

영화 추천 시스템 플랫폼 구축

3조

Bit academy Big data Marketing Team

민 병혁 최 현수

권 태양

목차

01

기획 단계

02

SYSTEM 구성

03

분석 및 구현

1. 기획 단계

NAVER 영화

영화 추천, 상영작, 예정작, 영화평점, 해설, **영화 리뷰**, 네이버 영화, 네이버 리뷰, 다운로드, 인기극장

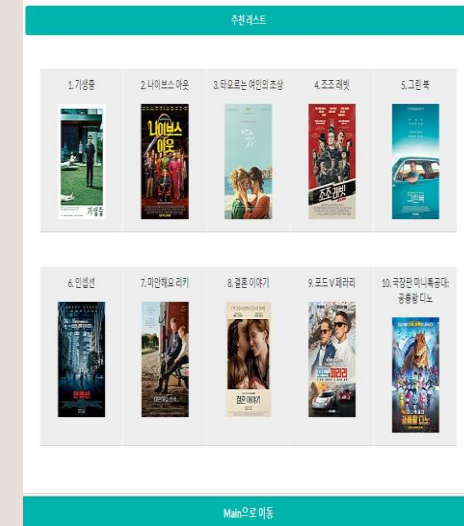
현재 리스트 • 총 12,893,198개의 항목이 있습니다.

번호	장르	공감	공감비율
36790385	경	★★★★★ 1	21세기에 아직도 달강행하러나... 신곡
36790284	경	★★★★★ 9	결혼 이야기 ★★★★★ 9 단독부의 주영이가 연가기 때문 출몰하다. 세상에 대한 갈증이 어떻게 고향에 고 어떻게 변화하는지에 대한 묘사가 너무 잘 드러나서 음 놀랐다. 훌륭한 배우 들이 등장한 시나리오를 탄탄 느낌이라서 편안하고 지루스럽지 않음 볼 수 있었다. 신곡
36790383	울	★★★★★ 9	올리버 트위스트 ★★★★★ 9 간단한 분위기에서 감동적인 스토리였습니다. 좋습니다. 신곡
36790281	인	★★★★★ 10	인생은 ★★★★★ 10 미 보다가 나온 영화가 있는가? 신곡
36790380	캐	★★★★★ 10	캐이프 피어 ★★★★★ 10 신곡
36790279	비	★★★★★ 2	비즈 오브 프레스(알다 칸의 황금한 해방) ★★★★★ 2 영화본사들이비밀하게주요한일까요? 주작이들이 남우최고 신곡
36790278	해	★★★★★ 1	해치지언어 ★★★★★ 1 이게뭐...7점이거... 신곡
36790277	비	★★★★★ 2	비즈 오브 프레스(알다 칸의 황금한 해방) ★★★★★ 2 웃음 붙어가는지 재미있었다 신곡
36790276	남	★★★★★ 7	남산의 부장들 ★★★★★ 7 후반 10분을 위한 2시간 애의 러닝타임이었지만 후반 10분이 훌륭하게했다 신 고
36790274	미	★★★★★ 5	미드소마 ★★★★★ 5 시작부분이 길어서 마지막에 뭔가 상상 이상의 것이 있을 줄 알았는데 그저 잔 잔하게 흘러갔다. 보면서 기분이 재미있었고 정신도나 사이버물론 인간이 아닐 줄 깨달았다. 신곡

네이버 영화
네티즌 평점



네이버 영화
현재 상영작



영화 추천 시스템

네티즌 평점에 의한 현재 상영작 사용자 별 추천 시스템 구현

2. SYSTEM구성



Web Crawling , Analysis



Operation Data 적재



Web Service View

작업 별 조원 간 역할 분담제 진행

Bit Academy Big data Marketing Team

3. 구현 과정 (1) 데이터 수집 단계

16700359	남산의 부장들 ★★★★★ 10 이 영화가 재미없다고 하는 사람들은역사적인 배경지식이 제로에 가깝다고 본 다안타깝다 젊은이들 신고	prec**** 20.02.11
16700358	결혼 이야기 ★★★★★ 8 여기서의 교훈은 부인과 남편은 분야가 겹치면안된다. ㅋㅋ특히 예술가끼리의 만남은 OO맞다.내가 분야겹치는 남자 만나기 싫어했던이유. 신고	kenj**** 20.02.11

```
# 영화 전체 리스트
from urllib.request import urlopen
from bs4 import BeautifulSoup

Main_url = "https://movie.naver.com/movie/running/current.nhn"
html = urlopen(Main_url)
main = BeautifulSoup(html, "html.parser", from_encoding="utf-8")
```



