|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**  **KHOA CÔNG NGHỆ THÔNG TIN**  **----🙚🕮🙘----** |
| **Logo  Description automatically generated** |
| **VÕ HOÀNG KHẢ DIỆU: 19133014**  **NGUYỄN QUỐC BẢO: 19133002** |
| Đề tài: |
| **TÌM HIỂU VỀ CHỦ ĐỀ VỀ HỆ THỐNG GỢI Ý VÀ ỨNG DỤNG** |
| **TIỂU LUẬN CHUYÊN NGÀNH**  **NGÀNH KỸ THUẬT DỮ LIỆU** |
| GIÁO VIÊN HƯỚNG DẪN  **ThS. Quách Đình Hoàng** |
| **KHÓA 2019 – 2023** |

Shape, square

Description automatically generated

|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**  **KHOA CÔNG NGHỆ THÔNG TIN**  **----🙚🕮🙘----** |
| **Logo  Description automatically generated** |
| **VÕ HOÀNG KHẢ DIỆU: 19133014**  **NGUYỄN QUỐC BẢO: 19133002** |
| Đề tài: |
| **TÌM HIỂU VỀ CHỦ ĐỀ VỀ HỆ THỐNG GỢI Ý VÀ ỨNG DỤNG** |
| **TIỂU LUẬN CHUYÊN NGÀNH**  **NGÀNH KỸ THUẬT DỮ LIỆU** |
| GIÁO VIÊN HƯỚNG DẪN  **ThS. Quách Đình Hoàng** |
| **KHÓA 2019 – 2023** |

Shape, square

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*\*** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  **\*\*\*\*\*\*\*\*** | |
| **PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN** | | |
| Họ và tên Sinh viên 1: Võ Hoàng Khả Diệu  Họ và tên Sinh viên 2: Nguyễn Quốc Bảo | | MSSV: 19133014  MSSV: 19133002 |
| Ngành: Kỹ thuật dữ liệu  Tên đề tài: TÌM HIỂU VỀ HỆ THỐNG GỢI Ý VÀ ỨNG DỤNG  Họ và tên giáo viên hướng dẫn: ThS.Quách Đình Hoàng | | |
| **NHẬN XÉT**  1. Về nội dung đề tài khối lượng thực hiện: ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 2. Ưu điểm: ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 3. Khuyết điểm  ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 4. Đề nghị cho bảo vệ hay không? 5. Đánh giá loại:  6. Điểm: | | |
| Tp*. Hồ Chí Minh, ngày    tháng     năm   2022*  Giáo viên hướng dẫn  *(Ký & ghi rõ họ tên)* | | |

|  |  |  |
| --- | --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*\*** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  **\*\*\*\*\*\*\*\*** | |
| **PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN** | | |
| Họ và tên Sinh viên 1: Võ Hoàng Khả Diệu  Họ và tên Sinh viên 2: Nguyễn Quốc Bảo | | MSSV: 19133014  MSSV: 19133002 |
| Ngành: Kỹ thuật dữ liệu  Tên đề tài: TÌM HIỂU VỀ HỆ THỐNG GỢI Ý VÀ ỨNG DỤNG  Họ và tên giáo viên phản biện: Ths Nguyễn Thành Sơn | | |
| **NHẬN XÉT**  1. Về nội dung đề tài khối lượng thực hiện: ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 2. Ưu điểm: ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 3. Khuyết điểm  ………………………………………………………………………………………………………  ………………………………………………………………………………………………………  ……………………………………………………………………………………………………… | | |
| 4. Đề nghị cho bảo vệ hay không? 5. Đánh giá loại:  6. Điểm: | | |
| Tp*. Hồ Chí Minh, ngày    tháng     năm   2022*  Giáo viên hướng dẫn  *(Ký & ghi rõ họ tên)* | | |

**LỜI CAM ĐOAN**

Tiểu luận này là công trình nghiên cứu của chúng em, được thực hiện dưới sự hướng dẫn khoa học của thầy Quách Đình Hoàng. Các số liệu, những kết luận nghiên cứu và sản phẩm được tạo ra bởi chúng em được trình bày trong khoá luận này là trung thực.

Chúng em xin hoàn toàn chịu trách nhiệm về lời cam đoan này.

|  |  |
| --- | --- |
| Sinh viên thực hiện  (Ký và ghi rõ họ tên) | Sinh viên thực hiện  (Ký và ghi rõ họ tên) |
| **Võ Hoàng Khả Diệu** | **Nguyễn Quốc Bảo** |

**LỜI CẢM ƠN**

Một kỳ thực hiện tiểu luận chuyên ngành đã trôi qua nhưng để lại trong chúng em rất nhiều cảm xúc. Chúng em xin được gửi lời cảm ơn chân thành nhất đến Thầy Quách Đình Hoàng. Thầy luôn theo dõi tiến độ và giải đáp, chia sẻ giúp chúng tôi vượt qua những khó khăn. Chúng em rất trân quý sự tâm huyết và trách nhiệm của Thầy trong công việc giảng dạy và truyền đạt kiến thức.

Chúng em cũng xin gửi lời cảm ơn sâu sắc đến Thầy Cô khoa Công nghệ Thông tin - Đại học Sư phạm Kỹ thuật TP.HCM đã đồng hành và hỗ trợ chúng em trong suốt quá trình học tập và thực hiện tiểu luận. Chúng em xin cảm ơn trường Đại học Sư phạm Kỹ thuật đã tạo nhiều điều kiện thuận lợi cho hoạt động phục vụ học tập của sinh viên chúng em, đặc biệt là thư viện số với nguồn tri thức vô tận. Chúng tôi cũng gửi lời cảm ơn chân thành đến các bạn khóa 2019 ngành Kỹ thuật Dữ liệu, cảm ơn những góp ý và chia sẻ quý giá từ tất cả các bạn. Cảm ơn sự động viên từ các bạn để nhóm chúng em có thể giữ vững tinh thần và thực hiện khóa luận đúng tiến độ.

Những giá trị cốt lõi nhà trường, Thầy Cô và bạn bè mang đến, chúng em sẽ luôn ghi nhớ để làm động lực thúc đẩy bản thân phát triển và hoàn thiện hơn nữa. Cuối cùng, chúng em xin cám ơn đến tác giả của những bài báo khoa học mà chúng em đã tham khảo. Các bài báo này giúp chúng em tiếp thu được nhiều kiến thức mới và quan trọng là hiểu rõ hơn về đề tài đang nghiên cứu. Chúng em cũng nhận thấy bản thân có những khuyết điểm và thiếu sót cần cố gắng cải thiện để tốt hơn, hướng tới mục tiêu lớn trong tương lai.

Xin chân thành cảm ơn!

