



데이터처리언어
최종발표

21512023 정주현
21512071 박순혁

01

Tasks 소개

02

Dataset 소개

03

분석 Framework

04

EDA

05

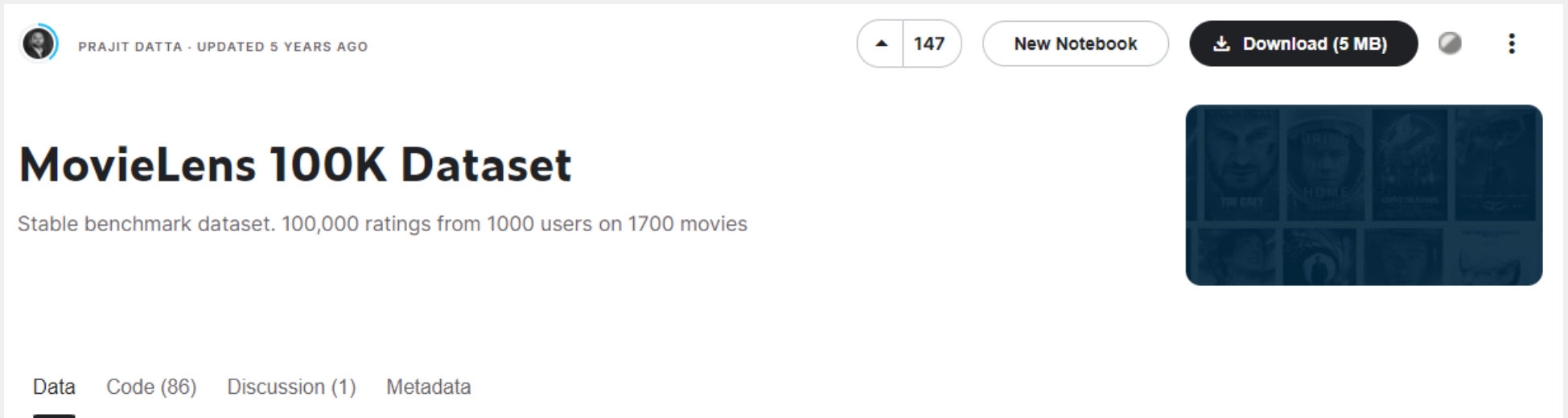
분석 방법

- User based
- Item based

06

분석 결과

01. Task 소개



The screenshot shows the Kaggle dataset page for 'MovieLens 100K Dataset' by Prajit Datta. The page header includes the user's profile, the dataset name, and update information. On the right, there are buttons for 'New Notebook', 'Download (5 MB)', and a view count of 147. The main title 'MovieLens 100K Dataset' is prominently displayed, followed by a description: 'Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies'. Below this, there are tabs for 'Data', 'Code (86)', 'Discussion (1)', and 'Metadata', with 'Data' being the active tab. A grid of movie posters is visible on the right side of the page.

PRAJIT DATTA · UPDATED 5 YEARS AGO

▲ 147 New Notebook Download (5 MB)

MovieLens 100K Dataset

Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies

Data Code (86) Discussion (1) Metadata

- Task : Movie Recommendation System
- MovieLens 100K : <https://www.kaggle.com/datasets/prajitdatta/movielens-100k-dataset>
- 분석 방법 및 역할 분담
 - ✓ User based collaborative filtering : 박순혁
 - ✓ Item based collaborative filtering : 정주현

02. Dataset 소개

MovieLens 100K

- GroupLens Research에서 MovieLens의 등급 dataset을 수집해서 제공한 자료
- 1682편의 영화에서 943명의 사용자로부터 100,000 개의 평가(1점에서 5점)가 포함되어 있음.
- 사용자들은 최소 20편의 영화에 대해 평가하였음.
- 1997년 9월 19일부터 1998년 4월 22일까지 데이터 포함.
- u.data, u.user, u.item 데이터로 이루어져있음.
- 이번 분석에는 u.data 와 u.item 데이터를 사용

Q 02. Dataset 소개

u.data : 943명의 사용자가 1682개 영화에 대한 rating 100,000개

Index	user_id	movie_id	rating	timestamp
0	196	242	3	881250949
1	186	302	3	891717742
2	22	377	1	878887116
3	244	51	2	880606923
4	166	346	1	886397596
5	298	474	4	884182806

