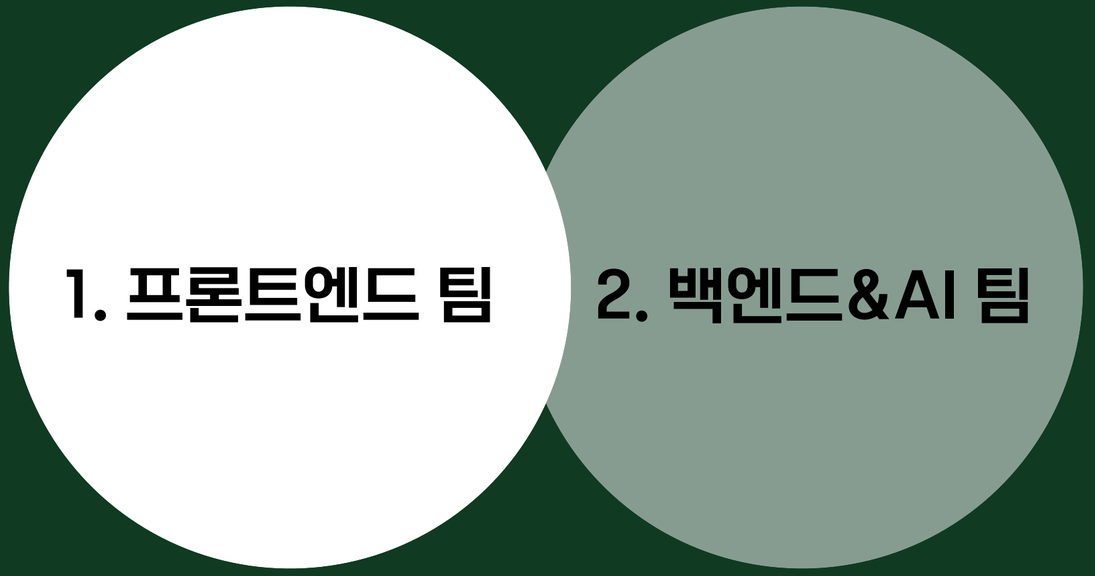
SKKU SCHOLARSHIP

| WEEKLY PROGRESS MEETING (1)

TEAM 스꿀라, SKKULAR

강병준 | 김규진 | 박진아 | 장이준 | 주재현

CONTENTS



TEAM. 스꿀라

SKKULAR

Brief Schedule

Based on the roles that were largely divided into two categories, we planned the project schedule by dividing it into AI and app parts. The schedule may vary depending on the situation for more efficient project progress.

Table 2. weekly project schedule

	weekly plan													
	2 3	4	5	6 7	8	9	10	11	12	13		14	15	
Overall	Project Proposa		Proposal Latex				Midterm					Finalterm	Final Latex	
AI			data collection ata preprocessin	Rule-Based			Q&A Model (BERT)							
App			UI/UX design Database setting	frontend work			backend work connect frontend and backend			Beta Test Debugging				

(a) : AI (b): Back-end (f) : Front-end

~ week 4
Project Proposal

week 5

1) Proposal Latex 작성

2) App :
UI/UX design (f)
Database setting (b)

3) AI :
Data collection (a,b)
Data preprocessing (a)

