HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐỖ THỊ HỒNG NGÁT

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG TỰ ĐỘNG TỔNG HỢP THÔNG TIN VỀ DỊCH BỆNH SỬ DỤNG SCRAPY FRAMEWORK

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2024

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Vũ Văn Thoả

Phản biện 1: PGS.TS. Phan Xuân Hiếu

Phản biện 2: PGS.TS. Hoàng Xuân Dậu

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 10 giờ 45 phút ngày 20 tháng 03 năm 2024

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Cùng với sự cải thiện đáng kể về chất lượng cuộc sống, ngày nay, sức khỏe cá nhân và của gia đình trở thành ưu tiên hàng đầu của mọi người. Vào thời đại số hóa, Internet đã trở thành một nguồn thông tin vô cùng quý báu về y tế và sức khỏe. Tuy nhiên, sự phát triển của công nghệ đã tạo điều kiện cho xuất hiện đông đảo các nguồn thông tin trên mạng, và điều này đặt ra một thách thức lớn: làm thế nào để phân biệt thông tin chính xác và đáng tin cậy từ những thông tin không chính thống hoặc giả mạo.

Năm 2020 chứng kiến một trong những biến đổi lớn nhất trong lịch sử loài người - đại dịch Covid-19. Tình hình khẩn cấp này đã buộc thế giới phải thích nghi nhanh chóng với mô hình trực tuyến. Việt Nam, không ngoại lệ, đã tận dụng công nghệ thông tin và Internet để đối phó với đại dịch này. Các công văn quan trọng của chính phủ đã được chuyển từ dạng giấy sang công điện, tiết kiệm thời gian và tài nguyên. Trong bối cảnh đó, sự lan rộng của thông tin chính xác và đáng tin cậy liên quan đến dịch bệnh trở nên cực kỳ quan trọng. Chính phủ và các cơ quan chức năng đã thiết lập các kênh truyền thông trực tuyến chính thức như các trang Facebook để đính chính và cung cấp thông tin đáng tin cậy về dịch bệnh. Điều này đã giúp người dân có nguồn thông tin tin cậy để nắm bắt tình hình dịch bệnh, biện pháp phòng chống, và hướng dẫn cách bảo vệ bản thân và cộng đồng.

Mặt khác, không chỉ trong thời của đại dịch, việc sử dụng công nghệ thông tin và truyền thông trực tuyến đã và đang trở thành một phần quan trọng của cuộc sống hiện đại. Nó mang lại sự linh hoạt, tiết kiệm thời gian và tạo điều kiện thuận lợi cho việc truyền tải thông tin đến mọi người một cách nhanh chóng và hiệu quả.

Trước những yêu cầu trên, hệ thống tự động tổng hợp thông tin về dịch bệnh ra đời để tổng hợp và cung cấp thông tin về dịch bệnh chính thống, đáng tin cậy và dễ dàng tiếp cận cho mọi người. Người dân có thể sử dụng hệ thống

này để tìm kiếm thông tin về tình hình dịch bệnh, những bài viết hướng dẫn sức khỏe hữu ích và thậm chí là các bệnh viện, cơ sở y tế có chuyên môn. Các bác sĩ và nhân viên y tế có thể sử dụng hệ thống để cung cấp thông tin đáng tin cậy cho bệnh nhân và cộng đồng. Chính quyền địa phương có thể tích hợp hệ thống này vào hệ thống quản lý y tế của họ để cung cấp dịch vụ tốt hơn cho người dân.

Với những lý do trên, học viên chọn đề tài "**NGHIÊN CỨU XÂY DỰNG HỆ THỐNG TỰ ĐỘNG TỔNG HỢP THÔNG TIN VỀ DỊCH BỆNH SỬ DỤNG SCRAPY FRAMEWORK**" làm đề án tốt nghiệp cao học của mình.

*** Mục đích, đối tượng và phạm vi nghiên cứu**

- Mục đích nghiên cứu:

+ Khảo sát về hệ thống thông tin, xu hướng phát triển hệ thống tự động tổng hợp thông tin về dịch bệnh trên thế giới và ở Việt Nam.

+ Nghiên cứu về cấu trúc, phương thức hoạt động của Scrapy framework phục vụ cho việc thu thập dữ liệu tự động và một số công cụ, công nghệ để phân tích, thiết kế, xây dựng hệ thống website.

+ Dự kiến sẽ xây dựng một hệ thống website theo kiến trúc tiểu dịch vụ hoàn thiện, với các chức năng cơ bản dành cho người dùng như: tìm kiếm bệnh viện, bài báo, đăng ký tài khoản, thay đổi thông tin cá nhân, đổi mật khẩu, ... Những chức năng cho quản trị viên như: quản lý danh sách bệnh viện, quản lý danh sách bài báo, quản lý người dùng, thu thập dữ liệu tự động tại các trang web khác…Ngoài ra hệ thống sẵn sàng để có thể mở rộng, nâng cấp để phục vụ nhiều người, tích hợp với các hệ thống có sẵn tại các cơ sở triển khai để quản lý dễ dàng hơn.

- Đối tượng nghiên cứu: Nghiên cứu Scrapy framework, những bệnh viện trên cả nước Việt Nam, các tin tức dịch bệnh bằng tiếng việt.

- ***Phạm vi nghiên cứu:*** cấu trúc, phương thức hoạt động của Scrapy framework, và đề xuất mô hình thử nghiệm hệ thống tại bệnh viện trên cả nước Việt Nam, song tập trung tại các thành phố lớn, nơi có nhiều cơ sở y tế đủ điều kiện triển khai hệ thống.

*** Phương pháp nghiên cứu:**

- ***Về mặt lý thuyết***: tập hợp, khảo sát, phân tích các tài liệu và thông tin có liên quan đến Scrapy framework, các trang web chính thống có thông tin về bệnh viện và tin tức dịch bệnh.

