코드스테이츠 **Project 1**

책 추천 시스템

AI_06_강지호

목차

01. 프로젝트 개요

■ 프로젝트 주제 및 목적

유저가 구입한 책 정보를 토대로 관심을 가질법한 다른 아이템 추천.
이를 통해 구매 유도 및 매출 증대에 기여

■ 프로젝트 선정 배경

개인화 추천 시스템은 재구매 유도, 충성고객 확보의 효과가 있고,
매출 증대에 직접적이며 강력한 영향력을 행사하므로 많은 기업에서
추천 시스템을 이용 중. 추천시스템에 대한 깊은 이해를 위해 기본적인
모델인 Content-Based Filtering(CB) 과 Collaborative Filtering(CF)을 사용하여
데이터를 분석해본다.

02. 프로젝트 진행 방법

■ 웹 크롤링 - 목표 데이터

yes24 홈페이지 > 베스트 > 카테고리별 > Top 1000개의 책에 대한
> 책 제목, 부제, 작가, 가격, 카테고리, 책 내용, 리뷰 수집

02. 프로젝트 진행 방법

1) 베스트 셀러

- 웹 크

-

2) 26개 카테고리

3) 50페이지 x 20개의 책 = 1000권

02. 프로[illegible]법

데이터 1) 책 ID
데이터 2) 책 제목
데이터 3) 부제
데이터 4) 작가
데이터 5) 출간일
데이터 6) 평점
데이터 7) 가격
+ 하단
데이터 8) 상위 카테고리
데이터 9) 하위 카테고리
데이터 10) 내용
소득공제
오늘의책
2021 올해의 책
어떻게 말해줘야 할까
오은영의 현실밀착 육아회화
오은영 저 | 차상미 그림 | 김영사 | 2020년 10월 25일
9.8
회원리뷰(219건) | 판매지수 912,978
베스트 국내도서 21위
구매혜택
[단독] 월간 채널예스 12월호 (포인트차감)
정가 17,500원
판매가 15,750원 (10% 할인)
북클럽머니 최대혜택가 14,250원
YES포인트 870원 (5% 적립)
5만원이상 구매 시 2천원 추가적립
결제혜택 카드/간편결제 혜택을 확인하세요
배송안내 서울특별시 영등포구 은행로 11(여의도동,일신빌딩) 지역변경
오은영의 현실밀착 육아회화
어떻게 말해줘야 할까
버럭 하지 않고 분명하게 알려주는 방법
<요즘 육아 금쪽같은 내 새끼> 채널A
오은영 박사의 최신작

02. 프로젝트 진행 방법

■ 웹

●

02. 프로젝트 진행 방법

df_book

	id	name	nameE	writer	category1	category2	summary	price	rate	date
0	105536311	귀멸의 칼날 외전	None	고토게 코요하루	만화/라이트노벨	판타지	수주(水柱) 토미오카가 만난 마타기 소녀 야에는 아버지의 원수를 갚기 위해 산으로 ...	6000.0	10.0	2021-12-13
1	105518025	도쿄 리벤저스 17		와쿠이 켄	만화/라이트노벨	액션	최신 타임 슬립 서스펜스 제17권!! 이누이의 청을 받아들여 11대 흑룡 총장이 된...	5000.0	10.0	2021-12-10
2	105518067	도쿄 리벤저스 18		와쿠이 켄	만화/라이트노벨	액션	최신 타임 슬립 서스펜스 제18권!!₩n₩n전의를 잃은 마이키가 빠진 상태에서 천축...	5000.0	9.7	2021-12-10

df_review

book_id	reviewer_id	reviewer	rate	review
105536311	9379785	용*	5.0	정발되기만을 손꼽아 기다렸습니다! 이렇게나마 렌고쿠를 볼 수 있어서 좋았어요.. 귀...
105536311	15192997	작***리	5.0	귀멸의 칼날 외전으로 다시 귀멸의 칼날을 만날 수 있어 좋다. 일단 만화는 성공적으...
105536311	16853388	쿠*	5.0	귀멸의 칼날의 애니메이션이 몇 달 전부터 시작되었고 2주 전부터는 새로운 유곽편이 ...
105536311	5970460	카**스	5.0	귀멸의 칼날 극장판의 흥행에 크게 성공하면서 귀멸의 칼날은 전국적으로 인기를 끌게 ...
105536311	9155449	b*****o	5.0	안녕하세요. 제가 이번에 리뷰할 만화는 고토게 코요하루 작가님의 '귀멸의 칼날' 외...

