TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Người thực hiện: **HỒ XUÂN BI – 20056951**

BÙI HOÀNG SƠN – 20053181

Lớp : **DHKHDL16A**

Khoá : **16**

Người hướng dẫn: **TS BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

Commented [TT2]: Bold, font size 14

Commented [TT3]: Bold, font size 14

Commented [CPP4]: Bold, font size 16, ghi rõ loại báo cáo (bài tập lớn hoặc đồ án cuối kì) và tên đầy đủ môn học?

Commented [TT5]: Bold, font 14

Commented [TT6]: Tên thầy cô hướng dẫn

Commented [TT7]: Đây là trang bìa, in trên bìa cứng màu xanh dương không hoa văn khi nộp.

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

Commented [TT8]: Bold, font size 14

Commented [TT9]: Bold, font size 14

ĐỒ ÁN CUỐI KÌ

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Commented [CPP10]: Bold, font size 16, ghi rõ loại báo cáo (bài tập lớn hoặc đồ án cuối kì) và tên đầy đủ môn học?

Người thực hiện: **HỒ XUÂN BI - 20056951**
BÙI HOÀNG SƠN - 20053181
Lớp : **DHKHDL16A**
Khoá : **16**
Người hướng dẫn: **TS. BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

Commented [TT11]: Bìa phụ, in giấy thường

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn đến Thầy TS. Bùi Thanh Hùng– Giảng viên hướng dẫn đã hỗ trợ về kiến thức, kinh nghiệm, kĩ năng giúp chúng tôi thực hiện Đồ án cuối kì này.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

Commented [TT12]: Bold, font size 16

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Commented [TT13]: Nếu là 2 sinh viên

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Hồ Xuân Bi

Commented [TT14]: Tên sinh viên 1

Bùi Hoàng Sơn

Commented [TT15]: Tên sinh viên 1

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

_____
_____
_____
_____
_____
_____
_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Bài toán hệ thống gợi ý phim là một trong những bài toán quan trọng trong lĩnh vực xử lý thông tin và máy học, nhằm cung cấp các đề xuất phim cá nhân hóa cho người dùng dựa trên sở thích và hành vi của họ. Đây là một vấn đề nghiên cứu quan trọng và có nhiều hướng tiếp cận.Mục tiêu là xây dựng một hệ thống gợi ý phim hiệu quả, cá nhân hóa và đa dạng để cung cấp các đề xuất phim chất lượng cho người dùng.

Một số hướng tiếp cận như Content-Based Filtering (CBF) dựa trên thông tin nội dung của phim như thể loại, diễn viên, đạo diễn, và từ khóa để tạo ra các đặc trưng mô tả phim. Sử dụng các thuật toán phân loại hoặc học máy để dự đoán sở thích của người dùng dựa trên thông tin nội dung. Collaborative Filtering (CF) dựa trên thông tin phản hồi từ người dùng về việc xếp hạng hoặc đánh giá các bộ phim để tìm ra các mẫu tương đồng giữa các người dùng hoặc phim. Sử dụng các phương pháp như K-Nearest Neighbors (KNN) hoặc Singular Value Decomposition (SVD) để dự đoán sở thích của người dùng. Hybrid Recommendation kết hợp cả CBF và CF để tận dụng lợi ích của cả hai phương pháp, cung cấp đề xuất phim đa dạng và cá nhân hóa hơn.

Xây dựng và tối ưu hóa các mô hình dự đoán sở thích của người dùng dựa trên các phương pháp CBF và CF. Kết hợp thông tin từ nhiều nguồn khác nhau như thông tin nội dung, xếp hạng người dùng, và thông tin ngữ cảnh để tạo ra các đề xuất phim chất lượng. Kết quả thu được là tính toán các đặc trưng nội dung phim hiệu quả, cải thiện việc phân loại và dự đoán sở thích của người dùng trong mô hình CBF. Áp dụng các phương pháp CF như KNN hoặc SVD để dự đoán sở thích của người dùng và tạo ra các đề xuất phim chất lượng.

Kết hợp hiệu quả giữa CBF và CF trong mô hình hybrid recommendation để cung cấp đề xuất phim tốt hơn với độ chính xác và đa dạng cao hơn, mang lại hiệu suất tốt hơn so với việc sử dụng mỗi phương pháp độc lập, cải thiện trải nghiệm người dùng khi xem phim và tạo ra các đề xuất phù hợp hơn.

