



응답하라 챗봇

정보컴퓨터공학과 201624597 최상준
정보컴퓨터공학과 201824448 김정민
불어불문학과 201902163 김채원



— 목차 —

1. 배경 및 목표
2. 설계 및 수정
3. 개발 결과
4. 역할 분담
5. 개발 일정

1

배경 및 목표

배경 및 목표

[KoAlpaca 실제 사용시의 문제점]

디자인에 관한 특허 하나 알려줘

- application design에 대한 미국 특허 제7,479,526호는 이동 전화와 컴퓨터 디스플레이 간의 그래픽 사용자 인터페이스 디자인 관련 디자인 특허입니다.

휴대폰 통신 시스템에 관련된 특허가 뭐가 있어?

- 미국 특허 제7,854,063호는 이동 전화와 컴퓨터 디스플레이 간의 그래픽 사용자 인터페이스 디자인 관련 디자인 특허이며, 컴퓨터와 다른 개체 간의 그래픽 사용자 인터페이스 디자인에 대한 미국 특허 제7,494,146호와 제7,504,147호는 각각 컴퓨터 및 TV 디스플레이 간의 그래픽 사용자 인터페이스 디자인에 대한 디자인 특허입니다.

▷ 전문적 지식이 요구되는 상황: 부적절한 답변 출력

배경 및 목표

['응답하라 챗봇' 의 목표]

