

유튜브로 알아보는 나의 독서 DNA

유튜브 시청기록 기반 도서 추천시스템

회귀분석팀

김보근



서유진



하хина



김민주



목차

1. 분석 배경 및 흐름
2. 데이터 수집
3. 클러스터링
4. 이슈추출
5. 유튜브 키워드 추출

01

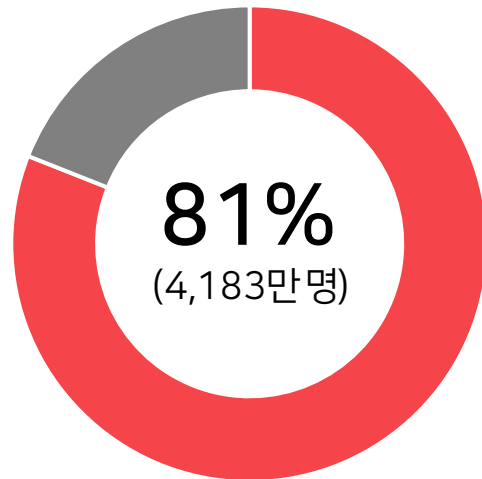
분석 배경 및 흐름

1. 분석 배경 및 흐름

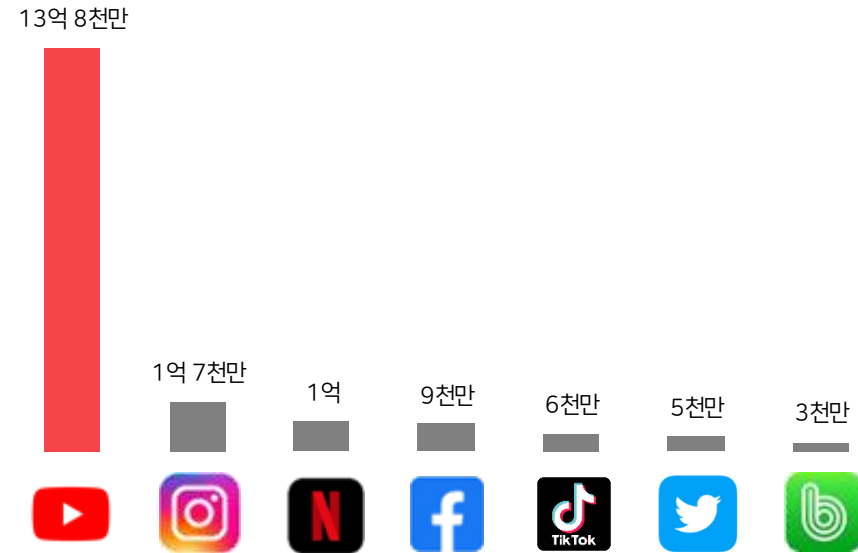
| 주제 선정 배경

한국인 YouTube 사용자 비율

한국인 5,163만 명 중



주요 앱 월간 총 사용시간



출처 : KOSIS 유튜브 사용시간 통계(2022)

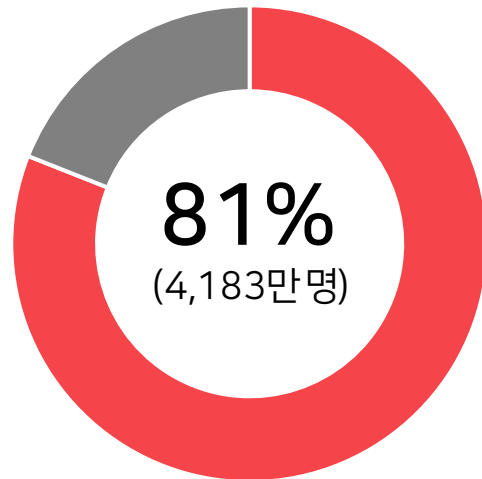
대한민국 전체 인구 중 80% 이상이 YouTube를 사용하고 1인당 월평균 32.9시간 사용하며
다른 어플리케이션과 비교하였을 때도 매우 높은 사용시간을 보임

1. 분석 배경 및 흐름

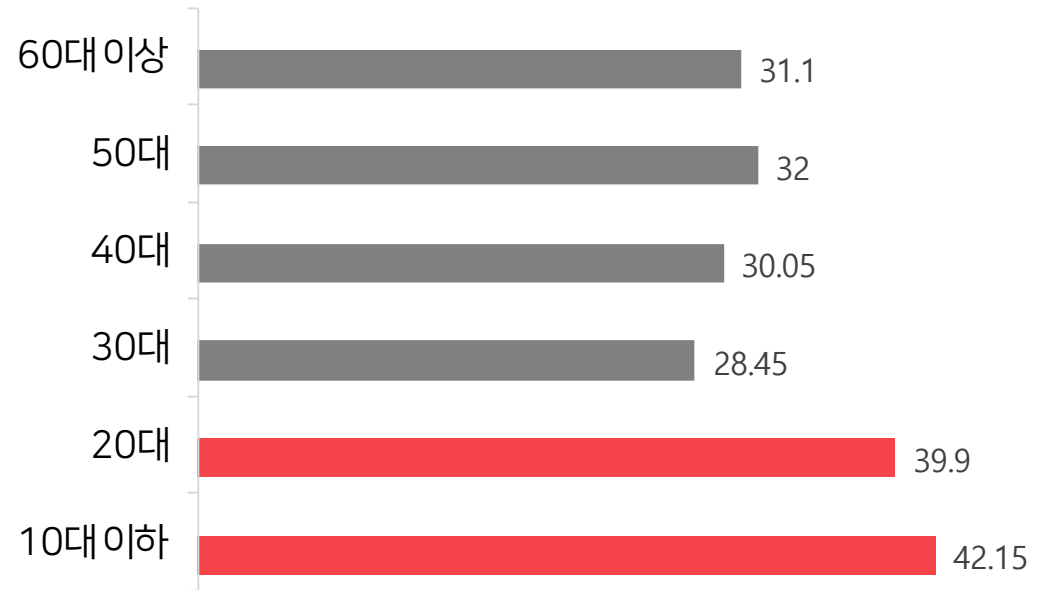
| 주제 선정 배경

한국인 YouTube 사용자 비율

한국인 5,163만 명 중



YouTube 연령별 1인당 월평균 사용시간



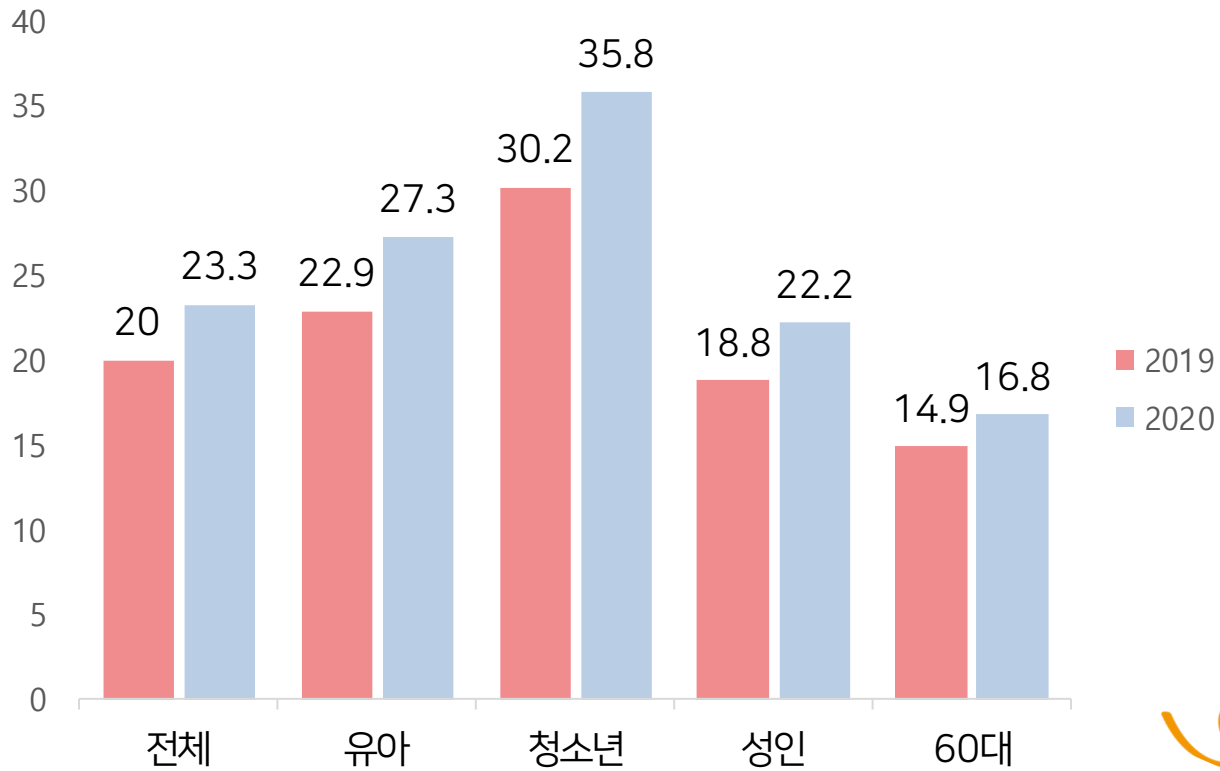
출처 : KOSIS 유튜브 사용시간 통계(2022)

특히 10대 및 20대 연령층에서 월평균 사용량이 높음

1. 분석 배경 및 흐름

| 주제 선정 배경

스마트폰 과의존 위험군 현황



출처 : 과학기술정보통신부, 한국지능정보사회진흥원

스마트폰 과의존

짧고 자극적인 영상 콘텐츠에
지속적으로 노출시키는 유튜브는
“스마트폰 과의존”으로 이어질 수 있다.

스마트폰 과의존 위험군은
매년 전연령에서 증가하는 추세!

1. 분석 배경 및 흐름

주제 선정 배경

스마트폰 과의존 진단 검사 일부

번호	항목	전혀 그렇지 않다	그렇지 않다	그렇다	매우 그렇다
1	스마트폰의 지나친 사용으로 학교성적이나 업무능률이 떨어진다.	1	2	3	4
2	스마트폰을 사용하지 못하면 온 세상을 잃을 것 같은 생각이 든다.	1	2	3	4
3	스마트폰을 사용할 때 그만해야지 라고 생각은 하면서도 계속한다.	1	2	3	4
4	스마트폰이 없어도 불안하지 않다.	1	2	3	4
5	수시로 스마트폰을 사용하다가 지적을 받은 적이 있다.	1	2	3	4
6	가족이나 친구들과 함께 있는 것보다 스마트폰을 사용하고 있는 것이 더 즐겁다.	1	2	3	4
7	스마트폰 사용시간을 줄이려고 해보았지만 실패한다.	1	2	3	4
8	스마트폰을 사용할 수 없게 된다면 견디기 힘들 것이다.	1	2	3	4
9	스마트폰을 너무 자주 또는 오래한다고 가족이나 친구들로부터 불평을 들은 적이 있다.	1	2	3	4
10	스마트폰 사용에 많은 시간을 보내지 않는다.	1	2	3	4
11	스마트폰이 옆에 없으면, 하루 종일 일(또는 공부)이 손에 안잡힌다.	1	2	3	4
12	스마트폰을 사용하느라 지금 하고 있는 일(공부)에 집중이 안 된 적이 있다.	1	2	3	4
13	스마트폰 사용에 많은 시간을 보내는 것이 습관화되었다.	1	2	3	4

이러한 증상에 해당하시나요??
그렇다면 당신은 이미 스마트폰 과의존 위험군!!



1. 분석 배경 및 흐름

| 주제 선정 배경

그렇다면 어떻게 해야 할까...?



최근 신경학자들은 읽기를 담당하는 특별한 부위를 밝혀냈는데

...

또 독서는 스트레스를 없앤다.

...

동일한 조건에서 독서를 하기 전보다

독서를 한 이후에 집중력이 약 10% 상승했고

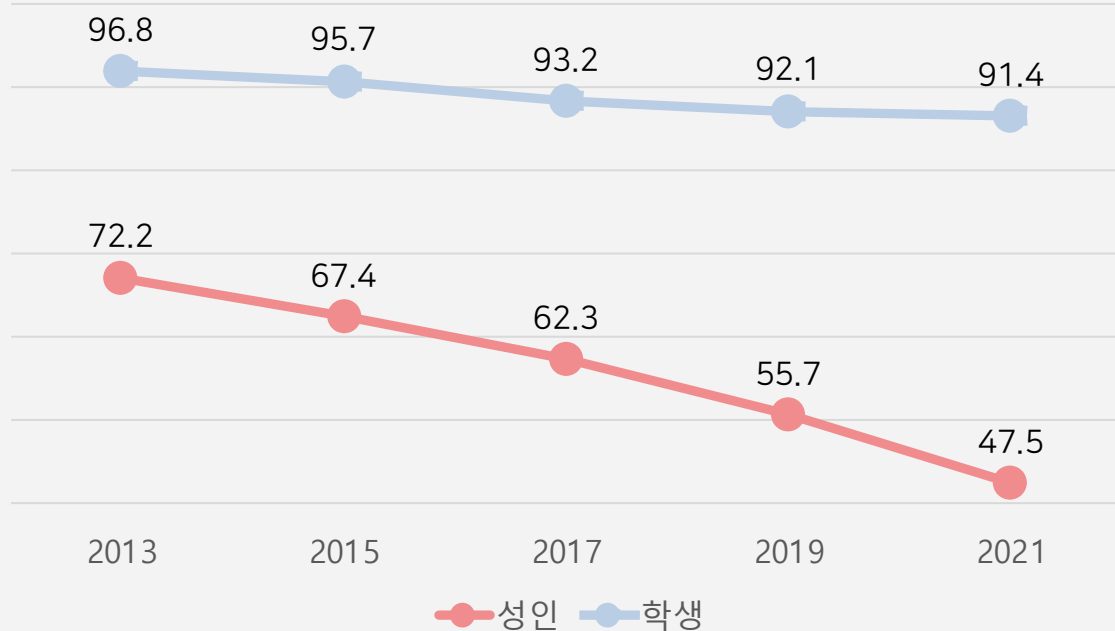
스트레스에 대한 저항도 역시 증가한다는 결과가 나왔다.

즉, 스마트폰 과의존에 대한 증가로 인한 집중력 감소라는 사회적 문제에, **독서**가 하나의 해답이 될 수 있을 것임!

1. 분석 배경 및 흐름

| 주제 선정 배경

독서율 변화 추이



출처 : KOSIS 국민도서 실태조사

모든 연령층 및 성별에 무관하게
독서량 추이가 지속적으로 감소
(종이책, 전자책, 오디오북 전부 포함)

⋮

특히 연간 성인 종합 독서율은 47.5%로
성인 중 절반 이상이 1년에 책을 한 권도 읽지 않음을 의미

1. 분석 배경 및 흐름

| 주제 선정 배경

유튜브가 우리를 중독시키는 원인 중 하나인 알고리즘의 원리에 착안하여 도서를 추천한다면?



기대효과 1 : 접근성

알고리즘의 원리를 바탕으로 관심분야에
부합한 책을 추천해줌으로써
유튜브를 통해 영상을 접하는 것처럼
책을 보다 쉽게 접할 수 있음

기대효과 2 : 독서량 증진

도서 접근성의 증대는
독서량 증진에 영향을 미칠 것이며
개인적, 사회적으로 집중력 증가 및
긍정적인 파급효과를 불러올 것임

유튜브 시청기록 기반 도서 추천 시스템!!

