[Spring 2023] DATA MINING THEORY AND APPLICATION (IIE 4102)

Practice-HW4
CBR / AR / CF

논문참조

강세정, 김규현, 신재욱, 장현우

Introduction

Contents

CBR / AR / CF

01. Research Objective

- 사례 기반 추론을 통해 영화의 흥행 유무와 사용자 평점을 예측해 보고자 함
- 연관분석을 통해 사용자들이 자주 함께 구매하는 제품들 간의 연관성을 분석하고자 함
- 협업 필터링을 통해 사용자들에게 영화 추천을 제공할 때, user-based로 나와 비슷한 사용자들이 공통적으로 선호하는 영화를 바탕으로 판단하고자 함

02. Data

데이터는 Kaggle에서 "The Movies Dataset"을 다운로드해 사용함

movies_metadata.csv : 2017년 7월 이전에 개봉한 영화들을 포함한 데이터

ratings_small.csv : 영화 관람객(user)에 대한 데이터

[movies_metadata.csv]

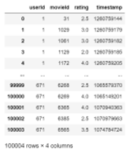
Variable	설명	Example(10번째 row의 데이터 예시)
adult	성인 영화 여부	FALSE
belongs_to_collection	id, name, poster_path, backdrop_path 등을 포함한 변수	{'id': 645, 'name': 'James Bond Collection', 'poster_path': '/HORpg5CSkmeQlAolx3bKMrKgfi.jpg', 'backdrop_path': '/6VcVl48kNKvdXOZfJPdarlUGOsk.jpg'}
budget	영화 예산	58000000
genres	영화 장르	[{'id': 12, 'name': 'Adventure'}, {'id': 28, 'name': 'Action'}, {'id': 53, 'name': 'Thriller'}]
homepage	영화 홈페이지	http://www.mgm.com/view/movie/757/Goldeneye/
id	아이디	710
imdb_id	imdb 아이디	tt0113189
original_language	영화 원본 언어	en
original_title	영화 제목	GoldenEye
overview	영화 요약	James Bond must unmask the mysterious head of the Janus Syndicate and prevent the leader from utilizing the GoldenEye weapons system to inflict devastating revenge on Britain.
popularity	일일 평점 투표 수, 관람 수, 좋아요 수, 개봉일, 총 평점 투표 수, 이전 점수, watchlist 등록 건수 등을 종합적으로 고려한 점수	14.686036
poster_path	영화 포스터 경로	/5c0ovjT41KnYIHYuF4AWsTe3sKh.jpg
production_companies	영화 출품 기업	[{'name': 'United Artists', 'id': 60}, {'name': 'Eon Productions', 'id': 7576}]
production_countries	영화 출품 국가	[{'iso_3166_1': 'GB', 'name': 'United Kingdom'}, {'iso_3166_1': 'US', 'name': 'United States of America'}]
release_date	영화 개봉일	1995-11-16
revenue	영화 수익	352194034
runtime	영화 상영 시간	130
spoken_languages	영화 사용 언어	'iso_639_1': 'en', 'name': 'English'}, {'iso_639_1': 'ru', 'name': 'Pусский'}, {'iso_639_1': 'es', 'name': 'Español'}
status	상태	Released
tagline	추가 설명	No limits. No fears. No substitutes.
title	영화 제목	GoldenEye
video	비디오 존재 여부	FALSE
vote_average	평점	6.6
vote_count	평점 투표 개수	1194

기존 데이터들 중에 numeric variable을 제외한 column들을 모두 drop

label이라는 새로운 column 생성 → popularity가 평균보다 작으면 0, 크면 1로 부여

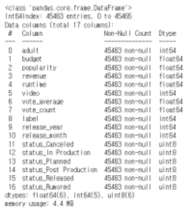
① data

[movies_metadata.csv]

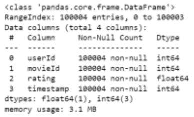


[ratings_small.csv]

② data.info()



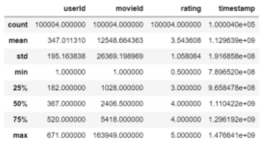
[movies_metadata.csv]



[ratings_small.csv]

