**TRƯỜNG ĐẠI HỌC THĂNG LONG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**-----□□□□-----**

Logo

Description automatically generated

**NHẬP MÔN KHOA HỌC DỮ LIỆU**

**Thu Thập Dữ Liệu Máy Tính Từ CellphoneS, Thegioididong và FPT Bằng Selenium**

|  |  |
| --- | --- |
| **GIÁO VIÊN HƯỚNG DẪN** | **NHÓM SINH VIÊN THỰC HIỆN** |
| khung doi **Đoàn Trung Phong** | A42171 - Nguyễn Thành Đạt |
|  |  |
|  |  |

**HÀ NỘI**

**MỤC LỤC**

[Chương I. Giới thiệu chung 3](#_Toc179062839)

[**1.1 Giới thiệu về bài toán** 3](#_Toc179062840)

[**1.2 Giới hạn đối tượng khách thể, đối tượng nghiên cứu** 3](#_Toc179062841)

[**1.3. Các nghiên cứu liên quan** 3](#_Toc179062842)

[Chương II. Cơ sở lý thuyết và công cụ sử dụng 3](#_Toc179062843)

[**2.1. Lý thuyết thu thập dữ liệu** 3](#_Toc179062844)

[**2.2. Các framework và công cụ sử dụng** 4](#_Toc179062845)

[**2.3. Quy trình thu thập dữ liệu** 4](#_Toc179062846)

[Chương III. Phương pháp thực hiện 4](#_Toc179062847)

[**3.1. Mô tả quy trình các bước thực hiện** 4](#_Toc179062848)

[**3.2. Mô tả quy trình các bước thực hiện** 5](#_Toc179062849)

[**3.2.1. Xác Định Mục Tiêu** 6](#_Toc179062850)

[**3.2.2. Nghiên Cứu Cấu Trúc HTML** 6](#_Toc179062851)

[**3.2.3. Tạo Kịch Bản Thu Thập Dữ Liệu** 6](#_Toc179062852)

[**3.2.4. Thực Hiện Thu Thập Dữ Liệu** 6](#_Toc179062853)

[**3.2.5. Phân Tích và Báo Cáo Dữ Liệu** 6](#_Toc179062854)

[**3.2.6. Đánh Giá và Cải Tiến** 6](#_Toc179062855)

[Chương IV. Kết quả thực nghiệm 6](#_Toc179062856)

[**4.1. Kết quả thu thập dữ liệu** 6](#_Toc179062857)

[**4.2. Phân tích khai thác dữ liệu** 8](#_Toc179062858)

[**4.3. Nhận xét về kết quả** 9](#_Toc179062859)

[**4.4. Định hướng phát triển bài làm trong tương lai** 10](#_Toc179062860)

# 

**MỤC LỤC HÌNH ẢNH**

[Hình 1: Quy trình các bước thực hiện 7](#_Toc179063966)

[Hình 2: Chuẩn bị dữ liệu 8](#_Toc179063967)

[Hình 3: Xử lý dữ liệu không hợp lệ 9](#_Toc179063968)

[Hình 4: Thay thế các giá trị không hợp lệ 9](#_Toc179063969)

[Hình 4: Xu hướng máy tính 10](#_Toc179063970)

[Hình 6: Xu hướng về giá 10](#_Toc179063971)

[Hình 7: Đánh giá sản phẩm 11](#_Toc179063972)

# **Chương I. Giới thiệu chung**

## **1.1 Giới thiệu về bài toán**

## Trong thời đại công nghệ số hiện nay, việc thu thập và phân tích dữ liệu từ các trang web thương mại điện tử trở nên ngày càng quan trọng. Đặc biệt, trong lĩnh vực bán lẻ điện tử, việc nắm bắt thông tin về sản phẩm như laptop, điện thoại, và các thiết bị công nghệ khác không chỉ giúp người tiêu dùng đưa ra quyết định mua sắm thông minh hơn mà còn hỗ trợ các doanh nghiệp trong việc cạnh tranh trên thị trường.

## Bài toán "Thu Thập Dữ Liệu Laptop Từ CellphoneS, Thegioididong và FPT Bằng Selenium" tập trung vào việc xây dựng một hệ thống tự động thu thập thông tin về các mẫu laptop từ ba trang web thương mại điện tử lớn tại Việt Nam. Mục tiêu chính của bài toán này là:

## **Thu thập thông tin sản phẩm**: Bao gồm tên sản phẩm, giá cả, đặc điểm kỹ thuật, và các thông tin liên quan khác.

## **Tự động hóa quy trình**: Sử dụng Selenium để tự động hóa việc truy cập và thu thập dữ liệu từ các trang web, giảm thiểu thời gian và công sức so với việc thu thập thủ công.