1. 분석 배경 및 흐름

| 주제 선정 배경

유튜브가 우리를 중독시키는 원인 중 하나인 알고리즘의 원리에 착안하여 도서를 추천한다면?



기대효과 1 : 접근성

알고리즘의 원리를 바탕으로 관심분야에
부합한 책을 추천해줌으로써
유튜브를 통해 영상을 접하는 것처럼
책을 보다 쉽게 접할 수 있음

기대효과 2 : 독서량 증진

도서 접근성의 증대는
독서량 증진에 영향을 미칠 것이며
개인적, 사회적으로 집중력 증가 및
긍정적인 파급효과를 불러올 것임

1. 분석 배경 및 흐름

| Cold Start!?

Cold Start

새로운 유저에 대해 충분한 정보가 수집된 상태가 아니기 때문에
추천 시스템이 적절한 추천을 제공하지 못하는 현상



교차 도메인으로 해결 가능!

도서 추천을 진행할 때 알지 못하는 유저의 도서 선호를
유튜브 시청 기록을 통해 예측할 수 있을 것임

1. 분석 배경 및 흐름

| 도서추천 알고리즘 흐름



1. 데이터 수집

사회 이슈, 대출 정보, 도서 정보 데이터 수집

2. 사용자별 선호도 분포 형성

유튜브 최근 시청 영상 분석 → 영상 키워드 추출 및 분야별 선호도 분포 형성

3. 최종 분포 형성

사회 이슈, 고객 클러스터 등 추가데이터 활용 → 분포 조정, 최종 분포 형성

4. 도서 추천

최종 분포 & 유튜브 시청기록 키워드 기반 도서 추천 진행

1. 분석 배경 및 흐름

| 도서추천 알고리즘 흐름



1. 데이터 수집

사회 이슈, 대출 정보, 도서 정보 데이터 수집

2. 사용자별 선호도 분포 형성

유튜브 최근 시청 영상 분석 → 영상 키워드 추출 및 분야별 선호도 분포 형성

3. 최종 분포 형성

사회 이슈, 고객 클러스터 등 추가데이터 활용 → 분포 조정, 최종 분포 형성

4. 도서 추천

최종 분포 & 유튜브 시청기록 키워드 기반 도서 추천 진행

1. 분석 배경 및 흐름

| 도서추천 알고리즘 흐름



1. 데이터 수집

사회 이슈, 대출 정보, 도서 정보 데이터 수집

2. 사용자별 선호도 분포 형성

유튜브 최근 시청 영상 분석 → 영상 키워드 추출 및 분야별 선호도 분포 형성

3. 최종 분포 형성

사회 이슈, 고객 클러스터 등 추가데이터 활용 → 분포 조정, 최종 분포 형성

4. 도서 추천

최종 분포 & 유튜브 시청기록 키워드 기반 도서 추천 진행

1. 분석 배경 및 흐름

| 도서추천 알고리즘 흐름



1. 데이터 수집

사회 이슈, 대출 정보, 도서 정보 데이터 수집

2. 사용자별 선호도 분포 형성

유튜브 최근 시청 영상 분석 → 영상 키워드 추출 및 분야별 선호도 분포 형성

3. 최종 분포 형성

사회 이슈, 고객 클러스터 등 추가 데이터 활용 → 분포 조정, 최종 분포 형성

4. 도서 추천

최종 분포 & 유튜브 시청기록 키워드 기반 도서 추천 진행

1. 분석 배경 및 흐름

배경지식 | 한국십진분류법 (KDC)

한국십진분류법 (KDC)

도서의 **모든 주제**를 10개(000~900)로 나눈 한국의 장서 분류법

총류, 철학, 종교, 사회과학, 자연과학, 기술과학, 예술, 언어, 문학, 역사

000 총류	100 철학	200 종교	300 사회과학	400 자연과학
010 도서학, 서지학	110 형이상학	210 비교종교	310 통계학	410 수 학
020 문헌정보학	120 인식론,인과론,인간학	220 불 교	320 경제학	420 물 리 학
030 백과사전	130 철학의 체계	230 기 독 교	330 사회학,사회문제	430 화 학
040 강연집,수필집,연설문집	140 경 학	240 도 교	340 정 치 학	440 천 문 학
050 일반연속간행물	150 동양철학,사상	250 천 도 교	350 행 정 학	450 지 학
060 일반학회,단체,협회,기관	160 서양철학	260 신 도	360 법 학	460 광 물 학
070 신문,연론,저널리즘	170 논 리 학	270 힌두교,브라마교	370 교 육 학	470 생명과학
080 일반전집,총서	180 심 리 학	280 미술참고(회교)	380 풍속,예절,민속학	480 식 물 학
090 항도자료	190 윤리학,도덕철학	290 기타 제종교	390 국방,군사학	490 동 물 학
500 기술과학	600 예술	700 언어	800 문학	900 역사
510 의 학	610 건 축 물	710 한 국 어	810 한국문학	910 아 시 아
520 농업,농학	620 조각,조형예술	720 중 국 어	820 중국문학	920 유 럽
530 공학,공업일반,포목공학,환경공학	630 공예,장식미술	730 일본어,기타아시아제어	830 일본문학,기타아시아문학	930 아프리카
540 건축공학	640 서 예	740 영 어	840 영미문학	940 북아메리카
550 기계공학	650 회화,도화	750 독 말 어	850 독일문학	950 남아메리카
560 전기공학,전자공학	660 사진예술	760 프랑스어	860 프랑스문학	960 오세아니아
570 화학공학	670 음 악	770 스페인어,포르투갈어	870 스페인,포르투갈문학	970 양국지방
580 제조업	680 공연예술,매체예술	780 이탈리아어	880 이탈리아문학	980 지 리
590 생활과학	690 오락,스포츠	790 기타제어	890 기타제문학	990 전 기

3__ : 대분류 ex. 사회과학

31_ : 중분류 ex. 사회과학 - 통계학

319 : 소분류 ex. 사회과학 - 통계학 - 인구통계

⋮

계층적 배열구조이기 때문에,
분류기호만으로도 **상하위** 개념을 알 수 있음!

1. 분석 배경 및 흐름

배경지식 | 한국십진분류법 (KDC)

한국십진분류법 (KDC)

도서의 **모든 주제**를 10개(000~900)로 나눈 한국의 장서 분류법

총류, 철학, 종교, 사회과학, 자연과학, 기술과학, 예술, 언어, 문학, 역사

000 총류	100 철학	200 종교	300 사회과학	400 자연과학
010 도서학, 서지학	110 형이상학	210 비교종교	310 통계학	410 수 학
020 문헌정보학	120 인식론, 인과론, 인간학	220 불 교	320 경 제 학	420 물 리 학
030 백과사전	130 철학의 체계	230 기 독 교	330 사회학, 사회문제	430 화 학
040 강연집, 수필집, 연설문집	140 경 학	240 도 교	340 정 치 학	440 천 문 학
050 일반연속간행물	150 동양철학, 사상	250 천 도 교	350 행 정 학	450 지 학
060 일반학회, 단체, 협회, 기관	160 서양철학	260 신 도	360 법 학	460 광 물 학
070 신문, 언론, 저널리즘	170 논 리 학	270 힌두교, 브라만교	370 교 육 학	470 생명과학
080 일반전집, 총서	180 심 리 학	280 이슬람교(회교)	380 풍속, 예절, 민속학	480 식 물 학
090 향토자료	190 윤리학, 도덕철학	290 기타 제종교	390 국방, 군사학	490 동 물 학
500 기술과학	600 예술	700 언어	800 문학	900 역사
510 의 학	610 건 축 물	710 한 국 어	810 한국문학	910 아 시 아
520 농업, 농학	620 조각, 조형예술	720 중 국 어	820 중국문학	920 유 럽
530 공학, 공업일반, 토목공학, 환경공학	630 공예, 장식미술	730 일본어, 기타아시아어	830 일본문학, 기타아시아문학	930 아프리카
540 건축공학	640 서 예	740 영 어	840 영미문학	940 북아메리카
550 기계공학	650 회화, 도화	750 독 일 어	850 독일문학	950 남아메리카
560 전기공학, 전자공학	660 사진예술	760 프랑수어	860 프랑스문학	960 오세아니아
570 화학공학	670 음 악	770 스페인어, 포르투갈어	870 스페인, 포르투갈문학	970 양국지방
580 제 조 업	680 공연예술, 매체예술	780 이탈리아어	880 이탈리아문학	980 지 리
590 생활과학	690 오락, 스포츠	790 기타제어	890 기타제문학	990 전 기

3__ : 대분류 ex. 사회과학

31_ : 중분류 ex. 사회과학 - 통계학

319 : 소분류 ex. 사회과학 - 통계학 - 인구통계

...

계층적 배열구조이기 때문에,
분류기호만으로도 **상하위** 개념을 알 수 있음!

1. 분석 배경 및 흐름

배경지식 | 한국십진분류법 (KDC)

한국십진분류법 (KDC)

도서의 **모든 주제**를 10개(000~900)로 나눈 한국의 장서 분류법

총류, 철학, 종교, 사회과학, 자연과학, 기술과학, 예술, 언어, 문학, 역사

000 총류 010 도서학, 서지학 020 문헌정보학 030 백과사전 040 강연집, 수필집, 연설문집 050 일반연속간행물 060 일반학회, 단체, 협회, 기관 070 신문, 언론, 저널리즘 080 일반전집, 총서 090 향토자료	100 철학 110 형이상학 120 인식론, 인과론, 인간학 130 철학의 체계 140 경 학 150 동양철학, 사상 160 서양철학 170 논 리 학 180 심 리 학 190 윤리학, 도덕철학	200 종교 210 비교종교 220 불 교 230 기 독 교 240 도 교 250 천 도 교 260 신 도 270 힌두교, 브라만교 280 이슬람교(회교) 290 기타 제종교	300 사회과학 310 통 계 학 320 경 제 학 330 사회학, 사회문제 340 정 치 학 350 행 정 학 360 법 학 370 교 육 학 380 풍속, 예절, 민속학 390 국방, 군사학	400 자연과학 410 수 학 420 물 리 학 430 화 학 440 천 문 학 450 지 학 460 광 물 학 470 생명과학 480 식 물 학 490 동 물 학
500 기술과학 510 의 학 520 농업, 농학 530 공학, 공업일반, 토목공학, 환경공학 540 건축공학 550 기계공학 560 전기공학, 전자공학 570 화학공학 580 제 조 업 590 생활과학	600 예술 610 건 축 물 620 조각, 조형예술 630 공예, 장식미술 640 서 예 650 회화, 도화 660 사진예술 670 음 악 680 공연예술, 매체예술 690 오락, 스포츠	700 언어 710 한 국 어 720 중 국 어 730 일본어, 기타아시아어 740 영 어 750 독 일 어 760 프 랑 스 어 770 스페인어, 포르투갈어 780 이탈리아어 790 기타 제어	800 문학 810 한국문학 820 중국문학 830 일본문학, 기타아시아문학 840 영미문학 850 독일문학 860 프랑스문학 870 스페인, 포르투갈문학 880 이탈리아문학 890 기타제문학	900 역사 910 아 시 아 920 유 럽 930 아프리카 940 북아메리카 950 남아메리카 960 오세아니아 970 양극지방 980 지 리 990 전 기

3__ : 대분류 ex. 사회과학

31_ : 중분류 ex. 사회과학 - 통계학

319 : 소분류 ex. 사회과학 - 통계학 - 인구통계

⋮

계층적 배열구조이기 때문에,
분류기호만으로도 **상하위** 개념을 알 수 있음!

1. 분석 배경 및 흐름

배경지식 | 국제 표준 도서 번호 (ISBN)

국제 표준 도서 번호 (ISBN)

개별 도서에 국제적으로 표준화하여 붙이는 **고유 도서번호**



각각의 도서는 하나의 고유값을 가짐

...



통계학원론
도서 25,740원

책 정보

카테고리

수학

ISBN

9791130301686

책을 식별하는 데에 활용 가능

+

크롤링 진행 시, 개별적인 접근 가능



1. 분석 배경 및 흐름

배경지식 | 국제 표준 도서 번호 (ISBN)

국제 표준 도서 번호 (ISBN)

개별 도서에 국제적으로 표준화하여 붙이는 **고유 도서번호**



각각의 도서는 하나의 고유값을 가짐

■
■
■
■
■



통계학원론
도서 25,740원

책 정보

카테고리

수학

ISBN

9791130301686

책을 식별하는 데에 활용 가능

+

크롤링 진행 시, 개별적인 접근 가능



02

데이터 수집

2. 데이터 수집

| 데이터 소개

책

파일이름	출처
인기대출	문화 빅데이터 플랫폼 (국립중앙도서관)
서울도서관 소장자료 현황정보	서울 열린데이터광장
책 소개, 제목, 분류기호 크롤링	YES24

유튜브

파일이름	출처
제목, 해시태그, 자막 크롤링	유튜브

사회 이슈

파일이름	출처
분야 별 뉴스 기사 (정치, 경제, 사회, 문화 외 4개)	빅카인즈

2. 데이터 수집

| 국립중앙도서관 성별 - 연령대별 인기 대출 도서 정보

책 제목, 저자, 출판사, 책 소개, KDC명, 연령, 성별, 분석 기간, 지역, 대출 수 등의 정보를 포함

...

책 제목	저자	KDC명	...	분석 기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	요비노리 다쿠미 지음 ;이지호 옮김	414	...	90일	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은이: 김상욱	420.13	...	90일	20대	남성	54
화폐전쟁	쑹홍빙 지음 ;차혜정 옮김	327.209	...	7일	20대	남성	3
	
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6	...	30일	40대	여성	103

2. 데이터 수집

Yes24 크롤링

ISBN

yes24

빠른분야찾기 베스트 신상품 0

yes24.com/product/search?query=9788958612131

9788958612131에 대한 검색결과

통합검색 (2) 중고매장 (1) 리뷰 (0)

분야
국내도서 (1)
중고샵 (1)

결과 내 재검색
검색어 입력

검색 조건
☐ 도서명
☐ 저자/역자
☐ 출판사

혜택
☐ 이벤트
☐ 쿠폰
☐ 사은품

상품 (1)

인기도순 정확도순 신상품순 최저가순 최고가순 평점순 리뷰순

[도서] 새빨간 거짓말, 통계 [개정판]

대렬 허프 저/박영훈 역 | 청년정신 | 2022년 01월

12,600원 (10% 할인) 700원

판매지수 8,046 | 회원리뷰(1건) ★★★★★ 7.0

19시까지 주문하면 내일(11/7, 화) 도착예정

관련상품 : 중고상품 11개

미리보기

책의 고유값인 ISBN을 이용하여
링크 접근 및 크롤링 가능



저자, 제목, 책 소개 등
약 200,000개 책의 정보 수집

크롤링 코드 짜는

팀장님의 간절한 손..



