Suprise  
 Surprise là một thư viện Python scikit dành cho việc xây dựng và phân tích các hệ thống gợi ý xử lý dữ liệu đánh giá rõ ràng.

Surprise được thiết kế với các mục đích sau:

1. Cung cấp cho người dùng sự kiểm soát hoàn hảo đối với các thí nghiệm của họ. Vì vậy, chúng tôi đặt một trọng tâm mạnh mẽ vào tài liệu, mà chúng tôi đã cố gắng làm cho nó càng rõ ràng và chính xác nhất có thể bằng cách chỉ ra mọi chi tiết của các thuật toán.

2. Giảm bớt nỗi đau trong việc xử lý Dữ liệu. Người dùng có thể sử dụng cả tập dữ liệu tích hợp sẵn (Movielens, Jester), và các tập dữ liệu tùy chỉnh của riêng họ.

3. Cung cấp các thuật toán dự đoán sẵn sàng sử dụng như các thuật toán cơ bản, phương pháp kề hàng, dựa trên phân rã ma trận (SVD, PMF, SVD++, NMF), và nhiều thuật toán khác. Ngoài ra, các phép đo tương tự khác nhau (cosine, MSD, pearson…) cũng được tích hợp sẵn.

4. Dễ dàng thực hiện các ý tưởng thuật toán mới.

5. Cung cấp các công cụ để đánh giá, phân tích và so sánh hiệu suất của các thuật toán. Các thủ tục chia dữ liệu kiểm tra chéo có thể được thực hiện một cách dễ dàng bằng cách sử dụng các bộ lặp CV mạnh mẽ (được lấy cảm hứng từ các công cụ xuất sắc của scikit-learn), cũng như tìm kiếm cận kề trên một tập hợp các tham số.

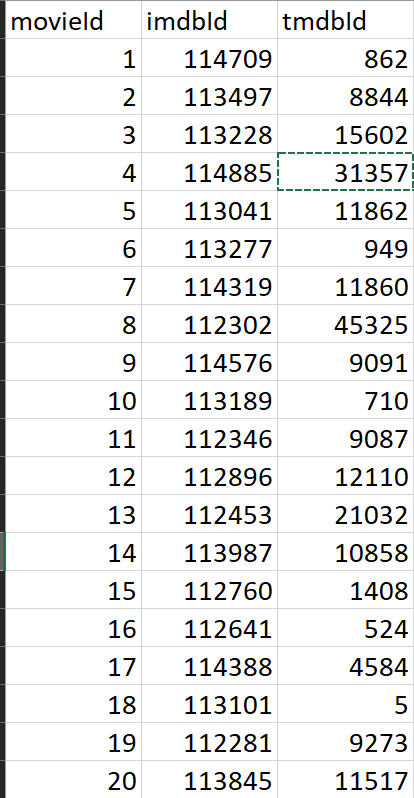
Dữ liệu đầu vào:

Bộ dữ liệu mô tả hoạt động đánh giá 5 sao và gắn thẻ văn bản miễn phí từ [MovieLens](http://movielens.org), một dịch vụ gợi ý phim. Nó bao gồm 100.004 đánh giá và 1.296 ứng dụng thẻ trên 9.125 bộ phim. Các dữ liệu này được tạo ra bởi 671 người dùng trong khoảng thời gian từ ngày 09 tháng 01 năm 1995 đến ngày 16 tháng 10 năm 2016. Bộ dữ liệu này được tạo ra vào ngày 17 tháng 10 năm 2016.

