



# PROJECT 2

Nhóm: BCH DOAN

## Đồ án kĩ thuật lập trình

GVHD: Lưu Quang Huân

Đại Học Quốc Gia TP Hồ Chí Minh  
Trường Đại học Bách Khoa

# Thành viên

STT	Họ và tên	MSSV
1	Nguyễn Thế Viễn	1814764
2	Nguyễn Ngọc Thuấn	1814217
3	Bùi Hữu Đang	1811828
4	Phạm Đức Duy Anh	1810814
5	Huỳnh Ngọc Tân	1813952

# Nội dung

01

## Giới thiệu

Mục đích, yêu cầu

02

## Tìm hiểu và hiện thực

Crawl data là gì?

03

## Khó khăn gấp phải

Khó khăn trong quá trình thực hiện

04

## Đề xuất ứng dụng

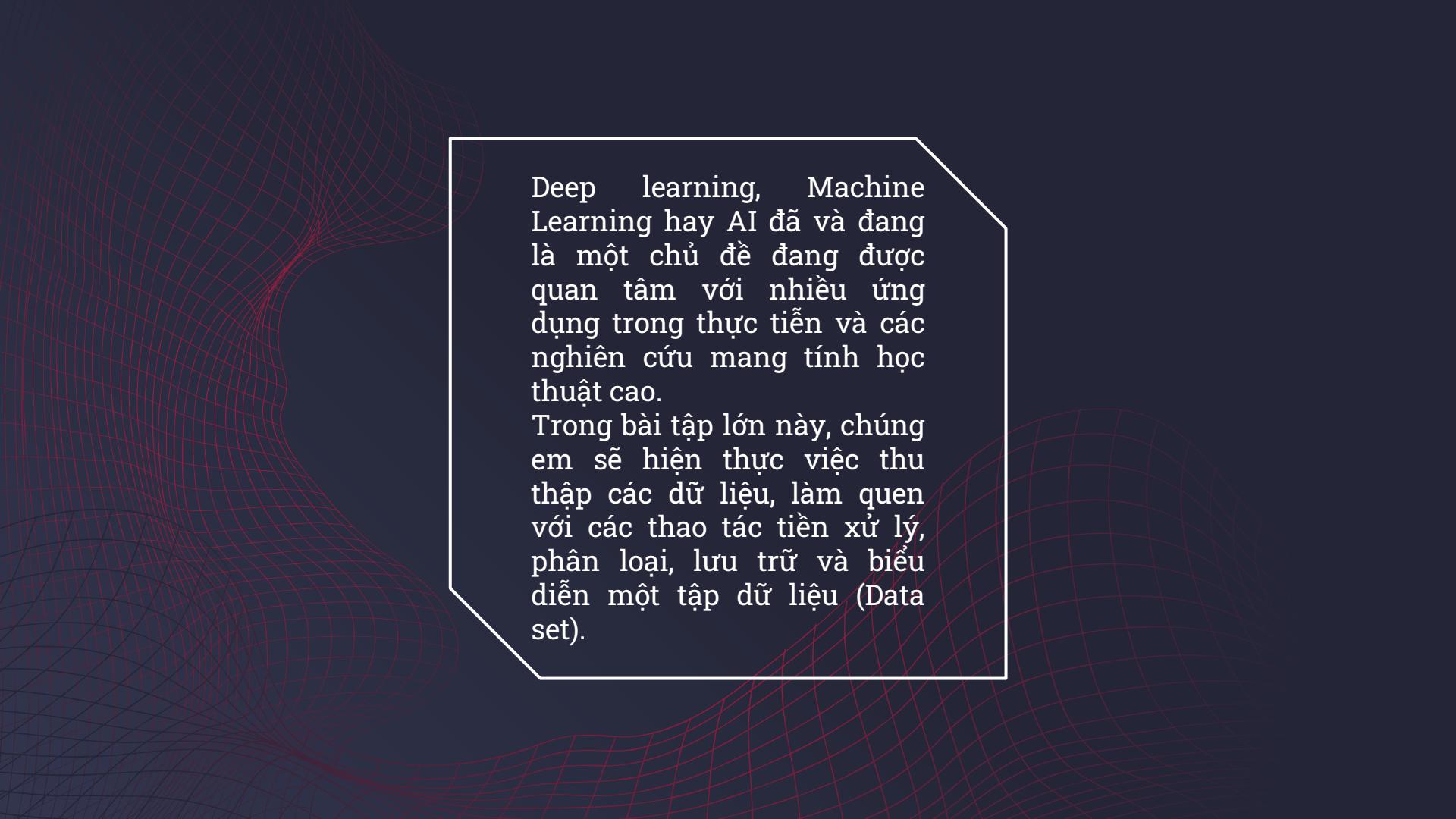
Mô tả ứng dụng

01

## Giới thiệu

The project's topic

---



Deep learning, Machine Learning hay AI đã và đang là một chủ đề đang được quan tâm với nhiều ứng dụng trong thực tiễn và các nghiên cứu mang tính học thuật cao.