- ***Về mặt thực nghiệm***: Khảo sát tình hình thực tế thu thập dữ liệu tại một số trang web chính thống và đưa ra đề xuất giải pháp phù hợp để triển khai hệ thống tổng hợp thông tin về dịch bệnh.

*** Cấu trúc của đề án gồm 3 chương chính:**

Chương 1: Tổng quan về hệ thống tự động tổng hợp thông tin về dịch bệnh và các vấn đề liên quan

Chương 2: Phân tích và thiết kế hệ thống

Chương 3: Triển khai và xây dựng hệ thống

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG TỰ ĐỘNG TỔNG HỢP THÔNG TIN VỀ DỊCH BỆNH VÀ CÁC VẤN ĐỀ LIÊN QUAN

Nội dung Chương 1 của đề án tập trung vào việc giới thiệu chung về hệ thống tổng hợp thông tin, khảo sát tổng quan về Scrapy framework và các vấn đề liên quan. Nội dung chương 1 sẽ làm cơ sở cho các nghiên cứu tiếp theo của đề án.

1.1. Giới thiệu chung về hệ thống tổng hợp thông tin

Trong mục này đề án sẽ khảo sát các khái niệm liên quan đến hệ thống tổng hợp thông tin, các hệ thống đã được triển khai trong thực tế, một số công nghệ được sử dụng, hình thức và đối tượng triển khai trong hệ thống tổng hợp thông tin. Từ đó thực hiện phân tích một số ưu điểm và hạn chế của hệ thống tổng hợp thông tin.

1.1.1. Khái niệm về hệ thống tổng hợp thông tin

Hệ thống tổng hợp thông tin là một cụm từ mô tả một loạt các công nghệ, quy trình và phương pháp được sử dụng để tổng hợp, xử lý và trình bày thông tin từ nhiều nguồn khác nhau một cách có tổ chức và hợp lý. Mục đích chính của hệ thống này là giúp người dùng thu thập và tiếp cận thông tin một cách hiệu quả [1]. Hệ thống tổng hợp thông tin thường bao gồm các thành phần sau: thu thập thông tin, xử lý và phân tích thông tin, trình bày thông tin, lưu trữ và quản lý dữ liệu

1.1.2. Các hệ thống tổng hợp thông tin đã được triển khai trong thực tế

- Google Search: Google sử dụng thuật toán phức tạp để tổng hợp và hiển thị các kết quả tìm kiếm từ hàng tỷ trang web trên Internet.

- RSS Readers: Các ứng dụng RSS như Feedly cho phép người dùng tổng hợp và đọc các tin tức từ nhiều nguồn khác nhau trong một nền tảng.

- Hệ thống quản lý kiến thức: Như Microsoft SharePoint, Confluence của Atlassian, giúp tổ chức và chia sẻ thông tin nội bộ doanh nghiệp.

- Hệ thống gợi ý nội dung cá nhân: Như Spotify, Netflix sử dụng dữ liệu cá nhân hóa để tổng hợp và gợi ý nội dung phù hợp với người dùng.

1.1.3. Một số công nghệ sử dụng trong hệ thống tổng hợp thông tin

- Thuật toán Machine Learning và AI
- Natural Language Processing (NLP)
- Công nghệ Big Data
- Web Scraping và Data Crawling

1.1.4. Hình thức triển khai và đối tượng của hệ thống tổng hợp thông tin

- Hình thức triển khai của hệ thống [1]: ứng dụng web, ứng dụng di động, phần mềm trên máy tính.

- Đối tượng của hệ thống [1]: cá nhân, doanh nghiệp, cơ quan chính phủ và tổ chức phi chính phủ, người tiêu dùng và công chúng.

1.1.5. Ưu điểm và hạn chế của các hệ thống tổng hợp thông tin

- Ưu điểm [1]: tăng hiệu suất tìm kiếm và tiếp cận thông tin, tổ chức thông tin hiệu quả, tính cá nhân hóa

- Hạn chế [1]: nguy cơ thông tin không chính xác hoặc thiếu trung thực, nguy cơ mất quyền riêng tư, hạn chế về phạm vi và độ chi tiết.

1.2. Xu hướng phát triển hệ thống tự động tổng hợp thông tin về dịch bệnh trên thế giới và tại Việt Nam.

1.2.1. Xu hướng phát triển hệ thống tổng hợp thông tin về dịch bệnh trên thế giới

Các quốc gia phát triển trên toàn cầu đã ứng dụng mạnh mẽ hệ thống tổng hợp thông tin về dịch bệnh để cung cấp thông tin, cập nhật tình hình dịch bệnh đang diễn ra cho quốc gia của mình.

Hệ thống tổng hợp thông tin về dịch bệnh tại Mỹ: CDC COVID Data Tracker, COVID Symptom Tracker, …

Hệ thống tổng hợp thông tin về dịch bệnh tại Trung Quốc: Trung Quốc có các hệ thống tổng hợp thông tin về nhiều loại dịch bệnh truyền nhiễm, từ cúm đến bệnh sốt xuất huyết.

Hệ thống tổng hợp thông tin về dịch bệnh tại Hàn Quốc: KCDC COVID-19 và các trang web của cơ quan y tế quốc gia.

1.2.2. Thực trạng phát triển và ứng dụng hệ thống tổng hợp thông tin về dịch bệnh tại Việt Nam

Nhiều ứng dụng di động được phát triển tại Việt Nam để cung cấp thông tin về dịch bệnh và hỗ trợ trong việc theo dõi sức khỏe cá nhân, như ứng dụng Bluezone và Vietnam Health Declaration.

Việt Nam đã hợp tác với các tổ chức quốc tế như WHO và CDC để chia sẻ dữ liệu và kinh nghiệm trong việc phát triển và ứng dụng hệ thống tổng hợp thông tin về dịch bệnh giúp nâng cao khả năng phát triển và cập nhật thông tin, đồng thời đảm bảo tính chính xác và tin cậy của dữ liệu.

