토마토주스

2025 전기 졸업과제

공공데이터를 활용한 KoBERT 파인튜닝과 한국어 키워드 분석 및 대시보드 시각화

박준혁
이차현
임성표
지도교수 : 조준수

목차

01

과제 배경

전 세계적으로 텍스트 분석 및 자연어 처리 시장은 급속도로 성장하고 있다. 빅데이터 시대에 접어들면서 대량의 텍스트 데이터에서 핵심 정보를 추출하는 키워드 추출 기술의 중요성이 크게 증가하고 있다. 국내에서도 정부 기관, 언론사, 기업들이 문서 요약, 검색 최적화, 콘텐츠 분류 등의 목적으로 키워드 추출 기술을 활용하고 있다.

01

문제 분석

1

계산 복잡도 문제

기존 KeyBERT는 전체 문서에 대한 CLS 토큰과 문서의 일부분에 대한 CLS 토큰 간의 벡터 유사도를 측정하는 방식으로, 입력 문장이 길어질수록 계산 시간이 기하급수적으로 증가한다.

2

도메인 특화 부족

KeyBERT에서 사용하는 사전학습 모델들은 키워드 추출이라는 특정 태스크에 최적화되어 있지 않아, 도메인별 특수성을 충분히 반영하지 못한다.

3

블랙박스 특성

신경망 기반 모델은 특정 단어를 왜 키워드로 선택했는지에 대한 판단 근거를 제공하지 못하는 특성을 가져 사용자가 신뢰를 가지는데 어려움이 있을 수 있다.

4

한국어 처리의 한계

형태소 단위로 의미가 구성되는 한국어의 특성상, 조사를 고려하지 않는 키워드 추출 방식은 정확도가 현저히 떨어진다.

