TEXT MINING

영화 리뷰를 통한 감성 분석 및
이를 기반한 비 흥행 명작 파악 서비스

CONTENTS

1. 개발 동기

1. 개발 동기

'어떤 영화 볼까?'

01

시발점

'평점 높아서 봤더니 영화 재미없어'

02

문제점

영화에 대한 네티즌 평점(별점)과
일치하지 않는 영화 내용의 재미도

03

문제점

영화에 대한 네티즌
평점(리뷰 작성)의 방대한 양

04

감성 분석

영화에 대해 작성된 리뷰를 이용해
영화의 긍정적 및 부정적 요소 분석

05

감성 분석

감성 분석(Sentiment Analysis) 기술을
이용해 더욱 자세한 영화 흥행에 대한
데이터 파악

문제점

출처: Naver, Daum

[SC초점]'82년생 김지영' 오늘 개봉, **별점테러**X악플 뚫고 '**영화**의 힘'으로 정...

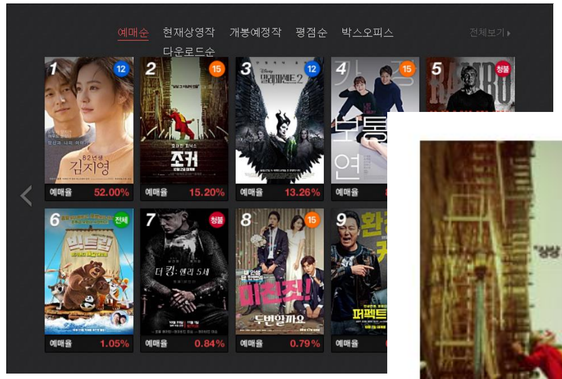
스포츠조선 | 12시간 전 | 네이버뉴스

하지만 **영화**는 그를 못된 사람이나 악당으로 그리지 않는다. 다만 따뜻하고 좋은 아빠이지만 싶었지만 무심했던 시대와 세대를 살아온 사람의 표상으로 그려질 뿐이다. 악플과 **별점테러**까지, 개봉 전부터 이해하기 힘든...

[리뷰] **별점 테러**와 N차 관람 사이, **영화**로 본 '82년생 김지영'

이로운넷 | 4일 전

한편에서는 **별점 테러**와 험오 댓글이 이어질 것이고, 다른 한편에서는 같은 **영화**를 보고 또 보는 'N차 관람'은 물론, 극장에 가지 않고도 티켓을 구매해 응원하는 '영혼 보내기' 운동을 펼칠 계획인 듯하다. 소설에 이어...



조커 (2019)

Joker

★★★★☆ 7.9/10

스릴러 | 미국, 캐나다

2019.10.02 개봉 | 123분, 15세이상관람가

(감독) 토드 필립스

(주연) 호아킨 피닉스

예매 2위 | 누적관객 4,642,360명

예매하기

minho ★☆☆☆☆ 1/10

재미없다. 반은 졸았다. 미친놈 연기하면 미친듯한 연기냐. 내 눈에는 미친 황제 코모두스만 보였다.

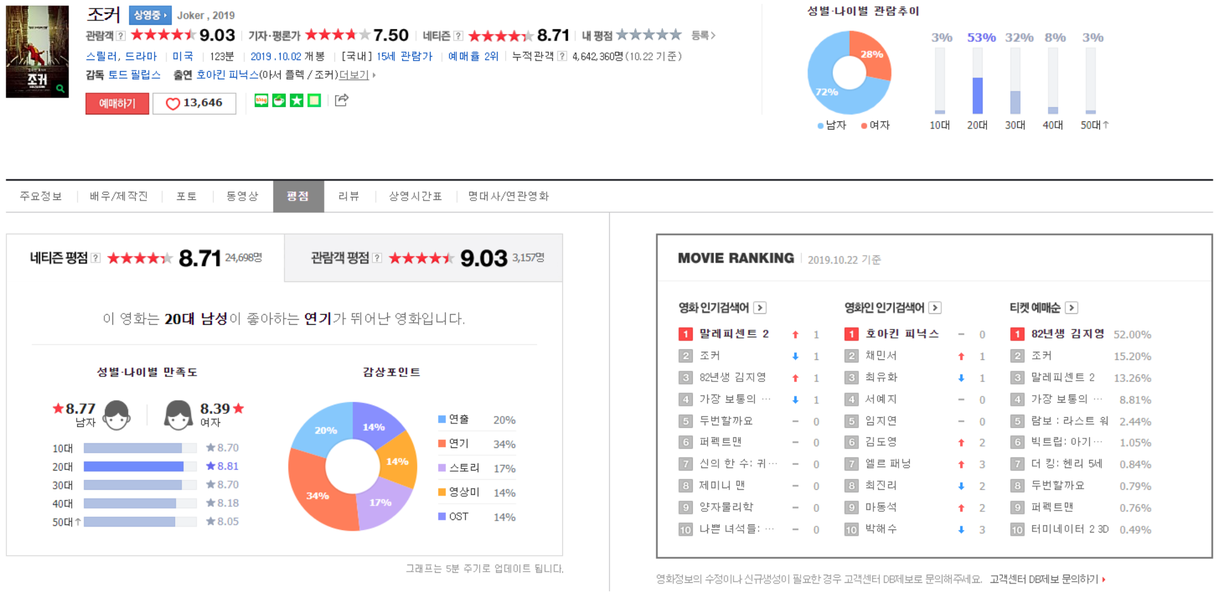
영화의 주된 평점(별점)과 상반되는 흥행 순위와 네티즌 리뷰

A. 네티즌이 주로 별점을 통해 영화를 선택하게 되는데 이런 현상은 영화 선택에 어려움을 줌

B. 사용자는 영화 내용 파악에 어려움을 겪음

단순 평점과 방대한 양의 리뷰 결과

출처: Naver 영화, Daum 영화



Naver 영화 '조커' 평점

Daum 영화 '조커' 평점

리뷰를 통해 감성 분석 기술 사용

출처: Naver 영화

★★★★★ 10 착하게 사는것은 높은 계단을 오르는것과 같지만 포기하고 내려갈때는 너무나도 빠르고 즐겁다.

