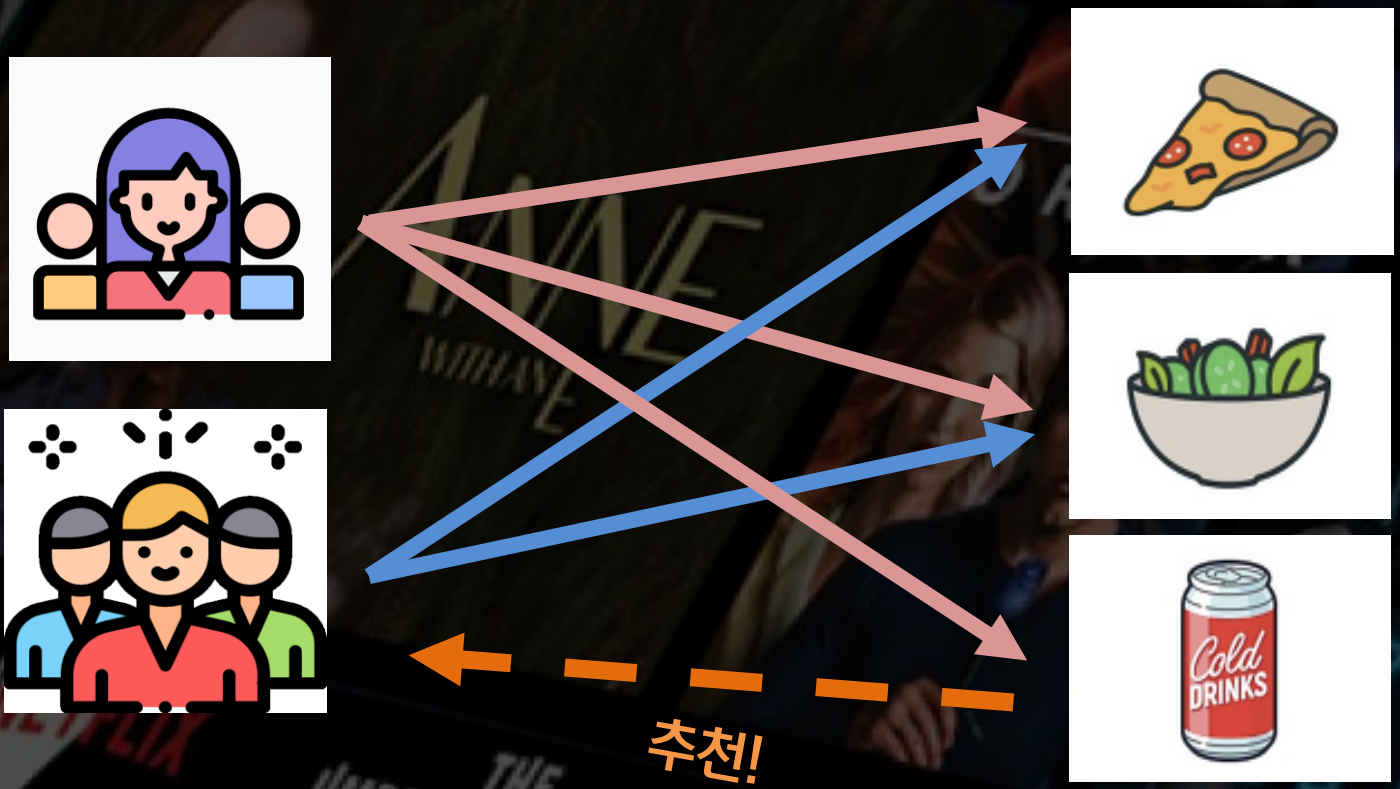
영화 추천 알고리즘

P5 _ 이세나 이정아 오세린 전지우 정윤지

Contents

1. 추천 알고리즘이란?

‘ 방대한 데이터를 일정한 규칙에 따라 분류하여
이용자가 선호할 만한 콘텐츠를 제안하는 기술 ’



1. 추천 알고리즘 _ 목표

‘ ***고객관계관리 (Customer Relationship Management) 의 극대화*** ’

↳ 고객이 누구인지 파악하고, 고객이 소비할 것
같은 상품이나 서비스를 제안하는 마케팅 활동

1. 추천 알고리즘 _ 중요성

ㅇㅇ • 1개월 전

알 수 없는 유튜브 알고리즘이 나를 꽤 괜찮은 곳으로 인도한 것 같다.