1. 프론트엔드 팀

기획 / UI,UX 구성

SKKULAR 핵심 기능

기능1

맞춤형
장학 검색

기능2

맞춤형
장학 알림

기능3

유사 키워드
장학 안내

기능4

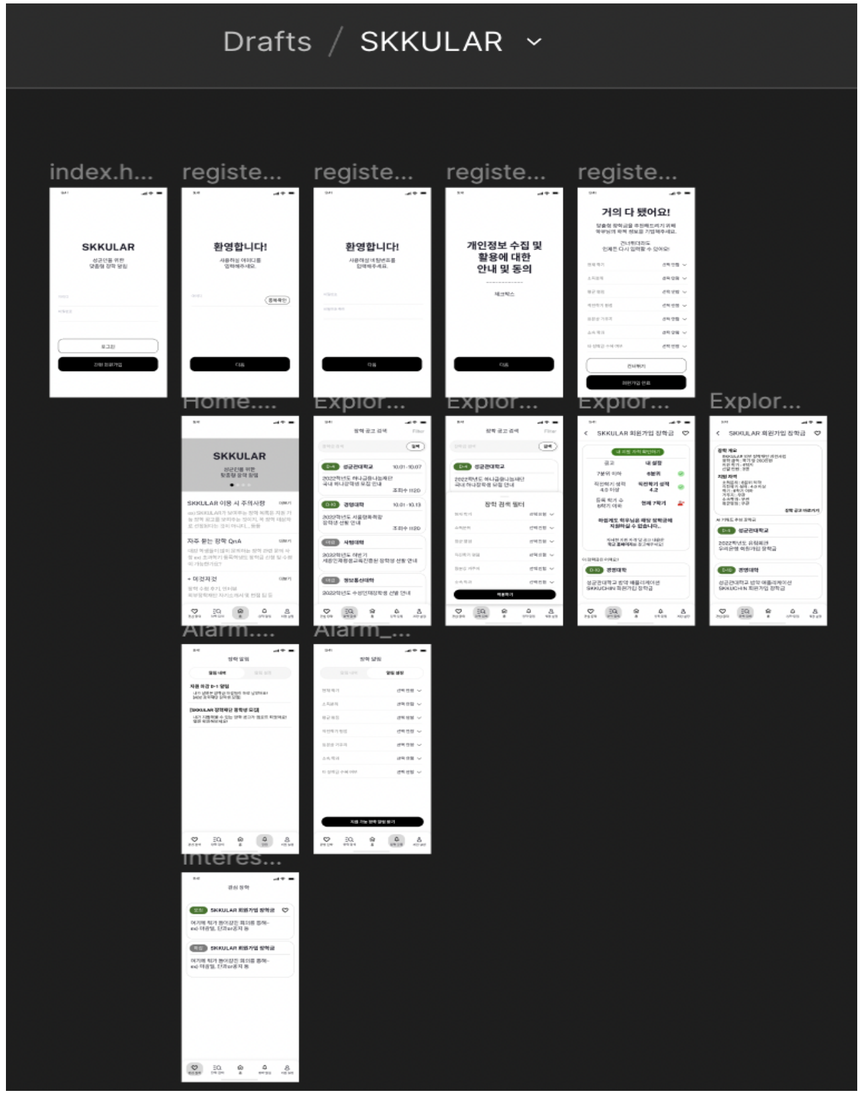
지원 마감일 알림

UI/UX 구성 | Figma 사용

전체적인 UI/UX 틀

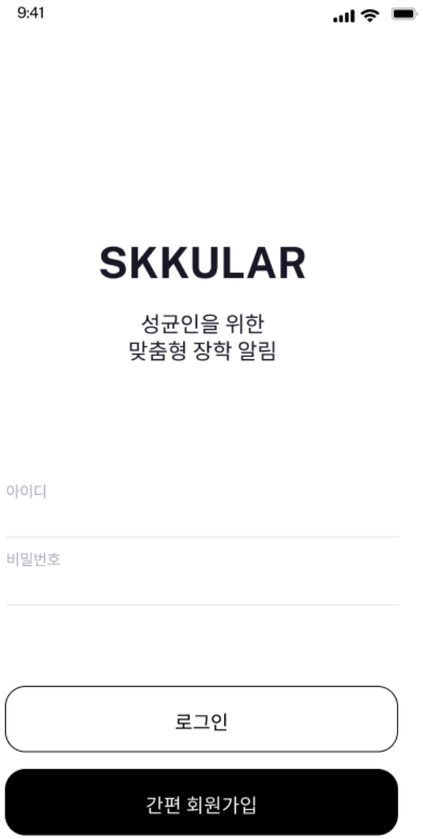
1)로그인 페이지
2) 회원가입 페이지
3) 홈 화면
4) 검색 페이지
5) 필터 적용 페이지
7) 맞춤형 추천 장학 상세 페이지
8) 관심 장학 페이지
9) 장학 알림 페이지

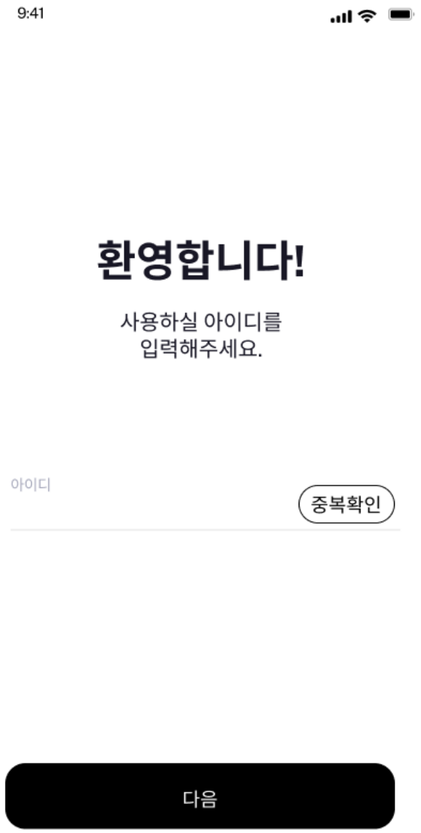
+ (10) 계정 설정 페이지)

