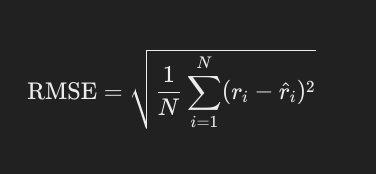
Tổng hợp những gì đã tìm hiểu được

Các cách evaluating một recommender system:  
  
Có nhiều loại tiêu chí

### + Dựa vào tiêu chí độ chính xác của một rating mà recommend system gợi ý: **1. RMSE (Root Mean Square Error - Sai số căn bậc hai của trung bình)**

* **Ý nghĩa**: RMSE đo lường độ chính xác của các dự đoán mà mô hình đưa ra. Nó tính toán sai số giữa đánh giá thực tế và đánh giá dự đoán từ mô hình.
* **Giải thích đơn giản**: Hãy tưởng tượng bạn đang đoán điểm số của bạn bè về các bộ phim. RMSE cho bạn biết mức độ sai lệch trung bình giữa điểm số bạn dự đoán và điểm số thật. Càng thấp, mô hình càng chính xác.
* **Kết quả**: 0.90 có nghĩa là trung bình, độ sai lệch giữa điểm số dự đoán và điểm thực tế của mô hình là 0.90 điểm.

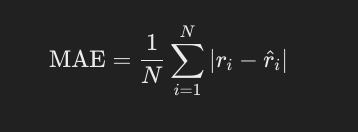


Trong đó:

* ri: Điểm đánh giá thực tế của người dùng đối với bộ phim thứ iii.
* r^i: Điểm đánh giá dự đoán của mô hình đối với bộ phim thứ iii.
* N: Tổng số bộ phim.

### **2. MAE (Mean Absolute Error - Sai số tuyệt đối trung bình)**

* **Ý nghĩa**: MAE cũng đo lường sai số giữa điểm số thực tế và dự đoán, nhưng thay vì bình phương sai số như RMSE, nó tính tổng các sai số tuyệt đối (không phân biệt dấu âm hay dương) và lấy trung bình.
* **Giải thích đơn giản**: MAE giống như việc bạn đoán điểm số nhưng lần này chỉ nhìn vào mức độ sai lệch tuyệt đối mà không quan tâm nó lớn hay nhỏ hơn. Cũng giống như RMSE, kết quả càng thấp càng tốt.
* **Kết quả**: MAE = 0.70 có nghĩa là trung bình, mỗi dự đoán của mô hình sai lệch khoảng 0.70 điểm so với điểm thật.

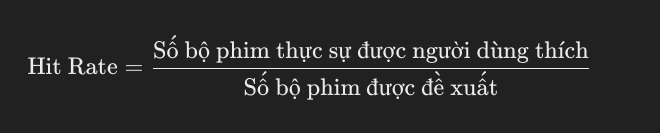


Trong đó:

* ri​: Điểm đánh giá thực tế của người dùng đối với bộ phim thứ iii.
* r^i​: Điểm đánh giá dự đoán của mô hình đối với bộ phim thứ iii.
* N: Tổng số bộ phim.

### **3. Hit Rate (Tỷ lệ phim đề xuất được người dùng yêu thích)**

* **Ý nghĩa**: Đây là tỷ lệ phần trăm của các bộ phim trong danh sách đề xuất mà người dùng thực sự thích (đánh giá cao, ví dụ 4 sao trở lên).
* **Giải thích đơn giản**: Giống như việc bạn có một danh sách các bộ phim được gợi ý và bạn xem thử bao nhiêu bộ phim trong danh sách đó mà bạn thực sự thích. Nếu tỷ lệ này là 0.03, tức là trong 100 bộ phim được đề xuất, chỉ có 3 bộ phim là bạn sẽ thích.
* **Kết quả**: Hit Rate = 0.03 có nghĩa là chỉ có 3% các bộ phim trong danh sách đề xuất là thực sự được người dùng yêu thích.

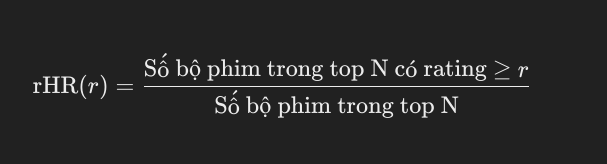


Trong đó:

* Số bộ phim thực sự được người dùng thích: Các bộ phim trong danh sách đề xuất có đánh giá thực tế cao (ví dụ, ≥ 4 sao).
* Số bộ phim được đề xuất: Tổng số bộ phim trong danh sách đề xuất.

