줄거리를 활용한 영화 관람등급 예측

NLP 프로젝트 : 이수현 이채영 한창헌

추진 배경

[영화 등급 분류 수 추이]

등급	평균 매출	평균 이익	평균투자수익률
전체관람가	3.105	615	24.7%
12세 관람가	8.462	914	12.1%
15세 관람가	8.262	1483	21.9%
청소년 관람 불가	3.451	-1384	-21.6%

[등급 별 한국 영화 매출, 이익, 수익률]

영화 줄거리를 통한 영화 관람 등급 예측

Data Set

Story

네이버 영화 '줄거리' 에 소개된 내용을 크롤링

Movie

영화에 관한 간략한 정보

title, e_title, running_time

Grade

영화 관람등급

	title	e_title	running_time	story	grade
0	시네마 천국	Cinema Paradiso , 1988	124.0	어린 시절 영화가 세상의 전부였던 소년 토토는 학교 수업을 마치면 마을 광장에 있는...	전체 관람가
1	백 투 더 퓨쳐	Back To The Future , 1985	120.0	힐 밸리(Hill Valley)에 사는 주인공 마티 맥플라이(Marty McFly:...	12세 관람가
2	백 투 더 퓨쳐 2	Back To The Future Part 2 , 1989	107.0	브라운 박사(Dr. Emmett 'Doc' L. Brown: 코리스토퍼 로이드 분)...	12세 관람가

