특허 문서 요약 웹 서비스

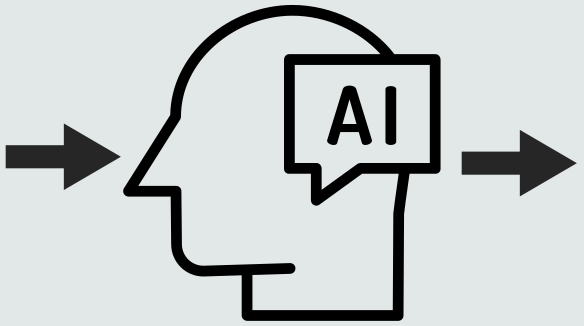
8조 팀장 심민기
김천구
이동훈
박종민

6/8(수) 최종발표 PPT

순서

수행배경 및 목표

기존 특허 문서의 구조



요약된 문서 제공

다양한 분야 도움

- 대부분 특허 요약문은 요약 없이 청구항을 그대로 삽입
- 해당 요약문만으론 이해가 어려움

시스템 요구사항

회사 측 요구 분석

웹서비스 제공

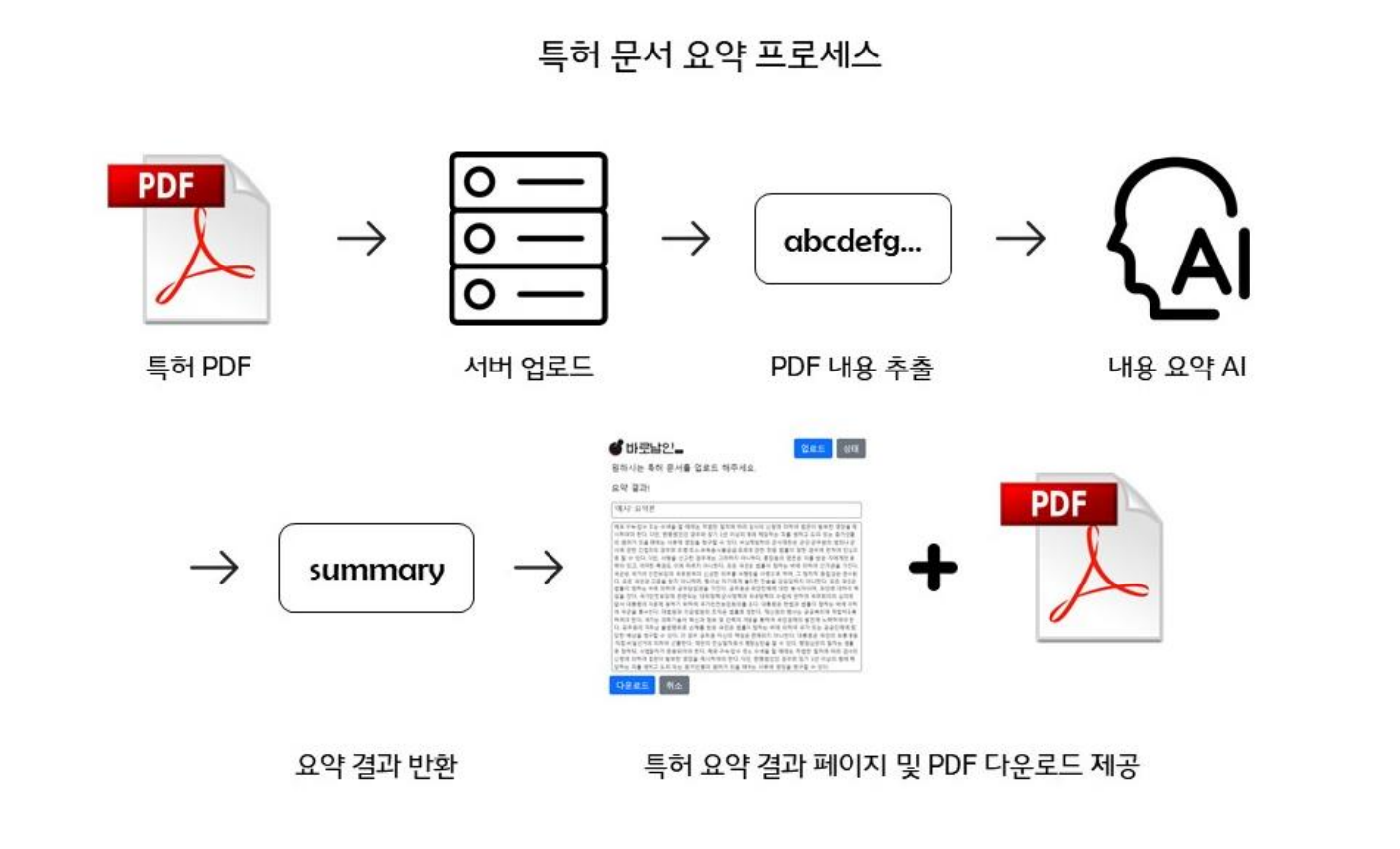
- 기존 회사 서비스 ‘바로날인’의 부가서비스 형태로 제공
- 기존 Spring에 ‘One Page Web’ 페이지 구현 요구