Mặc dù đã có sự phát triển, nhưng vẫn còn thách thức trong việc nâng cao tính chính xác và khả năng phản ứng của các hệ thống này trong bối cảnh các biến thể mới của virus và tình hình dịch bệnh biến động liên tục.

1.3. Các công nghệ sử dụng

Trong mục này đề án sẽ đi sâu và chi tiết về Scrapy framework, đồng thời trình bày một số nội dung về các công nghệ sử dụng liên quan để xây dựng hệ thống tự động tổng hợp thông tin về dịch bệnh.

1.3.1. Scrapy framework

Hệ thống tự động tổng hợp thông tin về dịch bệnh sử dụng công nghệ xử lý, thu thập dữ liệu phổ biến, mạnh mẽ Scrapy Framework để thu thập dữ liệu về dịch bệnh, cụ thể là dịch sốt xuất huyết.

1.3.1.1. Lịch sử hình thành

1.3.1.2. Công nghệ sử dụng trong Scrapy framework

Dưới đây là một số công nghệ chính được sử dụng trong Scrapy [12]:

- Python: Scrapy được viết bằng ngôn ngữ lập trình Python.

- Twisted: Twisted cung cấp khả năng xử lý đa luồng và không đồng bộ, giúp Scrapy thu thập dữ liệu từ nhiều trang web cùng một lúc.

- XPath và CSS Selectors: Scrapy hỗ trợ việc trích xuất dữ liệu từ các trang web thông qua việc sử dụng XPath và CSS Selectors.

- Middleware: Middleware cho phép tùy chỉnh và mở rộng khả năng của Scrapy bằng cách thêm các chức năng tùy chỉnh, như xử lý lỗi, lọc dữ liệu, và ghi log.

- Pipelines: Pipelines cho phép bạn thực hiện các xử lý như lọc dữ liệu, xử lý lỗi, và lưu trữ vào cơ sở dữ liệu.

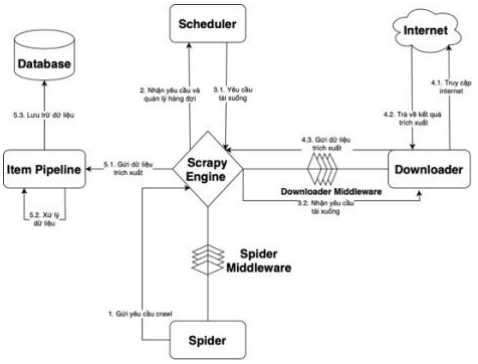
1.3.1.3. Các tính năng của Scrapy framework

Các tính năng chính của Scrapy framework [12]: crawling đa luồng (Multithreaded Crawling), cấu trúc cơ bản, thích ứng với thay đổi trang web, định vị dữ liệu (XPath và CSS Selectors), middleware và Pipelines, đa nền tảng, thiết kế modul và mở rộng

1.3.1.4. Các thành phần của Scrapy framework

1.3.1.5. Luồng dữ liệu của Scrapy framework

Hình 1.1 dưới đây mô tả luồng dữ liệu trong Scrapy framework:



Hình 1.1: Mô tả luồng dữ liệu trong Scrapy framework

1.3.2. Kiến trúc tiểu dịch vụ

Kiến trúc tiểu dịch vụ (Microservices[8]) là một kiểu kiến trúc mà cấu trúc hệ thống như một tập các dịch vụ. Vì vậy, nó giúp giảm thiểu quá trình phức tạp hóa trong các hệ thống lớn, như thuật toán chia để trị (divide and conquer), kiến trúc này chia hệ thống thành các phần nhỏ, độc lập để có thể dễ dàng đóng gói, triển khai, quản lý, kiểm thử, nâng cấp, bảo trì.

1.3.3. Ngôn ngữ Java, Spring Framework

Back-end của hệ thống tự động tổng hợp thông tin về dịch bệnh được phát triển dựa trên ngôn ngữ Java[4] kết hợp sử dụng Spring Framework[10] là một mã nguồn mở phát triển ứng dụng phổ biến cho Java Enterprise. Spring có rất nhiều các dự án con, giúp cho việc xây dựng hệ thống một cách dễ dàng: Spring MVC (thiết kế dành cho việc xây dựng web), Spring Security (cung cấp các cơ chế xác thực, phân quyền), Spring Boot (giúp phát triển, chạy ứng dụng một cách nhanh chóng), Spring Data (cung cấp các công nghệ để truy cập dữ liệu), Spring Cloud (cung cấp các công cụ để phát triển hệ thống phân tán),…

1.3.4. Ngôn ngữ TypeScript, thư viện Redux

Front-end của hệ thống được xây dựng dựa trên ngôn ngữ TypeScript[3] kết hợp với việc sử dụng Redux[2] là một thư viện độc lập, có thể sử dụng với bất kì khung giao diện (UI layer) hay framework nào.

1.3.5. Ngôn ngữ Python, Flask Framework

Thực thể dịch vụ này sử dụng Framework Flask[5] của Python[7] để tạo ra một ứng dụng web, giao tiếp với các ứng dụng web của thực thể dịch vụ triển khai bằng Java đã mô tả ở phía trên.

1.3.6. Hệ quản trị cơ sở dữ liệu MySQL

Hệ thống lưu trữ dữ liệu về thông tin người dùng, lịch sử thu thập dữ liệu của quản trị viên thông qua MySQL[9].

1.3.7. Elasticsearch

Ngoài lưu trữ dữ liệu qua MySQL, hệ thống lưu trữ dữ liệu bệnh viện, bài báo về dịch bệnh trên Elasticsearch [11].

