

토마토주스

2025 전기 졸업과제

공공데이터를 활용한 KoBERT 파인튜닝과 한국어 키워드 분석 및 대시보드 시각화

박준혁

이차현

임성표

지도교수 : 조준수

목차

01 문제 분석

02 개발 목표

03 시스템 설계

04 시스템 구현 결과

05 모델 성능 평가

06 주요 일정 및 역할분담

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.



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.

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

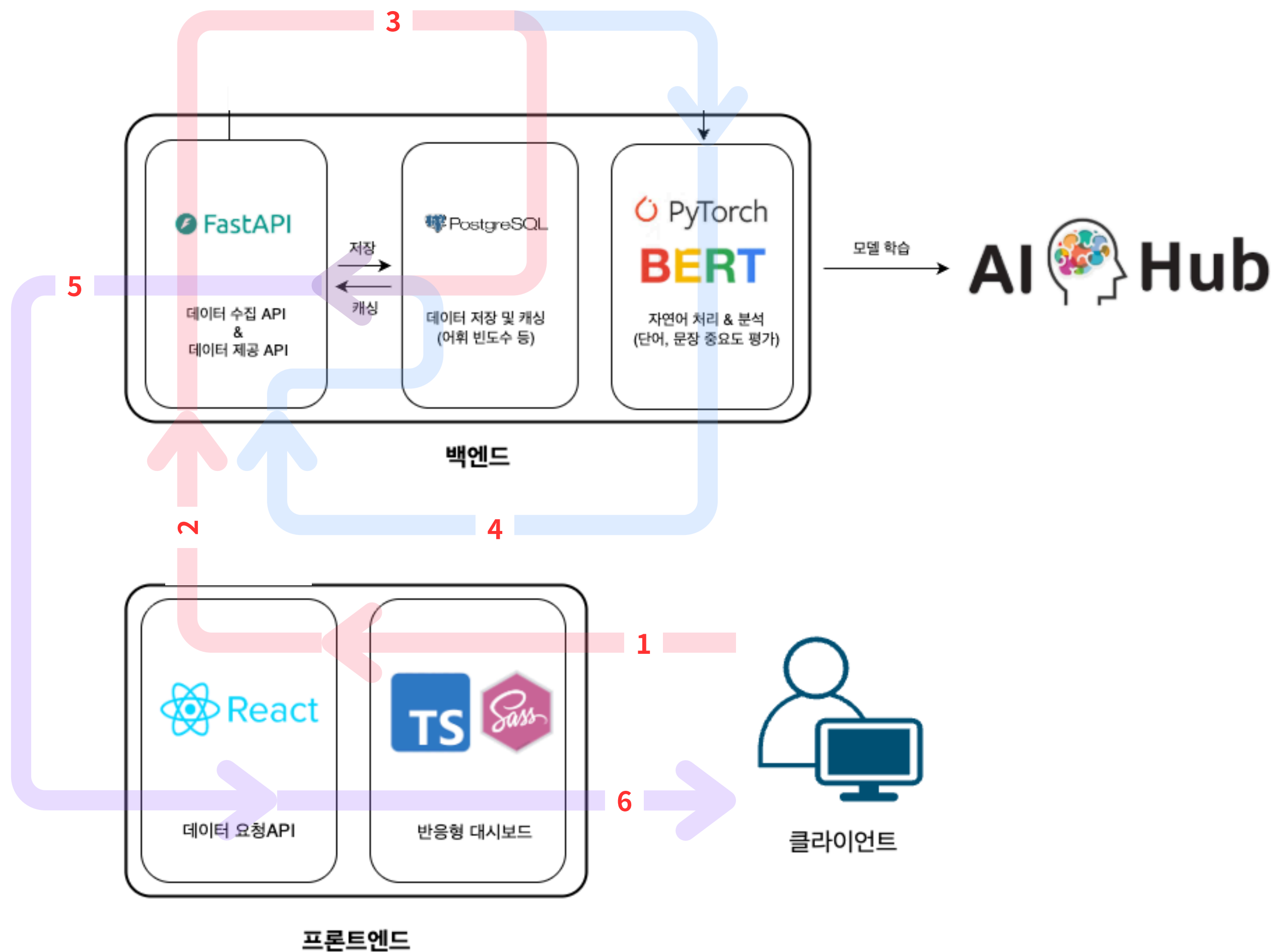
키워드 설명

품사 분석 시각화

실시간 분석

03

시스템 설계



시스템 구현 결과

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