버블껌(roac****) | 2019.10.02 12:54 | 신고

★★★★★ 10 하여간 역대 조커들은 너무 완벽해. 시저 로메로, 잭니콜슨, 로명수, 호아킨 피닉스..

꼬망스(jang****) | 2019.10.02 10:27 | 신고

★★★★★ 10 명작들만 골라서 번역하는 박지훈이야말로 이시대의 조커 아닐

김민수(msms****) | 2019.10.02 11:51 | 신고

★★★★★ 10 조커 분장 후 계단씬 지린다..

포동잉(epod****) | 2019.10.02 09:02 | 신고

★★★★★ 10 마블은 10년간 공들여 만든 "타노스"로 관객을 열광시키고 DC· 하나로 다시 부활한다.

엿먹어(cheo****) | 2019.10.02 14:23 | 신고

★☆☆☆☆ 1 볼거리는 주인공의 연기뿐인 영화. 조커는 없고, 호아킨 피닉스만이 남았다.

140(wow_****) | 2019.10.23 21:18 | 신고

0 0

★☆☆☆☆ 1 보다 정신병 걸리는 줄... 피해자라고 다 저러지는 않음

김아이(tjdm****) | 2019.10.23 19:19 | 신고

8 8

★☆☆☆☆ 1 스토리 너무 노잼이었다 결말도 별로고 돈이 아까웠다

lololol(play****) | 2019.10.23 19:18 | 신고

9 7

★☆☆☆☆ 1 너무 우울해서 ㅂㄹ 그냥저냥 ㅋ 내 스타일은 아니었음

리누누(kiuu****) | 2019.10.23 15:03 | 신고

8 8

★☆☆☆☆ 1 멋있는지도 모르고 이입도 잘 안 됩니다... 성폭력에 뿌리를 둔 영화라 그런가..

이주아(slow****) | 2019.10.23 14:38 | 신고

5 16

★☆☆☆☆ 1 성범죄자 OUT ! 낯짝도 두껍다. !

8383(yjhw****) | 2019.10.23 13:54 | 신고

2 8

★☆☆☆☆ 1 미안하지만 호아킨은 글래디에이터의 코모두스 캐릭터에서 한발짝에서 겨우 두세발 정도 나아갔다. 애정결핍 울분의 그 코모두스 연기. 아버지 황제를 죽이던그 모습 그대로..평해 이동 된 캐릭터 그대로였다.

아구아인(asra****) | 2019.10.23 00:34 | 신고

2 19

- 평점만으로는 알 수 없는 영화에 대한 평가를 감성 분석을 이용해 평가
- 이러한 점을 이용하여 텍스트 마이닝(Text Mining)
  감성 분석(Sentiment Analysis) 기술 사용해
  텍스트 추출 및 분석
- 해당 정보로 영화 흥행 파악 서비스 제공으로 응용

2. 데이터 수집 및 전처리, 군집화

크롤링을 통한 데이터 수집-1

분류 기준					독립변수					종속변수
A	B	C	D	E	F	G	H	I	J	K
영화코드	영화명	개봉일	대표국적	장르	관람객평점	평론가평점	네티즌평점	남자평점	여자평점	관객수
167697	신과함께-인과 연	180801	한국	모험	0.863	0.631	0.771	6.17	8.29	12274996
136315	어벤져스: 인피니티 워	180425	미국	액션	0.908	0.709	0.895	8.79	9.12	11212710
156464	보헤미안 랩소디	181031	미국	드라마	0.945	0.614	0.942	9.25	9.63	9224582
154222	미션 임파서블: 폴아웃	180725	미국	액션	0.915	0.756	0.912	8.98	9.37	6584915
85579	신과함께-죄와 벌	171220	한국	모험	0.873	0.592	0.783	6.4	7.9	5872007
154285	쥬라기 월드: 폴른 킹덤	180606	미국	액션	0.854	0.64	0.801	6.99	8.46	5661128
144330	앤트맨과 와스프	180704	미국	액션	0.885	0.613	0.862	8.08	8.92	5448134
163533	안시성	180919	한국	액션	0.863	0.622	0.799	6.72	8.91	5440186
137326	블랙 팬서	180214	미국	액션	0.833	0.667	0.749	6.56	7.87	5399227
167638	완벽한 타인	181031	한국	드라마	0.908	0.613	0.862	8.4	8.22	5293435
158191	1987	171227	한국	드라마	0.931	0.808	0.922	8.9	9.62	5290310
158178	독전	180522	한국	액션	0.841	0.529	0.753	6.34	8.17	5063684
153687	공작	180808	한국	드라마	0.786	0.693	0.684	6.16	6.96	4974520
119428	베놈	181003	미국	액션	0.824	0.46	0.79	7.38	8.39	3888096
167105	암수살인	181003	한국	액션	0.858	0.814	0.832	7.95	8.34	3789321
149236	데드풀 2	180516	미국	액션	0.909	0.586	0.869	8.12	8.56	3784602
164192	국가부도의 날	181128	한국	드라마	0.87	0.65	0.814	7.42	8.92	3747952
151728	코코	180111	미국	모험	0.92	0.8	0.926	9.19	9.51	3510017
151153	아쿠아맨	181219	미국	액션	0.877	0.683	0.835	8.13	8.38	3491857
158180	그것만이 내 세상	180117	한국	드라마	0.917	0.52	0.891	8.66	8.96	3419339
175322	마녀	180627	한국	공포	0.857	0.563	0.818	8.05	8.19	3189091
159892	탐정: 리턴즈	180613	한국	드라마	0.899	0.6	0.84	7.34	8.79	3152872
136990	인크레더블 2	180718	미국	모험	0.932	0.733	0.923	8.69	9.46	3033052
172425	서치	180829	미국	드라마	0.896	0.76	0.892	8.8	9.07	2949944
140652	너의 결혼식	180822	한국	멜로/로맨	0.9	0.657	0.84	8.06	8.23	2820969
172454	곤지암	180328	한국	공포	0.751	0.633	0.641	5.88	6.83	2675575
168298	지금 만나러 갑니다	180314	한국	멜로/로맨	0.898	0.55	0.872	8.35	9.05	2602273
169015	목격자	180815	한국	공포	0.756	0.525	0.483	3.51	5.44	2524720
165748	조선명탐정: 흡혈괴마의 비밀	180208	한국	모험	0.752	0.517	0.663	6.04	6.88	2444136
154255	신비한 동물들과 그린델왈드의	181114	미국	모험	0.746	0.533	0.616	5.1	6.35	2414062
149248	메이즈 러너: 데스 큐어	180117	미국	액션	0.815	0.5	0.784	6.84	8.81	2299732
164115	맘마미아!2	180808	미국	드라마	0.908	0.62	0.896	8.42	9.19	2293884
136898	레디 플레이어 원	180328	미국	액션	0.872	0.836	0.853	8.41	8.62	2254430
162981	명당	180919	한국	드라마	0.775	0.58	0.705	5.95	7.68	2087474