특히 도메인에 특화된
한국어를 사용하는

챗봇 구현

2

설계 및 수정

설계 및 수정



지식인

베스트 문답
분야별 문답

학습 가능 데이터 필터링
데이터 학습

키프리스

카테고리별 PDF 다운
PDF -> 텍스트 변환

데이터 바로 사용 가능
데이터 학습

배포 가능한 프롬프트 제작

설계 및 수정

① 지식IN



지식인

베스트 문답
분야별 문답

학습 가능 데이터 필터링
데이터 학습

배포 가능한 프롬프트 제작

키프리스

카테고리별 PDF 다운
PDF -> 텍스트 변환

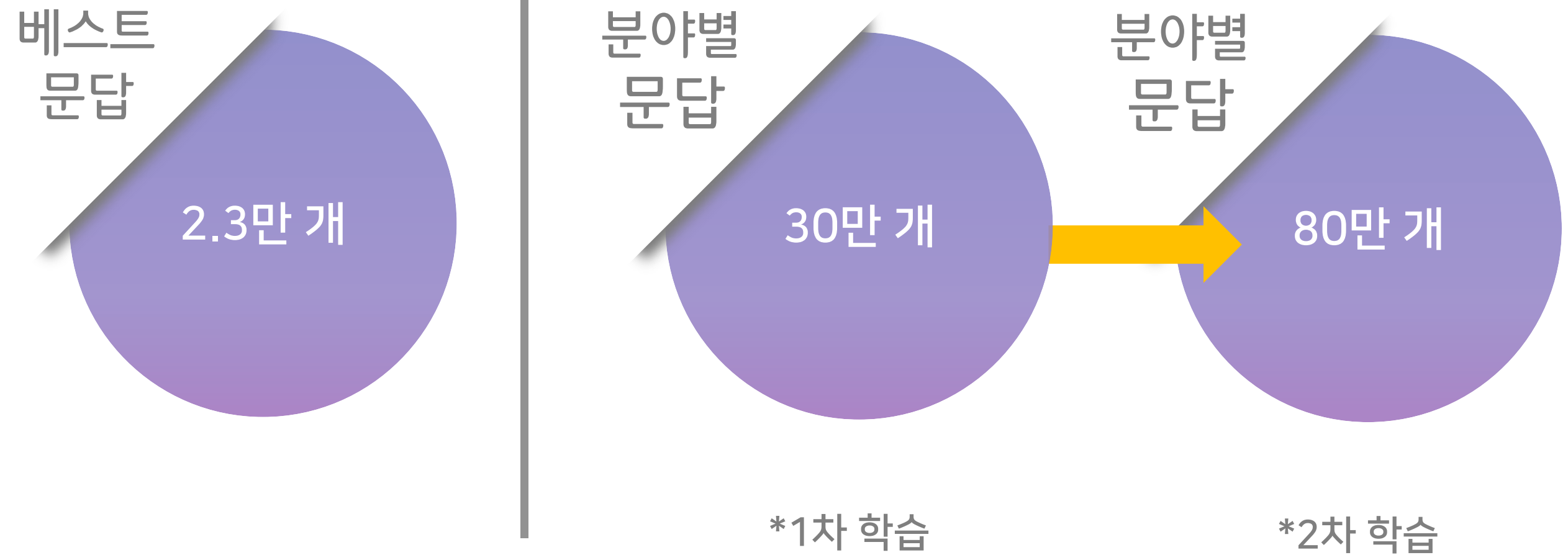
데이터 바로 사용 가능
데이터 학습

2

설계 및 수정

① 지식IN

[문답 크롤링]



설계 및 수정

① 지식IN



지식인

베스트 문답
분야별 문답

학습 가능 데이터 필터링
데이터 학습

배포 가능한 프롬프트 제작

키프리스

카테고리별 PDF 다운
PDF -> 텍스트 변환

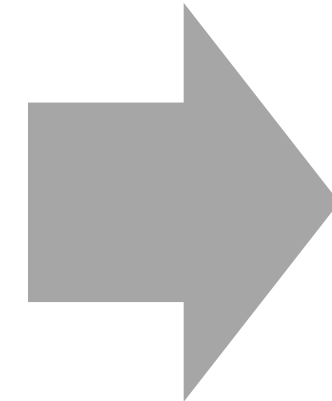
데이터 바로 사용 가능
데이터 학습

설계 및 수정

기존 계획

파이썬 Re 라이브러리

: 대량의 텍스트 사이에서
특정 부분을 검색 및 추출



문제점

1. 답변이 긴 경향
2. 특정 단어 사용 (내공)
3. 잘못된 / 편향된 답변

1. 지도개요

팀 명	응답하라 챗봇
과 제 명	도메인 지식에 특화된 한국어 대규모 언어모형 기반 챗봇
협력기관	(주)나라인포테크

2. 세부 지도 내용

데이터 전처리와 1 차 학습을 통해 보완점을 잘 파악함.
추가 전처리를 통해 현재 발견된 문제점들을 보완할 수 있을 것으로 파악됨.

아래는 이후 진행 상황에서 고려해 볼 만한 문제점 해결 방안입니다.
답변이 긴 경향: 학습 데이터의 길이를 제한하는 방식이 도움이 될 수 있음. Ex)
KoBertSum 을 사용해 핵심 문장만으로 구성된 축약된 답변을 생성하여 학습 진행.

'질문' '답변' '내용' 등의 단어 사용 또한 앞에서 진행한 전처리와 유사하게 해당 문장의 중요도를 고려한 문장 필터링을 거쳐 제거하면 답변의 완성도가 올라갈 것으로 예상됨.

잘못/편향된 답변 -> 학습 데이터에서 편향된 답변을 파악하고 사전에 제거 (ex> 성별, 지역, 인종 등에 대한 키워드 위주로 인식하여 필터링), K-StereoSet 과 같은 편향성 인식 모델을 활용.

