특허 문서 요약 웹 서비스

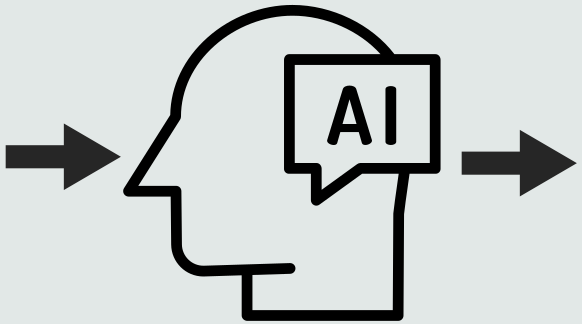
8조 팀장 심민기
김천구
이동훈
박종민

순서

1. 요구분석 및 정의
2. 설계 내용
3. 진행상황
4. 이슈사항 및 해결방안
5. 향후 일정

요구분석 및 정의

기존 특허 문서의 구조



요약된 문서 제공

다양한 분야 도움

- 청구항을 그대로 삽입
- 해당 요약문만으론 이해가 어려움

요구분석 및 정의

프로젝트 진행

Front-End

- 웹 사이트 디자인
- 파일 및 텍스트 업로드

Back-End

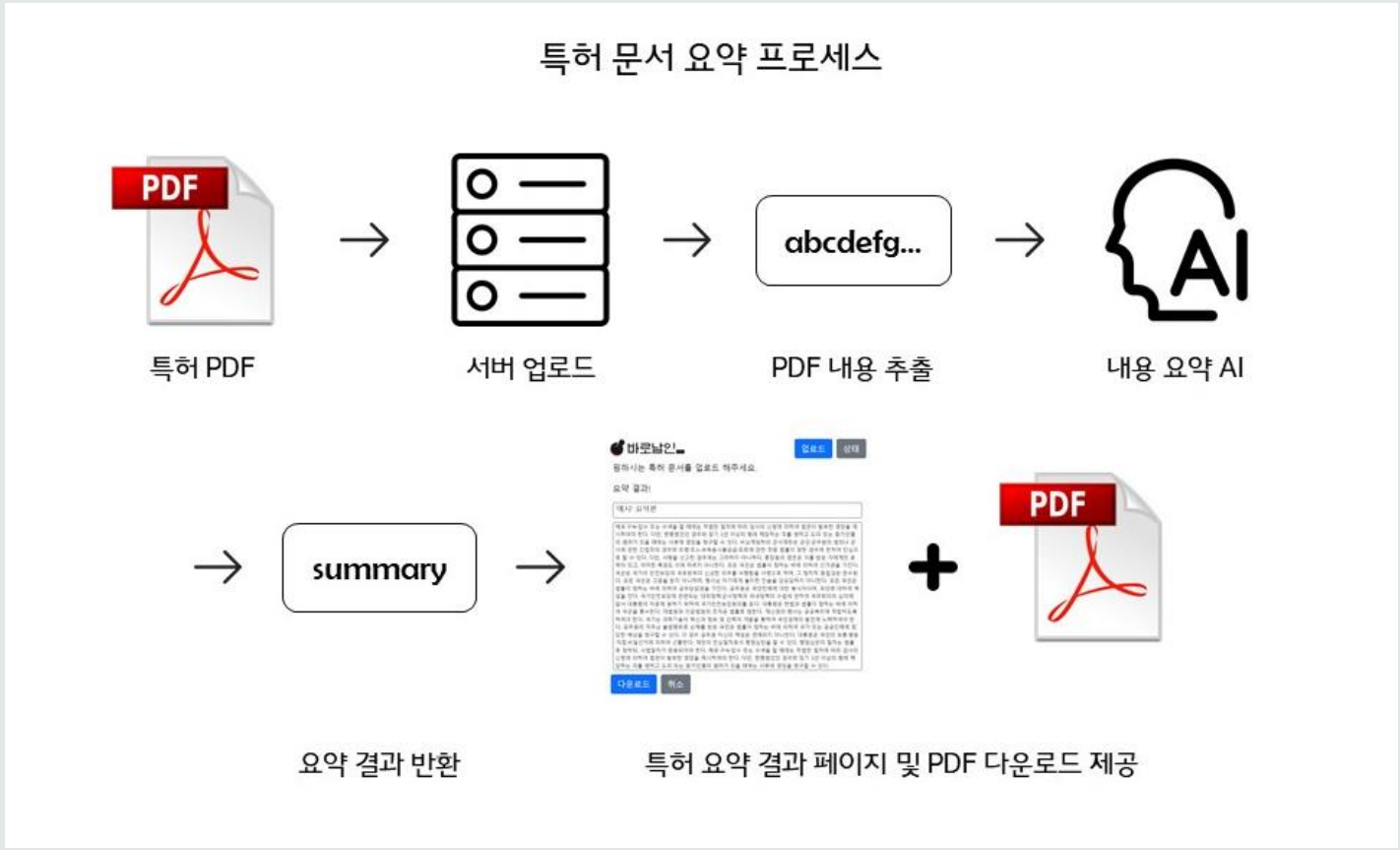
- PDF 텍스트 추출 및 PDF 다운로드
- AI 파이썬 프로그램 자바환경 실행

