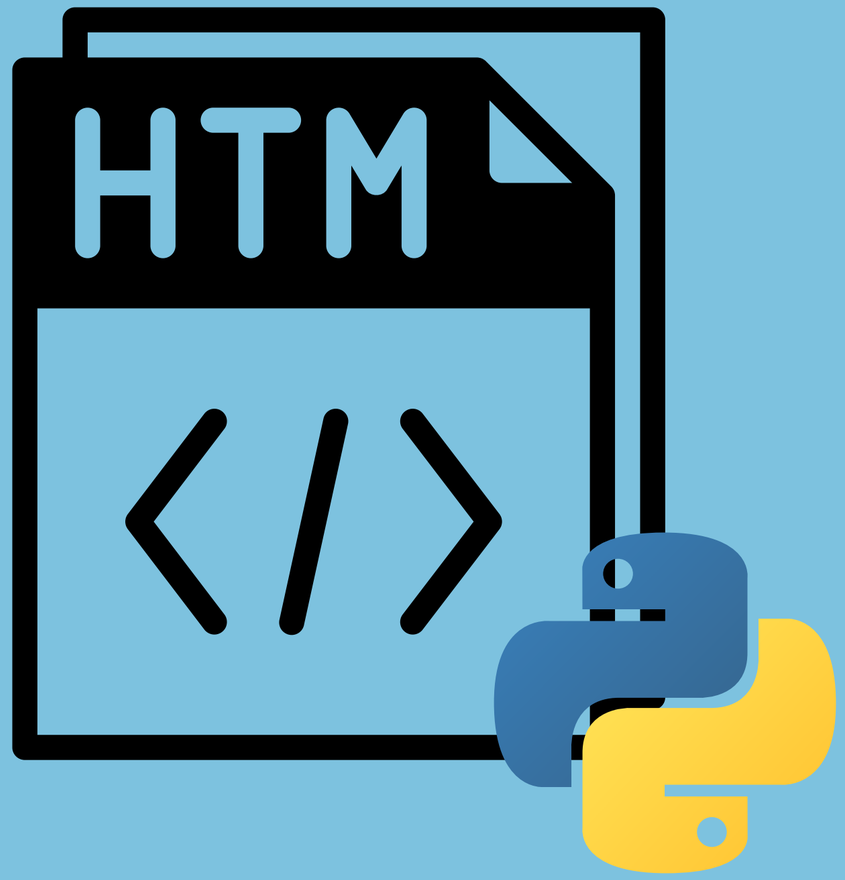
해외에서 흥행한

K-Content 흥행 요인 분석

github

팀명 : 북치고 장고 치우고

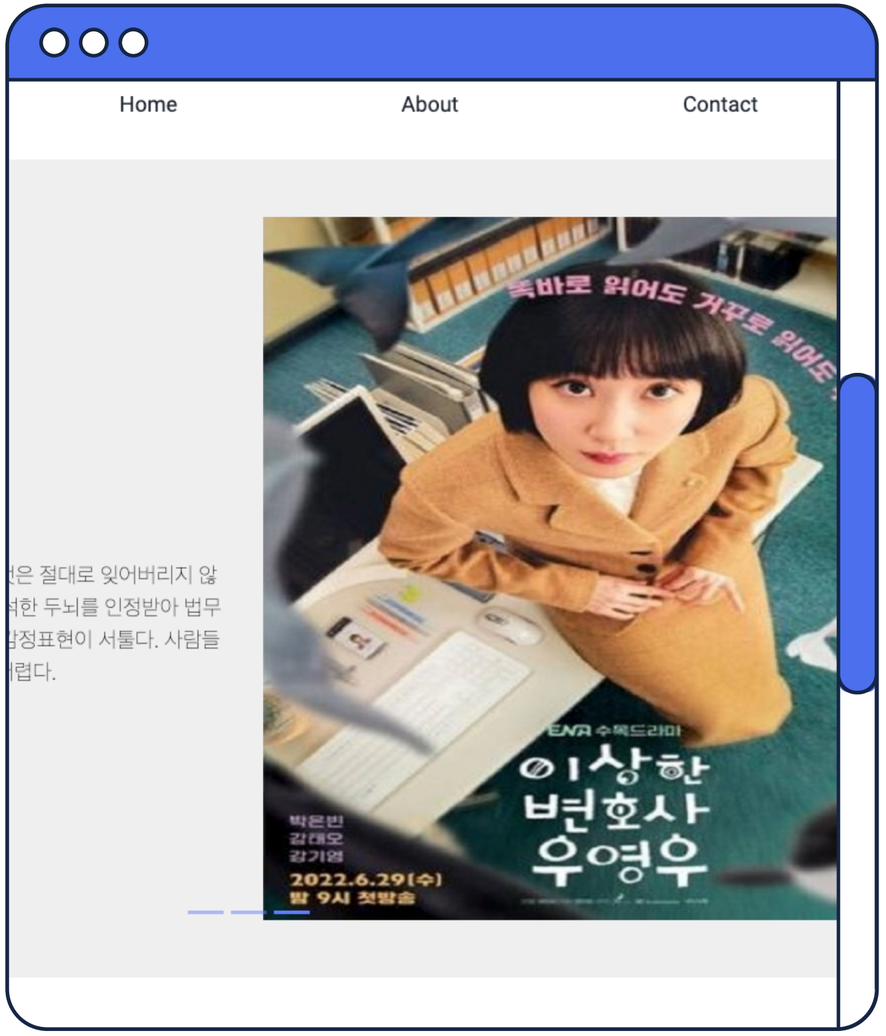
https://beatdrum.netlify.app/index.html



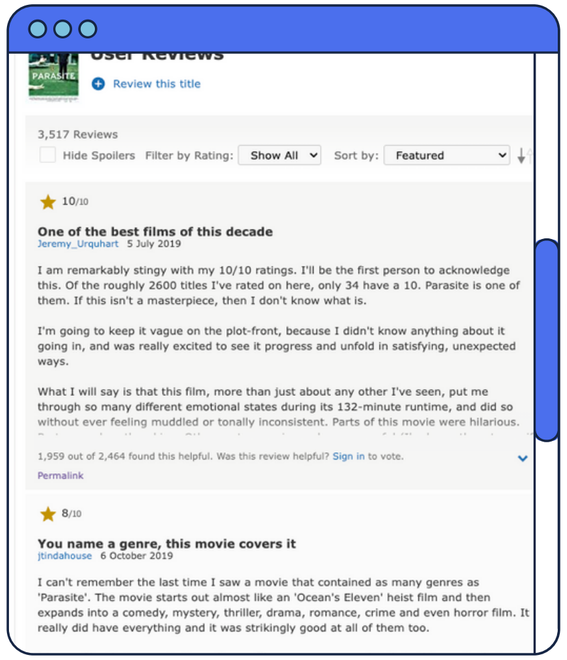
목차

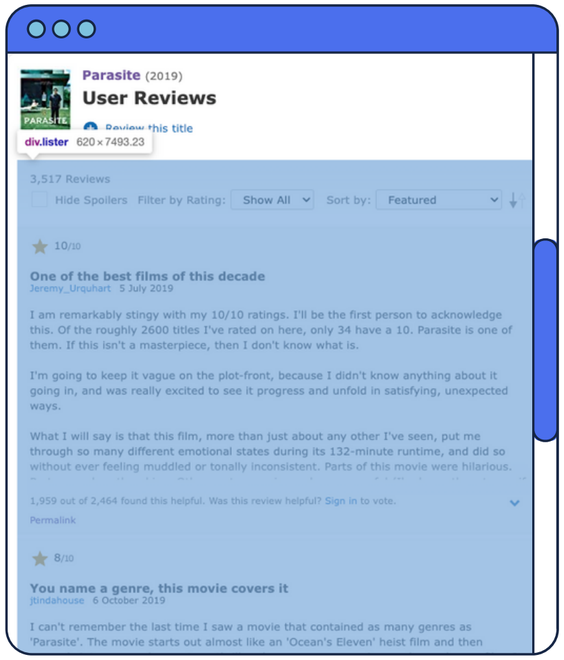
주제 선정 이유 및 방향

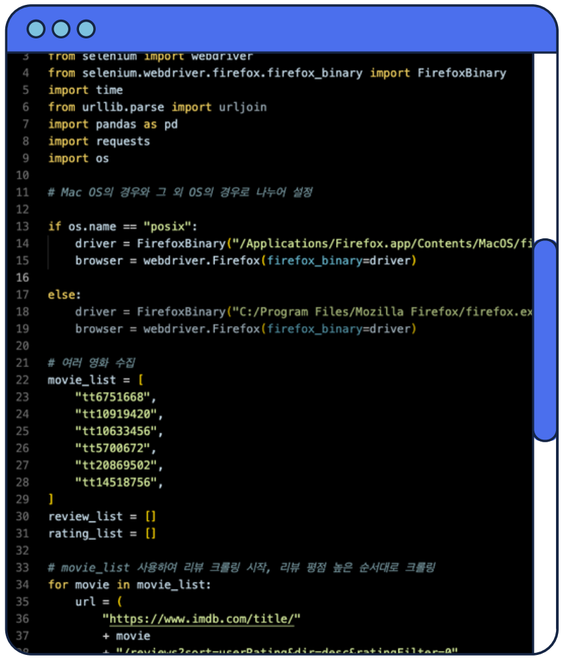
1. 코로나-19 발생 후 미디어 매체의 이용량 증가
2. K-컨텐츠의 해외 접속량 증가
3. 한국정서와 다른 외국인들의 리뷰를 바탕으로 흥행요인 분석



리뷰 크롤링







```
from selenium import webdriver
from selenium.webdriver.firefox.firefox_binary import FirefoxBinary
import time