MỤC LỤC

Commented [TT16]: Chứa các tiểu mục và số trang

LỜI CẢM ƠN i
PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iiiii
TÓM TẮT iv
MỤC LỤC 1
DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 3
DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4
CHƯƠNG 1 – GIỚI THIỆU BÀI TOÁN 6
1.1 Giới thiệu về bài toán 6
1.2 Ý nghĩa của bài toán 6
1.2 Ý nghĩa của bài toán 5
CHƯƠNG 2 – PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN 7
2.1 Yêu cầu của bài toán 7
2.2 Các phương pháp giải quyết bài toán 8
2.3 Phương pháp đề xuất giải quyết bài toán 9
CHƯƠNG 3 – PHƯƠNG PHÁP ĐỀ XUẤT 10
3.1 Mô hình tổng quát 10
3.2 Đặc trưng của mô hình đề xuất 10
3.2.1 Thu thập dữ liệu 10
3.2.2 Tiền xử lí dữ liệu 12
3.2.3 Xây dựng phương pháp 12
3.2.4 Kết hợp các mô hình 14
3.2.5 Phương pháp huấn luyện 14
CHƯƠNG 4 – THỰC NGHIỆM 17
4.1 Dữ liệu 17
4.2 Xử lý dữ liệu 17

4.3 Công nghệ sử dụng 17
4.4 Cách đánh giá 18
4.5 Kết quả đạt được 18
CHƯƠNG 5 – KẾT LUẬN 21
5.1 Kết luận 21
5.1.1 Kết quả 21
5.1.2 Hạn chế 21
5.2 Hướng phát triển 21
LÀM VIỆC NHÓM 25
TỰ ĐÁNH GIÁ 27

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

u	*user (người xem phim)*
i	*item (bộ phim)*
$sim(u_0, u_1)$	*hàm simirilaty*
$\hat{y}_i$	*giá trị dự đoán*
y_i	*giá trị thực tế*
n	*số lượng mẫu dữ liệu*

CÁC CHỮ VIẾT TẮT

RS	Recommendation System
CB	Content-Based
CBF	Content-Based Filtering
CF	Collaborative Filtering
HR	Hybrid Recommendation
NLP	Natural Language Processing
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
SVD	Singular Value Decomposition algorithm
KNN	K-Nearest Neighbor algorithm

DANH MỤC CÁC HÌNH VẼ

Commented [TT17]: Danh mục các bảng, biểu, hình vẽ, đồ thị (nếu có).

Hình 3.1: Mô hình tổng quan 11
Hình 3.2: Ma trận Cosine Similarity 11
Hình 4.1: Kết quả của RMSE và MAE 19
Hình 4.2: Biểu đồ boxplot của RMSE và MAE 19
Hình 4.3: Biểu đồ bar của RMSE và MAE 20

DANH MỤC CÁC BẢNG

Commented [TT18]: Danh mục các bảng, biểu, hình vẽ, đồ thị (nếu có).

Bảng 3.1: Ví dụ về utility matrix với hệ thống Gợi ý phim 13

Bảng 4.1: Mô tả công nghệ sử dụng cho thực nghiệm bài toán gợi ý phim ảnh 18

Bảng 4.2: So sánh kết quả thực nghiệm 20

CHƯƠNG 1
GIỚI THIỆU VỀ BÀI TOÁN

1.1 Giới thiệu về bài toán

Bài toán xây dựng *Hệ thống gợi ý xem phim (Recommander System) bằng phương pháp kết hợp giữa Content –Base và Collaborative Filtering tích hợp với phân tích ý kiến người dùng*, là một trong mảng khá rộng của Machine Learning và trí tuệ nhân tạo, có tuổi đời khá ngắn vì Internet mới thực sự bùng nổ khoảng 10 năm trở lại nhưng cơ hội phát triển lại rất lớn. Hệ thống giúp đề xuất những bộ phim có thể phù hợp và hấp dẫn với người dùng - users dựa trên sở thích cá nhân của họ.

Mô hình kết hợp Content-Based và Collaborative Filtering là một giải pháp hiệu quả. Content-Based - CB sử dụng thông tin về các thuộc tính của phim như thể loại phim, diễn viên, đánh giá, xếp hạng, ... để tạo ra các mô hình đề xuất dựa trên sự tương đồng về nội dung giữa các bộ phim. Trong khi đó, Collaborative Filtering tập trung vào mối quan hệ giữa người dùng và người dùng (user-user), phim và phim (item-item), dựa trên hành vi đánh giá - ranking, ... để tạo ra các gợi ý.

Việc kết hợp hai phương pháp này lại giúp hệ thống có thể cung cấp những đề xuất chính xác và mang tính cá nhân hóa cao hơn. Sự kết hợp này giúp bổ sung cho nhau, vượt qua những hạn chế riêng biệt của từng phương pháp, tạo nên một hệ thống gợi ý mạnh mẽ và linh hoạt để đáp ứng nhu cầu giải trí của người dùng khi xem phim.

