TRƯỜNG ĐẠI HỌC BÁCH KHOA ĐẠI HỌC QUỐC GIA TP.HCM KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH



KỸ THUẬT LẬP TRÌNH

GVHD: Lưu Quang Huân

SVTH: Nguyễn Văn Tiến 1814320

Nguyễn Hồ Phước Lộc 1812969 Nguyễn Hữu Thắng 1814096

Mục lục

1 Giới thiệu chung	4
2 Giới thiệu scrapy	
2.1 Scrapy:	
2.2 Crawl data bằng scrapy	
3 Xây dựng cơ sở dữ liệu crawl được	7
3.1 Xây dựng cơ sở dữ liệu	
3.2 Phân tích dữ liệu crawl được	
4 Xây dựng ứng dụng với cơ sở dữ liệu crawl được	
4.1 Giới thiệu Web	11
4.2 Một số tính năng của Web	12
4.2.1 Chức năng thêm dữ liệu vào database:	13
4.2.2 Chức năng tìm kiếm	
4.3.3 Chức năng cập nhật và xóa	17
5 Đánh giá kết quả thu được và hướng phát triển trong tương lai	18
5.1 Đánh giá kết quả thực hiện được	18
5.2 Hướng phát triển trong tương lai	
6 Tài liêu tham khảo	20

MÔ TẢ CHUNG

Đề tài: Xây dựng công cụ thu thập và phân loại tin tuyển dụng từ vieclam24h.com.vn, timviecnhanh.com...

MỤC ĐÍCH - YÊU CẦU

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. Tuy nhiên, việc triển khai các ứng dụng, các mô hình cũng như việc tiến hành nghiên cứu gặp nhiều khó khăn. Một mặt, do các bộ dữ liệu huấn luyện hạn chế, về cả số lượng và không đảm bảo về chất lượng. Các bộ dữ liệu hiện có cũng hạn chế, đa phần bằng các ngôn ngữ không phải tiếng Việt.

Xuất phát từ nhu cầu thực tế của việc thực hiện tác đề tài, đồ án môn học sau này hoặc khi làm LVTN liên quan đến dữ liệu, sinh viên gặp nhiều khó khăn trong việc thu thập dữ liệu mẫu (Dataset) để triển khai các mô hình Machine Learning, Deep Learning hoặc cơ bản hơn là thu thập dữ liệu làm dữ liệu mẫu cho một ứng dụng, hệ thống đã xây dựng. Trong bài tập lớn này, SV sẽ có cơ hội hiện thực việc thu thập các dữ liệu này, làm quen với các thao tác tiền xử lý, phân loại, lưu trữ và biểu diễn một tập dữ liệu (Data set).

1 Giới thiêu chung

Ta sử dụng công cụ scrapy để crawl data từ các website

- timviec365
- vieclamtuoitre
- careerbuilder
- vieclam24h
- jobsgo

Sau đó ta "làm sạch" dữ liệu, phân loại, lưu trữ và biểu diễn thành một cơ sở dữ liệu để phục vụ cho nhiều mục đích khác nhau, đồng thời ta sẽ phát triển một website sử dụng dữ liêu crawl được.

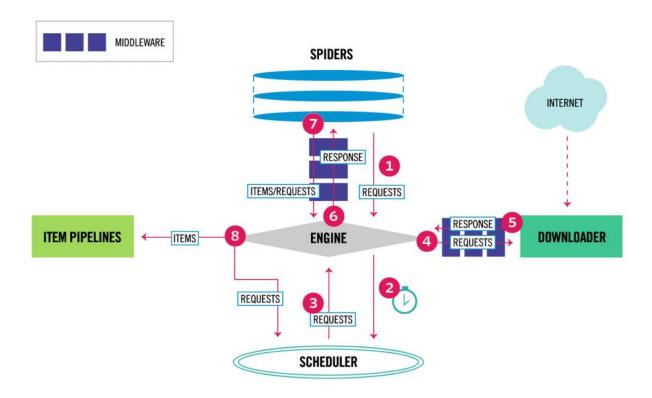
2 Giới thiệu scrapy

2.1 Scrapy:

Scrapy là một Python Framework mã nguồn mở hỗ trợ cho việc Crawling và Scraping dữ liệu từ các trang Web bằng việc download HTML và extract data từ chúng.

