BỘ GIÁO DỤC & ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

----ଓଓ📖ଌଌ----

HCMUTE

BÁO CÁO CUỐI KỲ

ĐỀ TÀI: BUILD A MUSIC RECOMMENDATION SYSTEM FOR USERS

GVHD: ThS. Quách Đình Hoàng

SVTH :

1. **Nguyễn Anh Đắc** **19133020**
2. **Nguyễn Thanh Tân Kỷ** **19133031**
3. **Đào Thị Cẩm Tiên** **19133055**
4. **Lại Hữu Trác** **19133059**

TP.Hồ Chí Minh, ngày 20 tháng 12 năm 2022

DANH SÁCH THÀNH VIÊN NHÓM VÀ NHIỆM VỤ

STT	Họ Tên	MSSV	Mức độ hoàn thành
17	Nguyễn Anh Đắc	19133020	100%
26	Nguyễn Thanh Tân Kỷ	19133031	100%
46	Đào Thị Cẩm Tiên	19133055	100%
48	Lại Hữu Trác	19133059	100%

ĐIỂM: .....

Nhận xét của giáo viên

.....

.....

.....

.....

.....

.....

.....

.....

Ngày….tháng…..năm 2022

Giáo viên chấm điểm

GVHD ThS. Quách Đình Hoàng

LỜI CẢM ƠN

Lời đầu tiên, để hoàn thành tốt đề tài và báo cáo cho môn *Big Data Applications: Machine Learning At Scale*, nhóm chúng em xin gửi lời cảm ơn chân thành đến giảng viên Quách Đình Hoàng, người đã trực tiếp giảng dạy về kiến thức, hỗ trợ chúng em trong suốt quá trình làm đề tài này. Nhờ thầy đưa ra những lời khuyên từ kinh nghiệm thực tiễn của mình để định hướng cho chúng em đi đúng hướng với đề tài đã chọn, thầy luôn tận tình giải đáp các thắc mắc một cách chi tiết trong suốt quá trình học để chúng em có thể có thêm kiến thức để thực hiện đề tài.

Vì đề tài của chúng em thực hiện trong thời gian không quá dài nên sẽ không thể tránh khỏi những sai sót và những mặt chưa hoàn thiện về mặt kỹ thuật cũng như là cách trình bày, chúng em mong thầy thông cảm bỏ qua những sai sót cho nhóm chúng em.

Cuối cùng, chúng em xin chúc quý thầy có thật nhiều sức khỏe, thành công hơn trên con đường sự nghiệp của mình. Chúng em xin chân thành cảm ơn.

MỤC LỤC

1. Tóm tắt (abstract)

1.1. Lý do chọn đề tài:

Hiện nhu cầu giải trí, thư giãn của con người ngày càng cao. Nghe nhạc cũng là một cách để giúp chúng ta thư giãn, nâng cao tinh thần sau những giờ làm việc mệt mỏi, tạo ra năng lượng tích cực.

Mặt khác việc bạn chỉ nghe đi nghe lại một số bài hát tạo ra sự nhàm chán cho chính mình. Vì thế để có thể mở rộng được gu âm nhạc của người dùng, thêm tính đa dạng thì nhóm em quyết định sẽ vận dụng kiến thức mà nhóm em đã được học ở môn này để xây dựng một *hệ thống gợi ý âm nhạc cho người dùng*.

1.2. Các phương pháp sử dụng:

- Nhóm em sẽ thực hiện theo cách tiếp cận Collaborative filtering - xây dựng hệ thống gợi ý dựa trên các ngươi dùng tương đồng.
- *Thuật toán sử dụng:* Alternating least squares (bình phương tối thiểu xen kẽ)
- *Độ đo:* RMSE

1.3. Kết quả:

- Kết quả đối với việc gợi ý bài hát cho người dùng là:

  Mô hình tốt nhất:

  - Rank: 20
  - Regularization: 0.1
  - RMSE: 6.26

  Kiểm tra trên tập test:

  - RMSE: 7.28

Gợi ý 5 bài hát khi truyền vào 1 user_id:

song_id	title	release	artist_name	year	prediction
SOJIPLZ12A6D4F6110	Right Where I Nee...	Greatest Hits	Gary Allan	1999	24.0
SOFJCCE12AB0183F96	Faith	Skunkworks	Bruce Dickinson	0	28.0
SODXXYB12AB0189FA6	Stratus [The Bott...	The Ultimate: Red...	Tommy Bolin	0	28.0
SOGHSMH12A8C137927	Skyway Avenue	We The Kings	We The Kings	2007	33.0
SOFXSLW12A6D4F7BF2	Another Great Divide	The Collection	Split Enz	1981	33.0

- Kết quả đối với việc gợi ý nghệ sĩ cho người dùng là:

  Mô hình tốt nhất:

  - Rank: 15
  - Regularization: 0.1
  - RMSE: 7.48

  Kiểm tra trên tập test:

  - RMSE: 10.34

Gợi ý 5 nghệ sĩ khi truyền vào 1 user_id:

artist_name	prediction
moe.	48.0
Black Crowes	53.0
Theatre Of Tragedy	54.0
Savatage	78.0
keller williams	91.0

2. Giới thiệu (introduction)

Hiện này hệ thống gợi ý (recommend system) đang trở thành một lĩnh vực nghiên cứu thú vị mà nhiều lập trình viên cũng nhà nghiên cứu quan tâm, bởi vì tính ứng dụng vào thực tiễn cao của nó, giúp người dùng đối phó với việc quá tải thông tin.

