홈

소설/문학

오디오북스

자기계발

경제/경영

인문/교양

역사

어린이(동화)

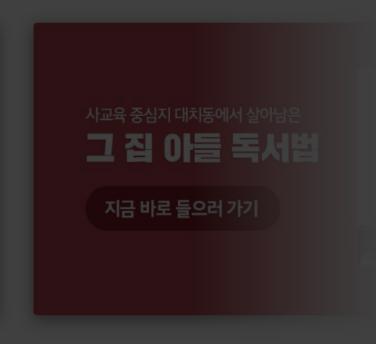
해외 오디오북

추천시스템 만들어주세요! 안돼요! 왜냐하면…



좋아요. 잘들었어요. 감사합니다.

지금 바로 들으러 가기



따끈따끈~ 새로 나온 오디오북











데잇걸즈 4기 | 크림빵 김연주, 김한나, 김현진, 박지현, 최미리



추천시스템 만들어주세요!



팟빵 오디오북

안녕하세요 크림빵 컨설팅의 김현진 선생님. 함께 일할 협력사가 필요하다는 소식 듣고 연락 드립니다. 현재 저희 회사에서 추천시스템이 필요한데, 개발 프로젝트 진행을 부탁드립니다. 🙂 (edited)

크림빵의 신입 컨설턴트. 팟빵 오디오북 콘텐츠 추천 시스템 개발 프로젝트를 맡게 되었다.



INDEX

Intro. 주제 소개

Part1. 1. 오디오북

2. 추천시스템 - 협업필터링

3. 팟빵 리뷰 데이터의 한계

Part2. 1. 오디오북 서비스별 댓글 텍스트 분석

2. 오디오북 서비스별 마케팅 방식 비교

3. 문제점 요약 및 해결책

Outro. 팀소개



한국 오디오북 시장의 성장성

[그림1] 2020년 오디오북 예상 규모



"책, 자주 들으세요?" 4조원 규모 '듣는 책' 오디오북 시장 뜬다 | 인터비즈

북유럽 오디오북 '스토리텔' 한국 상륙

△ 취재기자 배수진 │ ② 승인 2019.11.29 11:19 │ ഈ 댓글 0

교보문고도 오디오북 시장 진출

약 4천종 보유, '교보eBook' 앱에 전용 플레이어 탑재해 서비스 개시

김슬기 기자 | 입력 : 2020.07.22 11:07:14 🕡 0

직장인이 듣는 오디오북 '윌라', 기업서비스 1년 만에 700% 성장

입력: 2020,11,16 22:26









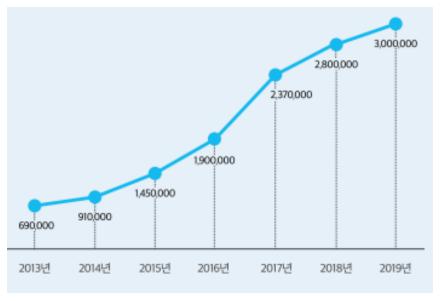






팟빵 오디오북 소개

[그림2] 팟빵 월 평균 이용자 수

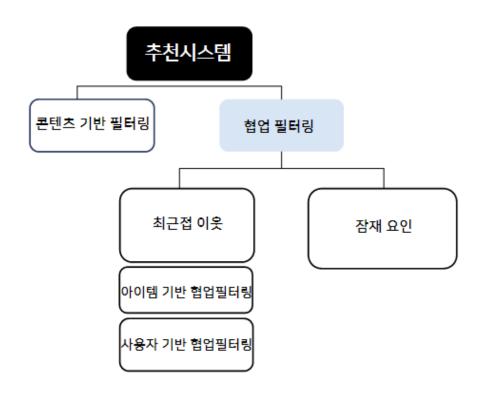


출처: 팟빵(2020)

- 팟빵은 국내 최초의 팟캐스트 플랫폼, 2019년 기준 이용자수 300만명 돌파
- 청각 자료를 이용한 정보 습득에 익숙한 사용자 보유 -> 오디오북으로의 전환 용이
- 라디오나 팟캐스트와 달리 오디오북은 유료서비스이기 때문에 유저 맞춤형 콘텐츠에 대한 수요가 있다.



