

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----------**



**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC**

**ĐỀ TÀI: Xây dựng hệ khuyến nghị sách trong bán hàng trực tuyến**

**sử dụng kỹ thuật lọc dựa trên nội dung**

**và kỹ thuật lọc cộng tác**

**Giảng viên hướng dẫn: Nguyễn Hoàng Anh**

**Sinh viên: Nguyễn Thị Lan Anh**

**Mã sinh viên: B16DCCN010**

**Lớp: D16CNPM1**

**Khóa: 2016-2021**

**Hệ đại học: Đại học chính quy**

**Hà Nội, Tháng 11/2020**

MỤC LỤC

[CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN 4](#_Toc56980512)

[1.1. Giới thiệu hệ khuyến nghị – Recommender System (nêu được lý do chọn đề tài) 4](#_Toc56980513)

[1.2. Giới thiệu hệ khuyến nghị sách (các use case thực tế) 5](#_Toc56980514)

[1.3. Phát biểu bài toán khuyến nghị sách (phân tích bài toán) 7](#_Toc56980515)

[1.4. Khó khăn và thách thức 8](#_Toc56980516)

[1.5. Các đóng góp của đồ án 8](#_Toc56980517)

[1.6. Kết luận chương 8](#_Toc56980518)

[**CHƯƠNG 2: CÁC KỸ THUẬT KHUYẾN NGHỊ SÁCH** 9](#_Toc56980519)

[**2.1.** **Kỹ thuật khuyến nghị lọc dựa trên nội dung** 9](#_Toc56980520)

[2.1.1. Khái niệm 9](#_Toc56980521)

[2.1.2. Phát biểu bài toán tư vấn theo nội dung 9](#_Toc56980522)

[2.1.3. Phương pháp trích chọn đặc trưng 9](#_Toc56980523)

[2.1.4. Xây dựng hồ sơ sản phẩm (Item profiles) 9](#_Toc56980524)

[**2.2.** **Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác** 10](#_Toc56980525)

[2.2.1. Khái niệm 10](#_Toc56980526)

[2.2.2. Phát biểu bài toán lọc cộng tác 12](#_Toc56980527)

[2.2.3. Phương pháp tư vấn lọc cộng tác dựa trên bộ nhớ 13](#_Toc56980528)

[2.2.4. Phương pháp tư vấn dựa lọc cộng tác dựa trên mô hình Matrix Factorization 14](#_Toc56980529)

[CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH 15](#_Toc56980530)

[3.1. Dữ liệu thực nghiệm 15](#_Toc56980531)

[3.2. Xử lý dữ liệu 16](#_Toc56980532)

[3.2.1. Tiền xử lý dữ liệu 16](#_Toc56980533)

[3.2.2. Chia dữ liệu thực nghiệm 16](#_Toc56980534)

[3.3. Thực nghiệm mô hình tư vấn 16](#_Toc56980535)

[3.3.1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc dựa trên nội dung (thực hiện, kết quả, đánh giá) 16](#_Toc56980536)

[3.4. Các thông số đánh giá giải thuật 17](#_Toc56980537)

[3.5. Kết quả thực nghiệm 17](#_Toc56980538)

[3.6. Phân tích lỗi trong quá trình thực nghiệm 17](#_Toc56980539)

[3.7. Kết luận chương 17](#_Toc56980540)

[CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN 18](#_Toc56980541)

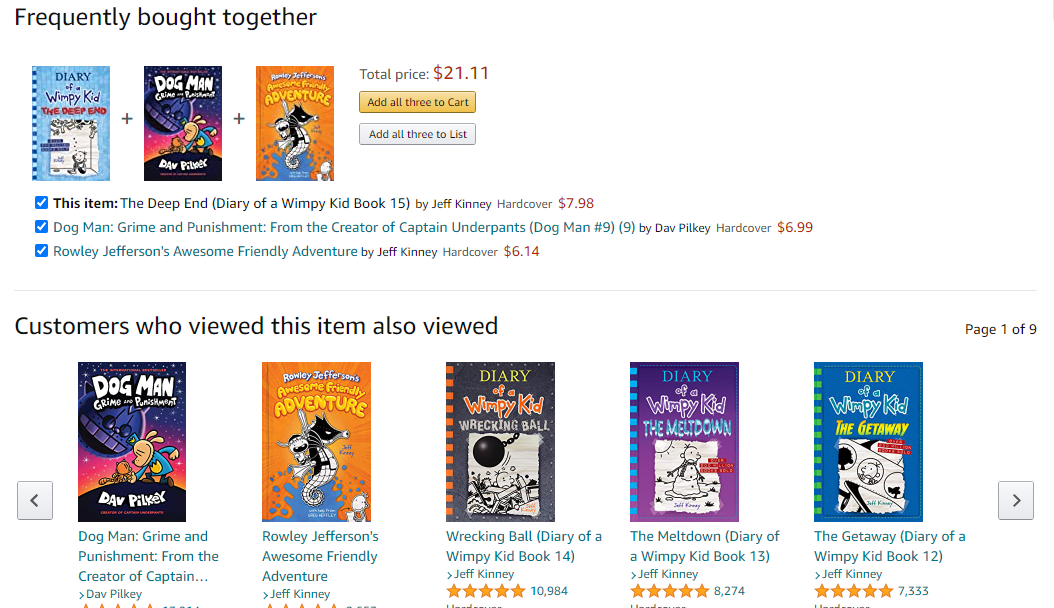
[4.1. Mô tả hệ thống 18](#_Toc56980542)

[4.2. Phân tích thiết kế hệ thống 18](#_Toc56980543)

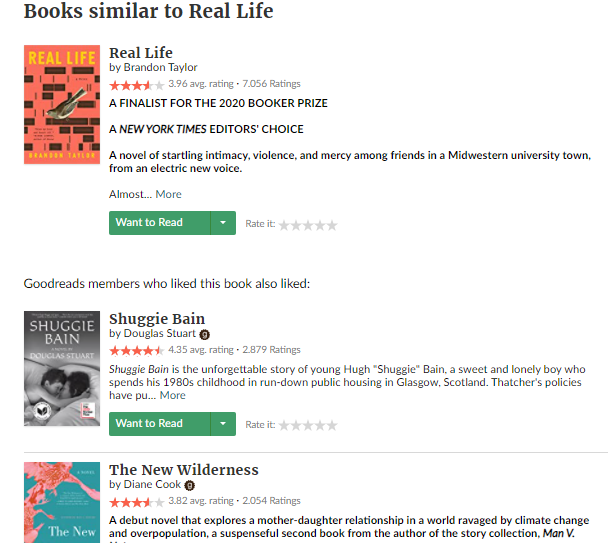
[4.3. Thiết kế hệ thống (công nghệ, API, giao diện) 18](#_Toc56980544)

CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN

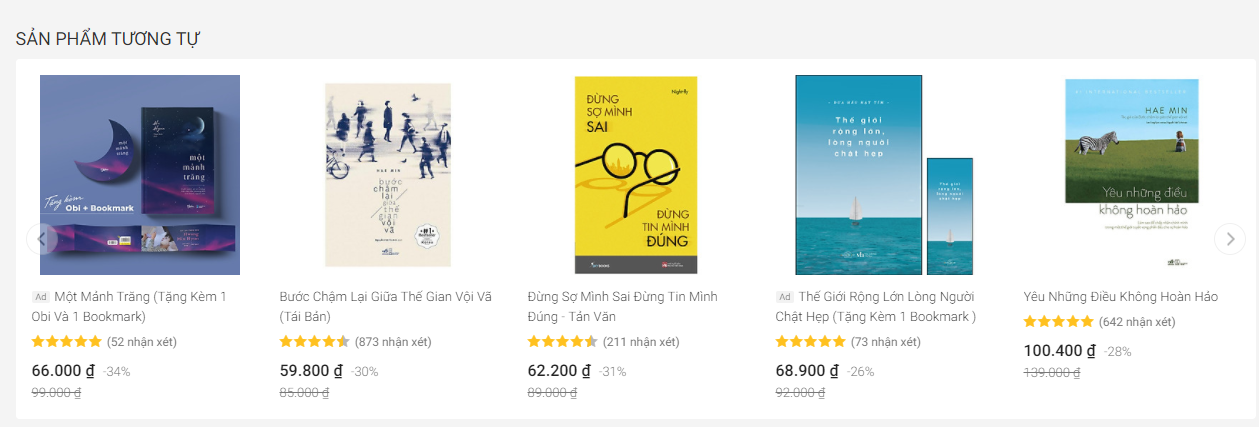
* 1. Giới thiệu hệ khuyến nghị – Recommender System (nêu được lý do chọn đề tài)
* Wikipedia định nghĩa, hệ thống khuyến nghị (Recommender System), hay còn gọi là hệ thống tư vấn là một hệ thống lọc thông tin nhằm dự đoán, đánh giá sở thích, mối quan tâm, nhu cầu của người dùng để đưa ra một hoặc nhiều mục, sản phẩm, dịch vụ mà người dùng có thể sẽ quan tâm với xác suất lớn nhất. Những gợi ý được cung cấp nhằm hỗ trợ cho người sử dụng đưa ra quyết định lựa chọn những sản phẩm, dịch vụ phù hợp với nhu cầu và thị hiếu của mình, chẳng hạn như: mua sản phẩm nào, nghe thể loại nhạc nào, hay tin tức nào nên đọc…
* Chức năng của hệ khuyến nghị:
* Đối với nhà cung cấp:
* Tăng số lượng mặt hàng bán ra cho các hệ thống thương mại điện tử: thay vì người dùng chỉ mua 1 sản phẩm mà họ cần, họ được gợi ý mua những sản phẩm ‘có thể họ cũng quan tâm’ mà bản thân họ không nhận ra. Bằng cách tìm ra những mối quan tâm ẩn của người dùng, hệ thống gợi ý làm gia tăng số lượng mặt hàng bán ra. Tương tự đối với các hệ thống phi thương mại, hệ thống gợi ý sẽ giúp người dùng tiếp cận với nhiều đối tượng thông tin hơn.
* Tăng sự hài lòng của người dùng: vai trò chủ đạo của hệ khuyến nghị là hiểu nhu cầu của người dùng, gợi ý cho họ những thứ họ cần. Họ sẽ tìm thấy các gợi ý thú vị, có hiệu quả, chính xác, gợi ý kịp thời và một giao diện đẹp có thể tối ưu việc sử dụng và tăng sự hài lòng của người dùng trong hệ thống.
* Tăng độ tin cậy, độ trung thực của người dùng: khi người dùng càng tin cậy vào hệ thống, đưa ra những đánh giá trung thực cho các sản phẩm, hệ thống càng mang lại cho người dùng những gợi ý chính xác hơn
* Đối với người sử dụng:
* Tìm ra một số sản phẩm tốt nhất: Hệ thống gợi ý tới người dùng một số sản phẩm được xếp hạng, từ đó người dùng có thể tìm được sản phẩm tốt nhất cho bản thân.
* Gợi ý liên tục: Thay vì tập trung vào gợi ý đơn, các hệ thống khuyến nghị tạo các gợi ý liên tục tới người dùng đến khi họ tìm được sản phẩm ưng ý
* Hệ thống khuyến nghị sẽ đưa ra các gợi ý dựa trên quá trình thu thập, xử lý và phân tích dữ liệu của người dùng. Dữ liệu đó được chia làm hai loại là phản hồi tường minh (explicit feedback) bằng cách yêu cầu người dùng phản hồi trực tiếp (đánh giá yêu thích, xếp hạng số sao, hành động thêm vào giỏ hàng,…) và phản hồi tiềm ẩn (implicit feedback) bằng cách tự động suy luận dựa trên những tương tác của người dùng với hệ thống (số lần nhấp chuột vào sản phẩm, thời gian quan sát…).
  1. Giới thiệu hệ khuyến nghị sách (các use case thực tế)
* Các trang web thương mại nói chung và website bán sách nói riêng hiện nay đều đã tích hợp hệ gợi ý.
* Quốc tế:
* Amazon.com