**KẾ HOẠCH THỰC HIỆN**

|  |  |  |
| --- | --- | --- |
| **Tuần** | **Thời gian** | **Nhiệm vụ**  **(Công việc dự kiến)** |
| 1+3 | 23/08 – 12/09 | Tìm hiểu đề tài hệ thống gợi ý |
| 4+5 | 13/09 – 26/09 | Hoàn thiện mô tả đề tài |
| 6 | 27/09 – 03/10 | Nghiên cứu Spark ml  Tìm hiểu nhóm thuật toán collaborative Filterring và Content based |
| 7 | 04/10 – 10/10 | Tìm hiểu về thuật toán ALS |
| 8 | 11/10 - 19/10 | Tìm Dataset, thực hiện các bước tiền xử lý dữ liệu |
| 9 | 18/10 – 24/10 | Hoàn thành các bước xử lý data |
| 10+11 | 25/10 – 07/11 | Tiến hành xây dựng model và debug |
| 12+13 | 8/11 – 21/11 | Debug xong final model |
| 14+15 | 22/11 – 05/12 | Viết báo cáo |
| 16 + 17 | 06/12 – 19/12 | Chuẩn bị file thuyết trình  Tinh chỉnh báo cáo |
| 18 | 20/12 – 26/12 | Trao đổi với thầy hướng dẫn về nội dung cần trình bày ở buổi báo cáo tiểu luận chuyên ngành. |

**MỤC LỤC**

[PHẦN 1: MỞ ĐẦU 11](#_Toc121693406)

[1. Tính cấp thiết của đề tài 11](#_Toc121693407)

[2. Mục tiêu và nhiệm vụ nghiên cứu 11](#_Toc121693408)

[3. Cách tiếp cận và phương pháp nghiên cứu 11](#_Toc121693409)

[4. Kết quả dự kiến đạt được 12](#_Toc121693410)

[5. Bố cục luận văn 12](#_Toc121693411)

[PHẦN 2: NỘI DUNG 13](#_Toc121693412)

[CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG GỢI Ý 13](#_Toc121693413)

[1.1 KHÁI NIỆM 13](#_Toc121693414)

[1.1.1. Hệ thống gợi ý là gì? 13](#_Toc121693415)

[1.1.2. Tại sao phải sử dụng hệ thống gợi ý 13](#_Toc121693416)

[1.2. UTILITY MATRIX 14](#_Toc121693417)

[1.3. CÁC LOẠI HỆ THỐNG GỢI Ý 15](#_Toc121693418)

[1.3.1. Popularity-Based Recommendation System 15](#_Toc121693419)

[1.3.2. Classification Model 16](#_Toc121693420)

[1.3.3. Content-Based Recommendation System 17](#_Toc121693421)

[1.3.4. Collaborative filtering Recommenders 18](#_Toc121693422)

[CHƯƠNG 2: DATASET 22](#_Toc121693423)

[2.1. TÌM HIỂU VỀ DATASET 22](#_Toc121693424)

[2.1.1. Nguồn dữ liệu 22](#_Toc121693425)

[2.1.2. Mô tả 22](#_Toc121693426)

[2.2. XỬ LÝ DỮ LIỆU 24](#_Toc121693427)

[2.2.1. Anime Dataset 24](#_Toc121693428)

[2.2.2. Ranting Dataset 24](#_Toc121693429)

[2.3. EDA (Exploratory Data Analysis) 24](#_Toc121693430)

[CHƯƠNG 3: XÂY DỰNG MÔ HÌNH RECOMMEND SYSTEM 29](#_Toc121693431)

[3.1. XÂY DỰNG MÔ HÌNH DỰA TRÊN THUẬT TOÁN ALS 29](#_Toc121693432)

[3.1.1. Thuật toán ALS 29](#_Toc121693433)

[3.1.2. Training Model 29](#_Toc121693434)

[3.2. XÂY DỰNG MÔ HÌNH DỰA TRÊN KỸ THUẬT TF-IDF 30](#_Toc121693435)

[3.2.1. Kỹ thuật TF-IDF 30](#_Toc121693436)

[PHẦN 3: KẾT LUẬN 31](#_Toc121693437)

[1. KẾT QUẢ ĐẠT ĐƯỢC 31](#_Toc121693438)

[2. HẠN CHẾ 31](#_Toc121693439)

[3. HƯỚNG PHÁT TRIỂN 31](#_Toc121693440)

[TÀI LIỆU THAM KHẢO 32](#_Toc121693441)

**DANH MỤC HÌNH ẢNH**

[Hình 1: Tổng quan về Recommendation System 13](#_Toc121695872)

[Hình 2. Hiển thị User-Item cho các webpage 14](#_Toc121695873)

[Hình 3. Ví dụ về utility matrix với hệ thống Gợi ý bài hát. 15](#_Toc121695874)

[Hình 4. Mô hình phân loại 16](#_Toc121695875)

[Hình 5. Khái niệm về lọc cộng tác. 18](#_Toc121695876)

[Hình 6. Công thức dự đoán User-user 19](#_Toc121695877)

[Hình 7. User-user Collaborative Filtering 19](#_Toc121695878)

[Hình 8. Công thức dự đoán Item-item 20](#_Toc121695879)

[Hình 9. File anime.csv 23](#_Toc121695880)

[Hình 10. file ranting.csv 23](#_Toc121695881)

[Hình 11. Master dataframe 25](#_Toc121695882)

[Hình 12. Top 10 anmie có group member cao nhất 25](#_Toc121695883)

[Hình 13. Top 10 Anime có tống review cao nhất 26](#_Toc121695884)

[Hình 14. Biểu đồ thể hiện rating 26](#_Toc121695885)

[Hình 15. Biểu đồ thể hiện thể loại của anime 27](#_Toc121695886)

[Hình 16. Biểu đồ thể hiện rating theo thể loại anime 27](#_Toc121695887)

**PHẦN 1: MỞ ĐẦU**

## **Tính cấp thiết của đề tài**

Với lượng thông tin ngày càng tăng lên hàng ngày, làm thế nào để nhanh chóng có được các Items thông tin để đáp ứng nhu cầu của mọi người trở thành vấn đề cấp bách hiện nay. Nỗ lực trong việc truy xuất thông tin đã mang lại sự tiện lợi lớn cho những người có xu hướng truy xuất thông tin bằng cách nhập truy vấn hoặc từ khóa. Nếu một số trang web có thể chủ động gợi ý các sản phẩm, mục thông tin mà người dùng có thể quan tâm thì sẽ nâng cao hiệu quả rất nhiều.

Trong những năm gần đây, ngành giải trí đã và đang phát triển cực kì mạnh mẽ. Không chỉ riêng các bộ phim điện ảnh được sản xuất công phu thì bên cạnh đó thể loại phim hoạt hình Anime cũng rất được mọi người ưa chuộng. Cụ thể là có rất nhiều bộ phim Anime đã và đang được chiếu tại các rap trên toàn thế giới và đạt được những con số doanh thu kỷ lục.

Vì thế, nhóm nhận thấy cần có một “**hệ thống để gợi ý các bộ phim Anime**” dựa theo thói quen xem phim của người dùng là thực sự cần thiết ngay lúc này, hệ thống này sẽ giảm thiểu thời gian mà người dùng phân vân không biết nên xem gì tiếp theo, và sẽ đưa ra một bộ phim với độ yêu thích dự đoán rất cao từ người dùng.

## **Mục tiêu và nhiệm vụ nghiên cứu**

Mục tiêu: Xây dựng được hệ thống gợi ý phim Anime với một số chức năng sau:

- Tạo đánh giá mới từ người dùng.

- Hiển thị top các bộ anime cho mỗi người dùng.

- Gợi ý một bộ anime mới cho người dùng dựa trên thể loại các anime người dùng đã đánh giá

Nhiệm vụ:

* Tìm hiểu các khái niệm về hệ thống gợi ý
* Tìm hiểu về các thuật toán sử dụng
* Tìm Dataset phù hợp với yêu cầu đề tài
* Xây dựng được hệ thống gợi ý

1. **Cách tiếp cận và phương pháp nghiên cứu**

Cách tiếp cận:

* Tìm các đặc trưng (features) có ảnh hưởng đến việc đánh giá của người dùng, thông qua việc phân tích và thăm dò dữ liệu
* Phân tích và áp dụng giải thuật filtering phù hợp
* Xử lí dữ liệu đầu vào trước khi huấn luyện.
* Tiến hành training mô hình và đánh giá hiệu suất.

## **Kết quả dự kiến đạt được**

Về mặt lí thuyết:

* Hiểu về các khái niệm, thuật toán, cách xử lý của một hệ thống gợi ý.

Về mặt sản phẩm: Xây dựng được mô hình của hệ thống gợi ý với kết quả và độ chính xác khả quan.