1.4. Kết luận chương 1

Chương 1 của đề án đã tiến hành một khảo sát toàn diện về hệ thống tổng hợp thông tin, Scrapy framework và các vấn đề liên quan. Đề án đã khảo sát xu hướng phát triển hệ thống tổng hợp thông tin về dịch bệnh trên toàn cầu cũng như tình hình phát triển và ứng dụng của hệ thống tổng hợp thông tin về dịch bệnh tại Việt Nam. Đề án cũng nghiên cứu về Scrapy framework và các công nghệ để triển khai hiệu quả hệ thống tự động tổng hợp thông tin về dịch bệnh trong thực tế.

Dựa vào nội dung chương 1, các vấn đề liên quan đến hệ thống tự động tổng hợp thông tin về dịch bệnh sẽ được phân tích và thiết kế chi tiết trong chương 2.

CHƯƠNG 2: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Trong chương 2, tác giả phân tích và thiết kế hệ thống tự động tổng hợp thông tin về dịch bệnh, xác định tổng quan hệ thống, đặc tả chi tiết và biểu đồ các ca sử dụng, xây dựng kiến trúc hệ thống và mẫu đặc tả dữ liệu để trích xuất thông tin về dịch bệnh cụ thể là dịch sốt xuất huyết.

2.1. Tổng quan về hệ thống

Về giao diện: Hệ thống sẽ cung cấp các cửa sổ cần thiết để người dùng tương tác, màu sắc mẫu nhập đơn giản dễ dàng nhận diện.

Về thao tác: Hệ thống cung cấp các thao tác đơn giản, tìm kiếm bệnh viện, các bài báo cho người dùng, cung cấp các công cụ đa dạng để quản trị viên quản lý nội dung trang web và thu thập dữ liệu.

2.1.1. Yêu cầu chức năng

Hệ thống có các chức năng chính sau:

Chức năng tìm kiếm: Người dùng có thể tìm kiếm bệnh viện theo tên, địa chỉ, số điện thoại, website, theo khoa và các dịch vụ khám bệnh. Hay tìm kiếm các bài báo theo tên bệnh, bệnh viện, nội dung bài báo…

Chức năng quản lý hệ thống: Quản trị viên có quyền quản lý dữ liệu về bệnh viện, bài báo, quản lý người dùng, cấp quyền quản trị viên cho người dùng. Hệ thống cho phép quản trị viên có thể tìm kiếm, xem chi tiết, sửa, xóa, thêm thông tin bệnh viện, bài báo, tìm kiếm, xem thông tin người dùng.

Chức năng hỗ trợ thu thập dữ liệu: Quản trị viên có thể thu thập dữ liệu từ các trang web tin cậy khác.

2.1.2. Yêu cầu phi chức năng

2.2. Xác định danh sách tác nhân và ca sử dụng

Hệ thống có 4 tác nhân sau đây:

Quản trị viên (admin): Là người chịu trách nhiệm quản lý hệ thống về mặt nội dung và kỹ thuật.

Người vãng lai (guest): Là người truy cập hệ thống nhưng chưa đăng ký tài khoản.

Người dùng (user): Là người đã đăng ký tài khoản và có thể sử dụng thêm một số chức năng của hệ thống mà người vãng lai không có.

Các hệ thống liên kết ngoài (third-party system): Là những hệ thống ngoài có kết nối với hệ thống.

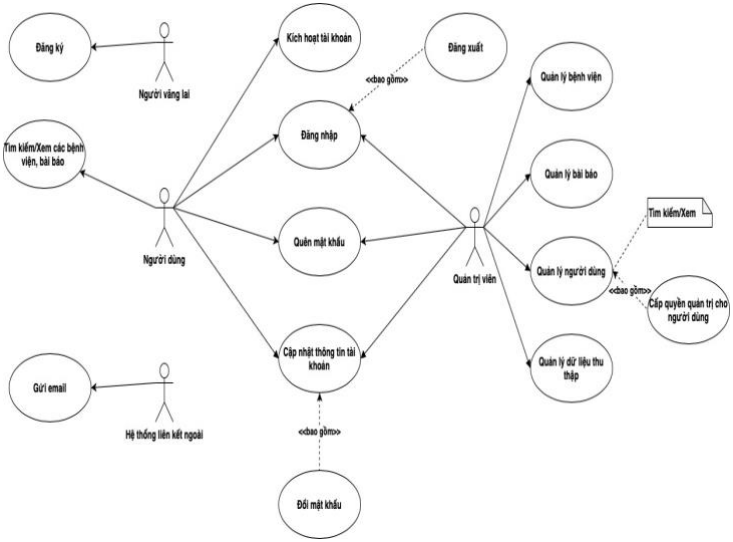
Hệ thống có 12 ca sử dụng chính, chia làm 3 nhóm:

Nhóm 1 là ca sử dụng liên quan đến tài khoản bao gồm: đăng ký, đăng nhập, quên mật khẩu, đổi mật khẩu, cập nhật thông tin tài khoản, đăng xuất.

Nhóm 2 là các ca sử dụng dùng cho người dùng bao gồm: tìm kiếm/xem bệnh viện, tìm kiếm/xem bài báo.

Nhóm 3 là các ca sử dụng dùng cho quản trị viên bao gồm: quản lý bệnh viện, quản lý bài báo, quản lý người dùng, quản lý dữ liệu thu thập.

Quan hệ giữa các tác nhân và ca sử dụng được mô tả ở hình 2.1.



Hình 2.1: Mô hình quan hệ các tác nhân và ca sử dụng

2.3. Đặc tả chi tiết và biểu đồ ca sử dụng

2.3.1. Đăng ký

Để sử dụng ứng dụng trước tiên tác nhân cần đăng ký tài khoản, ca sử dụng mô tả thao tác tác nhân đăng ký sử dụng.

2.3.2. Kích hoạt tài khoản

Để sử dụng tài khoản tác nhân cần xác thực email để kích hoạt tài khoản trước, ca sử dụng mô tả thao tác tác nhân kích hoạt tài khoản.