## **Phân tích dữ liệu**: Dữ liệu thu thập được sẽ được sử dụng để phân tích xu hướng thị trường, so sánh giá cả giữa các nhà cung cấp, và cung cấp thông tin hữu ích cho người tiêu dùng.

## **Lợi ích của bài toán**

## **Hỗ trợ quyết định mua sắm**: Cung cấp cho người tiêu dùng cái nhìn tổng quan về các sản phẩm trên thị trường, giúp họ dễ dàng so sánh và lựa chọn.

## **Phân tích thị trường**: Giúp các doanh nghiệp hiểu rõ hơn về cạnh tranh và xu hướng tiêu dùng.

## **1.2 Giới hạn đối tượng khách thể, đối tượng nghiên cứu**

## **1. Khách thể nghiên cứu**

## Khách thể nghiên cứu trong bài toán "Thu Thập Dữ Liệu Laptop Từ CellphoneS, Thegioididong và FPT Bằng Selenium" là các sản phẩm laptop được niêm yết trên ba trang web thương mại điện tử lớn tại Việt Nam. Việc xác định khách thể này rất quan trọng vì nó tạo ra một khuôn khổ rõ ràng cho việc thu thập và phân tích dữ liệu.

## **Các mẫu laptop**:

## Khách thể nghiên cứu bao gồm tất cả các mẫu laptop đang được bán trên ba trang web, từ các thương hiệu nổi tiếng như Dell, HP, Lenovo, ASUS, Acer, đến các thương hiệu ít phổ biến hơn. Mỗi thương hiệu thường có nhiều dòng sản phẩm khác nhau, phục vụ cho các phân khúc thị trường khác nhau, từ học sinh, sinh viên đến các chuyên gia và doanh nhân.

## Các mẫu laptop có thể được phân loại theo nhiều tiêu chí khác nhau như cấu hình (core i3, core i5, core i7), kích thước màn hình (13 inch, 15 inch, 17 inch), và mục đích sử dụng (gaming, văn phòng, thiết kế đồ họa).

## **Thông tin chi tiết**:

## Các thông tin về sản phẩm là rất đa dạng và phong phú. Chúng bao gồm:

## **Tên sản phẩm**: Tên đầy đủ của mẫu laptop, thường bao gồm thông tin về thương hiệu và dòng sản phẩm.

## **Giá cả**: Giá được niêm yết trên từng trang web, có thể thay đổi theo thời gian và khuyến mãi.

## **Thông số kỹ thuật**: Các thông số như CPU (thế hệ và loại), RAM, dung lượng ổ cứng (HDD/SSD), card đồ họa, và kích thước màn hình. Những thông số này sẽ giúp người tiêu dùng dễ dàng so sánh và lựa chọn sản phẩm phù hợp.

## **Đánh giá từ người tiêu dùng**: Các nhận xét và điểm số từ người dùng trước đó có thể cung cấp cái nhìn tổng quan về chất lượng và hiệu suất của laptop.

## **2. Đối tượng nghiên cứu**

## Đối tượng nghiên cứu là các yếu tố và thông tin cụ thể mà bài toán sẽ phân tích và thu thập. Trong trường hợp này, đối tượng nghiên cứu bao gồm:

## **Giá cả**:

## Việc thu thập thông tin về giá cả là một phần quan trọng trong nghiên cứu. Giá của các mẫu laptop không chỉ bị ảnh hưởng bởi thương hiệu mà còn bởi các yếu tố như cấu hình, tính năng và chương trình khuyến mãi. Việc so sánh giá giữa các trang web sẽ giúp xác định nơi nào cung cấp mức giá tốt nhất cho cùng một sản phẩm.

## Ngoài ra, việc phân tích sự biến động giá theo thời gian cũng có thể cung cấp cái nhìn sâu sắc về xu hướng thị trường và các yếu tố tác động đến giá cả.

## **Tính năng sản phẩm**:

## Các thông số kỹ thuật và tính năng của từng mẫu laptop sẽ được ghi nhận và phân tích. Các tính năng như thời gian pin, khả năng nâng cấp phần cứng, và thiết kế cũng đóng vai trò quan trọng trong quyết định mua sắm của người tiêu dùng.

## Sự đa dạng trong tính năng cũng phản ánh sự phát triển của công nghệ và nhu cầu ngày càng cao của người tiêu dùng. Việc phân tích các tính năng này sẽ giúp hiểu rõ hơn về sự cạnh tranh giữa các thương hiệu.

