멀티캠퍼스 데이터분석 8기 주간 프로젝트

코로나 이후 4번째 천만 영화

파묘

흥행 요인 분석

온라인 사용자 리뷰를 중심으로

김선규 정사라 김현우 박소현 이의준

흥행 요인 분석

프로젝트 개요

프로젝트 기획 배경 및 목표

파천내천

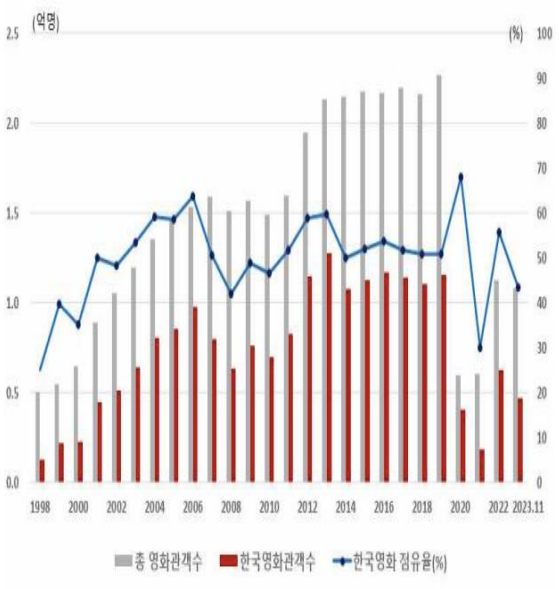
기획 배경: 코로나 이후 타격을 받은 영화 산업에서 이례적인 장르로 천만 관객을 돌파한 파묘의 평점과 리뷰를 분석

변화하는 영화 추세 : 입소문과 바이럴 마케팅이 일으키는 역주행으로 '개봉 첫 주 최대 관객 유입'에서 '개봉 2주차 이후 관객 확대'로 바뀌고 있다는 점 등이 있었다 * (논문, 영진위분석보고서)

구전은 영화흥행에 영향을 미치는 주요 요인으로 여러 연구에서 언급되어 왔는데, Henning-Thurau 외 3명(2004)은 관람객의 리뷰가 영화흥행에 중요 결정요인이며, 특히 평점은 첫 주의 관객수와 총 관객수에 영향을 준다고 밝혔다[22]. 정영호(2013)는 영화 선택 시 소비자들은 온라인 구전을 중요하게 고려하고, 다수의 의견이 제시된 경우 영화 선택에 긍정적 영향을 미친다고 말했다[26].

영화와 같이 관여도가 높은 상품일수록 구전의 수용과 확산이 활발하다는 연구결과도 존재한다[27][28].

<그림8> 국내 영화관 관객수 추이



자료: 영화진흥위원회 박스오피스 통계

```
boxoffice_sort = boxoffice.sort_values('증감관객수', ascending=False).iloc[:10]
boxoffice_sort
```

✓ 0.0s

	날짜	일일순위	증감순위	신규진입여부	일일매출	일일흥액대비매출	증감매출	증감매출비율	누적매출	일일관객수	증감관객수	증감관객수비율	누적관객수	일일스크린수	일일상영횟수
9	20240301	1	0	OLD	8368185017	67.5	4757382396	131.8	43442838683	851600	467306	121.6	4548834	2315	9700
3	20240224	1	0	OLD	7570365687	76.5	3914360449	107.1	14375475076	770974	396482	105.9	1481595	2355	9473
1	20240222	1	19	OLD	3091398151	73.6	3082796151	35838.1	3149104151	330118	329280	39293.6	336129	1944	7535
17	20240309	1	0	OLD	5539634220	68.1	3169536758	133.7	72848169427	559672	316437	130.1	7569650	2164	9143
24	20240316	1	0	OLD	3425655683	64.3	1943512680	131.1	86924949017	344954	194286	128.9	9013225	1959	7986
31	20240323	1	0	OLD	2641611585	59.3	1569539012	146.4	96257458340	265402	156513	143.7	9965290	1807	7402
7	20240228	1	0	OLD	3082112642	59.0	2454692666	8.7	31463851045	384594	86130	28.9	3312940	1885	7156
16	20240308	1	0	OLD	2370097462	67.4	857937120	56.7	67308535207	243235	82873	51.7	7009978	2098	8267
4	20240225	1	0	OLD	8041374368	77.8	4710086681	6.2	224168494444	819842	48868	6.3	2301437	2351	9262
23	20240315	1	0	OLD	1482143003	62.0	487654627	49.0	83499293334	150668	45603	43.4	8668271	1860	7090

프로젝트 개요

프로젝트 기획 배경 및 목표

파천내천

타겟

영화의 홍보의 목적을 가지고 있는 영화사 or 평점이 존재하지않은 리뷰 사이트

목적

실시간 리뷰의 분석을 통해

결과에 따라 사용자에게 차별된 홍보 방안 제시

기대효과

영화를 보고 나온 사람들에게 한번 더 영화를 홍보할 수 있는 수단

모델링

험한 것이 나왔다. 😊

분석과제 정의

Q1
온라인 사용자 반응과 관객수의 상관관계를 발견할 수 있을까?

가설

검색어 지수, 리뷰 수 등의 지표는 관객수 증가에 유의한 영향을 끼칠 것이다

분석 방법

구글과 네이버의 검색어 지수(트랜드 검색어) 와 리뷰 수를 카운트해 시각화

Q2
키워드 분석을 통해 새로운 인사이트를 발견할 수 있을까?

가설

이례적인 흥행이라고 생각되는 해당 영화의 독특한 특징이 존재 : 장르(오컬트), 반일 논란 등이 리뷰에서 긍정/부정의 형태로 드러날 것이다

분석 방법

긍정/부정 리뷰 로 label 설정 후 감성분석 진행 키워드 분석, 토픽화

Q3
구축한 모델을 어떻게 이용할 수 있을까?

가설

온라인 영화 평점 및 리뷰 데이터를 학습시킨 머신러닝 모델을 통해 파묘에 대한 짧은 의견의 실시간 감성분석이 가능할 것이다

분석 방법

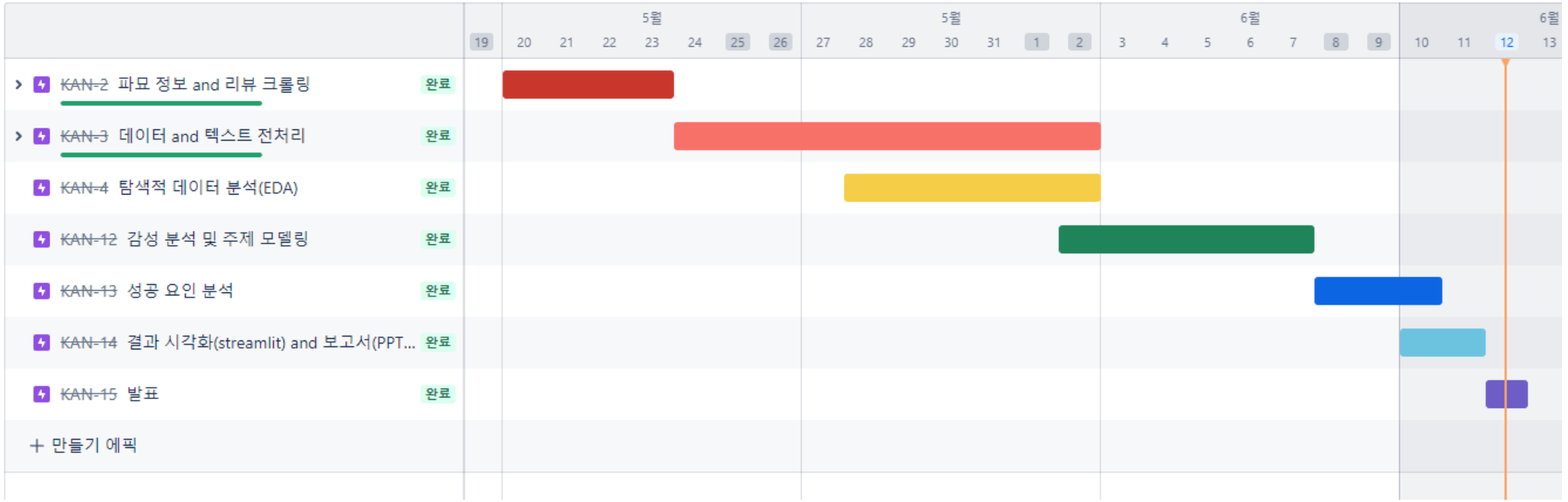
사용자 리뷰 입력 시 실시간 감성분석을 통해 사용자의 반응을 유추할 수 있을 것이다

