2022.05.30

Movie Recommendation

클라우드 서비스 기반 빅데이터 개발 분석

팀원 : 강보현 손혁진 박영미

Contents

Chapter

01. 프로젝트 과정 및 배경 지식

Processing

Day 1

주제 선정

2022 날씨빅데이터 콘테스트

음성인식으로 노래 장르 분류

사진으로 성별 나이 예측

✔ 영화별 관객수, 만족도를 통하여 재개봉 추천

Day 2

2011~2021년

연도별 영화 평점 수집

네이버 API를 활용하여

연도별 영화별 평점 수집

Day 3

영화별 관객수, 만족도를

통하여 재개봉 추천을 하려

하였으나 평점 및 관객수로

머신러닝을 하면 인기순으로

만 추천해줄것이라 판단하여

데이터 재수집

Day 4

데이터 전처리

원본데이터중 영화명, 장르, 국가, 감독명,
평점, 배우를 추출하고 배우에 있는
'<b>', '<b/>', |을 삭제

Day 5

전처리 및 머신러닝 가동
SVD 모델 활용

Day 6

모델 케이스 별
테스트 및 시각화

Processing

chapter 01

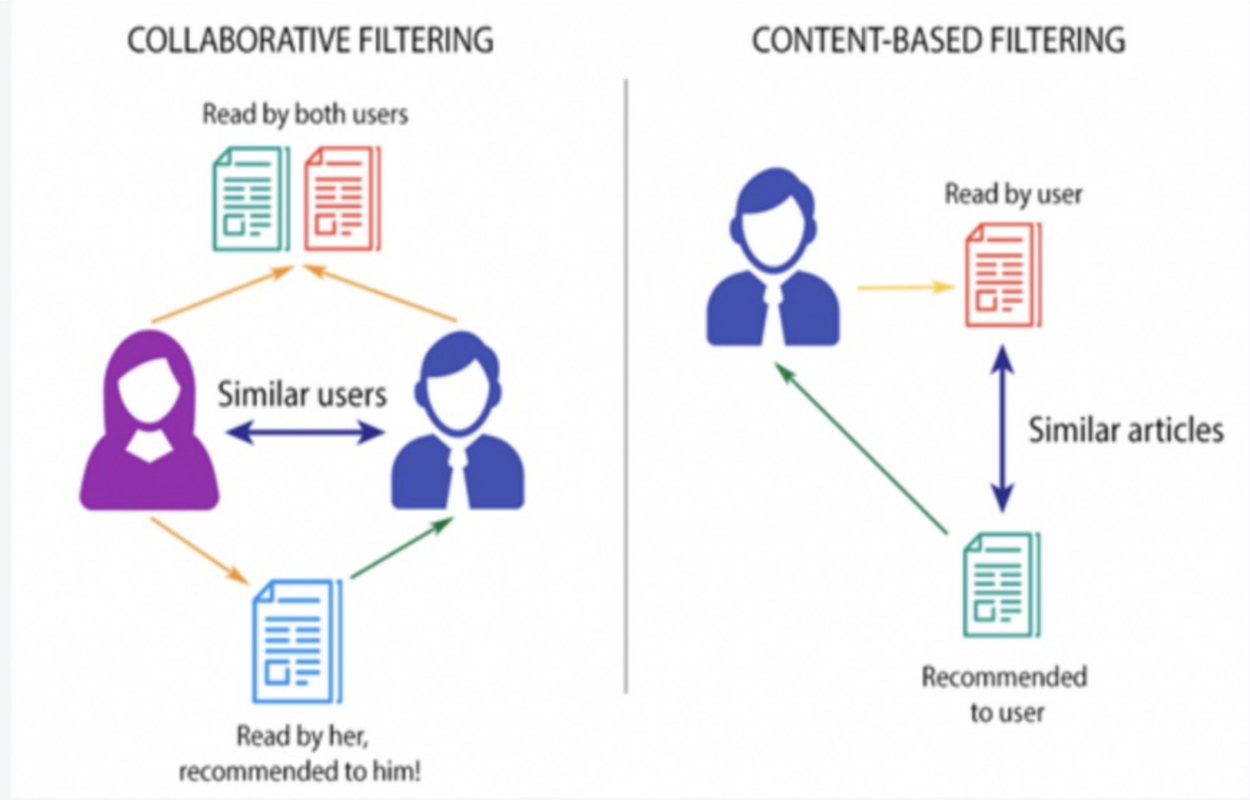
주제 재선정

영화별 관객수, 만족도를 통하여 재개봉 추천을 하려 하였으나 평점 및 관객수로 머신러닝을 하면 인기순으로만 추천해줄것이라 판단하여 영화 추천으로 변경
알고리즘 추천으로는 Contents Based Filtering과 Collaborative Filtering가 있는데 미니 프로젝트기간중 Contents Based Filtering만 사용하여 결과를 도출하기로함