16,675,535	부산행	8	edwa****
16,675,537	붉은 돼지	10	fbtj****
16,675,538	남산의 부장들	10	p273****
16,675,539	월요일이 사라졌다	10	wlgn****
16,675,540	양자물리학	10	curb****
16,675,541	아포칼립토	10	kimk****
16,675,542	아포칼립토	10	koop****
16,675,543	남산의 부장들	9	hyuk****

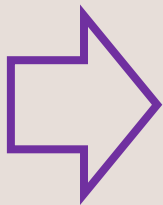
영화 제목 , 평점 , 사용자 이름에 대한 Crawling Code 작성

댓글 등록 순서를 Primary Key로 설정 후 영화, 평점, 유저 순으로 DB 적재

3. 구현 과정 (2) 데이터 전처리

		0	1	2
0	미운 오리 새끼	10	pig*****	
1	부산행	8	edw*****	
2	붉은 돼지	10	fbt*****	
3	남산의 부장들	10	p27*****	
4	월요일이 사라졌다	10	wlg*****	
...	
22732	터미네이터: 다크 페이트	6	sma*****	
22733	조조 래빗	10	ytm*****	
22734	기생충	9	017*****	
22735	인셉션	10	who*****	
22736	카센타	8	hia*****	

22737 rows x 3 columns

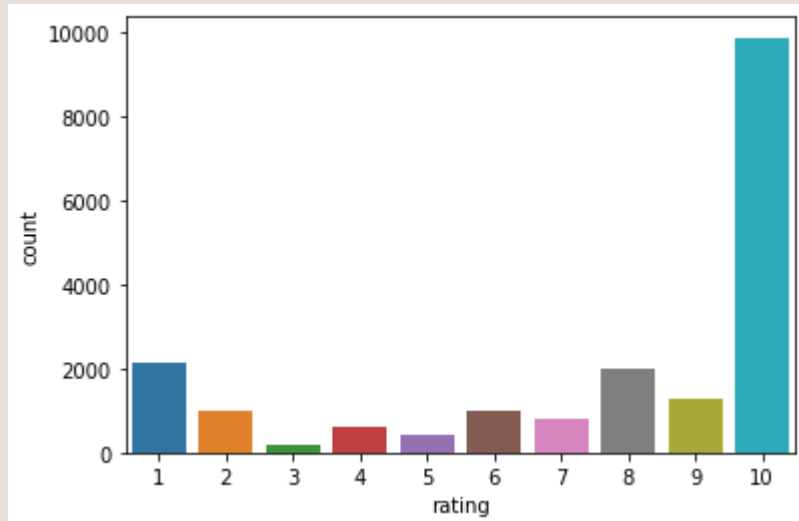


	userId	movieId	rating
0	pig	미운 오리 새끼	10
1	edw	부산행	8
2	fbt	붉은 돼지	10
3	p27	남산의 부장들	10
4	wlg	월요일이 사라졌다	10
...
22732	sma	터미네이터: 다크 페이트	6
22733	ytm	조조 래빗	10
22734	017	기생충	9
22735	who	인셉션	10
22736	hia	카센타	8

19221 rows x 3 columns

사용자 명이 일부 암호화되어 앞 3글자 가 동일시 한 사용자로 칭함.
이때 중복 값이 생기므로 처음 값을 제외한 나머지 값 제거.
Ex) pig*** / pig***** => pig

3. 구현 과정 (3) 데이터 탐색



- 주로 10점을 준 사용자가 대부분
- 다음으로 1점을 준 사용자가 많다.

```
len(df.userId.unique())
```

```
6224
```

사용자의 수 : 6224

```
len(df.moviId.unique())
```

```
2608
```

영화 수 : 2608

```
len(df) / (len(df.userId.unique()) * len(df.moviId.unique()))
```

```
0.0011841284282492468
```

데이터 비율 : 데이터 수 / (유저 수 * 영화 수)

- 0.001 이므로 희소행렬임을 알 수 있다.