③ data.describe()

[movies_metadata.csv]



[ratings_small.csv]

[ratings_small.csv]

Variable	설명	Example(1번째 row의 데이터 예시)
userId	유저 아이디	1
movieId	영화 아이디	31
rating	평점	2.5
timestamp	Unix timestamp(시간)	1260759144

Body

Contents

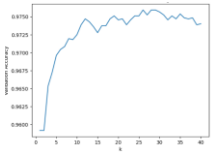
CBR / AR / CF

CBR

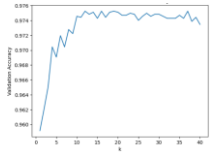
KNN(categorical target : 영화의 흥행 여부)

최적 파라미터 선정 : 각 K에 따른 오류율을 비교하여 K 선택, accuracy 기준

Weight : algorithm auto, metric euclidean 고정. Distance가 accuracy 최고

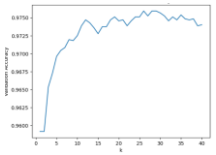


Distance : 0.9781150335422852

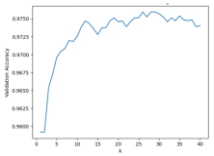


Uniform : 0.976575387660838

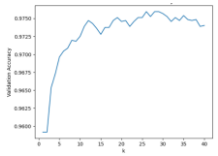
Algorithm : weights distance, metric euclidean 고정. 큰 차이가 없기에 Auto로 선정



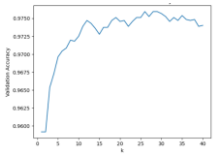
Auto : 0.9781150335422852



ball_tree : 0.9781150335422852

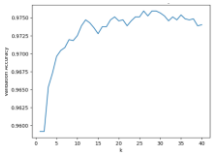


kd_tree : 0.9781150335422852

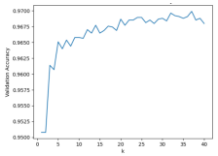


brute : 0.9781150335422852

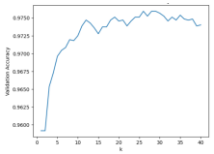
Metric : weights distance, algorithm auto 고정. Manhattan이 accuracy 최고



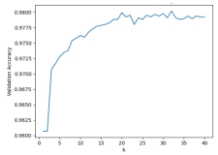
euclidean : 0.9781150335422852



chebyshev : 0.9712966017815903



minkowski : 0.9781150335422852



manhattan : 0.984053667656439

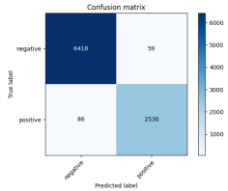
즉, accuracy가 가장 높은 weights=distance, algorithm=auto, metric=manhattan 선정 → 이때, n_neighbors = 32

Accuracy : 0.984053667656439

Precision : 0.977211278485902

Recall : 0.9671253822629969

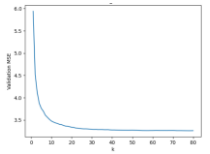
F1 score : 0.9721421709894332



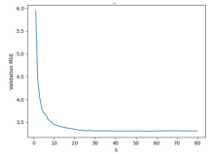
KNN(continuous target : 영화의 평점)

최적 파라미터 선정 : 각 K에 따른 오류율을 비교하여 K 선택, mse 기준

Weight : algorithm auto, metric euclidean 고정. Distance가 mse 최저

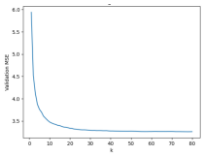


Distance : 3.1500569117149246

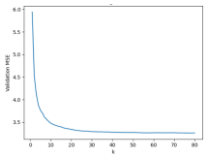


Uniform : 3.19798787796984

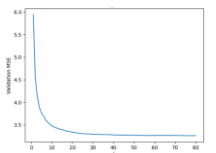
Algorithm : weights distance, metric euclidean 고정. 큰 차이가 없기에 Auto로 선정