Contents Based Filtering을 사용하여 결과를 도출하기로 하였으나 평점이 몇명 투표해서 나온수인지 크롤링을 할수 없어서 Collaborative Filtering으로 결과를 도출하기로 변경

chapter 01

배경 지식



협업 필터링(COLLABORATIVE FILTERING)

협업 필터링 알고리즘이란 기존 사용자의 행동 정보를 분석해 해당 사용자와 비슷한 성향의 사용자들이 기존에 좋아했던 항목을 추천하는 기술 예를 들어 영화 A를 본 사람들이 영화 B를 시청한 경우가 많으면 영화 A를 보는 사람에게 영화 B를 추천해 주는 방식이다. 이 알고리즘은 결과가 직관적이며 항목의 구체적인 내용을 분석할 필요가 없다는 장점이 있다.

내용 기반 필터링(CONTENT-BASED FILTERING)

내용 기반 필터링 알고리즘은 말 그대로 내용에 대한 분석을 기반으로 추천을 구현한다. 콘텐츠를 분석한 프로파일과 사용자의 선호도를 추출하고 유사성 분석을 통해 추천을 수행한다. 가령 슈퍼맨이라는 영화에 대한 사전 분석을 통해 SF, 영웅 등의 특징을 기록하고 슈퍼맨을 본 사람에게 배트맨이라는 비슷한 종류의 영화를 추천해주는 방식이다.

Chapter

02. 데이터 수집 및 전처리

데이터 크롤링 코드

```
import urllib.request
import pandas as pd
import time
from bs4 import BeautifulSoup

path = '/Users/bohyenkang/Downloads/2001_2.csv'

df = pd.read_csv(path)
df2 = df['영화명']
df3 = df['제작연도']
df4 = df['감독']
site_url =
'https://openapi.naver.com/v1/search/movie.xml'
cid = 'X-Naver-Client-Id'
csec = 'X-Naver-Client-Secret'
client_id = 'ICJJx6x7ygogtezSfUgx'
client_secret = 'Saaw1DiedV'
cnt = 0
name = []
rating = []
actor = []
print(len(df2))
```

```
try :
    for i in range(len(df2)):
        print(cnt)
        cnt += 1
        query = df2[i]
        year = str(df3[i])
        #print(query)
        if str(type(query)) == "<class 'float'>":   #영화제목없는오류
            name.append('nan')
            actor.append('nan')
            rating.append(0)
            continue

        q_param = "query=" + urllib.parse.quote(query)
        yf = 'yearfrom=' + year
        yt = 'yearto=' + year
        query_str = site_url + '?' + q_param + '&' + yf +'&' + yt
        request = urllib.request.Request(query_str)
        request.add_header(cid, client_id)
        request.add_header(csec, client_secret)
        response = urllib.request.urlopen(request)
```

```
time.sleep(1)

data = response.read().decode('utf-8')
#print(data)
hdoc = BeautifulSoup(data, 'html.parser')
#print(hdoc)
items = hdoc.find_all('item')
#print(items)

if len(items) == 0:          # itme에 아무것도 안잡히면 내용이 없는거

    name.append(query)

    actor.append('nan')

    rating.append(0)

    continue

j=0
```

```
while True:
    if (items[j].find('director').get_text() not in df4[i]):    #감독명이 비슷한지 비교
        name.append(query)
        rating.append(items[j].find('userrating').get_text())
        actor.append(items[j].find('actor').get_text())

        break

    j += 1

    if j == len(items):    #items의 크기와 같으면 종료
        name.append(query)
        actor.append('nan')
        rating.append(0)

        break

data = {
    '영화명': name,
    '평점': rating,
    '배우': actor
}
df = pd.DataFrame(data)
df.to_csv('2011.csv', index=True)
```