설계 및 수정

기존 계획

파이썬 Re 라이브러리

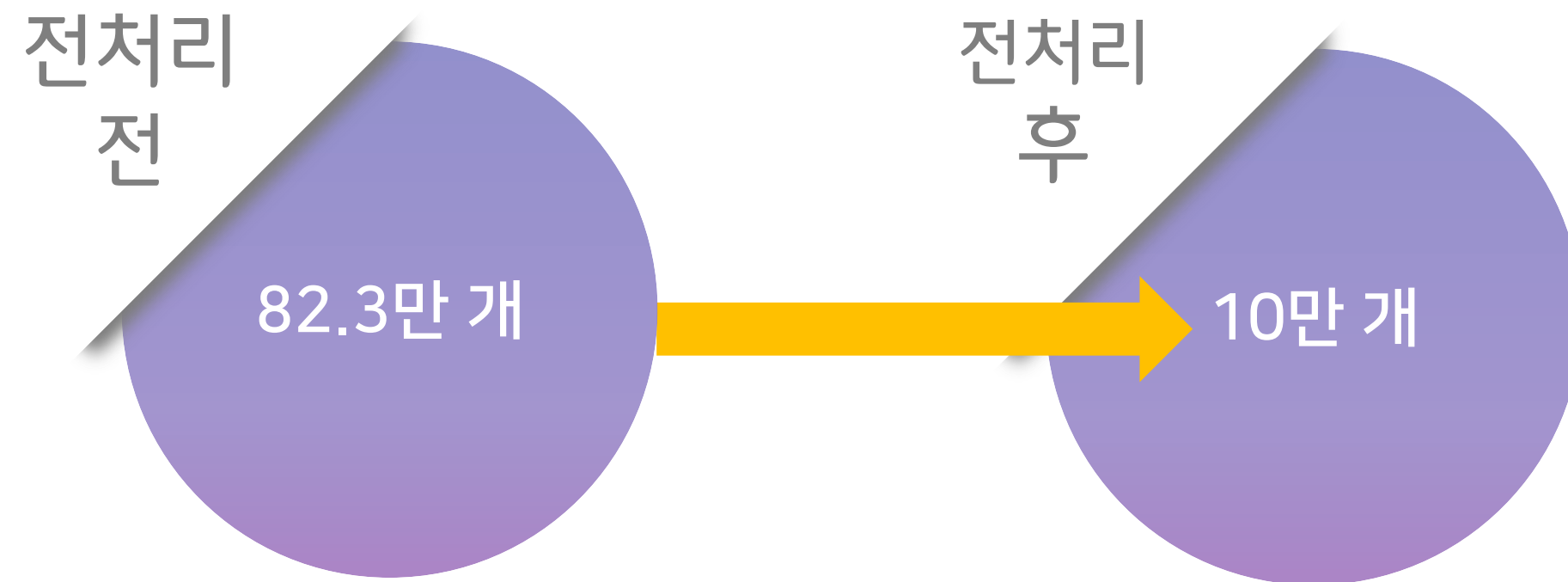
: 대량의 텍스트 사이에서
특정 부분을 검색 및 추출

수정 사항

GPT 이용

설계 및 수정

[문답 데이터 필터링]



설계 및 수정

② 키프리스



지식인

베스트 문답
분야별 문답

학습 가능 데이터 필터링
데이터 학습

배포 가능한 프롬프트 제작

키프리스

카테고리별 PDF 다운
PDF -> 텍스트 변환

데이터 바로 사용 가능
데이터 학습

설계 및 수정

[키프리스 카테고리 별 PDF 다운]

Smart Node에서의 All_IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템 < 통합행정정보 - 특허·실용신안 정보 — Mo...

kpat.kipris.or.kr/kpat/biblioa.do?method=biblioFrame

국영 상세보기 전환 | 프로그램 설치 안내 | 항목 전체 인쇄하기 | 인쇄하기 | 오류신고 | 도움말

Smart Node에서의 All_IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템
System and method for All_IP based Autonomous Mobile Communications on Smart Node

상세정보 | 공개전문 | 공고전문 | 등록사항 | **통합행정정보**

통합행정정보

본 '원본보기 서비스'는 참고용이므로, 일부 오류 및 누락이 발생할 수 있습니다.
정확한 서류를 확인하시려면 해당 웹사이트에서 조회하시기 바랍니다. (특히 **바로가기**: <http://www.patent.go.kr>)
해당 서비스는 점검으로 인해 **매주 일요일 00:00 ~ 02:00까지 이용이 중단됩니다.**

