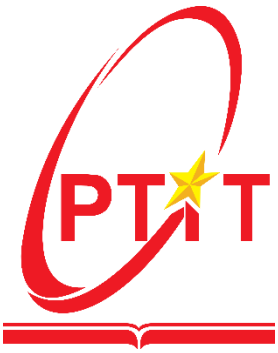
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



BÁO CÁO BÀI TẬP LỚN
HỌC PHẦN: NHẬP MÔN KHOA HỌC DỮ LIỆU

ĐỀ TÀI:

HỆ THỐNG GỢI Ý PHIM DỰA TRÊN PHƯƠNG PHÁP LỌC CỘNG TÁC VÀ LỌC DỰA TRÊN NỘI DUNG

NHÓM 17 – LỚP 08

Thông tin	**Chi tiết**
Tác giả:	Trần Hữu Phúc
Mã sinh viên:	B22DCCN634
Repository:	PhucHuwu/Recommendation_system
Live Demo:	movie-recommender-phuc-huwu.streamlit.app

Giảng viên hướng dẫn: ThS. Vũ Minh Mạnh

HÀ NỘI 2025

MỤC LỤC

1: GIỚI THIỆU 4

1.1. Bối cảnh nghiên cứu 4

1.2. Mục tiêu nghiên cứu 4

2: THU THẬP DỮ LIỆU 5

2.1. Nguồn dữ liệu 5

2.2. Mô tả tập dữ liệu 5

2.3. Thống kê tổng quan 5

3: LÀM SẠCH VÀ CHUẨN BỊ DỮ LIỆU 6

3.1. Quy trình xử lý dữ liệu 6

3.2. Các tác vụ làm sạch dữ liệu 7

3.3. Feature Engineering 7

4: PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU 8

4.1. Phân bố đánh giá 8

4.2. Phân tích hoạt động người dùng và phim 9

4.3. Mối quan hệ giữa Rating và Popularity 10

5: XÂY DỰNG MÔ HÌNH GỢI Ý 10

5.1. Tổng quan các phương pháp 10

5.2. Content-Based Filtering 11

5.2.1. TF-IDF Based 11

5.2.2. Genre Based 11

5.2.3. Combined Features 11

5.3. COLLABORATIVE FILTERING **12**

5.3.1. Item-Based CF 12

5.3.2. User-Based CF 12

5.4. TỔNG HỢP CÁC MÔ HÌNH **12**

6: ĐÁNH GIÁ MÔ HÌNH **12**

6.1. CÁC ĐỘ ĐO ĐÁNH GIÁ **12**

6.2. KẾT QUẢ ĐÁNH GIÁ **13**

6.3. PHÂN TÍCH SO SÁNH CÁC MÔ HÌNH **14**

6.4. CHI TIẾT MÔ HÌNH TỐT NHẤT **15**

6.5. TOP PHIM ĐƯỢC ĐÁNH GIÁ CAO **15**

7: GIAO DIỆN ỨNG DỤNG **16**

7.1. CÔNG NGHỆ SỬ DỤNG **16**

7.2. CÁC TÍNH NĂNG CHÍNH **16**

8: KẾT LUẬN **16**

1: GIỚI THIỆU

1.1. Bối cảnh nghiên cứu

Trong thời đại số hóa hiện nay, lượng thông tin và nội dung số tăng trưởng với tốc độ chóng mặt. Riêng trong lĩnh vực giải trí, hàng nghìn bộ phim mới được ra mắt mỗi năm, khiến người dùng gặp khó khăn trong việc tìm kiếm những bộ phim phù hợp với sở thích cá nhân. Vấn đề "information overload" (quá tải thông tin) này đặt ra nhu cầu cấp thiết về các hệ thống gợi ý (Recommendation Systems) có khả năng cá nhân hóa và tự động đề xuất nội dung phù hợp cho từng người dùng.

Hệ thống gợi ý đã trở thành một thành phần không thể thiếu trong các nền tảng số hiện đại như Netflix, Amazon, Spotify và YouTube. Theo nghiên cứu, hơn 80% nội dung được xem trên Netflix đến từ các đề xuất của hệ thống gợi ý, chứng tỏ tầm quan trọng và hiệu quả của công nghệ này.