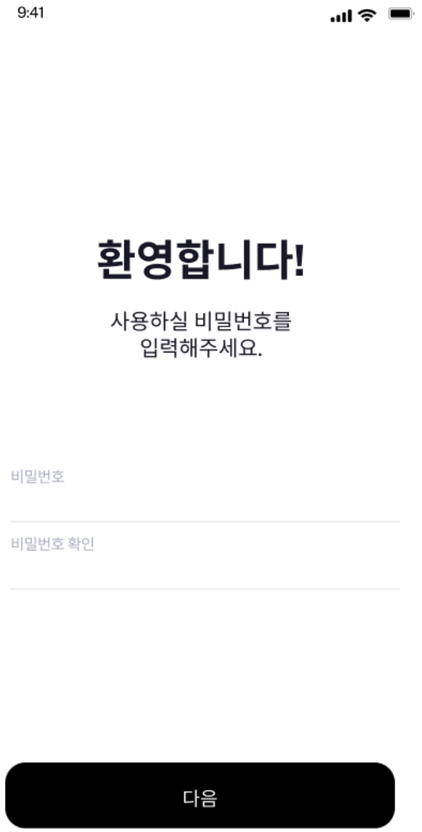


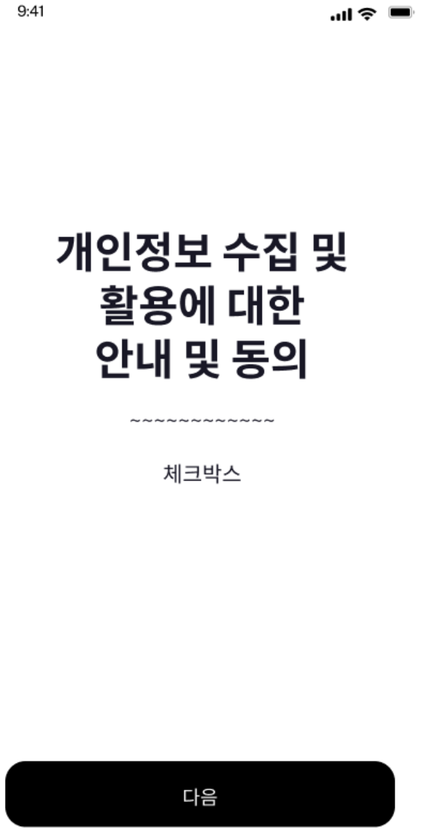
1. 로그인/회원가입

UI/UX 구성









로그인 혹은
회원가입 페이지로 이동

login.html

아이디 중복확인 / 비밀번호 일치확인 / 개인정보 수집 동의 / 초기 필터 설정

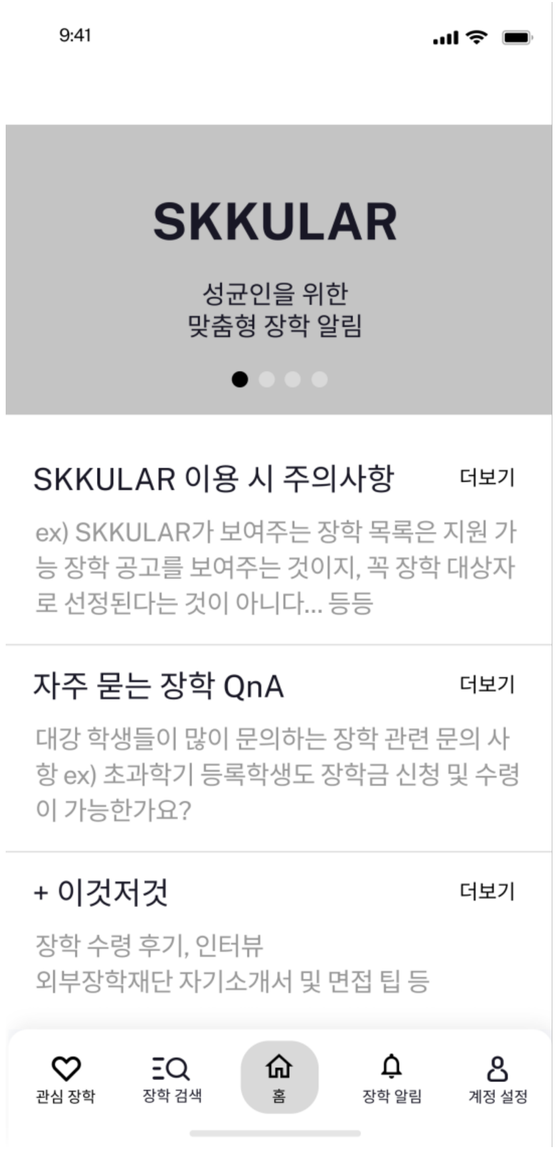
register1.html ~ register4.html

2. 홈 / 검색

UI/UX 구성

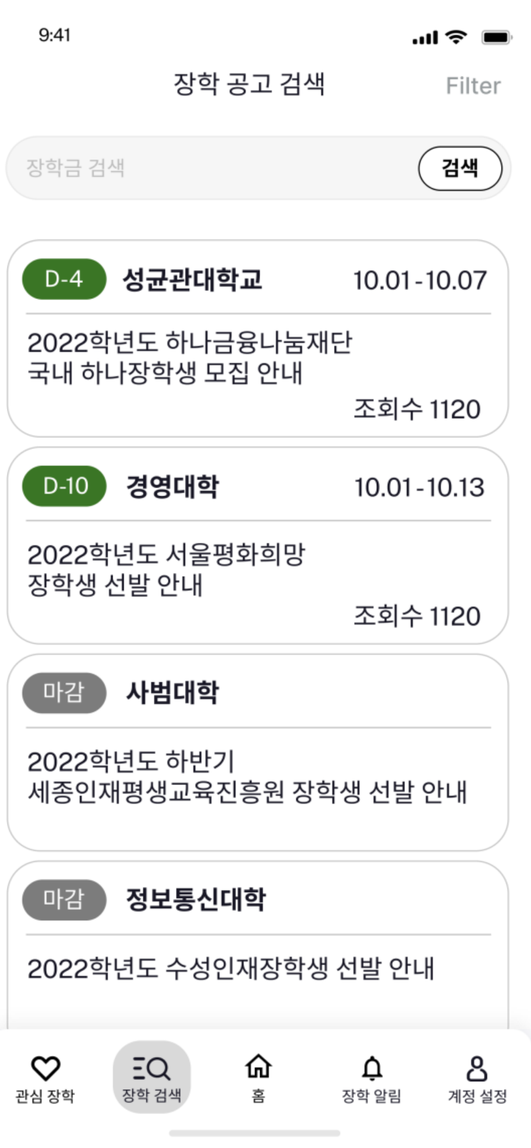
home.html

주의사항 및 Q&A 표시



search.html

전체 장학공지 리스트 표시

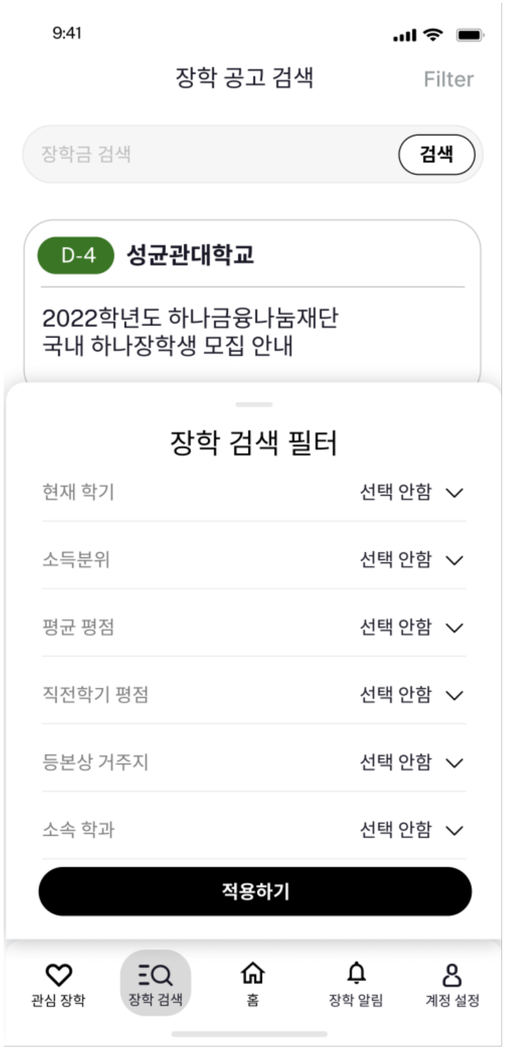


3. 필터 / 맞춤형 장학 추천 상세

UI/UX 구성

filter.html

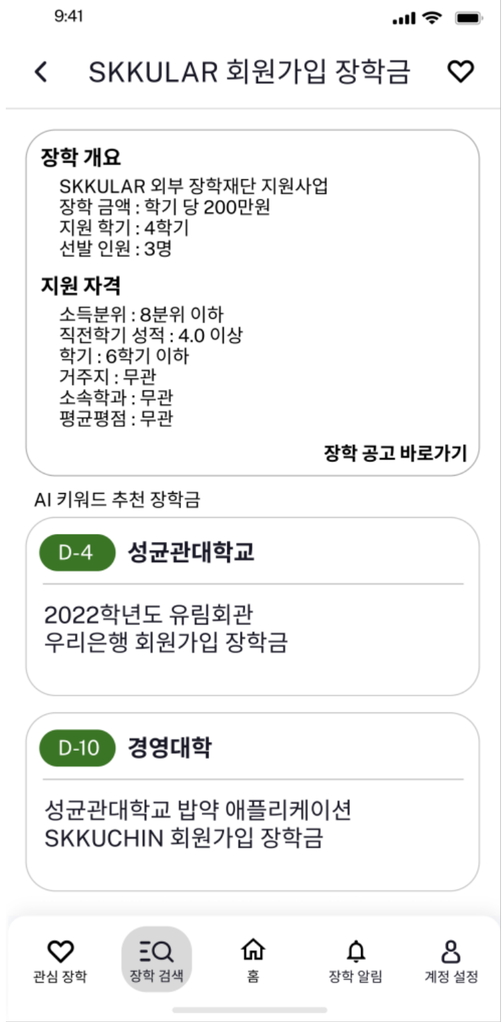
개인 맞춤형 필터 적용

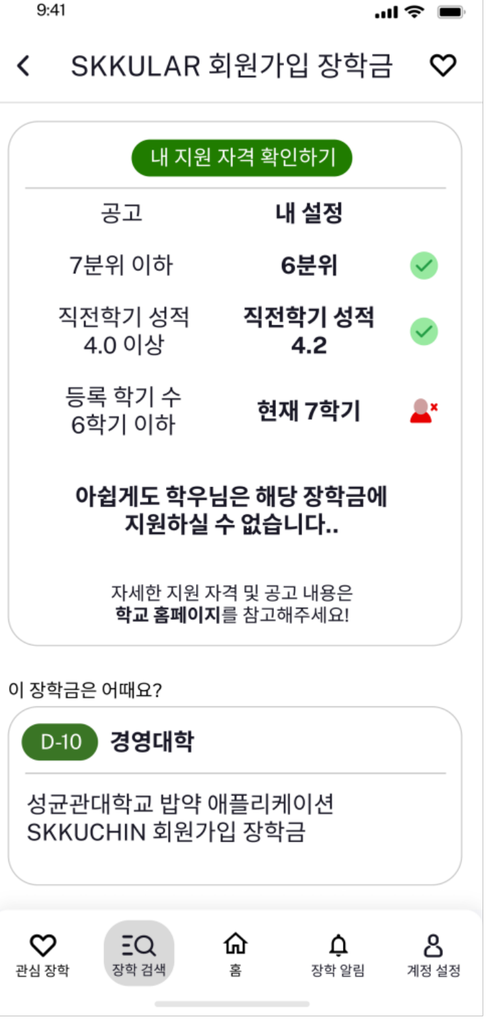


detail.html

맞춤형 장학 추천, 내용 상세 보기

with 유사 장학금 추천



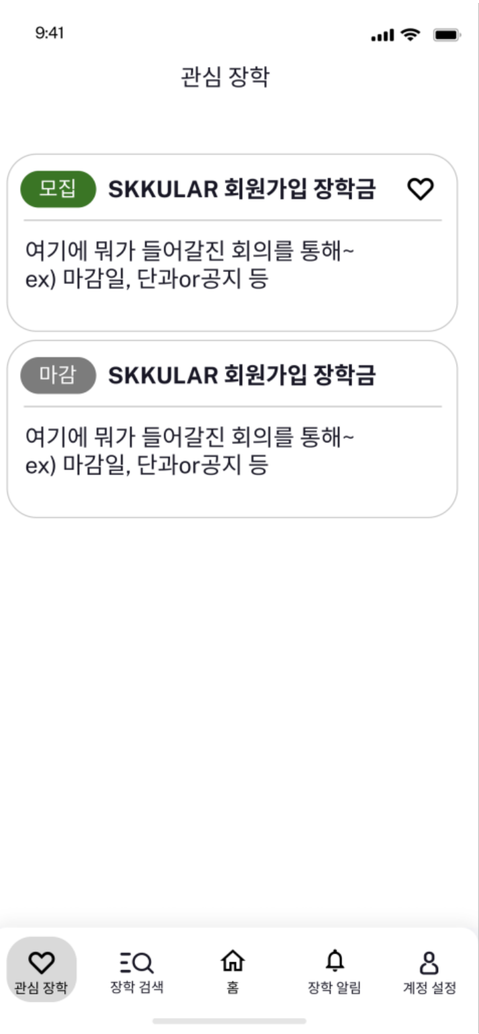


4. 관심 장학 / 장학 알림

UI/UX 구성

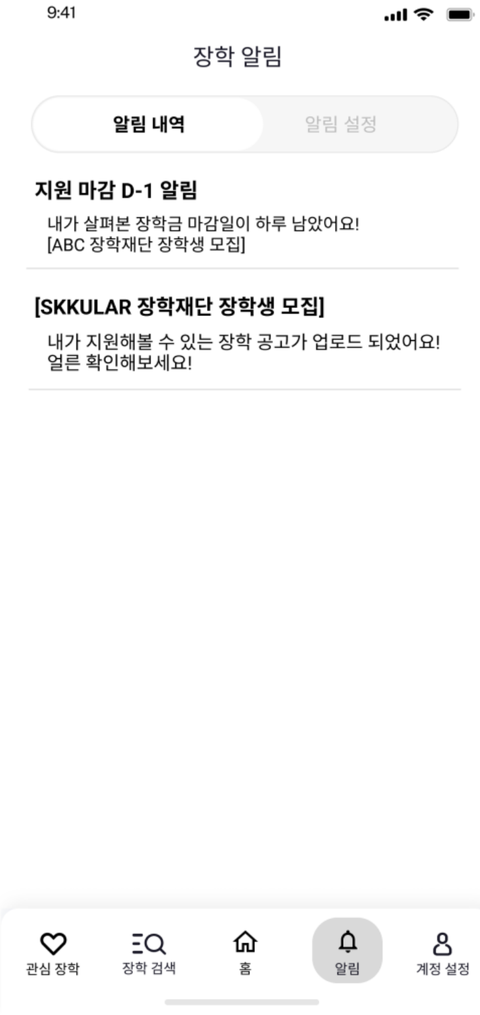
interested.html

관심 장학 스크랩 조회



alarm.html
alarm_setting.html

알림 내역 표시,
알림 on/off 설정



2. 백엔드 & AI 팀

크롤링

1. 크롤링 내용 _ 공지사항 게시판 목록

장학금
개별 게시판 URL

장학 공지
게시일

1. 크롤링 내용 _ 공지사항 게시판 목록

Ex) 0번째 게시물부터 64번째 게시물 크롤링하고 싶다면,

```
for page_num in crawling_number:
  # 크롤링 우회
  url="https://www.skku.edu/skku/campus/skk_comm/notice06.do?mode=list&&articleLimit=10&article.offset="+str(page_num)
  headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61 Safari/537.36'}
  html = requests.get(url, headers = headers)
  result = BeautifulSoup(html.content, "html.parser")

  post_numbers=[]
  # 수집해야 하는 데이터 목록:
  # 1) number/고유번호: No.1806(중복 거르기용)

  post_urls=result.select("#jwxe_main_content > div > div > div.container > div.board-name-list.board-wrap > ul > li > dl > dt > a")
  post_numbers=result.select("#jwxe_main_content > div > div > div.container > div.board-name-list.board-wrap > ul > li > dl > dd > ul > li:nth-child(1)")
```

1. crawling_number list에 [0,10,20,30,40,50,60]저장
2. For 문 이용하여 각 page_num 으로 들어가서 장학금 홈페이지 여러 페이지에 access
3. Beautifulsoup 이용하여 post_urls와 post_numbers에 같은 페이지에 있는 여러 게시물의 url과 게시물 번호를 저장

1. 크롤링 내용 _ 공지사항 게시판 목록

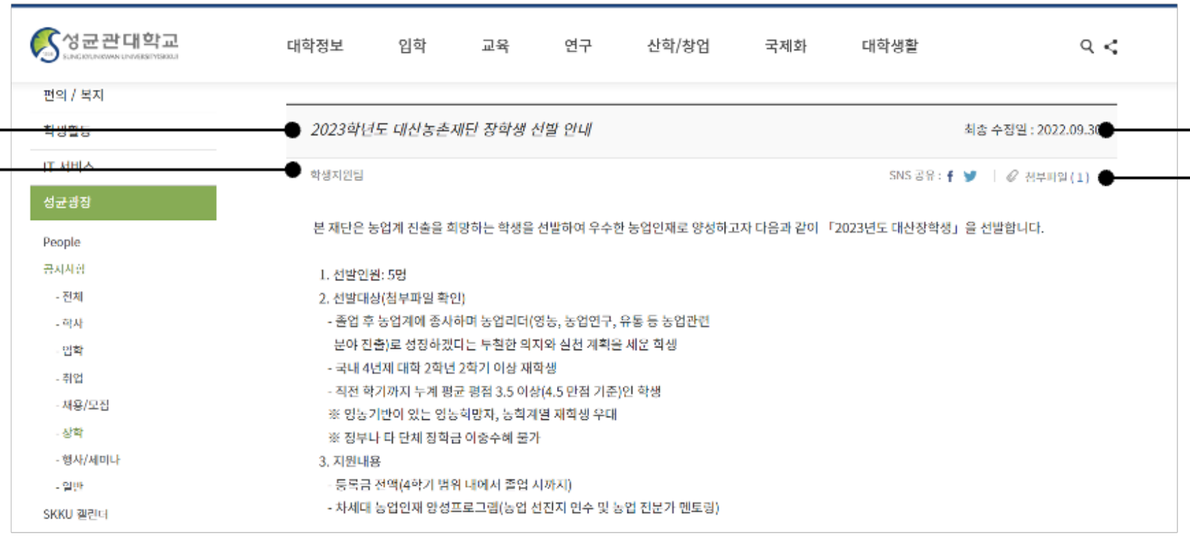
```
for i,post_url in enumerate(post_urls):

  print(str(i)+"번째 게시물입니다.")
  post_url=post_url["href"].replace("amp;","")
  post_url="https://www.skku.edu/skku/campus/skk_comm/notice06.do"+post_url
  headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61 Safari/537.36'}
  html = requests.get(post_url, headers = headers)
  result = BeautifulSoup(html.content, "html.parser")
```

4. For 문을 이용하여 각 게시물의 post_url을 얻기
5. Beautifulsoup을 이용하여 각 게시물의 url로 access 하기

2. 크롤링 내용 _ 세부 장학 게시글

공지사항
제목

장학금 담당
(학과/부서/人)



게시글
작성 일자

첨부파일
URL

2. 크롤링 내용 _ 세부 장학 게시글

```
#1) 홈페이지 고유 number If using all scalar values, you must pass an index
number=post_numbers[i].text

# 세부 페이지 접속: 수집해야 하는 데이터 목록
#2) date/작성날짜: 2022-09-20
date=result.select_one("#jwxe_main_content > div > div > div > div > table.board_view > thead > tr > th > span").text.replace("최종 수정일 : ","")
#3) title: 게시물 제목: 2022학년도 하나금융나눔재단 국내 하나장학생 모집 안내
title=result.select_one("#jwxe_main_content > div > div > div > div > table.board_view > thead > tr > th > em").text

#4) 게시물 내용
content=result.select_one("#jwxe_main_content > div > div > div > div > table.board_view > tbody > tr > td > dl > dd > pre").text
```

1. 홈페이지 고유 번호는 post_numbers list에 있는 번호 저장
2. 날짜, 제목, 및 게시물 내용은 각각 date, title, content에서 그 부분에 있는 text 크롤링