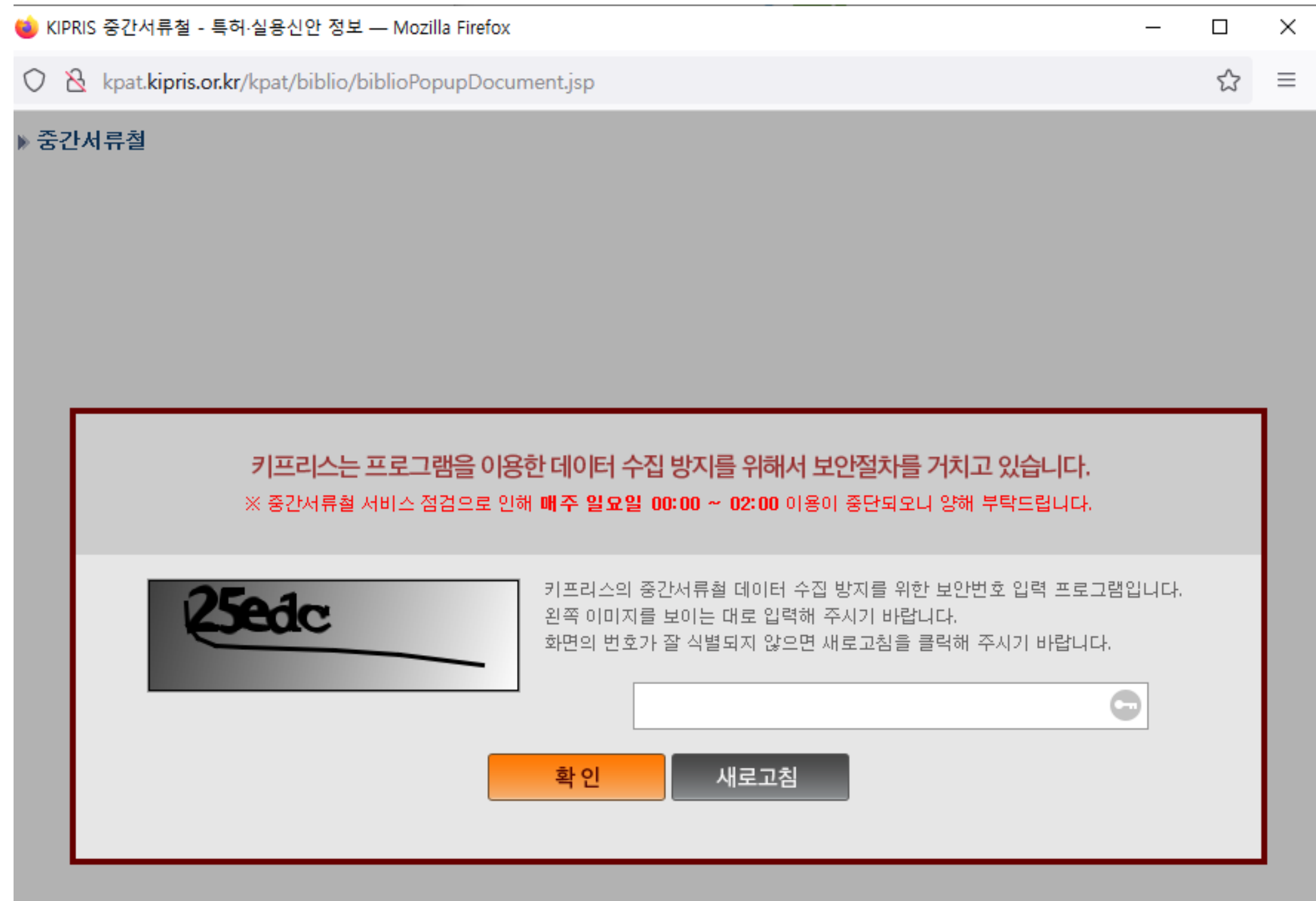
번호	서류명	접수/발송일자	처리상태	접수/발송번호
1	[특허출원]특허출원서 ([Patent Application] Patent Application)	2008.11.24	수리 (Accepted)	112008080867450
2	출원인정보변경(경정)신고서 (Notification of change of applicant's information)	2009.08.04	수리 (Accepted)	412009515089936
3	선행기술조사의뢰서 (Request for Prior Art Search)	2010.12.13	수리 (Accepted)	919999999999989
4	선행기술조사보고서 (Report of Prior Art Search)	2011.01.17	수리 (Accepted)	912011000331517
5	의견제출통지서 (Notification of reason for refusal)	2011.01.19	발송처리완료 (Completion of Transmission)	952011003481734
	[지정기가역장]기가역장(다중 경과구제)신청서			