⋮

user_id : 사용자 id(1~943)

movie_id : 영화 id(1~1682)

rating : 사용자가 영화에 매긴 1~5점까지의 평점

timestamp : 평가한 시간(유닉스 시간)

Q 02. Dataset 소개

u.item : 1682개의 영화에 대한 정보

Index	movie id	title	release date	video release date	IMDB URL	unknown	Action	Adventure
0	1	Toy Story (1995)	01-Jan-1995	nan	http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)	0	0	0
1	2	GoldenEye (1995)	01-Jan-1995	nan	http://us.imdb.com/M/title-exact?GoldenEye%20(1995)	0	1	1
2	3	Four Rooms (1995)	01-Jan-1995	nan	http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)	0	0	0
3	4	Get Shorty (1995)	01-Jan-1995	nan	http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995)	0	1	0
4	5	Copycat (1995)	01-Jan-1995	nan	http://us.imdb.com/M/title-exact?Copycat%20(1995)	0	0	0
5	6	Shanghai Triad (Yao a yao dao waipo qiao) (1995)	01-Jan-1995	nan	http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+qiao+(1995)	0	0	0

War	Western
0	0
0	0
0	0
0	0
0	0
0	0
0	0

...

:

movie_id : 영화 id

title : 영화 제목

release date : 개봉 날짜

video release date : 비디오 개봉 날짜

IMDB URL

unknown~Western : 장르 해당여부 (0: 비해당, 1 : 해당)

Q 03. 분석 Framework

EDA

Item

- Genre distribution
- Release date distribution(year/month/day)

Rating

- **Rating** distribution
- **Most rated movie**
- **Most preferred movie**
- Mean rating and volume by item