3.7천 25

답글 25개 보기

1. 추천 알고리즘 _ 종류

1) 콘텐츠 기반 필터링 (Content-based filtering)

2) 협업 필터링 (Collaborative filtering)

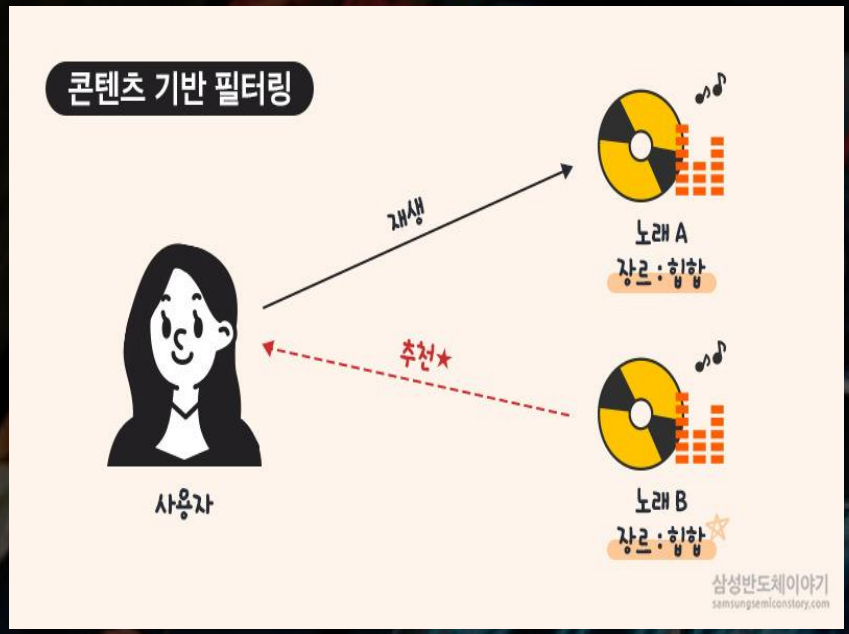
3) 하이브리드 추천 시스템
(Hybrid recommendation system)

1. 추천 알고리즘 _ 종류

1) 콘텐츠 기반 필터링

- 추천 기준 = 이용자가 '이미 소비한' 콘텐츠의 특성

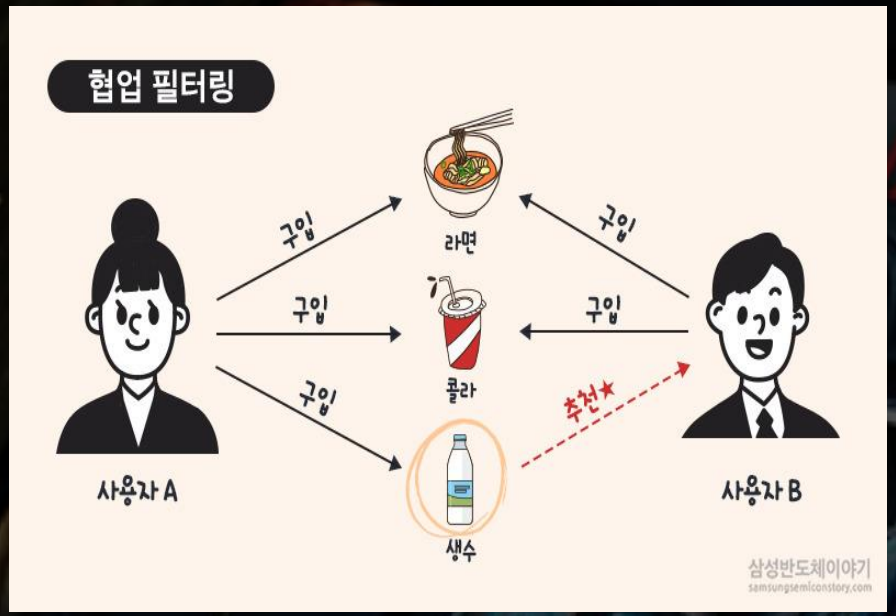
- 단점
  과거 소비 내역에서 벗어난 상품과 서비스의 추천이 어려울 수 있음



1. 추천 알고리즘 _ 종류

2) 협업 필터링

- 분석 기준 = 이용자, 유사 이용자군
- 단점 :
  -신규 사용자에게 아이템도 추천 불가능
  -사용자의 관심이 저조한 항목 추천
  가능성 낮음