- 군집화를 하기 위한 분류 기준 **영화 코드, 영화명, 개봉일 대표 국적**
- **종속변수**에 영향을 미치는 **독립변수 : 5가지의 평점**
- **독립변수**에 의해 영향을 받는 **종속변수 : 관객 수**

데이터 전처리

01

장르 군집화

29가지의 장르를
각각 5개의 장르로 통일
공포 - 드라마 - 액션 - 모험 - 로맨스

02

독립변수 정규화

평점의 양식이 각각 다름
최소 - 최대 정규화 사용

03

종속변수의 범위

독립영화 혹은 재개봉 영화의 경우
평점, 리뷰의 영향이 적음 따라서
관객 수 5만 명 이하의 영화를 절삭

크롤링을 통한 데이터 수집-2

출처: Naver 영화 , 한국영화진흥원

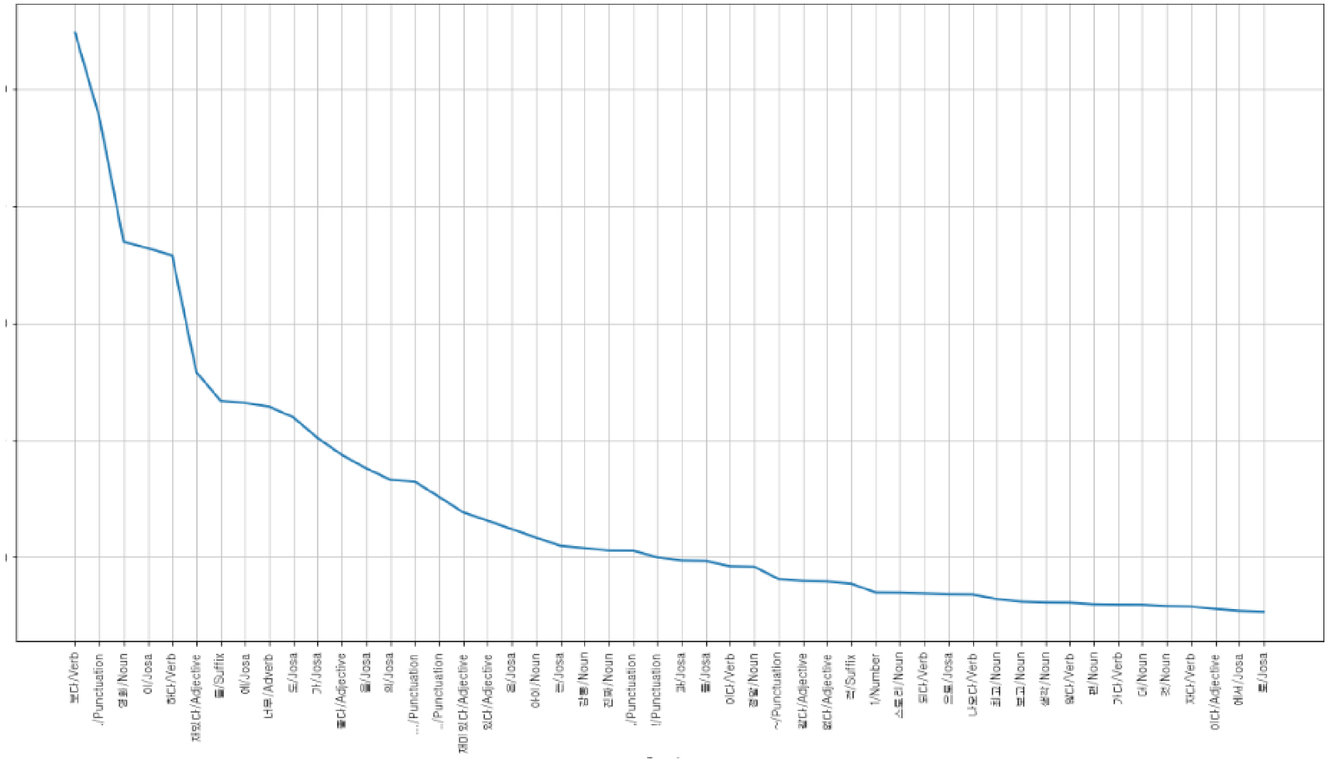
number	text	score
0	전편에서 김동욱이 터졌다면 이번편은 주지훈이 터진다	1
1	나만 좋았나....?	1
2	주지훈 ..터진다잉 앞머리 내렸을때 심쿵 ㅋㅋㅋ내용은 약간 여러 내용을 담을려고 너...	1
3	내가봣을땐 1점주는애들이 알바같은데	1
5	용서의 중요성 재강조_용서를 구하는 자에게 용기가 필요하고 , 용서를 해 주는 자에...	1
6	영화안본애들이꼭 저러더라—— 오전에 조조로봤는데 진짜 개꿀잼ㅇㅇ감동에 반전 소름인데...	1
7	대체 중간에 랩터들이랑 공룡들은 외 나온거지?	0
8	진짜 재미있게 봤는디... 댓글 알바 거리는데 여긴 거진뭐 미션 댓글 알바? 우리나...	1
9	영화 안본사람은 댓글달게하지마라.	1
11	어제 시사회 봤는데 개인적으로 1편보다는 더 여운있고 하정우연기도 좋았어요 다음편에...	1
12	공룡씬은 누가봐도 좀 많이 띠용스러웠다...너무 쥬라기월드(공원) 씬을 패러디?한게...	0
13	우리나라 흥행법칙=스크린독점	0
14	영화관 가면 진심 이것밖에 없어요 ㅜ 독과점 너무 심한 것 같아요 같은 배급산데 ;...	0
15	이 영화를 끝까지 본 내가 귀인이다.	0
16	그냥...스토리 탄탄하게 짰다는 생각이 듦.재판과정도 1편에서 짧게보여줬던 지옥들로...	1
18	재밌으면 좋아요좀 눌러줘요 보러가겡	1
20	1편에서도 초반에 안본사람들이 욕하고 평점 바닥치다가 진짜 본사람들이 평점올려놨는데...	1
21	영화 뒷 부분이 다 읽힘 내가 작가가 된 기분 시원한맛도 없음	0
22	왜 악플이 이리 많지 방금 남친이랑 보구 왔는데 난 잼잇게 봣눈뎅 남친두 넘 잼잇엇...	1
23	좀 뜬금없는게 많음 근데 나름 재밌게 봤는데..3차사 과거에서 뒤통수 7번 맞은 느낌	1
25	여기도 알바들 오지게있넼ㅋㅋㅋ 1탄에선 그나마 신파억지감동이 있어서 눈물샘이라도 자...	0
26	독점좀 하지마라 꼴보기싫다	0
27	자 미션임파서블 몰아웃 상영관 다시 늘려주세요	0
28	참 댓글 알바들 어이가 없네~ 영화 러닝타임이 141분인데 댓글 단 시간이 9시 ...	0
29	——네이버 평점 믿지 말라고 다는 댓글———오늘 조조9시꺼로 보고왔는데 주관적인 생...	1