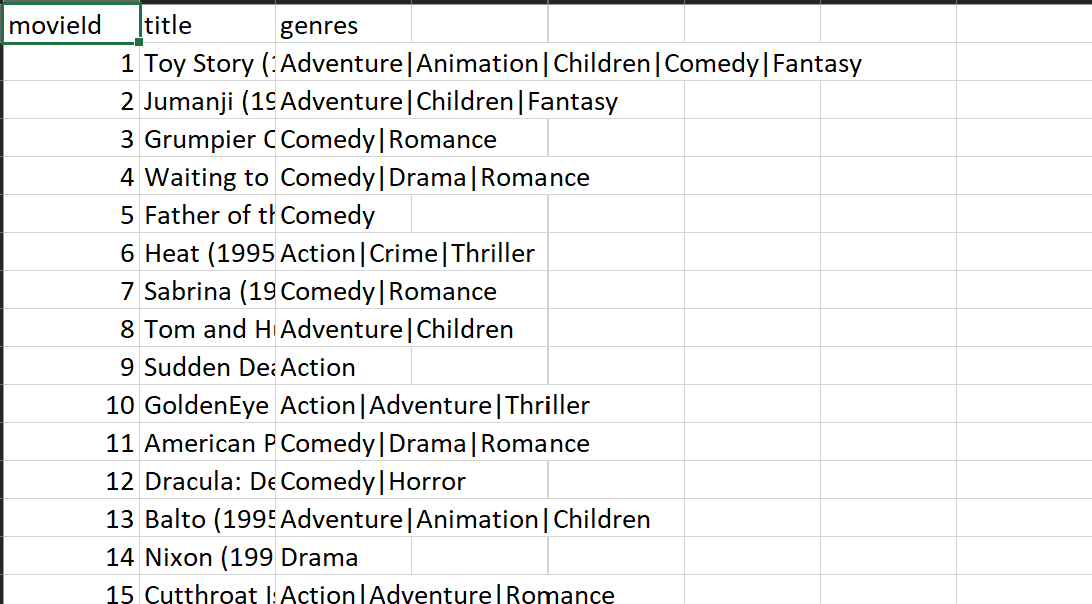
Người dùng được lựa chọn ngẫu nhiên để bao gồm vào bộ dữ liệu. Tất cả người dùng được chọn lựa có đánh giá ít nhất 20 bộ phim. Không có thông tin dân số được bao gồm. Mỗi người dùng được đại diện bằng một id, và không có thông tin nào khác được cung cấp.

Các dữ liệu được chứa trong các tệp `links.csv`, `movies.csv`, `ratings.csv` và `tags.csv`. Thêm thông tin về nội dung và cách sử dụng của tất cả các tệp này được liệt kê sau.

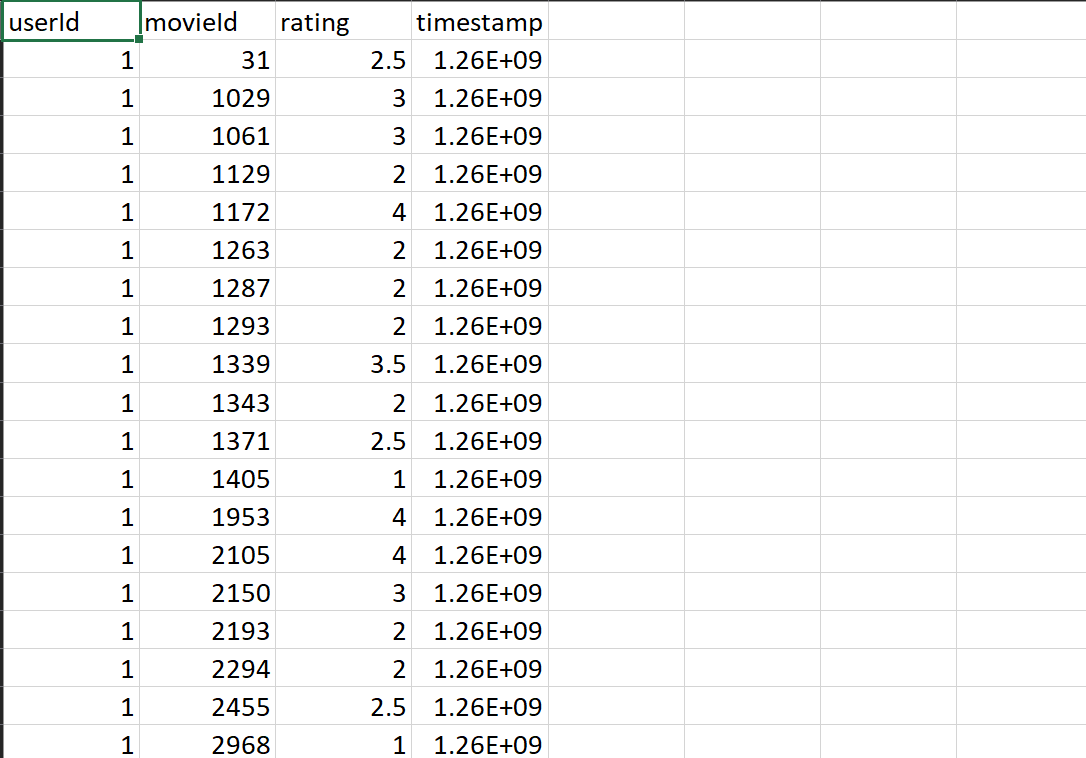
- links.csv: liên kết dữ liệu với nội dung thực trên Internet