1.2 Ý nghĩa của bài toán

Bài toán mang lại giá trị to lớn trong việc cải thiện trải nghiệm người dùng và tăng cường hoạt động kinh doanh cho các nền tảng giải trí, giúp tối ưu hóa trải nghiệm của người dùng khi xem phim trực tuyến trên các nền tảng:

- Cải thiện Trải Nghiệm Người Dùng: Hệ thống gợi ý phim giúp người dùng dễ dàng khám phá và truy cập những bộ phim mới và phù hợp với sở thích cá nhân.
- Tăng Khả năng Khám Phá: Thay vì chỉ dựa vào những bộ phim phổ biến, hệ thống gợi ý mở rộng tầm nhìn của người dùng, giúp họ khám phá các thể loại, đạo diễn, hoặc diễn viên mới mà có thể họ chưa biết.
- Tăng Sự Gắn Kết Khách Hàng: Việc gợi ý phim chính xác và hấp dẫn giúp tạo một môi trường tiêu dùng tích cực, tăng sự yêu thích và gắn kết của người dùng với nền tảng hoặc dịch vụ.
- Hiệu Quả Kinh Doanh: Hệ thống gợi ý xem phim giúp tối ưu hóa thu nhập của các nền tảng trực tuyến thông qua việc giữ chân người dùng ở lại lâu hơn và tăng cơ hội quảng cáo hoặc bán hàng.

CHƯƠNG 2
PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN

2.1 Yêu cầu của bài toán

Bài toán xây dựng hệ thống gợi ý xem phim kết hợp Content-Based và Collaborative Filtering tích hợp với phân tích ý kiến người dùng đặt ra một số yêu cầu cơ bản sau:

- Thu thập dữ liệu: cần thu thập thông tin về các thuộc tính của phim như tên phim, thể loại, năm xuất bản, tóm tắt, điểm đánh giá, diễn viên, đạo diễn, từ khóa và các đánh giá từ người dùng.
- Xây dựng mô hình Content-Based: phân tích nội dung, thể loại, ... của phim, dựa trên các thông tin thu thập được, để đo lường sự tương đồng giữa các bộ phim và đề xuất các phim có nội dung tương tự.
- Xây dựng mô hình Collaborative Filtering: phát triển mô hình để phân tích mối quan hệ giữa người dùng và phim dựa trên dữ liệu về đánh giá, xếp loại, nội dung để đưa ra gợi ý phù hợp cho người dùng.
- Tích hợp phân tích ý kiến người dùng: sử dụng các công cụ phân tích ý kiến người dùng để hiểu sâu hơn về cảm nhận, ý kiến và phản hồi của họ về các bộ phim. Tích hợp các phương pháp xử lý ngôn ngữ tự nhiên (NLP) để phân tích và hiểu được ý kiến, đánh giá của người dùng.
- Kết hợp các phương pháp: cần có một phương pháp kết hợp hiệu quả giữa Content-Based và Collaborative Filtering, có thể thông qua việc kết

hợp các điểm mạnh của mỗi phương pháp hoặc sử dụng các mô hình hybrid.

- Tối ưu hóa và đánh giá hiệu suất: đánh giá hiệu suất của hệ thống thông qua các chỉ số đánh giá như độ chính xác, độ phủ, hay các độ đo khác phù hợp với mục tiêu cụ thể.

Tổng quan, bài toán này đặt ra yêu cầu kỹ thuật cao, yêu cầu sự tích hợp linh hoạt giữa các phương pháp khác nhau cùng với việc hiểu sâu về ý kiến và hành vi người dùng để cung cấp đề xuất phim chất lượng và cá nhân hóa tốt.

2.2 Các phương pháp giải quyết bài toán

Tuy thời gian phát triển ngắn, nhưng các phương pháp giải quyết bài toán xây dựng hệ thống gợi ý xem phim kết hợp Content-Based và Collaborative Filtering tích hợp với phân tích ý kiến người dùng lại phát triển rất nhanh. Một số nghiên cứu đáng chú ý như bài báo "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model" bởi Koren đã đề xuất mô hình Matrix Factorization kết hợp với Collaborative Filtering, sử dụng tập dữ liệu Netflix Prize đạt được cải thiện đáng kể trong việc dự đoán xếp hạng của người dùng, nhưng không xử lý thông tin nội dung (Content-Based) của các bộ phim (Koren, 2008). Bài báo "Factorization Machines" của Rendle, với dataset Movielens, sử dụng Factorization Machines (FM) để kết hợp thông tin từ cả Content-Based và Collaborative Filtering (Rendle, 2010). FM có thể học được cả mối quan hệ giữa người dùng và phim (Collaborative Filtering) lẫn thông tin từ nội dung phim (Content-Based), cải thiện độ chính xác so với các phương pháp truyền thống. Nhưng FM vẫn có thể đòi hỏi nhiều dữ liệu và không hiệu quả với các hệ thống lớn. Bài báo "An enhanced semantic layer for hybrid recommender systems: Application to news recommendation", sử dụng mô hình kết hợp giữa Collaborative Filtering và Content-Based dựa trên ma trận phân tích Latent Semantic Analysis (LSA) (Cantador, Castells and Bellogín, 2011). Dữ liệu thực nghiệm sử dụng dữ liệu từ MovieLens. Mô hình hybrid này kết hợp thông tin về nội

dung phim và thông tin về sở thích người dùng, cải thiện đáng kể độ chính xác so với việc chỉ sử dụng một phương pháp. Nhưng mô hình cũng gặp khó khăn trong việc đối phó với sự hiếm hoi của đánh giá từ người dùng.

