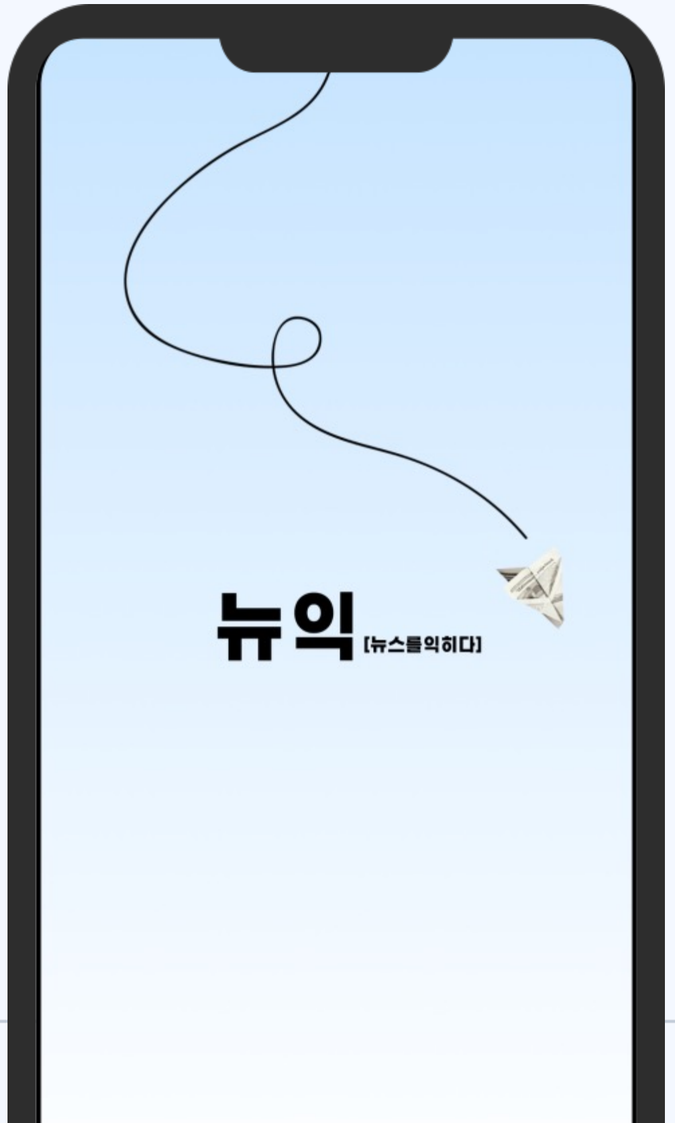
2022 캡스톤 디자인 29조

뉴익
뉴스를 익히다

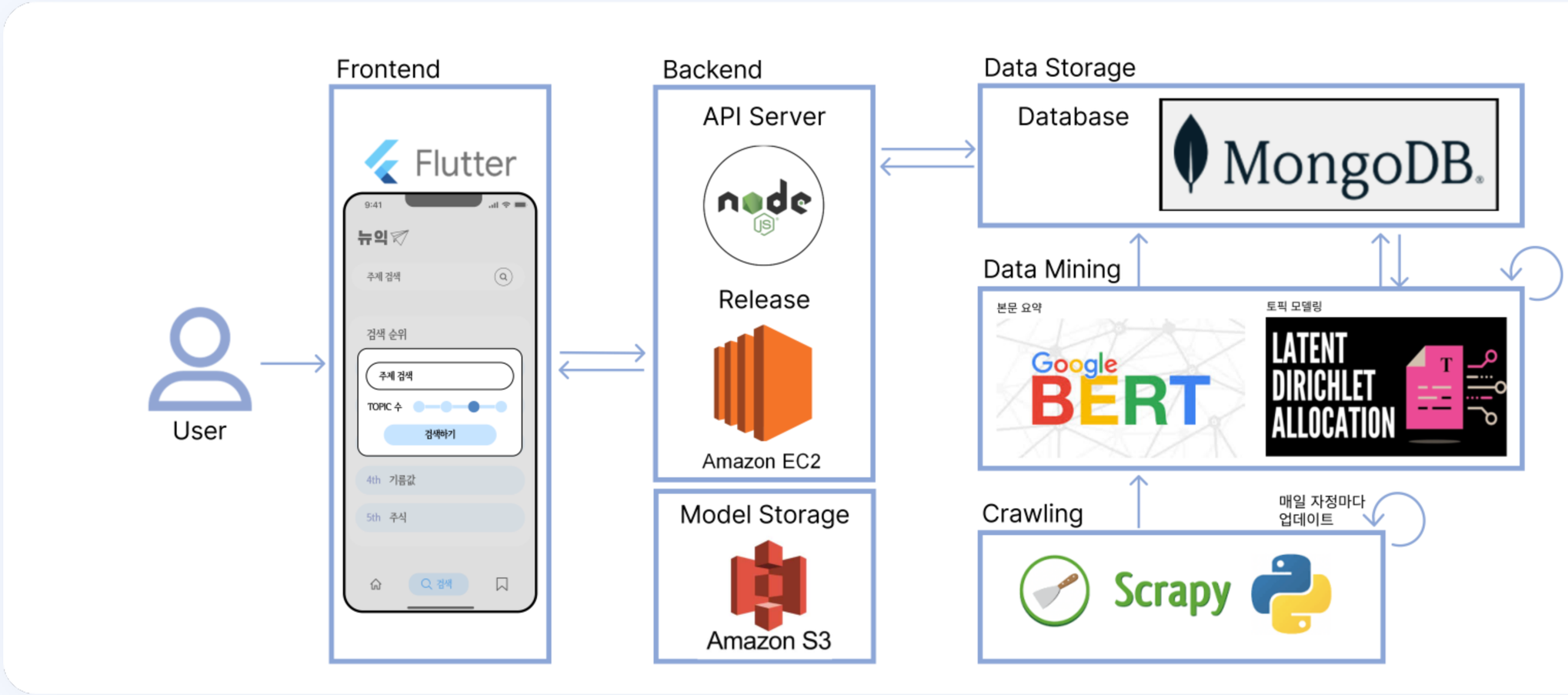
Contents

뉴익이란?

검색어에 대한 주요 토픽들의 흐름을
한 눈에 파악할 수 있게 도와주는 앱



시스템 구조

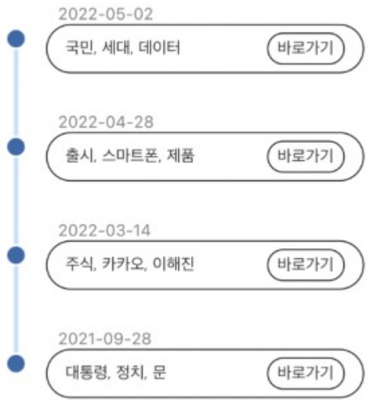


*이곳에 사업 및 서비스와 관련된 이미지를 넣어주세요.

주요 기능

사용자 맞춤 정보 제공

토픽 타임라인



뉴스 기사 요약

코로나19 백신 1차 접종자가 누적 3천만명을 넘어선 가운데 6일 오전 서울 중랑문화체육관에 마련된 접종센터에서 백신 접종을 마친 시민들이 이상반응 모니터링을 기다리고 있다. 연합뉴스 코로나19 잔여백신 대량 폐기 우려에 당국이 다음주께부터 2차 접종도 네이버·카카오톡 에스엔에스(SNS) 당일 예약이 가능하도록 시스템을 변경하기로 했다. 또한 잔여백신 1차 접종 대상을 늘리기 위해 우선 접종 대상을 희망자 전체로 확대했다.

