So sánh các mô hình gợi ý trên bộ dữ liệu Anime

Nguyễn Đoàn Kiều Liên - 18520298[1], Nguyễn Xuân Lộc - 18520087[1], Trần Xuân Thanh Mai - 18520098[1], and Ngô Tường Vy - 18520196[1]

Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh
{18520298, 18520087, 18520098, 18520196}@gm.uit.edu.vn

Tóm tắt nội dung : Anime là một hình thức truyền thông hoạt hình có nguồn gốc gắn liền với Nhật Bản. Một xu hướng gần đây của Google tiết lộ rằng có từ 10-100 triệu lượt tìm kiếm các chủ đề liên quan đến anime mỗi tháng. Con số này chỉ mới đạt đến đỉnh điểm trong những tháng gần đây do kết quả của lệnh cách ly trên toàn quốc và những nỗ lực tiếp theo nhằm tìm kiếm một phương tiện giải trí. Mục tiêu của nhóm trong báo cáo này là xây dựng các mô hình gợi ý những bộ phim phù hợp nhất dựa trên nhu cầu, mong muốn và sở thích của mỗi cá nhân người dùng. Với bộ dữ liệu về Anime được đề xuất từ 76.000 người dùng trên hơn 12.000 bộ phim tại myanimelist.net, nhóm thực hiện xây dựng những mô hình dựa trên ba phương pháp tiếp cận là: DemographicFiltering, Content-Based Filterting và Collaborative Filtering.

Keywords: Apache PySpark · hệ thống gợi ý · anime recommendation · Demographic Filtering · Content-Based Filterting · Collaborative Filtering

1 Giới thiệu

Các trang thương mại điện tử cũng như các trang web phục vụ cho nhu cầu giải trí xuất hiện để đáp ứng nhu cầu của người dùng, tuy nhiên số lượng các trang ngày càng nhiều khiến việc cạnh tranh trở nên khốc liệt. Để giữ chân được khách hàng, các trang cần phải phát triển các tính năng mới để thu hút khách hàng, nhờ đó mà các hệ thống gợi ý bắt đầu được chú ý tới và đầu tư để phát triển nhiều hơn.

Hiện nay hệ thống gợi ý được thường được sử dụng 3 loại: Demographic Filtering, Content-Based Filterting và Collaborative Filtering.

- Demographic Filtering: Đây là loại gợi ý chung cho tất cả người dùng, dựa vào độ nổi tiếng của phim hay thể loại đó.
- Content-Based Filterting: Là một loại gợi ý cho từng người dùng dựa trên nội dung, thể loại hay tác giả/diễn viên, hoặc nếu về sản phẩm thì sẽ là chất liệu tạo ra sản phẩm để từ đó đưa ra gợi ý phù hợp cho người dùng.
- Collaborative Filtering: Là hệ thống gợi ý được sử dụng nhiều hiện nay. Hệ thống sẽ dựa vào lựa chọn của các người dùng tương tự để đưa ra gợi ý. Nếu người dùng A có các xu hướng lựa chọn hay mua sản phẩm nào đó, và người B cũng có xu hướng giống người A, lúc này hệ thống sẽ gợi ý các sản phẩm người A đã mua và hài lòng cho người B.

Trong báo cáo này nhóm sẽ tập trung vào việc xây dựng các mô hình gợi ý về phim ảnh trên bộ dữ liệu Anime Recommendation Dataset trên Kaggle. Thông qua các mô hình gợi ý, người dùng có thể dễ dàng tiếp cận đến các bộ phim phù hợp với sở thích của mình mà mình chưa hề biết đến. Bài báo cáo sẽ được chia thành 5 phần: Phần 1 sẽ giới thiệu về đề tài của báo cáo cũng như thông tin sơ lược về hệ thống gợi ý; Phần 2 sẽ phân tích về bộ dữ liệu Anime Recommendation; Phần 3 sẽ đi sâu vào các bước tiền xử lý cũng như cách cài đặt của các mô hình này mà nhóm đã thực hiện; Phần 4 là kết quả và so sánh giữa một số mô hình đã được thực thi của nhóm và cuối cùng là kết luận.

