투 무빅스

<리뷰분석을 통한 영화 손익분기점 예측>

장재석, 오건우, 전종섭, 최문정, 황다솔

CONTENTS

주제 선정 배경

1 배경

- 2017년 개봉영화들의 누적 관객수

6,592,151명

6,879,844 명

5,653,270 명

3,849,087 명

3,279,296 명

주제 선정 배경

1 배경

- 이 영화들의 손익분기점은 ?

6,592,151 명

8,000,000명

6,879,844 명

2,000,000 명

5,653,270 명

2,000,000 명

3,849,087 명

5,000,000 명

3,279,296 명

1,800,000 명

주제 선정 배경

1 프로젝트 주제

영화 개봉 전, 배급사들은 영화의 반응을 알기 위해 시사회를 진행한다.

주제 선정 배경

1 프로젝트 주제

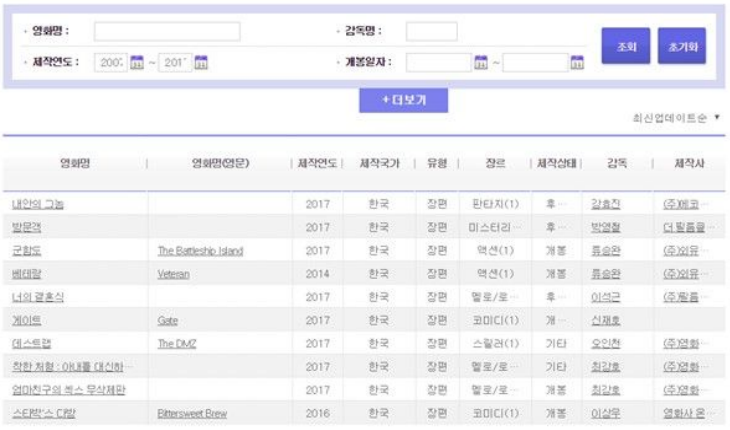
“그렇다면 이 개봉 전 반응으로 영화가 손익분기점을 넘길 지 알 수 있을까?”

데이터 수집

2 영화 목록

KOFIC 영화진흥위원회 Korean Film Council

-> 영화진흥위원회>영화정보센터>영화DB검색



영화명	영화명(영문)	제작연도	제작국가	유형	장르	제작상태	감독	제작사
내안의 그놈		2017	한국	장편	판타지(1)	후…	강효진	(주)제코…
방문객		2017	한국	장편	미스터리…	후…	방엘철	더 필름을…
군함도	The Battleship Island	2017	한국	장편	액션(1)	개봉	류승완	(주)외유…
베테랑	Veteran	2014	한국	장편	액션(1)	개봉	류승완	(주)외유…
너의 결혼식		2017	한국	장편	멜로/로…	후…	이석근	(주)필름…
게이트	Gate	2017	한국	장편	코미디(1)	개…	신재호	
데스트랩	The DMZ	2017	한국	장편	스릴러(1)	기타	오인천	(주)영화…
착한 처형 : 아내를 대신하…		2017	한국	장편	멜로/로…	기타	최강호	(주)영화…
엄마친구의 썩스 무삭제판		2017	한국	장편	멜로/로…	개봉	최강호	(주)영화…
스타맨'스 다방	Bittersweet Brew	2016	한국	장편	코미디(1)	개봉	이상우	영화사 온…

약 6500개의 영화목록 수집

2 네이버 영화 평점

범죄도시
THE OUTLAWS, 2017

관람객 9.28 기자·평론가 6.00
네티즌 9.15 내 평점 등록

개요 액션 | 한국 | 121분 | 2017.10.03 개봉
감독 강윤성
출연 마동석(마석도), 윤계상(장첸) 더보기
등급 [국내] 청소년 관람불가

