TRƯỜNG ĐẠI HỌC PHENIKAA

KHOA CÔNG NGHỆ THÔNG TIN

A logo for a university

Description automatically generated

**BÁO CÁO ĐỒ ÁN LIÊN NGÀNH**

|  |  |  |
| --- | --- | --- |
| Họ tên sinh viên | Mã sinh viên | Email |
| Nguyễn Quốc Trường | 21012325 | [21012325@phenikaa-uni.edu.vn](mailto:21012325@phenikaa-uni.edu.vn) |
| Người hướng dẫn: T.S Lương Văn Thiện | | |
| Người phản biện: T.S Lương Văn Thiện | | |

Đề tài: Thu thập, xử lý và phân tích dữ liệu

**Hà Nội, 5 tháng 3 năm 2024**

Mục lục

[I. Giới thiệu 2](#_Toc160590334)

[1. Dữ liệu là gì? 2](#_Toc160590335)

[2. Đặt vấn đề “Tại sao phải thu thập, xử lý và phân tích dữ liệu” 2](#_Toc160590336)

[II. Thu thập, xử lý và phân tích dữ liệu 3](#_Toc160590337)

[1. Mạng xã hội và trang web: 3](#_Toc160590338)

[1.1. Phương pháp thu thập dữ liệu từ mạng xã hội và trang web 3](#_Toc160590339)

[1.2. Thách thức và yếu tố cần cân nhắc 4](#_Toc160590340)

[2. Xử lý dữ liệu 5](#_Toc160590341)

[2.1. Tiền xử lý: 5](#_Toc160590342)

[2.2. Phân tích và biến đổi: 5](#_Toc160590343)

[2.3. Ghép nối và kết hợp: 6](#_Toc160590344)

[2.4. Bảo mật và quản lý dữ liệu: 6](#_Toc160590345)

[2.5. Tối ưu hóa hiệu suất: 7](#_Toc160590346)

[3. Phân tích dữ liệu 8](#_Toc160590347)

[3.1. Phân tích số liệu thống kê: 8](#_Toc160590348)

[3.2. Máy học và học máy: 8](#_Toc160590349)

[3.3. Khám phá dữ liệu: 8](#_Toc160590350)

[3.6. Ứng dụng của phân tích dữ liệu: 10](#_Toc160590351)

[4. Đánh giá, kết luận của đề tài. 10](#_Toc160590352)

[4.1. Đánh giá 10](#_Toc160590353)

[4.2. Kết luận 11](#_Toc160590354)

[III. Một số thành phần khác của đồ án 12](#_Toc160590355)

[1. Sơ đồ luồng đi của dữ liệu 12](#_Toc160590356)

# I. Giới thiệu

## 1. Dữ liệu là gì?

Dữ liệu là một yếu tố vô cùng quan trọng trong mọi lĩnh vực của cuộc sống hiện đại. Từ kinh doanh đến khoa học, từ chính trị đến y tế, dữ liệu đóng vai trò quyết định trong quá trình ra quyết định và phát triển các chiến lược. Tuy nhiên, dữ liệu không có giá trị nếu không được thu thập, xử lý và phân tích một cách hiệu quả. Trong bài luận này, chúng ta sẽ tìm hiểu về quá trình thu thập, xử lý và phân tích dữ liệu, cùng những thách thức và cơ hội mà chúng ta phải đối mặt trong quá trình này.

## 2. Đặt vấn đề “Tại sao phải thu thập, xử lý và phân tích dữ liệu”

Trong thời đại số hóa ngày nay, dữ liệu đã trở thành một tài nguyên quý báu có thể cung cấp thông tin và hiểu biết đáng giá để hỗ trợ quyết định và phát triển. Tuy nhiên, việc thu thập, xử lý và phân tích dữ liệu không chỉ đơn giản là một quá trình, mà còn là một thách thức đầy phức tạp và đa chiều.

Một trong những vấn đề chính là làm thế nào để hiệu quả thu thập dữ liệu từ nhiều nguồn khác nhau một cách đáng tin cậy và an toàn. Đây là một thách thức lớn đối với các tổ chức và cá nhân, đặc biệt là khi dữ liệu có tính nhạy cảm về quyền riêng tư hoặc phải được xử lý theo các quy định pháp luật nghiêm ngặt.