2. 크롤링 내용 _ 세부 장학 게시글

```
attachments=result.select("#jwxe_main_content > div > div > div > div > table.board_view > tbody > tr > td > div > div.fIR > ul > li:nth-child(2) > div > form >
attachment_content=""
attachment_url=""
try:
  for attach in attachments:
    # attach 제목에 attachments
    attach_title=attach.text
    if ("공고문" in attach_title) or ("안내" in attach_title):
      if not os.path.exists("/content/data/"):
        os.makedirs("/content/data/")
      else:
        try:
          os.remove("/content/data/report_temp.pdf")
        except:
          pass
      attachment_url="https://www.skku.edu/skku/campus/skk_comm/notice06.do"+attach["href"]
      print(1)
      wget.download(attachment_url,out="/content/data/report_temp.pdf")
      print(2)
      attachment_content= textractor("/content/data/report_temp.pdf")
      print(3)
      os.remove("/content/data/report_temp.pdf")
```

1. 첨부 파일에 있는 url attachments list에 저장 〉 attach_title에 첨부파일 제목 저장
2. Attach_title에서 “공고문“ or “안내” 가 있으면 pdf 다운로드
3. Textract 이용하여 다운로드된 pdf의 내용 저장
4. PDF 삭제

2. 크롤링 내용 _ 세부 장학 게시글

```
if crawling_start==False:

  post_temp=pd.DataFrame({"number":[number], "date":[date], "title":[title],"content":[content], "attachment_url":[attachment_url], "attachment_content":[attachment_
  crawling_start=True
else:
  post_t=post_temp.copy()
  #post_Series([number, date, title,content, attachment_utemp=post_t.append([number, date, title,content, attachment_url, attachment_content, current_url, department
  post_temp = post_t.append(pd.rl, attachment_content, current_url, department], index=post_temp.columns), ignore_index=True)

display(post_temp)
```

```
[ ]   1 ############ 구글 드라이브에 지정된 경로로 저장###############
      2 df_temp.to_csv("/content/drive/Shareddrives/skkular/data/성균관대학교_전체_장학금.csv",index=False, encoding="utf-8")

[ ]   1 df_temp
```

1. Pandas dataframe을 만들어서 크롤링한 값을 저장하거나 dataframe에 추가하기
2. 크롤링 된 결과를 저장한 dataframe을 csv file type으로 저장

3. 크롤링 결과

CSV 파일로 저장

number	date	title	content	attachment_url	attachment_content	current_url	department
No.1805	2022.09.20	2022학년도 2학기 DB	당 재단은 DB그룹(구 동부그룹) 설립자 김준기 회장이 기금 출연하여			https://www.skku.edu/	학생지원팀
No.1804	2022.09.16	2022학년도 (재)광산장	(재)광산장학회에서는 우리 지역의 우수한 인재를 발굴·육성하고 학생			https://www.skku.edu/	학생지원팀
No.1803	2022.09.16	2022학년도 2학기 안	안산인재육성재단에서는 교육비 부담 경감을 통한 지역 인재육성 및			https://www.skku.edu/	학생지원팀
No.1802	2022.09.16	2022학년도 2학기 고	고양시에서는 대학생들	https://www.skku.edu/	본문내용 입력 (2022년	https://www.skku.edu/	학생지원팀
No.1801	2022.09.15	2022학년도 2학기 63	2022년도 2학기 63장학금 장학생을 선발하고자 하오니 아래를 참고			https://www.skku.edu/	학생지원팀
No.1800	2022.09.14	2022학년도 하반기 (재	1. 재단 안내 ○ 명 경	https://www.skku.edu/	1111 재단법인 은평구	https://www.skku.edu/	학생지원팀
No.1799	2022.09.14	2022학년도 하반기 (재	가. 선발인원: 26명(성	https://www.skku.edu/skku/campus/skk_com		https://www.skku.edu/	학생지원팀
No.1798	2022.09.06	2022학년도 수림재단	수림재단에서 설립자인 동교 김희수 전 이사장의 교육이념을 받들어			https://www.skku.edu/	학생지원팀
No.1797	2022.09.06	2022학년도 2학기 지	본 재단은 국내 디자인 예술 산업 발전에 밑거름이 될 인재육성을 위			https://www.skku.edu/	학생지원팀
No.1796	2022.09.06	2022학년도 일문과학	1.선발분야 및 규모 : 0 명 내외 가. 전공 제한 없이 지원 가능 나			https://www.skku.edu/	학생지원팀
No.1795	2022.09.06	2022학년도 서물장학	서물장학재단에서는 ?	https://www.skku.edu/	2022년 <서물희망 디	https://www.skku.edu/	학생지원팀
No.1794	2022.09.06	2022학년도 광주광역	1. 지원자격 : 공고일 (	https://www.skku.edu/	재 (재)광주광역시 서?	https://www.skku.edu/	학생지원팀
No.1793	2022.09.05	2022학년도 하반기 혼	■ 신청기간: 2022. 9. 26.(월) ~ 9. 30.(금) (5일간)■ 접 수 처: 애향장			https://www.skku.edu/	학생지원팀
No.1792	2022.09.02	2022학년도 제 25기 '	「두을장학재단」은 삼성그룹의 창업주 故이병철 선대회장을 내조하신			https://www.skku.edu/	학생지원팀
No.1791	2022.09.02	2022학년도 경주시장	1. 선발 인원 및 신청자격□ 선발인원 - 총 450명(일반전형: 423명 ?			https://www.skku.edu/	학생지원팀
No.1790	2022.09.02	2022학년도 포항시장	1. 신청기간 : 2022. 08. 24(수) 09:00 ~ 2022. 09. 20(화) 18:002. (			https://www.skku.edu/	학생지원팀
No.1789	2022.09.02	2022학년도 인송문화	'仁松文化財團' 이 학업에 뜻이 있고 바른 사회관을 가진 대학생으로			https://www.skku.edu/	학생지원팀
No.1788	2022.08.31	2022학년도 2학기 도	도산장학회는 1995년 설립되어 매년 경제적 어려움이 있는 학생들(			https://www.skku.edu/	학생지원팀

- Number : 공지사항 번호
- Date : 공지사항 작성 날짜
- Title : 공지사항 제목
- Content : 공지사항 내용
- Attachment_url : 첨부파일 url
- Attachment_content : 첨부파일 내용
- Current_url : 공지사항 url
- Department : 공지사항 부서

