DEVIEW 2021

로그 없이 영끌 키워드 추천: Few-shot Learning and Sequential Pattern-based Keyword Recommendation

김세호 / 최민수 / 강상욱 News&Topic Recommendation

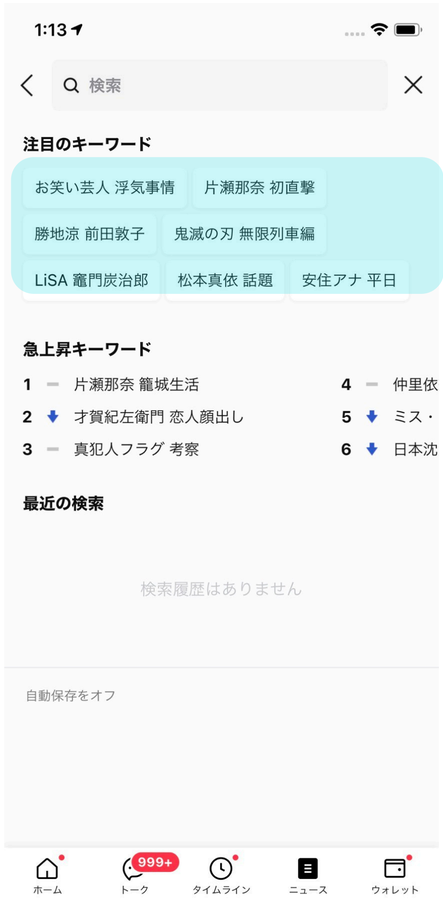
CONTENTS

1.키워드 추천 프로젝트 개요

1.1 라인 키워드 추천 개요

일본 라인 앱에 제공중인 다양한 키워드 추천

롤링키워드 추천 영역



추천 검색어 영역

연관 검색어 영역

1.1 라인 키워드 추천 개요

롤링키워드 추천 영역

- 상단 검색 창에 키워드가 자동으로 회전되며 추천 (이하, 롤링키워드)

- 주요 이슈를 검색 키워드로 노출

- 키워드 클릭 시 검색 결과로 이동 후 콘텐츠 소비 유도

click

1.2 2020 AiRS AB Test

AiRS AB Test 요약

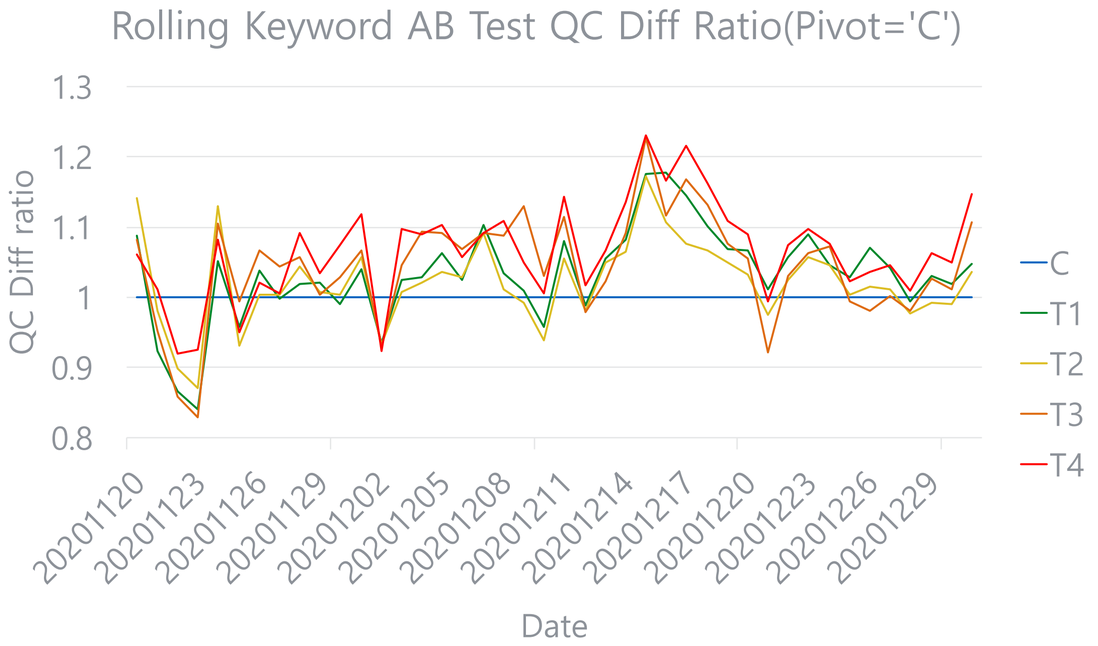
- 목표 : 롤링 키워드 개인화 랭킹을 통한 라인 검색 유도
- Control Group : 수동 편집된 키워드 리스트와 랭킹
- Test Group : 수동 편집된 키워드 리스트와 **개인화 랭킹**

검색 유도와 개인화 랭킹을 위한 노력

- 다양한 뉴스/검색 컨텍스트 사용
- 개인화 feature 사용

AB Test를 통해 가능성 확인!

- Query Count(이하, QC) 기준
  - Test group > Control group



11/20-12/31	평균 : QC	평균 : UU	평균 : SS	평균 : UQC
C	1.00	1.00	1.00	1.00
T1	1.03	1.03	1.04	1.11
T2	1.02	1.02	1.03	1.40
T3	1.04	1.03	1.04	4.00
T4	1.06	1.04	1.06	2.00

1.3 새로운 문제들

다양한 검색 니즈를 충족시키고 싶지만..

- 모든 키워드가 수동으로 생산(=인간지능)
- 뉴스 검색 키워드 위주로 소비
- NAVER에 비해 적은 규모의 검색 로그
- **검색 결과로 줄 수 있는 건 참 많은데... 노출할 방법을 찾고 싶다!**

검색 로그 없이 키워드를 만들어보자!

1.4 2021 AiRS TODO

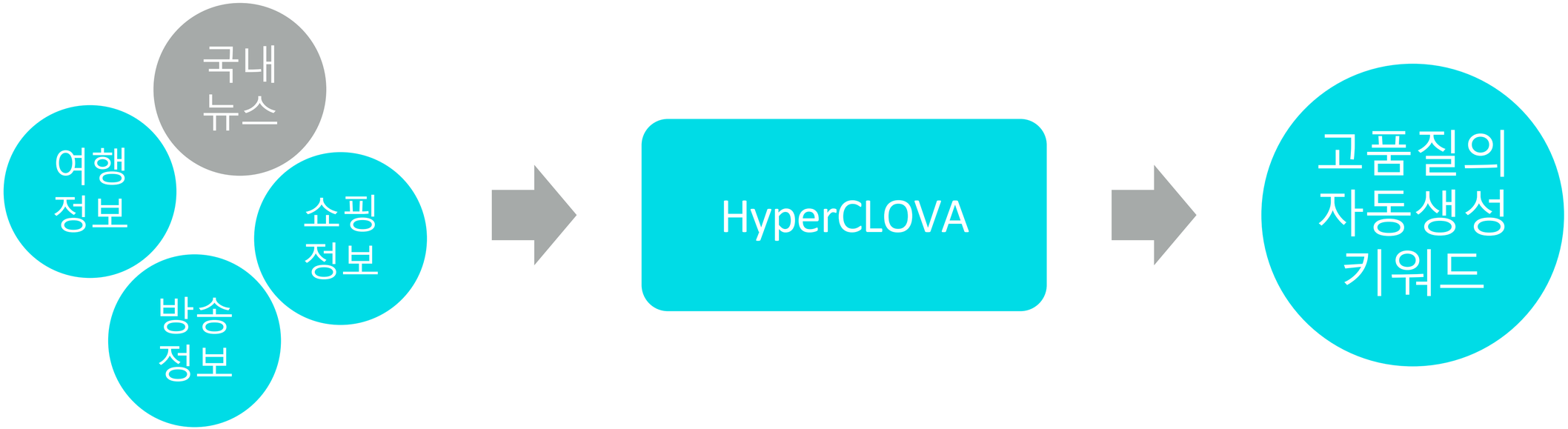
새로운 검색 키워드 만들기

- 1. 24시간 수동 관리 없이
- 2. 다양한 관심사에 대해서
- 3. 검색을 유도할 만한 고품질의 키워드를 생산

1.5 키워드 생성을 위한 모델 개요

HyperCLOVA를 이용한 키워드 생산

- 다양한 분야의 콘텐츠를 Hyper-scale Model의 입력으로 세팅
- **고품질의 키워드 생성** 후 사용자에게 노출



1.5 키워드 생성을 위한 모델 개요

DEVIEW 2021

큰 모델만이 정답은 아니다.

- Scalability 문제(Hyper-scale Model)
- 키워드 품질 문제

개인화 랭킹만이 정답은 아니다.

- 영역별로 서로 다른 사용성(개인화 vs 비개인화)

=> 상황과 목적에 맞는 **키워드 생성 모델**/랭킹 모델 사용하기

2. HyperCLOVA를 활용한 이슈 키워드 자동 생성

2.1 수동 편집 키워드를 자동화 해보자!

DEVIEW 2021

수동 편집 키워드는 어떻게 생성되는가?

- 실시간으로 주요 이슈를 파악 → 적절한 키워드 선정 → 적절한 Rank에 삽입

“적절한”(=인간지능)을 어떻게 자동화 할 것인가?

2.2 실시간 이슈 감지 모델

DEVIEW 2021

Latest Popular Model[1]

- 컨텐츠 소비자 관점에서, 사용자들이 최근에 많은 관심을 가지는 컨텐츠를 클릭 로그를 사용하여 탐지

Cluster Model[1]

- 컨텐츠 생성자 관점에서, 다수의 컨텐츠 제공자들이 공통적으로 발행하는 이슈를 포함하는 컨텐츠를 탐지

Future Impact Model[2]

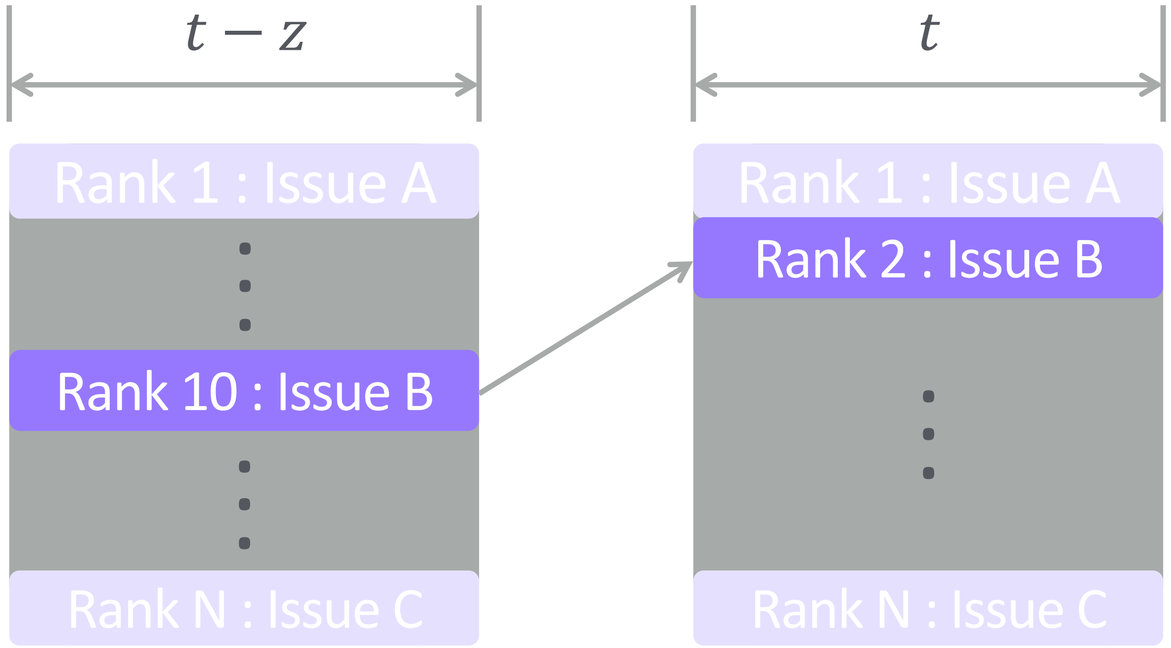
- 현재 컨텐츠에 대한 클릭을 기반으로, 미래의 클릭 수를 예측해서 이슈를 탐지

[1] https://blog.naver.com/naver_search/222439504418
[2] Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). Optimizing the recency-relevancy trade-off in online news recommendations. *26th International World Wide Web Conference, WWW 2017*, *i*, 837–846.