다운로드 8,415

주요정보 | 배우/제작진 | 포토 | 동영상 | 평점 | 리뷰 | 명대사/연관영화

개봉 전 개봉 후

기대지수

보고싶어요 7,567 1,129 글쎄요

2 각 영화별 리뷰

140자 평 | 총1,210건

베스트 평점 운영 정책 안내

호감순 최신순

★★★★★ 10 베스트 형사가 마동석 조폭이 윤계상. 조폭보다 형사가 무섭네

sork**** | 2017.03.05 00:28 | 신고

공감 354 비공감 26

★★★★★ 10 베스트 마동석 윤계상 ㄷㄷ 글구 대세 신스틸러 조재윤 최귀화까지.. 진짜 일단 믿고 극장가서 본다

JM(wjda****) | 2017.03.19 10:30 | 신고

공감 212 비공감 32

★★★★★ 10 베스트 마블리와 윤계상의 조화~!! 기대됩니다. ㅎㅎ

긍정의 힘(nati****) | 2017.03.22 14:25 | 신고

공감 172 비공감 20

★★★★★ 10 베스트 배우 윤계상씨의 악역으로 나온다고하니 더욱더 기대됩니다

seco**** | 2017.03.26 22:48 | 신고

공감 170 비공감 24

★★★★★ 10 베스트 윤계상 배우님 기대됩니다 ! 조폭연기 잘어울릴꺼같아요^^

댕댕(hyoj****) | 2017.03.28 09:18 | 신고

공감 151 비공감 22

2 영화 목록 필터링

6500 개의 영화

필터링

- 네이버 영화 내 평점 X
- 참여인원이 너무 적은 영화

1689 개의 영화

3 데이터 분할

TRAIN	VALIDATION	TEST
6	2	2

3 변수 설명

변수	설명		
글쎄요	수치형		
보고싶어요	수치형		
참여인원	수치형		
평점	수치형		
장르	SF, 판타지, 느와르, 로맨스 등	→	카테고리화
관람등급	전체관람가, 15세 이상 관람가, 청소년 관람 불가	→	카테고리화
배급사	롯데엔터테인먼트, NEW 등	→	배급사의 영화 배급 양에 따라 1~5

분석과정

3 기본 분석

	MEAN ACCURACY	MEAN F1 SCORE
RIDGE	0.80	0.65
LASSO	0.80	0.64
RANDOMFOREST	0.78	0.63
DECISIONTREE	0.80	0.53
KNN	0.76	0.52

3 F1 SCORE

		TRUE	
		POSITIVE	NEGATIVE
PREDICTED	POSITIVE	TRUE POSITIVE	FALSE POSITIVE
	NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{tp}{2tp + fp + fn}$$

3 영화 리뷰의 변수화

"각 영화의 리뷰들을 요약하여 변수로 만들자"

어떻게 만들까?

→ 1. COUNT 기반

→ 2. 감성사전 기반

분석과정

3 영화 리뷰 전처리

50만개의 리뷰 데이터

재미있을것같네요!
주인공연기가 너무좋았습니다.앞으로도 좋은연기부탁드립니다^^
4일에 보고 왔습니다. 정말 몰입해서 봤네요. 공감가는 내용도 많아 울컥했습니다.
아 완전 내가 원하던 내용의 영화!!! 개봉 꼭 해주세요!! 지구 반대편에서 개봉하더라도 보러갑니다 ♥
기대하구잇어요~~~
시간이가는줄모르고 봤습니다지금까지본 영화중 가장 현실적이면서 감정이입 잘 되는 영화!!
비정규직의 현실을 너무 디테일하게 잘 표현한 영화. 일주일에서 10분으로 바뀐 현실. 엔딩의 10분의 초
감정폭발의 표현 없이도 이런 영화가 나올수 있다는것을 보여주었다
18회 부국제에서 본 최고의 한국영화. 지극한 현실을 가지고 무섭도록 아픈 드라마를 재단해낸 솜씨.
꼭 실전 필수 관람! 적극 추천!
방금 영화의전당에서보고왔는데 정말 재밌게봤네요. 공감많이가고 너무현실적인 영화.. 잘봤습니다^^
부산영화제에서 봤는데 정말 대박 작품이네요. 시나리오, 연출, 배우 연기 모두 너무나 훌륭하고 시간가
18회 부산국제영화제에서 본영화중 2번째로 재미있는영화였음 공감이 많이가서 울컥
대박기원!! 재미날것 같아요~^^

