ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN

Đề tài: Xây dựng hệ khuyến nghị sách Tiki

GVHD: ThS. Đỗ Trọng Hợp
Lớp: IE212.M11

Nhóm sinh viên thực hiện: Nhóm 21
1. Bùi Quang Huy - 18520063

□ Tp. Hồ Chí Minh, 1/2022 □

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

……, ngày…….....tháng……năm 2022

Người nhận xét
(Ký tên và ghi rõ họ tên)

MỤC LỤC

I. ĐẶT VẤN ĐỀ:

Với sự bùng nổ mạnh mẽ của internet thì việc mua sắm, tìm kiếm sản phẩm, dịch vụ,.. qua các nền tảng trực truyến trở nên thông dụng. Trong tình hình dịch bệnh khi các dịch vụ giải trí offline ngừng hoạt động thì nhu cầu tìm kiếm, mua sách để phụ vụ cho học tập và giải trí càng tăng mạnh. Để giúp người dùng dễ dàng tìm kiếm và tiết kiệm thời gian khi mua sách cũng như giúp doanh nghiệp tăng doanh thu, cải thiện khả năng cạnh tranh thì việc nghiên cứu và xây dựng một hệ thống khuyến nghị sách là điều đáng được quan tâm.

Khi người dùng vào website cung cấp sách/tìm kiếm/xem 1 cuốn sách, hệ thống khuyến nghị sẽ trả về một danh sách ngắn(top-N) các quyển sách mà người dùng nhiều khả năng sẽ chọn, có thể bao gồm cả những quyển sách mà trước đó người mua sắm không biết. Hệ thống khuyến nghị sẽ chủ động đưa ra các dự đoán mà không cần người dùng yêu cầu, giúp tiết kiệm thời gian, tăng tốc độ tìm kiếm và giúp người dùng truy cập tới nội dung họ quan tâm một cách dễ dàng hơn, đồng thời, gợi ý tới người dùng những đề xuất mới mà trước đây họ chưa từng biết đến. Các sản phẩm khuyến nghị đưa ra dựa trên việc phân tích, tính tóan theo thuật tóan lọc cộng tác và lọc nội dung(khai thác thông tin, sở thích, lịch sử đánh giá, mua sắm, xem, tìm kiếm của người dùng).

Hệ thống khuyến nghị cũng giúp các nhà sách giới thiệu sản phẩm tới người mua, giúp gia tăng doanh số nhờ các ưu đãi, sản phẩm, dịch vụ được khuyến nghị một cách cá nhân hóa, làm nâng cao trải nghiệm của người mua,giúp việc mua sắm trở nên nhanh chóng, tăng khả năng giữ chân khách hàng, tăng doanh thu cho nhà sách.

II. BÀI TOÁN KHUYẾN NGHỊ:

Khuyến nghị sách cho người dùng: bài tóan có đầu vào là tập 1 hay nhiều người dùng và tập những quyển sách mà hệ thống quan sát được cùng với lịch sử đánh giá của người dùng. Hệ thống sẽ trả về danh sách các quyển sách tiềm năng mà người dùng có thể quan tâm, yêu thích ứng với một người dùng.

Đầu vào: • U = {u1, u2, u3, ..., un}: không gian người dùng. • P = {p1, p2, p3, ..., pm}: không gian sách. • R = {r1, r2, r3, ..., rk}: lịch sử đánh giá của người dùng.
Đầu ra: PTopN∈ khuyên nghị cho mỗi u∈ U $\forall pk \in PTopN, f(u, pk) \geq f(u, pk+1), với\ 1 \leq k \leq TopN - 1.$

Định nghĩa 1: Không gian người mua sắm (người dùng):

Không gian người mua sắm là tập tất cả những người dùng mà hệ thống cần thực hiện các phân tích, khuyến nghị. Ký hiệu là U, U = {u1, u2, u3, ..., un}.

Định nghĩa 2: Không gian sách (đối tượng khuyến nghị):

Không gian sách là tập tất cả những quyển sách mà hệ thống quan sát được. Tập những quyển sách này sẽ được dùng để phân tích, tinh tóan mức độ phù hợp, hữu ích tới quan tâm của mỗi người và khuyến nghị cho người dùng. Ký hiệu là P, P = {p1, p2, p3, ..., pm}.

Định nghĩa 3: Hàm hữu ích

Hàm hữu ích f là ánh xạ f: U x P ◊ R, dùng để ước lượng mức độ hữu ích của p∈P với u∈U. Với R là tập có thứ tự các số nguyên hoặc thực trong một khoảng nhất định.

Cho trước:

- U = {u1, u2, u3, ..., un}: không gian người dùng.
- P = {p1, p2, p3, ..., pm}: không gian sách.

Mục đích của hệ khuyến nghị là đi tìm hàm hữu ích f, ước lượng giá trị của f(u,p) (với u∈U, p∈P).Giá trị của f(u,p) giúp tiên đoán u sẽ thích quyển sách p nhiều hay ít, hay quyển sách p hữu ích đối với người dùng u như thế nào.

Đối với mỗi người dùng u∈U, hệ khuyến nghị cần chọn TopN quyển sách p∈P hữu ích nhất đối với người dùng u để khuyến nghị. Việc chọn TopN bao nhiêu là tùy thuộc vào nhu cầu thông tin của người dùng, cũng như mục đích cung cấp thông tin của hệ khuyến nghị.

Các quyển sách p∈PTopN, được chọn thỏa mãn các điều kiện ràng buộc sau:

1. $\forall pk \in PTopN, f(u, pk) \geq f(u, pk+1), với\ 1 \leq k \leq TopN - 1.$

   Tập sách khuyến nghị PTopN đã được sắp xếp có thứ tự.

   Những quyển sách đứng trước có giá trị của hữu ích f lớn hơn hoặc bằng những quyển sau, và được ưu tiên khuyến nghị hơn.

2. $\forall pk \in PTopN, \forall pi \in P \backslash PTopN, thì\ f(u, pk) \geq f(u, pi).$

   Giá trị hữu ích của các quyển sách được khuyến nghị, được xác định thông qua hàm f, phải lớn hơn hoặc bằng những quyển không được khuyến nghị.