## **Bố cục luận văn**

Chúng tôi định chia bố cục tiểu luận thành các phần như sau:

* Phần 1: Mở đầu
* Phần 2: Nội dung. Phần này gồm 4 chương
  + Chương 1: Tìm hiểu về hệ thống của gợi ý
  + Chương 2: Dataset
  + Chương 3: Xây dựng model hệ thống gợi ý
* Phần 3: Kết luận
* Tài liệu tham khảo

# **PHẦN 2: NỘI DUNG**

## **CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG GỢI Ý**

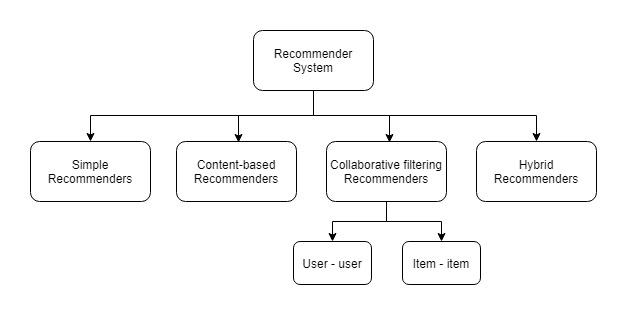
### **KHÁI NIỆM**

* + 1. **Hệ thống gợi ý là gì?**

Trong vài thập kỷ gần đây, với sự nổi lên của Youtube, Amazon, Netflix và nhiều dịch vụ web khác, hệ thống gợi ý ngày càng có nhiều vị trí hơn trong cuộc sống của chúng ta. Từ thương mại điện tử (gợi ý cho người mua những bài báo mà họ có thể quan tâm) đến quảng cáo trực tuyến (gợi ý cho người dùng những nội dung phù hợp, phù hợp với sở thích của họ), hệ thống giới thiệu ngày nay không thể tránh khỏi trong hành trình trực tuyến hàng ngày của chúng ta.

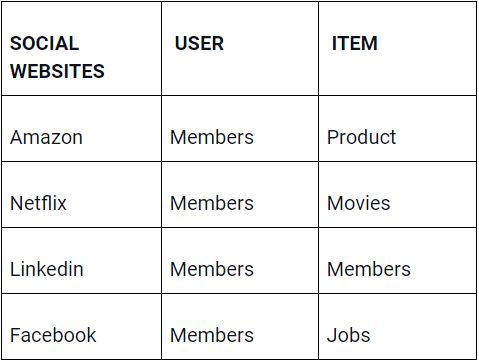
Nói một cách tổng quát, hệ thống gợi ý là các thuật toán nhằm đề xuất các mặt hàng có liên quan cho người dùng (mặt hàng là phim để xem, tin nhắn văn bản để đọc, sản phẩm cần mua hoặc bất kỳ thứ gì khác tùy theo ngành).

Hệ thống gợi ý xử lý một lượng lớn thông tin hiện có bằng cách lọc thông tin quan trọng nhất dựa trên dữ liệu do người dùng cung cấp và các yếu tố khác quan tâm đến sở thích và mối quan tâm của người dùng. Nó tìm ra sự phù hợp giữa người dùng và mặt hàng và đưa ra những điểm tương đồng giữa người dùng và mặt hàng để đề xuất.



Hình 1: Tổng quan về Recommendation System

* + 1. **Tại sao phải sử dụng hệ thống gợi ý**
* Mang lại lợi ích cho người dùng trong việc tìm kiếm các mặt hàng họ quan tâm.
* Giúp các nhà cung cấp mặt hàng trong việc phân phối các mặt hàng của họ đến đúng người dùng.
* Nhận dạng sản phẩm phù hợp nhất với người dùng.
* Nội dung được cá nhân hóa.
* Giúp các trang web cải thiện mức độ tương tác của người dùng.



Hình 2. Hiển thị User-Item cho các webpage

**1.2. UTILITY MATRIX**

Có hai thực thể chính trong các Recommendation Systems là users và items. Mỗi user sẽ có mức độ quan tâm (degree of preference) tới từng item khác nhau. Mức độ quan tâm này, nếu đã biết trước, được gán cho một giá trị ứng với mỗi cặp user-item. Giả sử rằng mức độ quan tâm được đo bằng giá trị user rate cho item, ta tạm gọi giá trị này là rating. Tập hợp tất cả các ratings, bao gồm cả những giá trị chưa biết cần được dự đoán, tạo nên một ma trận gọi là utility matrix.

Thông thường, có rất nhiều users và items trong hệ thống, và mỗi user thường chỉ rate một số lượng rất nhỏ các item, thậm chí có những user không rate item nào (với những users này thì cách tốt nhất là gợi ý các items phổ biến nhất).

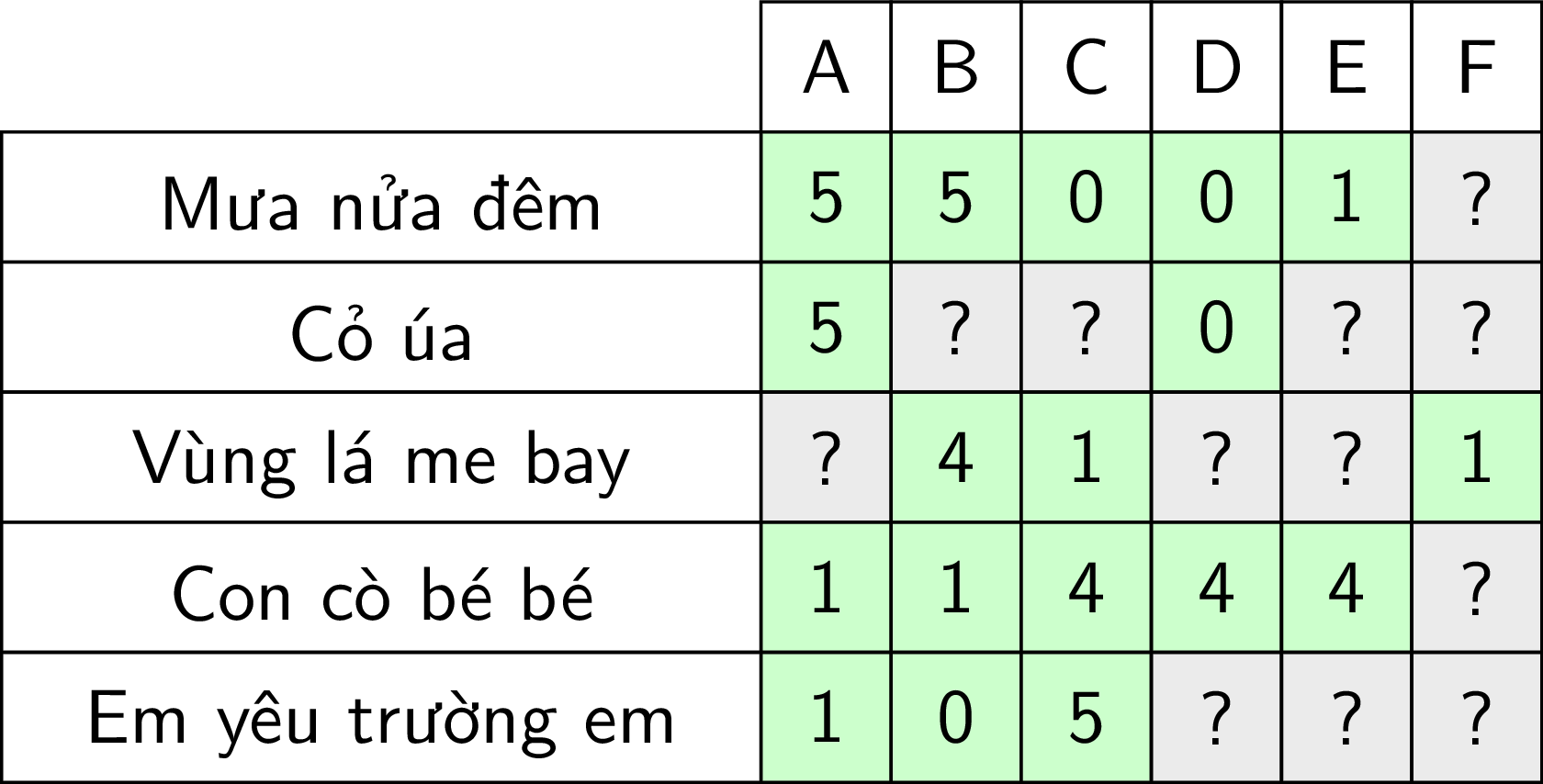
Càng nhiều ô được điền thì độ chính xác của hệ thống sẽ càng được cải thiện. Vì vậy, các hệ thống luôn luôn hỏi người dùng về sự quan tâm của họ tới sản phẩm, và muốn người dùng đánh giá càng nhiều sản phẩm càng tốt. Việc đánh giá các sản phẩm, vì thế, không những giúp các người dùng khác biết được chất lượng sản phẩm mà còn giúp hệ thống biết được sở thích của người dùng, qua đó có chính sách quảng cáo hợp lý.