1. 추천 알고리즘 _ 종류

3) 하이브리드 추천 시스템

- 협업 필터링 + 콘텐츠 기반 필터링
- 상호보완적 성격

  1) 신규 콘텐츠는 콘텐츠 기반 필터링 기술로 분석

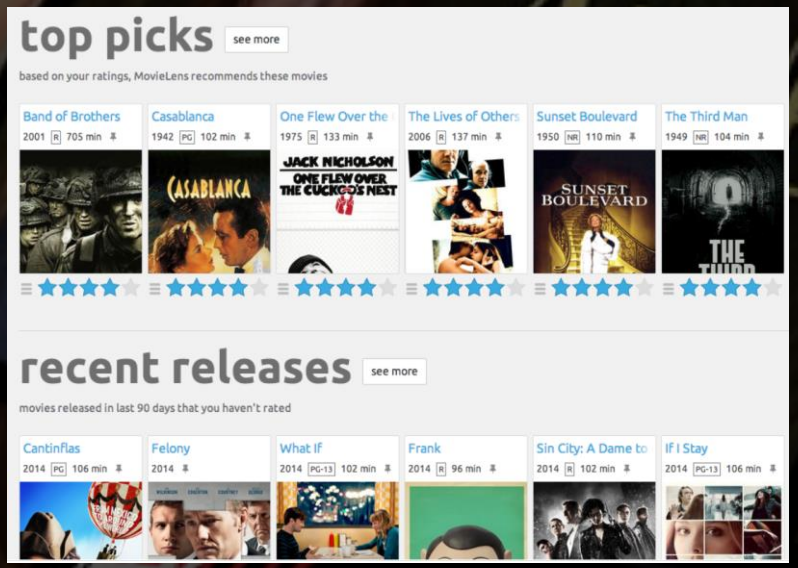
  2) 데이터가 쌓인 후부터 협업 필터링을 사용하여 정확성을 높임

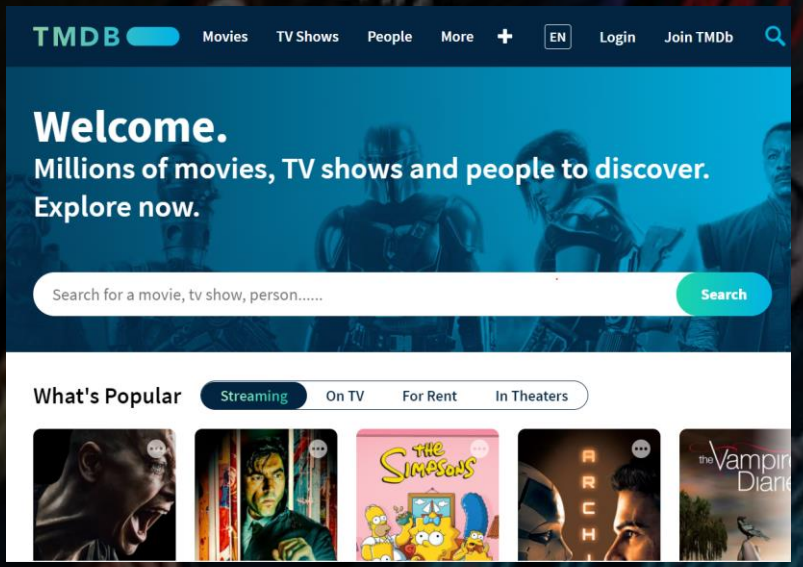
2. 주제 선정 이유

- 영화 산업이 왕성한 요즘 시대에 Netflix, Watcha 등은 어떤 추천 알고리즘을 이용할까?
- 영화 리뷰 데이터 다수

3. Data _ 소개

영화 평점 사이트 Movielens의 영화 목록, rating data set과 영화정보 사이트 TMDB의 영화 세부 정보, 크레디트, Keywords 정보를 합친 데이터





