

**MỞ ĐẦU**

Trong thời đại công nghệ số hiện nay, dữ liệu đã trở thành một nguồn tài nguyên quan trọng, đóng vai trò cốt lõi trong việc hỗ trợ ra quyết định, nghiên cứu và phát triển các ứng dụng thông minh. Đặc biệt, dữ liệu trên web – với sự phong phú và đa dạng từ các nền tảng thương mại điện tử, mạng xã hội đến các trang thông tin – đang mở ra nhiều cơ hội để khai thác và ứng dụng trong thực tiễn. Tuy nhiên, việc thu thập và xử lý dữ liệu web một cách hiệu quả đòi hỏi các kỹ thuật hiện đại như web scraping, xử lý dữ liệu lớn (Big Data) và phân tích dữ liệu, điều mà ngành Công nghệ Thông tin đang không ngừng nghiên cứu và phát triển.

Xuất phát từ nhu cầu thực tiễn và mong muốn nâng cao kỹ năng chuyên môn, nhóm sinh viên Khoa Công nghệ Thông tin, Trường Đại học Văn Lang đã lựa chọn thực hiện đồ án với chủ đề "Trích xuất và xử lý dữ liệu web" trong môn học 71ITDS40303 – Số hóa và Quản trị Thông số, học kỳ HK242. Đồ án tập trung vào việc thu thập dữ liệu sản phẩm laptop từ trang web thương mại điện tử Amazon, xử lý và phân tích để rút ra các thông tin hữu ích về thị trường, từ đó cung cấp giá trị cho cả người tiêu dùng và doanh nghiệp.

Báo cáo này trình bày toàn bộ quá trình thực hiện đồ án, bao gồm lý do chọn chủ đề, phương pháp thực hiện, kết quả đạt được và hướng phát triển trong tương lai. Thông qua dự án, nhóm không chỉ củng cố kiến thức lý thuyết mà còn rèn luyện kỹ năng lập trình, xử lý dữ liệu và làm việc nhóm, chuẩn bị hành trang cho các công việc thực tế sau này. Hy vọng rằng, đồ án không chỉ đáp ứng yêu cầu học thuật mà còn mang lại những đóng góp thiết thực trong lĩnh vực khoa học dữ liệu.

**MỤC LỤC**

[**CHƯƠNG 1. GIỚI THIỆU CHỦ ĐỀ ĐỒ ÁN** 5](#_Toc194697443)

[**1.1 Lý do chọn chủ đề** 5](#_Toc194697444)

[**1.2 Đối tượng và phạm vi tìm hiểu** 6](#_Toc194697445)

[**1.3 Mục tiêu của đồ án** 6](#_Toc194697446)

[**1.4 Ý nghĩa của đồ án** 7](#_Toc194697447)

[**1.5 Cấu trúc báo cáo** 7](#_Toc194697448)

[**CHƯƠNG 2. PHƯƠNG PHÁP THỰC HIỆN** 9](#_Toc194697449)

[**2.1 Tổng quan về phương pháp** 9](#_Toc194697450)

[**2.2 Các bước thực hiện** 9](#_Toc194697451)

[**2.2.1 Thu thập dữ liệu (Web Scraping)** 9](#_Toc194697452)

[**2.2.2 Xử lý dữ liệu** 10](#_Toc194697453)

[**2.2.3 Phân tích và trực quan hóa dữ liệu** 11](#_Toc194697454)

[**2.3 Công cụ và công nghệ sử dụng** 12](#_Toc194697455)

[**CHƯƠNG 3. KẾT QUẢ VÀ PHÂN TÍCH** 15](#_Toc194697456)

[**3.1 Kết quả thu thập dữ liệu** 15](#_Toc194697457)

[**3.2 Kết quả xử lý dữ liệu** 15](#_Toc194697458)

[**3.3.1 Phân tích giá cả** 16](#_Toc194697459)

[**3.3.2 Phân tích thương hiệu** 18](#_Toc194697460)

[**3.3.3 Phân tích đánh giá** 19](#_Toc194697461)

[**3.3.4 Phân tích đặc điểm kỹ thuật** 20](#_Toc194697462)

[**3.3.5 Tương quan giá-đánh giá** 21](#_Toc194697463)

[**3.4 Báo cáo phân tích** 22](#_Toc194697464)

[**CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 22](#_Toc194697465)

[**4.1 Kết luận** 22](#_Toc194697466)

[**4.2 Đánh giá đồ án** 23](#_Toc194697467)

[**4.3 Hướng phát triển** 23](#_Toc194697468)

[**KẾT LUẬN** 25](#_Toc194697469)

[**Ý nghĩa của đồ án:** 26](#_Toc194697470)

[**Đánh giá tổng quan:** 26](#_Toc194697471)

# **CHƯƠNG 1. GIỚI THIỆU CHỦ ĐỀ ĐỒ ÁN**

## **1.1 Lý do chọn chủ đề**

Trong bối cảnh công nghệ thông tin phát triển mạnh mẽ, dữ liệu trên web đã trở thành một nguồn tài nguyên phong phú và quan trọng, phục vụ cho nhiều lĩnh vực như nghiên cứu khoa học, kinh doanh, phân tích thị trường, trí tuệ nhân tạo (AI) và học máy. Chủ đề "Trích xuất và xử lý dữ liệu web" được nhóm lựa chọn dựa trên các lý do sau:

* Tính ứng dụng cao: Dữ liệu web chứa đựng thông tin đa dạng, từ giá cả sản phẩm, đánh giá người dùng đến thông số kỹ thuật, có thể được khai thác để hỗ trợ các quyết định kinh doanh, nghiên cứu thị trường hoặc phát triển sản phẩm mới.
* Sự phát triển của dữ liệu lớn (Big Data): Trong thời đại số hóa, khối lượng dữ liệu từ web ngày càng gia tăng, đòi hỏi các phương pháp tự động hóa hiệu quả để thu thập và xử lý thông tin một cách nhanh chóng và chính xác.
* Tính cấp thiết trong quản trị thông số: Việc thu thập và phân tích dữ liệu từ nhiều nguồn khác nhau giúp doanh nghiệp và tổ chức đưa ra các quyết định dựa trên cơ sở thực tiễn, từ đó tối ưu hóa chiến lược hoạt động.
* Ứng dụng trong trí tuệ nhân tạo và học máy: Dữ liệu web là nguồn tài nguyên quan trọng để huấn luyện các mô hình AI, chẳng hạn như hệ thống gợi ý sản phẩm, chatbot, hoặc phân tích ngữ nghĩa.
* Tích hợp với nhiều lĩnh vực: Kỹ thuật trích xuất và xử lý dữ liệu web có thể được áp dụng trong các ngành như thương mại điện tử, tài chính, y tế, giáo dục và truyền thông, mở ra nhiều cơ hội thực tiễn.
* Học hỏi công nghệ mới: Thực hiện đồ án này giúp nhóm làm quen với các công cụ và kỹ thuật hiện đại như web scraping (sử dụng BeautifulSoup, Scrapy), truy cập API, xử lý dữ liệu với Pandas, và lưu trữ dữ liệu bằng SQL hoặc NoSQL.

