2020년도 2학기 비정형데이터분석 Final Presentation

유튜브의 영화 예고편 영상 댓글 분석을 통한 영화 마케팅 전략 수립

Department of Data Science
Seoul National University of Science and Technology
강지철 황인원

SEOULTECH

Contents

연구 주제 소개

연구 주제 소개

- 유튜브 영화 영상 댓글에는 대개 콘텐츠에 대한 평가와 감성이 드러난다.
  - 유저의 감성 상태는 유저의 실제 영화 시청에 영향을 미칠 수 있다.
- 이 연구에서는 유튜브의 영화 예고편 댓글에 감성 분석을 수행하고, 다양한 영화의 흥행과 관련한 지표와 비교해봄으로써 어떤 방향의 영화 마케팅 전략을 수립하는 것이 적절한 지 탐색해본다.
- 유튜브의 경우 영상물인 영화와 직접적으로 관련이 있는 플랫폼이며, 다른 플랫폼에 비해서도 압도적인 이용률을 보이므로 조사의 대상으로 선정하였다.

https://www.koit.co.kr/news/articleView.html?idxno=78572

연구 주제 소개

주제 변경 이유

- Reddit은 커뮤니티 특성 상 특정 주제를 가진 게시판에 들어가도 주제와 관련 없는 데이터가 지나치게 많다.
- 미국의 대선이라는 주제로 분석을 수행할 때, 이미 결과가 나온 사후 분석의 한계도 가지고 있다.

➔ 비교적 명확한 글의 작성 의도를 가지고 있으며 분석이 의미를 갖는 유튜브 댓글 데이터를 분석 대상으로 선택했다.

연구 주제 소개

선행 연구 분석

- 영화의 댓글이나 리뷰 분석은 오래 전부터 다양한 방법으로 연구되고 있지만, 실제로 영화 산업에 이를 이용하는 방향으로 진행된 연구는 많지 않다.
- 영화라는 동일 주제에 대한 감성 분석을 다룬 오영택 외. (2019)의 연구에서는 한국어 공개 데이터셋인 Naver sentiment movie corpus(NSMC)에 감성 분석을 수행하고 RNN 모델을 적용함으로써 높은 성능을 보였다.

Parallel Stacked Bidirectional LSTM
모델을 이용한 한국어 영화리뷰 감성 분석
(Korean Movie-review Sentiment Analysis Using
Parallel Stacked Bidirectional LSTM Model)

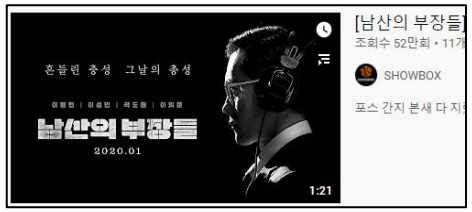
오 영 택 † 김 민 태 † 김 우 주 ††
(Yeongtaek Oh) (Mintae Kim) (Wooju Kim)

연구 과정

연구 과정

KOBIS
영화관입장권통합전산망

	영화명	매출액
1	이웃사촌	118,024,470 원
2	도굴	53,345,470 원
3	런	49,775,000 원
4	더 프롬	26,332,210 원
5	파티마의 기적	14,199,600 원
6	프리키 데스데이	13,224,050 원



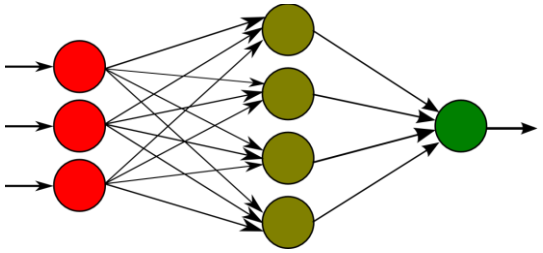
	Comment
173	갑자기영화'유령'이떠오르네ㅎㅎㅎ
383	위안부등수진놈들이누구더라?ㅋㅋ
172	"모든패는북환이꽉쥐고있습니다"봉딱섹스
26	대한민국대통령...한경재입니다...한마디하겠습니다...박근혜앞으로나와!!!!!!!!...
131	ㅋㅋ영화망함~~
180	2는멀손잡고지디콘서트가는거아냈나
435	정우성저역할하는동안아주신나있었겠네
129	0.5UBD응원합니다ㅎ
153	예고편만보고역겁기도힘든데...
157	유연석아닌줄ㅋㅋㅋㅋ넘마재밌겠다기대기대
346	뭐냐왜곽도원이북한군인으로나옴?