- 장르별로 분류 후 크롤링 작업 진행
  본 슬라이드에서는 모험 장르 감성 분석을 진행

- 네이버 영화 평점 및 작성 글을 크롤링
  - 평점 기준 9점 이상 : 긍정적인 반응 (1)
  - 평점 기준 4점 이하 : 부정적인 반응 (0)

크롤링을 통한 데이터 수집-3

출처: Naver 영화 , 한국영화진흥원



- 리뷰 크롤링을 통해 점수 부여한 Data 토큰화
- KoNLPy를 이용하여 품사를 매칭
- 자주 등장하는 단어 50개의 리스트
  (분석 시 5,000개의 토큰을 이용)

3. 단순 회귀 분석 및 감성 분석

단순 회귀 분석 -1

출처: Naver 영화 , 한국영화진흥원

In [1]:
```
import pandas as pd
from pandas import DataFrame as df
csv_data = pd.read_csv("2016to2018Movie.csv",encoding='euc-kr')
```

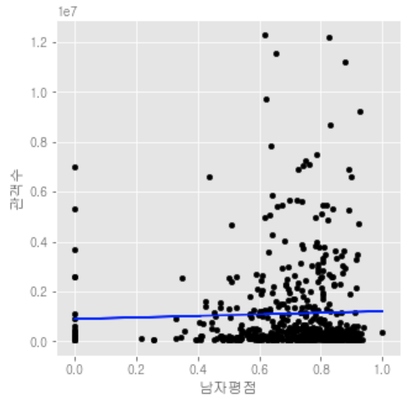
In [2]:
```
csv_data.head()
```

Out [2]:

	영화코드	영화명	개봉일	대표국적	장르	관람객평점	평론가평점	네티즌평점	남자평점	여자평점	관객수
0	167697	신과함께-인과 연	180801	한국	모험	0.875	0.631	0.809	0.617	0.839	12274996
1	136315	어벤져스: 인피니티 워	180425	미국	액션	0.921	0.709	0.939	0.879	0.923	11212710
2	156464	보헤미안 랩소디	181031	미국	드라마	0.958	0.614	0.988	0.925	0.975	9224582
3	154222	미션 임파서블: 폴아웃	180725	미국	액션	0.928	0.756	0.957	0.898	0.948	6584915
4	85579	신과함께-죄와 벌	171220	한국	모험	0.885	0.592	0.822	0.640	0.800	5872007

:
```
font_fname = 'c:/windows/fonts/gulim.ttc'
font_name = font_manager.FontProperties(fname=font_fname).get_name()
rc('font', family=font_name)
csv_data.plot(kind = "scatter", x = '남자평점', y = '관객수', figsize = (5,5), color = "black")

plt.plot(csv_data["남자평점"],prediction,color="blue")
```

:
```
[<matplotlib.lines.Line2D at 0x23b5c02f828>]
```



- 전처리 과정에서 준비했던 테이블을 불러와 train_set , test_set 분류 (150,000/50,000)
- JSON파일로 읽어 train_set의 토큰을 가져옴
- 문서 탐색용 nltk 라이브러리 사용

단순 회귀 분석 -2

출처: Naver 영화 , 한국영화진흥원

In [1]:
```
import pandas as pd
from pandas import DataFrame as df
csv_data = pd.read_csv("2016to2018Movie.csv",encoding='euc-kr')
```

In [2]:
```
csv_data.head()
```

Out [2]:

	영화코드	영화명	개봉일	대표국적	장르	관람객평점	평론가평점	네티즌평점	남자평점	여자평점	관객수
0	167697	신과함께-인과 연	180801	한국	모험	0.875	0.631	0.809	0.617	0.839	12274996
1	136315	어벤져스: 인피니티 워	180425	미국	액션	0.921	0.709	0.939	0.879	0.923	11212710
2	156464	보헤미안 랩소디	181031	미국	드라마	0.958	0.614	0.988	0.925	0.975	9224582
3	154222	미션 임파서블: 폴아웃	180725	미국	액션	0.928	0.756	0.957	0.898	0.948	6584915
4	85579	신과함께-죄와 벌	171220	한국	모험	0.885	0.592	0.822	0.640	0.800	5872007

```
linear_regression = linear_model.LinearRegression()
linear_regression.fit(X=pd.DataFrame(csv_data["관람객평점"]), y = csv_data["관객수"])
prediction = linear_regression.predict(X = pd.DataFrame(csv_data["관람객평점"]))
print('a value = ', linear_regression.intercept_)
print('b balue = ', linear_regression.coef_)
```

```
a value =  499828.2172515411
b balue =  [797218.5517032]
```

```
residuals = csv_data["관객수"] - prediction
residuals.describe()
```

```
count    5.410000e+02
mean     1.484780e-10
std      1.838609e+06
min     -1.237592e+06
25%     -1.038310e+06
50%     -7.091591e+05
75%      1.449092e+05
max      1.107760e+07
Name: 관객수, dtype: float64
```

```
SSE = (residuals**2).sum()
SST = ((csv_data["관객수"]-csv_data["관객수"].mean())**2).sum()
R_squared = 1 - (SSE/SST)
print('R_squared = ', R_squared)
```

```
R_squared =  0.011625552765804 57
```

