BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CẦN THƠ

KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

ĐỒ ÁN MÔN

KHAI KHOÁNG DỮ LIỆU

(CT31201)

ĐỀ TÀI

CÀO VÀ PHÂN TÍCH DỮ LIỆU TRANG UPWORK

Giáo viên hướng dẫn:
Lưu Tiến Đạo

Sinh viên thực hiện:
Trần Anh Khoa B1913240

Cần Thơ, 12/2021

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

NHẬN XÉT CỦA GIẢNG VIÊN 2
DANH MỤC HÌNH 4
GIỚI THIỆU 5
I. ĐẶT VẤN ĐỀ 5
II. PHƯƠNG PHÁP GIẢI QUYẾT VẤN ĐỀ 5
III. MỤC TIÊU CỦA ĐỀ TÀI 5
IV. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU 6
V. PHƯƠNG PHÁP NGHIÊN CỨU 6
VI. KẾT QUẢ ĐẠT ĐƯỢC 6
MÔ TẢ BÀI TOÁN 8
I. MÔ TẢ CHI TIẾT BÀI TOÁN 8
II. VẤN ĐỀ LIÊN QUAN ĐẾN BÀI TOÁN 8
III. GIẢI PHÁP CÓ LIÊN QUAN ĐẾN BÀI TOÁN 8
1. Cách tìm API của trang web upwork.com 8
2. Tìm hiểu về cấu trúc API của trang web 9
CÀI ĐẶT 12
I. CÀI ĐẶT CÁC THƯ VIỆN CẦN THIẾT 12
II. THIẾT LẬP HEADERS XÁC THỰC VỚI API VÀ HÀM NHẬN KẾT QUẢ 12
III. KHAI THÁC UPWORK: 12
PHÂN TÍCH DỮ LIỆU 14
I. TỈ LỆ PHÂN BỐ NGƯỜI TÌM VIỆC Ở KHU VỰC VIỆT NAM 14
II. THU NHẬP BÌNH QUÂN SAU 1 GIỜ LÀM VIỆC GIỮA CÁC TỈNH THÀNH CỦA VIỆT NAM 14
III. TỈ LỆ TÌM ĐƯỢC VIỆC LÀM TRÊN CẢ NƯỚC. 15
IV. TỈ LỆ TÌM ĐƯỢC VIỆC LÀM GIỮA CÁC TỈNH THÀNH CỦA VIỆT NAM. 15

DANH MỤC HÌNH

Hình 1. Ảnh minh hoạ file dữ liệu csv 6

Hình 2. Ảnh minh hoạ về thuộc tính của dữ liệu..... 6

Hình 3. Ảnh minh hoạ request và respone của trang web upwork.com..... 9

Hình 4. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1..... 10

Hình 5. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1..... 10

Hình 6. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1..... 11

Hình 7. Tỉ lệ phân bố người tìm việc ở các Tỉnh thành của Việt Nam. 14

Hình 8. Thu nhập bình quân sau 1 giờ làm việc giữa các Tỉnh thành của Việt Nam... 14

Hình 9. Tỉ lệ tìm được việc làm trên cả nước. 15

Hình 10. Tỉ lệ tìm được việc làm giữa các Tỉnh thành của Việt Nam. 15

GIỚI THIỆU

I. ĐẶT VẤN ĐỀ

Thu thập, thống kê và phân tích dữ liệu là hoạt động rất quan trọng trong thời đại cách mạng công nghệ 4.0. Từ dữ liệu có sẵn trong quá khứ, có thể dự đoán hành vi và ảnh hưởng đến hành vi của người dùng hoặc khách hàng trong tương lai.

Trong các hoạt động thu thập dữ liệu, một phương pháp rất phổ biến là thu thập dữ liệu từ các trang web trên Internet. Ngày nay, với sự phát triển mạnh mẽ của công nghệ, mọi hành vi và thông tin của người dùng đều được lưu trữ và thống kê đến mức chi tiết nhất. Khi công tác thống kê và nghiên cứu yêu cầu lượng dữ liệu lớn thì việc sử dụng lượng dữ liệu lớn là rất cần thiết.

Hiện nay, trong thời kỳ dịch bệnh phức tạp, kinh tế khủng hoảng trầm trọng, đời sống của nhiều người rất khó khăn. Nhu cầu tìm việc ngày càng cao, đặc biệt là nhu cầu làm việc từ xa. Do đó, các nhóm ngành liên quan đến kinh tế và marketing cần rất nhiều dữ liệu, bao gồm dữ liệu về mô tả công việc, kỹ năng nghề nghiệp, cấp bậc, vị trí… để thực hiện các nghiên cứu và thống kê đặc biệt về các nhu cầu tìm việc làm trong mùa dịch này. Việc thực hiện thu thập dữ liệu từ các trang web có dữ liệu lớn về thông tin tìm kiếm việc làm từ xa sẽ giúp việc nghiên cứu, thống kê dễ dàng đưa ra kết quả sát thực phản ánh đúng thực tế.