31743

Story

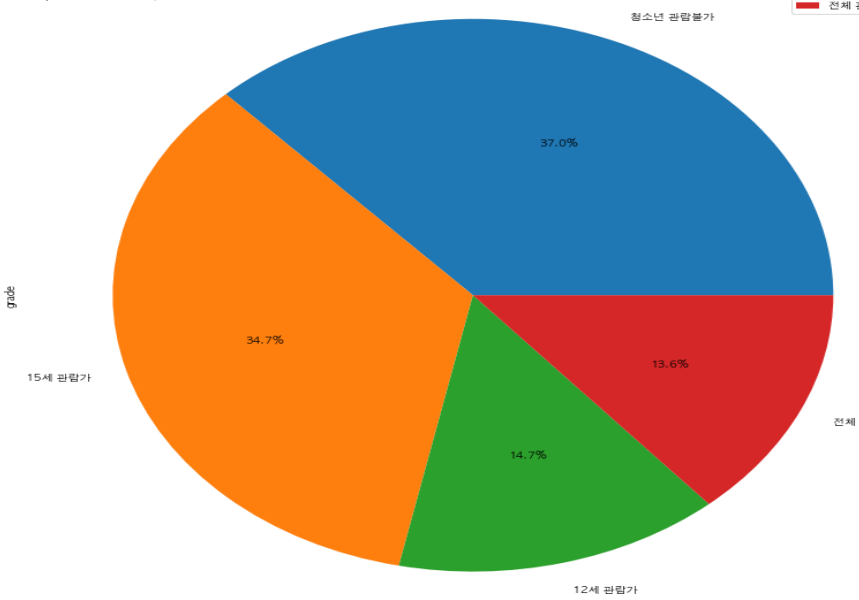
4

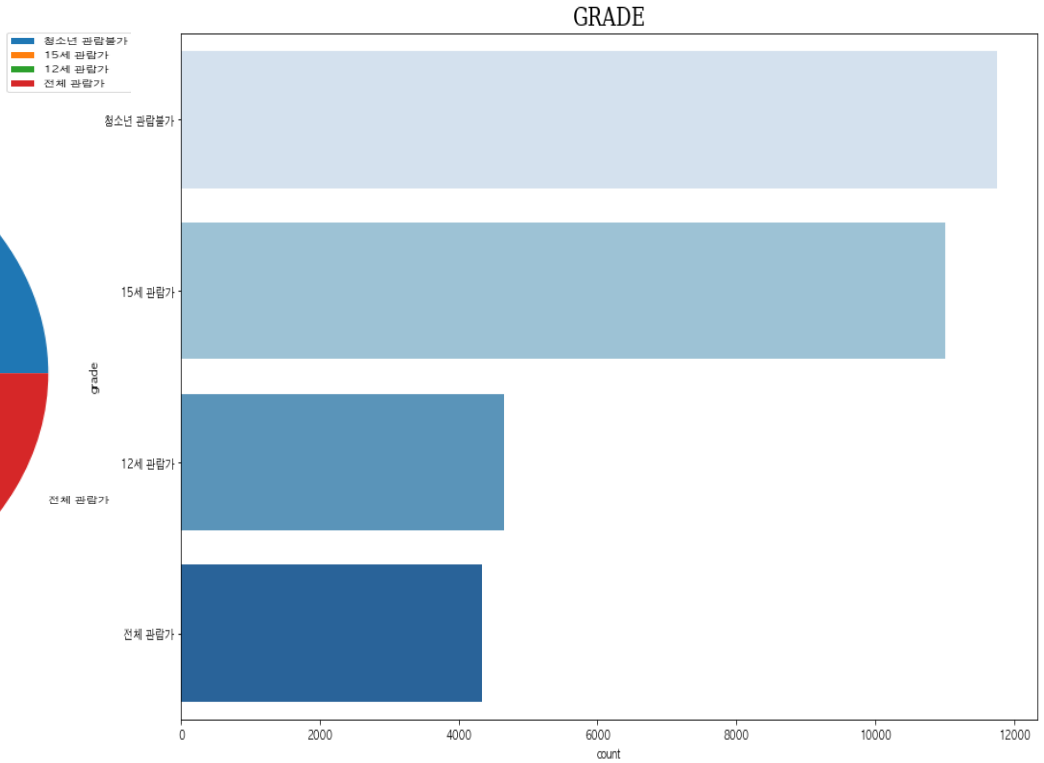
Grade

EDA

관람등급 관객수 분포

관람등급의 분포

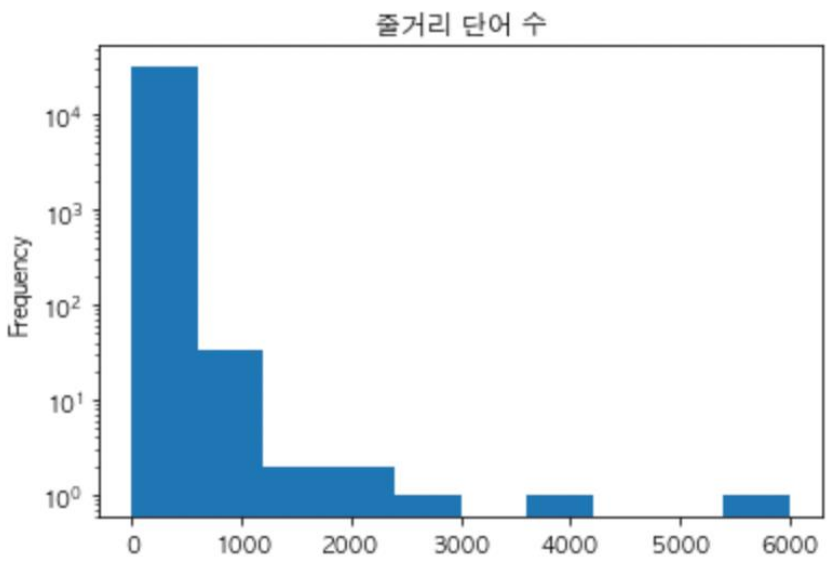


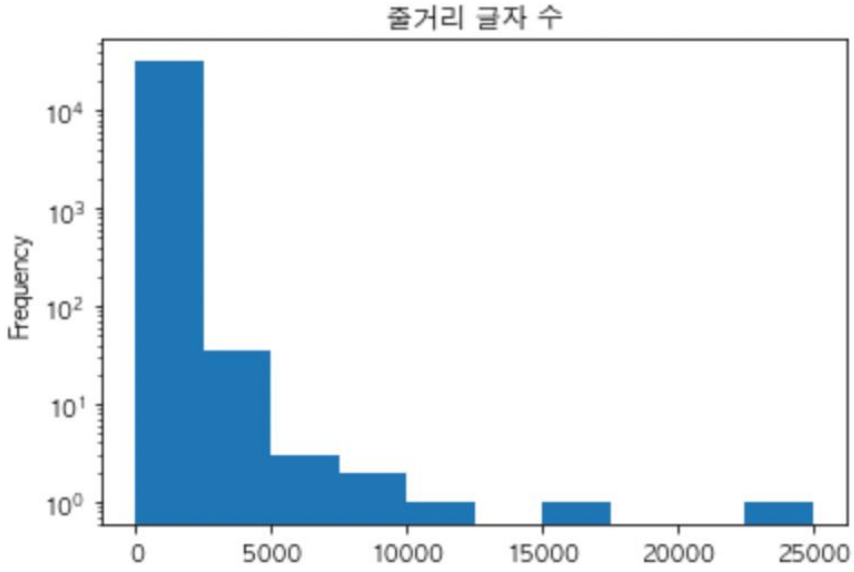


줄거리 길이 분포

평균 글자수 : 387.77

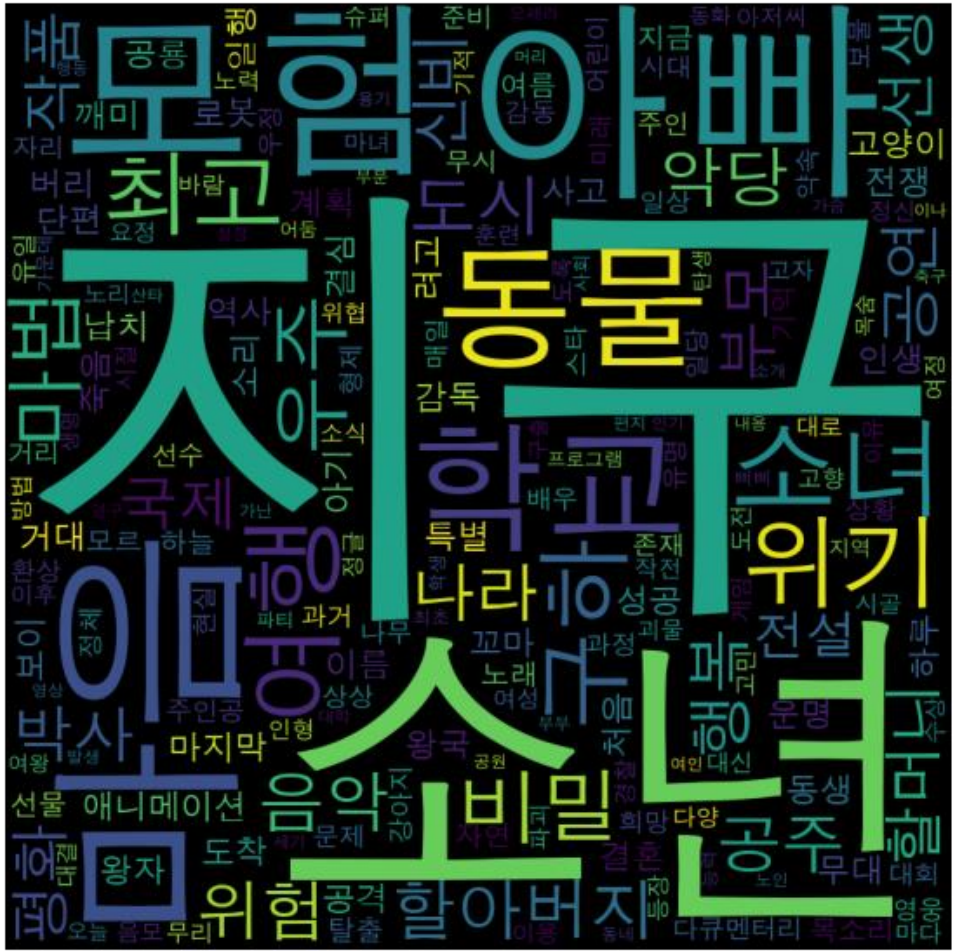
최대 글자수 : 24959





관람등급별 WordCloud

전체관람가 등급



관람등급별 WordCloud

12세 관람가 등급

관람등급별 WordCloud

15세 관람가 등급