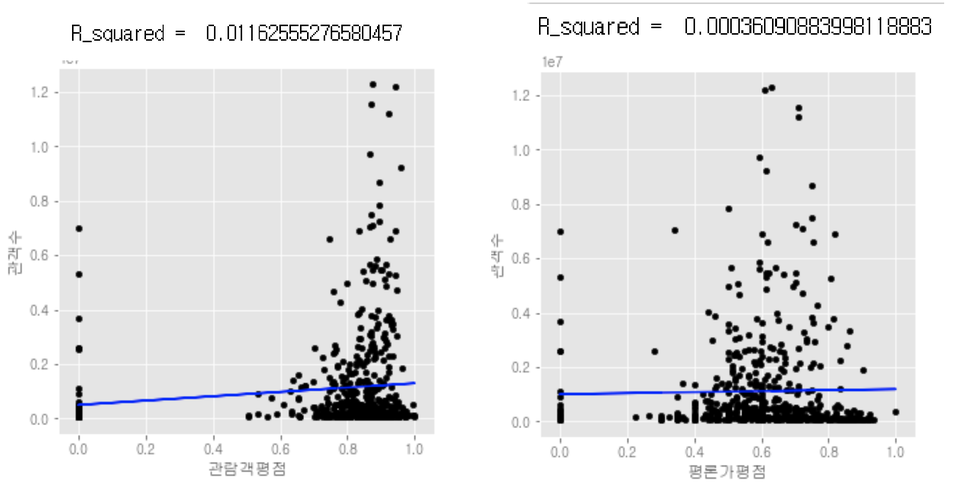
```
font_fname = 'c:/windows/fonts/gulim.ttc'
font_name = font_manager.FontProperties(fname=font_fname).get_name()
rc('font', family=font_name)
csv_data.plot(kind = "scatter", x = '관람객평점', y = '관객수', figsize = (5,5), color = "black")

plt.plot(csv_data["관람객평점"],prediction,color="blue")
```

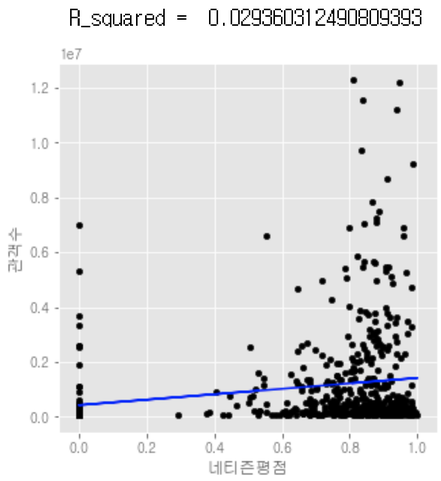
```
[<matplotlib.lines.Line2D at 0x23b5c0ae6d8>]
```

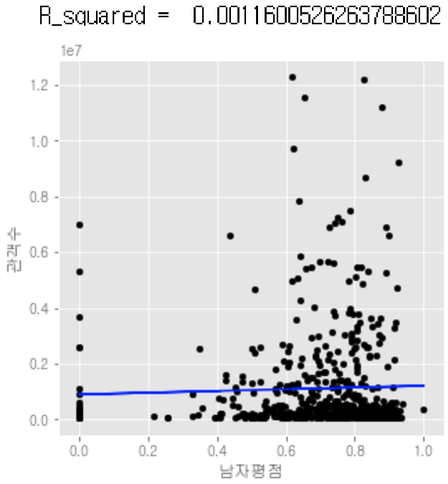
- 전처리 된 2016 ~ 2018년도 영화 정보 테이블
- 단순 회귀 분석으로 각각의 평점들과 관객 수 사이의 관계를 찾아보아 탐색
- 첫 시행으로 관람객 평점과 관객 수 사이의 관계를 찾고 그래프로 표현
- 결정 계수 R_squared 가 0에 가까우면 연관이 없음

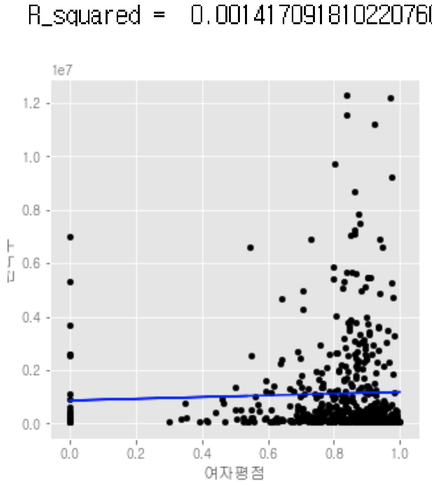
단순 회귀 분석 -3

출처: Naver 영화 , 한국영화진흥원









- 관람객 평점과 관객 수 사이의 관계를 그래프 표현 실시 선으로 표현하는 부분이 예측 값
- 같은 방법으로 나머지 평점에 대해 결정 계수와 그래프를 표현
- 그래프와 R 계수를 보면 알 수 있듯이 평점과 관객 수와 큰 연관은 없음

평점과 흥행 연관이 적다

좋은 영화임에도 불구하고 흥행을 못하는 경우가 있음

감성 분석을 통해 흥행에 상관없이 평가 가능

1.

영화 장르 설정 군집화 후 모험 장르의 영화를 머신 러닝을 이용한 모델링

2.