2.2.1 Latest Popular Model

사용자들이 최근에 많은 관심을 가지는 컨텐츠 탐지

- 단순히 클릭이 많은 컨텐츠를 사용하면 실시간 발생하는 이슈 탐지가 어려움
- 클릭 절대량이 높으면서 절대량의 랭크 변화가 큰 컨텐츠를 감지



2.2.1 Latest Popular Model

- 현재 시점 t 기준 최근 n분 동안의 컨텐츠 v_j의 클릭 수 산출
- 산출된 클릭 수를 기준으로 각 컨텐츠의 순위를 $r_j(t)$로 표현
- $H(t)$: $r_j(t)$ 기준 t시점의 가장 인기 있는 상위 k개의 컨텐츠 set
- $H(t)$ 의 각 컨텐츠 v_j에 대한 최신 인기(latest popular) 점수 $lp_j(t)$를 계산

$$lp_j(t) = \alpha \cdot pop_j(t) + \beta \cdot raise_j(t),$$

$$pop_j(t) = 1 - \frac{rank\left(avg\left(r_j(t-n), r_j(t)\right), H(t-n,t)\right)}{|H(t-n,\, t)|},$$

$H(t-n)$와 $H(t)$의 교집합

$$raise_j(t) = 1 - \frac{rank\left(r_{j(t-n)} - r_{j(t)},\, H(t-n,t)\right)}{|H(t-n,\, t)|},$$

https://blog.naver.com/naver_search/222439504418

2.2.1 Latest Popular Model

- 현재 시점 t 기준 최근 n분 동안의 컨텐츠 v_j의 클릭 수 산출
- 산출된 클릭 수를 기준으로 각 컨텐츠의 순위를 $r_j(t)$로 표현
- $H(t)$: $r_j(t)$ 기준 t시점의 가장 인기 있는 상위 k개의 컨텐츠 set
- $H(t)$ 의 각 컨텐츠 v_j에 대한 최신 인기(latest popular) 점수 $lp_j(t)$를 계산

$$lp_j(t) = \alpha \cdot pop_j(t) + \beta \cdot raise_j(t),$$

$$pop_j(t) = 1 - \frac{rank\left(avg\left(r_j(t-n), r_j(t)\right), H(t-n, t)\right)}{|H(t-n,\, t)|},$$

$t-n$시점과 t시점의 v_j 순위 평균값

$$raise_j(t) = 1 - \frac{rank\left(r_{j(t-n)} - r_{j(t)},\, H(t-n, t)\right)}{|H(t-n,\, t)|},$$

https://blog.naver.com/naver_search/222439504418

2.2.1 Latest Popular Model

- 현재 시점 t 기준 최근 n분 동안의 컨텐츠 v_j의 클릭 수 산출
- 산출된 클릭 수를 기준으로 각 컨텐츠의 순위를 $r_j(t)$로 표현
- $H(t)$: $r_j(t)$ 기준 t시점의 가장 인기 있는 상위 k개의 컨텐츠 set
- $H(t)$ 의 각 컨텐츠 v_j에 대한 최신 인기(latest popular) 점수 $lp_j(t)$를 계산

$$lp_j(t) = \alpha \cdot pop_j(t) + \beta \cdot raise_j(t),$$

$$pop_j(t) = 1 - \frac{rank\left(avg\left(r_j(t-n), r_j(t)\right), H(t-n, t)\right)}{|H(t-n, t)|},$$

$H(t-n, t)$내에서
$avg\left(r_j(t-n), r_j(t)\right)$
값의 순위

$$raise_j(t) = 1 - \frac{rank\left(r_{j(t-n)} - r_{j(t)}, H(t-n, t)\right)}{|H(t-n, t)|},$$

https://blog.naver.com/naver_search/222439504418

2.2.1 Latest Popular Model

- 현재 시점 t 기준 최근 n분 동안의 컨텐츠 v_j의 클릭 수 산출
- 산출된 클릭 수를 기준으로 각 컨텐츠의 순위를 $r_j(t)$로 표현
- $H(t)$: $r_j(t)$ 기준 t시점의 가장 인기 있는 상위 k개의 컨텐츠 set
- $H(t)$ 의 각 컨텐츠 v_j에 대한 최신 인기(latest popular) 점수 $lp_j(t)$를 계산

$$lp_j(t) = \alpha \cdot pop_j(t) + \beta \cdot raise_j(t),$$

$$pop_j(t) = 1 - \frac{rank\left(avg\left(r_j(t-n), r_j(t)\right), H(t-n,t)\right)}{|H(t-n,\, t)|},$$

$$raise_j(t) = 1 - \frac{rank\left(r_{j(t-n)} - r_{j(t)}, H(t-n,t)\right)}{|H(t-n,\, t)|},$$

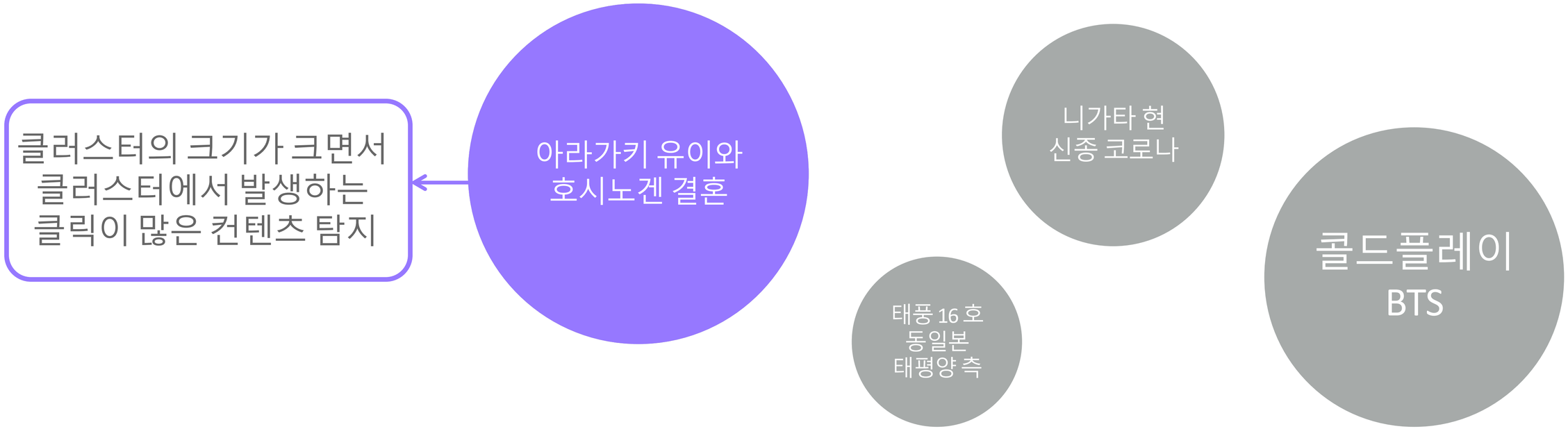
$t-n$시점과 t시점의 v_j 순위 차이값

https://blog.naver.com/naver_search/222439504418

2.2.2 Cluster-based Model

다수의 컨텐츠 생성자가 공통적으로 생성하는 컨텐츠 탐지

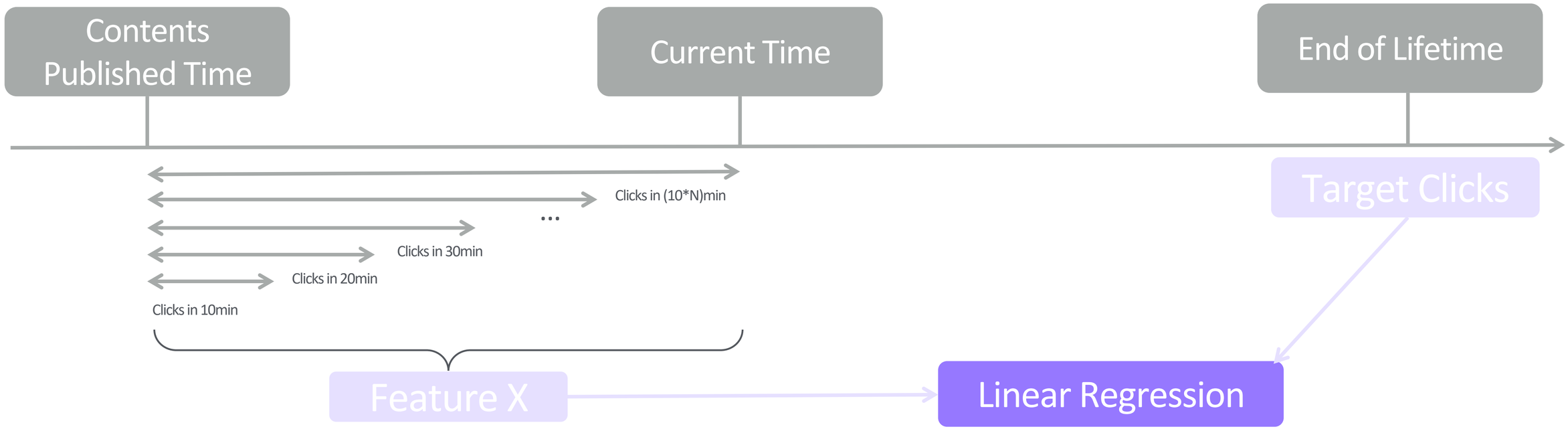
- 동일한 이슈를 다루는 컨텐츠를 클러스터링 기법으로 분류
- 컨텐츠의 제목과 본문을 이용하여 TF-IDF vector 생성
- Vector간의 유사도를 기반으로 Hierarchical clustering 수행



2.2.3 Future Impact Model

현재 클릭 수를 기반으로 미래의 클릭 수 예측

- 컨텐츠 생성 시점 이후 매 $(10 * N)$분 동안의 클릭 수를 조합해서 feature로 사용
- 컨텐츠 lifetime 동안의 전체 click 수를 예측



Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). Optimizing the recency-relevancy trade-off in online news recommendations. *26th International World Wide Web Conference, WWW 2017*, *i*, 837–846.

이슈 컨텐츠로부터
어떻게 키워드를 생성할 것인가?

2.3 키워드 생성 모델

DEVIEW 2021

HyperCLOVA란?