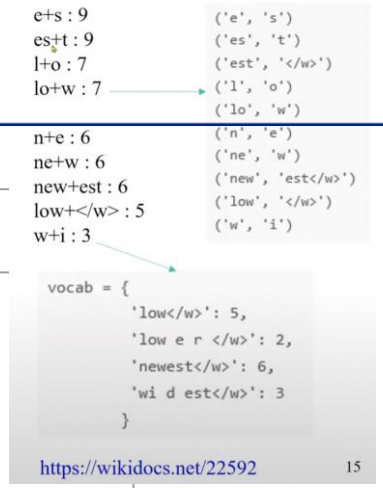
관람등급별 WordCloud

청소년 관람 불가 등급

Modeling

skt/kobert-base-v1

koBERT	BERT 한국어 모델
- 한국어 위키 기반으로 학습한 토크나이저 : 500만개의 문장과 5,400만 개의 단어를 학습 - sentencepiece model	- 충분치 못한 한국어 학습데이터 - wordpiece model : 단어 단위의 토큰화 인접한 두 토큰의 단어 모음에서 빈도를 계산해 가장 높은 빈도의 조합을 묶어주는 방식 병렬적으로 조합된 서양어와 달리 nonsegment 언어인 한국어에는 적합하지 못한 방식 ex) going -> go/ing (o) 킹받다 -> 킹받/다 ->oov



SentencePiece tokenizer

SPM: unigram language model

- Assumption

  A sentence is "a sequence of subwords"

$$\mathbf{x} = (x_1, \ldots, x_M)$$

- Sentence probability P(x) : unigram model

$$P(\mathbf{x}) = \prod_{i=1}^{M} p(x_i),$$

$$\forall i \; x_i \in \mathcal{V}, \; \sum_{x \in \mathcal{V}} p(x) = 1,$$

V: pre-determined vocabulary

Taku Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates", 2018. https://arxiv.org/pdf/1804.10959.pdf