Trong bài tập lớn này, chúng em sẽ hiện thực việc thu thập các dữ liệu, làm quen với các thao tác tiền xử lý, phân loại, lưu trữ và biểu diễn một tập dữ liệu (Data set).

## **Đề tài :**

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

- [www.vietnamworks.com/](http://www.vietnamworks.com/)
- [www.vietsingworks.com](http://www.vietsingworks.com)
- <http://careerbuilder.vn/>
- <http://mywork.com.vn/>
- <http://1001vieclam.com/itviec>

02

## Tìm hiểu và thực hiện

How does it work?

## Thư viện đã sử dụng

**Scrapy**

**Beautiful Soup**

**Selenium**

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

Hỗ trợ tích hợp  
để chọn và trích  
xuất dữ liệu từ  
các nguồn HTML /  
XML

**Scrapy**

Một bảng điều khiển  
trình bao tương tác  
(IPython nhận thức)

pip install Scrapy

Beautiful Soup là một thư viện Python để phân tích dữ liệu có cấu trúc, cho phép tương tác với HTML

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

**Beautiful Soup**

pip install requests  
pip install html5lib  
pip install bs4

Tự động chuyển đổi tài liệu đến sang Unicode và tài liệu đi sang UTF-8., nằm trên các trình phân tích cú pháp Python phổ biến như lxml và html5lib

Cung cấp các tiện ích mở rộng để mô phỏng tương tác của người dùng với các trình duyệt, một máy chủ phân phối.

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

- Máy tính cần có python (nhóm đang sử dụng Python 3) và pip.
- pip install selenium.
- Tải webdriver.

# Crawl Data Tool

Công cụ sử dụng trong quá trình crawl dữ liệu



Format data



File Json



Mô hình hóa  
cơ sở dữ liệu



Phân tích dữ liệu

## Format data

### *1. Data*

Data:

- Dữ liệu định tính
- Dữ liệu định lượng

### *2. Phân loại data*

Quan sát, thực nghiệm, mô phỏng, bắt nguồn hoặc biên dịch, tham chiếu hoặc chuẩn, định dạng tệp

# Định dạng JSON

JavaScript Object Notation (JSON) là một

giải pháp thay thế đang thu hút rất nhiều sự chú ý.

JSON là định dạng tệp phổ biến nhất cho dữ liệu “dạng cây” có khả năng có nhiều lớp

```
{  
  "users": [  
    {  
      "name": "Ravi Tamada",  
      "email": "ravi@androidhive.info",  
      "address": "XXX, XXXX, 1234"  
    }  
  ],  
  "posts": [  
    {  
      "id": 100,  
      "author": "Ravi Tamada",  
      "content": "This is awesome firebase realtime database...",  
      "timestamp": "13892733894"  
    }  
  ]  
}
```