영화의 예고편 댓글 정보 추출

네이버 영화 리뷰 감성 분석 데이터 학습

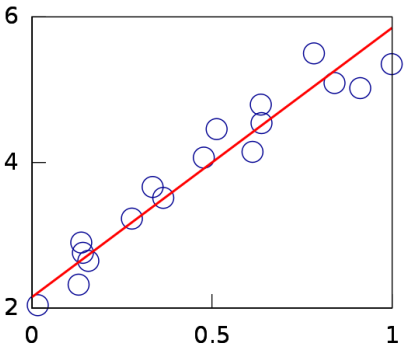
Naver sentiment movie corpus v1.0

```
$ head ratings_train.txt
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차
10265843        너무재밌었다그래서보는것을추천한다
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```



사전 학습된 모델로 예고편 댓글 감성 분석

영화의 흥행도, 상영관 수와의 관계 분석



데이터 수집

데이터 수집

데이터 수집

- 영화 박스오피스 집계 사이트인 KOBIS에서 2020년 출시 영화 약 1800개의 매출을 포함한 상세 데이터를 추출한 뒤, 분석 가능한 지표를 가진 상위 35개 영화를 선택했다.
- 유튜브에서 해당 영화의 예고편을 검색한 뒤, Selenium과 BeautifulSoup 패키지를 이용해 웹 크롤링으로 댓글을 수집한다.
- 총 35개 영화를 대상으로 7412개의 댓글이 수집되었다.

순위	영화명	개봉일	매출액	매출액	누적매출액	관객수	누적관객수	스크린수	상영횟수	대표국적	국적
18.0	결백	2020-06-10	7,859,629,340	1.7%	7,859,629,340	894,025	894,025	1,112	74,532	한국	한국
112.0	이 멋진 세계에 축복을! 붉은 전설	2020-02-06	152,082,580	0.0%	158,330,580	17,607	18,175	57	1,130	일본	일본