프로젝트 개요

구성원 및 개발 일정

파천내천

김선규	**(조장)** 일정관리, 업무분담 및 추진, 데이터 수집 및 통합 (크롤링 및 API), 전처리 및 모델링, 웹 구현
정사라	**(부조장)** 기획, 데이터 수집 및 통합 (크롤링 및 API), 전처리 및 모델링 고도화, 웹 구현, 발표자료 제작
김현우	데이터 수집 및 통합 (크롤링 및 API), 전처리 및 모델링, 웹 구현 프로토타이핑 및 고도화
박소현	데이터 수집 및 통합 (크롤링 및 API), 전처리 및 모델링, 웹 구현
이의준	데이터 수집 및 통합 (크롤링 및 API), 전처리 및 모델링, 웹 구현



프로젝트 개요

구성원 및 개발 일정

파천내천

```python
import requests, json, urllib.request

CLIENT_ID = ""
CLIENT_SECRET = ""


url = "https://openapi.naver.com/v1/datalab/search"
params = {
    "startDate": "2024-02-22",
    "endDate": "2024-03-24",
    "timeUnit": "date",
    "keywordGroups": [
        {"groupName": "파묘", "keywords": ["파묘", "파묘후기"]},
        {"groupName": "범죄도시4", "keywords": ["범죄도시4", "범죄도시"]}]}

headers = { "X-Naver-Client-Id": CLIENT_ID,
            "X-Naver-Client-Secret" : CLIENT_SECRET,
            "Content-Type" : "application/json"}

response = requests.post(url, data=json.dumps(params), headers=headers)
response.json()
```

```
{'startDate': '2024-02-22',
 'endDate': '2024-03-24',
 'timeUnit': 'date',
 'results': [{'title': '파묘',
   'keywords': ['파묘', '파묘후기'],
   'data': [{'period': '2024-02-22', 'ratio': 67.68141},
    {'period': '2024-02-23', 'ratio': 82.77883},
    {'period': '2024-02-24', 'ratio': 100},
    {'period': '2024-02-25', 'ratio': 95.67934},
```

API 사용(네이버)

리뷰 데이터 확인

```python
# 페이지 이동 함수
def move_to_page(page_number):
    try:
        page_link = driver.find_element(By.XPATH, f"//a[text()='{page_number}']")
        page_link.click()
        time.sleep(2)  # 페이지 로딩을 기다립니다.
        return True
    except Exception as e:
        print(f"{page_number}페이지 이동 실패: {e}")
        return False

# 페이지 이동 및 데이터 수집
page_number = 1
while page_number <= #마지막 페이지 숫자 입력:
    if move_to_page(page_number):
        # 현재 페이지의 리뷰 수집
        soup = bs(driver.page_source, "lxml")
        reviews = soup.select(".box-comment")
        nick_names = soup.select(".writer-name")
        days = soup.select(".day")
        recs = soup.select("#idLikeValue")


        # 리뷰 출력
        for name, review, day, rec in zip(nick_names, reviews, days, recs):
            name_text = name.text
            review_text = review.text
            day_text = day.text
            rec_text = rec.text
            print(f"닉네임: {name_text}₩n리뷰: {review_text}₩n작성일: {day_text}₩n추천수: {rec_text}")
            print()
        # 페이지 이동
        if page_number == 10:  # 10페이지일 때
            more = driver.find_element(By.CSS_SELECTOR, "#paging_point > li:nth-child(11) > button")
            more.click()
```

크롤링 사용(cgv)

프로젝트 개요

파천내천

구성원 및 개발 일정

- 영화 제목: 해당 영화 제목을 나타내는 것
  이유 : 필요
- 날짜(single_date_str): 해당 날짜를 나타내는 것
  이유 : 필요
- 일일순위(rank) : 해당일자의 박스오피스의 순위를 나타내는 것
  이유: 필요 땅땅땅
- 증감순위(rankInten) : 전일대비 순위의 증감분을 나타내는 것
  이유 :증감 순위는 일일 순위가 있기 때문에 없어도 된다고 판단
- 신규진입여부(rankOldAndNew) : new, old 는 랭킹신규진입 여부
  이유 : 그러므로 없어도 된다고 판단
- 일일매출(salesAmt): 해당일의 매출액을 나타내는 것
  이유 :이것가지고 할 수 있는 일이 매우 많음
- 일일총액대비매출(salesShare): 해당일의 매출총액 대비 해당 영화의 매출비율
  이유 :해당 일의 매출총액은 지나치게 많은 데이터 이므로 월별이 적당하다고 판단해서 지웁니다.
- 증감매출(salesInten): 전일 대비 매출액 증감분
  이유 :결측치 존재하기 때문에 삭제할것임
- 증감매출비율(salesChange): 전일 대비 매출액 증감 비율
  이유 :비율은 필요없어보임
- 누적매출(salesAcc): 누적매출액
  이유 : 필요하다고 판단
- 일일관객수(audiCnt): 해당일의 관객 수
  이유 : 영화 흥행을 판단하는 주요 요소이므로 보존
- 증감관객수(audiInten): 전일 대비 관객수 증감분을 나타냄
  이유 : 전일 대비 관객수는 데이터가 필요 없다고 판단
- 증감관객수비율(audiChange): 전일 대비 관객수 증감 비율
  이유 : 비율은 필요 없어 보이므로 삭제
- 누적관객수(audiAcc): 누적 관객 수
  이유 : 영화 흥행을 판단하는 주요 요소이므로 보존
- 일일스크린수(scrnCnt): 해당일자의 상영한 스크린 수
  이유 : 영화 흥행을 판단하는 주요 요소이므로 보존
- 일일상영횟수(showCnt): 해당일자에 상영된 횟수
  이유 : 영화 흥행을 판단하는 주요 요소이므로 보존

영어로 되어있는 변수 명을 해석 후 필요 없는 열 확인

날짜 통일

- 개봉일(2024-02-22)부터 천만관객 돌파일(2024-03-24)까지를 분석 기간으로 선정

```
# Set the integer dates

start_date = pd.to_datetime("2024-02-22")
end_date = pd.to_datetime("2024-03-24")

# Filter the DataFrame based on the date range
review_date = review_drop.loc[(review_drop['작성일'] >= start_date) & (review_drop['작성일'] <= end_date)].reset_index(drop=True)

```

분석 기준을 잡고 필요없는 날짜 제거

프로젝트 개요

구성원 및 개발 일정


```
df2.insert(0, '출처', 'CGV')  # 0은 열의 인덱스로, 첫 번째 위치를 의미함

df2.head()
```

	출처	닉네임	리뷰	작성일	추천수
0	CGV	so**20	나쁘지 않은 연기력...근데 크리처물이라 생각은 하지 않았는데 역시 판타지라 현실...	2024.05.22	0
1	CGV	디링다리리룽	2번봤어요 소재가 너무 신선하고 쫄리는 긴장감	2024.05.22	0
2	CGV	이화온	후반에 판타지되버려 아쉬움	2024.05.22	0
3	CGV	sg**0818	긴장의 끈을 놓치않는 좋은 영화	2024.05.22	0
4	CGV	kl**007	기대가 커서인지 너무 재미없었어요\n뭐랄까 명분이 없는 스토리 전개\n짜맞추기식 ...	2024.05.22	0

```
df2.info() #Nan값 없음
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54066 entries, 0 to 54065
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   출처      54066 non-null  object
 1   닉네임     54066 non-null  object
 2   리뷰      54066 non-null  object
 3   작성일     54066 non-null  object
 4   추천수     54066 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 2.1+ MB
```

```
# 출처, 닉네임,작성일, 리뷰 순서로 열 변경

df2 = df2[['출처','닉네임','작성일','리뷰','추천수']]
df2.head()
```