2.3.3. Quên mật khẩu

Ca sử dụng mô tả thao tác tác nhân quên mật khẩu.

2.3.4. Đăng nhập

Để sử dụng ứng dụng, tác nhân cần đăng nhập vào hệ thống, ca sử dụng mô tả thao tác tác nhân đăng nhập.

2.3.5. Đổi mật khẩu

Ca sử dụng mô tả thao tác tác nhân đổi mật khẩu.

2.3.6. Thay đổi thông tin cá nhân

Ca sử dụng mô tả thao tác tác nhân sửa thông tin cá nhân.

2.3.7. Đăng xuất

Ca sử dụng mô tả thao tác tác nhân đăng xuất khỏi hệ thống.

2.3.8. Tìm kiếm/xem bệnh viện/bài báo

Ca sử dụng mô tả thao tác tìm kiếm bệnh viện/bài báo.

2.3.9. Quản lý bệnh viện

Ca sử dụng mô tả thao tác tác nhân quản lý bệnh viện.

2.3.10. Quản lý bài báo

Ca sử dụng mô tả thao tác tác nhân quản lý bài báo.

2.3.11. Quản lý người dùng

Ca sử dụng mô tả thao tác tác nhân quản lý người dùng.

2.3.12. Quản lý dữ liệu thu thập

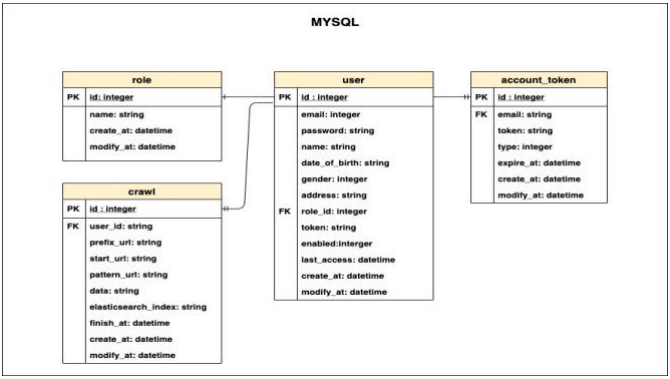
Ca sử dụng mô tả thao tác tác nhân thu thập dữ liệu từ các trang web khác về hệ thống.

2.4. Thiết kế cơ sở dữ liệu

2.4.1. Mô hình thực thể liên kết

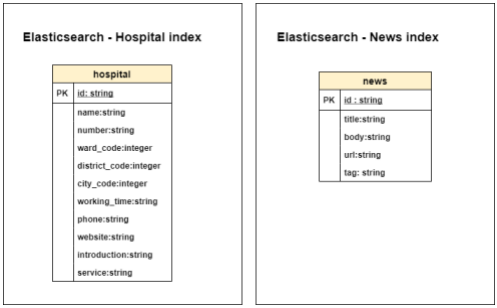
Cơ sở dữ liệu chứa thông tin người dùng, thông tin về lịch sử thu thập dữ liệu được lưu trong hệ quản trị cơ sở dữ liệu MySQL gồm 4 bảng: Role, User, Account_Token, Crawl.

Quan hệ của các bảng được mô tả như hình vẽ 2.14 bên dưới:



Hình 2.14: Mô hình thực thể liên kết

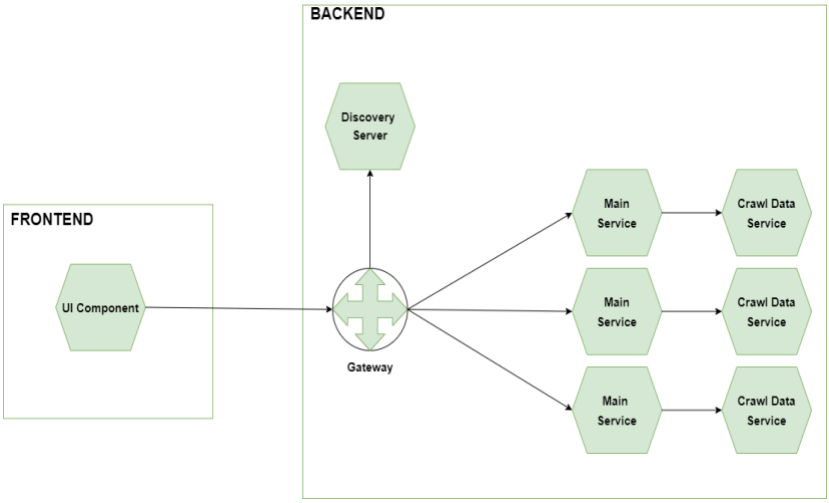
Dữ liệu bệnh viện, bài báo lưu trữ trên Elastricsearch như hình vẽ 2.15.