AI API 제공

- Spring 서버를 활용하여 HTTP API 구성(BM 활용)
- 인증 토큰을 활용한 선택적 결과 반환 대응

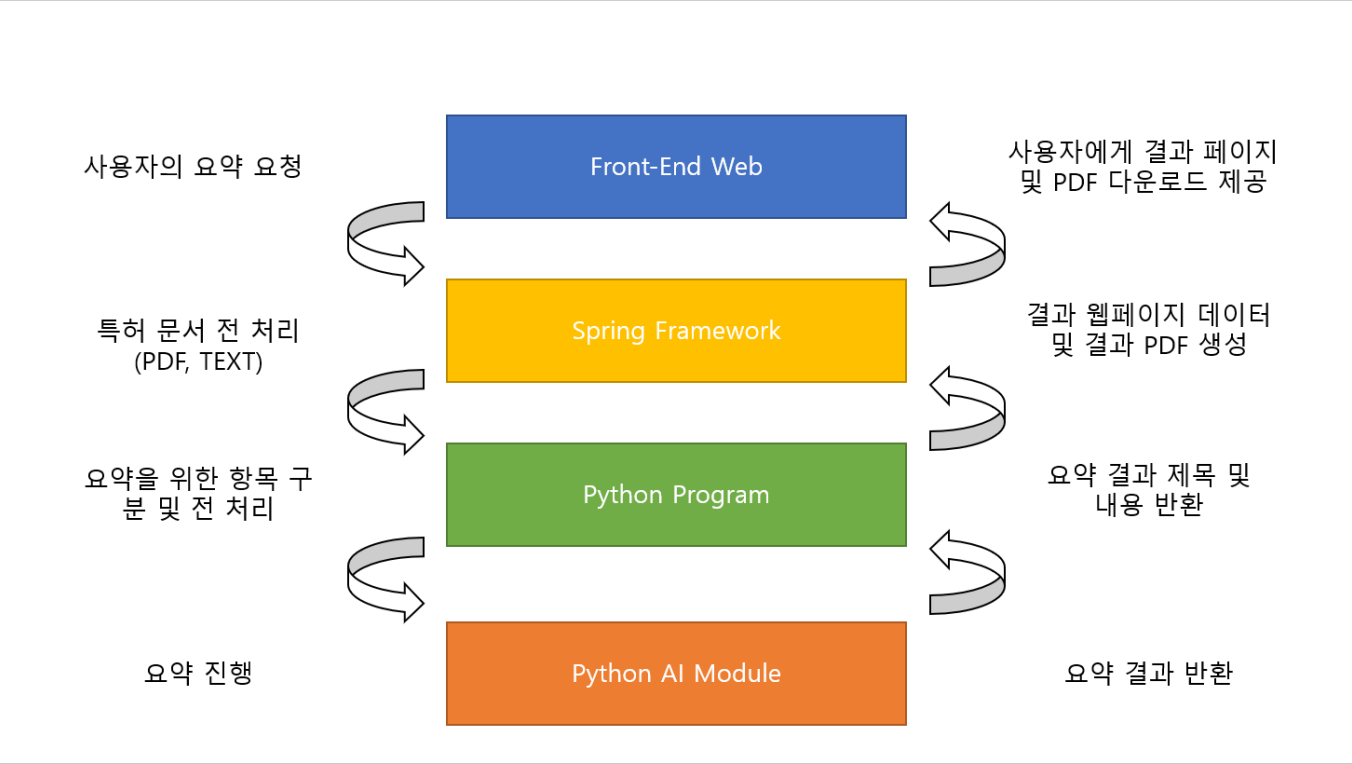
시스템 요구사항

전체 프로세스 정의