- movies.csv: tên + thể loại phim



- rating.csv: đánh giá số của người dùng cho phim



-tags.csv: từ khóa



Phương pháp gợi ý:

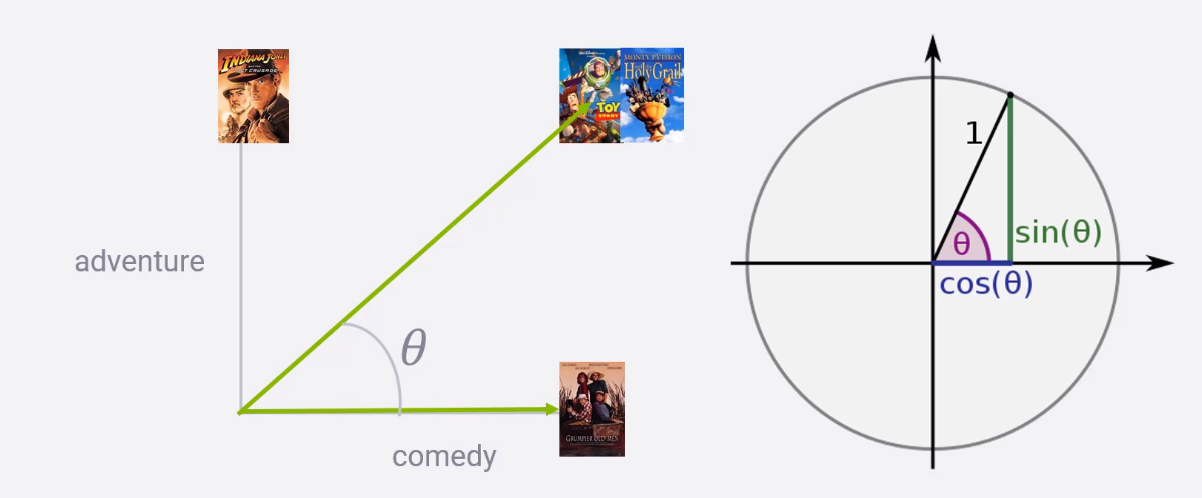
a. Content-based filtering:

- Biểu diễn nội dung của các bộ phim: Mỗi phim có 18 thể loại (genres) có thể thuộc vào, mỗi thể loại sẽ là 1 chiều của vector