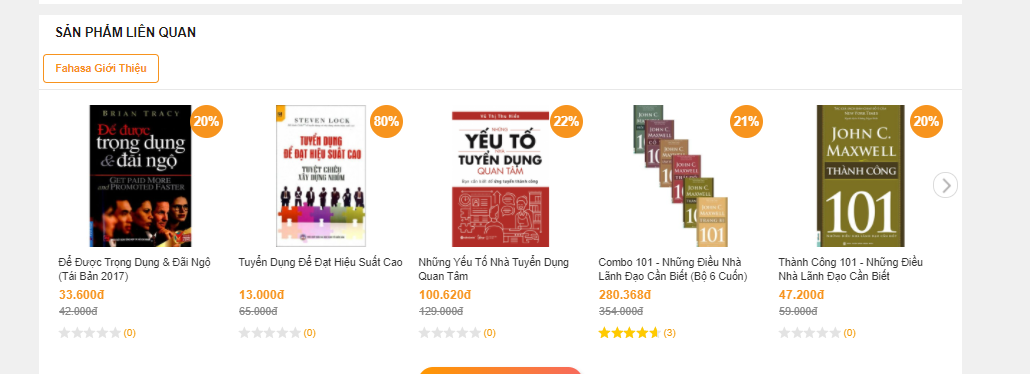
* Amazon được coi là 1 trong những hệ khuyến nghị tiên phong
* Goodreads.com



* Tại Việt Nam:
* Tiki.vn



* Fahasa.com



* Nhamam.com.vn



* Lí do chọn đề tài:
  1. Phát biểu bài toán khuyến nghị sách (phân tích bài toán)
* Một cách hình thức, bài toán khuyến nghị hay bài toán tư vấn được phát biểu như sau:
* Gọi U là tập tất cả người dùng, I là tập tất cả các sản phẩm có thể tư vấn. Tập I có thể rất lớn, từ hàng trăm ngàn đến hàng triệu bản ghi.
* Hàm r(u, i) đo độ phù hợp hay hạng của sản phẩm i với người dùng u. Với mỗi người dùng u thuộc U, cần tìm sản phẩm i’ sao cho hàm r(u, i’) đạt giá trị lớn nhất:
* Trong hệ thống gợi ý, độ phù hợp của một sản phẩm thường được đo bằng điểm, ví dụ người dùng A đánh giá cuốn sách B được 4/5 sao. Tuy nhiên, độ phù hợp có thể là một hàm bất kỳ phụ thuộc vào ứng dụng cụ thể. Giá trị của hàm r có thể được xác định bởi người dùng hoặc tính toán bằng một công thức nào đó.
* Mỗi người dùng trong không gian U được xác định bởi 1 hồ sơ (profile). Hồ sơ này bao gồm thông tin người dùng (id, tên). Tương tự mỗi sản phẩm s trong không gian I cũng được xác định bởi 1 tập đặc trưng. Trong hệ thống gợi ý sách, đặc trưng của tập I có thể là: id, tên sách, mô tả nội dung, giá bán,….
* Quá trình xây dựng hệ gợi ý bao gồm các bước sau:
* Tập hợp thông tin (Information Recollection)
* Lựa chọn, chuyển đổi, cấu trúc hóa thông tin
* Biểu diễn thông tin
  1. Khó khăn và thách thức
  2. Các đóng góp của đồ án
  3. Kết luận chương

# **CHƯƠNG 2: CÁC KỸ THUẬT KHUYẾN NGHỊ SÁCH**

## **Kỹ thuật khuyến nghị lọc dựa trên nội dung**

### Khái niệm

* Lọc dựa trên nội dung (Content-based filtering) là phương pháp dựa trên những đặc điểm chính của item. Nói cách khác, thuật toán cho phương pháp này cố gắng gợi ý những item có những điểm giống với những item mà người dùng đã từng quan tâm trong quá khứ. Gợi ý dựa trên nội dung khai thác những khía cạnh có liên quan đến nội dung thông tin sản phẩm người dùng đã từng sử dụng hay truy cập trong quá khứ để tạo nên gợi ý hay nói cách khác gợi ý trên nội dung được dựa trên mô tả có sẵn về đối tượng. Các phương pháp tiếp cận cho lọc nội dung có nguồn gốc từ lĩnh vực truy vấn thông tin, trong đó mỗi sản phẩm được biểu diễn bằng một hồ sơ người dùng.
* Hướng tiếp cận dựa trên nội dung bắt nguồn từ những nghiên cứu về thu thập thông tin và lọc thông tin. Những tiến bộ so với hướng tiếp cận cũ của thu thập thông tin là do việc sử dụng hồ sơ về người dùng. Hồ sơ này được xây dựng dựa trên những thông tin được người dùng cung cấp trực tiếp, hoặc gián tiếp
* Ưu điểm:
* Mô hình không cần bất cứ dữ liệu gì về người dùng khác, gợi ý riêng theo từng đối tượng người dùng. Do vậy, mô hình không yêu cầu số lượng người dùng lớn để đạt được độ chính xác đề nghị hợp lý
* Mô hình có thể nắm bắt được sở thích cụ thể của người dùng và có thể đề xuất các mặt hàng thích hợp mà rất ít người dùng khác quan tâm
* Nhược điểm:
* Biểu diễn đặc trưng của sản phẩm đòi hỏi các kiến thức liên quan về lĩnh vực sản phẩm được gợi ý, đòi hỏi chuyên gia phân tích và mang tính thủ công.
* Mô hình chỉ có thể đưa ra đề xuất dựa trên sở thích hiện có của người dùng. Nói cách khác mô hình có hạn chế trong việc hướng người dùng đến những sở thích mới. Thông thường những hệ thống gợi ý dựa trên nội dung đề xuất những đối tượng tương tự mà người dùng đã đánh giá trước đó. Trong một số trường hợp đặc biệt, đối tượng không nên được gợi ý vì chúng có độ tương tự gần như là tuyệt đối với sản phẩm người dùng vừa xem. Khi đó người dùng sẽ không quan tâm đến những sản phẩm đó nữa.
* Vấn đề người dùng mới, khi đó họ chưa cung cấp đủ dữ liệu để mô hình dự đoán, mô hình gặp vấn đề trong việc đưa ra gợi ý.