II. PHƯƠNG PHÁP GIẢI QUYẾT VẤN ĐỀ

Phương pháp được dựa vào các kiến thức được chia sẻ từ các nguồn uy tín:

- Cào dữ liệu từ trang Linkedin: https://www.youtube.com/watch?v=hfnBswCe4QE&t=153s
- Cào dữ liệu từ trang Facebook: https://www.youtube.com/watch?v=EawbYWaTP_k

III. MỤC TIÊU CỦA ĐỀ TÀI

Sử dụng các phương pháp cào dữ liệu, để lấy dữ liệu từ trang web upwork nhằm lấy được các thông tin sau:

- Mô tả về bản thân
- Các kỹ năng về nghề nghiệp
- Quốc gia (Việt Nam)
- Thành phố (HCM, Hà Nội, Đà Nẵng,…)
- Đơn vị tiền tệ (USD)
- Chi phí phải trả cho 1 giờ làm việc (USD/giờ)
- Tổng thu nhập trước thuế (USD)
- Tổng thu nhập sau thuế(USD)
- Tổng số công việc đã làm

- Tổng số giờ đã làm việc

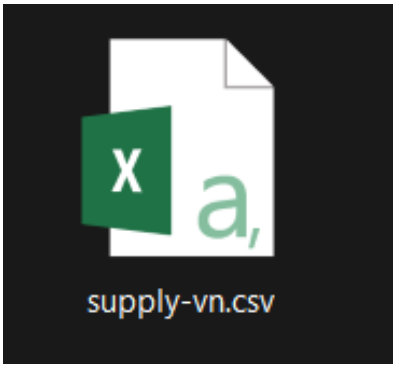
IV. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Chủ đề và phạm vi nghiên cứu: Dữ liệu thu thập được sẽ được cung cấp cho các nhà phân tích dữ liệu để xử lý và sử dụng giá trị trong dữ liệu thu thập được.

V. PHƯƠNG PHÁP NGHIÊN CỨU

- Crawl dữ liệu
- Phân tích dữ liệu
- Lấy thông tin chi tiết của từng cá nhân tìm việc.

VI. KẾT QUẢ ĐẠT ĐƯỢC



Hình 1. Ảnh minh hoạ file dữ liệu csv

top_skills	country	city
Autodesk 3ds Max, Autodesk AutoCAD, SketchUp, Chaos Group V-Ray, CGI,	Vietnam	Ho Chi Minh City
Microsoft PowerPoint, Presentation Design, Graphic Design, Microsoft Pow	Vietnam	Hanoi
Retouching, Fashion Retouch, Photo Editing, Portrait Photography, Model P	Vietnam	Ho Chi Minh
English to Vietnamese Translation, Data Entry, Market Research, Content W	Vietnam	Hanoi
English to Vietnamese Translation, Microsoft Office, Proofreading, Administ	Vietnam	Hanoi
Translation, English to Vietnamese Translation, Microsoft Word, Google Sea	Vietnam	Ho Chi Minh City
3D Modeling, 3D Rendering, Architectural Design, Architectural Rendering, I	Vietnam	HoChiMinh
English to Vietnamese Translation, Vietnamese to English Translation	Vietnam	Ho Chi Minh
Translation, Typeform, English to Japanese Translation, Vietnamese to Engl	Vietnam	Ho Chi Minh City
SDL Trados, English to Vietnamese Translation, Interpretation, Market Rese	Vietnam	Ho Chi Minh City

Hình 2. Ảnh minh hoạ về thuộc tính của dữ liệu

Mô tả về tập dữ liệu:

Tên cột	Kiểu dữ liệu	Ý nghĩa
ciphertext	str	Mã người dùng
username	str	Tên người dùng
title	str	Tiêu đề
description	str	Mô tả về bản thân
top_skills	str	Những kỹ năng của bản thân
country	str	Quốc gia
city	str	Thành Phố
state	str	Tỉnh
hourly_rate	int	Chi phí cho 1 giờ làm việc
primary_currency	str	Đơn vị tiền tệ
combinedTotalEarnings	float	Tổng thu nhập trước thuế
combinedTotalRevenue	float	Tổng thu nhập sau thuế
totalJobsWorked	int	Tổng số công việc đã hoàn thành
totalHourlyJobs	int	Tổng số giờ làm việc

MÔ TẢ BÀI TOÁN

I. MÔ TẢ CHI TIẾT BÀI TOÁN

Đề tài "lấy dữ liệu từ trang web upwork" thuộc dạng khai thác dữ liệu, sử dụng dữ liệu phong phú được cung cấp bởi trang web upwork (một trang web tìm kiếm việc làm nổi tiếng) để thống kê và nghiên cứu.