02. 프로젝트 진행 방법

■ 웹 크롤링 - 수집 데이터

- 14개 카테고리, 5,261권에 대한 정보 수집
- 리뷰가 있는 책의 갯수: 4,598권
- 리뷰 총 갯수: 246,753개
- 리뷰 남긴 사람 수: 89,865명

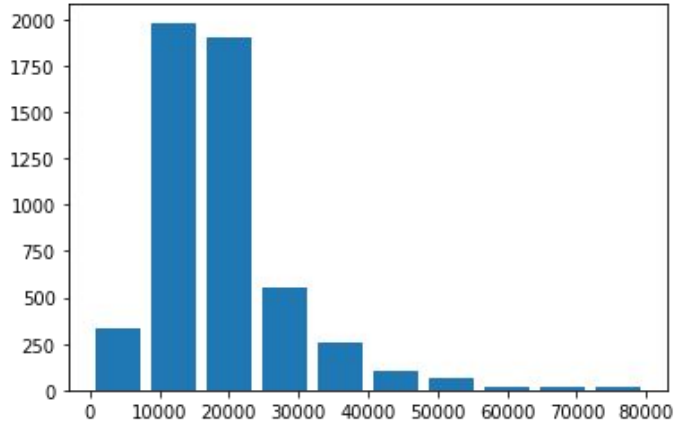
자연과학	849	역사	400
수험서 자격증	648	만화/라이트노벨	352
에세이	497	경제 경영	313
대학교재	471	자기 계발	159
유아	459	소설/시/희곡	138
인물	440	초등참고서	71
가정 살림	406	청소년	58
총합: 5261			

02. 프로젝트 진행 방법

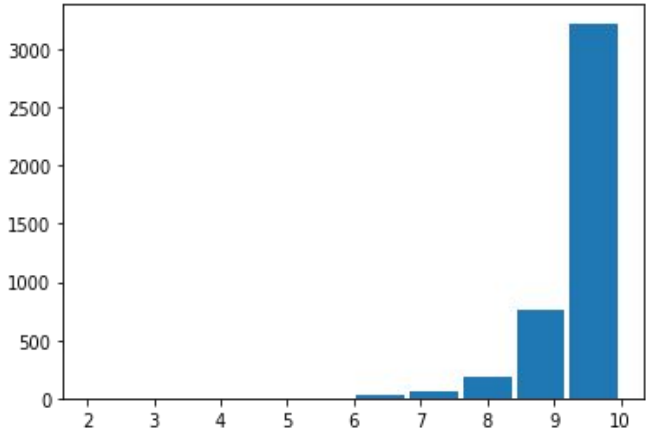
■ 웹 크롤링 - 수집 데이터

- 14개 카테고리, 5,261권에 대한 정보 수집
- 리뷰가 있는 책의 갯수: 4,598권
- 리뷰 총 갯수: 246,753개
- 리뷰 남긴 사람 수: 89,865명

[가격 분포 히스토그램]



[평점 분포 히스토그램]



02. 프로젝트 진행 방법

■ 웹 크롤링 - 수집 데이터

- 14개 카테고리, 5,261권에 대한 정보 수집
- 리뷰가 있는 책의 갯수: 4,598권
- 리뷰 총 갯수: 246,753개
- 리뷰 남긴 사람 수: 89,865명

