DS200 – Lab 1

Dữ liệu:

1. **movies.txt**
   - **Schema:** MovieID, Title, Genres
2. **ratings_1.txt, ratings_2.txt**
   - **Schema:** UserID, MovieID, Rating, Timestamp
3. **users.txt**
   - **Schema:** UserID, Gender, Age, Occupation, Zip-code

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.
  - **Output:** MovieTitle AverageRating: xx (TotalRatings: xx)

```
American Beauty (1999)  Average rating: 4.5 (Total ratings: 1)
Avatar (2009)   Average rating: 5.0 (Total ratings: 1)
Back to the Future (1985)       Average rating: 4.0 (Total ratings: 1)
Birdman (2014)  Average rating: 3.0 (Total ratings: 1)
Braveheart (1995)       Average rating: 4.5 (Total ratings: 1)
Coco (2017)     Average rating: 5.0 (Total ratings: 1)
Dunkirk (2017)  Average rating: 3.0 (Total ratings: 1)
Fight Club (1999)       Average rating: 5.0 (Total ratings: 1)
Forrest Gump (1994)     Average rating: 5.0 (Total ratings: 1)
Gladiator (2000)        Average rating: 3.0 (Total ratings: 1)
Good Will Hunting (1997)        Average rating: 3.0 (Total ratings: 1)
```

- **Yêu cầu thêm:**
  - Trong quá trình xử lý, tìm ra phim có điểm trung bình cao nhất (**chỉ xét những phim có tối thiểu 5 lượt đánh giá**).
  - Sử dụng biến lớp như maxMovie và maxRating trong reducer và xuất ra kết quả cuối cùng trong phương thức cleanup(), theo định dạng:

    MovieTitle is the highest rated movie with an average rating of *AverageRating* among movies with at least 5 ratings.

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 nếu cần) cho từng thể loại, dựa trên tất cả các phim thuộc thể loại đó.
- **Output:** Genre: AverageRating (TotalRatings)

```
Action  Avg: 3.72, Count: 20
Adventure       Avg: 3.85, Count: 30
Animation       Avg: 4.17, Count: 6
Biography       Avg: 4.14, Count: 14
Children        Avg: 4.25, Count: 2
Comedy  Avg: 3.96, Count: 12
Crime   Avg: 3.78, Count: 20
Drama   Avg: 3.81, Count: 76
Family  Avg: 4.25, Count: 2
Fantasy Avg: 3.95, Count: 10
History Avg: 4.00, Count: 6
Music   Avg: 3.88, Count: 4
Mystery Avg: 4.00, Count: 8
Romance Avg: 3.75, Count: 8
Sci-Fi  Avg: 3.85, Count: 20
Thriller        Avg: 3.88, Count: 16
War     Avg: 4.00, Count: 2
```

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, Female_Avg

```
American Beauty (1999)  Male: 3.00, Female: 4.50
Avatar (2009)   Male: 5.00, Female: 4.50
Back to the Future (1985)       Male: 4.00, Female: 4.00
Birdman (2014)  Male: 3.00, Female: 3.00
Braveheart (1995)       Male: 4.50, Female: 4.00
Coco (2017)     Male: 5.00, Female: 3.50
Dunkirk (2017)  Male: 4.00, Female: 3.00
Fight Club (1999)       Male: 5.00, Female: 3.50
Forrest Gump (1994)     Male: 2.50, Female: 5.00
Gladiator (2000)        Male: 3.50, Female: 3.00
Good Will Hunting (1997)        Male: 3.00, Female: 4.00
```

Bài 4 (Tùy Chọn): 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]

```
American Beauty (1999)  0-18: NA  18-35: 3.75  35-50: NA  50+: NA
Avatar (2009)   0-18: NA  18-35: 4.50  35-50: 5.00  50+: NA
Back to the Future (1985)       0-18: NA  18-35: 4.00  35-50: NA  50+: NA
Birdman (2014)  0-18: NA  18-35: 3.00  35-50: NA  50+: NA
Braveheart (1995)       0-18: NA  18-35: 4.00  35-50: 4.50  50+: NA
Coco (2017)     0-18: NA  18-35: 4.25  35-50: NA  50+: NA
Dunkirk (2017)  0-18: NA  18-35: 3.50  35-50: NA  50+: NA
Fight Club (1999)       0-18: NA  18-35: 4.25  35-50: NA  50+: NA
Forrest Gump (1994)     0-18: NA  18-35: 3.75  35-50: NA  50+: NA
Gladiator (2000)        0-18: NA  18-35: 3.00  35-50: 3.50  50+: NA
Good Will Hunting (1997)        0-18: NA  18-35: 3.50  35-50: NA  50+: NA
```