Để giải quyết vấn đề này, nghiên cứu này đã xây dựng một chương trình cho phép người dùng tạo ra một con bot, truy cập trực tiếp vào trang web hoạt động và truy xuất dữ liệu cần thiết. Ví dụ: Mô tả các kỹ năng chuyên môn của bạn, mức thu nhập trong 1 giờ làm việc, nơi bạn sống, ... Chương trình được xây dựng bằng ngôn ngữ lập trình Python. Cấu trúc chương trình đơn giản, dễ thực hiện và chỉnh sửa theo yêu cầu của người sử dụng.

II. VẤN ĐỀ LIÊN QUAN ĐẾN BÀI TOÁN

Các vấn đề liên quan đến cào dữ liệu trang upwork:

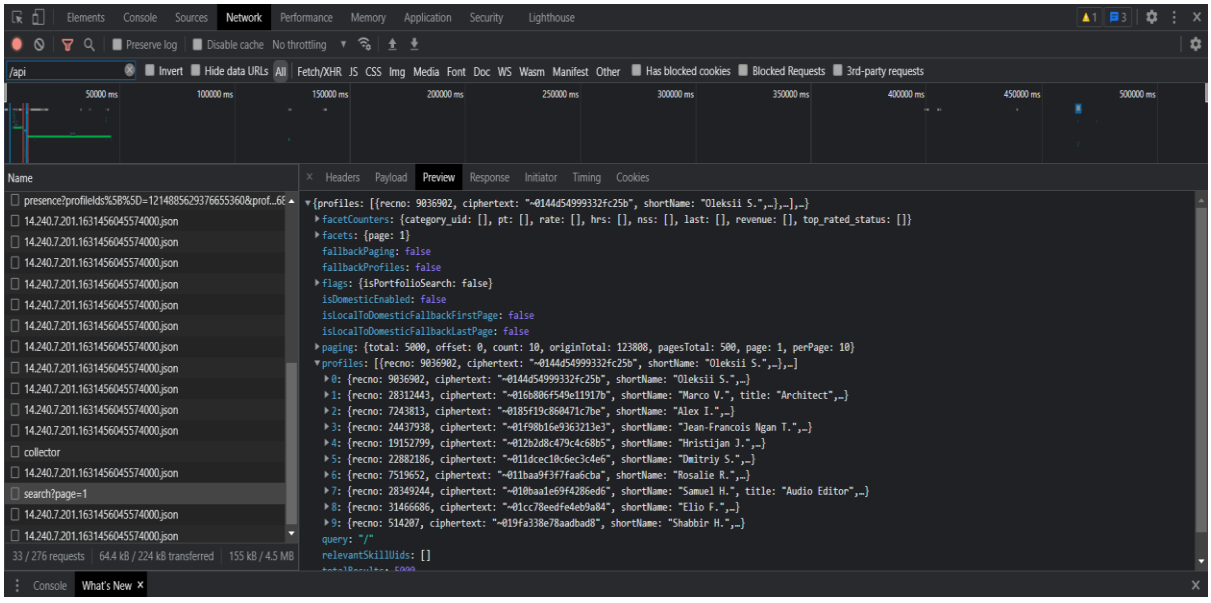
- Tìm API của website upwork.com.
- Tìm hiểu về cấu trúc API của trang web
- Phân tích và trích xuất các đặc điểm của dữ liệu được trả về từ API.

III. GIẢI PHÁP CÓ LIÊN QUAN ĐẾN BÀI TOÁN

1. Cách tìm API của trang web upwork.com

Để lấy được API của website upwork ta cần sử dụng công cụ devtool của trình duyệt Google Chrome hoặc Microsoft Edge để xem những truy vấn từ máy khách lên máy chủ. Công cụ này sẽ hiển thị tất cả các truy vấn khi ta bắt đầu truy cập bất kì một trang web nào, từ đó ta có thể tìm thấy được đường dẫn của API.

Để có thể lấy được request của trang web upwork.com ta cần mở thanh công cụ devtool từ hai trình duyệt phía bên trên bằng cách ấn phím F12 sau đó bật tab network và truy cập vào trang web upwork.com sau khi trang web đã được load đầy đủ tất cả request đều được hiển thị như hình.



