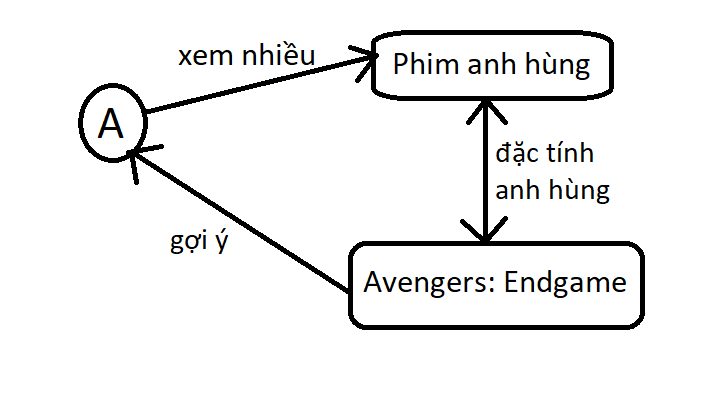
**Content-based Recommendation Systems (CB RS)**

**Hệ thống gợi ý phim và bộ dữ liệu MovieLens**

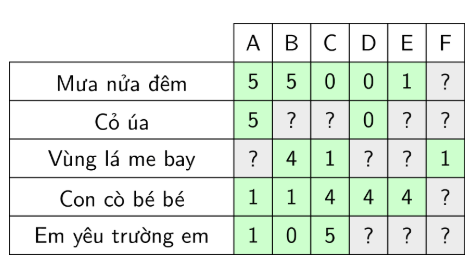
* Youtube tự chuyển đến các clip liên quan đến clip đang xem hoặc tự gợi ý những clip mà bạn sẽ thích
* Facebook quảng cáo liên quan đến thông tin vừa tìm kiếm; gợi ý kết bạn
* “Frequently bought together” (Sẽ sử dụng với **Luật kết hợp trong** Machine Learningsau)
* Hiện tượng Long Tail trong thương mại:
* *Nguyên lý Pareto* : Phần lớn kết quả gây ra bởi phần nhỏ nguyên nhân
* Phần lớn doanh thu (80%) đến từ phần nhỏ số sản phẩm phổ biến nhất (20%).
* Với các cửa hàng online, gian trưng bày là vô tận, việc sắp xếp linh hoạt, doanh thu vì thế có thể được tăng lên.
* **Recommendation Systems (RS) là một mảng khá rộng của Machine Learning và có tuổi đời ít hơn so với Classfication vì internet mới chỉ thực sự bùng nổ khoảng 10-15 năm đổ lại đây.**
* Có hai thực thể chính trong RS là **users** (người dùng) và **items** (sản phẩm).
* Mục đích chính của các RS là dự đoán **mức độ quan tâm** của một user tới một item nào đó, qua đó có chiến lược recommend phù hợp.

Dưới đây là một mô tả đơn giản về cách hoạt động của Content-based RS:



Cần sắp xếp các items vào từng nhóm hoặc đi tìm các đặc trưng của từng item.

1. **Utility matrix**

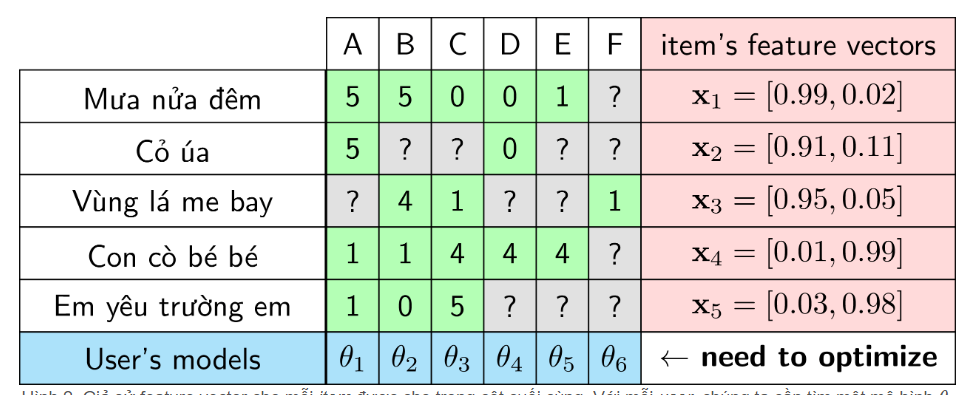
****

Hình 1: Ví dụ về utility matrix với hệ thống gợi ý bài hát.