3. 구현 과정 (4) 분석 - 모델 선택

1) SVD

- 특이 값 분해 알고리즘
- 희소 행렬에 적합
- 사용자, 영화의 중요 특징을 이용해 예측하는 알고리즘

2) Baseline Only

- User와 Item의 Baseline을 이용한 평점 예측 알고리즘

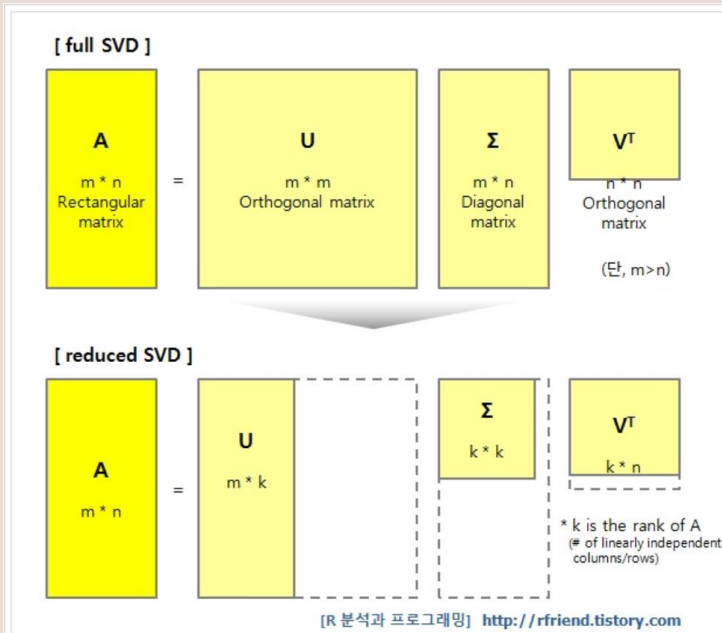
3) KNN Basic

- 사용자와 유사한 k명의 평점 데이터를 이용한 평점 예측 알고리즘

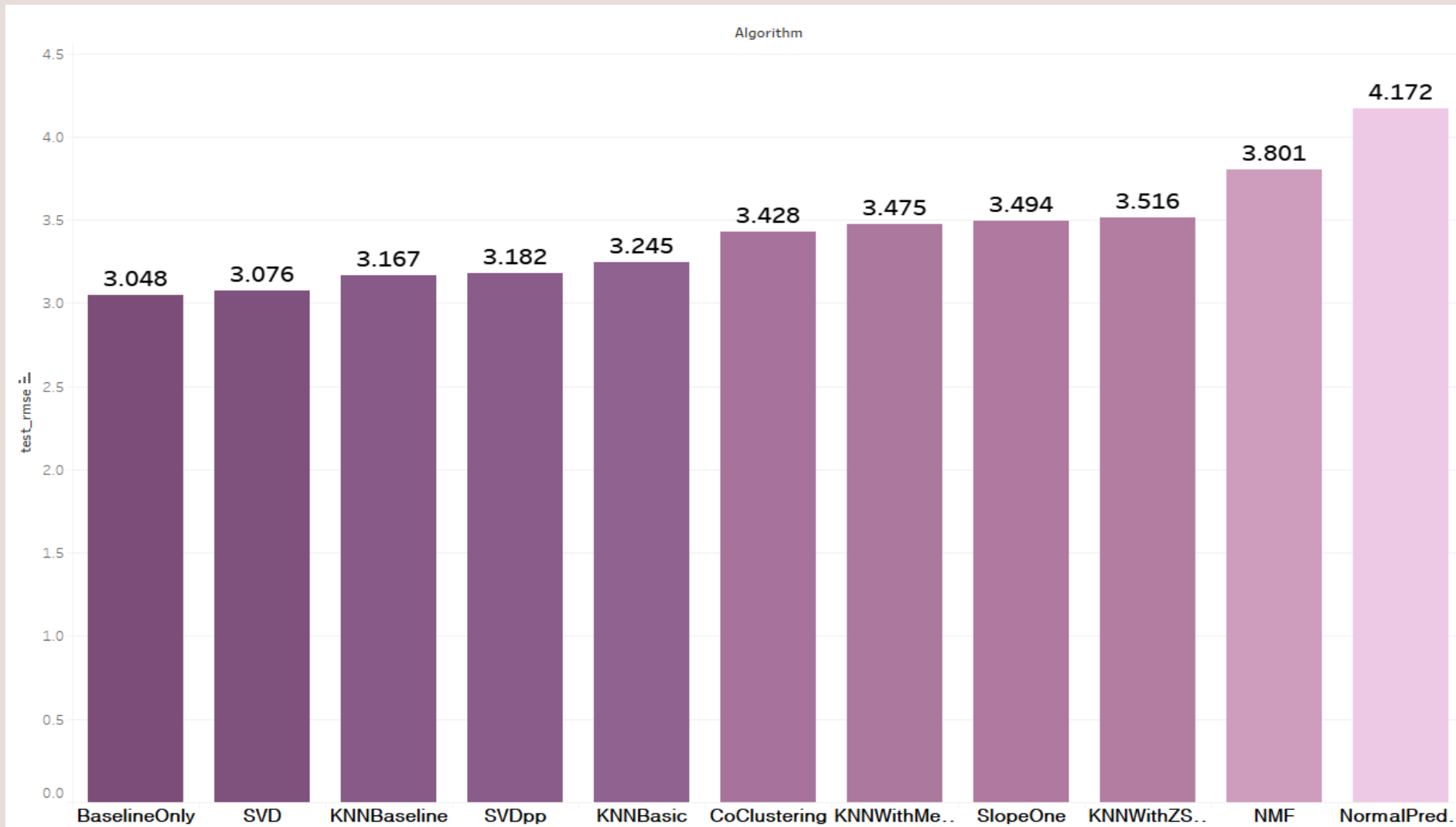
$$\hat{r}_{ui} = \mu + b_u + b_i$$

사용자 u가 I 라는 영화의 예측 값
= 평점 데이터 전체 평균
+ u라는 유저의 편향
+ i라는 영화의 편향

1. 사용자와 유사한 k명을 찾는다(사용자 별 유사도 비교 - default : msd)
2. K명의 유사 사용자들의 영화 평점의 가중합을 하고 정규화(유사도가 높을수록 가중치 증가)



3. 구현 과정 (4) 분석 - 모델 성능 평가



- 모델성능평가 : RMSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

Sum((예측값 - 실제값)^2)
의 값을 데이터 수로 나누고
루트를 씌운 값
:평균 오차에 대한 추정 값

RMSE를 기준으로 값이 낮은
두개의 모델 선택

- BaselineOnly
- SVD

3. 구현 과정 (4) 분석 - 모델 별 최적 파라미터 찾기

- BaselineOnly Default model RMSE : 3.048

```
# BaselineOnly 최적 파라미터 찾기
# defaults : als, 10, 15, 10
BaselineOnly_param_grid = {'bsl_options': {'method': ['als'],
                                             'n_epochs': [5,7,10],
                                             'reg_u': [15,20,30,40],
                                             'reg_i': [5,7,10]
                                             }}

```

After



```
print(baselineonly.best_score['rmse'])
print(baselineonly.best_params['rmse'])

3.041295344607057
{'bsl_options': {'method': 'als', 'n_epochs': 5, 'reg_u': 15, 'reg_i': 5}}
```

- SVD Default model RMSE : 3.076

```
# SVD 최적 파라미터 찾기
# defaults : 20, 100, 0.005, 0.02
svd_param_grid = {'n_epochs': [10, 20, 30],
                  'n_factors': [80,100,120],
                  'lr_all': [0.002, 0.005, 0.007],
                  'reg_all': [0.02, 0.05, 0.1, 0.15]}

```

After



```
print(svd.best_score['rmse'])
print(svd.best_params['rmse'])

3.0691171868356366
{'n_epochs': 20, 'n_factors': 80, 'lr_all': 0.005, 'reg_all': 0.1}
```