AI

- 입력 텍스트 요약 및 결과 텍스트 서버 반환
- 필요한 AI 모델 선정

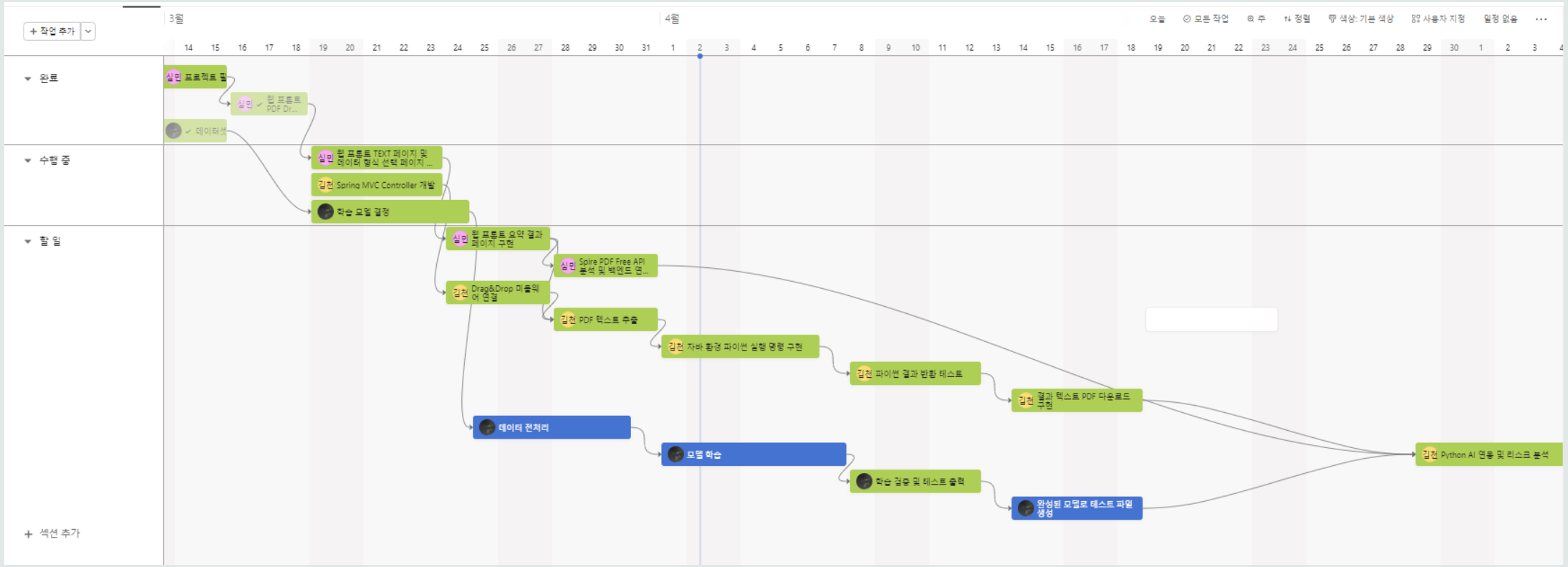
설계 내용

전체 프로세스 구성



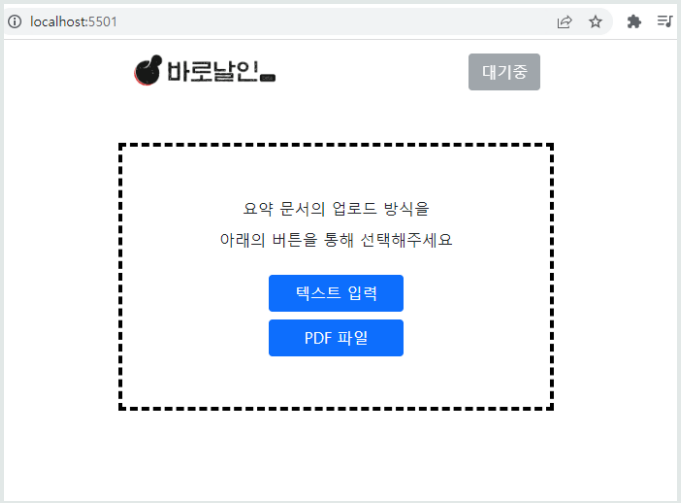
설계 내용

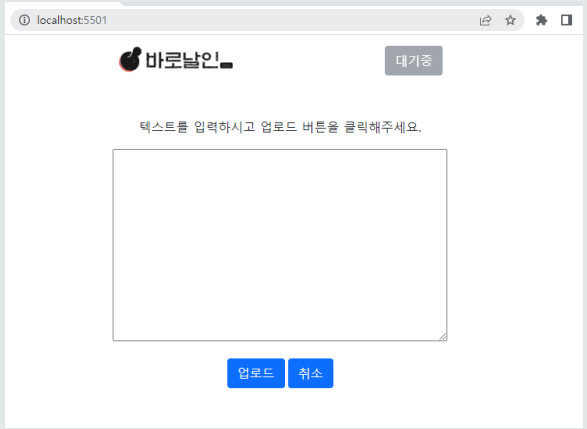
간트 차트


+ 작업 추가
3월
4월
오늘
모든 작업
주
정렬
색상: 기본 색상
사용자 지정
일정 없음
14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 1 2 3 4
완료
수행 중
할 일
+ 섹션 추가
심민 프로젝트 팀
심민 웹 프롬트 PDF Dr...
데이터셋
심민 웹 프롬트 TEXT 페이지 및 데이터 형식 선택 페이지 ...
김천 Spring MVC Controller 개발
학습 모델 결정
심민 웹 프롬트 요약 결과 페이지 구현