협업 필터링



협업 필터링

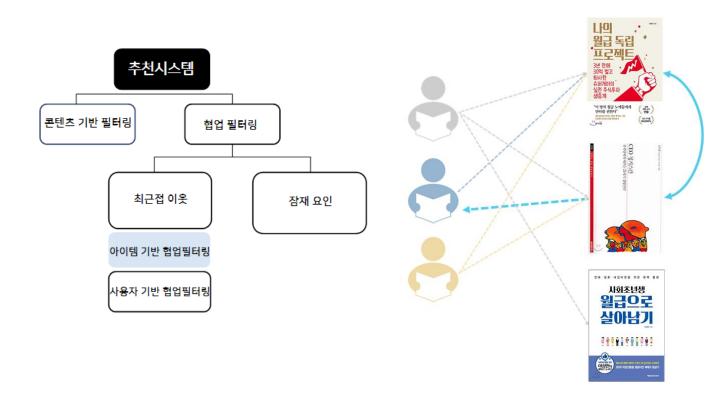
- 대규모의 유저 행동 정보를 분석하여 비슷한 유저 또는 비슷한 아이템을 추천하는 알고리즘
- 사용자 기반 협업 필터링 및 아이템 기반 협업 필터링

협업 필터링만의 장점

• 아이템의 특성에 의존하는 콘텐츠 기반 필터링과는 다르게, 상품에 대한 **사용자의 평점**만으로 구현 가능



협업 필터링/아이템 기반 협업 필터링



아이템 기반 협업 필터링

- 사용자가 선호하는 상품과 비슷한 상품 추천
- 사용자 기반 추천에 비해 정확도가 일반적으로 높음
- * 사용자 기반 협업 필터링: 사용자와 비슷한 사용자들이 선호하는 상품 추천



[표1] 유저- 아이템 행렬

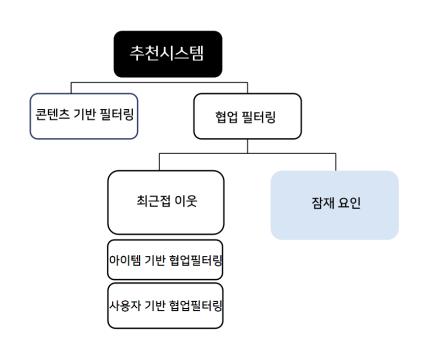
	연주	한나	현진	지현	미리
오발탄		5	1	3	
부의 추월차선	3			5	4
빨간 머리 앤		4	4		

유저-아이템 행렬:

협업 필터링은 유저와 아이템 간의 유사성을 사용하여 추천하는 방식 이를 위해 특정 유저가 여러 콘텐츠에 대해 남긴 피드백 필요 피드백을 토대로 위와 같은 유저-아이템 행렬을 구성하여 분석을 진행해야 함



협업 필터링/ 잠재요인 협업 필터링



	아이템(N)
사용자 (M)	M * N

		잠재요인(K)	
\$\$	사용자 (M)	M * K	*

		아이템(N)
k	잠재 요인 (K)	K*N

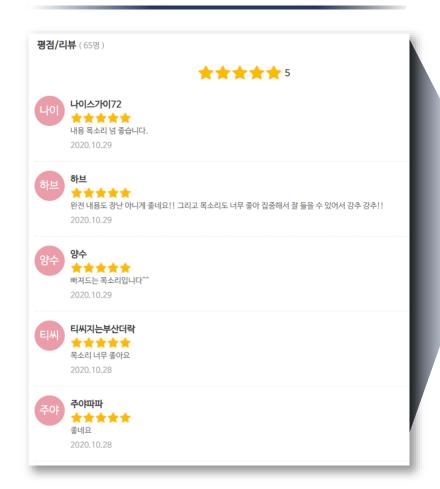
잠재요인 협업 필터링

- 행렬 분해를 기반으로 대규모 행렬을 SVD와 같은 차원 감소 기법으로 분해하는 과정에서 잠재요인을 찾아내는 방식
- 유저-아이템 행렬 데이터를 이용해 '잠재요인'을 찾아냄
- 즉, '유저-잠재요인', '아이템-잠재요인' 행렬로 분해 가능