Các nghiên cứu này cung cấp cái nhìn về cách mà các nhà nghiên cứu đã tiếp cận bài toán, thường kết hợp các phương pháp khác nhau để tối ưu hóa hiệu suất của hệ thống gợi ý phim. Tuy nhiên, hạn chế thường gặp phải là việc không thể tận dụng đầy đủ thông tin từ cả Content-Based và Collaborative Filtering hoặc không tích hợp phân tích ý kiến người dùng một cách toàn diện.

2.3 Phương pháp đề xuất giải quyết bài toán

Mô hình Hybrid Recommendation - HR kết hợp CF và CB là một hướng giải quyết phổ biến trong việc xây dựng hệ thống gợi ý xem phim. Giúp tận dụng thông tin đa dạng: CB tập trung vào nội dung của phim như tên phim, thể loại, nội dung để tạo ra đề xuất dựa trên sự tương đồng nội dung; còn CF dựa vào thông tin về đánh giá, thói quen của người dùng để tạo ra các gợi ý dựa trên sự tương đồng giữa người dùng. CB giải quyết được vấn đề cold start (khởi đầu lạnh) bằng cách đề xuất phim dựa trên thuộc tính của chúng mà không cần thông tin từ người dùng; CF cũng vượt qua hạn chế của CB khi có đủ dữ liệu đánh giá từ người dùng (Tian *et al.*, 2019).

Kết hợp hai phương pháp giúp cải thiện chính xác và đa dạng các đề xuất, tạo trải nghiệm cá nhân hóa, giúp người dùng khám phá nội dung mới mà vẫn phù hợp với sở thích cá nhân của họ (Sottocornola *et al.*, 2017). Việc kết hợp cũng mở ra cho việc áp dụng các phương pháp hybrid khác nhau, như kết hợp thông qua mạng neural, sử dụng kỹ thuật học sâu, ...

CHƯƠNG 3
PHƯƠNG PHÁP ĐỀ XUẤT

3.1 Mô hình tổng quát

Mô hình tổng quát giải quyết bài toán Hybrid Recommendation được trình bày theo sơ đồ dưới đây. Trong mô hình này gồm … phần chính :

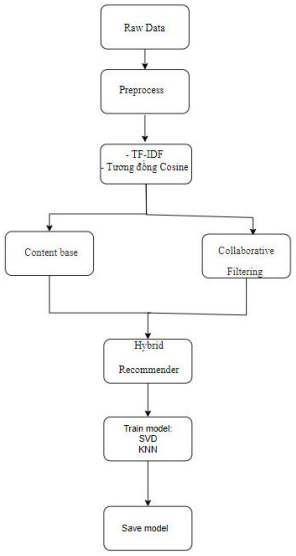
Phần 1: Thu thập dữ liệu

Phần 2: Tiền xử lí dữ liệu

Phần 3: Xây dựng các phương pháp

Phần 4: Kết hợp các phương pháp

Phần 5: Huấn luyện bằng các mô hình: SVD, KNN



Hình 3.1 : Mô hình tổng quan

3.2 Đặc trưng của mô hình đề xuất

3.2.1 Thu thập dữ liệu

Nhóm sử dụng bộ dữ liệu có sẵn, thu thập của Netflix, thu thập thêm ID của phim trên IMDb, TMDb.

3.2.2 Tiền xử lí dữ liệu

Nhóm sẽ tiến hành kiểm tra dữ liệu NaN, thay thế bằng giá trị rỗng để những quá trình xử lí sau không bị lỗi, vì những cột như tên phim, thể loại, nội dung, … không thể tự thêm vào được.

Sau đó tiến hành kết hợp dữ liệu ở các bộ dữ liệu khác nhau do có nhiều bộ dữ liệu, sử dụng hàm pandas.merge để tìm ra nội dung, đánh giá, xếp hạng của từng phim ở bộ dữ liệu chính.

3.2.3. Xây dựng phương pháp

Content-Base

Sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) hoặc kỹ thuật máy học để biểu diễn nội dung của các bộ phim thành các vectơ hoặc ma trận. Các kỹ thuật bao gồm mã hóa từ khóa, phân tích cú pháp văn bản, hoặc kỹ thuật học máy như TF-IDF, Word Embeddings (như Word2Vec hoặc GloVe) để biểu diễn văn bản.