사용자 맞춤 정보 제공

사용자의 검색어와 횟수를 버블 차트로 보여줌

사용자 맞춤 정보 제공

사용자의 키워드와 관련된 최신 뉴스를 홈에서 보여줌

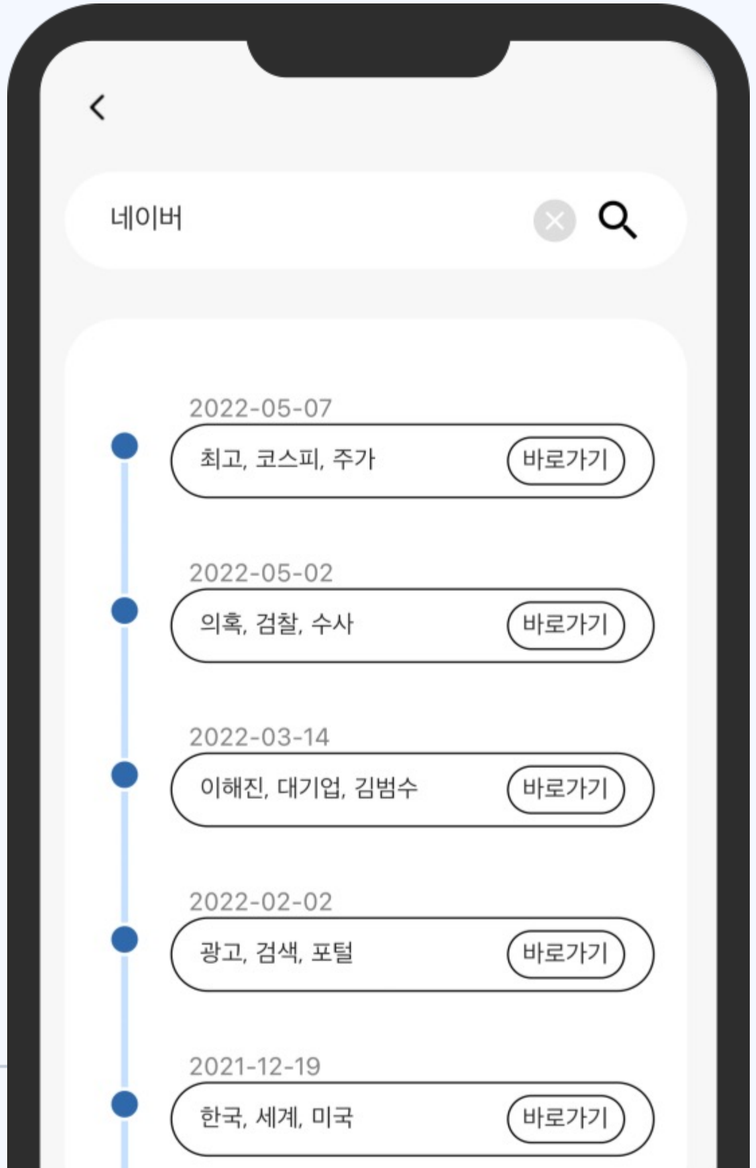
토픽 타임라인

뉴익은 사용자가 뉴스 검색어에 대한 흐름을 한 눈에 파악할 수 있도록 토픽 타임라인을 제공함

특정 검색어에 대한 주요 사건을 알기 위해 뉴스 포탈에 검색을 할 경우 중복되는 기사와 별로 중요치 않은 기사들이 쏟아져 나온다. 이를 방지하고, 주요 토픽을 한 눈에 볼 수 있게 하기 위하여 토픽 타임라인을 제작하였다.

Keyword 1	Keyword 2	Keyword 3
입력 데이터	**전처리**	**토픽 모델링**
title, summary, content	NLP	LDA