Ta muốn lấy dữ liệu từ các website nhưng dữ liệu đó quá lớn để copy rồi paste vào database, scrapy hỗ trợ ta làm điều đó. Việc lấy dữ liệu website hoàn toàn tự động nhanh chóng và việc sử dụng scrapy cũng rất đơn giản giúp ta tiếp kiệm được nhiều thời gian và công sức.

Ưu điểm của Scrapy là cung cấp sẵn 1 cấu trúc tương đối hoàn chỉnh để thực hiện việc crawl và scrape data, người dùng chỉ cần bổ sung thêm định nghĩa về dữ liệu cần lấy là xong (ví dụ như URL bắt đầu là gì, link chuyển qua trang mới là gì, các thông tin cần lấy ở mỗi trang là gì).



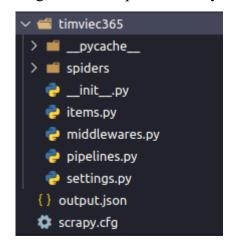
Hình 2.1: Quá trình từ lúc bắt đầu request cho đến lúc extract thông tin thành công ,luồng dữ liệu được thực thi qua các engine.

2.2 Crawl data bằng scrapy

Trong folder timviec365 (một project scrapy được tạo bằng lệnh *scrapy startproject timviec365*) sẽ có:

- scrapy.cfg: file config project scrapy
- settings.py: file chứa settings cho spiders
- middlewares.py: file chứa spider middlewares, là những framework được hook vào scrapy processing.
- items.py: file chứa cấu trúc của item mà bạn sẽ tách dữ liệu. Nói đơn giản thì nó là một cái khung, sau khi chúng ta lấy được dữ liệu từ spiders, chúng ta sẽ đặt dữ liệu vào đây và xử lý

- pipelines.py: sau khi chúng ta đặt dữ liệu vào cho items.py, chúng ta sẽ xử lý ở function process_items trong pipeline. Bạn có thể save vào database, hoặc chỉ trả về item đó
- folder spiders: là nơi chúng ta tạo file spider để chạy dữ liệu.



Trong chủ đề crawl data từ các trang web tìm kiếm việc làm ở đồ án này, ta sẽ lấy về những thông tin cơ bản sau:

- Tên công việc (Name)
- Tên công ty (Company)
- Mức lương (Salary)
- Địa điểm (Location

Ví dụ:



- Tên công việc (Name): Nhân Viên Kinh Doanh, Nhân Viên Văn Phòng
- Tên công ty (Company): BĐS Minh Quân
- Mức lương (Salary): 7 − 10 triệu
- Địa điểm (Location): Bình Dương

