**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**--------------------**



**ĐỒ ÁN 03**

**XÂY DỰNG HỆ THỐNG GỢI Ý**

**SINH VIÊN THỰC HIỆN:** **MSSV:**

**Lê Văn Quốc 17110212**

**Thạch Hoàng Duy 17110113**

**GIÁO VIÊN HƯỚNG DẪN: Cô Từ Tuyết Hồng**

**TP. HỒ CHÍ MINH - 2020**

Trang Bìa

**LỜI CẢM ƠN**

“Để hoàn thành đồ án môn học này, em xin gửi lời cảm ơn chân thành đến:

Ban giám hiệu trường Đại Học Sư phạm kỹ thuật Thành phố Hồ Chí Minh vì đã tạo điều kiện về cơ sở vật chất với hệ thống thư viện hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin.

Xin cảm ơn giảng viên bộ môn – c**ô Từ Tuyết Hồng** đã giảng dạy tận tình, chi tiết để em có đủ kiến thức và vận dụng chúng vào đồ án môn học này.

Do chưa có nhiều kinh nghiệm làm để tài cũng như những hạn chế về kiến thức, trong bài đồ án chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp, phê bình từ phía cô để bài đồ án môn học được hoàn thiện hơn.

Lời cuối cùng, em xin kính chúc cô nhiều sức khỏe, thành công và hạnh phúc.”

TP.HCM, ngày 15 tháng 12 năm 2020

Sinh viên thực hiện

**Thạch Hoàng Duy 17110113**

**Lê Văn Quốc 17110212**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên hướng dẫn

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 1**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ .........................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên phản biện 1

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 2**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên phản biện 2

**MỤC LỤC**

[Phần 1: Mở đầu 9](#_Toc31473)

[Lời mở đầu 9](#_Toc30140)

[Mục đích, phạm vi nghiên cứu 9](#_Toc20857)

[Phân chia công việc 10](#_Toc26326)

[Phần 2: Nội dung 11](#_Toc17969)

*[Chương 1:](#_Toc15929)* [Cơ sở lý thuyết 11](#_Toc15929)

[1.1. Một số khái niệm cơ bản 11](#_Toc30342)

[1.1.1. Định nghĩa hệ thống gợi ý 11](#_Toc1119)

[1.1.2. Các công năng của hệ thống gợi ý 11](#_Toc11972)

[1.1.3. Các hướng tiếp cận 11](#_Toc2871)

*[1.1.3.1.](#_Toc7335)* [Collaborative filtering 12](#_Toc7335)

[1.1.3.1.1. User-base 12](#_Toc1775)

[1.1.3.1.2. Item-base 13](#_Toc29565)

*[1.1.3.2.](#_Toc22028)* [Content-based filtering 13](#_Toc22028)

*[1.1.3.3.](#_Toc30311)* [Item cold start 13](#_Toc30311)

*[1.1.3.4.](#_Toc15153)* [Hybrid 13](#_Toc15153)

*[Chương 2:](#_Toc16752)* [Tổng quan thuật toán 13](#_Toc16752)

[2.1. Matrix Factorization 13](#_Toc7757)

[2.1.1. Giới thiệu 13](#_Toc12439)

[2.1.2. Hàm mất mát 14](#_Toc22535)

*[2.1.2.1.](#_Toc31748)* [Tối ưu hàm mất mát 14](#_Toc31748)

[2.2. User-user Collaborative Filtering 14](#_Toc24981)

[2.2.1. Similarity functions 14](#_Toc6906)

[2.2.2. Rating prediction 15](#_Toc21616)

[2.3. Item-item Collaborative Filtering 15](#_Toc19273)

*[Chương 3:](#_Toc16513)* [Triển khai đề tài 15](#_Toc16513)

[3.1. Tiền xử lí 15](#_Toc18754)

[3.1.1. Code hoàn chỉnh 15](#_Toc21679)

*[3.1.1.1.](#_Toc25739)* [Đọc dữ liệu 15](#_Toc25739)

*[3.1.1.2.](#_Toc29951)* [Class CF 15](#_Toc29951)

*[3.1.1.3.](#_Toc16443)* [Cập nhật Utility matrix 16](#_Toc16443)