Và cùng với sự phát triển rất mạnh mẽ của lĩnh vực truyền thông đa phương tiện như hiện nay thì âm nhạc đã trở thành nhu cầu không thể thiếu trong cuộc sống. Thế nhưng số lượng bài nhạc đang ngày càng tăng lên, đa dạng và phong phú cả về nội dung lẫn thể loại. Vì vậy, vấn đề xảy ra đó chính là người dùng sẽ khó khăn trong việc tìm được một bài hát phù hợp với sở thích âm nhạc của bản thân. Do đó, việc xây dựng một hệ thống gợi ý âm nhạc cho người dùng là rất cần thiết cho hiện nay.

- Bài toán gợi ý bài hát cho người dùng:
  - Input: Tập dữ liệu về user (user_id, song_id, listen_count (lượt nghe cho bài hát)).
  - Output: List bài hát phù hợp với người dùng.
- Bài toán gợi ý nghệ sĩ cho người dùng:
  - Input: user_id, artist_name, listen_count (lượt nghe cho nghệ sĩ).
  - Output: List nghệ sĩ phù hợp với người dùng.

3. Dữ liệu (data)

3.1. Giới thiệu với tập dữ liệu

Million Song là một bộ dữ liệu gồm các tính năng và siêu dữ liệu âm thanh được cung cấp miễn phí cho một triệu bản nhạc nổi tiếng đương thời.

Bộ dữ liệu *Million Song* bắt đầu là một dự án hợp tác giữa The Echo Nest và LabROSA. Nó được hỗ trợ một phần bởi NSF.

- Dataset Source: http://labrosa.ee.columbia.edu/millionsong/
- Paper: http://ismir2011.ismir.net/papers/OS6-1.pdf

Nhóm sẽ không sử dụng trực tiếp tập dữ liệu này, nhưng nhóm sẽ sử dụng một số phần của nó.

- Tập dữ liệu: Million Song - Recommendation Engines (Gồm có "10000.txt" và "song_data.csv")
- Link dataset: https://www.kaggle.com/code/mgmarques/million-song-recommendation-engines/data

Tập dữ liệu về user (10000.txt) có 3 cột và 2000000 dòng chứa số lượt phát của người dùng ẩn danh cho các bài hát có trong tập dữ liệu triệu bài hát

- *user_id:* ID của người dùng
- *song_id:* ID của bài hát
- *listen_count:* Lượt nghe của bài hát

	user_id	song_id	listen_count
0	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAKIMP12A8C130995	1
1	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBBMDR12A8C13253B	2
2	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBXHDL12A81C204C0	1
3	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBYHAJ12A6701BF1D	1
4	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SODACBL12A8C13C273	1

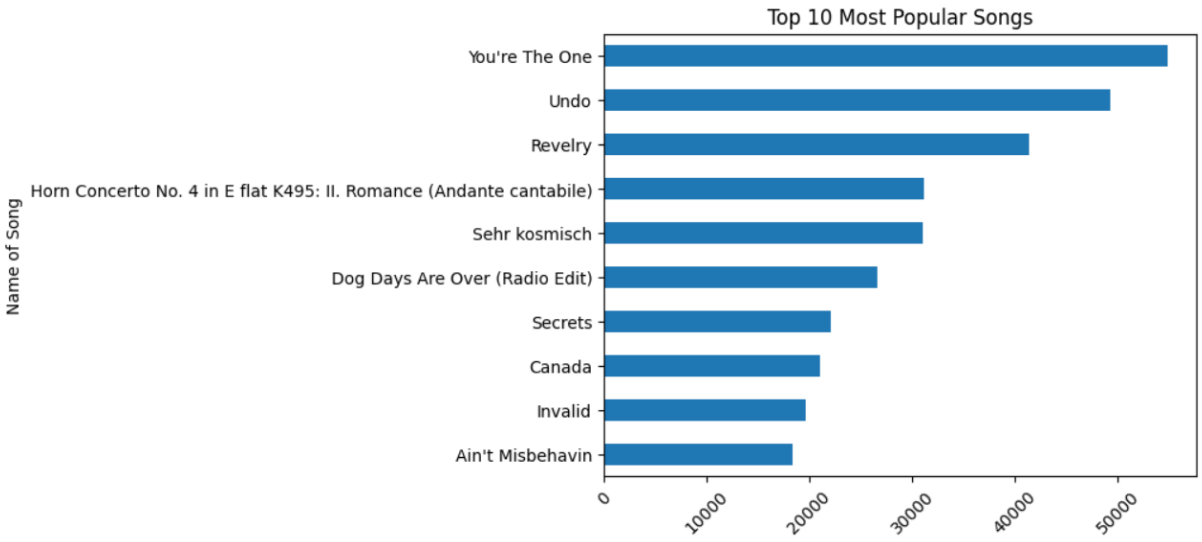
Tập dữ liệu về bài hát (song_data.csv) có 5 cột và 1000000 dòng chứa các thuộc tính của một triệu bài hát

- *song_id:* ID của bài hát
- *title:* Tên bài hát
- *release:* Tên ablum của bài hát đó (
- *artist_name:* Tên nghệ sĩ, ca sĩ
- *year:* Năm bài hát được phát hành