2 Bộ dữ liệu

2.1 Giới thiệu bộ dữ liệu

Bộ dữ liệu của chúng tôi là Anime Recommendations Database [1] được tải về từ Kaggle. Bộ dữ liệu này chứa thông tin từ 73.516 người dùng trên 12.294 bộ anime. Mỗi người dùng có thể thêm anime vào danh sách đã xem của họ và xếp hạng các bộ phim này. Bộ dữ liệu bao gồm 2 tập dữ liệu là Anime và Rating. Tập dữ liệu Anime bao gồm 7 cột: anime_id, name (tên bộ phim), genre (danh sách các thể loại của bộ phim), type (loại hình chiếu: Tivi, movie, v.v.), episodes (số tập), rating (xếp hạng trung bình của bộ phim trên thang điểm 10), members (số người theo dõi bộ phim). Tập dữ liệu Rating thể hiện đánh giá của mỗi người dùng đối với mỗi bộ phim mà người đó đã xem. Tập dữ liệu này có 3 côt: user_id, anime_id, rating (đánh giá của người dùng về bộ phim trên thang điểm 10, số điểm -1 thể hiện người đó đã xem phim nhưng chưa đánh giá)

Bảng 1: Ví dụ các dòng trong bộ dữ liệu anime

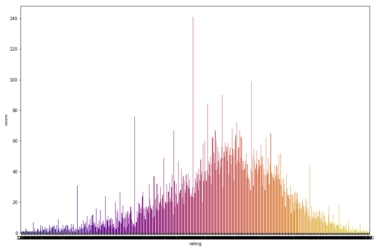
anime_id	name	genre	type	episodes	rating	members
32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
9253	Steins;Gate	Sci-Fi, Thriller	TV	24	9.17	673572
820	Ginga Eiyuu Densetsu	Drama, Military, Sci-Fi, Space	OVA	110	9.11	80679

Bảng 2: Ví dụ các dòng trong bộ dữ liệu rating

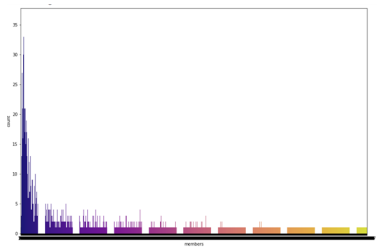
user_id	anime_id	rating
1	8074	10
2	11771	10
1	1692	-1

2.2 Phân tích sơ lược bộ dữ liệu

Tập dữ liệu Anime bao gồm 12294 dòng dữ liệu, trong đó có 230 dòng dữ liệu rỗng ở cột rating. Rating trung bình của các bộ phim là 6.5, bộ phim được đánh giá cao nhất là Taka no Tsume 8: Yoshida-kun no X-Files có rating bằng 10, bộ phim bị đánh giá thấp nhất là Platonic Chain: Ansatsu Jikkouchuu có rating bằng 1.67. Hình 1 thể hiện phân bổ rating của 12294 bộ phim.

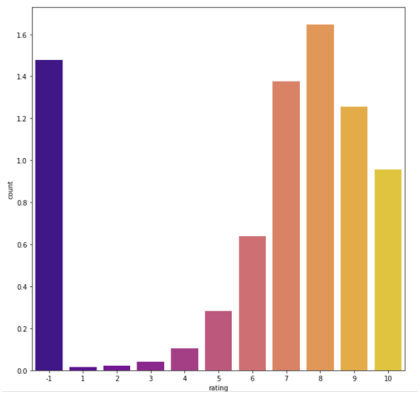


Hình 1: Phân bổ rating của 12294 bộ phim



Hình 2: Lượt người xem của các bộ Anime

Các bộ phim có trung bình khoảng 18070 người xem. Bộ phim có lượt người xem thấp nhất là Gou-chan: Moko to Chinjuu no Mori no Nakama-tachi với 5 lượt xem, bộ phim có lượt người xem cao nhất là Death Note với 1013917 lượt người xem. Hình 2 thể hiện lượt người xem của các bộ Anime. Đối với tập dữ liệu Rating, rating người dùng cho các bộ phim anime trung bình là 7.8 (không tính các hang có giá trị -1 vì các bộ đó chưa được người dùng xếp hạng), rating cao nhất người dùng cho 1 bộ phim là 10, rating thấp nhất người dùng cho 1 bộ phim là 1. Hình 3 thể hiện rating của người dùng cho các bộ anime.



Hình 3: Phân bố của các rating trong bộ dữ liệu

3 Tiền xử lý và Cài đặt

Trong báo cáo lần này nhóm sẽ thực thi 3 mô hình Demographic, Content-based và Collaborative Filtering. Ngoài ra, nhóm sẽ thực thi thêm một mô hình kết hợp giữa Demographic và Collaborative Filtering để tăng tính dự đoán của mô hình. Dưới đây là phần thiết lập và cài đặt 3 mô hình.