*[3.1.1.4.](#_Toc29354)* [Tính toán utility matrix và Similarity matrix 17](#_Toc29354)

*[3.1.1.5.](#_Toc16679)* [Chạy lại hàm khi có thêm dữ liệu 17](#_Toc16679)

*[3.1.1.6.](#_Toc32201)* [Dự đoán 18](#_Toc32201)

*[3.1.1.7.](#_Toc839)* [Tìm item gợi ý 19](#_Toc839)

*[3.1.1.8.](#_Toc26000)* [In kết quả 19](#_Toc26000)

*[3.1.1.9.](#_Toc14237)* [Kết quả user -user 19](#_Toc14237)

*[3.1.1.10.](#_Toc9004)* [Kết quả item-item 20](#_Toc9004)

[3.2. Huấn luyện mô hình 20](#_Toc5642)

[3.2.1. Chia tập dữ liệu 20](#_Toc19646)

[3.2.2. Đánh giá mô hình 20](#_Toc10983)

[Phần 3: kết luận 20](#_Toc14606)

*[1.](#_Toc18022)* [Đánh giá 20](#_Toc18022)

*[2.](#_Toc8676)* [Ưu điểm 20](#_Toc8676)

*[3.](#_Toc17417)* [Nhược điểm 20](#_Toc17417)

[Phần 4: Tài liệu tham khảo 20](#_Toc16943)

**DANH MỤC HÌNH ẢNH**

# Mở đầu

## Lời mở đầu

Hiện nay, các dịch vụ thương mại điện tử ngày càng phổ biến, cùng với đó là đa dạng trong sản phẩm, vì vậy sản phẩm, dịch vụ với số lượng cực lớn sẽ hạn chế thời gian tìm kiếm sản phẩm, dịch vụ thích hợp của khách hàng.

Vì vậy, hệ thống gợi ý xuất hiện nhằm đề xuất các sản phẩm được ưu chuộng nhất trên thị trường hiện nay, giúp giảm thời gian tìm kiếm của khách hàng cũng như tăng doanh thu cho doanh nghiệp.

## Mục đích, phạm vi nghiên cứu

Nhằm hỗ trợ người dùng tìm kiếm thông tin cần thiết, dự đoán sở thích người dùng và giới thiệu cho người dùng những sản phẩm, dịch vụ có thể lựa chọn mà họ chưa sử dụng.

Nội dung của báo cáo được trình bày theo 3 chương.

Tổ chức cấu trúc như sau:

**Chương 1** Tổng quan về học máy và hệ thống gợi ý: Chương này giới thiệu tổng quát về học máy và hệ thống gợi ý, bao gồm khái niệm, ứng dụng về học máy, các kỹ thuật.

* **Chương 2** Các thuật toán được sử dụng trong hệ thống gợi ý: chủ yếu về Collaborative filtering.

**Chương 3** Cài đặt, thử nghiệm và đánh giá thuật toán, bộ dữ liệu thử nghiệm và cài đặt chi tiết thuật toán đề cập ở chương 2. Phần cuối của chương trình bày kết quả thu được và đưa ra đánh giá về thuật toán..

## Phân chia công việc

|  |  |  |
| --- | --- | --- |
| **Nội dung công việc** | **Người thực hiện** | |
|  | **Lê Văn Quốc** | **Thạch Hoàng Duy** |
| Thu thập dữ liệu | **x** | **x** |
| Xử lí dữ liệu |  | **x** |
| Tìm hiểu nhóm thuật toán Collaborative filtering | **x** | **x** |
| Tìm hiểu Matrix Factorization | **x** |  |
| Tìm hiểu hàm mất mát | **x** | **x** |
| Triển khai giải thuật naive bayes | **x** | **x** |
| Thiết kế báo cáo | **x** |  |
| Kiểm tra, sửa lỗi báo cáo |  | **x** |

# Nội dung

## Cơ sở lý thuyết

### Một số khái niệm cơ bản

#### Định nghĩa hệ thống gợi ý

Hệ thống gợi ý là một lớp con của hệ thống lọc thông tin, tìm kiếm dự đoán “đánh giá” hoặc “ưu thích” của người dùng với một số sản phẩm hoặc đối tượng nào đó.