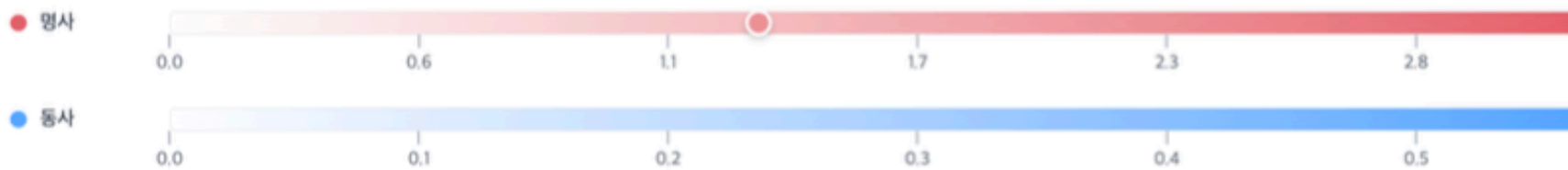
제품

공사계약

수의계약

그 조항에서는 우수제품의 구매에 대해서만 나와 있지 그렇다고 해서 그 제품을 생산한 회사에 공사계약까지 수의계약을 해야 된 다라고는 나와 있지 않습니다. 그렇다는 얘기는 제가 볼 때는 공사까지 수의계약을 하는 것은 잘못된 특혜다 이런 생각이 들고요. 전혀 상관없는데, 예를 들어서 그 제품 생산한다고 해서 그 회사가 반드시 그 공사도 가장 잘하고 가장 저렴하게 한다는 보장도 없고 오히려 자칫 잘못하면 공사 같은 경우에는 입찰을 하는 게 더 나은데도 불구하고 같이 함께 수의계약을 함으로써 가격이 훨씬 더 높아질 우려가 있습니다.

"공사계약" 키워드 점수 범례



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

품사별 분석 결과

품사 분석 과정에는 Mecab 라이브러리를 사용하였습니다.