## **Xu hướng mua sắm**:

## Dữ liệu thu thập được sẽ cho phép phân tích xu hướng tiêu dùng hiện tại. Những sản phẩm nào được ưa chuộng nhiều nhất? Có sự khác biệt nào giữa các phân khúc người tiêu dùng (học sinh, sinh viên, doanh nhân) không? Những câu hỏi này có thể được trả lời thông qua việc phân tích dữ liệu.

## Ngoài ra, việc theo dõi xu hướng theo thời gian có thể giúp dự đoán các sản phẩm nào sẽ trở nên phổ biến trong tương lai, giúp các nhà bán lẻ điều chỉnh chiến lược kinh doanh của họ.

## **Đánh giá và nhận xét**:

## Các phản hồi từ khách hàng về sản phẩm không chỉ giúp đánh giá mức độ hài lòng mà còn cung cấp thông tin quý giá về những điểm mạnh và điểm yếu của từng mẫu laptop. Những đánh giá tích cực có thể giúp tăng độ tin cậy của sản phẩm, trong khi những phản hồi tiêu cực có thể chỉ ra các vấn đề cần cải thiện.

## Việc phân tích các đánh giá này cũng có thể giúp hiểu rõ hơn về tâm lý tiêu dùng và các yếu tố ảnh hưởng đến quyết định mua sắm.

## **1.3. Các nghiên cứu liên quan**

Trong bối cảnh phát triển nhanh chóng của công nghệ thông tin và thương mại điện tử, nhiều nghiên cứu đã được tiến hành nhằm khám phá các khía cạnh khác nhau của việc thu thập và phân tích dữ liệu từ các trang web bán lẻ. Dưới đây là một số nghiên cứu liên quan có thể tham khảo:

* [**https://www.kaggle.com/datasets/muhammetvarl/laptop-price**](https://www.kaggle.com/datasets/muhammetvarl/laptop-price%20)
* [**https://www.kaggle.com/datasets/juanmerinobermejo/laptops-price-dataset**](https://www.kaggle.com/datasets/juanmerinobermejo/laptops-price-dataset%20)
* [**https://www.kaggle.com/datasets/rajugc/laptop-selection-dataset**](https://www.kaggle.com/datasets/rajugc/laptop-selection-dataset)

Các nghiên cứu liên quan đến thu thập và phân tích dữ liệu trong lĩnh vực thương mại điện tử không chỉ cung cấp kiến thức quý giá cho các nhà nghiên cứu mà còn giúp các doanh nghiệp tối ưu hóa quy trình kinh doanh của họ. Bằng cách tham khảo các nghiên cứu này, bài toán "Thu Thập Dữ Liệu Laptop Từ CellphoneS, Thegioididong và FPT Bằng Selenium" sẽ có thêm cơ sở lý thuyết vững chắc và các phương pháp thực tiễn để thực hiện hiệu quả.

# **Chương II. Cơ sở lý thuyết và công cụ sử dụng**

## **2.2. Các framework và công cụ sử dụng**

## Trong quá trình thu thập và phân tích dữ liệu từ các trang web thương mại điện tử, việc lựa chọn các công cụ và framework phù hợp là vô cùng quan trọng. Hai công cụ nổi bật được sử dụng trong bài toán này là Selenium và Pandas. Dưới đây là phần mô tả chi tiết về từng công cụ.

## **1. Selenium**

## Selenium là một công cụ mã nguồn mở mạnh mẽ, được thiết kế để tự động hóa các trình duyệt web. Nó cho phép người dùng thực hiện các tác vụ tự động hóa phức tạp trên các trang web một cách dễ dàng và hiệu quả. Dưới đây là một số điểm nổi bật của Selenium:

## Khả Năng Tương Tác Với Trình Duyệt: Selenium cho phép người dùng điều khiển trình duyệt để thực hiện các thao tác như nhấp chuột, nhập văn bản, cuộn trang, và thu thập thông tin. Điều này cực kỳ hữu ích khi làm việc với các trang web động, nơi mà nội dung có thể thay đổi dựa trên các tương tác của người dùng.

## Hỗ Trợ Nhiều Trình Duyệt: Selenium hỗ trợ một loạt các trình duyệt phổ biến như Google Chrome, Mozilla Firefox, Safari và Internet Explorer. Điều này giúp người dùng linh hoạt trong việc chọn trình duyệt phù hợp với nhu cầu của mình.

## Khả Năng Xử Lý JavaScript: Nhiều trang web hiện đại sử dụng JavaScript để tải nội dung. Selenium có khả năng tương tác với JavaScript, cho phép thu thập dữ liệu từ các trang web mà các công cụ khác có thể gặp khó khăn.