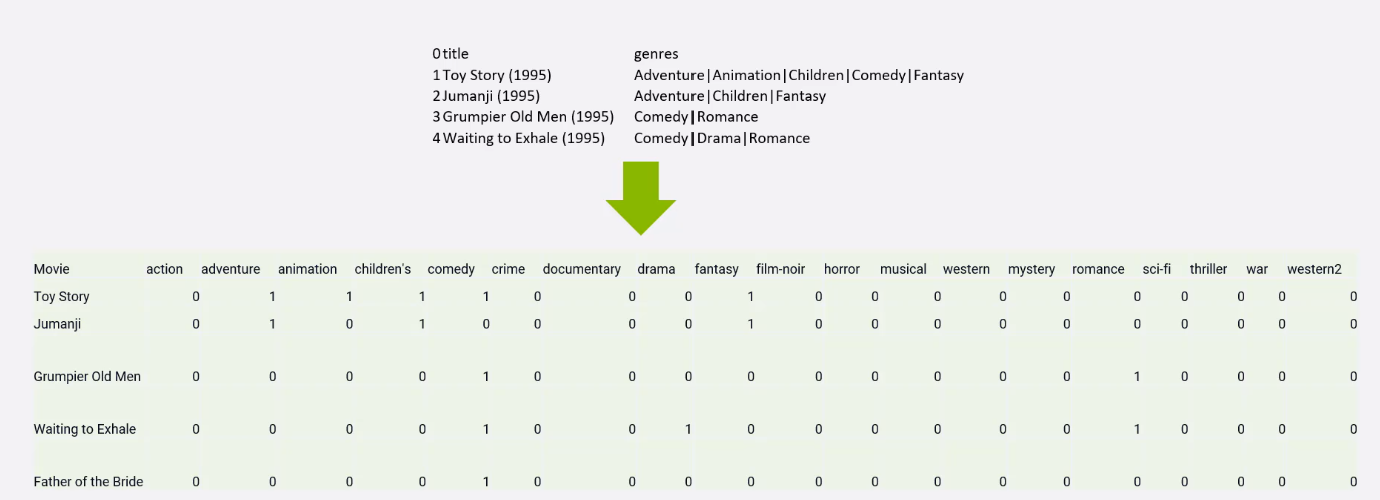


- Biểu diễn người dùng: Mỗi người dùng được biểu diễn bằng một hồ sơ (profile) của các bộ phim họ đã tương tác trước đó.

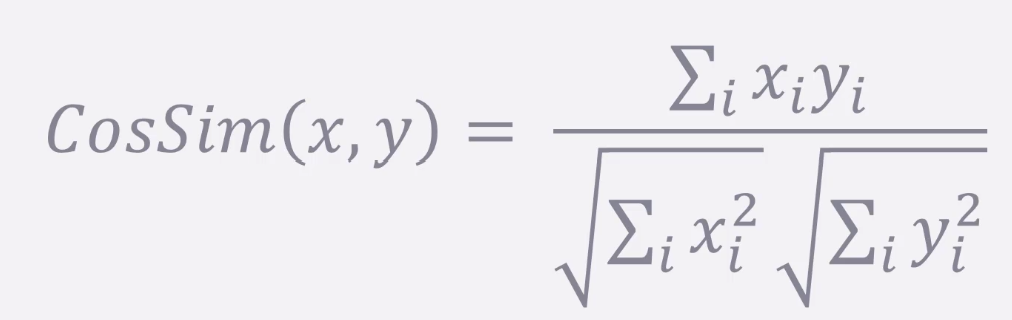
- Tính toán độ tương tự: Độ tương tự giữa các mặt hàng được tính bằng cách so sánh đặc trưng hoặc thuộc tính của chúng. Phương pháp tính độ tương tự được sử dụng bao gồm cosine similarity:

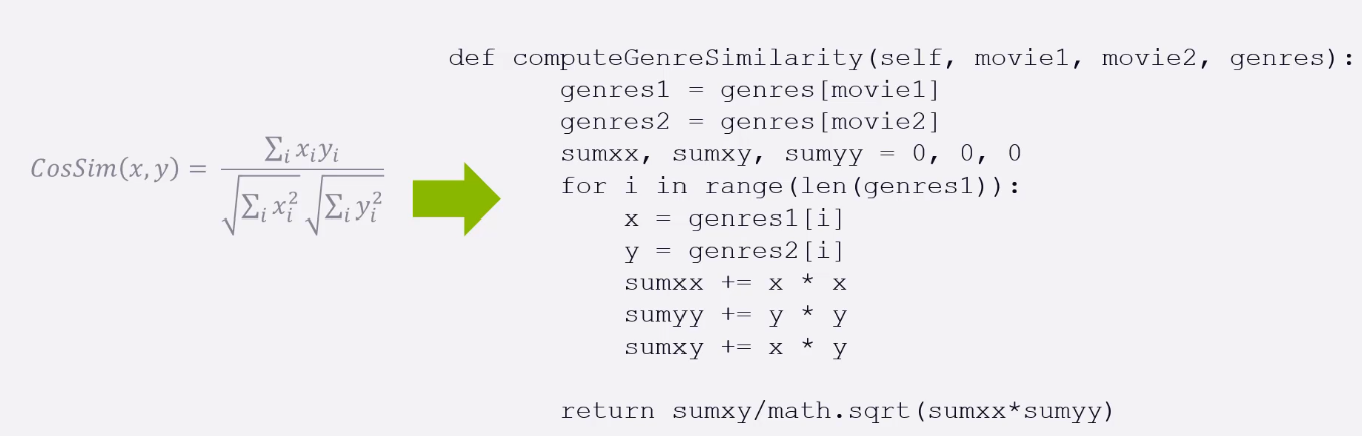


Chuyển phim thành các vector:



Công thức tính độ tương tự:

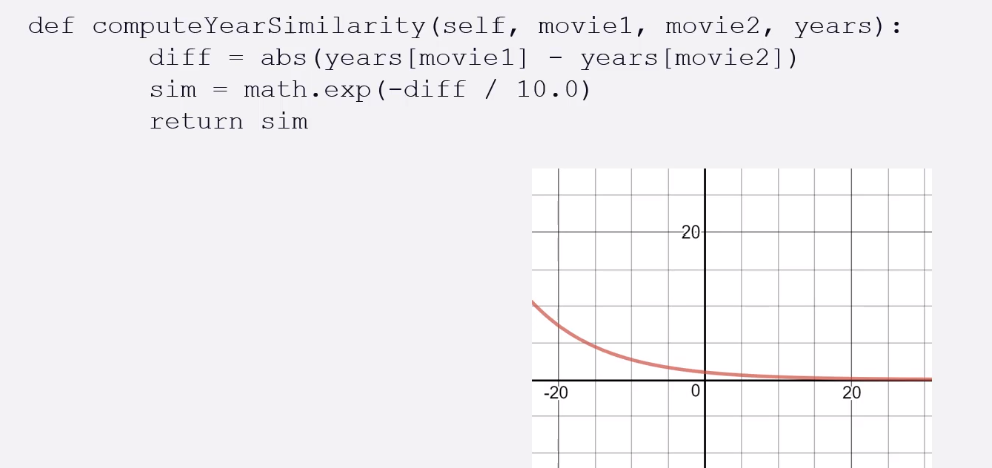




Hàm nhận vào 2 id bộ phim và 1 dict map id phim với tập thể loại của nó

Thời gian phát hành của phim:  


Đánh giá độ tương đồng dựa trên thời gian phát hành:

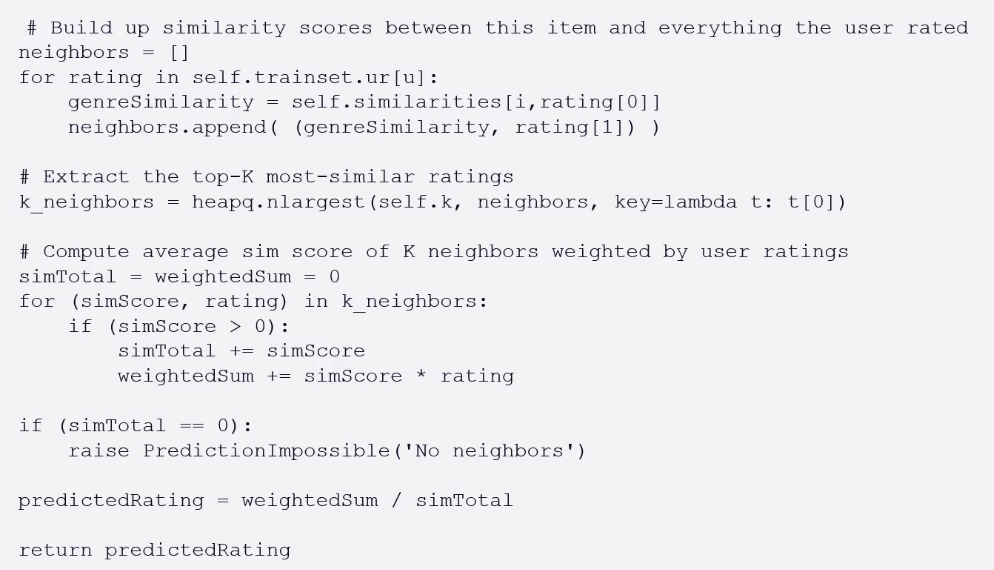


- Đề xuất mặt hàng tương tự: k-nearest-neighbors (đưa ra đánh giá dự đoán của một người dùng cho một bộ phim họ chưa đánh giá)