from urllib.parse import urljoin
import pandas as pd
import requests
import os

# Mac OS의 경우와 그 외 OS의 경우로 나누어 설정

if os.name == "posix":
    driver = FirefoxBinary("/Applications/Firefox.app/Contents/MacOS/fi
    browser = webdriver.Firefox(firefox_binary=driver)

else:
    driver = FirefoxBinary("C:/Program Files/Mozilla Firefox/firefox.ex
    browser = webdriver.Firefox(firefox_binary=driver)

# 여러 영화 수집
movie_list = [
    "tt6751668",
    "tt10919420",
    "tt10633456",
    "tt5700672",
    "tt20869502",
    "tt14518756",
]
review_list = []
rating_list = []

# movie_list 사용하여 리뷰 크롤링 시작, 리뷰 평점 높은 순서대로 크롤링
for movie in movie_list:
    url = (
        "https://www.imdb.com/title/"
        + movie
```

IMDB.com 리뷰사이트 접속

분석 할 컨텐츠(6가지)의 리뷰 페이지에 접속

리뷰 부분 크롤링

웹 페이지 검사 도구로 제목, 리뷰, 및 더보기 버튼의 XPATH를 탐색

imdb_review_movies.py

Selenium, BeautifulSoup, requests 등을 사용하여 크롤링

데이터 전처리

1. 모든 리뷰를 하나의 문자열로 합친다.
2. text_to_word_sequence()를 이용해 단어 단위로 분리
3. 불용어 제거
(영어가 아닌 단어, 길이가 4이하인 단어)
4. 표제어 추출

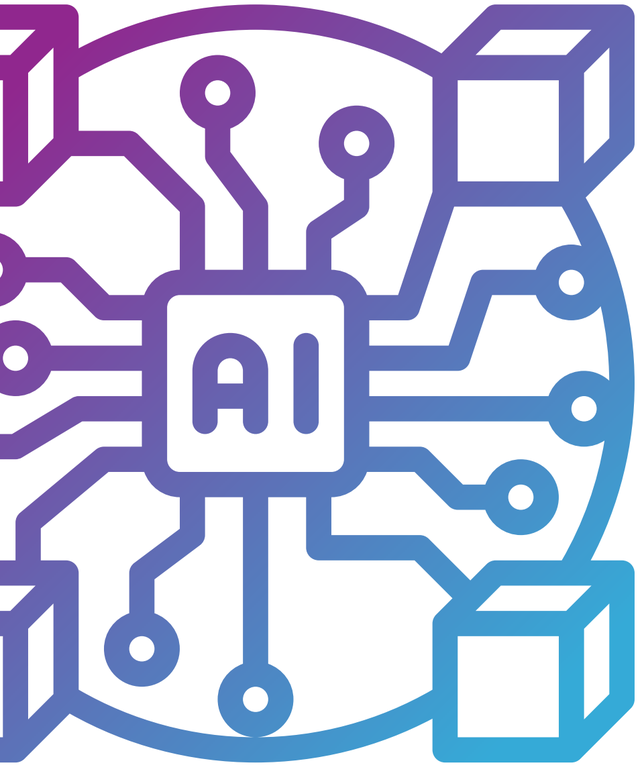
3. 품사 테그 후 분리

부록)