### **4. rHR (Hit Rate theo từng giá trị rating)**

* **Ý nghĩa**: Đây là tỷ lệ các phim trong danh sách đề xuất mà người dùng đánh giá từ một giá trị rating nhất định trở lên. Chẳng hạn, nếu người dùng đánh giá phim 4.0 sao hoặc cao hơn, mô hình sẽ kiểm tra xem trong các phim đề xuất, tỷ lệ phim đạt được mức đánh giá này là bao nhiêu.
* **Giải thích đơn giản**: Nếu bạn chỉ quan tâm đến các bộ phim có điểm số cao (chẳng hạn như 4 sao trở lên), rHR cho bạn biết có bao nhiêu bộ phim trong danh sách đề xuất đáp ứng điều đó.
* **Kết quả**: Với rHR = 0.07 cho rating 5.0 có nghĩa là 7% các bộ phim đề xuất có rating 5.0 (tức là những bộ phim rất được yêu thích).

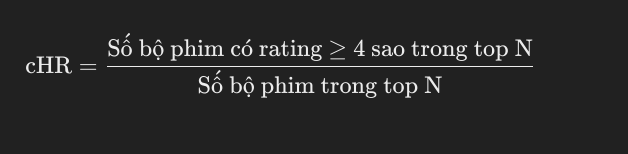


Trong đó:

* r: Mức đánh giá tối thiểu (ví dụ: 4 sao, 5 sao).
* Số bộ phim trong top N có rating ≥ rrr: Các bộ phim trong danh sách đề xuất có điểm đánh giá thực tế ≥ rrr.
* Số bộ phim trong top N: Tổng số bộ phim trong danh sách đề xuất.

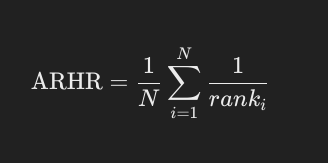
### **5. cHR (Cumulative Hit Rate - Tỷ lệ tích lũy các phim được đánh giá cao)**

* **Ý nghĩa**: Đây là tỷ lệ các phim trong danh sách đề xuất mà người dùng đánh giá từ 4 sao trở lên (phim chất lượng cao).
* **Giải thích đơn giản**: cHR cho bạn biết trong tất cả các phim được đề xuất, tỷ lệ phim chất lượng (4 sao trở lên) là bao nhiêu.
* **Kết quả**: cHR = 0.05 có nghĩa là 5% của tất cả các bộ phim được đề xuất có đánh giá từ 4 sao trở lên.



* Số bộ phim có rating ≥ 4 sao trong top N: Các bộ phim trong danh sách đề xuất có điểm đánh giá thực tế ≥ 4 sao.
* Số bộ phim trong top N: Tổng số bộ phim trong danh sách đề xuất.

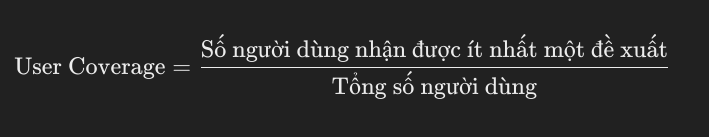
### **6. ARHR (Average Reciprocal Hit Rank - Trung bình vị trí nghịch đảo của điểm đánh giá)**



Trong đó:

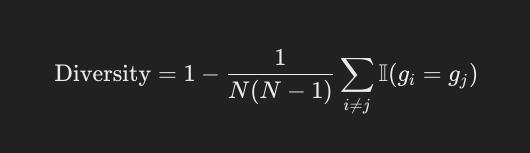
* rankii​: Vị trí của bộ phim yêu thích của người dùng trong danh sách đề xuất (1 là vị trí đầu tiên, 2 là vị trí thứ hai, v.v.).
* N: Tổng số bộ phim yêu thích trong danh sách đề xuất.
* **Ý nghĩa**: ARHR tính toán trung bình của các vị trí nghịch đảo trong danh sách đề xuất. Vị trí nghịch đảo càng nhỏ (phim ở đầu danh sách đề xuất) thì giá trị ARHR càng cao.
* **Giải thích đơn giản**: Nếu một bộ phim bạn thích nằm ở vị trí đầu tiên trong danh sách đề xuất, ARHR sẽ đánh giá cao điều này. Nó cho bạn biết phim yêu thích của người dùng có ở vị trí đầu hay không.
* **Kết quả**: ARHR = 0.01 có nghĩa là các bộ phim yêu thích của người dùng thường ở vị trí rất thấp trong danh sách đề xuất.