#### Các công năng của hệ thống gợi ý

Thực hiện tính toán của hệ thống gợi ý được xây dựng trên các thuật toán học máy, đưa ra các dự đoán tốt nhất về sản phẩm mà người dùng có thể thích.

Cải thiện trải nghiệm người dùng, tăng hiệu năng hoạt động bằng tự động hóa, biến khách hàng tiền năng trở thành khách hàng thật.

#### Các hướng tiếp cận

Các hệ thống gợi ý thường sử dụng nhiều thuật toán khác nhau, về cơ bản chúng ta có thể chia làm 2 nhóm lớn:

- Collaborative filtering

- Content-based filtering

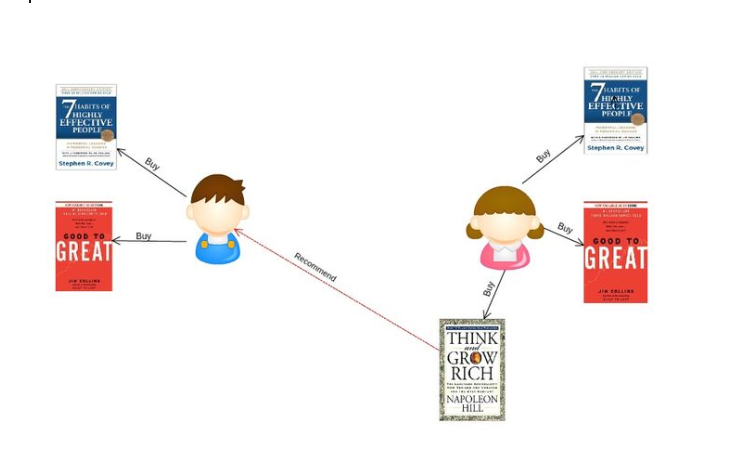
##### Collaborative filtering

Hoạt động dựa trên mô hình là những hành vi của người dùng trước đó, như: lịch sử giao dịch, để tìm quy luật tương tác giữa người dùng và các sản phẩm. Do đó những hệ thống gợi ý dựa trên cách tiếp cận này đều không quan tâm đến thuộc tính của sản phẩm, nó có khả năng khai thác thông tin ngoài phạm vi của các thuộc tính ản phẩm.

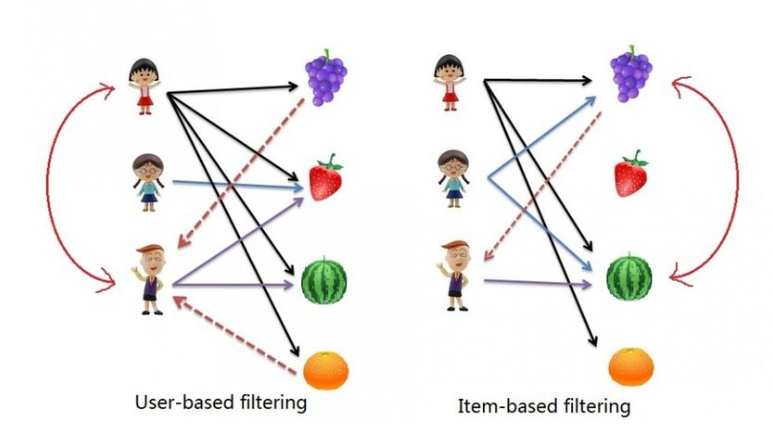
Mô hình huấn luyện có thể xây dựng dựa trên hành vi của một người dùng, hoặc nếu cần độ chính xác cao hơn thì có thể từ nhiều người dùng có cùng đặc điểm.

Khi làm việc với hành vi của người dùng khác, Collaborative filtering sử dụng kiến thức nhóm để tạo ra đề xuất dựa trene những người dùng tương tự.

Nó lọc trên nhũng người dùng có cùng sở thích, hay những người dùng có cùng những hành vi tương tự như: cùng bấm like, ch điểm đối với cùng một sản phẩm.



Phương pháp tiếp cận này có thể chia nhỏ ra thành 2 cách đó là: user-base và item-base



Hình 1

Hình 2: các phương pháp tiếp cận collaborative filtering

###### 1.1.3.1.1. User-base

Cách xây dựng:

- Biểu diễn mỗi người dùng bằng một vector các sản phẩm đã tương tác.