KONLPY
TWITTER

['재미있/Adjective', '을/PreEomi', '것/PreEomi', '같네/Adjective', '요/Eomi', '!/Punctuation']
['주인공/Noun', '연기/Noun', '가/Josa', '너무/Noun', '좋/Adjective', '았/PreEomi', '습니다/Eomi', './Punctuat
['4/Number', '일/Noun', '에/Josa', '보고/Noun', '왔/Verb', '습니다/Eomi', './Punctuation', '정말/Noun', '몰입,
['아/Exclamation', '완전/Noun', '내/Noun', '가/Josa', '원하던/Verb', '내용/Noun', '의/Josa', '영화/Noun', '!!!/
['기대하구/Adjective', '잇어/Verb', '요/Eomi', '~~~/Punctuation']
['시간/Noun', '이/Josa', '가는/Verb', '줄/PreEomi', '모르/Verb', '고/Eomi', '봤/Verb', '습니다/Eomi', '지금/No
['비정규직/Noun', '의/Josa', '현실/Noun', '을/Josa', '너무/Noun', '디테/Noun', '일하게/Verb', '잘/Verb', '표현
['감정/Noun', '폭발/Noun', '의/Josa', '표현/Noun', '없/Adjective', '이/PreEomi', '도/Eomi', '이런/Adjective', '
['18/Number', '회/Noun', '부/Noun', '국제/Noun', '에서/Josa', '본/Verb', '최고/Noun', '의/Josa', '한국영/No
['꼭/Noun', '실전/Noun', '필수/Noun', '관람/Noun', '!/Punctuation', '적극/Noun', '추천/Noun', '!/Punctuatio
['방금/Noun', '영화의전당/Noun', '에서/Josa', '보고/Noun', '왔/Verb', '는데/Eomi', '정말/Noun', '재밌게/Adj
['부산영화제/Noun', '에서/Josa', '봤/Verb', '는데/Eomi', '정말/Noun', '대박/Noun', '작품/Noun', '이네/Josa',
['18/Number', '회/Noun', '부산/Noun', '국제/Noun', '영화제/Noun', '에서/Josa', '본영/Noun', '화중/Noun', '
['대박/Noun', '기원/Noun', '!!/Punctuation', '재미/Noun', '날/Verb', '것/PreEomi', '같아/Adjective', '요/Eomi'
['직장인/Noun', '의/Josa', '공감/Noun', '백서/Noun', '!/Punctuation', '취준생/Noun', '은/Josa', '필수/Noun',

NOUN, VERB, ADJECTIVE , FOREIGN, KOREAN PARTICLE, PUNCTUATION

두 글자 이상의 단어

분석과정

3 COUNT 기반

- 계수 계산 방법

1000개

	기대됨/VERB	하는/VERB	입니/ADJECTIVE	시사회/NOUN	완전/NOUN	기대되/VERB	감독/NOUN	보러/VERB	...
총합	14755	14776	14861	15192	16834	16851	17403	17441	...
영화A	10	0	0	0	124	632	0	0	...
계수	6.7E-04	0	0	0	0.0073	0.0378	0	0	...