### **7. User Coverage (Mức độ bao phủ người dùng)**



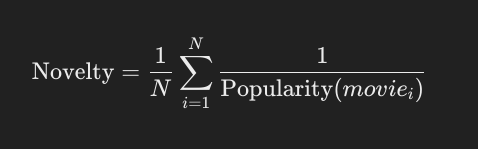
* **Ý nghĩa**: Đây là tỷ lệ người dùng mà hệ thống có thể đưa ra ít nhất một đề xuất. Nếu tỷ lệ này cao, điều đó có nghĩa là hệ thống có thể đưa ra đề xuất cho hầu hết người dùng.
* **Giải thích đơn giản**: Hệ thống có thể đề xuất phim cho bao nhiêu người dùng? Nếu User Coverage là 0.95, có nghĩa là 95% người dùng sẽ nhận được ít nhất một bộ phim trong danh sách đề xuất.
* **Kết quả**: User Coverage = 0.96 có nghĩa là hệ thống có thể đưa ra đề xuất cho 96% người dùng.

### **8. Diversity (Độ đa dạng của danh sách đề xuất)**



* N: Tổng số bộ phim trong danh sách đề xuất.
* g​: Thể loại của bộ phim thứ iii và bộ phim thứ jjj.
* II(gi = gj): Hàm chỉ thị, trả về 1 nếu thể loại của bộ phim iii và bộ phim jjj là giống nhau, ngược lại trả về 0.
* **Ý nghĩa**: Đây là một chỉ số đo lường sự đa dạng của các bộ phim trong danh sách đề xuất. Nếu chỉ có một thể loại phim được đề xuất, diversity sẽ thấp. Nếu nhiều thể loại, diversity sẽ cao.
* **Giải thích đơn giản**: Hệ thống không chỉ đưa ra các bộ phim giống nhau mà sẽ cố gắng cung cấp sự đa dạng (thể loại khác nhau, phim nổi tiếng và ít nổi tiếng).
* **Kết quả**: Diversity = 0.97 có nghĩa là danh sách đề xuất có sự đa dạng rất cao, bao gồm nhiều thể loại phim khác nhau.

### **9. Novelty (Mức độ mới mẻ của các bộ phim được đề xuất)**

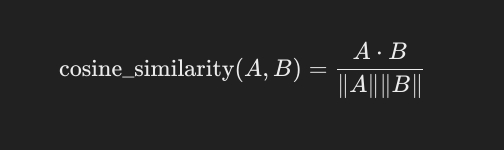


* N: Tổng số bộ phim trong danh sách đề xuất.
* Popularity(moviei): Vị trí phổ biến của bộ phim iii trong danh sách toàn cầu (ví dụ: xếp hạng bộ phim trong một cơ sở dữ liệu phim lớn).
* **Ý nghĩa**: Novelty đo lường mức độ phổ biến của các bộ phim trong danh sách đề xuất. Mức độ mới mẻ cao có nghĩa là hệ thống đề xuất các bộ phim ít phổ biến, giúp người dùng khám phá điều mới mẻ.
* **Giải thích đơn giản**: Nếu bạn đã xem quá nhiều phim nổi tiếng, hệ thống sẽ đưa ra các bộ phim ít người biết đến nhưng có thể bạn sẽ thích.
* **Kết quả**: Novelty = 491 có nghĩa là các bộ phim trong danh sách đề xuất có xu hướng ít nổi tiếng, với vị trí phổ biến trung bình là 491 trong danh sách xếp hạng phim toàn cầu.

Các phương pháp tính toán simularity giữa 2 thứ:

### **1. Cosine Similarity (Tương đồng cosine)**

**Ý nghĩa**: Cosine Similarity đo lường góc giữa hai vector trong không gian nhiều chiều. Độ tương đồng cosine càng lớn, nghĩa là hai đối tượng (ví dụ, người dùng hoặc bộ phim) càng giống nhau.