2

설계 및 수정

② 키프리스

[키프리스 카테고리 별 PDF 다운]



KIPRIS 중간서류철 - 특허·실용신안 정보 — Mozilla Firefox

kpat.kipris.or.kr/kpat/biblio/biblioPopupDocument.jsp

중간서류철

키프리스는 프로그램을 이용한 데이터 수집 방지를 위해서 보안절차를 거치고 있습니다.
※ 중간서류철 서비스 점검으로 인해 **매주 일요일 00:00 ~ 02:00** 이용이 중단되오니 양해 부탁드립니다.

키프리스의 중간서류철 데이터 수집 방지를 위한 보안번호 입력 프로그램입니다.
왼쪽 이미지를 보이는 대로 입력해 주시기 바랍니다.
화면의 번호가 잘 식별되지 않으면 새로고침을 클릭해 주시기 바랍니다.

25edc

확인 새로고침

▷ 보안문자 입력코드

이미지 소스를 문자로 복원한 후 captcha에 입력하는 방식

설계 및 수정

[키프리스 카테고리 별 PDF 다운]


Korean Intellectual Property Office

특허출원서

【참조번호】	0005
【출원구분】	특허출원
【출원인】	
【명칭】	한국전자통신연구원
【특허고객번호】	3-1998-007763-8
【대리인】	
【명칭】	특허법인 무한
【대리인번호】	9-2007-100061-4
【지정된변리사】	구기완
【포괄위임등록번호】	2007-052305-1
【발명(고안)의 국문명칭】	Smart Node에서의 All_IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템
【발명(고안)의 영문명칭】	System and method for All_IP based Autonomous Mobile Communications on Smart Node
【발명(고안)자】	
【성명】	박우구
【성명의 영문표기】	PARK Woo Goo
【주민등록번호】	정보보호를 위해 미공개
【우편번호】	정보보호를 위해 미공개
【주소】	정보보호를 위해 미공개

▷ PDF를 TXT로 변환

-> 필요한 부분만 추출

-> json 파일로 변환

```
{
  "발명(고안)의 국문명칭": "[ '다중 배합을 통한 3D 프린팅 방법 ' ]",
  "발명(고안)의 영문명칭": "[ '3D PRINTING METHOD THROUGH MULTIPLE', 'MULTIPLE-STAGE-MIXING' ]",
  "출원인": [
    {
      "명칭": "고려대학교 산학협력단"
    },
    {
      "명칭": "국민대학교 산학협력단"
    }
  ],
  "대리인": [
    {
      "명칭": "권성현"
    },
    {
      "명칭": "가익시"
    }
  ]
}
```

"요약서": "['본 발명은 해상 또는 해저에서 적어도 2개의 프린팅 원료를 혼합하여 제1혼합물을 생성',
'하는 제1혼합 단계; 상기 제1혼합물과 혼화제를 분리 이송하는 단계; 상기 제1혼합물과'

2

설계 및 수정

② 키프리스

[키프리스 크롤링 데이터]

크롤링
데이터

10만 개

A purple circle with a diagonal line passing through it from the top-left to the bottom-right. The text '10만 개' is centered inside the circle.