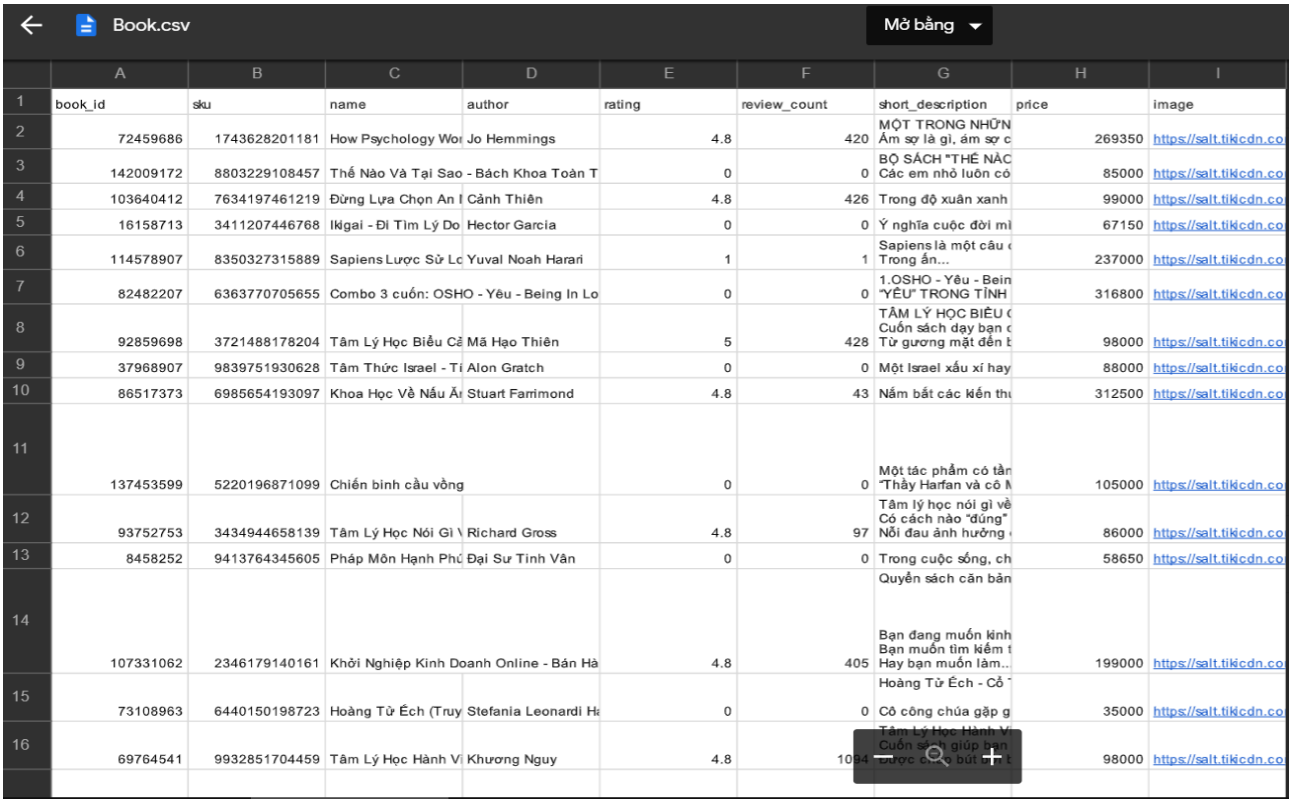
   Tùy vào phương pháp sử dụng ta có nhiều cách xây dựng hàm hữu ích f khác nhau.

III. DATASET:

Tập dữ liệu được sử dụng trong đồ án là tập dữ liệu về người dùng đã mua và đánh giá các quyển sách trong mục sách Tiếng Việt của trang web Tiki. Dữ liệu được

crawl và xử lý vào ngày 5/12/2021 từ địa chỉ https://tiki.vn/sach-truyen-tieng-viet/c316. Tập dữ liệu này bao gồm 3 file csv: Book.csv, Rating.csv, và User.csv.