39.0	뮬란	2020-09-17	2,079,737,620	0.4%	2,079,737,620	236,247	236,247	1,420	38,703	미국	미국
5.0	테넷	2020-08-26	18,385,436,230	4.0%	18,385,436,230	1,990,948	1,990,948	2,228	164,568	미국	미국

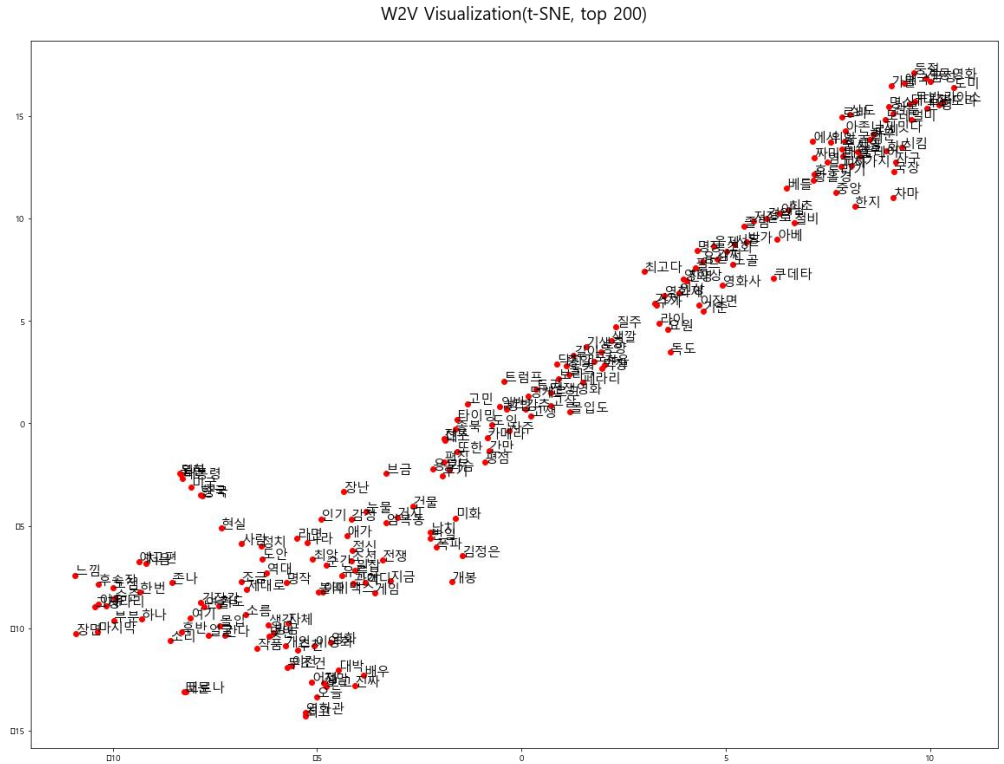
	Comment
173	갑자기영화'유령'이떠오르네ㅎㅎㅎ
383	위안부통수친놈들이누구더라?ㅋㅋ
172	"모든패는북환이꽉쥐고있습니다"뷰띵섹스
26	대한민국대통령...한경재입니다...한마디하겠습니다...박근혜앞으로나와!!!!!!!!...
131	ㅋㅋ영화망함~~
180	2는딸손잡고지디콘서트가는거아녔냐
435	정우성저역할하는동안아주신나있었겠네
129	0.5UBD응원합니다ㅎ
153	예고편만보고역겹기도힘든데...
157	유연석아닌줄ㅋㅋㅋㅋ넘마재밌겠다기대기대
346	뭐냐왜곽도원이북한군인으로나옴?

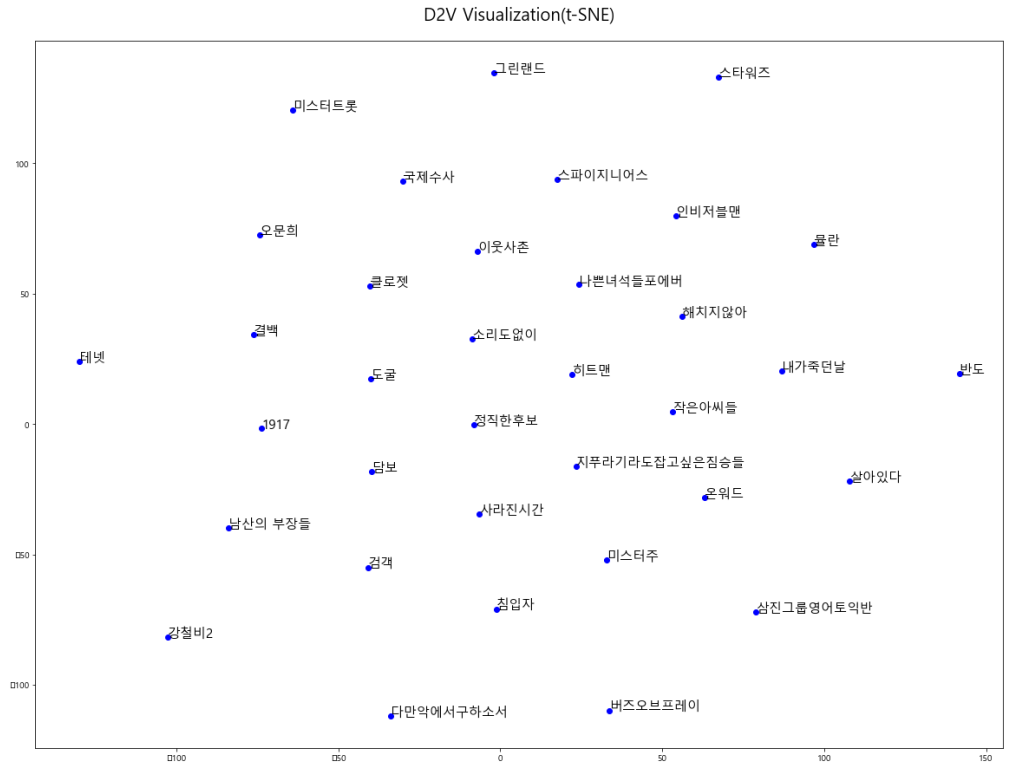
데이터 분석 및 결과

데이터 분석 및 결과

데이터 시각화

- Word2Vec 및 Doc2Vec(50 dimension, t-SNE)





데이터 분석 및 결과

데이터 시각화

- Word cloud

1917

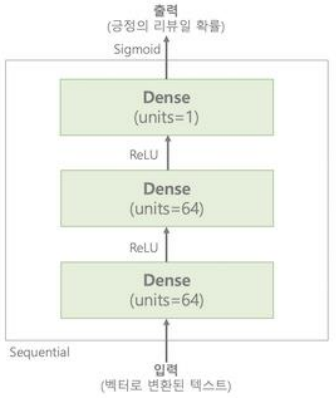
강철비

국제수사

데이터 분석 및 결과

네이버 영화 리뷰 감성 분석 데이터 학습

- Naver sentiment movie corpus(NSMC)
- 네이버 영화의 리뷰 중 영화당 100개의 리뷰를 모아 총 200,000개의 리뷰로 구성되어 있고, 학습에 15만개의 리뷰를 이용했다.