Công thức tính TF-IDF:

TF: $\boldsymbol{TF(t,d)} = \frac{\boldsymbol{sốlầntừtxuấthiệntrongvănbảnd}}{\boldsymbol{Tổngsốtừtrongvănbảnd}}$

IDF: $\boldsymbol{IDF(t,D)} = \boldsymbol{log}\left(\frac{\boldsymbol{Tổngsốvănbảntrongtậphợp}}{\boldsymbol{Sốvănbảnchứatừt+1}}\right)$

TF-IDF: $\boldsymbol{TF-IDF(t,d,D)} = \boldsymbol{TF(t,d)} \times \boldsymbol{IDF(t,D)}$

	A	B	C	D	Item's feature vectors
Jumanji	?	4	2	5	$\boldsymbol{x_1 = [0.85, 0.15]}$
Toy Story	5	4	?	?	$\boldsymbol{x_2 = [0.98, 0.02]}$
Waiting to Exhale	?	4	2	?	$\boldsymbol{x_3 = [0.05, 0.95]}$
Iron Man	3	5	?	5	$\boldsymbol{x_4 = [0.11, 0.89]}$
User's Models	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\theta_3}$	$\boldsymbol{\theta_4}$	$\boldsymbol{\leftarrow need\ to\ optimize}$

Bảng 3.1: Ví dụ về utility matrix với hệ thống Gợi ý phim. Các bộ phim được đánh giá theo từ 0 đến 5 sao. Các dấu '?' ứng với việc dữ liệu chưa tồn tại trong cơ sở dữ liệu. Recommendation Systems cần phải tự điền các giá trị này.

Sử dụng phương pháp tính toán sự tương đồng cosine similarity để đo lường sự tương đồng giữa các vectơ biểu diễn của các bộ phim. Dựa trên sự tương đồng tính được, tạo danh sách các bộ phim tương tự cho mỗi bộ phim trong cơ sở dữ liệu.

Collaborative Filtering

Xây dựng ma trận người dùng-phim, biểu diễn dữ liệu dưới dạng ma trận, trong đó hàng thể hiện người dùng, cột thể hiện các bộ phim và giá trị trong ma trận thể hiện sự tương tác giữa người dùng và phim (ví dụ: đánh giá, nội dung, lịch sử xem). Xác định tương đồng giữa phim bằng phương pháp tính toán tương đồng cosine để xác định sự tương đồng giữa người dùng hoặc giữa các bộ phim.

Một *similiarity function* tốt cần đảm bảo:

$$\boldsymbol{sim(u_0, u_1) > sim(u_0, u_1), \forall_i > 1}$$

Công thức tính Cosine similarity:

$$\boldsymbol{cosine\_similarrity(u_1, u_2) = cos(u_1, u_2) = \frac{{u_1}^T u_2}{|u_1|_2 . |u_2|_2}}$$

```
array([[1.        , 0.01801424, 0.        , ..., 0.        , 0.        ,
        0.01172966],
       [0.01801424, 1.        , 0.04893243, ..., 0.        , 0.00617224,
        0.01162237],
       [0.        , 0.04893243, 1.        , ..., 0.        , 0.00772206,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 1.        , 0.        ,
        0.        ],
       [0.        , 0.00617224, 0.00772206, ..., 0.        , 1.        ,
        0.00489843],
       [0.01172966, 0.01162237, 0.        , ..., 0.        , 0.00489843,
        1.        ]])
```

Hình 3.2: Ma trận Cosine Similarity

3.2.4. Kết hợp các mô hình

Sử dụng mô hình Hybrid kết hợp từ 2 mô hình trên, Content-Based tính toán đề xuất để tạo ra danh sách các bộ phim tương tự dựa trên nội dung của từng bộ phim, sau đó Collaborative Filtering dự đoán sở thích của người dùng cho các bộ phim mà họ chưa xem. Kết hợp danh sách đề xuất từ cả 2 mô hình, sử dụng trọng số để ưu tiên các đề xuất từ một mô hình hoặc kết hợp chúng một cách cân đối. Tinh chỉnh mô hình kết hợp dựa trên phản hồi từ việc đánh giá và kiểm tra hiệu suất của các đề xuất được tạo ra.

3.2.5. Phương pháp huấn luyện

SVD

Mô hình SVD (Singular Value Decomposition) là một phương pháp trong Collaborative Filtering được sử dụng trong hệ thống gợi ý phim. Nó tập trung vào việc giảm chiều dữ liệu và tìm ra các biểu diễn tiềm ẩn của người dùng và

mục tiêu (phim) dựa trên ma trận đánh giá. Cách thức hoạt động của SVD trong hệ thống gợi ý phim:

- Ma trận đánh giá:Bắt đầu với một ma trận, trong đó hàng biểu diễn người dùng, cột biểu diễn các bộ phim, và các giá trị trong ma trận là đánh giá của người dùng cho các bộ phim.
- Decompose ma trận:Áp dụng phép phân rã ma trận SVD để tách ma trận đánh giá thành các ma trận con, bao gồm ba ma trận: U (người dùng), $\boldsymbol{\sigma}$ (độ lớn của giá trị suy biến), và $\boldsymbol{V^T}$ (phim).
- Giảm chiều dữ liệu:Thông thường, chỉ giữ lại một số lượng lớn nhất các giá trị suy biến để giảm chiều dữ liệu. Điều này giúp giảm không gian chiều của ma trận ban đầu và tạo ra biểu diễn tiềm ẩn cho người dùng và phim.
- Dự đoán đánh giá: Sử dụng các ma trận U, $\boldsymbol{\sigma}$ và $\boldsymbol{V^T}$ đã được giảm chiều, áp dụng công thức SVD để dự đoán các đánh giá chưa được người dùng thực hiện cho các bộ phim mà họ chưa xem.
- Kết hợp với mô hình Content-Based:Kết hợp dự đoán từ mô hình SVD với các đề xuất từ mô hình Content-Based để tạo ra danh sách gợi ý phim cuối cùng.