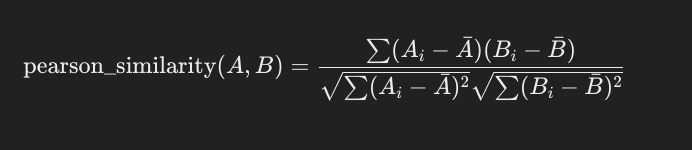


* A.B là tích vô hướng của hai vector A và B.
* ||A|| và ||B|| là độ dài (norm) của vector A và B.

Ngoài ra còn có:

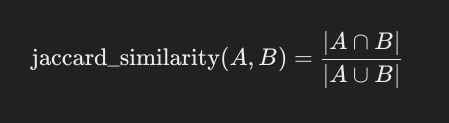
### **2. Pearson Correlation (Hệ số tương quan Pearson)**

**Ý nghĩa**: Pearson Correlation đo lường mức độ tương quan tuyến tính giữa hai vector. Đây là phương pháp phổ biến trong các hệ thống gợi ý dựa trên mối quan hệ giữa các người dùng hoặc các item.



### **3. Jaccard Similarity (Tương đồng Jaccard)**

**Ý nghĩa**: Jaccard Similarity đo lường sự tương đồng giữa hai tập hợp bằng cách tính tỉ lệ phần tử chung so với tổng số phần tử trong cả hai tập hợp.



......

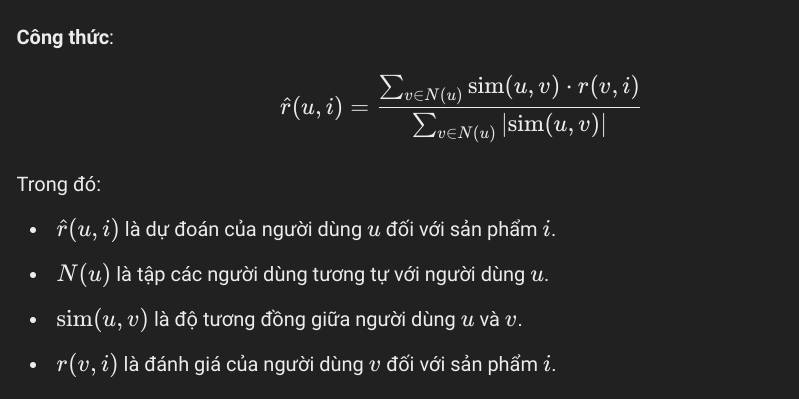
Các thuật toán recommend system:

### **1. Collaborative Filtering (Lọc cộng tác)**

**Collaborative Filtering (CF)** là phương pháp phổ biến nhất trong hệ thống gợi ý, dựa trên hành vi của người dùng và những tương tác trước đó với các sản phẩm. Collaborative Filtering không yêu cầu thông tin về sản phẩm mà chỉ dựa vào các sự kiện lịch sử (như đánh giá, lượt xem, hoặc mua).

#### **a. User-based Collaborative Filtering**

**Ý nghĩa**: Thuật toán này gợi ý sản phẩm cho người dùng dựa trên sự tương đồng giữa người dùng hiện tại và những người dùng khác. Cơ sở của thuật toán là giả định rằng nếu hai người dùng có hành vi tương tự trong quá khứ, họ sẽ thích những sản phẩm giống nhau trong tương lai.

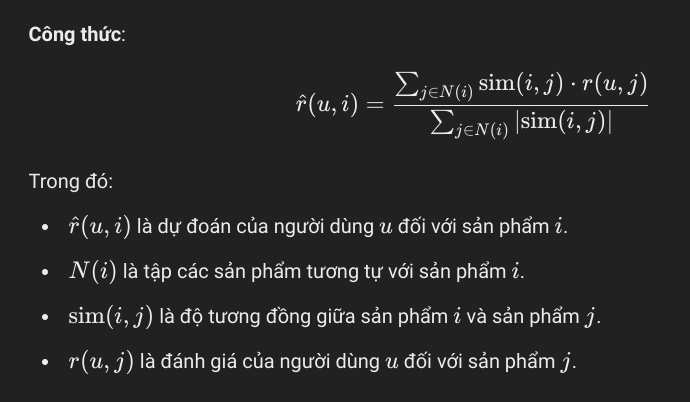


Giải thích:

* Các người dùng có hành vi giống nhau sẽ nhận được các đề xuất giống nhau. Ví dụ, nếu người dùng A và B cùng thích một số bộ phim, thì những bộ phim người dùng A thích nhưng người dùng B chưa xem sẽ được gợi ý cho B.

