The Saigon International University

Khóa luận tốt nghiệp

Website: www.siu.edu.vn

Thành phố Hồ Chí Minh - 2024

KHÓA LUẬN TỐT NGHIỆP

Ngành

KHOA HỌC MÁY TÍNH

Đề tài

XÂY DỰNG BIỂU ĐỒ PHÂN TÍCH HÀNH VI NGƯỜI DÙNG TRÊN FANPAGE TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN

Giảng viên hướng dẫn

Th.S Trần Hàm Dương

Sinh viên

Vũ Trung Nam

Mã sinh viên: **81012002090**

Lewis Campus

Email: admission@siu.edu.vn

Website: www.siu.edu.vn

LỜI CAM ĐOAN

Tôi xin cam đoan rằng đề tài "Xây Dựng Biểu Đồ Phân Tích Hành Vi Người Dùng Trên Trang Facebook SIU" là kết quả của quá trình tự mình nghiên cứu, xây dựng và phát triển cùng với sự hướng dẫn của giảng viên hướng dẫn. Toàn bộ nội dung và kết quả trong báo cáo này đều dựa trên kiến thức hiện có của bản thân và quá trình học tập từ các nguồn tài liệu trên mạng.

Tôi khẳng định rằng không có bất kỳ phần nào của đề tài này được sao chép từ các công trình nghiên cứu hay bài viết của người khác mà không được trích dẫn hoặc xin phép hợp pháp. Những thông tin, dữ liệu và kết quả phân tích đều được thực hiện một cách trung thực, cẩn thận và khoa học nhằm đảm bảo tính chính xác và khách quan.

Tôi hoàn toàn chịu trách nhiệm về lời cam đoan này và sẵn sàng chấp nhận mọi hình thức xử lý theo quy định của nhà trường nếu phát hiện có sự sao chép hoặc gian lận trong quá trình thực hiện đề tài.

Tp. Hồ Chí Minh, ngày 11 tháng 07 năm 2024

Sinh viên

Vũ Trung Nam

LỜI CẢM ƠN

Tôi xin được gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại học Quốc tế Sài Gòn nói chung, các thầy cô trong khoa Kỹ thuật & Khoa học máy tính nói riêng. Đặc biệt, để hoàn thành khóa luận tốt nghiệp này, tôi xin tỏ lòng biết ơn đến thầy Trần Hàm Dương, người đã tận tình hướng dẫn tôi trong suốt quá trình hoàn thành khóa luận tốt nghiệp để tôi có thể đạt kết quả tốt nhất.

Tôi cũng xin gửi lời cảm ơn đến các anh chị, thầy cô trong đơn vị thực tập đã tạo điều kiện cho tôi trang bị những kiến thức chuyên môn, có cơ hội được học hỏi, hiểu rõ hơn về cách tạo ra một sản phẩm hoàn chỉnh trong suốt 4 năm học tập và nghiên cứu. Từ đó, tạo cho tôi điều kiện thuận lợi và làm bước phát triển cho khóa luận tốt nghiệp trở nên chỉnh chu hơn.

Dù đã có cố gắng và nỗ lực trong quá trình thực hiện, song một số kiến thức vẫn còn thiếu và trình độ chuyên môn cũng như kinh nghiệm thực tế của tôi vẫn chưa có nhiều nên sẽ không tránh khỏi những hạn chế và thiếu sót. Vì vậy, tôi kính mong nhận được sự góp ý và nhận xét của quý thầy, cô để từ đó tôi có thể hoàn thiện khóa luận tốt nghiệp của mình.

Tôi xin chân thành cảm ơn.

