉ người dùng mới có thể nhận được phản hồi nhanh chóng của phản hồi của họ vì vậy trong trường hợp này phân tích cú pháp phía máy chủ JSON là điểm mạnh mà cho biết chúng tôi sử dụng JSON ở phía máy chủ.

- Hỗ trợ lược đồ:

Nó có nhiều loại trình duyệt được hỗ trợ tương thích với các hệ điều hành, vì vậy các ứng dụng được tạo bằng mã hóa JSON không đòi hỏi nhiều nỗ lực để làm cho nó tương thích với tất cả các trình duyệt. Trong quá trình phát triển, nhà phát triển nghĩ cho các trình duyệt khác nhau nhưng JSON cung cấp chức năng đó.

- Công cụ chia sẻ dữ liệu:

JSON là công cụ tốt nhất để chia sẻ dữ liệu ở bất kỳ kích thước nào kể cả âm thanh, video,... Điều này là do JSON lưu trữ dữ liệu trong các mảng nên việc truyền dữ liệu dễ dàng hơn. Vì lý do này, JSON là một định dạng tệp cao cấp cho các API web và để phát triển web.

### 3.4 Tìm hiểu các công cụ quản trị Hệ cơ sở dữ liệu

#### 3.4.1 Cách kết nối Mysql sử dụng python

- To update pip in the virtual environment, type the following command:

```
1 pip install -U pip
```

- Type the command for the package you want to install:

- To install the mysqlclient package, type the following command:

```
1 pip install mysqlclient
```

- To install the mysql-connector-python package, type the following command:

```
1 pip install mysql-connector-python
```

- To install the pymysql package, type the following command:

```
1 pip install pymysql
```



### 3.5 Tìm hiểu các thư viện của python phục vụ cho việc phân tích, thống kê dữ liệu

Để thuận lợi cho việc thống kê, phân tích dữ liệu sau khi đã được làm sạch, nhóm chúng em sử dụng các thư viện được cung cấp bởi Python như Pandas, Matplotlib, Numpy.

#### 3.5.1 Pandas

- Thư viện pandas trong Python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu này.
- Lý do sử dụng pandas trong Python:
  - DataFrame đem lại sự linh hoạt và hiệu quả trong thao tác dữ liệu và lập chỉ mục.
  - Là một công cụ cho phép đọc, ghi dữ liệu giữa bộ nhớ và nhiều định dạng file: csv, text, excel, sql database, hdf5.
  - Liên kết dữ liệu thông minh, xử lý được trường hợp dữ liệu bị thiếu. Tự động đưa dữ liệu lộn xộn về dạng có cấu trúc.
  - Hiệu quả cao trong trộn và kết hợp các tập dữ liệu.
- Cách cài đặt:

```
1 pip install pandas
```

#### 3.5.2 Matplotlib

- Thư viện matplotlib trong Python là một thư viện để thực hiện các suy luận thống kê cần thiết, trực quan hóa dữ liệu bằng đồ thị hoặc hình vẽ. Nó là một thư viện vẽ đồ thị rất mạnh mẽ hữu ích để phối hợp làm việc với Python và thư viện Numpy.
- Thư viện matplotlib cung cấp rất nhiều loại đồ thị, bao gồm cả đồ thị 2D và 3D như:
  - Line plot: Biểu đồ dạng đường.
  - Subplot: Hiển thị đa biểu đồ trong cùng một hàng.
  - Histograms: Biểu đồ phân phối tần suất.
  - Path: Biểu đồ dạng đường đi.



- Streamplot: Biểu đồ dạng vectơ.
- Ellipses: Biểu đồ các hình spacecraft.
- Bars chart: Biểu đồ cột.
- Pie chart: Biểu đồ tròn.
- Scatter plots: Biểu đồ dạng điểm.

- Cách cài đặt:

```
1 pip install matplotlib
```

### 3.5.3 NumPy

- Numpy là một thư viện cốt lõi của Python phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó.
- Numpy cung cấp rất nhiều chức năng cho việc tính toán khối lượng dữ liệu lớn như: mảng nhiều chiều, hàm xác suất thống kê, broadcasting, ...

- Cách cài đặt:

```
1 pip install numpy
```



## 4 Hiện thực việc crawl dữ liệu từ các trang web

### 4.1 Crawl dữ liệu từ trang <http://vietnamworks.com/>

- Sử dụng công cụ Scrapy (đã tìm hiểu ở mục 3.2.1) và Selenium (đã tìm hiểu ở mục 3.2.3) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON.
- Kết quả thu thập được:  $\approx 7000$  job items.

```
① jobdata.json x
201CO10313_BCH-DOAN > vietnamwork > datacrawler > datacrawler > spiders > ② jobdata.json > ...
1 [
2   {
3     "job_title": "Executive Assistant (English & Japanese N2)",
4     "company": "DAIKIN AIR CONDITIONING (VIETNAM) JOINT STOCK COMPANY",
5     "salary": "Thương lượng",
6     "location": "201-203 Cách Mạng Tháng 8, P.4, Q.3",
7     "position": "Executive Assistant (English & Japanese N2)",
8     "job_description": "Doing report as assigned. Doing other analysis regarding Daikin different sales channels & products. Provide, necessary",
9     "job_requirement": "Bachelor Degree in any field or engineer, related industry knowledge is a plus, Proficient in MS-Office, especially Excel",
10    "benefit": "Du Lịch, Phụ cấp, hưởng, Tăng lương, Chăm sóc sức khỏe, Đào tạo, Công tác phí, Phụ cấp thăm niệu, Chế độ nghỉ phép",
11    "quantity": "1"
12  },
13  {
14    "job_title": "Sales Support/ Nhân Viên Hỗ Trợ Bán Hàng",
15    "company": "ACE ANTENNA CO., LTD",
16    "salary": "$1500 - $2000",
17    "location": "KCN ĐÔNG VĂN -DUY TIẾN- HÀ NAM, Quốc lộ 38, tt. Đông Văn, Duy Tiên, Hà Nam, Việt Nam",
18    "position": "Sales Support/ Nhân Viên Hỗ Trợ Bán Hàng",
19    "job_description": "Receive P/O, discuss with customers to fix delivery schedule, Inform Production Planning/Material purchase/Logistics Team",
20    "job_requirement": "Having at least 2 years experiences on Sales/Sales Account/Sales Admin major, Good MS office computer skill, At least 2",
21    "benefit": "Lương tháng thứ 13, Miễn phí các bữa ăn tại Công ty, Xe đưa đón Hà Nội => Công ty",
22    "quantity": "1"
23  },
24  {
25    "job_title": "Test Engineer (Up to $1,500)",
26    "company": "LG VEHICLE COMPONENT SOLUTIONS DEVELOPMENT CENTER VIETNAM (LG VS DCV)",
27    "salary": "Thương lượng",
28    "location": "34F & 36F Keangnam Landmark 72, Pham Hung, Nam Tu Liem, Ha Noi",
29    "position": "Test Engineer (Up to $1,500)",
30    "job_description": "What you will do: Test software for automotive products., Essential Duties and Responsibilities:, Software testing car I",
31    "job_requirement": "Bachelor degree in Information Technology, Computer Science, Computer Engineering, Electronic Engineering, Telecommunications",
32    "benefit": "Competitive remuneration package (up to 16-month salary), Training & on-site opportunities abroad, 12 days of annual leave & 7 days of sick leave",
33    "quantity": "1"
34 }
```

Hình 1: Dữ liệu thu được khi crawl từ trang <http://vietnamworks.com/>

- Khó khăn gặp phải và cách khắc phục:

- Ban đầu nhóm sử dụng công cụ Scrapy để crawl dữ liệu, tuy nhiên trang web đã mã hóa để ẩn mã nguồn HTML của trang tìm kiếm việc làm nên không thể lấy được thông tin tuyển dụng đồng thời phải đăng nhập thì mới có thể lấy được mức lương của các tin tuyển dụng.

Cách khắc phục: Nhóm sử dụng Selenium để giả lập web, dùng để đăng nhập (để có thể lấy được mức lương của công việc ở các tin

tuyển dụng), lấy các đường link dẫn tới thông tin chi tiết của các tin tuyển dụng.

- Khi sử dụng Selenium để lấy các đường link dẫn tới thông tin chi tiết của các tin tuyển dụng thì ở mỗi trang chỉ lấy được đường link của 9 job items đầu tiên.

**Khắc phục:** Nhóm tìm ra được nguyên nhân của vấn đề (do khi vào mỗi trang, ta phải lướt xuống dưới thì các đường link của tất cả các tin tuyển dụng mới hiện ra, nếu không chỉ có 9 đường link của 9 tin đầu tiên) sau đó đã tìm ra phương pháp để giải quyết, cuối cùng có thể lấy được tất cả các đường link của các tin tuyển dụng.

#### 4.2 Crawl dữ liệu từ trang <http://vietsingworks.com/>

- Sử dụng công cụ Scrapy (đã tìm hiểu ở mục 3.1) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON (chia thành nhiều file để tăng tốc độ xử lý dữ liệu).
  - Kết quả thu được:  $\approx$  2000 job items.

```
① jobdata.json ×
201CO10313_BCH-DOAN > vietsingwork > vietsingwork > spiders > ① jobdata.json ...
```

1 [ ]  
2 {"job\_title": " OFFICE ADMINISTRATOR (\$2,200- \$2,400)", "salary": " SGD \$2,200- \$2,400", "location": " singapore", "position": " ", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " Discuss when interview", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 1", "is\_new": " 1"}  
3 {"job\_title": " Waitress", "salary": " Discuss when interview", "location": " singapore", "position": " Worker", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " Cong nhân sản xuất tại công ty may đắp (SGD \$2600)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 2", "is\_new": " 1"}  
4 {"job\_title": " FRONT DESK OFFICER (FOREIGN BANK/\$2500)", "salary": " SGD \$2500", "location": " singapore", "position": " Worker", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " JUNIOR SECRETARY (SGD \$2200)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 3", "is\_new": " 1"}  
5 {"job\_title": " RECEPTIONIST (Hotel / Front Line / #2350)", "salary": " SGD \$2200 - \$2300", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " Automotive technician (\$2600)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 4", "is\_new": " 1"}  
6 {"job\_title": " AUTOMOTIVE TECHNICIAN (\$2600)", "salary": " SGD \$2600", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " BARTENDER (RESTAURANT/ \$2200 + tip)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 5", "is\_new": " 1"}  
7 {"job\_title": " MARINE ENGINE SPECIALIST (\$2200 - \$3000)", "salary": " \$2200 - \$3000", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " Waiter / Waitress ( \$2200 SGD)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 6", "is\_new": " 1"}  
8 {"job\_title": " CHEF (SEAFOOD/VEGETARIAN) - SGD \$4000", "salary": " SGD \$4000", "location": " singapore", "position": " Supervisor/Manager", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " BUTLER - HOTEL SERVICES ( \$2200)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 7", "is\_new": " 1"}  
9 {"job\_title": " BUTLER - HOTEL SERVICES ( \$2200)", "salary": " SGD \$2200", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " RESTAURANT ASSISTANT MANAGER (Restaurant / \$2400 - \$2600)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 8", "is\_new": " 1"}  
10 {"job\_title": " EXPRESS SERVICE - CALL CENTRE HOTEL SERVICES (\$2200)", "salary": " SGD \$2200", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " IT Development Specialist ( SGD \$4500 - \$6000)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 9", "is\_new": " 1"}  
11 {"job\_title": " CUSTOMER SERVICE OFFICER ( SGD \$2300 - \$2400)", "salary": " SGD \$2300- \$2400", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " SALESMAN (SGD \$2300 - \$2600)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 10", "is\_new": " 1"}  
12 {"job\_title": " SALESMAN (SGD \$2300 - \$2600)", "salary": " SGD \$2300 - \$2600", "location": " singapore", "position": " Specialist", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " GLASS MANUFACTURING INDUSTRY - PRODUCTION SUPERVISOR", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 11", "is\_new": " 1"}  
13 {"job\_title": " FINANCE ANALYST ( SGD \$2,800)", "salary": " SGD \$2,800", "location": " singapore", "position": " ", "job\_type": " Permanent Staff", "age": " 21-30", "experience": " 0-1 years", "industry": " All Industries", "education": " Secondary School", "skills": " Microsoft Office", "description": " FINANCIAL ANALYST (\$2,800)", "status": " Active", "date": " 2023-09-11", "source": " JobStreet.com", "ref": " 12", "is\_new": " 1"}  
14 ]

Hình 2: Dữ liệu thu được khi crawl từ trang <http://vietsinaworks.com/>

- Khó khăn gặp phải:
    - Khi cài đặt thư viện scrapy của python bằng command trên terminal:

```
1 pip install scrapy
```
    - 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.microsoft.com/downloads/
-----
ERROR: Command errored out with exit status 1: 'c:\program files (x86)\python38-32\python.exe' -u -c 'import sys, setuptools, 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\\install-record.txt' --single-version-externally-managed --compile --install-headers 'c:\\\\program files (x86)\\\\python38-32\\\\Include\\\\Twisted' Check the logs for full command output.
```

Hình 3: Lỗi khi cài đặt thư viện scrapy của python

- Để khắc phục lỗi này, mở terminal cài đặt Twisted bằng command trên terminal:

```
1 pip install Twisted
```

### 4.3 Crawl dữ liệu từ trang http://careerbuilder.vn/

- Sử dụng công cụ Beautiful Soup (đã tìm hiểu ở mục 3.2) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON (chia thành nhiều file để tăng tốc độ xử lý dữ liệu).
- Kết quả thu được:  $\approx 10000$  job items.