	출처	닉네임	작성일	리뷰	추천수
0	CGV	so**20	2024.05.22	나쁘지 않은 연기력...근데 크리처물이라 생각은 하지 않았는데 역시 판타지라 현실...	0
1	CGV	디링다리리룽	2024.05.22	2번봤어요 소재가 너무 신선하고 쫄리는 긴장감	0
2	CGV	이화온	2024.05.22	후반에 판타지되버려 아쉬움	0

Null 값 확인 후 기준으로 잡은 열 순서로 변경

```
df = pd.concat([df1,df2],axis=0).reset_index(drop=True)
```

```
df
```

	출처	닉네임	작성일	리뷰	추천수
0	롯데시네마	강*윤	2024-05-24	재미지다	NaN
1	롯데시네마	최*성	2024-05-24	잼나게 잘 봤어요~^^	NaN
2	롯데시네마	차*우	2024-05-24	예상치 못한 일제강점기 내용과 장대같이 큰 무언가가 와닿진 않네요. 전반적인 재미는...	NaN
3	롯데시네마	김*연	2024-05-24	탄탄한 스토리와 연기력	NaN
4	롯데시네마	이*지	2024-05-23	뭔가 대단한게 나올것 같았는데...	NaN
...	...	...	...	...	...
61528	CGV	uriii79	2024.02.21	배우들의 연기가 미쳤다!!!	7.0
61529	CGV	je**ng75	2024.02.21	시사로 일찍보았다는게.정말 만족스럽다 간만에 대세감 있는 한국영화.나왔다	39.0
61530	CGV	camillee	2024.02.21	시간 가는 줄 모르고 봄.. 연기도 스토리도 기대보다 훨씬 미쳤다	38.0
61531	CGV	sq**k123	2024.02.21	그저 미쳤다는 말 밖에.... 진짜 오랜만에 영화다운 영화를 봤다!! 스토리, 배...	35.0
61532	CGV	ha**ysj03	2024.02.21	천만영화가 나왔다. 개꿀잼	5.0

61533 rows × 5 columns

변경 후 데이터 통합 및 확인

프로젝트 개요

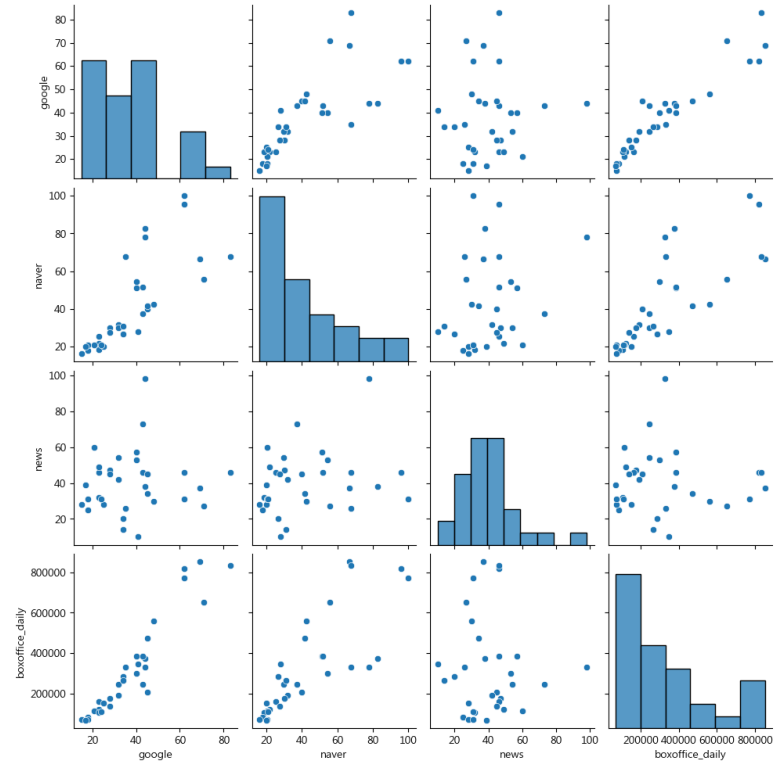
구성원 및 개발 일정

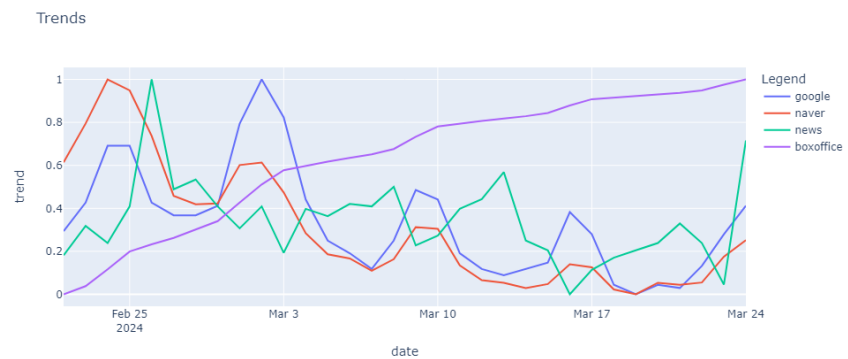
파천내천

```python
# boxoffice 포함용 normalizing

def normalize(data):
    # 최솟값과 최댓값 계산
    min_val = data.min()
    max_val = data.max()

    # 정규화 수행
    normalized_data = (data - min_val) / (max_val - min_val)

    return normalized_data

norm_columns = df[['google','naver','news','boxoffice']]
data_norm = normalize(norm_columns)

data_norm['date']=df['date']
data_norm.head(2)
```

	google	naver	news	boxoffice	date
0	0.294118	0.614558	0.181818	0.000000	2024-02-22
1	0.426471	0.794749	0.318182	0.037931	2024-02-23

정규화를 위한 normalizing 진행





모델링

험한 것이 나왔다. 😊

분석과제 정의

Q1
온라인 사용자 반응과 관객수의 상관관계를 발견할 수 있을까?

가설

검색어 지수, 리뷰 수 등의 지표는 관객수 증가에 유의한 영향을 끼칠 것이다

분석 방법

구글과 네이버의 검색어 지수(트랜드 검색어) 와 리뷰 수를 카운트해 시각화

Q2
키워드 분석을 통해 새로운 인사이트를 발견할 수 있을까?

가설

이례적인 흥행이라고 생각되는 해당 영화의 독특한 특징이 존재 : 장르(오컬트), 반일 논란 등이 리뷰에서 긍정/부정의 형태로 드러날 것이다

분석 방법

긍정/부정 리뷰 로 label 설정 후 감성분석 진행 키워드 분석, 토픽화

Q3
구축한 모델을 어떻게 이용할 수 있을까?