심민 Spire PDF Free API 분석 및 백엔드 연...
김천 Drag&Drop 미들웨어 연결
김천 PDF 텍스트 추출
김천 자바 환경 파이썬 실행 명령 구현
김천 파이썬 결과 반환 테스트
김천 결과 텍스트 PDF 다운로드 구현
데이터 전처리
모델 학습
학습 검증 및 테스트 출력
완성된 모델로 테스트 파일 생성
김천 Python AI 연동 및 리스크 분석

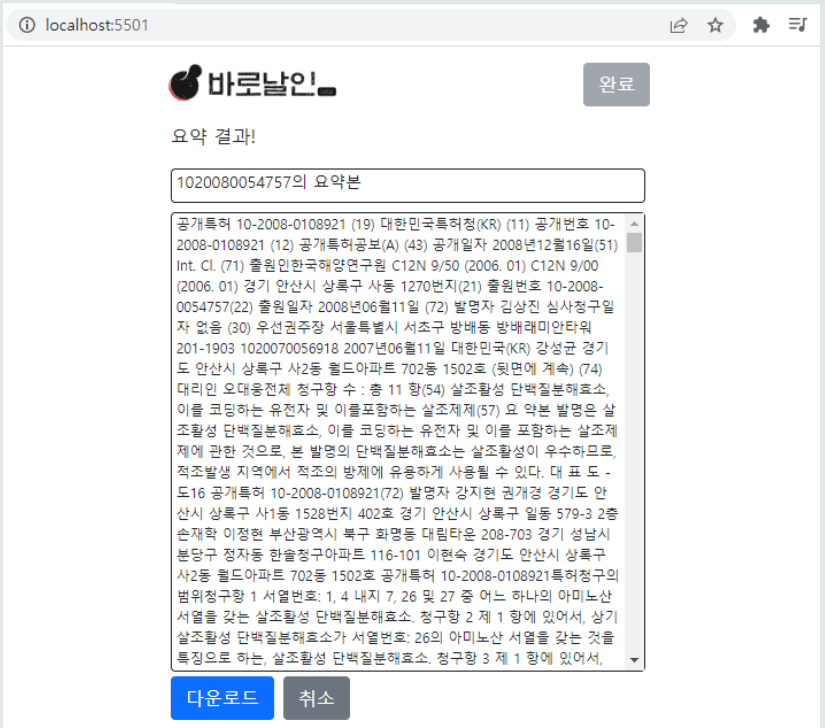
진행 상황

프론트 엔드


localhost:5501
바로날인
대기중
요약 문서의 업로드 방식을
아래의 버튼을 통해 선택해주세요
텍스트 입력
PDF 파일


localhost:5501
바로날인
대기중
텍스트를 입력하시고 업로드 버튼을 클릭해주세요.
업로드
취소


localhost:5501
바로날인
대기중
파일을 드래그 해서 넣어주시거나
아래 버튼을 클릭해서 파일을 선택해주세요.