Với những lý do trên, nhóm quyết định chọn chủ đề "Trích xuất và xử lý dữ liệu web" để không chỉ đáp ứng yêu cầu học tập mà còn phát triển kỹ năng thực tế, chuẩn bị cho các dự án trong tương lai**.**

## **1.2 Đối tượng và phạm vi tìm hiểu**

* Đối tượng nghiên cứu: Dữ liệu sản phẩm laptop được liệt kê trên trang web thương mại điện tử Amazon ([www.amazon.com](http://www.amazon.com)). Đây là một trong những nền tảng lớn nhất thế giới, cung cấp thông tin phong phú về sản phẩm, giá cả, đánh giá người dùng và các đặc điểm kỹ thuật.
* Phạm vi tìm hiểu:
  + Nguồn dữ liệu: 20 trang kết quả tìm kiếm trên Amazon với từ khóa "laptop".
  + Thông tin trích xuất: Bao gồm tiêu đề sản phẩm, thương hiệu, giá cả, đánh giá, số lượng đánh giá, tình trạng hàng, dung lượng RAM, dung lượng lưu trữ, kích thước màn hình và thông tin giảm giá.
  + Mục tiêu phân tích: Hiểu rõ xu hướng giá cả, phân phối thương hiệu, mối quan hệ giữa giá và chất lượng (dựa trên đánh giá), và các đặc điểm kỹ thuật phổ biến của laptop trên thị trường.

## **1.3 Mục tiêu của đồ án**

Đồ án hướng đến các mục tiêu cụ thể sau:

* Thu thập dữ liệu: Xây dựng một hệ thống tự động để trích xuất thông tin sản phẩm laptop từ Amazon một cách chính xác và hiệu quả.
* Xử lý dữ liệu: Làm sạch và chuẩn hóa dữ liệu thu thập được, loại bỏ các giá trị không hợp lệ để sẵn sàng cho phân tích.
* Phân tích dữ liệu: Cung cấp các thông tin chi tiết về thị trường laptop, bao gồm xu hướng giá cả, thương hiệu phổ biến, và đặc điểm sản phẩm được ưa chuộng.
* Trực quan hóa: Trình bày kết quả phân tích dưới dạng biểu đồ và báo cáo để dễ dàng hiểu và sử dụng.
* Ứng dụng thực tiễn: Đưa ra các nhận định có giá trị cho người tiêu dùng (lựa chọn sản phẩm phù hợp) và doanh nghiệp (hiểu rõ thị trường để định hướng kinh doanh).

## **1.4 Ý nghĩa của đồ án**

* Ý nghĩa khoa học: Đồ án góp phần nghiên cứu và ứng dụng các kỹ thuật trích xuất dữ liệu web, một lĩnh vực quan trọng trong khoa học dữ liệu và công nghệ thông tin.
* Ý nghĩa thực tiễn: Kết quả của đồ án có thể hỗ trợ người tiêu dùng trong việc lựa chọn laptop phù hợp với nhu cầu và ngân sách, đồng thời cung cấp thông tin thị trường cho các doanh nghiệp thương mại điện tử.
* Ý nghĩa học tập: Giúp nhóm sinh viên củng cố kiến thức về lập trình Python, xử lý dữ liệu, và trực quan hóa, đồng thời làm quen với quy trình thực hiện một dự án thực tế từ thu thập đến phân tích dữ liệu.

## **1.5 Cấu trúc báo cáo**

Báo cáo đồ án được chia thành các chương chính như sau:

* Chương 1. Giới thiệu chủ đề đồ án: Trình bày lý do chọn đề tài, đối tượng, phạm vi, mục tiêu và ý nghĩa của đồ án.
* Chương 2. Phương pháp thực hiện: Mô tả chi tiết các bước thực hiện, công cụ và công nghệ sử dụng.
* Chương 3. Kết quả và phân tích: Trình bày dữ liệu thu thập được, kết quả xử lý, phân tích và trực quan hóa.
* Chương 4. Kết luận và hướng phát triển: Tổng kết thành tựu, đánh giá đồ án và đề xuất hướng phát triển trong tương lai.

# **CHƯƠNG 2. PHƯƠNG PHÁP THỰC HIỆN**

## **2.1 Tổng quan về phương pháp**

Phương pháp thực hiện đồ án bao gồm ba giai đoạn chính: thu thập dữ liệu (web scraping), xử lý dữ liệu và phân tích/trực quan hóa dữ liệu. Mỗi giai đoạn được thiết kế để đảm bảo dữ liệu thu thập từ trang web Amazon được trích xuất chính xác, xử lý hiệu quả và phân tích một cách khoa học. Quy trình này không chỉ đáp ứng yêu cầu học thuật mà còn phù hợp với các ứng dụng thực tế trong lĩnh vực khoa học dữ liệu.



## **2.2 Các bước thực hiện**

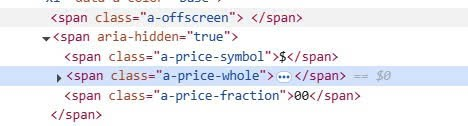
### **2.2.1 Thu thập dữ liệu (Web Scraping)**

* **Mục tiêu:** Trích xuất thông tin sản phẩm laptop từ 20 trang kết quả tìm kiếm trên Amazon với từ khóa "laptop".
* **Quy trình:** 
  1. Xác định URL: Tạo danh sách các URL cần thu thập (<https://www.amazon.com/s?k=laptop&ref=nb_sb_noss>).
  2. Gửi yêu cầu HTTP: Sử dụng thư viện **requests** để gửi yêu cầu GET tới từng URL. Để tránh bị chặn bởi hệ thống bảo vệ của Amazon, nhóm sử dụng danh sách **User-Agent** xoay vòng (ví dụ: Chrome, Safari) và thêm độ trễ ngẫu nhiên từ 1-5 giây giữa các yêu cầu bằng hàm random\_delay().
  3. Phân tích HTML: Sử dụng **BeautifulSoup** để phân tích cú pháp nội dung HTML của trang, trích xuất các liên kết sản phẩm (có lớp CSS a-link-normal s-no-outline).