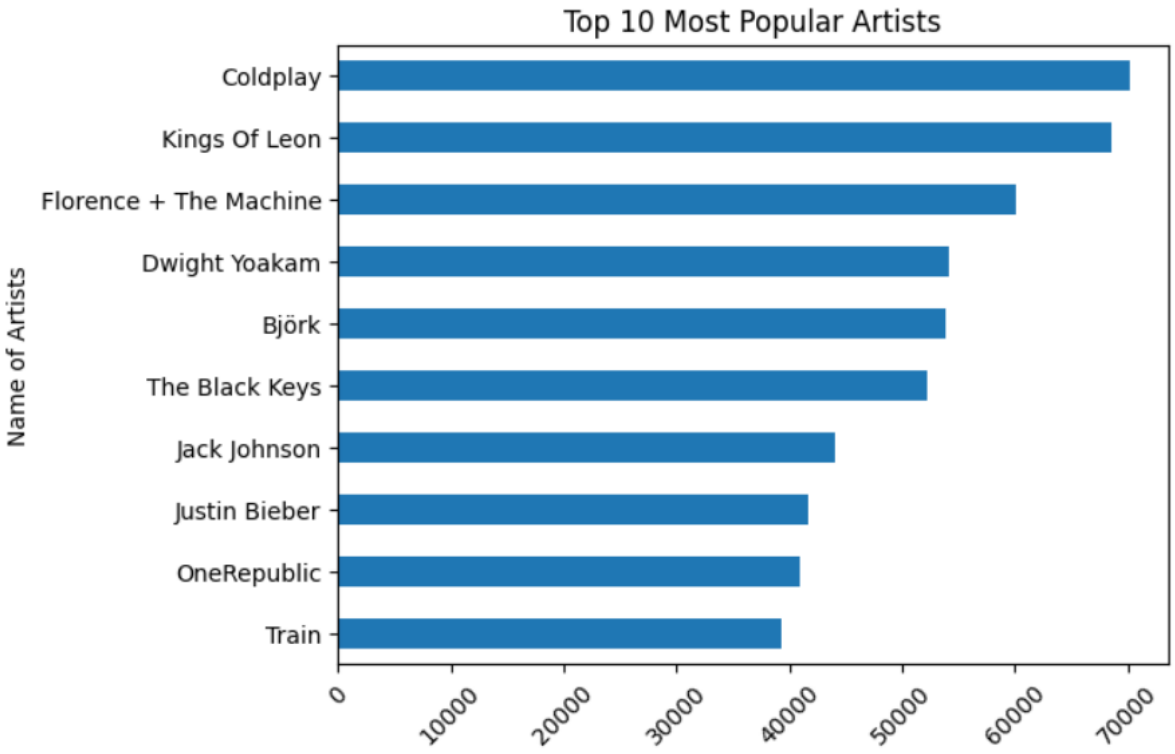
	song_id	title	release	artist_name	year
0	SOQMMHC12AB0180CB8	Silent Night	Monster Ballads X-Mas	Faster Pussy cat	2003
1	SOVFVAK12A8C1350D9	Tanssi vaan	Karkuteillä	Karkkiautomaatti	1995
2	SOGTUKN12AB017F4F1	No One Could Ever	Butter	Hudson Mohawke	2006
3	SOBNYVR12A8C13558C	Si Vos Querés	De Culo	Yerba Brava	2003
4	SOHSBXH12A8C13B0DF	Tangle Of Aspens	Rene Ablaze Presents Winter Sessions	Der Mystic	0

3.2. Trực quan hóa dữ liệu và EDA:

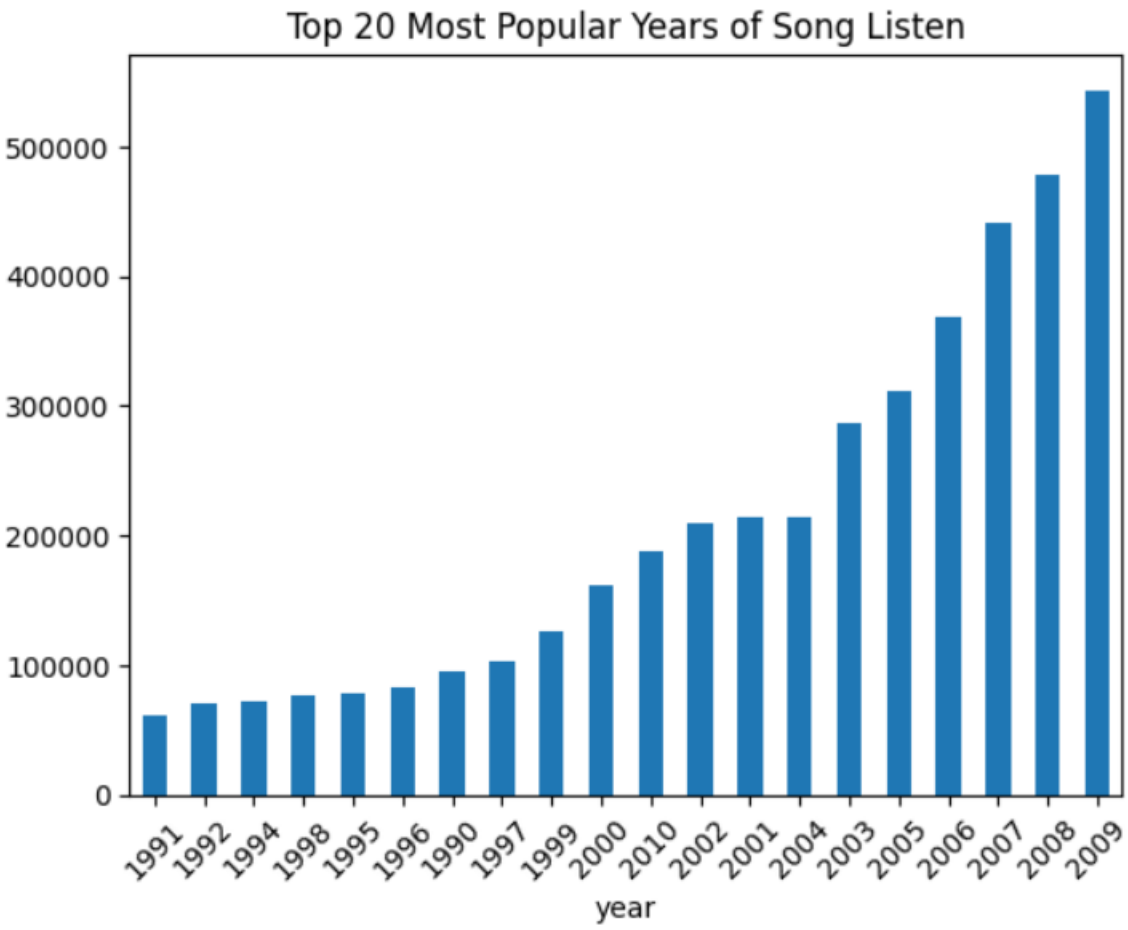
Biểu đồ thể hiện top 10 nghệ sĩ nổi tiếng nhất:



Biểu đồ thể hiện top 10 nghệ sĩ nổi tiếng nhất:



Biểu đồ thể hiện top 20 năm có lượt nghe nhiều nhất:



3.3. Chia tập dữ liệu

Để có thể tiến hành train mô hình, nhóm em quyết định chia tập dữ liệu thành 3 phần theo tỉ lệ như sau: 6 : 2 : 2

- Tập training: 60%
- Tập validation: 20%
- Tập test: 20%

4. Phương pháp (methods)

Có 3 cách tiếp cận hệ thống gợi ý:

- *Content-based:* Đưa ra các khuyến nghị mua bán cho người dùng dựa trên nội dung liên quan đến sản phẩm. Chẳng hạn một bài hát với các đặc điểm như: người biểu diễn, năm phát hành, thể loại,...
- *Collaborative filtering:* Hay còn gọi là lọc cộng tác, sử dụng sự tương tác qua lại trong hành vi mua sắm giữa các khách hàng để tìm ra sở thích của

một khách hàng đối với một sản phẩm. Từ đó sẽ khuyến nghị sản phẩm cho khách hàng dựa trên hành vi của các khách hàng khác liên quan nhất.

- *Kết hợp cả 2 phương pháp*: Ngoài ra chúng ta cũng có thể sử dụng kết hợp cả 2 phương pháp trên để tạo thành một thuật toán kết hợp. Ưu điểm của phương pháp này đó là vừa tận dụng được các thông tin từ phía sản phẩm và các thông tin về hành vi mua sắm của người dùng.

Dựa theo tập dữ liệu của nhóm thì tụi em quyết định sẽ thực hiện:

- Cách tiếp cận Collaborative filtering - xây dựng hệ thống gợi ý dựa trên các người dùng tương đồng.
- Thuật toán sử dụng: Alternating least squares (bình phương tối thiểu xen kẽ)
- Độ đo: RMSE

4.1. Collaborative Filtering

Collaborative filtering là thuật toán lọc tương tác tức là tìm ra sản phẩm mà khách hàng có khả năng ưa thích nhất dựa vào những sản phẩm mà những khách hàng khác có hành vi tương tự đã lựa chọn.

Thuật toán sẽ không cần sử dụng thông tin sản phẩm là đầu vào cho dự báo rating. Đầu vào của thuật toán là một ma trận tiện ích (ultility matrix) chứa giá trị rating của các cặp (user, item).

Mỗi cột là các rating mà một user đã rate và mỗi dòng là các rating của một item được rate.

Có 2 phương pháp chính được sử dụng trong collaborative filtering bao gồm: Neighborhood-based collaborative Filtering và Matrix Fractorization.

4.1.1. Neighborhood-Based Collaborative Filtering

Ý tưởng cơ bản của phương pháp này chính là xác định mức độ quan tâm của một user với một item dựa trên các users khác gần giống với user này..

Ở phương pháp này ta sẽ cần xây dựng ma trận hệ số tương quan của vector rating của các users để tìm ra nhóm users có cùng sở thích. Hệ số tương quan giữa các users càng lớn thì sở thích của họ càng giống nhau và trái lại thì họ càng có sở thích khác biệt.

Thuật toán sẽ dự đoán giá trị rating tại một cặp (user, item) chưa được rate bằng cách tính tổng có trọng số các giá trị rating của k users tương

quan nhất với user đó mà đã rate item trên. Trọng số thông thường sẽ bằng chính hệ số tương quan.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	3	?	?	0	?	?	?
i_2	?	4	1	?	?	1	2
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5

Hình 1. Utility matrix dựa trên số sao một user rate cho một item

Đặt *mức độ giống nhau* của hai *users* ui,uj là sim(ui,uj). Một *similiarity function* tốt cần đảm bảo:

$$\text{sim}(u_0, u_1) > \text{sim}(u_0, u_i),\ \forall i > 1.$$

Thuật toán sẽ trải qua lần lượt các bước sau đây:

- Chuẩn hóa dữ liệu ở ma trận Y tiện ích bằng cách trừ đi ở mỗi cột (là các rating của cùng 1 user) trung bình giá trị rating của cột. Giá trị rating dương thể hiện user ưa thích item và trái lại âm sẽ là không thích, bằng 0 là trung lập.
- Tính ma trận hệ số tương quan giữa các véc tơ cột. Hệ số tương quan dương và càng gần 1 chứng tỏ 2 users có sở thích giống nhau. Hệ số tương quan âm là 2 users có hành vi trái ngược.
  Đây là hàm tính độ similarity của hai vector:

$$\text{cosine\_similarity}(\mathbf{u}_1, \mathbf{u}_2) = \cos(\mathbf{u}_1, \mathbf{u}_2) = \frac{\mathbf{u}_1^T \mathbf{u}_2}{||\mathbf{u}_1||_2 . ||\mathbf{u}_2||_2}$$