KNN

Mô hình KNN (K-Nearest Neighbors) là một phương pháp trong Collaborative Filtering được sử dụng trong hệ thống gợi ý phim. Tập trung vào dự đoán sở thích của người dùng bằng cách tìm các người dùng tương tự nhau và dựa trên hành vi của họ để đưa ra các đề xuất cho người dùng hiện tại. Cách thức hoạt động của KNN trong hệ thống gợi ý phim:

- Bắt đầu với một ma trận, trong đó hàng biểu diễn người dùng, cột biểu diễn các bộ phim, và các giá trị trong ma trận là đánh giá của người dùng cho các bộ phim.

- Dựa trên mức độ tương đồng của đánh giá, sử dụng độ đo cosine similarity, Pearson correlation, hoặc Euclidean distance để tìm ra K-users tương tự nhất với người dùng hiện tại.
- Khi đã xác định được K-users tương tự, sử dụng hành vi xem phim hoặc đánh giá của họ để dự đoán sở thích của người dùng hiện tại cho các bộ phim mà họ chưa xem hoặc chưa đánh giá.
- Kết hợp dự đoán từ mô hình KNN với các đề xuất từ mô hình Content-Based để tạo ra danh sách gợi ý phim cuối cùng.

CHƯƠNG 4
THỰC NGHIỆM

4.1 Dữ liệu

Bộ dữ liệu xem phim của Netflix, được tham khảo và lấy từ github.

Gồm 3 file csv: file movies_metadata.csv chứa các thông tin, thuộc tính của bộ phim như: tên phim, thể loại, ID phim, thời lượng, trang web chính, ngôn ngữ, mô tả nội dung và 1 vài thông tin liên quan; file links_small.csv có thông tin về ID của phim và ID phim đó trên 2 kho dữ liệu IMDb và TMDb; file ratings_small.csv có ID user, ID của phim user đó đã xem, điểm đánh giá (rating) của phim mà user đó đánh giá. Số dòng …

4.2 Xử lý dữ liệu

Cần tiền hành tiền xử lý dữ liệu, vì thông tin của phim và điểm đánh giá của người dùng nằm ở các file khác nhau. Nhóm đã kiểm tra và xử lý dữ liệu missing, sau đó kết hợp các file lại thông qua ID của user, ID phim trên IMDb và TMDb để được bộ dữ liệu hoàn chỉnh.

4.3 Công nghệ sử dụng

Các công cụ sử dụng giải quyết bài toán

Ngôn ngữ	Python 3.8
Thư viện	Surprise, sklearn, numpy, pandas
Môi trường	Google Colab

Bảng 4.1: Mô tả công nghệ sử dụng cho thực nghiệm bài toán gợi ý phim ảnh

4.4 Cách đánh giá

Nhóm sử dụng độ đo chính là Root Mean Squared Error (RMSE) để đánh giá. RMSE tính độ lớn trung bình của bình phương của sự chênh lệch giữa giá trị dự đoán và giá trị thực tế. Việc lấy căn bậc hai của giá trị giúp chuyển đổi kết quả về cùng đơn vị với dữ liệu gốc.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Trong đó:

- $\hat{y}_i$ là giá trị dự đoán
- y_i là giá trị thực tế
- n là số lượng mẫu dữ liệu

RMSE càng nhỏ, tức là độ chính xác của mô hình dự đoán càng cao. Đây là một trong những độ đo phổ biến trong việc đánh giá hiệu suất của mô hình, đặc biệt trong các bài toán regression và hệ thống gợi ý khi đánh giá sự chính xác của việc dự đoán các đánh giá hoặc xếp hạng của người dùng đối với các mục tiêu như phim.

4.5 Kết quả đạt được

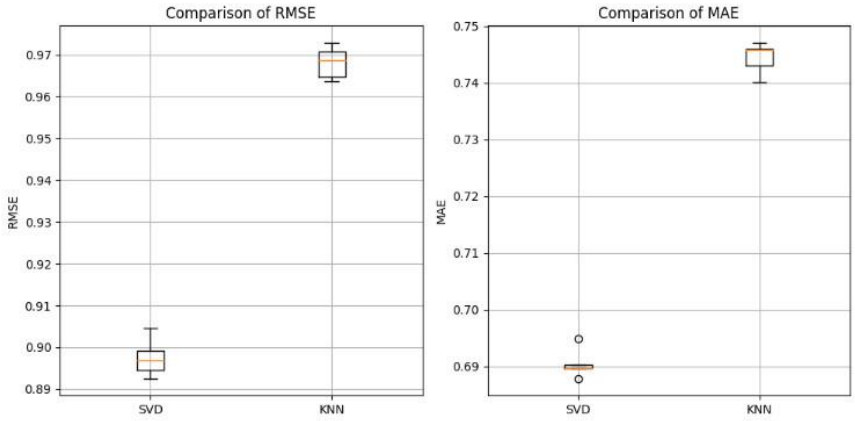
Đồ án sẽ tiến hành thực nghiệm trên tập dữ liệu phim của Netflix như đã trình bày ở phần trước. Tập dữ liệu sẽ được chia ra làm 8 phần cho huấn luyện mạng 2 phần để đánh giá trên tập huấn luyện.