2. 데이터 수집

유튜브 크롤링

해시태그



제목

유튜브 링크를 이용하여
개별 영상에 접근 후 크롤링 진행



개별 영상의 제목, 해시태그, 자막
정보 수집

03

클러스터링

3. 클러스터링

| 국립중앙도서관 성별 - 연령대별 인기 대출 도서 정보

책 제목, 저자, 출판사, 책 소개, KDC명, 연령, 성별, 분석 기간, 지역, 대출 수 등의 정보를 포함

...

책 제목	저자	KDC명	...	분석 기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	요비노리 다쿠미 지음 ;이지호 옮김	414	...	90일	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은이: 김상욱	420.13	...	90일	20대	남성	54
화폐전쟁	쑹홍빙 지음 ;차혜정 옮김	327.209	...	7일	20대	남성	3
	
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6	...	30일	40대	여성	103

3. 클러스터링

| 국립중앙도서관 성별 - 연령대별 인기 대출 도서 정보

책 제목, 저자, 출판사, 책 소개, KDC명, 연령, 성별, 분석 기간, 지역, 대출 수 등의 정보를 포함

책 제목	저자	KDC명	분석 기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	오바오리 다쿠미 지음 ;이지호 옮김	414 ...	9월	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은이: 김상욱	...	9월	20대	남성	54
화폐전쟁	쑹홍빙 지음 ;차혜정 옮김	327.209 ...	7월	20대		3
	
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6 ...	30일	40대		103

'분야 선호 클러스터'를
만들어보자!



3. 클러스터링

국립중앙도서관 성별  **연령 및 성별 데이터를 그대로 사용하지 않고**
새로운 클러스터를 형성한 이유

책 제목, 저자, 출판사, 책 소개, KDC명, 연령, 성별, 분석 기간, 지역, 대출 수 등의 정보를 포함

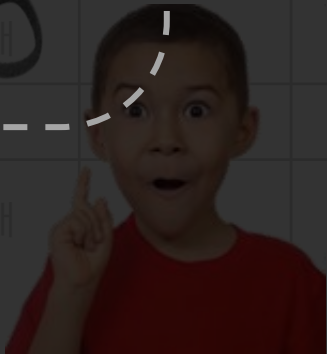
단순 **연령, 성별**로 사용자를 **구분**한다면

선호 분야 예측에 있어 오류 발생 가능성 존재

'분야 선호 클러스터'를

ex) 20대 남자라고 마냥 축구나 게임을 좋아하는 건 아님!

책 제목	저자	출판사	책 소개	KDC명	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	이진호	웅진	미적분	418.1	청소년 (14~19)	남성	34
김상욱의 양자 공부	김상욱	웅진	양자물리	327.209	20대	남성	54
완전히 새로운 현대 물리학 입문	송홍빈	지음	물리학	327.209	20대	남성	3
화폐전쟁	지은아	히가시노 게	소설	833.6	40대	남성	103
가면산장 살인사건	이교	웅진	소설	833.6	40대	남성	103



3. 클러스터링

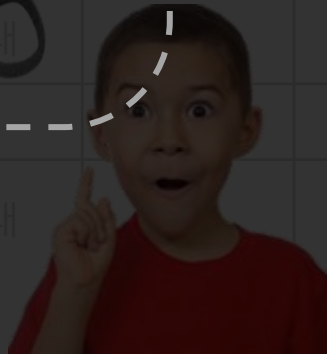
국립중앙도서관 성별  연령 및 성별 데이터를 그대로 사용하지 않고
새로운 클러스터를 형성한 이유

책 제목, 저자, 출판사, 책 소개, KDC명, 연령, 성별, 분석 기간, 지역, 대출 수 등의 정보를 포함

새로운 '분야 선호 클러스터'를 통해 분류

책 제목	저자	KDC명	분석기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	와타나베 다쿠미 지음 ;이지호 옮김	414 ...	9월	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은아	20대	남성	54
화폐전쟁	송홍병 지음 ;이재정 옮김	20대	...	3
가면산장 살인사건	지은아: 히가시노 게 이고 ; 옮긴이: 김난주	833.6 ...	30일	40대	...	103

'분야 선호 클러스터'를
사용자와 비슷한 성향의 사람들의
선호를 반영할 수 있도록 함



3. 클러스터링

| '분야 선호 클러스터' 활용 방안

1. 특정 연령대, 성별이 고려되지 않는 새로운 '분야 선호 클러스터' 형성

⋮

ex.

인문학분야가 두드러지는
철학자형

과학분야가 두드러지는
과학소년형

사회분야가 두드러지는
탐구형



3. 클러스터링

| '분야 선호 클러스터' 활용 방안

2. 유튜브 시청기록 영상의 키워드를 추출하여 사용자의 선호도 분포 형성

3. 각각의 '분야 선호 클러스터'와 유사도 계산 후 가장 가까운 클러스터의 분포와 가중합 해줌

⋮



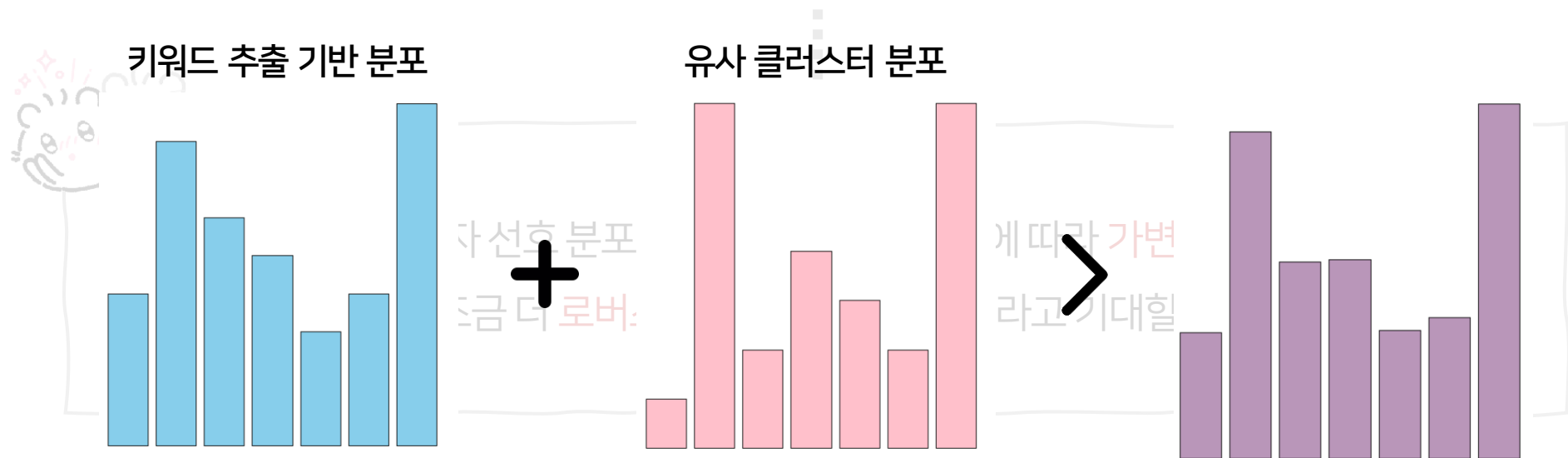
유튜브로 얻은 사용자 선호 분포는 최근 시청 기록 영상에 따라 **가변적**일 것이기 때문에,
분포를 조금 더 **로버스트**하게 만들어줄 것이라고 기대할 수 있음!

3. 클러스터링

| '분야 선호 클러스터' 활용 방안

2. 유튜브 시청기록 영상의 키워드를 추출하여 사용자의 선호도 분포 형성

3. 각각의 '분야 선호 클러스터'와 유사도 계산 후 가장 가까운 클러스터의 분포와 가중합 해줌



3. 클러스터링

| '분야 선호 클러스터' 활용 방안

2. 유튜브 시청기록 영상의 키워드를 추출하여 사용자의 선호도 분포 형성

3. 각각의 '분야 선호 클러스터'와 유사도 계산 후 가장 가까운 클러스터의 분포와 가중합 해줌

⋮




유튜브로 얻은 사용자 선호 분포는 최근 시청 기록 영상에 따라 **가변적**일 것이기 때문에,
분포를 조금 더 **로버스트**하게 만들어줄 것이라고 기대할 수 있음!

3. 클러스터링

| 데이터 전처리

책 제목	저자	KDC명	...	분석 기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	요비노리 다쿠미 지음 ;이지호 옮김	414	...	90일	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은이: 김상욱	420.13	...	90일	20대	남성	54
화폐전쟁	쑹홍빙 지음 ;차혜정 옮김	327.209	...	7일	20대	남성	3
...
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6	...	30일	40대	여성	103

 '총류', '철학', ..., '역사' 로 값 변경

3. 클러스터링

| 데이터 전처리

책 제목	저자	KDC명	...	분석 기간	연령대	성별	대출 수
(수학은 어렵지만) 미적분은 알고 싶어	요비노리 다쿠미 지음 ;이지호 옮김	414	...	90일	청소년 (14~19)	남성	34
김상욱의 양자 공부 :완전히 새로운 현대 물리학 입문	지은이: 김상욱	420.13	...	90일	20대	남성	54
화폐전쟁	쑹홍빙 지음 ;차혜정 옮김	327.209	...	7일	20대	남성	3
...
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6	...	30일	40대	여성	103



중복 집계를 제외하기 위해
'90일'인 경우만 사용

3. 클러스터링

데이터 전처리

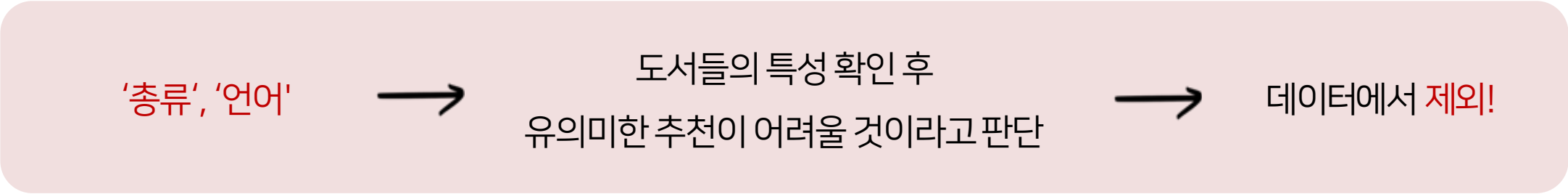
BOOK_TITLE_NM	AUTHR_NM	KDC_NM
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
(읽으면서 바로 써먹는) 어린이 사자성어	글·그림: 한날	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
신비아파트 한자 귀신 2 - 저주의 대가	김강현 (지은이), 김기수 (그림), 김경익, 박상우 (감수)	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어
마법천자문:손오공의 한자 대탐험	저자: 스튜디오시리얼,홍거북	언어

살려주세요

공공 이지 않는 마법천자문 이슈 ...

책 제목	저자	KDC명	...
21세기 희망의 경기포럼 : 강연집 / 경기도	최렬, 윤방부, 유상옥 외	총류	...
노션 =업무와 일상을 정리하는 새로운 방법 /Notion	전시진,이해봄 지음	총류	...
...
(읽으면서 바로 써먹는) 어린이 영단어	글·그림: 한날	언어	...

...



[총류 - 기타 분류 및 비도서 위주, 언어 - 토익 등 언어시험 위주]

3. 클러스터링

| 데이터 전처리

책 제목	저자	KDC명	...	분석 기간	연령대	성별	대출 수
(추리 천재) 엉덩이 탐정	트롤 글·그림 ;김정 화 옮김	833.6	...	90일	영유아(0~5)	여성	78
당근 유치원	지은아: 안녕달	813.70	...	90일	영유아(0~5)	여성	238
무지개 물고기	마르쿠스 피스터 글. 그림;공경희 옮김	853.00	...	90일	유아(6~7)	남성	58
...

⋮

'영유아', '유아'



유튜브를 잘 시청하지 않는 연령대이고,
대출 도서 특성 확인 후
추천 모델의 대상과 적합하지 않다고 판단



데이터에서 제외!

3. 클러스터링

| 최종 데이터셋

연령대, 성별, KDC명, 대출 수의 총합

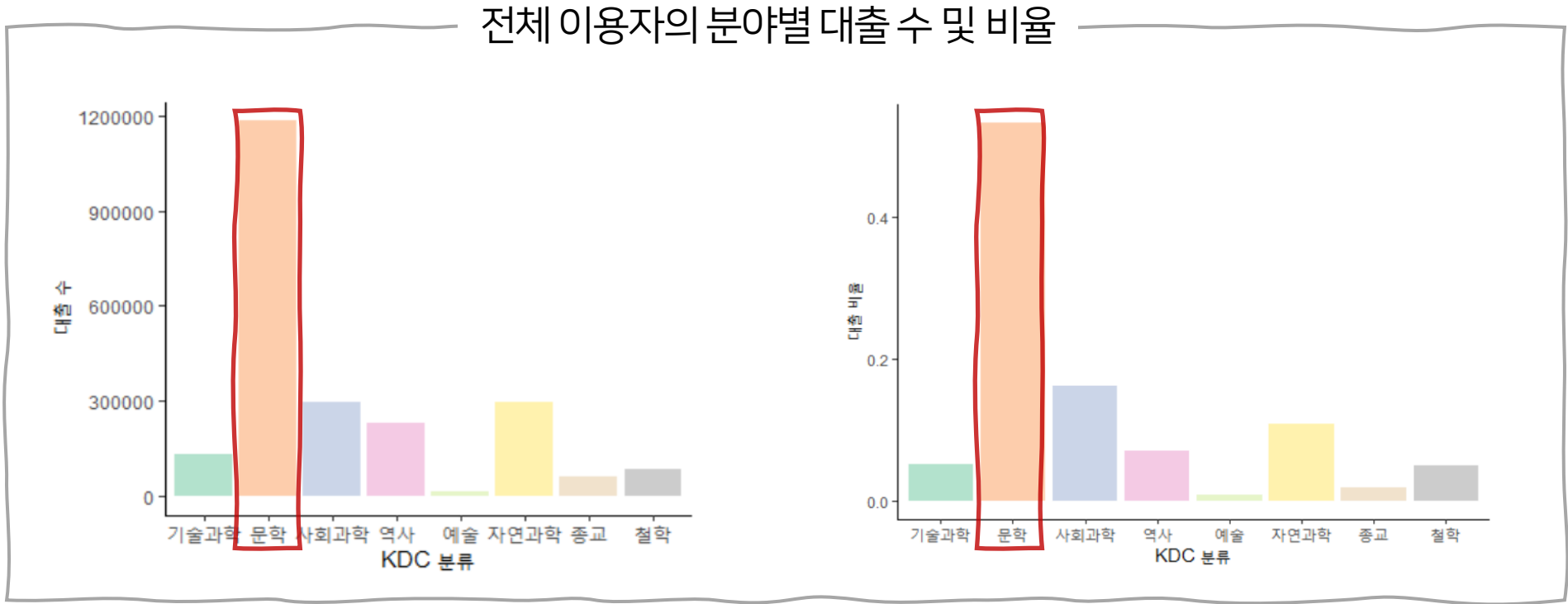
⋮

→ 대출 데이터가 없는 경우는 0으로 대체

연령대	성별	KDC명	총 대출 수
초등(8~13)	남성	철학	1747
초등(8~13)	남성	종교	7851
...
20대	남성	기술과학	1692
...
60대 이상	여성	역사	1250