가설

온라인 영화 평점 및 리뷰 데이터를 학습시킨 머신러닝 모델을 통해 파묘에 대한 짧은 의견의 실시간 감성분석이 가능할 것이다

분석 방법

사용자 리뷰 입력 시 실시간 감성분석을 통해 사용자의 반응을 유추할 수 있을 것이다

트렌드 분석

키워드 및 토픽 분석 / 감성 분석

모델링 및 분석 방법

지수와 관객수의
상관관계 분석

평점 있는 데이터로
평점 없는 데이터를 예측

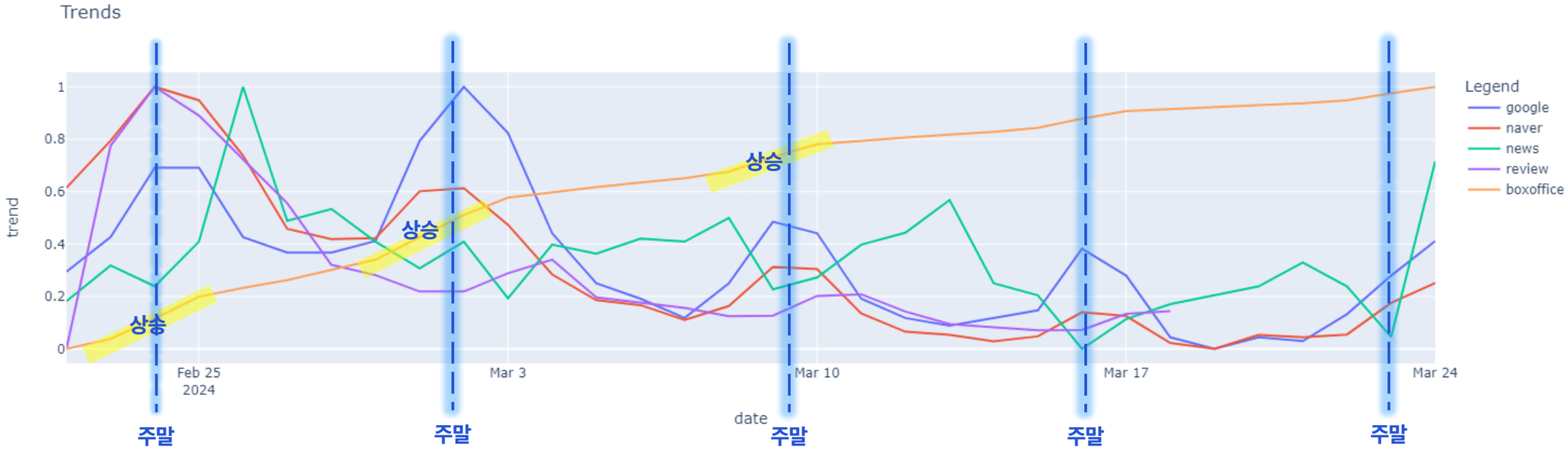
대표 키워드 및
긍정/부정 토픽 분석

분석결과

우리 아이가 달라졌어요! 😊

개봉일로부터 천만관객 돌파일까지의 Trend

박스오피스 지수는 개봉일로부터 천만관객 돌파일까지 완만한 상승곡선을 보임
일정 구간에서 미세한 기울기 변화가 관찰됨
리뷰와 네이버 구글의 두 검색어 지표는 비슷한 주기를 보이는데 비해 뉴스기사 수는 강하고 일정한 상관관계를 보이지는 않음



분석결과

우리 아이가 달라졌어요! 😊

파천내천

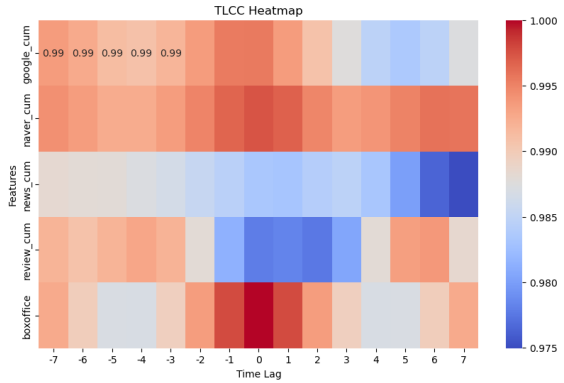
TLCC (Time Lagged Cross Correlation)

y : box office(관객수)를 기준으로 피쳐들의 인과 관계를 시각화

X google 검색어 지수 / naver 검색어 지수 / news 기사 수 / review 리뷰 수

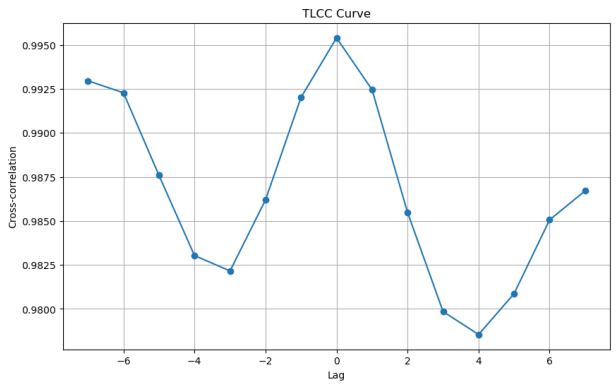
TLCC Heatmap

- Blue (파란색) : 음의 상관 관계
- Red (빨간색) : 양의 상관 관계
- White (흰색) : 0 또는 매우 낮은 상관 관계
- Lag (지연시간) : 음수일 경우 x가 선행할 경우의 관계, 양수일 경우 x가 후행 할 경우의 관계
- corr (상관계수) : 절대값이 1에 가까울수록 상관관계 높음, 양수일 경우 정비례, 음수일 경우 반비례



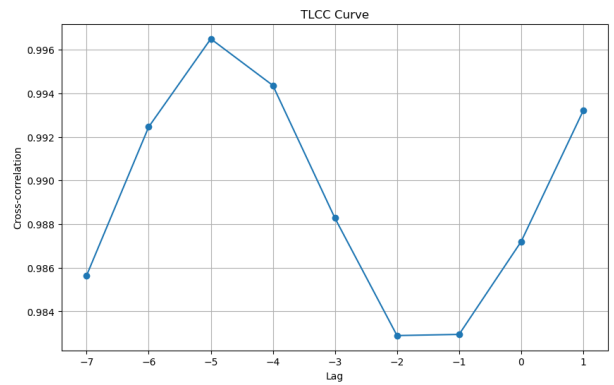
TLCC 네이버 검색어 지수

- 리뷰 데이터와 관객수 데이터는 일 주일 내내 높은 양의 상관관계를 보임
- 리뷰 데이터가 후행지표 역할을 하는 경우 관람객 수 증가 후 0~1일 사이에 리뷰 수가 증가함
- 리뷰 데이터가 선행지표 역할을 하는 경우 3에서 6일 전 리뷰가 관람객수에 가장 큰 영향을 미침



TLCC 리뷰 수

- 검색어 지수와 관객수는 일주일 내내 높은 양의 상관관계를 보임
- 당일의 검색어 지수와 관람객수가 가장 정비례관계가 높고, 3일 주기로 약해짐
- 검색어 지수와 관람객수는 기간이 길어질수록 상관관계가 약해짐.(2일 이전/이후의 검색지수는 관객수에 영향을 덜 미침)



모델링

우리 아이가 달라졌어요! 😊

파천내천

Sentimental Analysis

1단계	2단계	3단계
데이터셋 Cleaned	**데이터셋** SMOTE 활용 리샘플링	**데이터셋** okt 사용자사전 커스터마이징 데이터셋 수기 정제
사용 모델 Logistic Regression SVM, XGBoost Light GBM, Random Forest	**사용 모델** XGBoost SVM	**사용 모델** XGBoost SVM Ensemble
Fine Tuning Grid Search GV	**Fine Tuning** Grid Search GV 5-fold 교차검증 ROC_AUC 곡선 활용 최적의 임계 값 설정	**Fine Tuning** Grid Search GV 5-fold 교차검증 ROC_AUC 곡선 활용 최적의 임계 값 설정
평균 정확도 약 98% **평균 ROC AUC score** 0.5	**평균 정확도** 약 89% **평균 ROC AUC score** 0.7	**평균 정확도** 약 95% **평균 ROC AUC score** 0.95

모델링

우리 아이가 달라졌어요! 😊

파천내천

1단계

대부분의 모델이
부정 분류를 하지 않거나
거의 하지 않았음
(16/20000)

원인 데이터 불균형

총 38870개 데이터 중
긍정(7이상) 32715
부정(4이하) 1327
(약 100:4)