Tp. Hồ Chí Minh, ngày 11 tháng 07 năm 2024

Sinh viên

Vũ Trung Nam

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày 11 tháng 07 năm 2024

GIẢNG VIÊN HƯỚNG DẪN

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Từ viết tắt	Tên đầy đủ	Ý nghĩa
SIU	The SaiGon International University	Trường Đại học Quốc tế Sài gòn
SSR	Server-side rendering	Kết xuất giao diện phía máy chủ
CSR	Client-side rendering	Kết xuất giao diện phía máy khách
VS Code	Visual Studio Code	Trình soạn thảo mã
Npm	Node package module	Quản lý gói thư viện

DANH MỤC BẢNG

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH

CHƯƠNG 1. TỔNG QUAN

1.1 Thực trạng hiện nay

Hiện nay, với sự phát triển vượt bậc của công nghệ, các kênh truyền thông đang trở thành một phần không thể thiếu trong cuộc sống của chúng ta. Mạng xã hội đã thu hút một lượng người dùng lớn, tạo nên một không gian kết nối và chia sẻ thông tin một cách chưa từng có. Thông qua mạng xã hội, các công ty và tổ chức có thể dễ dàng tiếp cận thông tin, tin tức, nắm bắt tâm lý và xu hướng của người sử dụng. Điều này giúp họ hiểu rõ hơn về nhu cầu và mong muốn của khách hàng, từ đó cải thiện sản phẩm và dịch vụ của mình. Ngoài ra, thông tin từ mạng xã hội cũng có thể được sử dụng để phân tích xu hướng xã hội, đánh giá hiệu quả của chiến dịch tiếp thị và đưa ra các quyết định chiến lược. Việc thu thập dữ liệu từ mạng xã hội và các nguồn thông tin trực tuyến đã trở thành một yếu tố quan trọng để hiểu và phân tích hành vi người dùng, xu hướng thị trường và ý kiến công chúng. Từ những bình luận trên mạng xã hội, đánh giá sản phẩm, hay thậm chí các bài viết trên diễn đàn, thông tin này cung cấp một nguồn tài nguyên phong phú cho các công ty và tổ chức.

Để khai thác được tài nguyên thông tin này, việc tạo ra các công cụ thu thập dữ liệu là cần thiết. Công cụ này giúp tự động hóa quá trình thu thập, lưu trữ và xử lý dữ liệu từ mạng xã hội. Đây là một công cụ vô cùng cần thiết để hiểu rõ hơn về tâm lý và hành vi của người sử dụng. Các doanh nghiệp, tổ chức, và cả chính phủ đều có lợi từ việc thu thập thông tin này để nắm bắt xu hướng và nhu cầu của khách hàng, từ đó đưa ra các chiến lược và quyết định phù hợp.

1.2 Nhiệm vụ khóa luận

Khóa luận thực hiện các nhiệm vụ sau để giải quyết những vấn đề trên:

- Xây dựng công cụ thu thập dữ liệu trên trang facebook của trường Đại học Quốc tế Sài Gòn (SIU).
- Xây dựng giao diện trực quan hóa dữ liệu đã thu thập.

1.3 Phạm vi

Bao gồm các lượt thích, comment, bình luận, chia sẻ trên trang facebook của trường Đại học Quốc tế Sài Gòn từ 20/10/2023 đến 30/06/2024.

1.4 Đối tượng sử dụng

Phòng Tư vấn Tuyển sinh trường Đại học Quốc tế Sài Gòn.

1.5 Mục tiêu của ứng dụng

Ứng dụng mang mục tiêu giúp phòng tư vấn tuyển sinh SIU thu thập và trực quan hóa dữ liệu về thông tin của fanpage Trường Đại học Quốc tế Sài Gòn, bao gồm thông tin về các bài viết trên fanpage, lượt thích, cảm xúc, bình luận, chia sẻ của người dùng về các bài viết trên fanpage. Từ đó nắm bắt được xu hướng và tâm lý của người dùng để đưa ra các quyết định và chiến lược tuyển sinh phù hợp.

1.6 Các bước thực hiện khóa luận

Quá trình thực hiện khóa luận được chia thành các bước: đầu tiên là lập kế hoạch xây dựng hệ thống thu thập dữ liệu tự động. Tiếp theo, chúng tôi phân tích hệ thống thu thập dữ liệu tự động để hiểu rõ các yêu cầu và mục tiêu. Sau đó, chúng tôi thiết kế hệ thống một cách chi tiết và cụ thể. Cuối cùng, chúng tôi tiến hành cài đặt hệ

thống thu thập dữ liệu tự động, đảm bảo các thành phần hoạt động đúng chức năng và hiệu quả.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Công cụ thu thập dữ liệu thu thập dữ liệu

