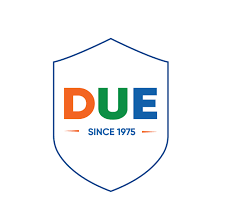
 **TRƯỜNG ĐẠI HỌC KINH TẾ** **ĐÀ NẴNG**

**KHOA THƯƠNG MẠI ĐIỆN TỬ**

**----------🙜 ✧ 🙞---------**

**BÁO CÁO TỐT NGHIỆP**

**Chuyên ngành Khoa học dữ liệu & Phân tích kinh doanh**

**XÂY DỰNG HỆ THỐNG ĐỀ XUẤT SÁCH**

Sinh viên thực hiện : Đoàn Huy Vũ

Lớp : 47K29.2

Đơn vị thực tập : Pandosima CO.LTD, Đà Nẵng

Cán bộ hướng dẫn : Lê Văn Khanh

Giảng viên hướng dẫn : TS. Phan Đình Vấn

***Đà Nẵng, ngày tháng năm 2025***

# **NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

# **LỜI CẢM ƠN**

Trong suốt thời gian học tập tại Đại học Kinh tế Đà Nẵng,cùng với sự tận tình giúp đỡ của thầy cô trong trường, em đã tích góp được cho mình những kiến thức những trải nghiệm. Với nền tảng kiến thức vững vàng, cũng như các kỹ năng cần thiết đối với tư duy phân tích, sẽ là hành trang vững chắc cho em thực hiện công việc em mong muốn

Đầu tiên, em xin gửi lời cảm ơn thầy Phan Đình Vấn, giáo viên phụ trách hướng dẫn trực tiép em trong suốt quá trình em thực hiện báo cáo tốt nghiệp.Sự tận tình khi chỉ ra những sai sót để có thể tối ưu cho đề tài, cũng như sát sao của thầy trong quá hướng dẫn đã giúp em hoàn thiện bài báo cáo của mình một cách chuyên nghiệp nhất

Em cũng xin gửi lời cảm ơn đến Pandosima Co.LTD Đà Nẵng, đã tạo cho em cơ hội để thực tập tại đơn vị. Được tiếp cận với một môi trường làm việc thực tế, cùng với sự chỉ bảo tận tình từ anh mentor Lê Văn Khanh giúp em có thêm những kiến thức hữu ích để hoàn thiện bài báo cáo

Mặc dù đã có sự nỗ lực từ bản thân, nhưng với quỹ thời gian hạn hẹp và chưa có nhiều kiến thức thực tế, em mong được nhận sự góp ý từ các thầy cô để có thể tự hoàn thiện bài báo cáo của mình. Em xin chân thành cảm ơn

# **LỜI CAM ĐOAN**

Em xin cam đoan đề tài “**Xây dựng hệ thống đề xuất sách”** được thực hiện bằng kiến thức của bản thân, thực hiện dưới sự giám sát và quản lý của giáo viên hướng dẫn Phan Đình Vấn và mentor Lê Văn Khanh, những số liệu thu thập cũng như kết quả của bài báo cáo hoàn toàn trung thực, không có việc sử dụng hay sao chép số liệu từ các bài khác. Nếu phát hiện có sự sao chép thì em xin chịu mọi trách nhiệm trước hội đồng

Đà Nẵng, ngày tháng năm 2025

Sinh viên thực hiện

Đoàn Huy Vũ

# **MỤC LỤC**

[**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP** i](#_Toc194509590)

[**LỜI CẢM ƠN** ii](#_Toc194509591)

[**LỜI CAM ĐOAN** iii](#_Toc194509592)

[**MỤC LỤC** iv](#_Toc194509593)

[**DANH MỤC HÌNH ẢNH** vii](#_Toc194509594)

[**DANH MỤC BẢNG BIỂU** viii](#_Toc194509595)

[**DANH MỤC CÁC TỪ VIẾT TẮT** ix](#_Toc194509596)

[**LỜI MỞ ĐẦU** 1](#_Toc194509597)

[**CHƯƠNG 1: BỐI CẢNH SỬ DỤNG HỆ THỐNG ĐỀ XUẤT TRONG CUỘC SỐNG HIỆN NAY VÀ GIỚI THIỆU ĐƠN VỊ THỰC TẬP** 4](#_Toc194509598)

[**1.1** **Bối cảnh sử dụng hệ thống đề xuất trong cuộc sống hiện nay** 4](#_Toc194509599)

[**1.2 Giới thiệu đơn vị thực tập** 5](#_Toc194509600)

[**1.2.1. Tầm nhìn và sứ mệnh** 5](#_Toc194509601)

[**1.2.2 Giá trị cốt lõi** 5](#_Toc194509602)

[**1.2.3** **Dịch vụ công ty** 6](#_Toc194509603)

[**1.3 Giới thiệu về ngành nghề thực tập** 7](#_Toc194509604)

[**1.3.1 Định Nghĩa về nhà khoa học dữ liệu( Data Scientist)** 7](#_Toc194509605)

[**1.3.2 Công việc tại Pandosima** 7](#_Toc194509606)

[**1.3.3 Các kỹ năng tối thiểu cần có để trở thành một Data Scientist** 9](#_Toc194509607)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 11](#_Toc194509608)

[**2.1. Hệ thống gợi ý** 11](#_Toc194509609)

[**2.1.1**  **Định nghĩa** 11](#_Toc194509610)

[**2.1.2 Sự Thay Đổi Của Quảng Cáo Và Vai Trò Của Hệ Thống Gợi Ý** 11](#_Toc194509611)

[**2.1.3 Nguyên lý hoạt động** 12](#_Toc194509612)

[**2.2. Các loại hệ thống gợi ý sản phẩm** 13](#_Toc194509613)

[**2.2.1. Hệ Thống Gợi Ý Dựa Trên Nội Dung (Content-Based Filtering)** 13](#_Toc194509614)

[**2.2.2. Hệ Thống Gợi Ý Dựa Trên Lọc Cộng Tác (Collaborative Filtering)** 14](#_Toc194509615)

[**2.2.3. Hệ Thống Gợi Ý Lai (Hybrid Recommendation Systems)** 16](#_Toc194509616)

[**CHƯƠNG 3: XÂY DỰNG HỆ THỐNG GỢI Ý SÁCH HỖ TRỢ BÁN HÀNG** 18](#_Toc194509617)

[**3.1. Thu thập và xử lý dữ liệu** 18](#_Toc194509618)

[**3.1.1. Nguồn dữ liệu** 18](#_Toc194509619)

[**3.1.2. Tiền xử lý dữ liệu** 18](#_Toc194509620)

[**3.1.3. Phân tích dữ liệu trước khi huấn luyện mô hình** 18](#_Toc194509621)

[**3.2. Huấn luyện mô hình AI** 18](#_Toc194509622)

[**3.2.1. Chọn phương pháp gợi ý phù hợp** 18](#_Toc194509623)

[**3.2.2. Cấu hình mô hình Content-Based Filtering** 18](#_Toc194509624)

[**3.3. Xây dựng API và tích hợp hệ thống** 18](#_Toc194509625)

[**CHƯƠNG 4: ĐÁNH GIÁ HIỆU SUẤT HỆ THỐNG** 19](#_Toc194509626)

[**4.1. Huấn luyện mô hình AI** 19](#_Toc194509627)

[**4.1.1. Tiêu chí đánh giá mô hình AI** 19](#_Toc194509628)

[**4.1.2. Thời gian phản hồi API** 19](#_Toc194509629)

[**4.2. Cải tiến và tối ưu hóa mô hình** 19](#_Toc194509630)

[**4.2.1. Dùng Reinforcement Learning để học từ phản hồi người dùng** 19](#_Toc194509631)

[**4.2.2. Tối ưu siêu tham số** 19](#_Toc194509632)

[**Tổng kết kết quả đạt được** 19](#_Toc194509633)

[**Hạn chế của hệ thống** 20](#_Toc194509634)

[**Hướng phát triển trong tương lai** 20](#_Toc194509635)

[**Tài liệu tham khảo** 20](#_Toc194509636)

# **DANH MỤC HÌNH ẢNH**

# **DANH MỤC BẢNG BIỂU**

# **DANH MỤC CÁC TỪ VIẾT TẮT**

# **LỜI MỞ ĐẦU**

1. **Lý do chọn đề tài:**

Lịch sử con người trải qua 4 thời kỳ trong đó 1.0 là phát minh ra lửa nơi đánh dấu bước ngoặt đầu tiên của con người khi mà chuyển từ việc săn bắt hái lượm sang chế tác các công cụ, thì thời kỳ 2.0 là động cơ hơi nước, hình thành nên một kỷ nguyên công nghiệp.Bùng nổ sau đó là thời đại 3.0, thời đại của internet và công nghệ số, thay đổi toàn bộ cục diện cũng như các mọi người giao tiếp và làm việc với nhau. Ngày nay thời đại 4.0 chính là thời đại của trí tuệ nhân tạo, định hình lại các công việc từ mọi lĩnh vực khác nhau. Càng khai thác và tìm tòi sâu về công nghệ, con người lại có những phát hiện thay thế đi những công nghệ,phát minh cũ. Tuy nhiên, có một đã song hành suốt chiều dài lịch sử phát triển công nghệ của loài người đó chính là dữ liệu. Không thể phủ nhận tầm quan trọng của dữ liệu khi mà trong suốt thời gian đó con người luôn cố gắng tìm cách lưu trữ và quản lý dữ liệu một cách hiệu quả nhất vì nó là cốt lõi để có được nền tảng trí tuệ nhân tạo vững chắc như ngày hôm nay từ đó tạo ra những sản phẩm AI như Chat GPT, Gemini, …Bên cạnh việc phát triển công nghệ về trí tuệ nhân tạo thì ở góc nhìn khác, các chủ doanh nghiệp luôn tìm cách để ứng dụng AI vào trong việc quản lý doanh nghiệp của minh.