고객사 리뷰 데이터 수집

팟빵 오디오북의 리뷰 데이터 예시



[표2] 팟빵 오디오북의 리뷰 크롤링 데이터 예시

	Nickname	Rating	Title	Voice	Author	Review	Created	Publish	Genre
0	나이스가이72	10	[단편]헤엄치는 밤	최인영	한유주	내용 목소리 넘 좋습니다	2020.10.29	미디어 창비	소설문학
1	하브	10	[단편]헤엄치는 밤	최인영	한유주	완전 내용도 장난 아니게 좋네요! 그리고 목 소리도 너무 좋아 집중해서 잘 들을 수…	2020.10.29	미디어 창비	소설문학
2	양수	10	[단편]헤엄치는 밤	최인영	한유주	빠져드는 목소리입니다 ^^	2020.10.29	미디어 창비	소설문학
3	티씨지는부산더락	10	[단편]헤엄치는 밤	최인영	한유주	목소리 너무 좋아요	2020.10.29	미디어 창비	소설문학
4	주야파파	10	[단편]헤엄치는 밤	최인영	한유주	좋네요	2020.10.29	미디어 창비	소설문학



고객사 리뷰 데이터 수집

[표3] 팟빵 오디오북의 유저-아이템 행렬 예시

	[단편]헤엄치는 밤	살고 싶다는 농담	오발탄
나이스가이72	10	10	
하브	10		
양수	10		10
티씨지는부산더락	10	10	
주야파파	10		



고객사 리뷰 데이터 수집

[표4] 메타 데이터

nickname	리뷰를 쓴 유저의 별명
rating	평점
title	오디오북의 제목
voice	오디오북의 목소리
author	책의 저자
review	오디오북에 대한 리뷰
created	리뷰를 쓴 날짜
publish	책의 출판사
genre	책의 장르

데이터 세부사항

리뷰데이터의 크기

15,533 건

리뷰데이터에 포함된 유저의 수 리뷰데이터에 포함된 책의 수

리뷰 수집 기간

4,450명

195권

2018.11.08

~2020.11.21



추천 시스템 구현 결과

Surprise Library 사용



명시적 추천 시스템 기능 구현을 위한 라이브러리.

Surprise Library의 머신러닝 프로세스는 아래와 같이 일반 머신러닝 프로세스와 동일하다.





팟빵 추천 시스템 구현 결과

[표5] 팟빵 추천 시스템의 정확도

Algorithm	test_rmse
SVD	1.044225
KNNBasic	1.094358

- * RMSE(평균 제곱근 편차):
- 모델의 예측치와 실제 관찰치의 차이의 제곱의 평균의 제곱근
- * SVD : 잠재 요인 필터링에 사용되는 계산 방식 중 하나
- 장점 : 계산 과정 단축
- 단점: 유저에게 추천 이유 설명할 수 없음

* K Nearest Neighbors:

- 높은 유사도를 가진 K명/개의 이웃을 구성하여 이웃의 평점을 토대로 유저가 경험하지 않은 콘텐츠의 평점을 예측

uid	iid	rui	est	details	lu	Ui	err	
0	youngie	우리 가족은 꽤나 진지합니다	8.0	9.336773	{'was_impossible': False}	3	95	1.336773
1	구니양	살고 싶다는 농담	10.0	9.731043	{'was_impossible': False}	7	130	0.268957
2	maxsy	최민식이 읽는 이범선의 오발탄	10.0	9.366055	{'was_impossible': False}	0	753	0.633945
3	bnb7870	사기를 읽다	10.0	9.434543	{'was_impossible': False}	0	113	0.565457
4	미미노리	어쩌면 내가 가장 듣고 싶었던 말 (요약)	10.0	9.850028	{'was_impossible': False}	2	119	0.149972
3442	헬로우멜로우	살고 싶다는 농담	10.0	9.813733	{'was_impossible': False}	0	130	0.186267
3443	고목나무맹이	담배만이 우리 세상	10.0	9.042141	{'was_impossible': False}	0	262	0.957859
3444	수원블루사나이	빨강머리 앤 : 초록지붕 집 이야기	8.0	8.656722	{'was_impossible': False}	6	174	0.656722
3445	이혼계약	열하일기 첫걸음	10.0	10.00000	{'was_impossible': False}	5	144	0.000000
3446	chichi04	익숙한 길의 왼쪽	8.0	9.013673	{'was_impossible': False}	18	53	1.013673