Một vấn đề khác là làm thế nào để xử lý và phân tích các tập dữ liệu lớn, phức tạp một cách hiệu quả và có ý nghĩa. Với sự gia tăng đáng kể về khối lượng dữ liệu được tạo ra hàng ngày từ nhiều nguồn khác nhau, việc áp dụng các công nghệ và phương pháp mới để xử lý dữ liệu trở nên cần thiết để đảm bảo rằng chúng ta có thể tận dụng hết tiềm năng của dữ liệu này.

Đồng thời, việc đảm bảo tính chính xác và đáng tin cậy của dữ liệu cũng là một thách thức quan trọng. Dữ liệu không chính xác hoặc không đáng tin cậy có thể dẫn đến việc ra quyết định sai lầm và thiếu chính xác, gây ra những hậu quả không mong muốn.

Trong bối cảnh này, việc nghiên cứu và phát triển các phương pháp, công cụ và kỹ thuật để thu thập, xử lý và phân tích dữ liệu trở nên cấp bách hơn bao giờ hết. Cần có sự hiểu biết sâu sắc về các thách thức và cơ hội trong quá trình làm việc với dữ liệu để đảm bảo rằng chúng ta có thể tận dụng hết tiềm năng của dữ liệu để đem lại giá trị và hiệu quả cho các tổ chức và cộng đồng hiệu quả?

# II. Thu thập, xử lý và phân tích dữ liệu

## 1. Mạng xã hội và trang web:

Mạng xã hội và các trang web đang ngày càng trở thành một nguồn thông tin phong phú và không thể thiếu trong cuộc sống hàng ngày của chúng ta. Từ việc chia sẻ suy nghĩ, ý kiến, đến việc mua sắm trực tuyến và tìm kiếm thông tin, mọi người dùng Internet thường có những hoạt động và dấu vết trên mạng. Quá trình thu thập dữ liệu từ mạng xã hội và các trang web có thể mang lại nhiều lợi ích, từ việc hiểu rõ hơn về người dùng đến việc phát hiện xu hướng thị trường và dự đoán hành vi tiêu dùng. Tuy nhiên, việc này không chỉ đặt ra những thách thức về quyền riêng tư và bảo mật mà còn đòi hỏi sự cẩn thận và phương pháp hợp lý.

### 1.1. Phương pháp thu thập dữ liệu từ mạng xã hội và trang web

Công cụ và API của mạng xã hội: Một số mạng xã hội như Facebook, Twitter, và Instagram cung cấp các công cụ và API cho phép các nhà nghiên cứu truy cập vào dữ liệu từ nền tảng của họ. Các công cụ này thường cho phép thu thập thông tin về hồ sơ người dùng, bài đăng, tương tác, và các yếu tố khác có liên quan.

A screen shot of a computer error

Description automatically generated

Hình 1: Một đoạn code cào dữ liệu qua API

Crawler và scraper web: Các công cụ tự động như crawler và scraper có thể được sử dụng để thu thập dữ liệu từ các trang web công cộng. Các công cụ này hoạt động bằng cách tự động duyệt qua các trang web và lấy dữ liệu từ các tài nguyên có sẵn. Tuy nhiên, việc sử dụng các công cụ này cần phải tuân thủ các quy định về bản quyền và chính sách quyền riêng tư của các trang web.

A screenshot of a computer program

Description automatically generatedA computer screen shot of text

Description automatically generated

Hình 2: Sử dụng selenium để cào dữ liệu

Phân tích dữ liệu trực tuyến: Sự phát triển của công nghệ và phương pháp phân tích dữ liệu trực tuyến cho phép nhà nghiên cứu theo dõi, thu thập và phân tích dữ liệu từ các hoạt động trực tuyến như diễn đàn, blog, và nhận xét sản phẩm. Các công cụ phân tích dữ liệu trực tuyến có thể tự động thu thập thông tin từ các nguồn này và tạo ra các báo cáo và đánh giá dựa trên dữ liệu thu thập được.

### 1.2. Thách thức và yếu tố cần cân nhắc

Quyền riêng tư và đạo đức: Việc thu thập dữ liệu từ mạng xã hội và các trang web đặt ra những vấn đề về quyền riêng tư của người dùng. Một số thông tin cá nhân như tài khoản ngân hàng, địa chỉ email, và thông tin địa chỉ có thể được người dùng chia sẻ trên mạng xã hội. Việc thu thập và sử dụng thông tin này đòi hỏi sự cẩn thận và tuân thủ các quy định về bảo vệ dữ liệu cá nhân.