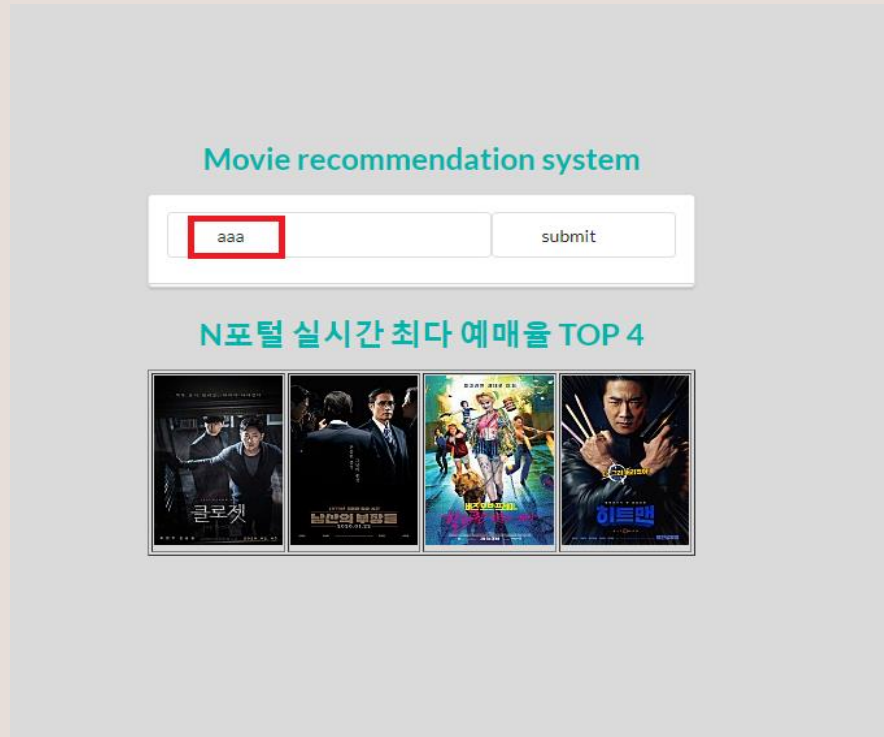
=> 최종 모델은 RMSE가 가장 낮은 값 3.041인 Baseline Only 선택(파라미터는 위와 동일)