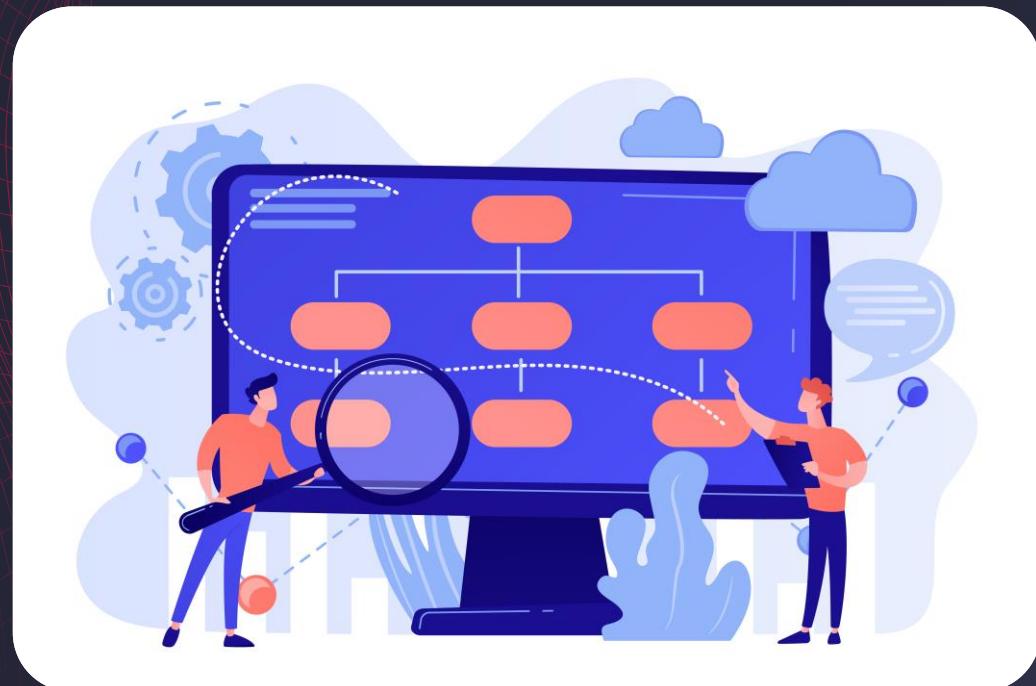
# Ưu điểm

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 rất nhỏ và trọng lượng nhẹ, đó là lý do mà JSON thực thi phản hồi nhanh.

Phân tích cú pháp phía máy chủ là phần quan trọng mà các nhà phát triển muốn nếu quá trình phân tích cú pháp nhanh ở phía máy chủ thì chỉ người dùng mới có thể nhận được phản hồi nhanh chóng từ phản hồi của họ

JSON 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 tương thích với tất cả các trình duyệt. JSON là công cụ tốt nhất để chia sẻ dữ liệu ở bất kỳ kích thước nào kể cả âm thanh, video, ...