Hiểu được vấn đề đó, Pandosima cung cấp một giải pháp kinh doanh thông minh ứng dụng AI vào quản lý doanh nghiệp, cung cấp cho những người mới hay đã làm lâu năm một hệ thống đơn giản để quản lý doanh nghiệp một cách dễ dàng và nhanh chóng, giúp họ kiểm soát được dữ liệu trong và ngoài công ty.

Đối với các doanh nghiệp bán hàng, thì việc họ phải quản lý các website, các dữ liệu về bài viết, các thông tin về nhân sự cũng như nhân viên bán hàng là vô cùng quan trọng nhằm mục đích đánh giá tình hình chung của doanh nghiệp.Với chi phí bỏ ra ít hơn nhiều so với thuê nhân viên quản lý, thì việc này cho phép người chủ doanh nghiệp tập trung vào chuyên môn của mình như là tìm kiếm nguồn hàng hay tìm các đối tác phát triển.

Với mục tiêu quan trọng nhất là giúp các người chủ doanh nghiệp tối ưu và gây ấn tượng với khách hàng trong với phương pháp quản lý chuyên nghiệp. Để làm được điều đó thì những sản phẩm, dịch vụ của doanh nghiệp phải đánh trúng nhu cầu của khách hàng. Hiểu được điều đó, xét thấy ứng dụng cần phải được cải tiến thêm ở một số tính năng đã thúc đẩy tôi thực hiện nghiên cứu với đề tài “Xây dựng hệ thống đề xuất sách”. Hệ thống sẽ giúp doanh nghiệp lựa chọn và đề xuất các cuốn sách phù hợp nhẩt với nhu cầu mua của khách hàng từ đó tăng doanh thu và hiệu suất bán hàng

**2. Mục tiêu nghiên cứu đề tài:**

Xây dựng hệ thống đề xuất sách khi khách hàng tìm kiếm, với hiển thị những cuốn sách liên quan để khách hàng có thể dễ dàng hơn trong việc ra quyết định. Đồng thời hình thành những từ khóa giúp doanh nghiệp hiểu hơn về xu hướng mua hàng của khách hàng

**3. Phương pháp nghiên cứu:**

**Phương pháp nghiên cứu lý thuyết:** Nghiên cứu về các tài liệu về các thuật toán gợi ý như Collaborative Filtering, Content-Based Filtering, Hybrid Model để có thể ứng dụng vào việc xây dựng các thuật toán đề xuất từ đó đưa ra những phán đoán chính xác nhất cho khách hàng về sản phẩm cần mua

**Phương pháp nghiên cứu thực nghiệm:** Dữ liệu được thu thập từ trang thương mại điện tử lớn Tiki, dữ liệu bao gồm các thông tin chi tiết về sách cũng như các doanh số bán và các đánh giá từ người mua. Dữ liệu sẽ được xử lý bằng NLP để tối ưu nhất trong việc đề xuất chúng cho khách hàng

**4. Đối tượng và phạm vi nghiên cứu:**

Đối tượng nghiên cứu đề tài: đối tượng nghiên cứu của đề tài là dữ liệu về sách được thu thập trang thương mại điện tử Tiki ,bao gồm các thông tin cũng như dữ liệu sản phẩm, và dữ liệu về bình luận đánh giá của khách hàng. Từ đó đưa ra một hệ thống đề xuất tối ưu, giúp khách hàng ra quyết định dễ hơn trong việc lựa chọn sản phẩm, đồng thời hỗ trợ doanh nghiệp phân loại phân khúc khách hàng

Phạm vi nghiên cứu của đề tài: được thực hiện trên dữ liệu về khách hàng mua hàng mua sách trên Tiki ở Việt Nam và ứng dụng AI để đề xuất sản phẩm phù hợp

**5. Nội dung của đề tài:**

Sau khi hiểu sơ lược về đề tài ở phần mở đầu, đề tài sẽ tập trung vào nghiên cứu nội dung trong 4 chương kế tiếp:

CHƯƠNG 1: BỐI CẢNH SỬ DỤNG HỆ THỐNG ĐỀ XUẤT TRONG CUỘC SỐNG HIỆN NAY VÀ GIỚI THIỆU ĐƠN VỊ THỰC TẬP

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

CHƯƠNG 3: XÂY DỰNG HỆ THỐNG GỢI Ý SÁCH

CHƯƠNG 4: ĐÁNH GIÁ HIỆU SUẤT HỆ THỐNG

KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

# **CHƯƠNG 1: BỐI CẢNH SỬ DỤNG HỆ THỐNG ĐỀ XUẤT TRONG CUỘC SỐNG HIỆN NAY VÀ GIỚI THIỆU ĐƠN VỊ THỰC TẬP**

## **Bối cảnh sử dụng hệ thống đề xuất trong cuộc sống hiện nay**

Trong bối cảnh phát triển nhanh chóng của công nghệ thông tin và dữ liệu lớn, hệ thống đề xuất ngày càng trở thành một công cụ không thể thiếu trong nhiều lĩnh vực của đời sống. Hệ thống này giúp cung cấp các gợi ý chính xác và phù hợp, góp phần cải thiện trải nghiệm người dùng trong môi trường số hóa. Với khả năng phân tích và dự đoán các sở thích cá nhân dựa trên hành vi, thói quen, và dữ liệu lịch sử, hệ thống đề xuất hiện diện mạnh mẽ trong các nền tảng thương mại điện tử, truyền thông xã hội, giải trí trực tuyến và giáo dục. Trong lĩnh vực bán lẻ, AI đóng vai trò quan trọng trong việc cá nhân hóa trải nghiệm khách hàng, dự đoán nhu cầu thị trường và tự động hóa dịch vụ. Các nền tảng thương mại điện tử sử dụng AI để gợi ý sản phẩm phù hợp với sở thích người tiêu dùng, từ đó tăng tỷ lệ chuyển đổi và doanh thu. Ngoài ra, AI cũng giúp doanh nghiệp quản lý hàng tồn kho hiệu quả, giảm thiểu tình trạng thiếu hụt hoặc dư thừa sản phẩm.

Trong xu hướng hiện nay, việc sử dụng hệ thống đề xuất không chỉ giúp tối ưu hóa quá trình tìm kiếm thông tin, mà còn nâng cao hiệu quả của các chiến lược marketing, từ đó tạo ra những trải nghiệm cá nhân hóa sâu sắc cho người dùng. Các công nghệ tiên tiến như học máy (machine learning), trí tuệ nhân tạo (AI), và phân tích dữ liệu lớn (big data) đóng vai trò quan trọng trong việc phát triển các hệ thống này, cho phép các nền tảng trực tuyến không chỉ hiểu và dự đoán sở thích của người dùng, mà còn thích ứng với những thay đổi trong hành vi và xu hướng tiêu dùng.

Tuy nhiên, bên cạnh những lợi ích rõ rệt, việc triển khai hệ thống đề xuất cũng đặt ra một số thách thức, đặc biệt liên quan đến vấn đề bảo mật và quyền riêng tư của người dùng. Việc thu thập và phân tích một lượng lớn dữ liệu cá nhân có thể tạo ra những lo ngại về việc xâm phạm quyền riêng tư và thiếu minh bạch trong việc sử dụng thông tin người dùng. Do đó, việc phát triển các hệ thống đề xuất không chỉ cần chú trọng vào hiệu quả và tính chính xác mà còn phải đảm bảo sự công bằng, bảo mật và tôn trọng quyền lợi của người tiêu dùng.

## **1.2 Giới thiệu đơn vị thực tập**

### **1.2.1. Tầm nhìn và sứ mệnh**

**Tầm nhìn:** Pandosima cung cấp các phương thức hỗ trợ trong việc quản lý doanh nghiệp. Pandosima hiểu rằng mỗi doanh nghiệp đều có những thách thức riêng trong các lĩnh vực như marketing, bán hàng, nhân sự, tài chính, chuỗi cung ứng, chăm sóc khách hàng hay quản lý hồ sơ nghiệp vụ. Vì vậy, Pandosima không chỉ cung cấp các công cụ hỗ trợ từng khía cạnh riêng lẻ mà còn xây dựng các giải pháp quản trị tích hợp, đáp ứng nhu cầu tổng thể của doanh nghiệp..Việc tích hợp AI và tự động hóa các mô hình học máy giúp doanh nghiệp có thể dự đoán với độ chính xác cao, giúp giảm thiểu tối đa nhân lực cần thiết để vận hành doanh nghiệp

**Sứ mệnh:** Với quyết tâm trở thành nhà cung cấp nền tảng quản lý doanh nghiệp hàng đầu, không chỉ trong nước mà còn trên thương trường quốc tế, Pandosima liên tục cải tiến các công nghệ giải pháp hỗ trợ doanh nghiệp tôi ưu hóa quy trình vận hành, cạnh tranh trực tiếp với đối thủ. Tin rằng trong bối cảnh chuyển biến công nghệ như vũ bão hiện nay, việc áp dụng các công nghệ hiện đại như trí tuệ nhân tạo, tự động hóa, giúp doanh nghiệp đưa ra những quyết định chính xác và kịp thời. Với mong muốn đưa ra mô hình quản lý phù với mọi quy mô doanh nghiệp từ công ty startup đến doanh nghiệp lớn giúp các doanh nghiệp phát triển một cách bền vững nhất

### **1.2.2 Giá trị cốt lõi**

**Chu đáo:** Đây là sự cam kết luôn luôn lắng nghe và thấu hiểu những vấn đề khách hàng đang phải đối mặt. Chỉ khi đặt mình vào vị trí khách hàng thì những giải pháp được đưa ra mới phù hợp và có tính thực tế cao. Pandosima không ngừng cải tiến để phù hợp với những thay đổi của khách hàng, mục tiêu của Pandosima là trở thành một người đồng hành trên hành trình phát triển của doanh nghiệp

**Toàn diện:** Biết rằng hệ thống vẫn chưa hoàn hảo, nhưng Pandosima luôn hướng đến điều đó, mỗi thành viên trong công ty đều sở hữu kiến thức chuyên môn ở nhiều lĩnh vực khác nhau, để có một góc nhìn đa chiều hơn khi phát triển dịch vụ.Chính sự kết hợp giữa công nghệ, kinh doanh và quản lý giúp Pandosima xây dựng các giải pháp không chỉ mạnh mẽ về kỹ thuật mà còn phù hợp với thực tế vận hành của doanh nghiệp.