Đồ án sẽ sử dụng mô hình SVD và KNN để áp dụng cho mô hình, sau đó tiến hành so sánh.

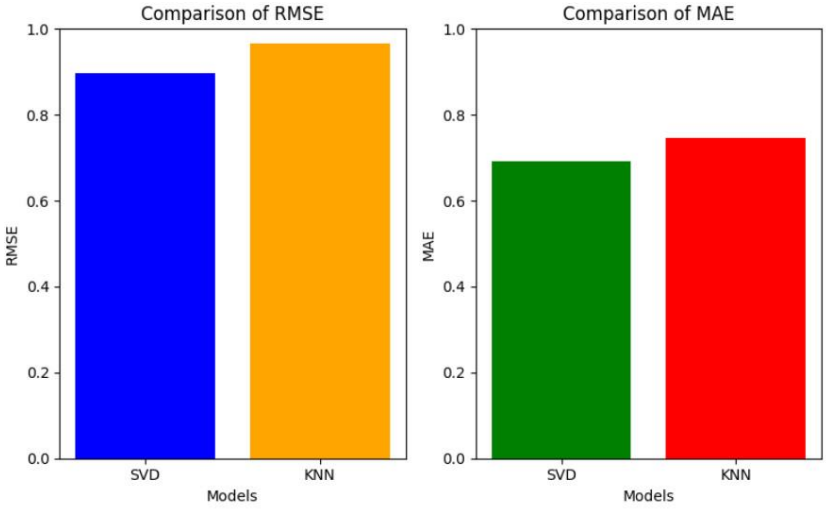
```
RMSE của SVD và KNN
RMSE: 0.8924
RMSE: 0.9664

MAE của SVD và KNN
MAE:  0.6880
MAE:  0.7444
```

Hình 4.1: Kết quả của RMSE và MAE



Hình 4.2: Biểu đồ boxplot của RMSE và MAE



Hình 4.3: Biểu đồ bar của RMSE và MAE

	RMSE	**MAE**
VSD	0.98	0.69
KNN	0.96	0.74

Bảng 4.2: So sánh kết quả thực nghiệm

Hai độ đo trên càng nhỏ, thì cho ra mô hình dự đoán càng chính xác. Kết quả cho thấy mô hình VSD cho ra kết quả tốt hơn so với KNN. VSD thường có hiệu suất tính toán tốt hơn trong việc xử lý dữ liệu lớn hơn so với KNN. VSD có khả năng tốt trong việc dự đoán sở thích của người dùng cho các mục tiêu mà họ chưa tương tác trước đó, đặc biệt là khi có dữ liệu thưa (sparse data).

CHƯƠNG 5
KẾT LUẬN

5.1 Kết luận

5.1.1. Kết quả

Về mặt lý thuyết, Đồ án đã tìm hiểu về các giải pháp cho bài toán gợi ý xem phim, đồng thời Đồ án cũng đề xuất phương pháp kết hợp Hybrid Recommendation cho hệ thống gợi ý xem phim.

Về mặt thực nghiệm, Đồ án đã sử dụng tập dữ liệu phim của Netflix cho mô hình đề xuất cùng với các thuật toán khác nhau của hệ khuyến nghị để so sánh. Kết quả thực nghiệm cho thấy mô hình đã đề xuất mang lại kết quả tốt hơn.

5.1.2. Hạn chế

Việc tích hợp các phương pháp khác nhau có thể đòi hỏi kiến thức chuyên sâu về nhiều thuật toán và phương pháp khác nhau, làm tăng độ phức tạp trong việc triển khai. Việc kết hợp thông tin từ nhiều nguồn khác nhau (như thông tin từ nội dung và thông tin từ hành vi người dùng) cũng có thể gặp khó khăn trong việc thống nhất và xử lý hiệu quả. Chưa tích hợp được phân tích ý kiến người dùng do nhóm còn thiếu kĩ năng, kinh nghiệm và thiếu dữ liệu.

5.2. Hướng phát triển

Recommendation system hay Hệ khuyến nghị vẫn còn khá mới so với nhiều mảng của ML, AI nên có rất nhiều hướng phát triển tiềm năng. Trong tương lai, Đồ án dự kiến sẽ tiếp tục nghiên cứu hướng phát triển chính như sau

- Phát triển, tích hợp các ý kiến cá nhân của người xem phim vào thuật toán để tối ưu hóa trải nghiệm người dùng.