명사 워드클라우드

회사 계약 애기 제품 공사
우수 계약 애기 제품 공사
특혜 계약 애기 제품 공사
우려 것 경우 생각 입찰 조항
예 생산 구매

동사 워드클라우드

하는
볼
나와있지
해야
들고요
대해서 하고
들어서하면

형용사 워드클라우드

없
나
높
상관없
그런
있

≡

연어린이집, 과속방지턱

2025-09-09

창원시, 도로, 이용료, 남산, 통행료

2025-09-09

마산합포구, 문화위생, 재직확인

2025-09-09

스타필드, 착공, 허가

2025-09-09

자동차, 운송사업, 허가

2025-09-09

키워드 없음

2025-09-09

키워드 없음

2025-09-09

복합쇼핑몰

2025-09-09

맥주, 수입, 전동주

2025-09-09

재난지원금, 지급

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

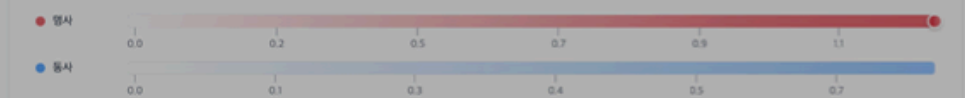
맥주

수입

전동주

지금 현재 모습이 작년 소위 때 모습하고 똑같은 겁니다. 그대로 지금 한 발짝도 앞으로 가지 않고 있어요. 한 두어 가지 말씀드릴 것은 맥주 수입을 시작한 것은 국내 맥주회사가 수입을 시작한 겁니다. 그리고 차관님께서 특정 회사를 보고 하는 건 아니라고 말씀했지만 특정 한 회사가 국내 시장의 60%를 차지하고 있어요. 회사가 여러 개 있고 하는 것 같으면 그것 하지만 그런 부분은 좀 살펴봐야 되고. 그다음에 말씀 중에 이것 이렇게 해도 전동주에 미치는 영향은 높지 않다. 누구 얘기인지는 모르겠어요. 제일 좋은 것은 뭐냐? 같이 **동맹** 해 주면 되는 거예요. 그렇지 않습니까? 왜 그걸 빼는 겁니까? **위스키** 때문에 같이할 수 없다. 도수 때문에? **소주**도, 전동주라는 게 45도짜리도 있지만 20도짜리도 많아요, 25도짜리도 있고. **일본**적으로 해 가지고 왜 그렇게 제끼고..... 예를 들면 전동주 **문배**다 그다음에 한산의 **소주**잖아 이것 **일본** 분이 맥주 쪽으로, 앞으로 또 막걸리도 **놀리니**까 이런 쪽으로 하면 **주요** 변화가 안 **일으켜지나**? 객관적으로 **일으킨다** 안 **일으킨다** 그 이전에 그 업계에서는 문제를 지적하고 있는 거예요. 그런데 작년하고 똑같이 이렇게 하면서 이번에는 그냥 **넘어가고** 내년에 **정책**을 마련한다? 나는 마련하는 무슨 근거가 없는 것 같아요. 영향이 없다? 그 **업계**는 있다는데

"전동주" 키워드 점수 범례



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

품사별 분석 결과

품사 분석 과정에는 Mecab 라이브러리를 사용하였습니다.

명사 워드클라우드

동사 워드클라우드

들면 되고 보고 미치는 할
문건
문건
문건

04 키워드별 분석 결과 비교

시스템 구현 결과

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

창원시

쓰레기

수거

음식물류

폐기물

창원시의 **음식물** 쓰레기 수거 **관련**해서 **지난번**에도 **민원** **올렸었는데** **개선된** 부분이 하나도 없네요. **날씨**가 추워서 **언** 것도 아니고, **음식물**은 저희 집에서 **내놓은** 음식물이 **맞고**, 다 수거가 되지 않았음에도 **칩은 빠져있네요**. **음식물류** 폐기물 수거 시 좀 더 세심한 주의를 기울여 **깨끗이** 수거가 될 수 **있도록** 수거업체를 지도 요청 **드립니다**.

"창원시" 키워드 점수 범례



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

창원시

쓰레기

수거

음식물류

폐기물

창원시의 **음식물** 쓰레기 수거 **관련**해서 **지난번**에도 **민원** **올렸었는데** **개선된** 부분이 하나도 없네요. **날씨**가 추워서 **언** 것도 아니고, **음식물**은 저희 집에서 **내놓은** 음식물이 **맞고**, 다 수거가 되지 않았음에도 **칩은 빠져있네요**. **음식물류** 폐기물 수거 시 좀 더 세심한 주의를 **기울**여 **깨끗이** 수거가 될 수 **있도록** 수거업체를 지도 요청 **드립니다**.

"쓰레기" 키워드 점수 범례



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

창원시

쓰레기

수거

음식물류

폐기물

창원시의 **음식물** 쓰레기 수거 **관련**해서 **지난번**에도 **민원** **올렸었는데** **개선된** 부분이 하나도 없네요. **날씨**가 추워서 **언** 것도 아니고, **음식물**은 저희 집에서 **내놓은** 음식물이 **맞고**, 다 수거가 되지 않았음에도 **칩은 빠져있네요**. **음식물류** 폐기물 수거 시 좀 더 세심한 주의를 **기울**여 **깨끗이** 수거가 될 수 **있도록** 수거업체를 지도 요청 **드립니다**.

"폐기물" 키워드 점수 범례



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

시스템 구현 결과

키워드 분석 결과

키워드를 클릭하시면 해당 키워드가 추출된 이유를 확인하실 수 있습니다.