Không có Utility matrix, gần như không thể gợi ý được sản phẩm tới ngừời dùng, ngoài cách luôn luôn gợi ý các sản phẩm phổ biến nhất. Vì vậy, trong các Recommender Systems, việc xây dựng Utility Matrix là tối quan trọng. Tuy nhiên, việc xây dựng ma trận này thường có gặp nhiều khó khăn. Có hai hướng tiếp cận phổ biến để xác định giá trị rating cho mỗi cặp user-item trong Utility Matrix:

- Nhờ người dùng rate sản phẩm.

- Dựa trên hành vi của users.

Ví dụ:



Hình 3. Ví dụ về utility matrix với hệ thống Gợi ý bài hát.

Các bài hát được người dùng đánh giá theo mức độ từ 0 đến 5 sao. Các dấu '?' nền màu xám ứng với việc dữ liệu chưa tồn tại trong cơ sở dữ liệu. Recommendation Systems cần phải tự điền các giá trị này.

Trong ví dụ đơn giản này, dễ thấy có 2 thể loại nhạc khác nhau: 3 bài đầu là nhạc Bolero và 2 bài sau là nhạc Thiếu nhi. Từ dữ liệu này, ta cũng có thể đoán được rằng A, B thích thể loại Bolero; C, D, E, F thích thể loại Thiếu nhi. Từ đó, một hệ thống tốt nên gợi ý Cỏ úa cho B; Vùng lá me bay cho A; Em yêu trường em cho D, E, F. Giả sử chỉ có hai thể loại nhạc này, khi có một bài hát mới, ta chỉ cần phân lớp nó vào thể loại nào, từ đó đưa ra gợi ý với từng người dùng.

**1.3. CÁC LOẠI HỆ THỐNG GỢI Ý**

**1.3.1. Popularity-Based Recommendation System**

Nó là một loại hệ thống gợi ý hoạt động dựa trên nguyên tắc phổ biến và bất cứ thứ gì có xu hướng. Các hệ thống này kiểm tra về sản phẩm hoặc bộ phim đang thịnh hành hoặc được người dùng yêu thích nhất và trực tiếp giới thiệu những sản phẩm hoặc bộ phim đó.

Ví dụ: nếu một sản phẩm thường được hầu hết mọi người mua thì hệ thống sẽ biết rằng sản phẩm đó phổ biến nhất, vì vậy đối với mỗi người dùng mới vừa nhấn vào sản phẩm đó, hệ thống sẽ giới thiệu sản phẩm đó cho người dùng đó và khả năng cao là người dùng mới cũng sẽ mua nó.

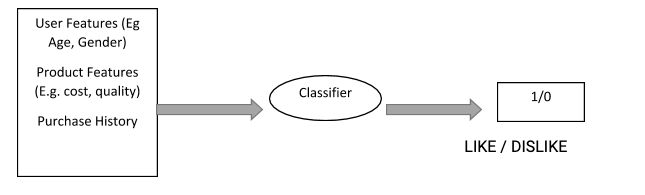
* Điểm mạnh của Popularity-Based Recommendation System
  + Không cá nhân hóa.
  + Hệ thống sẽ đề xuất cùng một loại sản phẩm / phim chỉ dựa trên mức độ phổ biến đối với mọi người dùng khác.

Ví dụ:

* + YouTube: Video thịnh hành.
  + Google News: Tin tức được lọc theo xu hướng và tin tức phổ biến nhất.

**1.3.2. Classification Model**

Mô hình sử dụng các tính năng của cả hai sản phẩm cũng như người dùng để dự đoán liệu người dùng có thích một sản phẩm hay không.



Hình 4. Mô hình phân loại

Đầu ra có thể là 0 hoặc 1. Nếu người dùng thích thì là 1 và ngược lại.

* Hạn chế của mô hình phân loại
  + Là một nhiệm vụ nghiêm ngặt để thu thập một lượng lớn thông tin về những người dùng khác nhau và cũng như các sản phẩm.
  + Ngoài ra, nếu việc thu thập được thực hiện thì cũng có thể khó phân loại.
  + Vấn đề về tính linh hoạt.

**1.3.3. Content-Based Recommendation System**

Đây là một loại hệ thống gợi ý khác hoạt động trên nguyên tắc nội dung tương tự. Nếu người dùng đang xem phim thì hệ thống sẽ kiểm tra các phim khác có nội dung tương tự hoặc cùng thể loại với phim mà người dùng đang xem. Có nhiều thuộc tính cơ bản khác nhau được sử dụng để tính toán mức độ giống nhau trong khi kiểm tra nội dung tương tự.

Nó đánh giá đặc tính của items được recommended.

Ví dụ: một user xem rất nhiều các bộ phim về cảnh sát hình sự, vậy thì gơi ý một bộ phim trong cơ sở dữ liệu có chung đặc tính hình sự tới user này, ví dụ phim Người phán xử.

Cách tiếp cận này yêu cầu việc sắp xếp các items vào từng nhóm hoặc đi tìm các đặc trưng của từng item. Tuy nhiên, có những items không có nhóm cụ thể và việc xác định nhóm hoặc đặc trưng của từng item đôi khi là bất khả thi.

Trong Content-based Recommendation Systems, chúng ta làm quen với một Hệ thống gợi ý sản phẩm đơn giản dựa trên đặc trưng của mỗi item. Đặc điểm của Content-based Recommendation Systems là việc xây dựng mô hình cho mỗi user không phụ thuộc vào các users khác mà phụ thuộc vào profile của mỗi items. Việc làm này có lợi thế là tiết kiệm bộ nhớ và thời gian tính toán. Đồng thời, hệ thống có khả năng tận dụng các thông tin đặc trưng của mỗi item như được mô tả trong bản mô tả (description) của mỗi item. Bản mô tả này có thể được xây dựng bởi nhà cung cấp hoặc được thu thập bằng cách yêu cầu users gắn tags cho items. Việc xây dựng feature vector cho mỗi item thường bao gồm các kỹ thuật Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP).