**Tư duy phát triển:** Ngày hôm nay tốt hơn ngày hôm trước 1%, Pandosima tin rằng một doanh nghiệp muốn đi được lâu dài thì phải liên tục đổi mới, thích nghi và cải tiến.Pandosima luôn cập nhật các công nghệ mới đồng thời học hỏi các giải pháp tiên tiến, không chỉ trong phát triển sản phầm mà còn trong cách làm việc hợp tác, nhằm mang lại cho khách hàng một trải nghiệm tốt nhất.

### **Dịch vụ công ty**

**Phát triển phần mềm:** Công ty cung cấp dịch vụ tư vấn và phát triển phần mềm theo yêu cầu, từ thiết kế hệ thống, xây dựng ứng dụng đến tích hợp trí tuệ nhân tạo và khai thác dữ liệu. Các giải pháp được thiết kế nhằm hỗ trợ doanh nghiệp tối ưu hóa quy trình vận hành, nâng cao hiệu suất làm việc và tạo ra lợi thế cạnh tranh trên thị trường.

**Đào tạo chuyên sâu:** Ngoài mảng công nghệ, Pandosima Co., Ltd. còn triển khai các chương trình đào tạo chuyên sâu về quản lý kinh doanh, tiếp thị kỹ thuật số và ứng dụng trí tuệ nhân tạo. Chương trình học được thiết kế phù hợp với nhu cầu thực tế, giúp học viên nâng cao kỹ năng chuyên môn và áp dụng hiệu quả vào công việc.

**Cung cấp laptop và linh kiện IoT:** Công ty chuyên phân phối laptop cũ chất lượng cao với giá hợp lý, đáp ứng nhu cầu sử dụng của sinh viên, cá nhân và gia đình. Bên cạnh đó, Pandosima Co., Ltd. cũng cung cấp các linh kiện và mạch IoT, hỗ trợ cộng đồng nghiên cứu và phát triển sản phẩm công nghệ.

**Cho thuê văn phòng chia sẻ:** Nhằm hỗ trợ các doanh nghiệp khởi nghiệp và cộng đồng làm việc tự do, công ty cung cấp dịch vụ cho thuê văn phòng chia sẻ với không gian làm việc yên tĩnh, đầy đủ tiện ích và chi phí hợp lý. Mô hình này giúp tối ưu hóa ngân sách hoạt động đồng thời tạo điều kiện thuận lợi cho kết nối và hợp tác giữa các doanh nghiệp.

## **1.3 Giới thiệu về ngành nghề thực tập**

### **1.3.1 Định Nghĩa về nhà khoa học dữ liệu( Data Scientist)**

Nhà khoa học dữ liệu là người chuyên phân tích dữ liệu, xây dựng mô hình dự đoán và đưa ra các quyết định dựa trên dữ liệu bằng cách sử dụng Machine Learning, AI, thống kê và kỹ thuật khai phá dữ liệu. Làm việc với các tập dữ liệu lớn, trích xuất thông tin giá trị và hỗ trợ doanh nghiệp tối ưu hóa quy trình kinh doanh.

### **1.3.2 Công việc tại Pandosima**

Tại Pandosima, việc khai thác và phân tích dữ liệu đóng vai trò cốt lõi trong quá trình phát triển các hệ thống trí tuệ nhân tạo và phần mềm thông minh. Để đảm bảo chất lượng dữ liệu đầu vào, công việc thu thập dữ liệu được thực hiện từ nhiều nguồn khác nhau, bao gồm hệ thống phần mềm, API và cơ sở dữ liệu (Database). Sau khi thu thập, dữ liệu cần trải qua quá trình tiền xử lý nhằm loại bỏ nhiễu, xử lý giá trị khuyết thiếu và chuẩn hóa định dạng. Các kỹ thuật Data Cleaning và Feature Engineering được áp dụng để trích xuất các đặc trưng quan trọng, giúp cải thiện chất lượng và tính hữu ích của dữ liệu trong các mô hình dự đoán.

Bên cạnh đó, Pandosima tập trung vào việc triển khai và xây dựng các hệ thống ứng dụng Machine Learning nhằm phát triển các giải pháp tự động hóa và hỗ trợ ra quyết định. Các thuật toán học máy được sử dụng để xây dựng mô hình dự đoán với mục tiêu tối ưu hóa độ chính xác và khả năng tổng quát hóa. Trong quá trình này, các thư viện phổ biến như scikit-learn và PyTorch đóng vai trò quan trọng trong việc phát triển, huấn luyện và đánh giá mô hình AI, giúp nâng cao hiệu suất xử lý dữ liệu và khả năng thích ứng của hệ thống.

Sau khi mô hình AI được phát triển, một bước quan trọng là tích hợp dữ liệu vào các sản phẩm phần mềm do Pandosima phát triển. Việc này giúp khai thác tối đa tiềm năng của dữ liệu để tạo ra các giá trị thực tiễn, từ đó nâng cao chất lượng sản phẩm. Đặc biệt, việc ứng dụng các kỹ thuật học máy để xây dựng hệ thống gợi ý (recommendation system) và chatbot AI giúp cá nhân hóa trải nghiệm người dùng, tối ưu hóa quy trình vận hành và tăng cường hiệu suất hoạt động của phần mềm.

Ngoài ra, để đảm bảo mô hình đạt hiệu suất cao nhất, quá trình theo dõi, đánh giá và tối ưu liên tục được thực hiện. Các chỉ số đo lường hiệu suất mô hình (model evaluation metrics) được sử dụng để đánh giá độ chính xác, độ nhạy và tính tổng quát của hệ thống. Việc tối ưu hóa hyperparameters và lựa chọn đặc trưng (feature selection) là những kỹ thuật quan trọng giúp cải thiện hiệu suất của mô hình, giảm thiểu sai số và đảm bảo khả năng áp dụng vào thực tế.

### **1.3.3 Các kỹ năng tối thiểu cần có để trở thành một Data Scientist**

Để trở thành một Data Scientist, người học cần trang bị một tập hợp kỹ năng chuyên sâu, kết hợp giữa lập trình, xử lý dữ liệu, thống kê và triển khai mô hình AI. Một trong những yêu cầu quan trọng nhất là thành thạo ngôn ngữ lập trình Python, do đây là ngôn ngữ phổ biến nhất trong lĩnh vực khoa học dữ liệu, cung cấp một hệ sinh thái phong phú với nhiều thư viện hỗ trợ phân tích, trực quan hóa và triển khai mô hình học máy.

Bên cạnh đó, kiến thức về Machine Learning và AI là nền tảng cốt lõi của một Data Scientist. Trong đó, thư viện scikit-learn đóng vai trò quan trọng trong việc phát triển các mô hình học máy, hỗ trợ nhiều thuật toán như phân loại (classification), hồi quy (regression) và phân cụm (clustering). Việc hiểu và vận dụng các kỹ thuật học máy giúp Data Scientist xây dựng các hệ thống dự đoán có độ chính xác cao.

Ngoài khả năng làm việc với dữ liệu thông thường, một Data Scientist cần có kỹ năng xử lý dữ liệu lớn để làm việc với các tập dữ liệu có quy mô lớn và phức tạp. Công nghệ Hadoop là một trong những công cụ phổ biến giúp quản lý và xử lý dữ liệu phân tán hiệu quả, hỗ trợ tối ưu hóa hiệu suất tính toán trên hệ thống dữ liệu lớn.

Trực quan hóa dữ liệu cũng là một kỹ năng không thể thiếu, giúp chuyển đổi dữ liệu thô thành các biểu đồ, bảng số liệu có ý nghĩa, giúp các bên liên quan dễ dàng hiểu và đưa ra quyết định. Các công cụ như Tableau và thư viện Matplotlib trong Python cung cấp nhiều phương pháp trực quan hóa khác nhau, từ biểu đồ cột, biểu đồ tán xạ đến biểu đồ nhiệt, giúp thể hiện xu hướng và mối quan hệ giữa các biến dữ liệu.

Sau khi xây dựng mô hình AI, một bước quan trọng là triển khai AI vào thực tế để tích hợp vào các hệ thống phần mềm và ứng dụng. Trong đó, FastAPI là một công cụ mạnh mẽ giúp triển khai các mô hình học máy dưới dạng dịch vụ web API, cho phép giao tiếp với các hệ thống khác một cách nhanh chóng và hiệu quả.