### Phát biểu bài toán tư vấn theo nội dung

* Cho là tập gồm n sản phẩm. Nội dung sản phẩm được ký hiệu là được biểu diễn thông qua tập K đặc trưng nội dung của I. Tập các đặc trưng sản phẩm I được xây dựng bằng các kỹ thuật truy vấn thông tin để thực hiện mục đích dự đoán những sản phẩm khác tương tự với
* Cho là tập gồm m người dùng. Với mỗi người dùng , gọi là hồ sơ người dùng u. Hồ sơ của người dùng u thực chất là lịch sử truy cập hoặc đánh giá của người đó đối với các sản phẩm mà người dùng u đã từng truy nhập hoặc đánh giá.
* Bài toán lọc theo nội dung khi đó là dự đoán những sản phẩm có nội dung thích hợp với người dùng dựa trên tập hồ sơ sản phẩm và hồ sơ người dùng

### Phương pháp trích chọn đặc trưng

* Phương pháp trích chọn đặc trưng hay còn gọi là biểu diễn hồ sơ sản phẩm. Để có một tập các đặc trưng đầy đủ, nội dung tài liệu phải được biểu diễn dưới dạng phù hợp để máy tính có thể tự động phân tích, tính toán trọng số các đặc trưng nội dung hoặc phải được thực hiện bán tự động.
* Phương pháp thường được sử dụng là phép đo tần suất kết hợp với tần suất xuất hiện ngược TF-IDF (Term Frequency/ Inverse Document Frequency)

### Xây dựng hồ sơ sản phẩm (Item profiles)

## Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác

### Khái niệm

* Không giống với phương pháp lọc theo nội dung, phương pháp lọc cộng tác gợi ý dựa trên sự tương quan (similarity) giữa các người dùng và/hoặc sản phẩm. Hệ thống sẽ so sánh, tính toán độ tương tự nhay giữa những người dùng hay mặt hàng, từ đó người dùng sẽ được gợi ý những thông tin, mặt hàng được ưa chuộng nhất bởi những người dùng có cùng sở thích. Trong phương pháp này, hệ thống thường xây dựng các ma trận đánh giá bởi người dùng lên các mặt hàng, bản tin. Từ đó tính toán độ tương tự giữa họ. Các hệ gợi ý dựa trên lọc cộng tác không yêu cầu quá nặng vào việc tính toán, do đó nó có thể đưa ra những gợi ý có độ chính xác cao và nhanh chóng cho một số lượng lớn người dùng. Hơn nữa, hệ gợi ý này không yêu cầu mô tả nội dung tường minh mà chỉ sự dụng đánh giá của người dùng để ước lượng, do đó những hệ này có khả năng gợi ý phong phú và thường tạo ra những gợi ý bất ngờ cho người dùng. Với phương pháp này, sở thích của người dùng là đầu vào duy nhất để quyết định kết quả gợi ý.
* Hệ thống lọc cộng tác dự đoán độ phù hợp của một sản phẩm i với người dùng u dựa trên độ phù hợp giữa người dùng và i, trong đó là người có cùng sở thích với u. Ví dụ, để gợi ý một quyển sách cho người dùng c, đầu tiên hệ thống cộng tác tìm những người dùng khác có cùng sở thích đọc sách với c. Sau đó, những cuốn sách được họ đánh giá cao sẽ được dùng để gợi ý cho c
* Đầu vào của bài toán là ma trận thể hiện những hành vi quá khứ, gọi là ma trận người dùng – sản phẩm. Các hàng trong ma trận đại diện cho người dùng, các cột đại diện cho sản phẩm, giá trị mỗi ô là đánh giá của người dùng lên sản phẩm đó.
* Tùy theo hệ thống mà đánh giá của người dùng được quy ước những giá trị nào. Trong ví dụ này, các đánh giá có giá trị từ 1->5.

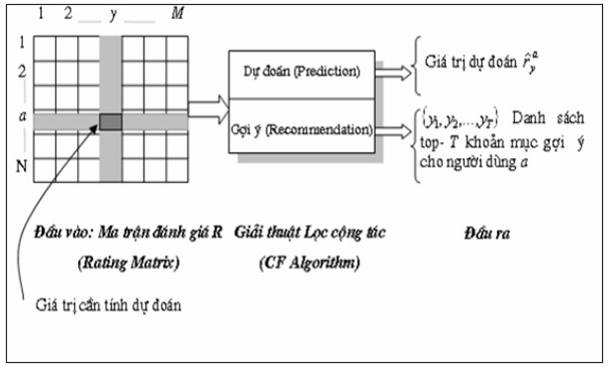
|  |  |  |  |
| --- | --- | --- | --- |
|  | Sản phẩm 1 | Sản phẩm 2 | Sản phẩm 3 |
| Người dùng 1 | 1 | 0 | 5 |
| Người dùng 2 | 4 | 2 | 2 |
| Người dùng 3 | 0 | 0 | 0 |

* Ở ma trận này, đánh giá của người dùng 1 đối với sản phẩm 1 là 1, sản phẩm 3 là 5, sản phẩm 2 chưa được đánh giá
* Đầu ra của bài toán là: đánh giá của người dùng lên những sản phẩm mà họ chưa đánh giá. Hệ thống gợi ý dựa trên các đánh giá này mà xếp hạng các sản phẩm và gợi ý cho người dùng.
* Trong ví dụ trên, hệ thống gợi ý phải đưa ra dự đoán: người dùng 1 đánh giá sản phẩm 2 là bao nhiêu, người dùng 3 đánh giá sản phẩm 1, 2, 3 là bao nhiêu.
* **Ưu điểm**
* Không giới hạn về loại đối tượng dùng để gợi ý: phương pháp này dựa hoàn toàn vào đánh giá của những người dùng để đưa ra các nhận định về sở thích của người dùng, chính vì thế các tính chất của đối tượng được gợi ý không có ảnh hưởng đến quá trình gợi ý
* Gợi ý đa dạng: Khắc phục được giới hạn của phương pháp tiếp cận dựa trên nội dung, phương pháp lọc cộng tác có thể đưa ra các đối tượng sản phẩm khuyến nghị hoàn toàn khác so với các sản phẩm mà người dùng u đã thích trong quá khứ.
* **Nhược điểm**
* Vấn đề người dùng mới: để phân bổ chính xác sản phẩm người dùng quan tâm, lọc cộng tác phải ước lượng được sở thích của người dùng đối với các sản phẩm thông qua những đánh giá của họ trong quá khứ. Trong trường hợp một người dùng mới, số đánh giá của người dùng cho các sản phẩm là 0. Khi đó, phương pháp lọc cộng tác khó để đưa ra những tư vấn chính xác cho người dùng này.
* Vấn đề sản phẩm mới: trong lọc thông tin, các sản phẩm thường xuyên được bổ sung, cập nhật vào hệ thống. Khi xuất hiện một sản phẩm mới, tất cả đánh giá người dùng cho sản phẩm này đều là 0. Do đó, lọc cộng tác không thể tư vấn sản phẩm cho bất kỳ người dùng nào trong hệ thống.
* Vấn đề dữ liệu thưa: Kết quả dự đoán của lọc cộng tác phụ thuộc chủ yếu vào số các đánh giá có trước của người dùng đối với các sản phẩm. Tuy nhiên, đối với các hệ thống thực tế, số lượng người dùng và sản phẩm là rất lớn, số những đánh giá biết trước thường rất nhỏ so với số lượng các đánh giá cần được dự đoán.