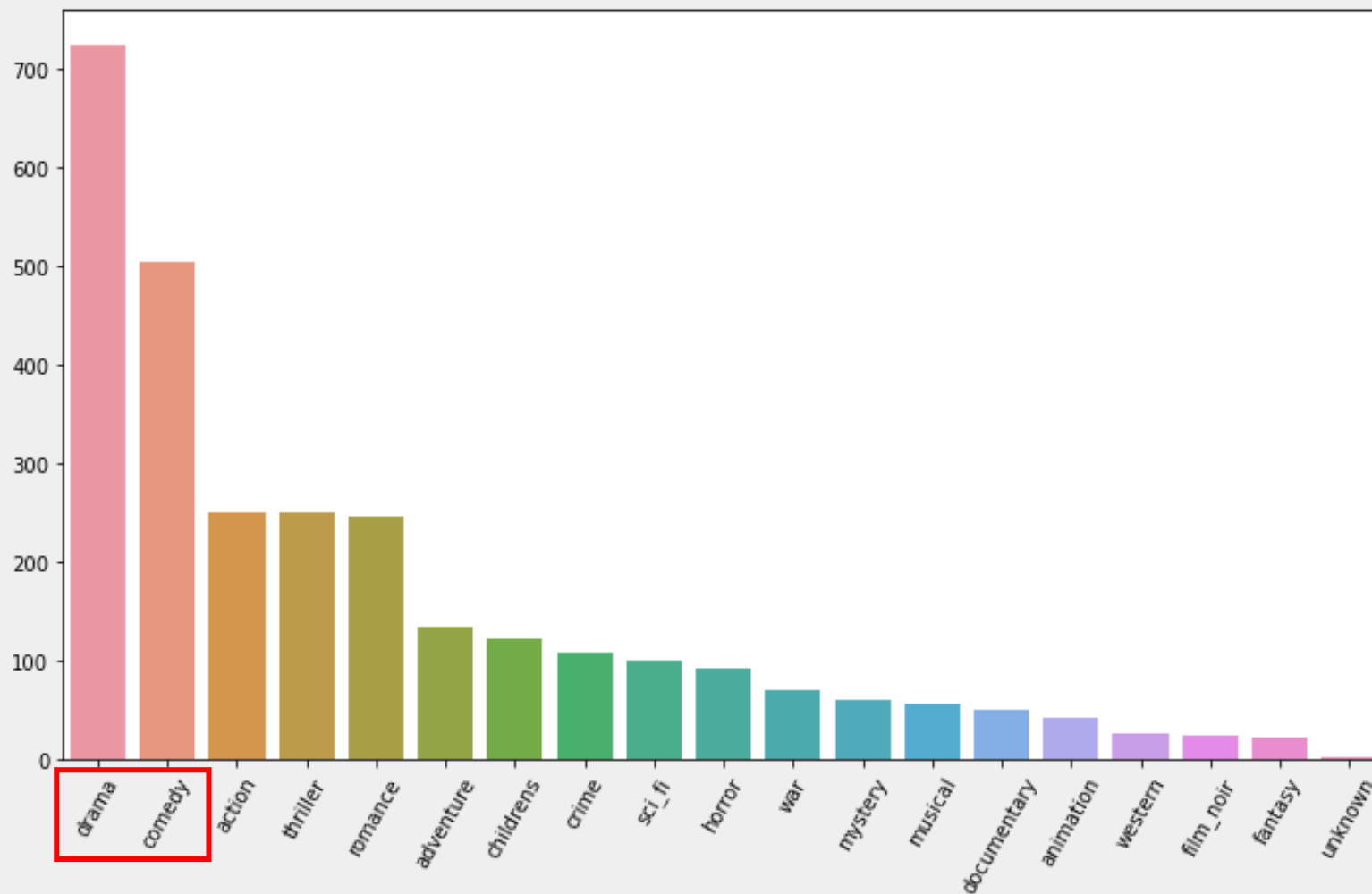


Recommend System

- User based collaborative filtering
- Item based collaborative filtering

Item

