

유튜브로 알아보는 나의 독서 DNA

유튜브 시청기록 기반 도서 추천시스템

회귀분석팀

김보근



서유진



하희나



김민주



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

| 도서추천 알고리즘 흐름



1. 데이터 수집

사회 이슈, 대출 정보, 도서 정보 데이터 수집

2. 사용자별 선호도 분포 형성

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

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 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, 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 일 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 양극지방
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 | 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



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

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 ;차해정 옮김	327.209	...	7일	20대	남성	3
	
가면산장 살인사건	지은이: 히가시노 게 이고 ;옮긴이: 김난주	833.6	...	30일	40대	여성	103

3. 클러스터링

| 분야별 선호 클러스터 생성

연령 및 성별 데이터를 그대로 사용하지 않고
새로운 클러스터 형성!



단순 연령, 성별로 사용자를 구분한다면
선호 분야 예측에 있어 오류 발생 가능성 존재
ex) 20대 남자라고 마냥 축구나 게임을 좋아하는 건 아님!

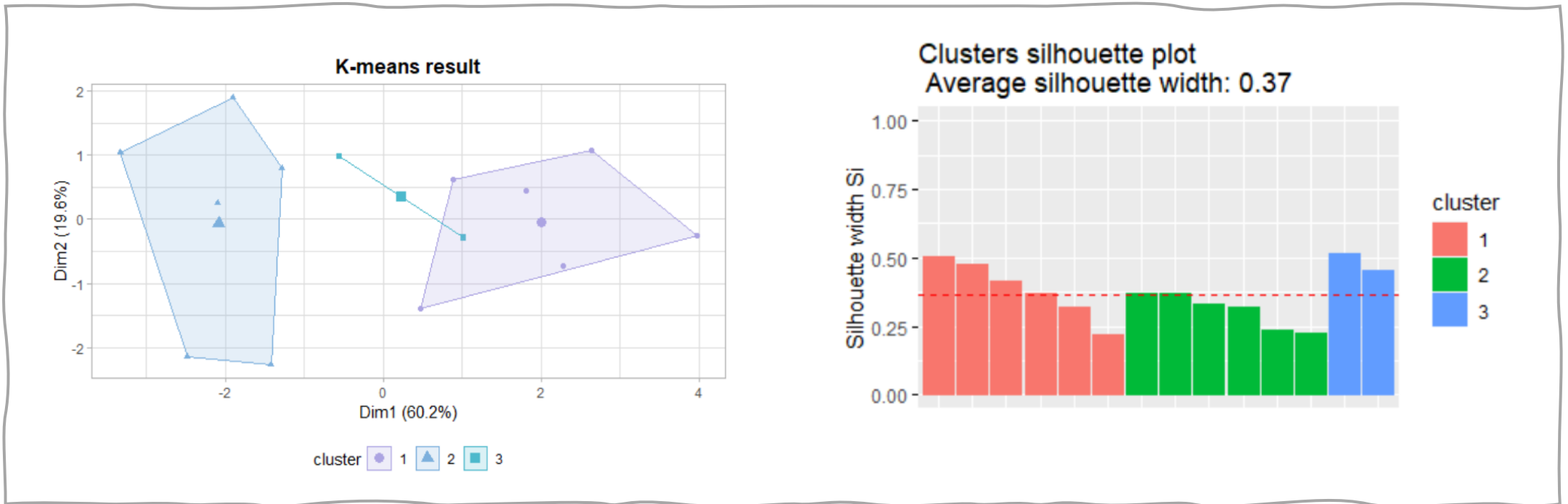


사용자와 비슷한 성향의 사람들의 선호를 반영!



3. 클러스터링

| 클러스터링 | K-means



...

최종 군집 개수 $K = 3$ 으로 클러스터링 진행

3. 클러스터링

| 클러스터링 | K-means

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 | K-means

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

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



유의미한 사회적 이슈 혹은 특징보다는
보편적으로 많이 나오는 단어들의 단순 나열
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

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

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

결과 비교

전체 : 기존 워드 클라우드



전체 : scoring 후 워드 클라우드



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 ...