Cuối cùng, một Data Scientist cần có kỹ năng thống kê để hiểu rõ bản chất của dữ liệu và đưa ra các phân tích chính xác. Các phương pháp thống kê như hồi quy (regression) giúp xác định mối quan hệ giữa các biến, trong khi phân cụm (clustering) hỗ trợ trong việc nhóm các dữ liệu có đặc điểm tương đồng. Việc nắm vững các kỹ thuật này giúp Data Scientist có thể xây dựng các mô hình phân tích mạnh mẽ và tối ưu hóa quy trình xử lý dữ liệu.

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **2.1. Hệ thống gợi ý**

### **2.1.1** **Định nghĩa**

Hệ thống gợi ý (Recommendation System) ngày càng trở thành một phần không thể thiếu trong các nền tảng số hiện nay. Đây là công cụ giúp cá nhân hóa trải nghiệm người dùng, mang lại các đề xuất phù hợp với nhu cầu và sở thích cá nhân của họ. Nhờ vào công nghệ này, doanh nghiệp có thể tối ưu hóa chiến lược tiếp thị, gia tăng doanh thu và cải thiện mức độ hài lòng của khách hàng. Từ các nền tảng thương mại điện tử, mạng xã hội đến dịch vụ phát trực tuyến, hệ thống gợi ý đóng vai trò quan trọng trong việc giữ chân người dùng và tối ưu hóa hiệu quả kinh doanh.

### **2.1.2 Sự Thay Đổi Của Quảng Cáo Và Vai Trò Của Hệ Thống Gợi Ý**

Trước đây, các chiến dịch quảng cáo truyền thống chủ yếu được triển khai trên các phương tiện truyền thông đại chúng như truyền hình và báo chí. Các doanh nghiệp phải chi trả một khoản lớn để hiển thị quảng cáo trên các khung giờ cao điểm với hy vọng tiếp cận càng nhiều người càng tốt. Tuy nhiên, phương thức này có một nhược điểm lớn: thiếu tính cá nhân hóa. Một quảng cáo xuất hiện trên truyền hình có thể không phù hợp với đa số người xem, dẫn đến hiệu quả thấp.

Sự phát triển của Internet và công nghệ dữ liệu lớn (Big Data) đã thay đổi hoàn toàn cách thức tiếp cận người dùng. Các nền tảng kỹ thuật số hiện nay không chỉ cung cấp nội dung chung cho mọi người mà còn sử dụng hệ thống gợi ý để hiển thị thông tin phù hợp với từng cá nhân. Chẳng hạn, nếu một người thường xuyên tìm kiếm các sản phẩm điện tử, hệ thống sẽ hiển thị các quảng cáo về thiết bị công nghệ thay vì các sản phẩm không liên quan. Nhờ vậy, doanh nghiệp có thể tối ưu hóa chi phí quảng cáo, trong khi người dùng cũng có trải nghiệm tốt hơn khi họ được tiếp cận với nội dung phù hợp.

### **2.1.3 Nguyên lý hoạt động**

Hệ thống gợi ý (Recommendation System) là một trong những ứng dụng quan trọng của trí tuệ nhân tạo, hoạt động dựa trên nguyên tắc phân tích dữ liệu để xác định các mẫu hành vi của người dùng, từ đó đưa ra các đề xuất phù hợp. Hiện nay, có ba phương pháp chính thường được sử dụng trong các hệ thống gợi ý: lọc cộng tác (collaborative filtering), lọc nội dung (content-based filtering) và hệ thống lai (hybrid system). Lọc cộng tác dựa vào phân tích hành vi của người dùng khác có sở thích tương đồng để đưa ra khuyến nghị, trong khi lọc nội dung sử dụng đặc điểm của sản phẩm hoặc nội dung để đề xuất cho người dùng. Phương pháp lai kết hợp cả hai kỹ thuật trên nhằm nâng cao độ chính xác và khả năng cá nhân hóa.

## **2.2. Các loại hệ thống gợi ý sản phẩm**

### **2.2.1. Hệ Thống Gợi Ý Dựa Trên Nội Dung (Content-Based Filtering)**

#### **2.2.1.1 . Định Nghĩa**

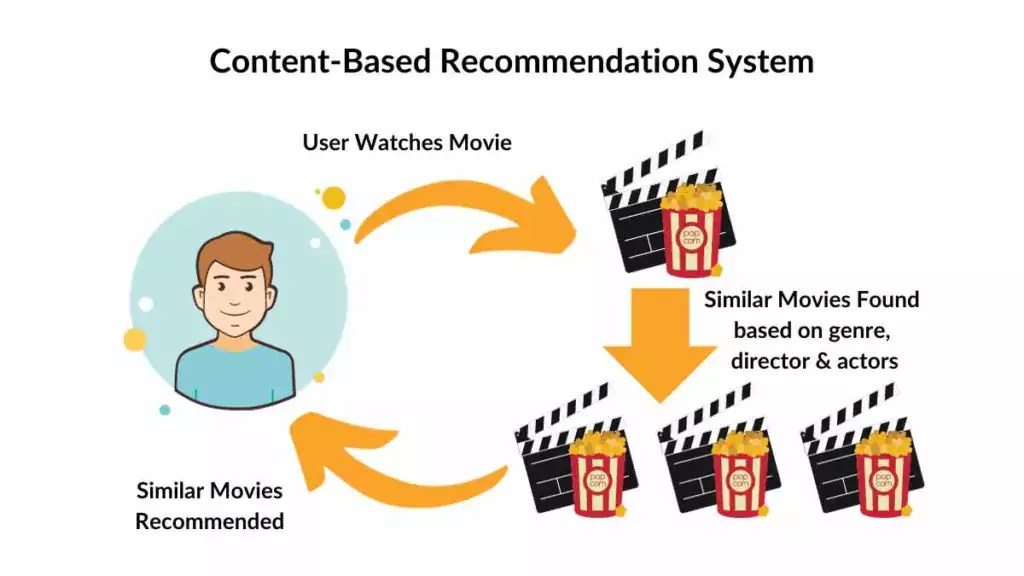


Figure 1 Content based Recommendation

Hệ thống gợi ý dựa trên nội dung (Content-Based Recommendation System – CBRS) là một phương pháp đề xuất thông tin được phát triển dựa trên việc phân tích đặc trưng của các mục (items) mà người dùng đã từng quan tâm hoặc tương tác trong quá khứ. Thay vì dựa vào hành vi của cộng đồng người dùng như phương pháp lọc cộng tác (Collaborative Filtering), CBRS tập trung xây dựng hồ sơ người dùng (user profile) bằng cách thu thập và học từ các đặc điểm nội tại của các mục đã được đánh giá tích cực bởi chính người dùng đó.

Các mục trong hệ thống được mô tả thông qua tập hợp các đặc trưng có thể là dữ liệu định lượng (như năm phát hành, số trang), dữ liệu định tính (như thể loại, tác giả, đạo diễn) hoặc đặc trưng được trích xuất từ nội dung (như từ khóa, tần suất xuất hiện của từ). Hồ sơ người dùng sau đó sẽ được so sánh với hồ sơ của các mục khác trong hệ thống để xác định mức độ tương đồng và từ đó đề xuất các mục phù hợp với sở thích cá nhân hóa của người dùng.

Phương pháp này được ứng dụng rộng rãi trong nhiều lĩnh vực như thương mại điện tử, phát trực tuyến, học trực tuyến và báo chí kỹ thuật số, giúp cá nhân hóa trải nghiệm người dùng và nâng cao hiệu quả tương tác.

#### **2.2.1.2. Cách Hoạt Động**

**a) Xây dựng hồ sơ người dùng (User Profile)**

Hồ sơ người dùng là biểu diễn các đặc điểm hoặc sở thích của người dùng dựa trên các mục mà họ đã tương tác, đánh giá hoặc yêu thích. Dữ liệu có thể bao gồm số lần xem, lượt thích, đánh giá số sao, hoặc các hành động khác như thêm vào danh sách yêu thích.

**b) Xây dựng hồ sơ mục (Item Profile)**

Hồ sơ mục là tập hợp các đặc trưng mô tả một mục cụ thể. Những đặc trưng này có thể là thuộc tính có cấu trúc (ví dụ: thể loại phim, tác giả sách, đạo diễn) hoặc đặc trưng trích xuất từ nội dung (ví dụ: từ khóa bài báo, mô hình “bag-of-words”). Ví dụ, trong hệ thống gợi ý sách, một cuốn tiểu thuyết có thể được mô tả thông qua các thuộc tính như thể loại (lãng mạn, phiêu lưu), tác giả, và ngôn ngữ.

Dưới dạng biểu diễn vector, mỗi mục sẽ được ánh xạ vào không gian đa chiều, trong đó mỗi chiều biểu thị một đặc trưng nội dung. Ví dụ, ba cuốn tiểu thuyết có thể được mô hình hóa như sau:

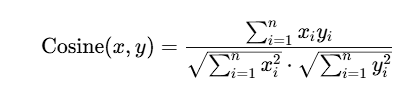
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tên sách | |  | | --- | |  |  |  | | --- | | **Phiêu lưu** | | |  | | --- | |  |  |  | | --- | | **Lãng mạn** | | |  | | --- | |  |  |  | | --- | | **Trưởng thành** | |
| Peter Pan | 1 | 0 | 0 |
| Little Women | 0 | 1 | 1 |
| Treasure Island | 1 | 0 | 0 |

Mỗi mục như vậy được biểu diễn bởi một vector nhị phân trong không gian đặc trưng, ví dụ Peter Pan là (1,0,0).