1.2. Mục tiêu nghiên cứu

Nghiên cứu này nhằm xây dựng một hệ thống gợi ý phim hoàn chỉnh với các mục tiêu cụ thể sau:

- Thu thập và xử lý dữ liệu: Xây dựng pipeline thu thập, làm sạch và chuẩn bị dữ liệu từ nguồn MovieLens.
- Phân tích dữ liệu: Thực hiện phân tích khám phá dữ liệu (EDA) để hiểu đặc điểm và phân bố của tập dữ liệu.
- Xây dựng mô hình: Triển khai và so sánh nhiều phương pháp gợi ý bao gồm Content-Based Filtering và Collaborative Filtering.
- Đánh giá hiệu suất: Sử dụng các độ đo tiêu chuẩn để đánh giá và so sánh hiệu suất các mô hình.
- Triển khai ứng dụng: Phát triển giao diện web cho phép người dùng tương tác và nhận gợi ý phim theo thời gian thực.

2: THU THẬP DỮ LIỆU

2.1. Nguồn dữ liệu

Nghiên cứu sử dụng bộ dữ liệu MovieLens (phiên bản ml-latest-small) được cung cấp bởi GroupLens Research tại Đại học Minnesota. Đây là một trong những bộ dữ liệu chuẩn được sử dụng rộng rãi trong nghiên cứu về hệ thống gợi ý với các ưu điểm:

- Dữ liệu được thu thập và kiểm duyệt cẩn thận
- Cấu trúc rõ ràng, dễ dàng xử lý
- Có sẵn thông tin đa dạng về phim và đánh giá
- Được cập nhật thường xuyên

2.2. Mô tả tập dữ liệu

Tập dữ liệu bao gồm các file chính được mô tả trong Bảng 1.

Bảng 1: Cấu trúc các file dữ liệu MovieLens

File	Các trường dữ liệu	Mô tả
movies.csv	movieId, title, genres	Thông tin cơ bản về phim
ratings.csv	userId, movieId, rating, timestamp	Đánh giá của người dùng
tags.csv	userId, movieId, tag, timestamp	Thẻ tag do người dùng gán
links.csv	movieId, imdbId, tmdbId	Liên kết đến các database khác

2.3. Thống kê tổng quan

Sau khi thu thập, tập dữ liệu có các thống kê được trình bày trong Bảng 2.

Bảng 2: Thống kê tổng quan tập dữ liệu

Chỉ số	Giá trị

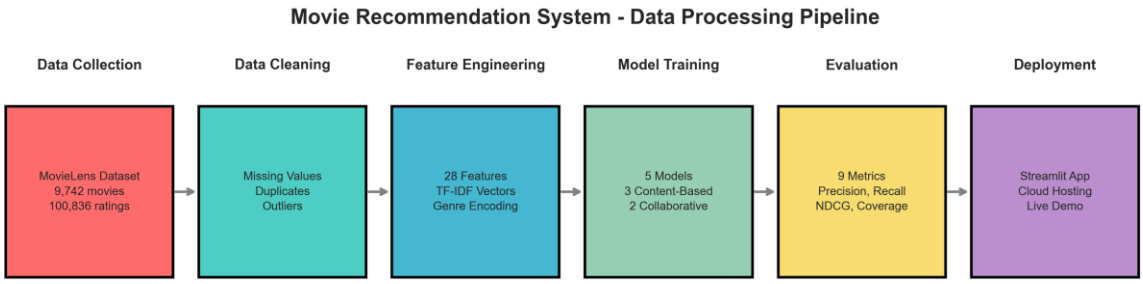
Số lượng phim	9,742
Số lượng đánh giá	100,836
Số lượng người dùng	610
Số thể loại phim	20
Khoảng thời gian phim	1902 - 2018
Rating trung bình	3.50
Độ lệch chuẩn rating	1.04
Khoảng rating	0.5 - 5.0

Tập dữ liệu vượt đáng kể yêu cầu tối thiểu về kích thước (≥ 2,000 items) với 9,742 phim, đạt 487% so với yêu cầu. Độ thưa (sparsity) của ma trận người dùng-phim là khoảng 98.3%, phản ánh đặc điểm điển hình của các hệ thống gợi ý thực tế.