토픽 모델링의 기법 중 하나인 LDA를 이용하여 기사 데이터에 숨어있는 토픽을 추출



토픽 타임라인 - 입력 데이터

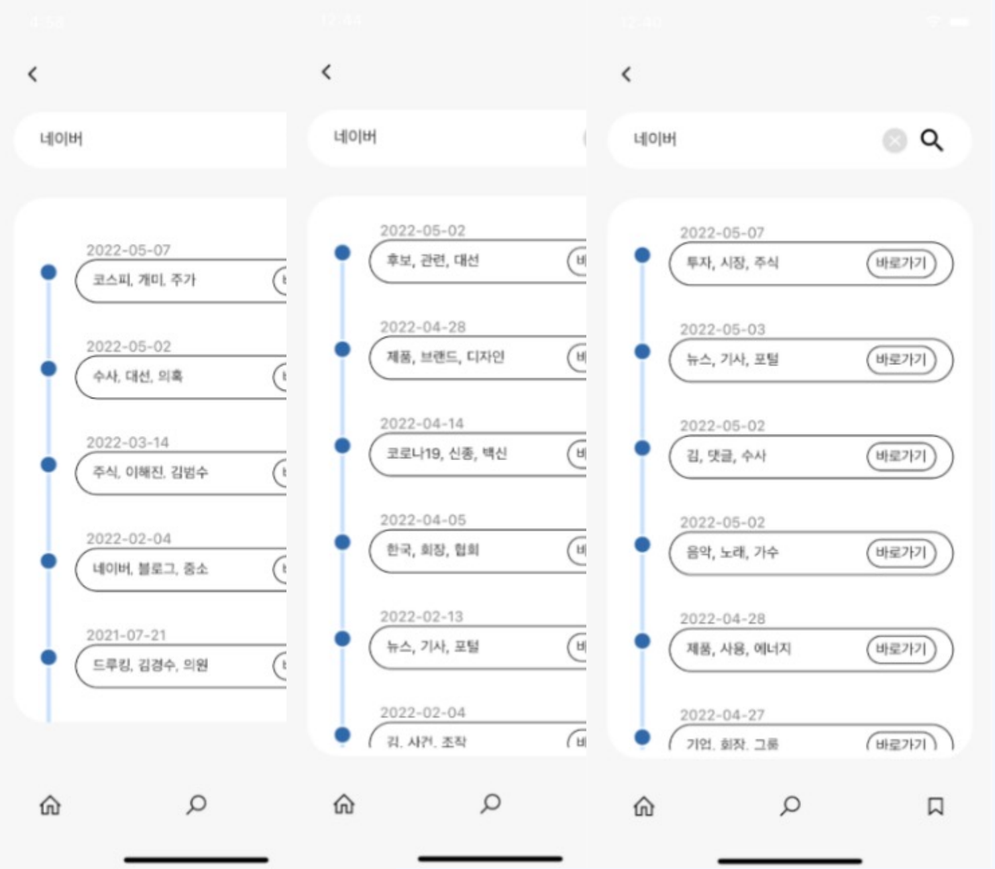

```
_id: ObjectId("62833f378f19a1c477ac33c6")
journal: "한겨레"
date: "2022-05-11"
title: "UEFA, 챔스 최종 개편안 승인…참가팀·경기수 모두 늘어난다"
url: "https://www.hani.co.kr/arti/sports/soccer/1042492.html"
        "기존 32팀에서 36팀으로, 조별리그 대신 풀리그

content: 리그에서 한 팀당 8팀과 8경기 치러 16강 티켓 경쟁

        유럽축구연맹(UE..."
summary: "유럽축구연맹(UEFA) 집행위원회 멤버들이 11일(현지시각) 오스트리아 빈에서 열린 46회 유에파 정기총회에 참석해 논의 중이..."
```


토픽 추출이 가능한 field: title, content, summary => 성능 비교

Keyword 1	Keyword 2	Keyword 3	Keyword 4
Coherence	Perplexity	소요 시간	사용자 시선

입력 데이터 - 성능 비교

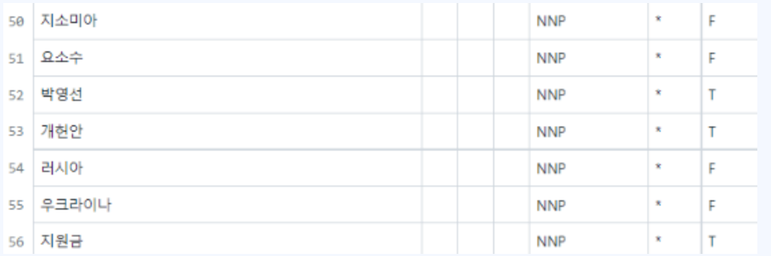
Coherence	content	>	summary	>	title
Perplexity	content	>	summary	>	title
소요 시간	title	X 7 >	summary	>	content X 26
사용자 시선	title	>	summary	>	content