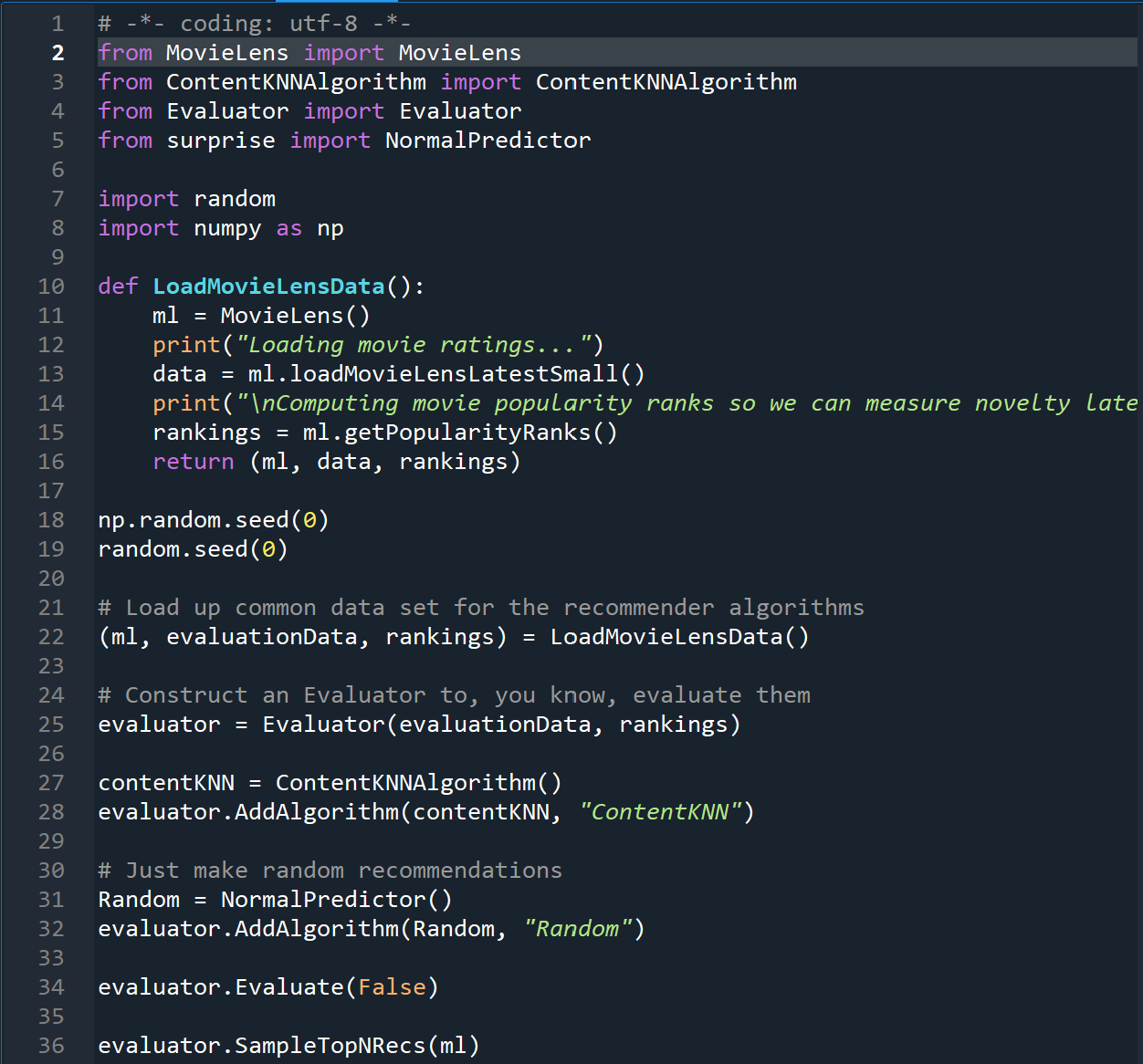
+ Tính toán độ tương tự của bộ phim đang xét đối với tất cả các bộ phim đã được người dùng đánh giá