3. Data _ 소개

cast	crew	id
[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gen	[{'credit_id': '52fe4284c3a36847f8024f49', 'department': 'Directing', 'gend	862
[{'cast_id': 1, 'character': 'Alan Parrish', 'credit_id': '52fe44bfc3a36847f80a7c73', 'gender':	[{'credit_id': '52fe44bfc3a36847f80a7cd1', 'department': 'Production', 'ger	8844
[{'cast_id': 2, 'character': 'Max Goldman', 'credit_id': '52fe466a9251416c75077a8d', 'gend	[{'credit_id': '52fe466a9251416c75077a89', 'department': 'Directing', 'genc	15602
[{'cast_id': 1, 'character': "Savannah 'Vannah' Jackson", 'credit_id': '52fe44779251416c91	[{'credit_id': '52fe44779251416c91011acb', 'department': 'Directing', 'genc	31357
[{'cast_id': 1, 'character': 'George Banks', 'credit_id': '52fe44959251416c75039eb9', 'genc	[{'credit_id': '52fe44959251416c75039ed7', 'department': 'Sound', 'gender	11862
[{'cast_id': 25, 'character': 'Lt. Vincent Hanna', 'credit_id': '52fe4292c3a36847f80291f5', 'g	[{'credit_id': '52fe4292c3a36847f802916d', 'department': 'Directing', 'genc	949
[{'cast_id': 1, 'character': 'Linus Larrabee', 'credit_id': '52fe44959251416c75039d97', 'gen	[{'credit_id': '52fe44959251416c75039da9', 'department': 'Directing', 'gen	11860
[{'cast_id': 2, 'character': 'Tom Sawyer', 'credit_id': '52fe46bdc3a36847f810f771', 'gender'	[{'credit_id': '52fe46bdc3a36847f810f797', 'department': 'Writing', 'gender	45325
[{'cast_id': 1, 'character': 'Darren Francis Thomas McCord', 'credit_id': '52fe44dbc3a3684	[{'credit_id': '52fe44dbc3a36847f80ae0f1', 'department': 'Directing', 'gend	9091
[{'cast_id': 1, 'character': 'James Bond', 'credit_id': '52fe426ec3a36847f801e10d', 'gender	[{'credit_id': '52fe426ec3a36847f801e14b', 'department': 'Directing', 'genc	710
[{'cast_id': 1, 'character': 'Andrew Shepherd', 'credit_id': '52fe44dac3a36847f80adf79', 'ge	[{'credit_id': '52fe44dac3a36847f80adfa3', 'department': 'Camera', 'gender	9087
[{'cast_id': 9, 'character': 'Count Dracula', 'credit_id': '52fe44b79251416c7503e811', 'genc	[{'credit_id': '52fe44b79251416c7503e7fb', 'department': 'Editing', 'gender	12110
[{'cast_id': 1, 'character': 'Balto (voice)', 'credit_id': '52fe4409c3a368484e00bb65', 'gende	[{'credit_id': '593f24b9c3a3680369002371', 'department': 'Production', 'ge	21032
[{'cast_id': 1, 'character': 'Richard Nixon', 'credit_id': '52fe43c59251416c7501d6e1', 'gend	[{'credit_id': '52fe43c59251416c7501d6f3', 'department': 'Directing', 'genc	10858
[{'cast_id': 1, 'character': 'Morgan Adams', 'credit_id': '52fe42f4c3a36847f802f65f', 'gende	[{'credit_id': '52fe42f4c3a36847f802f69f', 'department': 'Camera', 'gender'	1408
[{'cast_id': 4, 'character': "Sam 'Ace' Rothstein", 'credit_id': '52fe424dc3a36847f80139d1',	[{'credit_id': '52fe424dc3a36847f80139cd', 'department': 'Directing', 'genc	524
[{'cast_id': 6, 'character': 'Marianne Dashwood', 'credit_id': '52fe43cec3a36847f807103b',	[{'credit_id': '52fe43cec3a36847f807101f', 'department': 'Directing', 'gend	4584
[{'cast_id': 42, 'character': 'Ted the Bellhop', 'credit_id': '52fe420dc3a36847f80001b7', 'ge	[{'credit_id': '52fe420dc3a36847f800011b', 'department': 'Sound', 'gender	5
[{'cast_id': 1, 'character': 'Ace Ventura', 'credit_id': '52fe44dfc3a36847f80af279', 'gender':	[{'credit_id': '52fe44dfc3a36847f80af28b', 'department': 'Directing', 'gende	9273
[{'cast_id': 1, 'character': 'John', 'credit_id': '52fe44509251416c7503059b', 'gender': 2, 'id'	[{'credit_id': '52fe44509251416c750305a1', 'department': 'Directing', 'genc	11517