	영화명	영화명(영문)	제작연도	제작국가	유형	장르	제작상태	감독	제작사
0	마태복음	Il Vangelio Secondo Matteo	1964	이탈리아	장편	드라마,사극	개봉	피에르 파올로 파솔리니	NaN
1	월 · E	Wall · e	2008	미국	장편	애니메이션,코미디,액션,어드벤처,가족,SF	개봉	앤드류 스탠튼	NaN

< 초기 원본 데이터 >

https://www.kobis.or.kr

	영화명	평점	배우
0	마태복음	7.71	엔리케 이라조퀴|마르게리타 카루소|수잔나 파솔리니|
1	월 · E	9.42	벤 버트|엘리사 나이트|제프 갈린|프레드 윌러드|

< 웹크롤링 데이터 >

네이버영화 API

데이터 18,122개에서 11,455개로 전처리

	영화명	장르	제작국가	감독	평점	배우
0	마태복음	드라마,사극	이탈리아	피에르 파올로 파솔리니	7.71	엔리케 이라조퀴,마르게리타 카루소,수잔나 파솔리니
1	월 · E	애니메이션,코미디,액션,어드벤처,가족,SF	미국	앤드류 스탠튼	9.42	벤 버트,엘리사 나이트,제프 갈린,프레드 윌러드

< 초기 원본 데이터 + 웹크롤링 데이터 >

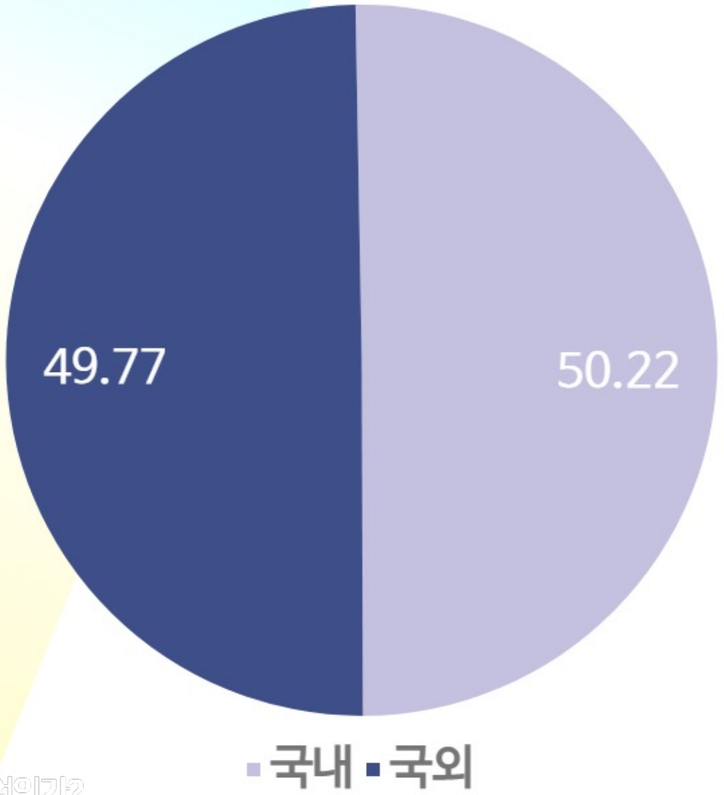
< user_id, movie_id, rating 생성 >

	movie_id	title	genre	country	director	avg_rating	actor	user_id	rating
0	0	마태복음	드라마,사극	이탈리아	피에르 파올로 파솔리니	7.71	엔리케 이라조퀴,마르게리타 카루소,수잔나 파솔리니	342	1
1	1	월 · E	애니메이션,코미디,액션,어드벤처,가족,SF	미국	앤드류 스탠튼	9.42	벤 버트,엘리사 나이트,제프 갈린,프레드 윌러드	832	3