- Tính độ tương tự giữa các vector đại diện cho người dùng

- Đối với người dùng A, ước tính độ phù hợp của sản phẩm dựa trên lịch sử của của nhóm người dùng tương tự A

- Chọn k người dùng giống A nhất, hoặc chon tất cả người dùng nhũng thêm trọng số để ưu tiên nhũng người giống A hơn

Ví dụ: Có rất nhiều người thích “cam”, đoán xem tôi có thích “cam” không.

###### 1.1.3.1.2. Item-base

Cách xây dựng:

- Biểu diễn mỗi sản phẩm bằng một vector người dùng.

- Tính độ tương tự giữa các sản phẩm.

- Đối với người dùng A, tìm các sản phẩm tượng tự với các sản phẩm A đã tương tác.

- Chọn sản phẩm cho A từ các sản phẩm nói trên, bằng các tiêu chí như trọng số cao, nhiều người tương tác.

Ví dụ: Tôi đã ăn nho, dưa hấu, đoán xem tôi có thích cam không

##### Content-based filtering

Content-based filtering gợi ý dựa trên hành vi của người dùng. Nó phụ thuộc vào thuộc tính của các sản phẩm, đặc biệt rất mạnh đối với các sản phẩm giàu nội dung như các sản phẩm thuộc lịnh vực truyền thông, quảng cáo, y tế,...

##### Item cold start

Các bước chính:

- Biểu diễn mỗi sản phẩm dưới dạng một vector thuộc tính.

- Gợi ý các sản phẩm tương tự nhau.

- Xây dựng profile người dùng theo cá thuộc tính sản phẩm và giới thiệu sản phẩm có thuộc tính phù hợp với profile người dùng

##### Hybrid

Cách tiếp cận bằng cách kết hợp giữa nhiều phương pháp: item-base, user-base, Metrix factorization, cũng có thể bao gồm cả content-base filtering.

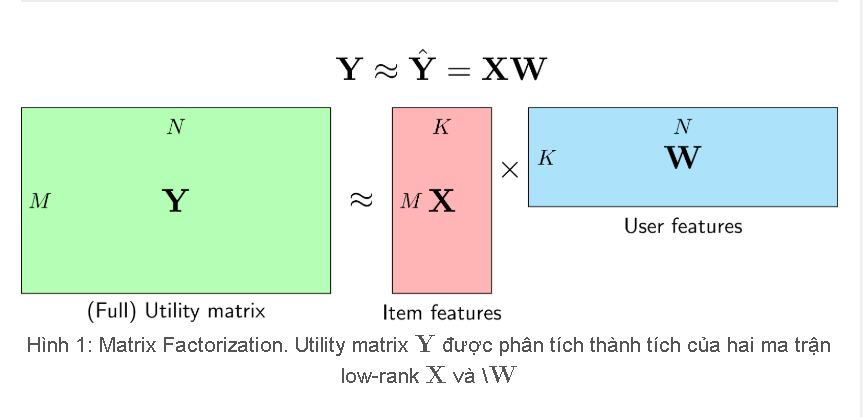
## Tổng quan thuật toán

### Matrix Factorization

#### Giới thiệu

- Ý tưởng của Matrix Factorization cho Recommendation Systems là tồn tại các latent features mô tả sự liên quan giữa các items và users. Ví dụ với hệ thống gợi ý các bộ phim, tính chất ẩn có thể là hình sự, chính trị, hành động,... Cũng có thể là sự kết hợp giữa các thể loại này.

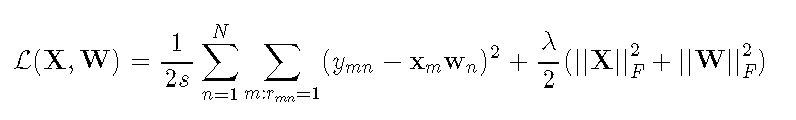
- Mỗi item sẽ tính chất ẩn ở một mức độ nào đó được mô tả tương ứng với hệ số. Mỗi user cũng thích một số tính chất ẩn nào đó. Hệ số cao tương ứng với việc user thích các bộ phim có tính chất ẩn đó.