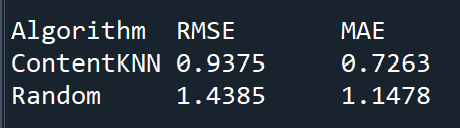
+ Các bộ phim gần nhất có độ tương đồng cao nhất

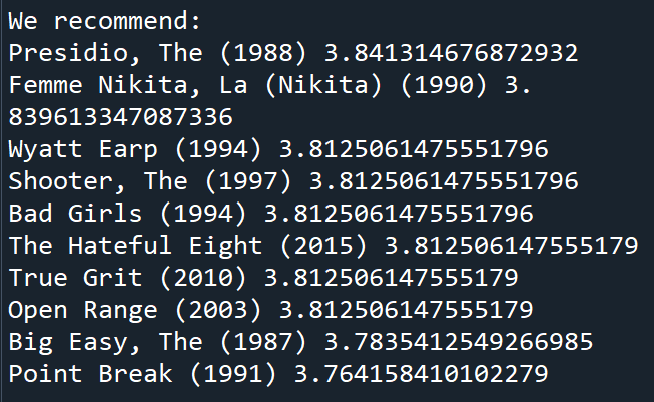
+ Lấy trung bình độ tương đồng với đánh giá của bộ phim nó tương đồng với

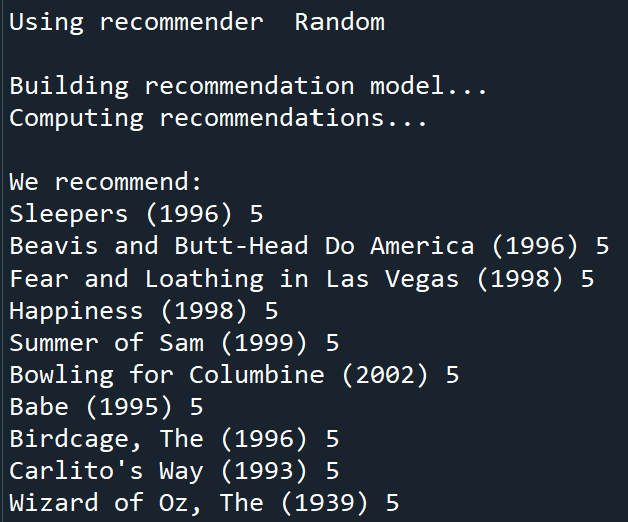


- Triển khai:



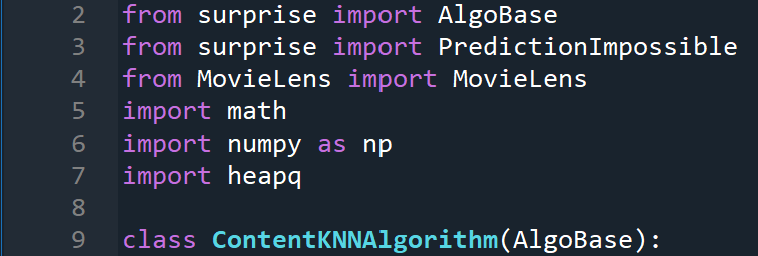


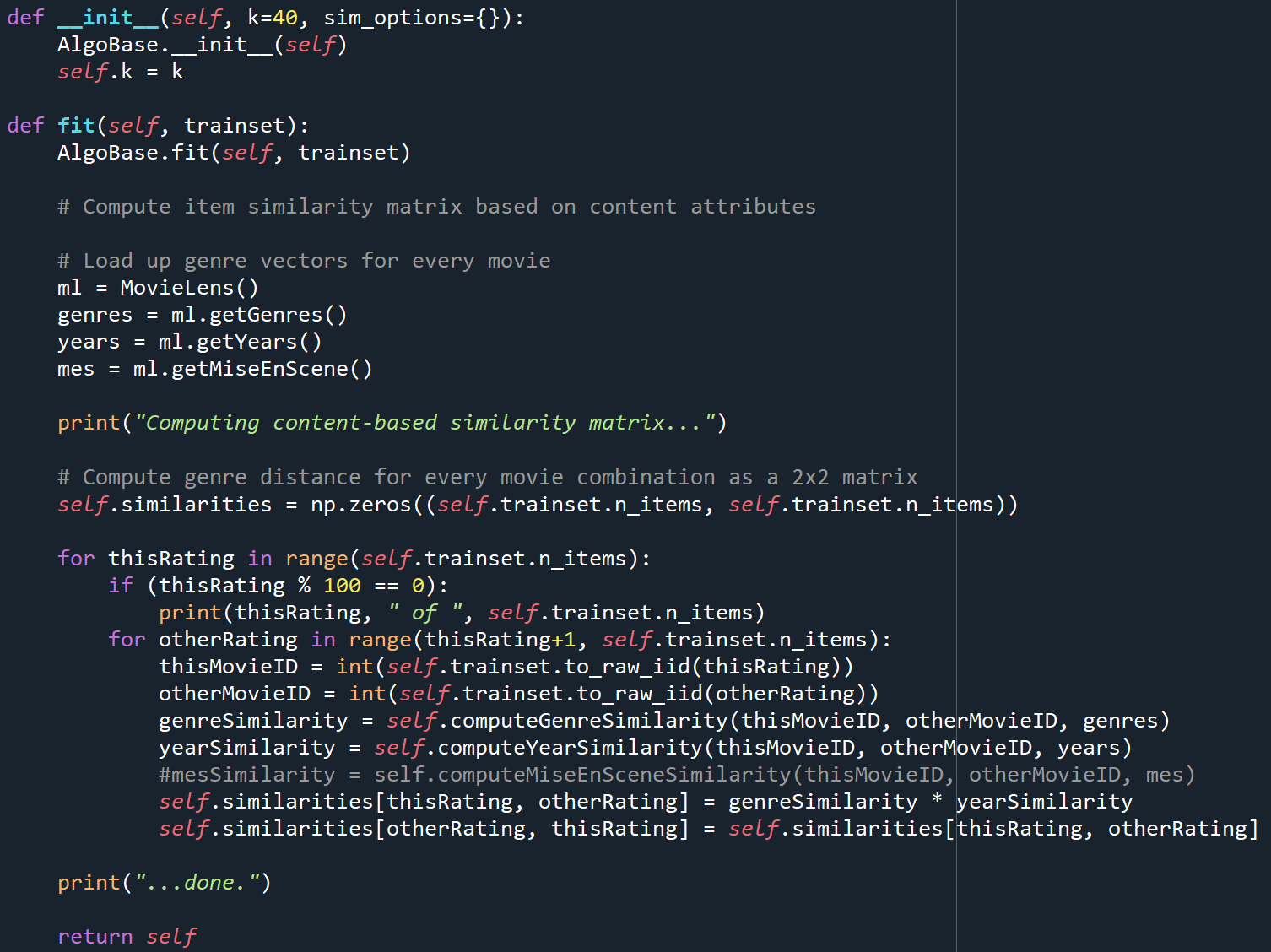




Class:

ContentKNNAlgorithm:





b. Neighborhood-based collaborative filtering: gợi ý dựa trên việc tính toán độ tương tự giữa các người dùng hoặc các mặt hàng để tạo ra các dự đoán về sở thích của người dùng.

- User-based collaborative filtering: Tính toán độ tương tự giữa các cặp người dùng dựa trên lịch sử đánh giá của họ và sau đó sử dụng các người dùng tương tự nhất để tạo ra các dự đoán cho người dùng hiện tại.

- Item-based collaborative filtering: Tính toán độ tương tự giữa các cặp mặt hàng dựa trên cách người dùng đã đánh giá chúng và sau đó sử dụng các mặt hàng tương tự nhất để tạo ra các dự đoán cho người dùng hiện tại.

Phương pháp đo sự tương đồng:

- Cosine similarity: phương pháp để đo độ tương đồng giữa hai vectơ trong không gian n chiều. Các chiều sẽ đại diện cho: người dùng có thích sản phẩm này hay không/sản phẩm có được người dùng này thích hay không

Khó khăn:

- Dữ liệu thưa thớt: phim có thể có it người dùng xem/người dùng có thể chưa xem nhiều phim