NHẬP MÔN CÔNG NGHỆ PHẦN MỀM

Kết Quả Thực Hiện

Giảng viên: Trần Văn Quý

Sinh viên: Lê Ngọc Huân

Mã số sinh viên: 22880226

Bộ môn Công nghệ phần mềm

Khoa Công nghệ thông tin

Đại học Khoa học tự nhiên TP HCM

MỤC LỤC

Bảng đánh giá thành viên

MSSV	Họ Tên	% đóng góp (tối đa 100%)	Chữ ký
22880226	Lê Ngọc Huân	100%	Huân

Môi trường phát triển và Môi trường triển khai

Môi trường phát triển ứng dụng

1. Hệ điều hành: Windows 10.
2. Hệ quản trị cơ sở dữ liệu: Microsoft SQL Server.
3. Công cụ dùng để phân tích, thiết kế: Không sử dụng công cụ phân tích thiết kế cụ thể nào do chỉ có 1 table, có thể tham khảo một vài như Microsoft Visio, Lucidchart.
4. Công cụ đã dùng để xây dựng ứng dụng: Python là ngôn ngữ lập trình chính, sử dụng một môi trường phát triển tích hợp (IDE) là PyCharm,
5. Các thư viện đã sử dụng: Flask cho web server, SQLAlchemy cho ORM, BeautifulSoup cho crawling, Pandas, numpy cho xử lý dữ liệu, pyodbc và SQLAlchemy để kết nối cơ sở dữ liệu.

Môi trường triển khai ứng dụng

1. Hệ điều hành: Giống như môi trường phát triển hoặc một hệ điều hành được hỗ trợ bởi Python và Flask.
2. Yêu cầu phần mềm:

   Python 3.12 hoặc cao hơn.

   Flask 3.0 hoặc cao hơn.

   SQLAlchemy cho ORM.

   Các thư viện phụ thuộc khác như BeautifulSoup, Pandas, pyodbc.

   Microsoft SQL Server 12
3. Yêu cầu cài đặt:

   Cài đặt Python và pip (quản lý gói Python).

   Sử dụng pip để cài đặt Flask, SQLAlchemy, BeautifulSoup, Pandas, và pyodbc qua lệnh pip install Flask SQLAlchemy beautifulsoup4 pandas pyodbc.

   Cài đặt một SQL Server và tạo cơ sở dữ liệu theo schema được định nghĩa trước
4. Cấu hình:

Đảm bảo rằng cổng mà Flask sử dụng (mặc định là 5000) không bị chặn bởi tường lửa.

Cấu hình chuỗi kết nối tới cơ sở dữ liệu trong mã nguồn.

Kết quả đạt được

1. Các Chức Năng Đã Phân Tích và Thiết Kế

   Crawling dữ liệu: Phân tích và thiết kế thuật toán để crawl dữ liệu từ hai trang web đặc biệt, đảm bảo thu thập dữ liệu một cách hiệu quả và chính xác, bao gồm tên hội nghị, nơi diễn ra và thời gian diễn ra.

   Xử lý dữ liệu: Thiết kế các cơ chế xử lý dữ liệu để định dạng lại cho phù hợp với cơ sở dữ liệu, bao gồm việc làm sạch, và chuẩn hóa dữ liệu.

   Lưu trữ dữ liệu: Thiết kế cơ sở dữ liệu để lưu trữ dữ liệu được thu thập, với cấu trúc phù cho việc truy xuất.

   Thiết kế giao diện để tương tác người dùng.

2. Các Chức Năng Đã Cài Đặt Hoàn Chỉnh

   Crawling dữ liệu: Hoàn thành chức năng crawl dữ liệu từ hai nguồn đã chọn, và hiện đang hoạt động mượt mà.

   Giao diện người dùng: Cài đặt giao diện web đơn giản cho phép hiển thị dữ liệu từ cơ sở dữ liệu và tương tác thông qua chức năng tìm kiếm.

   Lọc, xử lý và lưu trữ dữ liệu: Dữ liệu được craw về không phù hợp với định dạng nên đã thực hiện chỉnh sửa để lưu trữ dữ liệu vào cơ sở dữ liệu

3. Các Chức Năng Đã Cài Đặt Nhưng Chưa Hoàn Chỉnh

   Tối ưu hóa xử lý dữ liệu: Dù đã cài đặt chức năng xử lý dữ liệu, nhưng vẫn cần tối ưu thêm để cải thiện hiệu suất và giảm thời gian xử lý.

Dữ liệu được craw được nằm trong các tag khác nhau nên việc viết hàm chưa tối ưu sử dụng cho nhiều trang web khác, đồng thời dữ liệu khi craw về cần chỉnh sửa định dạng để phù hợp csdl

Cơ sở dữ liệu chưa nhiều, cần được mở rộng thêm như tên chủ trì hội nghị, số lượng người tham gia, …

Các dữ liệu được craw phù hợp với yêu cầu nhưng cần phong phú thêm

4. Điểm Đặc Sắc Của Đề Tài

Chính xác và hiệu quả: Các thuật toán crawl dữ liệu đã được tối ưu hóa để xử lý cụ thể cho từng trang web đã chỉ định, đảm bảo hiệu quả và độ chính xác cao trong việc thu thập dữ liệu dù nhiều dữ liệu không đúng định dạng hoặc lưu nhiều dạng khác nhau

Cơ sở dữ liệu tối ưu: Hoàn thành thiết lập cơ sở dữ liệu đơn giản để hỗ trợ hiệu quả cao trong việc lưu trữ và truy xuất dữ liệu, cho phép xử lý dữ liệu và nâng cấp thêm

Kiến trúc 3 lớp.

Giao diện người dùng thân thiện: Giao diện web được thiết kế để dễ dàng sử dụng, hỗ trợ hiển thị trực quan các kết quả dữ liệu, cần nâng cấp thêm

Hướng phát triển

1. Mở rộng nguồn dữ liệu

   Crawler linh hoạt hơn: Phát triển hàm crawler để nó có thể dễ dàng thích ứng với các cấu trúc HTML khác nhau, giúp mở rộng nguồn dữ liệu mà không cần sửa đổi nhiều trong code.

Thêm nguồn dữ liệu: Thêm các trang web mới vào danh sách nguồn dữ liệu để phong phú hóa dữ liệu thu thập, từ đó tăng tính ứng dụng của hệ thống.

2. Cải tiến xử lý và lưu trữ dữ liệu

   Tối ưu hóa cơ sở dữ liệu: Thêm các dữ liệu khác nhau như số lượng người tham gia, số ghế ngồi, giá vé, .. và tối ưu hóa cơ sở dữ liệu để cải thiện hiệu suất truy vấn và khả năng mở rộng.

   Tham khảo sử dụng Big Data Technologies: Xem xét áp dụng các công nghệ Big Data như Hadoop hoặc Spark để xử lý và phân tích dữ liệu lớn một cách hiệu quả hơn.

3. Nâng cao chức năng tìm kiếm

   Cải tiến giao diện tìm kiếm: Phát triển các tính năng tìm kiếm nâng cao như, tìm kiếm dựa trên hình ảnh tên hội nghị, hoặc tìm kiếm giọng nói.

4. Cải tiến giao diện người dùng

   Trải nghiệm người dùng: Cải tiến giao diện dễ nhìn để người dùng có trải nghiệm tốt hơn khi sử dụng hệ thống, có thể tham khảo thêm giao diện di động.