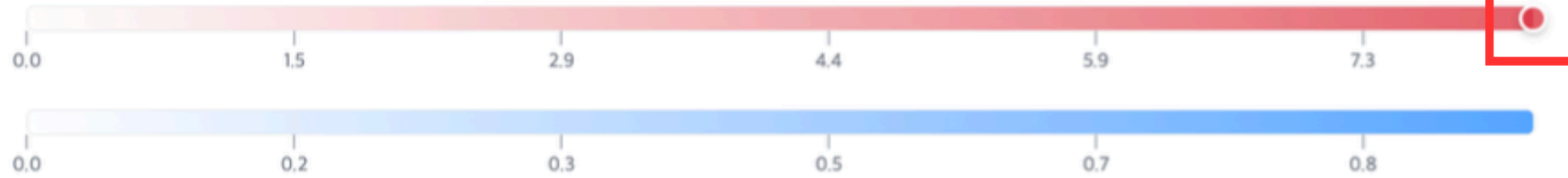
회의

정책

회의실에는 **책상**, 의자, 창문, 시계, 종이, 전화기, 가방, 신발, 컴퓨터, 모니터, 컵, 우산, 자동차, 달력, 사과, 기차가 있었다. 그러나 발표자는 이런 물건들과는 무관하게 새로운 정책 방향만을 강조했다.

"회의" 키워드 점수 범례

● 명사
● 동사



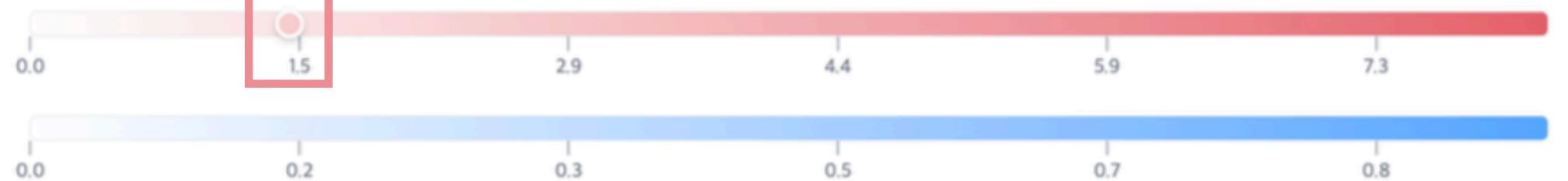
점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

회의 키워드에 대한 하이라이트 결과 중
회의실의 attention score 시각화 결과

회의 키워드에 대한 하이라이트 결과 중
책상의 attention score 시각화 결과

"회의" 키워드 점수 범례

● 명사
● 동사



점수가 높을수록 키워드 추출에 더 많이 기여한 단어입니다.

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)