### Phát biểu bài toán lọc cộng tác

* Ký hiệu là tập gồm m người dùng, là tập gồm n sản phẩm mà người dùng có thể lựa chọn.
* Ký hiệu R = . Trong đó mỗi người dùng đưa ra đánh giá của mình cho một số sản phẩm bằng một số . Giá trị phản ánh mức độ ưa thích của người dùng đối với sản phẩm , giá trị có thể được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị trong trường hợp người dùng chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm
* Với một người dùng cần được gợi ý (được gọi là người dùng hiện thời, người dùng cần được tư vấn, hay người dùng tích cực), bài toán lọc cộng tác là bài toán dự đoán đánh giá của đối với mặt hàng mà chưa đánh giá (), trên cơ sở đó gợi ý cho những sản phẩm được đánh giá cao.
* Bảng dưới thể hiện một ví dụ với ma trận đánh giá R trong hệ gồm 5 người dùng và 4 sản phẩm . Mỗi người dùng đều đưa ra ác đánh giá của mình về các sản phẩm theo thang bậc {0, 1, 2, 3, 4, 5}. Giá trị được hiểu là người dùng chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm . Giá trị là sản phẩm hệ thống cần dự đoán cho người dùng

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 2 | 1 | 3 | 5 |
|  | 4 | 2 | 1 | 0 |
|  | 3 | 0 | 2 | 4 |
|  | 4 | 4 | 0 | 0 |
|  | 4 | ? | 5 | 5 |



* Ma trận đánh giá R là thông tin đầu vào duy nhất của phương pháp lọc cộng tác. Dựa trên ma trận đánh giá, các phương pháp lọc cộng tác thực hiện hai tác vụ. Dựa trên ma trận đánh giá, các phương pháp lọc cộng tác thực hiện hai tác vụ: Dự đoán quan điểm của người dùng hiện thời (active user) về các sản phẩm mà họ chưa đánh giá, đồng thời đưa ra một danh sách các sản phẩm có đánh giá cao nhất phân bổ cho người dùng hiện thời.
* Phương pháp lọc cộng tác bao gồm các kỹ thuật như kỹ thuật láng giềng, kỹ thuật mạng Bayes, mạng neuron kết hợp SVD…
* Lọc cộng tác tiếp cận theo hai xu hướng chính: lọc cộng tác dựa trên bộ nhớ và lọc cộng tác dựa trên mô hình. Mỗi phương pháp tiếp cận có những ưu điểm và hạn chế riêng, khai thác các mỗi liên hệ trên ma trận đánh giá người dùng.

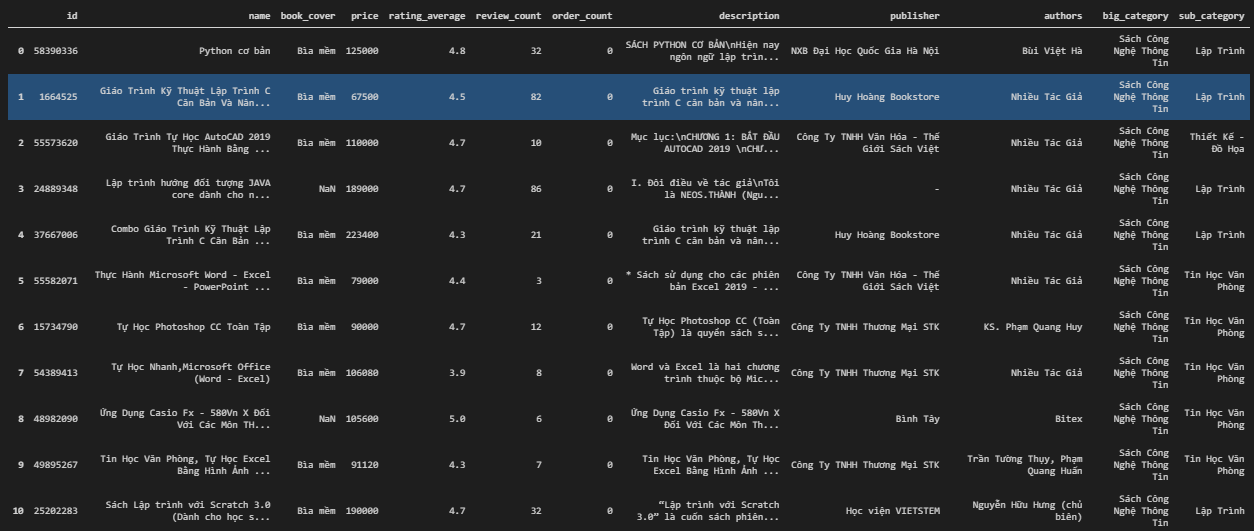
### Phương pháp tư vấn lọc cộng tác dựa trên bộ nhớ