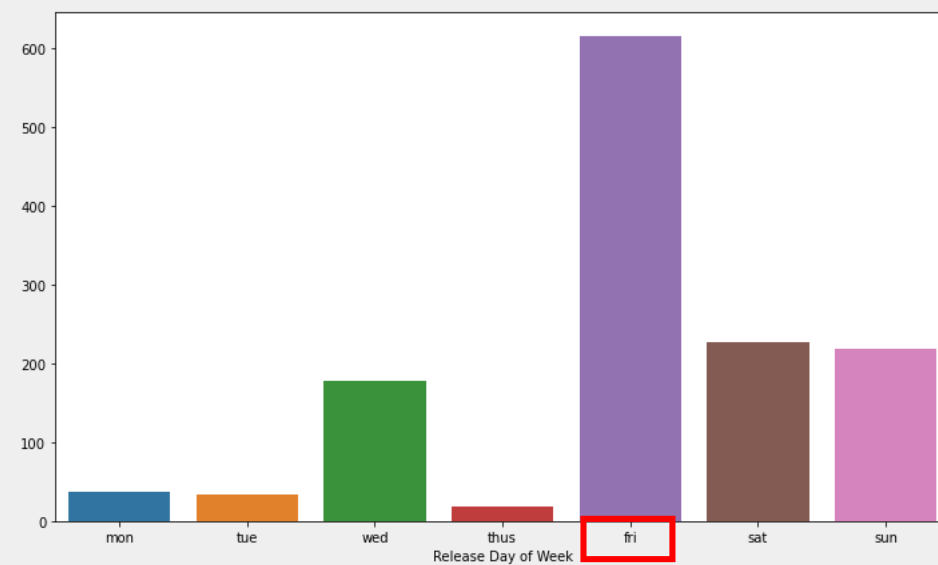
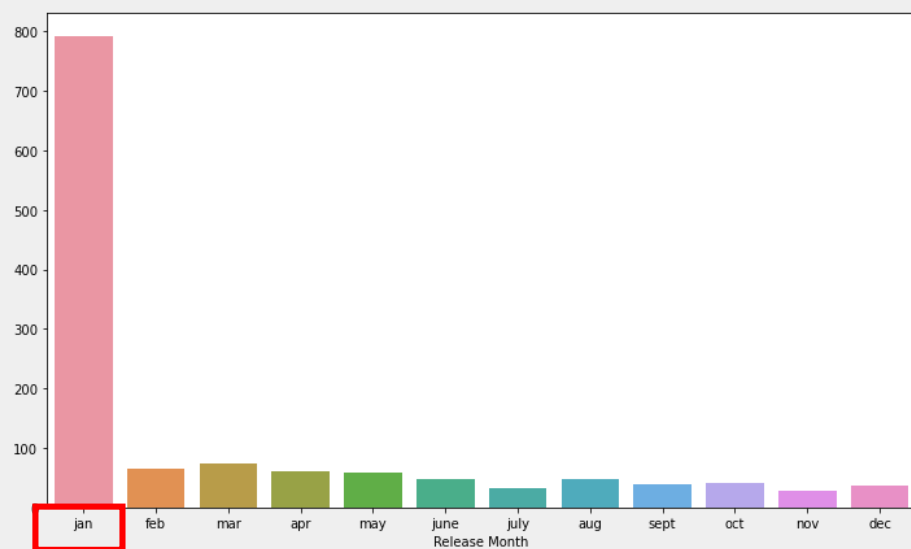
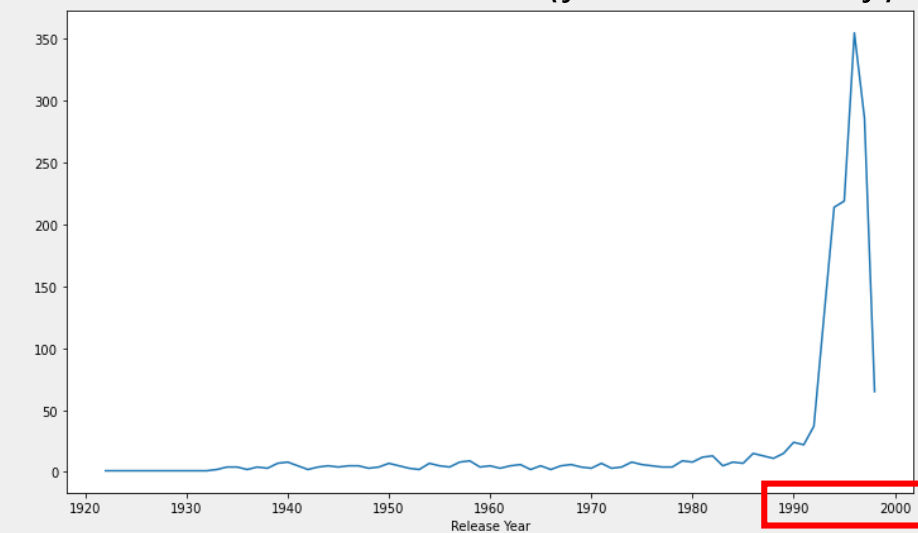
- Genre distribution