Hình 2.15: Mô tả dữ liệu chỉ mục trên elasticsearch

2.4.2. Các bảng cơ sở dữ liệu

2.5. Kiến trúc hệ thống

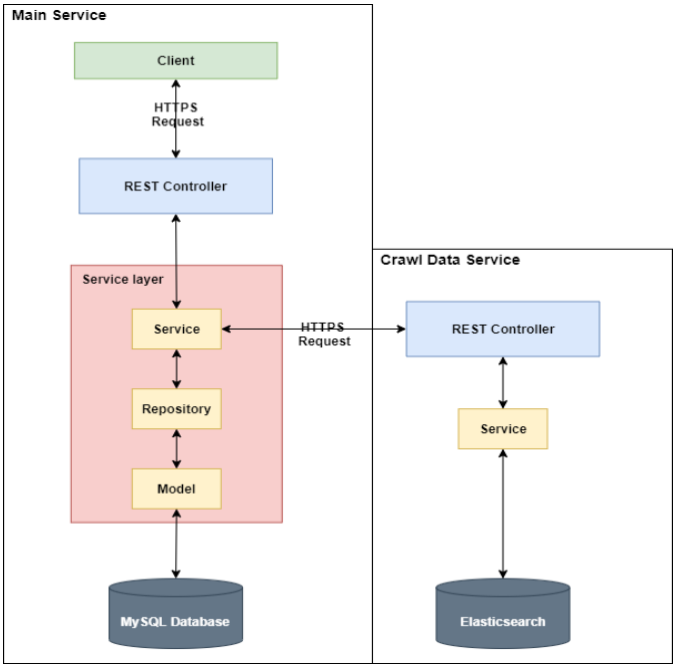


Hình 2.16: Kiến trúc tổng quan của hệ thống

Hệ thống tự động tổng hợp thông tin về dịch bệnh được xây dựng dựa trên kiến trúc tiểu dịch vụ gồm 2 tầng kiến trúc: tầng giao diện (Front-end), tầng ứng dụng (Back-end) như trong hình 2.16. Trong đó, tầng front-end chứa thực thể dịch vụ UI Compoment là hệ thống giao diện cho ứng dụng tự động tổng hợp thông tin về dịch bệnh. Tầng back-end gồm 4 thực thể dịch vụ: Gateway, Discovery server, Main Service, Crawl Data Service.

Discovery server sinh ra để lưu trữ thông tin và vị trí mạng của các tiểu dịch vụ này. Gateway sẽ xác thực, phân quyền, định tuyến và cân bằng tải các yêu cầu. Main Service là thực thể dịch vụ chứa các giao diện lập trình ứng dụng (Application Programming Interface - API) để xử lý logic cho hệ thống tự động tổng hợp thông tin về dịch bệnh, có thể có nhiều thể hiện (instances) để giúp hệ thống xử lý được nhiều yêu cầu một lúc. Crawl Data Service là thực thể dịch vụ chứa hệ thống thu thập dữ liệu và quản lý dữ liệu

trên Elasticsearch, giao tiếp với Main Service qua API để thực hiện chức năng quản lý bệnh viện, quản lý bài báo.



Hình 2.17: Luồng xử lý yêu cầu tại Main Service

Khi Gateway gửi yêu cầu đến thực thể dịch vụ Main Service. Luồng xử lý yêu cầu được mô tả như hình 2.17. Đầu tiên, yêu cầu sẽ được ánh xạ, phân tích và xử lý tại REST Controller bằng cách lấy dữ liệu từ tầng dịch vụ (Service). Tiếp theo, tầng dịch vụ chịu trách nhiệm kiểm tra dữ liệu đầu vào, xử lý logic và gọi xuống tầng quản lý dữ liệu (Repository) để lấy dữ liệu. Cuối cùng, tầng quản lý dữ liệu sẽ thực hiện các câu truy vấn dựa trên các thực thể (Model) và tương tác với cơ sở dữ liệu (Database) để lấy dữ liệu. Với các yêu cầu tìm kiếm, quản lý bệnh viện/bài báo, tầng dịch vụ sẽ gọi sang API của Crawl Data Service để lấy dữ liệu, Crawl Data Service sẽ giao tiếp với API của Elasticsearch để thực hiện yêu cầu.

2.6. Trích xuất thông tin dịch sốt xuất huyết

2.6.1. Luồng trích xuất thông tin

Khi trích xuất thông tin về dịch sốt xuất huyết tại một website bất kỳ, do cấu trúc DOM của mỗi website là khác nhau, quản trị viên cần phải tìm ra những giá trị của Xpath chứa thông tin cần thu thập và gửi lên hệ thống.

Sau khi dữ liệu được thu thập, nếu có dữ liệu đã tồn tại (bài viết có cùng đường dẫn đã tồn tại) thì hệ thống sẽ ghi đè dữ liệu mới lên dữ liệu cũ.

2.6.2. Mẫu đặc tả trích xuất thông tin về dịch sốt xuất huyết

prefixUrl: là tiền tố của đường dẫn, tránh truy cập sang trang web khác.

startUrl: là đường dẫn của trang đầu tiên khi thu thập

patternUrl: là mẫu của đường dẫn tại trang cần thu thập cuối cùng, định dạng theo chuẩn Python Regular Expression.

elasticsearchIndex: là tên của index muốn lưu trên elasticsearch, nếu muốn thu thập dữ liệu bệnh viện thì elasticsearchIndex là “hospital”, nếu muốn thu thập dữ liệu bài báo thì elasticsearchIndex là “news”. Khi đó, nên thu thập đầy đủ các trường theo định dạng mà hệ thống có sẵn. Ví dụ: dữ liệu bệnh viện sẽ có các trường: name, number, ward, district, city, workingTime, introduction, services, department, website, link, phone. Dữ liệu bài báo sẽ có các trường: title, body, tag, link.

data: gồm các trường muốn thu thập, định dạng là [tên trường] : [xpath của trường đó]. Cần quy định keyword có trong **body**, cụ thể trong đề án này **keyword** sẽ là “sốt xuất huyết” để tổng hợp thông tin về dịch sốt xuất huyết.

Mẫu đặc tả dữ liệu để trích xuất thông tin về dịch sốt xuất huyết cho trang web https://www.vinmec.com/vi/tin-tuc/ viết dưới dạng json như sau:

```
{
  "prefixUrl": "https://www.vinmec.com/vi/tin-tuc/",
  "startUrl":"https://www.vinmec.com/vi/tin-tuc/",
  "patternUrl":"^https://www.vinmec.com/vi/tin-tuc/(([a-z]|[-])+)/(([a-z]|[-])+)/((([a-z]|[-])+)/)*$",
```

```
    "elasticsearchIndex":"news",
    "data": {
      "title":"normalize-space(//*[@id=\"vue-
bootstrap\"]/div[2]/div[2]/h1/text())",
      "tag":"//*[@id=\"vue-bootstrap\"]/div[2]/div[3]/div[3]/a/text()",
      "body":"//*[@id=\"vue-bootstrap\"]/div[2]/div[3]/div[1]/div[1]",
      keyword:"Sốt xuất huyết"
    }
  }
```


2.7. Kết luận chương 2

Trong chương 2, đề án đã phân tích thiết kế hệ thống hệ thống tổng hợp thông tin về dịch bệnh và xây dựng được mẫu đặc tả dữ liệu để trích xuất thông tin về dịch sốt xuất huyết trên Scrapy framework.