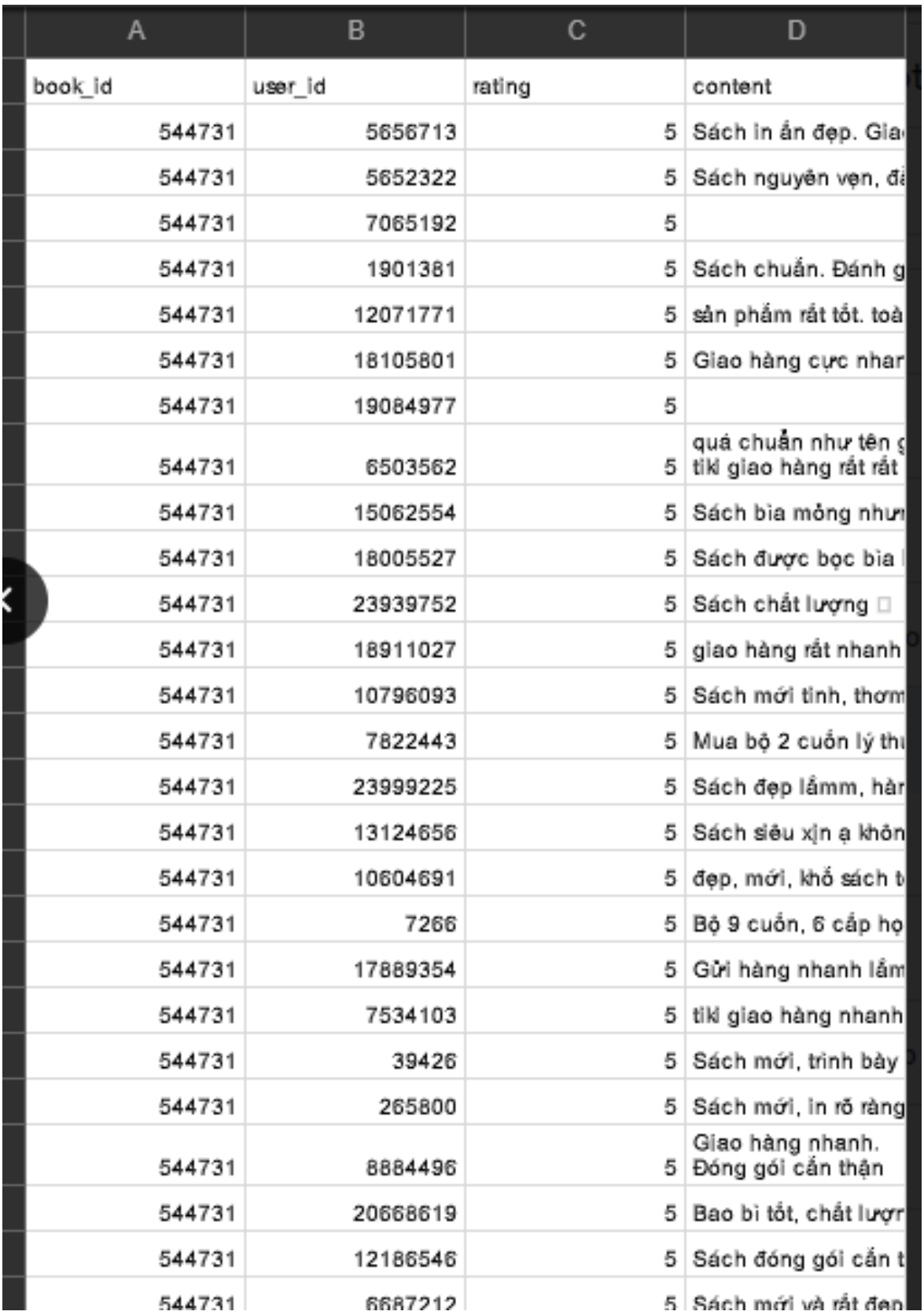
+ Book.csv: gồm 6000 dòng dữ liệu tương ứng với 6000 quyển sách với các thông tin sách được lưu trữ có 9 trường: id sách, mã sku, tên sách, tác giả, đánh giá trung bình, số lượt đánh giá, tóm tắt, giá cả và hình ảnh minh họa.



	A	B	C	D	E	F	G	H	I
1	book_id	sku	name	author	rating	review_count	short_description	price	image
2	72459686	1743628201181	How Psychology Wor	Jo Hemmings	4.8	420	MỘT TRONG NHỮN Âm sợ là gì, âm sợ c	269350	https://salt.tikicdn.co
3	142009172	8803229108457	Thế Nào Và Tại Sao - Bách Khoa Toàn T		0	0	BỘ SÁCH "THẾ NÀO Các em nhỏ luôn có	85000	https://salt.tikicdn.co
4	103640412	7634197461219	Đừng Lựa Chọn An I	Cảnh Thiên	4.8	426	Trong độ xuân xanh	99000	https://salt.tikicdn.co
5	16158713	3411207446768	Ikigai - Đi Tìm Lý Do	Hector Garcia	0	0	Ý nghĩa cuộc đời mỉ	67150	https://salt.tikicdn.co
6	114578907	8350327315889	Sapiens Lược Sử Lo	Yuval Noah Harari	1	1	Sapiens là một câu Trong ấn...	237000	https://salt.tikicdn.co
7	82482207	6363770705655	Combo 3 cuốn: OSHO - Yêu - Being In Lo		0	0	1.OSHO - Yêu - Bein "YÊU" TRONG TÌNH	316800	https://salt.tikicdn.co
8	92859698	3721488178204	Tâm Lý Học Biểu Cả	Mã Hạo Thiên	5	428	TÂM LÝ HỌC BIỂU Cuốn sách dạy bạn Từ gương mặt đến b	98000	https://salt.tikicdn.co
9	37968907	9839751930628	Tâm Thức Israel - Tí	Alon Gratch	0	0	Một Israel xấu xí hay	88000	https://salt.tikicdn.co
10	86517373	6985654193097	Khoa Học Về Nấu Ă	Stuart Farrimond	4.8	43	Nắm bắt các kiến thu	312500	https://salt.tikicdn.co
11	137453599	5220196871099	Chiến binh cầu vồng		0	0	Một tác phẩm có tầm "Thầy Harfan và cô M	105000	https://salt.tikicdn.co
12	93752753	3434944658139	Tâm Lý Học Nói Gì V	Richard Gross	4.8	97	Tâm lý học nói gì về Có cách nào "đúng" Nỗi đau ảnh hưởng	86000	https://salt.tikicdn.co
13	8458252	9413764345605	Pháp Môn Hạnh Phú	Đại Sư Tinh Vân	0	0	Trong cuộc sống, ch	58650	https://salt.tikicdn.co
14	107331062	2346179140161	Khởi Nghiệp Kinh Doanh Online - Bán Hà		4.8	405	Quyển sách căn bản Bạn đang muốn kinh Bạn muốn tìm kiếm t Hay bạn muốn làm..	199000	https://salt.tikicdn.co
15	73108963	6440150198723	Hoàng Tử Ếch (Truy	Stefania Leonardi Ha	0	0	Hoàng Tử Ếch - Cổ T Cô công chúa gặp g	35000	https://salt.tikicdn.co
16	69764541	9932851704459	Tâm Lý Học Hành Vi	Khương Nguy	4.8	1094	Tâm Lý Học Hành Vi Cuốn sách giúp bạn Bước chân bứt t	98000	https://salt.tikicdn.co

Hình 1 Flie book.csv

+ Rating.csv: gồm 241814 dòng dữ liệu , là đánh của người dùng tới 2000 cuốn sách. Thông tin lưu trữ có 4 trường: mã sách, mã người dùng thực hiện đánh giá, giá trị đánh giá của người dùng và comment của họ.