학교 공식 홈페이지 공지사항은 약 1,000개, 단과대 홈페이지 공지사항은 대략 30개씩 크롤링 진행

4. 크롤링 역할 분담

장이준

- 크롤링 전체적 구현

- 학교 공식 장학 공지, 학부대학, 유학대학, 문과대학, 사회과학대학 장학 공지

김규진

- Textract pdf 조사

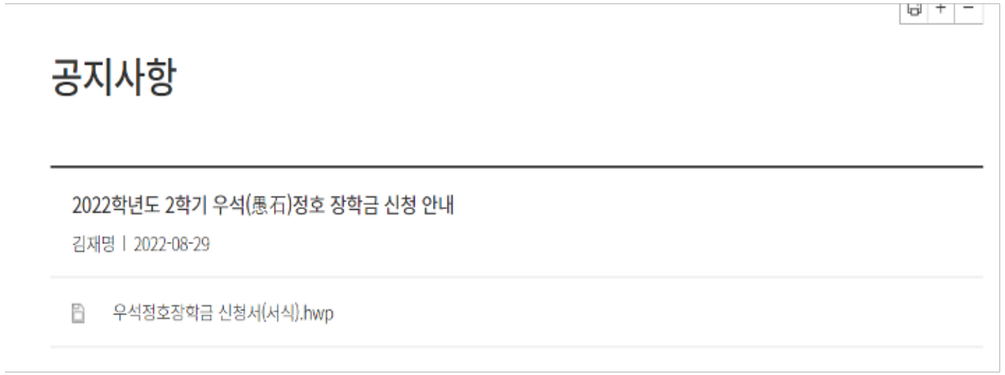
- 경제대학, 경영대학, 사범대학, 예술대학, 자연과학대학 장학 공지

강병준

- Javascript으로 인한 오류 수정

- 정보통신대학, 글융, 공과대학, 생명과학대학, 스포츠과학대학, 성균융합원 장학 공지

Trial and error _ textract pdf 저장




```
!wget -N https://github.com/neuml/txtai/releases/download/v3.5.0/tests.tar.gz
!tar -xvzf tests.tar.gz

# Install NLTK
import nltk
nltk.download('punkt')
```

```
%%capture

from txtai.pipeline import Textractor

# Create textractor model
textractor = Textractor()
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
pdf_path = "/content/drive/Sharedrives/skkular/data/a.pdf"
```

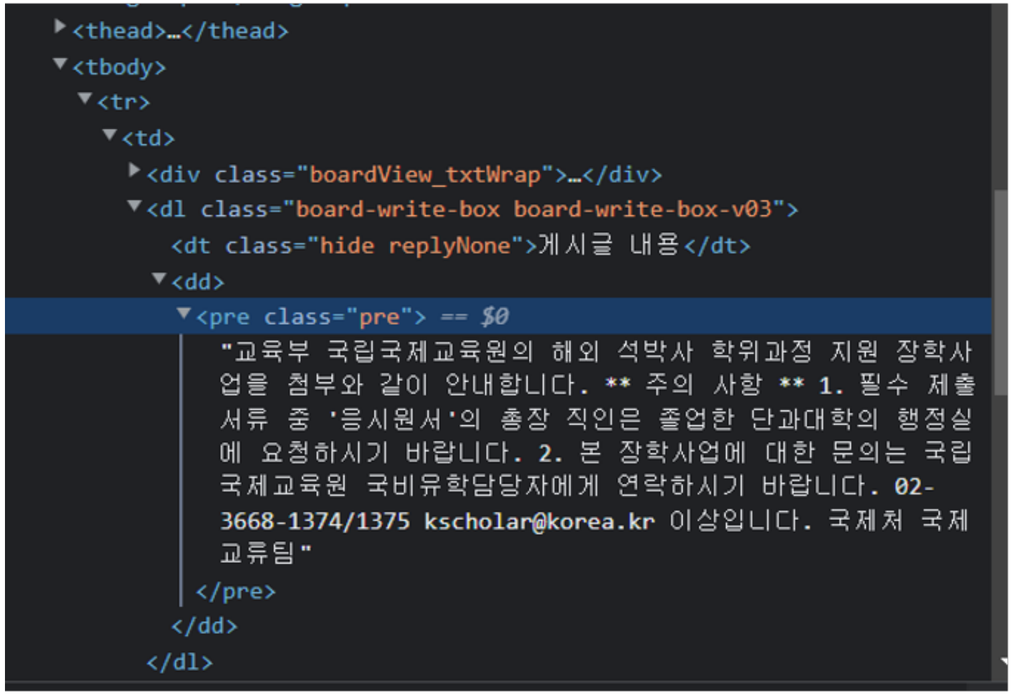
```
textractor("/content/drive/Shareddrives/skkular/data/a.pdf")
```

장학금 공지사항 내 첨부파일이 '신청서' 이면 크롤링 해서는 안 됨!