## Tính Năng Kiểm Thử: Ngoài việc thu thập dữ liệu, Selenium còn được sử dụng rộng rãi trong kiểm thử phần mềm. Các nhà phát triển có thể sử dụng Selenium để viết các kịch bản kiểm thử tự động cho các ứng dụng web, giúp tiết kiệm thời gian và nâng cao chất lượng sản phẩm.

## Tài Liệu Hỗ Trợ và Cộng Đồng: Selenium có một cộng đồng lớn và tài liệu phong phú, giúp người dùng dễ dàng tìm kiếm thông tin và hỗ trợ khi cần thiết. Điều này làm cho việc học hỏi và giải quyết vấn đề trở nên đơn giản hơn.

## Selenium đã trở thành công cụ tiêu chuẩn trong nhiều dự án thu thập dữ liệu và tự động hóa web, nhờ vào tính linh hoạt và khả năng tương tác cao của nó.

## **2. Pandas**

## Pandas là một thư viện Python nổi tiếng, được thiết kế để xử lý và phân tích dữ liệu. Được phát triển bởi Wes McKinney vào năm 2008, Pandas đã nhanh chóng trở thành một trong những công cụ không thể thiếu trong lĩnh vực khoa học dữ liệu và phân tích dữ liệu. Dưới đây là một số tính năng nổi bật của Pandas:

## Khả Năng Làm Việc Với Dữ Liệu Dạng Bảng: Pandas cung cấp hai cấu trúc dữ liệu chính là Series và DataFrame. DataFrame cho phép người dùng làm việc với dữ liệu dưới dạng bảng, tương tự như bảng trong cơ sở dữ liệu hoặc bảng tính Excel. Điều này giúp người dùng dễ dàng lưu trữ và tổ chức dữ liệu thu thập được.

## Công Cụ Phân Tích Mạnh Mẽ: Pandas cung cấp nhiều công cụ để xử lý và phân tích dữ liệu, bao gồm khả năng lọc dữ liệu, nhóm dữ liệu theo các tiêu chí khác nhau, và thực hiện các phép toán thống kê như tính trung bình, độ lệch chuẩn, và tổng. Những công cụ này giúp người dùng nhanh chóng rút ra thông tin giá trị từ dữ liệu.

## Tích Hợp Dễ Dàng Với Các Thư Viện Khác: Pandas có khả năng tích hợp tốt với các thư viện khác trong Python như NumPy, Matplotlib và SciPy. Điều này cho phép người dùng thực hiện các phân tích phức tạp và trực quan hóa dữ liệu một cách hiệu quả.

## Xử Lý Dữ Liệu Thiếu: Pandas cung cấp các chức năng để xử lý dữ liệu thiếu, cho phép người dùng dễ dàng điền, loại bỏ hoặc thay thế các giá trị thiếu trong bộ dữ liệu. Điều này rất quan trọng trong việc đảm bảo tính chính xác của phân tích dữ liệu.

## Tốc Độ và Hiệu Suất: Pandas được tối ưu hóa cho hiệu suất cao, cho phép xử lý các tập dữ liệu lớn mà không làm giảm tốc độ. Điều này rất hữu ích trong các ứng dụng cần xử lý dữ liệu thời gian thực hoặc phân tích dữ liệu quy mô lớn.

## Pandas đã trở thành công cụ không thể thiếu cho các nhà phân tích dữ liệu, nhà khoa học dữ liệu và lập trình viên Python, nhờ vào khả năng cung cấp các công cụ mạnh mẽ cho việc xử lý và phân tích dữ liệu.

## Việc sử dụng Selenium và Pandas trong bài toán thu thập dữ liệu laptop từ các trang web thương mại điện tử không chỉ giúp tối ưu hóa quy trình thu thập và phân tích dữ liệu, mà còn nâng cao hiệu quả và chất lượng của công việc. Selenium cho phép thu thập dữ liệu từ các trang web động một cách dễ dàng, trong khi Pandas cung cấp các công cụ mạnh mẽ để xử lý và phân tích dữ liệu thu thập được. Sự kết hợp giữa hai công cụ này tạo ra một nền tảng vững chắc cho việc nghiên cứu và phân tích thông tin trong lĩnh vực thương mại điện tử.

## **2.3. Quy trình thu thập dữ liệu**

# Quy trình thu thập dữ liệu từ các trang web thương mại điện tử là một chu trình có hệ thống, bao gồm nhiều bước liên quan đến việc truy cập, lấy thông tin, lưu trữ và phân tích dữ liệu. Dưới đây là mô tả chi tiết cho từng bước trong quy trình này:

# **1. Truy Cập Trang Web**

# Bước đầu tiên trong quy trình thu thập dữ liệu là truy cập vào trang web mà muốn thu thập thông tin. Để thực hiện điều này, chúng tôi sẽ sử dụng Selenium, một công cụ tự động hóa trình duyệt. Các bước cụ thể bao gồm:

# **Khởi Tạo WebDriver**: Đầu tiên, bạn cần khởi tạo WebDriver cho trình duyệt đã chọn (thường là Chrome hoặc Firefox). WebDriver sẽ cho phép tôi điều khiển trình duyệt từ mã Python .