3. 클러스터링

EDA | 분야별 대출 비율



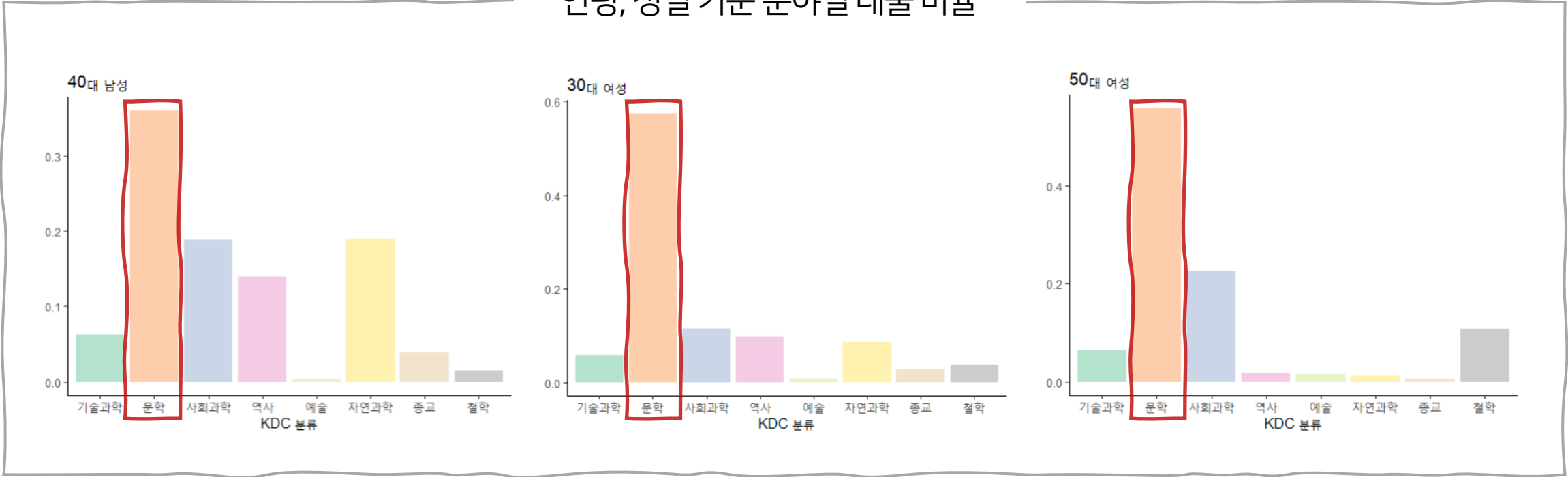
...

문학 대출 비율이 다른 분류보다 월등히 높음

3. 클러스터링

EDA | 분야별 대출 비율

연령, 성별 기준 분야별 대출 비율



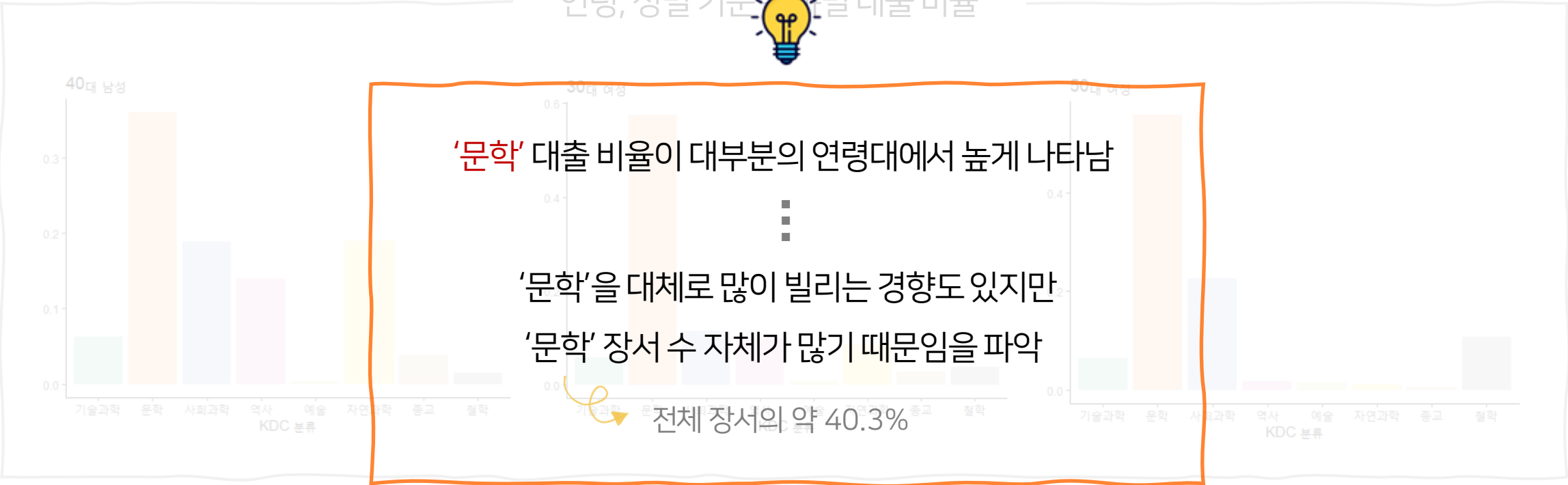
⋮

마찬가지로 문학대출 비율이 다른 분류보다 높음

3. 클러스터링

EDA | 분야별 대출 비율

연령, 성별 기준  분야별 대출 비율

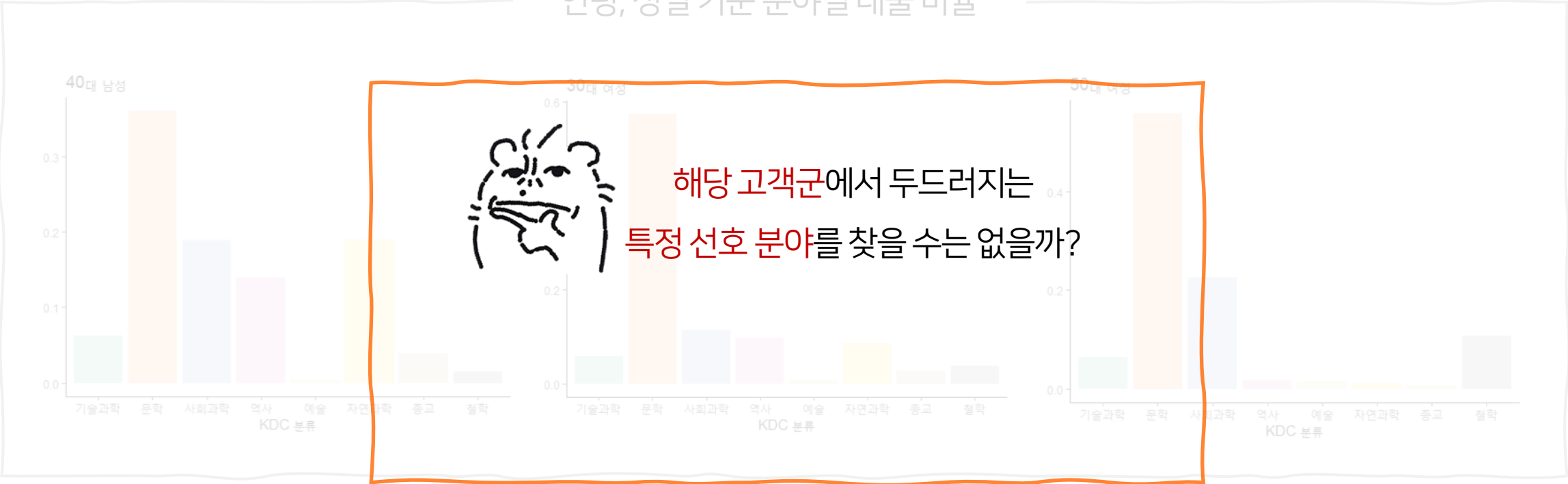


마찬가지로 문학대출 비율이 다른 분류보다 높음

3. 클러스터링

EDA | 분야별 대출 비율

연령, 성별 기준 분야별 대출 비율

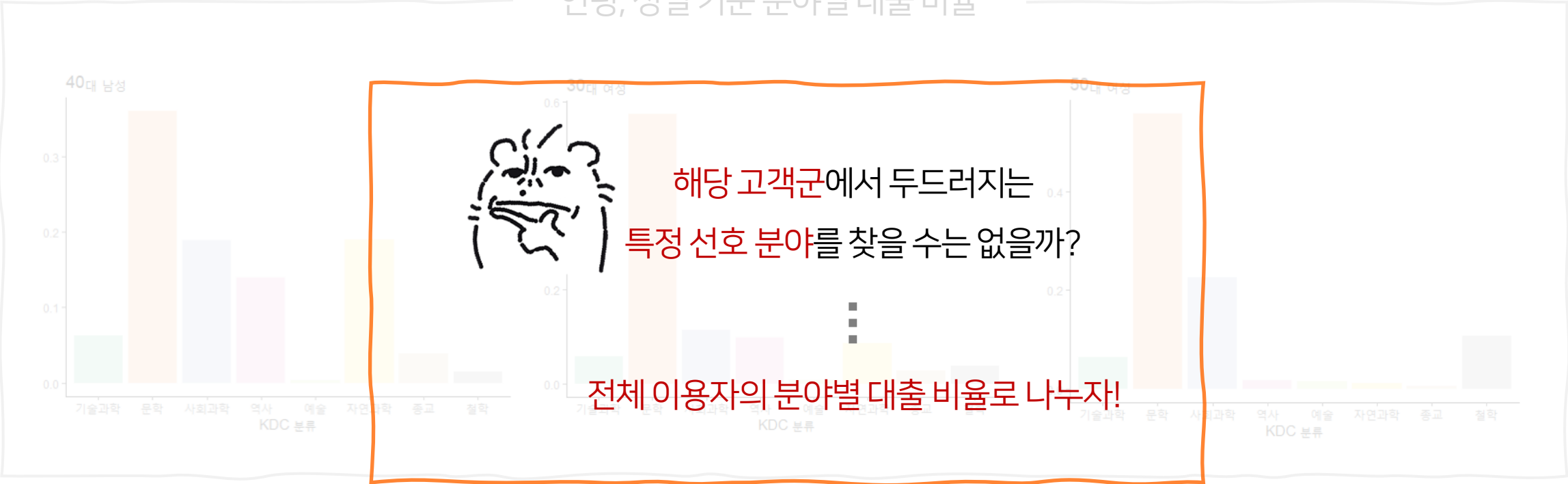


마찬가지로 문학대출 비율이 다른 분류보다 높음

3. 클러스터링

EDA | 분야별 대출 비율

연령, 성별 기준 분야별 대출 비율

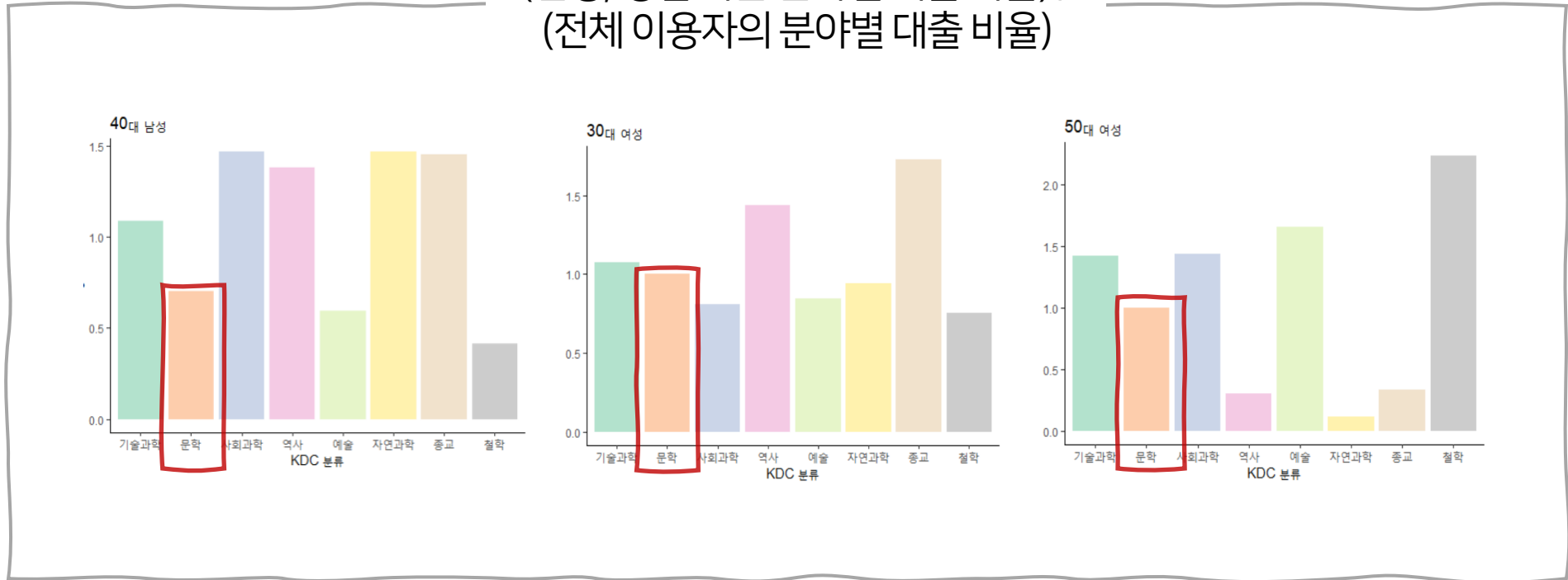


마찬가지로 문학대출 비율이 다른 분류보다 높음

3. 클러스터링

EDA | 분야별 대출 비율

(연령, 성별 기준 분야별 대출 비율) /
(전체 이용자의 분야별 대출 비율)



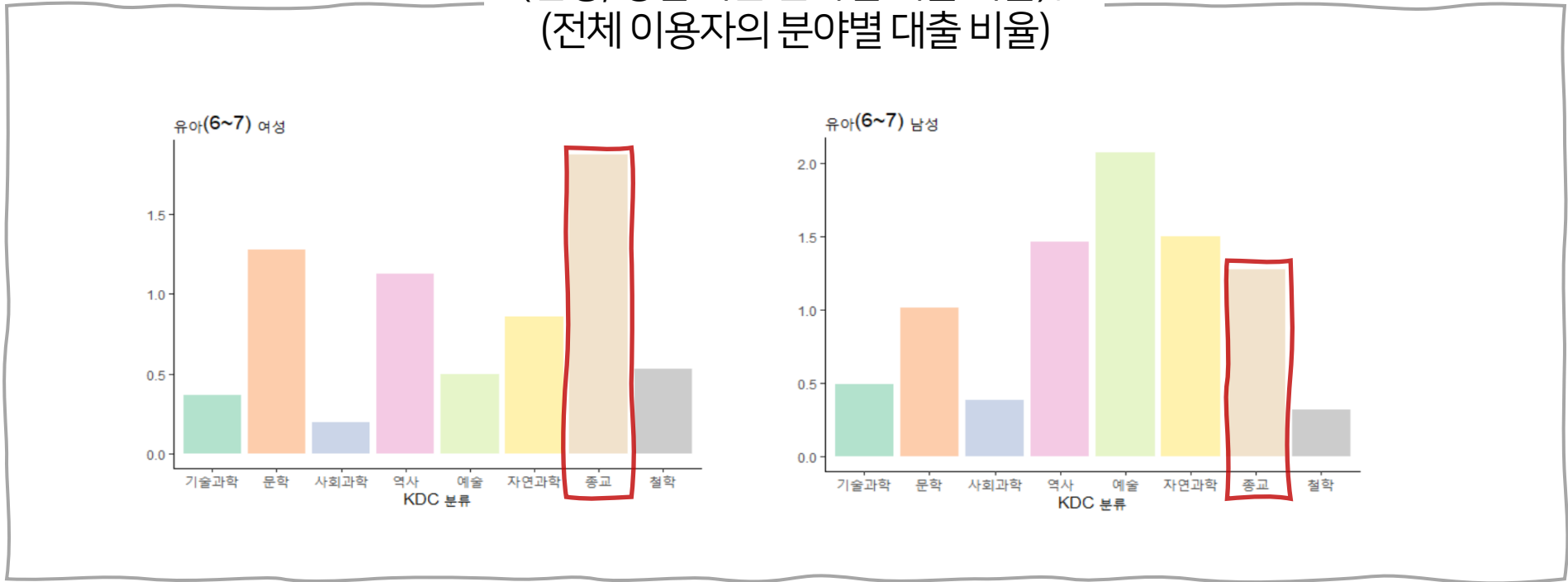
...