[카테고리별 데이터]

categoryl	price	rate	review
수험서 자격증	25609	9.80	13
초등참고서	12454	9.71	74
만화/라이트노벨	10059	9.66	56
가정 살림	17700	9.65	56
유아	13359	9.62	63
청소년	15333	9.44	83
에세이	15242	9.38	67
경제 경영	18232	9.38	82
대학교재	27365	9.37	5
자기계발	16294	9.36	112
역사	22192	9.31	39
자연과학	20407	9.28	25
소설/시/희곡	15509	9.24	245
인물	18615	8.82	16

02. 프로젝트 진행 방법 - CB

■ Content-Based Filtering (CB)

- CB: 아이템의 특성과 비슷한 특성을 가진 다른 아이템을 추천해주는 방식
- 기준 특성: 줄거리

name	summary
귀멸의 칼날 외전	수주(水柱) 토미오카가 만난 마타기 소녀 야에는 아버지의 원수를 갚기 위해 산으로 ...
도쿄 리벤저스 17	최신 타임 슬립 서스펜스 제17권!! 이누이의 청을 받아들여 11대 흑룡 총장이 된...
도쿄 리벤저스 18	최신 타임 슬립 서스펜스 제18권!!₩n₩n전의를 잃은 마이키가 빠진 상태에서 천축...

- 전처리: 한글, 숫자, 영어 제외 모두 제거, 불용어 제거

name	summary
귀멸의 칼날 외전	수주 토미 오카 만난 마타 기 소녀 야 에는 아버지 원수 갚다 위해 산 들어가다 그...
도쿄 리벤저스 17	최신 타임 슬립 서스펜스 제 17 권 이누이 청 을 받아들이다 11 대다 흑룡 총장...
도쿄 리벤저스 18	최신 타임 슬립 서스펜스 제 18 권 전의 잃다 마이키 빠지다 상태 에서 천축 치다...

03. 프로젝트 결과 - CB

■ Content-Based Filtering (CB)

- tf-idf로 전처리된 줄거리 특성 벡터화
- 코사인 유사도로 줄거리끼리의 유사도 계산

name	summary	score
실전 투자의 정석	주식 투자 로 성공하다 개인 투자가 슈퍼 개미 라고 부르다 한편 투자 수익 만으로 ...	1.000000
평생 부자로 사는 주식투자	주식 투자 고수 조언 을 귀 담 아 들다 이유 2020년 부터 불어 닥치다 주식 투...	0.576398
강방천존리와 함께하는 나의 첫 주식 교과서	대한민국 주식 투자 살 아 있다 전설 강 방천 존리 첫 합동 프로젝트 투자 시대 평...	0.515958
할 수 있다 퀀트 투자	가장 쉬다 편하다 안정 적 고 수익 내다 투자 비법 수백 편 논문 에서 뽑다 내다 ...	0.462399
부동산 투자 이렇게 쉬웠어	상승 장 물론 하락 장 에서도 수익 내다 방법 있다 기본 원리 만 알 면 부동산 투...	0.462120
절대수익 투자법칙	자산 저절로 늘어나다 마법 투자 왕 김 단테 추다 간 펀딩 2430 달성 성 공 올...	0.460918
투자는 디테일에 있다	슈퍼 개미 김정환 투자 바이블 드디어 추다 간 성공하다 1 투자자 어떻다 남다르다 ...	0.435887
돈의 흐름	인플레 위기 아니다 기회 다 앞 오르다 것 을 찾다 돈 을 벌 수 있다 43만 경제...	0.419732
다모다란의 투자 전략 바이블	모든 투자 전략 을 의심 하고 검증 하 라 그리고 반드시 이해 하 라 세계 적 석 ...	0.408329
전설로 떠나는 월가의 영웅	월 가의 전설 적 인 투자자 피터 린치 쓸다 주식 투자 고전 어떤 기업 이든 공부 ...	0.405965
심리투자 불변의 법칙	20년 넘다 투자 고수 추천 심리 투자 분야 최고 책 감정 흔들리다 않다 투자자 하...	0.402995