Kết quả thu được sẽ được lưu vào file json có dạng:

```
('Nhân viên kế toán', 'Cty Tnhh Phạm Hoàng Gia', '5 - 7 triệu ', 'Hồ Chí Minh'),

('Tuyển nhân viên mạng 3 SEO 1 ADW 3 FB ADS', 'Công Ty Tmhh Mạng Quảng Cáo', '7 - 10 triệu ', 'Hồ Chí Minh'),

('Kế toán kho', 'Cty ngọc giàu', '7 - 10 triệu ', 'Hồ Chí Minh'),

('Nhân viên Kinh doanh Bất Động Sản', 'Công ty Cổ Phần Địa ốc Tháng Lợi - HCM', '7 - 10 triệu ', 'Hồ Chí Minh'),

('Nhân viên kinh doanh Có lương cứng TPHCM', 'Công Ty Cổ Phần Tập Đoàn Địa Ốc Kim Tinh', '7 - 10 triệu ', 'Hồ Chí Minh'),

('10-15tr NHÂN VIÊN TƯ VĂN PV ĐI LÀM LUÔN - TP.HCM', 'Công ty CP dịch vụ tư vấn diamond star', '10 - 15 triệu ', 'Hồ Chí Minh'),

('Tuyển chuyên viên tư vấn thu nhập không giới hạn', 'SAIGONGOLDLAND', 'Trên 30 triệu ', 'Hồ Chí Minh'),

('Tổng đài viên', 'Công ty Cổ phần Công nghệ F.House', '10 - 15 triệu ', 'Hồ Chí Minh'),

('Sale Admin', 'Công ty Cổ phần Công nghệ F.House', '10 - 15 triệu ', 'Hồ Chí Minh'),

('Primary Campus Manager', 'Trường Tiều Học, Thcs Và Thpt Quốc Tế Trường Quốc Tế Tis', 'Trên 30 triệu ', 'Hồ Chí Minh'),

('NHÂN VIÊN KINH DOANH', 'Công Ty Tnhh Xong Hậu', '7 - 10 triệu ', 'Hồ Chí Minh'),
```

3 Xây dựng cơ sở dữ liệu crawl được

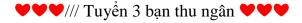
3.1 Xây dựng cơ sở dữ liệu

Sau khi đã crawl được data từ source code và lưu thành file json ta có thể tiến đến việc tạo thành một cơ sở dữ liệu lớn để dễ quản lý đồng thời dễ sử dụng:

Trước khi đưa dữ liệu vào database ta sẽ tiến hành tiền xử lý, kiểm tra và loại bỏ dữ liệu trùng nhau bằng cách sử dụng regular expression trên công cụ tìm kiếm của visual studio code và công cụ loại trừ trùng lặp của MS Excel, loại bỏ những items có thuộc tính null, loại bỏ những mô tả thừa (icon emotion, các ký hiệu đặc biệt,....) trong các thuộc tính của items.

VD:

Trước khi làm sach:



Sau khi làm sach:

Tuyển 3 bạn thu ngân

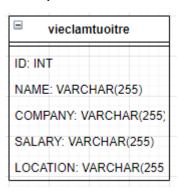
Kết quả thu được sau khi crawl:

Tên web	Dữ liệu crawl	Số dữ liệu trùng lặp đã	Số dữ liệu có thông tin dư
	được	phát hiện được	thừa, có thuộc tính null đã
			phát hiện được
Timviec365	11970	38	58
Vieclamtuoitre	11272	5	70
Jobsgo	14120	2399	42
Careerbuilder	19950	1300	30
Vieclam24h	1340	12	0

Sau đó, ta tạo 1 database tên là scrapy chứa các table chứa dữ liệu crawl được, mỗi table chứa hơn 10000 dữ liệu crawl được từ 1 trang web.

Table có các thuộc tính

- ID là khóa chính mã của item.
- NAME là tên việc làm.
- COMPANY là tên công ty tuyển dụng.
- SALARY là mô tả về lương.
- LOCATION là địa điểm làm việc.



Hình 3.1.1: Thiết kế của một table.

Sau đó ta đưa dữ liệu vào MySQL bằng lệnh INSERT INTO để tạo database phục vụ cho việc lập trình web

ID	NAME	COMPANY	SALARY	LOCATION
1	Chuyên viên phát triển khách hàng chiến lược	Công ty Cổ Phần Tập Đoàn Bizman	12 Tr - 50 Tr VND	Hà Nội, Hồ Chí Minh
2	Business Development Manager	Monroe Consulting Group Vietnam	Cạnh tranh	Đà Nẵng
3	Nhân Viên Kinh Doanh	Công Ty TNHH TM SX Me Non	Cạnh tranh	Hồ Chí Minh
4	Nhân Viên Kinh Doanh ngành vật liệu xây dựng	Công ty Cổ phần Đầu tư Kelsey	8 Tr - 20 Tr VND	Hồ Chí Minh
5	Sales Engineer	Ishida Vietnam Co Ltd	10 Tr - 13 Tr VND	Hồ Chí Minh

Hình 3.1.2: Một table trong MySQL.