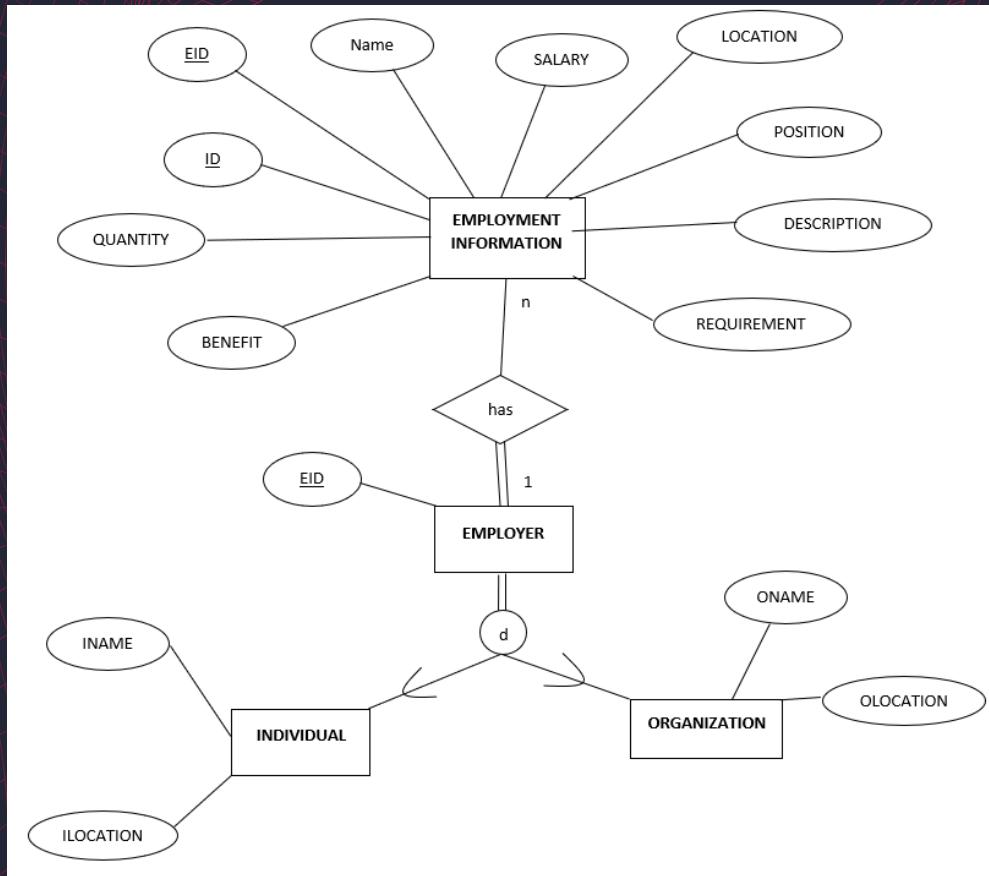
# *Data Model*



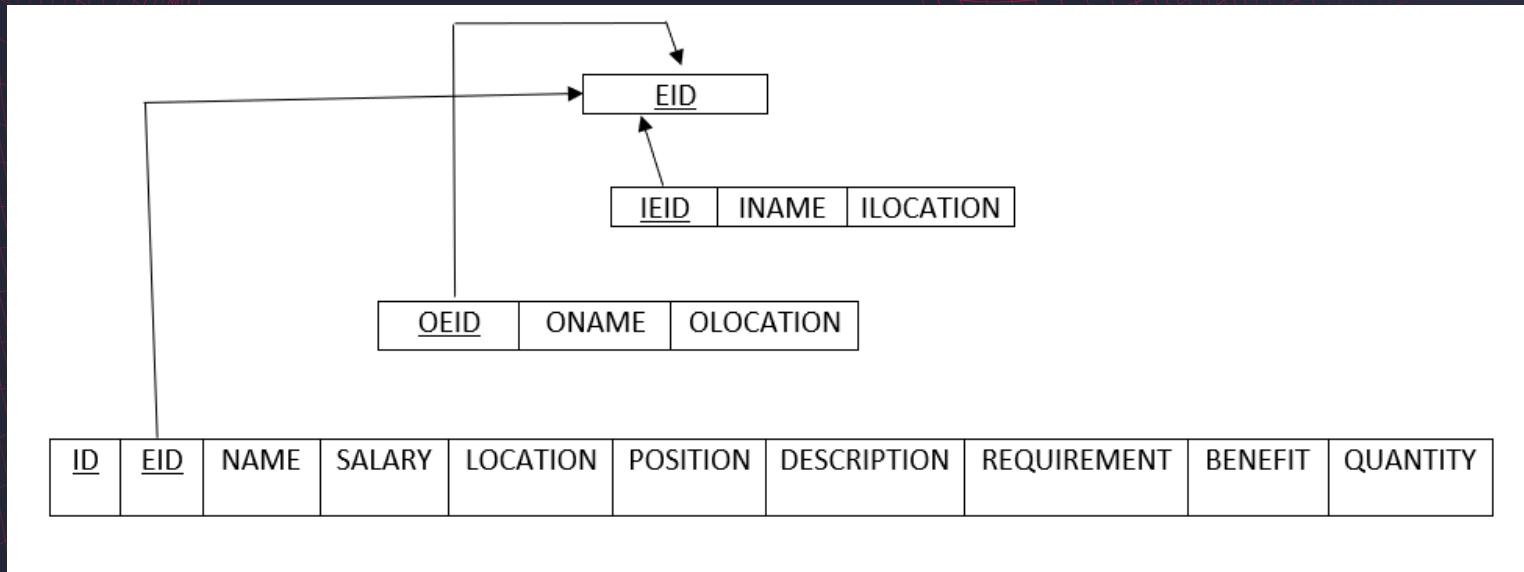
## 1. Mô tả

- 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 tên việc làm, mô tả công việc, mức lương, nơi làm việc, yêu cầu, lợi ích, số lượng...