03. 프로젝트 결과 - CB

■ Content-Based Filtering (CB)

- tf-idf로 전처리된 줄거리 특성 벡터화
- 코사인 유사도로 줄거리끼리의 유사도 계산

name	summary	score
아파트 청약 이렇게 쉬웠어	청약 최고 수 만난 수천 명 에게 집 생기다 가점 낮다 이미 집 있다 운 없다 청약...	1.000000
대한민국 재건축 재개발 지도	청약 문턱 높다 신축 비싸다 당신 에게 천국 가다 비상구 열리다 어렵다 숫자 모르다...	0.499322
왕초보도 바로 돈 버는 부동산 경매의 기술	오르다 부동산 로또 보다 힘드다 청약 당첨 누구 나 싸다 살 수 있다 경매 답 이다...	0.163376
쇼킹부동산 1	모두 관심 아파트 쏠리다 있다 무 주택 내 집 을 구 위해 유 주택 계속 가격 오르...	0.155031
수학은 어떻게 무기가 되는가	수학 삶 도움 되다 문과 생 모르다 수학 쓸모 책 세상 을 움직이다 것 숫자 이고 ...	0.119143
신은 주사위 놀이를 하지 않는다	통계학 대 영 제국 훈장 을 받다 데이비드 핸드 미스터리 사건 다섯 가지 법칙 을 ...	0.112332
딱 2년 안에 무조건 돈 버는 부동산 투자 시크릿	렘군 푸릉 부동산 사관학교 클래스 101 강의 입 소문 만으로 오픈 하루 만에 완판...	0.100321
거인의 포트폴리오	돈 버 사람 이렇게 주식 을 사다 따르다 수익 나 대가 검증 되다 투자 시스템 주식...	0.095673
돈의 속성	베스트셀러 종합 1 위 경제 경영 17 주 연속 1 위 유튜브 1100만 명 시청 ...	0.093374
오늘부터 건물주	이제 월급 대신 월세 받다 28 세 사회 초년 생 1년 만에 부동산 3 채 건물 주...	0.093148
이상하게 돈 걱정 없는 사람들의 비밀	따르다 해도 돈복 생기다 부자 마인드 27 수록 돈 기쁘다 쓰다 사람 에게 흐르다 ...	0.093009

03. 프로젝트 결과 - CB

■ Content-Based Filtering (CB)

- tf-idf로 전처리된 줄거리 특성 벡터화
- 코사인 유사도로 줄거리끼리의 유사도 계산

name	summary	score
도쿄 리벤저스 1	시리즈 누 계 100만 부일본 현지 에서 가장 자다 팔리다 타입 슬립 만화 새롭다 ...	1.000000
도쿄 리벤저스 3	학창시절 불량 배다 한심하다 남자 타 케미 치다 폭력배 연합 도쿄 만지다 희 에게 ...	0.266785
도쿄 리벤저스 5	드라켄 죽음 을 막다 과거 바꾸다 데 성공하다 타 케미 치 현대 에서 히나타 재회 ...	0.251080
도쿄 리벤저스 16	최신 타임 슬립 서스펜스 제 16 권 현대 에서 요코하마 천축 그 총장 쿠로 카 이...	0.229752
도쿄 리벤저스 14	최신 타임 슬립 서스펜스 제 14 권 임무 마치 고 12년 후 로 돌아오다 타 케미...	0.213946
도쿄 리벤저스 6	최신 타임 슬립 서스펜스 제 6 권도 만의 톱 되다 결심 타 케미 치바 지르다 다시...	0.207141
도쿄 리벤저스 10	최신 타임 슬립 서스펜스 제 10 권키 사키 에게 속 아 치 후유 잃다 타 케미 치...	0.198190
도쿄 리벤저스 18	최신 타임 슬립 서스펜스 제 18 권 전의 잃다 마이키 빠지다 상태 에서 천축 치다...	0.194611
도쿄 리벤저스 15	최신 타임 슬립 서스펜스 제 15 권 또다시 과거 로 돌아가다 타 케미 치다 만의 ...	0.184129
도쿄 리벤저스 11	최신 타임 슬립 서스펜스 제 11 권 만이 거대 악 되다 원인 인 흉악 폭주족 흑룡...	0.168565