localhost:5501
바로날인
완료
요약 결과!
1020080054757의 요약본
공개특허 10-2008-0108921 (19) 대한민국특허청(KR) (11) 공개번호 10-2008-0108921 (12) 공개특허공보(A) (43) 공개일자 2008년12월16일(51) Int. Cl. (71) 출원인한국해양연구원 C12N 9/50 (2006. 01) C12N 9/00 (2006. 01) 경기 안산시 상록구 사동 1270번지(21) 출원번호 10-2008-0054757(22) 출원일자 2008년06월11일 (72) 발명자 김상진 심사청구일자 없음 (30) 우선권주장 서울특별시 서초구 방배동 방배래미안타워 201-1903 1020070056918 2007년06월11일 대한민국(KR) 강성균 경기도 안산시 상록구 사2동 월드아파트 702동 1502호 (뒷면에 계속) (74) 대리인 오대웅전체 청구항 수 : 총 11 항(54) 살조활성 단백질분해효소, 이를 코딩하는 유전자 및 이를포함하는 살조제제(57) 요 약본 발명은 살조활성 단백질분해효소, 이를 코딩하는 유전자 및 이를 포함하는 살조제제에 관한 것으로, 본 발명의 단백질분해효소는 살조활성이 우수하므로, 적조발생 지역에서 적조의 방제에 유용하게 사용될 수 있다. 대 표 도 - 도16 공개특허 10-2008-0108921(72) 발명자 강지현 권개경 경기도 안산시 상록구 사1동 1528번지 402호 경기 안산시 상록구 일동 579-3 2층 손재학 이정현 부산광역시 북구 화명동 대림타운 208-703 경기 성남시 분당구 정자동 한솔청구아파트 116-101 이현숙 경기도 안산시 상록구 사2동 월드아파트 702동 1502호 공개특허 10-2008-0108921특허청구의 범위청구항 1 서열번호: 1, 4 내지 7, 26 및 27 중 어느 하나의 아미노산 서열을 갖는 살조활성 단백질분해효소. 청구항 2 제 1 항에 있어서, 상기 살조활성 단백질분해효소가 서열번호: 26의 아미노산 서열을 갖는 것을 특징으로 하는, 살조활성 단백질분해효소. 청구항 3 제 1 항에 있어서,
다운로드
취소

진행 상황

백엔드

드래그 & 드롭 파일 업로드

```java
@ResponseBody
@PostMapping(value = { "/pdf_upload" }, consumes = { "multipart/form-data" })
public JSONObject upload(@RequestParam(value="file", required=true) MultipartFile file) throws IOException
{
    String filename = file.getOriginalFilename();

    //사용자가 업로드한 파일 서버에 저장
    File result_file = fileService.MultipartFile_to_File(file);
    //pdf파일의 텍스트 추출
    String text = pdfService.getText_spire(result_file);
    //정규표현식으로 필요없는 띄어쓰기, 줄바꿈 제거
    text = text.replace(System.lineSeparator(), "");
    text = text.replaceAll("-\\s[0-9]+\\s-", "");
    text = text.replaceAll("\\[[0-9]+\\]", "");
    text = text.replace("                    ", "");
    text = text.replace(", ", ",");
    text = text.replace(". ", ".");
    text = text.replace(" .", ".");
    text = text.replace(". .", ".");

    String title = fileService.get_format_time()+filename.substring(0, filename.indexOf(".pdf"));

    //쓰레드 생성해서 파이썬에 전달해줄 텍스트 파일 만들기
    FileThread thread1 = new FileThread(title, text);

    thread1.start();
    try {
        //파일 완료는 동안 메인 쓰레드 대기
        thread1.join();
    } catch (InterruptedException e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }
    System.out.println("쓰레드 종료 : "+text);

    File textfile = new File("python" + File.separator+title+".txt");


    //파이썬 실행, 요약 텍스트 받아옴
    text = pythonService.execApachePy(textfile);

    textfile.delete();


    //json형태로 리턴
    HashMap<String,String> mymap = new HashMap<String,String>();
    mymap.put("title", filename.substring(0, filename.indexOf(".pdf"))+"의 요약본");
    mymap.put("body", text);
    JSONObject data = new JSONObject(mymap);


    return data;
}
```

PDF 파일 텍스트 추출

```java
public String getText_spire(File result_file )
{

    String text="";

    //if(file!=null && file.getOriginalFilename().contains(".pdf"))
    {
        try {
            System.out.println("파일이름 : "+result_file.getName());
            PdfDocument doc = new PdfDocument();
            doc.loadFromFile("temp"+File.separator+result_file.getName());
            PdfPageCollection pages = doc.getPages();

            StringBuilder sb=new StringBuilder();


            for(int i=0; i<pages.getCount();i++)
            {
                PdfPageBase page = pages.get(i);
                //SimpleTextExtractionStrategy strategy = new SimpleTextExtractionStrategy();
                text += page.extractText(true);
                sb.append(text);
            }
        /*
        FileWriter writer = new FileWriter("temp"+File.separator+"result.txt");

        writer.write(sb.toString());

        writer.flush();
        */
        doc.close();
        } catch (Exception e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
    }

    result_file.delete();

    return text;
}
```