* Các items được các users **rate** theo mức độ từ 0 đến 5 sao. Các dấu ‘?’ ứng với việc chưa có dữ liệu.
* RS cần phải dự đoán và tự điền các giá trị này. (Đôi khi đây cũng được coi là bài toán Matrix Completion)
* Trong ví dụ trên, 3 bài đầu là nhạc Bolero, 2 bài sau là nhạc thiếu nhi. Dễ dàng đoán được rằng, A, B thích nhạc Bolero; C, D, E, F thích nhạc thiếu nhi. Từ đó, hệ thống gợi ý “Cỏ úa” cho B, “Vùng lá me bay” cho A; “Em yêu trường em” cho D, E, F.
* Khi có một bài hát mới, ta chỉ cần phân lớp nó vào thể loại nào.
* Trong thực tế, lượng dấu ‘?’ là rất lớn.
* Xây dựng utility matrix là quan trọng. Có hai hướng tiếp cận: Nhờ người dùng rate sản phẩm và dựa trên hành vi của users.

1. **Item profiles**

* Item profiles được biểu diễn dưới dạng toán học là một **feature vector** (fv).



Hình 2: Feature vector cho mỗi item được cho trong cột cuối cùng

* Feature vector có hai chiều: chiều thứ nhất là mức độ Bolero, chiều thứ hai là mức độ Thiếu nhi.
* Với mỗi user, chúng ta cần tìm một mô hình **θi** tương ứng sao cho mô hình thu được là tốt nhất.

1. **Xây dựng hàm mất mát**

* Sử dụng mô hình **Ridge Regression**.
* Giả sử có **N** users, **M** items, utility matrix **Y**.
* Hàng **m**, cột **n** của **Y** là số sao **rated** của user **n** lên item **m.**
* **R** là ma trận *rated or not* thể hiện một user đã rated một item hay chưa; cụ thể, **rij** bằng 1 nếu item I đã được rated bởi user j và ngược lại.

**Mô hình tuyến tính:**

**…**

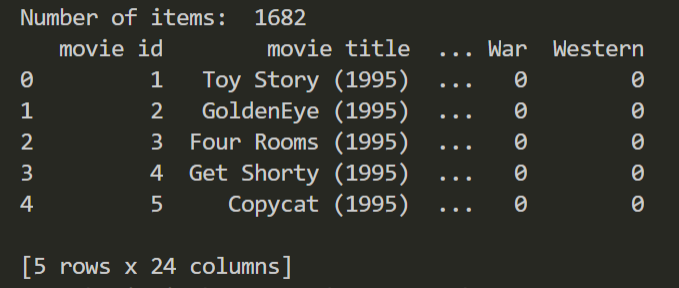
**Ví dụ về hàm mất mát cho user E:**

**…**

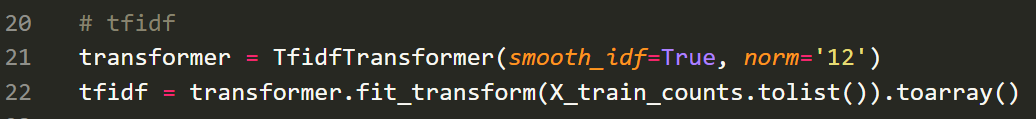
1. **Xây dựng item profile**

* Công việc quan trọng trong CB RS là xây dựng profile cho mỗi item, tức **fv** cho mỗi item.
* Load toàn bộ thông tin về các items vào biến items. Vì ta đang dựa trên thể loại của phim nên sẽ chỉ quan tâm tới 19 giá trị nhị phân ở cuối mỗi hàng.

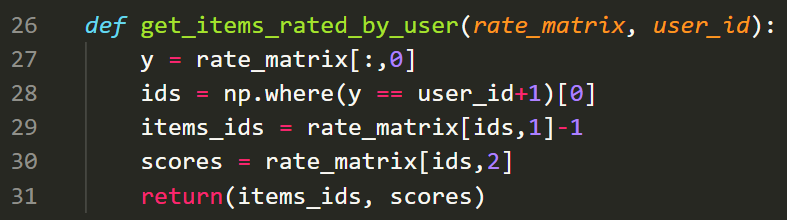




* Xây dựng **fv** cho mỗi item dựa trên ma trận thể loại phim và feature **TF-IDF** sử dụng **sklearn**. Mỗi hàng của **tfidf** tương ứng với feature vector của một bộ phim.

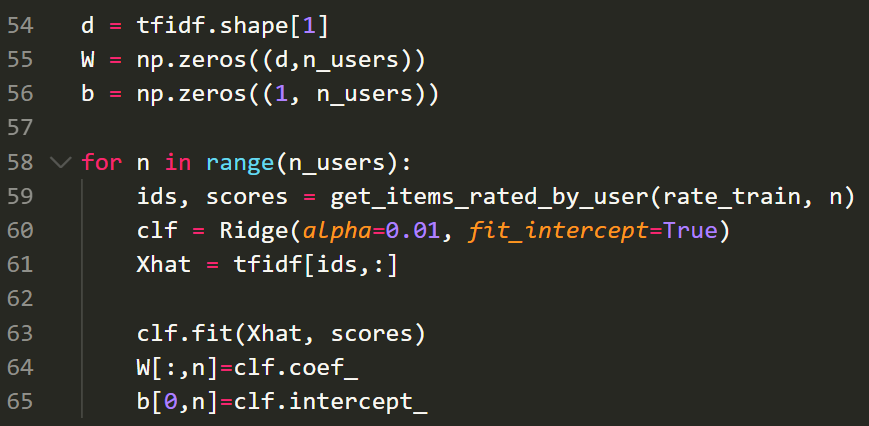


* Với mỗi user, cần tìm những bộ phim mà user đó đã rated và giá trị của các rating đó.