저희는 이 추천 시스템을 사용하지 않는 것을 추천드립니다.



명시적 평점의 한계

명시적 평점 (explicit rating)

- 사용자가 직접 입력한 평점 (ex: 별점, 좋아요)
- 직관적, 명시적인 만족/불만족 표시
- 특정한 경우에만 별점 입력해 왜곡의 우려
- 사용자가 평점을 입력하지 않는 경우도 많아 데이터가 적으며 유저-아이템 행렬의 희박성 높음

암묵적 평점 (implicit rating)

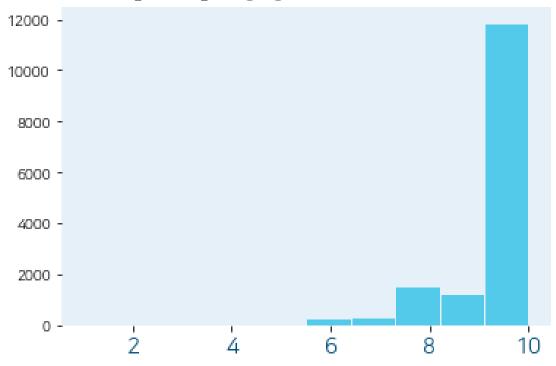
- 사용자의 행동을 바탕으로 추측한 평점 (ex: 클릭, 구매, 재구매 여부 등)
- 다양한 사용자 행동을 기록하고 각각에 따른 가 중치를 부여해 평점으로 변환
- 명시적 평점에 비해 상대적으로 희박성이 덜하다

팟빵의 명시적 평점만으로는 의미있는 추천 시스템을 만들기에 적절하지 않다.



1. 10점에 집중된 평점 분포

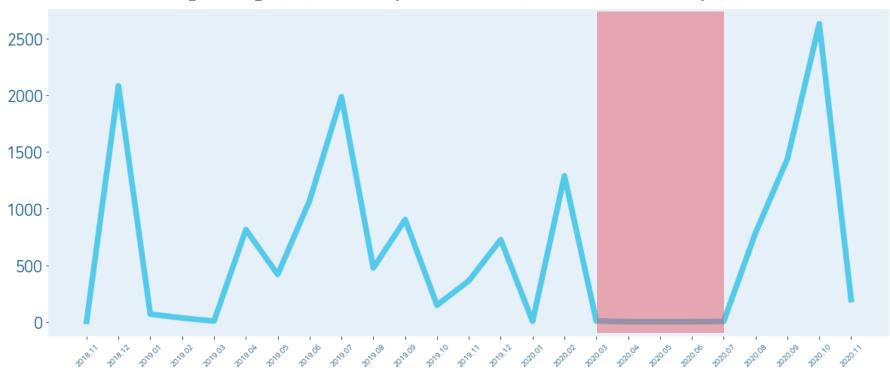
[그림3] 팟빵 평점 히스토그램





2. 특정 기간에만 작성된 리뷰





리뷰 작성 수가 매우 적은 3월~7월, 도대체 무슨 일이?