- 네이버가 국내 기업 최초로 자체 개발한 초대규모 AI
- OpenAI의 GPT-3(175B)를 뛰어넘는 204B(2,040억 개) 파라미터 규모로 개발
- GPT-3보다 한국어 데이터를 6,500배 이상 학습
- 현재 전세계에서 가장 큰 한국어 초거대 언어모델
- HyperCLOVA의 일본어 모델 중 6.7B 파라미터 크기의 모델 사용

https://www.navercorp.com/promotion/pressReleasesView/30546

2.4 HyperCLOVA

Zero-shot Learning vs. Few-shot Learning

- Zero-shot learning은 컨텐츠의 제목만을 입력으로 사용해서 결과 생성
- 사전 학습된 언어모델을 원하는 기능에 맞게 재 학습 필요

2.5 Few-shot Learning

DEVIEW 2021

모델에 여러 개의 예시를 제공하여 결과 도출

- 컨텐츠의 제목과 키워드를 쌍으로 하는 예시 사용
- 모델의 가중치 업데이트는 하지 않음

Few-shot Learning Example

本文:阪神·俊介「感謝の言葉しか…」引退会見に原口、新井コーチがサプライズで駆けつけ涙あふれる\n検索語:阪神·俊介 引退会見
本文:アンミカ、夫の不正受給疑惑がスルーされる事情\n検索語:アンミカ 夫
本文:田中圭、中谷美紀との夫婦役に「ハラハラ」\n検索語:田中圭 中谷美紀
本文:中川翔子、婚活アプリ初体験「こんなハイスペックみたいな人がいるの!?」\n検索語:中川翔子 婚活アプリ初体験
本文:小倉優子、一部週刊誌記事は「作り話」と否定　「このような事が許されない時代になって」\n検索語:小倉優子 一部週刊誌記事否定
本文:「今日好き」“ふたみら”酒寄楓太&横田未来カップル、破局を報告\n検索語:酒寄楓太&横田未来カップル 破局報告
本文:コロナ感染·A.B.C-Z橋本良亮、体調回復で活動再開へ\n検索語: 橋本良亮 活動再開へ
本文:渡辺美奈代、家族４人分の豪華ローストビーフ弁当を公開「盛りつけと彩りの天才」「感動です」\n検索語:渡辺美奈代 ローストビーフ弁当
本文:「元日本留学生、脅迫受け逃げ回っている」アフガンでの窮状訴え\n検索語:元日本留学生 窮状
本文:SNSでデザインが「ダサい」と不評も…新1万円札に込められた配慮\n検索語:新1万円札 デザイン
本文:【巨人記録室】危険球で退場、山口俊が最多４度…桑田真澄、浅尾拓也、内海哲也を上回る\n検索語:巨人 危険球で退場
本文:PCR検査「陽性」を「陰性」と誤通知 保健所職員が検査結果を見誤る 愛知県\n検索語:PCR検査結果 誤通知
本文:しまパト大成功！【しまむら】着回しできる「アウター」買ってみた！\n検索語:しまむら アウター

本文:トレエン斎藤司、カラコン装着姿を公開「煉獄さんみたい」「オッドアイになっとる」\n検索語: トレエン斎藤 カラコン

2.6 자동 생성 품질 고도화 이슈

DEVIEW 2021

키워드 자동 생성에서 발생하는 문제점들

- 정보 왜곡(부정확한 키워드 생성)

薬丸裕英、コロナ感染体験語る
野々村真に「ちょっと震えが来るほどですよね」
야쿠마루 히로히데, 코로나 감염 경험을 이야기하는
노노무라 마코토에게 "약간 몸이 떨릴 정도죠”

薬丸裕英 コロナ感染
야쿠 마루 히로히데
코로나 감염

- 정보량이 적은 키워드 생성

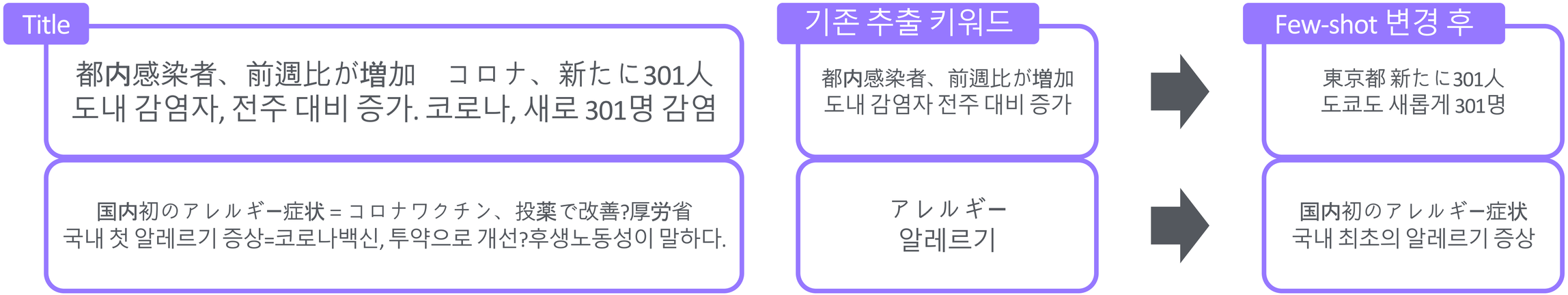
焼きたてサンマが１匹１００円！ コロナ禍に活力を
갓 구운 꽁치가 1마리 100엔! 코로나 재난에 활력을

サンマ
꽁치

2.7 Few-shot Example Tuning

컨텐츠의 카테고리별 Few-shot Example을 다르게 사용

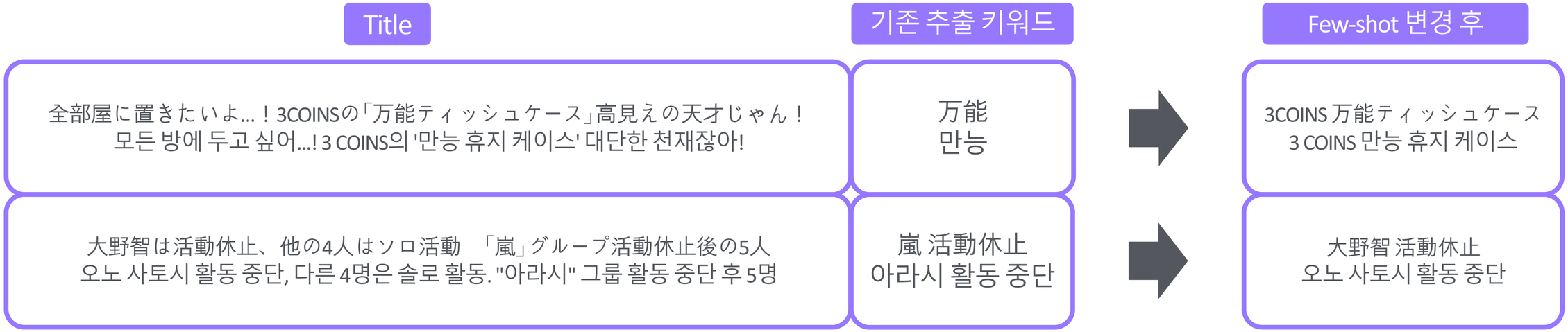
- 가장 간단하고 빠르게 tuning 가능
- 카테고리 별 제목의 형식이나 자주 사용되는 단어들이 다름(e.g. 코로나 → 백신)
- Few-shot example을 코로나 관련 컨텐츠로 구성한 후의 결과 변화



2.7 Few-shot Example Tuning

사람이 수동 편집하는 키워드에 더 가깝게! 더 많이!

- 수동 편집 키워드는 좀 더 길고 정보량이 많은 경향
- 수동 편집 키워드와 유사하게 few-shot example을 구성 + 개수 증가

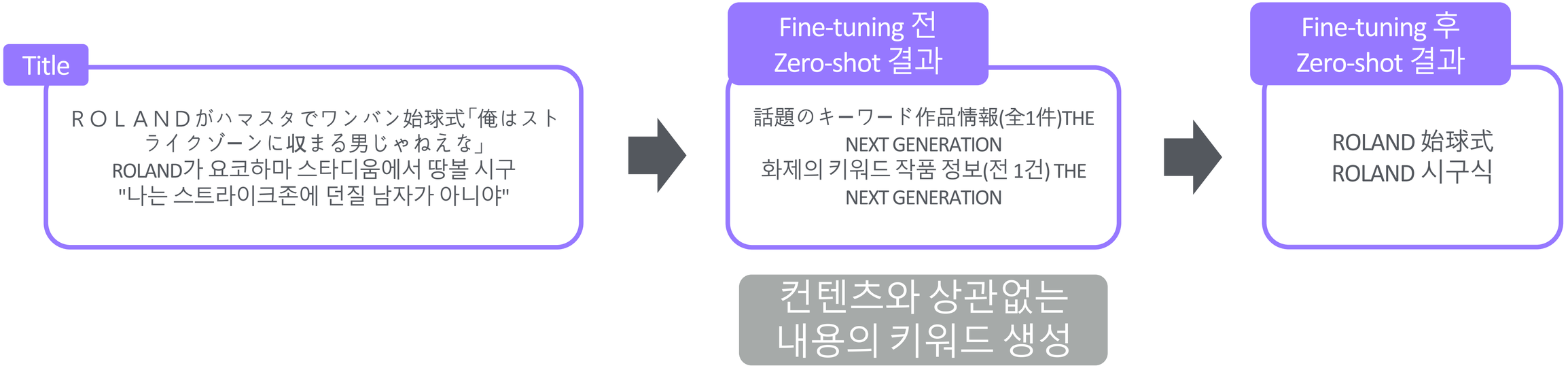


- Input size가 커질 수록 필요한 GPU memory 증가 → 가성비 문제

2.8 Fine-tuning

Few-shot example을 사용하지 않고 더 좋은 성능을 내려면?

- 학습 데이터: 컨텐츠의 타이틀,키워드 쌍 약 15,000건
- 학습 대상 모델: 사전 학습된 6.7B 파라미터 크기의 일본어 모델
- 결과: Zero-shot Learning으로도 기존과 유사한 수준으로 동작



2.9 Evaluation Result

정성평가 결과

- 1,077개의 Test Set 사용
- 원어민 사용자가 직접 Title과 Keyword를 비교해서 평가
- 약 81%의 키워드는 바로 사용 가능

2.10 키워드 자동 생성의 한계

DEVIEW 2021

키워드 생성 비용

- 대용량 생성에 많은 GPU 리소스 필요
- 실시간 이슈 컨텐츠에서 키워드를 추출하는 용도로는 적합

자동 생성 키워드 품질 이슈

- 약 20%의 품질이 낮은 키워드를 필터링하는 로직 필요
- 혹은 생성된 키워드의 정보량을 측정할 수 있는 새로운 모델 추가 필요

3. 검색 다양화를 위한 키워드 확장과 개인화 추천

3.1 개요

추천 키워드를 다양하게 만들어보자!

- **다양한 키워드 검색**으로, 라인 앱에서 **풍부한 경험**을 할 수 있도록 유도
- 유저의 관심 키워드와 **연관된 단어들**을 찾아보자

  예) BTS + 신곡 / 멤버 / 댄스 영상 / ...

3.2 구상

컨텐츠 A

쉴 틈이 없어요 쭉쭉쭉~방탄소년단, 유엔총회→콜드플레이 콜라보(종합)[...

어나더 월드 클래스 **BTS**이기에 가능한 일이다. 방탄소년단은 24일 오전 문재인 대통령의 특별 사절(특사)... 이날 오후 콜드플레이와 **콜라보**레이션 한 곡 ‘My Universe’가 공개되자 공식 ...

컨텐츠 B

BTS·콜드플레이, 콜라보 음원 발매..."My Universe, 한국어·영어 가사"

방탄소년단과 영국 밴드 콜드플레이가 역대급 **콜라보**레이션을 공개했다. 방탄소년단과 콜드플레이가 24일 오후 1시 **콜라보**레이션 싱글 ‘마이 유니버스’(My Universe)를 전 세계에 발표...