02

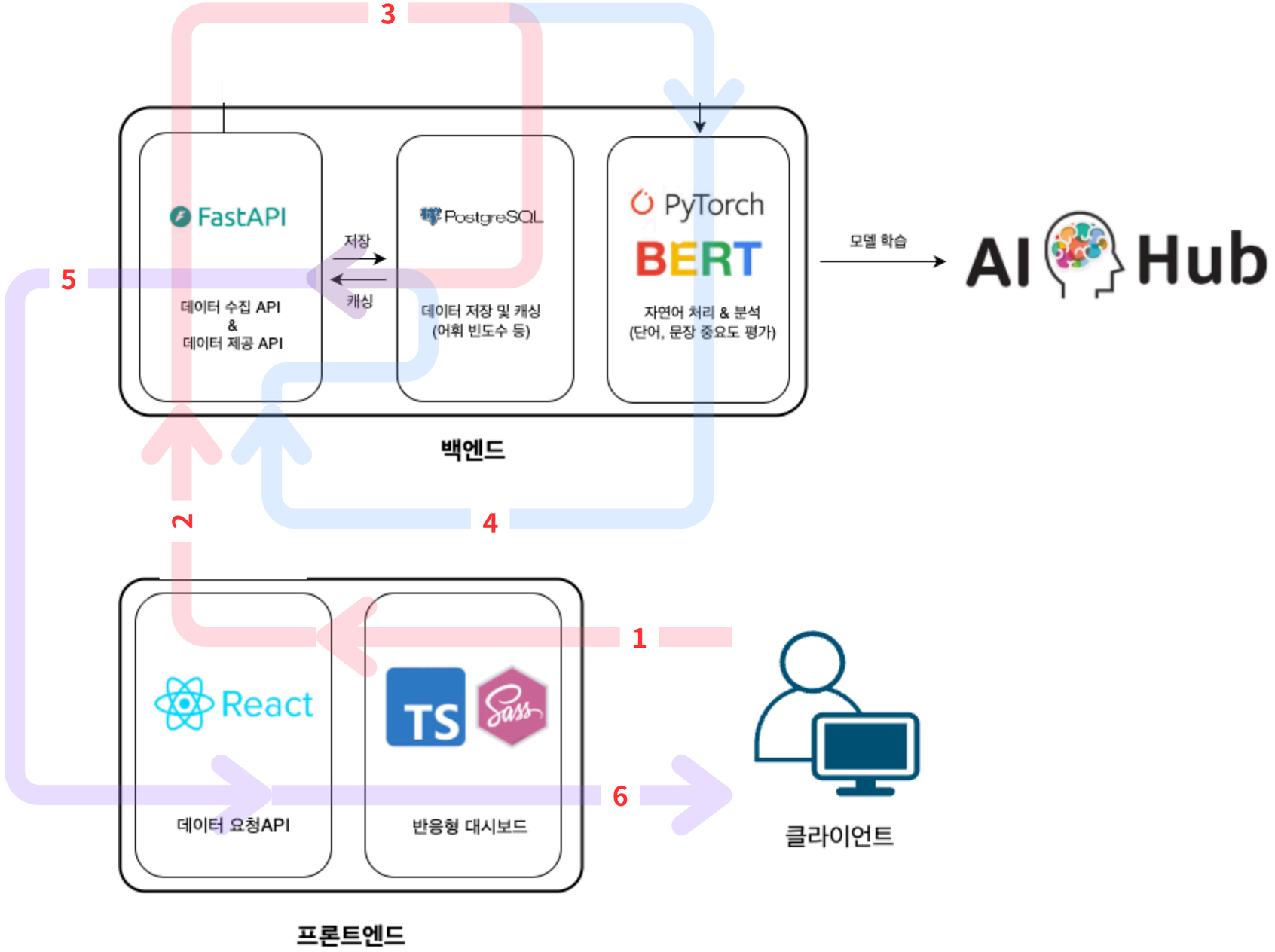
개발 목표 및 주요 기능

한국어에 특화된 고성능 키워드 추출 모델을 개발하고, 웹 기반 텍스트 분석 서비스를 구현하여 사용자가 쉽게 접근할 수 있는 한국어 텍스트 분석 플랫폼을 제공한다.

키워드 추출	키워드 설명	품사 분석 시각화	실시간 분석

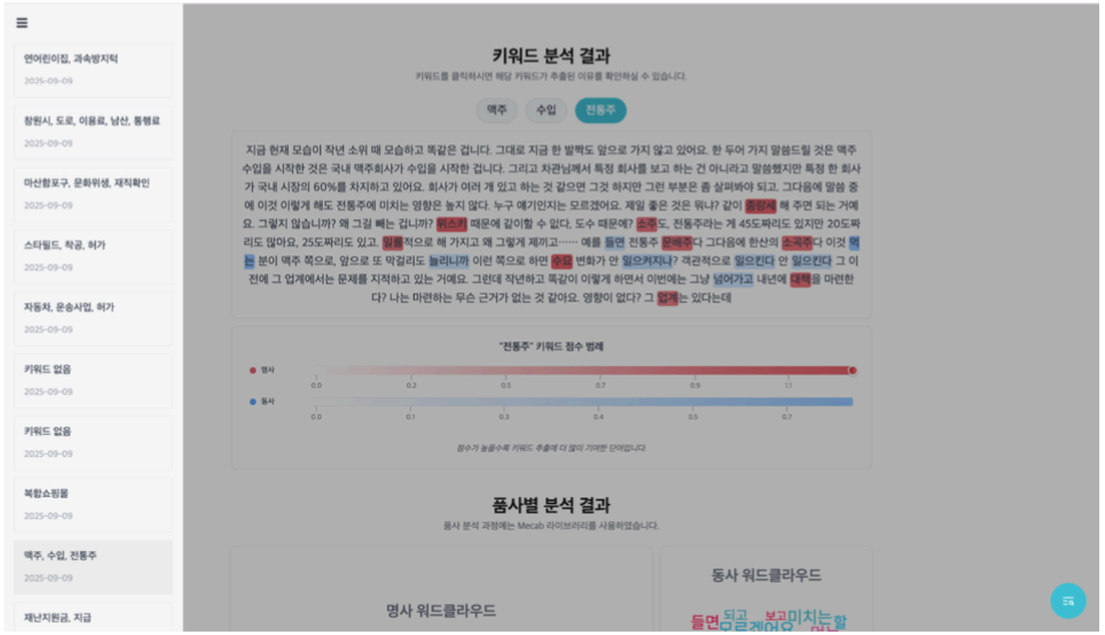
03

시스템 설계


3
FastAPI
저장
캐싱
PostgreSQL
PyTorch
BERT
모델 학습
AI Hub
5
데이터 수집 API
&
데이터 제공 API
데이터 저장 및 캐싱
(어휘 빈도수 등)
자연어 처리 & 분석
(단어, 문장 중요도 평가)
백엔드
2
4
1
React
TS
Sass
6
데이터 요청API
반응형 대시보드
클라이언트
프론트엔드