AI 데모 버전 실행

```java
public String execApachePy(File text)
{
    int[] exitvalues = {0,1};
    String[] command = new String[4];
    command[0] = "python";
    command[1] = "python/kobart_demo.py";
    command[2] = text.getAbsolutePath();


    String result = "요약결과 : ";
    try {
        CommandLine commandLine = CommandLine.parse(command[0]);
        for (int i = 1, n = command.length; i < n; i++) {
            commandLine.addArgument(command[i]);
        }

        ByteArrayOutputStream outputStream = new ByteArrayOutputStream();
        PumpStreamHandler pumpStreamHandler = new PumpStreamHandler(outputStream);
        DefaultExecutor executor = new DefaultExecutor();
        executor.setStreamHandler(pumpStreamHandler);
        executor.setExitValues(exitvalues);//파이썬 프로그램 에러 리턴 시 exit() 종료 번호를 입력
        ExecuteWatchdog watchdog = new ExecuteWatchdog(100000);
        executor.setWatchdog(watchdog);

        executor.execute(commandLine);

        //System.out.println("output: " + outputStream.toString());
        result += outputStream.toString("euc-kr");
    } catch (Exception e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }


    return result;
}
```

진행 상황

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.'
```

이슈사항 및 해결방안

- **이슈사항**

AI 데모 버전에 필요한 기존 사용하던 버전의 라이브러리 설치가 안되는 문제

```
C:₩Users₩TZ1009>pip3 install torch==1.10.0
ERROR: Could not find a version that satisfies the requirement torch==1.10.0 (from versions: 1.11.0)
ERROR: No matching distribution found for torch==1.10.0
```

- **해결방안**

-최신버전을 사용하여도 동작함

-버전 차이로 인한 호환성에 문제 있는지 검사

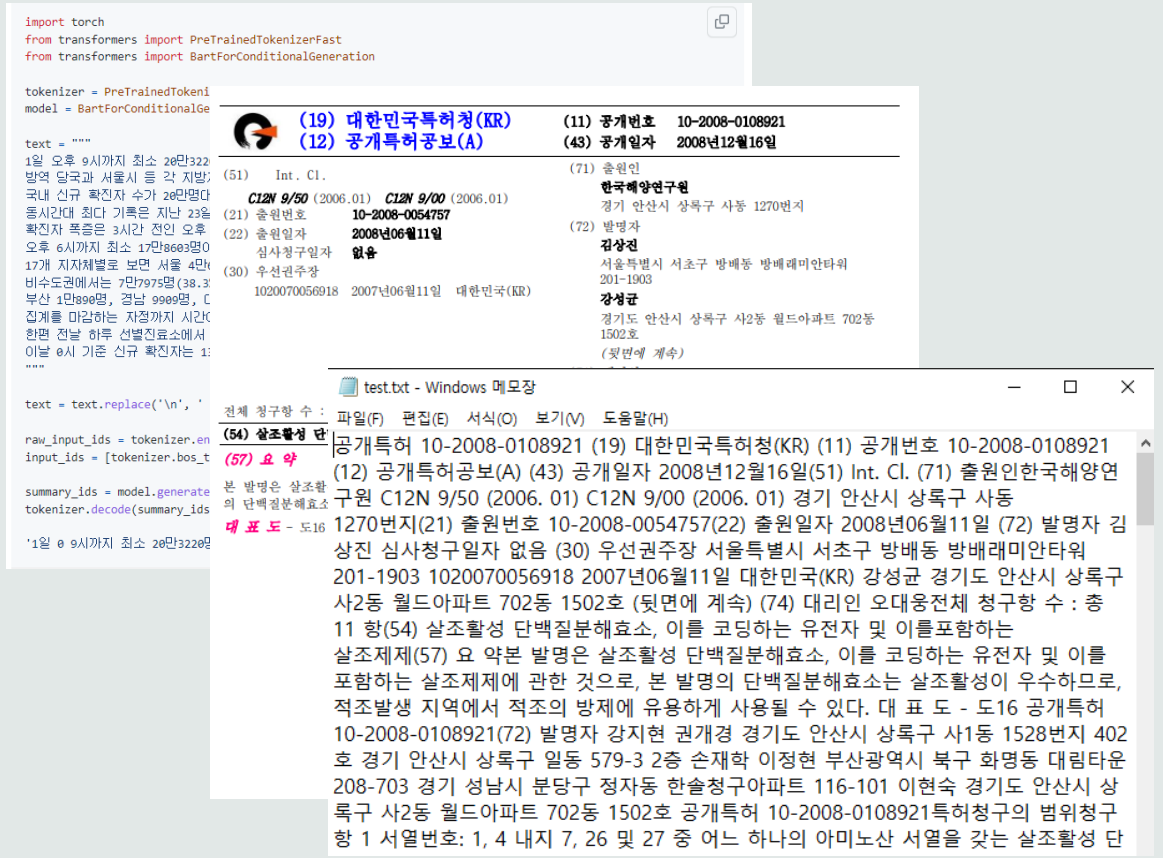
이슈사항 및 해결방안

- **이슈사항**

  PDF 추출 텍스트가 사용하던 모델과 호환되지 않는 문장구조

- **해결방안**

  -호환되는 구조를 위해 전처리 과정 추가



이슈사항 및 해결방안

- **이슈사항**

PDF 추출 텍스트가 KoBART가 처리할 수 있는 최대 토큰보다 긴 index range 문제