```
0 data_01.json ×
201CO10313_BCH-DOAN > careerbuilder > 0 data_01.json > ...
1 [
2   {
3     "job_title": "Digital marketing executive (Upto 15M + Nghỉ T7,CN)",
4     "company": "Homedy Inc.",
5     "salary": "10 Tr - 18 Tr VND",
6     "location": "Tầng 3, Khu A (văn phòng), Tòa nhà Imperia Garden, 203 Nguyễn Huy Tưởng, Thanh Xuân, Hà Nội",
7     "position": "Digital marketing executive (Upto 15M + Nghỉ T7,CN)",
8     "job_description": "- Lên kế hoạch Marketing phù hợp với mục tiêu hoạt động cho từng giai đoạn của công ty- Triển khai các chiến dịch Digi",
9     "job_requirement": "1 - 3 Năm",
10    "benefit": "Chế độ bảo hiểm, Du Lịch, Chế độ thưởng, Chăm sóc sức khỏe, Đào tạo, Tăng lương",
11    "quantity": ""
12  },
13  {
14    "job_title": "Test Engineer- Electronics US Company- VSIP2- BD",
15    "company": "East West Industries Vietnam LLC.",
16    "salary": "Cạnh tranh",
17    "location": "No. 26A Street No. 5, VSIP 2, Hoa Phu Ward, Thu Dau Mot City, Binh Duong, Viet Nam",
18    "position": "Test Engineer- Electronics US Company- VSIP2- BD",
19    "job_description": "Report to the PCB Engineering ManagerOnsite support testing and debug of product, system on productionsupport define",
20    "job_requirement": "1 - 3 Năm",
21    "benefit": "Laptop, Chế độ bảo hiểm, Du Lịch, Phụ cấp, Xe đưa đón, Chế độ thưởng, Chăm sóc sức khỏe, Đào tạo, Tăng lương, Công t",
22    "quantity": ""
23  },
24  {
25    "job_title": "Trưởng phòng Kinh Doanh (Ô Tô Mitsubishi)",
26    "company": "Trung tâm Ô Tô Mitsubishi Nam Miền Trung -Ninh Thuận",
27    "salary": "15 Tr - 30 Tr VND",
28    "location": "QL 1A, Vĩnh Tân, Tuy Phong, Bình Thuận",
29    "position": "Trưởng phòng Kinh Doanh (Ô Tô Mitsubishi)",
30    "job_description": "1.Xây dựng kế hoạch, chiến lược bán hàng ngắn, trung và dài hạn của Công ty.2.Xây dựng các chương trình marketing, quâ",
31    "job_requirement": "3 - 5 Năm",
32    "benefit": "Laptop, Chế độ bảo hiểm, Du Lịch, Phụ cấp, Đồng phục, Chế độ thưởng, Đào tạo, Tăng lương, Công tác phí, Nghỉ phép nă",
33    "quantity": ""
34  },
35]
```

Hình 4: Dữ liệu thu được khi crawl từ trang http://careerbuilder.vn/



#### 4.4 Crawl dữ liệu từ trang <http://mywork.com.vn/>

- Sử dụng công cụ Scrapy (đã tìm hiểu ở mục 3.1) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON (chia thành nhiều file để tăng tốc độ xử lý dữ liệu).
- Kết quả thu được:  $\approx 1000$  job items.

```
data.json x
201CO10313_BCH-DOAN > mywork > mywork > data.json ...
1
2 { "Position": "Chuyên Viên Tư Vấn Bán Hàng Online", "Company": "Công Ty TNHH Thương Mại Dịch Vụ Điện Tử Hà Nguyễn", "Salary": "7 triệu - 10 tri
3 { "Position": "Nhân Viên Kinh Doanh Tại Hải Phòng", "Company": "Cty CP Thủ Sân Anh Minh", "Salary": "10 triệu - 12 triệu", "Location": "Hải Ph
4 { "Position": "Nhân Viên Sale Nội Thất", "Company": "Công Ty TNHH Xuất Nhập Khẩu Được Phẩm Đức Minh", "Salary": "7 triệu - 10 triệu", "Location
5 { "Position": "Nhân Viên Sale Admin (hỗ Trợ Kinh Doanh)", "Company": "Công Ty TNHH Ngô Thành Thành", "Salary": "7 triệu - 10 triệu", "Location"
6 { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Môi Trường Sen Vàng", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh,
7 { "Position": "Nhân Viên Tư Vấn Visa Lương Trên 15 Triệu", "Company": "Công Ty TNHH Thương Mại Dịch Vụ Quốc Tế Việt Thế Giới", "Salary": "15 tr
8 { "Position": "Nhân Viên Kinh Doanh", "Company": "Silver City Apartment", "Salary": "7 triệu - 10 triệu", "Location": "Hà Nội" },
9 { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Orai Việt Nam", "Salary": "7 triệu - 10 triệu", "Location": "Hồ Chí Minh" },
10 { "Position": "Nhân Viên Phát Triển Thị Trường (tp Hà Nội)", "Company": "Công Ty Cổ Phần Waytech Việt Nam", "Salary": "6 triệu - 30 triệu", "Lo
11 { "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
12 { "Position": "Nhân Viên Bán Hàng Thời Trang - Quận 9 Hcm", "Company": "Công Ty TNHH May Vĩnh Phú", "Salary": "5 triệu - 7 triệu", "Location": "
13 { "Position": "Bản Hiểm Manulife Tim Dối Tắc Khởi Nghiệp Part Time / Full Time Tại Hà Nội (dự án Khởi Nghiệp VÀ Phát Triển Lãnh Đạo Fta Kế Cán
14 { "Position": "Nhân Viên Kinh Doanh", "Company": "Công Ty TNHH Bảo An Sính", "Salary": "7 triệu - 10 triệu", "Location": "Bình Định, Đà Nẵng, Q
15 { "Position": "Nhân Viên Sale Online (mảng Nha Khoa Thẩm Mỹ)", "Company": "Công Ty TNHH Thiên Phúc - Nha Khoa Dora", "Salary": "8 triệu - 10 tr
16 { "Position": "Nhân Viên Marketing Facebook", "Company": "Công Ty Cổ Phần Thương Mại Và Dịch Vụ Asiatech Việt Nam", "Salary": "6 triệu - 9 triệ
17 { "Position": "Nhân Viên Tư Vấn Chăm Sóc Khách Hàng", "Company": "Công Ty TNHH Trang Thiết Bị Ô Tô Việt Nam", "Salary": "8 triệu - 12 triệu", "
18 { "Position": "Chuyên Viên Tư Vấn (thu Nhập Trên 20tr / Tháng)", "Company": "Infinox", "Salary": "20 triệu - 30 triệu", "Location": "Hà Nội", H
19 { "Position": "Nhân Viên Trực Fanpage VÀ Trang Thương Mại Điện Tử", "Company": "Công Ty TNHH Xuất Nhập Khẩu Hs Trí Tuệ", "Salary": "6 triệu - 1
20 { "Position": "Nhân Viên Phát Triển Thị Trường", "Company": "Công Ty Cổ Phần L.q Joton", "Salary": "8 triệu - 15 triệu", "Location": "Hồ Chí Mi
21 { "Position": "Nhân Viên Kinh Doanh Ngành Điện Máy và Nội Thất - Không Cần Kinh Nghiệm", "Company": "công Ty TNHH KDT-89", "Salary": "7 triệu -
22 { "Position": "Nhân Viên Văn Phòng _ Không Telesales (không Yêu Cầu Bằng & Kinh Nghiệm) _ Hcm5", "Company": "Công Ty TNHH Khu Du Lịch Vịnh Thiê
23 { "Position": "Chuyên Viên Tư Vấn _hcm", "Company": "công Ty Cổ Phần Đầu Tư Và Kinh Doanh Nhà Thời Đại", "Salary": "7 triệu - 10 triệu", "Locat
24 { "Position": "Nhân Viên Kinh Doanh Thiết Bị Môi Trường Tại Hà Nội (10-15tr / Tháng)", "Company": "Công Ty Cổ Phần Đầu Tư Thương Mại Va Ký Thuậ
25 { "Position": "Tổng Đài Viên (tư Vấn & Đặt Lịch Hẹn) - Kênh Liên Kết Công Đồng Đầu", "Company": "Công Ty TNHH Manulife (việt Nam)", "Salary": "
26 { "Position": "Nhân Viên Kinh Doanh Thiết Bị Môi Trường (10-15tr / Tháng)", "Company": "Công Ty Cổ Phần Đầu Tư Thương Mại Va Ký Thuật Lương Gia
27 { "Position": "Tổng Đài Viên (tư Vấn & Đặt Lịch Hẹn) - Hà Nội", "Company": "Công Ty TNHH Manulife (việt Nam)", "Salary": "5 triệu - 7 triệu", "
28 { "Position": "Nhân Viên Tư Vấn - Telesales Làm Việc Lại Văn Phòng", "Company": "Công Ty Tài Chính Trách Nhiệm Hữu Hạn Một Thành Viên Shinhan V
29 { "Position": "Nhân Viên Chăm Sóc Khách Hàng", "Company": "Công Ty Cổ Phần Đầu Tư VÀ Kinh Doanh Nhà Thời Đại", "Salary": "7 triệu - 10 triệu",
30 { "Position": "Nhân Viên Hồ Trợ Kinh Doanh Tp. Hồ Chí Minh", "Company": "Công Ty Tài Chính TNHH MTV Home Credit Việt Nam", "Salary": "10 triệu
31 { "Position": "Nhân Viên Bán Hàng (tp.hcm)", "Company": "Công Ty TNHH Thương Mại Và Đầu Tư Phát Triển Xuân Phúc", "Salary": "5 triệu - 7 triệu"
32 { "Position": "Nhân Viên Tổng Đài Ngân Hàng Mb - 21 Cát Linh - Quận Đống Đa - Tp Hà Nội", "Company": "Công Ty Cổ Phần Truyền Thông Kim Cương",
```

Hình 5: Dữ liệu thu được khi crawl từ trang <http://mywork.com.vn/>

- Khó khăn gặp phải:

- Khi crawl đến trang 50 thì đường dẫn đến trang 51 trở đi không thay đổi (dữ liệu vẫn ở trang 50), do đó lượng data crawl được tương đối ít.
- Hiện tại nhóm chưa khắc phục được vấn đề này.

#### 4.5 Crawl dữ liệu từ trang <http://1001vieclam.com/itviec/>

- Sử dụng công cụ Beautiful Soup (đã tìm hiểu ở mục 3.2) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON (chia thành nhiều file để tăng tốc độ xử lý dữ liệu).
- Kết quả thu được:  $\approx 10000$  job items.



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

```
(1) data.json x
201CO10313_BCH-DOAN > 1001viec > (1) data.json > ...
1  [
2    {
3      "job_title": "Nhân Viên Kinh Doanh Thiết Bị Y Tế",
4      "company": "Công ty TNHH Khoa Học Và Kỹ Thuật Olympic",
5      "salary": "8.000.000 - 15.000.000 / tháng",
6      "location": "Hồ Chí Minh",
7      "position": "Bán hàng, ', phâm/chẩn đoán, ', Kinh doanh/Phụ trách KH, ', Marketing/tiếp thị, ",
8      "job_description": "Kinh doanh, tư vấn cho khách hàng các sản phẩm về thiết bị y tế. ",
9      "job_requirement": "Trình độ Đại học/Cao đẳng chuyên ngành Quản trị kinh doanh, kinh tế thương mại, Ngoại thương, Marketing. Yêu
10     "benefit": "Thu nhập: Lương cơ bản + Trợ cấp công tác + Điện thoại + Hoa hồng 23% doanh số",
11     "quantity": "1"
12   },
13   {
14     "job_title": "07 PHP Developers (Magento, JavaScript)",
15     "company": "Sutunam Việt Nam",
16     "salary": "2.500 USD / tháng",
17     "location": "Hà Nội",
18     "position": "CNTT/ Viễn thông",
19     "job_description": "Configuration and management of website architecture, including application, and database layers.;', 'Participate in
20     "job_requirement": "Minimum 2year experiences in PHP and/or JavaScript web development, systems integration (for Middle);', 'Minimum 4yea
21     "benefit": "Because you want to be part of a multiwinning award digital Agency;', 'Because you want to work with talented colleagues;', '
22     "quantity": "1"
23   },
24   {
25     "job_title": "Senior Frontend Dev (ReactJS/VueJS)",
26     "company": "Citigo Software",
27     "salary": "2.000 USD / tháng",
28     "location": "Hồ Chí Minh",
29     "position": "CNTT/ Viễn thông",
30     "job_description": "Bạn sẽ quyết định frontend của một trong những sản phẩm thành công của chúng tôi, tạo ra những tác động nổi bật sẽ tr
31     "job_requirement": "Trải nghiệm với React stack (ReactJS, Redux, State Management, Functional Programming)', 'Hiểu biết nâng cao về Java
32     "benefit": "Mức lương khởi điểm hấp dẫn, cạnh tranh, tương xứng với năng lực và kinh nghiệm làm việc;', 'Được xét duyệt tăng lương định k
33   },
34 ]
```

Hình 6: Dữ liệu thu được khi crawl từ trang <http://1001vieclam.com/itviec/>

## 4.6 Crawl dữ liệu từ trang <http://chotot.vn/>

- Sử dụng công cụ Beautiful Soup (đã tìm hiểu ở mục 3.2) để crawl dữ liệu về, dữ liệu được lưu về dưới định dạng JSON (chia thành nhiều file để tăng tốc độ xử lý dữ liệu).
- Kết quả thu được:  $\approx 10000$  job items.