Cách làm trên có hai nhược điểm cơ bản. Thứ nhất, khi xây dựng mô hình cho một user, các hệ thống Content-based không tận dụng được thông tin từ các users khác. Những thông tin này thường rất hữu ích vì hành vi mua hàng của các users thường được nhóm thành một vài nhóm đơn giản; nếu biết hành vi mua hàng của một vài users trong nhóm, hệ thống nên suy luận ra hành vi của những users còn lại. Thứ hai, không phải lúc nào chúng ta cũng có bản mô tả cho mỗi item. Việc yêu cầu users gắn tags còn khó khăn hơn vì không phải ai cũng sẵn sàng làm việc đó; hoặc có làm nhưng sẽ mang xu hướng cá nhân. Các thuật toán NLP cũng phức tạp hơn ở việc phải xử lý các từ gần nghĩa, viết tắt, sai chính tả, hoặc được viết ở các ngôn ngữ khác nhau

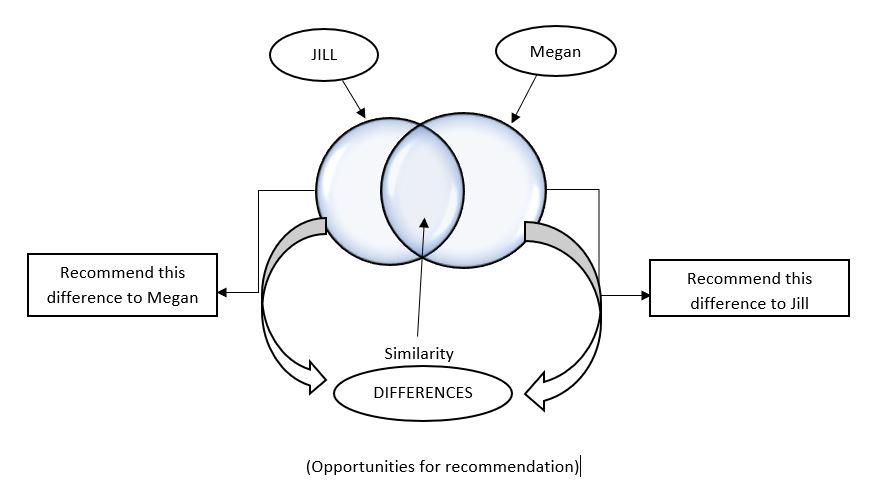
**1.3.4. Collaborative filtering Recommenders**

Nó được coi là một trong những hệ thống gợi ý rất thông minh, hoạt động dựa trên sự tương đồng giữa những người dùng khác nhau và cũng có những mặt hàng được sử dụng rộng rãi như một trang web thương mại điện tử và cả các trang web phim trực tuyến. Nó kiểm tra thị hiếu của những người dùng tương tự và đưa ra các đề xuất.

Sự giống nhau không bị hạn chế đối với sở thích của người dùng hơn nữa có thể xem xét sự giống nhau giữa các items khác nhau. Hệ thống sẽ đưa ra các đề xuất hiệu quả hơn nếu chúng ta có một lượng lớn thông tin về người dùng và items.

đang hiệu quả với tập dữ liệu.

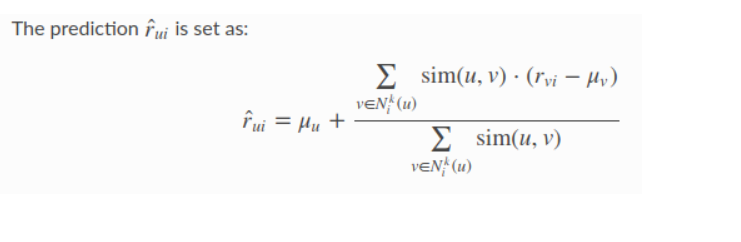
Ví dụ:



Hình 5. Khái niệm về lọc cộng tác.

Hình trên cho thấy hai người dùng khác nhau và sở thích của họ cùng với sự tương đồng giữa khẩu vị của cả hai người dùng. Người ta thấy rằng cả Jil và Megan đều có sở thích giống nhau nên sự quan tâm của Jill được gợi ý cho Megan và ngược lại.

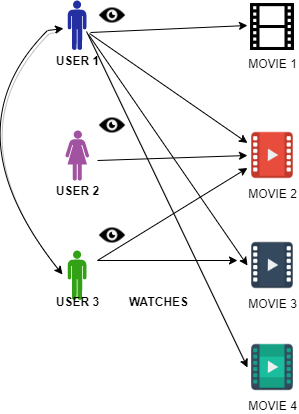
**\*** **User-user Collabarative Filtering**

****

Hình 6. Công thức dự đoán User-user

**Ghi chú:**

* + Đầu ra là dự đoán về xếp hạng của người dùng u đối với mục i.
  + Chúng tôi sử dụng thước đo độ tương đồng giữa người dùng u và người dùng v trong trường hợp này.

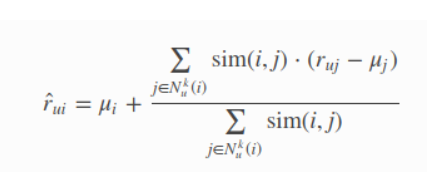


Hình 7. User-user Collaborative Filtering

Hình 1.5 cho thấy bộ lọc cộng tác giữa người dùng và người dùng trong đó có ba người dùng 1, 2 và 3 tương ứng và sự quan tâm của họ đối với phim. Hệ thống tìm ra những người dùng có cùng sở thích xem phim và sự tương đồng giữa những người dùng được tính toán dựa trên hành xem phim. Người dùng 1 và người dùng 3 tương tự nhau vì họ đã xem các phim tương tự.

**\* Item-item Collaborative Filtering**

Ở phương pháp này thay vì sử dụng mức độ tương tự của người dùng, chúng tôi sử dụng thước đo độ tương tự của mặt hàng để tính toán dự đoán.

****

Hình 8. Công thức dự đoán Item-item

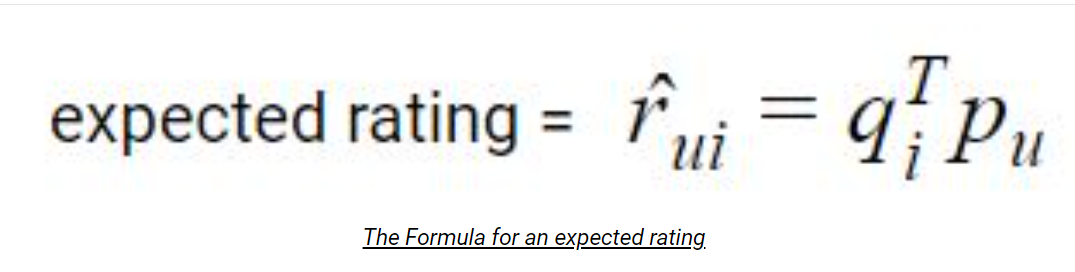
Ưu điểm của lọc dựa trên mục so với lọc dựa trên người dùng:

* + Thang đo tốt hơn : Tính năng lọc dựa trên người dùng không mở rộng quy mô cũng như lượt thích / sở thích của người dùng có thể thay đổi thường xuyên. Do đó, khuyến nghị cần được đào tạo lại thường xuyên.
  + Rẻ hơn về mặt tính toán : Trong nhiều trường hợp, có nhiều người dùng hơn các mặt hàng. Sẽ rất hợp lý khi sử dụng tính năng lọc dựa trên mục trong trường hợp này.