Q 04. EDA

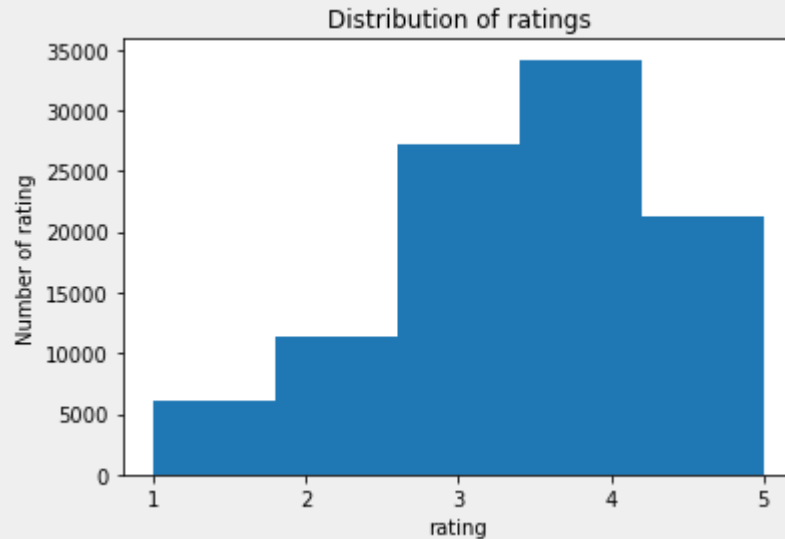
Item

- Release date distribution(year/month/day)