# **Mở Trình Duyệt**: Sử dụng phương thức get() của WebDriver để mở trang web mục tiêu. Ví dụ, nếu bạn muốn thu thập dữ liệu từ CellphoneS, bạn sẽ sử dụng driver.get("https://cellphones.com.vn/laptop.html").

# **Chờ Tải Trang**: Sau khi mở trang, cần đợi cho nội dung trang được tải hoàn toàn. Có thể sử dụng các phương thức chờ của Selenium để đảm bảo rằng bạn chỉ bắt đầu thu thập dữ liệu khi tất cả các phần tử cần thiết đã sẵn sàng.

# **2. Lấy Thông Tin**

# Sau khi truy cập vào trang web, bước tiếp theo là lấy thông tin cần thiết. Để thực hiện điều này, bạn sẽ sử dụng các selector CSS hoặc XPath để xác định các thành phần cụ thể mà bạn muốn thu thập. Các bước cụ thể bao gồm:

# **Xác Định Selector**: Sử dụng công cụ phát triển của trình duyệt (như Chrome DevTools) để tìm hiểu cấu trúc HTML của trang và xác định các selector cho các thành phần cần thu thập. Ví dụ, bạn có thể lấy tên sản phẩm bằng cách sử dụng một selector cụ thể như h3.product-title.

# **Sử Dụng Selenium Để Lấy Dữ Liệu**: Sử dụng các phương thức như find\_element\_by\_css\_selector() hoặc find\_elements\_by\_xpath() để lấy các phần tử tương ứng. Bạn có thể lặp qua danh sách các sản phẩm trên trang và thu thập tên, giá cả, mô tả, và đánh giá cho từng sản phẩm.

# **Xử Lý Dữ Liệu**: Đôi khi, dữ liệu thu thập được có thể cần được xử lý thêm, chẳng hạn như loại bỏ khoảng trắng, chuyển đổi kiểu dữ liệu, hoặc thay đổi định dạng để dễ sử dụng hơn.

# **3. Lưu Trữ Dữ Liệu**

# Sau khi thu thập thông tin, bước tiếp theo là lưu trữ dữ liệu vào một định dạng dễ dàng xử lý. Pandas là một thư viện Python rất hữu ích cho việc này. Các bước cụ thể bao gồm:

# **Tạo DataFrame**: Sử dụng Pandas để tạo một DataFrame từ dữ liệu đã thu thập. DataFrame cho phép bạn tổ chức dữ liệu theo dạng bảng, với các cột đại diện cho các thuộc tính như tên sản phẩm, giá cả, mô tả, và đánh giá.

# **Lưu Dữ Liệu**: Sau khi tạo DataFrame, bạn có thể lưu dữ liệu vào các định dạng phổ biến như CSV hoặc Excel bằng cách sử dụng các phương thức như to\_csv() hoặc to\_excel(). Điều này giúp bạn dễ dàng chia sẻ và sử dụng dữ liệu trong các ứng dụng phân tích khác.

# **Kiểm Tra Dữ Liệu**: Trước khi hoàn tất, hãy kiểm tra dữ liệu đã lưu để đảm bảo rằng mọi thông tin cần thiết đều được thu thập và lưu trữ đúng cách.

# **4. Phân Tích Dữ Liệu**

# Bước cuối cùng trong quy trình thu thập dữ liệu là phân tích dữ liệu đã thu thập. Đây là giai đoạn quan trọng giúp bạn rút ra các kết luận và thông tin giá trị từ dữ liệu. Một số hướng phân tích có thể bao gồm:

# **Khám Phá Dữ Liệu**: Sử dụng các công cụ và phương pháp thống kê cơ bản để khám phá dữ liệu, như tính toán các số liệu thống kê mô tả (trung bình, độ lệch chuẩn, v.v.) để hiểu rõ hơn về các thuộc tính của sản phẩm.

# **Phân Tích Xu Hướng**: Sử dụng các biểu đồ và đồ thị để trực quan hóa dữ liệu, giúp dễ dàng nhận thấy các xu hướng trong thị trường laptop, như sự biến động của giá cả theo thời gian hoặc sự phổ biến của các thương hiệu khác nhau.

# **So Sánh và Đánh Giá**: So sánh các sản phẩm dựa trên các tiêu chí như giá cả, tính năng, và đánh giá từ người tiêu dùng. Điều này có thể giúp bạn xác định sản phẩm nào có giá trị tốt nhất cho người tiêu dùng.