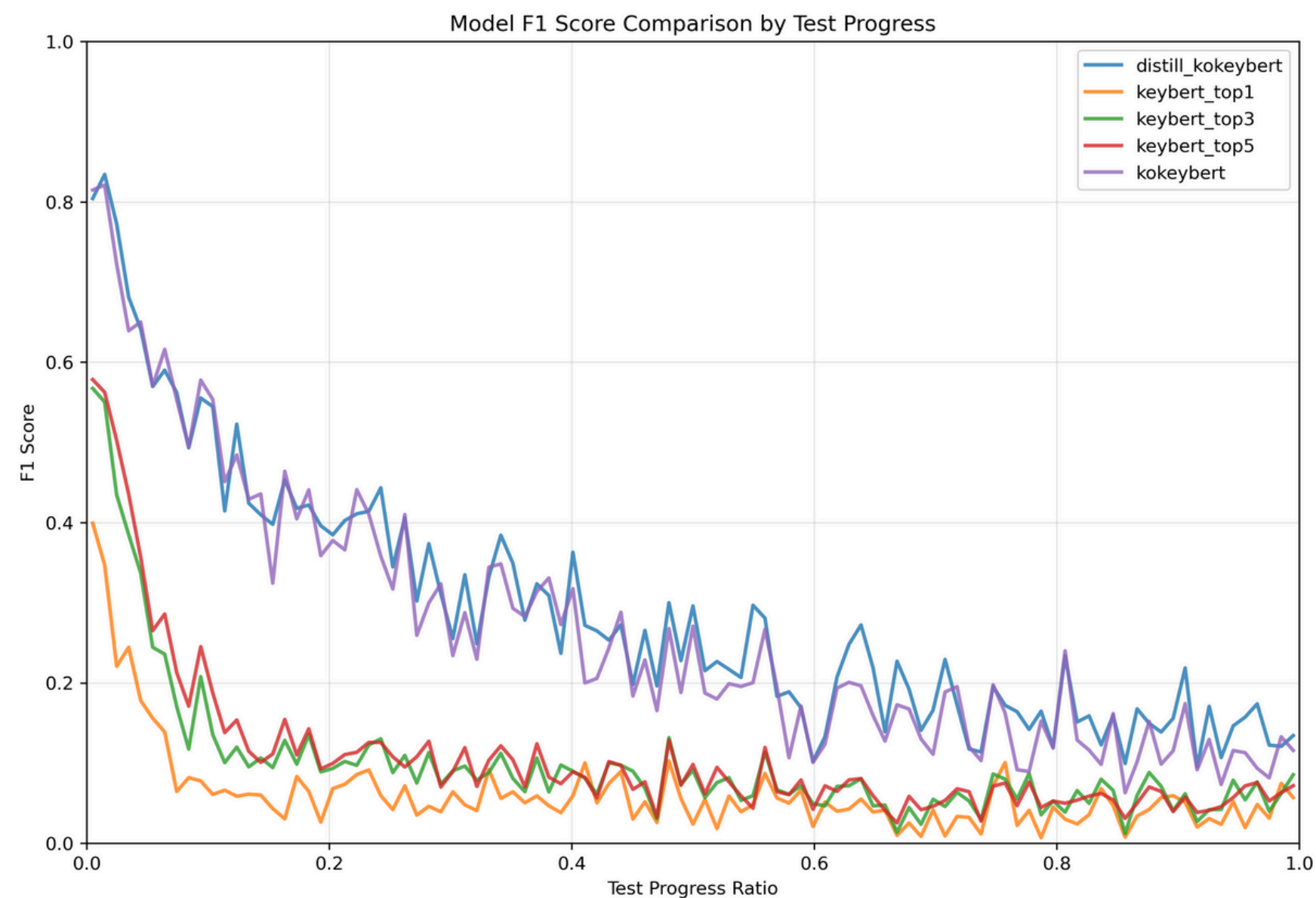
균형성

Precision과 Recall의 균형 잡힌 성능

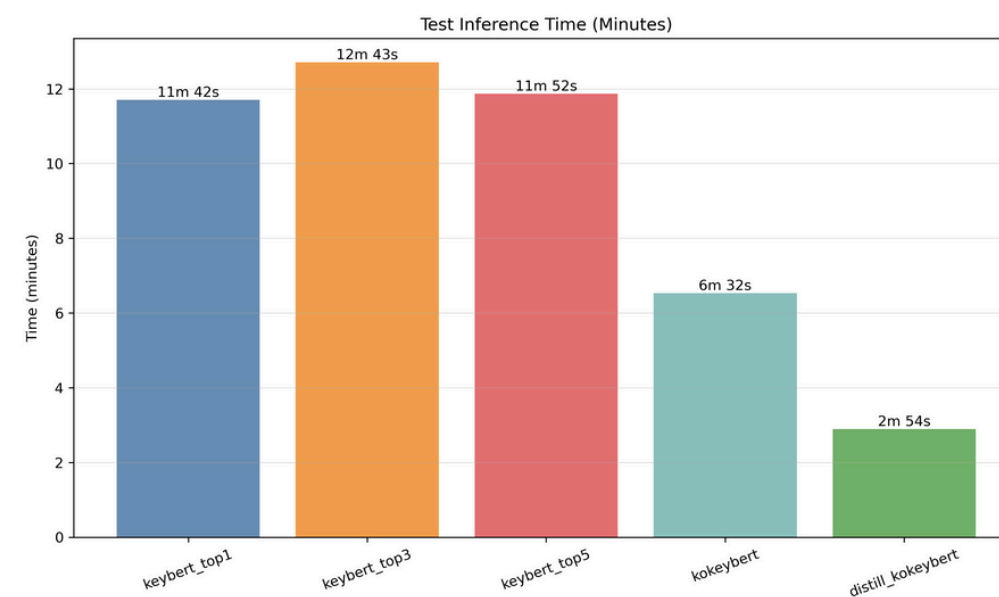
경량화

Teacher 모델 대비 5.75배 빠른 추론 속도

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.