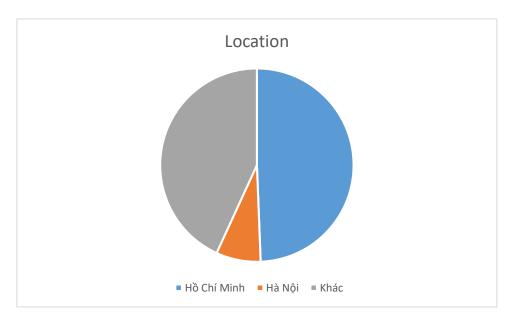
3.2 Phân tích dữ liệu crawl được

Từ dữ liệu crawl được ta đã xây dựng được một database có gần 60000 items, từ đó ta có thể nhân xét được một số khía cạnh sau:

Về địa điểm làm việc:

Hồ Chí Minh	Hà Nội	Khác
33819	11343	13419

Sử dụng công cụ biểu đồ của Excel ta sẽ có được biểu đồ hình quạt thể hiện tỉ số % nhu cầu tuyển dụng ở các tỉnh thành:



Hình 3.2.1: Biểu đồ quạt thể hiện tỉ trọng nhu cầu tuyển dụng theo khu vực

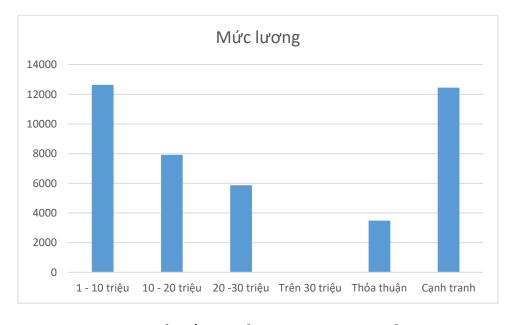
Nhận xét:

Từ biểu đồ ta thấy được nhu cầu tuyển dụng của thành phố Hồ Chí Minh chiếm cao nhất (chiếm 58%), thứ hai là nhu cầu tuyển dụng từ các tỉnh thành khác trên cả nước như: Hải Phòng, Tiền Giang, Đà Nẵng, Lâm Đồng,.... (chiếm 23%), còn nhu cầu tuyển dụng của Hà Nội (chiếm 19%) đứng ở vị trí cuối cùng.

Qua đó ta có thể nhận xét nhu cầu tuyển dụng lao động ở nước ta những năm gần đây thì thành phố Hồ Chí Minh có xu thế áp đảo những địa phương còn lại. Nguyên do vì chênh lệch trình độ phát triển khoa học kỹ thuật, trình độ kinh tế giữa các vùng miền với nhau, vốn đầu tư nước ngoài mở rộng thị trường, khuynh hướng lựa chọn nơi làm việc của người lao động, mức sống, mức lương của công việc... từ đó dẫn đến những khác biệt về nhu cầu tuyển dụng lao động phục vụ phát triển kinh tế - xã hội.

Về mức lương:

1 – 10 triệu	10 – 20 triệu	20 – 30 triệu	Trên 30	Thỏa thuận	Cạnh tranh
			triệu		
12632	7921	5873	2	3485	12453



Hình 3.2.2: Biểu đồ cột thể hiện mức lương tuyển dụng.

Nhân xét:

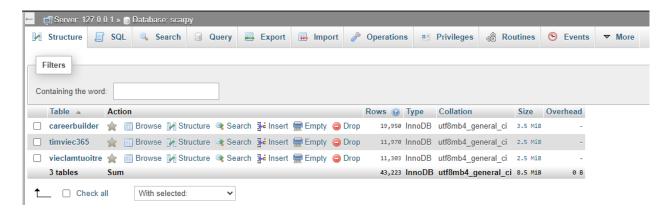
Từ biểu đồ trên ta thấy mức lương cạnh tranh chiếm cao nhất (12453), thứ hai là mức lương 1-10 triệu (12632), thứ ba là mức lương 10-20 triệu (7921), tiếp theo là mức lương 20-30 triệu (5873), kế đó là mức lương thỏa thuận (3485), đứng cuối là mức lương trên 30 triệu (2).