(class tittle)

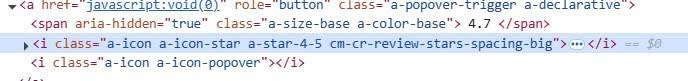
* 1. Thu thập chi tiết sản phẩm: Với mỗi liên kết sản phẩm, thực hiện yêu cầu HTTP mới để lấy trang chi tiết và trích xuất các thông tin như tiêu đề (get\_title), giá (get\_price), đánh giá (get\_rating), v.v.



(class price)



(class ram)



(class rating)

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

(screen size)

* Xử lý lỗi: Xử lý các trường hợp ngoại lệ như mã trạng thái HTTP không phải 200 hoặc dữ liệu không tồn tại bằng cách ghi log lỗi và tiếp tục với sản phẩm tiếp theo.

### **2.2.2 Xử lý dữ liệu**

* **Mục tiêu:** Chuẩn hóa dữ liệu thô để sẵn sàng cho phân tích.
* **Quy trình:** 
  1. Tạo DataFrame: Dữ liệu thu thập được lưu vào một từ điển (all\_data) và chuyển thành DataFrame bằng pandas.
  2. Làm sạch dữ liệu:
     + Loại bỏ các bản ghi có tiêu đề trống (title).
     + Chuẩn hóa giá (price\_clean) bằng hàm clean\_price, loại bỏ ký hiệu tiền tệ và chuyển thành số thực.
     + Chuẩn hóa đánh giá (rating\_clean) và số lượng đánh giá (review\_count\_clean) bằng các hàm clean\_rating và clean\_review\_count.
  3. Phân loại: Phân loại laptop theo mức giá (Budget: < $200, Mid-range: $200-$500, High-end: $500-$800, Premium: > $800) bằng hàm categorize\_by\_price.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

* 1. Loại bỏ trùng lặp: Xóa các sản phẩm trùng lặp dựa trên tiêu đề để đảm bảo tính duy nhất của dữ liệu.
* Lưu trữ: Dữ liệu thô lưu vào amazon\_raw\_data.csv, dữ liệu đã xử lý lưu vào amazon\_processed\_data.csv.

### **2.2.3 Phân tích và trực quan hóa dữ liệu**

* Mục tiêu: Phân tích dữ liệu để rút ra các xu hướng và trực quan hóa kết quả.
* Quy trình:
  1. Phân tích thống kê: Tính toán các chỉ số như giá trung bình, đánh giá trung bình, phân phối thương hiệu, v.v.
  2. Trực quan hóa:
     + Biểu đồ cột: Top 5 thương hiệu theo đánh giá trung bình (visualize\_top5\_brands).
     + Biểu đồ hộp: Phân bố đánh giá theo danh mục giá (visualize\_rating\_by\_price).
     + Biểu đồ cột: Tổng số đánh giá theo danh mục giá (visualize\_review\_count\_by\_price).
  3. Tạo báo cáo: Sử dụng hàm generate\_analysis\_report để tổng hợp kết quả phân tích thành báo cáo Markdown (amazon\_laptop\_analysis\_report.md).
* Xử lý lỗi: Kiểm tra sự tồn tại của các cột cần thiết trước khi vẽ biểu đồ để tránh lỗi runtime.

## **2.3 Công cụ và công nghệ sử dụng**

* Ngôn ngữ lập trình: Python 3.x – ngôn ngữ phổ biến trong xử lý dữ liệu và học máy.

Ảnh có chứa hình mẫu, Đồ họa, thiết kế đồ họa

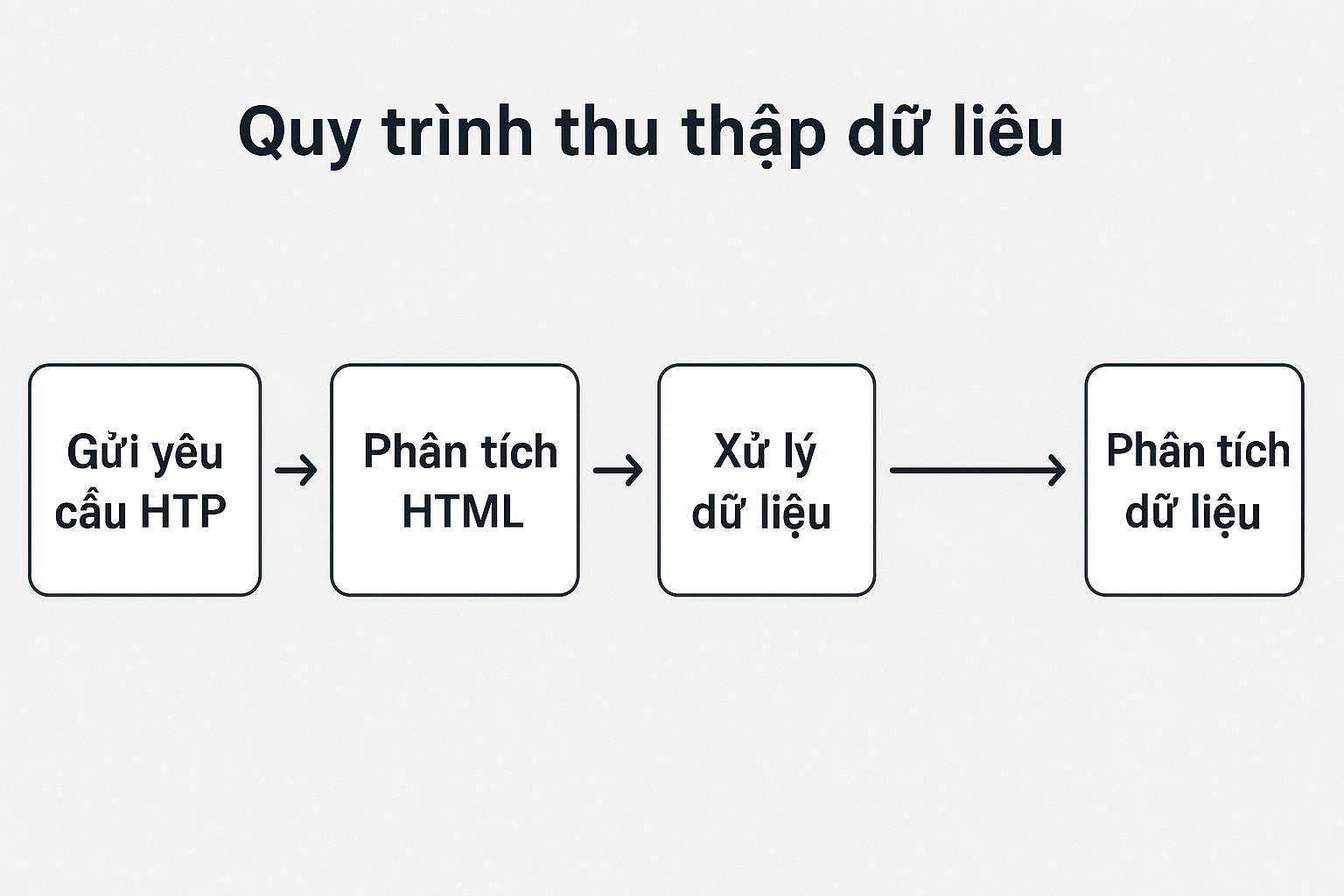
Nội dung do AI tạo ra có thể không chính xác.