02. 프로젝트 진행 방법 - CF

■ Collaborative Filtering (CF)

- CF: 유저와 비슷한 성향의 다른 유저가 읽은 책을 추천해주는 방법
- 기준: 평점
- item-based: 유저의 수가 아이템의 수보다 현저히 많았기 때문에 사용
- 코사인 유사도 사용

reviewer_id	100415	100691	100732	100867	101349	101419	101601	101653	102012
book_id									
599	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
712	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
773	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
784	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

03. 프로젝트 결과 - CF

■ Collaborative Filtering (CF)

- CF: 유저와 비슷한 성향의 다른 유저가 읽은 책을 추천해주는 방법
- 기준: 평점
- item-based: 유저의 수가 아이템의 수보다 현저히 많았기 때문에 사용
- 코사인 유사도 사용

실전 투자의 정석	1.0000
오피스 누나 이야기	0.1478
차트의 기술	0.1290
내 안에 삶의 나침반이 있다	0.1207
세계 최고의 여행기 열하일기 하	0.1207
나의 문화유산답사기 일본편 5 교토의 정원과 다도	0.0922
할 수 있다! 퀀트 투자	0.0865
투자는 디테일에 있다	0.0840
세상에서 가장 쉬운 상대성이론	0.0823
부의 시작	0.0775

아파트 청약 이렇게 쉬웠어?	1.0000
싱글맘 부동산 경매로 홀로서기	0.1273
엑시트 EXIT	0.1108
자기계발과 PR의 선구자들	0.0804
대한민국 땅따먹기	0.0685
경매 권리분석 이렇게 쉬웠어?	0.0622
알기 쉬운 선형대수	0.0569
부동산 상승 신호 하락 신호	0.0501
민법총칙	0.0421
지성의 돈되는 부동산 1인법인	0.0414

도쿄 리벤저스 1	1.0000
도쿄 리벤저스 4	0.7333
도쿄 리벤저스 3	0.6628
도쿄 리벤저스 5	0.6149
도쿄 리벤저스 6	0.5377
도쿄 리벤저스 8	0.5091
도쿄 리벤저스 7	0.4929
도쿄 리벤저스 10	0.3343
도쿄 리벤저스 11	0.3254

04. 셀프 피드백

■ 프로젝트 관리 측면

업무 중요도에 따른 시간 분배

크롤링 코드를 구현하는데에 4일을 사용함.
본 프로젝트의 목표는 추천 시스템의 기본적인 모델에 대한 이해도 향상이기 때문에 데이터 수집보다는 분석에 더 많은 시간을 투자했어야 함.

■ 모델 측면

1. CB의 경우, 소비자의 취향이 변할 경우 새로운 책을 추천해주지 못함.
   CF의 경우, 유저-아이템의 평점 매트릭스가 sparse하기 때문에 정확도가 떨어지며, 신규 유저의 경우 정보가 없으므로 추천이 어려움.
2. 특성을 CB는 줄거리, CF는 평점으로 각 하나씩만 사용함. 다양한 특성을 고려한 모델을 만든다면 정확도가 높아질 것.
3. CB와 CF를 합친 Hybrid model로 예측한다면 더 정확한 예측이 가능할 것.

04. 참고 자료

https://blog.naver.com/PostView.naver?blogId=myincizor&logNo=221829075434&parentCategoryNo=&categoryNo=6&viewDate=&isShowPopularPosts=false&from=postView

https://pearlluck.tistory.com/668

https://velog.io/@suminwooo/%EC%B6%94%EC%B2%9C-%EC%8B%9C%EC%8A%A4%ED%85%9C3

https://www.kaggle.com/gspmoreira/recommender-systems-in-python-101