Hình 3. Ảnh minh hoạ request và respone của trang web upwork.com

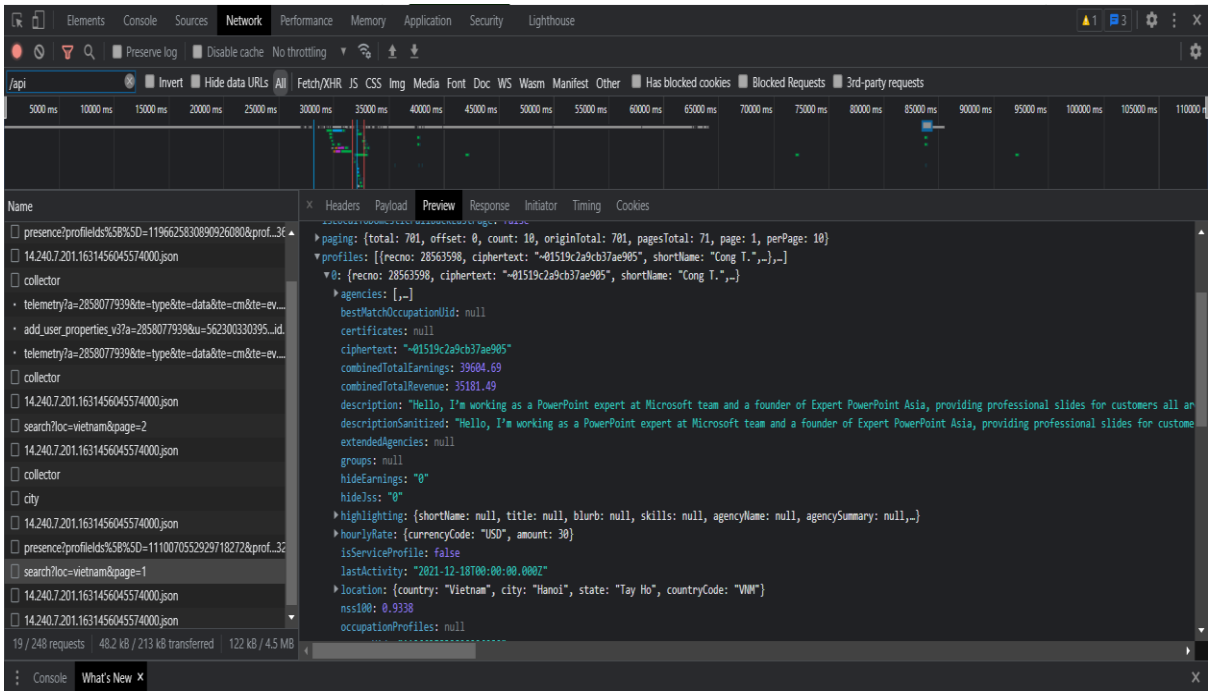
Dựa vào request và response ta có thể thấy rằng để lấy được danh sách các user trên upwork thì truy vấn đến API sau:
https://www.upwork.com/search/profiles/api/search?page=1

2. Tìm hiểu về cấu trúc API của trang web

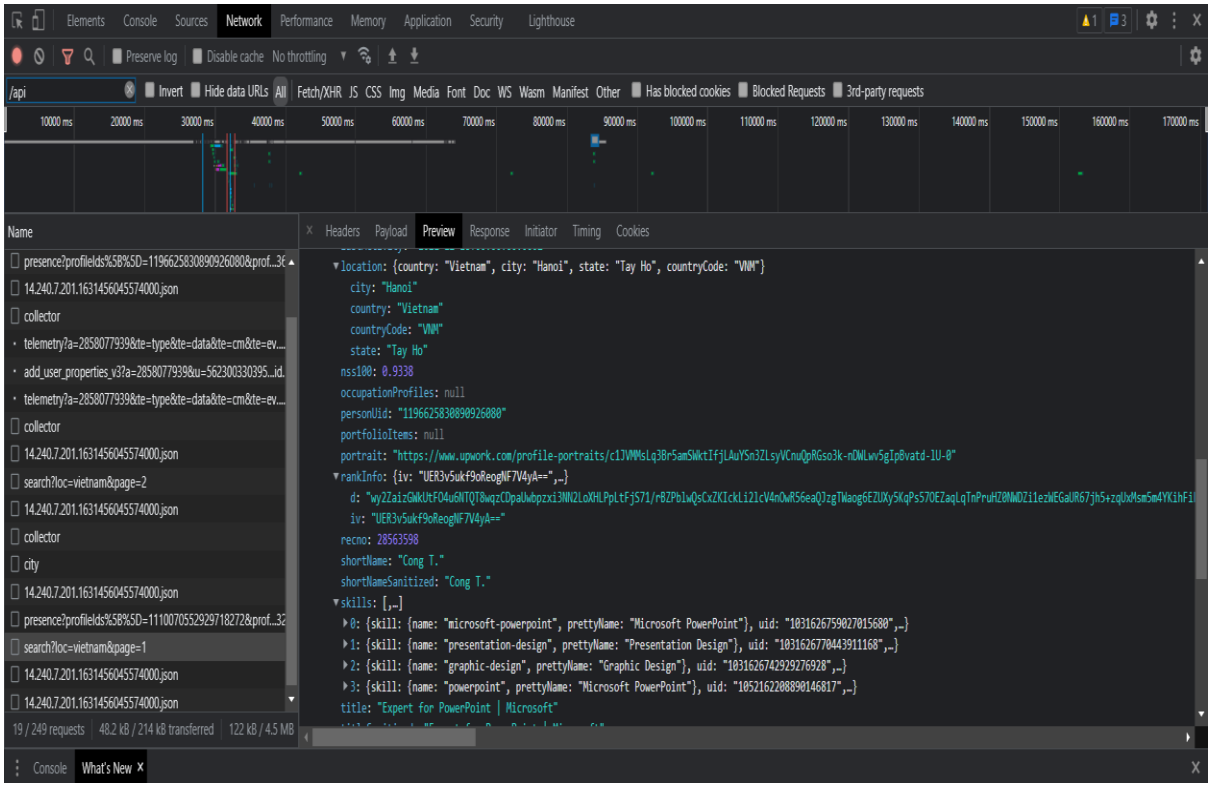
Để lấy thông tin về người tìm việc làm ở khu vực Việt Nam ta cần phải truy vấn đến API sau: https://www.upwork.com/search/profiles/?loc=vietnam&page=1 API sẽ trả về danh sách các user ở khu vực Việt Nam và đầy đủ các thuộc tính sau:

- ciphertext: bản mã của từng người dùng
- username: tên người dùng
- title: tiêu đề
- description: mô tả
- top_skills: danh sách các kỹ năng
- country: quốc gia
- city: thành phố
- state: tỉnh/ tiểu bang
- hourly_rate: chi phí hàng giờ
- combinedTotalEarnings: tổng thu nhập
- combinedTotalRevenue: tổng doanh thu
- totalJobsWorked: tổng số công việc đã làm
- totalHourlyJobs: tổng số công việc làm hàng giờ

Đề tài: Cào và phân tích dữ liệu trang upwork.com

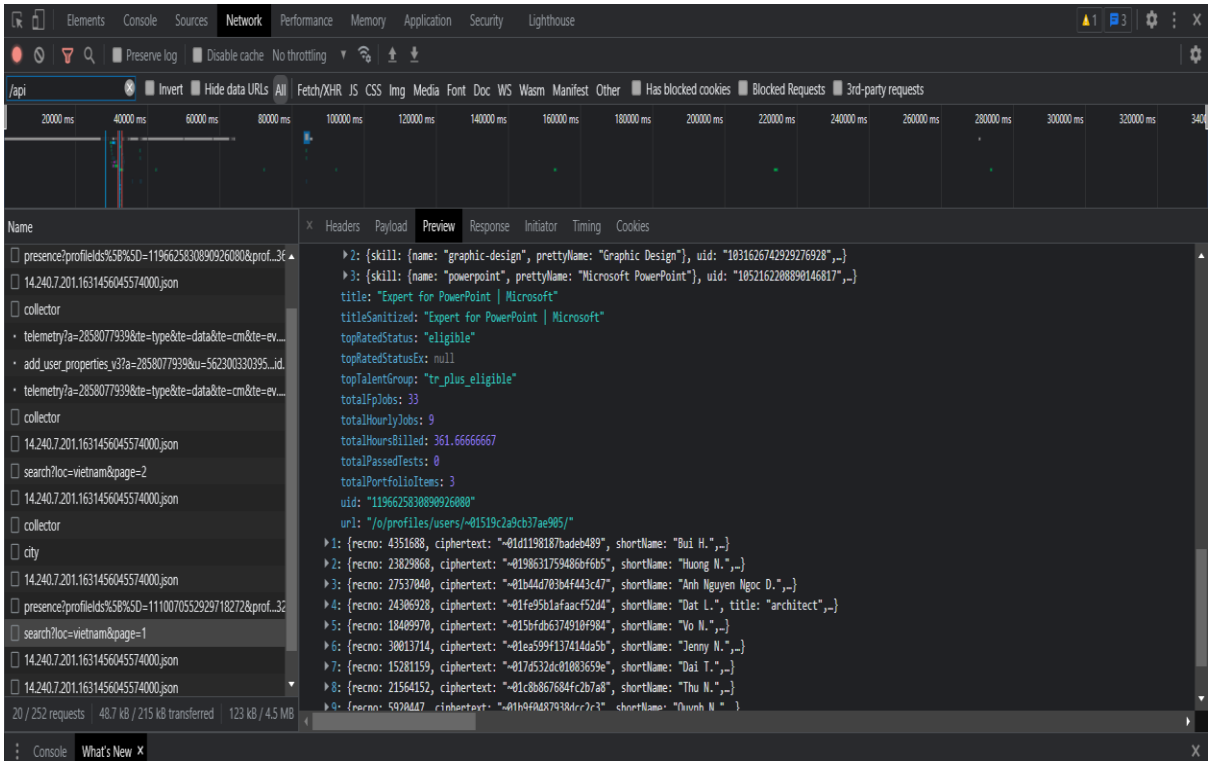


Hình 4. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1



Hình 5. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1

Đề tài: Cào và phân tích dữ liệu trang upwork.com



Hình 6. Kết quả của API https://www.upwork.com/search/profiles/?loc=vietnam&page=1