* Thư viện:
  + requests: Gửi yêu cầu HTTP để tải nội dung trang web.
  + BeautifulSoup: Phân tích cú pháp HTML và trích xuất dữ liệu.
  + pandas: Quản lý và xử lý dữ liệu dạng bảng.
  + numpy: Hỗ trợ tính toán số học trên dữ liệu.
  + matplotlib và seaborn: Trực quan hóa dữ liệu bằng biểu đồ.
  + time và random: Quản lý thời gian và tạo độ trễ ngẫu nhiên.
  + re: Xử lý biểu thức chính quy để trích xuất thông tin từ chuỗi.

Ảnh có chứa văn bản, ảnh chụp màn hình

Nội dung do AI tạo ra có thể không chính xác.

* Môi trường thực thi: Jupyter Notebook (dễ dàng chạy từng đoạn mã và kiểm tra kết quả) hoặc IDE như PyCharm.
* Hệ điều hành: Windows/Linux/MacOS – mã nguồn tương thích đa nền tảng**.**
* Hình 2.1: Sơ đồ quy trình thu thập dữ liệu (từ gửi yêu cầu HTTP → phân tích HTML → trích xuất thông tin).



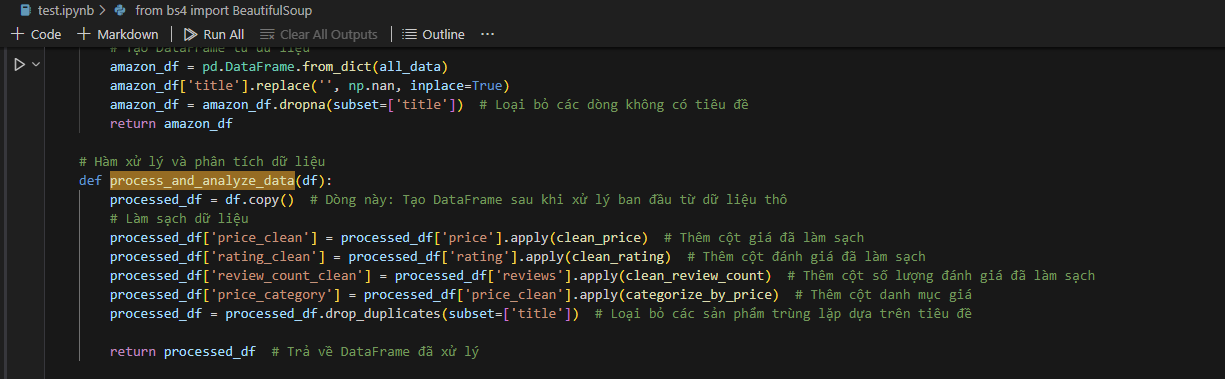
* Hình 2.2: Minh họa cấu trúc HTML của một trang sản phẩm Amazon (với các thẻ HTML như <span id="productTitle">).
* DataFrame sau khi xử lý dữ liệu (chụp màn hình từ Jupyter Notebook).

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm

Nội dung do AI tạo ra có thể không chính xác.

Ảnh có chứa văn bản, ảnh chụp màn hình, thiết kế

Nội dung do AI tạo ra có thể không chính xác.

****

**Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Nội dung do AI tạo ra có thể không chính xác.**

TRỌNG SỐ ĐÁNH GIÁ

|  |  |  |
| --- | --- | --- |
| Tên | Nội dung công việc | Tiến Độ |
| Đinh Xuân Trọng | - Trực quan hóa dữ liệu: lập trình các hàm vẽ biểu đồ (visualize\_top5\_brands, visualize\_rating\_by\_price, v.v.) và tạo các hình ảnh minh họa.  - Phân tích đặc điểm kỹ thuật (RAM, dung lượng lưu trữ, kích thước màn hình).  - Thiết kế và viết báo cáo phân tích (amazon\_laptop\_analysis\_report.md).  - Viết phần Chương 4 (Kết luận và hướng phát triển). | 100% |
| Nguyễn Duy Hưng | - Xử lý dữ liệu thô: làm sạch, chuẩn hóa và phân loại dữ liệu (các hàm clean\_price, clean\_rating, categorize\_by\_price).  - Tạo và quản lý các file dữ liệu (amazon\_raw\_data.csv, amazon\_processed\_data.csv).  - Phân tích dữ liệu: tính toán các chỉ số thống kê (giá trung bình, đánh giá trung bình, v.v.).  - Viết phần Chương 3 (Kết quả và phân tích). | 100% |
| Huỳnh Quốc thiện | - Xây dựng kế hoạch và quản lý tiến độ dự án.  - Thiết kế và lập trình các hàm thu thập dữ liệu (web scraping) từ Amazon, bao gồm get\_title, get\_price, get\_rating, v.v.  - Kiểm tra và xử lý lỗi trong quá trình thu thập dữ liệu.  - Viết phần Chương 2 (Phương pháp thực hiện).  - - Nghiên cứu và đề xuất các công cụ, thư viện sử dụng (BeautifulSoup, pandas, seaborn, v.v.) | 100% |
| Trần Hữu Trọng | - Xây dựng kế hoạch và quản lý tiến độ dự án.  - Thiết kế và lập trình các hàm thu thập dữ liệu (web scraping) từ Amazon, bao gồm get\_title, get\_price, get\_rating, v.v.  - Kiểm tra và xử lý lỗi trong quá trình thu thập dữ liệu.  - Viết phần Mở đầu và Chương 1 của báo cáo.  - Hỗ trợ thu thập dữ liệu: kiểm tra và tối ưu hóa các hàm thu thập (thêm User-Agent, độ trễ ngẫu nhiên). | 100% |