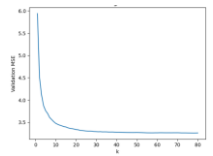
Auto : 3.1500569117149246



ball_tree : 3.1500569117149246

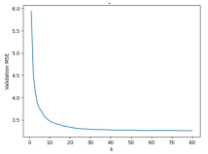


kd_tree : 3.1500569117149246

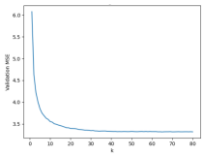


brute : 3.150056911715222

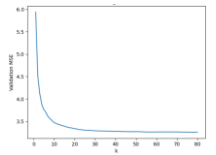
Metric : weights distance, algorithm auto 고정. Manhattan이 mse 최저



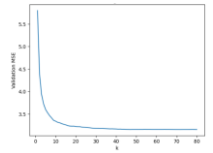
euclidean : 3.1500569117149246



chebyshev : 3.1997020942971712



minkowski : 3.1500569117149246



manhattan : 3.0543789404181876

즉, mse가 가장 낮은 weights=distance, algorithm=auto, metric=manhattan 선정 → 이때, n_neighbors = 72

MSE : 3.0543789404181876

RMSE : 1.7476781569894921

MAE : 1.2219950111848765

MAPE : 1323497180607568.0

y_true = 0인 케이스 때문에 MAPE 값이 매우 크게 나타난다. (scikit-learn 설명 참조)
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html

Body

Contents

CBR / AR / CF

AR

추가 전처리

Rating 데이터에서 movieId로 영화 제목이 확인 가능한 데이터만 사용

(100004, 4)

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205

→

(44994, 4)

	userId	movieId	rating	title
0	1	1371	2.5	Rocky III
1	4	1371	4.0	Rocky III
2	7	1371	3.0	Rocky III
3	19	1371	4.0	Rocky III
4	21	1371	3.0	Rocky III

총 2,794개의 영화가 평점이 매겨졌으며 각 영화당 평점의 수는 1회~324회

이 중에서 최소 30회 이상 평점이 메겨진 영화 403개 사용 / 평점을 매긴 사용자 수는 총 670명

	title	totalRatings
1959	Terminator 3: Rise of the Machines	324
2306	The Million Dollar Hotel	311
1837	Solaris	305
1967	The 39 Steps	291
1374	Monsoon Wedding	274

(403, 670)

userId	1	2	3	4	5	6
title						
20,000 Leagues Under the Sea	0.0	0.0	0.0	3.0	0.0	2.0
2001: A Space Odyssey	0.0	3.0	0.0	0.0	0.0	0.0
28 Weeks Later	0.0	0.0	0.0	0.0	0.0	0.0
300	0.0	0.0	3.0	0.0	0.0	0.0
48 Hrs.	0.0	5.0	0.0	0.0	4.0	0.0

각 사용자가 해당 영화에 평점을 매겼는지에 따라 T/F 로 변환 후 연관분석 시행

(1) Support 값 비교

	support	itemsets
112	0.483582	(Terminator 3: Rise of the Machines)
131	0.464179	(The Million Dollar Hotel)
106	0.455224	(Solaris)
114	0.434328	(The 39 Steps)
70	0.408955	(Monsoon Wedding)
...	...	...
4651	0.100000	(Rain Man, Sissi, 48 Hrs., Syriana)
4648	0.100000	(Rain Man, Silent Hill, 48 Hrs., Three Colors:...
383	0.100000	(Reservoir Dogs, Arlington Road)
4647	0.100000	(Rain Man, Silent Hill, The Million Dollar Hot...
5913	0.100000	(Rain Man, Monsoon Wedding, Psycho, Big Fish)

	support	itemsets
112	0.483582	(Terminator 3: Rise of the Machines)
131	0.464179	(The Million Dollar Hotel)
106	0.455224	(Solaris)
114	0.434328	(The 39 Steps)
70	0.408955	(Monsoon Wedding)
83	0.364179	(Once Were Warriors)
144	0.340299	(Three Colors: Red)
66	0.334328	(Men in Black II)
1209	0.334328	(Solaris, Terminator 3: Rise of the Machines)
1256	0.326866	(The Million Dollar Hotel, Terminator 3: Rise ...
133	0.325373	(The Passion of Joan of Arc)
102	0.320896	(Silent Hill)
104	0.317910	(Sissi)