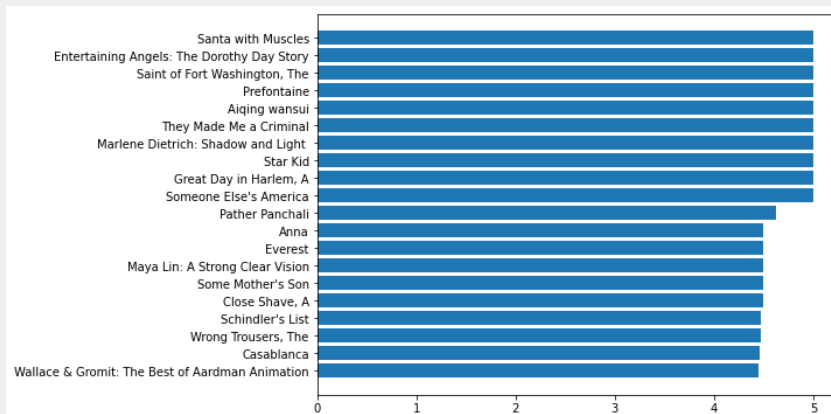


Rating

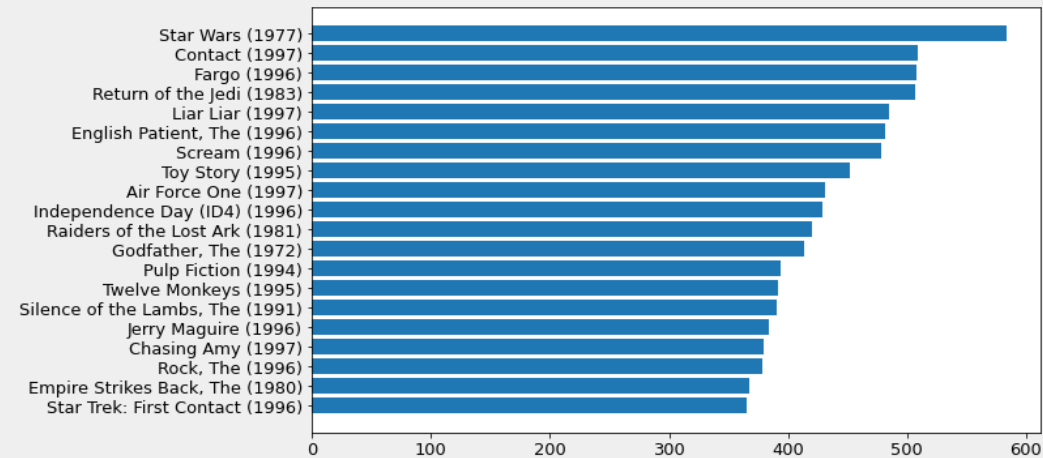
Rating distribution



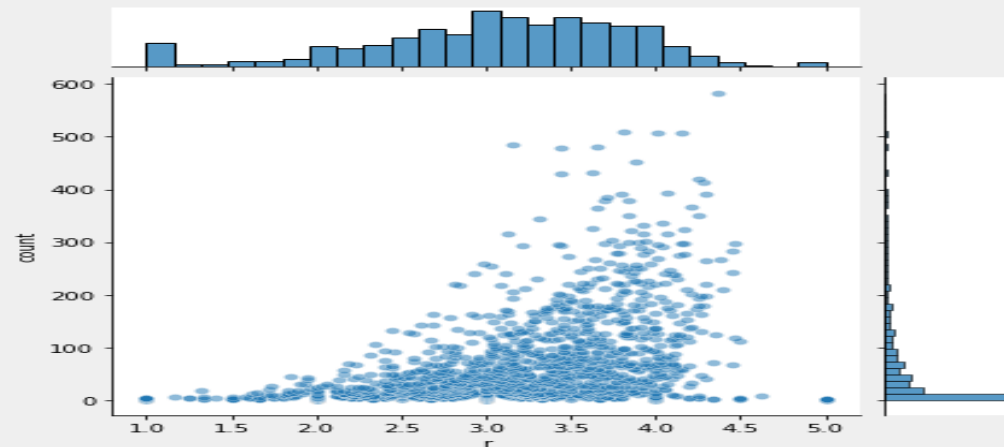
Most preferred movie



Most rated movie



Mean rating and volume by item

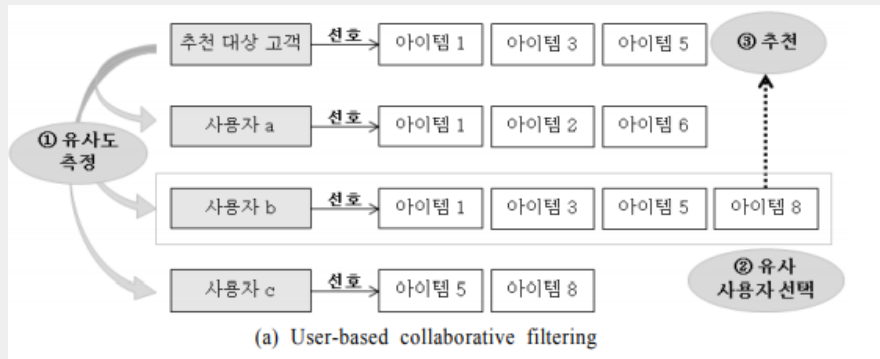


05. 분석 방법

collaborative filtering recommendation system

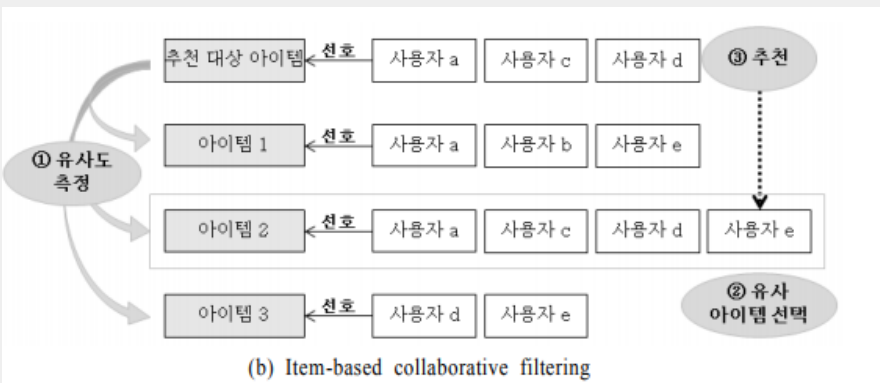
User-based filtering

고객 간의 유사도를 계산하여 나와 유사한 성향의 고객이 좋아한 상품/콘텐츠를 추천하는 기법



Item-based filtering

아이템들의 유사도와 사용자의 선호도를 기준으로 추천하는 기법 ex) 선호 장르를 분석해서 영화 추천



05. 분석 방법

User based filtering

- User-based CF의 대표적인 유사도

➤ Cosine(COS) :
$$\frac{\sum_{j \in I_a \cap I_b} (R_{a,j})(R_{b,j})}{\sqrt{\sum_{j \in I_a \cap I_b} (R_{a,j})^2} \sqrt{\sum_{j \in I_a \cap I_b} (R_{b,j})^2}}$$

➤ Pearson Correlation Coefficient(PCC) :
$$\frac{\sum_{j \in I_a \cap I_b} (R_{a,j} - \bar{R}_a)(R_{b,j} - \bar{R}_b)}{\sqrt{\sum_{j \in I_a \cap I_b} (R_{a,j} - \bar{R}_a)^2} \sqrt{\sum_{j \in I_a \cap I_b} (R_{b,j} - \bar{R}_b)^2}}$$

➤ Jaccard(JAC) :
$$\frac{|I_a \cap I_b|}{|I_a \cup I_b|}$$

➤ Mean Squared Difference(MSD) :
$$1 - \frac{\sum_{j \in I_a \cap I_b} (R_{a,j} - R_{b,j})^2}{|I_a \cap I_b|}$$

- 평점 예측

➤
$$\hat{R}_{a,j} = \bar{R}_a + \frac{\sum_{b \in RN} Sim(a,b)(R_{b,j} - \bar{R}_b)}{\sum_{b \in RN} |Sim(a,b)|}$$

I_a : 사용자 a 가 평가한 아이템 집합
 I_b : 사용자 b 가 평가한 아이템 집합
 $R_{a,j}$: 아이템 j 에 대한 사용자 a 의 평점
 $R_{b,j}$: 아이템 j 에 대한 사용자 b 의 평점
 \bar{R}_a : 사용자 a 의 전체 평점 평균
 \bar{R}_b : 사용자 b 의 전체 평점 평균
 RN : 아이템 j 를 평가한 상위 유사도 사용자