2. 특정 기간에만 작성된 리뷰

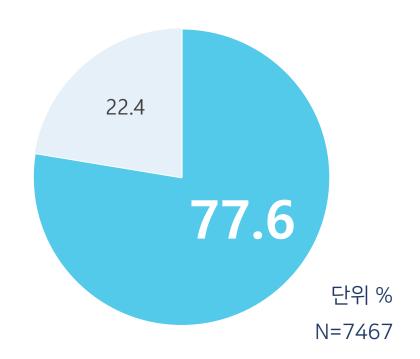
[그림5] 팟빵 오디오북 이벤트 스케쥴





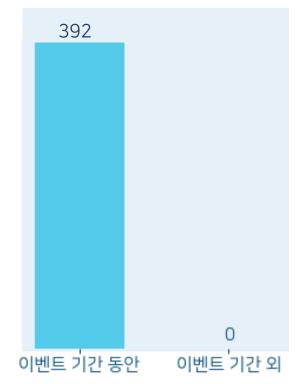
2. 특정 기간에만 작성된 리뷰

[그림6] 이벤트 기간 동안 이벤트 도서에 작성된 댓글 비율



■ 이벤트 도서에 작성된 댓글 ■ 그 외 도서에 작성된 댓글 * 이벤트 기간: 2019년 12월 ~ 2020년 12월

[그림7] 이벤트 도서 <살고 싶다는 농담> 댓글 수



- * 콘텐츠 서비스 시작일 2020.09.16
- * 댓글 작성 이벤트- 2020.10.30 ~ 2020.11.0520

끝…?!



팟빵 리뷰 데이터 구성

팟빵 리뷰 = 평점 + 댓글

팡빵의 리뷰 데이터 명점 댓글

유용한 댓글의 필요성

"제품을 직접 경험해본 사람만 알 수 있는 정보를 제공, 타 사용자의 구매 결정에 도움"

유용한 댓글은 서비스의 사용자 경험 개선에 기여한다



팟빵 및 경쟁사 특징 파악

	⊘ 팟빵	월라	audioclip
런칭	2018.07	2018.02	2018.12
특징	오디오 콘텐츠에 익숙한 팟캐스트 유저를 타겟	오디오북 콘텐츠 특화 플랫폼으로 오디오북과 강연 서비스	네이버 웹툰, 웹소설 등을 원작으로 하는 자체 콘텐츠 제작 및 서비스
과금 형태	개별 콘텐츠 구매/대여	정액제 구독형 (월)	개별 콘텐츠 구매/대여
리뷰 수집 건수 (수집일)	15,533 (2020/11/21)	4,433 (2020/12/01)	11,022 (2020/11/28)