- 중립적인 평점(5~8)은 제외하고 긍정(9~10점)과 부정(1~4점) 리뷰만을 데이터에 포함했다.
- KoNLPy 라이브러리의 Okt(Open Korean Text)클래스를 이용해 형태소 분석 및 품사 tagging을 수행한다.
- 자주 사용되는 토큰 10,000개를 사용해 데이터를 벡터화한다.
- 64개의 유닛을 가지는 2개의 Dense 층으로 구성된 RNN 모델로 학습을 수행. 처음 두 개의 층은 relu, 마지막 층은 sigmoid 활성 화 함수를 사용해 긍정의 리뷰일 확률을 출력한다.
- 손실함수로는 binary_crossentropy를 이용하고, RMSProp optimizer를 이용해 경사하강법을 수행했다.
- 배치사이즈는 512, epoch은 10회로 학습한다.



데이터 분석 및 결과

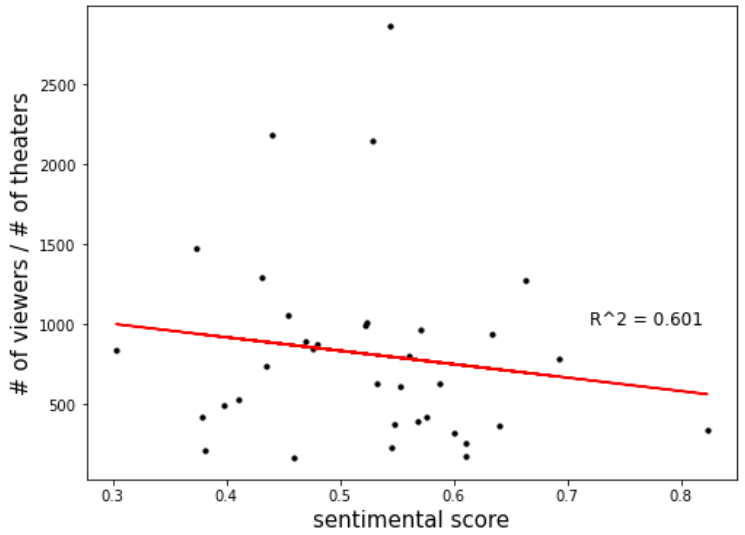
유튜브 영화 댓글에 대해 감성 분석 수행

source	comment	score
테넷	이영화는그냥영화가아닙니다.	0.290406
그린랜드	마지막에다살아요	0.675281
그린랜드	C-QC-Q여기는그린란드누군가있습니까?	0.525467
반도	강동원...좀좋은작품컨택하지...매번ㅜㅜ	0.921800
그린랜드	이시국이라보면스트레스받을듯...	0.217929
살아있다	천만가즈아	0.054916
반도	좋아요	0.923208
반도	보고오니깐예고편알겠다예고편에다담겨있네잘봅았당	0.062034
스타워즈	2019년12월이라썼는데,우리나라만2020년1월이라읽는다...ㅠㅠ	0.187496
살아있다	이거코믹영화라던데	0.202146