```
(1) sample1_50.json x
201CO10313_BCH-DOAN > chotot > (1) sample1_50.json > ...
1  [
2    {
3      "job_title": "Quê Anh Water Tuyển Giao Hàng Xe Máy Tại Quận 2",
4      "company": "QUẾ ANH WATER",
5      "salary": "Từ 7.500.000 đ/tháng",
6      "location": "Phường Bình Trưng Tây, Quận 2, Tp Hồ Chí Minh",
7      "position": "Tài xế/Giao nhận xe ô tô",
8      "job_description": "-Công ty TNHH Thương Mại Dịch Vụ VÀ Đầu Tư Vương Anh Phát ', '-', '-Cần tuyển 5 nhân viên giao hàng nước uống cao cấp
9      "job_requirement": "Toàn thời gian",
10     "quantity": "10"
11   },
12   {
13     "job_title": "Tuyển Phụ Kho + Phụ Xe Nước Uống - Quê Anh Water",
14     "company": "QUẾ ANH WATER",
15     "salary": "6.500.000 - 7.500.000 đ/tháng",
16     "location": "Phường Bình Trưng Tây, Quận 2, Tp Hồ Chí Minh",
17     "position": "Tài xế/Giao nhận xe ô tô",
18     "job_description": "-Công ty TNHH Thương Mại Dịch Vụ VÀ Đầu Tư Vương Anh Phát ', '-', '-Cần tuyển 01 phụ kho và 01 phụ xe.', '-',
19     "job_requirement": "Toàn thời gian",
20     "quantity": "3"
21   },
22   {
23     "job_title": "Tuyển Người Giữ Em Bé Không Ở Lại",
24     "company": "cá nhân",
25     "salary": "7.000.000 đ/tháng",
26     "location": "Phường Thạnh Mỹ Lợi, Quận 2, Tp Hồ Chí Minh",
27     "position": "Giúp việc/tập sự",
28     "job_description": "-Hiện tại cần tuyển người chăm em bé gái 8 tháng tuổi.', '-Làm việc tại 02 phòng văn phòng, quận 2', '-Thời gian: 7h30-19
29     "job_requirement": "Toàn thời gian",
30     "benefit": "thưởng dịp lễ tết",
31     "quantity": "1"
32   },
33   {
34     "job_title": "CÔNG TY VƯƠNG ANH PHÁT Tuyển Dịch Vụ Khách Hàng",
35     "company": "CÔNG TY VƯƠNG ANH PHÁT"
36     "salary": "5.500.000 - 6.000.000 đ/tháng",
37     "location": "Phường Bình Trưng Đông, Quận 2, Tp Hồ Chí Minh"
```

Hình 7: Dữ liệu thu được khi crawl từ trang <http://chotot.vn/>

- Khó khăn gấp phải:
    - Khi ta gọi request trả về nội dung của file HTML, ta thu được như hình:

**Hình 8:** NỘI DUNG FILE HTML TRẢ VỀ KHI QUỐC REQUEST ĐẾN TRANG WEB

- Trong trường hợp này, file HTML mà beautiful soup trả về bị lỗi. Để khắc phục, ta đổi nội dung file sang dạng chuỗi và dùng các method của python để lấy các trường dữ liệu tương ứng.



## 4.7 Tổng kết bước tiến hành crawl dữ liệu

- Nhóm đã tiến hành crawl được tổng cộng  $\approx 40000$  items job, cụ thể là:

STT	Trang web	Số lượng
1	<a href="http://vietnamworks.com/">http://vietnamworks.com/</a>	7000
2	<a href="http://vietsingworks.com/">http://vietsingworks.com/</a>	2000
3	<a href="http://careerbuilder.vn/">http://careerbuilder.vn/</a>	10000
4	<a href="http://mywork.com.vn/">http://mywork.com.vn/</a>	1000
5	<a href="http://1001vieclam.com/itviec/">http://1001vieclam.com/itviec/</a>	10000
6	<a href="http://chotot.vn/">http://chotot.vn/</a>	11000

- Dữ liệu crawl về còn ở dạng thô, cần phải qua quá trình làm sạch để có thể đem đi phân tích, xử lý.
- Vẫn còn một số vấn đề nhóm chưa giải quyết được, như trang <http://mywork.com.vn/>



## 5 Hiện thực làm sạch dữ liệu

Để tiến hành làm sạch dữ liệu crawl về được từ các trang web, nhóm tiến hành thực hiện làm sạch ở 2 bước: làm sạch ở mức unit (cá nhân) và làm sạch ở mức program (chương trình).

### 5.1 Tiến hành làm sạch dữ liệu ở mức unit (cá nhân)

- Trong quá trình crawl dữ liệu về máy và lưu trữ ở file JSON, nhóm đã thực hiện làm sạch dữ liệu ở mức unit, nghĩa là các thành viên tự làm sạch dữ liệu và lưu vào các trường đã được thống nhất.
- Các trường dữ liệu cần được làm sạch là: Job title, Company, Salary, Location, Position, Job description, Job requirement, Benefit, Quantity.
- Để có thể làm sạch dữ liệu, nhóm đã sử dụng các method xử lý chuỗi trong Python.
- Kết quả trả về: dữ liệu được phân thành các trường và lưu về file JSON.
- Quá trình thực hiện việc làm sạch ở mức unit được tích hợp trong quá trình crawl dữ liệu (mục 4.1) nên nhóm không đề cập thêm ở đây.

### 5.2 Tiến hành làm sạch dữ liệu ở mức program (chương trình)

- Sau khi đã thực hiện làm sạch dữ liệu ở mức unit, nhóm tiến hành tập trung dữ liệu vào một file JSON chung và chuyển sang định dạng CSV (để phục vụ cho quá trình phân tích, thống kê ở bước sau).
- Tiếp theo, nhóm tiến hành định dạng lại các trường dữ liệu theo ràng buộc sau:
  - Job title: string, độ dài bất kì.
  - Salary: string, có định dạng là x/y (x là giá trị lương, y là thời hạn trả tiền, VD: tháng, năm, ...) hoặc là "Thương lượng".
  - Company: string, độ dài bất kì.
  - Location: string, độ dài bất kì, phải là tỉnh thành của Việt Nam.
  - Position: string, độ dài bất kì.
  - Job description: string, độ dài bất kì.



- Job requirement: string, độ dài bất kì.
  - Benefit: string, độ dài bất kì.
  - Quantity: int.
- Mục đích của việc làm sạch và ràng buộc dữ liệu như trên là để việc phân tích, thống kê trở nên dễ dàng và thuận lợi hơn.
  - Để có thể làm sạch dữ liệu theo định dạng trên, nhóm đã sử dụng biểu thức chính quy (regular expression) trong thư viện **re**, đồng thời kết hợp các method xử lý chuỗi trong Python.
  - Sau khi đã định dạng lại dữ liệu, nhóm tiến hành lọc qua 2 bước: lọc những dữ liệu bị thiếu trường giá trị và lọc những dữ liệu bị trùng.

### 5.3 Tiến hành lọc dữ liệu bị thiếu trường dữ liệu - Missing data

- Để tiến hành lọc các dữ liệu bị trùng, ta sử dụng thư viện pandas của python. Pandas cho phép thao tác linh hoạt với missing data trong Series, DataFrame như tìm giá trị bị thiếu (missing value), xác định giá trị tồn tại (không bị thiếu), loại bỏ giá trị bị thiếu, chèn giá trị bị thiếu, điền vào giá trị bị thiếu.
- Đầu tiên ta tiến hành đọc filexlsx, sử dụng thư viện pandas để lưu dữ liệu vào DataFrame. Sau đó ta tiến hành lọc những dữ liệu bị thiếu trường giá trị bằng cách sử dụng method pandas.isna(). Hàm pandas.isna() trả về giá trị False nếu có dữ liệu và trả về giá trị True nếu thiếu hoặc dữ liệu bị sai.

The screenshot shows a code editor with two files open: 'analyze.py' and 'missing\_data.py'. The 'missing\_data.py' file contains the following code:

```
analyze.py missing_data.py
data_analysis > missing_data.py > {} np
1 import pandas as pd
2 import numpy as np
3 import re
4 header_list = ["job_title", "company", "salary", "location", "position", "job_description", "job_requirement", "benefit", "quantity"]
5 df = pd.read_excel('data_1.xlsx', engine='openpyxl')
6 df_check = df.isna()
7 df1 = df[df.isna().any(axis=1)]
8 print(df_check)
```

The terminal below shows the execution of the script:

```
PS D:\BK\201\ĐÁN KÌI\Assignment_2\201C010313_BCH-DOAN\data_analysis> python missing_data.py
job_title company salary location position job_description job_requirement benefit quantity
0 False False False False False False False False False
1 False False False False False False False False False
2 False False False False False False False False False
3 False False False False False False False False False
4 False False False False False False False False False
...
9994 False False False False False False False False True
9995 False False False False False False False False True
9996 False False False False False False False False True
9997 False False False False False True True True True
9998 False True True True True True True True True
```

[9999 rows x 9 columns]

Hình 9: Kết quả khi gọi hàm `pandas.isna()`



- Mặt khác, ta có thể lọc cái hàng dữ liệu bị thiếu trực tiếp với tham số df.isna().any(axis=1). Method này sẽ trả về các record dữ liệu có ít nhất một trường giá trị bị thiếu.

```
data_analysis > missing_data.py > ...
3 import re
4 header_list = ["job_title", "company", "salary", "location", "position", "job_description", "job_requirement", "benefit", "quantity"]
5 df = pd.read_excel('data_1.xlsx', engine='openpyxl')
6 df_check = df.isna()
7 df1 = df[df.isna().any(axis=1)]
8 print(df1)

TERMINAL PROBLEMS OUTPUT DEBUG CONSOLE
1: powershell
```

	job_title	company	benefit	quantity
20	Trưởng nhóm kinh doanh	CÔNG TY TNHH DV ; TM ROYAL FINANCE	...	Nan 1.0
21	Chuyên viên nhân sự	Meey Land	...	Nan 1.0
22	Giám đốc kinh doanh trực tiếp	Meey Land	...	Nan 1.0
24	Tuyển Chuyên Viên Phòng Dịch Vụ Chứng Khoán	NaN ...	...	Nan 1.0
25	NHÂN VIÊN KINH DOANH KHU VỰC MIỀN BẮC	NaN ...	Thưởng hoàn thành kế hoạch cuối năm dựa vào kế...	1.0
...	...	...	...	...
9994	Senior Account Executive	CÔNG TY CỔ PHẦN HỢP TÁC VÀ PHÁT TRIỂN TRUYỀN T...	Chế độ bảo hiểm, Du Lịch, Chế độ thường, D...	Nan
9995	STRATEGIC PLANNER	CÔNG TY CỔ PHẦN HỢP TÁC VÀ PHÁT TRIỂN TRUYỀN T...	Chế độ bảo hiểm, Du Lịch, Chế độ thường, D...	Nan
9996	Kế Toán Tổng Hợp	CÔNG TY TNHH MỘT THÀNH VIÊN TEX VIỆT NAM	Chế độ bảo hiểm, Du Lịch, Chế độ thường, D...	Nan
9997	Senior Python Developer (Odoo)	Kyanon Digital	...	Nan Nan
9998	\t\tDevelop/customize Odoo modules and APIs in...	NaN ...	...	Nan

[7378 rows x 9 columns]