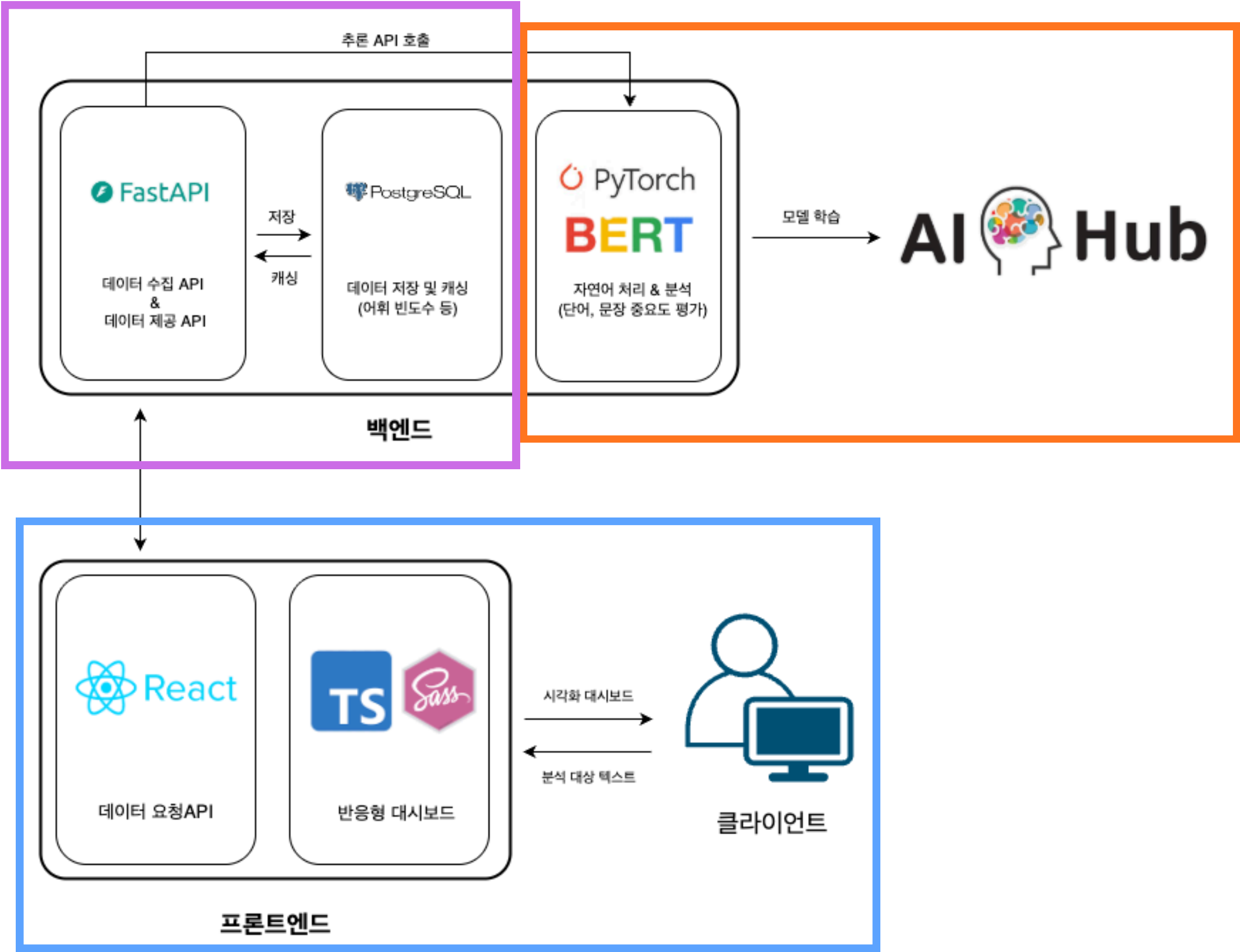


실험환경: MacBook M3 Air

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

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