Chất lượng dữ liệu: Dữ liệu thu thập từ mạng xã hội và các trang web có thể chứa nhiều yếu tố nhiễu và không chính xác. Việc xác định và loại bỏ thông tin không chính xác hoặc không cần thiết là một thách thức đối với nhà nghiên cứu. Sự chính xác của dữ liệu thu thập được ảnh hưởng trực tiếp đến tính chính xác của các phân tích và dự đoán sau này.

Bảo mật thông tin: Bảo mật thông tin cá nhân của người dùng là một yếu tố quan trọng cần được xem xét khi thu thập dữ liệu từ mạng xã hội và các trang web. Việc đảm bảo rằng dữ liệu được lưu trữ và xử lý một cách an toàn và bảo mật là một ưu tiên hàng đầu.

Pháp luật và chính sách: Việc thu thập dữ liệu từ mạng xã hội và các trang web cần phải tuân thủ các quy định pháp luật và chính sách về quyền riêng tư và bảo vệ dữ liệu cá nhân của từng quốc gia. Việc vi phạm các quy định này có thể gây ra hậu quả pháp lý nghiêm trọng cho tổ chức hoặc cá nhân thu thập dữ liệu.

## 2. Xử lý dữ liệu

Sau khi dữ liệu đã được thu thập, quá trình tiếp theo là xử lý dữ liệu để làm cho nó có ý nghĩa và hữu ích. Quá trình này bao gồm các bước sau đây:

### 2.1. Tiền xử lý:

Tiền xử lý dữ liệu là bước quan trọng nhằm loại bỏ dữ liệu không hợp lệ hoặc thiếu sót, kiểm tra tính nhất quán và chuẩn hóa dữ liệu. Các bước cụ thể có thể bao gồm:

Loại bỏ dữ liệu trùng lặp: Xác định và loại bỏ các bản ghi hoặc quan sát trùng lặp trong tập dữ liệu.

Xử lý dữ liệu thiếu: Điền giá trị thiếu hoặc loại bỏ các quan sát có dữ liệu thiếu, tùy thuộc vào mức độ ảnh hưởng của dữ liệu thiếu đối với phân tích.

Kiểm tra tính nhất quán: Kiểm tra tính nhất quán của dữ liệu bằng cách phát hiện và sửa chữa các lỗi trong dữ liệu như lỗi nhập sai hoặc không phù hợp.

A screenshot of a computer screen

Description automatically generated

Hình 3: Loại bỏ dữ liệu không nhất quán

### 2.2. Phân tích và biến đổi:

Sau khi tiền xử lý đã hoàn thành, dữ liệu được sẵn sàng cho quá trình phân tích và biến đổi. Các bước trong quá trình này bao gồm:

Tính tổng doanh thu từng mặt hàng:

A screen shot of a computer

Description automatically generated

### 2.3. Ghép nối và kết hợp:

Dữ liệu được thu thập từ nhiều nguồn khác nhau. Để tạo ra một bộ dữ liệu hoàn chỉnh và đa chiều, quá trình ghép nối và kết hợp được thực hiện. Các bước cụ thể bao gồm:

Ghép nối dữ liệu: Kết hợp các bảng dữ liệu với nhau

A screenshot of a computer

Description automatically generated

Hình 4: Ghép nối dữ liệu của các mặt hàng khác nhau

A screenshot of a computer

Description automatically generated

Hình 5: Ghép nối dữ liệu trong 1 ngày để phân tích

### 2.4. Bảo mật và quản lý dữ liệu:

Bảo vệ dữ liệu khỏi việc truy cập trái phép và quản lý dữ liệu là một phần quan trọng trong quá trình xử lý dữ liệu. Các biện pháp bảo mật và quản lý dữ liệu bao gồm:

Phân quyền truy cập: Xác định và quản lý quyền truy cập vào dữ liệu để đảm bảo chỉ những người được ủy quyền mới có thể truy cập và sử dụng dữ liệu.

Mã hóa dữ liệu: Sử dụng các thuật toán mã hóa để bảo vệ thông tin cá nhân và quan trọng trong dữ liệu.

Sao lưu và phục hồi: Thực hiện việc sao lưu dữ liệu định kỳ và thiết lập các kế hoạch phục hồi để đảm bảo rằng dữ liệu có sẵn khi cần thiết.

### 2.5. Tối ưu hóa hiệu suất:

Quá trình xử lý dữ liệu cũng đòi hỏi việc tối ưu hóa hiệu suất để đảm bảo rằng quá trình thu thập dữ liệu được thực hiện một cách hiệu quả và nhanh chóng

Tối ưu hóa cấu trúc dữ liệu: Tối ưu hóa cấu trúc dữ liệu để giảm thiểu thời gian truy cập và xử lý dữ liệu.