04 주요 기능 소개

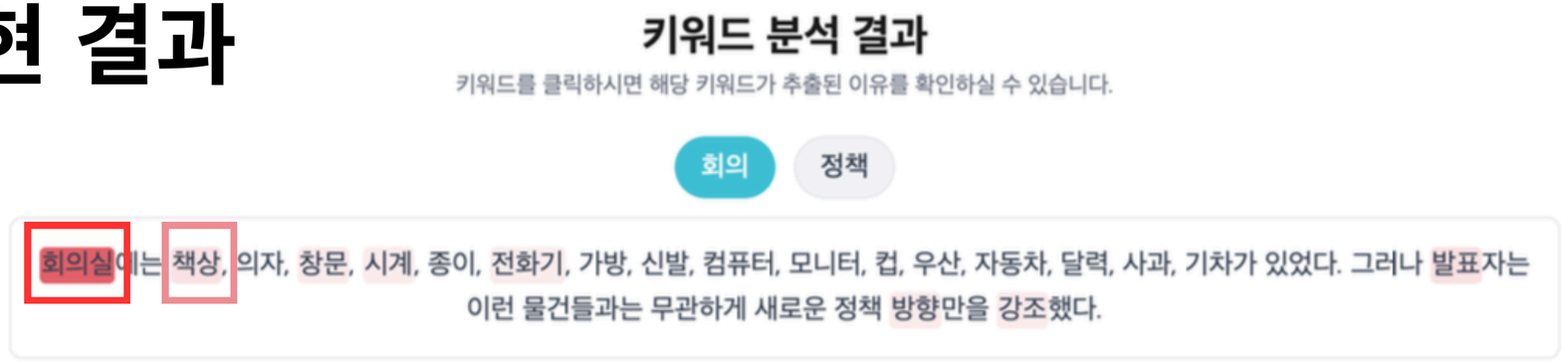
시스템 구현 결과



04 키워드별 분석 결과 비교

시스템 구현 결과

시스템 구현 결과



회의 키워드에 대한 하이라이트 결과 중 **회의실**의 attention score 시각화 결과

회의 키워드에 대한 하이라이트 결과 중 **책상**의 attention score 시각화 결과

05

모델 성능 평가

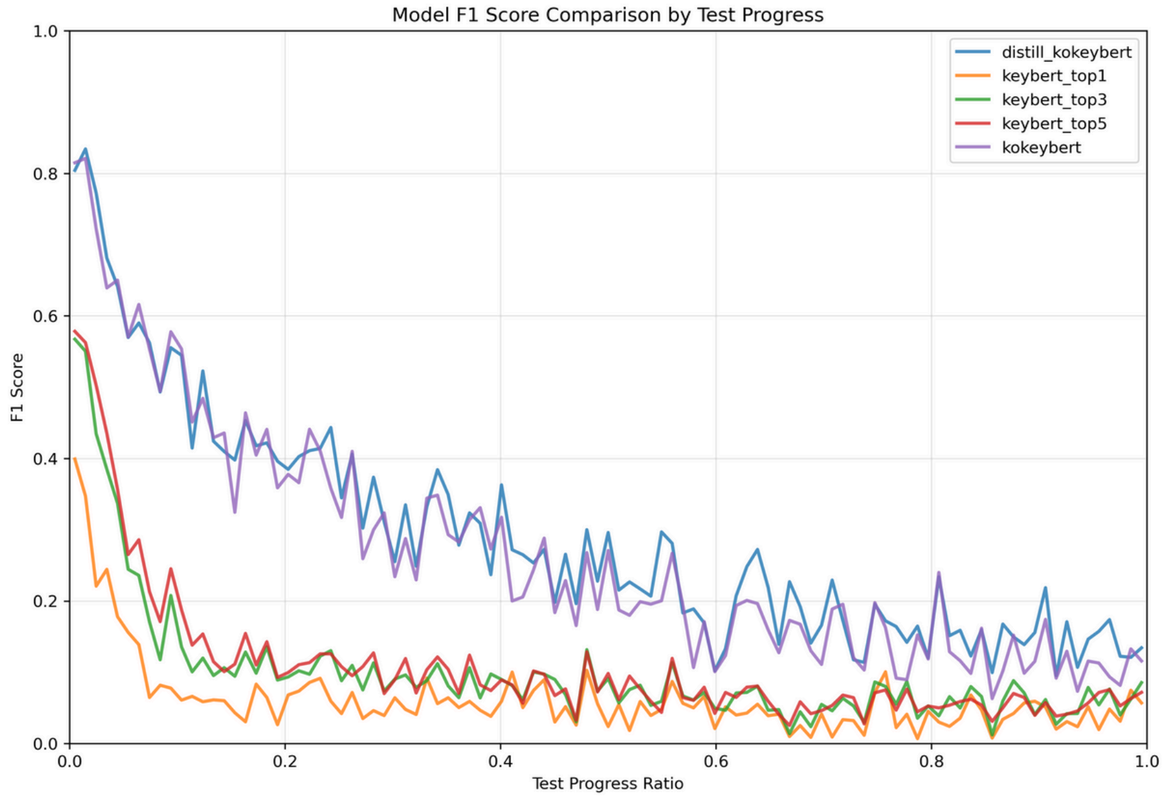
모델별 성능 비교 (테스트 데이터: 4,845개 샘플)