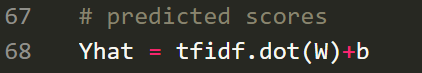


1. **Tìm mô hình cho mỗi user**

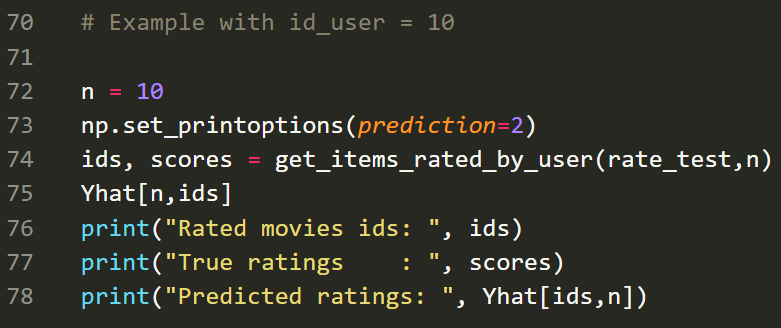
* Tìm các hệ số của Ridge Regression cho mỗi user.

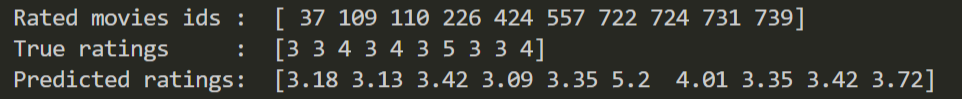


* Sau khi tính được các hệ số W và b, ratings cho mỗi items được dự đoán bằng cách:



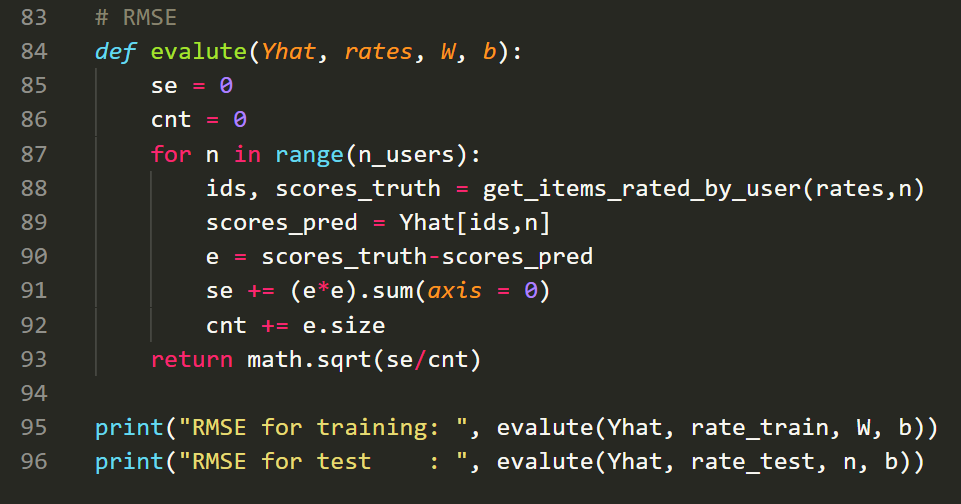
* Ví dụ với user có id là 10.

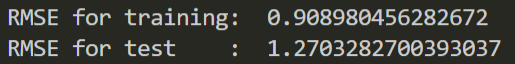




1. **Đánh giá mô hình**

* Sử dụng Root Mean Squared Error (RMSE).
* Lỗi được tính là **hiệu true rating và predicted rating**.





* Với tập training, sai số khoảng 0.9 sao; với tập test, sai số khoảng 1.3 sao.
* **Kết quả chưa thực sự tốt vì đã đơn giản hóa mô hình đi rất nhiều.**

1. **Thảo luận**

* Content-based Recommendation Systems là phương pháp đơn giản nhất trong các hệ thống RS.
* Đặc điểm của phương pháp này là việc xây dựng mô hình cho mỗi user không phụ thuộc vào các users khác.
* Item profile không phụ thuộc vào user mà phụ thuộc vào các đặc điểm mô tả của item hoặc cách yêu cầu người dùng gắn tag.
* Kết quả tồi hơn so với phương pháp Collaborative Filtering.