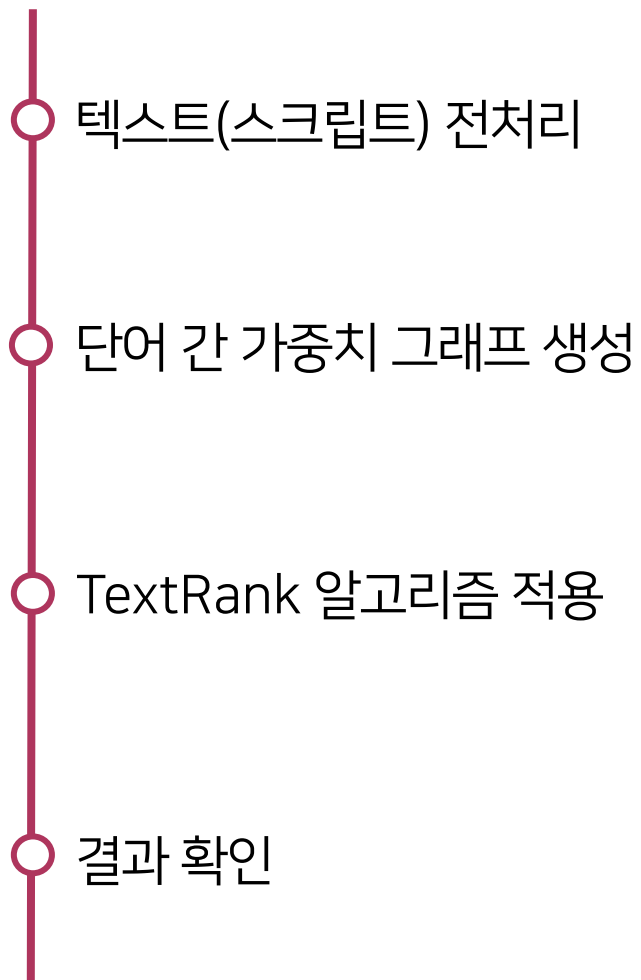
자막



길이가 길고 불필요한 정보가 있기 때문에 **문서 요약** 필요

5. 유튜브 키워드 추출

| TextRank 알고리즘을 적용한 유튜브 키워드 추출



문서를 문장 단위로 분리 후, 형태소 토큰화
품사 태깅을 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) ...

2. 키워드 추출 개선

| 분석 대상 영상 선별 | ② 정보성 영상 vs 오락성 영상



김상욱 교수의 '오펜하이머'에 관한 이야기 | 전쟁을 끝낸 원자폭탄의 원리와 위력 | 맨해튼 프로젝트

슬통강 전용 강의

베이지안 통계학
개념 잡기

1강. 이산형 분포

32:07

베이지 통계학 1강 - 사전분포, 사후분포, 베이지 통계학의 큰 그림 그리기



text 유사도 계산을 통해 제목과 자막을 모두 활용,
제목에 정보를 많이 담고 있으므로 가중치 부여!

2. 키워드 추출 개선

| 분석 대상 영상 선별 | ② 정보성 영상 vs 오락성 영상



처음부터 결말까지 개지리는 영화 ㅎㅎㅎㅎ...



구마유시 : 와 상혁이형 ...



게임, 영화, 스포츠와 관련된 영상의 경우
영상 속 자막을 통해 추출된 키워드만을 통해 주제 파악!
제목이 주제를 반영하지 못하는 문제 존재

2. 키워드 추출 개선

| 정보성 영상 | ① 키워드 추출

EXAMPLE) [케도X김상욱] 우주와 물리학 기막힌 콜라보

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!