모델명	Precision	Recall	F1 Score	추론 시간 (테스트 데이터셋에 대한 macbook m3 air 기준)
KeyBERT(5words)	0.0880	0.1296	0.1048	11분 52초
KoKeyBERT	0.4848	0.2263	0.3086	6분 42초
DistillKoKeyBERT	**0.3640**	**0.2877**	**0.3214**	**2분 54초**

정확도 F1 Score 0.3214 (KeyBERT 대비 22%p 향상)
처리 속도 2분 54초 (KeyBERT 대비 약 4배 빠름)
균형성 Precision과 Recall의 균형 잡힌 성능
경량화 Teacher 모델 대비 5.75배 빠른 추론 속도

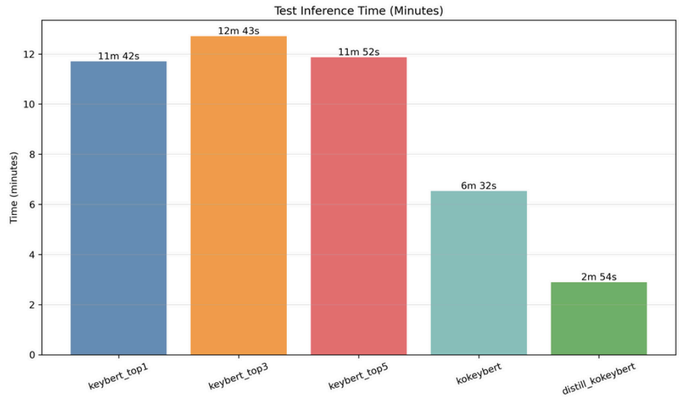
05

모델 성능 평가



- 교사모델 - kokeybert / 학생 모델 - distill_kokeybert

- 교사모델과 학생 모델의 F-1 score는 테스트 전반에 걸쳐 keyBERT 계열 모델들보다 일관되게 높은 성능을 유지한다.
- 학생 모델은 교사 모델과 거의 동일한 성능 궤적을 그리며, 이는 지식 증류를 통해 교사 모델의 핵심 성능이 학생 모델에 효과적으로 이전되었음을 시사
- 이를 통해 제안 모델들이 특정 데이터 구간에만 국한되지 않고, 전체 테스트 데이터에 걸쳐 안정적으로 우수한 성능을 보임을 알수있다.