* Các phương pháp lọc dựa trên bộ nhớ sử dụng toàn bộ ma trận đánh giá để sinh ra dự đoán các sản phẩm cho người dùng hiện thời. Phương pháp thực hiện theo hai bước: Tính toán mức độ tương tự và bước tạo nên dự đoán
* Tính toán độ tương tự sim(x,y): mô tả khoảng cách, sự liên quan, hay trọng số giữa hai người dùng x và y hoặc giữa hai sản phẩm x và y
* Dự đoán: đưa ra dự đoán cho người dùng cần được tư vấn bằng cách xác định tập láng giềng của người dùng này. Tập láng giềng của người dùng cần tư vấn được xác định dựa trên mức độ tương tự giữa các cặp người dùng hoặc sản phẩm.
* Các phương pháp tính toán mức độ tương tự: có nhiều phương pháp khác nhau tính toán mức độ tương tự sim(x,y) giữa các cặp người dùng hoặc các cặp sản phẩm. Hai phương pháp phổ biến được sử dụng là độ tương quan Pearson và giá trị cosin giữa hai vector (trình bày kỹ công thức ở dưới)
* Phương pháp dự đoán mức độ thích hợp của sản phẩm p chưa được người dùng u đánh giá được tính toán dựa trên tập những người dùng khác đã đánh giá p. Gọi là tập N người dùng tương tự nhất đối với u. Khi đó, mức độ phù hợp của người dùng u đối với sản phẩm mới p được xác định như một hàm các đánh giá của tập láng giềng. Dưới đây là một số phương pháp thông dụng nhất để dự đoán mức độ phù hợp của sản phẩm p đối với người dùng u. (trình bày ở dưới)

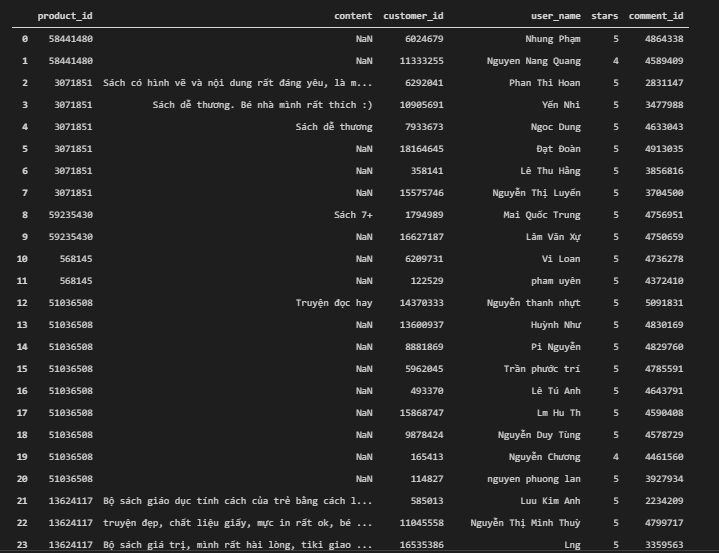
### Phương pháp tư vấn dựa lọc cộng tác dựa trên mô hình Matrix Factorization

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH

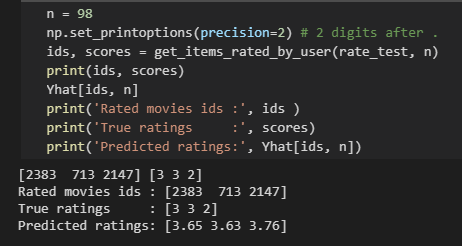
* 1. **Dữ liệu thực nghiệm**
* Dữ liệu thực nghiệm bao gồm 2 tập dữ liệu: 1 bộ dữ liệu thông tin sách và 1 bộ dữ liệu lưu trữ đánh giá bình luận của người dùng cho sách
* Dữ liệu sách gồm 10330 bản ghi sách, gồm các đặc trưng: id, tên, loại bìa, mô tả, giá bán, đánh giá trung bình, thể loại, thể loại con, tác giả, nhà xuất bản, số lượt bán



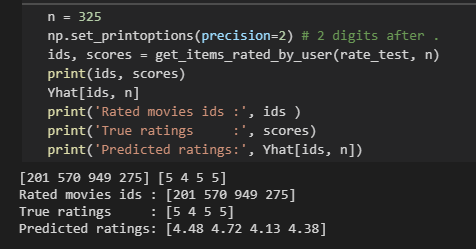
* Dữ liệu tương tác giữa người dùng và sách bao gồm 20849 bản ghi, tối thiểu số lượt đánh giá của 1 người là 10 để kết quả gợi ý chính xác hơn.



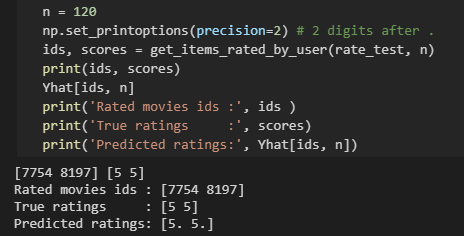
* 1. **Xử lý dữ liệu**
     1. Tiền xử lý dữ liệu
* Đối với dữ liệu dạng text tiến hành xử lý, ví dụ một số trường trong dữ liệu sách như: phần mô tả sách, còn tồn tại các ký tự đặc biệt, thẻ html, từ dùng
* Làm sạch dữ liệu bằng cách loại bỏ một số dấu câu, kí tự không cần thiết (trừ dấu hỏi chấm), loại bỏ thẻ html
* Chuẩn hoá từ: Chuẩn hoá dạng ký tự viết hoa về ký tự viết thường.
* Loại bỏ từ dừng
  + 1. Chia dữ liệu thực nghiệm
* Đối với bộ dữ liệu rating, chia dữ liệu thành 2 phần: 1 phần để huấn luyện mô hình (80%) và 1 phần để đánh giá (20%)
  1. **Thực nghiệm mô hình tư vấn**
     1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc dựa trên nội dung (thực hiện, kết quả, đánh giá)
* Đặc trưng: TFIDF trên trường mô tả của tập dữ liệu sách làm item profile (thư viện sklearn)
* Với mỗi user, đi tìm những cuốn sách mà user đó đã đánh giá, và giá trị các đánh giá đó
* Thuật toán sử dụng để tìm mô hình cho mỗi user: linear model (Ridge). Từ đó ta tìm các hệ số của mô hình -> từ đó tính được ratings của từng user cho mỗi item
* Tiếp tục tối ưu các tham số của mô hình và thử nghiệm với các mô hình khác
* Ví dụ userid = 98