	A	B	C	D
	book_id	user_id	rating	content
	544731	5656713	5	Sách in ấn đẹp. Gia
	544731	5652322	5	Sách nguyên vẹn, đ
	544731	7065192	5	
	544731	1901381	5	Sách chuẩn. Đánh g
	544731	12071771	5	sản phẩm rất tốt. toà
	544731	18105801	5	Giao hàng cực nhar
	544731	19084977	5	
	544731	6503562	5	quá chuẩn như tên g tiki giao hàng rất rất
	544731	15062554	5	Sách bìa mỏng như
	544731	18005527	5	Sách được bọc bìa
	544731	23939752	5	Sách chất lượng
	544731	18911027	5	giao hàng rất nhanh
	544731	10796093	5	Sách mới tinh, thơm
	544731	7822443	5	Mua bộ 2 cuốn lý thu
	544731	23999225	5	Sách đẹp lắmm, hàr
	544731	13124656	5	Sách siêu xịn ạ khôn
	544731	10604691	5	đẹp, mới, khổ sách t
	544731	7266	5	Bộ 9 cuốn, 6 cấp họ
	544731	17889354	5	Gửi hàng nhanh lắm
	544731	7534103	5	tiki giao hàng nhanh
	544731	39426	5	Sách mới, trình bày
	544731	265800	5	Sách mới, in rõ ràng
	544731	8884496	5	Giao hàng nhanh. Đóng gói cẩn thận
	544731	20668619	5	Bao bì tốt, chất lượn
	544731	12186546	5	Sách đóng gói cẩn t
	544731	6687212	5	Sách mới và rất đẹp

Hình 2 File Rating.csv

+ User.csv: có 170941 dòng dữ liệu, tương ứng với 170941 người dùng đã đánh giá 2000 quyển sách. Thông tin lưu trữ có: id người dùng, tên, hình ảnh đại diện.

A	B	C
user_id	user_name	user_avatar
5656713	Dang Mai Ha	//tiki.vn/assets/img/av
5652322	Lê Tú	//tiki.vn/assets/img/av
7065192	Trần Thị Mai Ngọc	//tiki.vn/assets/img/av
1901381	Nguyễn Quân	//tiki.vn/assets/img/av
12071771	Tú Hảo	https://platform-looka
18105801	Phạm Phương Quỳn	//tiki.vn/assets/img/av
19084977	Tran Thi Kim Yến	//tiki.vn/assets/img/av
6503562	Hồ Tiến Đức	//tiki.vn/assets/img/av
15062554	Hoàng Hoa	//tiki.vn/assets/img/av
18005527	Võ Phương Anh	//tiki.vn/assets/img/av
23939752	Phương Anh Nguyễr	//tiki.vn/assets/img/av
18911027	Linh Chung	//tiki.vn/assets/img/av
10796093	Thảo Quyên	https://graph.faceboc
7822443	Đoàn Thị Mỹ Hoa	//tiki.vn/assets/img/av
23999225	Phương Anh	//tiki.vn/assets/img/av
13124656	Lê Hằng	//tiki.vn/assets/img/av
10604691	Binh ng	//tiki.vn/assets/img/av
7266	TRINH VÕ	//tiki.vn/assets/img/av
17889354	Yếnn Linhh	//tiki.vn/assets/img/av
7534103	Thu Thuỷ	//tiki.vn/assets/img/av
39426	Nhu Quynh	//tiki.vn/assets/img/av
265800	Lê Trường	//tiki.vn/assets/img/av

Hình 3 Flie user.csv

IV. PHƯƠNG PHÁP:

1. Phương pháp Lọc cộng tác:

Nhóm sử dụng phương pháp lọc cộng tác dựa trên bộ nhớ (Memory-base Collaborative Filtering). Phương pháp này sử dụng toàn bộ dữ liệu có được về người dùng và sách để tạo ra dự đoán, tìm ra tập người dùng - những người mà đã có lịch sử mua, đánh giá sách, sau đó sử dụng thuật toán để tính toán dự đoán đánh giá sách.

Có hai cách tiếp cận dựa trên bộ nhớ phổ biến đó là: **lọc dựa trên người dùng** và **lọc dựa trên sản phẩm**. Hệ thống lọc dựa trên người dùng sẽ tìm ra tập người dùng tương tự với người dùng đang xét dựa trên các sản phẩm mà các người dùng đó cùng đánh giá, sau đó sẽ dự đoán đánh giá của người dùng 1 với sản phẩm p dựa trên đánh giá của nhóm người dùng tương tự.

Tương tự như vậy, hệ thống lọc dựa trên sản phẩm sẽ tìm ra các sản phẩm tương tự nhau dựa vào nhóm người cùng đánh giá các sản phẩm đó, hệ thống dự đoán đánh giá của người dùng u với sản phẩm p dựa trên đánh giá của các sản phẩm tương tự.

a. Phương pháp tính toán độ tương đồng:

Việc đo độ tương tự giữa người dùng hoặc sản phẩm quyết định đến hiệu quả của phương pháp này, do đó cần chọn phương pháp đo độ tương tự phù hợp.Có nhiều phương pháp tính độ tương đồng như: hệ số tương quan Pearson, Hệ số tương tự cosine, khoảng cách Eclidean,… Nhóm chọn sử dụng hệ số tương tự cosine, hệ số tương quan Pearson để tính toán và xây dựng ma trận tương đồng của người/sách.

Cho:

- u, v là 2 người dùng trong bài toán khuyến nghị.
- r_{up}, r_{vp} , r_{ui}, r_{vi} lần lượt là đánh giá của người dùng u và v cho sản phẩm p, i.
- P_u, P_v, lần lượt là tập sản phẩm mà người dùng u và v đã đánh giá, m là tổng số sản phẩm chung của a và v cùng đánh giá.
- $\overline{ru}, \overline{rv}$ là trung bình tất cả đánh giá của người dùng từ u và v.