* Credits Data : 영화 제작 등에 참여한 사람들의 이름에 관한 데이터

3. Data _ 소개

id	keywords
862	[{'id': 931, 'name': 'jealousy'}, {'id': 4290, 'name': 'toy'}, {'id': 5202, 'name': 'boy'}, {'id': 6054, 'name': 'friendship'}, {'id': 9713, 'name': 'friends'}, {'id': 9823, 'name
8844	[{'id': 10090, 'name': 'board game'}, {'id': 10941, 'name': 'disappearance'}, {'id': 15101, 'name': "based on children's book"}, {'id': 33467, 'name': 'new home'},
15602	[{'id': 1495, 'name': 'fishing'}, {'id': 12392, 'name': 'best friend'}, {'id': 179431, 'name': 'duringcreditsstinger'}, {'id': 208510, 'name': 'old men'}]
31357	[{'id': 818, 'name': 'based on novel'}, {'id': 10131, 'name': 'interracial relationship'}, {'id': 14768, 'name': 'single mother'}, {'id': 15160, 'name': 'divorce'}, {'id': 33
11862	[{'id': 1009, 'name': 'baby'}, {'id': 1599, 'name': 'midlife crisis'}, {'id': 2246, 'name': 'confidence'}, {'id': 4995, 'name': 'aging'}, {'id': 5600, 'name': 'daughter'}, {'id':
949	[{'id': 642, 'name': 'robbery'}, {'id': 703, 'name': 'detective'}, {'id': 974, 'name': 'bank'}, {'id': 1523, 'name': 'obsession'}, {'id': 3713, 'name': 'chase'}, {'id': 7281, 'n
11860	[{'id': 90, 'name': 'paris'}, {'id': 380, 'name': 'brother brother relationship'}, {'id': 2072, 'name': 'chauffeur'}, {'id': 9398, 'name': 'long island'}, {'id': 9492, 'name':
45325	[]
9091	[{'id': 949, 'name': 'terrorist'}, {'id': 1562, 'name': 'hostage'}, {'id': 1653, 'name': 'explosive'}, {'id': 193533, 'name': 'vice president'}]
710	[{'id': 701, 'name': 'cuba'}, {'id': 769, 'name': 'falsely accused'}, {'id': 1308, 'name': 'secret identity'}, {'id': 2812, 'name': 'computer virus'}, {'id': 3268, 'name': 'se
9087	[{'id': 833, 'name': 'white house'}, {'id': 840, 'name': 'usa president'}, {'id': 1605, 'name': 'new love'}, {'id': 33476, 'name': 'widower'}, {'id': 211505, 'name': 'wildl
12110	[{'id': 3633, 'name': 'dracula'}, {'id': 11931, 'name': 'spoof'}]
21032	[{'id': 1994, 'name': 'wolf'}, {'id': 6411, 'name': 'dog-sledding race'}, {'id': 9880, 'name': 'alaska'}, {'id': 15162, 'name': 'dog'}, {'id': 15169, 'name': 'goose'}, {'id': 1
10858	[{'id': 840, 'name': 'usa president'}, {'id': 2946, 'name': 'presidential election'}, {'id': 4240, 'name': 'watergate scandal'}, {'id': 5565, 'name': 'biography'}, {'id': 608
1408	[{'id': 911, 'name': 'exotic island'}, {'id': 1454, 'name': 'treasure'}, {'id': 1969, 'name': 'map'}, {'id': 3799, 'name': 'ship'}, {'id': 5470, 'name': 'scalp'}, {'id': 12988, 'n
524	[{'id': 383, 'name': 'poker'}, {'id': 726, 'name': 'drug abuse'}, {'id': 1228, 'name': '1970s'}, {'id': 2635, 'name': 'overdose'}, {'id': 33625, 'name': 'illegal prostitution
4584	[{'id': 420, 'name': 'bowling'}, {'id': 818, 'name': 'based on novel'}, {'id': 964, 'name': 'servant'}, {'id': 2755, 'name': 'country life'}, {'id': 7564, 'name': 'jane auste
5	[{'id': 612, 'name': 'hotel'}, {'id': 613, 'name': "new year's eve"}, {'id': 616, 'name': 'witch'}, {'id': 622, 'name': 'bet'}, {'id': 922, 'name': 'hotel room'}, {'id': 2700, 'r
9273	[{'id': 409, 'name': 'africa'}, {'id': 1551, 'name': 'indigenous'}, {'id': 2526, 'name': 'human animal relationship'}, {'id': 5155, 'name': 'bat'}]
11517	[{'id': 380, 'name': 'brother brother relationship'}, {'id': 1552, 'name': 'subway'}, {'id': 14512, 'name': 'new york city'}, {'id': 155735, 'name': 'new york subway'},