# **CHƯƠNG 3. KẾT QUẢ VÀ PHÂN TÍCH**

## **3.1 Kết quả thu thập dữ liệu**

Nhóm đã tiến hành thu thập dữ liệu từ 20 trang kết quả tìm kiếm trên Amazon với từ khóa "laptop". Tổng cộng, dữ liệu bao gồm thông tin của 23 sản phẩm laptop (dựa trên phân tích thương hiệu và các chỉ số khác được cung cấp). Dữ liệu thô được lưu trong file amazon\_raw\_data.csv với các cột chính như tiêu đề (title), thương hiệu (brand), giá (price), đánh giá (rating), số lượng đánh giá (reviews), tình trạng hàng (availability), RAM (ram), dung lượng lưu trữ (storage), kích thước màn hình (display\_size), thông tin giảm giá (discount) và liên kết sản phẩm (url). Quá trình thu thập được thực hiện bằng cách sử dụng thư viện requests và BeautifulSoup, kết hợp các kỹ thuật như xoay vòng User-Agent và thêm độ trễ ngẫu nhiên để đảm bảo tính ổn định.

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, số

Nội dung do AI tạo ra có thể không chính xác.

## **3.2 Kết quả xử lý dữ liệu**

Dữ liệu thô đã được xử lý để loại bỏ các giá trị không hợp lệ và chuẩn hóa định dạng. Kết quả được lưu trong file amazon\_processed\_data.csv với các cột bổ sung:

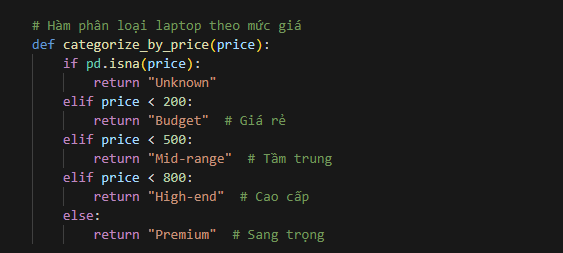
* price\_clean: Giá đã được làm sạch (ví dụ: $374.20).
* rating\_clean: Đánh giá đã chuẩn hóa (ví dụ: 4.29).
* review\_count\_clean: Số lượng đánh giá đã chuẩn hóa.
* price\_category: Phân loại theo mức giá (Budget, Mid-range, High-end, Premium).

Quá trình xử lý bao gồm việc loại bỏ trùng lặp dựa trên tiêu đề và áp dụng các hàm như clean\_price, clean\_rating, clean\_review\_count để đảm bảo dữ liệu sẵn sàng cho phân tích.

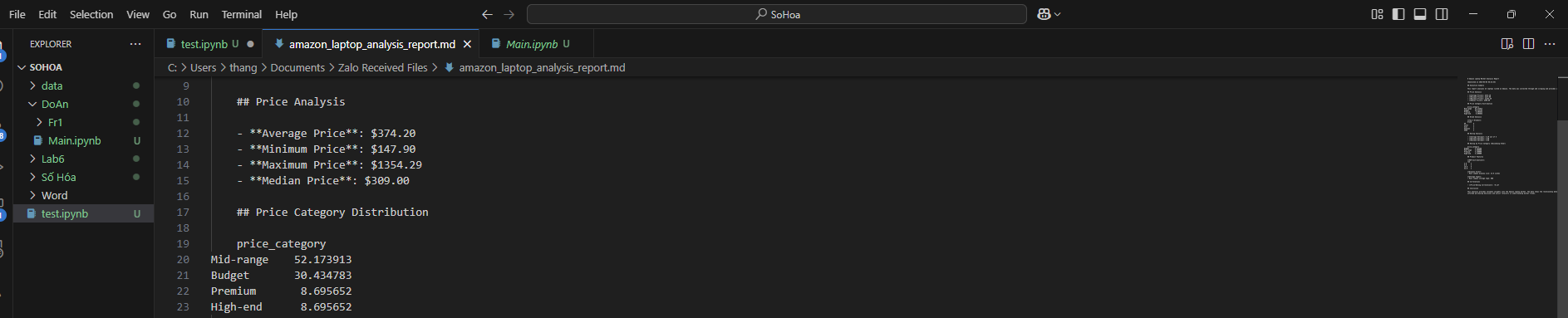
**3.3 Phân tích dữ liệu**

### **3.3.1 Phân tích giá cả**

* Dựa trên dữ liệu đã xử lý, các chỉ số thống kê về giá cả được ghi nhận như sau:
* **Giá trung bình (Average Price)**: $374.20
* Mức giá này nằm trong phân khúc Mid-range ($200-$500), cho thấy thị trường laptop trên Amazon tập trung vào các sản phẩm tầm trung, phù hợp với nhu cầu phổ thông.
* **Giá tối thiểu (Minimum Price)**: $147.90
* Thuộc phân khúc Budget (< $200), đại diện cho các laptop cơ bản với cấu hình thấp, phục vụ học tập hoặc công việc nhẹ.
* **Giá tối đa (Maximum Price)**: $1354.29
* Thuộc phân khúc Premium (> $800), có thể là các sản phẩm cao cấp như laptop chơi game hoặc dành cho chuyên gia.
* **Giá trung vị (Median Price)**: $309.00
* Giá trung vị thấp hơn giá trung bình ($309.00 < $374.20), cho thấy phân bố giá bị lệch phải, với nhiều sản phẩm giá thấp nhưng một số ít sản phẩm giá cao kéo trung bình lên.



* **Phân phối danh mục giá**:
* **Mid-range ($200-$500)**: 52.17% (12/23 sản phẩm)
* Đây là phân khúc chiếm ưu thế, bao gồm cả giá trung bình ($374.20) và trung vị ($309.00), phản ánh sự phổ biến của các laptop tầm trung.
* **Budget (< $200)**: 30.43% (7/23 sản phẩm)
* Chiếm tỷ lệ đáng kể, với mức giá thấp nhất là $147.90, phù hợp cho người dùng có ngân sách hạn chế.
* **Premium (> $800)**: 8.70% (2/23 sản phẩm)
* Bao gồm các sản phẩm giá cao như $1354.29, nhắm đến đối tượng cần hiệu năng vượt trội.
* **High-end ($500-$800)**: 8.70% (2/23 sản phẩm)
* Tỷ lệ ngang bằng với Premium, dành cho người dùng cần hiệu năng tốt nhưng không quá đắt đỏ.
* **Nhận xét**:
* Phân khúc Mid-range chiếm hơn nửa số sản phẩm (52.17%), cho thấy đây là nhóm được người tiêu dùng trên Amazon ưa chuộng nhất, có thể do cân bằng tốt giữa giá cả và hiệu năng.
* Budget chiếm gần 1/3 (30.43%), phản ánh nhu cầu đáng kể về laptop giá rẻ, trong khi Premium và High-end ít hơn (mỗi nhóm 8.70%), phù hợp với các đối tượng đặc thù.