스페이스 허브 TV (Space Hub TV)
조회수 85만회 · 5개월 전

2. 키워드 추출 개선

| 정보성 영상 | ② 제목과 키워드 간 유사도 계산

임베딩 (Embedding)

임베딩이란 자연어 처리에서 사람이 쓰는 자연어를 기계가 이해할 수 있도록
숫자 형태인 vector로 바꾸는 과정으로 단어 간 유사도 계산을 위해 필요

⋮

BERT

양방향 학습을 지원하는 알고리즘으로, 사전 훈련된 모델을 이용하여
언어 모델을 학습하는 방법



2. 키워드 추출 개선

정보성 영상 | ③ 최종 score 계산

EXAMPLE) [레도X김상욱] 우주와 물리학 기막힌 콜라보


$$score = similarity^2 * textrank$$

keywords	제목 간 유사도(similarity)	textrank	score
우주	0.5363	9.75	$0.5365^2 * 9.75 = 2.8055$
물리	0.2915	5.05	$0.2915^2 * 5.05 = 0.4294$
물리학	0.5057	2.83	$0.5057^2 * 2.83 = 0.9788$
중요	0.0700	2.81	$0.0700^2 * 2.81 = 0.0137$
...
설명	0.1070	1.38	$0.1070^2 * 1.38 = 0.0157$

2. 키워드 추출 개선

| 정보성 영상 | ④ 상위 8개 키워드 추출



금강산, 하늘, 영상, 일제,
조선, 독립, 여성, 시대



칸트, 명제, 사물, 이성,
인간, 공간, 비판, 인식

2. 키워드 추출 개선

오락성 영상 | 고유명사 제거

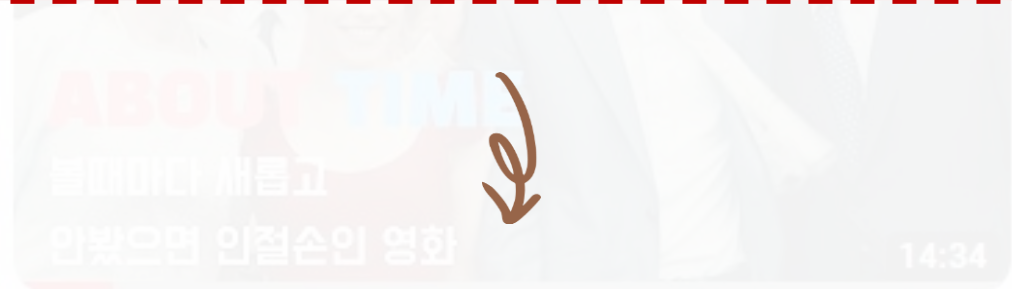
EXAMPLE) [결말포함] 시간여행자가 알려주는 인생의 교훈 '어바웃타임'

Mecab 키워드 추출 결과

팀 (5.56)	마음 (1.50)
시간 (3.76)	여행 (1.27)
아버지 (2.40)	인사 (1.27)
동생 (2.18)	결국 (1.27)
샬롯 (1.72)	과거 (1.27)
행복 (1.72)	날 (1.27)
친구 (1.50)	힘 (1.27)



영화 주인공 이름과 같은 고유명사들이 다수 존재
→ 이 경우 mecab에서 NNP(고유명사)로 분류



NNP(고유명사)로 분류된 키워드 제거

2. 키워드 추출 개선

오락성 영상 | 상위 8개 키워드 추출



DA CINEMA [결말포함] 시간여행자가 알려주는 인생의 교훈 :
'어바웃타임'
다시네마 · 조회수 41만회 · 4년 전



팀, 시간, 아버지, 동생,
행복, 친구, 마음, 여행



movie Puree 세계 제일 지적이고 환상적인 영화를 찾는다면 꼭 :
봐야 할 명작 영화 [SF 영화 추천] [인셉션] [...]
무비퓨레, moviepuree · 조회수 18만회 · 3년 전



꿈, 단계, 자신, 의식,
아리아드네, 사이트, 임무, 마음

3. 토픽 모델링

| Labeled LDA (L-LDA)

L-LDA

LDA의 대표적인 지도학습 버전!

여러 개의 **주제가 label된** 문헌들의 집합을 분석 → 주제별 단어 분포를 학습

⋮

새로운 문헌이 입력되어도,
해당 문헌에 **어떤 주제가 얼마나 포함**되어 있는지 계산 가능!