```
PS D:\BK\201\Đồ án KTLT\Assignment_2\201C010313_BCH-DOAN\data_analysis> []
```

Hình 10: Kết quả khi gọi hàm pandas.isna().any(axis=1)

- Sau khi lọc được các kết quả bị thiếu dữ liệu như trên, nhóm tiến hành xóa các record dữ liệu.
- Kết quả đạt được:  $\approx 33000$  job items có đầy đủ các trường giá trị.

## 5.4 Tiến hành lọc dữ liệu trùng lặp

- Khi làm việc với dữ liệu để phân tích, chắc hẳn sẽ gặp phải những dòng dữ liệu bị trùng nhau, và cần lọc để xóa đi hoặc thay đổi.
- Sử dụng phím tắt trong Excel để loại bỏ giá trị trùng lặp.
  - Bước 1: Ctrl + A: Chọn tất cả bảng dữ liệu này.
  - Bước 2: Alt + A + M: Remove Duplicate – loại bỏ giá trị trùng bằng tổ hợp phím tắt trên.
  - Click OK – Loại bỏ các giá trị trùng lặp.
- Sau khi thực hiện các bước trên ta được  $\approx 25000$  job items có đầy đủ các trường giá trị và không bị trùng lặp.



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

job_title	company	salary
Nhân Viên Kinh Doanh Thiết Bị Y Tế	Công ty TNHH Khoa Học Và Kỹ Thuật Olympic	8.000.000 15.000.000 / tháng
07 PHP Developers (Magento, JavaScript)	Sutunam Việt Nam	2.500 USD / tháng
Senior Frontend Dev (ReactJS/VueJS)	Citigo Software	2.000 USD / tháng
Senior .NET Dev (ASP.NET, C#)	Citigo Software	2.000 USD / tháng
Senior Data Engineer	Biểu trưng của người bán NVG Technology	2.500 USD
02 Lead/Senior QA Engineers	Biểu trưng của người bán NVG Technology	2.000 USD / tháng
Engineering Manager	Biểu trưng của người bán NVG Technology	3.500 USD / tháng
Senior Business Analyst (eCommerce)	Dat Xanh Group	2.500 USD / tháng
FullStack Developer (Python,Ruby,Golang)	Công Ty TNHH MÁY NÔNG NGHIỆP CAO ĐẠT	2.000 USD / tháng
Chuyên Viên Công Nghệ Thông Tin	Công Ty TNHH Hoàng Kim Minh	1.500 USD
IT / Project Support (Networking / SQL)	Công Ty Cổ Phần Thương Mại Tổng Hợp V&V	1.500 USD
Project Manager ( Branch Transformation Project) Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	3.500 USD
.Net Developer	CÔNG TY TNHH THIẾT BỊ DÂN DỤNG THUẬN PHONG	1.500 USD / tháng
Project Manager Quản Lý Dự Án Khối Quản Trị Rủi Ro Hà Nội	CÔNG TY TNHH DV ; TM ROYAL FINANCE	3.500 USD
Information Security Specialist – Application Security Engineer Hà Nội	Meey Land	2.500 USD / tháng
Chuyên Viên Cao Cấp Quản Lý Dữ Liệu Hà Nội	Meey Land	1.500 USD / tháng
Giám đốc Cao cấp Quản lý Khách Hàng Corporate Banking Hà Nội	Công Ty Cổ Phần TM & DV Du Lịch Đất Việt	4.000 USD / năm
Development Expert Collection System Hà Nội	CÔNG TY TNHH MÁY NÔNG NGHIỆP CAO ĐẠT	2.000 USD / tháng
Chuyên Viên Hỗ Trợ Dịch Vụ Công Nghệ Thông Tin Hà Nội	Công Ty TNHH Hoàng Kim Minh	20.000.000 VND / năm
Senior Developer (Crm, Erp) HCM	CÔNG TY TNHH THIẾT BỊ DÂN DỤNG THUẬN PHONG	30.000.000 VND / tháng
Trưởng nhóm kinh doanh	CÔNG TY TNHH DV ; TM ROYAL FINANCE	15.000.000 VND / tháng
Chuyên viên nhân sự	Meey Land	10.000.000 VND / tháng
Giám đốc kinh doanh trực tiếp	Meey Land	15.000.000 VND / tháng
THỰC TẬP SINH TEAM BUILDING	Công Ty Cổ Phần TM & DV Du Lịch Đất Việt	Thương lượng Thương lượng
Tuyển Chuyên Viên Phòng Dịch Vụ Chứng Khoán	CÔNG TY TNHH MÁY NÔNG NGHIỆP CAO ĐẠT	9.000.000 VND / tháng
NHÂN VIÊN KINH DOANH KHU VỰC MIỀN BẮC	Công Ty TNHH Hoàng Kim Minh	10.000.000 VND / tháng
Tuyển Lái Xe Tài Lượng Chế Độ Hấp Dẫn Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp V&V	13.500.000 VND / tháng
Tuyển Dụng Lái Xe Tài Lượng Chế Độ Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng
Tuyển Gấp Lái Xe Tài Lượng Thường Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng
Tuyển Gấp Lái Xe Tài Lượng Thường Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng

Hình 11: Tiến hành chọn các trường để remove duplicates.

job_title	company	salary
Nhân Viên Kinh Doanh Thiết Bị Y Tế	Công ty TNHH Khoa Học Và Kỹ Thuật Olympic	8.000.000 15.000.000 / tháng
07 PHP Developers (Magento, JavaScript)	Sutunam Việt Nam	2.500 USD / tháng
Senior Frontend Dev (ReactJS/VueJS)	Citigo Software	2.000 USD / tháng
Senior .NET Dev (ASP.NET, C#)	Citigo Software	2.000 USD / tháng
Senior Data Engineer	Biểu trưng của người bán NVG Technology	2.500 USD
02 Lead/Senior QA Engineers	Biểu trưng của người bán NVG Technology	2.000 USD / tháng
Engineering Manager	Biểu trưng của người bán NVG Technology	3.500 USD / tháng
Senior Business Analyst (eCommerce)	Dat Xanh Group	2.500 USD / tháng
FullStack Developer (Python,Ruby,Golang)	Công Ty TNHH MÁY NÔNG NGHIỆP CAO ĐẠT	2.000 USD / tháng
Chuyên Viên Công Nghệ Thông Tin	Công Ty TNHH Hoàng Kim Minh	1.500 USD
IT / Project Support (Networking / SQL)	Công Ty Cổ Phần Thương Mại Tổng Hợp V&V	1.500 USD
Project Manager ( Branch Transformation Project) Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	3.500 USD
.Net Developer	CÔNG TY TNHH DV ; TM ROYAL FINANCE	1.500 USD / tháng
Project Manager Quản Lý Dự Án Khối Quản Trị Rủi Ro Hà Nội	Meey Land	3.500 USD
Information Security Specialist – Application Security Engineer Hà Nội	Meey Land	2.500 USD / tháng
Chuyên Viên Cao Cấp Quản Lý Dữ Liệu Hà Nội	Công Ty TNHH DV ; TM ROYAL FINANCE	1.500 USD / tháng
Giám đốc Cao cấp Quản lý Khách Hàng Corporate Banking Hà Nội	Meey Land	4.000 USD / năm
Development Expert Collection System Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	2.000 USD / tháng
Chuyên Viên Hỗ Trợ Dịch Vụ Công Nghệ Thông Tin Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	20.000.000 VND / năm
Senior Developer (Crm, Erp) HCM	CÔNG TY TNHH THIẾT BỊ DÂN DỤNG THUẬN PHONG	30.000.000 VND / tháng
Trưởng nhóm kinh doanh	CÔNG TY TNHH DV ; TM ROYAL FINANCE	15.000.000 VND / tháng
Chuyên viên nhân sự	Meey Land	10.000.000 VND / tháng
Giám đốc kinh doanh trực tiếp	Meey Land	15.000.000 VND / tháng
THỰC TẬP SINH TEAM BUILDING	Công Ty Cổ Phần TM & DV Du Lịch Đất Việt	Thương lượng Thương lượng
Tuyển Chuyên Viên Phòng Dịch Vụ Chứng Khoán	CÔNG TY TNHH MÁY NÔNG NGHIỆP CAO ĐẠT	9.000.000 VND / tháng
NHÂN VIÊN KINH DOANH KHU VỰC MIỀN BẮC	Công Ty TNHH Hoàng Kim Minh	10.000.000 VND / tháng
Tuyển Lái Xe Tài Lượng Chế Độ Hấp Dẫn Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp V&V	13.500.000 VND / tháng
Tuyển Dụng Lái Xe Tài Lượng Chế Độ Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng
Tuyển Gấp Lái Xe Tài Lượng Thường Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng
Tuyển Gấp Lái Xe Tài Lượng Thường Cao Và Phụ Xe Giao Hàng	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	13.500.000 VND / tháng

Hình 12: Kết quả thu được.



## 5.5 Tổng kết bước làm sạch dữ liệu thu được

- Sau quá trình làm sạch dữ liệu từ unit đến program, tiến hành lọc thiếu và trùng lặp dữ liệu, nhóm thu được kết quả:  $\approx 25000$  job items có đầy đủ các trường giá trị và không bị trùng lặp.
- Các record thu được tương đối sạch, thỏa mãn được ràng buộc đã nêu ở phần 5.2

job_title	company	salary	location
Project Manager ( Branch Transformation Project) Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	3.500 USD / tháng	Hà Nội
.Net Developer	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	1.500 USD / tháng	Hà Nội
Project Manager Quản Lý Dự Án Khối Quản Trí Rủi Ro Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	3.500 USD / tháng	Hà Nội
Information Security Specialist – Application Security Engineer Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	2.500 USD / tháng	Hà Nội
Chuyên Viên Cao Cấp Quản Lý Dữ Liệu Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	1.500 USD / tháng	Hà Nội
Giám Đốc Cao Cấp Quan Hệ Khách Hàng Corporate Banking Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	4.000 USD / năm	Hà Nội
Development Expert Collection System Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	2.000 USD / tháng	Hà Nội
Chuyên Viên Hỗ Trợ Dịch Vụ Công Nghệ Thông Tin Hà Nội	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	20.000.000 VND / năm	Hà Nội
Senior Developer (Crm, Erp) HCM	CÔNG TY TNHH THIẾT BỊ DẪN DUNG THUẬN PHONG	30.000.000 VND / tháng	Hồ Chí Minh
Trưởng nhóm kinh doanh	CÔNG TY TNHH DV ; TM ROYAL FINANCE	15.000.000 VND / tháng	Bắc Ninh
Chuyên viên nhân sự	Meey Land	10.000.000 VND / tháng	Hà Nội
Giám đốc kinh doanh trực tiếp	Meey Land	15.000.000 VND / tháng	Hà Nội
THỰC TẬP SINH TEAM BUILDING	Công Ty Cổ Phần ĐT TM DV Du Lịch Đất Việt	Thương lượng	Hồ Chí Minh
Tuyển Lái Xe Tài Lương Chế Độ Hấp Dẫn Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Bắc Ninh
Tuyển Dụng Lái Xe Tài Lương Chế Độ Cao Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Bắc Giang
Tuyển Cấp Lái Xe Tài Lương Thủ Thường Cao Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hà Nam
Tuyển Cấp Lái Xe Tài Lương Thủ Thường Cao Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hà Nam
Tuyển Cấp Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hà Nam
Tuyển Cấp Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hai Dương
Tuyển Cấp Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hoà Bình
Nhân viên thiết kế 3D nội thất	Công ty TNHH VMD Việt Nam	815 triệu / tháng	Hà Nội
Tuyển NV Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng Lương Cao	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Hưng Yên
Tuyển NV Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng Lương Cao	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Lạng Sơn
Tuyển NV Lái Xe Tài (nhận bằng lái mới có bổ túc ) Và Phụ Xe Giao Hàng Lương 9. 13 Triệu / tháng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Nam Định
Tuyển NV Lái Xe Tài Giao Hàng Trong Tỉnh Và Phụ Xe Giao Hàng Lương 9. 13 Triệu / tháng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Ninh Bình
Tuyển Cấp Lái Xe Tài Thủ Thung Kín HuynDai Và Phụ Xe Giao Hàng Lương 9. 13 Triệu / tháng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Phú Tho
Tuyển Cấp Lái Xe Tài Thủ Thung Kín HuynDai Và Phụ Xe Giao Hàng Trong Tỉnh	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Quảng Ninh
Tuyển Cấp Lái Xe Tài(bổ túc lái mới) Và Phụ Xe Giao Hàng Trong Tỉnh	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Thái Bình
Tuyển Cấp Lái Xe Tài Lương Thủ Thường Cao Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Thái Nguyên
Tuyển NV Lái Xe Tài Cố Nhận Đào Tạo Và Phụ Xe Giao Hàng	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Vĩnh Phúc
Tuyển NV Lái Xe Tài Cố Nhận Đào Tạo Và Phụ Xe Giao Hàng Lương Cao	Công Ty Cổ Phần Thương Mại Tổng Hợp Và Vận Tải Minh Thành	13.500.000 VND / tháng	Tuyên Quang
Học Viện Âm Nhạc SEAMI Tuyển Dụng Nhân Viên Giáo Vụ Fulltime 2020	CÔNG TY CỔ PHẦN ĐẦU TƯ VÀ PHÁT TRIỂN GIÁO DỤC NGHỆ THUẬT ĐÉ	5.500.000 VND / tháng	Hồ Chí Minh
QA Engineers (Automation and Manual)	EVIZI LLC	Thương lượng	Đà Nẵng

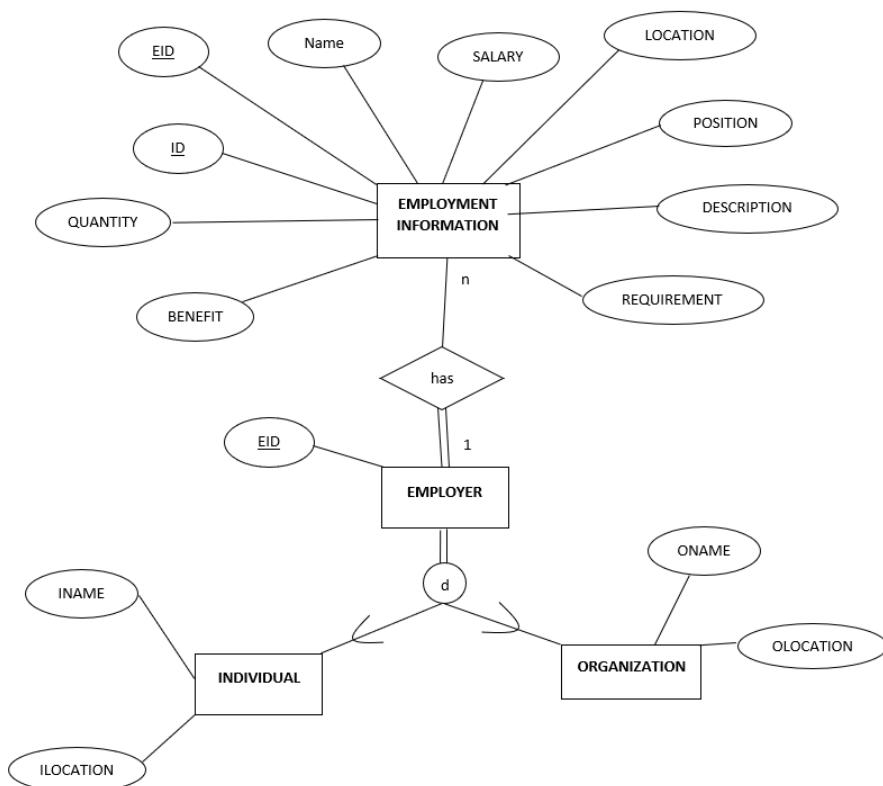
Hình 13: Dataset thu được sau quá trình làm sạch

## 6 Mô hình hóa dữ liệu và xây dựng Hệ cơ sở dữ liệu để lưu trữ

- Sau quá trình làm sạch dữ liệu, nhóm tiến hành mô hình hóa dữ liệu và lưu chúng vào Hệ cơ sở dữ liệu.
- Nhóm sử dụng xampp để giả lập một server tại local, sử dụng DBMS MySQL trên Maria DB được cung cấp sẵn trên xampp/phpmyadmin để lưu trữ tập dữ liệu.

### 6.1 Mô hình hóa dữ liệu

#### 6.1.1 Sơ đồ quan hệ thực thể (ER Diagram)



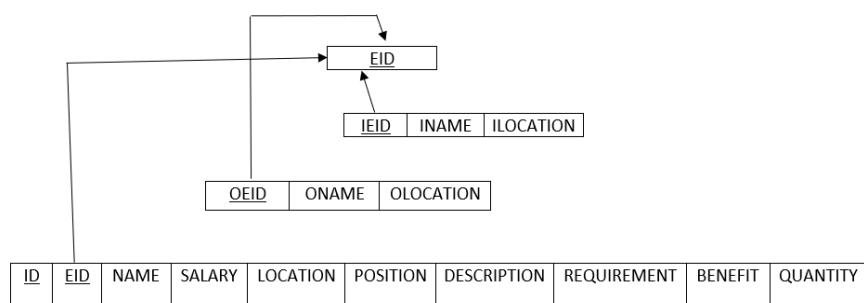
Hình 14: Sơ đồ quan hệ thực thể - ERD



### 6.1.2 Mô tả

- Bộ dữ liệu thông tin tuyển dụng gồm 2 đối tượng chính là nhà tuyển dụng và thông tin tuyển dụng, trong đó nhà tuyển dụng có thể là tổ chức hoặc cá nhân.
- Tổ chức có thể là một công ty, tập đoàn. Cần lưu trữ các thông tin bao gồm mã nhà tuyển dụng, tên tổ chức, địa chỉ.
- Cá nhân cần lưu trữ các thông tin bao gồm mã nhà tuyển dụng, tên, địa chỉ.
- Về thông tin tuyển dụng, cần lưu trữ các thông tin về việc làm cần tuyển như mã số, tên việc làm, mô tả công việc, mức lương, nơi làm việc, yêu cầu, lợi ích, số lượng...

### 6.1.3 Ánh xạ lược đồ liên kết thực thể sang lược đồ dữ liệu quan hệ - Mapping



Hình 15: Ánh xạ lược đồ

## 6.2 Xây dựng hệ cơ sở dữ liệu để lưu trữ

- Sau khi kết nối với MySQL phpmyadmin (đã đề cập ở mục 3.4), nhóm sử dụng các method của Python để tạo các câu truy vấn (query) thực hiện việc insert từng record vào Hệ cơ sở dữ liệu.



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