title summary content

토픽 타임라인 - 전처리

Step 1 > Step 2 > Step 3 > Step 4 > Step 5

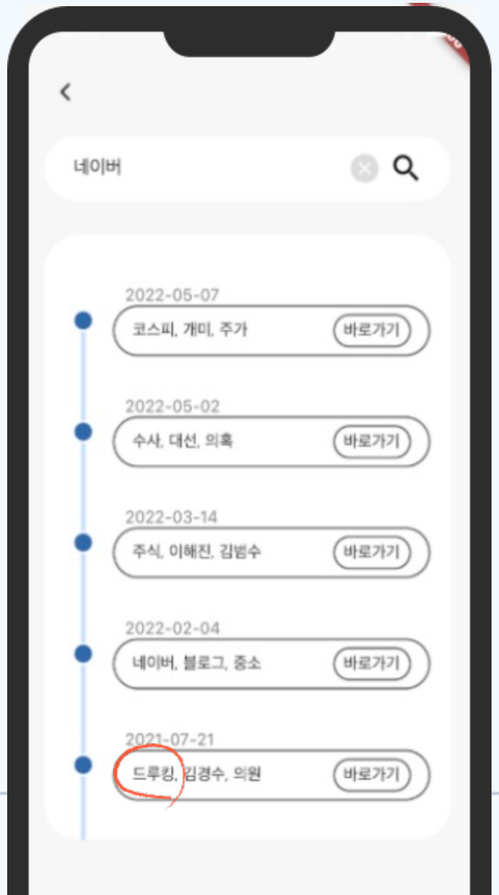
정규표현식	알파벳 소문자화	토큰화 & 품사 태깅	불용어 처리	사용자 사전
[이진순 칼럼] [포토] 박병수 선임기자 suh@hani.co.kr		외국어(SL), 일반 명사(NNG), 고유 명사(NNP), 동사(VV), 형용사(VA), 어근(XR), 한자(SH)		
• 토픽 파악에 도움이 되지 않는 중복된 요소 제거	• APPLE과 apple은 같은 키워드이므로 혼란도를 줄이기 위해 소문자로 통일	• 토픽일 가능성이 있는 품사만 남기고 제거	• 불용어 처리 및 Mecab의 user-dic을 업데이트하여 필요 없는 단어 제거 및 잘못 분리되는 단어 없도록 개선	

전처리 - 성능 개선

품사 필터링
하지 않았을 때

사용자 사전에
추가하지 않았을 때

둘다 만족



토픽 타임라인 - 토픽 모델링(LDA)

Step 1 > Step 2 > Step 3 > Step 4

토픽 개수 정하기

- LDA에서 토픽의 개수는 하이퍼 파라미터

-> perplexity와 coherence를 이용하여 최적의 토픽 개수를 구함

iteration 횟수 정하기

- iteration은 학습 반복 횟수

-> 커질수록 소요시간 증가 perplexity와 coherence 및 소요시간을 고려하여 정함

θd threshold

- θd threshold는 문서(기사)가 해당 토픽일 확률

-> 이를 이용하여 토픽에 해당하는 기사 중 관련도가 낮은 기사 필터링

num of news threshold

- num of news는 토픽 타임라인에 들어가기 위해 토픽 당 하루에 등장해야하는 뉴스의 최소 개수

-> 이 값을 조절하여 토픽 타임라인의 자세함 정도를 설정 가능

02 주요 기능 및 구현 방법 3

뉴스 기사 요약

사용자가 뉴스 원문의 중요한 내용을 보다 빠르고 편리하게 읽을 수 있도록 본문 요약 서비스를 제공함

뉴스마다 원문의 길이가 다양해 어떤 기사는 내용이 너무 많고, 또 어떤 기사는 내용이 너무 간결하다. 뉴익에서는 그러한 점을 보완하고자, 중요한 문장들이 포함된 최대한 비슷한 길이의 기사를 제공하고자 노력하였다.