3.1 Demographic Filtering

Tiền xử lý: Mô hình Demographic Filtering sẽ sử dụng cả hai bộ dữ liệu Rating và Anime. Bên cạnh việc loại bỏ các dòng dữ liệu bị rỗng hoặc khuyết một số giá trị, đối với bộ dữ liệu rating sẽ thực thi việc loại bỏ các dòng dữ liệu rating = -1 (tương ứng với người dùng không đánh giá về phim). Tiếp theo, sẽ tiến hành thống kê số lượng bình chọn đối với từng phim. Cuối cùng, nhóm sẽ gộp bộ dữ liệu thống kê số lượng đánh giá phim với bộ dữ liệu anime để có một bộ dữ liệu như bảng dưới đây.

Bảng 3: Ví dụ về cơ sở dữ liệu đầu vào cho Demographic Filtering

anime_id	name	rating	count
8516	Baka to Test to Shoukanjuu Ni	7.98	6548
20965	Wonder Garden	6.48	55
9253	Steins;Gate	9.17	17151
6547	Angel Beats!	8.39	23565
18029	Shingeki no Kyojin	8.54	25290

Đầu vào cần thiết sẽ bao gồm 2 cột **rating**: thể hiện đánh giá trung bình của tất cả người dùng về bộ phim và cột **count**: thể hiện số lượng người dùng thực hiện đánh giá phim.

Cài đặt: Demographic Filtering được sử dụng đối với các người dùng mới, khi hệ thống chưa có dữ liệu gì để có thể dự đoán theo sở thích cho người dùng. Lúc này ta có thể sử dụng mô hình Demographic Filtering để đưa ra các gợi ý, ví dụ như ở đây mô hình của nhóm sẽ đưa ra gợi ý các bộ phim nổi tiếng trong các người dùng đang sử dụng ứng dụng. Công thức hỗ trợ cho việc tính toán và sắp xếp của mô hình này sẽ dựa vào công thức weighted rating của IMDB [3].

$$WR = (\frac{v}{v+m} * R) + (\frac{v}{v+m} * C) \quad (1)$$

Trong đó,
WR (weighted rating): số điểm được sử dụng để đánh giá điểm trung bình thực tế của phim.
v: số lượng người dùng đánh giá phim.
m: số lượt đánh giá tối thiểu để có thể nằm trong danh sách gợi ý cho người dùng.

R: điểm của phim được đánh giá bởi người dùng.
C: điểm đánh giá trung bình của toàn bộ bộ dữ liệu.

Đối với số lượt đánh giá tối thiểu (m) sẽ được xác định bằng việc tính toán 90% số lượng đánh giá tối đa trong bộ dữ liệu. Dựa vào weighted rating mới được tính toán phía trên, nhóm sẽ có các rating mới được dùng để sắp xếp thứ tự gợi ý cho các người dùng mới.

3.2 Content-based Filtering

Tiền xử lý: Tất cả các hàng có giá trị null trong một hay nhiều cột sẽ bị loại bỏ, các cột genre và type được biến đổi thành các array để thuận tiện bước tiếp theo, cụ thể là biến đổi hai cột này theo hướng MultiLabelBinarizer, mỗi array sẽ biến thành các giá trị nhị phân 0 – 1 cho biết sự hiện diện của các giá trị có trong array (0 là không và 1 là có).

Cài đặt: Nhóm sử dụng thuật toán KNN trong mô hình Content-based cho phép người dùng nhập vào một bộ phim tùy ý mà họ thích, sau đó đề xuất một số bộ phim có nhiều điểm tương đồng nhất với bộ phim đầu vào. Để mô hình hóa điều này, nhóm sẽ trích xuất một vector đặc trưng A từ bộ phim mà người dùng nhập vào và tính KNN dựa trên vector này để có thể nắm bắt được bản chất đặc trưng về sở thích của người dùng đối với bộ phim đó. Nhóm sử dụng các thông tin có sẵn trong bộ dữ liệu để trích xuất nên vector đặc trưng từ các bộ phim: Thứ nhất là thể loại (genre) đã được gán nhãn cho từng bộ phim, thứ hai là số điểm đánh giá (rating) của bộ phim đó, thứ ba là số tập phim (episodes) và cuối cùng là loại hình thức chiếu bộ phim đó (type).