CÀI ĐẶT

I. CÀI ĐẶT CÁC THƯ VIỆN CẦN THIẾT

Để mô phỏng lại các HTTP request và respond trên Python, ta sử dụng thư viện **requests**. Cài đặt thư viện **requests** bằng cách nhập vô console lệnh sau:

```
!pip install requests
```

II. THIẾT LẬP HEADERS XÁC THỰC VỚI API VÀ HÀM NHẬN KẾT QUẢ

Do API được xác thực bằng: Cách xác thực bên dưới sẽ nâng số lượng page cào được khoảng 500 pages (nếu không được xác thực các trường như bên dưới sẽ bị giới hạn số lượng page cào còn 70 pages).

```
#Set header fields.
Headers = {
             'sec-ch-ua': '"Chromium";v="93", "\"Not\\A;Brand";v="99"',
             'X-Odesk-User-Agent': 'oDesk LM',
             'sec-ch-ua-mobile': '?0',
             'Authorization': 'Bearer
oauth2v2_f867e03ce6d7185c2256de1c25b6f53e',
             'accept': 'application/json, text/plain, */*',
             'X-Oauth2-Required': 'True',
             'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.111 Safari/537.36',
             'X-Requested-With': 'XMLHttpRequest',
             'X-Odesk-Csrf-Token': '61408d93fe0f8fa8d9b0ff889de3422a',
             'Sec-Ch-Ua-Platform': 'windows',
             'sec-fetch-site': 'same-origin',
             'sec-fetch-mode': 'cors',
             'Sec-Fetch-Dest': 'empty',
             'Referer': 'https://www.upwork.com/ab/find-work/recommended',
             'upgrade-insecure-requests': '1',
             'Accept-Encoding': 'gzip, deflate',
             'Accept-Language': 'en-US,en;q=0.9'
    }
```

Do kết quả trả về của requests là dạng JSON nên ta cần viết hàm **checkExists** để kiểm tra thông tin của 1 người đã được lấy chưa thông qua cột cirphertext được đọc từ file dữ liệu đang thu thập nếu thông tin đã có thì bỏ qua.

```
df = pd.read_csv(filename_supply) #filename_supply = ‘supply-vn.csv’
def checkExists(data):
   return data in df.ciphertext.values
```

III. KHAI THÁC UPWORK:

Do upwork chỉ cho phép truy cập tối đa 500 pages nên ta gán cứng giá trị 500 ở trong vòng lặp. Đồng thời kiểm tra xem người dùng đã được được cào chưa. Nếu chưa thì thêm vào file csv dữ liệu thu thập.

```
for page in range(1, 500):
    response =
session.get('https://www.upwork.com/search/profiles/api/search?loc=vietnam&page={}
'.format(page))
    profiles = response.json()['profiles']
    for profile in profiles:
        if(checkExists(profile['ciphertext'])):
            continue
        else:
            print(profile['ciphertext'])
            save_profile_to_csv(profile)
```

Hàm lưu dữ liệu được truyền vào file csv và chuẩn hoá lại tên của các cột dữ liệu.

```
def save_profile_to_csv(profile):
    result = {}
    result['ciphertext']            = profile['ciphertext']
    result['username']              = profile['shortName']

    result['title']                 = profile['title']
    result['description']           = profile['description']

    result['top_skills']            = get_top_skills(profile['skills'])
    result['country']               = profile['location']['country']
    result['city']                  = profile['location']['city']
    result['state']                 = profile['location']['state']
    result['hourly_rate']           = profile['hourlyRate']['amount']
    result['primary_currency']      = profile['hourlyRate']['currencyCode']
    result['combinedTotalEarnings']      = profile['combinedTotalEarnings']
    result['combinedTotalRevenue']        = profile['combinedTotalRevenue']
    result['totalJobsWorked']       = profile['totalFpJobs']
    result['totalHourlyJobs'] = profile['totalHourlyJobs']
    write_to_csv(filename_supply, result)
```

Các hàm hỗ trợ quá trình chuẩn hoá tên cột dữ liệu.
Hàm **get_top_skills** dùng để lấy từng skill từ mảng top_skills dạng JSON và trả về mảng skills.