## 2. Entity Relationship Diagram



### 3. Ánh xạ lượt đồ





Kết nối Mysql

## *Cách kết nối Mysql sử dụng python*

`pip install -U pip`

`pip install mysqlclient`

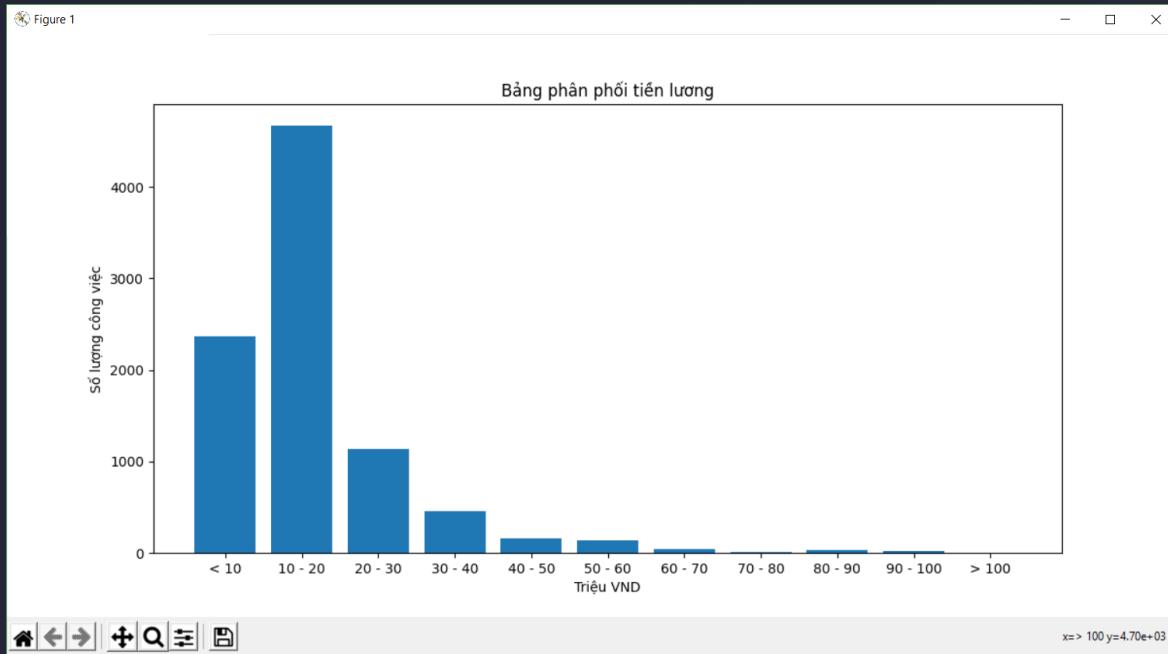
`pip install mysql - connector - python`