모험 장르의 영화 중 흥행에 성공한 영화와 실패한 영화의 감성 분석을 실시

감성 분석 -1

출처: Naver 영화 , 한국영화진흥원

In [18]:
```
csv_data_train = csv_data[1:150000]
```

In [19]:
```
csv_data_test = csv_data[150000:200000]
```

In [21]:
```
csv_data_train.head(10)
```

Out[21]:

In [1]:
```
import json
import os
from pprint import pprint
with open('train_docs.json' ,encoding="utf-8") as f:
    train_docs = json.load(f)
with open('test_docs.json', encoding="utf-8") as f:
    test_docs = json.load(f)
```

In [2]:
```
tokens = [t for d in train_docs for t in d[0]]
print(len(tokens))
```

```
2232216
```

:
```
import nltk
text = nltk.Text(tokens, name='NMSC')
print(text)
```

```
<Text: NMSC>
```

- 전처리 과정에서 준비했던 테이블을 불러와 train_set , test_set 분류 (150,000/50,000)
- JSON파일로 읽어 train_set의 토큰을 가져옴
- 문서 탐색용 nltk 라이브러리 사용

감성 분석 -2

출처: Naver 영화 , 한국영화진흥원

In [4]:
```
selected_words = [f[0] for f in text.vocab().most_common(5000)]
```

In [5]:
```
def term_frequency(doc):
    return [doc.count(word) for word in selected_words]

train_x = [term_frequency(d) for d, _ in train_docs]
test_x = [term_frequency(d) for d, _ in test_docs]
train_y = [c for _, c in train_docs]
test_y = [c for _, c in test_docs]
```

In [6]:
```
import numpy as np

x_train = np.asarray(train_x).astype('float32')
x_test = np.asarray(test_x).astype('float32')

y_train = np.asarray(train_y).astype('float32')
y_test = np.asarray(test_y).astype('float32')
```

:
```
from tensorflow.keras import models
from tensorflow.keras import layers
from tensorflow.keras import optimizers
from tensorflow.keras import losses
from tensorflow.keras import metrics

model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(5000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))

model.compile(optimizer=optimizers.RMSprop(lr=0.001),
              loss=losses.binary_crossentropy,
              metrics=[metrics.binary_accuracy])

model.fit(x_train, y_train, epochs=10, batch_size=512)
results = model.evaluate(x_test, y_test)
```

- 가장 빈도가 높은 토큰 5,000개를 가져와 각각 data_set에 매칭을 시켜줌
- 매칭시킨 데이터를 float로 형 변환 과정 진행
- Tensorflow를 이용해 모델을 세우고 에포크 10번 배치 사이즈 512 학습 진행

감성 분석 -3

출처: Naver 영화 , 한국영화진흥원

In [26]:
```
def predict_pos_neg(review):
    token = tokenize(review)
    tf = term_frequency(token)
    data = np.expand_dims(np.asarray(tf).astype('float32'), axis=0)
    score = float(model.predict(data))
    if(score > 0.5):
        print("[{}]는 {:.2f}% 확률로 긍정 리뷰예측 ₩n".format(review, score * 100))
    else:
        print("[{}]는 {:.2f}% 확률로 부정 리뷰예측 ₩n".format(review, (1 - score) * 100))
```

In [27]:
```
predict_pos_neg("이영화 진짜재밋네")
```

```
[이영화 진짜재밋네]는 97.52% 확률로 긍정 리뷰예측
```

- 리뷰를 입력하면 해당 리뷰를 토큰화하여 스코어를 예측하여 0.5 기준으로 긍정 부정을 나눠줌

감성 분석 -4

출처: Naver 영화 , 한국영화진흥원

number	text	score
0	전편에서 김동욱이 터졌다면 이번편은 주지훈이 터진다	1
1	나만 좋았나....?	1
2	주지훈 ..터진다잉 앞머리 내렸을때 심쿵 ㅋㅋㅋ내용은 약간 여러 내용을 담을려고 너...	1
3	내가봣을땐 1점주는애들이 알바같은데	1
5	용서의 중요성 재강조_용서를 구하는 자에게 용기가 필요하고 , 용서를 해 주는 자에...	1
6	영화안본애들이꼭 저러더라~~ 오전에 조조로봤는데 진짜 개꿀잼ㅇㅇ감동에 반전 소름인데...	1
7	대체 중간에 랩터들이랑 공룡들은 왜 나온거지?	0
8	진짜 재미있게 봤는디... 댓글 알바 거리는데 여긴 거진뭐 미션 댓글 알바? 우리나...	1
9	영화 안본사람은 댓글달게하지마라.	1
11	어제 시사회 봤는데 개인적으로 1편보다는 더 여운있고 하정우연기도 좋았어요 다음편에...	1
12	공룡씬은 누가봐도 좀 많이 띠용스러웠다...너무 쥬라기월드(공원) 씬을 패러디?한게...	0
13	우리나라 흥행법칙=스크린독점	0
14	영화관 가면 진심 이것밖에 없어요 ㅜ 독과점 너무 심한 것 같아요 같은 배급산데 ;...	0
15	이 영화를 끝까지 본 내가 귀인이다.	0
16	그냥...스토리 탄탄하게 짰다는 생각이 듦.재판과정도 1편에서 짧게보여줬던 지옥들로...	1
18	재밌으면 좋아요좀 눌러줘요 보러가장	1

```
from tensorflow.keras import models
from tensorflow.keras import layers
from tensorflow.keras import optimizers
from tensorflow.keras import losses
from tensorflow.keras import metrics

model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(5000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))

model.compile(optimizer=optimizers.RMSprop(lr=0.001),
              loss=losses.binary_crossentropy,
              metrics=[metrics.binary_accuracy])

model.fit(x_train, y_train, epochs=10, batch_size=512)
results = model.evaluate(x_test, y_test)
```

1. 해당 스코어가 0 아니면 1로 이루어져 있어 선형 자료에 비선형 특성을 주기 위해 Relu 함수로 활성화 / 더욱 성능을 높이기 위해 2번 활성화