가장 많은 평점이 남겨진 영화는

'Terminator 3: Rise of the Machines' (support = 0.483582),

'The Million Dollar Hotel' (support = 0.464179),

'Solaris' (support = 0.455224) 등이 있음

'Solaris', 'Terminator 3: Rise of the Machines' 에 같이

평점을 남긴 경우가 많음 (support = 0.334328)

(2) Confidence 값 비교

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
23936	(Rain Man, Romeo + Juliet, 48 Hrs.)	(Sissi)	0.104478	0.317910	0.104478	1.000000	3.145540	0.071263	inf
20744	(Rain Man, 48 Hrs., A Nightmare on Elm Street)	(Monsoon Wedding)	0.126866	0.408955	0.125373	0.988235	2.416488	0.073491	50.238806
60676	(Rain Man, Sissi, 48 Hrs., A Nightmare on Elm ...	(Monsoon Wedding)	0.116418	0.408955	0.114925	0.987179	2.413906	0.067316	46.101493
60736	(Rain Man, 48 Hrs., A Nightmare on Elm Street,...	(Monsoon Wedding)	0.110448	0.408955	0.108955	0.986486	2.412211	0.063787	43.737313
20674	(Back to the Future Part II, 48 Hrs., A Nightm...	(Monsoon Wedding)	0.107463	0.408955	0.105970	0.986111	2.411294	0.062023	42.555224
...	...	...	...	...	...	...	...	...	...
15624	(Terminator 3: Rise of the Machines)	(The Million Dollar Hotel, My Name Is Bruce)	0.483582	0.128358	0.100000	0.206790	1.611039	0.037928	1.098879
4434	(Terminator 3: Rise of the Machines)	(The Passion of Joan of Arc, 5 Card Stud)	0.483582	0.129851	0.100000	0.206790	1.592522	0.037207	1.096998
11807	(Terminator 3: Rise of the Machines)	(Dawn of the Dead, Rope)	0.483582	0.131343	0.100000	0.206790	1.574425	0.036485	1.095116
12228	(Terminator 3: Rise of the Machines)	(The Passion of Joan of Arc, Grill Point)	0.483582	0.131343	0.100000	0.206790	1.574425	0.036485	1.095116
56788	(Terminator 3: Rise of the Machines)	(The Passion of Joan of Arc, The 39 Steps, Rope)	0.483582	0.137313	0.100000	0.206790	1.505972	0.033598	1.087589

(Rain Man, Romeo + Juliet, 48 Hrs.)에 평점을 남긴 사용자는 전부 (Sissi)에도 평점을 남겼음 (confidence = 1)

(Rain Man, 48 Hrs., A Nightmare on Elm Street)에 평점을 남긴 사용자는 거의 모두

(Monsoon Wedding)에도 평점을 남겼음 (confidence = 0.988235)

(3) Lift 값 비교

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
52	(Waiter)	(Muxmäuschenstill)	0.120896	0.156716	0.105970	0.876543	5.593180	0.087024	6.830597
53	(Muxmäuschenstill)	(Waiter)	0.156716	0.120896	0.105970	0.676190	5.593180	0.087024	2.714881
7786	(Solaris, Psycho)	(Big Fish, Terminator 3: Rise of the Machines)	0.134328	0.143284	0.100000	0.744444	5.195602	0.080753	3.352369
7785	(Big Fish, Terminator 3: Rise of the Machines)	(Solaris, Psycho)	0.143284	0.134328	0.100000	0.697917	5.195602	0.080753	2.865672
7762	(Rain Man, Psycho)	(Titanic, Big Fish)	0.131343	0.150746	0.101493	0.772727	5.126013	0.081693	3.736716
...	...	...	...	...	...	...	...	...	...
12248	(Rain Man, Monsoon Wedding, Solaris)	(The Million Dollar Hotel, 48 Hrs.)	0.177612	0.204478	0.108955	0.613445	3.000061	0.072638	2.057982
20071	(Silent Hill, Monsoon Wedding, Terminator 3: R...	(Sissi, The Conversation)	0.177612	0.204478	0.108955	0.613445	3.000061	0.072638	2.057982
12257	(The Million Dollar Hotel, 48 Hrs.)	(Rain Man, Monsoon Wedding, Solaris)	0.204478	0.177612	0.108955	0.532847	3.000061	0.072638	1.760424
20078	(Sissi, The Conversation)	(Silent Hill, Monsoon Wedding, Terminator 3: R...	0.204478	0.177612	0.108955	0.532847	3.000061	0.072638	1.760424
12439	(Sissi, Monsoon Wedding, Terminator 3: Rise of...	(The Hours, 48 Hrs.)	0.200000	0.201493	0.120896	0.604478	3.000000	0.080597	2.018868