# **Rút Ra Kết Luận**: Cuối cùng, dựa trên các phân tích đã thực hiện, bạn sẽ rút ra các kết luận có thể hỗ trợ người tiêu dùng trong việc đưa ra quyết định mua sắm và giúp các doanh nghiệp điều chỉnh chiến lược của họ.

# **Chương III. Phương pháp thực hiện**

## **3.1. Mô tả quy trình các bước thực hiện**

## Quy trình thu thập dữ liệu từ các trang web thương mại điện tử như CellphoneS, Thegioididong và FPT Shop được thực hiện qua một chuỗi các bước có hệ thống. Mỗi bước đều đóng vai trò quan trọng trong việc đảm bảo rằng dữ liệu thu thập được chính xác, đầy đủ và có thể sử dụng cho các mục đích phân tích. Dưới đây là mô tả chi tiết cho từng bước trong quy trình này:

## **Bước 1: Xác Định Mục Tiêu Thu Thập Dữ Liệu**

## **Mục tiêu**: Bước đầu tiên là xác định rõ ràng các thông tin cần thu thập từ các trang web. Những thông tin này có thể bao gồm:

## **Tên sản phẩm**: Để biết sản phẩm nào đang được cung cấp.

## **Giá cả**: Để so sánh giá giữa các sản phẩm và các nhà cung cấp.

## **Mô tả sản phẩm**: Để hiểu rõ hơn về các tính năng và thông số kỹ thuật.

## **Đánh giá từ người tiêu dùng**: Để có cái nhìn tổng quan về chất lượng sản phẩm và sự hài lòng của khách hàng.

## **Lập danh sách**: Sau khi xác định được các thông tin cần thu thập, bạn cần tạo một danh sách các sản phẩm cụ thể hoặc các danh mục sản phẩm mà bạn muốn thu thập thông tin. Danh sách này có thể được tổ chức theo từng thương hiệu, loại sản phẩm hoặc theo các tiêu chí khác nhau để dễ dàng theo dõi và quản lý.

## **Bước 2: Nghiên Cứu Cấu Trúc HTML của Trang Web**

## **Phân tích HTML**: Sử dụng các công cụ phát triển của trình duyệt, như "Inspect Element" trong Google Chrome hoặc Firefox, để nghiên cứu cấu trúc HTML của trang web. Bước này giúp bạn hiểu cách mà dữ liệu được tổ chức trên trang và cách mà bạn có thể truy cập nó.

## **Ghi chú các selector**: Trong quá trình phân tích, hãy ghi lại các selector CSS hoặc XPath cho từng phần tử cần thu thập thông tin. Ví dụ:

## Selector cho tên sản phẩm có thể là h3.product-title.

## Selector cho giá cả có thể là .price.

## Selector cho mô tả có thể là .product-description.

## Việc ghi chú này sẽ giúp bạn thực hiện các bước thu thập dữ liệu sau này một cách chính xác và hiệu quả.

## **Bước 3: Tạo Kịch Bản Thu Thập Dữ Liệu**

## **Lên kế hoạch**: Trước khi bắt đầu viết mã, hãy xác định quy trình tự động hóa mà bạn sẽ sử dụng với Selenium. Điều này bao gồm việc quyết định thứ tự các hành động mà bạn sẽ thực hiện trên trình duyệt.

## **Mô tả các hành động**: Hãy xác định các hành động cần thực hiện. Điều này có thể bao gồm:

## Mở trang web chính xác mà bạn muốn thu thập dữ liệu.

## Nhấp vào các liên kết sản phẩm để truy cập vào từng trang chi tiết.

## Lấy thông tin từ các phần tử mà bạn đã xác định ở bước 2.

## Mô tả các hành động này sẽ giúp bạn dễ dàng viết mã và đảm bảo rằng quy trình thu thập dữ liệu diễn ra suôn sẻ.

## **Bước 4: Thực Hiện Thu Thập Dữ Liệu**

## **Mở trình duyệt**: Sử dụng Selenium để mở trình duyệt và truy cập các trang web đã xác định trước đó. Bạn cần khởi tạo WebDriver và sử dụng phương thức get() để mở từng trang.

## **Lấy dữ liệu**: Sử dụng các selector đã ghi chú để lấy thông tin từ các sản phẩm. Bạn có thể lặp qua danh sách các sản phẩm trên trang và sử dụng các phương thức như find\_element\_by\_css\_selector() để lấy dữ liệu.