절대적으로 높던 '문학'의 비율이 줄어든 것을 보아
각 연령대, 성별별 두드러지는 선호 분야를 올바르게 찾을 수 있을 것이라 판단

3. 클러스터링

EDA | 분야별 대출 비율

(연령, 성별 기준 분야별 대출 비율) /
(전체 이용자의 분야별 대출 비율)

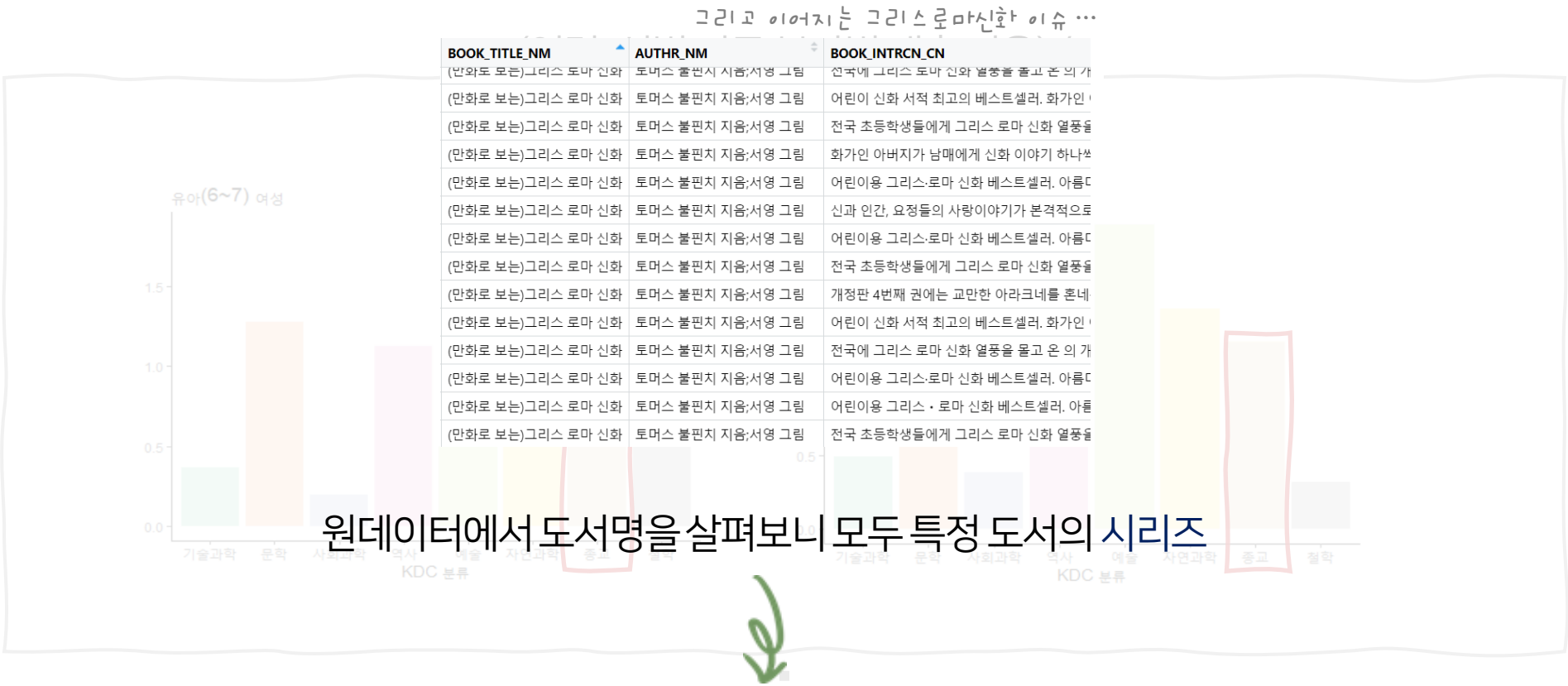


⋮

유아(6~7)의 경우 '종교'의 비율이 높음

3. 클러스터링

EDA | 분야별 대출 비율



앞에서 제외한 연령대이긴 하지만
고객군별 두드러지는 특징을 잘 잡을 수 있음을 확인함!



3. 클러스터링

| 클러스터링 기법

K-means

반복적으로 클러스터의
평균을 업데이트하며
가장 가까운 점들을 군집화

K-medoids

K-means를 변형한 것으로,
데이터의 **중간점**을 사용해
이상치에 덜 민감하도록 군집화

계층적 군집화

가장 유사도가 높은 군집 두 개를
하나로 합치면서 **군집의 개수를**
줄여 나가는 방식으로 군집화

DBSCAN

데이터가 기하학적 특징을
가질 때 유용한 방법으로
데이터의 **밀도**를 활용해 군집화

GMM

데이터가 여러 다른 모양의 가우시안 분포로
결합되어 있다는 가정 하에
개별 데이터를 **동일한 가우시안 분포**별로
묶어주는 비지도 학습 알고리즘

3. 클러스터링

| 클러스터링 기법

K-means

반복적으로 클러스터의
평균을 업데이트하며
가장 가까운 점들을 군집화

K-medoids

K-means를 변형한 것으로,
데이터의 **중간점**을 사용해
이상치에 덜 민감하도록 군집화

계층적 군집화

가장 유사도가 높은 군집 두 개를
하나로 합치면서 **군집의 개수를**
줄여 나가는 방식으로 군집화

DBSCAN

데이터가 기하학적 특징을
가질 때 유용한 방법
데이터의 **밀도**를 활용해 군집화



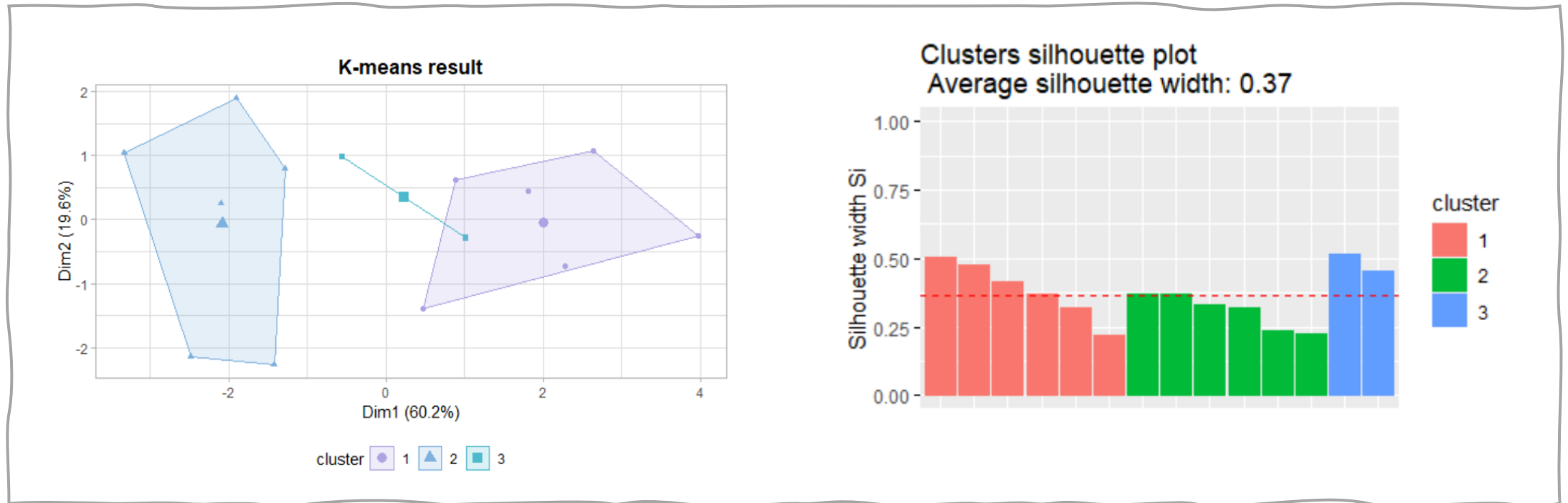
위 세 개의 클러스터링 기법을 사용해봄!

GMM

데이터가 여러 다른 모양의 가우시안 분포로
결합되어 있다는 가정 하에
개별 데이터를 **동일한 가우시안 분포**별로
묶어주는 비지도 학습 알고리즘

3. 클러스터링

| 클러스터링 | K-means



⋮

최종 군집 개수 $K = 3$ 으로 클러스터링 진행

3. 클러스터링

| 클러스터링 | K-medoids, 계층적 군집화

K-means와 클러스터링 결과 동일



데이터의 특성을 고려했을 때 DBSCAN과 GMM은 적절하지 않다고 판단



'분야 선호 클러스터' 결과가 타당하다고 판단!



3. 클러스터링

| 클러스터링 결과

클러스터1

연령대	성별	철학	...	예술	문학	역사
30대	남성	0.75	...	0.67	0.72	1.27
40대	여성	0.7	...	0.69	0.89	1.63
⋮			⋮			⋮
초등 (8~13)	여성	0.5	...	0.41	0.99	1.67

클러스터2

연령대	성별	철학	...	예술	문학	역사
20대	남성	2.3	...	1.2	0.67	0.25
50대	남성	1.6	...	1.16	0.61	0.7
⋮			⋮			⋮
60대 이상	여성	1.678	...	1.79	1.3	0.5

클러스터3

연령 대	성별	철학	...	예술	문학	역사
청소년 (14~19)	남성	0.96	...	0.78	0.75	0.87
청소년 (14~19)	여성	0.95	...	1.26	1.08	0.36

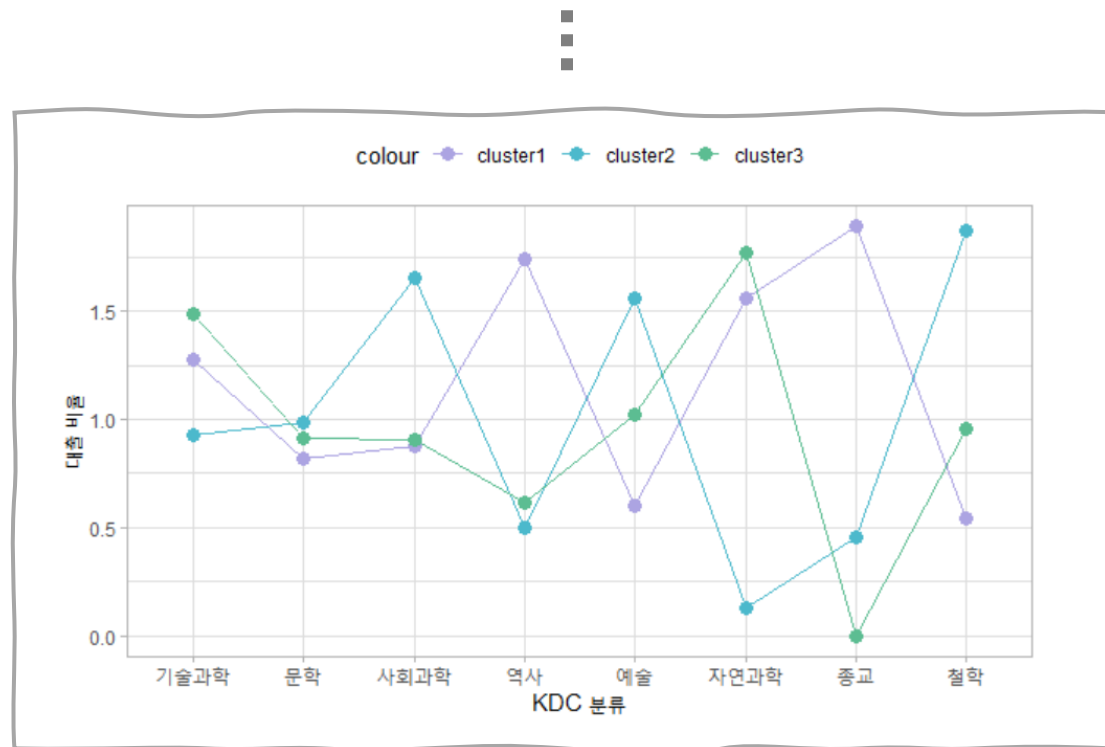
새로운 '분야 선호 클러스터' 도출



3. 클러스터링

| 클러스터링 결과

각 클러스터에 속하는 고객군의 대출 비율의 **평균**을 구해서 시각화



▲군집별로 두드러지는 선호 분야 파악

3. 클러스터링

| 클러스터링 결과

각 클러스터에 속하는 고객군의 대출 비율의 **평균**을 구해서 시각화

⋮

클러스터 1

'종교', '역사' 분야 선호

클러스터 2

'철학', '사회과학', '예술' 분야 선호

클러스터 3

'기술과학', '자연과학' 분야 선호

▲ 군집별로 두드러지는 선호 분야 파악

04

이슈 추출

4. 이슈 추출

| 필터 버블

필터 버블

추천 시스템이 개인화된 정보를 제공하는 과정에서 이용자에게 특정 정보만을 추천하는 현상



확증 편향

이용자는 수많은 정보 가운데 기존 관심에 부합하는
일부 정보만을 받게 되어 그 편향성이 강화됨

4. 이슈 추출

| 필터 버블

필터 버블

추천 시스템이 개인화된 정보를 제공하는 과정에서 이용자에게 특정 정보만을 추천하는 현상



확증 편향

이용자는 수많은 정보 가운데 기존 관심에 부합하는
일부 정보만을 받게 되어 그 편향성이 강화됨

4. 이슈 추출

| 필터 버블

필터 버블

추천 시스템이 개인화된 정보를 제공하는 과정에서 이용자에게 특정 정보만을 추천하는 현상

유튜브 또한 필터 버블 현상 존재,
유튜브 기반 도서 추천 또한 필터
버블이 발생할 수 있음



도서 추천 시
사회 이슈 활용!

4. 이슈 추출

| 필터 버블

필터 버블

추천 시스템이 개인화된 정보를 제공하는 과정에서 이용자에게 특정 정보만을 추천하는 현상

유튜브 또한 필터 버블 현상 존재,
유튜브 기반 도서 추천 또한 필터
버블이 발생할 수 있음



도서 추천 시
사회 이슈 활용!