(Waiter) 에 평점을 남긴 사용자가 (Muxmäuschenstill)에 평점을 남길 확률에는 양의 상관관계가 있으며, 그 역도 성립함

(lift = 5.593180, 양방향 모두 동일)

Body

Contents

CBR / AR / CF

CF

User-based collaborative filtering : 실습 코드(item based)에 없는 user based로 진행

① Movie metadata, rating, link 데이터를 불러온 후 결합

	movieId	userId	rating	timestamp	imdbId	genres	overview	release_date	title
38811	1552	502	3.0	868217926	118880.0	[{'id': 28, 'name': 'Action'}, {'id': 53, 'nam...	When the government puts all its rotten crimin...	1997-06-01	Con Air
96474	99114	176	4.0	1364721906	1853728.0	[{'id': 18, 'name': 'Drama'}, {'id': 37, 'name...	With the help of a German bounty hunter, a fre...	2012-12-25	Django Unchained
54131	4022	99	3.0	982280624	162222.0	[{'id': 12, 'name': 'Adventure'}, {'id': 18, '...	Chuck, a top international manager for FedEx, ...	2000-12-22	Cast Away
48847	36401	624	3.0	1332087753	355295.0	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	Folklore collectors and con artists, Jake and ...	2005-08-26	The Brothers Grimm
60226	1393	93	3.5	1304992382	116695.0	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	Jerry Maguire used to be a typical sports agen...	1996-12-06	Jerry Maguire
84281	4826	353	3.0	1137410897	80437.0	[{'id': 18, 'name': 'Drama'}, {'id': 10752, 'n...	A veteran sergeant of the World War I leads a ...	1980-05-28	The Big Red One

② Filtered_items라는 함수를 사용하여 특정 등급 미만의 영화 및 사용자를 필터링

	userId	movieId	rating
147	4	10	4.0
148	4	34	5.0
149	4	112	5.0
150	4	141	5.0
151	4	153	4.0
...	...	...	...
99686	665	5479	2.0
99688	665	5502	4.0
99689	665	5679	3.0
99690	665	5952	5.0
99691	665	5991	4.0

61805 rows × 3 columns

③ 사용자를 행으로, 영화를 열로 사용하여 등급 피벗 테이블 제작

movieId	1	2	3	4	5	6	7	9	10	11	...	132046	134130	134853	136864	138036	139385	139644	142488	148626	152081
userId																					
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	2.0	2.0	NaN	NaN	4.5	4.0	NaN	NaN	3.0	2.5	...	0.5	3.5	1.0	3.0	1.0	2.5	3.0	3.5	3.5	3.0
17	NaN	NaN	NaN	NaN	NaN	4.5	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19	3.0	3.0	3.0	3.0	NaN	3.0	3.0	3.0	3.0	3.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 2794 columns

④ 코사인 유사도를 사용하여 지정된 사용자와 유사한 사용자 탐색

⑤ 지정된 사용자가 아직 평가하지 않은 영화에 대해 이러한 유사한 사용자의 평균 평점을 기반으로 영화를 추천

⑥ 추천할 영화의 수인 n과 추천할 사용자의 ID인 userId 입력 → 해당 사용자의 User-based 최고 추천 영화 출력

```
cosine_re_userBased(n=20, userId=119)
```