## **Lưu trữ dữ liệu**: Sau khi thu thập xong, dữ liệu cần được lưu trữ vào định dạng như CSV hoặc Excel. Sử dụng Pandas để tạo DataFrame từ dữ liệu thu thập được và sau đó lưu nó vào file bằng các phương thức như to\_csv() hoặc to\_excel(). Điều này không chỉ giúp dễ dàng xử lý và phân tích dữ liệu sau này mà còn giúp bạn quản lý dữ liệu một cách có tổ chức.

## **Bước 5: Phân Tích và Báo Cáo Dữ Liệu**

## **Phân tích dữ liệu**: Sau khi lưu trữ dữ liệu, bước tiếp theo là sử dụng Pandas và các công cụ phân tích khác để xử lý và phân tích dữ liệu thu thập được. Bạn có thể thực hiện các phân tích thống kê cơ bản, tính toán các số liệu tổng hợp hoặc khám phá các mối quan hệ giữa các biến.

## **Trình bày kết quả**: Cuối cùng, hãy tạo một báo cáo tổng hợp về các xu hướng giá cả, sản phẩm phổ biến và các thông tin quan trọng khác mà bạn đã rút ra từ phân tích. Bạn có thể sử dụng các biểu đồ và đồ thị để trực quan hóa dữ liệu, giúp người đọc dễ dàng nắm bắt thông tin.

## **Bước 6: Đánh Giá và Cải Tiến**

## **Đánh giá quy trình**: Sau khi hoàn thành quy trình thu thập dữ liệu, hãy xem xét hiệu quả của các bước đã thực hiện. Đánh giá xem có vấn đề gì phát sinh trong quá trình thu thập dữ liệu không, chẳng hạn như dữ liệu không đầy đủ, lỗi trong việc xác định selector, hoặc thời gian thu thập dữ liệu quá lâu.

## **Cải tiến**: Dựa trên những đánh giá này, hãy đưa ra các điều chỉnh cần thiết để cải thiện quy trình thu thập dữ liệu trong các lần thực hiện sau. Điều này có thể bao gồm việc tối ưu hóa mã, cải thiện khả năng xử lý lỗi hoặc thay đổi cách thức thu thập dữ liệu để nâng cao độ chính xác và hiệu quả.

## **3.2. Mô tả quy trình các bước thực hiện**

**A diagram with black text

Description automatically generated**

Hình 1: Quy trình các bước thực hiện

### **3.2.1. Xác Định Mục Tiêu**

* **Lập danh sách sản phẩm cụ thể**: Xác định các sản phẩm cần thu thập thông tin.
* **Thông tin cần thu thập**: Bao gồm tên, giá, mô tả và đánh giá của sản phẩm…

### **3.2.2. Nghiên Cứu Cấu Trúc HTML**

* **Sử dụng "Inspect Element"**: Kiểm tra mã nguồn HTML của trang web.
* **Ghi chú selector**: Lưu lại các selector cho từng thông tin cần thu thập.

### **3.2.3. Tạo Kịch Bản Thu Thập Dữ Liệu**

* **Xác định trình tự các bước**: Mở trang, nhấp chuột và lấy thông tin.
* **Phương án xử lý lỗi**: Đề xuất cách xử lý khi không tìm thấy thông tin.

### **3.2.4. Thực Hiện Thu Thập Dữ Liệu**

* **Mở trình duyệt**: Truy cập trang web đã xác định.
* **Thực hiện các hành động**: Nhấp chuột, cuộn trang để thu thập dữ liệu.
* **Lưu thông tin**: Ghi lại dữ liệu vào danh sách tạm thời.

### **3.2.5. Phân Tích và Báo Cáo Dữ Liệu**

* **Làm sạch dữ liệu**: Loại bỏ các thông tin trùng lặp và không hợp lệ.
* **Sử dụng Pandas**: Phân tích dữ liệu và tạo biểu đồ minh họa.
* **Tạo báo cáo tổng hợp**: Trình bày kết quả phân tích một cách rõ ràng.

### **3.2.6. Đánh Giá và Cải Tiến**

* **Phân tích hiệu suất**: Đánh giá thời gian thu thập và độ chính xác của dữ liệu.
* **Ghi chú vấn đề phát sinh**: Lưu lại những vấn đề gặp phải và cách khắc phục.
* **Đề xuất cải tiến**: Đưa ra các giải pháp nâng cao quy trình thu thập trong tương lai.

