

교환학생 후기 데이터 분석

-학생들의 전공 학점인정과 해외 생활 적응을 위해서-

안영진(2015190122)

교환학생 후기 데이터 분석

해외로 교환학생을 가는 연세대 학생을 위해서 전공 학점인정이 가능한 대학을 분류하고, 해외 생활정보를 시각화함.

연구 질문

- 교환학생을 준비하는 연세대학교 학생들은 전공 학점 인정, 해외 생활 환경에 관심이 많음.
- 하지만 연세대학교 후기 게시판에서 **국가 별로 / 대학 별로 후기들을 열람하기는 번거로움.**

연구 설계

- 종합대학과 단과대학의 분류:** 해외 파견대학에서 "전공 학점을 채우기 힘든 상황"을 방지하기 위해서 전공 학과가 고르게 존재하는 종합대학과, 특정 학과만이 존재하는 단과대학을 분류함.
- 해외 생활정보 시각화:** 물가, 시내와의 근접성, 도심/교외, 기후 등을 각 대학별로 워드클라우드로 시각화함.

데이터 선정 및 특징

- 연세대학교 대학생 후기들 8326개를 전수조사함.
- 후기 데이터는 비정형 자연어(문어체)이며, 기후/주변환경/식사 등 9개 항목들로 분류되어 있음.

데이터 처리 및 연구 기법

종합대학과 단과대학의 분류

- BeautifulSoup4와 Selenium으로 후기 데이터 8326개 크롤링.
- Pandas, statistic을 이용해서 파견 인원 수와 학과 별 인원의 분산값으로 2차원 df를 제작함.
- Scikit-learn, matplotlib을 이용해서, 각 파견 대학들을 K-Means로 4가지 군집으로 분류함.

해외 생활정보 시각화

- KONLPY를 이용해서 문서에서 자주 등장하는 명사를 워드클라우드로 제작함.
- TF-IDF와 KONLPY를 이용해서 같은 대학의 9개 항목 문서들 간에 공통적으로 등장하는 명사는 비중을 낮추고, 각 문서들에서 자주 등장하는 명사들을 갖고 워드클라우드를 제작함.

결과 분석

- 종합대학 / 단과대학을 분류하는 정확도(accuracy)는 93%로 나옴.
- 워드클라우드는 각 파견대학의 차이점들을 시각화한 것에서 의의가 있음.

토의

- Pandas 라이브러리를 이용해서 데이터를 전처리하는 것이 까다로웠음.
- 지금껏 소외된 교환학생 데이터를 corpus로 만들고, 분류와 시각화를 한 것에서 의의가 있음.

교환학생 후기 데이터 분석

해외로 교환학생을 가는 연세대 학생을 위해서 전공 학점인정이 가능한 대학을 분류하고, 해외 생활정보를 시각화함.

- 연세대학교 교환후기 데이터 수집 코드

https://github.com/snoop2head/OIA_Text_Wrangling/blob/master/data_collect.py

- 교환후기 Corpus (CSV Format, 364 files, 96MB)

<https://drive.google.com/drive/u/0/folders/1iXS9UVp0d0J-s2hRVKZGUTWfYOl4-f0>

- 해외 교환대학을 종합대학, 단과대학으로 분류하기 (Jupyter Notebook)

https://colab.research.google.com/drive/1-ALy9nUHQ9ioJYkDHdbd_0BOTiadhbpm

- 대학 별 특징 워드클라우드로 시각화하기 (Jupyter Notebook)

https://colab.research.google.com/drive/1_DYkc6sJqcUF7DRRZHUrOjsMan0uD5zi

An aerial night view of a city skyline, likely Hong Kong, with numerous illuminated skyscrapers and a body of water in the background. A large yellow rectangular box is centered over the image, containing the text '01 문제 상황'.

01 문제 상황

연세대학교 학생들이 해외 교환 대학에서 전공 학점을 인정 받기 위해서는
문제점이 두 가지 존재함.



단과대학 존재



특정어문용
대학 존재

I. 종합 대학인 연세대학교와는 다르게, 학과가 하나만 존재하는 단과대학들이 존재함.



Aalto University
- 종합대학이겠지?