movieId	0	title
1222	0.770829	Full Metal Jacket
2959	0.761705	Fight Club
1307	0.726349	When Harry Met Sally...
2791	0.724587	Airplane!
2396	0.698609	Shakespeare in Love
1288	0.684089	This Is Spinal Tap
4993	0.682841	The Lord of the Rings: The Fellowship of the Ring
1500	0.673632	Grosse Pointe Blank
150	0.659549	Apollo 13
380	0.650902	True Lies
293	0.633143	Leon: The Professional
4027	0.630034	O Brother, Where Art Thou?
5952	0.625285	The Lord of the Rings: The Two Towers
1380	0.607529	Grease
1393	0.607423	Jerry Maguire
3578	0.602818	Gladiator
2599	0.599981	Election
34	0.598540	Babe
4226	0.597880	Memento
357	0.571655	Four Weddings and a Funeral

User-based collaborative filtering

시청한 영화의 평점을 기반으로 사용자의 유사도 평가

→ 유사한 사용자가 좋게 평가한 영화 추천

Item-based collaborative filtering

user들 간 평점이 비슷한 정도를 기반으로 영화들의 유사도 평가

→ 좋게 평가한 영화와 유사한 영화 추천

Conclusion

Contents

CBR / AR / CF

01. Result

(1) CBR

- categorical target (영화의 흥행 여부), continuous target (영화의 평점) 두 가지 상황에서 분석 진행
- KNN의 최적 파라미터 결정을 위해 weight, algorithm, metric 3가지 parameter 값에 따른 정확도를 비교하였음
- 두 상황 모두 **(weight, algorithm, metric) = (distance, auto, manhattan)** 일 때 가장 좋은 성능을 보임

▶ 영화의 흥행 여부 : k = 32일 때 정확도가 가장 높았음

Accuracy = 0.9841, Precision = 0.9772, Recall = 0.9671, F1 Score = 0.9721

▶ 영화의 평점 : k = 72일 때 MSE가 가장 낮았음

MSE = 3.0544, RMSE = 1.7477, MAE = 1.2220. 그러나 y_true = 0인 케이스에 의해 MAPE 값은 매우 높게 나타났음

(2) AR

- 670명의 사용자가 매긴 403개의 영화에 대해 연관분석을 시행함
- support 값을 통해 어떤 영화가 같이 평점이 남겨진 경우가 많은 지 확인할 수 있었으며 confidence, lift 값을 통해 특정 영화에 평점을 남긴 사용자가 다른 어떤 영화에 평점을 많이 남겼는지 분석할 수 있었음

(3) CF

- 실습에서 다루지 않은 user based collaborative filtering을 사용함
- 영화 평점을 기반으로 사용자 간의 유사도를 평가하여 각 사용자에 대해 유사한 사용자가 좋게 평가한 영화를 추천함

02. Limitation

(1) Data 관련 한계

- 연관 규칙을 수행할 때, 2017년 공개된 Kaggle의 The Movies Dataset을 이용함
- Metadata on over 45,000 movies, 26 million ratings from over 270,000 users 데이터가 존재
- 이 중 일부만을 임의로 선정해 연구를 진행하여 관계를 온전히 파악하지 못함

(2) Parameter 선정의 한계 (KNN Classifier)

- KNN Classifier의 parameter 중, **weights / algorithm / metric** 을 선정할 때, 가능한 조합의 경우의 수가 많아 모두 비교하지 못함
- K 역시 특정 범위 **(1~40 / 1~80)** 내에서만 비교를 수행함

03. Development

(1) Data 관련 개선점

- 데이터의 특성을 왜곡하지 않도록 대표성 있는 Sampling을 수행하거나, 샘플링 과정에서 중요한 변수를 고려하여 샘플링을 수행하는 것이 중요하겠음

(2) Parameter 선정 관련 개선점

- GridSearchCV를 이용하여 최적의 parameter를 찾는다면 더 좋은 성능(높은 accuracy / 낮은 mse)을 얻으리라 기대됨

Data Source

[데이터 출처]

(1) Kaggle (The Movies Dataset)
- https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=ratings_small.csv

[참고 문헌]

(1) Ahn, Shinhyun and Chungkon Shi. “Exploring Movie Recommendation System Using Cultural Metadata.” Transactions on Edutainment (2008).