컨텐츠 C

BTS·콜드플레이 콜라보 곡 '마이 유니버스' 오늘 공개

그룹 방탄소년단(**BTS**)이 세계적 브릿팝 밴드 콜드플레이(Coldplay)와의 컬래버레이션 곡 '마이 유니버스'를... 부분을 **BTS** 멤버들이 불렀습니다. **BTS**는 한국어와 영어 가사를 통해 우주...

컨텐츠 D

베일 벗은 콜드플레이·BTS의 콜라보 '마이 유니버스'… 한국어 가사도 귀...

영국의 세계적 록 밴드 콜드플레이(Coldplay)가 현재 최고 인기를 달리고 있는 그룹 방탄소년단(**BTS**)과 함께... 콜드플레이 멤버들과 **BTS**의 RM, 슈가, 제이홉이 작사·작곡자에 이름을 ...

Sequential Patterns

...

(BTS, 콜드플레이)

(BTS, 콜라보)

(BTS, My Universe)

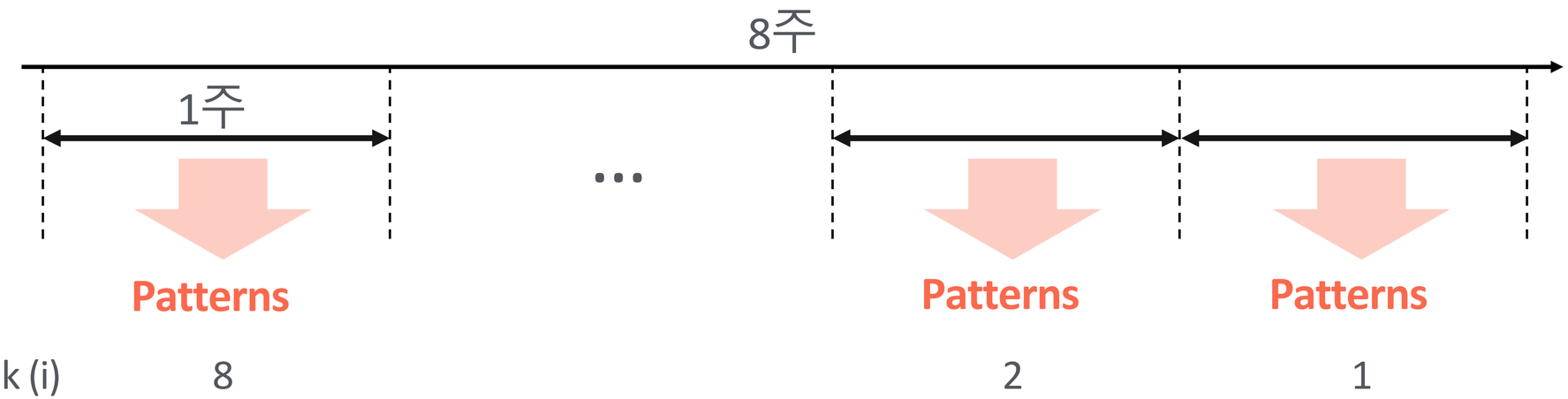
(BTS, 콜드플레이, 콜라보)

(콜드플레이, My Universe)

...

3.3 Sequential Pattern Mining

- 너무 시의성 짙은 화젯거리는 지양함 → HyperCLOVA 모델에서 수행
- 어느 정도 steady한 패턴을 찾으려 함
- 패턴의 trendy한 정도를 고려하려 함



각 패턴에 대한 점수 계산

$$\sum_{i=1}^{8} \frac{support}{\log_2(i+1)}$$

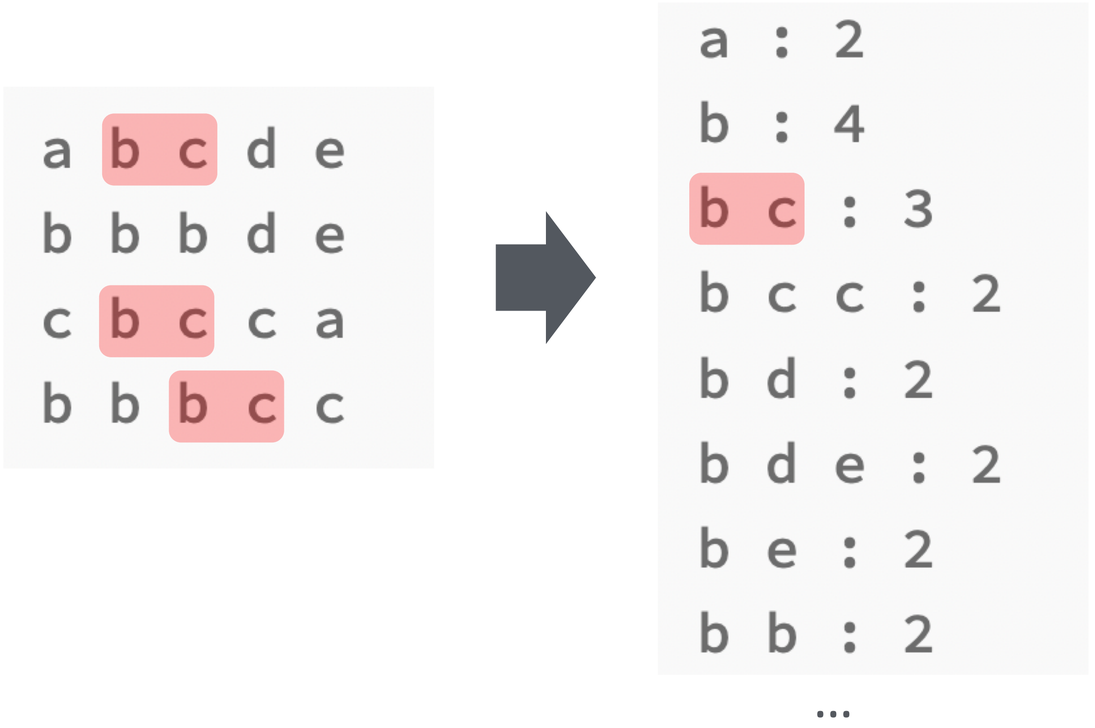
여러 구간에서 뽑힐수록,
각 구간에서의 **support (출현 빈도)값이 클수록**,
최근에 뽑힐수록 값이 커짐

3.3 Sequential Pattern Mining

DEVIEW 2021

첫 시도

- PrefixSpan 알고리즘 사용
- 컨텐츠 제목으로부터 토큰화 수행



Mining sequential patterns by pattern-growth: The prefixspan approach, Pei Jian et al., TKDE, 2004