Item based filtering

- Item-based CF의 대표적인 유사도

➤ Cosine(COS) :
$$\frac{\sum_{u \in U_x \cap U_y} r_{u,x} * r_{u,y}}{\sqrt{\sum_{u \in U_x \cap U_y} r_{u,x}^2} * \sqrt{\sum_{u \in U_x \cap U_y} r_{u,y}^2}}$$

➤ Pearson Correlation Coefficient(PCC) :
$$\frac{\sum_{u \in U_x \cap U_y} (r_{u,x} - \bar{r}_x) * (r_{u,y} - \bar{r}_y)}{\sqrt{\sum_{u \in U_x \cap U_y} (r_{u,x} - \bar{r}_x)^2} \sqrt{\sum_{u \in U_x \cap U_y} (r_{u,y} - \bar{r}_y)^2}}$$

➤ Jaccard(JAC) :
$$\frac{|U_x \cap U_y|}{|U_x \cup U_y|}$$

➤ Mean Squared Difference(MSD) :
$$1 - \frac{\sum_{u \in U_x \cap U_y} (r_{u,x} - r_{u,y})^2}{|U_x \cap U_y|}$$

- 평점 예측

$$p_{u,a} = \bar{r}_a + \frac{\sum_{i=1}^n w_{a,i} * (r_{u,i} - \bar{r}_i)}{\sum_{i=1}^n w_{a,i}}$$

U_x : 아이템 x 를 평가한 사용자 집합

U_y : 아이템 y 를 평가한 사용자 집합

$r_{u,x}$: 아이템 x 에 대한 사용자 u 의 평점

$r_{u,y}$: 아이템 y 에 대한 사용자 u 의 평점

\bar{r}_x : 아이템 x 의 전체 평점 평균

\bar{r}_y : 아이템 y 의 전체 평점 평균

\bar{r}_u : 사용자 u 의 전체 평점 평균

Q 05. 분석 방법

Evaluation Metric

- $MAE(\text{Mean Absolute Error}) = \frac{\sum_{j=1}^N |R_{a,j} - \hat{R}_{a,j}|}{N}$
- $\text{Precision} = \frac{TP}{TP+FP}$
- $\text{Recall} = \frac{FN}{TP+FP}$
- $F1 - \text{Measure} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$
- 본 연구에서는 성능척도로 MAE와 F1 만을 사용

N : the total number of test ratings

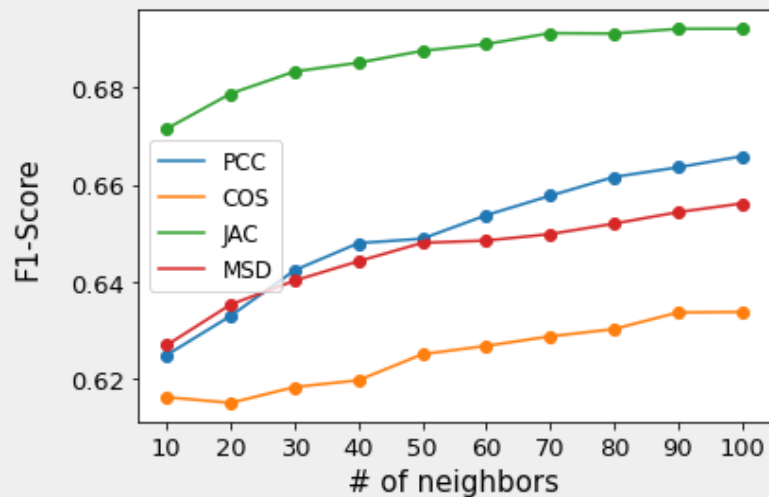
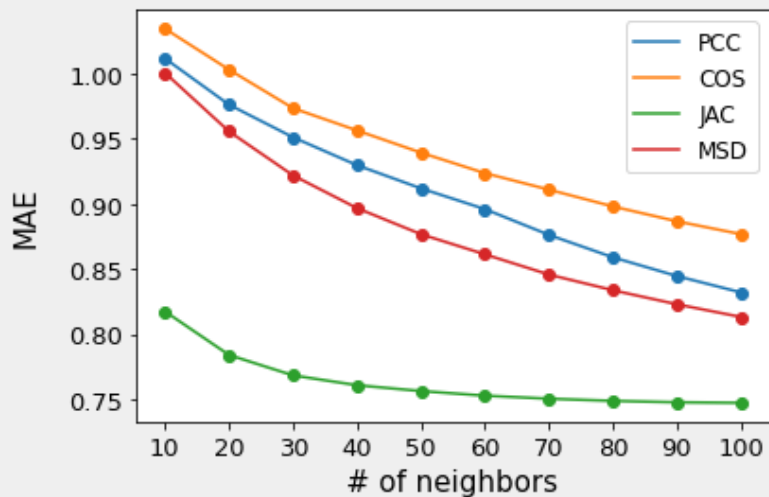
$TP : R_{a,j} \geq \bar{R}_a \ \& \ \hat{R}_{a,j} \geq \bar{R}_a$