```
# else:
#     NAME = -1
if item.get("company") is None:
    NAME = -1
else:
    NAME = item.get("company")

LOCATION = item.get("location")
POSITION = item.get("position")
DESCRIPTION = item.get("job_description")
BENEFIT = item.get("benefit")
SALARY = item.get("salary")
REQUIREMENT = item.get("job_requirement")
QUANTITY = item.get("quantity")

cursor.execute(
    "insert into EID(EID) values(%s)",
    (EID))

cursor.execute(
    "insert into Individual(IEID, INAME, ILOCATION) values(%s, %s, %s)",
    (IEID, INAME, ILOCATION))

cursor.execute(
    "insert into Organization(OEID, ONAME, OLOCATION) values(%s, %s, %s)",
    (OEID, ONAME, OLOCATION))
```

Hình 16: Các method của Python để tạo câu truy vấn và hiện thực thay đổi trên MySQL

- Kết quả thu được: ta có Hệ cơ sở dữ liệu như sau:

+ Tùy chọn												
	ID	EID	NAME	SALARY	LOCATION	POSITION	DESCRIPTION	REQUIREMENT	BENEFIT	QUANTITY		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000001	EID000001	Công ty TNHH Khoa Học Và Kỹ Thuật Olympic	8.000.000 15.000.000 / tháng	Bán hàng, , phản/Chẩn doán, , Kinh doanh/Phu ...	Kinh doanh, tư vấn cho khách hàng các sản phẩm v... Quản trị k...	Trình độ Đại học/Cao đẳng chuyên ngành ... ...	Thu nhập: Lương cơ bản + Trợ cấp công tác + Điện ... ...	1		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000002	EID000002	Sulunam Việt Nam	2.500 USD / tháng	Hà Nội	CNTT/ Viễn thông	Configuration and management of website architectu...	Because you want to be part of a multowinning awar...	1		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000003	EID000003	Citigo Software	2.000 USD / tháng	Hồ Chí Minh	CNTT/ Viễn thông	Ban sẽ quyết định frontend của m... trong những sản...	Mức lương khởi diểm hấp dẫn, cạnh tranh, tương xứng...	1		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000004	EID000004	Citigo Software	2.000 USD / tháng	Hồ Chí Minh	CNTT/ Viễn thông	Tham gia phát triển sản phẩm phần mềm quản lý bán ...	Citigo tự hào luôn là 1 môi trường làm việc CÔNG B...	1		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000005	EID000005	Biểu trưng của người bán NVG Technology	2.500 USD	Hồ Chí Minh	CNTT/ Viễn thông	You are our specialist for data provision', 'Prov...	Good command of English communication (both written...	1		
<input type="checkbox"/>	<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	ID000006	EID000006	Biểu trưng của người bán NVG Technology	2.000 USD / tháng	Hồ Chí Minh	CNTT/ Viễn thông	Responsible for test design, test plan and test ma...	Good skills in English communication (both written...	1		

Hình 17: Bảng employee\_information



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

The screenshot shows a MySQL database interface with the following details:

- Table: individual
- Columns: IEDID, INAME, ILOCATION
- Data:

IEDID	INAME	ILocation
EID000001	Nhân Viên Kinh Doanh Thiết Bị Y Tế	Hồ Chí Minh
EID000002	07 PHP Developers (Magento, JavaScript)	Hà Nội
EID000003	Senior Frontend Dev (ReactJS/VueJS)	Hồ Chí Minh
EID000004	Senior .NET Dev (ASP.NET, C#)	Hồ Chí Minh
EID000005	Senior Data Engineer	Hồ Chí Minh
EID000006	02 Lead/Senior QA Engineers	Hồ Chí Minh
EID000007	Engineering Manager	Hồ Chí Minh
EID000008	Senior Business Analyst (eCommerce)	Hồ Chí Minh
EID000009	FullStack Developer (Python/Ruby/Golang)	Hồ Chí Minh
EID000010	Chuyên Viên Công Nghệ Thông Tin	Hà Nội
EID000011	IT / Project Support (Networking / SQL)	Hà Nội
EID000012	Project Manager (Branch Transformation Project) ...	Hà Nội
EID000013	.Net Developer	Hà Nội
EID000014	Project Manager: Quản Lý Dự Án Khối Quản Trị R&D	Hà Nội
EID000015	Information Security Specialist – Application Secu...	Hà Nội
EID000016	Chuyên Viên Cao Cấp Quản Lý Dữ Liệu Hà Nội	Hà Nội
EID000017	Giám Đốc Cao Cấp Quản Hỗn Khách Hàng Corporate Ban...	Hà Nội
EID000018	Development Expert - Collection System - Hà Nội	Hà Nội

Hình 18: Bảng individual

The screenshot shows a MySQL database interface with the following details:

- Table: organization
- Columns: OEDID, ONAME, OLOCATION
- Data:

OEDID	ONAME	OLocation
EID000001	Công ty TNHH Khoa Học Và Kỹ Thuật Olympic	Hồ Chí Minh
EID000002	Sutunam Việt Nam	Hà Nội
EID000003	Citigo Software	Hồ Chí Minh
EID000004	Citigo Software	Hồ Chí Minh
EID000005	Biểu trưng của người bán NVG Technology	Hồ Chí Minh
EID000006	Biểu trưng của người bán NVG Technology	Hồ Chí Minh
EID000007	Biểu trưng của người bán NVG Technology	Hồ Chí Minh
EID000008	Dat Xanh Group	Hồ Chí Minh
EID000009	OnPoint Vietnam	Hồ Chí Minh
EID000010	NGÂN HÀNG TMCP NGOẠI THƯƠNG VIỆT NAM VIETCOMBANK	Hà Nội
EID000011	MICROTEC VIETNAM CO., LTD	Hà Nội
EID000012	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội
EID000013	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội
EID000014	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội
EID000015	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội
EID000016	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội
EID000017	Ngân hàng TMCP Việt Nam Thịnh Vượng VPBank	Hà Nội

Hình 19: Bảng organization



### 6.3 Tổng kết bước mô hình hóa dữ liệu và xây dựng Hệ cơ sở dữ liệu

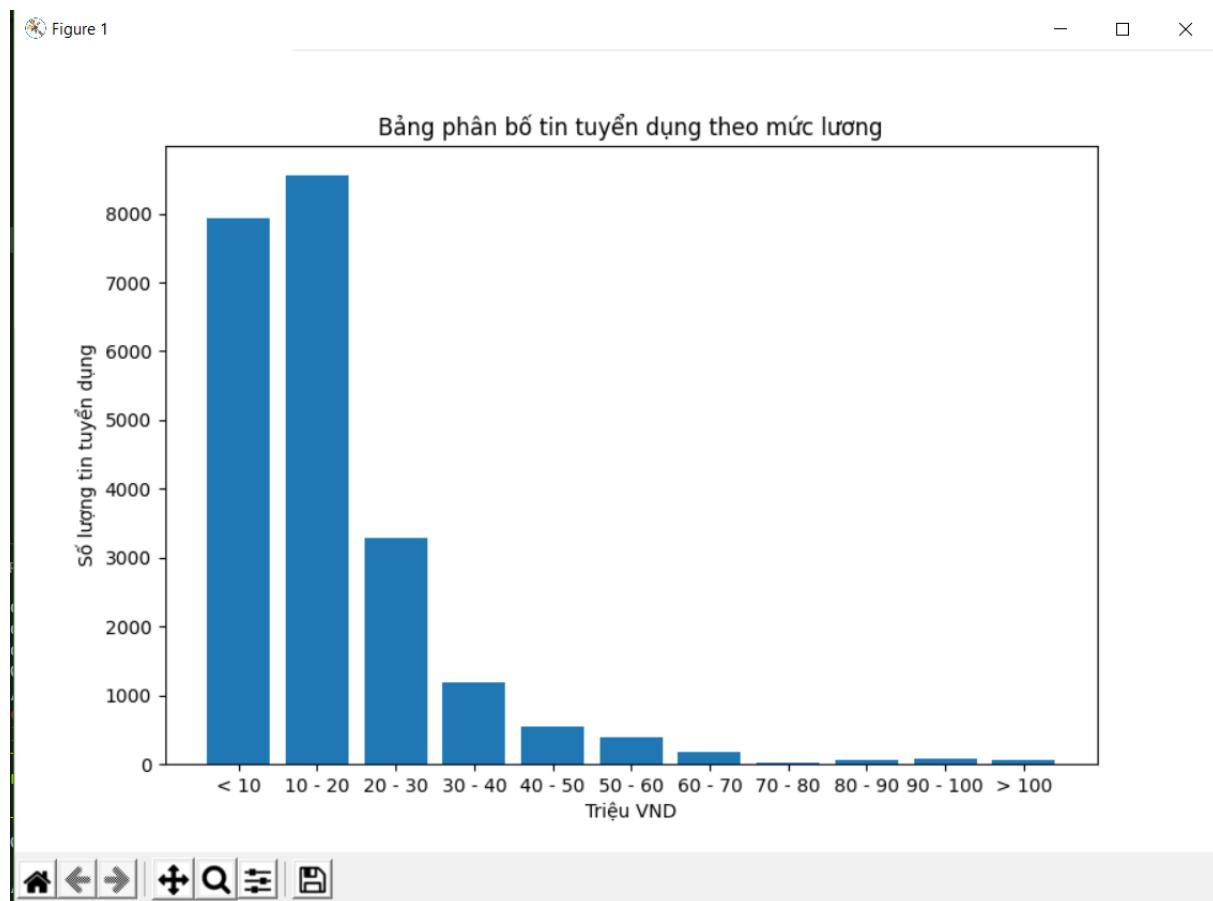
- Sau quá trình mô hình hóa dữ liệu, nhóm đã phân dữ liệu thành 3 bảng: employee\_information, individual, organization
- Hoàn thành việc cập nhật các record sau khi làm sạch lên Hệ cơ sở dữ liệu MySQL phpmyadmin.



## 7 Hiện thực phân tích dữ liệu thu thập được từ các trang

- Sau khi trải qua quá trình làm sạch dữ liệu từ các trang web, nhóm tiến hành phân tích dữ liệu thu thập được. Bên cạnh đó, nhóm sử dụng công cụ vẽ biểu đồ của thư viện matplotlib trong python để trực quan hóa dữ liệu cũng như có cái nhìn tổng quan hơn.
- Dựa vào tập dữ liệu thu thập được, nhóm tiến hành phân tích dữ liệu theo các mục sau:

### 7.1 Phân tích thông tin tuyển dụng theo giá trị tiền lương

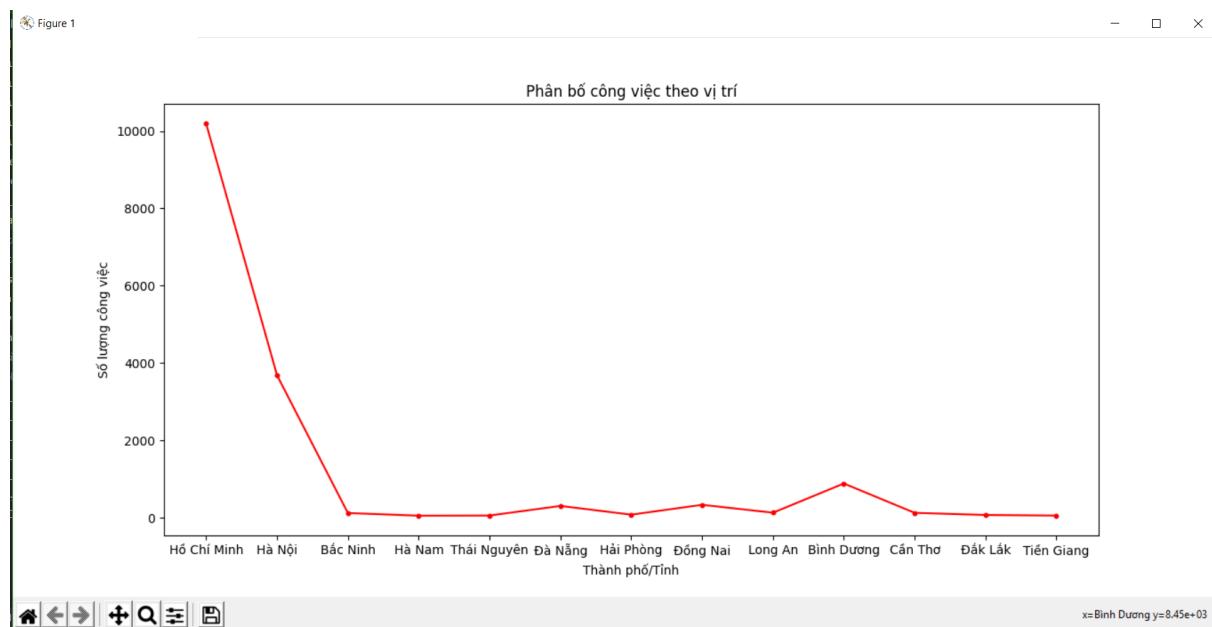


Hình 20: Sơ đồ phân bố tin tuyển dụng theo giá trị tiền lương