#### **b. Item-based Collaborative Filtering**

**Ý nghĩa**: Thay vì tìm kiếm người dùng tương tự, thuật toán này tìm kiếm các sản phẩm tương tự nhau. Khi người dùng thích một sản phẩm nào đó, hệ thống sẽ gợi ý các sản phẩm tương tự dựa trên các hành vi chung của những người dùng khác.



Giải thích:

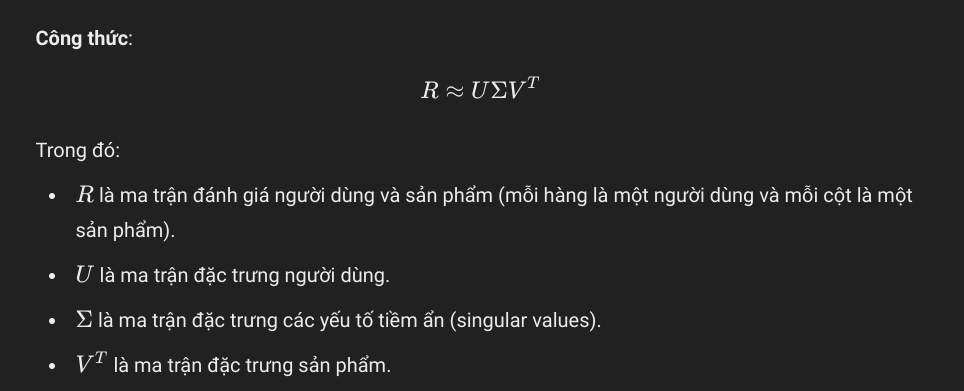
* Ví dụ, nếu người dùng thích bộ phim "Avengers", hệ thống sẽ gợi ý các bộ phim tương tự, như "Iron Man" hay "Thor".

### **2. Matrix Factorization (Phân tách ma trận)**

**Matrix Factorization** là một phương pháp mạnh mẽ để giảm chiều của ma trận đánh giá, giúp tìm ra các yếu tố ẩn (latent factors) mô tả sở thích của người dùng và các đặc tính của sản phẩm. Phương pháp này thường được sử dụng trong **Singular Value Decomposition (SVD)** và các biến thể của nó.

#### **a. Singular Value Decomposition (SVD)**

**Ý nghĩa**: SVD là một kỹ thuật phân tách ma trận để tìm ra các yếu tố tiềm ẩn mô tả người dùng và sản phẩm. Phương pháp này giúp giảm số chiều trong ma trận đánh giá, từ đó tìm kiếm các mối quan hệ ẩn giữa người dùng và sản phẩm.



**Giải thích**:

* Phương pháp này giúp giảm chiều của ma trận đánh giá lớn và tìm ra các yếu tố tiềm ẩn mà người dùng và sản phẩm có thể chia sẻ (ví dụ, thể loại phim, sở thích chung).

#### **b. Alternating Least Squares (ALS)**

**Ý nghĩa**: ALS là một biến thể của Matrix Factorization, sử dụng phương pháp tối ưu hóa để phân tách ma trận đánh giá thành các ma trận người dùng và sản phẩm. ALS thường được sử dụng trong các hệ thống gợi ý lớn.

**Giải thích**:

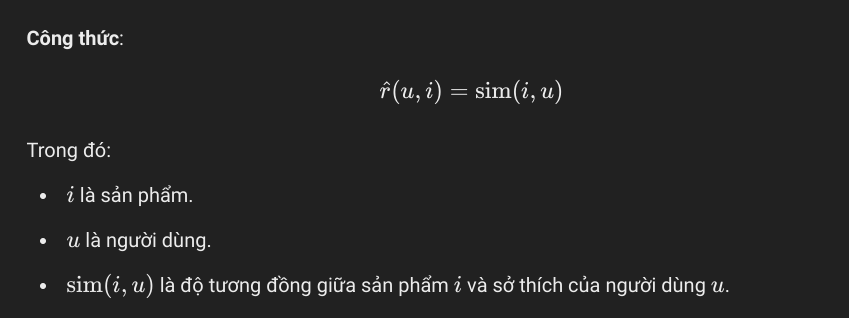
* ALS tối ưu hóa các ma trận bằng cách luân phiên tính toán các giá trị người dùng và sản phẩm, sử dụng một thuật toán giảm dần bình phương lỗi.