Qua đó ta nhận xét thấy các doanh nghiệp hiện nay có xu hướng đưa ra mức lương cạnh tranh với nhau để thu hút người tài về với mình, những công ty có mức lương đưa ra cao hơn sẽ thu hút nhiều người hơn. Trong việc tuyển dụng online thì những công việc với mức lương trung bình (1- 10 triệu) chiếm đa số, chủ yếu là những việc như: thực tập sinh, làm part-time. Tương tự nên nhu cầu tuyển dụng online với mức lương (10 – 20 triệu, 20 – 30 triệu) cũng giảm dần, những việc làm gắn với mức lương này thường là nhân viên kỹ thuật, nhân viên sale,... Mức lương thỏa thuận được một số doanh nghiệp đưa ra trong quá trình phỏng vấn và ký kết hợp đồng lao động. Cuối cùng là những việc làm với mức lương cao trên 30 triệu thường ít được tuyển dụng online, những việc làm này chủ yếu như là chuyên viên bất đông sản (lương 50 triêu),...

Lưu ý: số liệu được đem ra phân tích không phải sử dụng hoàn toàn 100% số liệu từ 5 trang web jobsgo, careerbuilder, vieclamtuoitre, timviec365, vieclam24h. Ví dụ với việc phân tích mức lương thì một số trường hợp như: Mức lương 5 – 30 triệu, Trên 9 triệu sẽ không được lấy để xây dựng biểu đồ vì khó xác định nằm trong nhóm lương nào,....

4 Xây dựng ứng dụng với cơ sở dữ liệu crawl được 4.1 Giới thiêu Web

Sau khi đã xây dựng một database chứa dữ liệu đã crawl được thì ta tiến hành xây dựng web UI thao tác với database này.

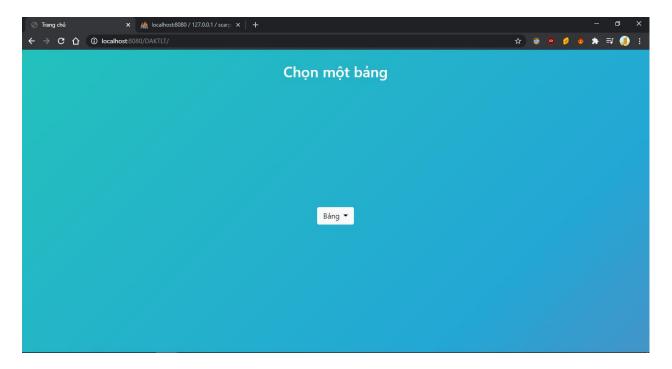


Hình 4.1.1: Database scarpy

Cơ sở dữ liệu scarpy với mỗi table chứa khoảng hơn 10000 dữ liệu từ các trang web crawl được và ta sẽ tiến hành xây dựng một web để thao tác với dữ liệu này bằng ngôn ngữ HTLM/CSS/PHP.

4.2 Một số tính năng của Web

Web UI có các chức năng chính: thêm/xóa/cập nhật, tìm kiếm, phân loại dữ liệu dựa trên các thuộc tính,.... các dữ liệu trong cơ sở dữ liệu đã xây dựng từ những tuần trước.