Hệ số tương quan Pearson:

Phương pháp này tính toán thống kê đánh giá chung của 2 người dùng để xác định sự giống nhau giữa 2 người dùng. Công thức tính hệ số tương quan pearson như sau:

$$S_{Pearson}(u,v) = \frac{\sum_{i \in P_u \cap P_v} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt{\sum_{i \in P_u \cap P_v} (r_{ui} - \overline{r}_u)^2} \sqrt{\sum_{i \in P_u \cap P_v} (r_{vi} - \overline{r}_v)^2}}$$

Hệ số tương quan pearson có giá trị nằm trong đoạn [-1,1]. Giá trị S càng lớn thì mức độ giống nhau càng cao. Nếu S>0 thì người dùng có xu hướng đánh giá giống nhau, ngược lại nếu S<0 thì người dùng có xu hướng đánh giá trái ngược nhau.

Hệ số tương tự Cosine:

Độ tương tự cosin là một cách đo độ tương tự (measure of similarity) giữa hai vectơ khác không của một không gian tích vô hướng. Độ tương tự này được định nghĩa bằng giá trị cosine của góc giữa hai vectơ, và cũng là tích vô hướng của cùng các vectơ đơn vị để cả hai đều có chiều dài 1.

Độ tương tự giữa 2 người dùng u và v có thể được đo dựa trên tính toán Độ tương tự cosin. Hệ số tương tự Cosine mô phỏng người dùng trong không gian vector (ru, rv là vector đánh giá của u, v) sau đó lấy cosine các góc giữa 2 vector để tính toán độ tương tự giữa 2 người dùng.

$$S_{Cos}(u,v) = \frac{r_u . r_v}{\| r_u \|_2 \| r_v \|_2} = \frac{\sum_{i \in P_u \cap P_v} r_{ui} r_{vi}}{\sqrt{\sum_{i \in P_u \cap P_v} (r_{ui})^2} \sqrt{\sum_{i \in P_u \cap P_v} (r_{vi})^2}}$$

Cũng giống hệ số tương quan pearson, hệ số tương tự cosine có giá trị nằm trong đoạn [-1,1]. Giá trị S càng lớn thì mức độ giống nhau càng cao. Nếu S>0 thì người dùng có xu hướng đánh giá giống nhau, ngược lại nếu S<0 thì người dùng có xu hướng đánh giá trái ngược nhau.

Phương pháp lọc cộng tác dựa trên người dùng:

Đầu vào: U = {u}, tập các người dùng quan sát được P = {p}, tập các sách của hệ thống. R = {r}, tập đánh giá của các người dùng quan sát được **Đầu ra:** ∀u∈U, trả về TopN những p∈P dựa trên giá trị hữu ích(giá trị đánh giá) tiên đoán.
Bước 1: Tiền xử lý dữ liệu thu thập: Dữ liệu thu thập được rất lớn với số lượng đánh giá của 1 người dùng rất ít và số gây ra ma trận thưa. Để xây dựng hệ khuyến nghị nhanh chóng và hiệu quả cần xử lý dữ liệu thực nghiệm bằng cách lọc lại dự liệu theo điều kiện người dùng phải đánh giá tối thiểu 10 cuốn sách và 1 cuốn sách phải có tối thiểu 100 lượt đánh giá. **Bước 2: Xây dựng Ma trận đánh giá:** Ma trận đánh giá được xây dựng từ dữ liệu rating sau khi lọc. Hàng là danh sách người dùng, cột là danh sách sách. **Bước 3: Tìm Top_N những người có đồng sở thích với người dùng X cần khuyến nghị:** Sử dụng độ đo cosine hoặc hệ số tương quan pearson để tính độ tương đồng của tất cả người dùng trong tập U và sắp xếp kết quả theo thứ tự từ cao đến thấp. **Bước 4: Dự đoán rating của người dùng X cho tập sách mà X chưa rating.**

Dựa trên Top_N người dùng tương đồng đã tìm được dự đoán rating của người dùng X với những quyển sachs chưa được đánh giá bằng phương pháp trung bình đánh giá và sắp xếp kết quả dự đoán rating theo thứ tự từ cao đến thấp

Bước 5: Khuyến nghị cho người dùng:

Khuyến nghị cho người dùng top_20 quyển sách mà họ chưa mua, đánh giá có khả năng phù hợp với sở thích của của họ dựa vào các dự đoán của hệ thống.

2. Phương pháp Lọc dựa trên nội dung:

Lọc dựa trên nội dung (Content-based) dựa trên mô tả của sách và thông tin của người dùng. Thuật toán này cố gắng đề xuất các quyển sách tương tự như các sách mà người dùng đã mua và đánh giá tốt trong quá khứ (hoặc đang kiểm tra trong hiện tại). Vấn đề chính của phương pháp này là bị giới hạn bởi nội dụng của sách, chỉ tư vấn được các sách tương tự trong cùng nội dung.

Phương pháp tiền xử lý dữ liệu:

Dữ liệu tiêu đề và tóm tắt của sách sẽ được sử dụng để xây dựng hệ khuyến nghị sách dựa trên nội dung nên cần phải xử lý, làm sạch các thông tin này:

+ Loại bỏ các thẻ html, các ký tự đặt biệt.

+ Chuẩn hóa từ, dấu câu

+ Vì tập dữ liệu sách Tiếng Việt nên sử dụng thư viện stopword vietnamese để loại bỏ stopword.

+ Sử dụng thư viện ViTokenizer để tách từ.

Phương pháp biểu diễn hồ sơ sách:

TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector (VSM), hoặc thành những vector thưa thớt.

TF (Term Frequency): là tần suất xuất hiện của một từ trong một đoạn văn bản. Với những đoạn văn bản có độ dài khác nhau, sẽ có những từ xuất hiện nhiều ở những đoạn văn bản dài thay vì những đoạn văn bản ngắn. Vì thế, tần suất này thường được chia cho độ dài của đoạn văn bản như một phương thức chuẩn hóa (normalization). TF được tính bởi công thức:

$$tf(t) = \frac{f(t,d)}{T}$$

Với *t* là một từ trong đoạn văn bản.

f(t,d) là tần suất xuất hiện của t trong đoạn văn bản d.

T là tổng số từ trong đoạn văn bản đó.