4. 이슈 추출

| 빅카인즈 특성 추출

날짜: 2023.10.06 ~ 2023.11.06

분야: 정치, 경제, 사회, 국제, 스포츠, IT_과학

* 사진, 만평 등은 분석대상에서 제외



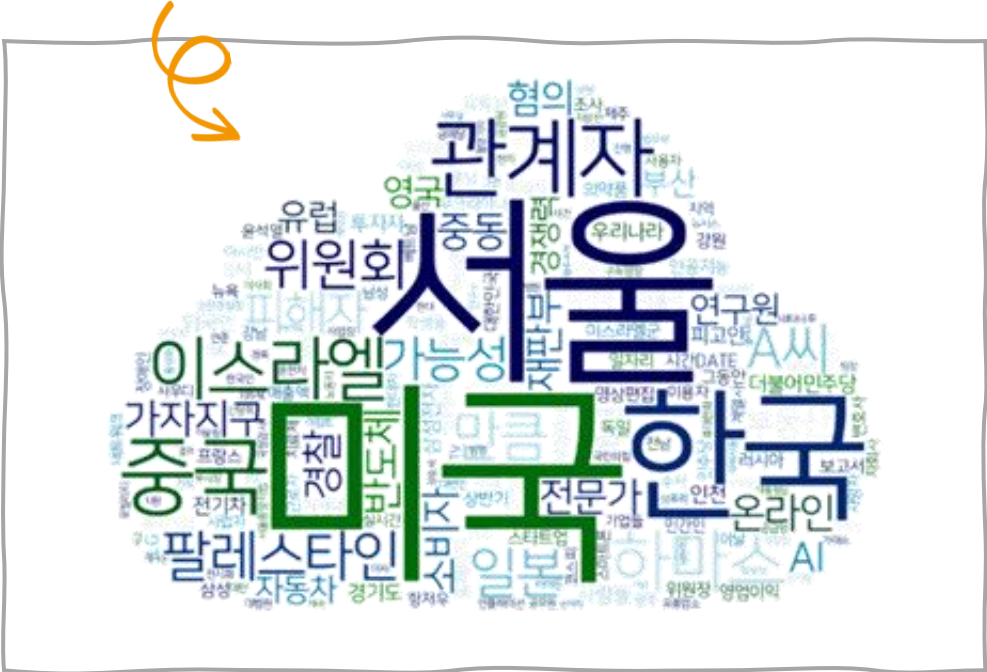
뉴스 식별자	일자	언론사	...	키워드	특성추출 (가중치순 상위 50개)	본문	...
02100501.20 2310311350 33001	20231031	파이낸셜뉴스	...	화학,산업,은택훈장,대 표,KPX,케미칼,최재호 ...	화학산업,최재호,은택,관 계자,정동건,장수영,장영 진...	[파이낸셜뉴스]산업통상 자원부가...	...
02100501.20 2310311350 32001	20231031	파이낸셜뉴스	...	공사비,초과,달라,쌍용건 설,신사옥,KT,판교,시위 ...	공사비,쌍용건설,kt,판교 신사옥,국토부...	10월 31일 쌍용건설과 하도급 업체 직원들이
...

4. 이슈 추출

| 빅카인즈 특성 추출

뉴스 식별자	일자	언론사	...	키워드	특성추출 (가중치순 상위 50개)	본문	...
02100501.20 2310311350 33001	20231031	파이낸셜뉴스	...	화학,산업,은택훈장,대 표,KPX,케미칼,최재호 ...	화학산업,최재호,은택,관 계자,정동건,장수영,장영 진...	[파이낸셜뉴스]산업통상 자원부가...	...
02100501.20 2310311350 32001	20231031	파이낸셜뉴스	...	공사비,초과,달라,쌍용건 설,신사옥,KT,판교,시위 ...	공사비,쌍용건설,kt,판교 신사옥,국토부...	10월 31일 쌍용건설과 하도급 업체 직원들이

전체 기사 본문 키워드 바탕으로 추출한
특성 상위 50개의 단순 빈도로 워드 클라우드 생성



4. 이슈 추출

빅카인즈 특성 추출

뉴스 식별자	일자	언론사	...	키워드	특성추출 (가중치순 상위 50개)	본문	...
02100501.20 2310311350 33001	20231031	파이낸셜뉴스	...	화학,산업,은택훈장,대 표,KPX,케미칼,최재호 ...	화학산업,최재호,은택,관 계자,정동건,장수영,장영 진...	[파이낸셜뉴스]산업통상 자원부가...	...
02100501.20 2310311350 32001	20231031	파이낸셜뉴스	...	공사비,초과,달라,쌍용건 설,신사옥,KT,판교,시위 ...	공사비,쌍용건설,kt,판교 신사옥,국토부...	10월 31일 쌍용건설과 하도급 업체 직원들이



유의미한 사회적 이슈 혹은 특징보다는
보편적으로 많이 나오는 단어들이 단순 나열됨
ex) 나라 이름, 도시 이름 등



4. 이슈 추출

빅카인즈 특성 추출

뉴스 식별자	일자	언론사	...	키워드	특성추출 (가중치순 상위 50개)	본문	...
02100501.20 2310311350 33001	20231031	파이낸셜뉴스	...	화학,산업,은택훈장,대 표,KPX,케미칼,최재호 ...	화학산업,최재호,은택,관 계자,정동건,장수영,장영 진...	[파이낸셜뉴스]산업통상 자원부가...	...
02100501.20 2310311350 32001	20231031	파이낸셜뉴스	...	공사비,초과,달라,쌍용건 설,신사옥,KT,판교,시위 ...	공사비,쌍용건설,kt,판교 신사옥,국토부...	10월 31일 쌍용건설과 하도급 업체 직원들이



유의미한 사회적 이슈 혹은 특징보다는
보편적으로 많이 나오는 단어들이 단순 나열됨
ex) 나라 이름, 도시 이름 등

보편적인 단어들은 모든 분야에서
고르게 많이 등장하는 반면
이슈를 담은 키워드는 하나의 분야에
밀집하여 나오지 않을까?



기사의 분야(섹션)를 하나의 문서로 보고
TF-IDF의 아이디어 적용!

4. 이슈 추출

빅카인즈 특성 추출

뉴스 식별자	일자	언론사	...	키워드	특성추출 (가중치순 상위 50개)	본문	...
02100501.20 2310311350 33001	20231031	파이낸셜뉴스	...	화학,산업,은택훈장,대 표,KPX,케미칼,최재호 ...	화학산업,최재호,은택,관 계자,정동건,장수영,장영 진...	[파이낸셜뉴스]산업통상 자원부가...	...
02100501.20 2310311350 32001	20231031	파이낸셜뉴스	...	공사비,초과,달라,쌍용건 설,신사옥,KT,판교,시위 ...	공사비,쌍용건설,kt,판교 신사옥,국토부...	10월 31일 쌍용건설과 하도급 업체 직원들이



유의미한 사회적 이슈 혹은 특징보다는
보편적으로 많이 나오는 단어들이 단순 나열됨
ex) 나라 이름, 도시 이름 등

보편적인 단어들은 모든 분야에서
고르게 많이 등장하는 반면
이슈를 담은 키워드는 하나의 분야에
밀집하여 나오지 않을까?



기사의 분야(섹션)를 하나의 문서로 보고
TF-IDF의 아이디어 적용!

4. 이슈 추출

| 빅카인즈 특성 추출

TF-IDF

단어의 빈도와 문서의 빈도를 사용하여 단어마다 중요한 정도에 따라 가중치를 부여하는 방법으로,
모든 문서에 등장하는 단어는 중요도가 낮고, **특정 문서에만 등장**하는 단어는 중요도가 높음

$$TF(t, d) = \frac{\text{문서 } d \text{에 등장하는 단어 } t \text{의 빈도}}{\text{문서 } d \text{의 총 단어 개수}}$$

$$IDF(t, D) = \log \frac{\text{총 문서 개수}}{\text{단어 } t \text{를 포함하는 문서의 개수}}$$

⋮

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

4. 이슈 추출

| 빅카인즈 특성 추출

TF-IDF

단어의 빈도와 문서의 빈도를 사용하여 단어마다 중요한 정도에 따라 가중치를 부여하는 방법으로,
모든 문서에 등장하는 단어는 중요도가 낮고, **특정 문서에만 등장**하는 단어는 중요도가 높음

$$TF(t, d) = \frac{\text{문서 } d \text{에 등장하는 단어 } t \text{의 빈도}}{\text{문서 } d \text{의 총 단어 개수}}$$

$$IDF(t, D) = \log \frac{\text{총 문서 개수}}{\text{단어 } t \text{를 포함하는 문서의 개수}}$$

⋮

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

4. 이슈 추출

빅카인즈 특성 추출

TF-IDF

이미 특성 추출 및 WordCount, 필터링이 진행된 데이터였기 때문에,

식을 그대로 적용하기는 부적절하다고 판단

단어의 빈도와 문서의 빈도를 사용하여 단어마다 중요한 정도에 따라 가중치를 부여하는 방법,
모든 문서에 등장하는 단어는 중요도가 낮으며, 특정 문서에만 등장하는 단어는 중요도가 높다

최초 Custom Score

$$score = \frac{\log(1 + \text{해당 섹션에서 등장 빈도})}{\log(1 + \text{단어를 포함하는 섹션의 개수})}$$

$$TF(t, d) = \frac{\text{문서 } d \text{에 단어 } t \text{의 등장 횟수}}{\text{문서 } d \text{의 총 단어 개수}}$$

$$IDF(t, D) = \log \frac{\text{총 문서 개수}}{\text{단어 } t \text{를 포함하는 문서의 개수}}$$

단어의 등장 횟수에 대한 가중치를 주고,

여러 문서에서 등장하는 단어에 대해 페널티를 주는 아이디어 자체는 TF-IDF와 동일

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

① 분야별 단어 단순 등장 빈도 추출, 상위 150개 선정(필터링)

Word	IT_과학	경제	국제	사회	스포츠	정치	지역
미국	2526	7646	8878	1564	844	2890	393
관계자	1941	1394	1394	3795	174	1774	3845
한국	1854	5967	3263	1698	2393	2477	1192
스타트업	938	1318	0	0	0	0	0

② IT 분야 외의 분야에서 각 단어 등장 유무 판단

Word	IT_과학	경제	국제	사회	스포츠	정치	지역	SUM
미국	2526	1	1	1	1	1	1	6
관계자	1941	1	1	1	1	1	1	6
한국	1854	1	1	1	1	1	1	6
스타트업	938	1	0	0	0	0	0	1

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

① 분야별 단어 단순 등장 빈도 추출, 상위 150개 선정(필터링)

Word	IT_과학	경제	국제	사회	스포츠	정치	지역
미국	2526	7646	8878	1564	844	2890	393
관계자	1941	1394	1394	3795	174	1774	3845
한국	1854	5967	3263	1698	2393	2477	1192
스타트업	938	1318	0	0	0	0	0

② IT 분야 외의 분야에서 각 단어 등장 유무 판단

Word	IT_과학	경제	국제	사회	스포츠	정치	지역	SUM
미국	2526	1	1	1	1	1	1	6
관계자	1941	1	1	1	1	1	1	6
한국	1854	1	1	1	1	1	1	6
스타트업	938	1	0	0	0	0	0	1

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

① 분야별 단어 단순 등장 빈도 추출, 상위 150개 선정(필터링)

Word	IT_과학	경제	국제	사회	스포츠	정치	지역
미국	2526	7646	8878	1564	844	2890	393
관계자	1941	1394	1394	3795	174	1774	3845
한국	1854	5967	3263	1698	2393	2477	1192
스타트업	938	1318	0	0	0	0	0

② IT 분야 외의 분야에서 각 단어 등장 유무 판단

Word	IT_과학	경제	국제	사회	스포츠	정치	지역	SUM
미국	2526	1	1	1	1	1	1	6
관계자	1941	1	1	1	1	1	1	6
한국	1854	1	1	1	1	1	1	6
스타트업	938	1	0	0	0	0	0	1

한 번이라도 등장한 경우 1,
그렇지 않은 경우는 그대로 0

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

③ 각 빈도 유무를 합한 후 minmax scaling 진행

Word	IT_과학	경제	국제	사회	스포츠		SUM	minmax
미국	2526	1	1	1	1	...	6	1
관계자	1941	1	1	1	1		6	1
한국	1854	1	1	1	1		6	1
스타트업	938	1	0	0	0		1	0.1666



Min-max scaling 진행 이유

Scaling을 진행하지 않은 경우 섹션 등장 횟수에 대한 페널티가 심하게 커짐

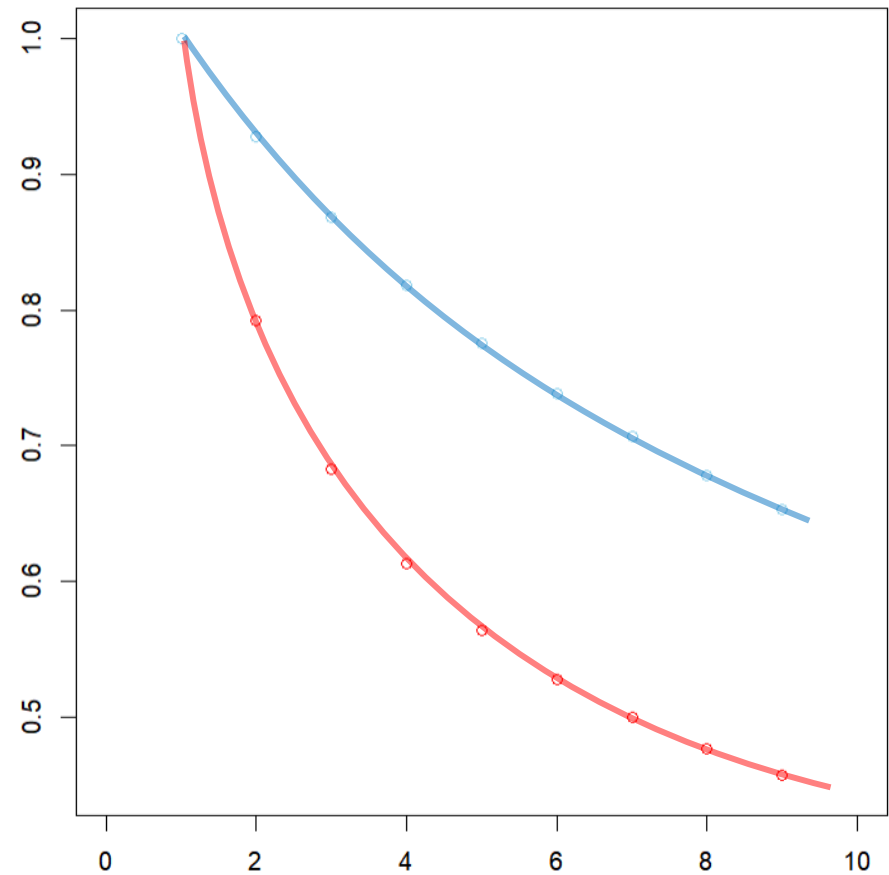
→ 페널티를 완화해주는 역할

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

③ 각 빈도 유무를 합한 후 minmax scaling 진행



국제	사회	스포츠		SUM	minmax
1	1	1		6	1
1	1	1		3	0.5
0	0	0		1	0.1666

1개의 섹션에만 있는 단어에 비해 x개의 문서에서 등장한 단어의 Score가 어느 정도의 비율로 낮아지는지 나타낸 Plot

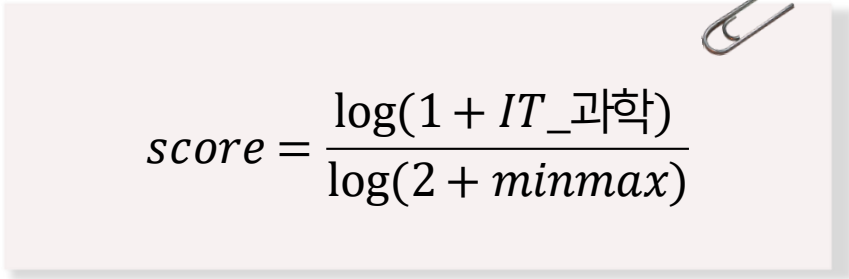
최초 cumstom 식(빨간색)에 비해 min-max 적용 식(파란색)이 페널티를 덜 주고 있는 모습을 확인 가능
→ 페널티를 완화해주는 역할