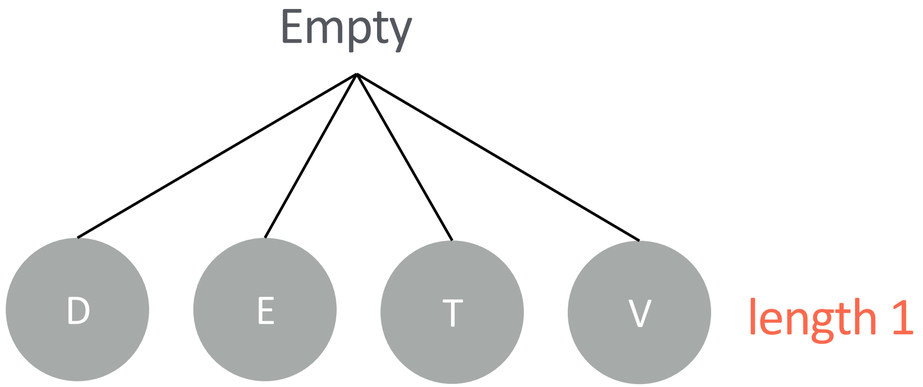
3.3 Sequential Pattern Mining

DEVIEW 2021

PrefixSpan

- Pattern Growth 방식

Seq. ID	Sequence
1	C, D, E, T, V, E
2	D, S, P, E, A, H, T, V



3.3 Sequential Pattern Mining

DEVIEW 2021

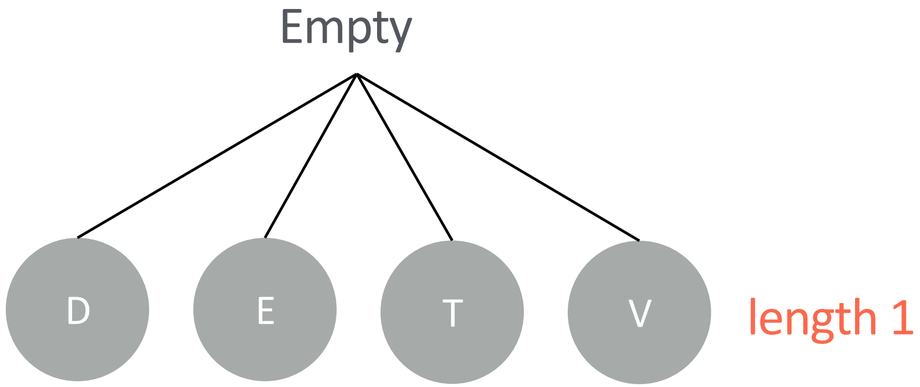
PrefixSpan

- Pattern Growth 방식

Seq. ID	Sequence
1	C, D, E, T, V, E
2	D, S, P, E, A, H, T, V

<D>-projected DB

Seq. ID	Sequence
1	E, T, V, E
2	S, P, E, A, H, T, V



3.3 Sequential Pattern Mining

DEVIEW 2021

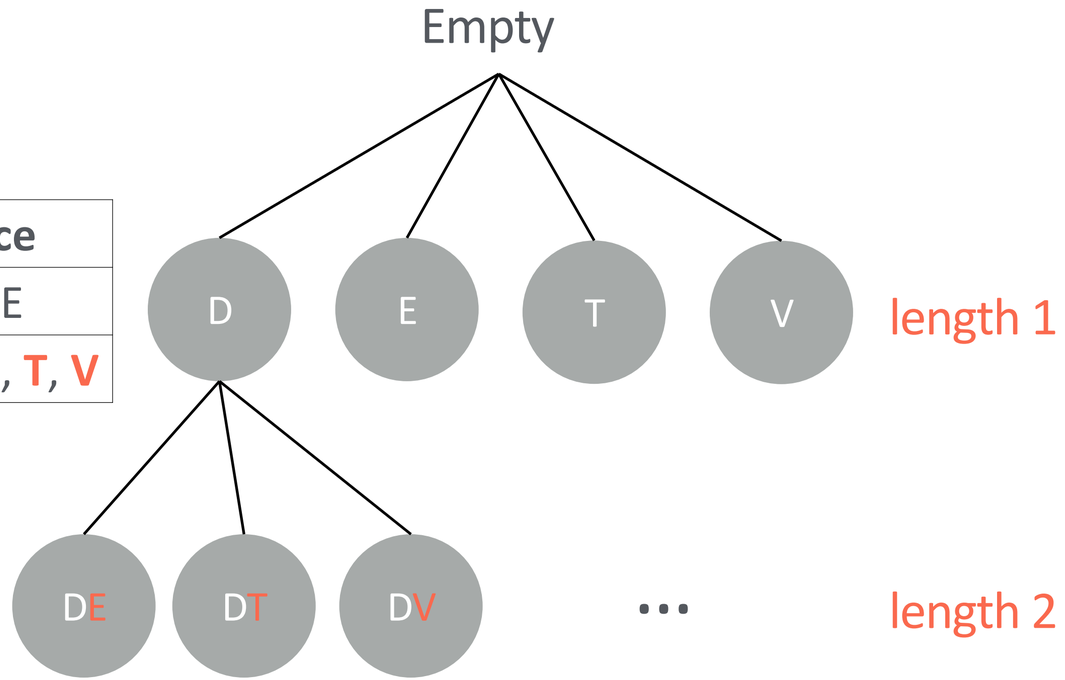
PrefixSpan

- Pattern Growth 방식

Seq. ID	Sequence
1	C, D, E, T, V, E
2	D, S, P, E, A, H, T, V

<D>-projected DB

Seq. ID	Sequence
1	**E**, **T**, **V**, E
2	S, P, **E**, A, H, **T**, **V**



3.3 Sequential Pattern Mining

DEVIEW 2021

PrefixSpan

- Pattern Growth 방식

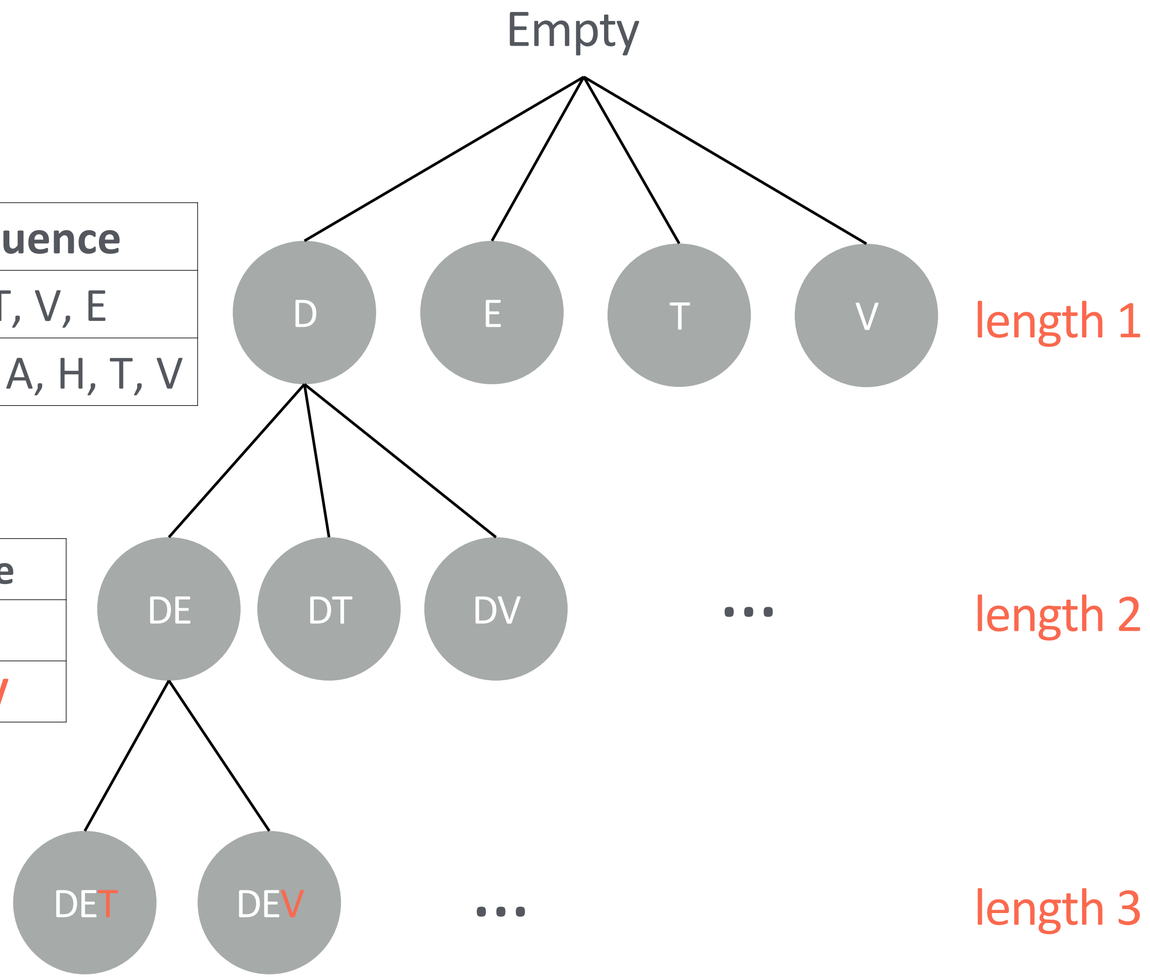
Seq. ID	Sequence
1	C, D, E, T, V, E
2	D, S, P, E, A, H, T, V

<D>-projected DB

Seq. ID	Sequence
1	E, T, V, E
2	S, P, E, A, H, T, V

<DE>-projected DB

Seq. ID	Sequence
1	**T, V**, E
2	A, H, **T**, **V**



3.4 부적절한 단어 제거

DEVIEW 2021

인물명 뒤에 부적절 단어들 붙을 시 제거

- 본 과정에서 걸러지는 패턴 예시

(예) 인물명 + **왕따** / **감염** / **불륜** 등

청소년 유해어나 부정적인 단어 포함시 제거

- 본 과정에서 걸러지는 패턴 예시

(예) 오사카 **하반신 노출** 등

반드시 걸러져야하는 단어 포함시 제거

- 본 과정에서 걸러지는 패턴 예시

(예) **자살** 미수

FYI

부적절 단어 패턴매칭

: aho corasick 알고리즘

(ahocorapy 패키지)

https://github.com/abusix/ahocorapy

3.5 발견패턴 결과

DEVIEW 2021

트와이스 관련 패턴

('twice', '公開'): 5.660
('twice', 'ジョンヨン'): 4.616
('twice', 'ナヨン'): 4.554
('twice', 'モモ'): 4.264
('twice', 'photo'): 4.058
('twice', 'ツウィ'): 2.709
('twice', 'ファン'): 2.684
('twice', 'サナ'): 2.609
('twice', '姿'): 2.381
('twice', 'ショット'): 2.231
('twice', '話題'): 2.178
('twice', 'jyp'): 2.085
('twice', 'アルバム'): 2.022
('twice', '出演'): 1.969

('t wice', '공개') : 5.660
('t wice', '정연') : 4.616
('t wice', '나연') : 4.554
('t wice', '모모') : 4.264
(' twice ' , ' photo ') : 4 . 058
('t wice', '쯔위'): 2.709
('t wice', '팬') : 2.684
('t wice', '사나') : 2.609
('t wice', '모습') : 2.381
('t wice', '샷') : 2.231
('t wice', '화제') : 2.178
(' twice ' , ' jyp ') : 2 . 085
('t wice', '앨범') : 2.022
('t wice', '출연') : 1.969

유니클로 관련 패턴

('ユニクロ', 'コーデ'): 22.017
('ユニクロ', 'アイテム'): 18.905
('ユニクロ', '新作'): 16.982
('ユニクロ', 'パンツ'): 16.690
('ユニクロ', 'tシャツ'): 16.029
('ユニクロ', '大人'): 15.613
('ユニクロ', 'コラボ'): 15.283
('ユニクロ', 'デニム'): 14.408
('ユニクロ', 'gu'): 13.011
('ユニクロ', '神'): 12.256
('ユニクロ', 't'): 11.818
('ユニクロ', 'トップス'): 11.290
('ユニクロ', '優秀'): 11.039
('ユニクロ', '夏'): 10.419

('유니클로', '코디') : 22.017
('유니클로', '아이템'): 18.905
('유니클로', '신상') : 16.982
('유니클로', '팬츠'): 16.690
('유니클로', 't셔츠'): 16.029
('유니클로', '어른') : 15.613
('유니클로', '콜라보') : 15.283
('유니클로', '데님'): 14.408
('유니클로', 'gu'): 13.011
('유니클로', '신'): 12.256
('유니클로', 't') : 11.818
('유니클로', '상의') : 11.290
('유니클로', '우수') : 11.039
('유니클로', '여름'):10.419

3.6 품질 개선 시도

다소 어색한 패턴이 발생하는 경우 존재했음 「TWICE」 영어 싱글 「The Feels」 뮤직비디오 티저 2탄 공개!

- 'Twice 공개' 등의 패턴은 일반적으로 컨텐츠 제목 상 멀리 떨어진 단어의 조합
- 유저들에게 궁금증을 유발할 수도 있으나, 다소 어색한 느낌을 줄 수도 있음
- 패턴 발견할 때, Window를 고려해보자!

	Window Size: 5	Window Size: 6	Window Size: 7	Window 고려 안함
1	(twice, 일본)	(twice, itzy)	(twice, itzy)	**(twice, 공개)**
2	(twice, 정연)	(twice, 정연)	(twice, 정연)	(twice, itzy)
3	(twice, 나연)	(twice, 나연)	(twice, 나연)	(twice, 정연)
4	(twice, 일본, 데뷔)	twice, 일본, 데뷔)	(twice, 일본, 데뷔)	(twice, 나연)
5	(twice, 동생뻘)	(twice, 동생뻘, itzy)	(twice, 동생뻘, itzy)	(twice, 멤버)
6	(twice, itzy, 일본)	(twice, 모모)	(twice, 모모)	(twice, 일본, 데뷔)
7	(twice, 모모)	(twice, perfectworld)	(twice, perfectworld)	(twice, 동생뻘, itzy)
8	(twice, perfectworld)	(twice, 쯔위)	(twice, 쯔위)	(twice, 모모)
9	(twice, 쯔위)	(twice, 일본)	(twice, 일본)	(twice, 주목)
10	(twice, 워너뮤직재팬, 일본)	(twice, 사나)	**(twice, 공개)**	(twice, thefeels)
11	(twice, 사나)	(twice, 앨범)	(twice, 사나)	(twice, 쯔위)
12	(twice, 앨범)	(twice, 다현)	(twice, 팬)	(twice, 팬)
13	(twice, 다현)	(twice, 팬)	(twice, 앨범)	(twice, perfectworld)
14	(twice, 트와이스)	(twice, 트와이스)	(twice, 다현)	(twice, 일본)
15	(twice, 신곡)	(twice, 신곡)	(twice, jyp)	(twice, 앨범)

Frequent Sequential Pattern을
키워드 개인화 추천에 어떻게 사용할까?