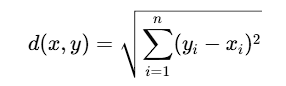
**c) Đo lường mức độ tương đồng (Similarity Measurement)**

Để xác định các mục tương đồng với sở thích của người dùng, hệ thống sử dụng các độ đo khoảng cách hoặc độ tương đồng giữa các vector. Một số phương pháp phổ biến bao gồm:

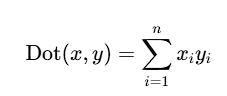
**- Cosine Similarity (Độ tương đồng Cosine)** Giá trị cosine dao động từ 0 đến 1, trong đó giá trị gần 1 biểu thị rằng hai vector có hướng gần giống nhau, tức là mức độ tương đồng cao



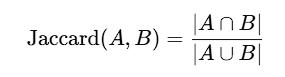
**- Euclidean Distance** đo khoảng cách tuyến tính giữa hai điểm dữ liệu Khoảng cách càng nhỏ thì hai điểm càng giống nhau. Tuy nhiên, phương pháp này dễ bị ảnh hưởng nếu dữ liệu không được chuẩn hóa.



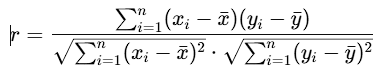
- **Dot Product** (tích vô hướng) là một phép nhân hai vector, cho kết quả là một số thể hiện mức độ tương quan. Kết quả lớn hơn cho thấy hai vector có xu hướng cùng chiều, tuy nhiên phương pháp này không phản ánh tốt mức độ tương đồng nếu độ lớn vector chênh lệch quá lớn.



- **Jaccard Similarity** thường được áp dụng cho dữ liệu dạng nhị phân hoặc tập hợp, như việc người dùng có xem một bộ phim hay không. Công thức là:



**- Pearson Correlation** được dùng khi dữ liệu là các giá trị xếp hạng (ratings). Phương pháp này đo lường mối quan hệ tuyến tính giữa hai vector sau khi đã loại bỏ ảnh hưởng của trung bình. Giá trị r nằm trong khoảng từ -1 đến 1; giá trị dương cao thể hiện tương quan mạnh mẽ giữa hai vector.



**d) Dự đoán tương tác người dùng – mục và đề xuất**

Dựa trên vector hóa của hồ sơ người dùng và các mục mới, hệ thống sẽ sử dụng mô hình học máy như hồi quy hoặc phân loại để dự đoán khả năng người dùng sẽ tương tác tích cực với một mục cụ thể. Những mục có xác suất tương tác cao nhất sẽ được đưa vào danh sách đề xuất.

### **2.2.2. Hệ Thống Gợi Ý Dựa Trên Lọc Cộng Tác (Collaborative Filtering)**

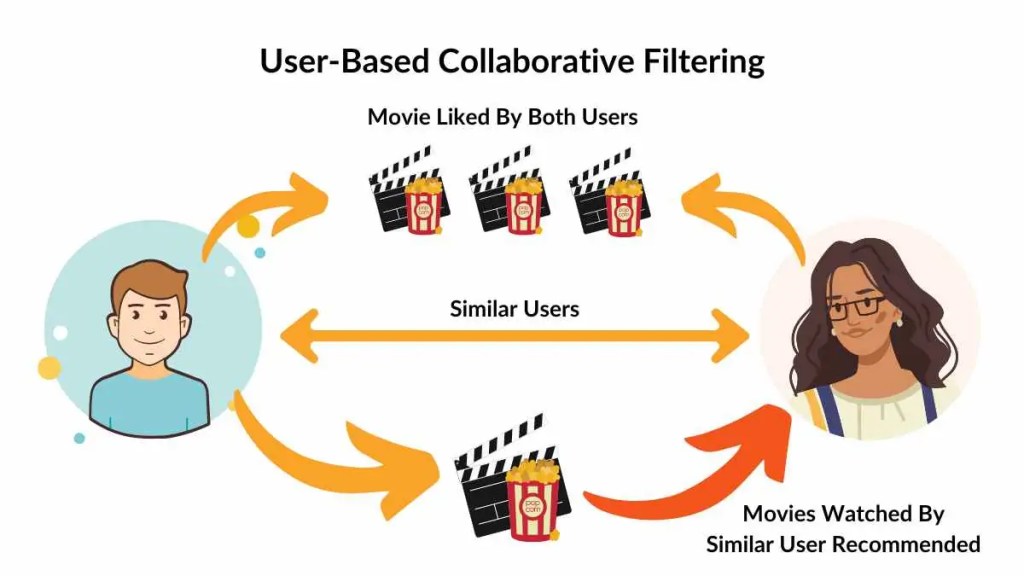


Figure 2 Collaborative Filtering

#### **2.2.2.1. Định Nghĩa**

Hệ thống gợi ý dựa trên lọc cộng tác (Collaborative Filtering - CF) là một trong những kỹ thuật phổ biến và lâu đời nhất được ứng dụng trong các hệ thống đề xuất. Phương pháp này hoạt động dựa trên nguyên lý: “Người dùng có hành vi tương tự trong quá khứ có khả năng sẽ có sở thích giống nhau trong tương lai”. Thay vì phân tích đặc điểm nội dung như phương pháp gợi ý dựa trên nội dung (Content-Based Filtering), lọc cộng tác tập trung vào mối quan hệ giữa người dùng và các mục (item), thông qua việc khai thác lịch sử tương tác (như lượt thích, đánh giá, mua hàng, v.v.) để đưa ra đề xuất phù hợp.

Phương pháp này đặc biệt phát huy hiệu quả trong những hệ thống có cơ sở dữ liệu người dùng lớn và phong phú về hành vi, chẳng hạn như nền tảng thương mại điện tử (Amazon), dịch vụ phát trực tuyến (Netflix, Spotify) hay mạng xã hội (Facebook, TikTok). CF không yêu cầu thông tin về đặc điểm của mục, do đó nó có khả năng gợi ý cả những mục mới mẻ với người dùng – những nội dung mà bản thân người dùng có thể chưa từng biết đến.

#### **2.2.2.2. Cách Hoạt Động**

Lọc cộng tác chia thành hai nhóm chính:

**Lọc cộng tác dựa trên người dùng (User-Based Collaborative Filtering):** Hệ thống tìm kiếm những người dùng có hành vi tương tự với người dùng mục tiêu, sau đó đề xuất những mục mà những người dùng này đã đánh giá cao nhưng người dùng mục tiêu chưa từng tương tác. Quá trình này bao gồm:

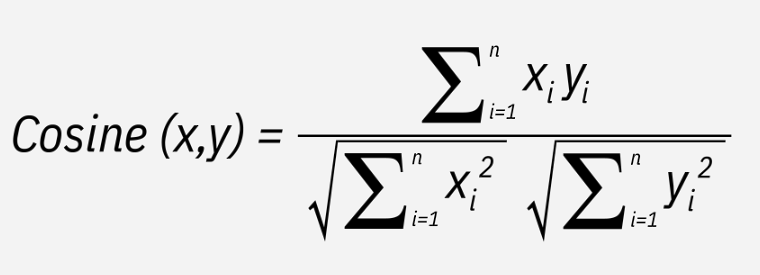
* Tính toán độ tương đồng giữa các người dùng dựa trên lịch sử đánh giá, sử dụng các độ đo như Cosine Similarity hoặc Pearson Correlation.
* Lựa chọn một nhóm nearest neighbors.
* Dự đoán mức độ yêu thích của người dùng mục tiêu với các mục mới dựa trên đánh giá của nhóm hàng xóm.

**Lọc cộng tác dựa trên mục (Item-Based Collaborative Filtering):** Thay vì tìm người dùng tương tự, hệ thống tìm các mục tương tự nhau dựa trên mô hình đánh giá từ tất cả người dùng. Nếu người dùng đã thích mục A, hệ thống sẽ gợi ý mục B nếu nhiều người dùng khác cũng đồng thời thích cả A và B. Quá trình này bao gồm:

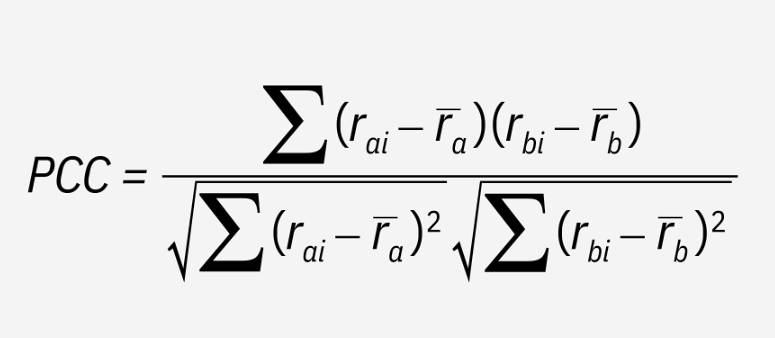
* Tính toán độ tương đồng giữa các mục.
* Xác định những mục liên quan đến các mục mà người dùng đã tương tác.
* Dự đoán và đề xuất các mục mới.

#### **2.2.2.3. Các phương pháp tính độ tương đồng**

**Cosine Similarity:** Tính toán góc giữa hai vector trong không gian đặc trưng. Sử dụng khi dữ liệu đánh giá là các vector hiếm (sparse vectors). Công thức:



**Pearson Correlation Coefficient (PCC):** Đo độ tương quan tuyến tính giữa hai tập đánh giá. Thích hợp khi muốn loại bỏ độ lệch về quy mô đánh giá của từng người dùng. PCC nằm trong khoảng [-1, 1].



#### **2.2.2.4. Phân loại hệ thống CF**

**Memory-Based CF (Dựa trên bộ nhớ):** Lưu trữ toàn bộ ma trận người dùng - mục (user-item matrix), và thực hiện phép tính trực tiếp trên dữ liệu lịch sử. Loại này bao gồm cả user-based và item-based.