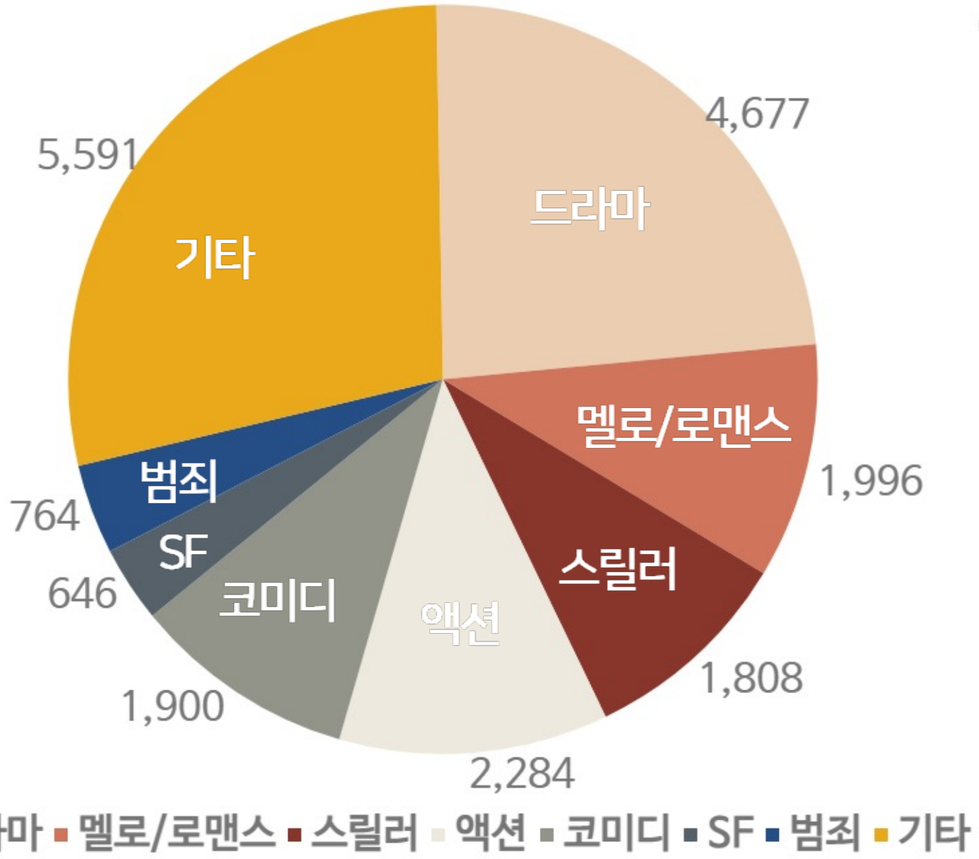
chapter 02

분포도

국내/외



장르



기회는 무엇인가?

Chapter

03. 머신러닝 및 결과

chapter 03

SVD 모델

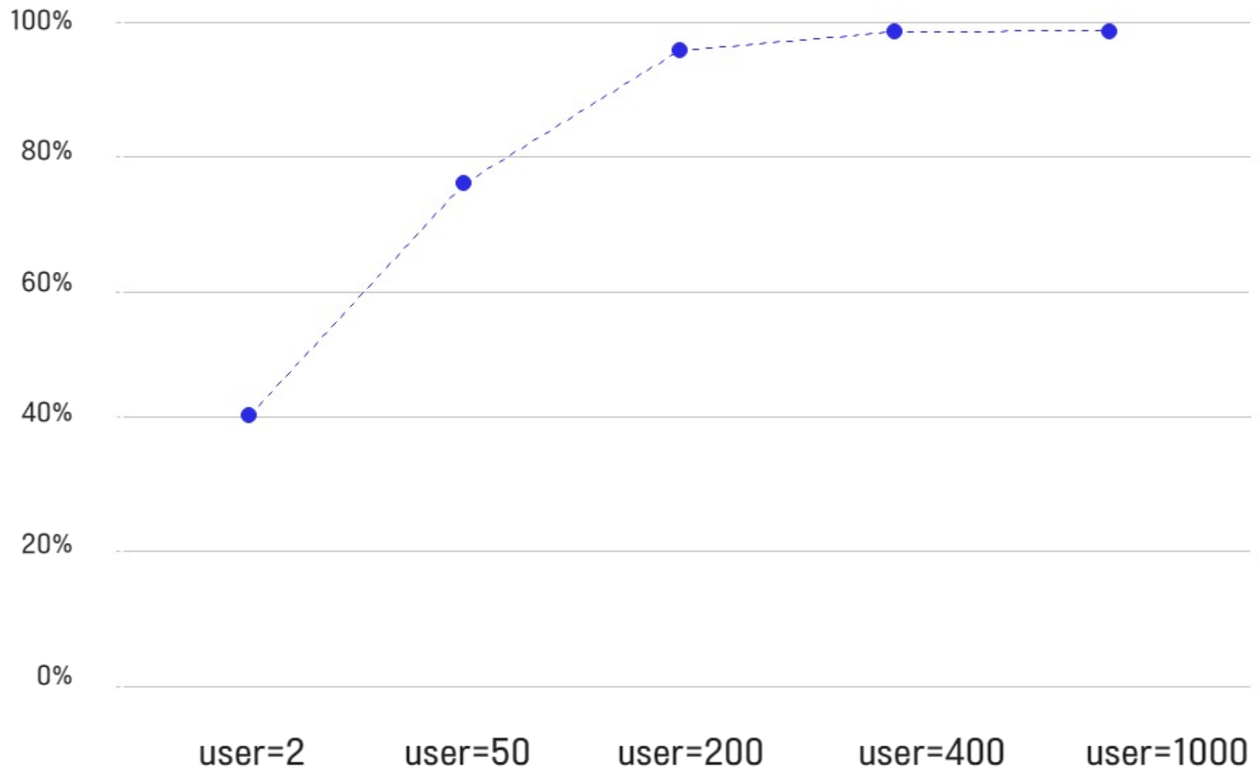
```
import pandas as pd
data =pd.read_csv('/Users/bohyenkang/df1000.csv')
data =data.drop(data.columns[0],axis=1)

from surprise import SVD, Dataset, Reader, accuracy
from surprise.model_selection import train_test_split
# reader 생성
reader = Reader(rating_scale = (1,5))

# 데이터를 생성한다.
a1 = data[["user_id", "movie_id", "rating"]]
data2 = Dataset.load_from_df(a1, reader)
train_data = data2.build_full_trainset()

# 모델 생성 및 학습
model = SVD()
model.fit(train_data)
```