**University of Jyväskylä,
School of Business & Economics**
- 상경대인가?

I. 종합 대학인 연세대학교와는 다르게, 학과가 하나만 존재하는 단과대학들이 존재함.



Aalto University
경영학과만 존재하는 대학.



**University of Jyväskylä,
School of Business & Economics**
경영, 경제, 문헌정보, 기계공학, 신문방송,
언론홍보, 심리학, 응용통계학...

예를 들어서 Aalto University로 교환을 간 연세대학교 학생들 중 경영학과를 제외한 다른 학과 학생들은 전공 학점인정을 받지 못함.

Aalto University

No	제목	학과
73	알토대학교에서의 한학기	UIC 창의기술경영
72	Aalto, Mikkeli: 우주의 밤바닥에서	경영학과
71	꿈같았던 Mikkeli에서의 한 학기:)	교육학과
70	마음의 고향, 핀란드	철학
69	가장 낯설었던 나라에서 보낸, 잊을 수 없는 1년	사회학과
68	생각의 폭을 넓혀준 교환학생의 1년!	경영학과
67	미켈리에서의 한 학기	정치외교학과
66	Mikkeli Spirit	경영학과

II. 해외 교환 대학들 중에서는 특정 어문계열만이 갈 수 있는 학교들이 존재함.



St. Petersburg State University
- 주립대니까 종합대학이겠지?



Fudan University
- 중어중문과만 갈 수 있나?

II. 해외 교환 대학들 중에서는 특정 어문계열만이 갈 수 있는 학교들이 존재함.



St. Petersburg State University는
러시아어를 배우는 어학당.

“...필자는 인문대학으로 신청했다가 도저히 버틸
수가 없어서 어학당으로 옮겼다..”



Fudan University는 중국 3위인
종합대학이며, 다양한 학과 학생들이 지원함.

연세대학교 국제처에서 제공하는 자료는 한계가 있음.

밑에 표에서 볼 수 있듯, 대학교 홈페이지를 일일이 방문하거나 과사무실에 직접 문의해야 함.


Country	Program	University/Institution	가능학과/Available area of study	불가능학과/Unavailable area of study
Australia	교환 ESP	Australian National University	Almost all of our undergraduate courses are open to exchange	If a course is not offered for exchange students it would be
Australia	교환 ESP	Bond University		
Australia	교환 ESP	Curtin University	Use our course ("unit") finder - http://handbook.curtin.edu.au	Restricted- OT, nursing, midwifery, medicine, speech pathology
Australia	교환 ESP	Deakin University	http://www.deakin.edu.au/students/enrolment-fees-and-	Medicine and optometry
Australia	교환 ESP	Griffith University	https://www.griffith.edu.au/international/global-mobility/inbound/study-	
Australia	교환 ESP	La Trobe University		No one faculty, but some individual classes are not open
Australia	교환 ESP	Monash University	The majority of courses are open to exchange students. Details	Fourth year and honours, Off-campus learning
Australia	교환 ESP	Murdoch University	Any unit, as long as students meet the pre-requisites for this unit	Veterinary Sciences and Teaching/Practicum units
Australia	교환 ESP	University of Adelaide	https://www.adelaide.edu.au/inbound-study-abroad/choosing/	External Courses are not available. These courses are identified

연세대학교 국제처에서 제공하는 자료는 한계가 있음.

밑에 표에서 볼 수 있듯, 대학교 홈페이지를 일일이 방문하거나 과사무실에 직접 문의해야 함.

Country	Program	University/Institution	가능학과/Available area of study	불가능학과/Unavailable area of study
Australia	교환 ESP	Australian National University	Almost all of our undergraduate courses are open to exchange	If a course is not offered for exchange students it would be
Australia	교환 ESP	Bond University		
Australia	교환 ESP	Curtin University	Use our course unit finder http://handbook.curtin.edu.au/	Education, OT, nursing, midwifery, medicine, speech pathology
Australia	교환 ESP	Deakin University	http://www.deakin.edu.au/study/undergraduate/enrolment-fees	Medicine and optometry
Australia	교환 ESP	Griffith University	https://www.griffith.edu.au/international/global-mobility/inbound/study-abroad	
Australia	교환 ESP	La Trobe University		No one faculty, but some individual classes are not open
Australia	교환 ESP	Monash University	The majority of courses are open to exchange students. Details https://www.monash.edu.au/international/global-mobility/inbound/study-abroad/choosing/	Fourth year and honours, Off-campus learning
Australia	교환 ESP	Murdoch University	Any unit, as long as students meet the pre-requisites for this unit https://www.adelaide.edu.au/inbound-study-abroad/choosing/	Veterinary Sciences and Teaching/Practicum units
Australia	교환 ESP	University of Adelaide	https://www.adelaide.edu.au/inbound-study-abroad/choosing/	External Courses are not available. These courses are identified

“종합대학과 단과대학 분류하기”를
연구 목표로 설정함.