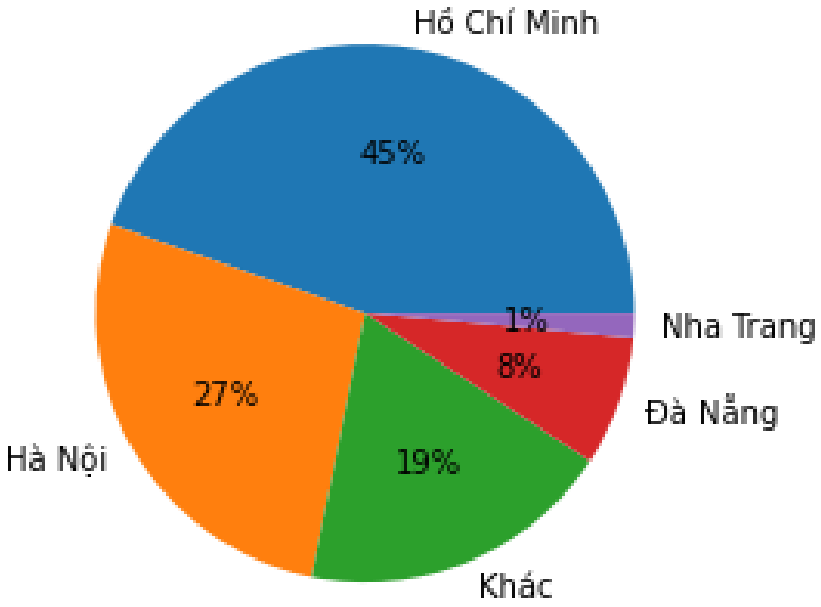
```
def get_top_skills(top_skills):
    string = ""
    for skill in top_skills:
        string = string +", "+ str(skill['skill']['prettyName'])
    return string[2:]
```

Hàm **write_to_csv** dùng để ghi thông tin sau khi đã chuẩn hoá tên cột dữ liệu và lưu dữ liệu vào file csv đã được quy định từ trước.

```
def write_to_csv(filename, info):
    with open(filename, 'a', encoding="utf-8", newline='') as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(info.values())
```