개선방안 데이터 리샘플링

SMOTE를 통한 오버 샘플링

2단계

수기검사 결과 총 **9929행** 중
단 **12개만 부정 리뷰**로 분류,
그 중 **5개 값을 오분류**

원인 원본 데이터의 노이즈

긍정리뷰로 판단되는 텍스트에 **부정적인 점수**를
달거나, 그 반대로 입력된 행이 상당수 존재,
혹은 **의미 없는 텍스트의 나열**

개선방안 데이터 수기 정제

긍정리뷰 10000행 수기 정제:
데이터 행 61개 삭제 후 극단 긍정 값인 **10점만 추출**
긍정:부정 10:1 비율로 정제 후 SMOTE

리뷰	
부모님 보여드렸는데 좋아하세요	0
부모님 보여드렸는데 재미있었다 합니다	0
부모님 선물로 보여드렸는데 좋아하셨어요	0
산으로 산으로 산으로	0
개인적인 의견으로 내용이 산으로 가는기대이하ㅜㅜ	0
너무 기대했었나 생각보다 그저그랬어요	0
생각보다 무섭진않아요	0
생각보다 꿀잼 볼만합니다	0

*2단계 SVM의 오분류 내용 (0은 부정)

8	중반부 까지는 괜찮았으나 후반부로 갈수록 내용이 산으로 감
9	잼 있기는 한데 cg가 영~~~ 그닥
10	이외로 무섭지가 않고 화면이 너무 어두워서 귀신도 안보이네요 사운드만 크고.. 다른 영화에 비해 색
8	초반까지는 좋은데 일본귀신무사가 살아움직이는건 좀...
10	스토리가 좀 허술하긴 하나 배우들 연기력으로 커버됐음
10	소재가 되게 신선하고 좋아서 이건 드라마로 만들었어도 잘 됐겠다 싶었는데 영화여서 뭔가 용두사미
8	굳굳 초반엔 재밌었는데 후반으로 갈수록 조금 지루해지네요

*제거한 원본 데이터의 예시 (사용자입력 리뷰점수, 리뷰내용)

모델링

우리 아이가 달라졌어요! 😊

사용자사전 커스터마이징 코드 일부 (최종코드는 Github 업로드 예정)

뉴스 빅데이터 플랫폼 빅카인즈의 '파묘' 키워드를 수집 후 명사만 골라 명사 사전에 등록
Modifier 등 수식어 사전에서 "여라문" 등 현대에 사용되지 않는 단어를 삭제하고 현대적인 단어만 남김

```
import os

os.chdir('C:/Users/PnM_Media_3/AppData/Roaming/Python/Python311/site-packages/konlpy/java')
os.getcwd()

!jar xvf open-korean-text-2.1.0.jar
```

```
print(okt.pos('천만영화를사랑해오컬트가 좋아요 김고은짱!'))
print(okt.pos('예상치 못한 일제강점기 내용과 장대같이 큰 무언가가 와닿진 않네요 전반적인 재미는 있지만 끼워맞
print(okt.pos('영화보기전에 유퀴즈에서 살짝 스포당했지만 그래도 예상못한 전개와 배우들의 미친 연기력 너무 잘봤
print(okt.pos('한국 오컬트와 역사적 의미가 혼재된 영화해석 미리 안보고 갔음 혼란스러웠을수도'))
print(okt.pos('왜 세워진 관이 있었는가 뭣 땜에 그 위에 묘터를 알려 주었을까 봉은사 처사님은 왜 죽었을까 이해

'''
[('천만', 'Noun'), ('영화', 'Noun'), ('를', 'Josa'), ('사랑', 'Noun'), ('해', 'Noun'), ('오컬트', 'Noun
[('예', 'Modifier'), ('상치', 'Noun'), ('못', 'Noun'), ('한', 'Josa'), ('일제강점기', 'Noun'), ('내용',
[('영', 'Modifier'), ('화보', 'Noun'), ('기전', 'Noun'), ('에', 'Josa'), ('유', 'Noun'), ('퀴즈', 'Noun
[('한국', 'Noun'), ('오컬트', 'Noun'), ('와', 'Josa'), ('역사', 'Noun'), ('적', 'Suffix'), ('의미', 'No
[('왜', 'Noun'), ('세워진', 'Verb'), ('관', 'Noun'), ('이', 'Josa'), ('있었는가', 'Adjective'), ('뭣',
'''

# '영화'를 명사로 학습시켰음에도 불구하고 분리를 제대로 못 함. 한국어 형태소 문법 학습에 관한 문제로 보임.
```

```
0       영화
1       파묘
2       관객
3       감독
4       개봉
      ...
995    일본어
996     조작
```

* 빅카인즈 ' 파묘' 연관 키워드

```
그냥
그따위
그딴
그런
그런그런
그런저런
```

* 형태소 사전 수정

모델링

우리 아이가 달라졌어요! 😊

3단계 모델 코드 일부 (최종코드는 Github 업로드 예정)

Grid Search GV, 5-fold 교차검증(가장 일반적이고 빠른 학습속도), ROC_AUC 곡선 활용 최적의 임계 값 설정, ensemble

```python
import numpy as np
import xgboost as xgb
from sklearn.model_selection import KFold, G
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,

# HyperParameter setting
xgb_param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [3, 6, 9],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0]
}

svm_param_grid = {
    'C': [0.1, 1, 10, 100],
    'gamma': [1, 0.1, 0.01, 0.001],
    'kernel': ['rbf']
}

X = X_resampled
y = y_resampled

# k-fold setting
k_folds = 5
kf = KFold(n_splits=k_folds, shuffle=True, r

xgb_models = []
svm_models = []
```

```python
# Optimal Threshold
optimal_threshold_xgb = 0.44898483
optimal_threshold_svm = 0.6180014186887699

# k-fold
y_pred_proba_xgb_avg = np.zeros(X_resampled.shape[0])
y_pred_proba_svm_avg = np.zeros(X_resampled.shape[0])

for i, (train_index, test_index) in enumerate(kf.split(X_resampled)):
    xgb_model = xgb_models[i]
    svm_model = svm_models[i]

    X_test = X_resampled[test_index]

    y_pred_proba_xgb_fold = xgb_model.predict_proba(X_test)[:, 1]
    y_pred_proba_svm_fold = svm_model.predict_proba(X_test)[:, 1]

    y_pred_proba_xgb_avg[test_index] = y_pred_proba_xgb_fold
    y_pred_proba_svm_avg[test_index] = y_pred_proba_svm_fold

# Ensemble
y_pred_proba_ensemble = (y_pred_proba_xgb_avg + y_pred_proba_svm_avg)
optimal_threshold_ensemble = (optimal_threshold_xgb + optimal_thresho
```

```
XGBoost:
accuracy:  0.9029243292131444
precision: 0.8731155778894473
recall:    0.94287006331022
F1:        0.9066531381359618
ROC AUC:   0.9029243292131444


SVM:
accuracy:  0.9825896894784444
precision: 0.9917063431116572
recall:    0.9733192643955382
F1:        0.9824267782426778
ROC AUC:   0.9825896894784444


Ensemble:
accuracy:  0.9796502864033766
precision: 0.9905951279679309
recall:    0.9684956285800422
F1:        0.9794207317073171
ROC AUC:   0.9796502864033765
```

모델링 및 분석

우리 아이가 달라졌어요! 😊

파천내천

3단계

총1000개 행 수기 검사 결과 오분류 개수

XGB	**26**
SVM	**62**
Ensemble	**42**

원인 과적합

SVM의 모델 평가 지수가 유난히 높았던 것으로 보아
과적합 가능성이 크며,
SVM과의 앙상블도 SVM의 영향을 받아
오분류율이 오히려 높아짐

	XGB	SVM	Ensemble
배우분들 연기짱역사적인 의미가 숨어있는게 대단합니다	1	1	1
이게 왜 천만이야 ㅋ	0	0	0
음양오행 참 신기하기도 특이하기도	1	1	1
기대한 것 만큼은 아닌	0	1	1
정말 긴장감 있게 잘 만들었습니다	1	1	1
긴 말 필요없이 번 봤음	1	1	1
그냥 너무 좋아요 모든게 다 좋아요	1	1	1
조금 무섭지만 잼나게봤어요	1	1	1
마지막 파트가 좀 뜬금없지만 그 전까지는 완벽합니다	1	1	1
너무너무 좋아요 재밌게 잘 봤어요	1	1	1
너무재밋게봤지요오김고은짱이도현짱최민식짱유해진짱	1	0	0
그냥 단순한 귀신얘기가 아니라 재미있게 봤습니다 징글징글한 ㅇㅂ	1	1	1