4. 이슈 추출

TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

④ 최종 Custom Score


$$score = \frac{\log(1 + IT_과학)}{\log(2 + minmax)}$$

- *it_과학*: IT_과학 분야에서 해당 단어 등장 횟수
- *minmax* : IT_과학 분야를 제외한 분야에서의 등장 유무 합에 min-max scaling

Word	IT_과학	경제	국제	사회	스포츠		SUM	minmax	score
미국	2526	1	1	1	1	...	6	1	7.1315
관계자	1941	1	1	1	1		6	1	6.8918
한국	1854	1	1	1	1		6	1	6.8501
스타트업	938	1	0	0	0		1	0.1666	8.852

4. 이슈 추출

| TF-IDF 아이디어 적용 과정

EXAMPLE) IT_과학 분야 뉴스 크롤링 데이터

⑤ 결과 확인

Word	IT_과학	경제	국제	사회	스포츠		SUM	minmax	score
구글	476	0	0	0	0	...	1	0.1666	8.897
빅데이터	463	0	0	0	0		0	0	8.857
스타트업	938	1	0	0	0		1	0	8.852
삼성전자	910	1	0	0	0		1	0.1666	8.813

⋮

score가 높을수록 해당 분야(IT_과학)의 이슈를 대표한다고 판단,
score가 높은 순으로 정렬

4. 이슈 추출

결과

경제



IT_과학



4. 이슈 추출

결과

국제



스포츠



I 결과

결과 비교



전체 워드 클라우드에서도 보편적인 단어보다
사회 이슈를 반영한 단어 등장 빈도 ↑

전체 : scoring 후 워드 클라우드



05

유튜브 키워드 추출

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보



제목



댓글



해시태그




자막

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보




제목




시청자가 썸네일과 함께 영상에 대해 파악할 수 있는 요소
흥미를 유발하고 호기심을 자극하는 역할

EX



[레도X김상욱] 우주와 물리학 기막힌 콜라보

와..... 진짜 제대로 만든 영화를 보시려거든 이런 작품을 보세요



댓글

해시태그

자막


지식, 정보 전달 관련 영상은 주요 소재가 표시되는 반면,
오락(영화, 게임 등) 관련 영상은 키워드가 잘 표시되지 않는 경우 多

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보




제목




시청자가 썸네일과 함께 영상에 대해 파악할 수 있는 요소
흥미를 유발하고 호기심을 자극하는 역할

EX



[레도X김상욱] 우주와 물리학 기막힌 콜라보



와..... 진짜 제대로 만든 영화를 보시려거든 이런 작품을 보세요

댓글

해시태그




자막

정보 전달성 영상과 오락성 영상을 구분한 후,
정보 전달성 영상에 대해 제목 활용

구분은 부록 참고!

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보

		 영상의 시청자가 영상을 보고 남긴 반응 <div data-bbox="1375 596 2333 931"><p>궤도의 안광이 여전히 빛나고 있어서 너무 좋다</p><p>.....</p><p>2부 언제 나오냐 개재밌다 진심...</p><p>궤도 김상욱 교수 미친 조합</p></div>	
제목	댓글	해시태그	자막

영상이나 출연자에 대한 본인의 **감상**을 남긴 경우가 대부분,
영상의 내용을 파악하기에 적합하지 않음 → 제외

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보

업로더가 사용자의 유입을 위해 설정한 검색 연관어

궤도 # 김상욱 # 우주물리학 # 스페이스허브
스페이스하이커 # 한화

.....
ASMR # 떡볶이먹방 # 먹거리 # 맛집

제목

댓글

#

해시태그



자막

영상의 출연자가 언급되거나,
영상 내용과 관련 없이 유입만을 위한 해시태그가 있는 경우도 多

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보

업로더가 사용자의 유입을 위해 설정한 검색 연관어

궤도 # 김상욱 # 우주물리학 # 스페이스허브
스페이스하이커 # 한화

.....
ASMR # 떡볶이먹방 # 먹거리 # 맛집

제목

댓글

#

해시태그



자막

영상 키워드 추출에는 사용하지 않기로 결정!

이후 정보 전달성 영상과 오락성 영상을 구분할 때 활용

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보

업로더가 직접 업로드하거나,
따로 업로드하지 않은 경우 동영상의 **음성과 소리를 그대로 입력**



[음악] 세상에 없던 우주자파 지식 토크쇼, 스페이사이코신 여러분 환영합
니다. 저는 여러분의 우주여행을 도울

...

적용받고 움직인다 이것이 물리학의 핵심이죠 그래서 ...



자막

음성이 그대로 입력되는 경우 질이 다소 떨어지지만 영상의 내용, 소재를 충분히 담고 있음

5. 유튜브 키워드 추출

| 유튜브에서 얻을 수 있는 정보

업로더가 직접 업로드하거나,
따로 업로드하지 않은 경우 동영상의 **음성과 소리를 그대로 입력**

[음악] 세상에 없던 우주자파 지식 토크쇼, 스페이사이코신 여러분 환영합
니다. 저는 여러분의 우주여행을 도울

...

적용받고 움직인다 이것이 물리학의 핵심이죠 그래서 ...



자막

길이가 길고 불필요한 정보가 있기 때문에 **문서 요약** 필요

5. 유튜브 키워드 추출

| 문서 요약

문서 요약

문서 내용을 짧게 축약하여 핵심적인 내용만을 담는 과정

추출적 요약

주어진 문서 집합 내에서 이를 대표할 수 있는
단어들이나 문장들을 **선택**하는 방법

생성적 요약

실제 사람이 요약문을 만드는 것처럼,
문서의 내용을 기반으로 요약문을
새롭게 생성하는 방법

5. 유튜브 키워드 추출

| 문서 요약

문서 요약

문서 내용을 짧게 축약하여 핵심적인 내용만을 담는 과정

추출적 요약

주어진 문서 집합 내에서 이를 대표할 수 있는
단어들이나 문장들을 **선택**하는 방법

생성적 요약

주어진 문서 데이터 내에서 단어와 문장을 선택하므로,
터무니없는 요약 결과를 만들어 낼 가능성은 적음
But, 가능한 표현이 제한된다는 단점

5. 유튜브 키워드 추출

| 문서 요약

문서 요약

문서 내용을 짧게 축약하여 핵심적인 내용만을 담는 과정

추출적 요약

원문의 내용을 보다 자연스럽게 요약할 수 있으며,
문서에 없던 단어나 스타일이 반영 가능

생성적 요약

실제 사람이 요약문을 만드는 것처럼,
문서의 내용을 기반으로 요약문을
새롭게 생성하는 방법

5. 유튜브 키워드 추출

| 문서 요약

문서 요약

문서 내용을 짧게 축약하여 핵심적인 내용만을 담는 과정



추출형 요약

학습 데이터를 기반으로 한 **Supervised Learning**

주어진 문서 집합 내에서 이를 대표할 수 있는

특정 도메인의 문서를 요약하는 모델을 만들기 위해

해당 도메인을 요약한 학습 데이터가 반드시 필요

생성적 요약

실제 사람이 요약문을 만드는 것처럼,

문서의 내용을 기반으로 요약문을

새롭게 생성하는 방법

5. 유튜브 키워드 추출

| 문서 요약

문서 요약

문서 내용을 짧게 축약하여 핵심적인 내용만을 담는 과정

요약하고자 하는 스크립트가 어떤 도메인인지 알 수 없는 상황

학습 데이터의 부재



생성적 요약

학습 데이터를 기반으로 한 Supervised Learning

실제 사람이 요약문을 만드는 것처럼,

특정 도메인의 문서를 요약하는 모델을 만들기 위해
해당 도메인을 요약한 학습 데이터가 반드시 필요

비지도 학습인 TextRank (추출적 요약) 알고리즘 사용!

문서의 내용을 기반으로 요약문을
새롭게 생성하는 방법

5. 유튜브 키워드 추출

| PageRank

PageRank

특정 페이지를 인용하는 다른 페이지가 얼마나 있는지를 정규화하여 세는 방법

$$PR(A) = \frac{(1-d)}{N} + d \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

$PR(A)$: 웹페이지 A의 PageRank

$C(T_1)$: T1이라는 페이지가 가지고 있는 링크의 총 개수 (T1이 인용한 페이지 수)

d (damping factor) : 웹서핑을 하는 사람이 그 페이지에 만족하지 못하고 다른 인용 링크를 클릭할 확률

↪ 크게 중요하지는 않음! 보통 0.85로 설정

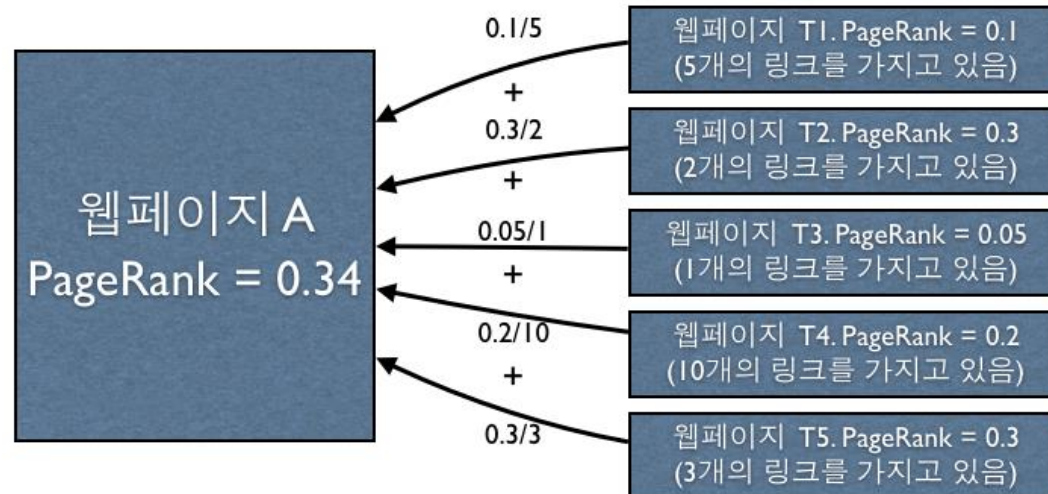
5. 유튜브 키워드 추출

PageRank

PageRank

특정 페이지를 인용하는 다른 페이지가 얼마나 있는지를 정규화하여 세는 방법

$$PR(A) = \frac{(1-d)}{N} + d \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$



5. 유튜브 키워드 추출

PageRank

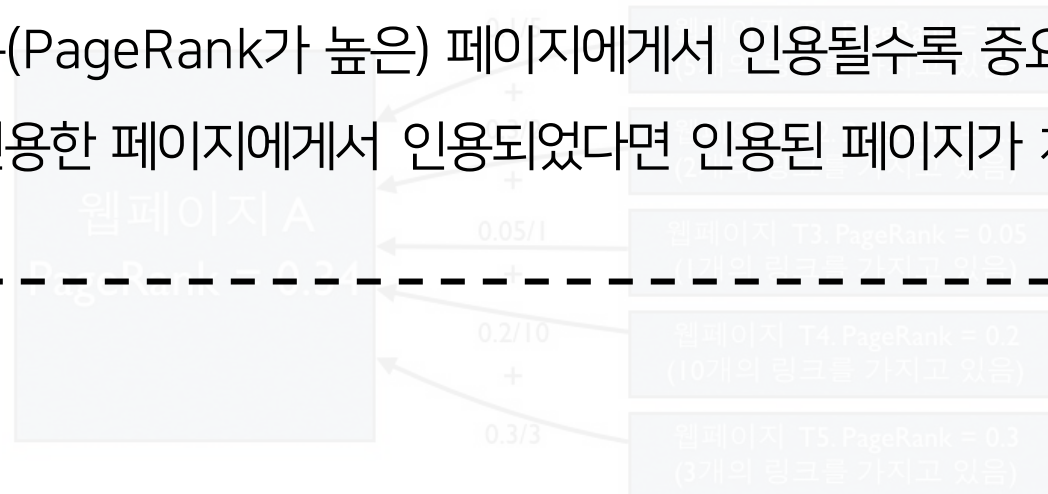
PageRank

특정 페이지를 인용하는 다른 페이지가 얼마나 있는지를 정규화하여 세는 방법

기본적인 아이디어

영향력 있는(PageRank가 높은) 페이지에게서 인용될수록 중요도가 올라감

다른 많은 페이지를 인용한 페이지에게서 인용되었다면 인용된 페이지가 차지하는 비중이 낮아짐



5. 유튜브 키워드 추출

| PageRank

PageRank

특정 페이지를 인용하는 다른 페이지가 얼마나 있는지를 정규화하여 세는 방법

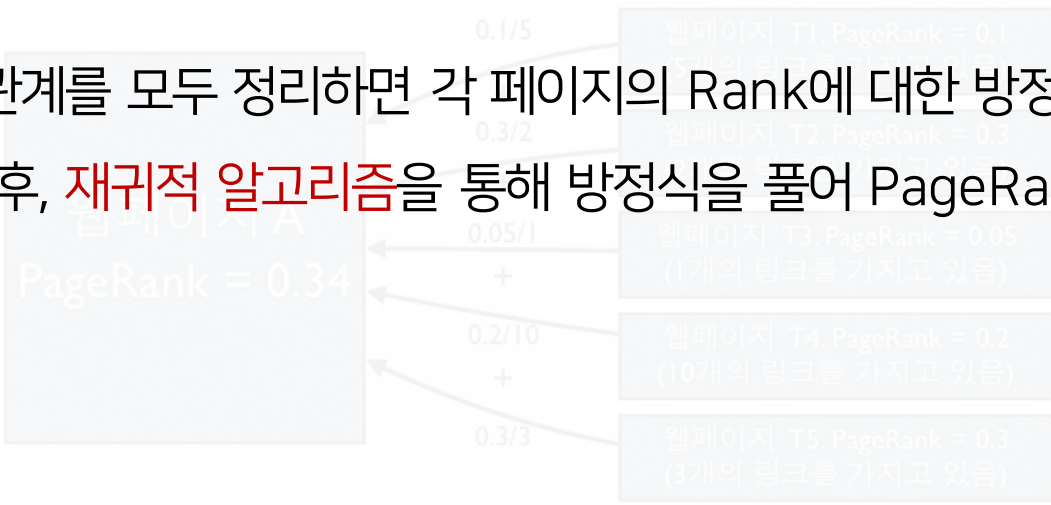


다른 페이지들의 PageRank가 계산되어 있지 않을 텐데 A의 PageRank는 어떻게 계산?

$$PR(A) = \frac{1}{N} + d \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$



페이지들의 관계를 모두 정리하면 각 페이지의 Rank에 대한 방정식 설정 가능
초기값 설정 후, 재귀적 알고리즘을 통해 방정식을 풀어 PageRank 해를 구함



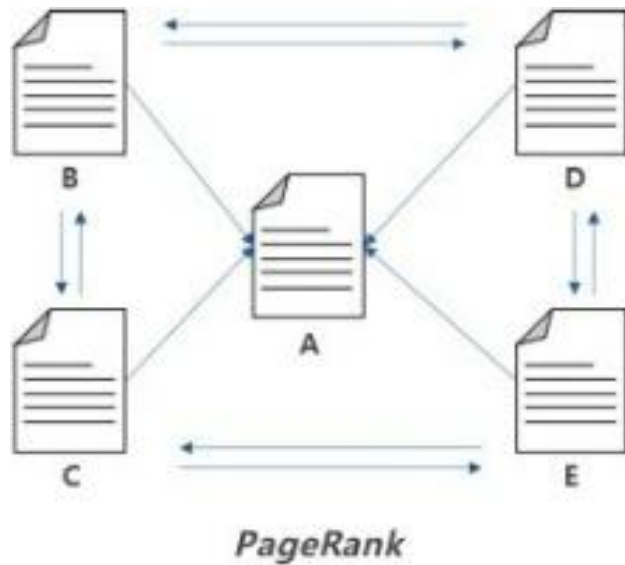
통계계산입문 수업에서 배우는
Gauss-Seidel 알고리즘 참고!