* Ví dụ userid = 325



* Ví dụ userid = 120



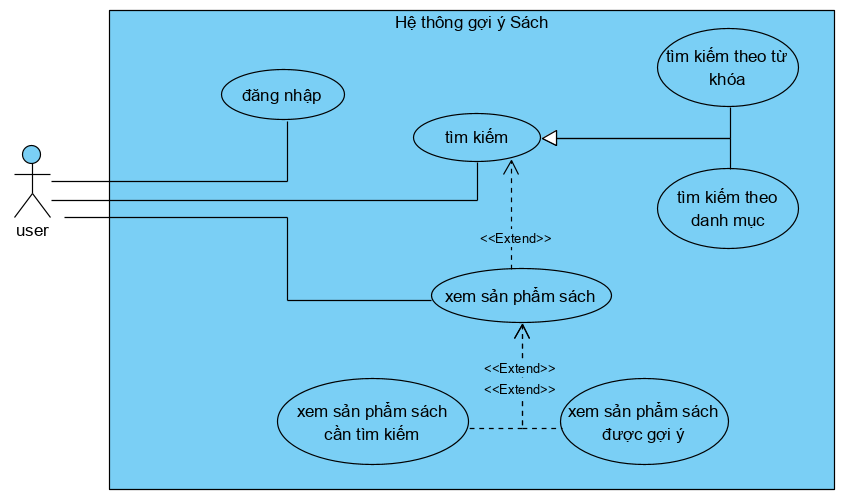
* Tổng quát: đo RMSE trên tập training và tập test

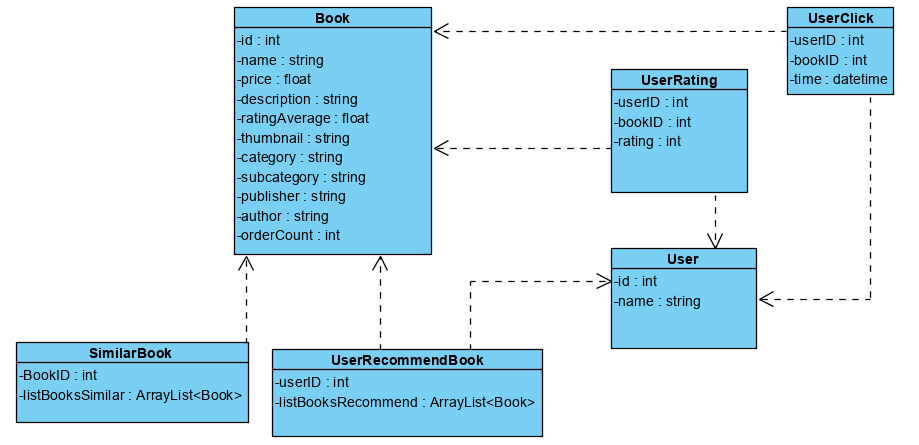


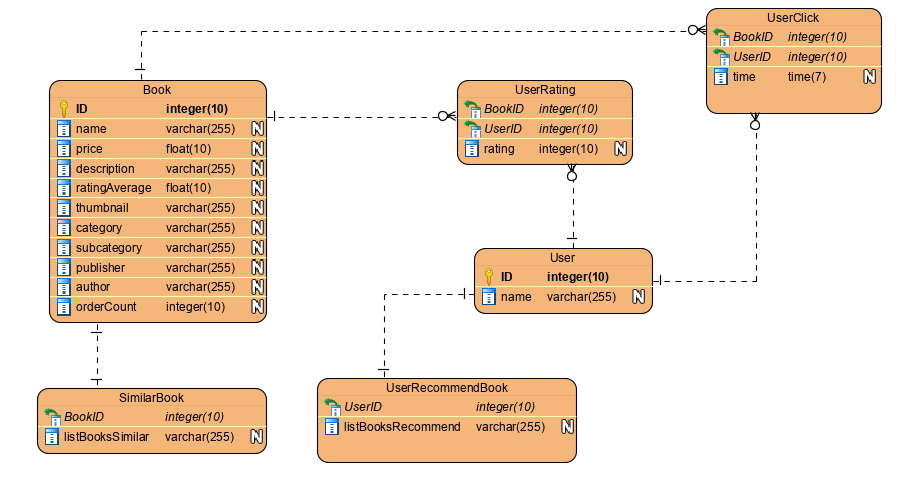
* + 1. Thực nghiệm mô hình tư vấn sách theo kỹ thuật lọc cộng tác (thực hiện, kết quả, đánh giá)
  1. **Các thông số đánh giá giải thuật**
  2. **Kết quả thực nghiệm**
  3. **Phân tích lỗi trong quá trình thực nghiệm**
  4. **Kết luận chương**

CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN

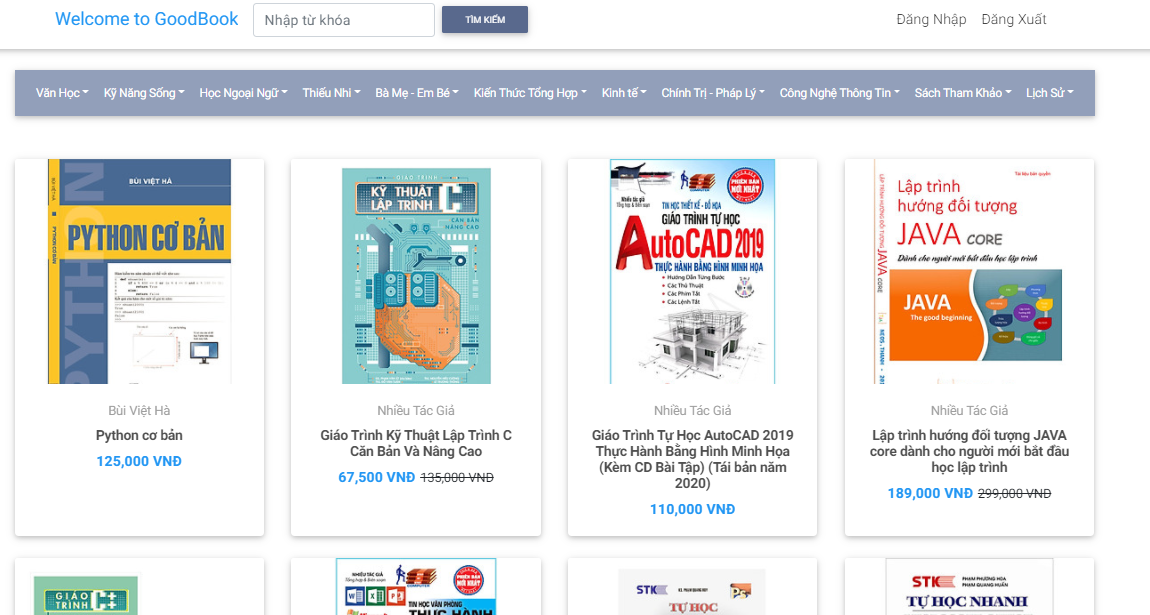
* 1. Mô tả hệ thống
* Trang web được xây dựng với mục đích tìm hiểu, nghiên cứu hoạt động của hệ khuyến nghị
* Hệ thống giới thiệu và gợi ý sách cho phép khách hàng bất kỳ có thể tìm kiếm hoặc xem các sản phẩm theo thể loại, danh sách sách được ưa thích nhất …
* Hệ thống cho phép người dùng đăng nhập qua Facebook
* Trong quá trình xem sản phẩm, hệ thống sẽ gợi ý cho người dùng các loại sách trong quá trình chọn sản phẩm sử dụng kỹ thuật lọc theo nội dung và kỹ thuật lọc cộng tác, hiển thị các cuốn sách tương tự với cuốn sách mà người dùng đang xem sử dụng các thuộc tính tương tự về thể loại sách của sách đó.
  1. Phân tích thiết kế hệ thống



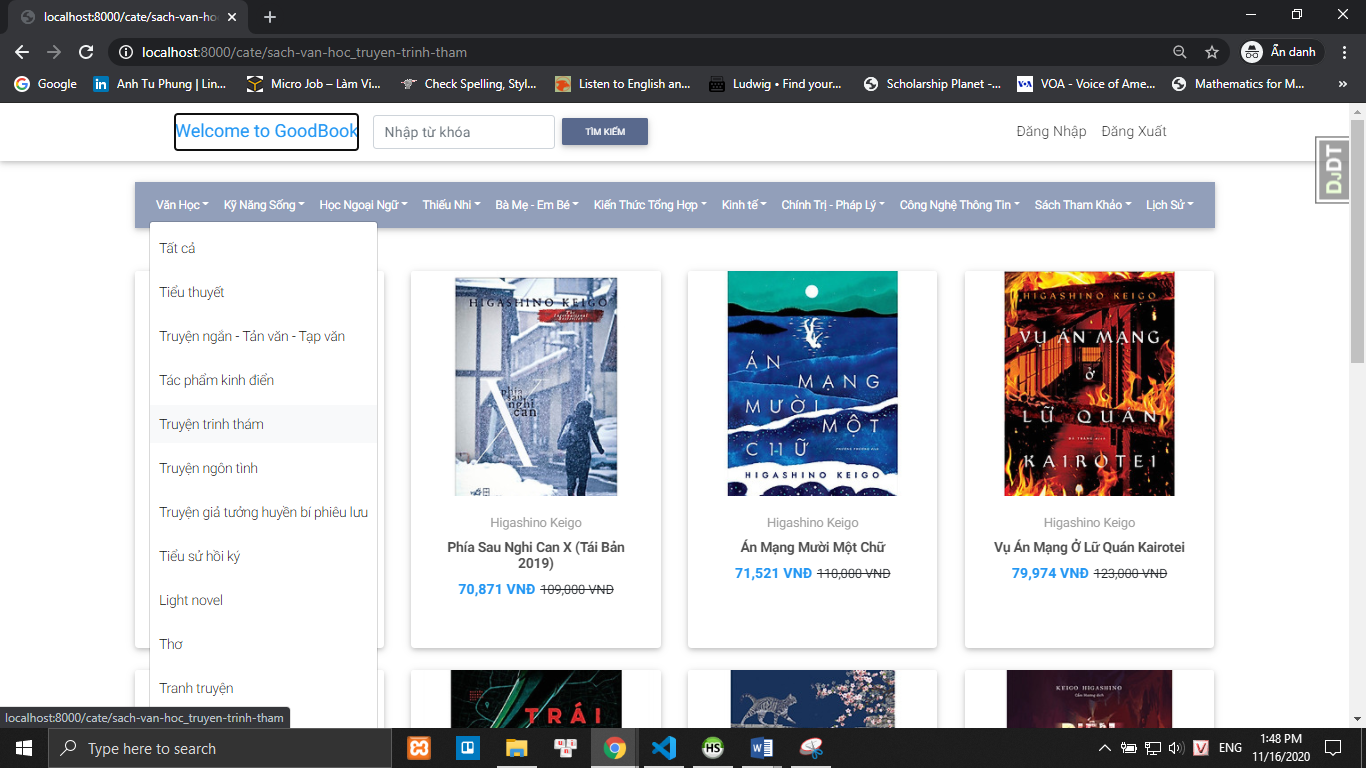




* 1. Thiết kế hệ thống (công nghệ, API, giao diện)
     1. Các công nghệ sử dụng
* Ngôn ngữ lập trình thuật toán: Python
* Ngôn ngữ lập trình server: framework Django
* Giao diện: HTML, Bootstrap, CSS, Javascripts
* Database: SQLite
  + 1. Kiến trúc hệ thống
    2. Triển khai ứng dụng thực tế
* Một số giao diện minh họa của hệ thống
* Trang chủ (mục đích hiển thị top bán chạy và gợi ý theo lọc cộng tác – tuy nhiên phần gợi ý lọc cộng tác vẫn còn chưa hoàn thiện)



* Trang hiển thị sách theo phân loại.



* Trang chi tiết sản phẩm: bao gồm thông tin chi tiết về sản phẩm và KẾT HỢP GỢI Ý THEO NỘI DUNG (gợi ý các sản phẩm có chỉ số cosine similarity theo 2 tiêu chí: Tiêu đề và Tóm tắt truyện lớn nhất)