모델링

우리 아이가 달라졌어요! 😊

키워드 및 토픽 분석

키워드 분석	군집 수 결정	토픽 LDA
벡터화 CountVectorizer	**벡터화** 불용어 제거 후 Countvectorizer (불용어로 인해 군집화가 잘 되지 않음)	**벡터화** 불용어 제거 후 Countvectorizer (불용어로 인해 군집화가 잘 되지 않음)
분석 방법 Logistic Regression Coefficient	**분석 방법** Elbow Silhouette	**분석 방법** Latent Dirichlet Allocation
시각화 Word Cloud	**Fine Tuning** Elbow Silhouette	**시각화** pyLDAvis

모델링

우리 아이가 달라졌어요! 😊

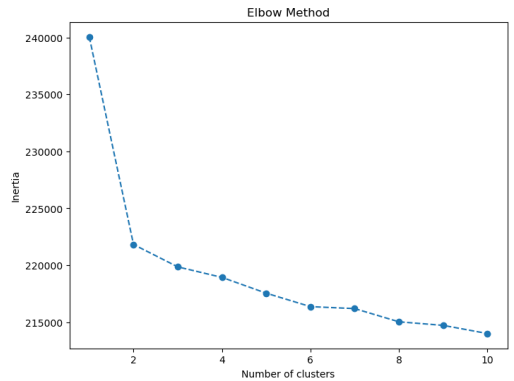
파천내천

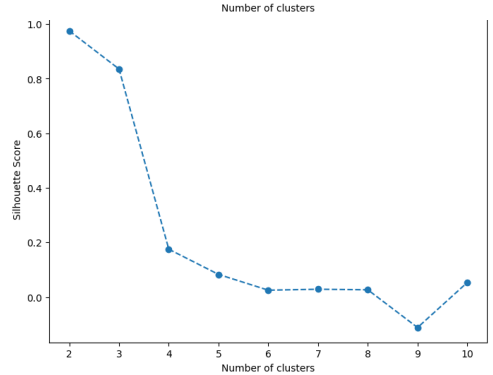
키워드 분석

LR 분류 모델의 상관계수 값으로
단어를 긍정/부정으로 나누어 정렬

군집 수 결정

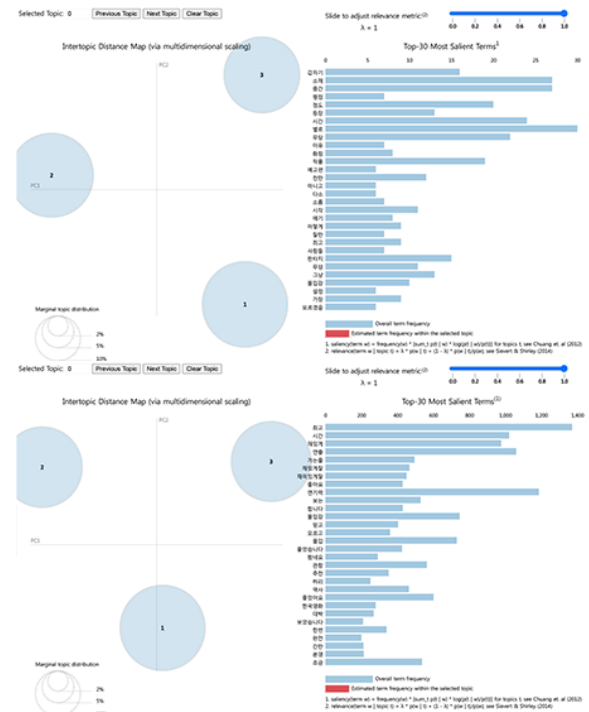
긍정 리뷰의 Elbow, Silhouette 계수 적용
시각화를 통해 후보 군집 수 선정





토픽 LDA

후보 군집수를 바탕으로 시뮬레이션,
3개 군집 선택



분석결과

우리 아이가 달라졌어요! 😊

파천내천

토픽 분류 결과

토픽 분류 결과 긍정 / 부정 감정의 사용자의 키워드 파악, 향후 각각 다른 홍보 전략 구축 가능

Positive Review Topics

Topic # 0 스토리에 대한 긍정리뷰 : 😱몰입감 😳긴장감

[시간 연기력 몰입감 가는줄 몰입 긴장감 재미있게잘 관람 좋아요 합니다]

Topic # 1 소재에 대한 긍정리뷰: KR애국 🤡귀신

[재밌게잘 장르 한국 역사 귀신 이야기 일본 소재 중반 좋았어요]

Topic # 2 연출과 연기에 대한 긍정리뷰 : 🎬장재현 😎김고은 😎최민식

[최고 재밌게 연출 연기력 후반부 보는 한국 믿고 좋았습니다 역시]

Negative Review Topics

Topic # 0 스토리에 대한 부정리뷰 : 😒후반부 👹일본귀신

[후반부 장르 중간 곡성 귀신 전개 중반 장재현]

Topic # 1 감독의 이전 작품과 비교 : 🎬사바하 🎬검은사제들

[별로 초반 중반 시간 갑자기 최민식 사바하 좋은]

Topic # 2 유사 작품과 비교 : 🎬곡성 👺무당

[귀신 장르 정도 연출 한국 작품 무당 재미 곡성]

긍정 리뷰

장르	소재	인물
# 몰입감 #긴장감	# 애국 # 귀신	# 장재현 # 김고은 # 최민식

부정 리뷰

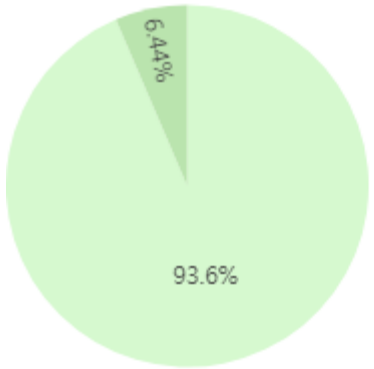
스토리	필모그래피	장르
# 일관성 # 일본	# 사바하 # 검은사제들	# 곡성 # 무당

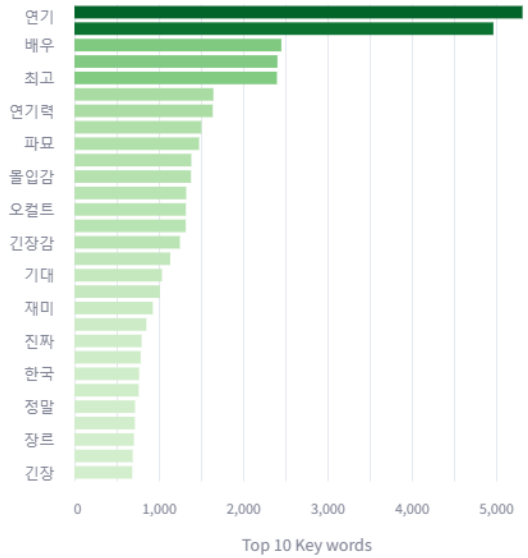
분석결과

우리 아이가 달라졌어요! 😊

리뷰 분석결과 시각화

ML 감성분석 결과를 합해 총 10만개 리뷰 데이터 분류: 긍정리뷰 94% 대 부정리뷰 6%로 구성됨
전체 리뷰의 최빈 명사 키워드 추출 : 배우, 연기, 영화, 스토리에 대한 리뷰가 긍정/부정 리뷰 모두에 압도적으로 많음
‘공포’, ‘오컬트’는 긍정 키워드에도 높은 빈도로 등장, 흥행요인으로 볼 수 있음





