**DS200 – Lab 2**

**Dữ liệu:**

1. **movies.txt**

* **Schema:** MovieID, Title, Genres

1. **ratings\_1.txt, ratings\_2.txt**

* **Schema:** UserID, MovieID, Rating, Timestamp

1. **users.txt**

* **Schema:** UserID, Gender, Age, Occupation, Zip-code

1. **occupation.txt**

* **Schema**: ID, Occupation

**Thực hiện các bài tập dưới đây sử dụng RDD và PySpark.**

**Bài 1: Tính Điểm Đánh Giá Trung Bình và Tổng Số Lượt Đánh Giá Cho Mỗi Phim**

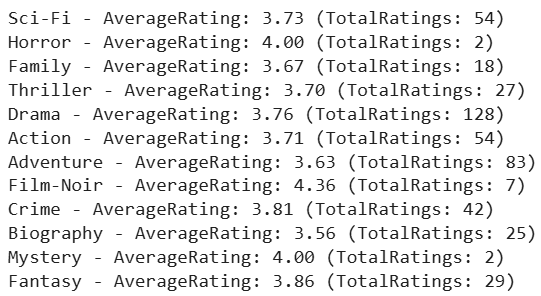
* **Mục tiêu:**
  + Tính điểm trung bình cho từng phim từ cả 2 file ratings (ratings\_1.txt và ratings\_2.txt).
  + Tính tổng số lượt đánh giá cho mỗi phim.
  + Tìm ra phim có điểm trung bình cao nhất (**chỉ xét những phim có tối thiểu 5 lượt đánh giá**).
  + **Output:**   
      
    MovieTitle AverageRating: xx (TotalRatings: xx)  
      
    *MovieTitle* is the highest rated movie with an average rating of *AverageRating* among movies with at least 5 ratings.

A screenshot of a computer

AI-generated content may be incorrect.

**Bài 2: Phân Tích Đánh Giá Theo Thể Loại**

* **Mục tiêu:**
  + Vì một phim có thể thuộc nhiều thể loại (Genres được phân tách bằng dấu “|”), mapper cần tách riêng từng thể loại của phim đó.
  + Tính điểm trung bình (và tổng số lượt đánh giá) cho từng thể loại, dựa trên tất cả các phim thuộc thể loại đó.
* **Output:** Genre - AverageRating (TotalRatings)



**Bài 3: Phân Tích Đánh Giá Theo Giới Tính**

* **Mục tiêu:**
  + Thực hiện join dữ liệu giữa ratings và users (dựa trên UserID) để lấy thông tin giới tính của người đánh giá.
  + Với mỗi phim, tính riêng điểm trung bình từ người dùng nam và nữ.
* **Output:** MovieTitle - Male\_Avg: xx, Female\_Avg: xx

A white text with black text

AI-generated content may be incorrect.

**Bài 4: Phân Tích Đánh Giá Theo Nhóm Tuổi**

* **Mục tiêu:**
  + Phân nhóm người dùng theo độ tuổi (ví dụ: 0-18, 18-35, 35-50, 50+).
  + Với mỗi phim, tính điểm trung bình cho mỗi nhóm tuổi.
* **Output:** MovieTitle - [0-18: AvgRating, 18-35: AvgRating, 35-50: AvgRating, 50+: AvgRating]

A white background with black numbers and letters

AI-generated content may be incorrect.

**Bài 5: Phân Tích Đánh Giá Theo Occupation (Nghề nghiệp) Của Người Dùng**

* **Mục tiêu:**
  + Tính trung bình rating và tổng số lượt đánh giá cho từng Occupation.
* **Output:** Occupation - TotalRatings: xx, AverageRating: xx

A screenshot of a computer

AI-generated content may be incorrect.

**Bài 6 (Tùy chọn): Phân Tích Đánh Giá Theo Thời Gian**

* **Mục tiêu:**
  + Tính tổng số lượt đánh giá và điểm trung bình cho mỗi năm.
* **Output:** Year - TotalRatings: xx, AverageRating: xx