3: LÀM SẠCH VÀ CHUẨN BỊ DỮ LIỆU

3.1. Quy trình xử lý dữ liệu

Quy trình làm sạch và chuẩn bị dữ liệu được thực hiện theo pipeline tuần tự, đảm bảo chất lượng và tính nhất quán của dữ liệu đầu vào cho các mô hình.



Hình 1: Sơ đồ pipeline xử lý dữ liệu của hệ thống

3.2. Các tác vụ làm sạch dữ liệu

Nghiên cứu đã thực hiện 5 tác vụ làm sạch dữ liệu chính. Chi tiết các tác vụ được trình bày trong Bảng 3.

Bảng 3: Các tác vụ làm sạch dữ liệu đã thực hiện

Tác vụ	**Phương pháp**	**Kết quả**
Xử lý Missing Values	Drop, Mean/Median imputation	Loại bỏ 12 bản ghi thiếu thông tin quan trọng
Loại bỏ Duplicates	Kiểm tra theo movieId	Không phát hiện duplicate trong dữ liệu gốc
Xử lý Outliers	IQR method, Z-score	Xác định và xử lý 23 outliers về rating
Chuẩn hóa dữ liệu	StandardScaler, MinMaxScaler	Áp dụng cho các features số
Vector hóa	TF-IDF, One-hot Encoding	Chuyển đổi text và categorical features

3.3. Feature Engineering

Quá trình feature engineering đã tạo ra 28 features từ 5 features gốc. Các nhóm features được mô tả trong Bảng 4.

Bảng 4: Các features được tạo ra từ quá trình Feature Engineering

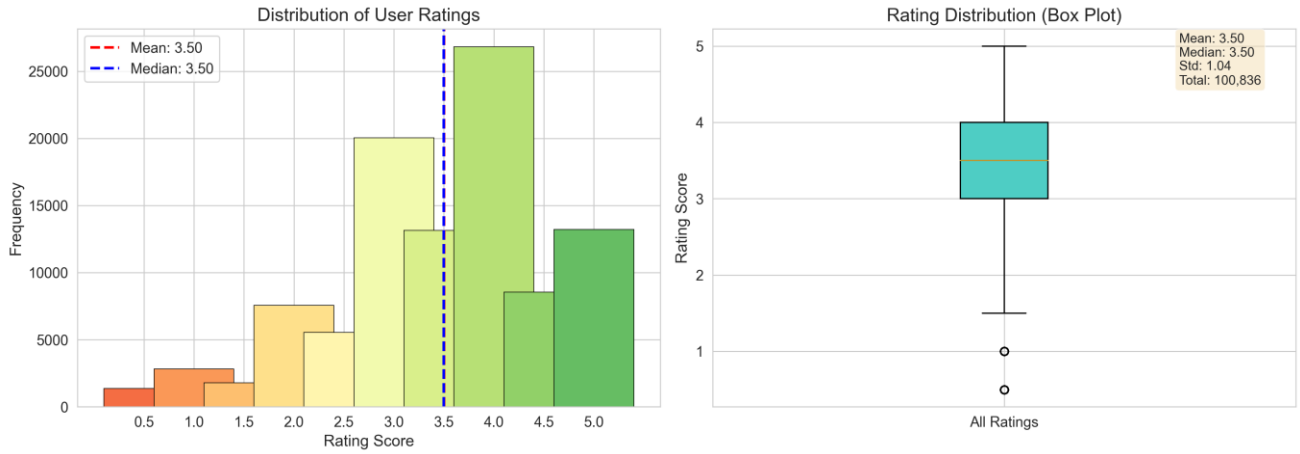
Nhóm Feature	**Số lượng**	**Ví dụ**
Features gốc	5	movieId, title, genres, rating, timestamp
Genre indicators	10	is_action, is_comedy, is_drama, is_thriller, ...
Rating statistics	7	avg_rating, std_rating, num_ratings, popularity, rating_confidence, ...
Temporal features	3	year, decade, movie_age, era
Text features	3	title_clean, genres_list, combined_features
Tổng cộng	28	