Có hai công thức để tính độ tương đồng giữa các bộ phim mà nhóm đã xem xét trong việc thiết lập mô hình. Công thức đầu tiên được nhóm sử dụng là công thức tính độ tương tự Cosine. Công thức Cosine được biểu diễn như sau:

$$\cos\theta = \frac{\overrightarrow{A}.\overrightarrow{B}}{||\overrightarrow{A}||.||\overrightarrow{B}||} \tag{2}$$

Công thức Cosine là công thức được sử dụng để đo độ tương tự (measure of similarity) giữa hai vectơ trong một không gian tích vô hướng. Trong đó,
A: vector đặc trưng được trích xuất từ bộ phim mà người dùng nhập vào.
B: vector đặc trưng được trích xuất từ từng bộ phim có trong bộ dữ liệu.

Để việc so sánh trở nên dễ dàng hơn, chúng ta phải thay đổi công thức Cosine sao cho giá trị Cosine thấp hơn sẽ tương ứng với góc thấp hơn. Thì ở đây, nhóm tác giả gốc mà nhóm tham khảo đã giải quyết vấn đề này bằng cách thay đổi lại công thức đại diện cho công thức Cosine như sau:

$$1 - \cos\theta = 1 - \frac{\overrightarrow{A}.\overrightarrow{B}}{||\overrightarrow{A}||.||\overrightarrow{B}||} \tag{3}$$

Trong đó,
A: vector đặc trưng được trích xuất từ bộ phim mà người dùng nhập vào.
B: vector đặc trưng được trích xuất từ từng bộ phim có trong bộ dữ liệu.

Công thức thứ hai được nhóm sử dụng là công thức Euclide. Công thức Euclide được biểu diễn như sau:

$$d(x, y) = \sqrt{||\overrightarrow{A} - \overrightarrow{B}||^2} \tag{4}$$

Trong đó,
A: vector đặc trưng được trích xuất từ bộ phim mà người dùng nhập vào.
B: vector đặc trưng được trích xuất từ từng bộ phim có trong bộ dữ liệu.

Khác với công thức Cosine, công thức Euclide tương tự như việc đo khoảng cách thực giữa hai vector, và do đó kết quả thu được sẽ bị ảnh hưởng bởi góc và độ lớn của các vector. Ở đây, nhóm triển khai công thức Euclide như một phép đo thay thế vì nhóm muốn xem xét kết quả trả về từ các công thức tính khác nhau sẽ có sự khác biệt như thế nào.

3.3 Collaborative Filtering

Tiền xử lý: Đối với mô hình Collaborative Filtering và mô hình kết hợp giữa Demographic và Collaborative Filtering, bộ dữ liệu Rating sẽ được chia thành ba phần là training, valid và test theo tỉ lệ 7:2:1. Bên cạnh đó các dòng dữ liệu có cột rating = -1 (người dùng đã xem phim nhưng không đánh giá) sẽ bị loại bỏ.

Cài đặt: Sử dụng module pyspark.ml.recommendation ta có thể dễ dàng xây dựng một mô hình recommendation theo phương pháp ALS.

Các tham số khi khởi tạo mô hình đó là:

- userCol: tên cột chứa id của user
- itemCol: tên cột chứa id của item
- ratingCol: tên cột chứa giá trị của rating
- coldStartStrategy = "drop": loại bỏ bất kỳ hàng nào trong DataFrame của các dự đoán có chứa giá trị NaN.

Mô hình được đánh giá thông qua chỉ số RMSE (Root mean squared error). Lỗi trung bình bình phương (RMSE) là độ lệch chuẩn của phần dư (lỗi dự đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy; RMSE là thước đo mức độ lan truyền của những phần dư này. Nói cách khác, RMSE cho biết mức độ tập trung của dữ liệu xung quanh đường hồi quy phù hợp nhất.

Về phần tinh chỉnh, nhóm sẽ thực thi Grid Search bằng cách sử dụng ParamGridBuilder trên 3 tham số: rank, maxIter và regParam.