```
  File "C:\Users\TZ1009\AppData\Local\Programs\Python\Python310\lib\site-packages\t
    return torch.embedding(weight, input, padding_idx, scale_grad_by_freq, sparse)
IndexError: index out of range in self
```

- **해결방안**

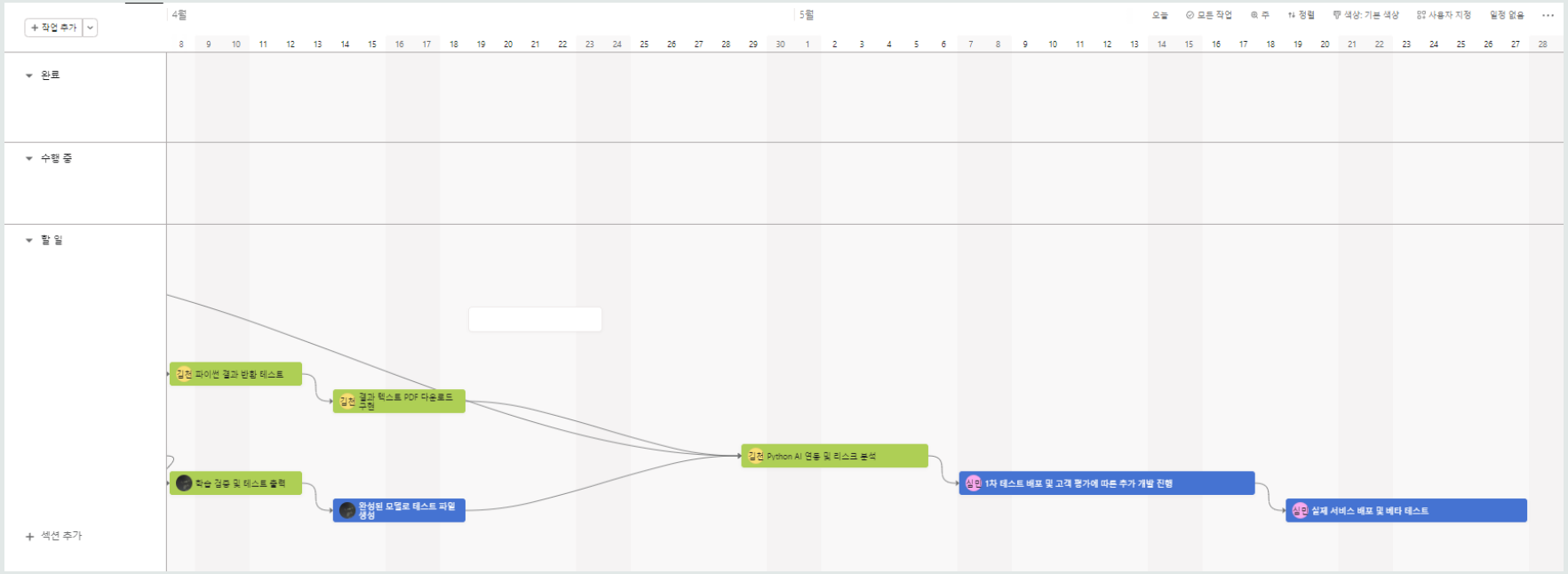
-textrank와 같은 알고리즘을 활용하여 의미있는 문장을 먼저 선별

-선별된 결과를 KoBART 입력으로 활용

향후 일정

간트 차트

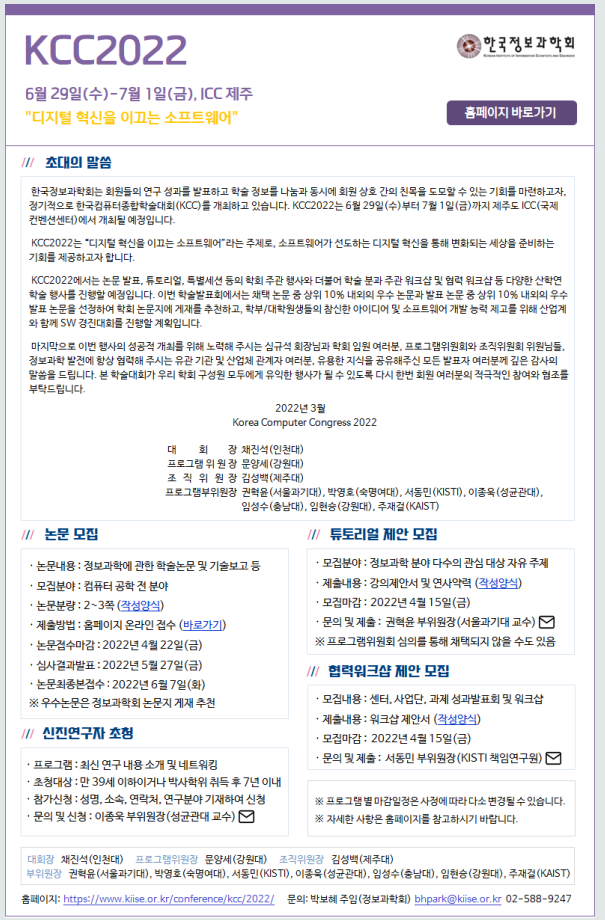
- AI 결과 반환 확인
- PDF파일 다운로드 구현
- 학습 검증



향후 일정

- **2022 한국컴퓨터종합학술대회 (KCC 2022) 발표**
  - 논문접수마감: 2022-04-22(금)
  - 심사결과발표: 2022-05-27(금)

- **AI Hub 학습데이터 활용 사업화 아이디어 해커톤**
  - 참가신청서 : 2022-03-31(목)
  - 해커톤 제안서 제출: 2022-04-11(월)
  - 결과발표: 2022-04-22(금)



KCC2022

한국정보과학회

6월 29일(수)~7월 1일(금), ICC 제주

"디지털 혁신을 이끄는 소프트웨어"

홈페이지 바로가기

/// 초대의 말씀

한국정보과학회는 회원들의 연구 성과를 발표하고 학술 정보를 나눔과 동시에 회원 상호 간의 친목을 도모할 수 있는 기회를 마련하고자, 정기적으로 한국컴퓨터종합학술대회(KCC)를 개최하고 있습니다. KCC2022는 6월 29일(수)부터 7월 1일(금)까지 제주도 ICC(국제컨벤션센터)에서 개최될 예정입니다.

KCC2022는 "디지털 혁신을 이끄는 소프트웨어"라는 주제로, 소프트웨어가 선도하는 디지털 혁신을 통해 변화되는 세상을 준비하는 기회를 제공하고자 합니다.

KCC2022에서는 논문 발표, 튜토리얼, 특별세션 등의 학회 주관 행사와 더불어 학술 분과 주관 워크샵 및 협력 워크샵 등 다양한 산학연 학술 행사를 진행할 예정입니다. 이번 학술발표회에서는 채택 논문 중 상위 10% 내외의 우수 논문과 발표 논문 중 상위 10% 내외의 우수 발표 논문을 선정하여 학회 논문지에 게재를 추천하고, 학부/대학원생들의 참신한 아이디어 및 소프트웨어 개발 능력 제고를 위해 산업계와 함께 SW 경진대회를 진행할 계획입니다.