$FP : R_{a,j} \geq \bar{R}_a \ \& \ \hat{R}_{a,j} < \bar{R}_a$

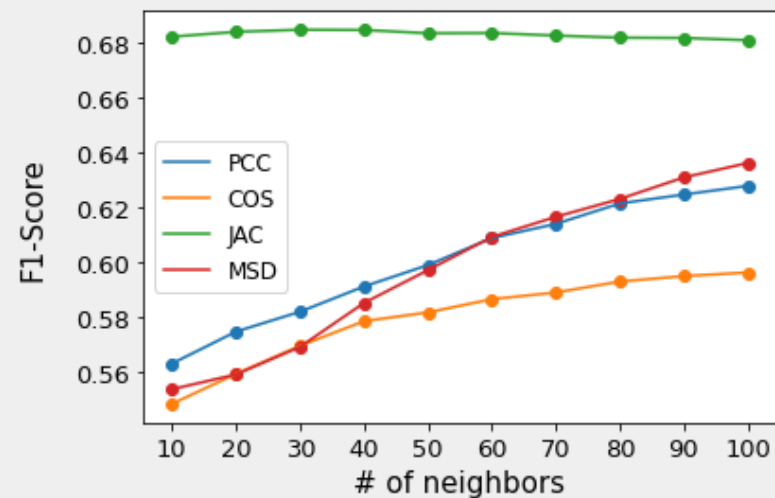
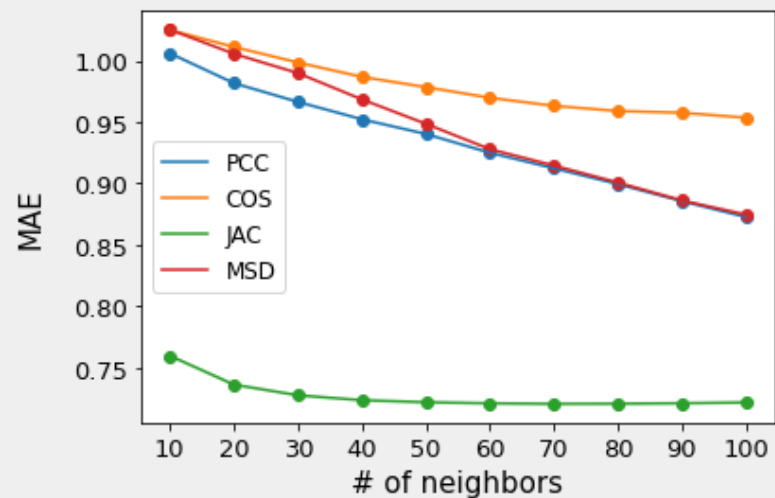
$FN : R_{a,j} < \bar{R}_a \ \& \ \hat{R}_{a,j} \geq \bar{R}_a$

06. 분석결과

User based filtering



Item based filtering



Q 06. 분석 결과

Conclusion

- User based CF 와 Item based CF 각각의 유사도 성능 순서가 비슷
 - JAC 유사도가 MAE와 F1에서 둘 다 가장 좋은 성능을 나타냄
- User based CF & Item based CF JAC 유사도를 최종 추천시스템 모델로 선정
- 타 유사도와 달리 JAC 유사도가 계산 복잡도 낮고, 단순함.
- MovieLens100K 뿐만 아니라 다른 데이터셋에서도 이러한 유사도의 성능이 나오는 지 검증 필요



감사합니다