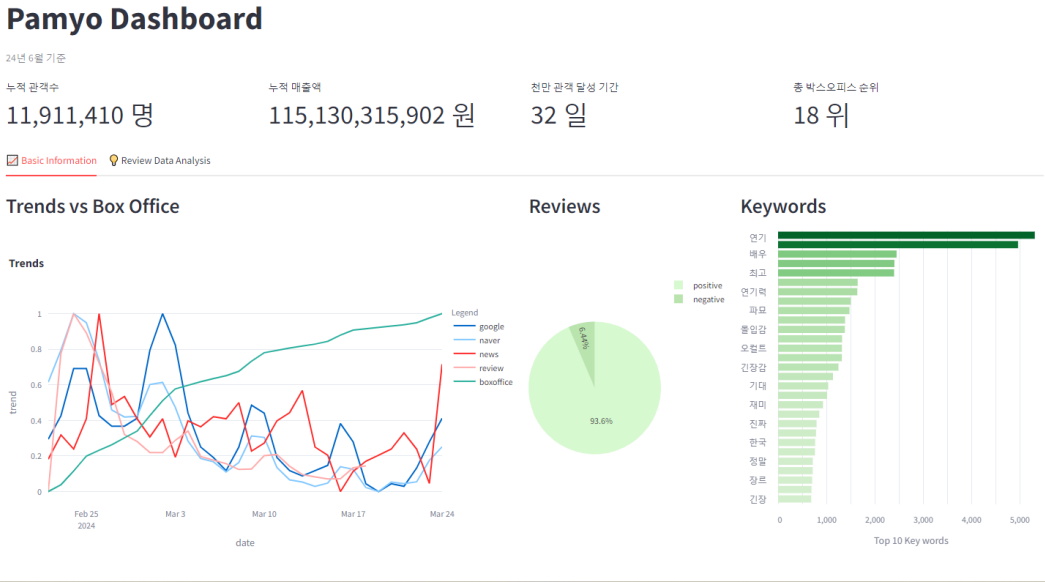
분석결과

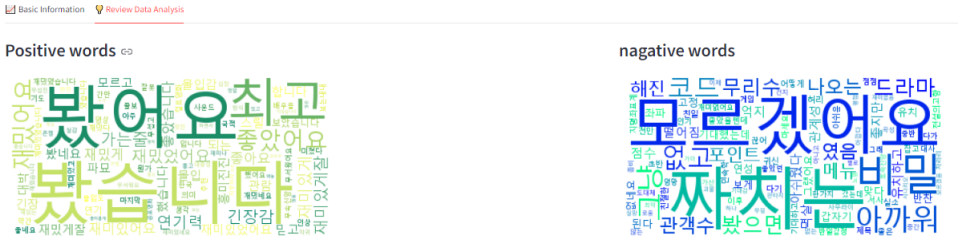
우리 아이가 달라졌어요! 😊

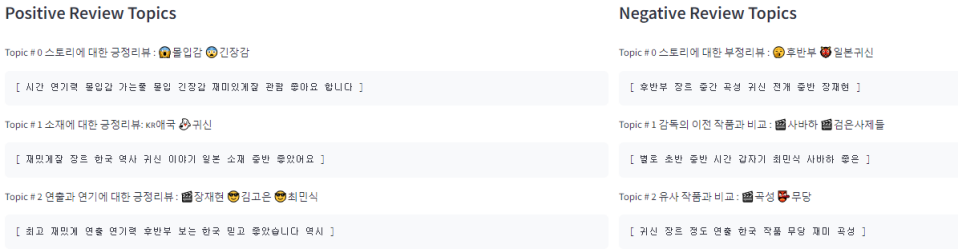
파천내천

대시보드 제작

API연동으로 관객수, 매출액을 한눈에 파악
리뷰 분석 정보를 제공하여 다른 관객들은 어떻게 봤는지 분석 결과 정리







	리뷰	평점	Topic	Percentage
0	오랜만에 극장에서 볼 만한 영화였다.	10	0	0.7456
1	파묘다...!! 키즈네무서워	10	2	0.5525
2	2가지 맛으로 즐기는 새로운 오컬트	10	2	0.5247
3	듀마 어벤져스 4명 ,연기보는 맛이 있네요	9	2	0.7775
4	너무 재밌어서 3번 봄	10	0	0.7775
5	새로운 관점 그것을 넘어 섬	10	2	0.5817
6	볼만합니다재미있습니다	10	0	0.772
7	갑자기 도깨비의 등장으로 몰입이 깨졌지만 배우들의 연기는 인상적	8	2	0.6665
8	신기하고 흥미로워요 재밌어요	10	1	0.8319
9	굿뜨. 다들 연기가미침	10	0	0.8576

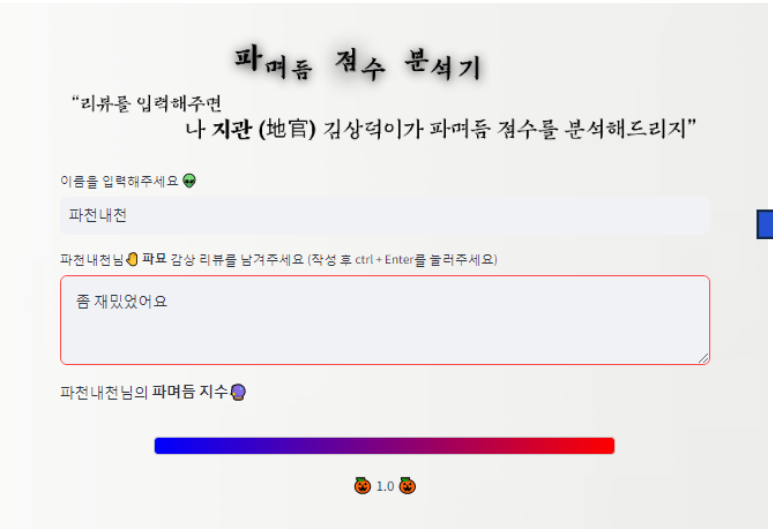
	리뷰	평점	Topic	Percentage
0	기대가 너무 컸나 봐요	2	1	0.8536
1	인기 많아서 봤는데 별로	3	1	0.6054
2	생각보다많이심심해용	1	1	0.8593
3	연기는 좋은데 후반부는 별로	1	2	0.897
4	음 홍보에 낚인 기분..., 별로 아니 그 이하 였어요	2	0	0.8772
5	강렬한 장면장면과 이미지, 배우들의 열연에 비해	4	2	0.8286
6	난 너무 재미없던데 —— 내용 어이상실——	1	0	0.9049
7	글쎄요... 스토리가 좀 별로네요	3	1	0.8462
8	진짜 개노잼입니다 왜 천만인지 이해가안감	1	1	0.7617
9	이도저도 아닌 영화	1	0	0.8232

분석결과

우리 아이가 달라졌어요! 😊

리뷰 감성분석 웹앱 프로토타입

파묘를 본 관람객을 대상으로 리뷰를 입력하면 감성분석 및 분류 서비스를 제공
사전 토픽 분석 결과를 활용하여 관심이 있을 영화 정보를 추가로 제공