3.7 가지고 있는 데이터

DEVIEW 2021

Frequent Sequential Patterns

StarSpace Embedding Vectors

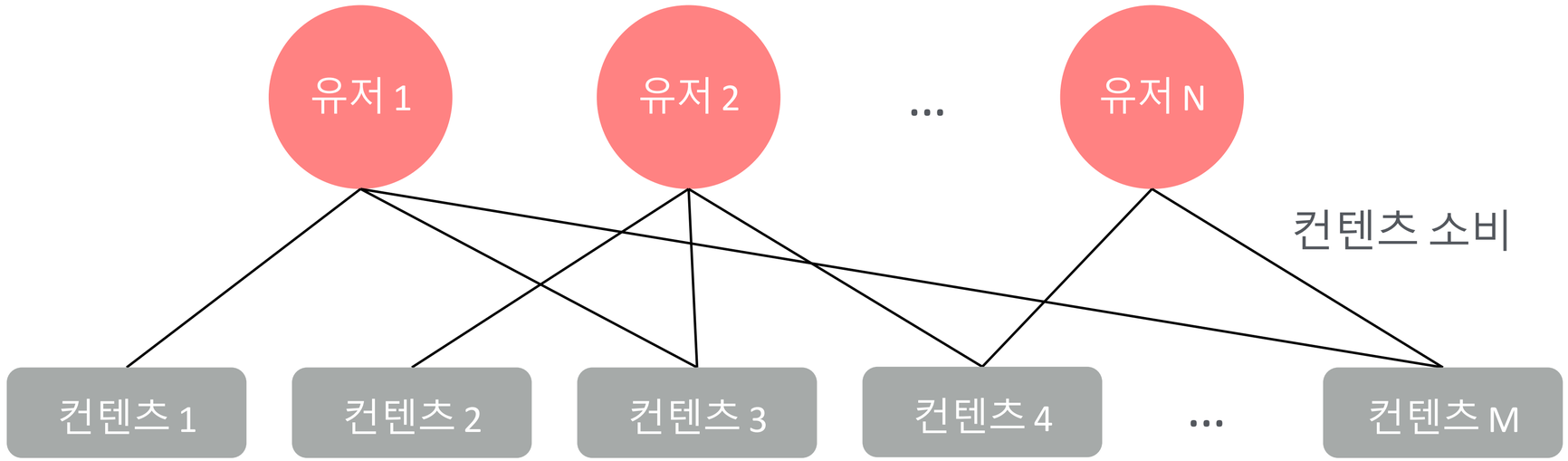
- User Embeddings
- Content Embeddings

StarSpace: Embed all the things!, Facebook Research, AAAI 2018

3.7 가지고 있는 데이터

StarSpace Embedding

- Entity 타입에 관계없이 (*) 같은 공간 (Space) 상에 매핑

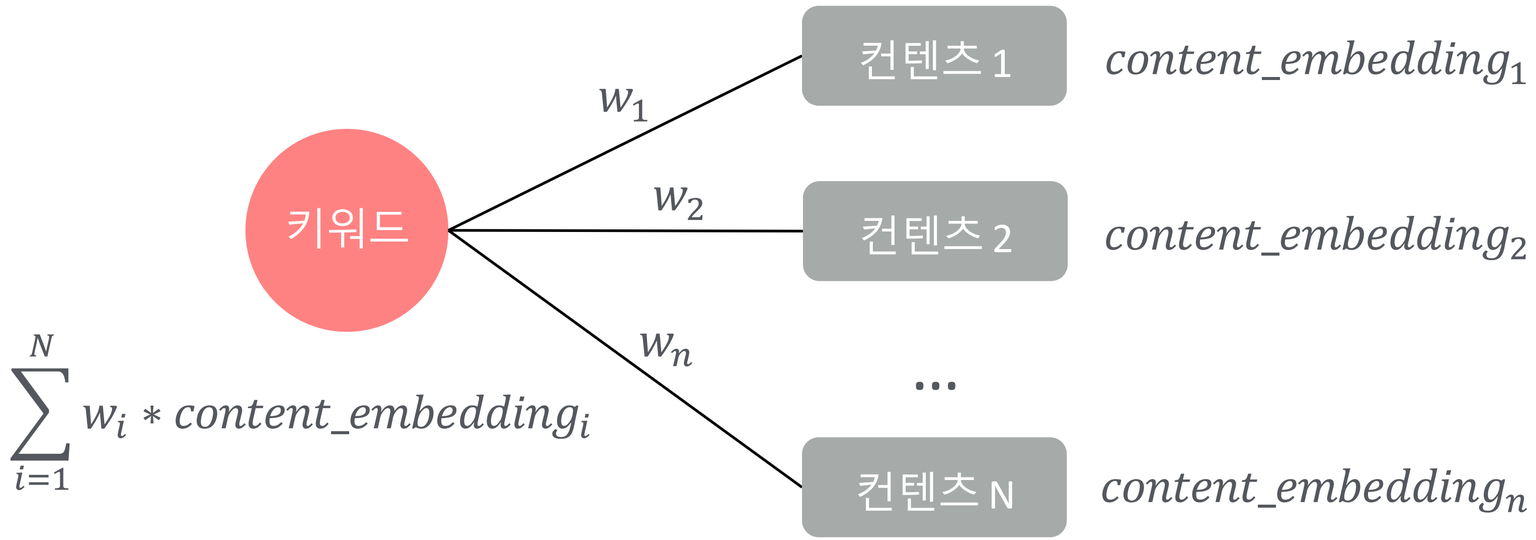


Loss Function $\sum_{\substack{(a,b)\in E^{+} \\ b^{-}\in E^{-}}} L^{batch}(sim(a,b), sim(a,b_1^{-}), \ldots, sim(a,b_k^{-}))$

3.8 키워드 임베딩

연관 컨텐츠의 임베딩으로부터 계산

- 각 컨텐츠와의 연관도 기반 Weighted Sum



3.8 키워드 임베딩

BTS 연관 키워드

BTS 멤버, 곡명

K-pop 그룹이나 연예인 키워드들

JUNG KOOK : 0.903029

SUGA : 0.867230

Butter : 0.820302

BLACKPINK : 0.695677

Wanna One : 0.666787

Red Velvet : 0.653172

IZ * ONE : 0.556477

미야와키 사쿠라 : 0.538116

권은비 : 0.517248

TWICE : 0.506854

리사: 0.493148

장원영: 0.430021

…

iPhone 연관 키워드

Apple 관련이나

전자제품 키워드들

Apple Watch : 0.630689

Google Pixel : 0.596355

헤드셋: 0.592960

Apple Music : 0.577275

스마트 스피커: 0.569488

KDDI : 0.566779

Apple TV : 0.554590

Microsoft Surface : 0.549453

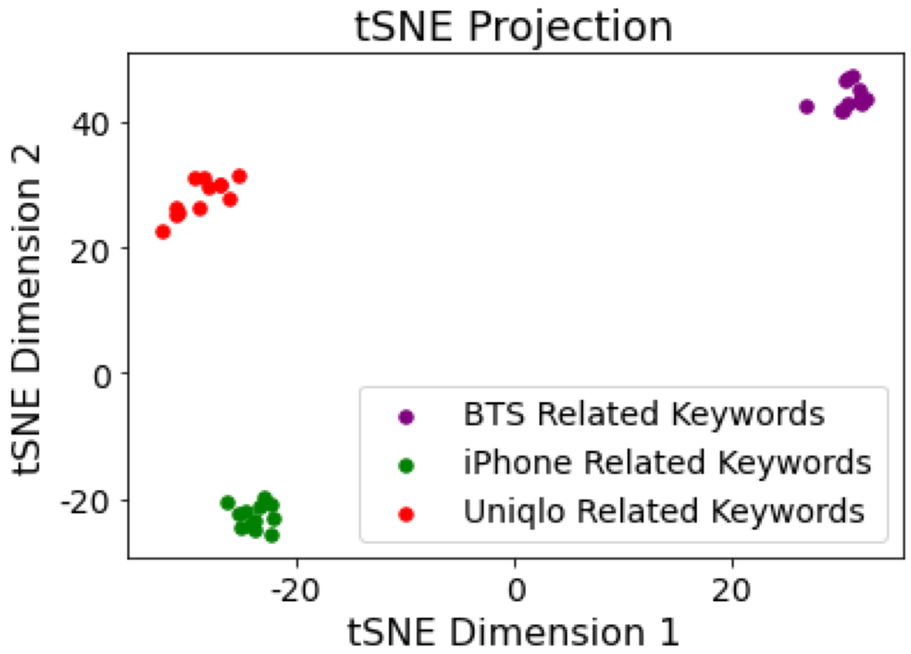
OPPO : 0.546837

헤드폰: 0.540111

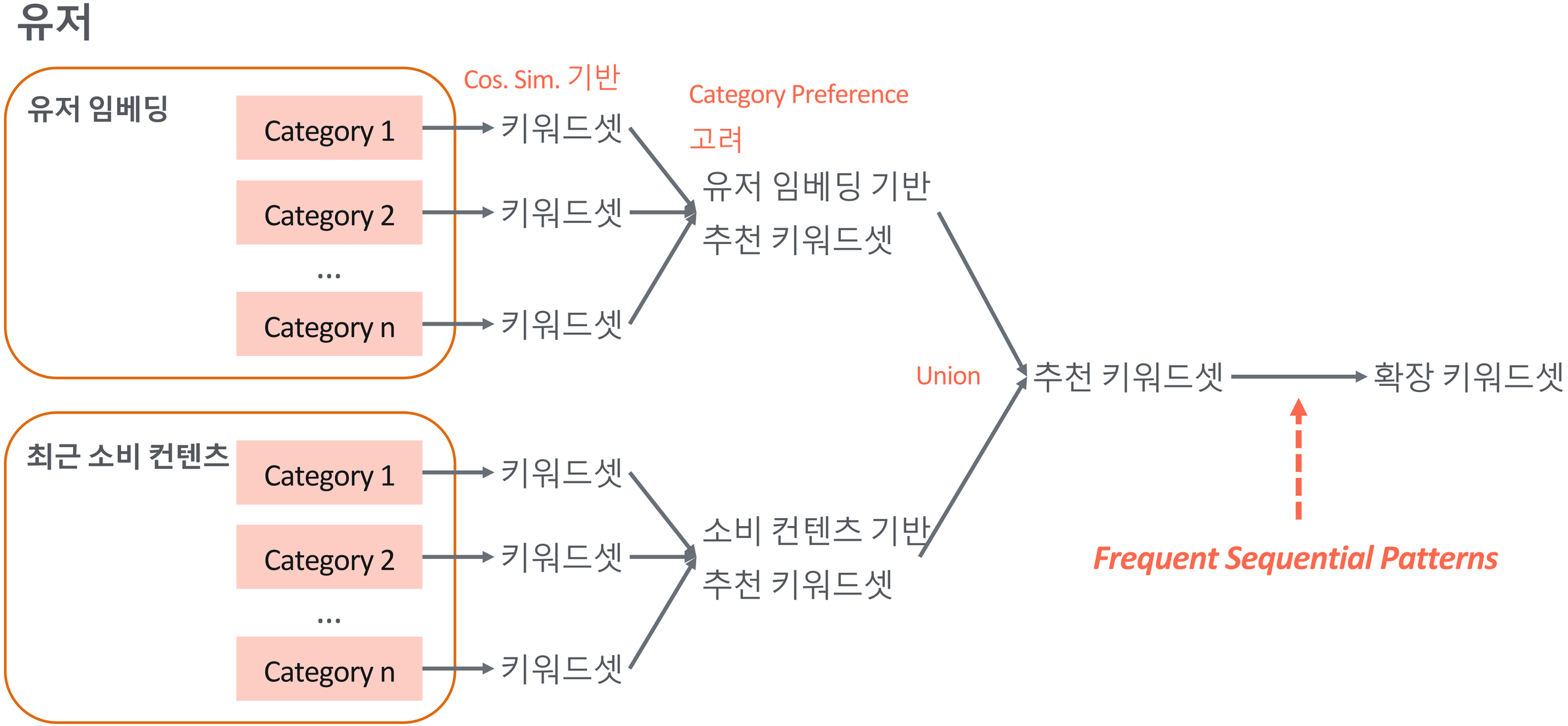
파나소닉 : 0.537820

닛산·노트: 0.527332

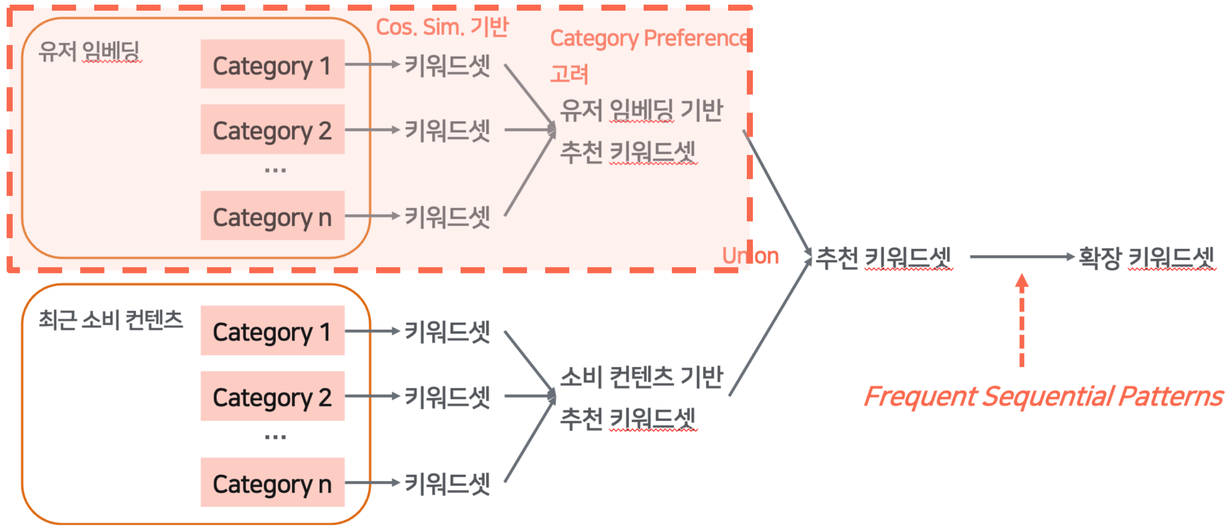
…



3.9 키워드 추천 흐름도 (WIP)