Dựa trên nội dung đã được phân tích và thiết kế trong chương 2, đề án sẽ triển khai xây dựng hệ thống tổng hợp thông tin về dịch bệnh cụ thể là dịch sốt xuất huyết.

CHƯƠNG 3: TRIỂN KHAI VÀ XÂY DỰNG HỆ THỐNG

Trong chương 3 đề án đi sâu và chi tiết về việc xây dựng hệ thống tự động tổng hợp thông tin về dịch bệnh theo kiến trúc tiểu dịch vụ sử dụng Scrapy framework và các công cụ lập trình web đã giới thiệu ở chương 1. Dựa trên mẫu đặc tả dữ liệu để trích xuất thông tin ở chương 2, đề án sẽ triển khai thu thập dữ liệu mẫu từ hai trang web https://timbenhvien.vn và https://www.vinmec.com.

3.1 Kịch bản triển khai

3.1.1 Các nội dung triển khai

Hệ thống tự động tổng hợp thông tin về dịch bệnh sẽ triển khai với các nội dung và tính năng chính sau:

- Thu thập dữ liệu tự động: Thực hiện triển khai trích xuất các thông tin liên quan về bệnh sốt xuất huyết trên hai trang web: https://timbenhvien.vn/ và https://www.vinmec.com/vi/tin-tuc/.

- Cập nhật thông tin liên tục

- Tính năng tìm kiếm và lọc thông tin

- Tính năng quản lý người dùng, dữ liệu thu thập về bệnh viện và tin tức cho quản trị viên.

- Triển khai hệ thống theo mô hình kiến trúc tiểu dịch vụ để dễ dàng sửa, mở rộng, tích hợp với hệ thống khác sẵn có.

3.1.2. Thu thập dữ liệu tự động

3.1.2.1. Thu thập dữ liệu bệnh viện

Dữ liệu bệnh viện được thu thập từ trang web: https://timbenhvien.vn/ Dữ liệu được thu thập thông qua Scrapy framework, tuy nhiên phần lấy số điện thoại đã bị trang web che đi, cần phải thực hiện click vào nút “Hiển thị số điện thoại”, nên phần thu thập số điện thoại được thực hiện bằng Selenium framework [14], mô phỏng lại một chorme driver cho các thao tác truy cập đường dẫn, bấm vào nút “Hiển thị số điện thoại” để lấy dữ liệu.

Kết quả thu thập được: Tổng có 8157 bản ghi, được lưu trên elasticseach với các thông tin được mô tả dưới bảng 3.1.

Bảng 3.1: Bảng mô tả thu thập dữ liệu bệnh viện

STT	Tên cột	Giải thích	Ví dụ
1	Id	Mã định danh	42ATD3kBZLAYSuhRazKy
2	Name	Tên bệnh viện	Bảo Hà Spa - CS3
3	Number	Tên đường/phố	109/6 Nguyễn Bỉnh Khiêm
4	Ward	Tên phường/xã	Phường Đa Kao
5	District	Tên quận/huyện	Quận 1
6	City	Tên Tỉnh/Thành Phố	Hồ Chí Minh
7	workingTime	Thời gian làm việc	T2,T3,T4,T5,T6: 09:00 - 18:00
8	Introduction	Giới thiệu về bệnh viện	Bảo Hà Spa luôn làm việc bằng cả tấm lòng…
9	Services	Dịch vụ của bệnh viện	Dịch vụ chăm sóc mẹ bầu, chăm sóc mẹ sau sinh,
10	Department	Các khoa của bệnh viện	Khoa Da Liễu, Nhi Khoa, Sản Khoa
11	Website	Trang web của bệnh viện	http://baohaspa.vn
12	Link	Trang web thu thập dữ liệu	https://timbenhvien.vn/chi-tiet/bao-ha-spa---cs3/7364
13	Phone	Số điện thoại	0941.958.186

3.1.2.2. Thu thập dữ liệu bài báo

Dữ liệu bài báo được thu thập tại: https://www.vinmec.com/vi/tin-tuc/

Kết quả thu thập được: Tổng có 14138 bản ghi, trong đó có 626 bản ghi có nội dung liên quan đến dịch sốt xuất huyết được lưu trên elasticseach với các thông tin được mô tả dưới bảng 3.2.

Bảng 3.2: Bảng mô tả thu thập dữ liệu bài báo

STT	Tên cột	Giải thích	Ví dụ
1	Id	Mã định danh	5k3OvY0BiLF8oB1QQbTR
2	Title	Tiêu đề bài báo	Sốt xuất huyết không được uống thuốc gì?
3	Body	Nội dung bài báo	Đặc điểm bệnh sốt xuất huyết là không có thuốc đặc trị....
4	Link	Địa chỉ bài báo	https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/suc-khoe-tong-quat/sot-xuat-huyet-khong-duoc-uong-thuoc-gi/
5	Tag	Chủ đề bài báo	["Paracetamol","Thuốc hạ sốt", "Oresol","Điều trị sốt xuất huyết","Truyền nhiễm","Sốt xuất huyết"]

3.1.3 Các yêu cầu cần đạt của hệ thống

Hệ thống cần đạt các yêu cầu sau: Độ chính xác và đáng tin cậy, tính cập nhật, dễ sử dụng, tính linh hoạt, bảo mật thông tin