IDF (Inverse Document Frequency): tính toán độ quan trọng của một từ. Khi tính toán TF, mỗi từ đều quan trọng như nhau, nhưng có một số từ trong tiếng Anh như "is", "of", "that",... xuất hiện khá nhiều nhưng lại rất ít quan trọng. Vì vậy, chúng ta cần một phương thức bù trừ những từ xuất hiện nhiều lần và tăng độ quan trọng của những từ ít xuất hiện những có ý nghĩa đặc biệt cho một số đoạn văn bản hơn bằng cách tính IDF:

$$idf(t) = \log(N/(df + 1))$$

$$Tf\text{-}idf(t) = tf(t) \times idf(t)$$

Phương pháp lọc dựa trên nội dung:

Đầu vào: R = {r}, lịch sử đánh giá sách của người dùng quan sát được P = {p}, tập các sách của hệ thống. **Đầu ra:** ∀r∈R, trả về TopN những p∈P dựa trên giá trị hữu ích tiên đoán.
Bước 1:Tiền xử lý dữ liệu: Kết hợp tên sách, tác giả, tóm tắt làm feature biểu diễn nội dung sách. ∀p∈P. • Rút trích phần tiêu đề, tên tác giả và tóm tắt. Loại bỏ stopwords , chuẩn hóa câu, dấu câu, kí tự đặc biệt, tách từ. **Bước 2: Xây dựng ma trận vector TFIDK** biểu nội dung của các quyển sách • Xây dựng vector biểu diễn nội dung sách p, là $\vec{fp}$, dùng phương pháp gán trọng số TFIDF. • Xây dựng ma trận TFIDF cho tất cả 6000 quyển sách. **Bước 4: Thực hiện khuyến nghị** : Từ nội dung các sách mà người dùng đã mua và đánh giá tốt khuyến nghị top N những quyển sách tương tự với những quyển sách này cho người dùng. Sử dụng độ đo Cosine để tìm ra topN khuyến nghị.

V. THỰC NGHIỆM, ĐÁNH GIÁ HIỆU QUẢ:

1. Tập dữ tập liệu và thiết lập thực nghiêm:

Trong thực nghiệm, nhóm tiến hành thực nghiệm trên tập dữ liệu sách Tiki đã crawl. Tuy nhiên tập dữ liệu này có kích thước lớn (2000 cuốn sách với 241814 đánh giá của 170941 người dùng) và qua khảo sát thấy rằng có rất nhiều người dùng chỉ mua và đánh giá một vài sách và có những quyển sách chưa có đánh giá/ số lượt đánh giá thấp. Để giảm thưa cho ma trận , tăng độ tin cậy của kết quả khuyến nghị của mô hình, nhóm tiến hành xây dựng lại tập dữ liệu thực nghiệm cho mô hình theo điều kiện chỉ chọn những người dùng đánh giá từ 10 quyển sách trở lên và những quyển sách được đánh giá ít nhất bởi 100 người dùng. Sau khi thực hiện các thao tác chọn lọc, nhóm đã có ma trận nhị phân cho thực nghiệm có kích thước 766 x 304.(dữ liệu đánh giá của 766 cuốn sách của 304 người dùng – 4095 dòng đánh giá)

```
rating_matrix
```

user_id	2426	9777	23348	26806	36210	39401	51302	88523	88707	96
book_id										
185906	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
339308	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
365780	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
365794	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
381234	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	
117595535	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
121236564	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
124780633	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	
126131817	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
134846265	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

766 rows × 304 columns

Vì không thể thu thập dữ liệu đánh giá sách theo thời gian hay tương tác (xem, thích, thêm vào giỏ) của người dùng nên nhóm sẽ sử dụng tập các sách mà 1 người dùng đã mua và có đánh giá để xây dựng tập GroundTruth đánh giá độ chính xác của khuyến nghị.

Bi : danh sách rating sách mà người dùng i đã thực hiện mua,đánh giá.

Bi sẽ được chia thành 2 tập nhỏ.1 tập làm lịch sử mua của người dùng đưa và mô hình khuyến nghị để khuyến nghị và tập còn lại sẽ ẩn khỏi mô hình khuyến ghị dùng làm GroundTruth để đánh giá. Tỉ lệ 2 tập này là 30-70

2. **Phương pháp đánh giá:**

Độ đo Precision , recall, F-mearsure:

$$\text{Precion} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F} = 2.\frac{\text{Precion,Recall}}{\text{Precion+Recal}}$$

Lựa chọn của người dùng	**Khuyến nghị của mô hình**	
	Khuyến nghị	Không khuyến nghị

Có đánh giá	TP	FN
Không đanh giá	FP	TN

TP: những sách hệ thống khuyến nghị có trong GroundTruth.

FN: số sách có trong GroundTruth, nhưng mô hình không khuyến nghị (bỏ sót).

FP: số sách không có trong GroundTruth, nhưng được mô hình khuyến nghị (nhận nhầm)

Precision và Recall có giá trị trong [0,1], hai giá trị này càng gần với 1 thì mô hình càng chính xác. Precision càng cao đồng nghĩa với các điểm được phân loại càng chính xác. Recall càng cao cho thể hiện cho việc ít bỏ sót các điểm dữ liệu đúng.

3. **Kết quả:**

Đánh giá phương pháp lọc cộng tác dựa trên người dùng:

Bảng 1 Lọc cộng tác dựa trên nội dung, sử dụng độ đo cosine

User_id	***precision***	***Recall***	***F-measure***
5253296	***0,040***	***0,103***	***0,058***
1895509	***0,050***	***0,156***	***0,076***
22456806	***0,040***	***0,133***	***0,062***
1053107	***0,020***	***0,080***	***0,032***
19514032	***0,080***	***0,320***	***0,128***
	0,046	***0,158***	***0,072***

Bảng 2 - Lọc cộng tác dựa trên người dùng sử dụng độ đo pearson

User_id	***precision***	***Recall***	***F-measure***
5253296	***0,030***	***0,077***	***0,043***
1895509	***0,010***	***0,032***	***0,015***
22456806	***0,020***	***0,067***	***0,030***
1053107	***0,020***	***0,080***	***0,032***
19514032	***0,030***	***0,120***	***0,048***
	0,022	***0,079***	***0,034***

Đánh giá phương pháp lọc dựa trên nội dung:

User_id	***precision***	***Recall***	***F-measure***
5253296	***0,076***	***0,163***	***0,104***

18837650	***0,04***	***0,429***	***0,073***
9909635	***0,1***	***0,231***	***0,140***
26806	***0,167***	***0,385***	***0,233***
19514032	***0.133***	***0,32***	***0,188***
	0,1032	***0,3056***	***0,1476***

4. So sánh, nhận xét:

	precision	***Recall***	***F-measure***
Lọc cộng tác sử dụng pearson	***0,022***	***0,079***	***0,034***
Lọc cộng tác sử dụng cosin	***0,046***	***0,158***	***0,072***
Lọc dựa trên nội dung	***0,1032***	***0,3056***	***0,1476***

+ Lọc cộng tác sử dụng độ đo Cosine cho hiệu quả tốt hơn lọc cộng tác sử dụng độ đo Pearson, lọc cộng tác sử dụng độ đo Pearson cho kết quả kém nhất.
+ Mô hình lọc dựa trên nội dung có kết quả tốt hơn lọc cộng tác.
+ Kết quả của đánh của mô hình khuyến nghị thấp.
+ Ưu, khuyết điểm của 2 phương pháp xây dựng hệ khuyến nghị:

	Lọc cộng tác	Lọc dựa trên nội dung
Ưu	+Có thể đa dạng khuyến nghị	+Giải quyết được vấn đề ma trận thưa. + Giải quyết được vấn đề người sản phẩm mới.
Khuyết	+Ma trận thưa +Không khuyến nghị được cho người dùng mới khi chưa có lịch sử hành vi, tương tác đánh giá. + Không khuyến nghị các sản phẩm mới khi chưa được ai xem, đánh giá. + Thời gian tính toán cao khi ma trận lớn, khó đáp ứng được thời gian thực. + Kết quả khuyến nghị không hiệu quả khi có các đánh giá sai (đối thủ cạnh tranh)	+Không đa dạng được khuyến nghị. + Khuyến nghị,tư vấn các sản phẩm quen thuộc +Không khuyến nghị được cho người dùng mới khi chưa có lịch sử hành vi, tương tác đánh giá.

VI. DEMO:

Lọc cộng tác dựa trên người dùng:

```
[68] recommend_book(19514032,15,100,'cosine')
```

```
('Thư Viện Nửa Đêm', 117238177, 5.0),
('Lược Sử Vật Lý Lượng Tử - Chúa Có Gieo Xúc Xắc Cho Bạn?', 109018532, 5.0),
('Spy X Family - Tập 5 - Tặng Kèm Standee PVC', 108956811, 5.0),
('Ổn Định Hay Tự Do', 105483727, 5.0),
('Hung Khí Hoàn Mỹ', 103550551, 5.0),
('Người Hùng Mang Ngàn Gương Mặt', 102775969, 5.0),
('\xa0The Having Tư Duy Thịnh Vượng', 102411866, 5.0),
('Bài Học Cuộc Sống (Tặng Kèm 01 Bookmark)', 101995510, 5.0),
('Tan (Tặng Kèm 12 Postcard Nhân Vật)', 98239315, 5.0),
('7 Thói Quen Hiệu Quả (The 7 Habits Of Highly Effective People) (Tái Bản)',
 91947689,
 5.0),
('Overlord 6 (Phiên Bản Manga)', 91027731, 5.0),
('"Cậu" Ma Nhà Xí Hanako Tập 4', 89179384, 5.0),
('Lối Tắt Khởi Nghiệp - Con Đường Ngắn Từ Tay Trắng Đến Thành Công Bền Vững',
 87941401,
 5.0),
('Tôi Tự Học (Tái Bản)', 87391047, 5.0),
('Giữa Thế Gian Ồn Ào Sống Một Đời Giản Đơn', 86540440, 5.0),
('"Cậu" Ma Nhà Xí Hanako Tập 3', 85256403, 5.0),
('Óc Sáng Suốt (Tái Bản 2021)', 82936273, 5.0),
('"Cậu" Ma Nhà Xí Hanako Tập 2', 81506284, 5.0),
('"Cậu" Ma Nhà Xí Hanako - Tập 1', 78015570, 5.0),
('Học Viện - The Institute (Stephen King)', 78015041, 5.0),
('Nghệ Thuật Tập Trung - Nâng Cao Năng Suất, Tối Ưu Thời Gian, Hiệu Quả Bất Ngờ',
 77774625,
```

Hình 4 Danh sách các cuốn sách khuyến nghị cho người dùng id = 19514032 bằng phương pháp lọc cộng tác dựa trên người dùng và sử dụng độ tương tự cosine

```
recommend_book(19514032,15,100,'pearson')
```

```
( Horimiya - 06 , 124780633, 0),
('PHÂN TÍCH MẪU HÌNH BIỂU ĐỒ - Những Bí Quyết Giúp Nhà Giao Dịch Siêu Hạng DAN ZANGER Biến 11 Ngh
 121236564,
 0),
('Kẻ Ngoại Cuộc', 117595535, 0),
('Thư Viện Nửa Đêm', 117238177, 0),
('Quý Cô Thịnh Vượng -\xa0Khi Phụ Nữ Tư Duy Đúng Về Tiền', 112333133, 0),
('Nhớ Ra Tên Tôi Chưa (Tập 1+2)', 111981249, 0),
('Miền Đất Hứa 20', 111536237, 0),
('\xa0[Manga] Diệt Slime Suốt 300 Năm, Tôi Levelmax Lúc Nào Chẳng Hay (Tập 5)\xa0',
 111155169,
 0),
('Dr. Stone - Tập 4: Phòng Thí Nghiệm Của Senku', 110594177, 0),
('Tội Ác Và Hình Phạt', 110448031, 0),
('Combo 8 cuốn: Bộ sách IQ vượt trội - IQ booster', 109862191, 0),
('Combo 8 cuốn Ehon Kỹ năng sống: Bon và Gia đình, Bạn bè', 109322300, 0),
('Sách - Thiên Hồn - Mệnh Do Trời Định, Vận Do Ta Sinh( Cao Minh )',
 109295448,
 0),
('Dragon Ball Super Anime Comics Bảy Viên Ngọc Rồng Siêu Cấp Broly [Tặng Kèm Ngẫu Nhiên 1 Trong 3
 109089738,
 0),
('Horimiya 4', 109019773, 0),
('Lược Sử Vật Lý Lượng Tử - Chúa Có Gieo Xúc Xắc Cho Bạn?', 109018532, 0),
('Combo 3 cuốn: Miếng dán Thông minh cho bé - My First Sticker Book',
 108957065.
```