# **Chương IV. Kết quả thực nghiệm**

## **4.1. Kết quả thu thập dữ liệu**

**Sau khi thực hiện quy trình thu thập dữ liệu từ các trang web CellphoneS, Thegioididong và FPT Shop, chúng tôi đã thu thập được:**

* **Số lượng sản phẩm**: Tổng cộng 291 sản phẩm từ trang web Thegioididong.
* **Thông tin thu thập**: Bao gồm tên sản phẩm(data\_name), giá cả(price), loại(brand), giá cũ(discount\_list), phần trăm giảm giá (discount\_precent), danh sách thưởng(Gift\_list), RAM và SSD(RAM\_SSD), đánh giá theo 5 sao (Evaluate\_list), số lượt đánh giá( Evaluate\_count\_list)

**Quá trình làm sạch dữ liệu**:

**Chuẩn bị dữ liệu**: Thu thập các dữ liệu bằng selenium xong tập hợp thành một fie csv để phân tích A screenshot of a computer

Description automatically generated

Hình 2: Chuẩn bị dữ liệu

* **Xử lý dữ liệu không hợp lệ**: Sửa lại các kiểu dự liệu của các cột như price, discount\_list… thành các dữ liệu kiểu số cho tiện phân tích A screenshot of a computer program

  Description automatically generated

Hình 3: Xử lý dữ liệu không hợp lệ

* **Thay thế các giá trị** : Thay thế các giá trị của các cột sao cho phù hợp với dữ liệu: A screenshot of a computer program

  Description automatically generated

Hình 4: Thay thế các giá trị không hợp lệ

* **Khai thác dữ liệu**: Sau khi dữ liệu được làm sạch (kể cả về mặt dữ liệu và logic) thì sẽ đưa vào khai thác dữ liệu

**Vấn đề phát sinh**:

* Một số trang web có cơ chế bảo vệ, khiến việc thu thập dữ liệu gặp khó khăn (ví dụ: CAPTCHA hoặc chặn IP).
* Thời gian tải nhiều trang cùng một lúc có thể ảnh hưởng đến hiệu suất thu thập.

## **4.2. Phân tích khai thác dữ liệu**

Sau khi làm sạch, dữ liệu đã được phân tích để rút ra các kết luận quan trọng:

* **Xu hướng loại máy tính**:
  + **HP** là thương hiệu đứng đầu với số lượng cao nhất.
  + **Asus** và **Dell** cũng có số lượng đáng chú ý, nhưng thấp hơn HP.
  + Các thương hiệu khác như **Acer**, **MSI**, **Lenovo**, **MacBook**, **Samsung**, **ASUS**, và **Apple** có số lượng thấp hơn nhiều so với ba thương hiệu đầu tiên. A graph of different brands

    Description automatically generated

Hình 4: Xu hướng máy tính

Điều này cho thấy HP, Asus và Dell hiện đang chiếm ưu thế trong thị trường máy tính.

* **Xu hướng về giá cho từng loại máy tính** :
  + Biểu đồ cột hiển thị giá của các thương hiệu máy tính khác nhau, giúp người xem dễ dàng so sánh giá giữa các thương hiệu. A graph of different colored bars

    Description automatically generated

Hình 6: Xu hướng về giá

* **Đánh giá và phản hồi**:
  + Nghiên cứu cho thấy rằng các sản phẩm có đánh giá trên 4 sao thường có tỷ lệ bán hàng cao hơn. A graph with blue bars

    Description automatically generated

Hình 7: Đánh giá sản phẩm

## **4.3. Nhận xét về kết quả**

Kết quả thu thập dữ liệu và phân tích cho thấy quy trình thực hiện đã diễn ra khá suôn sẻ. Chúng tôi đã thu thập được một lượng dữ liệu đáng kể và có giá trị từ các trang web thương mại điện tử. Dữ liệu thu thập được không chỉ mang lại cái nhìn tổng quan về thị trường mà còn giúp phát hiện các xu hướng và nhu cầu của người tiêu dùng.

Tuy nhiên, vẫn còn một số vấn đề cần cải thiện, như tối ưu hóa quy trình thu thập và xử lý lỗi khi gặp các cơ chế bảo vệ từ trang web.

## **4.4. Định hướng phát triển bài làm trong tương lai**

Trong tương lai, dự định:

* **Mở rộng phạm vi thu thập dữ liệu**: Bao gồm các sản phẩm công nghệ khác như mobile phone, tablet và phụ kiện.
* **Sử dụng các công nghệ học máy**: Để phân tích sâu hơn về hành vi người tiêu dùng và dự đoán xu hướng thị trường.
* **Xây dựng hệ thống tự động hóa**: Để thu thập dữ liệu một cách liên tục và cập nhật theo thời gian thực.

**5. Bảng phân công nhiệm vụ trong quá trình thực hiện đề tài**

| **Tên thành viên** | **Nhiệm vụ** |
| --- | --- |
| A42171-Nguyễn Thành Đạt | Thu Thập Dữ Liệu , Làm sạch dữ liệu, Viết báo cáo, Làm PowerPoint |
|  |  |