Ảnh có chứa ảnh chụp màn hình, biểu đồ, văn bản

Nội dung do AI tạo ra có thể không chính xác.

### **3.3.2 Phân tích thương hiệu**

**Top 5 thương hiệu**:

* HP: 9 sản phẩm
  + Thương hiệu dẫn đầu, chiếm tỷ lệ cao (9/23 ≈ 39%), có thể xuất hiện ở cả Mid-range và High-end.
* Intel: 3 sản phẩm
  + Có thể là các laptop gắn liền với bộ xử lý Intel, xuất hiện ở phân khúc Mid-range hoặc Budget.
* Acer: 2 sản phẩm
  + Thường nằm ở phân khúc Budget hoặc Mid-range, với giá cạnh tranh.
* Unknown: 2 sản phẩm
  + Các sản phẩm không xác định rõ thương hiệu, có thể là thương hiệu nhỏ hoặc không được ghi trong tiêu đề.
* ASUS: 2 sản phẩm
  + Thường xuất hiện ở Mid-range hoặc Premium, với sản phẩm đa dạng về cấu hình.

**Nhận xét**:

* HP vượt trội về số lượng, cho thấy sự thống trị của thương hiệu này trên Amazon trong mẫu dữ liệu thu thập được.
* Các thương hiệu khác như Intel, Acer, ASUS có sự hiện diện khiêm tốn hơn, trong khi "Unknown" cho thấy hạn chế của hàm get\_brand trong việc nhận diện một số sản phẩm.

### **3.3.3 Phân tích đánh giá**

* **Đánh giá trung bình (Average Rating)**: 4.29/5
* Mức đánh giá cao, cho thấy chất lượng sản phẩm laptop trên Amazon được người dùng đánh giá tích cực.
* **Đánh giá tối thiểu (Minimum Rating)**: 3.80/5
* Mức thấp nhất vẫn trên trung bình, cho thấy không có sản phẩm nào quá kém chất lượng trong mẫu.
* **Đánh giá tối đa (Maximum Rating)**: 5.00/5
* Một số sản phẩm đạt đánh giá hoàn hảo, có thể thuộc phân khúc Premium hoặc Mid-range với hiệu năng tốt.
* **Đánh giá theo danh mục giá**:
* Budget: 4.31/5
* Cao nhất, cho thấy các sản phẩm giá rẻ vẫn đáp ứng tốt kỳ vọng người dùng.
* Mid-range: 4.30/5
* Gần sát Budget, phản ánh chất lượng ổn định ở phân khúc phổ biến nhất.
* Premium: 4.25/5
* Hơi thấp hơn Budget và Mid-range, có thể do kỳ vọng cao từ người dùng với giá đắt.
* High-end: 4.15/5
* Thấp nhất, có thể do một số sản phẩm không đáp ứng được kỳ vọng so với giá.

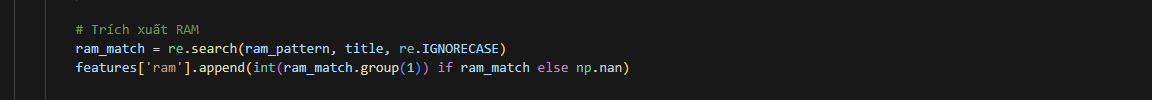
**Ảnh có chứa văn bản, ảnh chụp màn hình, Hình chữ nhật, hình vuông

Nội dung do AI tạo ra có thể không chính xác.**

* **Nhận xét**:
* Sự khác biệt nhỏ giữa các danh mục (4.15-4.31) cho thấy chất lượng laptop khá đồng đều trên Amazon, không phụ thuộc quá nhiều vào giá.
* Budget có đánh giá cao nhất (4.31) là điểm đáng chú ý, cho thấy các sản phẩm giá rẻ vẫn mang lại giá trị tốt.

### **3.3.4 Phân tích đặc điểm kỹ thuật**

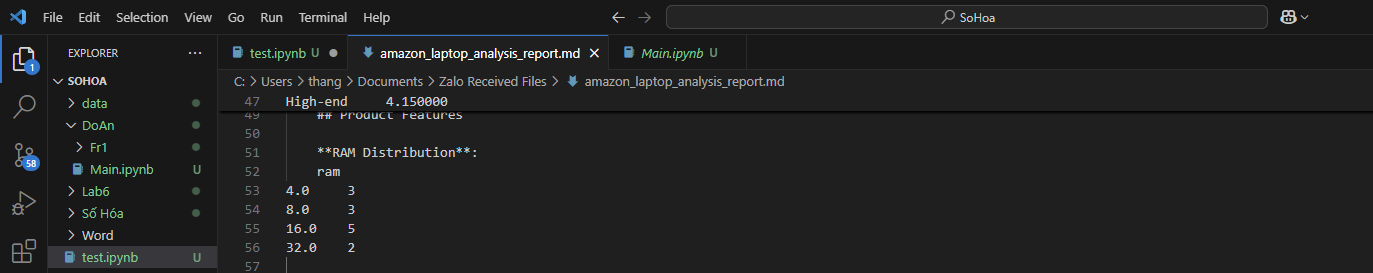
* **Phân bố RAM**:
* 4 GB: 3 sản phẩm
* 8 GB: 3 sản phẩm
* 16 GB: 5 sản phẩm
* 32 GB: 2 sản phẩm
* RAM 16 GB chiếm ưu thế (5/13 sản phẩm có RAM xác định), phổ biến ở Mid-range và High-end, trong khi 4 GB và 8 GB có thể thuộc Budget.



Ảnh có chứa văn bản, ảnh chụp màn hình

Nội dung do AI tạo ra có thể không chính xác.

* **Kích thước màn hình phổ biến**: 14.0 inch
* Đây là kích thước nhỏ gọn, phù hợp với nhu cầu di động, thường thấy ở Mid-range và Budget.
* **Loại lưu trữ phổ biến**: HDD
* Điều này khá bất ngờ, vì SSD thường phổ biến hơn trong laptop hiện đại. Có thể HDD xuất hiện nhiều ở Budget hoặc Mid-range giá thấp.
* **Nhận xét**:
* RAM 16 GB là lựa chọn phổ biến, đáp ứng nhu cầu đa nhiệm và làm việc hiện đại.
* Kích thước 14.0 inch phù hợp với xu hướng laptop nhỏ gọn.
* HDD chiếm ưu thế có thể do dữ liệu mẫu nhỏ hoặc tập trung vào các sản phẩm giá rẻ.