개발 결과

4

역할 분담

4

역할 분담

최상준

데이터셋 수집 및 전처리

김정민

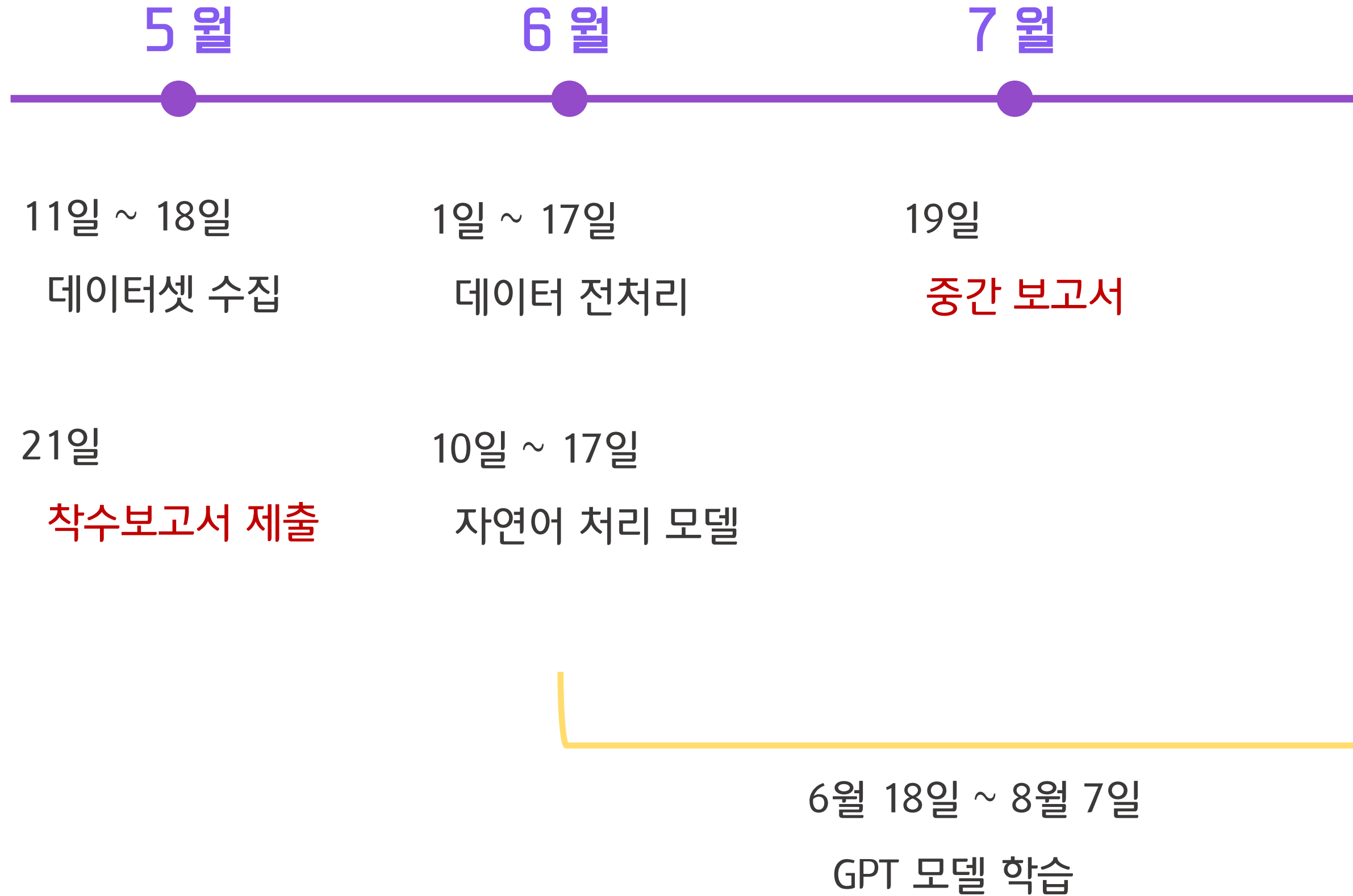
자연어 처리 모델 생성, 프롬프트 학습

김채원

web 챗봇 서비스 구현

6

추진 일정



6

추진 일정

8 월

9월 ~10월

테스트 및 보완
과제 마무리

11월 3일

졸업과제 발표회

6월 18일 ~ 8월 7일

GPT 모델 학습

8월 8일 ~ 9월 15일

개발 계획 수정 반영

3

현재 진행 상황

② 키프리스

[키프리스 카테고리 별 PDF 다운]

```
# 이미지링크를 가져와서 captcha 문자 가져오기
print("link:", 'http://kpat.kipris.or.kr'+img['src'])
driver.switch_to.default_content()

time.sleep(5)
bs4 = BeautifulSoup(driver.page_source, 'lxml')
txt = str(bs4.find_all('script')[2].get_text())

# 필요없는 문자제거및 문서고유번호 추출
src = txt.split('document.getElementById("pdfViewFrame").src = ')[1].split(';')[0]
src =src.replace('amp;', '')

# captcha 우회하는 스크립트 jquery 이동
script = '$("#pdfViewFrame").show();#
$("#bgBox").css("display","none");#
$("#simpleCaptcha").css("display","none");#
showPopLoadingBar();#
document.getElementById("pdfViewFrame").src = ''+src+'';#
resizeH();'
driver.execute_script(script)
time.sleep(5)
```

▷ 보안문자 입력코드

이미지 소스를 문자로 복원한 후 captcha에 입력하는 방식

현재 진행 상황

```
def convert_pdf_to_txt(path):
    rsrcmgr = PDFResourceManager()
    retstr = StringIO()
    codec = 'UTF-8'
    laparams = LAParams()
    device = TextConverter(rsrcmgr, retstr, codec=codec, laparams=laparams)
    fp = file(path, 'rb')
    interpreter = PDFPageInterpreter(rsrcmgr, device)
    password = ""
    maxpages = 0
    caching = True
    pagenos = set()
    for page in PDFPage.get_pages(fp, pagenos, maxpages=maxpages, password=password, caching=caching,