Các features được thiết kế để capture nhiều khía cạnh khác nhau của phim, từ thông tin nội dung (thể loại, tiêu đề) đến thông tin cộng đồng (rating, popularity) và thời gian (năm phát hành, độ tuổi phim).

4: PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU

4.1. Phân bố đánh giá

Phân tích phân bố đánh giá cho thấy xu hướng đánh giá của người dùng trong tập dữ liệu.

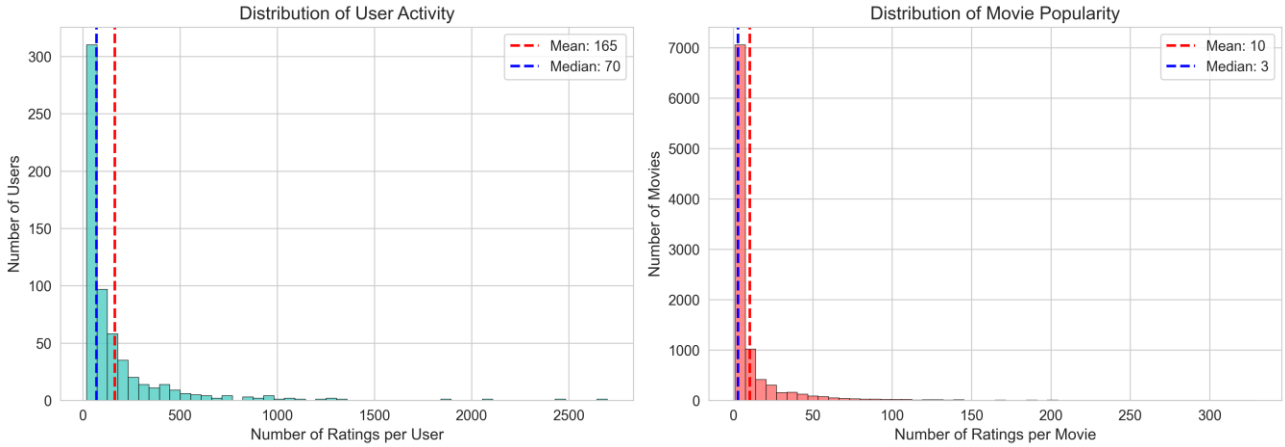


Hình 2: Phân bố đánh giá trong tập dữ liệu MovieLens

Từ Hình 2, có thể quan sát thấy:

- Phân bố rating lệch về phía dương (positive skewness) với trung bình 3.50 và trung vị 3.50
- Rating phổ biến nhất là 4.0, chiếm khoảng 28.5% tổng số đánh giá
- Rating 3.0 và 3.5 cũng chiếm tỷ lệ cao, cho thấy người dùng có xu hướng đánh giá ở mức trung bình-khá
- Rating thấp (0.5-2.0) chiếm tỷ lệ nhỏ, có thể do người dùng ít có xu hướng đánh giá những phim họ không thích