**Model-Based CF (Dựa trên mô hình học máy):** Sử dụng các thuật toán học máy như Decision Tree, KNN, Bayesian, Matrix Factorization, hoặc các mô hình học sâu như Neural Collaborative Filtering để xây dựng mô hình từ dữ liệu lịch sử, từ đó dự đoán tương tác tiềm năng.

### **2.2.3. Hệ Thống Gợi Ý Lai (Hybrid Recommendation Systems)**

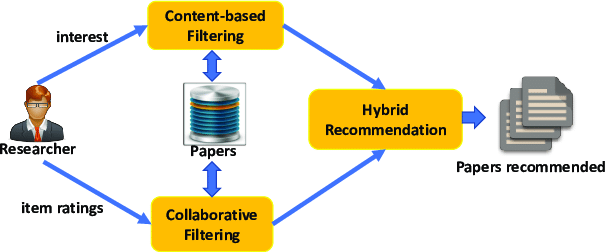


Figure 3 Hybrid Recommendation System

#### **2.2.3.1. Định Nghĩa**

Hệ thống gợi ý lai (Hybrid Recommendation System) là một phương pháp tiên tiến kết hợp giữa lọc dựa trên nội dung (Content-Based Filtering) và lọc cộng tác (Collaborative Filtering), nhằm tận dụng ưu điểm của cả hai phương pháp trong việc cá nhân hóa nội dung. Trong khi lọc dựa trên nội dung phân tích đặc điểm của các mục mà người dùng đã quan tâm để đưa ra gợi ý tương tự, lọc cộng tác lại dựa vào hành vi của nhiều người dùng để xác định xu hướng chung. Việc kết hợp hai phương pháp này giúp khắc phục những hạn chế riêng lẻ của từng phương pháp, chẳng hạn như vấn đề cold start (thiếu dữ liệu ban đầu) của lọc cộng tác hoặc tính hạn chế trong phạm vi đề xuất của lọc nội dung. Nhờ đó, hệ thống gợi ý lai không chỉ tăng độ chính xác của dự đoán mà còn giúp cải thiện trải nghiệm người dùng bằng cách cung cấp các đề xuất phong phú và đa dạng hơn.

#### **2.2.3.2. Cách Hoạt Động**

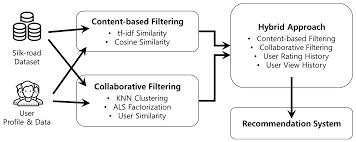


Figure 4 Cách tiếp cận của thuật toán Hybrid

Có nhiều cách tiếp cận khác nhau để xây dựng hệ thống gợi ý lai, mỗi cách có cách thức vận hành riêng để tối ưu hóa kết quả đề xuất. Một phương pháp phổ biến là kết hợp tuyến tính, trong đó điểm số dự đoán từ cả hai mô hình được tính toán và gộp lại theo một trọng số phù hợp để đưa ra đề xuất cuối cùng. Một cách tiếp cận khác là kết hợp theo tầng, trong đó một phương pháp được sử dụng trước để lọc dữ liệu hoặc thu hẹp tập hợp đề xuất, sau đó kết quả sẽ tiếp tục được xử lý bằng phương pháp còn lại để nâng cao độ chính xác. Ngoài ra, kết hợp theo đặc trưng là một phương pháp mạnh mẽ khác, trong đó thông tin từ cả hai mô hình được sử dụng để xây dựng các đặc trưng tổng hợp, giúp thuật toán học sâu (Deep Learning) hoặc các mô hình học máy (Machine Learning) có thể tận dụng toàn bộ dữ liệu để tạo ra đề xuất chất lượng hơn. Với những lợi thế vượt trội này, hệ thống gợi ý lai đã trở thành một trong những giải pháp hiệu quả nhất trong các nền tảng thương mại điện tử, dịch vụ phát trực tuyến và các hệ thống quản lý nội dung số hiện nay.

## **2.3 Giới thiệu tổng quan về Django**

### **2.3.1 Giới thiệu**

Django là khung công tác mã nguồn mở dùng Python, hỗ trợ phát triển ứng dụng web nhanh, bảo mật và dễ mở rộng. Ra mắt năm 2005 bởi Adrian Holovaty và Simon Willison, Django được chọn cho hệ thống đề xuất sách nhờ khả năng tích hợp học máy, quản lý dữ liệu và tạo giao diện thân thiện

### **2.3.2 Tính năng chính của Django**

Django sử dụng kiến trúc MVT (Model-View-Template), tách biệt dữ liệu, logic và giao diện. Trong hệ thống, Model quản lý dữ liệu sách (tên, thể loại, giá), View xử lý gợi ý với TfidfVectorizer và cosine\_similarity, Template tạo giao diện bằng HTML và Tailwind CSS [1].

ORM của Django cho phép thao tác cơ sở dữ liệu MySQL bằng Python, hỗ trợ nhập và truy vấn dữ liệu sách hiệu quả, tương thích với nhiều hệ quản trị.

Django kết hợp với pandas, scikit-learn để xử lý dữ liệu và gợi ý, hỗ trợ mở rộng qua các gói như Django REST Framework

### **2.3.3 Cấu trúc của một ứng dụng Django**

Cấu trúc ứng dụng Django trong hệ thống đề xuất sách bao gồm các thành phần chính sau. Phần cốt lõi là ứng dụng, nơi cấu hình mô hình, định tuyến URL và tích hợp thư viện học máy. Mô hình (Model) định nghĩa dữ liệu sách như tên, giá, mô tả, được lưu trữ trong MySQL qua ORM. View xử lý yêu cầu người dùng, từ tìm kiếm đến gợi ý, kết nối dữ liệu với giao diện. Template sử dụng HTML và Tailwind CSS để hiển thị giao diện động, bao gồm tìm kiếm, lọc và phân trang, đảm bảo tính tái sử dụng và thân thiện với người dùng [1].

# **CHƯƠNG 3: XÂY DỰNG HỆ THỐNG GỢI Ý SÁCH HỖ TRỢ BÁN HÀNG**

## **3.1. Thu thập và xử lý dữ liệu**

### **3.1.1. Nguồn dữ liệu**

#### **3.1.1.1 Sự quan trọng của dữ liệu đối với hệ thống đề xuất**

Dữ liệu đóng vai trò cốt lõi trong mọi hệ thống gợi ý, quyết định khả năng cung cấp các đề xuất chính xác và phù hợp với nhu cầu người dùng. Giống như một đầu bếp cần nguyên liệu chất lượng để tạo ra món ăn ngon, hệ thống gợi ý phụ thuộc vào dữ liệu đầy đủ, chính xác và đa dạng để xây dựng các gợi ý mang tính cá nhân hóa.

Dữ liệu gợi ý thường được chia thành hai loại chính: dữ liệu người dùng và dữ liệu sản phẩm. Dữ liệu người dùng bao gồm thông tin về hành vi và sở thích, như lịch sử mua sắm, lượt xem sản phẩm, đánh giá hoặc tương tác khác (thêm vào giỏ hàng, lượt thích). Những thông tin này giúp hệ thống nắm bắt thị hiếu cá nhân, dù trong nghiên cứu này, dữ liệu người dùng chưa được khai thác. Trong khi đó, dữ liệu sản phẩm – cụ thể là sách trong đề tài – bao gồm các thuộc tính như tiêu đề, thể loại, nhà xuất bản, mô tả nội dung, điểm đánh giá, và số lượt mua. Đây là nguồn thông tin chính để hệ thống xây dựng mô hình gợi ý dựa trên nội dung (Content-Based Filtering).

Chất lượng dữ liệu ảnh hưởng trực tiếp đến hiệu quả hệ thống. Dữ liệu thiếu sót, không đồng nhất hoặc chứa nhiễu sẽ làm giảm khả năng phân tích và dự đoán, dẫn đến các gợi ý kém giá trị. Ngược lại, dữ liệu phong phú và đáng tin cậy tạo điều kiện để thuật toán học đặc trưng sản phẩm và đưa ra gợi ý phù hợp.

**3.1.1.2 Thu thập dữ liệu**

Để xây dựng hệ thống gợi ý sách hiệu quả, nghiên cứu đã thu thập dữ liệu từ nền tảng thương mại điện tử Tiki.vn thông qua giao diện lập trình ứng dụng (API). Phương pháp này đảm bảo dữ liệu có cấu trúc, dễ tự động hóa và giảm thiểu sai sót so với thu thập thủ công.

Quá trình thu thập được chia thành hai giai đoạn chính:

**Giai đoạn 1: Thu thập danh sách sản phẩm**

Sử dụng API tìm kiếm của Tiki với từ khóa "sách" để truy xuất danh sách các đầu sách. Từ khóa này được chọn để bao quát nhiều thể loại, từ sách văn học đến giáo khoa. Kết quả trả về theo từng trang (pagination), nên nghiên cứu đã truy vấn liên tiếp nhiều trang để thu thập danh sách mã sản phẩm (product ID). Các mã này đóng vai trò là chìa khóa để lấy thông tin chi tiết ở giai đoạn sau, đồng thời giúp đảm bảo không bỏ sót các đầu sách ít phổ biến nhưng có giá trị.

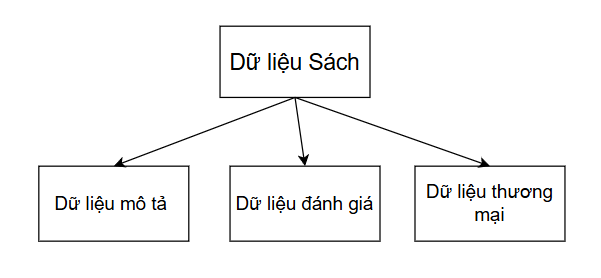
**Giai đoạn 2: Lấy thông tin chi tiết sản phẩm**

Dựa trên danh sách mã sản phẩm, nghiên cứu tiếp tục truy vấn API chi tiết của Tiki để thu thập thông tin đầy đủ về từng cuốn sách, bao gồm tiêu đề, thể loại, mô tả, giá, đánh giá, và các thuộc tính khác. Dữ liệu được lưu trữ dưới dạng cấu trúc nhất quán, tạo điều kiện thuận lợi cho việc xử lý sau này.