마지막으로 이번 행사의 성공적 개최를 위해 노력해 주시는 심규석 회장님과 학회 임원 여러분, 프로그램위원회와 조직위원회 위원님들, 정보과학 발전에 항상 협력해 주시는 유관 기관 및 산업체 관계자 여러분, 유용한 지식을 공유해주신 모든 발표자 여러분께 깊은 감사의 말씀을 드립니다. 본 학술대회가 우리 학회 구성원 모두에게 유익한 행사가 될 수 있도록 다시 한번 회원 여러분의 적극적인 참여와 협조를 부탁드립니다.

2022년 3월

Korea Computer Congress 2022

대 회 장 채진석(인천대)
프로그램위원장 문양세(강원대)
조 직 위 원 장 김성백(제주대)
프로그램부위원장 권혁윤(서울과기대), 박영호(숙명여대), 서동민(KISTI), 이종욱(성균관대), 임성수(충남대), 임현승(강원대), 주재걸(KAIST)

/// 논문 모집

· 논문내용 : 정보과학에 관한 학술논문 및 기술보고 등
· 모집분야 : 컴퓨터 공학 전 분야
· 논문분량 : 2~3쪽 (작성양식)
· 제출방법 : 홈페이지 온라인 접수 (바로가기)
· 논문접수마감 : 2022년 4월 22일(금)
· 심사결과발표 : 2022년 5월 27일(금)
· 논문최종본접수 : 2022년 6월 7일(화)
※ 우수논문은 정보과학회 논문지 게재 추천

/// 튜토리얼 제안 모집

· 모집분야 : 정보과학 분야 다수의 관심 대상 자유 주제
· 제출내용 : 강의제안서 및 연사약력 (작성양식)
· 모집마감 : 2022년 4월 15일(금)
· 문의 및 제출 : 권혁윤 부위원장(서울과기대 교수)

/// 협력워크샵 제안 모집

· 모집내용 : 센터, 사업단, 과제 성과발표회 및 워크샵
· 제출내용 : 워크샵 제안서 (작성양식)
· 모집마감 : 2022년 4월 15일(금)
· 문의 및 제출 : 서동민 부위원장(KISTI 책임연구원)

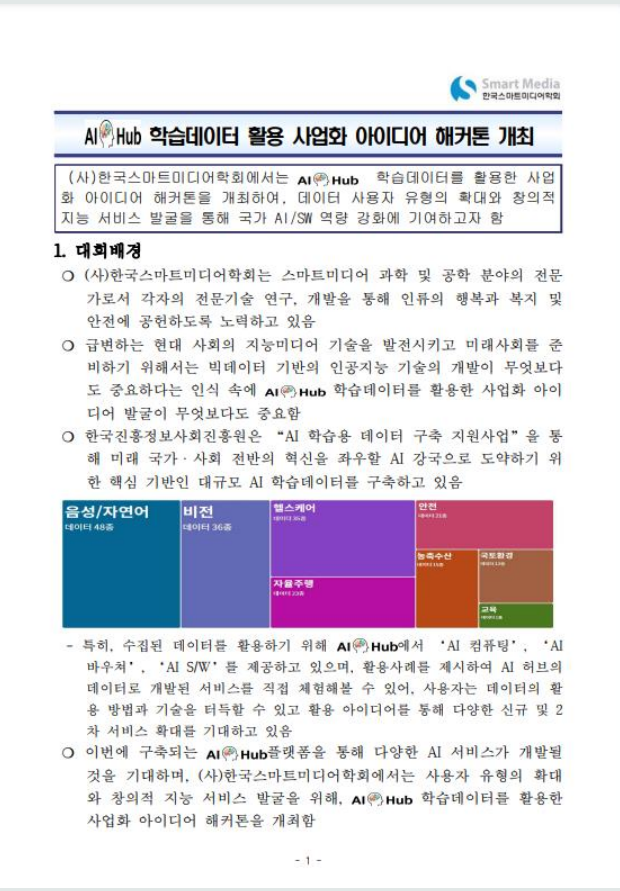
/// 신진연구자 초청

· 프로그램 : 최신 연구 내용 소개 및 네트워킹
· 초청대상 : 만 39세 이하이거나 박사학위 취득 후 7년 이내
· 참가신청 : 성명, 소속, 연락처, 연구분야 기재하여 신청
· 문의 및 신청 : 이종욱 부위원장(성균관대 교수)