텍스트 자동 분류나 자동 태깅, 키워드 추출 등에 유용



3. 토픽 모델링

| Labeled LDA (L-LDA)

우리의 목적!

유튜브 키워드가 들어왔을 때, 그 키워드가 어느 KDC분류에 대한 정보를 갖고 있는지 분류해주는 것

⋮

L-LDA

주제분포를 모르는 새로운 문서가 입력되어도,
토픽을 이용해 다시 따로 매칭해주는 과정 없이
해당 문서에 어떤 주제가 얼마만큼 속해 있는지를 효율적이고 정확하게 생성 가능

3. 토픽 모델링

| Labeled LDA (L-LDA) | 주제별 단어 분포

철학 (100)

철학, 삶, 자신, 마음, 인간, 심리학, 인생, 감정, 말, 생각 ...

기술과학 (500)

치료, 음식, 기술, 의학, 요리, 레시피, 만들, 기술, 설명 ...

종교 (200)

하나님, 삶, 성경, 종교, 불교, 기독교, 사람, 사랑, 신앙 ...

예술 (600)

미술, 예술, 영화, 디자인, 사진, 이야기, 작가, 이해, 세계 ...

사회과학 (300)

사회, 경제, 기업, 투자, 교육, 정치, 분석, 제시, 변화 ...

문학 (800)

소설, 사랑, 이야기, 시인, 아이, 마음, 삶, 친구, 엄마 ...

자연과학 (400)

과학, 수학, 우주, 지구, 자연, 식물, 생명, 연구, 설명, 알 ...

역사 (900)

역사, 조선, 문화, 나라, 일본, 전쟁, 유럽, 여행, 이야기 ...

4. 분포 업데이트 및 최종 분포 형성

| Motivation

클러스터 분포를 반영하지 않고 영상이 하나만 입력되었을 때

⋮



영상이 갖고 있는 정보 하나에 의해서만 분포가 결정되는데,
그것이 실제 사용자의 선호도 분포라고 단정할 수 없음

4. 분포 업데이트 및 최종 분포 형성

| Bayesian 통계 분석

베이즈 패러다임

경험에 기반한 정보를 기반으로 하여, 추가 정보(관측)를 바탕으로 확률을 갱신

유튜브 기록으로부터 얻은 분포

$$P(\theta|y) = \frac{P(y|\theta) \cdot P(\theta)}{P(y)}$$

클러스터 분포

최종 선호도 분포

4. 분포 업데이트 및 최종 분포 형성

| Bayesian 통계 분석

켈레 사전분포(Conjugate prior)

사후 확률이 사전 확률 분포와 **같은 분포 계열**에 속하는 경우

켈레 사전분포를 이용하면 사전 분포의 parameter를 업데이트하는 방식으로
사후 확률을 계산할 수 있게 되어 계산이 간편해짐



클러스터 분포와 선호도 분포 모두 **디리클레 분포**를 따르고
디리클레 분포는 디리클레 분포에 대한 Conjugate Prior이므로,
효과적으로 분포를 업데이트할 수 있음!

4. 분포 업데이트 및 최종 분포 형성

계층적 모형과 경험적 베이즈

계층적 모형

사전 분포 역시 다른 모수(hyperparameter)에 의존하는 모형

$$y_1, \dots, y_n \mid \theta_1, \dots, \theta_n, \lambda \sim P(y_i \mid \theta_i, \lambda)$$

$$\theta_1, \dots, \theta_n \mid \lambda \sim P(\theta_i \mid \lambda)$$

Hyperparameter

데이터에 의해 사전 분포가 결정됨
hyperparameter는 보통 MLE로 추정
(Maximum Likelihood Estimator)