3.2. Triển khai xây dựng tầng ứng dụng hệ thống

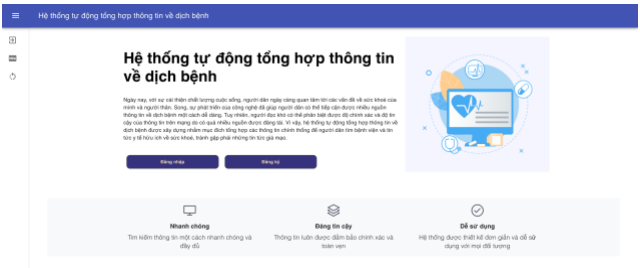
3.2.1. Môi trường

3.2.2. Ứng dụng hệ thống

3.3. Triển khai xây dựng tầng giao diện hệ thống

3.3.1. Giao diện trang chủ hệ thống

Hình 3.1 mô tả giao diện trang chủ của hệ thống



Hình 3.1: Giao diện trang chủ

3.3.2. Giao diện đăng nhập

3.3.3. Giao diện đăng ký tài khoản, kích hoạt tài khoản

3.3.4. Giao diện quên mật khẩu

3.3.5. Giao diện trang chủ sau khi đăng nhập

3.3.6. Giao diện đổi mật khẩu, thay đổi thông tin tài khoản

3.3.7. Giao diện tìm kiếm bệnh viện/bài báo của người dùng

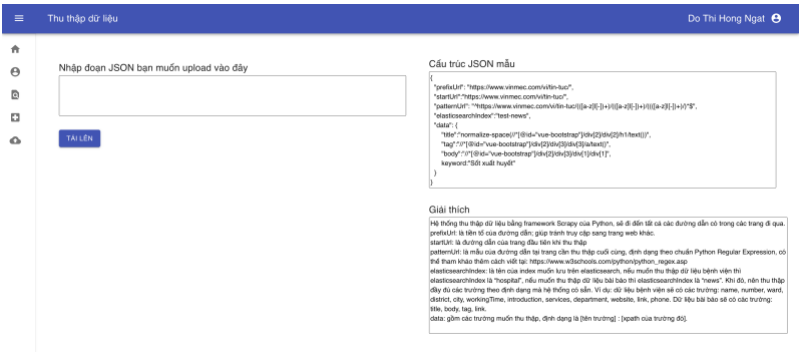
3.3.8. Giao diện quản lý bệnh viện

3.2.9. Giao diện quản lý bài báo

3.3.10. Giao diện quản lý người dùng

3.3.11. Giao diện quản lý dữ liệu thu thập

Giao diện tạo yêu cầu thu thập dữ liệu được mô tả như hình 3.32.



Hình 3.32: Giao diện tạo yêu cầu thu thập dữ liệu

3.4. Đánh giá hệ thống

Hệ thống tổng hợp thông tin về ý tế có giao diện khá trực quan và dễ sử dụng, các chức năng hoạt động chính xác với độ trễ đúng như thiết kế đặt ra. Tuy nhiên, hệ thống còn một số hạn chế về mặt chức năng như: chức năng đăng tải thông tin bài báo, bệnh viện chỉ nhận định dạng CSV, chưa hỗ trợ các loại định dạng khác, ngoài ra độ lớn của tệp tin cũng không quá 500MB. Email gửi thông tin các loại mã xác thực chưa được thiết kế đẹp, chuyên nghiệp. Hiện tại hệ thống chỉ thu thập dữ liệu tại 2 trang web https://timbenhvien.vn và https://www.vinmec.com. như mô tả bên trên nên chưa có sự đa dạng dữ liệu.

3.5. Kết luận chương 3

Trong chương 3 đề án đã xây dựng hệ thống tự động tổng hợp thông tin về dịch bệnh theo kiến trúc tiểu dịch vụ sử dụng Scrapy framework và các công cụ lập trình web với dữ liệu dựa trên hai trang web https://timbenhvien.vn và https://www.vinmec.com. Trên cơ sở đó, đề án đề xuất giải pháp triển khai hệ thống phù hợp trong hoạt động quản lý thông tin, cung cấp công cụ tìm kiếm cho người dân trên toàn quốc.

KẾT LUẬN

Các kết quả đạt được của đề án

Với mục tiêu nghiên cứu Scrapy framework và xây dựng hệ thống tự động tổng hợp thông tin về dịch bệnh, đề án đã đạt được một số kết quả sau đây:

- Khảo sát tổng quan về hệ thống tổng hợp thông tin, thực tế triển khai hệ thống tổng hợp thông tin trên thế giới và tại Việt Nam cùng các các vấn đề liên quan.

- Khảo sát tổng quan về xu hướng phát triển Hệ thống tự động tổng hợp thông tin về dịch bệnh trên thế giới và tại Việt Nam.

- Nghiên cứu về Scrapy framework.

- Nghiên cứu một số công nghệ phát triển web: Kiến trúc tiểu dịch vụ, thư viện Redux, Spring framework, Flask framework, MySQL, Elasticsearch

- Tiến hành phân tích và thiết kế hệ thống hệ thống tự động tổng hợp thông tin về dịch bệnh. Xây dựng được mẫu đặc tả để trích xuất thông tin về dịch bệnh trên trang web https://www.vinmec.com/

- Tiến hành triển khai và xây dựng hệ thống tự động tổng hợp thông tin về dịch bệnh và trích xuất dữ liệu thành công trên hai trang web https://timbenhvien.vn và https://www.vinmec.com.

Các kết quả nghiên cứu của đề án có thể sử dụng như một tài liệu tham khảo trong quá trình nâng cấp và mở rộng của hệ thống tự động tổng hợp thông tin.

Hướng phát triển tiếp theo

Tiếp tục nghiên cứu, đề xuất và triển khai thêm chức năng như tích hợp với các hệ thống có sẵn ở bệnh viện triển khai để giảm thiểu việc quản lý, kết nối trực tuyến giữa bác sỹ, nhân viên y tế và người bệnh, giải đáp trực tuyến 24/7.