팟빵 및 경쟁사 특징 파악





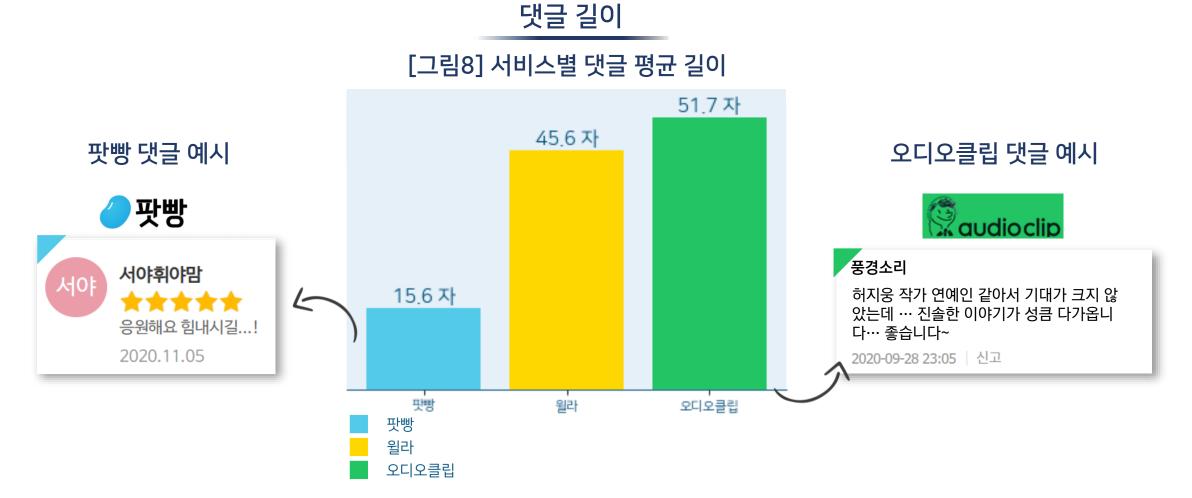






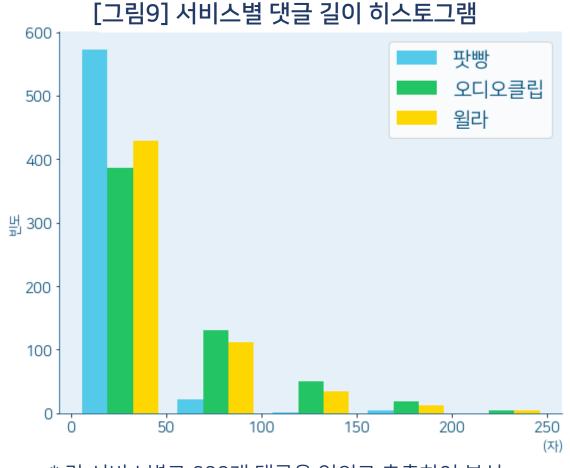








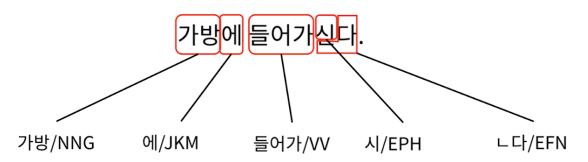
댓글 길이





명사 추출을 통한 키워드 다양성 확인

형태소 분석



문장을 가장 작은 말의 단위인 형태소 단위로 구분하고, 언어적인 구조를 파악하는 것

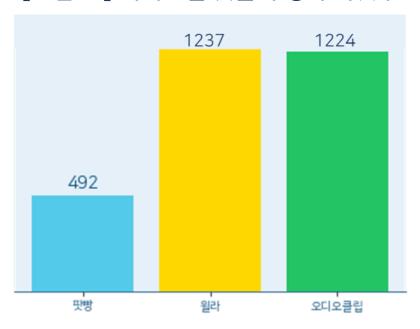
Mecab 형태소 분석기 사용

- 1. 자주 쓰이는 파이썬 한국어 형태소 분석 패키지인 KoNLPy 중의 하나
- 2. 다른 한국어 형태소 분석기보다 속도가 빠르고, 분석 품질이 뛰어남



명사 추출을 통한 키워드 다양성 확인

[그림10] 서비스별 댓글의 명사 가짓수



팟빵 오디오북에서 사용되는 명사 가짓수가 현저히 적음



팟빵 댓글 데이터 형태소 분석

[표6] 형태소별 빈도수 TOP10

팟빵		
형태소	빈도	
좋	228	
들	130	
네요	130	
습니다	129	
잘	118	
아요	99	
이	94	
었	90	
는	75	
목소리	68	

팟	빵	팟	빵
형태소	빈도	형태소	빈도
좋	228	좋	228
들	130	들	130
네요	130	네요	130
습니다	129	습니다	129
잘	118	잘	118
아요	99	아요	99
이	94	O	94
었	90	었	90
는	75	는	75
목소리	68	목소리	68

좋네요 좋아요 잘 들었습니다

^{*} 팟빵 전체 댓글 중 600개를 임의로 추출하여 분석



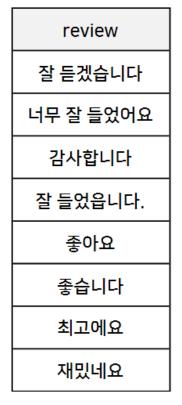
단문 댓글 심화 분석

필터링 조건

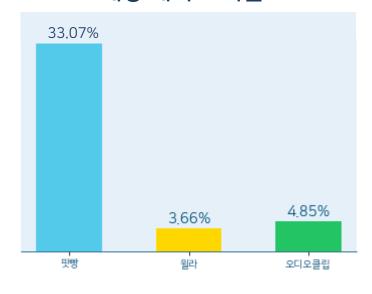
- 1. 댓글 길이가 10자 미만이며,
- 2. 다음과 같은 단어를 포함한 경우