스타워즈	한국어번역부탁	0.429704
강철비2	ㄹㅇ나만알고있는줄알았는데댓글에서다마카롱티비얘기하니까신기하네	0.348720
반도	드디어	0.711828
지푸라기라도잡고싶은짐승들	내인생이네.	0.698613
작은아씨들	시얼샤사랑해.....벌써눈물그리는중ㅠㅠ	0.994066
살아있다	Woahhh,ican'twait	0.419945

데이터 분석 및 결과

예고편 댓글 감성 분석 점수와 스크린 수 대비 관객 수와의 관계

- R^2 value: 0.601

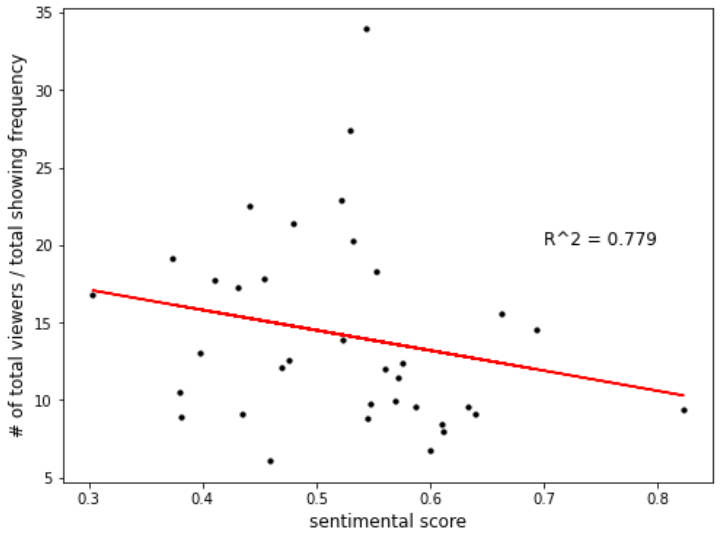


스크린 수 대비 관객 수

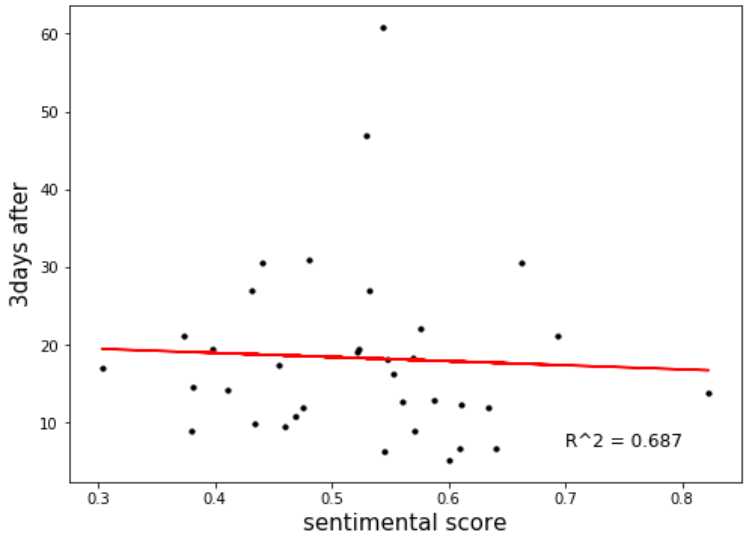
데이터 분석 및 결과

예고편 댓글 감성 분석 점수와 상영횟수 대비 관객 수와의 관계

- R^2 value: 0.779, 0.687



상영횟수 대비 관객 수(전체 기간)

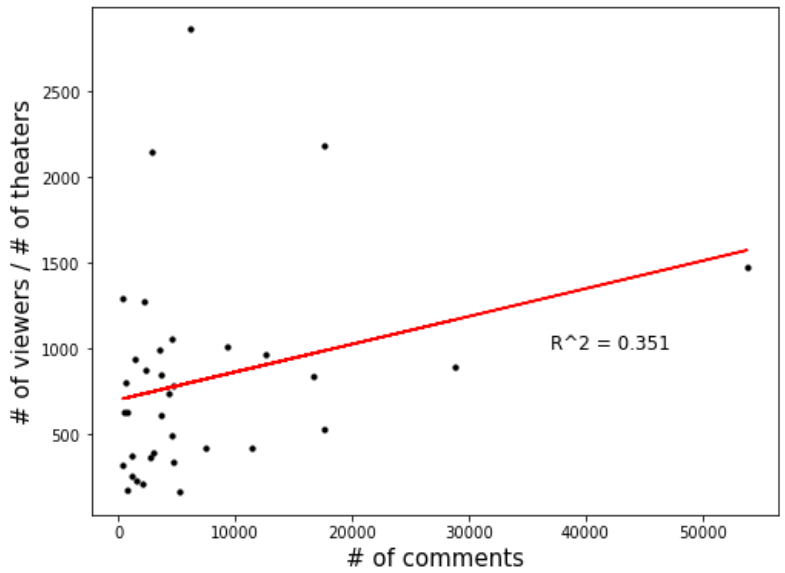


상영횟수 대비 관객 수(개봉 후 3일)

데이터 분석 및 결과

예고편 댓글의 총 길이와 스크린 수 대비 관객 수와의 관계

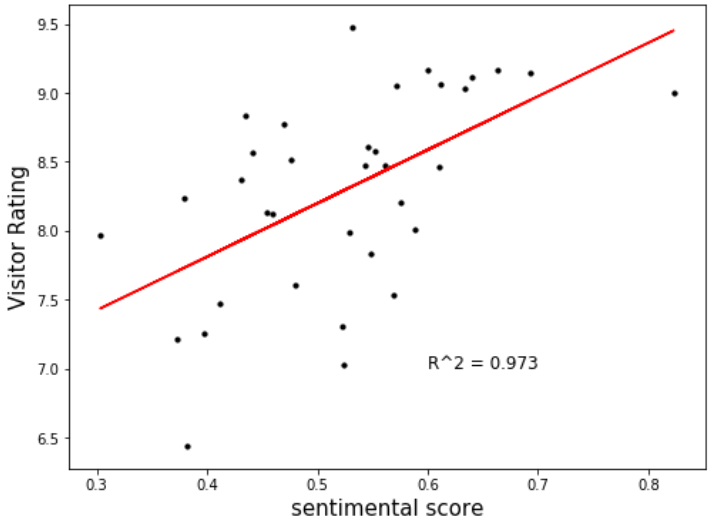
- R^2 value: 0.351



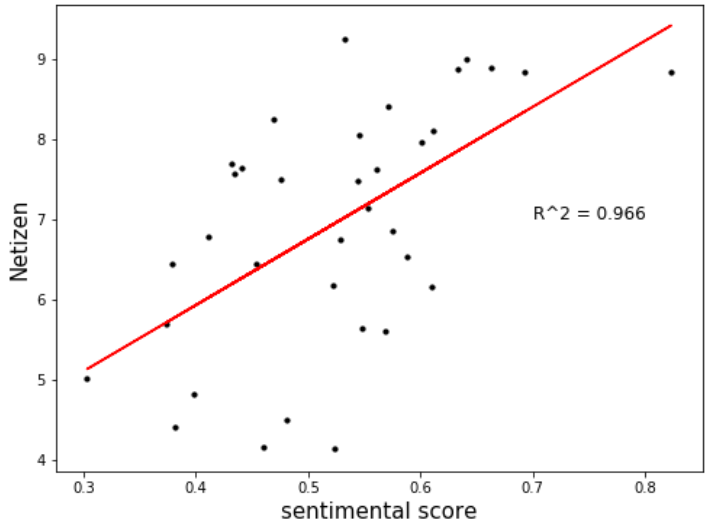
데이터 분석 및 결과

예고편 댓글 감성 분석 점수와 포털사이트 영화 평점의 관계

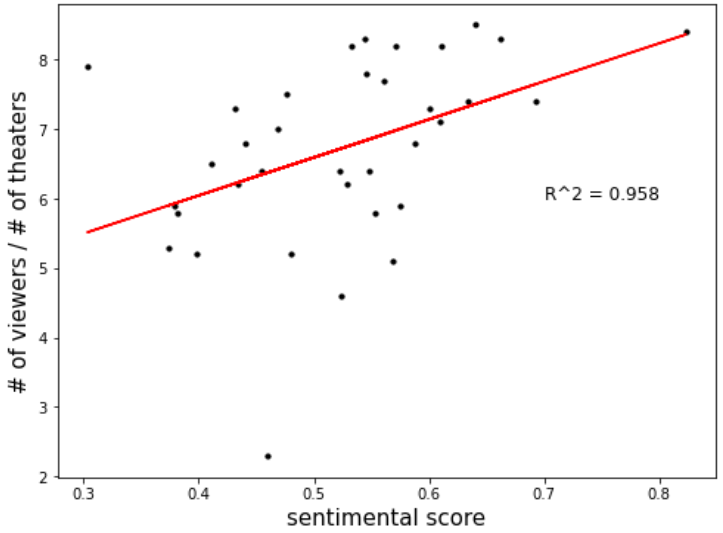
- R^2 value: 0.973, 0.966, 0.958



네이버 관람객 평점



네이버 네티즌 평점



다음 영화 평점