시스템 요구사항

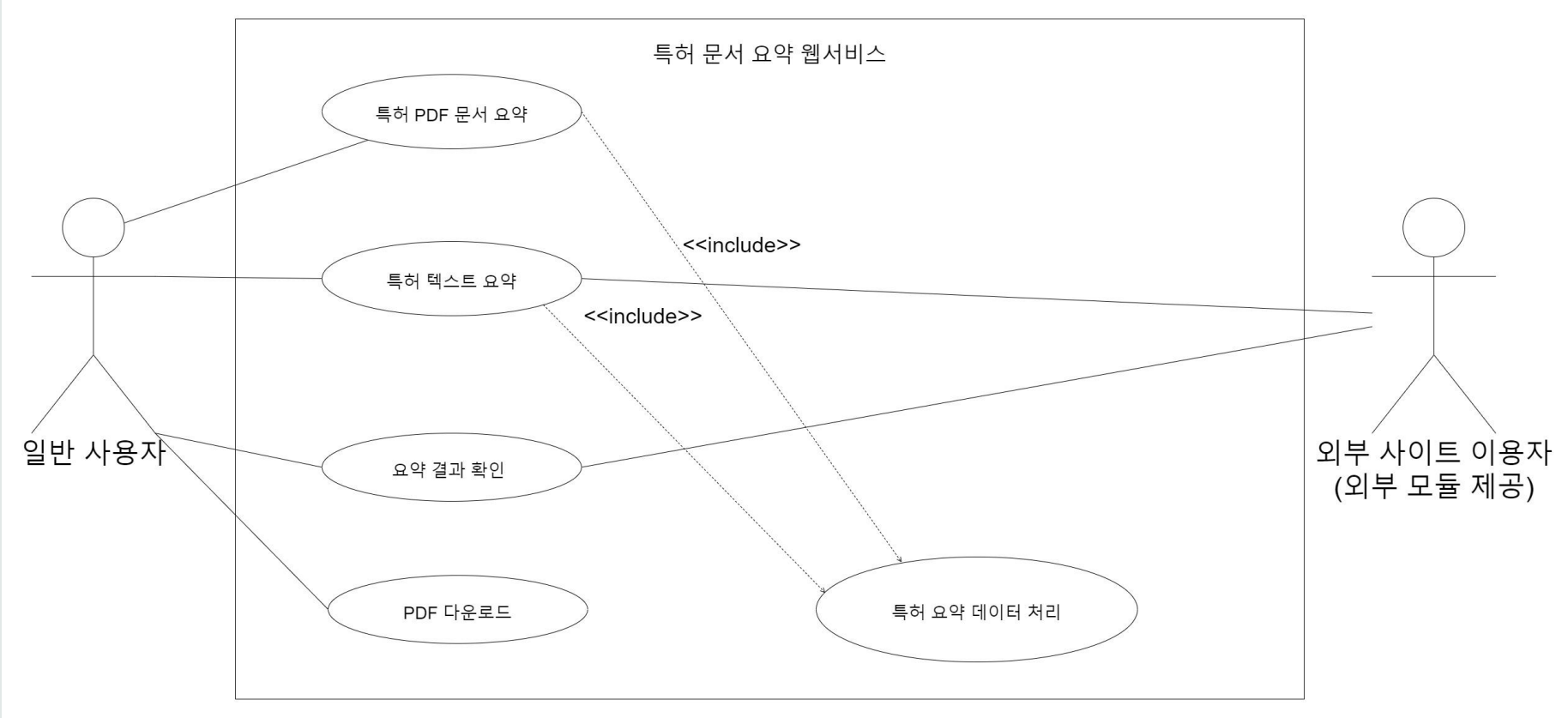
계층 분석



전체 시스템의 계층 분석

시스템 설계내용

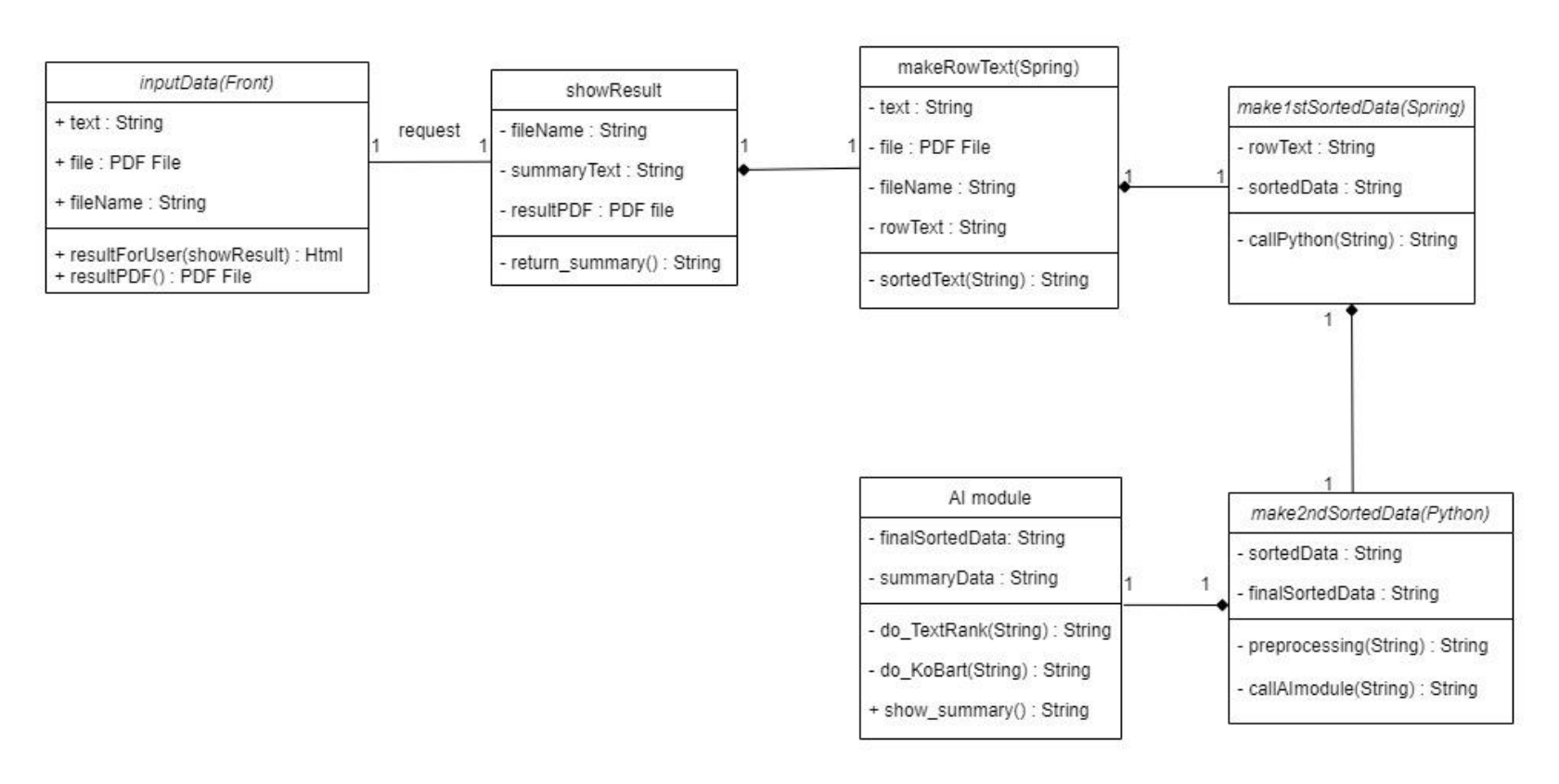