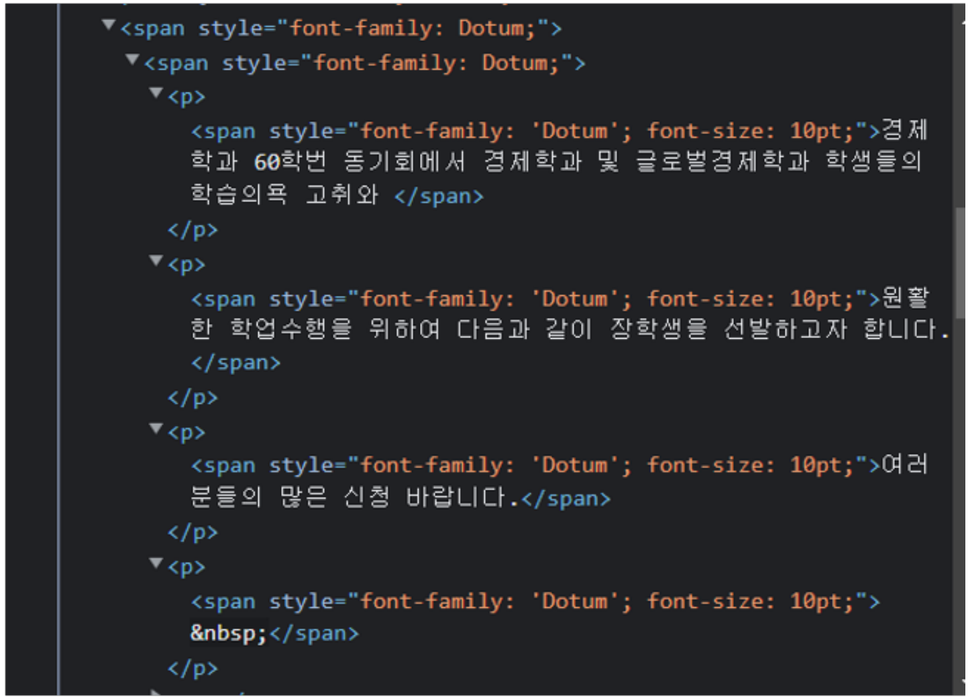
↓

첨부파일 제목에 "공고문" or "안내" 라는 단어가 있으면 textract 이용하여 pdf 내용 저장

Trial and error _ 전체 공지 vs 단과대 공지



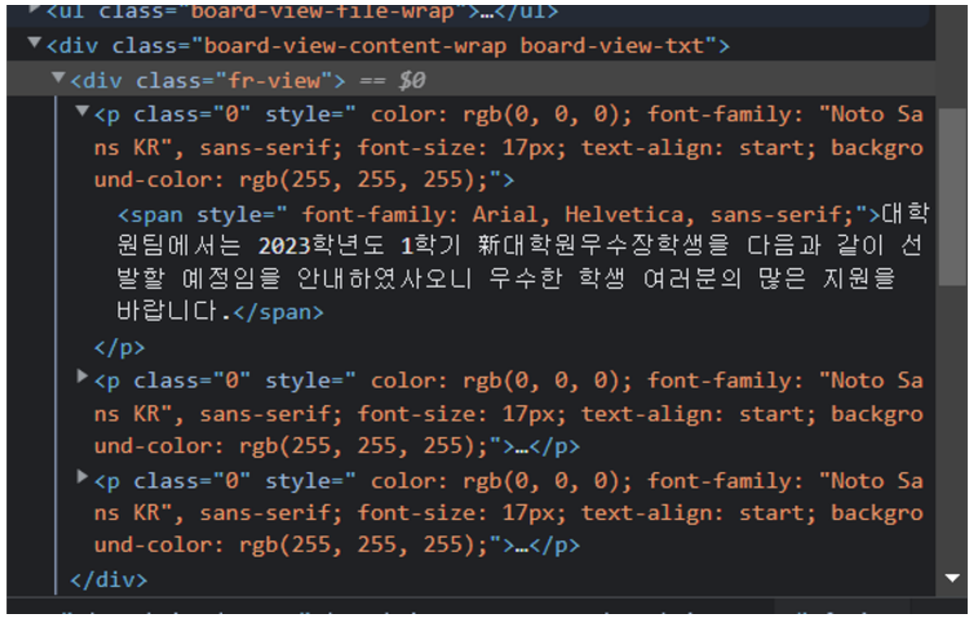
전체 공지사항

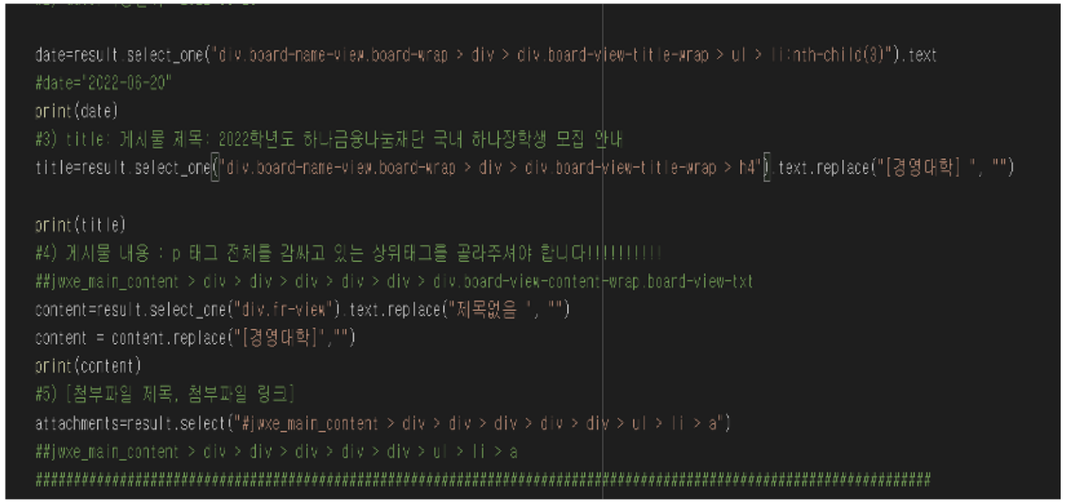


단과대 공지사항

전체 공지사항은 〈dd〉에 있는 text인 반면, 단과대 공지사항은 〈p〉에 있는 text를 추출

Trial and error _ 단과대 공지





경영대학 등 일부 대학은 javascript에서 selector 값이 보이는 것과 다르게 나오기 때문에
html에서 값을 직접 select 하는 방식을 취함

Brief Schedule [next week]

Based on the roles that were largely divided into two categories, we planned the project schedule by dividing it into AI and app parts. The schedule may vary depending on the situation for more efficient project progress.

Table 2. weekly project schedule

	plan											
	2 3 4	5	6 7 8		10	11	12	13	14	15		
Overall	Project Proposal	Proposal Latex			Midterm				Finalterm	Final Latex		
AI		data collection data preprocessin	Rule-Based		Q&A Model (BERT)							
App		UI/UX design Database setting	frontend work		backend work connect rontend and backend			Beta Test Debugging				

(a) : AI (b): Back-end (f) : Front-end

week 6

1) APP :

Database setting(b)

UI/UX design 보완(f)

Frontend 개발 시작(f)

2) AI :

Data preprocessing(a)

Data labeling (a)

Rule-Based 구현 시작(a)

Q&A

지금까지 TEAM. SKKULAR 였습니다.
감사합니다.