- Giá trị dự báo rating của user u sẽ được tính bằng tổng có trọng số của các rating trong tập k users tương quan nêu trên theo công thức bên dưới:

$$\hat{y}_{i,u} = \frac{\sum_{u_j \in \mathcal{N}(u,i)} \bar{y}_{i,u_j} \text{sim}(u, u_j)}{\sum_{u_j \in \mathcal{N}(u,i)} |\text{sim}(u, u_j)|}$$

- Chuyển giá trị dự báo ở ma trận chuẩn hóa sang giá trị dự báo rating bằng cách cộng các giá trị ở ma trận chuẩn hóa với giá trị trung bình của mỗi cột.

4.1.2. Matrix Factorization Collaborative Filtering

Ngoài ra còn một phương pháp collaborative filtering khác dựa trên một phép phân rã ma trận (matrix fractorization). Tức là chúng ta sẽ phân tích ma trận tiện ích thành tích của các ma trận items và ma trận users.

$$\mathbf{Y} \approx \begin{bmatrix} \mathbf{x}_1\mathbf{w}_1 & \mathbf{x}_1\mathbf{w}_2 & \dots & \mathbf{x}_1\mathbf{w}_M \\ \mathbf{x}_2\mathbf{w}_1 & \mathbf{x}_2\mathbf{w}_2 & \dots & \mathbf{x}_2\mathbf{w}_M \\ \dots & \dots & \ddots & \dots \\ \mathbf{x}_N\mathbf{w}_1 & \mathbf{x}_N\mathbf{w}_2 & \dots & \mathbf{x}_N\mathbf{w}_M \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_N \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_M \end{bmatrix} = \mathbf{XW}$$

Chúng ta sẽ đi tìm X và W bằng cách thay phiên cố định 1 biến rồi tối ưu hàm theo biến còn lại. Việc này lặp lại cho đến khi các điểm hội tụ (Y xấp xỉ X*W)

- Y: chứa mức độ quan tâm(ưa thích) của người dùng x với item w.
- X: ma trận yếu tố tiềm ẩn user
- W: ma trận yếu tố tiềm ẩn item

Xây dựng hàm mất mát cho Matrix Factorization:

$$\mathcal{L}(\mathbf{X}, \mathbf{W}) = \frac{1}{2s} \sum_{n=1}^{N} \sum_{m:r_{mn}=1} (y_{mn} - \mathbf{x}_m \mathbf{w}_n)^2 + \frac{\lambda}{2} (||\mathbf{X}||_F^2 + ||\mathbf{W}||_F^2)$$

4.2. Phương pháp Alternating least squares (ALS)

ALS hay còn gọi là phương pháp bình phương tối thiểu thay phiên nhau. Gọi là thay phiên nhau bởi hàm mất mát trên có 2 biến khiến hàm không lồi (khó để tìm cực tiểu) chính vì vậy ta thay phiên cố định 1 biến rồi tối ưu hàm theo biến còn lại.

Khi cố định 1 biến hàm trở thành hàm bậc 2 và có thể giải bằng phương pháp bình phương tối thiểu. Việc cố định 1 biến rồi tính toán lại biến còn lại được thực hiện lần lượt cho đến khi các điểm hội tụ lại tại điểm cực tiểu.

Có 2 trường hợp mà ALS lại được ưu tiên hơn.

- Trường hợp 1: Là hệ thống có khả năng chạy các tiến trình 1 cách song song.

Ta thấy khi cố định 1 trong 2 cột công thức hàm cần tối ưu sẽ trở thành

$$\sum_{(u,i) \in K} (r_u i - q_i^\tau p_u)^2 + \lambda ||q_i||^2$$

Từ đó bài toán có thể chuyển thành việc tối ưu từng cột 1 của q (hoặc p) cho nên việc sử dụng trong hệ thống song song sẽ dễ dàng hơn.

- Trường hợp thứ 2: Là khi sử dụng với hệ thống tập trung vào các dữ liệu ẩn.

5. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussions)

5.1. Bài toán gợi ý bài hát cho người dùng dựa vào số lượng lượt nghe

Nhóm em thực hiện ALS với cái tham số là: rank, maxIter, regularizations

Xử lý dữ liệu đầu vào: string → number

```
+----------------------------------------+------------------+------------+-------------+-------------+
|user_id                                 |song_id           |listen_count|user_id_index|song_id_index|
+----------------------------------------+------------------+------------+-------------+-------------+
|b80344d063b5ccb3212f76538f3d9e43d87dca9e|SOAKIMP12A8C130995|1           |15203.0      |2244.0       |
|b80344d063b5ccb3212f76538f3d9e43d87dca9e|SODACBL12A8C13C273|1           |15203.0      |205.0        |
|b80344d063b5ccb3212f76538f3d9e43d87dca9e|SONSAEZ12A8C138D7A|1           |15203.0      |1002.0       |
|b80344d063b5ccb3212f76538f3d9e43d87dca9e|SORQHCG12A58A7EEBA|1           |15203.0      |9485.0       |
|b80344d063b5ccb3212f76538f3d9e43d87dca9e|SOVQEYZ12A8C1379D8|1           |15203.0      |3351.0       |
+----------------------------------------+------------------+------------+-------------+-------------+
only showing top 5 rows
```

Train mô hình:

- Thực hiện train mô hình với: ranks = [5, 10, 15, 20]; regularizations = [0.1, 0.05]
- Đánh giá RMSE

```
For rank 5, regularization parameter 0.1 the RMSE is 7.0107423339157675
For rank 10, regularization parameter 0.1 the RMSE is 6.401843392713502
For rank 15, regularization parameter 0.1 the RMSE is 6.2826838735159685
For rank 20, regularization parameter 0.1 the RMSE is 6.26138702853378
For rank 5, regularization parameter 0.05 the RMSE is 7.841366074425806
For rank 10, regularization parameter 0.05 the RMSE is 6.678715426708202
For rank 15, regularization parameter 0.05 the RMSE is 6.403918716528148
For rank 20, regularization parameter 0.05 the RMSE is 6.363069432040251
For rank 5, regularization parameter 0.01 the RMSE is 11.6969161993882
For rank 10, regularization parameter 0.01 the RMSE is 8.306544364588797
For rank 15, regularization parameter 0.01 the RMSE is 7.034754907585259
For rank 20, regularization parameter 0.01 the RMSE is 6.675167082808299
```

- Chọn mô hình tốt nhất: Ở đây chúng ta thấy bộ tham số tốt nhất cho mô hình là: rank = 20, regularization = 0.1, RMSE = 6.26

```
The best model was trained with regularization parameter 0.1
The best model was trained with rank 20
```

Kiểm tra lại trên tập test, ta có:

```
The model have RMSE on the test:7.28147108325175
```

Kết quả gợi ý 5 bài hát cho 1 user (user_id = "b80344d063b5ccb3212f76538f3d9e43d87dca9e"):

song_id	title	release	artist_name	year	prediction
SOJIPLZ12A6D4F6110	Right Where I Nee...	Greatest Hits	Gary Allan	1999	24.0
SOFJCCE12AB0183F96	Faith	Skunkworks	Bruce Dickinson	0	28.0
SODXXYB12AB0189FA6	Stratus [The Bott...	The Ultimate: Red...	Tommy Bolin	0	28.0
SOGHSMH12A8C137927	Skyway Avenue	We The Kings	We The Kings	2007	33.0
SOFXSLW12A6D4F7BF2	Another Great Divide	The Collection	Split Enz	1981	33.0

Kết quả gợi ý 5 bài hát cho 5 user:

user_id_index	recommendations
26	[{8214, 190.94931}, {8410, 147.03552}, {8003, 70.20171}, {9171, 66.13743}, {1118, 57.16788}]
27	[{8214, 59.767647}, {5161, 57.633797}, {9316, 47.130135}, {8410, 40.74507}, {5269, 40.279724}]
28	[{8214, 209.73056}, {1738, 107.643425}, {2494, 60.912422}, {8654, 52.939365}, {7165, 46.89265}]
31	[{1738, 99.383934}, {8214, 78.16559}, {9729, 31.826645}, {5840, 29.07843}, {1223, 26.124235}]
34	[{3749, 60.647232}, {8214, 46.66249}, {7012, 45.589367}, {7629, 38.59021}, {4059, 38.23446}]

only showing top 5 rows

5.2. Bài toán gợi ý nghệ sĩ cho người dùng dựa vào số lượng lượt nghe của nghệ sĩ theo từng user:

Tương tự ở trên nhóm em cũng thực hiện ALS với cái tham số là: rank, maxIter, regularizations

Trước khi train mô hình, nhóm em thực hiện tính số lượng lượt nghe của nghệ sĩ theo từng user:

```
+--------------------+--------------------+------------+
|             user_id|         artist_name|listen_count|
+--------------------+--------------------+------------+
|c9c44b6a8e8728e70...|           Seu Jorge|           3|
|8d4f1822e21f0a91f...|           Seu Jorge|           1|
|233036b55aad72c53...|        Jimi Hendrix|           1|
|39fdb665ce16d51ef...|               Feist|           3|
|39ba95440835a0c43...|           Daft Punk|           2|
|fe979a7b199de3ee8...|        The Ventures|           1|
|bb21097bd36275b07...|     Curtis Mayfield|           1|
|7bc3c518191756ab3...|         The Prodigy|           2|
|a883f0da4b362a7ca...|            Paramore|           1|
|6a944bfe30ae8d6b8...|          Ryan Adams|          43|
|cd47f11d66541dab1...|       Faith No More|          14|
|d3ea5a391823b95d2...|     Aretha Franklin|           1|
|95e8810b2f4234c90...|   The Lonely Island|           3|
|14e25113a07fe7960...|Five Finger Death...|           1|
|9442f74726a2f682c...| Simian Mobile Disco|           4|
|1e3483eb876e42290...|                 Can|           1|
|bc4ac44e63282902a...|   Empire Of The Sun|           3|
|24bedcf0dde551385...|         James Blunt|           2|
|629b64c113fedb5b1...|           Metallica|           1|
|a06c2f59c56b356db...|             Incubus|           2|
+--------------------+--------------------+------------+
only showing top 20 rows
```