Sử dụng BeautifulSoup: Sử dụng BeautifulSoup để tăng cường hiệu suất xử lý dữ liệu.

A screenshot of a computer code

Description automatically generated

Hình 6: Một đoạn code sử dụng BeautifulSoup để thu thập dữ liệu nhanh chóng

Tối ưu hóa thuật toán: Tối ưu hóa thuật toán và quá trình xử lý để giảm thiểu thời gian và tài nguyên cần thiết.

A computer screen shot of a program code

Description automatically generated

Hình 7: Tách từng phần code để tối ưu

## 3. Phân tích dữ liệu

Phân tích dữ liệu là một phần quan trọng trong quá trình khai thác thông tin từ dữ liệu đã được thu thập và xử lý. Đây là quá trình tìm kiếm các mẫu, mối quan hệ và thông tin có ý nghĩa từ dữ liệu, giúp đưa ra các nhận định, dự đoán và quyết định chiến lược. Dưới đây là một số phương pháp phân tích dữ liệu phổ biến:

### 3.1. Phân tích số liệu thống kê:

Phân tích số liệu thống kê là quá trình sử dụng các kỹ thuật thống kê để mô tả và suy luận từ dữ liệu. Các phương pháp thống kê cơ bản như trung bình, phương sai, phân phối tần suất, và kiểm định giả thuyết thường được sử dụng để mô tả tính chất của dữ liệu và đưa ra các kết luận về các biến quan trọng. Ngoài ra, các phương pháp thống kê nâng cao như hồi quy, phân tích phương sai (ANOVA), và phân tích chuỗi thời gian cũng được áp dụng để hiểu sâu hơn về mối quan hệ giữa các biến.

### 3.2. Máy học và học máy:

Máy học và học máy là các lĩnh vực của trí tuệ nhân tạo tập trung vào việc phát triển các thuật toán và mô hình máy tính có khả năng học từ dữ liệu và thực hiện các dự đoán hoặc phân loại tự động. Các thuật toán máy học như cây quyết định, máy vector hỗ trợ (SVM), mạng nơ-ron, và học sâu được sử dụng để phát hiện mẫu ẩn và dự đoán kết quả từ dữ liệu. Máy học có thể áp dụng cho nhiều mục đích khác nhau như dự đoán giá cả, phân loại ảnh, và phát hiện gian lận.

### 3.3. Khám phá dữ liệu:

Khám phá dữ liệu là quá trình sử dụng các công cụ và kỹ thuật để khám phá dữ liệu và tìm kiếm thông tin mới và không biết đến trước. Các phương pháp khám phá dữ liệu bao gồm phân tích nhóm, phát hiện chuỗi, phân tích đồng tham chiếu, và phân tích ngữ cảnh. Mục tiêu của khám phá dữ liệu là tìm ra các mẫu mới, xu hướng, và thông tin tiềm ẩn trong dữ liệu mà không cần có các giả định trước.

3.4. Visualizations:

Visualizations là quá trình sử dụng biểu đồ và đồ họa để trực quan hóa dữ liệu và hiểu rõ hơn về các mẫu và mối quan hệ trong dữ liệu. Các loại biểu đồ phổ biến bao gồm biểu đồ cột, biểu đồ đường, biểu đồ tròn, biểu đồ phân tán, và biểu đồ bức xạ. Việc sử dụng visualizations giúp tạo ra những hình ảnh rõ ràng và dễ hiểu từ dữ liệu, giúp người dùng nắm bắt thông tin một cách nhanh chóng và hiệu quả.

A graph of blue bars with white text

Description automatically generated

Hình 8: Một biểu đồ đã được trực quan hóa từ dữ liệu thu được

A graph with numbers and a bar

Description automatically generated

### 3.6. Ứng dụng của phân tích dữ liệu:

Trong kinh doanh: Phân tích dữ liệu giúp doanh nghiệp hiểu rõ hơn về thị trường và khách hàng, từ đó đưa ra các chiến lược tiếp thị và sản phẩm hiệu quả.

