

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN



BÁO CÁO LAB 2: SPARK

GVHD: CN. Nguyễn Hiếu Nghĩa

Lớp: IE212.Q11

Sinh viên thực hiện:

Họ và tên: Trương Hoài Bảo

MSSV: 22520126

Thành phố Hồ Chí Minh, 11/2025

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

....., ngày tháng năm 2025

Người nhận xét

(Ký tên và ghi rõ họ tên)

MỤC LỤC

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	ii
DANH MỤC HÌNH ẢNH.....	iv
NỘI DUNG THỰC HÀNH.....	5
1.1Bài 1: Tính Điểm Đánh Giá Trung Bình và Tổng Số Lượt Đánh Giá Cho Mỗi Phim.....	5
1.2Bài 2: Phân Tích Đánh Giá Theo Thể Loại	5
1.3Bài 3: Phân Tích Đánh Giá Theo Giới Tính.....	6
1.4Bài 4 : Phân Tích Đánh Giá Theo Nhóm Tuổi	6
1.5Bài 5: Phân Tích Đánh Giá Theo Occupation (Nghề nghiệp) Của Người Dùng.....	7
1.6Bài 6: Phân Tích Đánh Giá Theo Thời Gian	7
TÀI LIỆU THAM KHẢO	8

DANH MỤC HÌNH ẢNH

Hình 1.1-1 Kết quả câu 1.....	5
Hình 1.2-1 Kết quả câu 2.....	5
Hình 1.3-1 Kết quả câu 3.....	6
Hình 1.4-1 Kết quả câu 4.....	6
Hình 1.5-1 Kết quả câu 5.....	7
Hình 1.6-1 Kết quả câu 6.....	7

NỘI DUNG THỰC HÀNH

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

The screenshot shows a Jupyter Notebook interface with two tabs: 'bai5.ipynb' and 'bai1.ipynb'. The 'bai1.ipynb' tab is active, displaying Python code and its execution output. The code prints the highest rated movie from a dataset. The output shows several movies with their average ratings and total ratings, concluding with 'Sunset Boulevard (1950) is the highest rated movie with an average rating of 4.36 among movies with at least 7 ratings.' Below the notebook is a terminal window showing the command to activate a virtual environment and the current directory.

```
print(f"{title} is the highest rated movie with an average rating of {avg:.2f} among movies with at least {total} ratings.")

... After merging: 184
E.T. the Extra-Terrestrial (1982): AverageRating:3.67 (TotalRatings:18)
Mad Max: Fury Road (2015): AverageRating:3.47 (TotalRatings:18)
Sunset Boulevard (1950): AverageRating:4.36 (TotalRatings:7)
The Lord of the Rings: The Return of the King (2003): AverageRating:3.82 (TotalRatings:11)
Lawrence of Arabia (1962): AverageRating:3.44 (TotalRatings:18)
Fight Club (1999): AverageRating:3.58 (TotalRatings:7)
Gladiator (2000): AverageRating:3.61 (TotalRatings:18)
The Social Network (2010): AverageRating:3.86 (TotalRatings:7)
Psycho (1960): AverageRating:4.00 (TotalRatings:2)
The Silence of the Lambs (1991): AverageRating:3.14 (TotalRatings:7)
Sunset Boulevard (1950) is the highest rated movie with an average rating of 4.36 among movies with at least 7 ratings.

# Clear
spark_context.stop()
spark_session.stop()

# Kết quả
for (genre, (avg, cnt)) in genre_avg.take(10):
    print(f"{genre} - AverageRating:{avg:.2f} (TotalRatings:{cnt})")
```

```
PROBLEMS 4 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER SONARQUBE 4
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ source "/d/UIT_Courses/Bigdata/Lab 2/.venv/scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ 
```