Hình ảnh 1: Ma trận Y=XW

#### Hàm mất mát

Việc xây dựng hàm mất mát cũng được dựa trên các thành phần đã được điền của ma trận Y. Điểm khác là không có thành phần bias và biến tối ưu cả X và W.

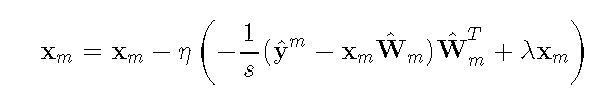
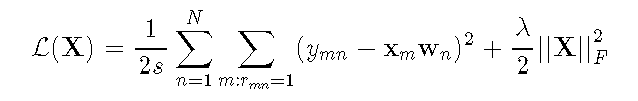


Trong đó rmn=1 nếu item thứ m đã được đánh giá bởi user thứ n, s là toàn bộ số ratings đã có. Thành phần thứ nhất là trung bình sai số của mô hình, thành phần thứ hai trong hàm mất mát là l2 regularization, giúp tránh overfitting

Giá trị ratings thường là cá giá trị đã được chuẩn hóa, bằng cách trừ mỗi hàng của Y đi trung bình cộng của các giá trị đã biết của hàng đó(item-based) hoặc trừ đi trung bình cộng của các giá trị đã biết trong cột đó(user-based).

##### Tối ưu hàm mất mát

Khi cố định W, việc tối ưu được đưa về tối ưu hàm:



### User-user Collaborative Filtering

#### Similarity functions

Công việc quan trọng nhất phải làm trước tiên trong thuật toán này là phải xác định sự giống nhau giữa hau users. Sự giống nhau này phải được xác định dựa trên các cột tương ứng với hai users trong ma trận.

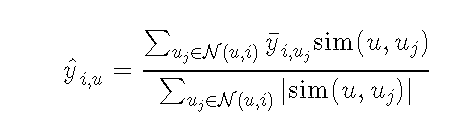
#### Rating prediction

Cần xác định mức độ quan tâm của một user lên một item dựa trên các user gần nhất này.

- Missing ratings được xác định dựa trên thông tin về neighbor user.

- predicted rating được xác định là trung bình cộng có trộng số của các ratings đã được chuẩn hóa.

Công thức được sử dụng để dự đoán rating của user u cho item i là:



Trong đó: N(u,i) là tập hợp k users trong neighborhood của u mà đã rated i.

### Item-item Collaborative Filtering

Nếu ta tính toán similarity giữa các items rồi recommend những items gần giống với item yêu thích của một user thì sẽ có những lợi iichs sau:

- thuận lợi cho việc lưu trữ và tính toán ở các bước sau, do items thường nhỏ hơn user.

- mỗi hàng của ma trận sẽ có nhiều phần tử đã biết hơn số phần tử đã biết trong cột vì số items(hàng) ít hơn số users(cột).

Về mặt tính toán, item-item CF có thể nhân được từ User-user CF bằng cách chuyển vị ma trận và coi như items đang rate user. Sau khi tính ra kết quả cuối cùng, ta lại chuyển vị một lần nữa để thu được kết quả.

## Triển khai đề tài

### Tiền xử lí

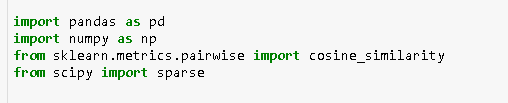
Vì hiện nay trên trên internet có rất phong phú thông tin của sản phẩm vì vậy ta có thể downloads trực tiếp từ internet.

Sau đó ta đọc dữ liệu từ file bằng thư viện pandas.