Xử lý dữ liệu đầu vào: string → numbe

user_id	artist_name	listen_count	user_id_index	artist_name_index
c9c44b6a8e8728e70...	Seu Jorge	3	18994.0	1627.0
8d4f1822e21f0a91f...	Seu Jorge	1	801.0	1627.0
233036b55aad72c53...	Jimi Hendrix	1	13319.0	397.0
39fdb665ce16d51ef...	Feist	3	877.0	160.0
39ba95440835a0c43...	Daft Punk	2	2775.0	5.0
fe979a7b199de3ee8...	The Ventures	1	68615.0	1734.0
bb21097bd36275b07...	Curtis Mayfield	1	1108.0	2149.0
7bc3c518191756ab3...	The Prodigy	2	6577.0	240.0
a883f0da4b362a7ca...	Paramore	1	12094.0	37.0
6a944bfe30ae8d6b8...	Ryan Adams	43	27.0	600.0
cd47f11d66541dab1...	Faith No More	14	24010.0	153.0
d3ea5a391823b95d2...	Aretha Franklin	1	1461.0	1673.0
95e8810b2f4234c90...	The Lonely Island	3	14732.0	285.0
14e25113a07fe7960...	Five Finger Death...	1	8793.0	193.0
9442f74726a2f682c...	Simian Mobile Disco	4	4754.0	164.0
1e3483eb876e42290...	Can	1	20410.0	1148.0
bc4ac44e63282902a...	Empire Of The Sun	3	15169.0	419.0
24bedcf0dde551385...	James Blunt	2	20561.0	313.0
629b64c113fedb5b1...	Metallica	1	7715.0	14.0
a06c2f59c56b356db...	Incubus	2	4791.0	94.0

only showing top 20 rows

Train mô hình:

- Thực hiện train mô hình với: ranks = [5, 10, 15, 20]; regularizations = [0.1, 0.05]
- Đánh giá RMSE

```
For rank 5, regularization parameter 0.1 the RMSE is 8.632506581834008
For rank 10, regularization parameter 0.1 the RMSE is 7.693745345753791
For rank 15, regularization parameter 0.1 the RMSE is 7.479916213851013
For rank 20, regularization parameter 0.1 the RMSE is 7.48158516062916
For rank 5, regularization parameter 0.05 the RMSE is 10.019462589349459
For rank 10, regularization parameter 0.05 the RMSE is 8.26054025851876
For rank 15, regularization parameter 0.05 the RMSE is 7.7630352777825555
For rank 20, regularization parameter 0.05 the RMSE is 7.698677613080277
For rank 5, regularization parameter 0.01 the RMSE is 18.69897427007977
For rank 10, regularization parameter 0.01 the RMSE is 11.575792598210501
For rank 15, regularization parameter 0.01 the RMSE is 9.657117824353776
For rank 20, regularization parameter 0.01 the RMSE is 8.70701627323634
```

- Chọn mô hình tốt nhất: Ở đây chúng ta thấy bộ tham số tốt nhất cho mô hình là: rank = 15, regularization = 0.1, RMSE = 7.48

```
The best model was trained with regularization parameter 0.1
The best model was trained with rank 15
```

Kiểm tra lại trên tập test, ta có:

```
The model have RMSE on the test:10.344412762045025
```

Kết quả gợi ý 5 nghệ sĩ cho 1 user (user_id = " b80344d063b5ccb3212f76538f3d9e43d87dca9e"):

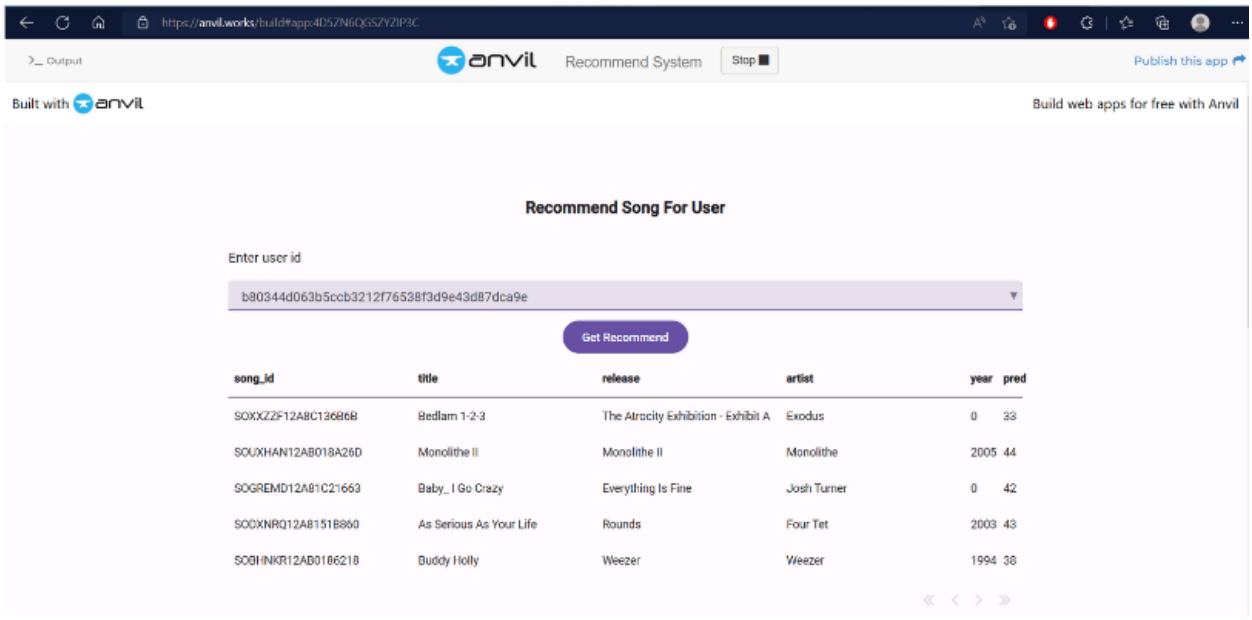
artist_name	prediction
moe.	48.0
Black Crowes	53.0
Theatre Of Tragedy	54.0
Savatage	78.0
keller williams	91.0

Kết quả gợi ý 5 nghệ sĩ cho 5 user:

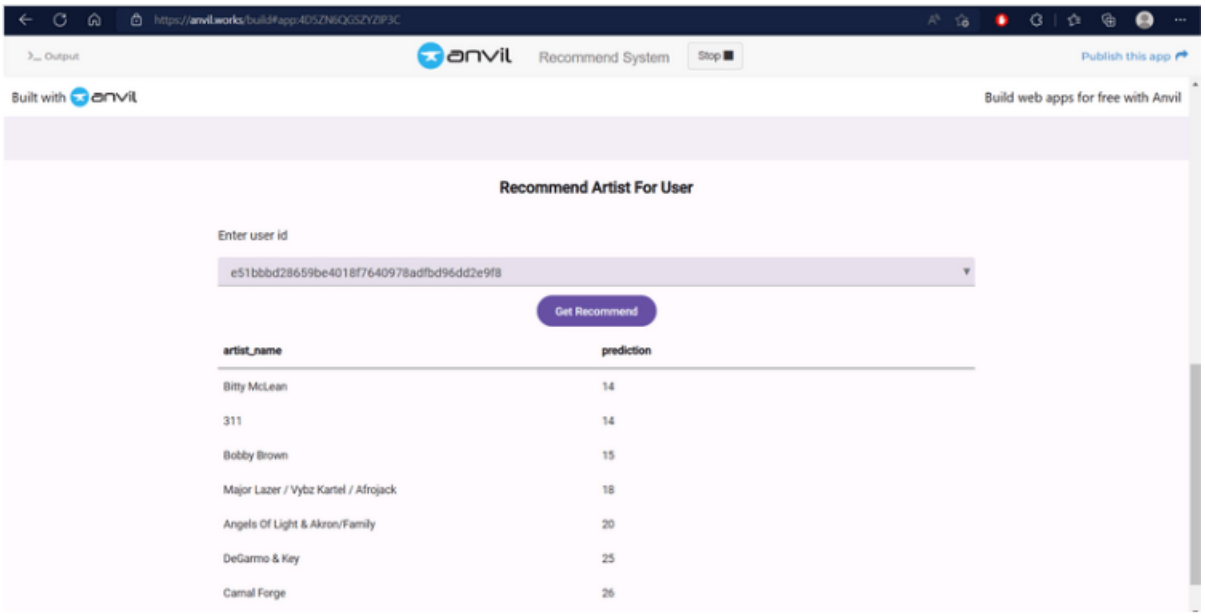
user_id_index	recommendations
26	[{2244, 43.732998}, {1874, 30.248865}, {2164, 29.663994}, {2025, 27.709517}, {2588, 27.577105}]
27	[{2756, 295.4883}, {1337, 203.3607}, {2025, 173.85187}, {2256, 117.48979}, {2195, 109.77941}]
28	[{2588, 158.41528}, {479, 131.66039}, {1411, 91.18557}, {2189, 89.42696}, {2988, 81.30814}]
31	[{1780, 177.97258}, {834, 98.000305}, {3143, 75.89802}, {2472, 75.86464}, {3275, 68.426926}]
34	[{3275, 303.1888}, {2256, 284.3488}, {2756, 244.3869}, {1337, 178.99048}, {2461, 158.01068}]

5.3. Giao diện hệ thống gợi ý trên WebApp Anvil

Giao diện gợi ý bài hát cho user



Giao diện gợi ý nghệ sĩ cho user



6. Kết luận (conclusion)

6.1. Đánh giá

- Dự đoán mô hình dựa trên thuật toán ALS và phương pháp Collaborative Filtering trên các tập.

- Nhóm sử dụng độ đo RMSE (Root Mean Square Error) để đánh giá độ chính xác của mô hình.
- Kết quả đánh giá:

Bài toán	RMSE	
	Best on validation	**Test**
Gợi ý bài hát cho người dùng	6.26	7.28
Gợi ý nghệ sĩ cho người dùng	7.48	10.34

Sau khi kiểm tra độ chính xác dựa vào RMSE, ta thấy:

- Ở bài toán gợi ý bài hát cho người dùng, trên tập validation là 6.28 và trên tập test là 7.28, có thể thấy kết quả này là tốt, mô hình dự đoán chính xác.
- Ở bài toán gợi ý nghệ sĩ cho người dùng, trên tập validation là 7.48 và trên tập test là 10.34, có thể thấy kết quả này khá tốt, không chênh lệch nhiều, mô hình dự đoán tương đối chính xác.

6.2. Hướng phát triển

- Cải thiện tốc độ xử lý của hệ thống gợi ý
- Chỉnh sửa giao diện web đẹp hơn

7. Đóng góp (contributions)

STT	MSSV	Họ Tên	Nhiệm vụ
17	19133020	Nguyễn Anh Đắc	Xây dựng model, đánh giá bài toán gợi ý bài hát cho người dùng
27	19133031	Nguyễn Thanh Tân Kỷ	Tìm data, xử lý dữ liệu, làm báo cáo silde, hỗ trợ xây dựng model
46	19133055	Đào Thị Cẩm Tiên	Xây dựng model, đánh giá bài toán gợi ý nghệ sĩ cho người dùng
48	19133059	Lại Hữu Trác	Trực quan hóa dữ liệu, EDA, build web

8. Tham khảo (references)

[1]. https://machinelearningcoban.com/2017/05/24/collaborativefiltering/
[2]. https://machinelearningcoban.com/2017/05/31/matrixfactorization/
[3].https://phamdinhkhanh.github.io/2019/11/04/Recommendation_Compound_Part1.html
[4]. https://viblo.asia/p/cong-nghe-matrix-factorization-cho-he-thong-goi-y-naQZRJe0Zvx