전체 프로세스의 다이어그램화



USE CASE Diagram

시스템 설계내용

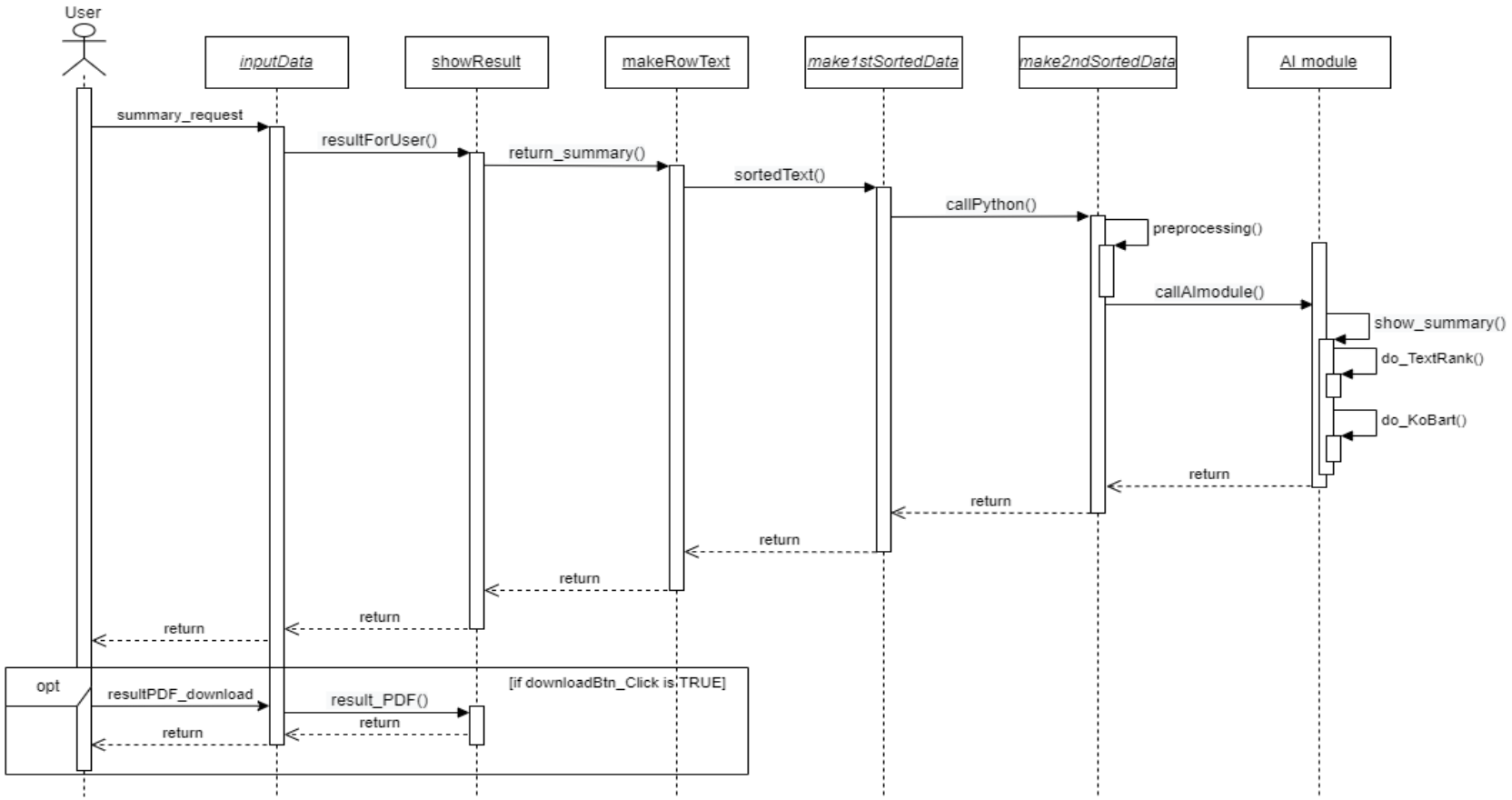
전체 프로세스의 다이어그램화



CLASS Diagram(구조)

시스템 설계 (구조 및 동작)