### **3. Content-Based Filtering (Lọc dựa trên nội dung)**

**Content-Based Filtering** là phương pháp gợi ý dựa trên các đặc tính của sản phẩm. Khi người dùng đã thích một sản phẩm, hệ thống sẽ gợi ý những sản phẩm có đặc tính tương tự.

#### **Ý nghĩa:**

* Thay vì dựa vào hành vi của người dùng khác, thuật toán này sử dụng thông tin mô tả về sản phẩm để tìm ra những sản phẩm tương tự với những gì người dùng đã yêu thích.



**Giải thích**:

* Ví dụ, nếu người dùng thích các bộ phim hành động, hệ thống sẽ gợi ý các bộ phim khác thuộc thể loại hành động.

### **4. Hybrid Methods (Phương pháp kết hợp)**

**Hybrid Methods** kết hợp các phương pháp khác nhau (Collaborative Filtering và Content-Based Filtering) để tận dụng ưu điểm của từng phương pháp. Điều này giúp cải thiện độ chính xác của các đề xuất và giảm thiểu các hạn chế của từng phương pháp riêng biệt.

#### **Ví dụ:**

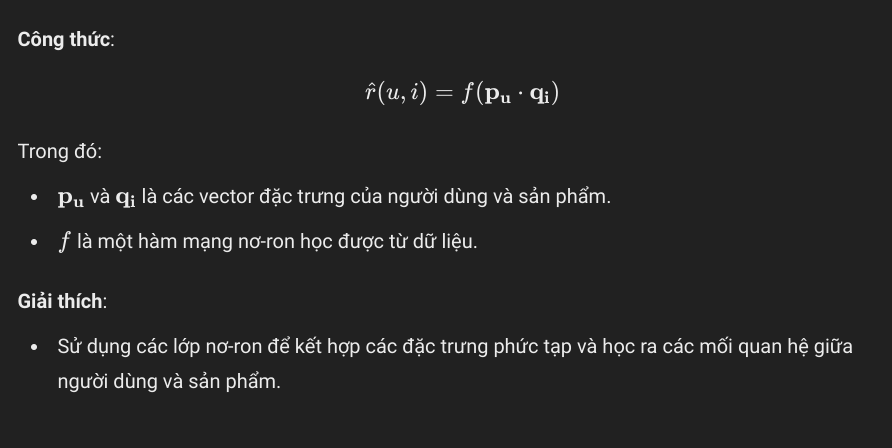
* **Weighted Hybrid**: Kết hợp các phương pháp khác nhau bằng cách gán trọng số cho mỗi phương pháp. Ví dụ, 70% độ chính xác có thể được tính từ Collaborative Filtering và 30% từ Content-Based Filtering.
* **Switching Hybrid**: Chọn phương pháp phù hợp nhất cho từng người dùng hoặc từng tình huống cụ thể.

### **5. Deep Learning-Based Approaches (Phương pháp dựa trên học sâu)**

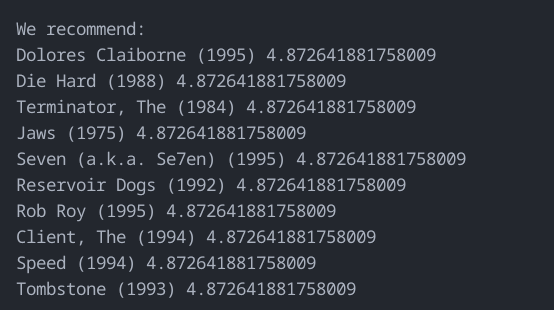
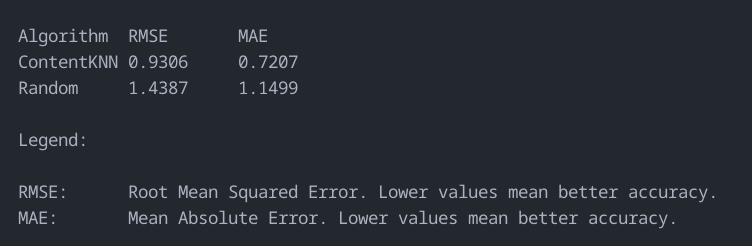
**Deep Learning** trong hệ thống gợi ý áp dụng các mô hình học sâu như **Neural Collaborative Filtering (NCF)** để học các biểu diễn phức tạp của người dùng và sản phẩm từ dữ liệu lớn.