※ 프로그램별 마감일정은 사정에 따라 다소 변경될 수 있습니다.
※ 자세한 사항은 홈페이지를 참고하시기 바랍니다.

대회장 채진석(인천대) 프로그램위원장 문양세(강원대) 조직위원장 김성백(제주대)
부위원장 권혁윤(서울과기대), 박영호(숙명여대), 서동민(KISTI), 이종욱(성균관대), 임성수(충남대), 임현승(강원대), 주재걸(KAIST)

홈페이지: https://www.kiise.or.kr/conference/kcc/2022/ 문의: 박보혜 주임(정보과학회) bhpark@kiise.or.kr 02-588-9247



Smart Media 한국스마트미디어학회

AI Hub 학습데이터 활용 사업화 아이디어 해커톤 개최

(사)한국스마트미디어학회에서는 AI Hub 학습데이터를 활용한 사업화 아이디어 해커톤을 개최하여, 데이터 사용자 유형의 확대와 창의적 지능 서비스 발굴을 통해 국가 AI/SW 역량 강화에 기여하고자 함

1. 대회배경

○ (사)한국스마트미디어학회는 스마트미디어 과학 및 공학 분야의 전문가로서 각자의 전문기술 연구, 개발을 통해 인류의 행복과 복지 및 안전에 공헌하도록 노력하고 있음

○ 급변하는 현대 사회의 지능미디어 기술을 발전시키고 미래사회를 준비하기 위해서는 빅데이터 기반의 인공지능 기술의 개발이 무엇보다도 중요하다는 인식 속에 AI Hub 학습데이터를 활용한 사업화 아이디어 발굴이 무엇보다도 중요함

○ 한국진흥정보사회진흥원은 "AI 학습용 데이터 구축 지원사업"을 통해 미래 국가·사회 전반의 혁신을 좌우할 AI 강국으로 도약하기 위한 핵심 기반인 대규모 AI 학습데이터를 구축하고 있음

- 특히, 수집된 데이터를 활용하기 위해 AI Hub에서 'AI 컴퓨팅', 'AI 바우처', 'AI S/W'를 제공하고 있으며, 활용사례를 제시하여 AI 허브의 데이터로 개발된 서비스를 직접 체험해볼 수 있어, 사용자는 데이터의 활용 방법과 기술을 터득할 수 있고 활용 아이디어를 통해 다양한 신규 및 2차 서비스 확대를 기대하고 있음

○ 이번에 구축되는 AI Hub플랫폼을 통해 다양한 AI 서비스가 개발될 것을 기대하며, (사)한국스마트미디어학회에서는 사용자 유형의 확대와 창의적 지능 서비스 발굴을 위해, AI Hub 학습데이터를 활용한 사업화 아이디어 해커톤을 개최함

- 1 -

발표를 마치겠습니다.

감사합니다.