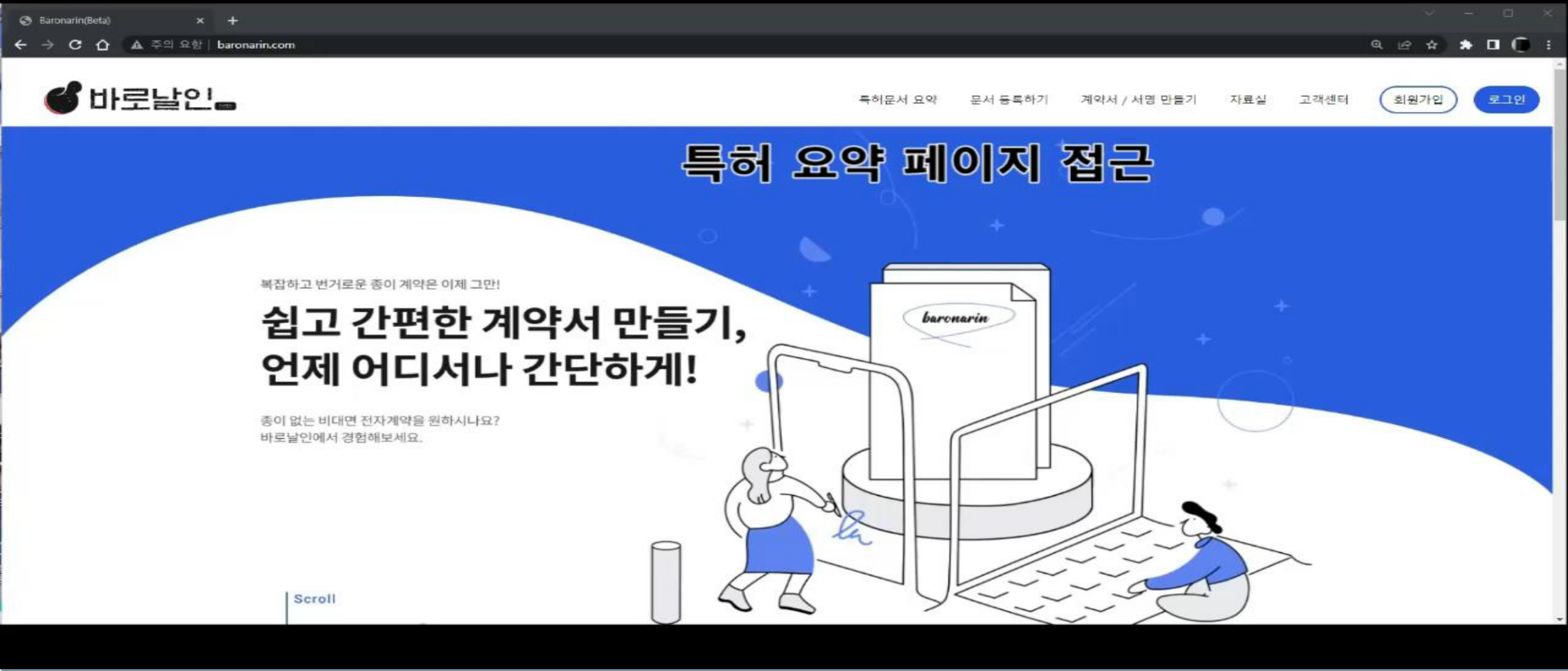
전체 프로세스의 다이어그램화



SEQUENCE Diagram(동작)

수행 결과(시연 영상)

https://youtu.be/-jgqZz2Kaq4



수행 결과

FRONT-END & BACK-END

- 기존 웹서버에 개발 코드를 이식해야 하므로 ‘ONE-PAGE’ 이지만 React 활용 불가.

  (∵ 개발 코드 때문에 JSP 기반 웹서버를 React로 모두 변경하는 것은 비용 大)

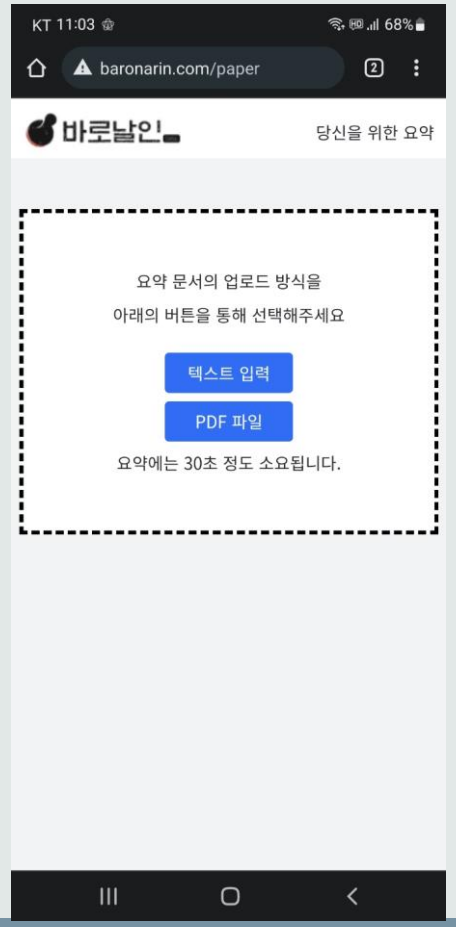
- JSP에 Bootstrap CSS를 활용하고, style 속성 변경 함수를 활용해 React의 DOM을 약소하게 모방하는 방식으로 ‘ONE-PAGE’ 구현

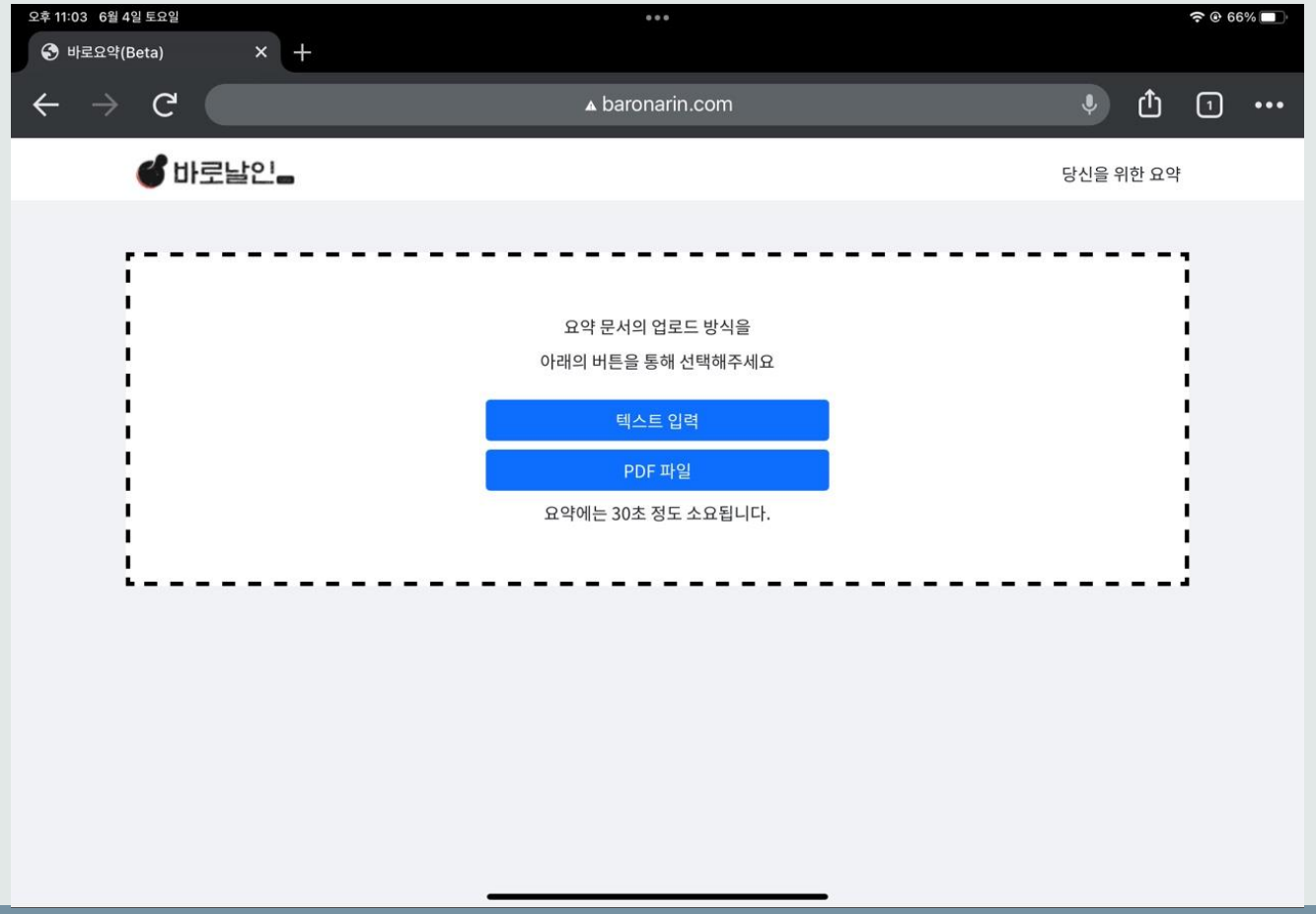
  (∴ 모바일,태블릿,PC 모두 반응형 Web으로 대응)