4 Kết quả

4.1 Demographic Filtering

Mô hình Demographic sẽ đưa ra gợi ý đa phần hướng tới các đối tượng người dùng mới. Mô hình sẽ gợi ý các bộ phim có sức hút lớn tại thời điểm hiện tại, hoặc các bộ phim được đánh giá cao. Sau khi huấn luyện mô hình trả về kết quả top 5 phim được yêu thích nhất như bảng dưới đây:

Bảng 4: Top 5 bộ phim gợi ý bởi mô hình Demographic Filtering

anime_id	name	rating	predicted_rating
1535	Death Note	8.71	7.71
5114	Fullmetal Alchemist: Brotherhood	9.26	7.69
9253	Steins;Gate	9.17	7.51
16498	Shingeki no Kyojin	8.54	7.47
6547	Angel Beats!	8.39	7.37

Ta có thể thấy mặc dù điểm số đánh giá bởi người dùng cao, tuy nhiên số lượng người dùng thực hiện đánh giá khác nhau nên kết quả dự đoán bằng mô hình Demographic sẽ tạo nên một số sự khác biệt.

4.2 Content-based Filtering

Mô hình content-based sẽ đưa ra gợi ý dựa vào đặc điểm thuộc tính của bộ phim. Mô hình sẽ sử dụng thuật toán KNN gợi ý các bộ phim tương tự với bộ phim mà người dùng nhập vào. Dưới đây là kết quả thu được khi nhập vào phim 'Naruto: Akaki Yotsuba no Clover wo Sasage'.

Bảng 5: Top 5 bộ phim gợi ý bởi mô hình Content-based Filtering theo phương pháp tính độ tương tự cosine

anime_id	name	genre	rating	cosine distance
2558	Wakakusa Monogatari: Nan to Jo-sensei	Drama, Historical, School, Slice of Life	7.73	7.09e-10
22877	Seirei Tsukai no Blade Dance	Action, Comedy, Ecchi, Fantasy, Harem, Romance, School, Supernatural	7.15	7.71e-10
1122	Highlander: The Search for Vengeance	Action, Horror, Sci-Fi	7.1	7.77e-10
2178	Pokemon Advanced Generation: Rekkuu no Houmonsha Deoxys	Action, Adventure, Comedy, Fantasy, Kids, Sci-Fi	6.9	7.96e-10
12823	Shakugan no Shana III (Final) Specials	Comedy, Parody	7.25	8e-10

Bảng 6: Top 5 bộ phim gợi ý bởi mô hình Content-based Filtering theo phương pháp tính khoảng cách Euclidean

anime_id	name	genre	rating	euclidean distance
4444	Zero Duel Masters	Adventure, Comedy, Game, Sports	6.25	11.31
956	Daikuu Maryuu Gaiking	Adventure, Mecha, Sci-Fi, Shounen	7.02	12.37
9750	Itsuka Tenma no Kuro Usagi	Comedy, Ecchi, Romance, Shounen, Supernatural, Vampire	6.83	24.78
2994	Death Note Rewrite	Mystery, Police, Psychological, Supernatural, Thriller	7.84	34.31
2576	Araiguma Rascal	Drama, Historical, Slice of Life	6.82	56.15

4.3 Collaborative Filtering

- **Trước khi tinh chỉnh:** Chỉ số RMSE của mô hình là **1.169** với các tham số đầu vào mặc định: Rank = 10, MaxIter = 10, RegParamL = 0.1
- **Sau khi tinh chỉnh:** Mô hình đạt được hiệu suất cao nhất với các tham số: Rank = 20, MaxIter = 20, RegParamL = 0.1. Chỉ số RMSE của mô hình là **1.133**.

Mô hình ALS sẽ đưa ra gợi ý các bộ phim cho từng người dùng dựa vào đánh giá của người dùng với các bộ phim đã xem. Bảng 7 thể hiện top 5 bộ phim được mô hình gợi ý cho người dùng có user_id = 5 dựa vào danh sách các bộ phim mà người dùng này đã xem và đánh giá (Bảng 8)

Bảng 7: Top 5 bộ phim gợi ý cho người dùng có user_id = 5 bởi mô hình ALS

anime_id	name	rating prediction	genre
820	Ginga Eiyuu Densetsu	8.14	Drama, Military, Sci-Fi, Space
4639	Gosenzo-sama Banbanzai!	8.13	Comedy, Sci-Fi
20671	Shashinkan	7.98	Drama, Historical, Slice of Life
2921	Ashita no Joe 2	7.98	Drama, Sports
18029	Hello Kitty no Suteki na Kyoudai	7.84	Fantasy, Kids