2.1.1 Công cụ thu thập dữ liệu là gì?

Công cụ thu thập dữ liệu, hay còn gọi là trình thu thập dữ liệu (crawler), là một chương trình máy tính tự động duyệt qua các trang web để thu thập và trích xuất dữ liệu một cách có hệ thống. Nó được sử dụng rộng rãi trong nhiều lĩnh vực, từ các công cụ tìm kiếm như Google đến các hệ thống giám sát và phân tích thông tin, công cụ thu thập dữ liệu đóng vai trò quan trọng trong việc tổng hợp và lập chỉ mục nội dung web. Với khả năng tải và phân tích hàng loạt trang web, công cụ này giúp thu thập thông tin về các sản phẩm, giá cả, tin tức, và nhiều loại dữ liệu khác, tạo nền tảng cho các hoạt động nghiên cứu, kinh doanh và phát triển công nghệ. Bằng cách tự động hóa quy trình thu thập thông tin, công cụ thu thập dữ liệu giúp tiết kiệm thời gian, công sức, và cung cấp dữ liệu chính xác, cập nhật, hỗ trợ cho việc ra quyết định hiệu quả.

2.1.2 Các công cụ thu thập dữ liệu phổ biến

Trên thế giới, có nhiều công cụ thu thập dữ liệu phổ biến được sử dụng trong các lĩnh vực khác nhau, từ nghiên cứu học thuật đến kinh doanh và tiếp thị. Dưới đây là một số công cụ thu thập dữ liệu nổi bật:

- **Googlebot**:
  - Mục đích: Thu thập dữ liệu của Google để phân tích.
  - Chức năng: Lập chỉ mục các trang web trên Internet để cung cấp kết quả tìm kiếm chính xác.

- Đặc điểm: Sử dụng thuật toán phức tạp để đánh giá và xếp hạng các trang web dựa trên nội dung, cấu trúc và các yếu tố khác.

- **Octoparse**:
  - Mục đích: Công cụ thu thập dữ liệu không cần mã hóa, phù hợp cho người dùng không chuyên về kỹ thuật.
  - Chức năng: Hỗ trợ kéo thả để xây dựng quy trình thu thập dữ liệu, tự động hóa và quản lý dữ liệu.
  - Đặc điểm: Giao diện trực quan, dễ sử dụng, và hỗ trợ nhiều tính năng nâng cao như xử lý Captcha, quản lý phiên đăng nhập.
- **WebHarvy**:
  - Mục đích: Công cụ thu thập dữ liệu tự động với giao diện thân thiện.
  - Chức năng: Hỗ trợ trích xuất dữ liệu bằng cách chỉ định các mẫu trực tiếp trên trang web mà không cần viết mã.
  - Đặc điểm: Dễ sử dụng, hỗ trợ nhiều định dạng xuất dữ liệu như CSV, XML, JSON và SQL.
- **ParseHub**:
  - Mục đích: Công cụ thu thập dữ liệu trực quan với khả năng xử lý các trang web phức tạp.
  - Chức năng: Hỗ trợ trích xuất dữ liệu từ các trang web động sử dụng AJAX hoặc JavaScript.
  - Đặc điểm: Giao diện đồ họa, hỗ trợ nhiều định dạng xuất và khả năng xử lý dữ liệu phức tạp.

Các công cụ thu thập dữ liệu này cung cấp nhiều tính năng và khả năng phù hợp với các nhu cầu khác nhau của người dùng, từ các nhà phát triển chuyên nghiệp đến người dùng không chuyên về kỹ thuật. Với sự hỗ trợ của các công cụ này, quá trình thu thập, xử lý và phân tích dữ liệu từ web trở nên dễ dàng và hiệu quả hơn.