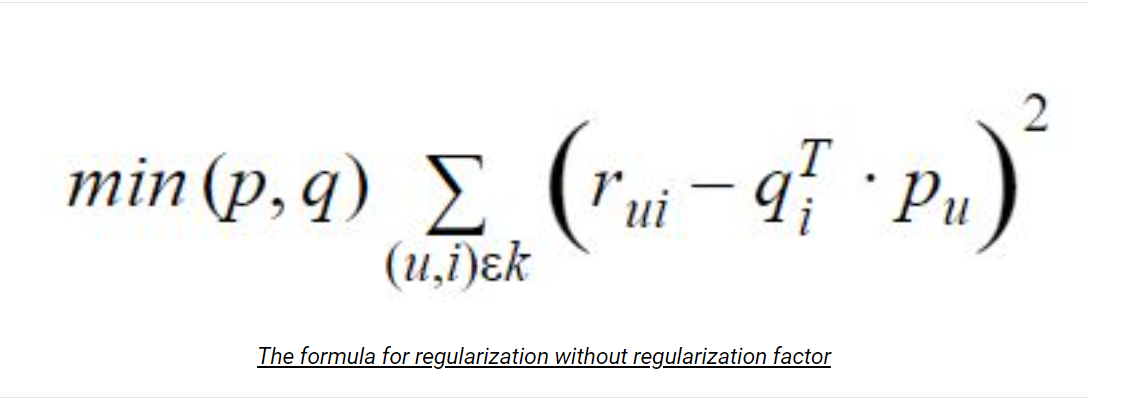
**\* Singular value decomposition and matrix-factorization**

Singular value decomposition còn được gọi là thuật toán SVD được sử dụng như một phương pháp lọc cộng tác trong các hệ thống gợi ý. SVD là một phương pháp phân tích nhân tử ma trận được sử dụng để giảm các đặc trưng trong dữ liệu bằng cách giảm các kích thước từ N đến K trong đó (K <N).

Đối với một phần của gợi ý, phần duy nhất được quan tâm là phân tích nhân tử ma trận được thực hiện bằng ma trận xếp hạng mục người dùng. Matrix factorizationlà tất cả về việc lấy 2 ma trận có tích là ma trận ban đầu. Các vectơ được sử dụng để đại diện cho mục 'qi' và 'pu' của người dùng sao cho sản phẩm chấm của họ là xếp hạng dự kiến.

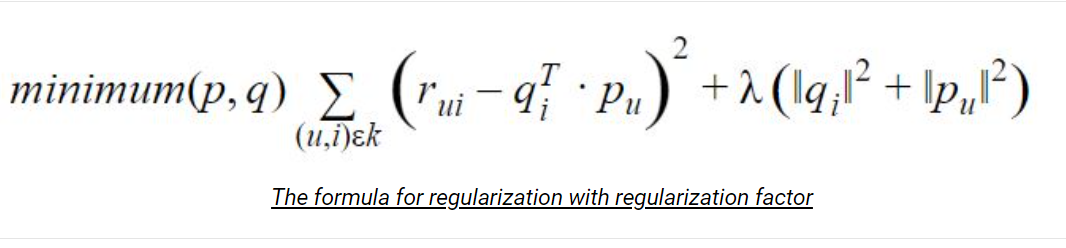


'qi' và 'pu' có thể được tính theo cách sao cho chênh lệch sai số bình phương giữa sản phẩm chấm của người dùng và vật phẩm và xếp hạng ban đầu trong ma trận người dùng-mặt hàng là ít nhất.



Quy định hóa: Tránh trang bị quá nhiều mô hình là một khía cạnh quan trọng của bất kỳ mô hình học máy nào vì nó dẫn đến độ chính xác của mô hình thấp. Quy định hóa loại bỏ nguy cơ mô hình bị trang bị quá mức.

Vì mục đích này trong chính quy hóa, một điều khoản phạt được đưa vào phương trình tối thiểu hóa ở trên. λ là hệ số chính quy được nhân với tổng bình phương độ lớn của vectơ người dùng và vật phẩm.



**CHƯƠNG 2: DATASET**

**2.1. TÌM HIỂU VỀ DATASET**

**2.1.1. Nguồn dữ liệu**

Tập dữ liệu được lấy từ trang kaggle:

<https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database>

Thu thập dữ liệu qua API về Anime và rating của user của web https://myanimelist.net/

Tập dữ liệu này chứa thông tin về sở thích của người dùng từ 73516 người dùng trên 12294 bộ Anime.

Mỗi người dùng có thể thêm một bộ Anime vào danh sách yêu thích của họ và đánh giá xếp hạng.

**2.1.2. Mô tả**

**\*File Anime.csv**

Tệp này chứa thông tin về các bộ phim anime

- anime\_id : id của một anime (unique).

- Name: tên đầy đủ của một anime.

- Genre: danh sách các thể loại.

- Type: Loại (movie, TV, OVA, etc).

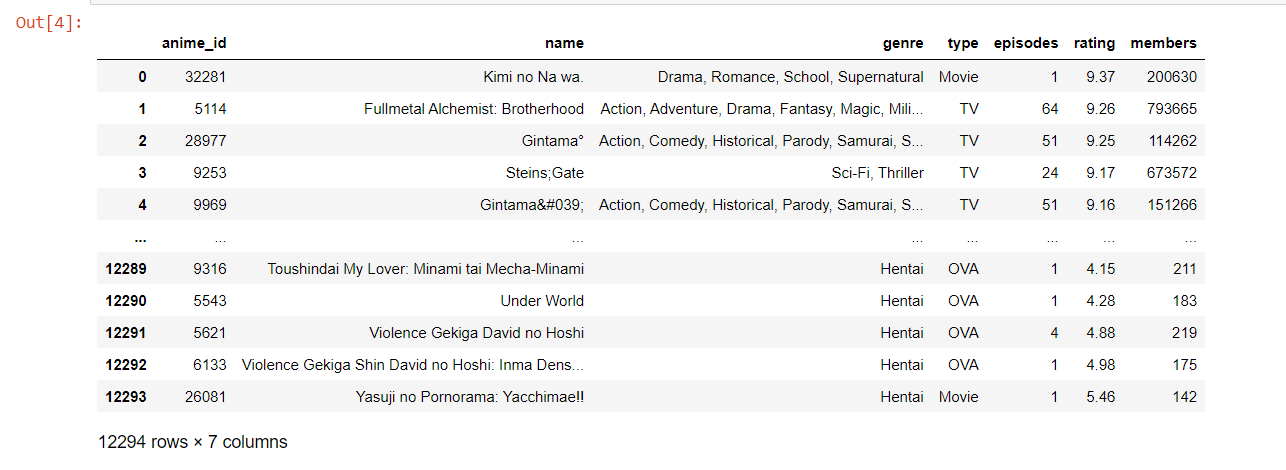
- Episodes: số tập của 1 bộ phim

- Rating: đánh giá rating.

- members - number of community members that are in this anime's

- "group“: số lượng thành viên của nhóm.

Shape of The Anime Dataset : (12294, 7)



Hình 9. File anime.csv

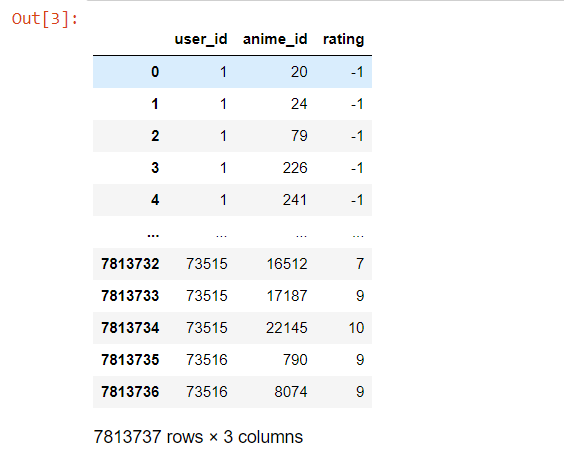
**\*File ranting.csv**

Tệp này chứa các đánh giá xếp hạng của người dùng.

- user\_id: id người dùng

- anime\_id: id anime mà người dùng này đã đánh giá.