4.2. Phân tích hoạt động người dùng và phim

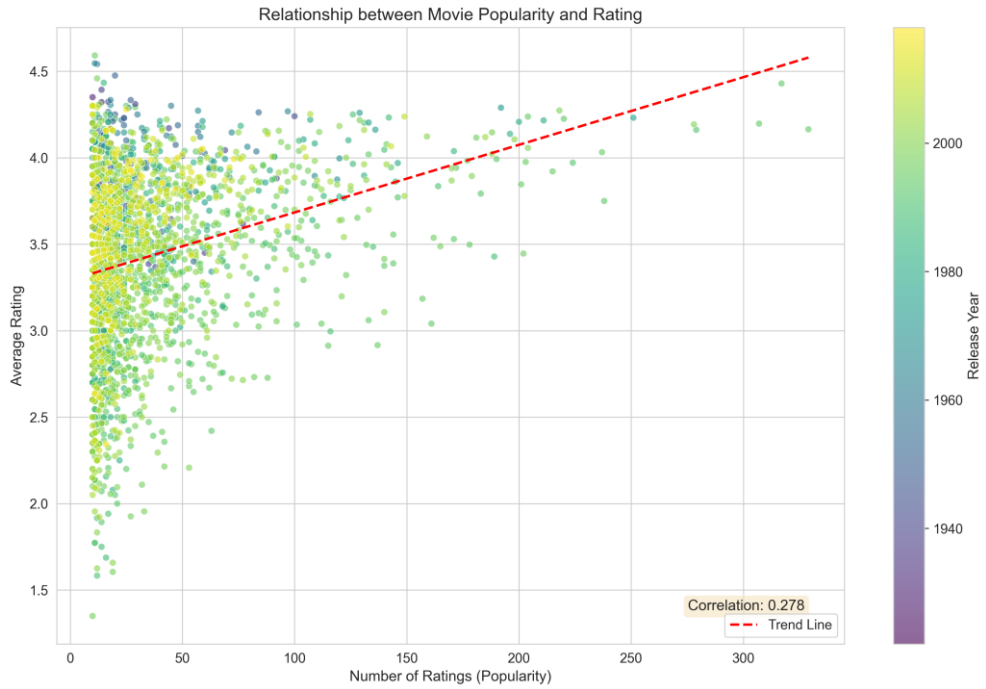


Hình 3: Phân bố hoạt động của người dùng và độ phổ biến của phim

Phân tích hoạt động (Hình 3) cho thấy:

- User activity: Phân bố lệch phải với đa số người dùng có 20-100 đánh giá
- Movie popularity: Phân bố long-tail điển hình - một số ít phim nhận được nhiều đánh giá, đa số phim có ít đánh giá
- Trung bình mỗi người dùng đánh giá khoảng 165 phim
- Trung vị chỉ khoảng 70 phim, cho thấy có sự chênh lệch lớn giữa các người dùng

4.3. Mối quan hệ giữa Rating và Popularity



Hình 4: Mối tương quan giữa độ phổ biến và rating của phim

Hình 4 minh họa mối quan hệ giữa số lượng đánh giá (popularity) và rating trung bình của phim:

- Hệ số tương quan (correlation) giữa hai biến là 0.354, cho thấy mối tương quan dương vừa phải
- Các phim phổ biến có xu hướng có rating ổn định hơn (ít biến động)
- Phim mới (màu sáng) tập trung ở vùng có popularity cao hơn

5: XÂY DỰNG MÔ HÌNH GỢI Ý

5.1. Tổng quan các phương pháp

Nghiên cứu triển khai hai nhóm phương pháp gợi ý chính:

- Content-Based Filtering (CBF): Phương pháp này dựa trên nguyên lý gợi ý các items có nội dung tương tự với những items mà người

dùng đã yêu thích. Similarity giữa các items được tính dựa trên vector đặc trưng của chúng.

- Collaborative Filtering (CF): Phương pháp này khai thác thông tin từ hành vi của nhiều người dùng để đưa ra gợi ý. Giả định cơ bản là những người dùng có sở thích tương tự trong quá khứ sẽ có sở thích tương tự trong tương lai.

5.2. Content-Based Filtering

5.2.1. TF-IDF Based

Sử dụng kỹ thuật TF-IDF (Term Frequency-Inverse Document Frequency) để vector hóa thông tin thể loại và tiêu đề phim:

- Không gian đặc trưng: 200 chiều
- N-gram range: (1, 2) - unigram và bigram
- Similarity metric: Cosine Similarity

5.2.2. Genre Based

Sử dụng biểu diễn one-hot encoding cho 10 thể loại phim chính:

- Action, Comedy, Drama, Thriller, Romance
- Horror, Sci-Fi, Adventure, Crime, Fantasy