2. 나오는 결과값을 0~1로 표현하기 위해 sigmoid로 활성화

3. 자료가 0 ~ 1 범주형 자료 이므로 binary_croseentropy 사용

4. 자료가 불규칙적으로 등장하므로 RMSprop 옵티마이저를 사용

감성 분석 -5

출처: Naver 영화 , 한국영화진흥원

	Unnamed: 0	영화코드	영화명	개봉일	대표국적	장르	관람객평점	평론가평점	네티즌평점	남자평점	여자평점	관객수
0	0	167697	신과함께-인과 연	180801	한국	모험	0.875	0.631	0.809	0.617	0.839	12274996
146	532	154133	백설공주 : 사라진 아빠를 찾아서	160929	중국	모험	0.765	0.737	0.000	0.640	0.755	55355

text
전편에서 김동욱이 터졌다면 이번편은 주지훈이 터진다
나만 좋았나....?
주지훈 ..터진다잉 앞머리 내렸을때 심쿵 ㅋㅋㅋ내용은 약간 여러 내용을 담을려고 너무 노력함..,근데 평점이 왜케 낮아용...,재밌는뎅 그래서 10점 드림
내가 ㅂㅏㅅㅇ을땐 1점주는애들이 알바같은데
할 뭐지 난 1편이 대놓고 신파라서 너무 싫었는데 볼까말까 하다가 걍 봤는데 2편이 더 좋았음..반응 이래서 개놀람...조반은 엄청 지루해서 나갈까 고민했는데 나중엔 몰입 엄청해서 봄. 저는 좋았어요 개취!@
용서의 중요성 재강조_용서를 구하는 자에게 용기가 필요하고 , 용서를 해 주는 자에게는 더 큰 용기가 필요하다_용서를 구하여 용서를 받기 전까지는 무거운 짐을 지고 살아갈 수밖에 없다_용서를 구하는 것이 실종된 현 ...
영화안본애들이꼭 저러더라—— 오전에 조조로봤는데 진짜 개꿀잼ㅇㅇ감동에 반전 소름인데 다 안보고 일찍나간애들이 불쌍ㅋㅋ
대체 중간에 랩터들이랑 공룡들은 외 나온거지?
진짜 재미있게 봤는디... 댓글 알바 거리는데 여긴 거진뭐 미션 댓글 알바? 우리나라 사람들 마술보러가도 마술 감상은 안하고 자신들이 이미 마술사라더니 영화도 이미 작가.. 평점 1점주고 테러 할정도 아닌듯.. 1편...
영화 안본사람은 댓글달게하지마라.
슬직히 1편보다는 지루한감이있었고 성주신나올때 조금웃기고슬프고 반전은 신선했고 원일병 마지막에 쫌 웃겼고 영화가끝나면 반전 대박이다 하는영화 근데 마지막빼곤 지루하다
어제 시사회 봤는데 개인적으로 1편보다는 더 여운있고 하정우연기도 좋았어요 다음편에 도경수 기대해도 되는거죠?
공룡씬은 누가봐도 좀 많이 띠용스러웠다...너무 쥬라기월드(공원) 씬을 패러디?한게 티가 남ㅠㅠ그래도 전생기억하는 부분은 재밌었다. 전생 기억하는 부분만..아 김수홍 깐죽대는 거 진심 한 대 치고싶었음^^„
우리나라 흥행법칙=스크린독점
영화관 가면 진심 이것밖에 없어요 ㅜ 독과점 너무 심한 것 같아요 같은 배급산데 ;국내영화라고 개봉 이주도 안 된 미션 아이맥스 4D관 다 내려버리고 심지어 일반관까지 반토막내버리는 게 어디있나요 ㅠ 이렇게까지 하는...

4살여자아이 보고나서 엉엉울었네요 ㅎㅎ아무래도 아빠를 찾는 장면에서 뭉클했던거같아요~^^
괜찮음ㅎㅎ 아수라보단
나쁘진않았지만 뭔가 짬뽕된것같고 그래픽도 좀유치. . 아무리 애니라도 내용에 설득력이 많이부족해서 그다지 재밌다거나한 느낌은 없었네요. 좀 막만든. . 그냥 애라 같이 본데 의의를~ 잘봤어요
아이들 관점에서 잼있었어요~ 저희 아이는 눈물도 흘렸네요^^
재밌어여.여아들이 조아할 만한 스토리.그리 길지않아 지루하지 않게 볼 수 있었어요.
왠만하면 6살짜리 우리딸 깔깔 웃으며 보는데 단한번도 안웃었다. 그래픽도 어설프고 음악도 엉망임. 어른은 다 자고 있고 애들은 언제 재밌는 장면 나오나 기다리다가 그냥 끝남
제눈엔허접스러워요 7살아이는 재밌다네요 ㅋㅋ
가족과 재미있는 영화로 좋은 시간 보냈습니다 아이에게 의미있는 영화로 남을것같아요
아이들과 함께가서 즐거운시간 보내고왔습니다. 아이들이 아빠를 다시 한번 생각하게되는 좋은기회였던것같습니다~^^
두려워 하지마. 잘 봤습니다.
아이들의 눈높이~재미있게 보았네요.
대학교 졸업작품같은 완성도...돈이 아까워요
등장인물들의 움직임들이 부자연스러워 보이는게 아쉽습니다 제대로 홍보도 되질 않았는지 하루에 12시 딱 한번 상영하는것도 아쉬었네요 내용은 6~7세 여자아이가 딱 좋아할 내용이였습니다
아이와 같이 잘봤어요. 아빠의 소중함을 느끼길바라며~~~
아이들이 보기에 잔잔한 감동이 있네요

1. 모험 장르 중 관객 수가 많은 영화 1개 / 관객 수가 적은 영화 1개 선정

2. 선정 후 리뷰만 크롤링 작업 진행 후 저장
   흥행 성공 adv_review1 / 흥행 실패 adv_review2

감성 분석 -6

출처: Naver 영화 , 한국영화진흥원

```
import pandas as pd
from pandas import DataFrame as df
csv_data1 = pd.read_csv("adv_review1.csv",encoding='euc-kr')
csv_data2 = pd.read_csv("adv_review2.csv",encoding='euc-kr')
```

```
def predict_score(review):
    token = tokenize(review)
    tf = term_frequency(token)
    data = np.expand_dims(np.asarray(tf).astype('float32'), axis=0)
    score = float(model.predict(data))
    if(score > 0.5):
        return 1
    else:
        return 0
```

```
u=0
l=[]
for i in csv_data1["text"]:
    l.append(predict_score(i))
```