잘	재밌
들었습니다	감사
좋아요	굿
좋습니다	고마
최고	고맙

[표7] 예시



[그림11] 서비스별 전체 댓글 대비 해당 케이스 비율

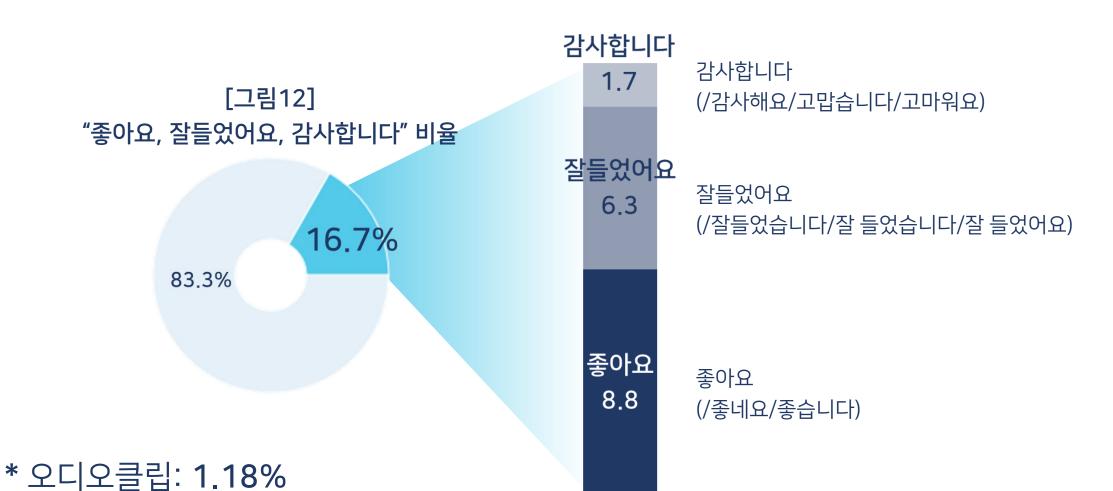


^{*} 각 서비스별로 600개 댓글을 임의로 추출하여 분석

* 윌라: 1.14%



좋아요, 잘들었어요, 감사합니다

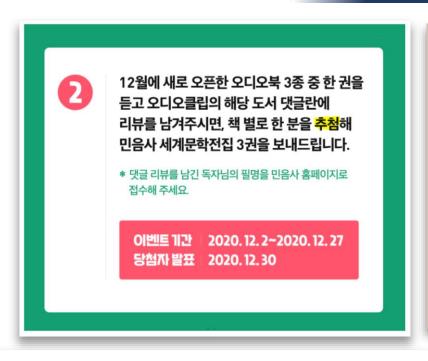


31



상이한 이벤트 방식

네이버 오디오클립 이벤트





김현정

건조한 문장을 툭툭 던지듯 읽어서 제가 갖고 있던 노인의 바다 이미지에 딱 들어맞는 는데, 예전에 읽었던 익숙한 문장,익숙한 격려가 그리울 때가 있더군요. 민음사 홈페이 🔘 😾 🗲 🖂 ≓ 🖒 🗝 간 귀호강 마음 호강을 얻었습니다. 감사합니다.