- AJAX를 활용해 불필요한 페이지 전체 리로드를 줄이고, 부드럽게 동작하도록 구현

수행 결과

FRONT-END(반응형 Web)





수행 결과

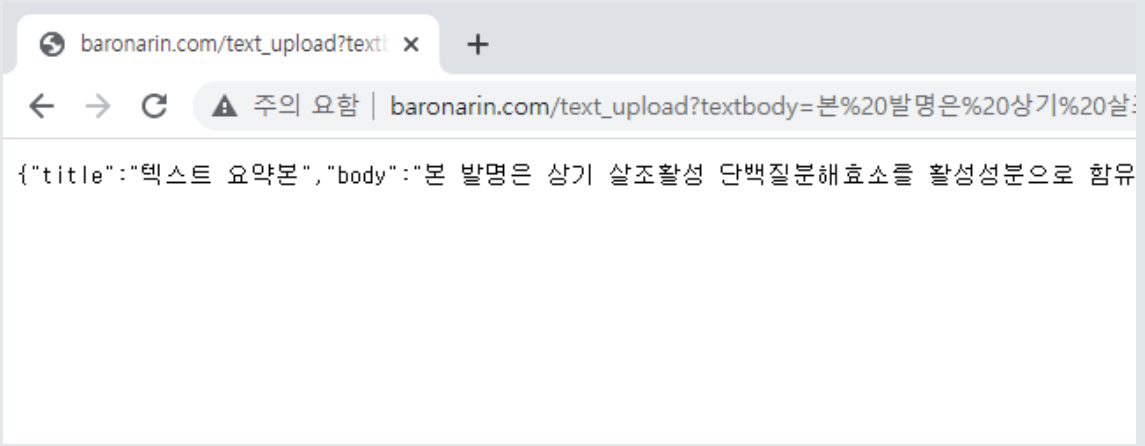
FRONT-END & BACK-END(Spring boot)

- PDF 텍스트 추출
- AI 실행 및 결과 반환
- JSP에 결과 전달
- PDF 다운로드

활용 패키지

- GET 메소드를 활용하여 외부에서 JSON 형태로 특허 텍스트 입력

  -> HTTP API 형태로 활용할 수 있도록 구현



수행 결과

AI

- AI 모델 결정 : KoBART
- KoBART 모델 환경 구축 및 실습
- 데이터 셋에 맞게 학습 및 파인튜닝

Data

- Dacon 한국어 문서 생성요약 AI 경진대회 의 학습 데이터를 활용함
- 학습 데이터에서 임의로 Train / Test 데이터를 생성함
- 데이터 탐색에 용이하게 tsv 형태로 데이터를 변환함
- Data 구조
  - Train Data : 34,242
  - Test Data : 8,501
- default로 data/train.tsv, data/test.tsv 형태로 저장함