`pip install pymysql`

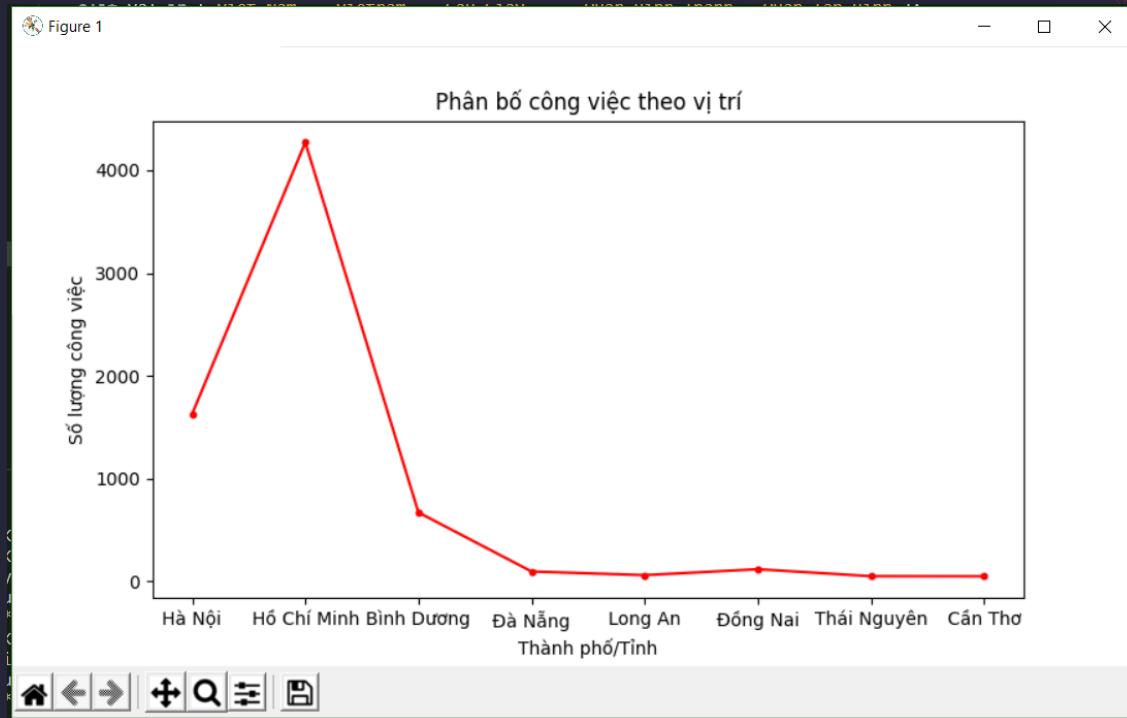
# Phân tích dữ liệu



## Sơ đồ phân bố tin tuyển dụng theo giá trị tiền lương



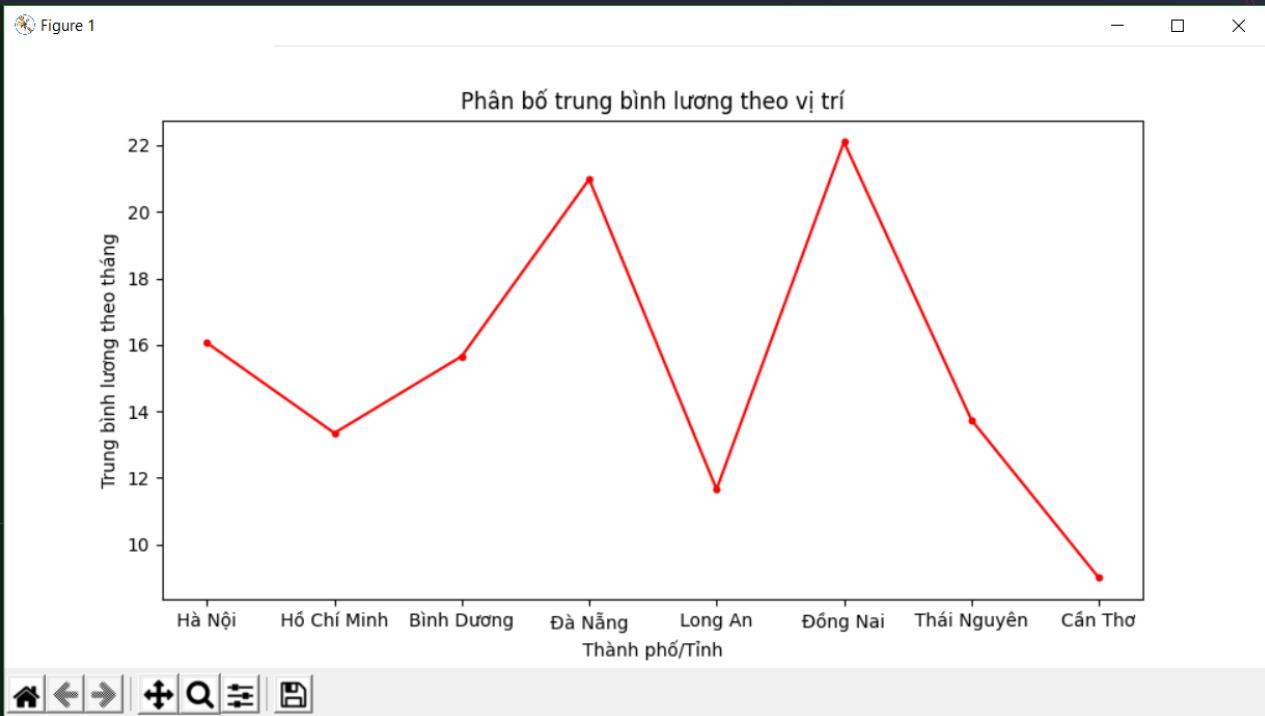
## Sơ đồ phân bố tin tuyển dụng theo vị trí địa lý