- Theo như sơ đồ ở trên, mức lương mà các nhà tuyển dụng đưa ra nằm chủ yếu trong phân khúc từ 10.000.000 VND - 20.000.000 VND, tiếp theo là phân khúc dưới 10.000.000 VND. Các phân khúc tiền lương còn lại trải dài đều và thấp hơn hẳn các phân khúc.
- Quan sát nhóm đối tượng tuyển dụng nằm trong phân khúc tiền lương 10.000.000 VND - 20.000.000 VND là những đối tượng cho tri thức: kỹ sư, kế toán, chuyên viên thiết kế, chuyên viên sales, ...
- Các đối tượng tuyển dụng nằm trong phân khúc tiền lương dưới 10.000.000 VND là những đối tượng không cần trình độ chuyên môn cao, hoặc tuyển dụng theo dạng hợp đồng như: bảo vệ, bảo mẫu, quét dọn, ...
- Đặc biệt, các đối tượng có mức lương trên 20.000.000 VND là những công việc yêu cầu trình độ chuyên môn cao, cũng như kinh nghiệm và bằng cấp như giám đốc bộ phận, quản lý phòng ban, senior IT, ...

## 7.2 Phân tích thông tin tuyển dụng theo vị trí địa lý



Hình 21: Sơ đồ phân bố tin tuyển dụng theo vị trí địa lý

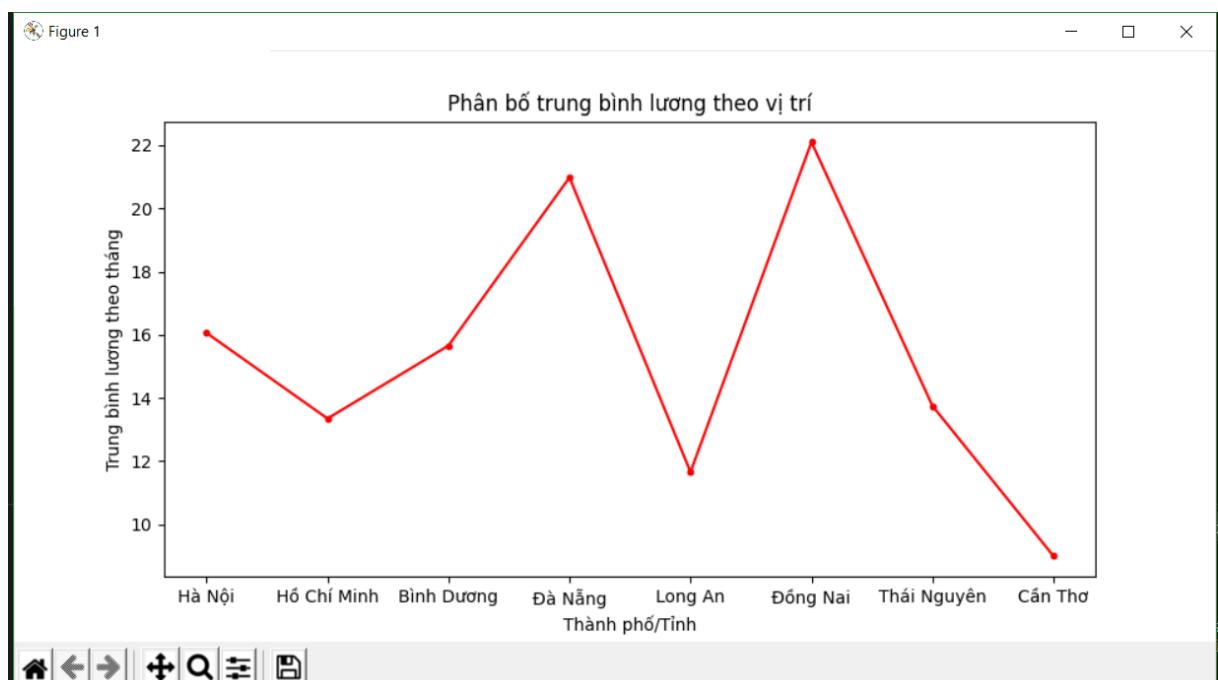
- Theo như sơ đồ ở trên, hai trung tâm kinh tế lớn nhất của cả nước là Thành phố Hồ Chí Minh và Hà Nội có mức tuyển dụng cao nhất, đặc biệt là Thành



phố Hồ Chí Minh với trên 10000 tin tuyển dụng. Điều này có thể lý giải vì ngoài các tin tuyển dụng của công ty còn các tin tuyển dụng tư nhân như: phục vụ quán, bảo mẫu, người giúp việc, ...

- Bên cạnh đó, các trung tâm công nghiệp như Đồng Nai, Bình Dương, Quảng Ninh, Đà Nẵng, ... có mức tuyển dụng đứng thứ hai sau 2 trung tâm kinh tế của cả nước. Các công việc ở các trung tâm công nghiệp phần lớn là làm nhân viên cho các xí nghiệp, nhà máy của các công ty có xưởng sản xuất tại Việt Nam.
- Ngoài ra, các tỉnh thành khác có mức tuyển dụng tương đối đồng đều so với nhau.

### 7.3 Phân tích giá trị tiền lương trung bình theo vị trí địa lý



**Hình 22:** Sơ đồ phân bố tiền lương trung bình theo vị trí địa lý

- Theo sơ đồ ở trên, hai trung tâm kinh tế lớn nhất của cả nước là Thành phố Hồ Chí Minh và Hà Nội tuy có mức tuyển dụng cao, nhưng mặt bằng lương không cao so với các trung tâm công nghiệp như Đồng Nai, Đà Nẵng,...



Có thể quan sát thấy rất nhiều tin tuyển dụng của Thành phố Hồ Chí Minh và Hà Nội có mức lương < 10.000.000 VND, do tư nhân và đặc thù nghề nghiệp (VD: phục vụ, bảo mẫu, ...)

- Các trung tâm công nghiệp như Đồng Nai, Bình Dương hay trung tâm du lịch như Đà Nẵng tuy số lượng công việc ít như đặc thù công việc nên giá trị lương nằm chủ yếu trong phân khúc 10.000.000 - 20.000.000 VNĐ.
- Các nơi khác thì mức lương không ổn định do sự đa dạng về tin tuyển dụng cũng như đặc thù nghề nghiệp yêu cầu

#### 7.4 Tổng kết bước phân tích dữ liệu

- Nhóm đã sử dụng ngôn ngữ Python để phân tích dữ liệu, bên cạnh đó áp dụng thư viện **pandas** từ pip để đọc dữ liệu từ file xlsx và **matplotlib.pyplot** để vẽ đồ thị minh họa.
- Qua việc phân tích dữ liệu trên, ta thấy được thi trường việc làm một các tổng quan hơn, các địa điểm có nhu cầu tuyển dụng ra sao, mức lương dao động như thế nào và đặc thù nghề nghiệp tại mỗi nơi.



## 8 Đề xuất ứng dụng

### 8.1 Ý tưởng ứng dụng

Thiết kế website giới thiệu việc làm:

- Trang chủ
- Giới thiệu
- Thông tin dịch vụ - việc làm
- Mô tả chi tiết việc làm
- Liên hệ

Các tính năng cơ bản:

- Khách: xem các thông tin public trên trang web, cho phép đăng ký, đăng nhập.
- Thành viên (sau khi đã đăng nhập): cho phép thực hiện một số hàm chức năng cơ bản: thay đổi thông tin cá nhân, mật khẩu, tìm kiếm thông tin tuyển dụng, xem các xu hướng hiện tại.
- Quản trị viên: hiện thực các tính năng quản lý:
  - Quản lý thành viên (xem thông tin, xóa thành viên,...)
  - Tính năng quản lý (xem, thêm, sửa, xoá) các tài nguyên của ứng dụng web như thông tin tuyển dụng, các công việc đang có nhu cầu lớn,...
  - Hiện thực phân trang hiển thị cho các tính năng quản lý.

Thực hiện kiểm tra dữ liệu đầu vào (sử dụng cả kiểm tra bằng javascript (client side) và PHP (server side).

Tính năng tìm kiếm tài nguyên đơn giản trên trang web.

### 8.2 Công cụ hiện thực ứng dụng

Các công cụ hiện thực ứng dụng:

- HTML (viết tắt của từ Hypertext Markup Language, hay là "Ngôn ngữ Đánh dấu Siêu văn bản") là một ngôn ngữ đánh dấu được thiết kế ra để tạo nên các trang web trên World Wide Web. Cùng với CSS và JavaScript, HTML là một trong những ngôn ngữ quan trọng trong lĩnh vực thiết kế website.



HTML được định nghĩa như là một ứng dụng đơn giản của SGML và được sử dụng trong các tổ chức cần đến các yêu cầu xuất bản phức tạp. HTML đã trở thành một chuẩn mực của Internet do tổ chức World Wide Web Consortium (W3C) duy trì. Phiên bản chính thức mới nhất của HTML là HTML 4.01 (1999). Sau đó, các nhà phát triển đã thay thế nó bằng XHTML. Hiện nay, phiên bản mới nhất của ngôn ngữ này là HTML5.

- CSS - được dùng để miêu tả cách trình bày các tài liệu viết bằng ngôn ngữ HTML và XHTML.[1] Ngoài ra ngôn ngữ định kiểu theo tầng cũng có thể dùng cho XML, SVG, XUL. Các đặc điểm kỹ thuật của CSS được duy trì bởi World Wide Web Consortium (W3C). Thay vì đặt các thẻ quy định kiểu dáng cho văn bản HTML (hoặc XHTML) ngay trong nội dung của nó, bạn nên sử dụng CSS.
- JavaScript, theo phiên bản hiện hành, là một ngôn ngữ lập trình thông dịch được phát triển từ các ý niệm nguyên mẫu. Ngôn ngữ này được dùng rộng rãi cho các trang web (phía người dùng) cũng như phía máy chủ (với Nodejs). Nó vốn được phát triển bởi Brendan Eich tại Hãng truyền thông Netscape với cái tên đầu tiên Mocha, rồi sau đó đổi tên thành LiveScript, và cuối cùng thành JavaScript. Giống Java, JavaScript có cú pháp tương tự C, nhưng nó gần với Self hơn Java. .js là phần mở rộng thường được dùng cho tập tin mã nguồn JavaScript.
- AJAX là một nhóm các công nghệ phát triển web được sử dụng để tạo các ứng dụng web động hay các ứng dụng giàu tính Internet (rich Internet application). Từ Ajax được ông Jesse James Garrett đưa ra và dùng lần đầu tiên vào tháng 2 năm 2005 để chỉ kỹ thuật này, mặc dù các hỗ trợ cho Ajax đã có trên các chương trình duyệt từ 10 năm trước. Ajax là một kỹ thuật phát triển web có tính tương tác cao bằng cách kết hợp nhiều ngôn ngữ.