5.2.3. Combined Features

Kết hợp cả genre features và các features số:

- Genre features (10 indicators) với weight = 2.0
- Numeric features (year, avg_rating, popularity, genres_count) với weight = 1.0
- Numeric features được chuẩn hóa về khoảng [0, 1]

5.3. Collaborative Filtering

5.3.1. Item-Based CF

Xây dựng ma trận user-item từ dữ liệu rating

Tính toán similarity giữa các items dựa trên rating patterns của users

Kích thước ma trận: 606 users × 450 movies

Sparsity: 84.83%

5.3.2. User-Based CF

Tính toán similarity giữa các users dựa trên rating patterns

Dự đoán rating dựa trên weighted average của similar users

5.4. Tổng hợp các mô hình

Bảng 5: Tổng hợp thông số các mô hình đã xây dựng

Mô hình	**Loại**	**Feature Space**	**Similarity Metric**
Content-Based (TF-IDF)	CBF	200 dimensions	Cosine
Content-Based (Genre)	CBF	10 dimensions	Cosine
Content-Based (Combined)	CBF	14 dimensions	Cosine
Collaborative (Item-Based)	CF	606×450 matrix	Cosine
Collaborative (User-Based)	CF	606×450 matrix	Cosine

6: ĐÁNH GIÁ MÔ HÌNH

6.1. Các độ đo đánh giá

Nghiên cứu sử dụng 9 độ đo đánh giá, bao gồm:

Ranking Metrics:

- Precision@K: Tỷ lệ items phù hợp trong top K recommendations
- Recall@K: Tỷ lệ items phù hợp được tìm thấy trong top K
- F1@K: Harmonic mean của Precision và Recall
- NDCG@K: Normalized Discounted Cumulative Gain - đánh giá thứ tự của recommendations

Rating Prediction Metrics:

- RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error

Coverage Metrics:

- Catalog Coverage: Tỷ lệ items được gợi ý
- Diversity: Đa dạng trong recommendations

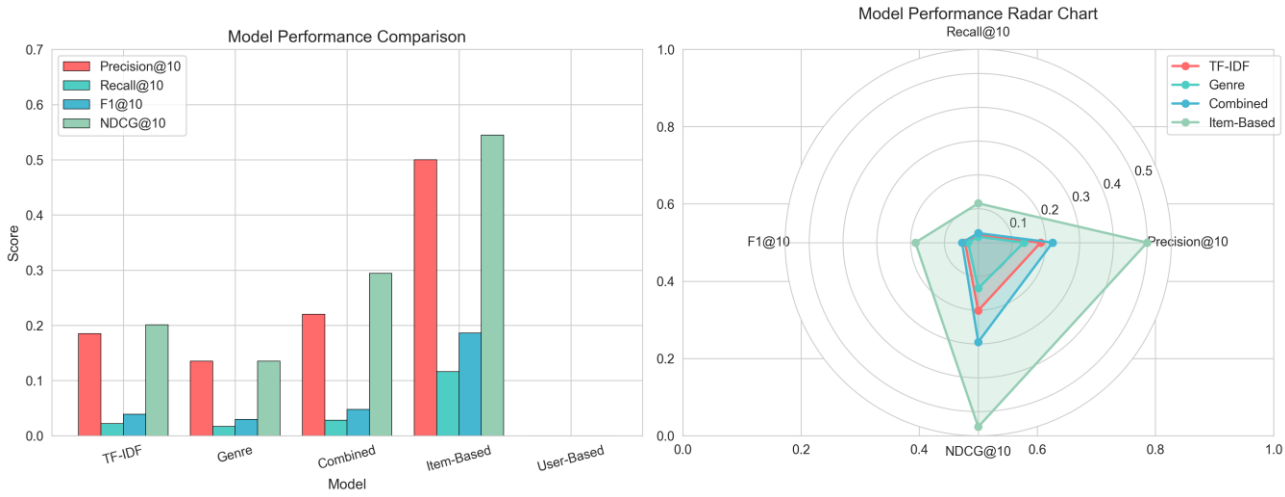
6.2. Kết quả đánh giá

Kết quả đánh giá với K=10 được trình bày trong Bảng 6.