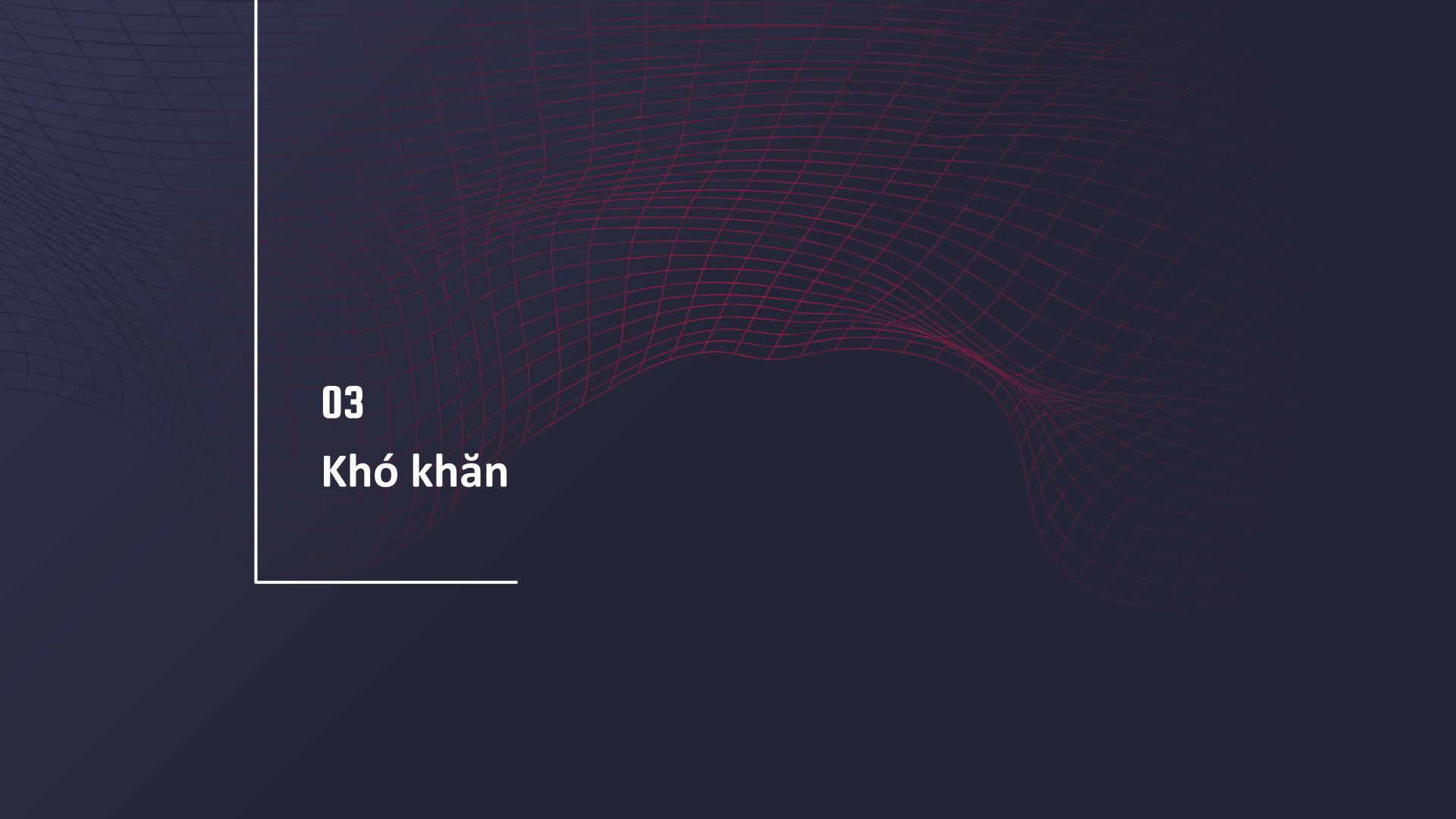


'Hà Nội': 1631, 'Hồ Chí Minh': 4270, 'Bình Dương': 671, 'Đà Nẵng': 97,

'Long An': 61, 'Đồng Nai': 119, 'Thái Nguyên': 51, 'Cần Thơ': 50}

## Sơ đồ phân bố tiền lương trung bình theo vị trí địa lý





03

Khó khăn

## Cài đặt thư viện scrapy của python:

- 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\Code Bishwas\AppData\Local\Temp\pip-install-4feurday\Twisted\setup.py'"'"'; __file__ = '"'"'C:\Users\Code Bishwas\AppData\Local\Temp\pip-install-4feurday\Twisted\setup.py'"'"'; open(__file__, 'r', encoding='utf-8').read().replace('"'"'\r\n'"'"', '"'"'\n'"'); f.close(); exec(compile(open(__file__, 'r', encoding='utf-8').read(), __file__, 'exec'))' install --record 'C:\Users\Code Bishwas\AppData\Local\Temp\pip-record-3zhefghj\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
```

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

The screenshot shows a Visual Studio Code interface with the following details:

- File Explorer (Left):** Shows the project structure:
  - OPEN EDITORS:** scraper.py, temp.html, content.html
  - CRAWL DATA:** tutorial (selected), spiders, items.py, pipelines.py, settings.py, scrapy.cfg
- Content Area (Right):** Displays the content.html file with the following code:

```
17 .cRiaYe{display: -webkit-box; display: -webkit-flex; display: flex; -webkit-align-items: center; align-items: center; justify-content: space-between; -webkit-align-items: center; align-items: center; margin-bottom: 10px; border: 1px solid transparent; white-space: nowrap; padding: 8px 10px; line-height: 1.42857; border-radius: 4px; -webkit-user-select: none; -moz-user-select: none; -ms-user-select: none; user-select: none; font-size: 13px; -webkit-transition: background-color 200ms ease-in-out; transition: background-color 200ms ease-in-out; color: #fff; background-color: #33a837;}.kH1Ckq:hover{background-color: #388122;}@media (max-width: 991px){.kH1Ckq{height: 0px; border-radius: 0px; font-weight: 600;}}
```

```
18 /* sc-component-id: sc_gcmjmt */
```

```
19 .hdIvun{display: inline-block; height: 24px; vertical-align: middle; margin-right: 10px;}
```

```
20 /* sc-component-id: sc_VigVT */
```

```
21 /* sc-component-id: sc_jTzLTM */
```

```
22 /* sc-component-id: sc_fjhpx */
```

```
23 .kH1Ckq{display: block; margin-bottom: 0; text-align: center; vertical-align: middle; touch-action: manipulation; cursor: pointer; background-image: none; border: 1px solid transparent; white-space: nowrap; padding: 8px 10px; line-height: 1.42857; border-radius: 4px; -webkit-user-select: none; -moz-user-select: none; -ms-user-select: none; user-select: none; font-size: 13px; -webkit-transition: background-color 200ms ease-in-out; transition: background-color 200ms ease-in-out; color: #fff; background-color: #33a837;}.kH1Ckq:hover{background-color: #388122;}@media (max-width: 991px){.kH1Ckq{height: 0px; border-radius: 0px; font-weight: 600;}}
```