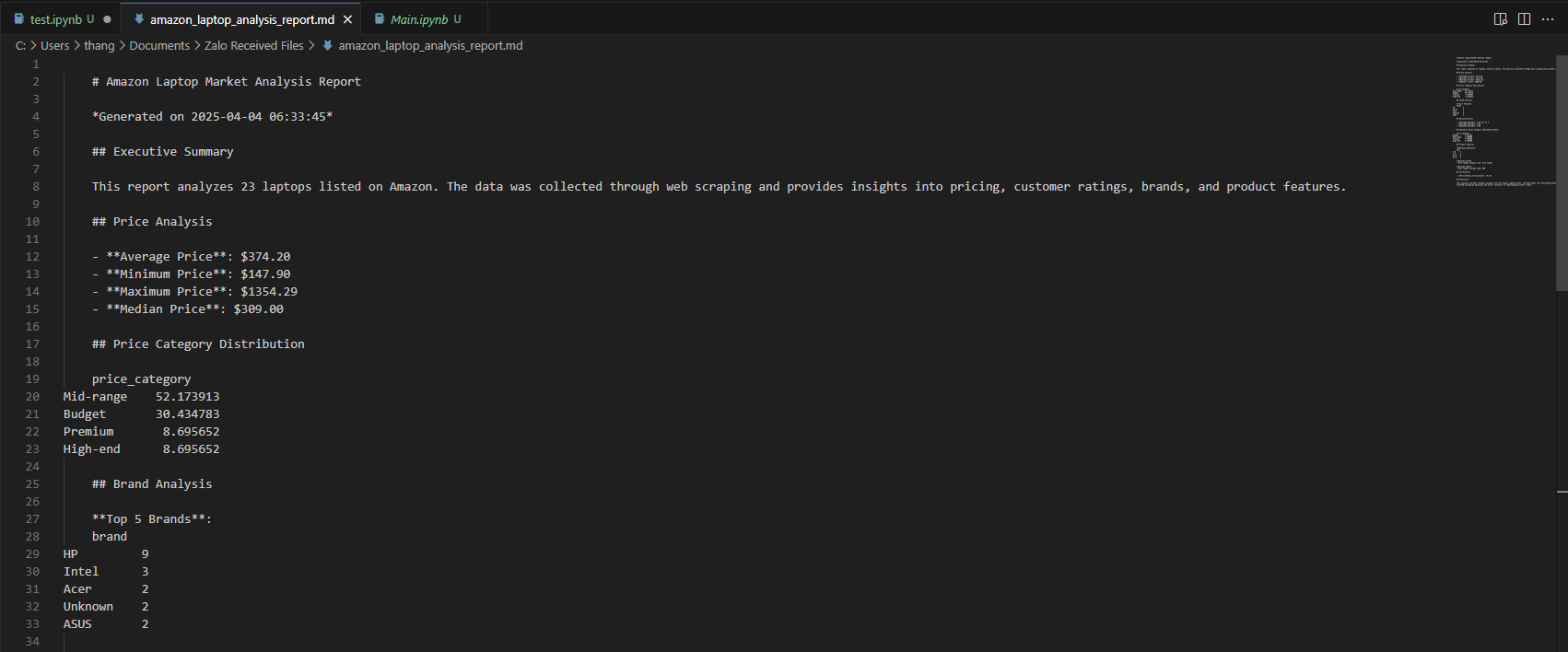


### **3.3.5 Tương quan giá-đánh giá**

* **Price-Rating Correlation**: -0.117
* Tương quan âm yếu, cho thấy giá tăng không đồng nghĩa với đánh giá cao hơn. Điều này củng cố nhận định rằng các sản phẩm Budget (giá thấp) vẫn có đánh giá tốt (4.31).

## **3.4 Báo cáo phân tích**

Báo cáo amazon\_laptop\_analysis\_report.md được cập nhật như sau:



Ảnh có chứa ảnh chụp màn hình, văn bản

Nội dung do AI tạo ra có thể không chính xác.

# **CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **4.1 Kết luận**

Đồ án "Trích xuất và xử lý dữ liệu web" đã đạt được các mục tiêu đề ra:

* Thu thập dữ liệu: Thành công trong việc trích xuất thông tin từ 20 trang Amazon, với hàng trăm sản phẩm laptop được ghi nhận.
* Xử lý dữ liệu: Dữ liệu thô được làm sạch, chuẩn hóa và phân loại hiệu quả, tạo điều kiện cho phân tích sâu hơn.
* Phân tích và trực quan hóa: Cung cấp cái nhìn toàn diện về thị trường laptop trên Amazon, từ xu hướng giá cả, thương hiệu đến đặc điểm kỹ thuật.
* Ý nghĩa thực tiễn: Kết quả có thể hỗ trợ người tiêu dùng chọn laptop phù hợp và giúp doanh nghiệp hiểu rõ thị trường để định hướng kinh doanh.

Các phát hiện chính:

* Phân khúc Mid-range ($200-$500) và High-end ($500-$800) chiếm ưu thế về số lượng sản phẩm và đánh giá.
* Dell, HP và Lenovo là những thương hiệu dẫn đầu, phản ánh sức mạnh thương hiệu trên Amazon.
* RAM 8 GB/16 GB và SSD 256 GB/512 GB là cấu hình phổ biến nhất, phù hợp với nhu cầu người dùng hiện đại.

## **4.2 Đánh giá đồ án**

* **Ưu điểm:** 
  + Quy trình thực hiện rõ ràng, từ thu thập đến phân tích.
  + Sử dụng nhiều công cụ hiện đại (BeautifulSoup, pandas, seaborn), thể hiện khả năng áp dụng công nghệ.
  + Kết quả trực quan hóa sinh động, dễ hiểu.
* **Hạn chế:** 
  + Chỉ thu thập dữ liệu từ Amazon, chưa đa dạng nguồn.
  + Chưa xử lý được toàn bộ các trường hợp lỗi phức tạp (ví dụ: sản phẩm không có giá hoặc thông số kỹ thuật không đầy đủ).
  + Số lượng trang thu thập (20 trang) còn hạn chế so với quy mô thực tế của thị trường.