Để đảm bảo tính ổn định, quá trình thu thập được kiểm soát chặt chẽ. Tốc độ gửi yêu cầu được giới hạn với các khoảng nghỉ hợp lý nhằm tránh gây áp lực lên máy chủ Tiki hoặc bị chặn truy cập. Đồng thời, hệ thống giám sát liên tục các lỗi tiềm ẩn như kết nối gián đoạn hoặc dữ liệu trả về không hợp lệ, từ đó đảm bảo chất lượng và tính cập nhật của dữ liệu. Kết quả là một tập dữ liệu phong phú, sẵn sàng cho các bước phân tích và xây dựng mô hình gợi ý.

#### **3.1.1.2. Phân loại dữ liệu thu thập**

Dữ liệu thu thập từ nền tảng Tiki.vn được phân loại dựa trên nội dung và vai trò trong hệ thống gợi ý, nhằm tối ưu hóa quá trình xử lý và xây dựng mô hình. Theo cách tiếp cận này, dữ liệu được chia thành ba nhóm chính như sau:



**Figure 5 Phân loại dữ liệu thu thập**

**Dữ liệu mô tả sản phẩm** bao gồm các thông tin đặc trưng của từng cuốn sách, giúp hệ thống hiểu rõ nội dung và thuộc tính cơ bản. Nhóm này bao gồm tiêu đề sách, thể loại (ví dụ: văn học, khoa học, thiếu nhi), mô tả ngắn, mô tả chi tiết, nhà xuất bản, công ty phát hành, loại bìa (bìa cứng, bìa mềm), và số trang. Những thông tin này đóng vai trò quan trọng trong việc xây dựng mô hình gợi ý dựa trên nội dung, vì chúng phản ánh đặc điểm riêng biệt của từng sản phẩm, từ đó giúp hệ thống xác định sự tương đồng giữa các cuốn sách.

**Dữ liệu đánh giá sản phẩm** tập trung vào phản hồi của người dùng về chất lượng sách, bao gồm điểm đánh giá trung bình (thang điểm từ 0 đến 5) và số lượng đánh giá. Những thông tin này không chỉ thể hiện mức độ phổ biến và độ tin cậy của sản phẩm mà còn hỗ trợ hệ thống ưu tiên các đầu sách có chất lượng cao trong quá trình gợi ý. Mặc dù dữ liệu đánh giá có thể phản ánh sở thích của người dùng, nghiên cứu này chủ yếu sử dụng chúng để bổ trợ cho việc mô tả sản phẩm, thay vì phân tích hành vi cá nhân.

**Dữ liệu thương mại** bao gồm các thông tin liên quan đến giao dịch và tồn kho, như giá bán, giá gốc, tỷ lệ giảm giá, số lượng đã bán, trạng thái tồn kho (còn hàng, hết hàng), và tên nhà cung cấp hiện tại. Nhóm dữ liệu này cung cấp cái nhìn tổng quan về hiệu suất bán hàng và tính sẵn có của sản phẩm, hỗ trợ hệ thống cân nhắc các yếu tố thực tế khi đưa ra gợi ý, chẳng hạn ưu tiên sách đang có sẵn hoặc đang được giảm giá.

Việc phân loại dữ liệu như trên giúp nghiên cứu tổ chức thông tin một cách khoa học, tạo điều kiện thuận lợi cho các bước xử lý và phân tích tiếp theo. Mỗi nhóm dữ liệu đều đóng góp vào việc nâng cao độ chính xác và tính thực tiễn của hệ thống gợi ý, đồng thời đảm bảo rằng các đề xuất không chỉ dựa trên nội dung mà còn phù hợp với bối cảnh thương mại điện tử.

#### **3.1.1.3. Cấu trúc dữ liệu lưu trữ**

Để đảm bảo dữ liệu thu thập được quản lý hiệu quả và sẵn sàng cho các bước phân tích, nghiên cứu đã thiết kế một cấu trúc lưu trữ thống nhất và có tổ chức. Dữ liệu được lưu trong cơ sở dữ liệu quan hệ, với bảng chính chứa thông tin về sách được xây dựng dựa trên các thuộc tính thu thập từ Tiki.vn. Cấu trúc này không chỉ hỗ trợ việc truy xuất nhanh mà còn tạo điều kiện thuận lợi cho việc tích hợp với hệ thống gợi ý.

Bảng dữ liệu chính, được gọi là bảng sản phẩm, bao gồm các cột tương ứng với các thuộc tính của sách. Cụ thể, mỗi bản ghi đại diện cho một cuốn sách và chứa các thông tin sau:

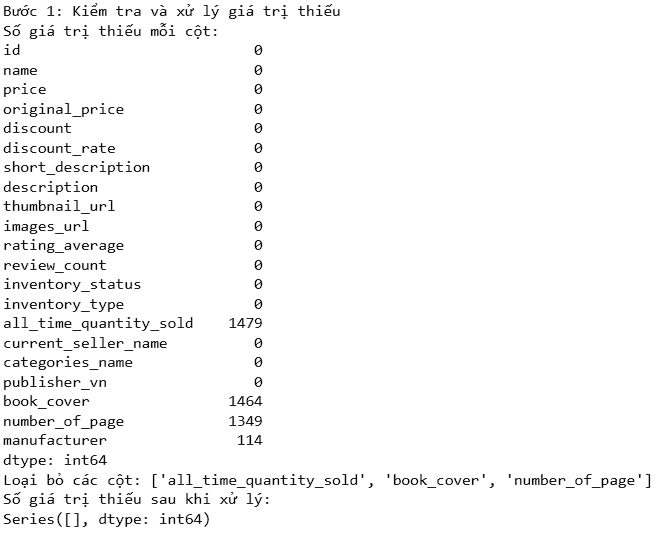
|  |  |  |
| --- | --- | --- |
| Tên trường | Kiểu dữ liệu | Mô tả |
| Id | Số nguyên | Mã định danh duy nhất cho mỗi cuốn sách. |
| Name | Chuỗi | Tên hoặc tiêu đề của cuốn sách. |
| Price | Số nguyên | Giá bán hiện tại của cuốn sách (tính bằng VND). |
| Original\_price | Số nguyên | Giá gốc của cuốn sách trước khi giảm giá (tính bằng VND). |
| Discount | Số nguyên | Số tiền giảm giá áp dụng cho cuốn sách (tính bằng VND). |
| Discount\_rate | Số nguyên | Tỷ lệ phần trăm giảm giá áp dụng cho cuốn sách. |
| Short\_description | Chuỗi | Tóm tắt ngắn gọn hoặc nội dung quảng bá về cuốn sách. |
| Description | Chuỗi | Mô tả chi tiết về nội dung sách và thông tin tác giả. |
| Thumpnail\_url | Chuỗi | Đường dẫn URL đến hình ảnh bìa chính của cuốn sách. |
| Images\_url | Chuỗi | Danh sách các URL hình ảnh bổ sung, phân tách bằng dấu chấm phẩy. |
| Rating Average | Số thực | Điểm đánh giá trung bình của cuốn sách (ví dụ: từ 0 đến 5). |
| Review\_count | Số nguyên | Số lượng đánh giá của cuốn sách. |
| Inventory\_status | Chuỗi | Trạng thái tồn kho của cuốn sách (ví dụ: available, backorder, instock). |
| Inventory\_type | Chuỗi | Loại quản lý tồn kho (ví dụ: backorder, instock). |
| All\_time\_quanity\_sold | Chuỗi | Tổng số lượng sách đã bán |
| Current\_seller\_name | Chuỗi | Tên nhà bán hoặc nhà sách cung cấp cuốn sách. |
| Categories\_name | Chuỗi | Thể loại hoặc danh mục của cuốn sách (ví dụ: Sách tư duy - Kỹ năng sống). |
| Publisher\_vn | Chuỗi | Tên nhà xuất bản tại Việt Nam. |
| Book\_cover | Chuỗi | Loại bìa sách (ví dụ: Bìa mềm, Bìa gập). |
| Number\_of Page | Chuỗi | Số trang của cuốn sách (lưu dưới dạng chuỗi, có thể cần chuyển đổi). |
| Manufacturer | Chuỗi | Đơn vị sản xuất sách (thường là nhà xuất bản). |

**Figure 6 Bảng mô tả dữ liệu**

Cơ sở dữ liệu được thiết kế để cho phép các trường có thể để trống, nhằm xử lý trường hợp thông tin từ Tiki.vn không đầy đủ (ví dụ: một số sách thiếu mô tả chi tiết hoặc số trang). Điều này đảm bảo tính linh hoạt và khả năng mở rộng khi dữ liệu được cập nhật hoặc bổ sung từ các nguồn khác trong tương lai.

Dữ liệu được lưu trữ dưới dạng bảng quan hệ giúp tối ưu hóa các thao tác truy vấn, đặc biệt khi cần lọc sách theo thể loại, giá, hoặc tìm kiếm theo từ khóa. Ngoài ra, cấu trúc này hỗ trợ tích hợp trực tiếp với hệ thống gợi ý, cho phép thuật toán truy xuất nhanh các thuộc tính như tiêu đề, thể loại, và mô tả để phân tích nội dung và tạo ra các đề xuất phù hợp.