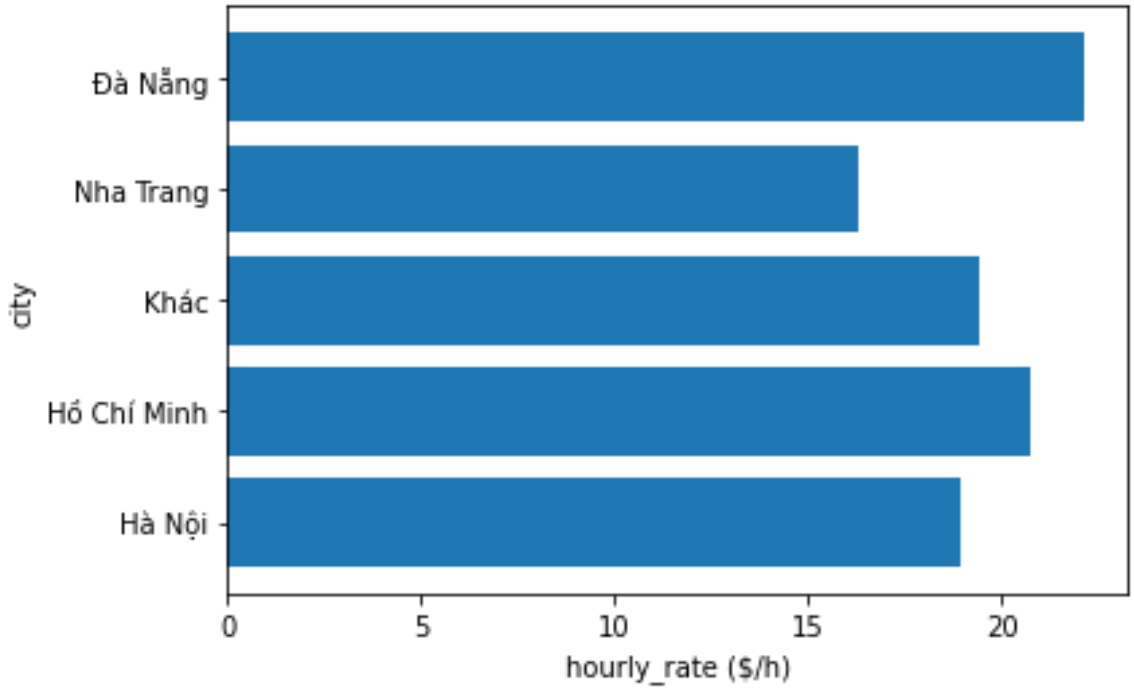
PHÂN TÍCH DỮ LIỆU

I. TỈ LỆ PHÂN BỐ NGƯỜI TÌM VIỆC Ở KHU VỰC VIỆT NAM .



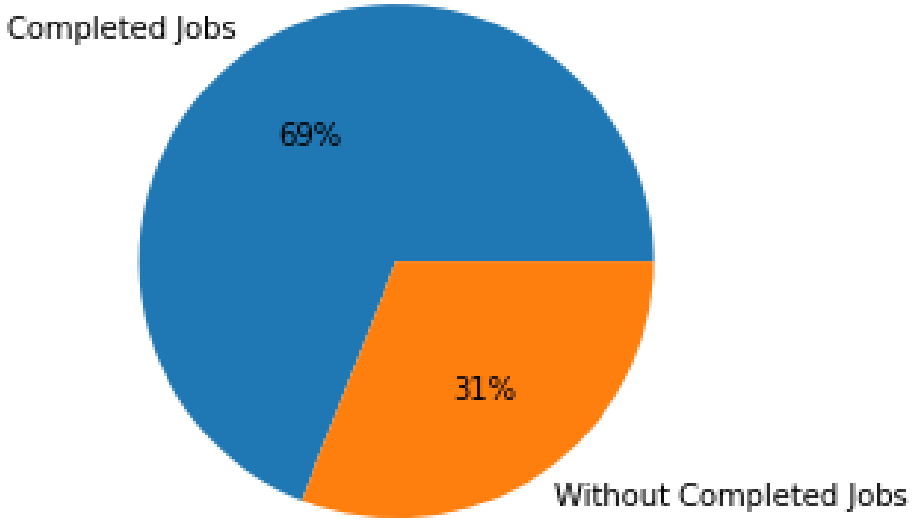
Hình 7. Tỉ lệ phân bố người tìm việc ở các Tỉnh thành của Việt Nam.

II. THU NHẬP BÌNH QUÂN SAU 1 GIỜ LÀM VIỆC GIỮA CÁC TỈNH THÀNH CỦA VIỆT NAM.



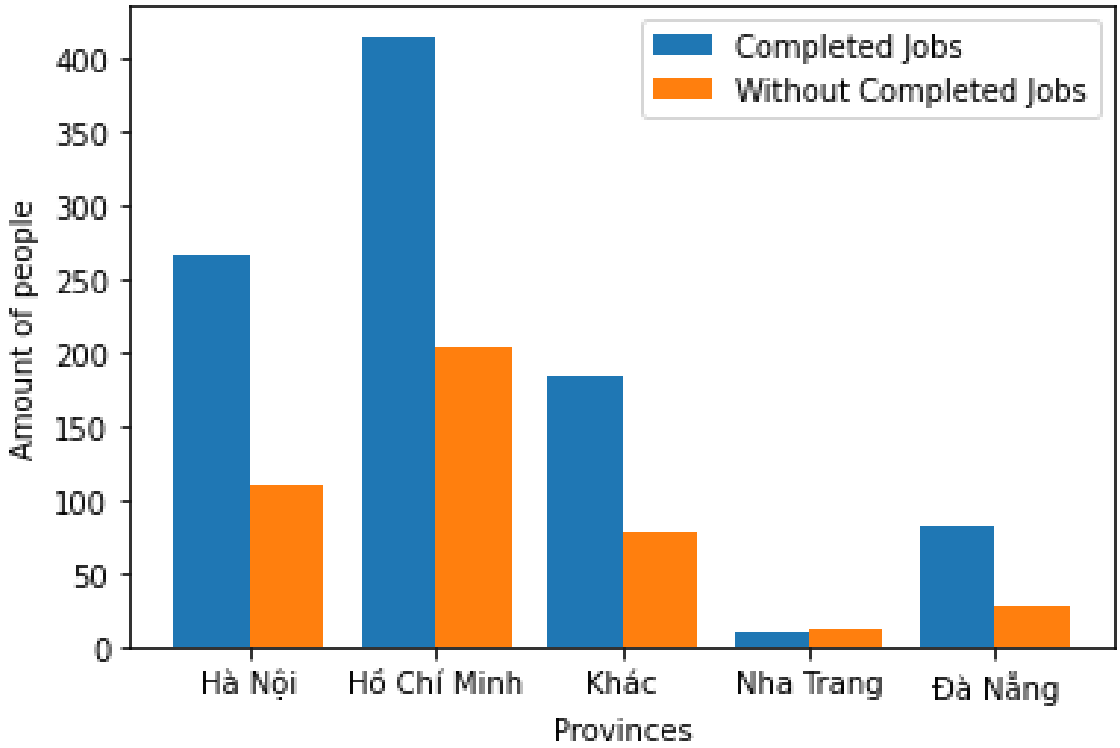
Hình 8. Thu nhập bình quân sau 1 giờ làm việc giữa các Tỉnh thành của Việt Nam

III. TỈ LỆ TÌM ĐƯỢC VIỆC LÀM TRÊN CẢ NƯỚC.



Hình 9. Tỉ lệ tìm được việc làm trên cả nước.

IV. TỈ LỆ TÌM ĐƯỢC VIỆC LÀM GIỮA CÁC TỈNH THÀNH CỦA VIỆT NAM.



Hình 10. Tỉ lệ tìm được việc làm giữa các Tỉnh thành của Việt Nam.