#### Code hoàn chỉnh

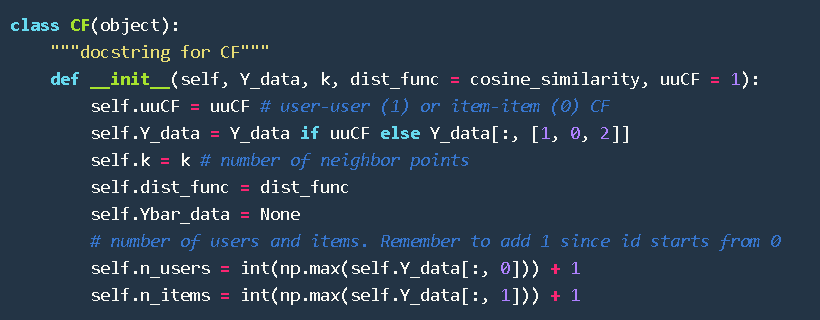
##### Đọc dữ liệu

Import các thư viện cần dùng



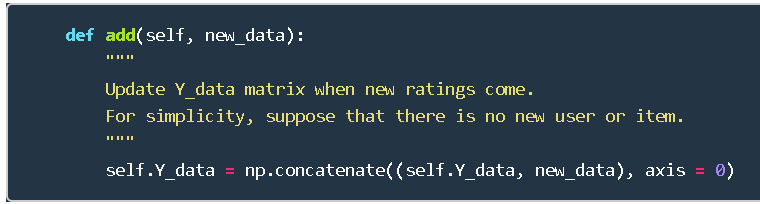
Hình ảnh 2: Đọc file

##### Class CF



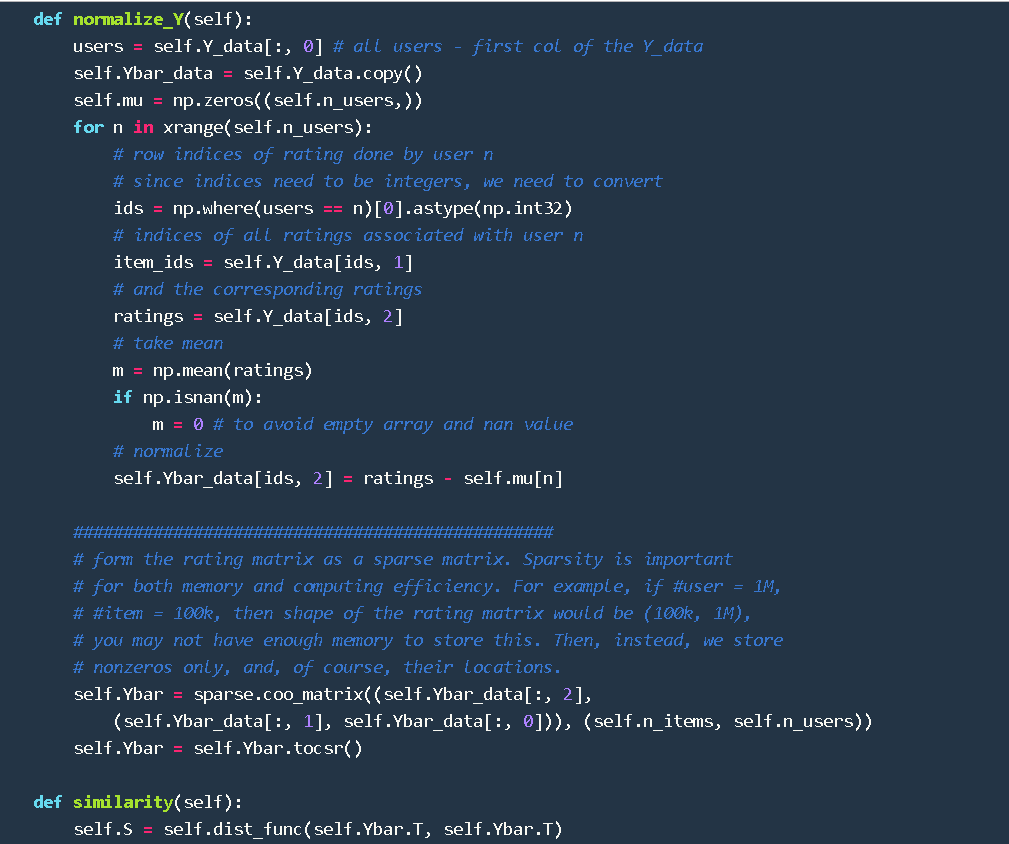
Hình ảnh 3: Class CF

##### Cập nhật Utility matrix



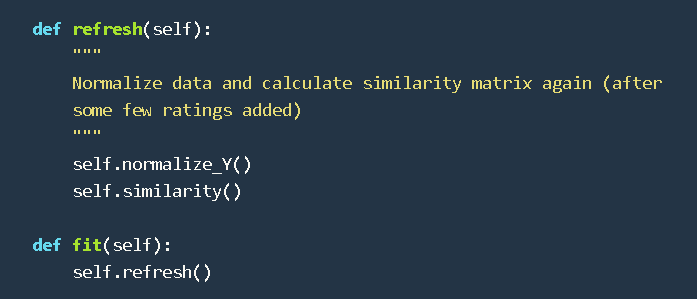