## 

A graph of blue bars with white text

Description automatically generated

Hình 9: Trong ảnh việc là 10 mặt hàng doanh thu lớn nhất. Dựa vào đây giúp doanh nghiệp hiểu rõ hơn về thị trường

## 4. Đánh giá, kết luận của đề tài.

### 4.1. Đánh giá

Quá trình nghiên cứu và viết về đề tài "Thu thập, Xử lý và Phân tích Dữ liệu" đã mang lại nhiều kiến thức quý báu và cái nhìn sâu sắc về quy trình từ việc thu thập dữ liệu ban đầu đến việc xử lý và phân tích dữ liệu để đưa ra các kết luận và quyết định. Dưới đây là một số đánh giá và nhận xét về từng phần của đề tài:

Phần 1: Thu thập dữ liệu

Phần này giới thiệu về quy trình thu thập dữ liệu từ các nguồn khác nhau như cuộc khảo sát, phỏng vấn, cơ sở dữ liệu, mạng xã hội và trang web. Nó cung cấp cái nhìn tổng quan về các phương pháp, công cụ và kỹ thuật để thu thập dữ liệu một cách hiệu quả và đáng tin cậy. Tuy nhiên, để phần này trở nên đầy đủ hơn, có thể thêm vào ví dụ cụ thể và kịch bản thực tế về quá trình thu thập dữ liệu từ các nguồn khác nhau.

Phần 2: Xử lý dữ liệu

Phần này mô tả quá trình xử lý dữ liệu sau khi nó đã được thu thập, bao gồm tiền xử lý, phân tích và biến đổi, ghép nối và kết hợp, và bảo mật và quản lý dữ liệu. Nó cung cấp cái nhìn tổng quan về các bước cần thiết để làm cho dữ liệu có ý nghĩa và hữu ích cho quá trình phân tích và ra quyết định. Tuy nhiên, có thể cung cấp thêm ví dụ và kịch bản cụ thể về các thách thức và giải pháp trong quá trình xử lý dữ liệu.

Phần 3: Phân tích dữ liệu

Phần này trình bày các phương pháp phân tích dữ liệu như phân tích số liệu thống kê, máy học và học máy, khám phá dữ liệu, và visualizations. Nó cung cấp cái nhìn tổng quan về cách sử dụng các công cụ và kỹ thuật để tìm kiếm mẫu, mối quan hệ và thông tin từ dữ liệu đã được xử lý. Tuy nhiên, để phần này trở nên đầy đủ hơn, có thể cung cấp thêm các ví dụ cụ thể và ứng dụng trong các lĩnh vực khác nhau.

### 4.2. Kết luận

Qua quá trình nghiên cứu và viết về đề tài "Thu thập, Xử lý và Phân tích Dữ liệu", chúng tôi nhận thấy sự quan trọng và cần thiết của quá trình này trong việc đưa ra quyết định và nghiên cứu trong nhiều lĩnh vực khác nhau. Việc hiểu rõ về các phương pháp, công cụ và kỹ thuật trong quá trình này không chỉ giúp chúng ta khai thác thông tin từ dữ liệu một cách hiệu quả mà còn giúp chúng ta đưa ra các quyết định có căn cứ và chiến lược. Tuy nhiên, để phát triển thêm, việc áp dụng các kiến thức và kỹ năng này vào các dự án và thực tế là cực kỳ quan trọng. Hiện nay các web thương mại điện tử cũng đã tăng cường chặn truy cập tự động, chặn API, … từ phía người dùng để tránh bị cào data. Điển hình của tiki từ cuối tháng 1 đã đổi sang giao diện khác với trước đó, lazada chặn truy cập qua API, shopee chặn truy cập tự động bằng cách đăng nhập. Để tiếp tục có thể cào dữ liệu từ phía các trang web cần phân tích, cần hiểu sâu cách hoạt động của web.

Đồng thời, việc tiếp tục nghiên cứu và theo dõi các xu hướng và công nghệ mới cũng sẽ giúp nâng cao hiệu quả và hiệu suất của quá trình này trong tương lai.

Trong kết luận, chúng tôi hy vọng rằng đề tài này đã cung cấp cho độc giả một cái nhìn tổng quan và sâu sắc về quy trình thu thập, xử lý và phân tích dữ liệu, và cung cấp các kiến thức và cái nhìn cần thiết để áp dụng vào thực tế và nghiên cứu trong tương lai.

# III. Một số thành phần khác của đồ án

## 1. Sơ đồ luồng đi của dữ liệu

- Loại mặt hàng: Dữ liệu cào về sẽ nằm ở đây

- Danh mục mặt hàng: Tổng hợp dữ liệu của nhiều loại mặt hàng

- Ngày: Tổng hợp dữ liệu của nhiều danh mục mặt hàng

- Sàn thương mại điện tử: Dữ liệu tổng hợp của nhiều ngày

A diagram of a flowchart

Description automatically generated