A high-angle, nighttime photograph of a dense urban skyline, likely Hong Kong. The image is filled with numerous skyscrapers and buildings, all of which are brightly lit with various colors of lights (yellow, white, blue, red). The lights create a vibrant, glowing effect against the dark night sky. The buildings are packed closely together, and the overall scene conveys a sense of a bustling, modern metropolis. A large, semi-transparent yellow rectangle is overlaid in the center of the image, containing the text.

02

데이터 선정 및 수집 과정

모든 교환학생들은 교환 복귀 이후 7주 내에 필수적으로 후기를 작성해야 함.

따라서 전수조사를 진행할 수 있음. 교환학생들의 학과 정보와 후기 정보를 연구 대상으로 선정함.

교환대학의 크기, 지리적 위치, 기후 등

* Aalto University school of Economics Mikkeli campus, 이 대학은 헬싱키를 기준으로 북쪽으로 200km쯤에 있는 작은 도시 미켈리에 있습니다. 헬싱키와 달리 이곳은 지하철이나 트램이 없고 다른 도시, 특히 여행을 위해 공항으로 가려면 기차나 버스를 타고 4~5시간 이동해야 합니다. 그래서 주말에 헬싱키를 제외한 다른 도시, 나라로의 여행은 비용과 시간을 고려했을 때 사실상 어렵습니다. 학기 전후나 아니면 한 코스를 비우고 3주동안 여행하시는 것이 합리적입니다. (저 또한 이 같은 방식으로 유럽투어를 했습니다)

이 지역의 기후는 다른 보고서에도 보셨듯이 매우 Extreme합니다. 처음 이 도시에 왔을 때가 8월이었는데 가디건은 입어야 할 정도로 쌀쌀했고 학기가 시작되고 9월이 채 지나기 전에는 눈이 내립니다. 또한 10월 말부터는 해가 짧아지기 시작해 12월이 가까이 올 때쯤에는 11시에 해가 떠서 2시면 해가 지는 광경을 목격하실 겁니다. 저는 한 학기까지만 해서 12월 말에 나왔지만 같이 있던 친구들 말을 빌리면 하루 3시간의 해, 영하 20도의 추위, 그리고 무릎높이의 눈이 4월초까지 이어진다고 합니다. 이처럼 미켈리의 기후는 한국과 매우 다르며 적응하기에 쉽지 않을 겁니다. 특히 소도시에 해가 안 뜨고 추운 날씨로 인해 기숙사에만 있다 보면 정신적인 스트레스로 우울증에 시달릴 수도 있습니다. 저 역시도 기후부분을 생각하지 않고 가서 적응하는데 많은 고생이 있었습니다.

교환학교의 크기는 정말 아담합니다. 5층건물 두 개가 있는데 그 중 하나는 교직원과 필요 없는 도서관이어서 모든 수업은 건물 하나에서 진행됩니다. 연세대학교의 광활한 백양로나 다큐멘터리에서 보던 미국대학교의 위용을 생각하면 조금 아니 매우 초라하다고도 느껴질 수 있습니다. 하지만 외형에 비해 내부시설이나 구조는 편리합니다. 강의실, 컴퓨터실, 강당, 그리고 식당이 있어 추운 날 밖에 나가지 않고 모든 것을 해결할 수 있습니다. 특히 지하에는 BAR가 있는데 매주 수요일 각기 다른 주제로 파티가 열립니다. 그리스 신화나 할로윈 등 여러 주제로 파티가 열리니까 자주 나가서 학생들과 친하게 지내는 게 좋습니다.