Hình ảnh 4: Cập nhật Utility matrix

##### Tính toán utility matrix và Similarity matrix



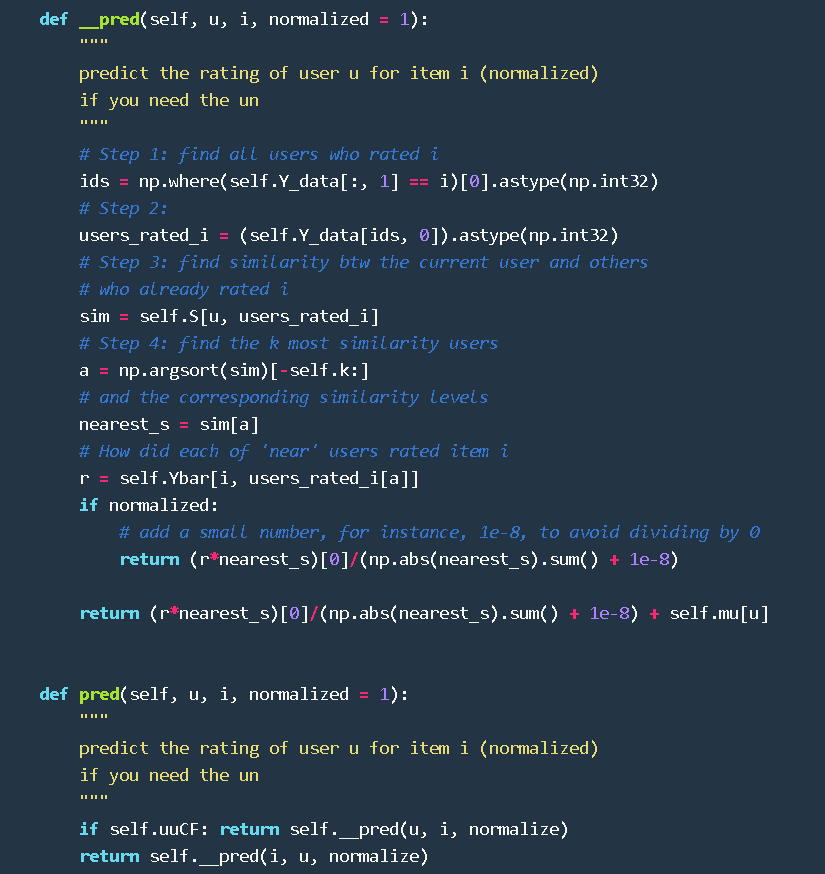
Hình ảnh 5: Tính toán utility matrix và Similarity matrix

##### Chạy lại hàm khi có thêm dữ liệu



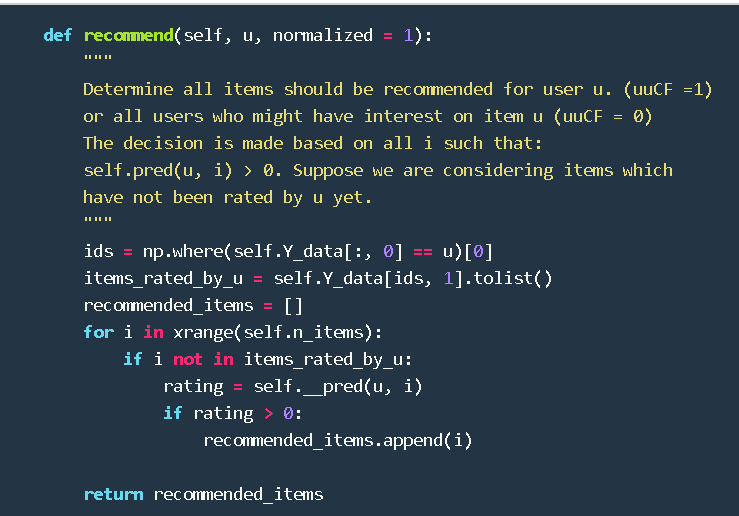
Hình ảnh 6: Chạy lại hàm khi có thêm dữ liệu