In [18]:
```
print(len(l))
sum=0
for a in l:
    sum = sum+a
print(sum)
```

```
19984
13343
```

```
u=0
k=[]
for i in csv_data2["text"]:
    k.append(predict_score(i))
```

:
```
print(len(k))
sum=0
for a in k:
    sum = sum+a
print(sum)
```

```
130
110
```

1. 크롤링 진행 후 구한 모델을 이용하여 총 리뷰 중 긍정적인 부분이 몇 %를 차지하는지 확인

2. 흥행에 실패한 영화의 경우에도 긍정적인 비율이 총 리뷰 130개 중 110개 나올 정도로 관람객들의 반응이 좋았지만 실제로 흥행에는 성공하지 못했음

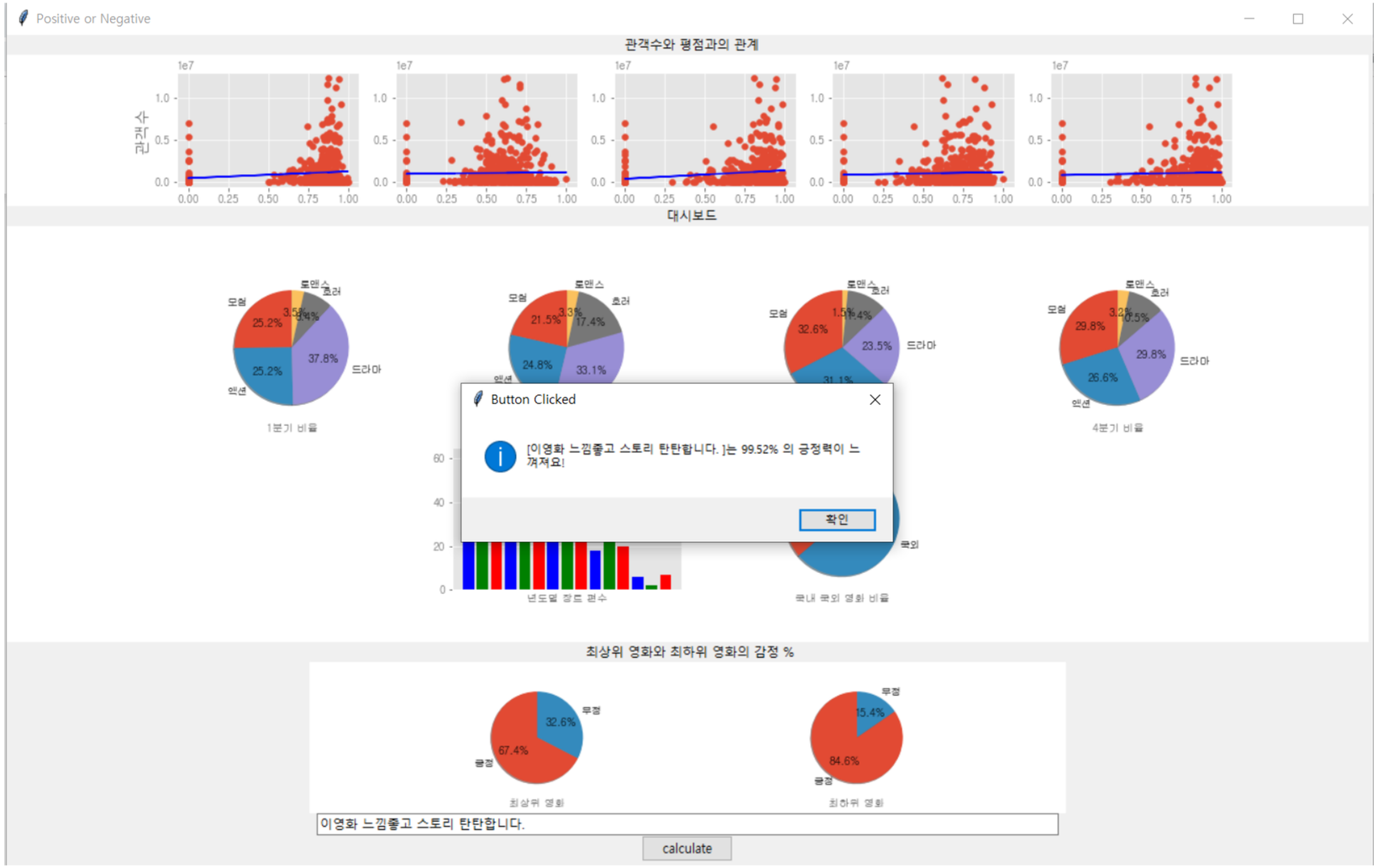
3. 흥행 성공한 영화의 경우에도 긍정적인 비율이 대개 높겠지만 무조건적으로 흥행 실패한 영화보다 높지는 않음(19,984개 중 13,343개가 긍정적인 비율)

UI

출처: Naver 영화 , 한국영화진흥원

UI

출처: Naver 영화 , 한국영화진흥원



4. 한계점 및 보완점

4. 한계점 및 보완점

01

문제점 : pyspark 미사용

기본 영화 100개만 해도
리뷰 양이 50만 개를 넘어가
파이선 자체로 감당이 안 되었음

02

문제점 : 프로그램의 용량

문제점 1번과 같이 리뷰 양이 많아짐에
따라 프로그램도 무거워져 실행에 지연
및 여러 가지로 컴퓨터에 영향을 줌

03

보완점 : 여러가지 장르에 대해 분석 실시

같은 단어라고 해도 장르마다
긍정적인 의미와 부정적인 의미로
상이하게 받아 들여질 수 있음

04

다양한 크롤링 진행

한국 영화뿐 아니라 외국 로튼 토마토
사이트에 언급된 자료 크롤링 혹은 트위터를
이용한 반응 조사를 통해 다양한 분석 과정
진행으로 보완

5. 참고 사이트

5. 참고사이트 및 자료 출처

한국어 정보 처리 :

https://konlpy-ko.readthedocs.io/ko/v0.4.3/

네이버 리뷰 분석 :

https://cyc1am3n.github.io/2018/11/10/classifying_korean_movie_review.html

영화 박스 오피스 :

https://www.kofic.or.kr/

리뷰 및 영화 자료 출처

: https://movie.naver.com/

Thank You