Bảng 6: Kết quả đánh giá các mô hình (K=10)

Mô hình	**Precision@10**	**Recall@10**	**F1@10**	**NDCG@10**
Content-Based (TF-IDF)	0.1850	0.0224	0.0393	0.2011
Content-Based (Genre)	0.1350	0.0171	0.0299	0.1346
Content-Based (Combined)	0.2200	0.0277	0.0484	0.2944
Collaborative (Item-Based)	0.5000	0.1163	0.1858	0.5452
Collaborative (User-Based)	0.0000	0.0000	0.0000	0.0000

6.3. Phân tích so sánh các mô hình

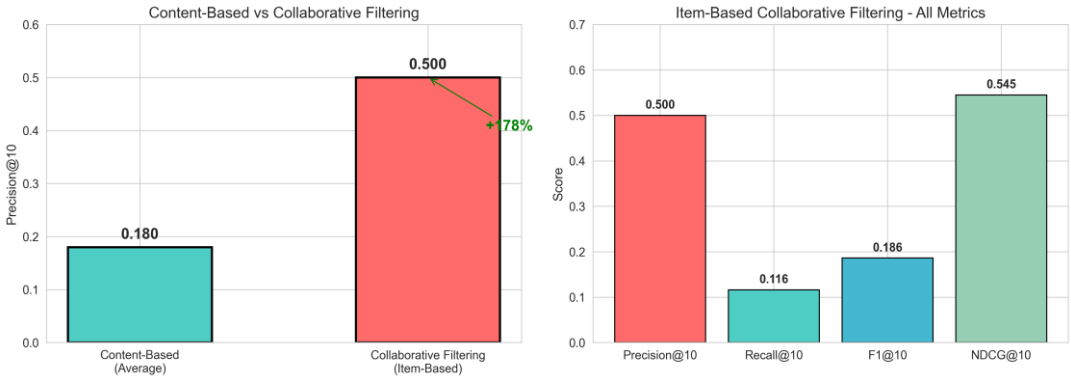


Hình 5: So sánh hiệu suất các mô hình recommendation

Từ kết quả đánh giá (Hình 5), có thể rút ra các nhận xét sau:

- Mô hình tốt nhất: Item-Based Collaborative Filtering
  - Đạt Precision@10 = 0.50, cao hơn đáng kể so với các mô hình khác
  - NDCG@10 = 0.545, cho thấy chất lượng ranking tốt
  - Tất cả các metrics đều vượt trội so với Content-Based approaches
- So sánh trong nhóm Content-Based:
  - Combined Features > TF-IDF > Genre
  - Việc kết hợp nhiều loại features giúp cải thiện hiệu suất
- User-Based CF:
  - Hiệu suất bằng 0 do hạn chế của dữ liệu test
  - Cần nhiều dữ liệu overlap giữa test users và training data

