Xây dựng một content-based recommendation system đơn giản cho dataset Movilens Lastest Small(100k ratings)

1. Content-based: hệ gợi ý sử dụng thông tin về thể loại(genres) của mỗi bộ phim

Table

Description automatically generated

Có tổng cộng tất cả 19 thể loại được nêu ra trong dataset này. Mỗi bộ phim có thể thuộc một hay nhiều thể loại cùng lúc, cũng có 1 số bộ phim chưa xác định được thể loại (no genres listed)

A picture containing table

Description automatically generated

=> Ta biểu diễn thuộc tính thể loại của phim dưới dạng 1 vector 20 chiều (19 thể loại + 1 (no genres listed))

A picture containing background pattern

Description automatically generated

Dữ liệu bảng rating chia làm 2 phần cho training và testing với tỉ lệ 90/10

2. Features selection

Từ tập các vector thu được ở trên, biến đổi thành features TF-IDF sử dụng thư viện sklearn

Table

Description automatically generated

3. Training

Từ feature vector của các bộ phim đã tính được ở trên và sử dụng bảng rating, với mỗi user ta sẽ tìm 1 mô hình để ước lượng điểm rating của user đó trên từng bộ phim.

Nhóm chọn 1 mô hình đơn giản là Ridge Regression cho tất cả users.

Kết quả được lưu trong biến Yhat:

Text

Description automatically generated

Kết quả minh họa cho 1 user:

Text

Description automatically generated

4. Đánh giá mô hình

Sử dụng RMSE(root mean squared error)

Text

Description automatically generated

Với tập training, sai số dự đoán là khoảng 0.8 sao, trong khi đó với tập test sai số cao hơn, khoảng 0.94 sao