Keyword 1

추출 요약

KoBertSum

Keyword 2

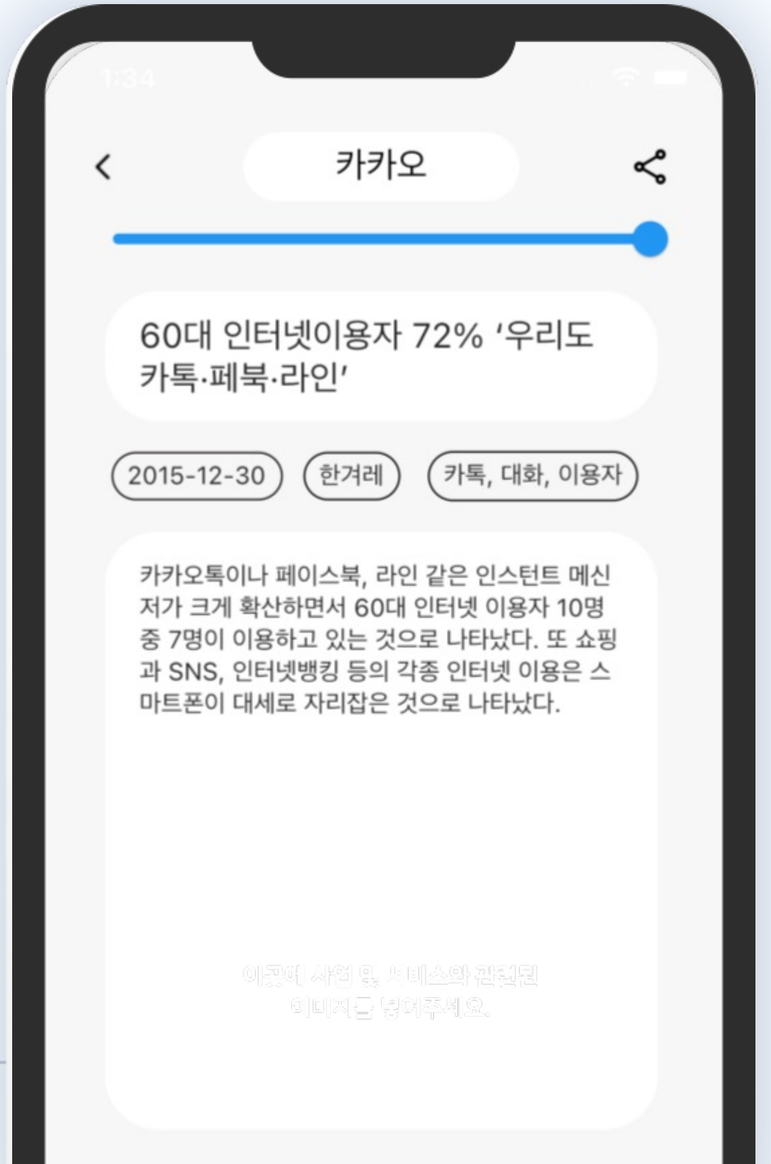
ROUGE

성능 평가 지표

Keyword 3

문단 단위 요약

추출 요약 모델인 KoBertSum으로 학습하여, ROUGE로 평가



요약 모델 프로세스

Step 1 > Step 2 > Step 3

AI HUB 문서요약 텍스트 신문기사 Data | Model | 요약문

01 csv 파일

02 json 파일

03 .pt 파일

지난 10일 오후 부산 해운대구 벡스코 제1전시장에서 열렸던 '나훈아 AGAIN 테스형' 콘서트를 찾은 관람객들이 발열 체크를 하며 입장하고 있다. 이날 부산에선 코로나19 확진자가 319명 발생했다. 연합뉴스 부산에서 사흘 연속 코로나19 확진자가 최다 발생하면서 누적 확진자가 2만명을 넘었다. 경북에서도 이틀 연속 코로나19 하루 확진자가 최다 발생했다.