데이터 분석 및 결과

R^2 Table

	감성 분석 점수와의 관계(R^2 score)
총 관객 수 / 스크린 수	0.601
총 관객 수 / 상영횟수(전체 기간)	0.779
총 관객 수 / 상영횟수(개봉 직후 3일)	0.687
네이버 관람객 평점	0.973
네이버 네티즌 평점	0.966
다음 영화 평점	0.958

기타) 예고편 댓글의 총 길이와 스크린 수 대비 관객 수와의 관계: 0.351

고찰

고찰

결과 분석

- 감성 분석 점수와 스크린 수 대비 상영관 수와는 크게 관계가 없으며 오히려 음의 관계를 보이는 경향이 있다.
- 상영 횟수 대비 감성 분석 점수는 스크린 수 대비 상영관 수보다는 조금 더 높은 관계를 보이지만, 음의 관계를 보이며 그 관계의 정도가 높다고 볼 수 없다.
- 포털사이트의 관람객 또는 네티즌 평점과는 상당히 높은 양의 관계를 보이며, 두 종류의 대형 포털에서 유사한 결과를 보인다.

영화 마케팅 전략 수립 방안

- 조사한 지표가 실제 관람객 수와 관련이 있는 지표들과 관계가 높다면, 스크린 수와 상영 횟수를 더 확보하거나 오프라인 마케팅을 강화할 수 있다.
- 조사한 지표가 포털사이트의 평점과 관계가 높다면, 각종 온라인 마케팅이나 영화관 상영이 끝난 후 OTT 플랫폼에 대한 마케팅과 관련이 있다고 볼 수 있다.
- 영화 예고편 댓글 영상의 감성 분석 점수 지표는 포털사이트의 평점과 관계가 높으므로, 이 지표만으로 상영관이나 스크린 수를 무리하게 확보하기 보다는 추후 OTT 플랫폼에 대한 마케팅에 집중하는 전략을 수립할 수 있다.

고찰

한계

- 예고편 영상의 댓글을 막아둔 영화가 존재한다.