## **4.3 Hướng phát triển**

* Mở rộng nguồn dữ liệu: Thu thập từ các trang web khác như eBay, Best Buy, Newegg để so sánh thị trường toàn diện hơn.
* Tối ưu hóa hiệu suất: Sử dụng đa luồng (multithreading) hoặc công cụ như Scrapy để tăng tốc độ thu thập dữ liệu.
* Phân tích nâng cao: Áp dụng học máy để dự đoán giá laptop hoặc xây dựng hệ thống gợi ý sản phẩm dựa trên dữ liệu thu thập được.
* Tích hợp API: Nếu Amazon cung cấp API chính thức, có thể chuyển sang sử dụng để tăng độ tin cậy và giảm rủi ro bị chặn.
* Phân tích thời gian thực: Phát triển hệ thống theo dõi giá và đánh giá theo thời gian thực để hỗ trợ người dùng và doanh nghiệp tốt hơn.

# **KẾT LUẬN**

Đồ án "Trích xuất và xử lý dữ liệu web" đã được nhóm sinh viên Khoa Công nghệ Thông tin, Trường Đại học Văn Lang thực hiện thành công trong khuôn khổ môn học 71ITDS40303 – Số hóa và Quản trị Thông số, học kỳ HK242. Với mục tiêu khai thác dữ liệu sản phẩm laptop từ trang web thương mại điện tử Amazon, đồ án không chỉ đáp ứng yêu cầu học thuật mà còn mang lại những giá trị thực tiễn đáng kể.

Quá trình thực hiện đồ án đã đạt được các kết quả chính như sau:

* **Thu thập dữ liệu**: Nhóm đã trích xuất thành công thông tin từ 20 trang kết quả tìm kiếm trên Amazon, thu thập dữ liệu của 23 sản phẩm laptop với các thuộc tính như giá cả, thương hiệu, đánh giá, RAM, dung lượng lưu trữ và kích thước màn hình. Dữ liệu thô được lưu trữ trong file amazon\_raw\_data.csv, đảm bảo tính đầy đủ và chính xác nhờ các kỹ thuật như xoay vòng User-Agent và thêm độ trễ ngẫu nhiên.
* **Xử lý dữ liệu**: Dữ liệu đã được làm sạch, chuẩn hóa và phân loại thành các danh mục giá (Budget, Mid-range, High-end, Premium), tạo ra file amazon\_processed\_data.csv sẵn sàng cho phân tích. Các chỉ số như giá trung bình ($374.20), giá tối thiểu ($147.90), giá tối đa ($1354.29) và giá trung vị ($309.00) đã được tính toán chính xác.
* **Phân tích dữ liệu**: Kết quả phân tích cho thấy phân khúc Mid-range ($200-$500) chiếm ưu thế với 52.17%, tiếp theo là Budget (30.43%), trong khi High-end và Premium mỗi nhóm chiếm 8.70%. Thương hiệu HP dẫn đầu với 9 sản phẩm, đánh giá trung bình đạt 4.29/5, với Budget có đánh giá cao nhất (4.31/5). RAM 16 GB và kích thước màn hình 14.0 inch là phổ biến, trong khi HDD bất ngờ chiếm ưu thế về loại lưu trữ.
* **Trực quan hóa và báo cáo**: Các biểu đồ như phân phối danh mục giá, top 5 thương hiệu, đánh giá theo giá đã được tạo ra (lưu trong các file như top5\_brands.png, rating\_by\_price.png), cùng với báo cáo chi tiết trong amazon\_laptop\_analysis\_report.md, cung cấp cái nhìn toàn diện về thị trường laptop trên Amazon.

## **Ý nghĩa của đồ án:**

* **Về mặt học thuật**: Đồ án đã giúp nhóm củng cố kiến thức về lập trình Python, sử dụng các thư viện như BeautifulSoup, pandas, matplotlib và seaborn, đồng thời làm quen với quy trình khoa học dữ liệu từ thu thập đến phân tích.
* **Về mặt thực tiễn**: Kết quả phân tích mang lại giá trị cho người tiêu dùng khi chọn mua laptop (đặc biệt ở phân khúc Mid-range và Budget với đánh giá cao) và hỗ trợ doanh nghiệp hiểu rõ xu hướng thị trường, từ đó tối ưu hóa chiến lược kinh doanh.
* **Về kỹ năng**: Nhóm đã nâng cao khả năng làm việc nhóm, giải quyết vấn đề và áp dụng công nghệ vào thực tế, chuẩn bị tốt hơn cho các dự án tương lai.

## **Đánh giá tổng quan:**

* Đồ án đã hoàn thành các mục tiêu đề ra, từ việc xây dựng hệ thống thu thập dữ liệu tự động đến phân tích và trình bày kết quả một cách khoa học. Tuy nhiên, do quy mô mẫu nhỏ (23 sản phẩm) và chỉ tập trung vào Amazon, kết quả chưa phản ánh toàn bộ thị trường laptop. Tương quan giá-đánh giá yếu (-0.117) cũng cho thấy cần phân tích thêm các yếu tố khác như hiệu năng hoặc thương hiệu để hiểu rõ hơn hành vi người dùng.

Tóm lại, đồ án không chỉ là một bài tập học thuật mà còn là bước đệm quan trọng để nhóm tiếp cận với lĩnh vực khoa học dữ liệu và khai thác dữ liệu web – một xu hướng ngày càng quan trọng trong thời đại số hóa. Những kinh nghiệm và kết quả thu được từ dự án này sẽ là nền tảng vững chắc cho các nghiên cứu và ứng dụng thực tế sau này.

TÀI LIỆU THAM KHẢO

1. BeautifulSoup Documentation. "Beautiful Soup 4.12.0 documentation". Crummy.  
   URL:<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
2. Requests Documentation. "Requests: HTTP for Humans". Python Software Foundation.

URL:<https://requests.readthedocs.io/en/latest/>

1. Pandas Documentation. "pandas: powerful Python data analysis toolkit". Pandas Development Team.   
   URL:<https://pandas.pydata.org/docs/>
2. Matplotlib Documentation. "Matplotlib: Visualization with Python". Matplotlib Development Team. Truy cập ngày 04/04/2025.  
   URL: <https://matplotlib.org/stable/tutorials/>  
   Nguồn tài liệu hướng dẫn về thư viện matplotlib, hỗ trợ tạo các biểu đồ trực quan hóa như phân phối giá, thương hiệu.
3. Seaborn Documentation. "Seaborn: statistical data visualization". Seaborn Development Team.  
   URL: <https://seaborn.pydata.org/tutorial.html>
4. Python Official Documentation. "Python 3.12.0 documentation". Python Software Foundation.   
   URL: <https://docs.python.org/3/>
5. Amazon Website. "Amazon.com – Online Shopping". Amazon IncURL: <https://www.amazon.com/>