#### **a. Neural Collaborative Filtering (NCF)**

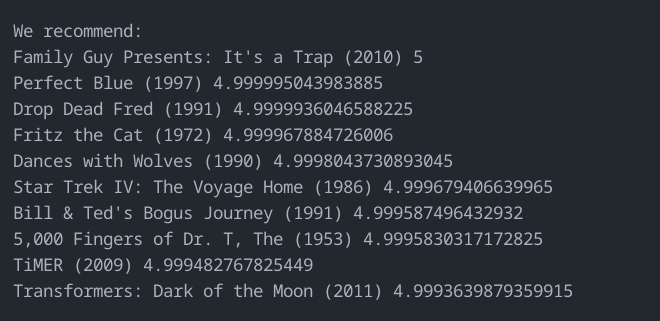
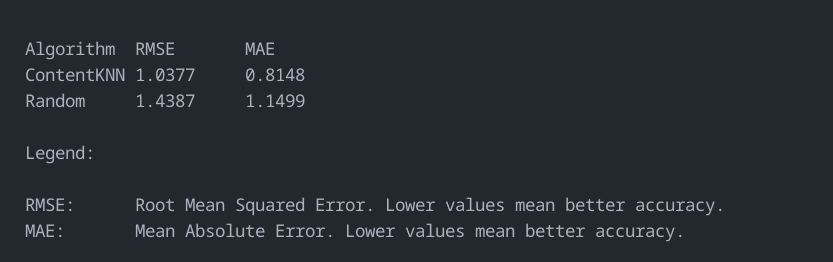
**Ý nghĩa**: NCF sử dụng mạng nơ-ron để học các mối quan hệ phức tạp giữa người dùng và sản phẩm. Các đặc tính của người dùng và sản phẩm được đưa vào mạng nơ-ron để tạo ra các dự đoán chính xác hơn.



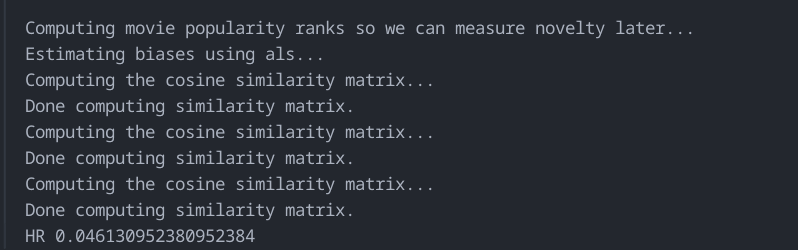
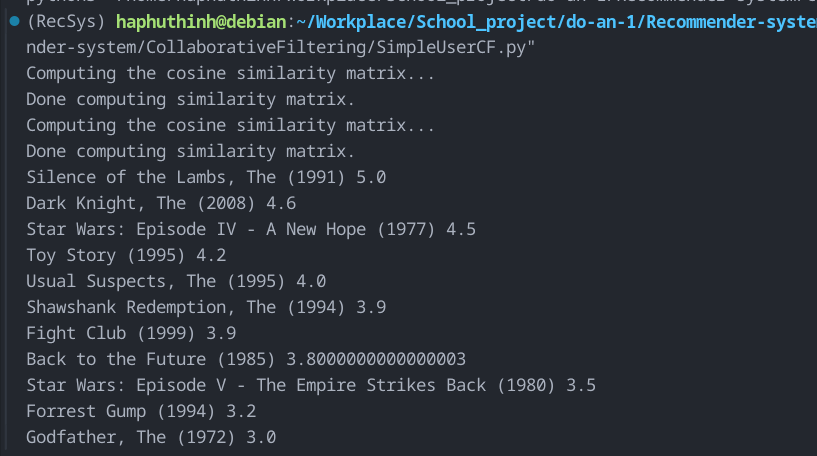
Test thuật toán:  
  
Khi chưa dùng visual feature, mà chỉ kết hợp giữa genre và date release:



Sau khi áo dụng visual base:



Kết quả khi áp dụng Collaborative Filtering (User based):



Kết quả khi áp dụng Collaborative Filtering (Item based):