- 댓글 수의 차이에 대한 해석 차이가 있을 수 있다.
- 감성 분석 결과를 검증하는 것에 어려움이 있다.
- 전염병의 영향으로 영화에 대한 기대치가 실제 관람으로 이어지지 않을 가능성이 존재한다.

Open issues

- 해당 영상들에 대한 조회수나 좋아요 수, 다른 관련 영상의 데이터를 추가적으로 수집해 볼 수 있다.
- 학습 모델에 대해 여러 변수를 조정해볼 수 있다.
- 포털사이트의 평점과 OTT 플랫폼의 소비율의 관계를 구체적으로 조사해볼 수 있다.
- 실제 관람객 수와 관련한 지표는 어떤 지표와 관련이 있을 지 조사해볼 수 있다.

참고 문헌

- 오영택 외. (2018). Parallel Stacked Bidirectional LSTM 모델을 이용한 한국어 영화리뷰 감성 분석. 정보과학회논문지 46(1), 2019.1, 45-49.
- https://www.koit.co.kr/news/articleView.html?idxno=78572

NSMC dataset

- github.com/e9t/nsmc

RNN model reference

- nbviewer.jupyter.org/github/cyc1am3n/Deep-Learning-with-Python/blob/master/Chap03-getting_started_with_neural_networks/Chap03-Extra-classifying_korean_movie_review.ipynb

Images

- commons.wikimedia.org/wiki/File:MultiLayerPerceptron.png
- magoosh.com/data-science/what-is-a-regression-model