* Keywords Data : 영화의 plot과 관련된 키워드를 정리한 데이터

3. Dataset

Data ①. Full Dataset

27,000명의 사용자

45,000개의 영화

26,000,000개의 평점

750,000개의 태그

links 데이터(영화 제목)

movieId	imdbId	tmdbId
1	114709	862
2	113497	8844
3	113228	15602
4	114885	31357
5	113041	11862
6	113277	949
7	114319	11860
8	112302	45325
9	114576	9091
10	113189	710
11	112346	9087
12	112896	12110
13	112453	21032
14	113987	10858
15	112760	1408
16	112641	524
17	114388	4584
18	113101	5
19	112281	9273
20	113845	11517
21	113161	8012
22	112722	1710
23	112401	9691
24	114168	12665
25	113627	451
26	114057	16420
27	114011	9263
28	114117	17015
29	112682	902

Ratings 데이터(평점)

userId	movieId	rating	timestamp
1	110	1	1425941529
1	147	4.5	1425942435
1	858	5	1425941523
1	1221	5	1425941546
1	1246	5	1425941556
1	1968	4	1425942148
1	2762	4.5	1425941300
1	2918	5	1425941593
1	2959	4	1425941601
1	4226	4	1425942228
1	4878	5	1425941434
1	5577	5	1425941397
1	33794	4	1425942005
1	54503	3.5	1425941313
1	58559	4	1425942007
1	59315	5	1425941502
1	68358	5	1425941464
1	69844	5	1425942139
1	73017	5	1425942699
1	81834	5	1425942133
1	91500	2.5	1425942647
1	91542	5	1425942618
1	92439	5	1425941424
1	96821	5	1425941382
1	98809	0.5	1425942640
1	99114	4	1425941667

3. Dataset

Data ②. Small Dataset

700명의 사용자

9,000개의 영화

100,000개의 평점

1,300개의 태그

links 데이터(영화 제목)

movieId	imdbId	tmdbId
1	114709	862
2	113497	8844
3	113228	15602
4	114885	31357
5	113041	11862
6	113277	949
7	114319	11860
8	112302	45325
9	114576	9091
10	113189	710
11	112346	9087
12	112896	12110
13	112453	21032
14	113987	10858
15	112760	1408
16	112641	524
17	114388	4584
18	113101	5
19	112281	9273
20	113845	11517
21	113161	8012
22	112722	1710
23	112401	9691
24	114168	12665
25	113627	451
26	114057	16420

Ratings 데이터(평점)

userId	movieId	rating	timestamp
1	31	2.5	1260759144
1	1029	3	1260759179
1	1061	3	1260759182
1	1129	2	1260759185
1	1172	4	1260759205
1	1263	2	1260759151
1	1287	2	1260759187
1	1293	2	1260759148
1	1339	3.5	1260759125
1	1343	2	1260759131
1	1371	2.5	1260759135
1	1405	1	1260759203
1	1953	4	1260759191
1	2105	4	1260759139
1	2150	3	1260759194
1	2193	2	1260759198
1	2294	2	1260759108
1	2455	2.5	1260759113
1	2968	1	1260759200
1	3671	3	1260759117
2	10	4	835355493
2	17	5	835355681
2	39	5	835355604
2	47	4	835355552
2	50	4	835355586
2	52	3	835356031

4. 분석 계획

Hybrid recommendation system

Why?

- 최근 연구에서 collaborative filtering 과 content-based filtering을 섞은 hybrid 접근법이 더 효과적일 수 있다고 설명
- Netflix가 Hybrid 방식 사용

4. 분석 계획

- 컨텐츠 필터링과 협업 필터링의 이론 및 코딩 학습 필요…!

- 우리 데이터에 대한 추천 알고리즘 적용 방식 탐색
  - Full / Small dataset 이용 방식
  - Hybrid recommendation system 구현 방식

- R과 Python을 이용해 실제 모델 구축

4. 분석 계획

Hybrid recommendation system

Content Based Methods + Collaborative Filtering

1. Output integration
2. Algorithm integration

* 참고 사이트

1. 넷플릭스의 영화 추천 알고리즘(https://brunch.co.kr/@cysstory/159)
2. 추천 알고리즘 SVD (https://seing.tistory.com/67)
3. 추천 알고리즘, 내 취향을 어떻게 그렇게 잘 알아?
   (https://brunch.co.kr/@biginsight/15)
4. [Python 머신러닝] 9장. 추천시스템 (Recommendation System)
(https://joyfuls.tistory.com/66)


Q & A


THANK YOU!