과보면 글자는 도저히 읽고 싶지 않은데, 위로가 필요한 때 **책장에 좀처럼 손이 가지 않**

2020-05-26 23:55 | 신고

답글





상이한 이벤트 방식

팟빵 오디오북 이벤트



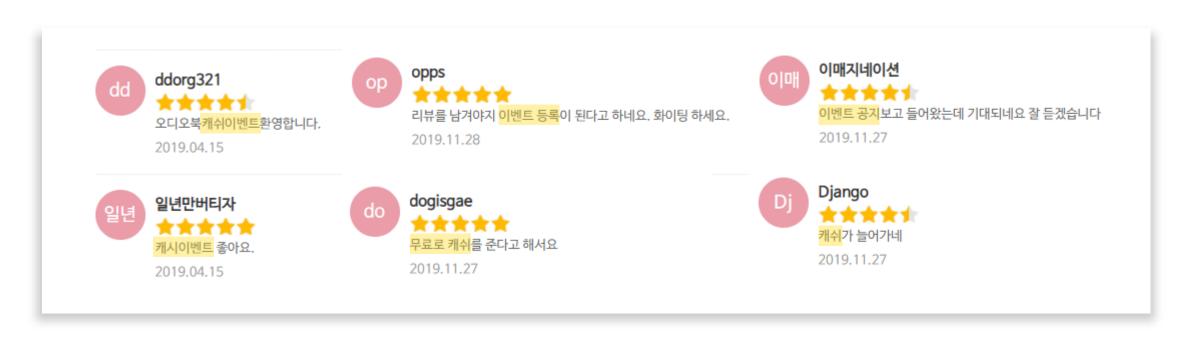






상이한 이벤트 방식

팟빵 오디오북 이벤트 참여 댓글





리뷰 기능의 맹점

듣지 않은 오디오북에 대해서도 누구나 리뷰를 남길 수 있음





AARRR로 보는 팟빵 오디오북 리뷰 이벤트

ACQUISITION

ACTIVATION

RETENTION

REVENUE

REFERRAL

획득

이벤트 기간 동안 많은 수의 유저 유입

활성화

'오디오북 청취' 경험을 하지 않은 상태에서도 리뷰 작성 가능

재방문

이벤트로 유입된 유저는 오디오북 청취를 위해 **재방문할 가능성 낮음**. 타 고객 재방문에도 부정적인 영향 끼칠 위험

매출

오디오북을 구매하지 않고 리뷰 작성 가능해 실제 매출로 이어지지 않을 가능성

추천

유저간 입소문 효과가 활발히 일어나기 어려움



SUMMARY

고객사 리뷰 데이터의 특징

- 1. 10점에 집중된 평점 분포
- 15,533개의 데이터 중 11,878개가 10점
- 2. 특정 기간에만 작성된 리뷰 대다수
- 1년 동안 이벤트 대상 도서에 작성된 댓글이 동기간 전체 댓글의 **77%**에 달함
- 3. 경쟁사 대비 낮은 댓글 퀄리티
- 댓글의 길이가 짧으며, 경쟁사에 비해 댓글 길이의 분포가 좁음
- **"좋아요", "잘들었어요", "감사합니다"**와 같은 유형이 댓글의 **17%**를 구성하며 키워드 다양성 또한 낮음



SUMMARY

PROBLEM

SOLUTION

유저의 실제 선호를 반영하지 못한 평점으로 인해 명시적 평점만을 활용한 추천 시스템 구현의 한계 추천 시스템 도입을 위해

오디오북 사용 로그를 활용한 암묵적 평점 기반 추천

- 1. 댓글을 쓰기만 하면 캐시를 주는 이벤트 방식
- 2. 오디오북을 듣지 않아도 리뷰를 남길 수 있는 서비스 특징



"좋아요, 잘들었어요, 감사합니다"

리뷰 퀄리티 개선을 위해

- <기대평> 게시판과 실구매자의 후기를 위한 <리뷰> 게시판 **분리하여 인센티브 제공**
- 잠재 고객 타깃 이벤트 기획
- 한시적 가격 할인 이벤트 도입

유저 경험을 개선하는 데이터를 쌓기 위해서는, 전사가 함께 노력해야



TEAM

https://github.com/dataitgirls4/team_3

김연주



twinduswn@ewhain.net linkedin/yeonju-kim

Data Scraping Management

Material Design Management

Data Analysis

Project Management

김한나



khnn09@yonsei.ac.kr linkedin/hannakim0924

Tech Lead

Recommendation System Development

Data Analysis

Data Visualization

김현진



hjkim0603@yonsei.ac.kr linkedin/hyunjinkim0603

Project Management

Data Analysis

Natural Language Processing

External Communications

박지현



jihyunn1919@gmail.com linkedin/jihyun-park-4501a61ab

Data Scraping

Data Analysis

Attribute Analysis

최미리



miree0612@gmail.com linkedin/miree-choi

Data Management

Data Analysis

Natural Language Processing

Web Development

데모데이 프로젝트 발표