2.1.3 Nguyên tắc hoạt động của công cụ thu thập dữ liệu

Công cụ thu thập dữ liệu được thiết kế dựa trên nguyên tắc tự động duyệt qua các trang web trên Internet và thu thập thông tin từ những trang đó. Công cụ hoạt động theo các bước sau:

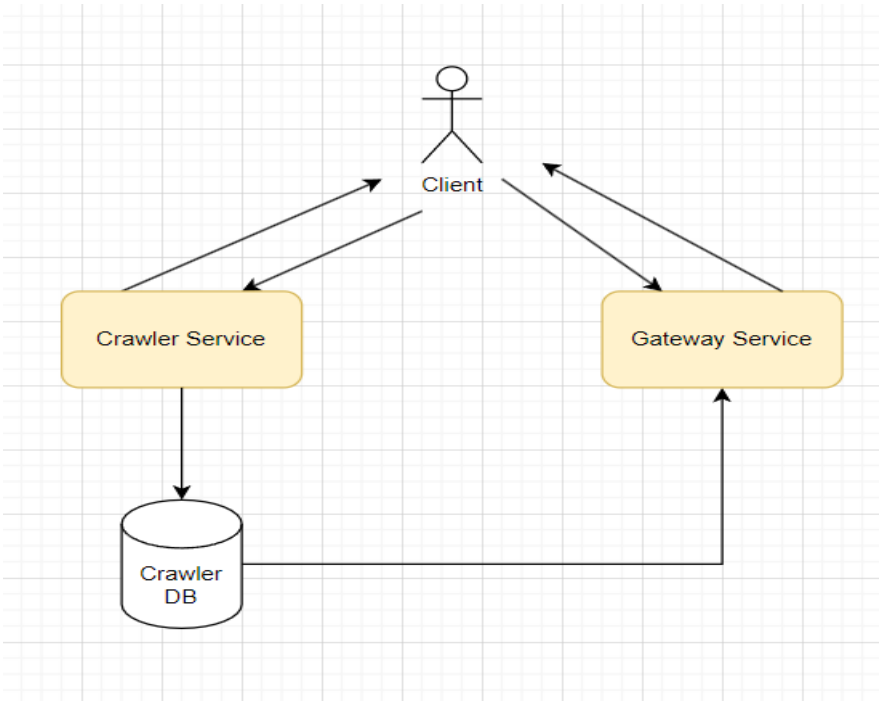
1. **Xác định các ĐƯỜNG LIÊN KẾT cần truy cập**: Công cụ bắt đầu bằng việc chọn một số ĐƯỜNG LIÊN KẾT ban đầu để bắt đầu quá trình duyệt web. Các ĐƯỜNG LIÊN KẾT này có thể được cung cấp sẵn hoặc được tạo ra tự động dựa trên tiêu chí nhất định.
2. **Tải các trang web**: Công cụ gửi yêu cầu HTTP đến các ĐƯỜNG LIÊN KẾT đã chọn và tải các trang web về để phân tích nội dung. Trang web có thể được tải bằng cách sử dụng HTTP GET request, các yêu cầu này đều được thực hiện trên trình duyệt.
3. **Phân tích nội dung của trang web**: Sau khi tải về, crawler phân tích nội dung HTML của từng trang để tìm các liên kết (ĐƯỜNG LIÊN KẾT) và thông tin cần thiết khác như văn bản, hình ảnh, video, …
4. **Lưu trữ thông tin**: Crawler lưu trữ thông tin thu thập được theo cấu trúc dữ liệu nhất định, ví dụ như cơ sở dữ liệu hoặc các tệp lưu trữ.

5. **Tiếp tục duyệt các liên kết**: Sau khi phân tích một trang, crawler sẽ tìm các liên kết khác trên trang này và lặp lại quá trình tải về và phân tích nội dung cho các liên kết mới này. Quá trình này có thể tiếp tục đến khi tất cả các liên kết thỏa mãn điều kiện dừng đã được đặt ra.

Các nguyên tắc này giúp định hướng hoạt động của crawler và đảm bảo tính hiệu quả và lâu dài của quá trình thu thập dữ liệu từ Internet.

2.2 Hệ thống thu thập dữ liệu

2.2.1 Tổng quan hệ thống



Hình 1. Tổng quan hệ thống thu thập dữ liệu.

Trong đề tài này, hệ thống được thiết kế với bốn phần chính bao gồm: client, dịch vụ thu thập dữ liệu (crawler service), cơ sở dữ liệu (crawler database) và dịch vụ cổng (gateway service). Client có chức năng tương tác với dịch vụ thu thập dữ liệu để khởi động hoặc dừng quá trình thu thập dữ liệu tự động. Khi nhận lệnh từ client, dịch vụ