3.10 키워드 추천 예시



Category Preference

읽은 컨텐츠 수 / 날짜 수에 기반

엔터테인먼트: 0.557922

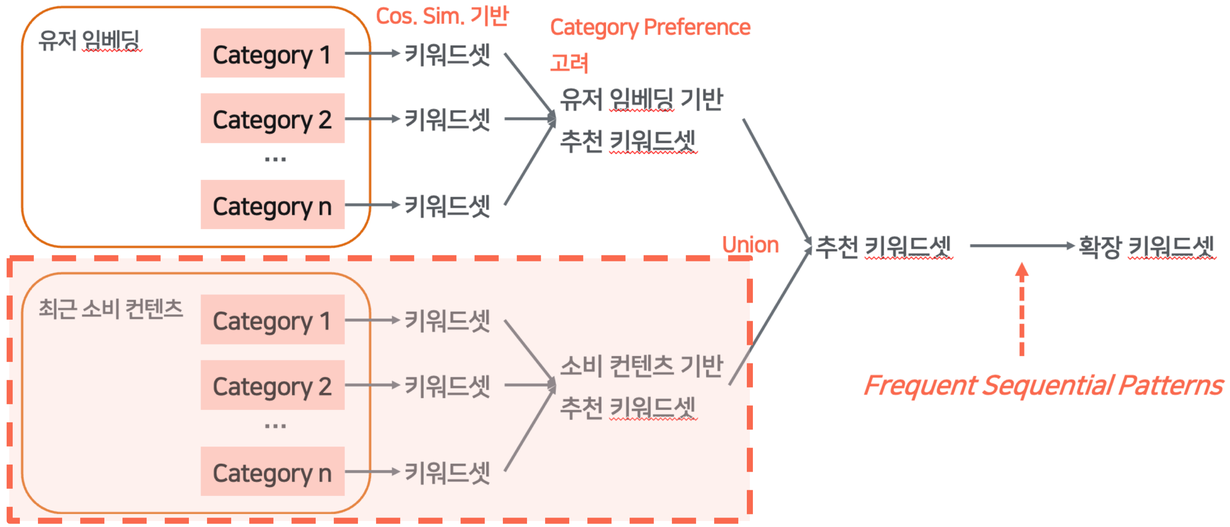
인터넷IT: 0.162259

문화: 0.147509

음식: 0.132310

엔터테인먼트	음식	인터넷 IT
BTS	로손	iPhone
JUNG KOOK	초밥	Apple Music
SUGA	카레	Google Pixel
BLACKPINK	미니스톱	Apple Watch
Wanna One	식빵	Microsoft Surface
Butter	코메다	ASUS
…	…	…

3.10 키워드 추천 예시



엔터테인먼트

BTS: 0.764609

BLACKPINK: 0.745327

SUGA: 0.727481

JUNG KOOK: 0.686835

리사: 0.675646

Wanna One: 0.622867

Butter: 0.615348

Red Velvet: 0.577691

...

인터넷 IT

Apple Watch: 0.679642

iPhone: 0.613610

Apple TV: 0.552783

Apple Music: 0.500514

Google Pixel: 0.474261

헤드폰: 0.436094

Samsung Galaxy: 0.425419

...

음식

스시: 0.625859

갓파스시: 0.623755

카레: 0.611984

로손: 0.601496

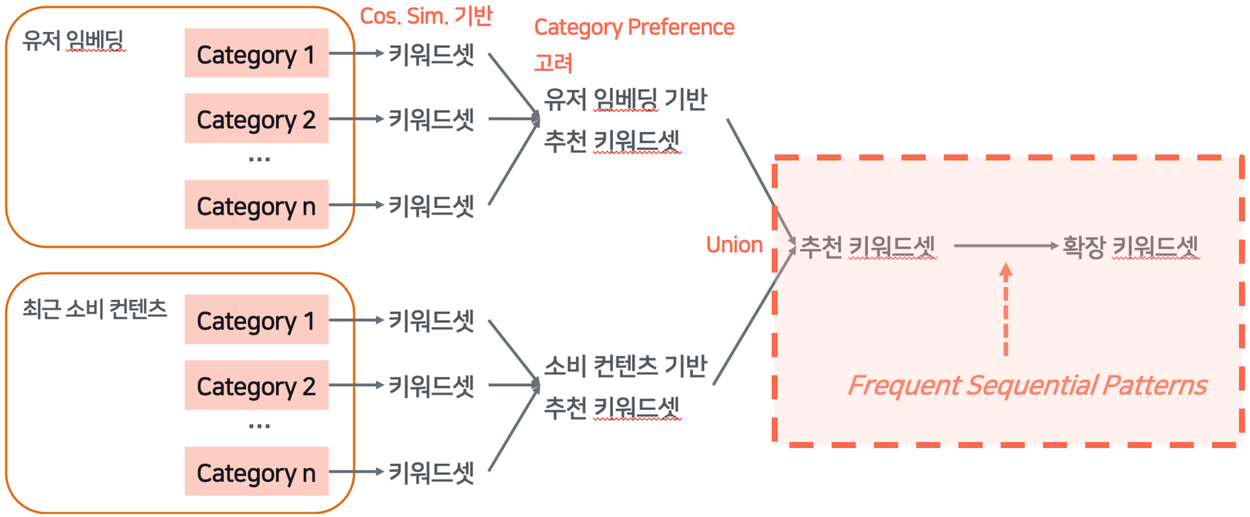
불고기: 0.546414

미니스톱: 0.528504

크로와상: 0.521363

...

3.10 키워드 추천 예시



시드

Apple Watch

크로와상

Samsung Galaxy

햄버거

초밥

Red Velvet

코메다

Pokémon GO

TWICE

GOT 7

...

확장된 키워드셋

applewatch 밴드

...

redvelvet 새 드라마

redvelvet photo

redvelvet 미니앨범 queendom

redvelvet queendom

redvelvet 패션

redvelvet 조이 crush

redvelvet crush

redvelvet 조이

...

...

twice mv 재생 횟수

twice 일본인 멤버

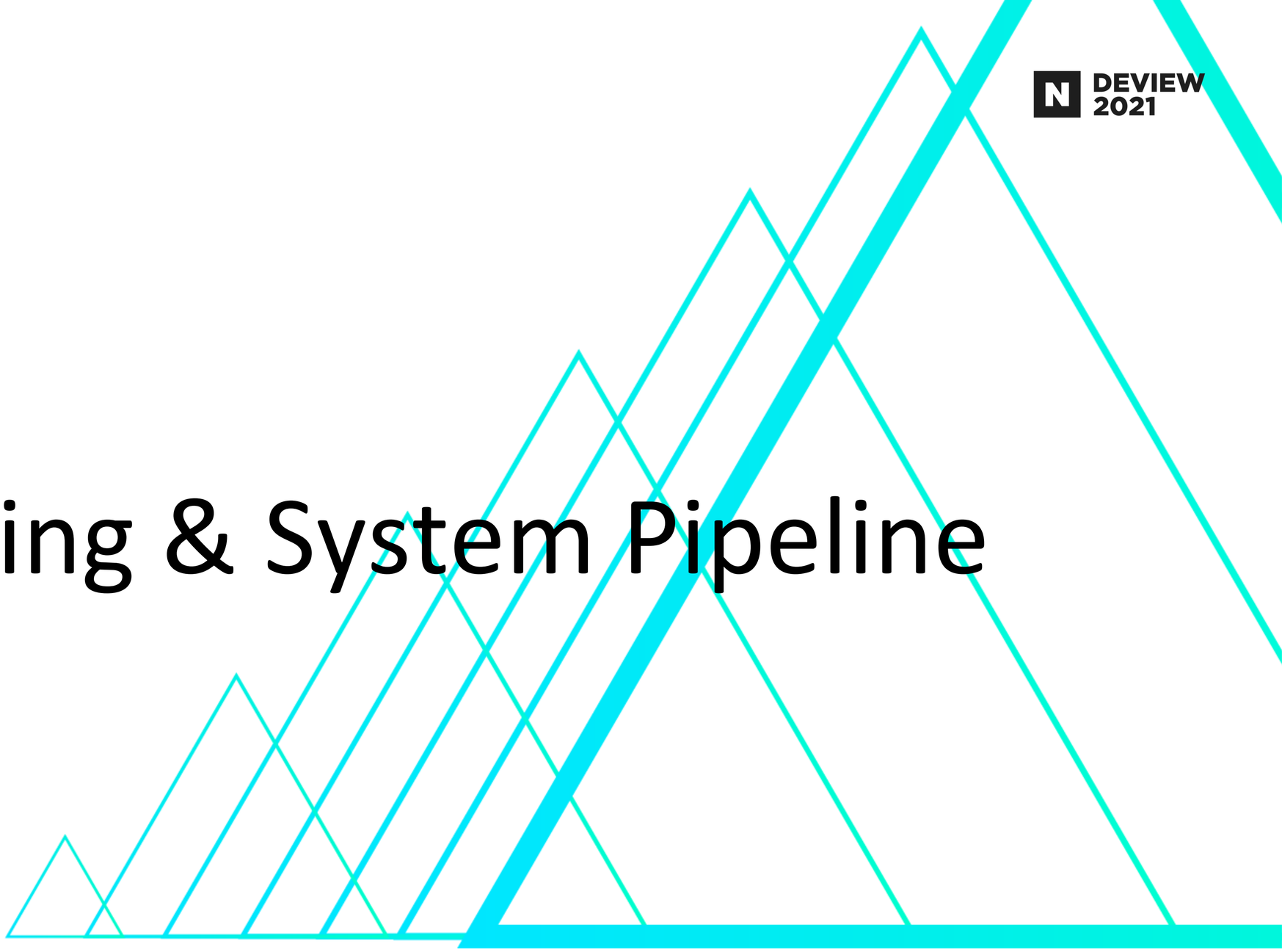
twice 영상

twice 모모 공개

twice 근황 공개

twice 트와이스 메이크업

...