news	summary
뉴스원문	요약문

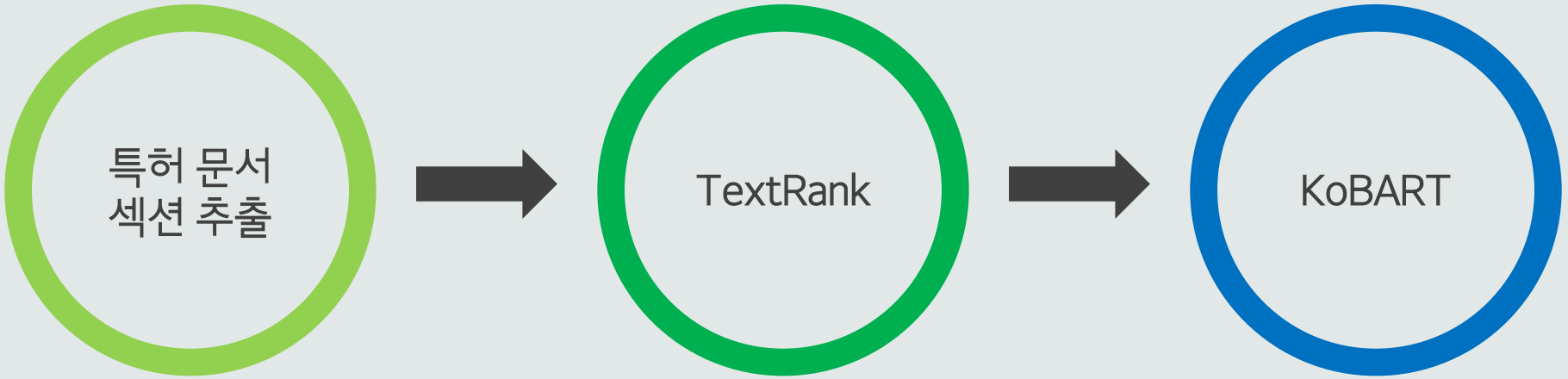
- 참조 데이터
  - AIHUB 문서 요약 데이터 (https://aihub.or.kr/aidata/8054)

```
import torch
from transformers import PreTrainedTokenizerFast
from transformers import BartForConditionalGeneration

tokenizer = PreTrainedTokenizerFast.from_pretrained('digit82/kobart-summarization')
model = BartForConditionalGeneration.from_pretrained('digit82/kobart-summarization')

text = """
1일 오후 9시까지 최소 20만3220명이 코로나19에 신규 확진됐다. 또다시 동시간대 최다 기록으로, 사상 처음 20만명대에 진
방역 당국과 서울시 등 각 지방자치단체에 따르면 이날 0시부터 오후 9시까지 전국 신규 확진자는 총 20만3220명으로 집계
국내 신규 확진자 수가 20만명대를 넘어선 것은 이번이 처음이다.
동시간대 최다 기록은 지난 23일 오후 9시 기준 16만1389명이었는데, 이를 무려 4만1831명이나 웃돌았다. 전날 같은 시간
확진자 폭증은 3시간 전인 오후 6시 집계에서도 예견됐다.
오후 6시까지 최소 17만8603명이 신규 확진돼 동시간대 최다 기록(24일 13만8419명)을 갈아치운 데 이어 이미 직전 0시 기
17개 지자체별로 보면 서울 4만6938명, 경기 6만7322명, 인천 1만985명 등 수도권이 12만5245명으로 전체의 61.6%를 차지
비수도권에서는 7만7975명(38.3%)이 발생했다. 제주를 제외한 나머지 지역에서 모두 동시간대 최다를 새로 썼다.
부산 1만890명, 경남 9909명, 대구 6900명, 경북 6977명, 충남 5900명, 대전 5292명, 전북 5150명, 울산 5141명, 광주 5
집계를 마감하는 자정까지 시간이 남아있는 만큼 2일 0시 기준으로 발표될 신규 확진자 수는 이보다 더 늘어날 수 있다. 이
한편 전날 하루 선별진료소에서 이뤄진 검사는 70만8763건으로 검사 양성률은 40.5%다. 양성률이 40%를 넘은 것은 이번이
이날 0시 기준 신규 확진자는 13만8993명이었다. 이를 연속 13만명대를 이어갔다.
"""

text = text.replace('\n', ' ')

raw_input_ids = tokenizer.encode(text)
input_ids = [tokenizer.bos_token_id] + raw_input_ids + [tokenizer.eos_token_id]

summary_ids = model.generate(torch.tensor([input_ids]),  num_beams=4,  max_length=512,  eos_token_id=1)
tokenizer.decode(summary_ids.squeeze().tolist(), skip_special_tokens=True)

'1일 0 9시까지 최소 20만3220명이 코로나19에 신규 확진되어 역대 최다 기록을 갈아치웠다.'
```