상위 1000개의 단어를 이용해 각 영화들의 계수를 계산

분석과정

3 COUNT 기반

- 분석

	MEAN ACCURACY	MEAN F1 SCORE
LASSO	0.85	0.70
RANDOM FOREST	0.84	0.68
RIDGE	0.85	0.67
DECISION TREE	0.85	0.63
KNN	0.81	0.54

3 감성사전 기반
- 긍/부정 라벨링

리뷰 별 평점을 이용하여 긍/부정 라벨링

분석과정

3 감성사전 기반

- 감성 사전 만들기

1000개

	단어A	단어B	단어C	단어D	단어E	단어F	...
리뷰1	0	0	0	0	1	0	...
리뷰2	1	1	0	0	0	0	...
...							

50만개

TF-IDF 생성

100 차원

	1	2	3	4	5	6	...
단어A	0.012	0.551	-0.102	0.432	0. 357	-0.230	
단어B	-0.982	-0.234	-0.475	0.452	0.032	0.098	
...							

100개

WORD VECTOR 생성

예) 리뷰2가 단어A와 단어B만을 가지고 있다면?

3 COUNT 기반
- 배우 이름 제거

~~최승현, 차승원, 소지섭, 서인국, 박서준, 이선균, 신세경 ....~~

상위 1000개의 단어에 많은 배우 이름들이 포함
배우 이름들을 제거 전 후를 비교

분석과정

3 COUNT 기반

- 배우 이름 제거 후 분석

	MEAN ACCURACY	MEAN F1 SCORE
RIDGE	0.84	0.69
LASSO	0.85	0.68
RANDOM FOREST	0.83	0.64
DECISION TREE	0.85	0.59
KNN	0.81	0.54

분석과정

3 감성사전 기반

- 감성 사전 만들기

100 차원

	1	2	3	4	5	6	...
리뷰1	...	...	...	...	...	...	...
리뷰2	-0.485	0.185	-0.288	0.442	0.194	-0.066	...
...							

분석과정

3 감성사전 기반

- 감성 사전 만들기

437개

	영화	너무	기대	배우	개봉	연기	정말	보고	...
계수	0.019	0.020	0.065	0.003	0.043	-0.047	0.063	0.041	...

LASSO를 통해서 각 단어의 긍/부정 계수를 계산하여 감성사전 구축

3 감성사전 기반

- 분석

	MEAN ACCURACY	MEAN F1 SCORE
RANDOM FOREST	0.84	0.73
RIDGE	0.84	0.70
LASSO	0.85	0.67
KNN	0.82	0.54
DECISION TREE	0.84	0.49

3 감성사전 기반

- 배우 이름 제거 후 분석

	MEAN ACCURACY	MEAN F1 SCORE
RIDGE	0.85	0.68
RANDOM FOREST	0.85	0.68
LASSO	0.85	0.67
KNN	0.82	0.66
DECISION TREE	0.85	0.63

4 한계점

손익 분기점

- 총 제작비 = 순 제작비 + 마케팅 비용
→ 배급사의 규모로 대체 했으나 부족

영화 필터링

- 성인영화 등

리뷰

- 리뷰들 대부분이 긍정적인 반응을 기대하는 댓글. 이런 부분이 분석에 좋지 않은 영향

결론

4 총합

순위	방법론 + 데이터	MEAN ACCURACY	MEAN F1 SCORE	TEST ACCURACY	TEST F1 SCORE
1	RANDOM FOREST + DATA5	0.85	0.68	0.84	0.72
2	LASSO + DATA3	0.85	0.68	0.85	0.69
3	LASSO + DATA2	0.85	0.7	0.85	0.68
4	RANDOM FOREST+ DATA4	0.84	0.73	0.84	0.68
5	RIDGE + DATA1	0.85	0.67	0.84	0.65

*DATA1: 기본 데이터, DATA2: COUNT기반, DATA3: COUNT기반(배우제거), DATA4: 감성사전기반, DATA5: 감성사전기반(배우제거)

배우를 제거한 데이터에서 더 높은 F1 SCORE를 관찰 할 수 있음.

방법론적으로는 RANDOM FOREST 와 LASSO가 좋은 성능을 보임.

THANK YOU