Hình 1.1-1 Kết quả câu 1

1.2 Bài 2: Phân Tích Đánh Giá Theo Thể Loại

The screenshot shows a Jupyter Notebook interface with two tabs: 'bai5.ipynb' and 'bai2.ipynb'. The 'bai2.ipynb' tab is active, displaying Python code and its execution output. The code prints the average rating for each movie genre. The output shows genres like Sci-Fi, Action, Family, Drama, Biography, Horror, Thriller, Adventure, Film-Noir, and Mystery, each with their average rating and total ratings. Below the notebook is a terminal window showing the command to activate a virtual environment and the current directory.

```
# Kết quả
for (genre, (avg, cnt)) in genre_avg.take(10):
    print(f"{genre} - AverageRating:{avg:.2f} (TotalRatings:{cnt})")

... After merging: 184
Sci-Fi - AverageRating:3.73 (TotalRatings:54)
Action - AverageRating:3.71 (TotalRatings:54)
Family - AverageRating:3.67 (TotalRatings:18)
Drama - AverageRating:3.76 (TotalRatings:128)
Biography - AverageRating:3.56 (TotalRatings:25)
Horror - AverageRating:4.00 (TotalRatings:2)
Thriller - AverageRating:3.70 (TotalRatings:27)
Adventure - AverageRating:3.63 (TotalRatings:83)
Film-Noir - AverageRating:4.36 (TotalRatings:7)
Mystery - AverageRating:4.00 (TotalRatings:2)

# Clear
spark_context.stop()
spark_session.stop()

# Kết quả
for (genre, (avg, cnt)) in genre_avg.take(10):
    print(f"{genre} - AverageRating:{avg:.2f} (TotalRatings:{cnt})")
```

```
PROBLEMS 7 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER SONARQUBE 7
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ source "/d/UIT_Courses/Bigdata/Lab 2/.venv/scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ 
```

Hình 1.2-1 Kết quả câu 2

IE108.O21 – Báo cáo đồ án cuối môn

1.3 Bài 3: Phân Tích Đánh Giá Theo Giới Tính

The screenshot shows the Jupyter Notebook interface with two open files: bai5.ipynb and bai3.ipynb. The bai3.ipynb file is active and displays a list of movies with their average ratings for males and females. The code used is:

```
print(f"\n{title} - Male_Avg: {formatter(male_avg)}, Female_Avg: {formatter(female_avg)}")
```

The output shows the following movie ratings:

- Gladiator (2000) - Male_Avg: 3.59, Female_Avg: 3.64
- The Terminator (1984) - Male_Avg: 3.93, Female_Avg: 4.14
- Lawrence of Arabia (1962) - Male_Avg: 3.55, Female_Avg: 3.31
- Mad Max: Fury Road (2015) - Male_Avg: 4.00, Female_Avg: 3.32
- No Country for Old Men (2007) - Male_Avg: 3.92, Female_Avg: 3.83
- Psycho (1960) - Male_Avg: NA, Female_Avg: 4.00
- E.T. the Extra-Terrestrial (1982) - Male_Avg: 3.81, Female_Avg: 3.55
- Fight Club (1999) - Male_Avg: 3.50, Female_Avg: 3.50
- The Godfather: Part II (1974) - Male_Avg: 4.06, Female_Avg: 3.94
- The Lord of the Rings: The Fellowship of the Ring (2001) - Male_Avg: 4.00, Female_Avg: 3.80
- The Silence of the Lambs (1991) - Male_Avg: 3.33, Female_Avg: 3.00
- Sunset Boulevard (1950) - Male_Avg: 4.33, Female_Avg: 4.50
- The Lord of the Rings: The Return of the King (2003) - Male_Avg: 3.75, Female_Avg: 3.90
- The Social Network (2010) - Male_Avg: 4.00, Female_Avg: 3.67

The terminal tab shows the command to stop the spark context:

```
# clear  
spark_context.stop()
```

The bottom of the interface includes tabs for PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL, PORTS, JUPYTER, and SONARQUBE.

Hình 1.3-1 Kết quả câu 3

1.4 Bài 4 : Phân Tích Đánh Giá Theo Nhóm Tuổi

The screenshot shows the Jupyter Notebook interface with two open files: bai5.ipynb and bai4.ipynb. The bai4.ipynb file is active and displays a list of movies grouped by age categories. The code used is:

```
for _, (title, age_dict) in movie_age_group.take(10):  
    print(  
        f"\n{title} - "  
        f"[0-18: {formatter(age_dict.get('0-18'))}, "  
        f"18-35: {formatter(age_dict.get('18-35'))}, "  
        f"35-50: {formatter(age_dict.get('35-50'))}, "  
        f"50+: {formatter(age_dict.get('50+'))}]"  
    )
```

The output shows the following movie age groups:

- Gladiator (2000) - [0-18: NA, 18-35: 3.44, 35-50: 3.81, 50+: 3.50]
- The Terminator (1984) - [0-18: NA, 18-35: 4.17, 35-50: 4.05, 50+: 3.75]
- Lawrence of Arabia (1962) - [0-18: NA, 18-35: 3.60, 35-50: 3.29, 50+: 4.50]
- Mad Max: Fury Road (2015) - [0-18: NA, 18-35: 3.36, 35-50: 3.64, 50+: NA]
- No Country for Old Men (2007) - [0-18: NA, 18-35: 3.81, 35-50: 3.94, 50+: 4.00]
- Psycho (1960) - [0-18: NA, 18-35: 4.50, 35-50: 3.50, 50+: NA]
- E.T. the Extra-Terrestrial (1982) - [0-18: NA, 18-35: 3.56, 35-50: 3.83, 50+: 3.00]
- Fight Club (1999) - [0-18: NA, 18-35: 3.50, 35-50: 3.50, 50+: 3.50]
- The Godfather: Part II (1974) - [0-18: NA, 18-35: 3.78, 35-50: 4.25, 50+: NA]
- The Lord of the Rings: The Fellowship of the Ring (2001) - [0-18: NA, 18-35: 4.00, 35-50: 3.83, 50+: NA]

The terminal tab shows the command to source the virtual environment:

```
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2  
$ source "/d/UIT_Courses/Bigdata/Lab 2/.venv/Scripts/activate"  
(.venv)  
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2  
$
```

The bottom of the interface includes tabs for PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL, PORTS, JUPYTER, and SONARQUBE.

Hình 1.4-1 Kết quả câu 4

IE108.O21 – Báo cáo đồ án cuối môn

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

The screenshot shows a Jupyter Notebook interface with a Python script named `bai5.ipynb`. The code defines a `formatter` function and prints the top 10 occupations with their average rating and total ratings. The output shows various professions like Nurse, Artist, Manager, etc., with their respective average ratings and counts.

```
def formatter(x):
    return f'{x:.2f}' if x is not None else "NA"

for occupation_name, (avg, count) in rating_occupation_avg.take(10):
    print(f'{occupation_name} - AverageRating: {formatter(avg)} (TotalRatings: {count})')
```

... After merging: 184
Nurse - AverageRating: 3.86 (TotalRatings: 11)
Artist - AverageRating: 3.73 (TotalRatings: 11)
Manager - AverageRating: 3.47 (TotalRatings: 16)
Programmer - AverageRating: 4.25 (TotalRatings: 10)
Designer - AverageRating: 4.00 (TotalRatings: 13)
Teacher - AverageRating: 3.70 (TotalRatings: 5)
Salesperson - AverageRating: 3.65 (TotalRatings: 17)
Engineer - AverageRating: 3.56 (TotalRatings: 18)
Consultant - AverageRating: 3.86 (TotalRatings: 14)
Student - AverageRating: 4.00 (TotalRatings: 8)

TERMINAL

```
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ source "/D/UIT_Courses/Bigdata/Lab 2/.venv/Scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$
```

Hình 1.5-1 Kết quả câu 5

1.6 Bài 6: Phân Tích Đánh Giá Theo Thời Gian

The screenshot shows a Jupyter Notebook interface with a Python script named `bai6.ipynb`. The code defines a `rating_year_avg` function using `mapValues` and prints the top 10 years with their average rating and total ratings. The output shows the year 2020 with an average rating of 3.75.

```
#(year, (avg, count))
rating_year_avg = ratings_stats.mapValues(
    lambda x: (x[0] / x[1], x[1])
)

def formatter(x):
    return f'{x:.2f}' if x is not None else "NA"

for year, (avg, count) in rating_year_avg.take(10):
    print(f'{year} - (TotalRatings: {count}), AverageRating: {formatter(avg)}')
```

... After merging: 184
2020 - (TotalRatings: 184), AverageRating: 3.75

```
# Clear
spark_context.stop()
spark_session.stop()
```

TERMINAL

```
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$ source "/D/UIT_Courses/Bigdata/Lab 2/.venv/Scripts/activate"
(.venv)
DELL@TruongHoaiBao-22520126 MINGW64 /d/UIT_Courses/Bigdata/Lab 2
$
```

Hình 1.6-1 Kết quả câu 6

TÀI LIỆU THAM KHẢO