Hình 5 Danh sách các cuốn sách khuyến nghị cho người dùng id = 19514032 bằng phương pháp lọc cộng tác dựa trên người dùng và sử dụng độ tương tự pearson

Lọc dựa trên nội dung:

```
recommend_for_user(5253296)
```

```
[('Hạt Giống Tâm Hồn 3 - Từ Những Điều Bình Dị (New Edition 2020)',
  4957,
  5.0,
  0.4439983597239661),
 ('Hạt Giống Tâm Hồn - Tuyển Chọn Những Câu Chuyện Hay Nhất (Tái Bản)',
  4897,
  4.5,
  0.2769431218010265),
 ('Hạt Giống Tâm Hồn  - Tuyển Chọn Những Câu Chuyện Hay Nhất (Bìa Mềm)',
  2553,
  4.8,
  0.2691920950572343),
 ('Combo Trọn Bộ Bubu (59 Cuốn)(Tái Bản)', 1649, 5.0, 0.22237498371765194),
 ('Hạt Giống Tâm Hồn 1 (Tái Bản 2020)', 2416, 4.8, 0.2162203267353206),
 ('Xuyên Thành Phản Diện Biết Sống Sao Đây, Tập 4',
  802,
  5.0,
  0.6607868572633117),
 ('Xuyên Thành Phản Diện Biết Sống Sao Đây', 748, 4.8, 0.44916657740238974),
 ('Trọn Bộ 6 Tập: Tru Tiên (Tái Bản)', 1325, 4.8, 0.18249203168406908),
 ('Nhật Ký Sống Sót Của Nữ Phụ Phản Diện - Tập 1',
  4554,
  5.0,
  0.14983933477685005),
 ('Hệ Thống Tự Cứu Của Nhân Vật Phản Diện - Tập 3',
  1243,
  5.0,
```

Hình 6 - Danh sách các cuốn sách khuyến nghị cho người dùng id = 19514032 bằng phương pháp lọc dựa trên nội dung

LINK DATASET VÀ DEMO:

https://drive.google.com/drive/folders/16uL2MssnZLzqZv0ImpzsFNeuUdmiM8rq?usp=sharing

VII. KẾT LUẬN:

1. **Kết quả đạt được:**

- Có kiến thức cơ bản về hệ khuyến nghị, các kỹ thuật, thuật toán, phương pháp xây dựng và đánh giá một hệ khuyến nghị.
- Tìm hiểu và xây dựng được hệ khuyến nghị bằng 2 phương pháp truyền thống: lọc cộng tác dựa trên người dùng và lọc nội dung.

2. **Khó khăn, thách thức:**

- Dữ liệu lớn, dữ liệu thu thập được rất thưa, dù đã xử lý và chọn lọc dữ liệu nhưng ma trận đánh giá vẫn khá thưa.
- Không thu thập được dữ liệu thích hợp để xây dựng được tập dữ liệu chuẩn cho việc đánh giá. Việc lựa chọn phương pháp đánh giá chưa tốt.
- Độ chính xác của hệ khuyến nghị chưa cao. Chưa thực hiện tốt việc đánh giá.

- Mô hình lọc dựa trên nội dung chưa thực sự hiệu quả vì danh sách khuyến nghị có thể trùng, cần xây dựng hồ sơ người dùng.
- Mới tìm hiểu về hệ khuyến nghị,thời gian ngắn phải tiếp cận nhiều kiến thức mới, kiến thức còn hạn hẹp nên còn nhiều thiếu sót.

3. Hướng cải thiện, phát triển mô hình khuyến nghị cho đồ án:

- Sử dụng phương pháp lọc cộng tác kết hợp với lọc nội dung để khắc phục vấn đề ma trận thưa, tăng độ tin cậy cho hệ khuyến nghị .Sử dụng các phương pháp giảm số chiều như: SVD, PCA để giảm độ lớn của ma trận.
- Thực nghiệm trên nhiều phương pháp tính độ tương đồng để tìm ra phương pháp lọc cộng tác hiệu quả.
- Sử dụng thêm các phương pháp đánh giá MRR, NDCG để đánh giá chất lượng danh sách khuyến nghị.
- Sử dụng thêm các phương pháp embedding như vec2word để cải thiện mô hình khuyến nghị bằng phương pháp lọc nội dung.
- Triển khai hệ thống, xây dựng khuyến nghị trực tiếp, và tìm hiểu cách đánh giá cho hệ thống online.

VIII. TÀI LIỆU THAM KHẢO:

[1] Wikipedia: Hệ số tương tự Cosine, Hệ số tương quan Pearson, MAE:
Địa chỉ: https://vi.wikipedia.org/ [Truy cập lần cuối 30/12/2020]

[2] viblo.asia: Trích chọn thuộc tính trong đoạn văn bản với TF-IDF

Địa chỉ: https://viblo.asia/p/trich-chon-thuoc-tinh-trong-doan-van-ban-voi-tf-idf-Az45bAOqlxY [Truy cập lần cuối 30/12/2020]

[3] tutorials.aiclub.cs.uit.edu.vn: Làm thế nào để đánh giá một mô hình Máy học?

Địa chỉ: http://tutorials.aiclub.cs.uit.edu.vn/index.php/2021/05/18/evaluation/ [Truy cập lần cuối 30/12/2020]

[4] viblo.asia: Introduction to Recommender Systems

Địa chỉ: https://viblo.asia/p/introduction-to-recommender-systems-aWj53LQ8K6m

[Truy cập lần cuối 30/12/2020]