##### Dự đoán



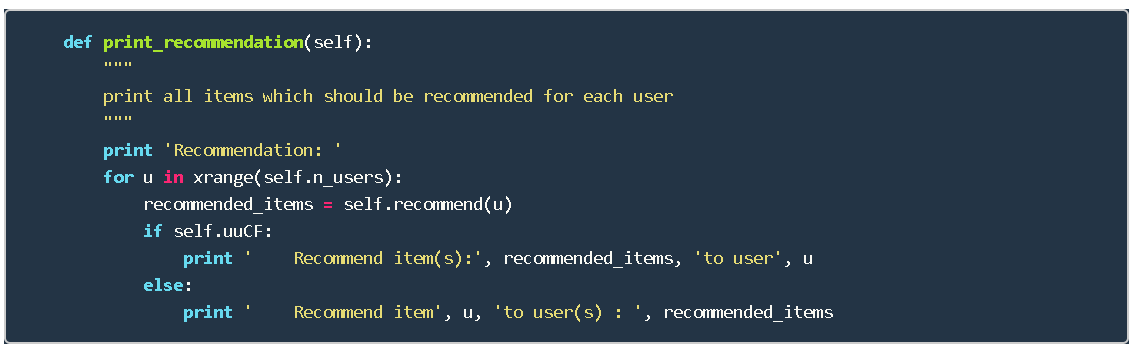
Hình ảnh 7: Dự đoán

##### Tìm item gợi ý



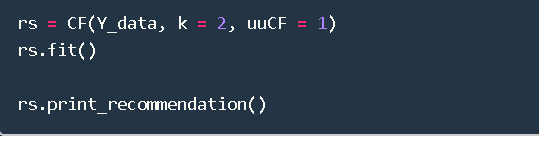
Hình ảnh 8: Tìm item gợi ý

##### In kết quả



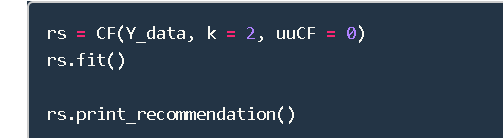
Hình ảnh 9: In kết quả

##### Kết quả user -user



Hình ảnh 10: Kết quả user-user

##### Kết quả item-item



Hình ảnh 11: Kết quả item-item

### Huấn luyện mô hình

#### Chia tập dữ liệu

Dữ liệu được chia thành 2 phần là train và test

Train: 80%

Test: 20%

#### Đánh giá mô hình

Dùng giải thuật để đánh giá mô hình có đạt yêu cầu hay không.

# kết luận

1. Đánh giá

Collaborative Filtering là một phương pháp gợi ý sản phẩm với ý tưởng chính dựa trên các hành vi của các *users* khác (collaborative) cùng trên một *item* để suy ra mức độ quan tâm (filtering) của một *user* lên sản phẩm. Việc suy ra này được thực hiện dựa trên Similarity matrix đo độ giống nhau giữa các *users*.

Để tính được Similarity matrix, trước tiên ta cần chuẩn hoá dữ liệu. Phương pháp phổ biến là *mean offset*, tức trừ các *ratings* đi giá trị trung bình mà một *user* đưa ra cho các \_items.

1. Ưu điểm

Giảm thời lượng tìm kiếm thông tin của khách hàng.

Tăng hiệu suất công việc,

Giảm chi phí quảng cáo sản phẩm.

1. Nhược điểm

Hạn chế khi lượng dữ liệu lớn

# Tài liệu tham khảo

<https://viblo.asia/p/matrix-factorization-phuong-phap-goi-y-dua-tren-ky-thuat-phan-ra-ma-tran-p1-Az45bapNlxY>

<https://machinelearningcoban.com/2017/05/31/matrixfactorization/>

[https://datajobs.com/data-science-repo/Recommender-Systems-%5BNetflix%5D.pdf](https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf)

<https://viblo.asia/p/cong-nghe-matrix-factorization-cho-he-thong-goi-y-naQZRJe0Zvx>

<https://machinelearningcoban.com/2017/05/24/collaborativefiltering/#-user-user-collaborative-filtering>