- rating: xếp hạng trong số 10 người dùng này đã chỉ định (-1 nếu người dùng đã xem nhưng không chỉ định xếp hạng)

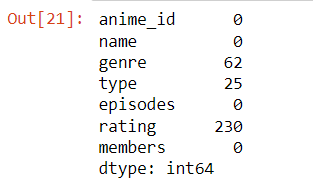


Hình 10. file ranting.csv

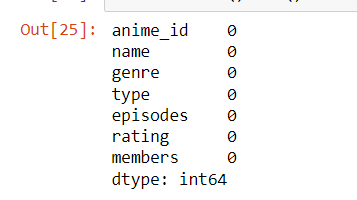
**2.2. XỬ LÝ DỮ LIỆU**

**2.2.1. Anime Dataset**

Kiểm tra missing values

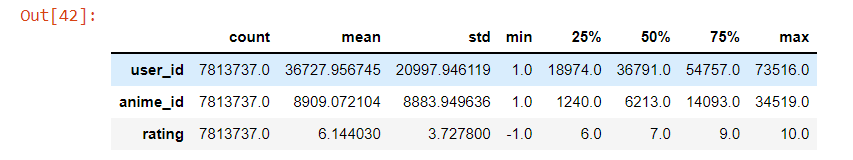


Xóa các missing values và kiểm tra lại dữ liệu

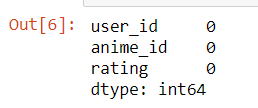


* + 1. **Ranting Dataset**

Summary ranting.csv



* + Kiểm tra missing values

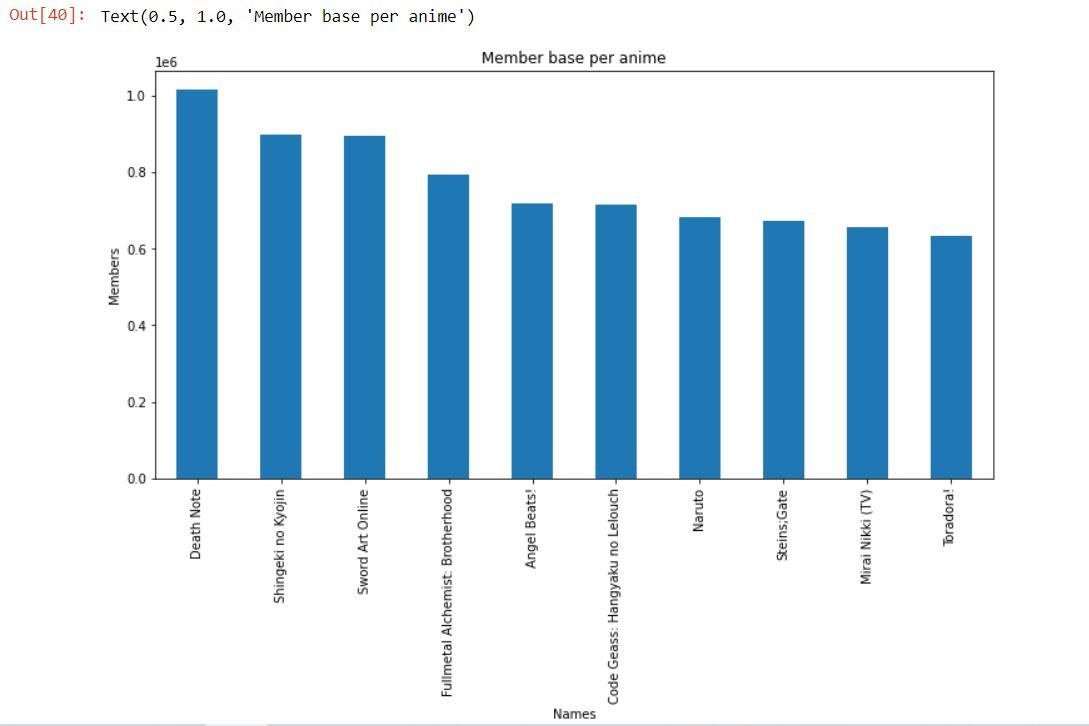


* 1. **EDA (Exploratory Data Analysis)**

Nối 2 dataframe lại để tạo thành một master dataframe phục vụ cho nhiệm vụ trực quan hóa dữ liệu.

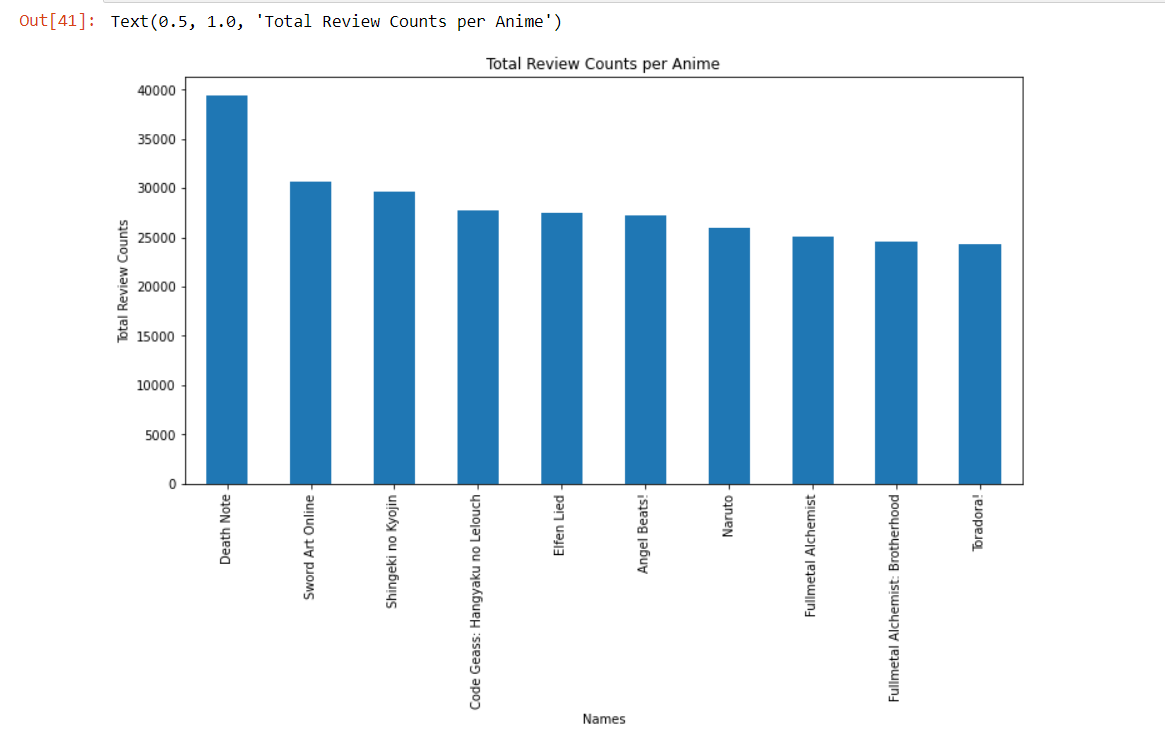
****

Hình 11. Master dataframe

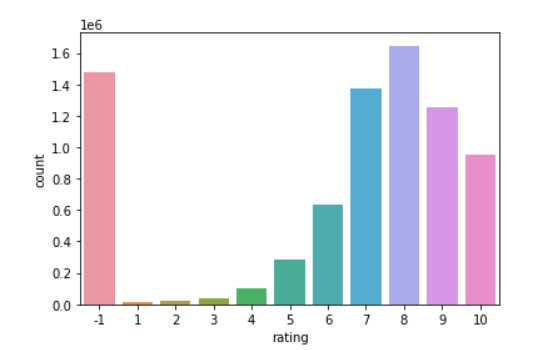
****

Hình 12. Top 10 anmie có group member cao nhất

Ta thấy được ở hình 2.3.2 cho thấy top 10 bộ anime có số lượng thành viên tham gia nhóm cộng đồng cao nhất. Và cao nhất là bộ Death Note, tiếp theo là Shingeki no Kyojin và Sword Art Online.