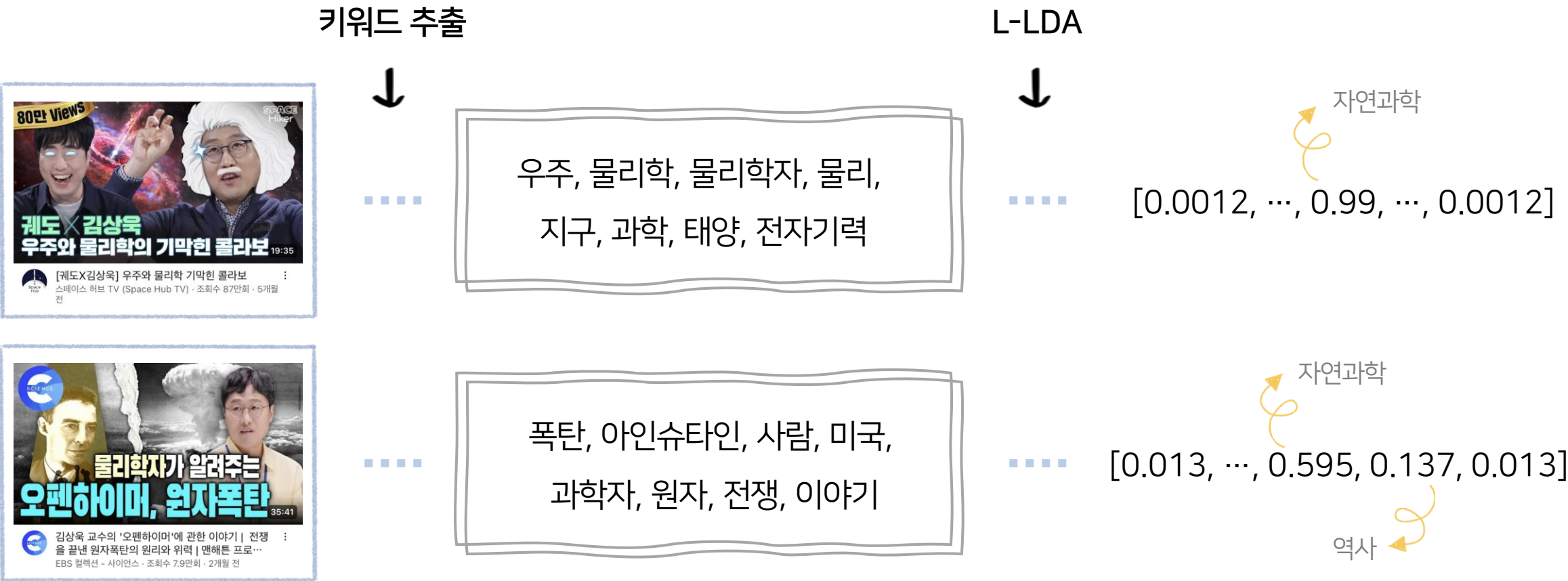
경험적 베이즈(EB)