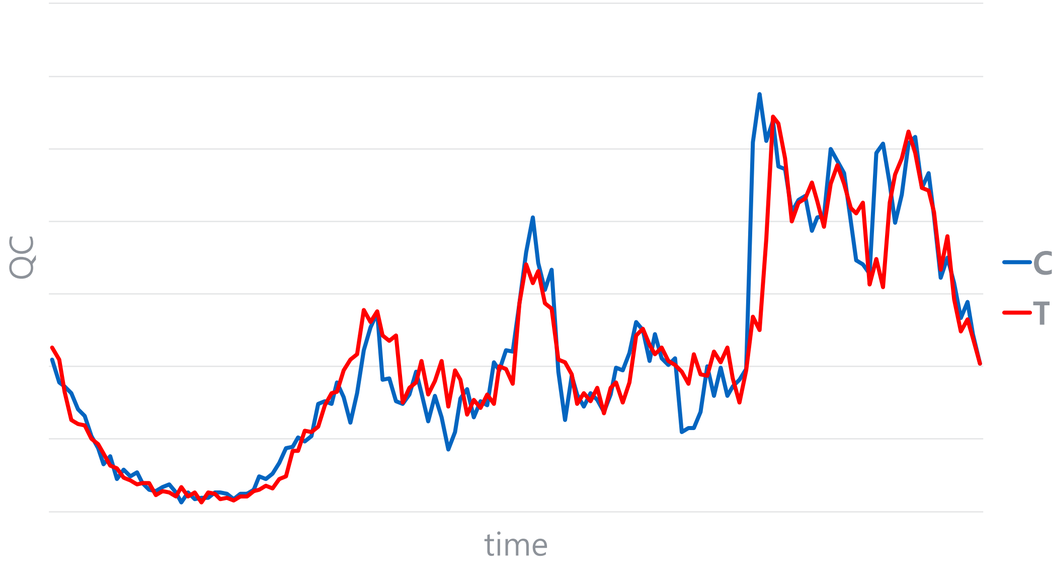
4.Ranking & System Pipeline

4.1 2020년의 교훈

하단에 노출하는 키워드일수록 개인화가 잘 동작한다

- C : Control Group(편집국의 수동 랭킹 모델) / T : Test Group(AiRS의 개인화 랭킹 모델)
- 상위 랭크는 모두가 관심있어 할 만한 키워드 : 화제의 키워드
- 하위 랭크로 내려갈수록 개인화 강화 : 개인화 키워드

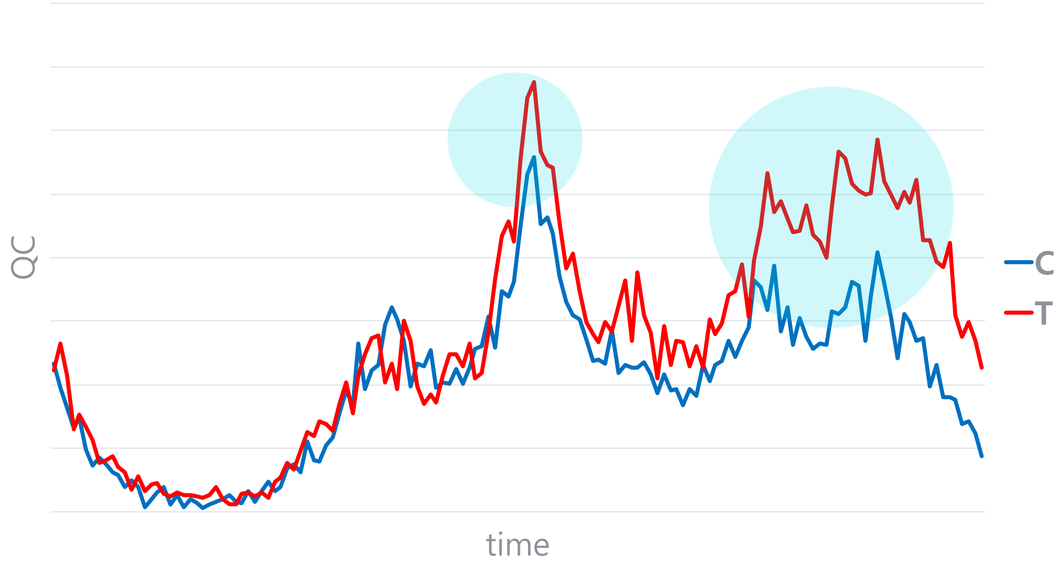
QC Graph of Keyword from Rank1 to 3 (Control vs Test)



상위 랭크에서 Control Group과의 유의미한 차이 확인 어려움

Control Group 대비 Test Group QC의 비율: 99.6%로 비슷한 수치

QC Graph of Keyword after Rank4 (Control vs Test)



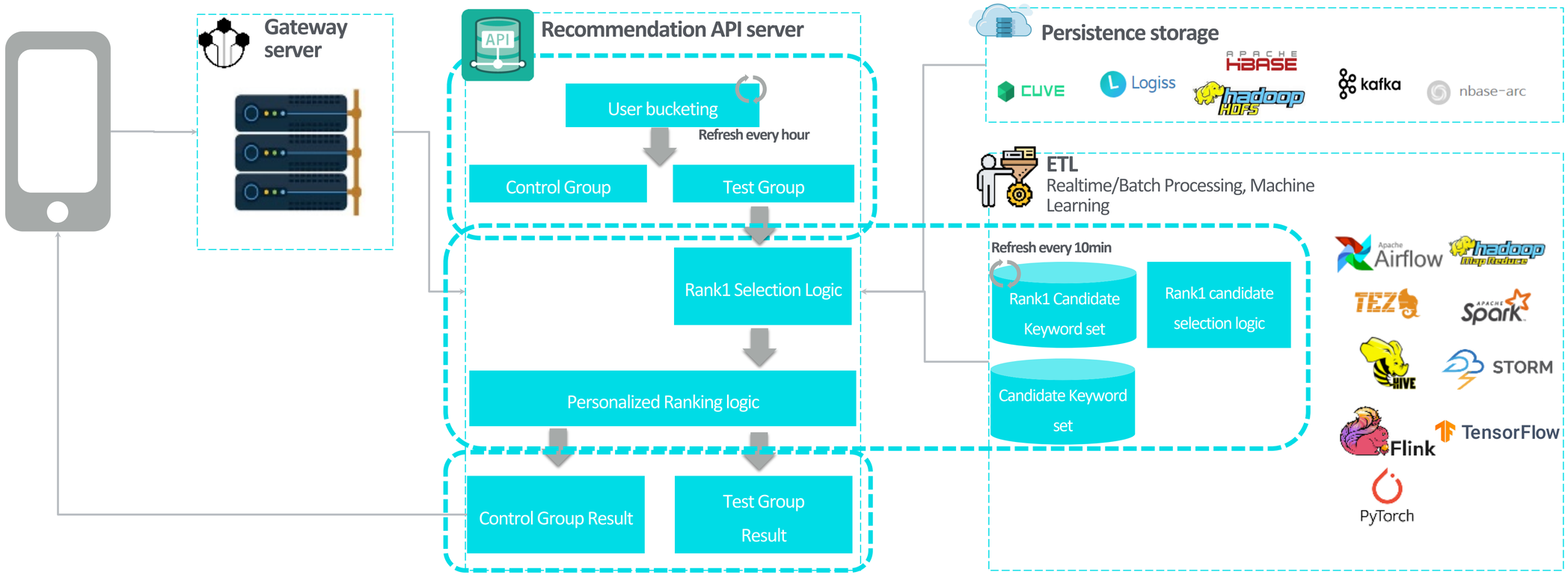
하위 랭크에서 Test Group이 Control Group 대비 우세

Control Group 대비 Test Group QC의 비율 : 132.4%로 T4우세

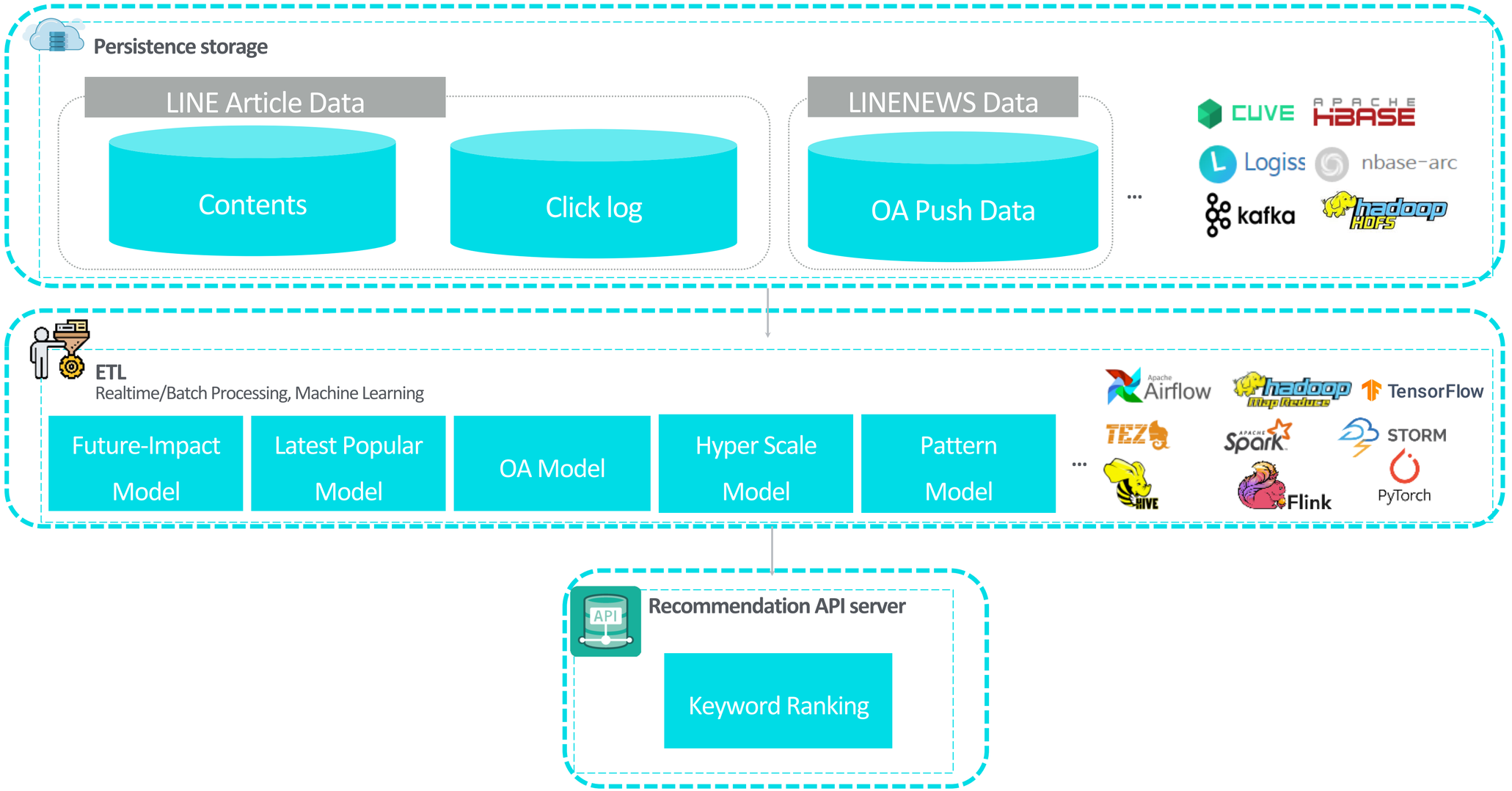
4.2 Rank1 키워드 선정 파이프라인

피드백 로그를 활용한 rank1 키워드 선정(bandit-like)

- Test group을 대상으로 대상 키워드 노출, 피드백 로그를 통해 최적의 키워드 선정
- 선정된 키워드는 Control Group 사용자들의 rank1 키워드로 노출



4.3 키워드 추천을 위한 파이프라인



5. Future Works

5.1 Future Works

이슈 키워드 자동 생성 모델 개선

- 생성된 키워드의 신뢰도를 측정할 수 있는 모델 학습

키워드 확장 모델 개선

- 더 다양한 Sequence 데이터로부터 추출
- 추출되는 패턴들의 품질 개선
- 패턴 마이닝 기법 외 State-of-the-art 방법론 적용

검색 만족도를 고려한 키워드 추천 모델 개발

- LINE News이외의 영역과 서비스로 키워드 추천 영역 확장
- 검색 결과 개선으로의 선순환 고리 생성

Thank You