21

SPM: unigram language model

Maximize the probability of a subword sequence $x = (x_1, \ldots, x_M)$

- x* is obtained by Viterbi algorithm

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{S}(X)} P(\mathbf{x}),$$

S(X): set of segmentation candidates

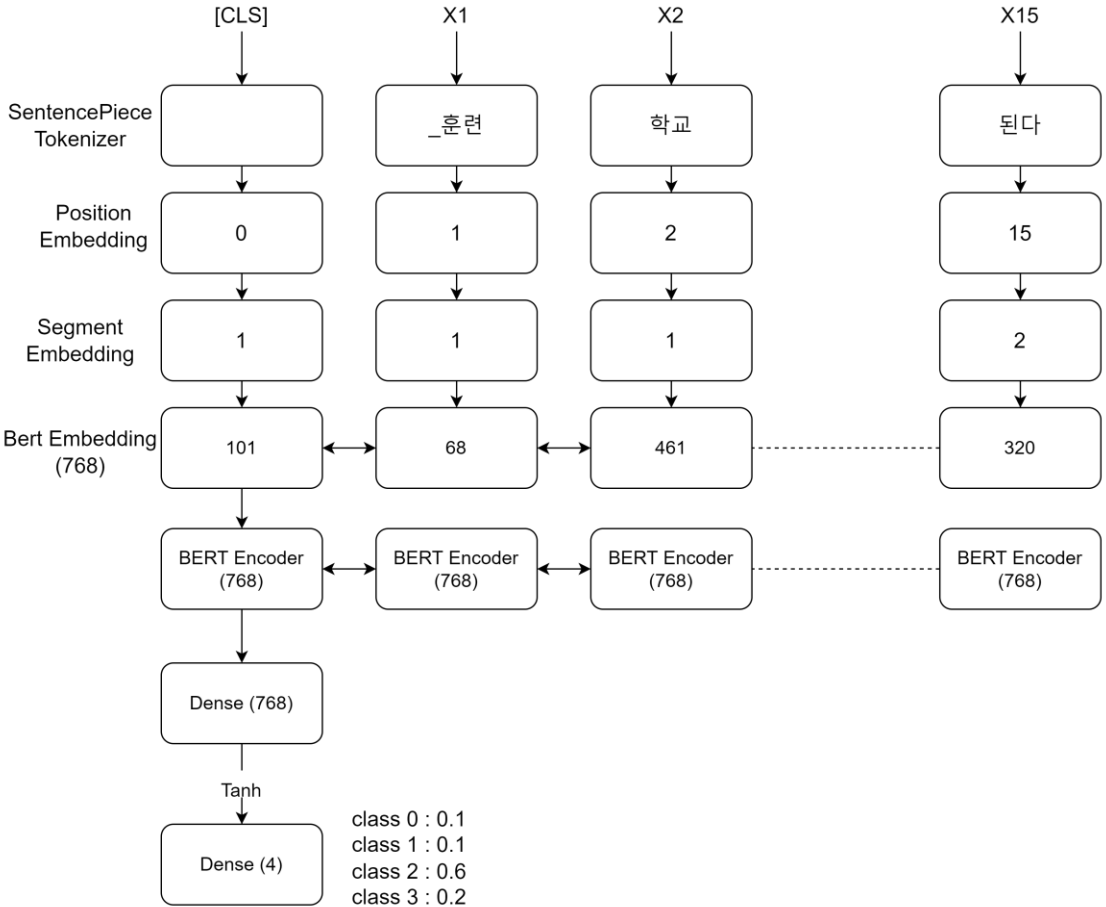
Ex) Set of segmentation candidates

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Table 1: Multiple subword sequences encoding the same sentence "Hello World"