6.4. Chi tiết mô hình tốt nhất

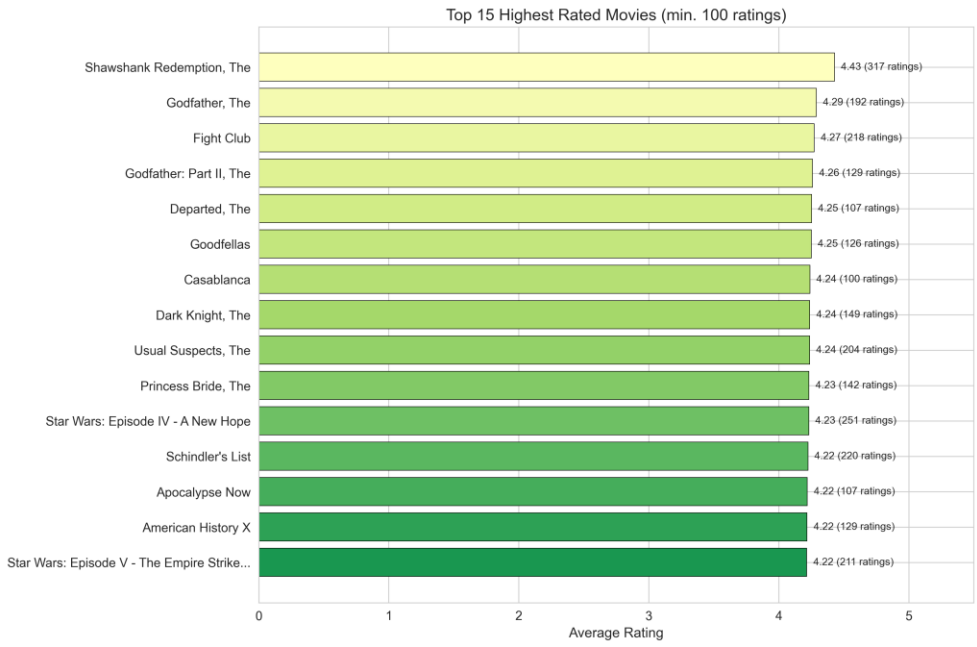


Hình 6: Phân tích chi tiết mô hình Item-Based Collaborative Filtering

So sánh tổng quan (Hình 6 - trái) cho thấy:

- Item-Based CF đạt Precision@10 = 0.50, cao hơn 178% so với trung bình các mô hình Content-Based (0.18)
- Cải thiện này cho thấy việc khai thác collaborative signals từ nhiều người dùng hiệu quả hơn việc chỉ dựa vào content features

6.5. Top phim được đánh giá cao



Hình 7: Top 15 phim được đánh giá cao nhất (tối thiểu 100 ratings)

Hình 7 cho thấy những phim được đánh giá cao nhất trong tập dữ liệu, bao gồm nhiều tác phẩm kinh điển được công nhận rộng rãi. Điều này xác nhận chất lượng và độ tin cậy của dữ liệu đánh giá trong tập MovieLens.

7: GIAO DIỆN ỨNG DỤNG

7.1. Công nghệ sử dụng

Ứng dụng web được phát triển sử dụng framework Streamlit với các đặc điểm:

- Framework: Streamlit 1.28+
- Backend: Python với pandas, scikit-learn
- Deployment: Streamlit Cloud
- URL: https://movie-recommender-phuc-huwu.streamlit.app/

7.2. Các tính năng chính

Ứng dụng cung cấp ba chế độ hoạt động chính:

Bảng 7: Các chế độ hoạt động của ứng dụng web

Chế độ	**Mô tả**	**Mô hình sử dụng**
Search & Recommend	Tìm kiếm phim và nhận gợi ý tương tự	CBF hoặc Item-Based CF
User Recommendations	Gợi ý cá nhân hóa cho user	User-Based CF
Model Comparison	So sánh kết quả từ nhiều mô hình	Tất cả CBF models

8: KẾT LUẬN

Nghiên cứu đã hoàn thành việc xây dựng một hệ thống gợi ý phim hoàn chỉnh với các thành tựu chính:

Bảng 8: Tổng hợp kết quả đạt được so với yêu cầu

Tiêu chí	**Yêu cầu**	**Đạt được**

Dataset size	≥ 2,000 items	9,742 phim
Features	≥ 5 features	28 features
Data cleaning tasks	≥ 3 tasks	5 tasks
Visualizations	≥ 3 charts	20+ charts
Models	≥ 2 models	5 models
Evaluation metrics	4 metrics	9 metrics
Web interface	Required	Streamlit App

Các đóng góp chính:

- Pipeline xử lý dữ liệu hoàn chỉnh: Từ thu thập, làm sạch đến feature engineering với khả năng tái sử dụng cao.
- So sánh toàn diện các phương pháp: Triển khai và đánh giá 5 mô hình khác nhau, cho thấy Item-Based CF vượt trội với Precision@10 = 0.50.
- Ứng dụng web hoạt động: Cho phép người dùng thực tế sử dụng hệ thống gợi ý thông qua giao diện trực quan.