### **3.1.2. Tiền xử lý dữ liệu**

****

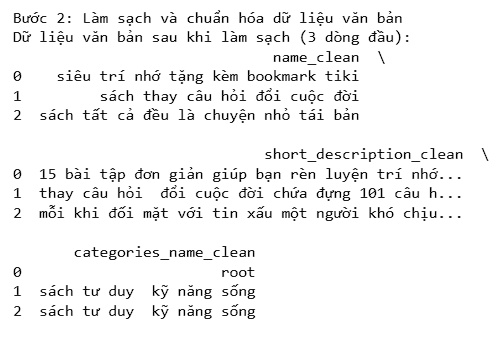
**Figure 7 Xử lý dữ liệu bị thiếu**

**Bước 1: Loại bỏ các giá trị thiếu**

Kiểm tra các giá trị thiếu trên các cột cho thấy các cột như all\_time\_quantity\_sold, book\_cover và number\_of\_page có số lượng giá trị thiếu lần lượt là 1479, 1464 và 1349.Vì thiếu quá nhiều dữ liệu nên loại bỏ ba cột này để tránh ảnh hưởng tiêu cực đến hiệu suất của mô hình, đồng thời giảm thiểu sự phức tạp trong quá trình xử lý dữ liệu

Dữ liệu về manufacturer vẫn được dữ lại và fill bằng giá trị mặc định là 'Unknown' để để đảm bảo tính đầy đủ của dữ liệu mà không làm mất đi ý nghĩa của cột

**Bước 2: Làm sạch và chuẩn hóa dữ liệu văn bản**



**Figure 8 Làm sạch dữ liệu văn bản**

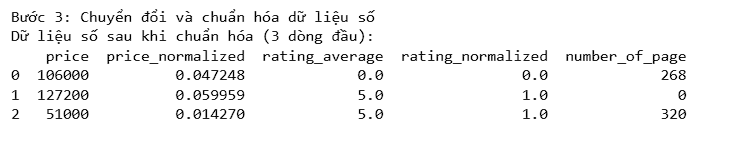
Để phục vụ cho việc làm sạch và chuẩn hóa dữ liệu văn bản, hàm clean\_text đóng vai trò thực hiện nhiều thao tác quan trọng bao gồm:

Đầu tiên, kiểm tra và chuyển đổi các giá trị không phải chuỗi thành chuỗi rỗng để tránh lỗi đồng thời loại bỏ các URL bằng cách sử dụng biểu thức chính quy nhằm đảm bảo nội dung văn bản không chứa các liên kết không liên quan.

Sau đó, loại bỏ các ký tự đặc biệt, chỉ giữ lại chữ cái, số và khoảng trắng để làm sạch dữ liệu; cuối cùng, chuyển toàn bộ văn bản về chữ thường và loại bỏ khoảng trắng thừa ở đầu hoặc cuối chuỗi để chuẩn hóa định dạng.

Hàm clean\_text được áp dụng cho các cột văn bản quan trọng trong tập dữ liệu, bao gồm name, short\_description, description, categories\_name và publisher\_vn

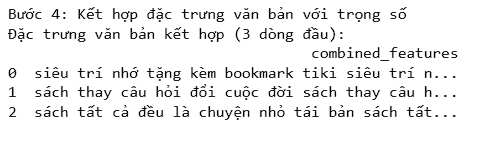
**Bước 3: Chuyển đổi và chuẩn hóa dữ liệu số**



**Figure 9 Chuyển đổi dữ liệu dạng số**

Chuyển đổi cột number\_of\_page thành kiểu số nguyên, đồng thời chuẩn hóa cột price và rating\_average về khoảng giá trị [0, 1] để giảm sự chênh lệch về độ lớn, tạo ra các cột mới là price\_normalized và rating\_normalized nhằm tăng tính nhất quán, hỗ trợ mô hình gợi ý hoạt động hiệu quả hơn.

**Bước 4: Kết hợp đặc trưng văn bản với trọng số**



**Figure 10 Kết hợp trọng số với đặc trưng văn bản**

Kết hợp các cột name\_clean, short\_description\_clean, description\_clean, categories\_name\_clean và publisher\_vn\_clean thành một cột mới combined\_features, trong đó áp dụng trọng số cao hơn cho name\_clean và description\_clean bằng cách lặp lại giá trị của chúng để tăng tầm quan trọng. Sau khi hoàn thành, tôi hiển thị ba dòng đầu tiên của cột combined\_features để kiểm tra kết quả.Điều này giúp tạo ra một đặc trưng văn bản tổng hợp, nhấn mạnh các thông tin quan trọng, giúp mô hình gợi ý dựa trên TF-IDF và độ tương đồng cosine hoạt động hiệu quả hơn, từ đó nâng cao chất lượng gợi ý sách.

## **3.2. Xây dựng thuật toán đề xuất**

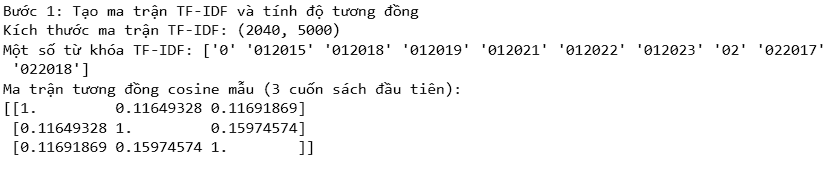
### **3.2.1. Chọn phương pháp gợi ý phù hợp**

Việc lựa chọn phương pháp gợi ý cần được thực hiện dựa trên đặc điểm của dữ liệu hiện có và mục tiêu tối ưu hóa trải nghiệm khách hàng trên nền tảng thương mại điện tử Tiki. Dữ liệu được thu thập trong nghiên cứu này chủ yếu bao gồm thông tin chi tiết về sản phẩm sách, như tiêu đề, thể loại, mô tả ngắn, mô tả chi tiết, nhà xuất bản, giá bán, điểm đánh giá trung bình, và trạng thái tồn kho.

Chính vì thế, hệ thống phù hợp nhất để triển khai chính

### **3.2.2. Cấu hình mô hình Content-Based Filtering**

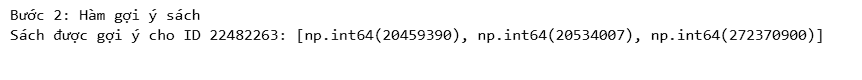
**Bước 1: Tạo ma trận TF-IDF và tính độ tương đồng**



**Figure 11 Tạo ma trận TF-IDF và tính độ tương đồng**

Tạo ma trận TF-IDF và tính độ tương đồng để chuyển đổi dữ liệu văn bản thành dạng số hóa phục vụ mô hình. Cụ thể, tôi sử dụng TF-IDF Vectorizer với giới hạn 5000 đặc trưng và danh sách stop words tiếng Việt để tạo ma trận TF-IDF từ cột combined\_features, sau đó tính toán ma trận độ tương đồng cosine giữa các cuốn sách. Kết quả được kiểm tra thông qua kích thước ma trận TF-IDF, một số từ khóa tiêu biểu và độ tương đồng của ba cuốn sách đầu tiên.

**Bước 2: Hàm gợi ý sách**



**Figure 12 Tạo hàm đề xuất sách**

Trên cơ sở đó, hệ thống xây dựng hàm gợi ý với mục tiêu tìm ra những cuốn sách có nội dung tương tự với một sản phẩm cụ thể mà người dùng đang quan tâm. Tuy nhiên, để đảm bảo chất lượng đề xuất, thuật toán còn áp dụng một số tiêu chí lọc bổ sung. Chỉ những sách hiện đang còn trong kho (inventory\_status là "available") và có điểm đánh giá trung bình từ người dùng lớn hơn hoặc bằng 3.0 mới được xem xét. Nhờ đó, hệ thống không chỉ đưa ra các đề xuất phù hợp về nội dung mà còn đảm bảo các lựa chọn này có sẵn và được đánh giá tích cực từ cộng đồng độc giả.

Khi người dùng lựa chọn một cuốn sách cụ thể, hệ thống sẽ xác định vị trí của sản phẩm đó trong tập dữ liệu, tra cứu các sản phẩm có độ tương đồng cao nhất, sau đó lọc và chọn ra danh sách các cuốn sách được đánh giá cao và có sẵn. Cuối cùng, một danh sách đề xuất được trả về, mang tính cá nhân hóa và có khả năng thỏa mãn sở thích cũng như nhu cầu thực tế của người dùng.

## **3.3. Xây dựng API và tích hợp hệ thống**

# **CHƯƠNG 4: ĐÁNH GIÁ HIỆU SUẤT HỆ THỐNG**

## **4.1. Huấn luyện mô hình AI**

### **4.1.1. Tiêu chí đánh giá mô hình AI**

### **4.1.2. Thời gian phản hồi API**

## **4.2. Cải tiến và tối ưu hóa mô hình**

# **Tổng kết kết quả đạt được**

# **Hạn chế của hệ thống**

# **Hướng phát triển trong tương lai**

# **Tài liệu tham khảo**

<https://viblo.asia/p/tim-hieu-ve-content-based-filtering-phuong-phap-goi-y-dua-theo-noi-dung-phan-1-V3m5WGBg5O7>

<https://www.ibm.com/think/topics/content-based-filtering>

<https://viblo.asia/p/xay-dung-content-based-filtering-rs-recommender-system-co-ban-phan-2-bWrZnVovZxw>

<https://www.ibm.com/think/topics/collaborative-filtering#:~:text=Collaborative%20filtering%20is%20an%20information,have%20interacted%20with%20that%20item>.

[1] Django Software Foundation, “Django Documentation: Models and Databases,” Django Project, 2023.  
[2] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.