22

koBERT

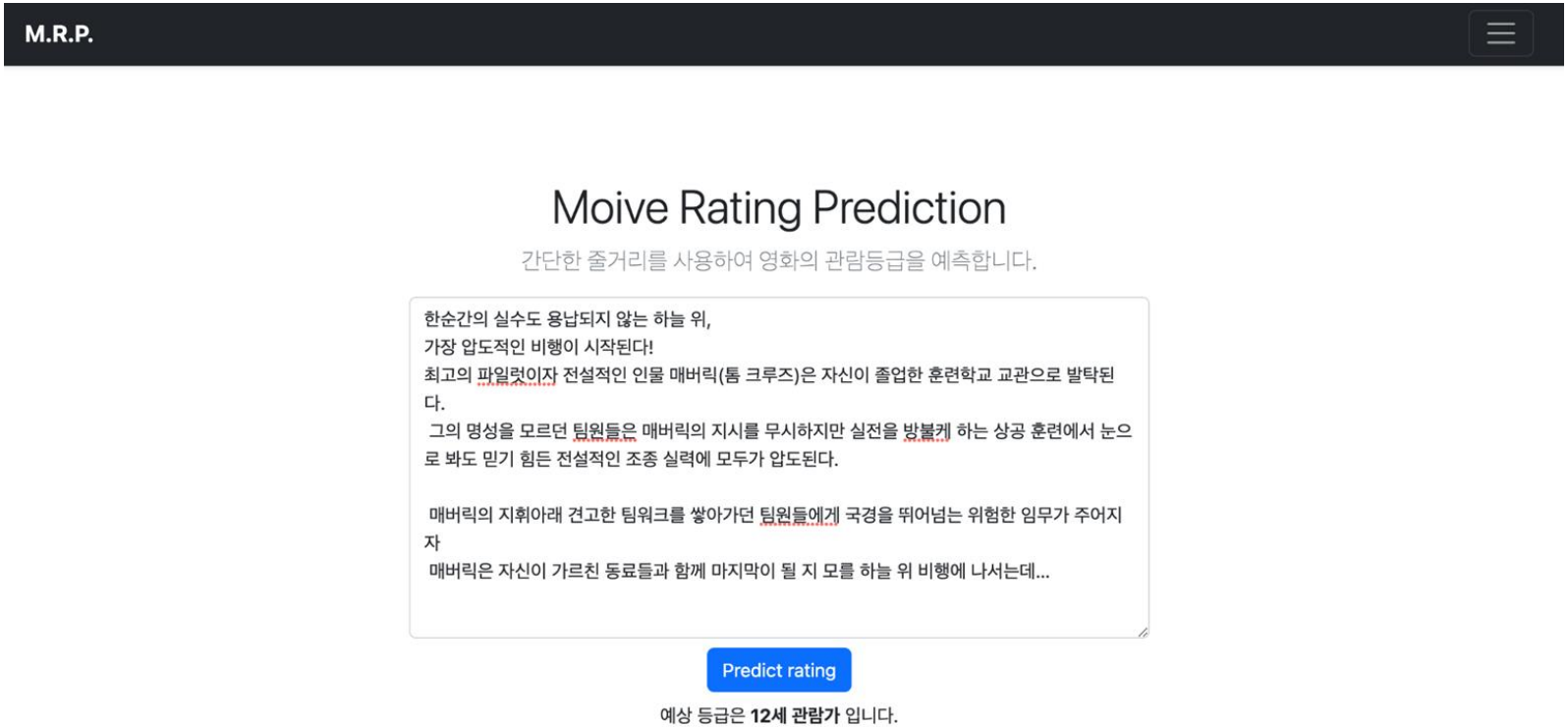

[CLS]
X1
X2
X15
SentencePiece Tokenizer
_훈련
학교
된다
Position Embedding
0
1
2
15
Segment Embedding
1
1
1
2
Bert Embedding (768)
101
68
461
320
BERT Encoder (768)
BERT Encoder (768)
BERT Encoder (768)
BERT Encoder (768)
Dense (768)
Tanh
Dense (4)
class 0 : 0.1
class 1 : 0.1
class 2 : 0.6
class 3 : 0.2

모델 실험

	okt + one-hot vector	okt + fasttext	ELMO	KoBERT
test acc	0.455	0.455	0.48	0.59

Web

Web


M.R.P.
Moive Rating Prediction
간단한 줄거리를 사용하여 영화의 관람등급을 예측합니다.
한순간의 실수도 용납되지 않는 하늘 위,
가장 압도적인 비행이 시작된다!
최고의 파일럿이자 전설적인 인물 매버릭(톰 크루즈)은 자신이 졸업한 훈련학교 교관으로 발탁된다.
그의 명성을 모르던 팀원들은 매버릭의 지시를 무시하지만 실전을 방불케 하는 상공 훈련에서 눈으로 봐도 믿기 힘든 전설적인 조종 실력에 모두가 압도된다.
매버릭의 지휘아래 견고한 팀워크를 쌓아가던 팀원들에게 국경을 뛰어넘는 위험한 임무가 주어지자
매버릭은 자신이 가르친 동료들과 함께 마지막이 될 지 모를 하늘 위 비행에 나서는데...
Predict rating
예상 등급은 12세 관람가 입니다.

Web

탑건: 매버릭 상영중

Top Gun: Maverick, 2021

관람객 ★★★★★ 9.60 기자·평론가 ★★★★☆ 8.44

네티즌 ★★★★★ 9.77 내 평점 ★★★★★ 등록>

개요 액션 | 미국 | 130분 | 2022 .06.22 개봉

감독 조셉 코신스키

출연 톰 크루즈(매버릭), 마일즈 텔러(루스터), 제니퍼 코넬리(페니) 더보기

등급 [국내] 12세 관람가

흥행 누적관객 7,224,385명(08.03 기준)

Web

title	year	grade
폴리스 아카데미 2 - 첫임무	1987.0	PG-13
매드 맥스 3	1985.0	PG-13
아파치	1990.0	PG-13
백야	1986.0	PG-13
나의 성공의 비밀	1990.0	PG-13
...	...	...
네메시스	2003.0	PG-13
애니매트릭스 - 오시리스 최후의 비행	2003.0	PG-13
페이퍼 하트	2010.0	PG-13
블라인드 가이	2007.0	PG-13
조조 래빗	2020.0	PG-13

Q&A