SVD를 사용하면 사용자의 특징과 아이템의 특징 그리고 이 두 가지를 대표하는 대각 행렬을 추출할 수 있습니다. 대각 행렬을 축소하여 전체 처리해야 하는 데이터의 양을 줄일 수 있습니다. 또한 아직 평가하지 않은 데이터에 대해서 평균 값을 이용해 결측치를 채운 후 SVD를 이용해 예상 점수를 추측할 수 있습니다. 추천 시스템에서는 이 추측한 점수 중 높은 아이템에 대해서 추천을 할 수 있습니다.

chapter 03

모델 비교분석

user 수	RMSE
2 users	0.4043
50 users	0.7557
200 users	0.9565
400 users	0.9856
1000 users	0.9864



예상 평점

```
영화 제목 : 시크릿 슈퍼스타
장르 : 코미디,드라마
국가 : 인도
배우 : 아미르 칸,자이라 와심
평균 평점 : 9.24
예상 평점 : 4.1
--------------------------------------------------
영화 제목 : 세븐 파운즈
장르 : 드라마
국가 : 미국
배우 : 윌 스미스,로사리오 도슨,우디 해럴슨
평균 평점 : 7.54
예상 평점 : 4.0
--------------------------------------------------
영화 제목 : 지구침략: 기계들의 반란
장르 : 액션,SF
국가 : 미국
배우 : 조 랜도,리사 로시세로
평균 평점 : 4.26
예상 평점 : 3.9
--------------------------------------------------
영화 제목 : 럭키 루크: 전설의 무법자
장르 : 어드벤처,코미디,서부극(웨스턴)
국가 : 프랑스
배우 : 멜빌 푸포,장 뒤자르댕
평균 평점 : 6.8
예상 평점 : 3.8
--------------------------------------------------
```

Problem & Solution

Problem	Solution
수집가능한 데이터 부족	주제 재선정
데이터 전처리 중 다수의 예외 발생	IF문 or Try~ Except

전체적인 느낀점

처음 선정한 주제에 맞는 데이터 수집이 제한적이어서 여러 번 주제를 변경하면서 시간이 많이 소요되었다.

주제 선정에 대한 방향성을 제대로 정하지 못하였고, 이로 인해서 데이터 탐색하는 시간이 길어졌다.

프로젝트를 진행하면서 데이터 탐색 기준 선별 능력이 다소 부족하다는 것을 많이 느끼게 되었고, 다음부터는 주제에 맞는 기준을 잘 선정해야겠다.

Thank you

여기까지 들어주셔서 감사합니다.