- Nghiên cứu, phát triển các thuật toán và phương pháp tính toán hiệu quả hơn, đặc biệt là trong việc kết hợp nhiều phương pháp.
- Phát triển các phương pháp để diễn giải quyết định của hệ thống gợi ý kết hợp để người dùng có thể hiểu rõ hơn lý do vì sao một mục được gợi ý.

TÀI LIỆU THAM KHẢO

Cantador, I., Castells, P. and Bellogín, A. (2011) 'An enhanced semantic layer for hybrid recommender systems: Application to news recommendation', *International Journal on Semantic Web and Information Systems (IJSWIS)*, 7(1), pp. 44–78.

Koren, Y. (2008) 'Factorization meets the neighborhood: a multifaceted collaborative filtering model', in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434.

Rendle, S. (2010) 'Factorization machines', in *2010 IEEE International conference on data mining*. IEEE, pp. 995–1000.

Sottocornola, G. *et al.* (2017) 'Towards a deep learning model for hybrid recommendation', in *Proceedings of the International Conference on Web Intelligence*, pp. 1260–1264.

Tian, Y. *et al.* (2019) 'College library personalized recommendation system based on hybrid recommendation algorithm', *Procedia CIRP*, 83, pp. 490–494.

PHỤ LỤC

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

LÀM VIỆC NHÓM

Tóm tắt quá trình làm việc nhóm:

1. Đề xuất ý kiến và lên kế hoạch: Bắt đầu bằng việc đề xuất ý kiến và lên kế hoạch công việc cụ thể. Các thành viên sẽ thảo luận và đưa ra ý kiến của mình, sau đó tìm ra giải pháp tốt nhất và xác định mục tiêu cụ thể cho nhóm.

2. Phân chia công việc: Nhóm sẽ phân chia công việc cho từng thành viên dựa trên khả năng và kỹ năng của mỗi người.

3. Gặp nhau và tiến hành làm việc: Mỗi buổi gặp nhau sẽ kéo dài trong khoảng 2 giờ để tiến hành làm việc. Trong các buổi họp, hai bên sẽ báo cáo tiến độ công việc, chia sẻ thông tin và cập nhật nhau về các vấn đề liên quan. Thông qua sự gặp gỡ này, nhóm có thể giải quyết các khó khăn, đánh giá lại kế hoạch và điều chỉnh công việc nếu cần thiết.

4. Đánh giá và hoàn thiện: Sau khi hoàn thành công việc, nhóm sẽ đánh giá kết quả và rút ra bài học từ quá trình làm việc.

Phân chia công việc của các thành viên trong nhóm:

1. Hồ Xuân Bi: Đảm nhiệm việc tìm các bài báo tham khảo, tìm dữ liệu và viết báo cáo, hỗ trợ tìm hiểu code cùng bạn Sơn.

2. Bùi Hoàng Sơn: Đảm nhiệm việc lên kế hoạch, tìm các bài báo tham khảo, tìm dữ liệu và xử lý code, bạn Sơn có hỗ trợ viết báo cáo cùng bạn Bi.

Tổng số lần gặp nhau: 7 buổi

Tổng thời gian gặp nhau: 14 giờ

MỘT SỐ CHÚ Ý KHI VIẾT BÁO CÁO

1. Thống nhất kích cỡ chữ, kiểu chữ trong toàn bộ báo cáo. Không tô màu chữ, chỉ dùng màu đen
2. Các công thức phải tự gõ và đánh số theo Chương, ví dụ 1.1, 2.1, 2.2, 2.3
3. Các hình và Bảng phải đánh số theo chương, ví dụ Hình 1.1, Hình 2.1, Bảng 3.1, Bảng 3.2
4. Các hình nếu lấy ở ngoài phải đề footnote chú thích nguồn ở dưới
5. Hình mô hình tổng quát phải tự vẽ bằng Word, không dán hình
6. Các tài liệu tham khảo phải đính vào luận văn theo thứ tự từ nhỏ tới lớn, bắt đầu từ 1, ít nhất phải từ 5-15 tài liệu tham khảo, lựa chọn các tài liệu tham khảo mới
7. Tóm tắt trình bày được các nội dung sau: giới thiệu, phương pháp làm, kết quả, nhận xét (không dùng hình, bảng ở mục này)

TỰ ĐÁNH GIÁ

Chương	Nội dung	Điểm chuẩn	Tự chấm	Ghi chú
1	**Giới thiệu về bài toán**	0.5	0.5	
2	**Phân tích yêu cầu của bài toán**	1.5	1.5	
3	**Phương pháp giải quyết bài toán**	2.5	2.25	
4	**Thực nghiệm**	4.5	4	
5	**Kết luận**	0.5	0.5	
Nhóm	**Điểm nhóm**	0.5	0.5	
Tổng điểm			9.25	