Bảng 8: Top 5 bộ phim được người dùng có user_id = 5 đánh giá cao nhất

anime_id	name	rating	genre
15335	Gintama Movie: Kanketsu-hen - Yorozuya yo Eien Nare	10.0	Action, Comedy, Sci-Fi
14719	JoJo no Kimyou na Bouken (TV)	9.0	Action, Adventure
918	Gintama	9.0	Action, Comedy, Sci-Fi
28891	Haikyuu!! Second Season	9.0	Comedy, Drama, School
9969	Gintama	9.0	Action, Comedy, Sci-Fi

Nhìn vào bảng 8, có thể thấy người dùng có user_id = 5 khá thích các thể loại phim hành động, phiêu lưu, khoa học viễn tưởng, hài hước. Mô hình ALS đã gợi ý được cho người dùng một số phim có thể loại tương tự.Tuy nhiên vẫn xuất hiện một vài bộ phim dành cho trẻ em, trong khi các bộ phim người dùng này xem không có thể loại đó.

4.4 Demographic và Collaborative Filtering

- **Trước khi tinh chỉnh:** Chỉ số RMSE của mô hình là **1.157** với các tham số đầu vào mặc định: Rank = 10, MaxIter = 10, RegParamL = 0.1
- **Sau khi tinh chỉnh:** Mô hình đạt được hiệu suất cao nhất với các tham số: Rank = 16, MaxIter = 20, RegParamL = 0.1. Chỉ số RMSE của mô hình là **1.12**.

  Mô hình Demographic kết hợp với ALS sẽ đưa ra gợi ý các bộ phim cho từng người dùng dựa vào đánh giá của người dùng với các bộ phim đang có sức hút tại thời điểm hiện tại hoặc các bộ phim được cộng đồng đánh giá cao. Bảng 6 thể hiện top 5 bộ phim được mô hình gợi ý cho người dùng có user_id = 5

Bảng 9: Top 5 bộ phim gợi ý cho người dùng có user_id = 5 bởi mô hình Demographic kết hợp mô hình ALS

anime_id	name	rating prediction	genre
8598	Yuki	8.28	Adventure, Demons, Fantasy
820	Ginga Eiyuu Densetsu	7.76	Drama, Military, Sci-Fi, Space
2920	Ashita no Joe (Movie)	7.45	Action, Drama, Sports
22583	Uchuu Kyoudai: Number Zero	7.43	Comedy, Sci-Fi, Seinen, Slice of Life
2921	Ashita no Joe 2	7.41	Drama, Sports

5 Kết luận

Trong báo cáo này, nhóm đã đề xuất một số phương án thực thi việc gợi ý phim anime thông qua các mô hình gợi ý khác nhau. Mỗi mô hình được xử lý theo nhiều cách khác nhau để phù hợp với đầu vào cũng như để nâng cao kết quả huấn luyện. Thông qua việc huấn luyện và so sánh, nhóm đề xuất mô hình kết hợp Demographic và Collaborative Filtering đã đem lại kết quả tốt hơn so với việc sử dụng riêng lẻ Collaborative Filtering (kết quả RMSE mô hình kết hợp là 1.12, thấp hơn so với mô hình Collaborative). Ngoài ra mô hình Demographic cũng đưa lại các kết quả khả quan khi gợi ý các bộ phim được đánh giá là nổi tiếng trong anime.

Về cơ bản, nhóm đã áp dụng các hướng đi phù hợp đối với các đối tượng khác nhau. Hướng phát triển trong tương lai đối với bài toán này nhóm muốn kết hợp giữa Content-based và ALS để có thể tăng mức độ dự đoán của mô hình. Ngoài ra, với mô hình Content-based, nhóm muốn tập trung vào việc cải thiện tiền xử lý dữ liệu để việc dự đoán trở nên đúng gần nhất.

Tài liệu

1. Kaggle Anime Recommendations Database, `https://www.kaggle.com/CooperUnion/anime-recommendations-database`. Last accessed 18 Jan 2022.
2. Koren, Y., Bell, R., Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), 30–37. https://doi.org/10.1109/mc.2009.263
3. IMDb, Weighted Average Ratings, `https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK?ref_=helpsect_pro_2_8`, Last accessed: 22 Jan 2022
4. Apache Spark, ALS, `https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.recommendation.ALS.html`, Last accessed: 22 Jan 2022
5. Jamen Long, ALS parameters and hyperparameters, `https://campus.datacamp.com/courses/recommendation-engines-in-pyspark/how-does-als-work?ex=13`, Last accessed: 22 Jan 2022