수행 결과

AI

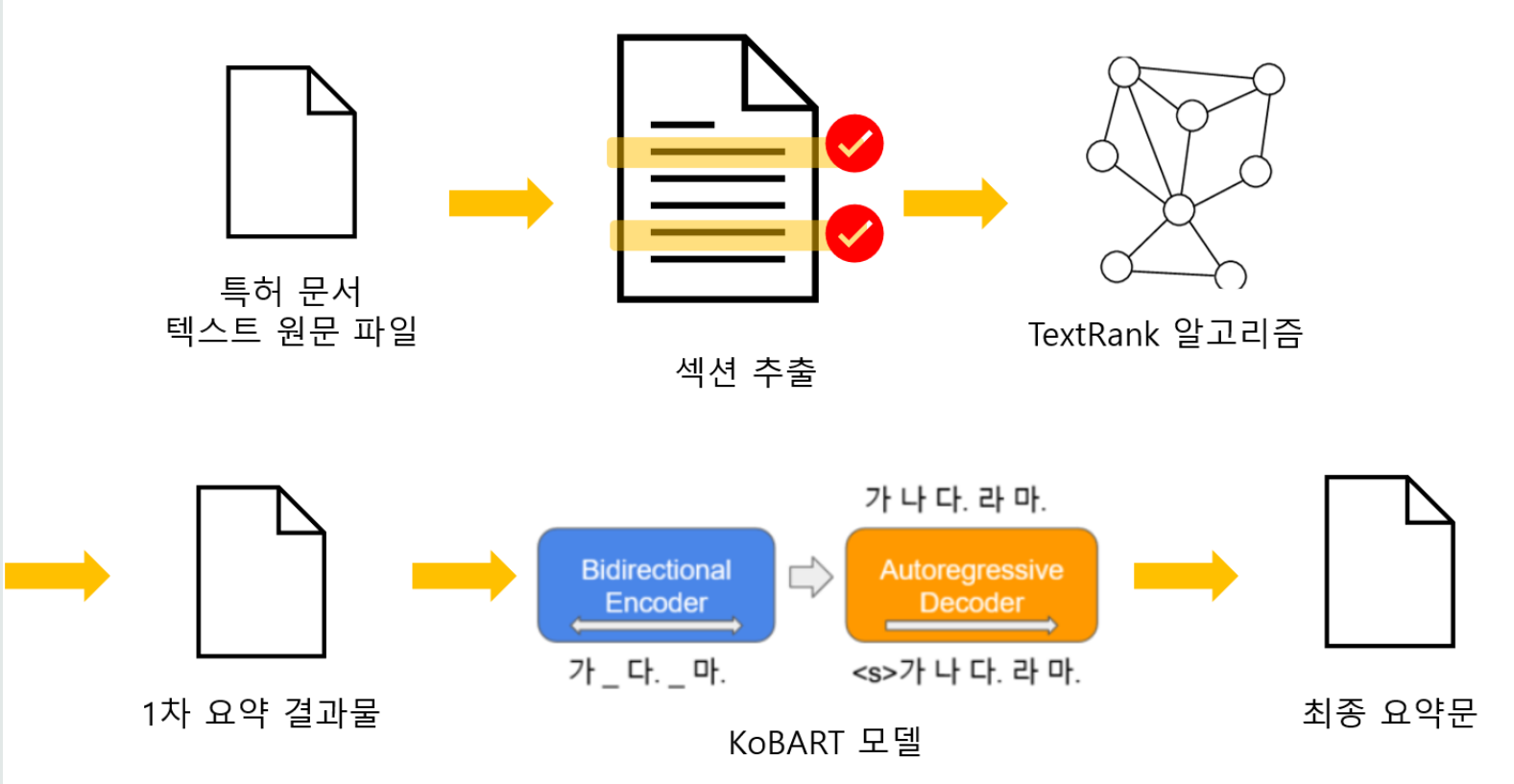
- KoBART에 넣기에는 원문 길이가 길다(Max Length문제)
- 추가적인 전처리 : 섹션추출, TextRank
- 의미있는 섹션만 추출한뒤 이 섹션을 한번 더 Textrank를 넣어 추출요약 진행 후, 튜닝한 KoBART모델을 사용해 생성요약을 진행함



수행 결과

AI

- 최종 AI 프로세스



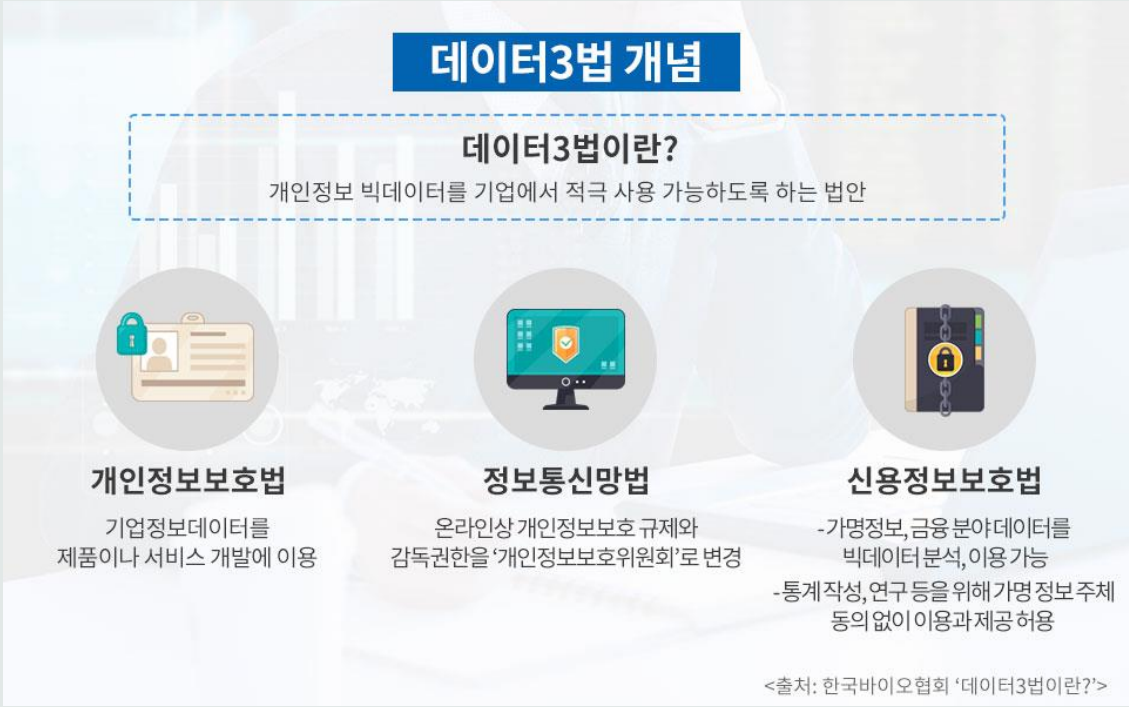
기대효과 및 활용방안

- 특허 문서 비교에 많은 도움

- 시간 및 비용 절약으로 사회적 후생 증가

- 최근 개정된 데이터 3법에 대응하여 많은 데이터 처리에 도움

- DB를 사용하지 않고 개발하여 높은 이식성이 장점. (ex. 바로날인)



바로날인 beta

http://baronarin.com/

발표를 마치겠습니다.

감사합니다.