5. 유튜브 키워드 추출

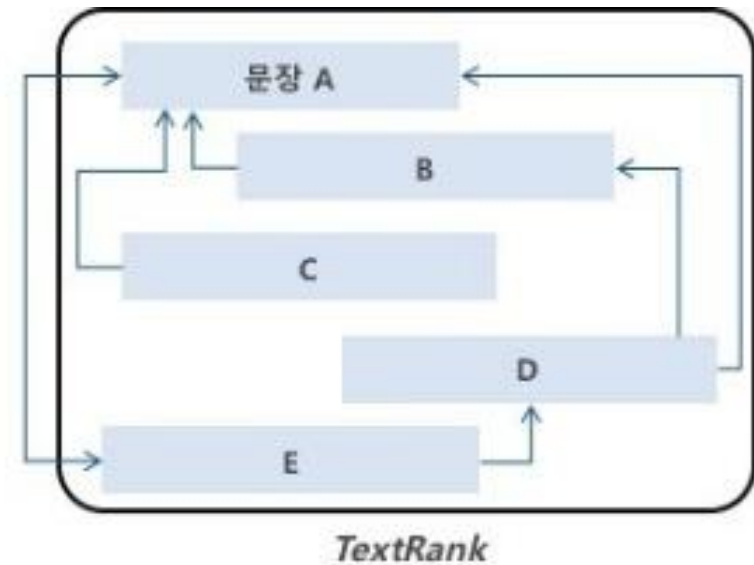
| TextRank

TextRank

PageRank 알고리즘을 응용하여, 문서 내 문장 또는 단어의 Ranking을 계산하는 알고리즘

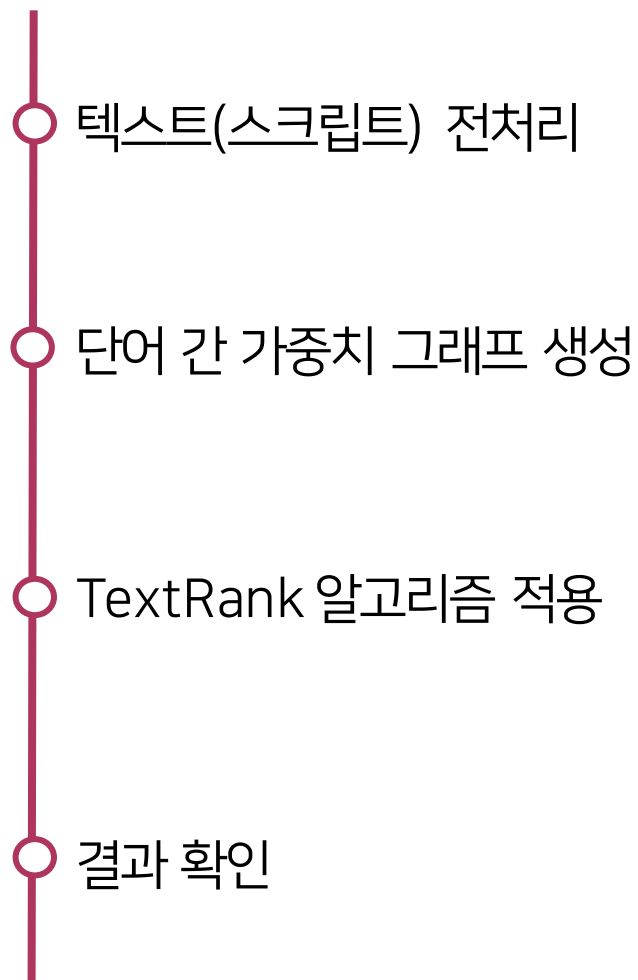


문서 내
문장에 적용



5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출



문서를 문장 단위로 분리 후, 형태소 토큰화
품사태깅을 통해 어근, 명사만 추출

TF-IDF 모델 생성

Sentence-Term Matrix의 상관계수 행렬을 가중치로 사용

가중치 그래프를 이용해 TextRank 알고리즘 적용

Rank가 높은 순으로 정렬 후 요약할 단어 개수만큼 출력

우주(9.75), 물리(5.05), 물리학(3.83), 중요(2.81) ...

5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출

○ 텍스트(스크립트) 전처리

문서를 문장 단위로 분리 후, 형태소 토큰화
품사태깅을 통해 어근, 명사만 추출

○ 단어 간 가중치 그래프 생성

TF-IDF 모델 생성

Sentence-Term Matrix의 상관계수 행렬을 가중치로 사용

— 토큰화(Tokenization)란? —

주어진 코퍼스(corpus)에서 **토큰(token)**이라 불리는 단위로 나누는 작업

토큰의 단위는 상황에 따라 다르지만, 보통 의미 있는 단위로 토큰을 정의

○ 결과 확인

우주(9.75), 물리(5.05), 물리학(3.83), 중요(2.81) ...

5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출

○ 텍스트(스크립트) 전처리

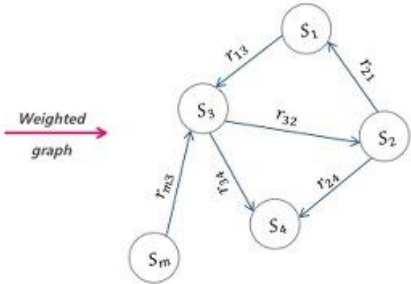
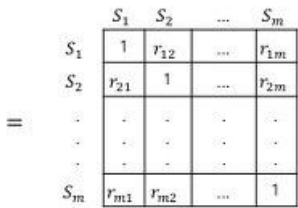
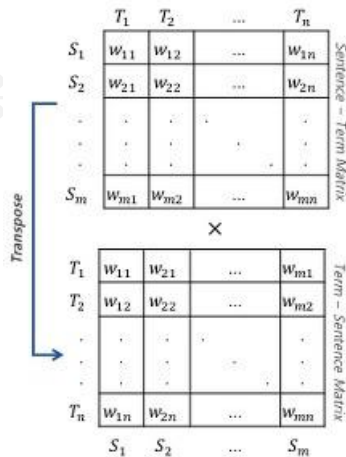
○ 단어 간 가중치 그래프 생성

○ TextRank 알고리즘 적용

○ 결과 확인

문서를 문장 단위로 분리 후, 형태소 토큰화
품사 태깅을 통해 어근, 명사만 추출

TF-IDF 모델 생성
Sentence-Term Matrix의 상관관계수 행렬을 가중치로 사용



TextRank 알고리즘 적용
→ 요약할 단어 개수만큼 출력

리학(3.83), 중요(2.81) ...

5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출

○ 텍스트(스크립트) 전처리

○ 단어 간 가중치 그래프 생성

○ TextRank 알고리즘 적용

○ 결과 확인

문서를 문장 단위로 분리 후, 형태소 토큰화
품사태깅을 통해 어근, 명사만 추출

TF-IDF 모델 생성

Sentence-Term Matrix의 상관계수 행렬을 가중치로 사용

가중치 그래프를 이용해 TextRank 알고리즘 적용

Rank가 높은 순으로 정렬 후 요약할 단어 개수만큼 출력

우주(9.75), 물리(5.05), 물리학(3.83), 중요(2.81) ...

5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출

○ 텍스트(스크립트) 전처리

문서를 문장 단위로 분리 후, 형태소 토큰화
품사태깅을 통해 어근, 명사만 추출

○ 단어 간 가중치 그래프 생성

TF-IDF 모델 생성

Sentence-Term Matrix의 상관계수 행렬을 가중치로 사용

○ TextRank 알고리즘 적용

가중치 그래프를 이용해 TextRank 알고리즘 적용

Rank가 높은 순으로 정렬 후 요약할 단어 개수만큼 출력

○ 결과 확인

우주(9.75), 물리(5.05), 물리학(3.83), 중요(2.81) ...

5. 유튜브 키워드 추출

| 토큰나이저별 키워드 추출 결과 비교

KOMORAN	KKMA	OKT	mecab
우주 (8.96)	우주 (9.18)	우주 (7.9)	우주 (9.75)
물리 (4.56)	물리 (6.24)	물리 (3.85)	물리 (5.05)
⋮	⋮	⋮	⋮
중요 (2.75)	중요 (2.88)	수 (3.18)	시작 (2.19)
전자 (2.55)	별 (2.67)	별 (2.34)	전자기력 (1.99)
⋮	⋮	⋮	⋮
과학 (1.75)	이해 (1.41)	여기 (1.83)	말 (1.38)

[궤도X김상욱] 우주와 물리학 기막힌 콜라보' 영상 기준

5. 유튜브 키워드 추출

| 토큰나이저별 키워드 추출 결과 비교

KOMORAN	KKMA	OKT	mecab
우주 (8.96)	우주 (9.18)	우주 (7.9)	우주 (9.75)
물리 (4.56)	물리 (6.24)	물리 (3.85)	물리 (5.05)
			⋮
‘전자기력’을 ‘전자’와 ‘기력’ 으로 구분하지 않고 그대로 토큰화 하는 등			시작 (2.19)
중요 (2.75)	중요 (2.89)	수 (3.18)	전자기력 (1.99)
전자 (2.55)	별 (2.67)	별 (2.34)	⋮
⋮	⋮	⋮	말 (1.38)
과학 (1.75)	이해 (1.41)	여기 (1.83)	

5. 유튜브 키워드 추출

| TextRank 활용 키워드 추출 결과

Mecab 키워드 추출 결과

우주 (9.75)	시작 (2.19)
물리 (5.05)	생각 (1.99)
물리학 (2.83)	전자기력 (1.99)
중요 (2.81)	과학 (1.78)
얘기 (2.4)	태양 (1.78)
사람 (2.4)	물리학자 (1.58)
지구 (2.4)	설명 (1.38)



[케도X김상욱] 우주와 물리학 기막힌 콜라보

스페이스 허브 TV (Space Hub TV)

조회수 85만회 • 5개월 전

5. 유튜브 키워드 추출

| TextRank 활용 키워드 추출 결과

Mecab 키워드 추출 결과

우주 (9.75)	시작 (2.19)
물리 (5.05)	생각 (1.99)
물리학 (2.83)	전자기력 (1.99)
중요 (2.81)	과학 (1.78)
얘기 (2.4)	태양 (1.78)
사람 (2.4)	물리학자 (1.58)
지구 (2.4)	설명 (1.38)



전반적으로 핵심 소재가 잘 추출되었지만,
영상의 핵심 내용과는 크게 관련이 없지만 Rank가 높거나
중요한 소재임에도 Rank가 낮은 경우가 생김



영상의 제목과 유사도 계산을 통해 한번 더 필터링!

4주차 예정

5. 유튜브 키워드 추출

| TextRank 활용 키워드 추출 결과

Mecab 키워드 추출 결과

우주 (9.75)	시작 (2.19)
물리 (5.05)	생각 (1.99)
물리학 (2.83)	전자기력 (1.99)
중요 (2.81)	과학 (1.78)
얘기 (2.4)	태양 (1.78)
사람 (2.4)	물리학자 (1.58)
지구 (2.4)	설명 (1.38)



전반적으로 핵심 소재가 잘 추출되었지만,
영상의 핵심 내용과는 크게 관련이 없지만 Rank가 높거나
중요한 소재임에도 Rank가 낮은 경우가 생김



영상의 제목과 유사도 계산을 통해 한번 더 필터링!

4주차 예정

4주차 예고

1. 텍스트 유사도 계산
2. 토픽 모델링
3. 분포 유사도 계산
4. 추천 시스템
5. 기대 효과

감사합니다!



자유까지
단 한 번 남았다 ...

06

Appendix

6. 부록

| 오락성 영상과 정보 전달성 영상의 구분

Motivation

보다 정확한 키워드 추출을 위해 제목을 활용할 예정이지만,
제목이 동영상의 내용을 대표하지 않는 경우가 있었음



영화 리뷰, 게임, 스포츠 등 오락성 영상에 대해 그런 경향성이 높음을 확인하여
오락성 영상에 대해서는 키워드 추출 시 제목을 활용하지 않기로 결정

6. 부록

| 오락성 영상과 정보 전달성 영상의 구분

Motivation

보다 정확한 키워드 추출을 위해 제목을 활용한 예정이었으나

다수의 영상에 대해 직접 오락성 여부를 labeling한 후 **지도 학습을 통해 분류**하는 방법도 고려하였으나,
시간적 한계로 rough하게만 구분하기로 결정..

+

또한 ASMR, 플레이리스트 등 사용자의 선호 분야가 드러나지 않을 것이라고 생각되는 영상은
분석에 포함하지 않기로 함

오락성 영상에 대해서는 키워드 추출시 제목을 활용하지 않기로 결정



6. 부록

| 오락성 영상과 정보 전달성 영상의 구분

1. 게임 : 게임 목록 리스트를 이용하여 제목 및 해시태그에 해당 게임 목록 리스트가 존재하는 경우 오락성 영상으로 판단
2. 드라마 및 영화 : [영화리뷰, 결말 포함, 몰아보기, 드라마] 등의 내용이 있는 경우 오락성 영상으로 판단
3. 스포츠 : 빅카인즈 '스포츠' 분류 내 특성 키워드에서 추출된 단어들이 있는 경우 오락성 영상으로 판단



아예 동영상을 분석에 포함시키지 않을 주제

'Vlog', 'ASMR', '먹방', 'playlist', 'GRWM', '~with me' 가 제목에 포함된 경우