                                  check_extractable=True):
        interpreter.process_page(page)
    fp.close()
    device.close()
    str = retstr.getvalue()
    retstr.close()
    return str
```

▷ PDF를 TXT로 변환

3

현재 진행 상황

② 키프리스

[키프리스 카테고리 별 PDF 다운]

```
if __name__ == "__main__":
    # pdf폴더에있느걸 모두변환하여 txt폴더에
    folder_root = os.getcwd() + "/pdf/"
    txt_root = os.getcwd() + "/txt/"
    filelist = os.listdir(folder_root)
    # pdf to txt
    for idx in range(len(filelist)):
        if filelist[idx].split('.')[1]=='pdf':
            txt = convert_pdf_to_txt(folder_root + filelist[idx])
            text_file = open(txt_root + filelist[idx].split('.')[0] + '.txt', "w")
            text_file.write(txt)
            text_file.close()

    # txt parsing
    txtlist = os.listdir(txt_root)
    for idx in range(len(filelist)):
        with open(txt_root + txtlist[idx], 'r') as f:
            read_data = f.readlines()
            # 제출일 8 , 발명국문 10, 발명영문 12, 출원인성명코드 14,15 줄에있음
            # 아니면 정규표현식으로 추출하는 방법도있는데 특허출원서 양식이
            # 동일해서 대부분 이경우에서 걸린다.
            date = read_data[8].strip()
            korName = read_data[10].strip()
            engName = read_data[12].strip()
            name = read_data[14].strip()
            code = read_data[15].strip()
            print date, korName, engName, name, code
            # 이런식으로 데이터를 추출해서
            # 아래 코드를 이용 DB에 넣으면된다
            # 위치는 프로그램 특성에따라 조절
            # Connect to the database
```

▷ TXT의 필요한 부분만 필터링