```
24 /* sc-component-id: sc_fjhpx */
```

```
25 .blfMza{display: -webkit-box; display: -webkit-flex; display: flex; -webkit-align-items: center; align-items: center; justify-content: space-between; -webkit-align-items: center; align-items: center; -webkit-justify-content: center; justify-content: center; -webkit-align-items: center; align-items: center; height: 100%;}
```

```
26 /* sc-component-id: sc_jcRNIG */
```

```
27 .eeRHVh{display: inline-block; width: 24px; height: 24px; vertical-align: middle; margin-right: 10px;}
```

```
</style><meta name="viewport" content="width=device-width,initial-scale=1.0,minimum-scale=1.0,maximum-scale=1.0,viewport-fit=cover">
```

```
<meta charset="utf-8"/><link rel="canonical" href="https://www.chatbot.com/toan-quoc/viec-lam"/><title>Tim Viec Lam Nhanh, Thong Tin Tuyen Dung Viec Lam 11/2020</title><meta name="description" content="Tim viec lam, Tuyen dung Viec lam nhanh, hieu quoc trang 24h tren Toan Quoc" /> Cập nhật việc làm mới tháng 11/2020 | Tim viec vang tung ngay hom nay!</meta><meta name="next-head-count" content="5"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/commons.b2347fe.chunk.css" as="style"/><link rel="stylesheet" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/commons.b2347fe.chunk.css"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/styles.db737d7.chunk.css" as="style"/><link rel="stylesheet" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/styles.db737d7.chunk.css"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/styles.a91963929948011003495a69a933c3e84599f3e0.css" as="style"/><link rel="stylesheet" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/a91963929948011003495a69a933c3e84599f3e0.css" as="style"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/b872196.chunk.css" as="script"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/css/b872196.chunk.css"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/main-366fc1c4374dd316f794.js" as="script"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/main-366fc1c4374dd316f794.js"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/framework.231eb8819e0afbf19ce.js" as="script"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/framework.231eb8819e0afbf19ce.js"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/4adfbca5bd804bcf2d2088abfafe97dab2df5d.f5b34f4870323c5038893.js" as="script"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/9c00fdec875320c0761a2b6ba9ca3b96eccf311.0682a5541486006a04a.js" as="script"/><link rel="preload" href="https://static.chatbot.com/storage/C2C_V2_PRODUCTION/2d0176d9/next/static/chunks/9c00fdec875320c0761a2b6ba9ca3b96eccf311.0682a5541486006a04a.js"/>
```
- Terminal (Bottom Left):** Shows the command: `scrapy crawl spiders`
- Output (Bottom Middle):** Shows the log message: `INFO: Spider closed (finished)`
- Status Bar (Bottom Right):** Shows the power shell icon and other status indicators.

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

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

Cụ thể ở trang web <https://vietnamworks.com/>, khi ta gọi request trả về nội dung của file HTML, web chỉ trả về một số lượng trường dữ liệu nhất định và số lượng link trả về cũng bị hạn chế. Nhóm tìm hiểu thêm về selenium để giải quyết vấn đề này



04

## Đề xuất ứng dụng

*Thiết kế webside giới thiệu việc làm:*

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

*Các tính năng cơ bản:*

- \* **Khách:** xem các thông tin public trên trang web, cho phép đăng kí, 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.*

# Phân công công việc

STT	Họ và tên	Công việc	Hoàn thành
1	Nguyễn Thế Viễn	Viết báo cáo tổng kết, push data lên mysql, phân tích dữ liệu, Crawl data, Làm sạch dữ liệu	100%
2	Nguyễn Ngọc Thuấn	Thiết kế data model, phân tích dữ liệu, Crawl data, Làm sạch dữ liệu	100%
3	Bùi Hữu Đang	Làm slide thuyết trình, thiết kế data model, Crawl data, Làm sạch dữ liệu, phân công công việc	100%
4	Phạm Đức Duy Anh	Crawl data, làm sạch dữ liệu, phân tích dữ liệu	100%
5	Huỳnh Ngọc Tân	Đề xuất ứng dụng, Crawl data, làm sạch dữ liệu	100%

Cảm ơn thầy và các bạn đã  
chú ý lắng nghe!

---