- Thư viện Wow.js là một thư viện giúp làm hiệu ứng cho website tuyệt vời nhất mà mình từng sử dụng trong xuất chặng đường phát triển website của mình. Nó giúp cho website của bạn trở lên sinh động hơn, đẹp mắt hơn và giúp cho người dùng chú ý đến website của bạn hơn. Wow.js không làm ảnh hưởng đến hiệu năng trong website của bạn và cũng không làm hạ điểm khi bạn kiểm tra tốc độ bằng google developer speed
- phpMyAdmin là một công cụ nguồn mở miễn phí được viết bằng PHP dự định để xử lý quản trị của MySQL thông qua một trình duyệt web. Nó có thể thực hiện nhiều tác vụ như tạo, sửa đổi hoặc xóa bỏ cơ sở dữ liệu, bảng, các trường hoặc bản ghi; thực hiện báo cáo SQL; hoặc quản lý người dùng và cấp phép.



The screenshot shows the phpMyAdmin interface for the 'jobdata' database. The left sidebar lists various databases and tables. The main area displays the 'newjob' table with the following data:

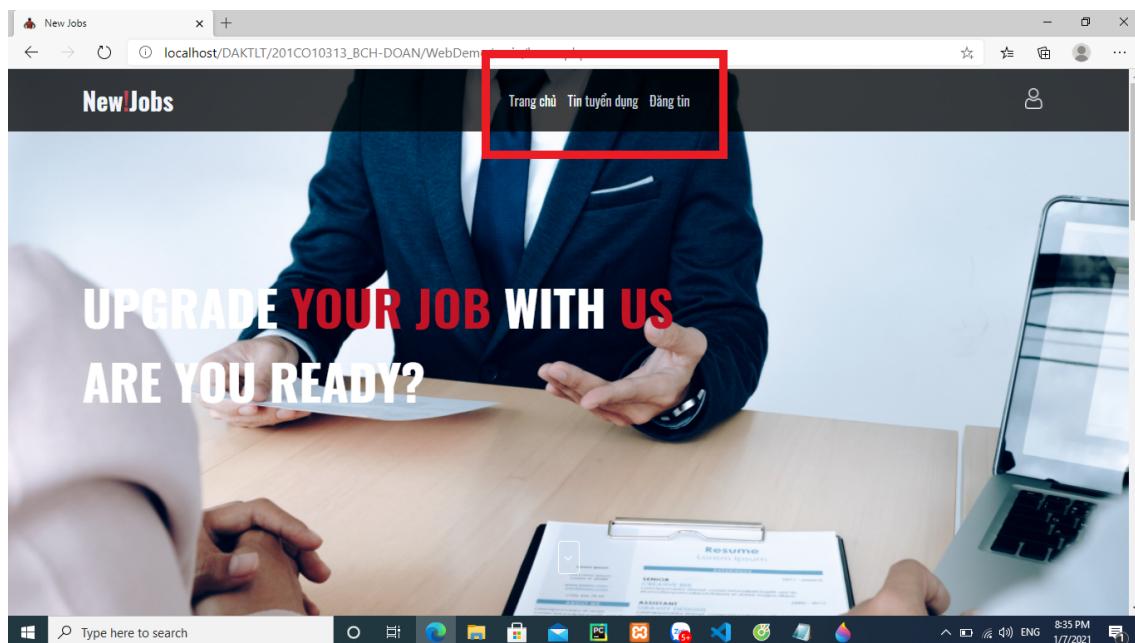
	id	name	salary	location	position	quantity	description	requirement	benefit
	2	Công ty Bất động sản	9 triệu	Hồ Chí Minh	nhan vien bán hàng	6	Bán lô đất	có bằng trung học trở lên	tuần nghỉ 4 ngày
	3	Công ty C	12 triệu	Hà Nội	Nhân viên	3	Nhập liệu	Tốt nghiệp phổ thông	Bảo hiểm

Hình 23: Database trong phpMyAdmin

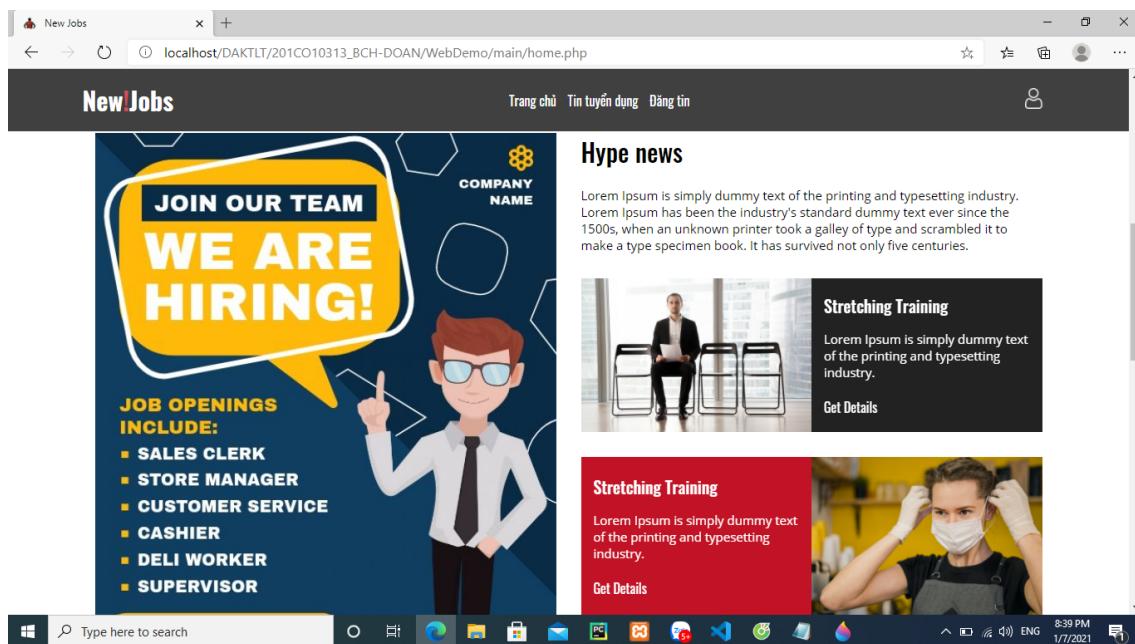
## 8.3 Web tuyển dụng

Link video mô tả: <https://youtu.be/wYZkLOuQ-G4>

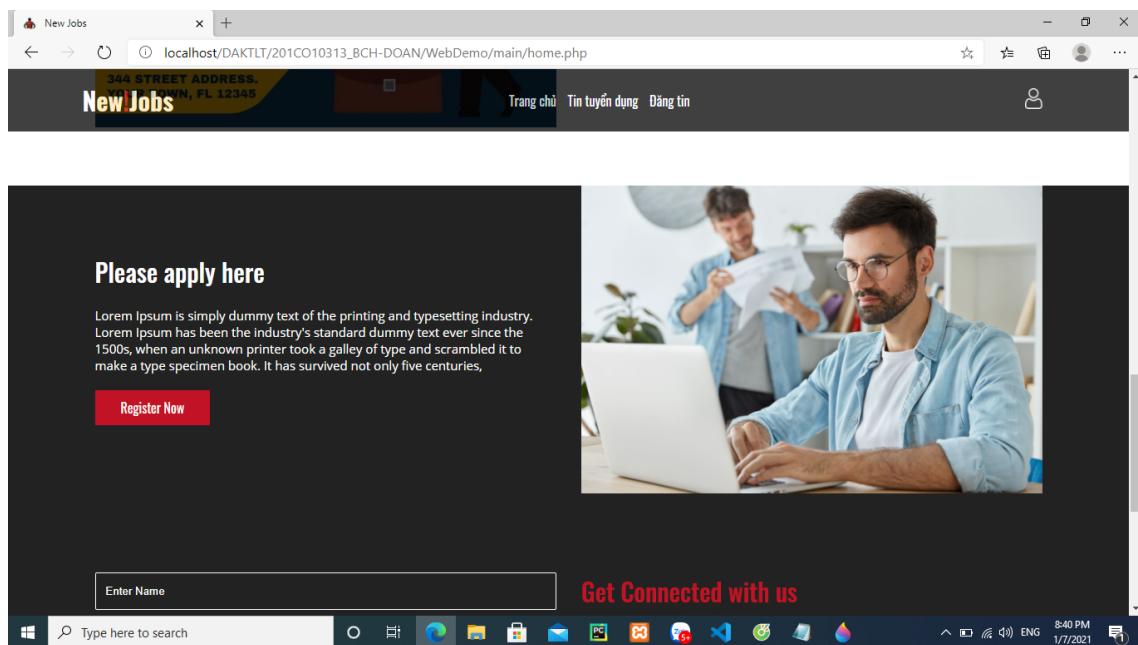
### 8.3.1 Trang chủ



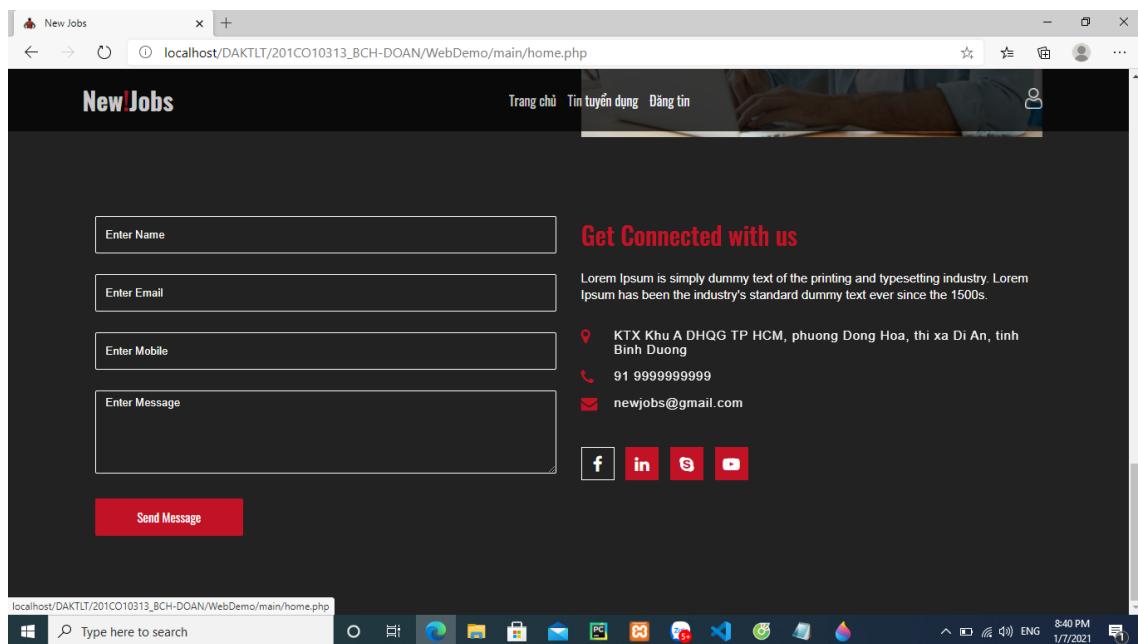
Hình 24: Trang chủ của ứng dụng



Hình 25: Trang chủ của ứng dụng



Hình 26: Trang chủ của ứng dụng



Hình 27: Liên hệ của ứng dụng



### 8.3.2 Thông tin công việc

Trang thông tin công việc bao gồm:

- Thanh tìm kiếm.
- Số lượng công việc.
- Các công việc.
- Chuyển trang.

The screenshot shows a web browser window titled "NewJobs" with the URL "localhost/DAKTLT/201CO10313\_BCH-DOAN/WebDemo/main/newjob.html". The page has a dark header with the title "NewJobs" and navigation links for "Trang chủ", "Tin tuyển dụng", and "Đăng tin". Below the header is a search bar with three input fields: "Tên công ty", "Vị trí tuyển dụng", and "Địa điểm", followed by a search button. A link "32533 việc làm" is visible. The main content area displays four job listings:

- 1. Công ty TNHH Khoa Học Và Kỹ Thuật Olympic**  
Lương : 8.000.000 15.000.000 / tháng  
Vị trí : Bán hàng, 'phẩm/Chẩn đoán,' Kinh doanh/Phụ trách KH, 'Marketing/tiếp thị,'  
Hồ Chí Minh
- 2. Sutunam Việt Nam**  
Lương : 2.500 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hà Nội
- 3. Citigo Software**  
Lương : 2.000 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hồ Chí Minh
- 4. Citigo Software**  
Lương : 2.000 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hồ Chí Minh

At the bottom of the browser window, the Windows taskbar is visible with icons for File Explorer, Edge, File, Mail, Photos, Task View, and others. The system tray shows the date and time as "8:42 PM 1/7/2021".

Hình 28: Trang thông tin của ứng dụng



The screenshot shows a web browser window titled 'NewJobs' with a search bar at the top. The search bar contains fields for 'Tên công ty', 'Vị trí tuyển dụng', and 'Địa điểm', with a red box highlighting the search button. Below the search bar, a message '33533 việc làm' is displayed. The main content area lists four job postings:

- 1. Công ty TNHH Khoa Học Và Kỹ Thuật Olympic**  
Lương : 8.000.000 15.000.000 / tháng  
Vị trí : Bán hàng, 'pharm/Chẩn đoán,' 'Kinh doanh/Phụ trách KH,' 'Marketing/tiếp thị,'  
Hồ Chí Minh
- 2. Sutunam Việt Nam**  
Lương : 2.500 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hà Nội
- 3. Citigo Software**  
Lương : 2.000 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hồ Chí Minh
- 4. Citigo Software**  
Lương : 2.000 USD / tháng  
Vị trí : CNTT/ Viễn thông  
Hồ Chí Minh

The taskbar at the bottom of the screen shows various application icons.

Hình 29: Trang thông tin của ứng dụng

The screenshot shows a web browser window titled 'NewJobs' with a search bar at the top. The main content area lists four company entries:

- 97. CÔNG TY TNHH THIẾT BỊ GIÁO DỤC BAVICO**  
Lương : 6.000.000 VND / tháng  
Vị trí : Khác  
Hồ Chí Minh
- 98. CÔNG TY TNHH THIẾT BỊ GIÁO DỤC BAVICO**  
Lương : 6.000.000 VND / tháng  
Vị trí : Bán hàng  
Hồ Chí Minh
- 99. CÔNG TY TNHH THIẾT BỊ GIÁO DỤC BAVICO**  
Lương : 6.000.000 VND / tháng  
Vị trí : Dịch vụ khách hàng  
Hồ Chí Minh
- 100. CÔNG TY TNHH THIẾT BỊ GIÁO DỤC BAVICO**  
Lương : 6.000.000 VND / tháng  
Vị trí : Kế toán/kiem toán  
Hồ Chí Minh

At the bottom of the page, there are navigation buttons: '< back' (left), 'Trang 2' (center), and 'next >' (right), all of which are highlighted with red boxes. The taskbar at the bottom of the screen shows various application icons.

Hình 30: Trang thông tin của ứng dụng



Khi nhấn tìm kiếm với các thông tin như bất động sản, nhân viên, Hồ Chí Minh.

New Jobs    +  
localhost/DAKTLT/201CO10313\_BCH-DOAN/WebDemo/main/newjob.html

NewJobs    Trang chủ    Tin tuyển dụng    Đăng tin   

bất động sản    nhân viên    Hồ chí minh   

**44 việc làm**

**1. Công Ty Cổ Phần Dịch Vụ Bất Động Sản Vietin House**  
Lương : 5.000.000 - 99.000.000 đ (tương khoản)  
Vị trí : Nhân viên kinh doanh      Phường 10, Quận Gò Vấp, TP Hồ Chí Minh

**2. CÔNG TY TNHH BẤT ĐỘNG SẢN KHANG THỊNH PHÁT**  
Lương : 5.000.000 - 20.000.000 đ/tháng  
Vị trí : Nhân viên kinh doanh      Phường 25, Quận Bình Thạnh, TP Hồ Chí Minh

**3. Công Ty Xây Dựng Đầu Tư Và Phát Triển Bất Động Sản Đại Vũ**  
Lương : 4.000.000 - 20.000.000 đ/tháng  
Vị trí : Nhân viên kinh doanh      Phường 12, Quận Tân Bình, TP Hồ Chí Minh

**4. CTY CỔ PHẦN ĐẦU TƯ VÀ PHÁT TRIỂN BẤT ĐỘNG SẢN SAIGONGOLDLAND-HƯNG ĐẠO VƯƠNG**  
Lương : 15.000.000 - 30.000.000 đ/tháng  
Vị trí : Nhân viên kinh doanh      Phường 2, Quận 3, TP Hồ Chí Minh

Type here to search    ENG 9:19 PM 1/7/2021