감성분석 모델

분석결과 적용

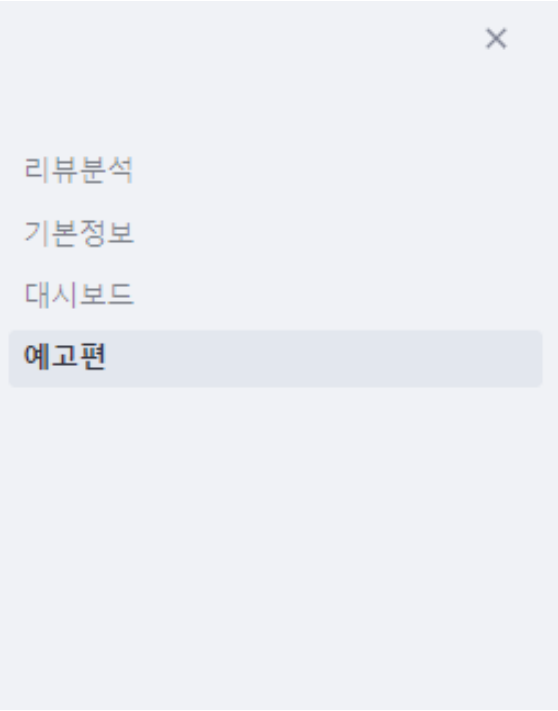
토픽 LDA

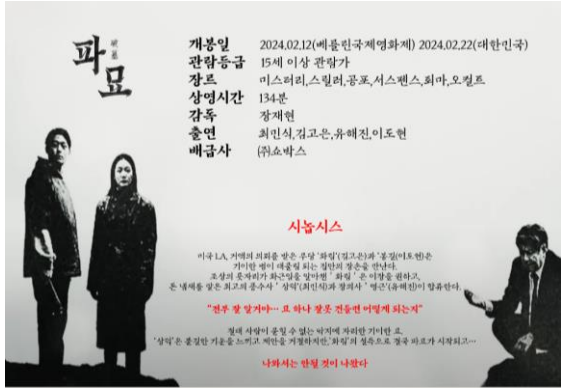
분석결과

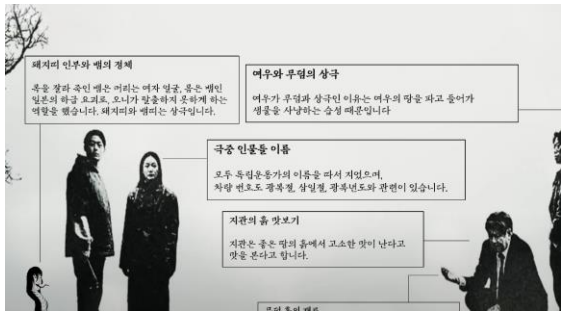
우리 아이가 달라졌어요! 😊

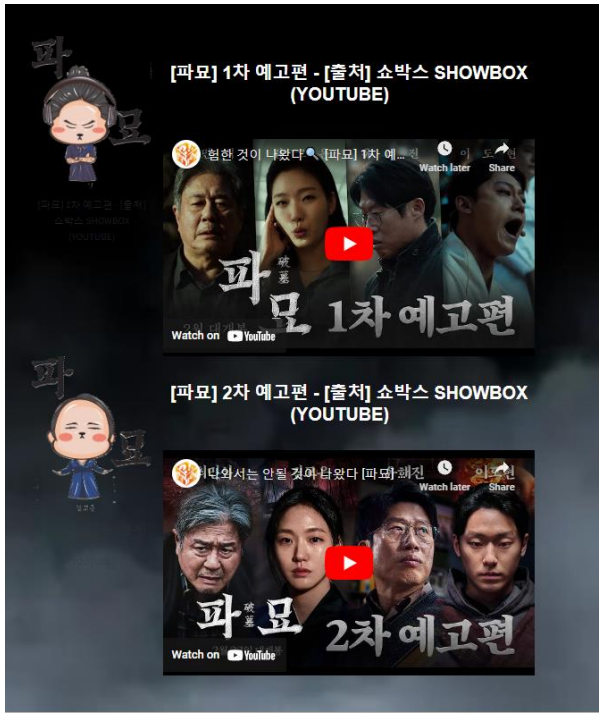
기본 정보 페이지 제작

시놉시스, 영화 정보, 비하인드 등 기본 정보 페이지 제작
예고편 등 소개 페이지 제작









모델링

우리 아이가 달라졌어요! 😊

분석결과 summary

Q1 온라인 사용자 반응과 관객수의 상관관계를 발견할 수 있을까?

가설

- 검색어 지수, 리뷰 수 등의 지표는 관객수 증가에 유의한 영향을 끼칠 것이다

분석결과

- 검색어와 리뷰 지수 모두 관객수에 정비례함
- 일정한 기간을 두고 선행지표가 되기도 하고 후행지표가 되기도 하며 높은 상관관계를 보여줌
- 기사수는 뚜렷한 상관관계를 보여주지 않음

Q2 키워드 분석을 통해 새로운 인사이트를 발견할 수 있을까?

가설

- 이례적인 흥행이라고 생각되는 해당 영화의 독특한 특징이 존재 : 장르(오컬트), 반일 논란 등이 리뷰에서 긍정/부정의 형태로 드러날 것이다

분석결과

- '공포'에 관련한 키워드는 긍정/ 부정 모두에 높게 나타남 (공포 키워드는 긍정, 즉 흥행요인)
- '일본, 반일'에 관련한 키워드는 부정 키워드에 주로 등장
- 역사, 우리나라, 한국 등은 긍정 키워드에 등장

Q3 구축한 모델을 어떻게 이용할 수 있을까?

가설

- 온라인 영화 평점 및 리뷰 데이터를 학습시킨 머신러닝 모델을 통해 파묘에 대한 짧은 의견의 실시간 감성분석이 가능할 것이다

분석결과

- 감성분류기 웹앱을 구현하여 실시간 리뷰를 받고, 긍정/부정의 결과에 따라 홍보 방안을 다르게 적용
- 파묘 대시보드를 제작하여 비즈니스 관계자들이 분석 결과를 쉽게 확인할 수 있도록 함

모델링

우리 아이가 달라졌어요! 😊

자체 피드백

학습 데이터가 '한줄평' 형식의 리뷰에는 아주 잘 작동.
하지만 '댓글 분석'에서 부정리뷰에 대한 정확도가 매우 낮았음.

=> 두 플랫폼의 사용자 태도의 차이가 존재함.
리뷰 사이트의 한줄평은 긍정 리뷰의 경우 형태소 중 서술어를 기준으로 강한 상관계수를 갖고있었음.
어미가 경어체 ('했습니다' '재밌었어요' '봤습니다')로 끝나는 경우 강한 긍정 확률을 보임.

-유튜브 댓글의 경우 긍정 반응이어도 대부분 평어체를 사용,
'기대평'과 '후기' 사이에 사용되는 용어의 차이가 컸음. '기대된다' '보고싶다' / '재밌었습니다' '봤습니다'

보완 방안

<파묘>뿐 아니라 유사한 장르의 다양한 영화를 분석할 수 있도록 학습 데이터를 구축하고 모델을 확대

모델링

우리 아이가 달라졌어요! 😊

Reference

도서(e-book포함)

<딥 러닝을 이용한 자연어 처리 입문> 유원준 외, 위키독스
<파이썬 라이브러리를 활용한 텍스트 분석> 젠스 알브레히트 외, 한빛미디어
<데이터커뮤니케이션: 텍스트마이닝> 안도현
<데이터 사이언스 스쿨>, 김도형

논문

영화

풍원, 공포영화를 보는 즐거움과 괴로움에 대한 수용자 반응의 차이 연구, 2019
한상윤, 한국 공포영화의 오컬트 장르 초기 수용 양상 연구, 2017
정경석 외, 한국 공포영화의 흥행요인 분석 - <곤지암>을 중심으로 2019
이은의, 공포인자의 특징에 따른 한국 공포영화의 분류 및 흥행과의 연관성 연구, 2006
ZHAO WENJIN, 호러(Horror)시청 공포감이 시청 즐거움에 미치는 영향: 심리적 거리감의 매개효과 중심으로, 2019
김연형, 영화 흥행 결정요인과 흥행성과 예측 연구, 2011
황예나 외,흥행영화의 온라인 구전패턴과 관객수의 관계,2019

모델링

우리 아이가 달라졌어요! 😊

파천내천

Reference

논문

모델링

전성현, 영화흥행 예측변수로서 온라인 구전변수의 효과, 2016
우종필 외, 빅데이터 분석을 통한 천만 관객 영화 예측 모델, 2018
전은정, LDA 토픽모델링의 적정 표본크기 분석 연구 고교학점제 뉴스기사를 중심으로, 2023
이강우 외, 센서별 시간지연 교차 상관관계를 이용한 GCN기반의 시계열 데이터 이상 탐지 방법, 2023
김점구 외, 랜덤 포레스트 분류기를 이용한 감성어 사전 기반 감성 분석 개선. 2024

전처리

고광호, 딥러닝을 위한 텍스트 전처리에 따른 단어벡터 분석의 차이 연구, 2023
이상훈 외, 영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석, 2016