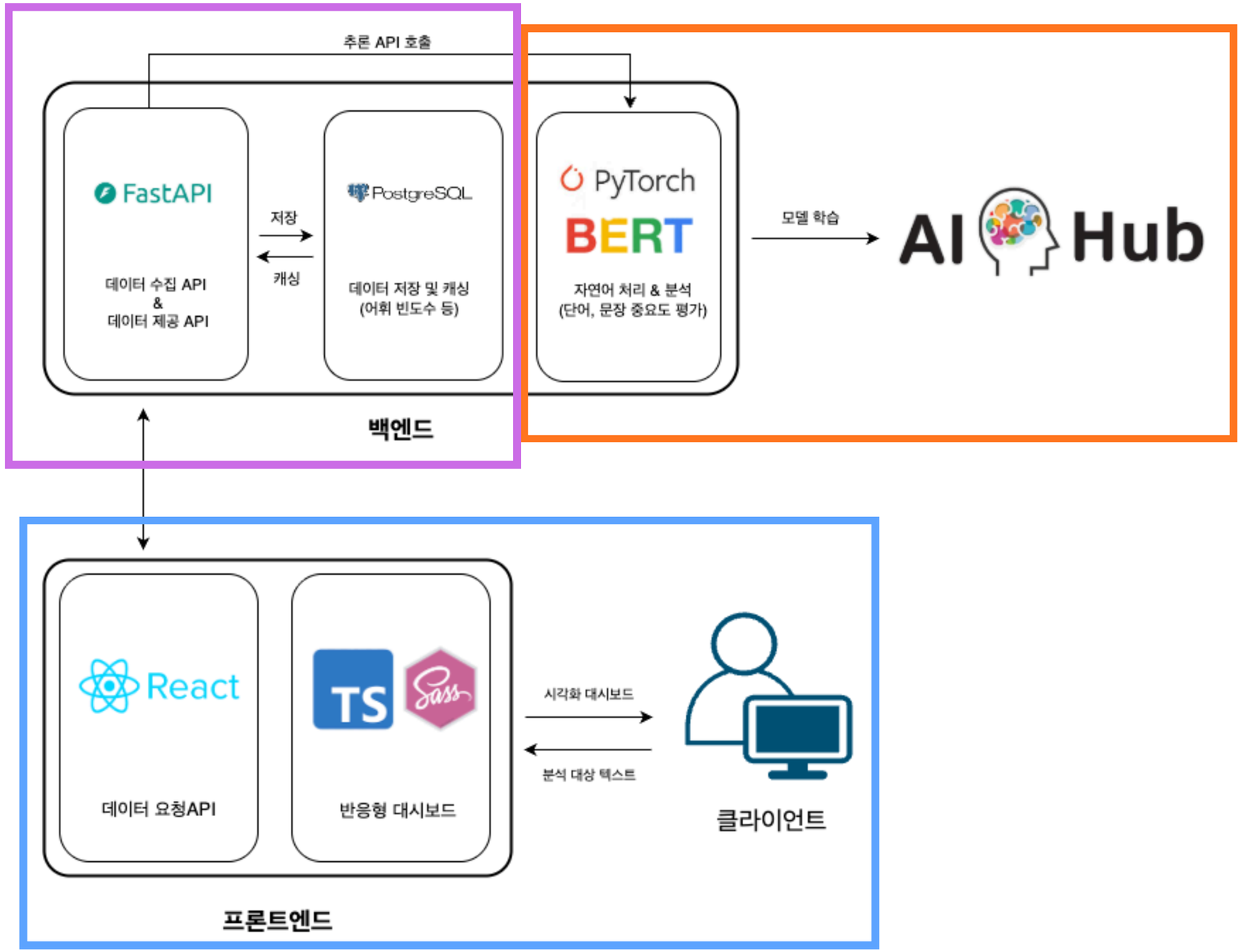
실험환경: MacBook M3 Air

평가 데이터셋에 대한 총 추론 시간 비교. distill_kokeybert 모델은 교사 모델보다 2.3배, KeyBERT보다 4배 이상의 시간 효율성을 증명했다.

05

주요 일정 및 역할 분담

구분	개발 일정											
	1주차	2주차	3주차	4주차	5주차	6주차	7주차	8주차	9주차	10주차	11주차	12주차
데이터 수집 및 전처리												
학습 데이터 구조 설계												
키워드 추출 알고리즘 설계 및 구현												
기여도 계산 및 정량화 지표 정의												
어텐션 스코어 추출 모듈 개발												
시각화 방식 및 UX 전략 설계												
프론트엔드 UI/UX 개발												
기여도 시각화 기능 구현												
백엔드 API 개발												
데이터베이스 설계 및 연결												
프론트엔드-백엔드 연동												
기능 통합 및 테스트 (공통)												
서비스 배포 및												



감사합니다