Hình 4.2.1: Trang chủ của trang web

4.2.1 Chức năng thêm dữ liệu vào database:

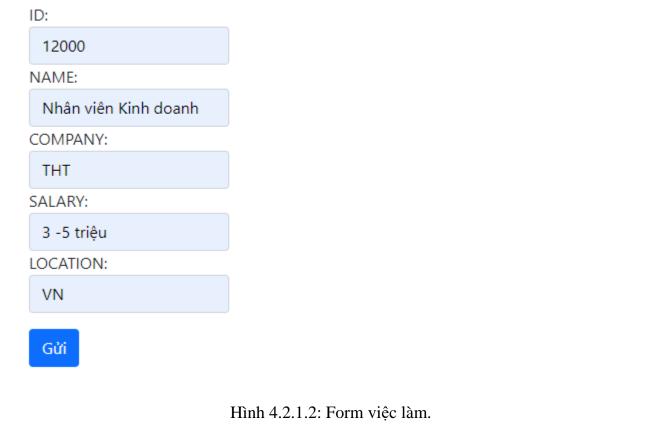
Thêm một bản ghi cho careerbuilder

id:	
NAME:	
COMPANY:	
SALARY:	
LOCATION:	
Gửi	

Hình 4.2.1.1: Thao tác thêm dữ liệu vào database

Ta nhập dữ liệu vào form và nhấn nút Gửi, Database sẽ tự động cập nhật dữ liệu

Thêm một bản ghi cho timviec365

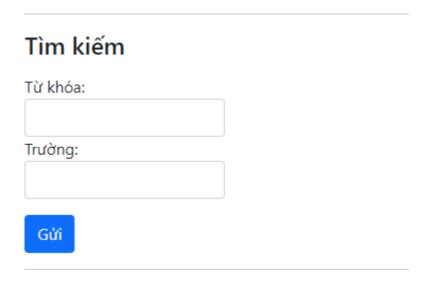


12000 Nhân viên Kinh doanh THT 3 -5 triệu VN

Hình 4.2.1.3: Dữ liệu đã được cập nhật trên database.

4.2.2 Chức năng tìm kiếm

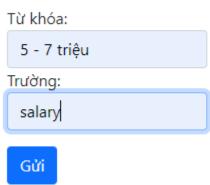
Giúp lọc ra thông tin người dùng muốn



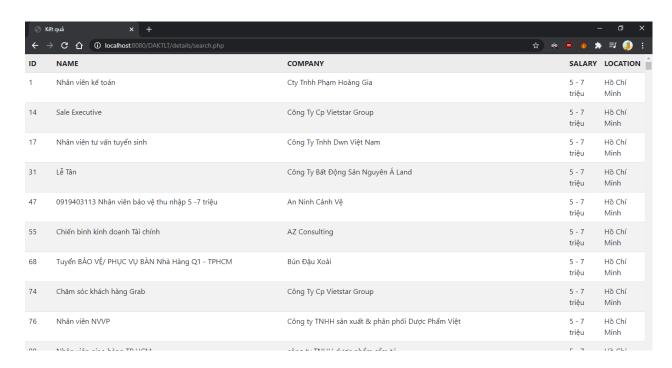
Hình 4.2.2.1: Thao tác tìm kiếm dữ liệu trên database.

VD: Ta muốn tìm kiếm những công việc có mức lương 5-7 triệu, ta sẽ nhập như sau:

Tìm kiếm



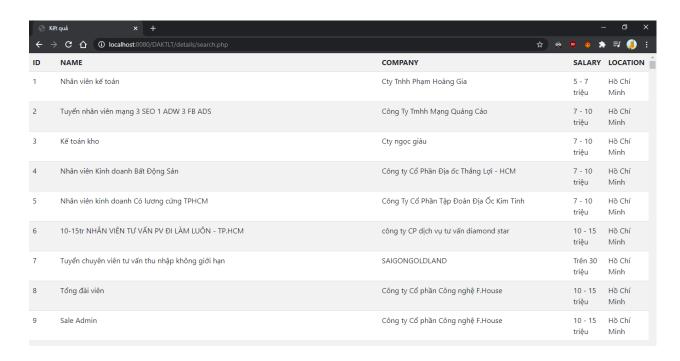
Kết quả trả về như sau:



hoặc tìm kiếm những công việc ở thành phố Hồ Chí Minh

Từ khóa: hồ chí minh Trường: location

Kết quả trả về như sau:



4.3.3 Chức năng cập nhật và xóa

Ở trang thao tác có hiển thị thông tin của từng bản ghi như sau:



Hình 4.3.3.1: Trang thao tác với database.