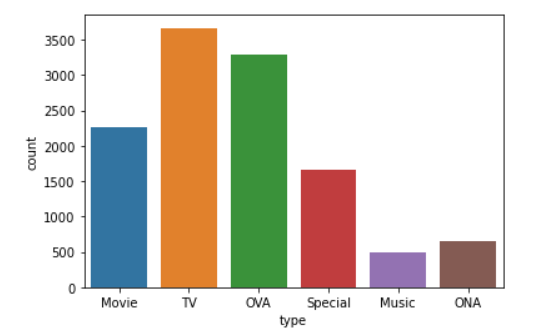


Hình 13. Top 10 Anime có tống review cao nhất

Tương tự ở hình 2.3.2 thể hiện tổng số lượng review của 10 bộ anime có số review cao nhất. Đứng ở đầu tiên là Death Note, tiếp theo là Sword Art Online và Shingeki noKyojin.

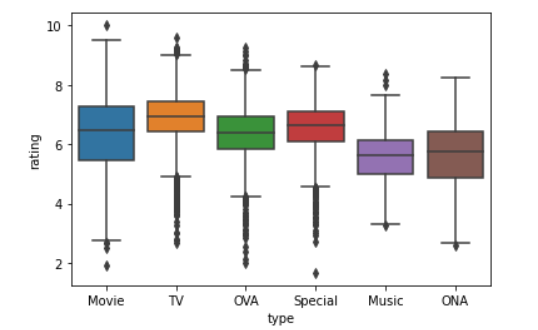
Hình 14. Biểu đồ thể hiện rating

Hình 2.3.3 cho thấy rating phân bố đa số phân bố ở mức từ 6-10. Cao nhất là ở mức 8. Sau đó là ở mức 7, 9, 10.



Hình 15. Biểu đồ thể hiện thể loại của anime

Biểu đồ hình 3.3.4 thể hiện loại của anime. Ta thấy số lượng anime thuộc thể loại TV là có số lượng cao nhất khoảng trên 3500 bộ. Tiếp theo là các thể loại OVA, Movie.

****

Hình 16. Biểu đồ thể hiện rating theo thể loại anime

Hình 3.3.5 Thể hiện mức rating của từng thể loại phim. Ở thể loại TV thì mức rating chủ yếu ở khoảng 7, Movie thì khoảng 5,5 đến 7. Tương tự cho các thể loại còn lại, ta có OVA thì ở khoảng mức 6 đến 7, Special ở khoảng 6 đến 7, Music khoảng 5 đến 6, ONA ở khoảng 5 đến 6.5.

**CHƯƠNG 3: XÂY DỰNG MÔ HÌNH RECOMMEND SYSTEM**

**3.1. XÂY DỰNG MÔ HÌNH DỰA TRÊN THUẬT TOÁN ALS**

**3.1.1. Thuật toán ALS**

Alternating Least Square (ALS) là một thuật toán phân tích matrix factorization và nó tự chạy theo kiểu song song. ALS được triển khai trong Apache Spark ML và được xây dựng cho các vấn đề larges-scale collaborative filtering. ALS đang thực hiện khá tốt công việc giải quyết khả năng mở rộng và độ thưa thớt của dữ liệu ranting, đồng thời nó đơn giản và chia tỷ lệ tốt cho các tập dữ liệu rất lớn.

Cũng giống như các thuật toán học máy khác, ALS có bộ siêu tham số riêng. Chúng tôi có thể muốn điều chỉnh các siêu tham số của nó thông qua hold-out validation or cross-validation.

Các siêu tham số quan trọng nhất trong Alternating Least Square (ALS):

* maxIter: số lần lặp tối đa để chạy (mặc định là 10)
* rank: số lượng các yếu tố tiềm ẩn trong mô hình (mặc định là 10)
* regParam: tham số chuẩn hóa trong ALS (mặc định là 1.0)

Hyper-parameter tuning là một nhiệm vụ có tính lặp lại cao trong nhiều dự án máy học. Chúng ta có thể mã hóa nó trong một chức năng để tăng tốc độ lặp lại điều chỉnh.

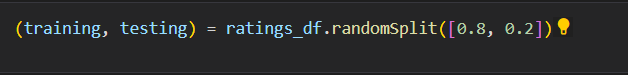
**3.1.2. Training Model**

Trước khi đi vào thử nghiệm từng mô hình, chúng ta tiến hành chia dữ liệu thành 2 tập riêng biệt: Tập training và tập test như sau:

- Chia tập dữ liệu làm 2 phần theo tỉ lệ 8:2:

+ 2 phần là dùng cho tập test.

+ 8 phần là dùng cho tập train.



* + Xây dựng mô hình recommendation sử dụng thuật toán ALS trên tập dữ liệu huấn luyện

**Chỗ này tụi em đang hoàn thiện về code nên tụi em sẽ bổ sung sau ạ**

**3.2. XÂY DỰNG MÔ HÌNH DỰA TRÊN KỸ THUẬT TF-IDF**

**3.2.1. Kỹ thuật TF-IDF**

TF-IDF viết tắt của term frequency-inverse document frequency, là một con số thể hiện tầm quan trọng thống kê của bất kỳ từ cụ thể nào đối với toàn bộ bộ sưu tập tài liệu.

TF-IDF được tính bằng cách nhân term frequency và inverse document frequency.

****

* TF: Số lần một từ xuất hiện trong tài liệu/số từ trong tài liệu.
* IDF: log(Số tài liệu/Số tài liệu chứa từ).

**PHẦN 3: KẾT LUẬN**

1. **KẾT QUẢ ĐẠT ĐƯỢC**

Tìm hiểu được cách hoạt động và các khái niệm của một hệ thống gợi ý. Cũng như tìm hiểu về các thuật toán dự đoán.

Chúng em đã xây dựng được một mô hình gợi ý phim dựa trên thuật toán Collaborative Filtering (lọc cộng tác) với độ chính xác đạt gần ~93%.

Xử lý được model trên tập dữ liệu lớn áp dụng Apache Spark RDD trong quá trình xử lý dữ liệu, train và test model.

1. **HẠN CHẾ**

Nhược điểm của hệ thống:

* + Dữ liệu quá lớn và xây dựng mô hình phức tạp nên thời gian xử lý chưa đáp ứng tốc độ cao.
  + Chưa phát triển lên thành giao diện Web cho hệ thống

1. **HƯỚNG PHÁT TRIỂN**

Hướng phát triển của nhóm:

+ Xây dựng giao diện người dùng cho hệ thống.

+ Tối ưu hoá hệ thống giúp gợi ý nhanh hơn và chính xác hơn. Gợi ý cho người dùng chính xác đến các tầng dữ liệu chi tiết hơn ví dụ là quốc gia, giới tính, khu vực và độ tuổi. Xây dựng gợi ý phim trên mô hình spark streaming để năng cao trải nghiệm và xử lý của hệ thống.

+ Tối ưu hoá hơn thuật toán Collaborative Filtering trên tập dữ liệu về đánh giá của người dùng trên thể loại phim

+ Hoàn thiện hệ thống gợi ý phim và đóng gói thành một package, giúp cho các trang web xem phim có thể nâng cao khả năng tương tác người dùng như: Netflix, Disney,…

+ Xây dựng thêm các API như là: comment, lọc comment, thời lượng xem phim của mỗi người dùng.

# **TÀI LIỆU THAM KHẢO**

[1] Rohit Dwivedi, What Are Recommendation Systems in Machine Learning? Retrieved December 10, 2022, from <https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning>  
[2] Michael D. Ekstrand, Boise State University, Matrix Factorization from <https://md.ekstrandom.net/teaching/recsys/lecture-notes.pdf>  
[3] Kevin Liao, Nov 17, 2018, Alternating Least Square (ALS) from https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1  
[5] Congcong Wang, Collaborative Filtering Algorithms from <https://wangcongcong123.github.io/files/bechelor-thesis/report.pdf>   
[6] Tiep Vu Huu, Content-based Recommendation Systems, from https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/