- 데이터 셋 274,105개에 대해 학습 진행
- KoBertSum 모델로 학습
- 문단 단위로 요약문 완성

성능 - ROUGE

ROUGE-1, 2, L

- Rouge: 모델이 자동 요약한 요악문과 모범 답안 요약문이 얼마나 유사한지 비교해 성능을 계산하는 지표
- Rouge-n: 연속된 n개의 단어를 하나의 단위로 보는 것을 기반으로 두 요약문 사이의 유사도를 계산
- Rouge-L: 최장 공통 부분 수열로 평가

- Rouge-R: Recall
- Rouge-P: Precision
- Rouge-F: F1 Score

성능

- F1 Score Pre-trained vs 학습 model

	BERTSUM+Classifier	BERTSUM+Transformer
ROUGE-1	43.23	43.25
ROUGE-2	20.24	20.24
ROUGE-L	39.64	39.59

	Fine-tune-Classifier	Fine-tune-Transformer
ROUGE-1	58.51	63.49
ROUGE-2	41.77	45.42
ROUGE-L	58.44	63.43

- Recall Pre-trained vs 학습 model

	BERTSUM+Classifier
ROUGE-1	46.66
ROUGE-2	26.35
ROUGE-L	42.62

	Fine-tune-Classifier	Fine-tune-Transformer
ROUGE-1	79.40	75.92
ROUGE-2	56.66	54.52
ROUGE-L	79.33	75.84

02 주요 기능 및 구현 방법 3

뉴스 데이터 요약문 생성

Step 1 > Step 2 > Step 3

MongoDB Data | 저장한 Model | 요약문 생성

01 Dictionary Data

02 json 파일

03 .pt 파일

- 검색어가 title과 원문에 포함된 Data set을 Dictionary 형태로 받아옴

- 학습한 모델 Load

02

시연 영상

기대 효과

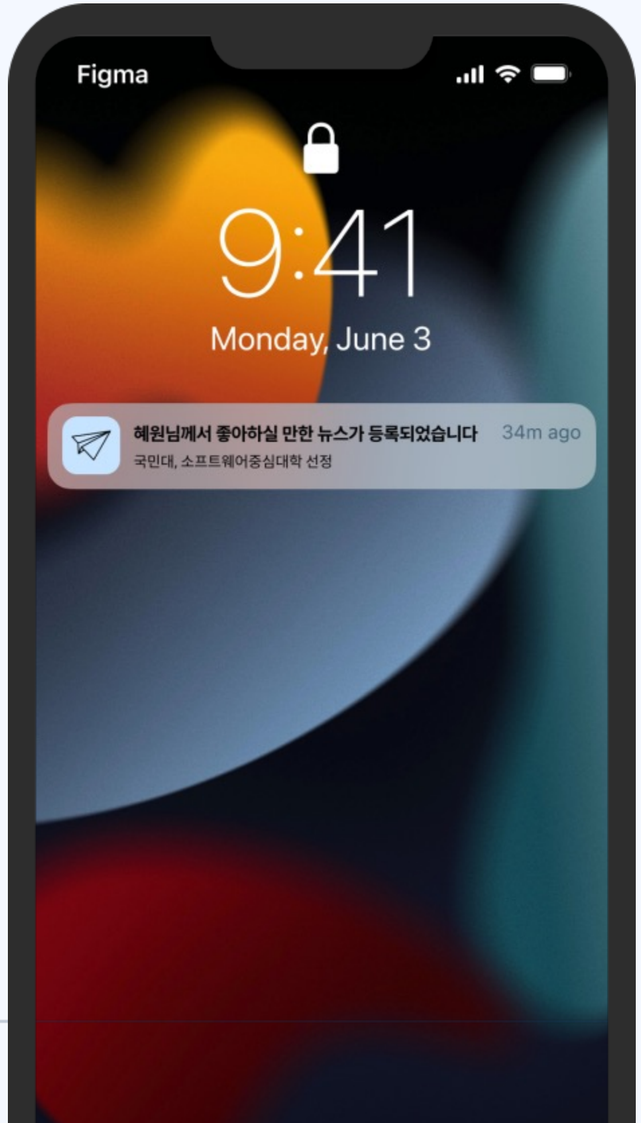
용이한
자료 조사

빠른
정보 습득

전 연령대
뉴스 구독 장려

발전 방향

- 다양한 언론사 추가
- 유사한 기사 묶기
- 추천 알고리즘 적용



뉴익

Thank you