text_to_word_sequence()

	from nltk import word_tokenize	from tensorflow.keras.preprocessing.text import text_to_word_sequence
예시 문장	Don't be fooled by the dark sounding name, Mr. Jone's Orphanage is as cheery as cheery goes for a pastry shop.	
실행 결과	['Do', "n't", 'be', 'fooled', 'by', 'the', 'dark', 'sounding', 'name', ',', 'Mr.', 'Jone', "'s", 'Orphanage', 'is', 'as', 'cheery', 'as', 'cheery', 'goes', 'for', 'a', 'pastry', 'shop', '.']	["don't", 'be', 'fooled', 'by', 'the', 'dark', 'sounding', 'name', 'mr', "jone's", 'orphanage', 'is', 'as', 'cheery', 'as', 'cheery', 'goes', 'for', 'a', 'pastry', 'shop']
특징 요약	구두점, 어퍼스트로피(')가 보존되지 않음	구두점, 어퍼스트로피(')가 보존됨

기술 사용



1. GloVe를 통하여 모든 단어 벡터화
2. 평균벡터를 연관성이 높은 단어벡터로 판단 및 추출
3. 평균벡터와 각 단어 벡터 간의 코사인 유사도 비교
4. 코사인 유사도가 높은 순으로 단어를 정렬
5. 품사가 명사인 단어만 추출 및 저장

부록) Glove 사용 이유

Word2Vec 와 GloVe의 공통점과 차이점

자연어 처리 방법	장점	단점
LSA(카운트 기반 방법)	문서 전체의 통계적인 정보를 활용	단어간 유사도 측정 어려움
Word2Vec(예측 기반 방법)	단어간 유사도 측정 가능	주변 단어 몇개만 활용하여 결과가 도출되어, 문서 전체의 단어 정보가 반영되기 힘듦
Glove = LSA + Word2Vec	단어간 유사도 측정 가능, 문서 전체의 통계적인 정보 활용 가능	동음이의어를 동일한 벡터로 임베딩하여 성능이 좋지 않음

웹페이지 결과 시연 (코드)


Home
About
Contact
오징어 게임
데스게임 | 서바이벌 | 스릴러
456억 원의 상금이 걸린 의문의 서바이벌에 참가한 사람들이 최후의 승자가 되기 위해 목숨을 걸고 극한의 게임에 도전하는 이야기를 담은 넷플릭스 시리즈 빚에 쫓기는 수백 명의 사람들이 서바이벌 게임에 뛰어든다. 거액의 상금으로 새로운 삶을 시작하기 위해. 하지만 모두 승자가 될 순 없는 법. 탈락하는 이들은 끔찍한 결과를 각오해야 한다.
456억,
오징어 게임
영화 | 드라마
해외에서 흥행한 국뽕차오르는 드라마와 영화들
PARASITE

느낌점

유명상

데이터 분야를 포함한 다른분야로도
정말 큰 배움이 있었습니다.

박정현

새로운것을 배워가며 프로젝트를 완성
할 수 있어 뜻깊었습니다!

보완점

1. 장고를 사용하지 않아 아쉽다.
(정적 웹페이지 구현으로 장고의 기능이 필요하지 않았다.)

2. Glove의 단어간 유사도 측정 부분을,
평균벡터와 각 단어 벡터 간의 코사인 유사도 비교로 대체하였다.
향후 이 부분을 제대로 사용할 예정이다.

2. 트위터 API를 사용하여 영화 언급 수를 날짜별로 카운트하여
시각화 하려 했다.
트위터 API 등급을 업그레이드 하였음에도 request
허용치를 초과하여 더 이상 진행 할 수 없었다.

```
  File "/Users/bagjeonghyeon/miniforge3/envs/baseDeep/lib/python3.8/site-packages/tweepy/api.py"
    raise TooManyRequests(resp)
tweepy.errors.TooManyRequests: 429 Too Many Requests
88 - Rate limit exceeded

 ~/Week_Project_1/Week_Project_1  on  jeonghyeon's ↓1 *1 !1 ?5
```

Q&A

Q.

A.