Ta nhấn vào nút xóa để xóa dữ liệu mình muốn



Hình 4.3.3.2: Một item trong database được biểu diễn trên web.

Tương tự nhấn vào nút Cập nhật để điền lại form:



Hình 4.3.3.3: Form cập nhật item.

5 Đánh giá kết quả thu được và hướng phát triển trong tương lai
5.1 Đánh giá kết quả thực hiện được
Ưu điểm:

- Crawl được hơn 10000 items từ các trang jobsgo, vieclamtuoitre, timviec365, careerbuilder và khoảng 1500 items từ vieclam24h.
- Xây dựng được database với những dữ liệu crawl được từ các trang web trên.

• Xây dựng được một website sử dựng database đã được xây dựng.

Nhược điểm:

- Trong quá trình crawl và xây dựng database chỉ phát hiện được những dữ liệu giống nhau 100% về ngữ pháp, còn những dữ liệu giống nhau về mặt ý nghĩa vẫn không phát hiện được. Ví dụ: ta có 2 việc làm giống nhau như "Nhân viên bán hàng qua điện thoại" và "Nhân viên telesale" thì ta vẫn không thể phát hiện là trùng lặp được.
- Một số trường hợp đặc biệt vẫn chưa tìm được cách xử lý như có một công ty tuyển 2 nhân viên phục vụ, thay vì chỉ cần đăng 1 tin với nội dung là tuyển 2 nhân viên phục vụ nhưng công ty đó lại lần lượt đăng 2 tin, mỗi tin tuyển 1 nhân viên phục vụ với mô tả như nhau, trường hợp này nếu xử lý trùng lặp theo mặt ngữ pháp thì 2 tin này là trùng lặp, nhưng thực tế thì không phải.
- Vì chưa được trang bị nhiều kiến thức về lập trình web nên giao diện trang web còn thô sơ, chưa được tối ưu, đẹp mắt.

5.2 Hướng phát triển trong tương lai

Từ database ta xây dựng được ta có thể tiến hành xây dựng một trang web hoặc app di động gợi ý công việc cho những đối tượng đang có nhu cầu tìm kiếm việc làm. Xa hơn ta có thể kết hợp với AI để xây dựng những con bot trò chuyện có khả năng giải đáp thắc mắc, gợi ý cho nhà tuyển dụng đưa ra thông tin tuyển dụng phù hợp lẫn giúp đỡ người lao động tìm kiếm việc làm phù hợp với nhu cầu công việc, mức lương, nơi sinh sống.

6 Tài liệu tham khảo

- [1] Bộ slide Cơ sở dữ liệu, ĐH Bách Khoa, ĐHQG-TPHCM.
- [2] Scrapy công cụ bóc tách dữ liệu siêu mạnh mẽ.

https://www.hocpython.com/blog/scrapy-cong-cu-boc-tach-du-lieu-sieu-manh/

[3] Bộ video lập trình PHP&MySQL qua ví dụ thiết kế web

 $\frac{https://www.youtube.com/watch?v=7x1PDHsQyGw\&list=PLzrVYRai0riRzNtnHgc1OE}{OmRsJ_9lN3p}$

BẢNG PHÂN CÔNG CÔNG VIỆC

Crawl data từ vieclamtuoitre, timviec365.	
Xây dựng database trên mysql.	
Lập trình web.	
Video demo.	
Viết báo cáo.	
Crawl data từ careerbuilder, vieclam24h, jobgo.	
Xây dựng database trên mysql.	
Crawl data từ careerbuilder, vieclam24h, jobgo.	
Xây dựng database trên mysql.	
Slide.	