3. 구현 과정 (4) 분석 - 모델 학습 및 산출 결과

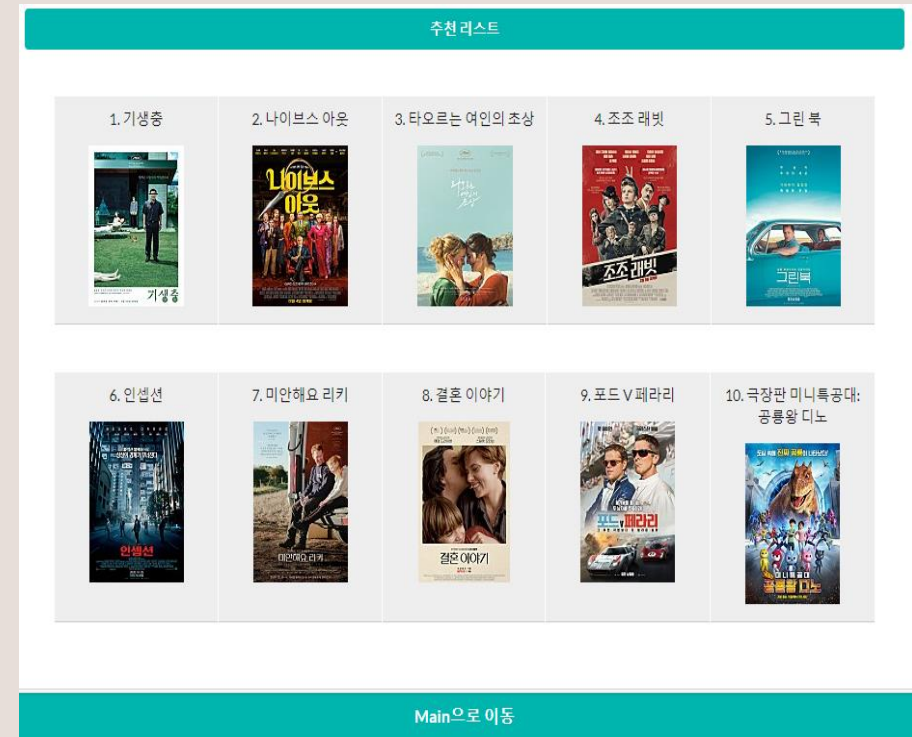
기생충 예상 평점 : 10.000	포드 V 페라리 예상 평점 : 9.657	피아니스트의 전설 예상 평점 : 8.998	핑크퐁 시내마 콘서트: 우주대탐험 예상 평점 : 8.609	리틀 큐 예상 평점 : 8.300	러브레터 예상 평점 : 8.294	경계선 예상 평점 : 8.253	카잔자키스 예상 평점 : 8.248	페인 앤 글로리 예상 평점 : 8.214	에릭 클렘튼: 기타의 신 예상 평점 : 8.210	82년생 김지영 예상 평점 : 8.185
나이트비트 아웃 예상 평점 : 10.000	극장판 미니특공대: 공룡왕 디노 예상 평점 : 9.584	파바로티 예상 평점 : 8.912	성혜의 나라 예상 평점 : 8.582							
타오르는 여인의 초상 예상 평점 : 10.000	스파이 지니어스 예상 평점 : 9.462	울지마 몬즈 2: 슈크란 바바 예상 평점 : 8.818	두 교황 예상 평점 : 8.574	메기 예상 평점 : 8.159	호름 예상 평점 : 8.060	아이리시맨 예상 평점 : 8.058	러브 라이브! 더 스쿨 아이돌 무비 예상 평점 : 8.014	썩시드 예상 평점 : 8.004	사마에게 예상 평점 : 7.895	조커 예상 평점 : 7.888
		그녀에게 예상 평점 : 8.152								
조조 래빗 예상 평점 : 9.955		벌룬 예상 평점 : 9.370	윤희에게 예상 평점 : 8.790	미드웨이 예상 평점 : 8.508	겨울왕국 2 예상 평점 : 8.141	나쁜 녀석들: 포에버 예상 평점 : 7.880	가장 따뜻한 색, 블루 예상 평점 : 7.732	이태원 예상 평점 : 7.731	공기인형 예상 평점 : 7.688	신의 은총으로 예상 평점 : 7.498
그린 북 예상 평점 : 9.938		하이큐!! 땅 VS 하늘 예상 평점 : 9.344	흑소 고지 예상 평점 : 8.723	울지마 몬즈 예상 평점 : 8.388						
인생선 예상 평점 : 9.787		이 멋진 세계에 축복을! 붉은 전설 예상 평점 : 9.299	날씨의 아이 예상 평점 : 8.681	잃어버린 세계를 찾아서 예상 평점 : 8.356	파비안느에 관한 진실 예상 평점 : 8.130	로마 예상 평점 : 7.851	버즈 오브 프레이(할리 권의 황홀한 해방) 예상 평점 : 7.320	목적자: 눈이 없는 아이 예상 평점 : 6.447	나이트 헌터 예상 평점 : 6.296	산상수호 예상 평점 : 6.020
미안해요, 리키 예상 평점 : 9.762		고흐, 영원의 문에서 예상 평점 : 9.183	벌새 예상 평점 : 8.630	21 브릿지: 테러 섯다운 예상 평점 : 8.334	작은 빛 예상 평점 : 8.103		원스 어폰 어 타임... 인 할리우드 예상 평점 : 7.787			
겉은 이야기 예상 평점 : 9.699		천문: 하늘에 묻는다 예상 평점 : 9.015	신비아파트 극장판 하늘도깨비 대 요르모간드 예상 평점 : 8.623	아이 인 더 스카이 예상 평점 : 8.301	기억할 만한 지나침 예상 평점 : 8.074	남산의 무장들 예상 평점 : 7.743	닥터 두리들 예상 평점 : 6.647	해치지않아 예상 평점 : 5.238	백두산 예상 평점 : 4.959	미스터 주: 사라진 VIP 예상 평점 : 4.418
							히트맨 예상 평점 : 6.583			

- 모델 학습 후 현재 상영작들에 대한 예측 평점(사용자 : 'aaa')
- 가장 높은 평점인 왼쪽 위에서부터 가장 낮은 평점인 오른쪽 아래까지 내림차순 정렬
- 왼쪽 위에서부터 상위 10개를 추천

3. 구현 과정 (5) Web 화면 출력



Main 화면(사용자 입력 화면)



사용자 'aaa'를 위한 현재 상영작 10개 추천

감사합니다.

코드 참고

https://github.com/sunnight9507/Bit_Academy/tree/master/12%EC%A3%BC%EC%B0%A8/mini_project

Bit Academy Big data Marketing Team