Hình 31: Tìm kiếm công việc



### 8.3.3 Chi tiết công việc

Khi người dùng click vào title của 1 job bên trang "thông tin" thì sẽ được chuyển sang trang "chi tiết công việc" để xem thông tin chi tiết của công việc đó.

The screenshot shows a web browser window with the URL [localhost/DAKTLT/201CO10313\\_BCH-DOAN/WebDemo/main/datail.html](http://localhost/DAKTLT/201CO10313_BCH-DOAN/WebDemo/main/datail.html). The page has a dark header with 'NewJobs' and navigation links 'Trang chủ', 'Tin tuyển dụng', and 'Đăng tin'. A user icon is on the right. Below the header, there's a green button labeled 'Nộp đơn ứng tuyển'. The main content area displays a job listing for 'Citigo Software'. It includes details like 'Địa điểm: Hồ Chí Minh', 'Lương: 2.000 USD / tháng', 'Vị trí: CNTT/ Viễn thông', and 'Số lượng: 1'. A section titled 'Phúc lợi:' lists various benefits such as 'Mức lương khởi điểm hấp dẫn, cạnh tranh, tương xứng với năng lực và kinh nghiệm làm việc.', 'Được xét duyệt tăng lương định kỳ 2 lần/năm & lương tháng 13 theo kết quả công việc.', 'Thưởng dự án, thường xuyên.', 'Làm việc từ Thứ 2 – Thứ 6, ca làm việc linh hoạt.', 'Có hội được đào tạo và thăng tiến tốt trong công việc.', 'Lộ trình phát triển công việc rõ ràng.', 'Được hưởng các quyền lợi và chế độ theo luật quy định (Các ngày nghỉ lễ, BHXH, BHYT...)', 'Thưởng lễ, Tết, sinh nhật các chế độ đãi ngộ dành cho người thân, du lịch hàng năm.', 'Làm việc trong môi trường trẻ trung, năng động.', 'Được cung cấp máy tính cấu hình cao, trang thiết bị làm việc hiện đại.'

#### Mô tả công việc:

Bạn sẽ quyết định frontend của một trong những sản phẩm thành công của chúng tôi, tạo ra những tác động nổi bật sẽ trao quyền cho hàng ngàn Thương mại Điện tử., 'Để xuất giải pháp cho các yêu cầu để hoàn thành công việc', 'Đảm bảo UI / UX của các trang và phần mềm được phát triển với thiết kế pixelperfect', 'Xây dựng các components/modules có thể tái sử dụng cho các trang web responsive design', 'Hợp tác với nhóm phụ trợ chuyên gia để phát triển các plugin, tính năng, mới cho ứng dụng Thương mại điện tử phục vụ hàng ngàn người dùng'

#### Yêu cầu công việc:

Trải nghiệm với React stack (ReactJS, Redux, State Management, Functional Programming), 'Hiểu biết nâng cao về Javascript ES5 & ES6 +: Vue stack (Vuex, Vue Router, SPA, State Management, Functional Programming)', 'Kiến thức tốt với HTML5, CSS3, SASS / SCSS & framework UI xây dựng trên JavaScript, HTML, CSS', 'Thành thạo JavaScript, bao gồm thao tác DOM và ObjectJavaScript Model,' Có kinh nghiệm với công cụ Git / luồng Git: UI / UX & Hành vi người dùng: API RESTful', '1. Học vấn, kiến thức và trình độ chuyên môn: Tốt nghiệp cao đẳng, đại học; ưu tiên tốt nghiệp chuyên ngành CNTT, hoặc chuyên ngành liên quan', '2.Kinh nghiệm: Ít nhất 3 năm kinh nghiệm với Frontend bằng cách sử dụng ngôn ngữ lập trình', '3.Những kỹ năng cần thiết cho công việc', '4.Quan hệ công việc ( Bên trong công ty, bên ngoài công ty).

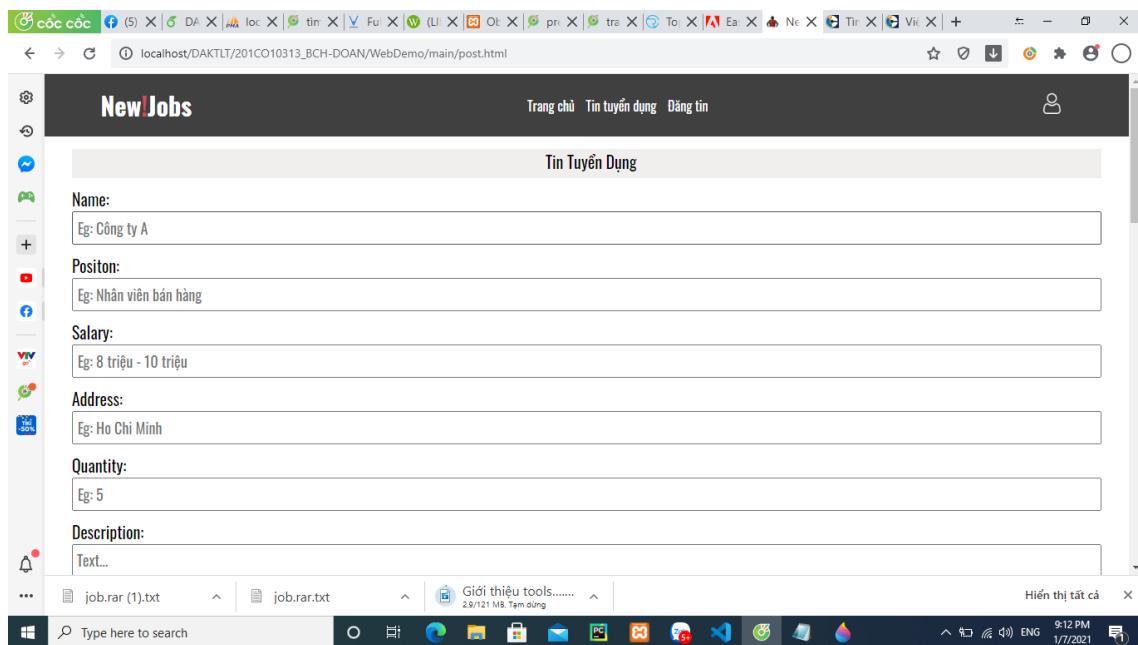
**Hình 32:** Trang chi tiết công việc



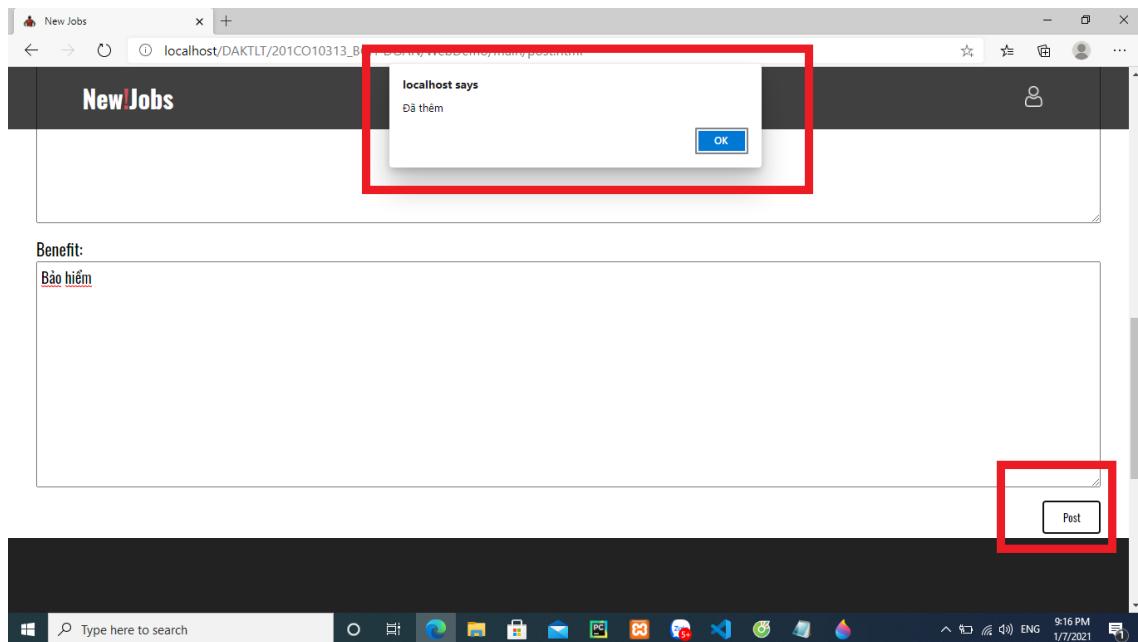
#### 8.3.4 Trang đăng tin tuyển dụng

Trang đăng tin tuyển dụng được thiết kế cho nhà tuyển dụng, để đăng thông tin tuyển dụng của công ty mình, thông tin sẽ được kiểm duyệt bởi admin và thêm vào trang "thông tin công việc" khi được duyệt.

Thông tin chưa duyệt sẽ nằm trong file newjob của database (mô tả bên dưới).



Hình 33: Trang đăng tin tuyển dụng



Hình 34: Trang đăng tin tuyển dụng



The screenshot shows the phpMyAdmin interface with the following details:

- Database:** jobdata
- Table:** newjob
- Table Structure:**

	id	name	salary	location	position	quantity	description	requirement	benefit
<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	2	Công ty Bất động sản	9 triệu	Hồ Chí Minh	nhân viên bán hàng	6	Bán lô đất có bằng trung học trở lên	tuần nghỉ 4 ngày	
<a href="#">Sửa</a> <a href="#">Chép</a> <a href="#">Xóa bỏ</a>	3	Công ty C	12 triệu	Hà Nội	Nhân viên	3	Nhập liệu	Tốt nghiệp phổ thông	Bảo hiểm
- SQL Query:** SELECT \* FROM `newjob`
- Toolbar Buttons:** Duyệt, Cấu trúc, SQL, Tìm kiếm, Chèn, Xuất, Nhập, Đặc quyền, Thảo tác, Theo dõi, Bấy

Hình 35: Trang đăng tin tuyển dụng



## 9 Tổng kết về dự án

### 9.1 Các kết quả đã đạt được

- Thông qua dự án lần này, nhóm đã biết cách sử dụng các công cụ của python để crawl dữ liệu từ các trang web. Đặc biệt đã giải quyết được vấn đề của trang <http://vietnamworks.com/> trong lần báo cáo trước.
- Sau đó, nhóm đã học được cách làm sạch dữ liệu, mô hình hóa dữ liệu sau khi làm sạch và đồng thời xây dựng Hệ cơ sở dữ liệu để lưu trữ dataset thu được.
- Bên cạnh đó, nhóm cũng tìm hiểu thêm cách phân tích dữ liệu sau khi đã làm sạch và xây dựng một ứng dụng để áp dụng tập dataset đã thu được.

### 9.2 Những điểm hạn chế

- Bên cạnh các kết quả đã đạt được, nhóm còn những hạn chế mà chưa xử lý được.
  - Quá trình làm sạch dữ liệu còn chưa áp dụng nhiều kĩ thuật để lọc dữ liệu một cách chính xác hơn.
  - Quá trình phân tích dữ liệu chỉ mới tập trung ở các trường dữ liệu dễ quan sát, dễ nhận biết. Các trường giá trị hỗn loạn vẫn chưa tìm ra cách để phân tích chính xác.
  - Ứng dụng tập dataset thu được chỉ dừng ở mức lọc dữ liệu, chưa thể áp dụng mô hình training dữ liệu để dự đoán một tập dữ liệu cho vào.

## 10 Link Project

Link project: [https://github.com/ThuanNguyen1210/201C010313\\_BCH-DOAN](https://github.com/ThuanNguyen1210/201C010313_BCH-DOAN)



## Tài liệu

- [1] Scrapy [online], from: <https://scrapy.org/>, viewed 22/11/2020.
- [2] Selenium [online], from: <https://www.selenium.dev/>, viewed 23/11/2020.
- [3] Beatifulsoup [online], from: <https://pypi.org/project/beautifulsoup4/>, viewed 24/11/2020.
- [4] Pandas [online], from: <https://pandas.pydata.org/docs/>, viewed 24/12/2020.
- [5] NumPy [online], from: <https://numpy.org/doc/stable/>, viewed 23/12/2020.
- [6] Matplotlib [online], from: <https://matplotlib.org/>, viewed 23/12/2020.