관측치에 근거해 초모수 λ 를 추정한 후
이를 대입해 추론에 사용하는 방법

4. 분포 업데이트 및 최종 분포 형성

| 최종 분포 생성 과정

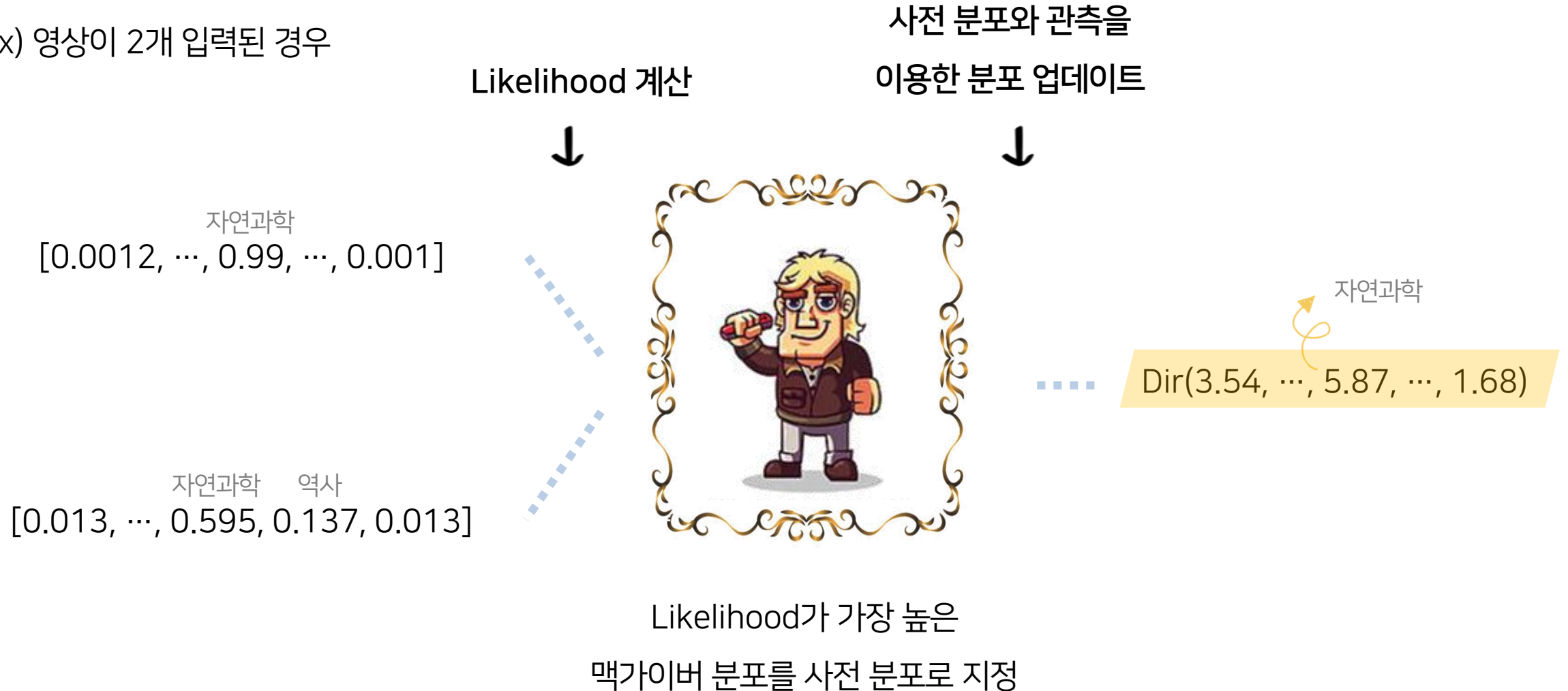
Ex) 영상이 2개 입력된 경우



4. 분포 업데이트 및 최종 분포 형성

| 최종 분포 생성 과정

Ex) 영상이 2개 입력된 경우



4. 분포 업데이트 및 최종 분포 형성

| 최종 분포 생성 과정

Ex) 영상이 2개 입력된 경우

선호도 분포 생성



자연과학

Dir(3.54, ..., 5.87, ..., 1.68)



[0.2, ..., 0.36, 0.02]

사회 이슈 키워드로부터
L-LDA 분포 생성

[0.001, 0.257, ..., 0.592, 0.001]

종교

역사



가중합(선호도 0.7, 이슈 0.3)

최종 분포 및 샘플링 진행

[0.14, ..., 0.28, 0.28]



5. 추천시스템

1. 추천 대상 도서 선택

Dialog

유튜브로 알아보는 나의 독서 DNA

추천 대상 도서 선택

서울도서관 인기 도서

이달의 신간

성균관대학교 핫북



입력 완료

- 1) 서울도서관 인기 도서
 - 2) 이달의 신간
 - 3) 성균관대학교 핫북
- 중 선택 가능

5. 추천시스템

| 2. 유튜브 링크 입력

Form

유튜브 링크를 입력해주세요

링크 추가

입력 완료

각 영상들로부터 키워드를 추출하고

L-LDA를 이용한 1차 분포 형성

⋮

EB를 이용한 분포 업데이트 및

사회 이슈 분포와 결합해 최종 분포 형성

5. 추천시스템

3. 책 추천

Form

당신에게 맞는 책 추천



이슬람과 유럽 문명의 종말

유해석 지음

ISBN: 9791196654689



불편한 편의점

김호연 지음

ISBN: 9791161571188

추천 종료

선택된 추천 대상 도서 데이터에서
최종 분포로부터 샘플링하여 책 추천 진행



감사합니다!