대학 주변 환경

* 도시가 소도시다 보니까 학교 주변에 딱히 설명할 만한 유명한 곳은 없습니다. 하지만 편의시설은 생활하는데 크게 불편한 점은 없을 정도로 갖추고 있습니다. 우선 학교에서 10분 거리에 위치한 다운타운엔 K-mart, S-market과 같은 마트와 음식점, H&M과 같은 상점이 있습니다. H&M은 미켈리에 몇 없는 메이커 의류이라 자주 갔었는데 한국 H&M에 비해 전반적으로 가격이 싸고, 날을 잘 맞춰가면 10유로 이내에 머플러나 바지도 살 수 있습니다. 또한 학교에서 20분정도 거리에 미켈리의 또 다른 대학교인 MAMK가 있습니다. 이 학교는 그래도 대학교라고 부를 수 있을만한 캠퍼스 크기에 실내체육관, 자전거 대여소, 그리고 헬스클럽이 있습니다. 특히 자전거 대여와 같은 경우 학생 수에 비해 턱없이 적기 때문에 미켈리에 오자마자 학교에 물어봐서 관련 절차를 밟고 자전거를 받으시길 바랍니다. 자전거 유무에 따라 활동 반경과 삶의 만족도에 큰 차이가 있기 때문에 빌리시는걸 추천합니다. 마지막으로 호수의 나라답게 도시를 조금만 벗어나면 곳곳에 숲과 호수가 자리잡고 있습니다. 호숫가를 따라 자전거를 타거나 조깅을 하는 것도 미켈리의 매력이 아닐까 생각합니다.

Beautifulsoup, Selenium으로 학과 정보 및 후기 크롤링(10시간)

	A	B	C	D	E	F
1	0	1	2	3	4	5
2	Australian	Bond Univ	Curtin Univ	La Trobe U	Monash U	University
3	UD IS	??????????	??????????	??????????	UIC	GSIS
4	??????????	??????????	??????????	??????????	UIC	??????????
5	??????????	??????????	??????????	??????????	??????????	??????????
6	??????????	??????????	??????????	??????????	??????????	??????????
7	??????????	??????????	??????????	??????????	??????????	??????????
8	??????????	??????????	??????????	??????????	??????????	??????????
9	??????????	??????????	??????????	??????????	??????????	??????????
10	??????????	??????????	??????????	??????????	??????????	??????????
11	??????????	??????????	????????	??????????	??????????	??????????
12	??????????	??????????	??????????	??????????	??????????	??????????
13	??????????	??????????	??????????	??????????	??????????	??????????
14	??????????	??????????	??????????	??????????	??????????	??????????
15	??????????	??????????	??????????	??????????	??????????	??????????
16	??????????	??????????	??????????	??????????	??????????	??????????
17	??????????	??????????	??????????	??????????	??????????	??????????

```
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
from selenium import webdriver
import csv
from urllib.parse import urlparse

def crawl_as_csv(univ_query):
    page = 1
    dummy_data1 = {}
    df = pd.DataFrame(dummy_data1)
    while page:
        url = "https://oia.yonsei.ac.kr/partner/expReport.asp?page=" + str(page) + "&cur_pack=0&ucode="+str(univ_query)+"&bgbn=A"
        res = requests.get(url)
        soup = BeautifulSoup(res.content, 'lxml')
        table = soup.find_all('table')[0]
        df_crawl = pd.read_html(str(table), encoding='utf-8', header=0)[0]
        df_crawl['href'] = [np.where(tag.has_attr('href'), tag.get('href'), "no link") for tag in table.find_all('a')]
        if not df_crawl.empty:
            page += 1
            df = pd.concat([df, df_crawl], sort=False)
        else:
            print(df)
            break
    df_without_index = df.reset_index()
    print(df_without_index)
    df_without_index.to_csv(r'C:/Users/pc/Documents/GitHub/OIA_Text_Wrangling/dataf/'+univ_query+'.csv', index=False, encoding="utf-8")
```

```
def crwl_as_csv(univ_query):
    page = 1
    dummy_data1 = {}
    df = pd.DataFrame(dummy_data1)
    while page:
        url = "https://oia.yonsei.ac.kr/partner/expReport.asp?page=" + str(page)+"&cur_pack=0&ucode="+str(univ_query)+"&bgbn=A"
        res = requests.get(url)
        soup = BeautifulSoup(res.content,'lxml')
        table = soup.find_all('table')[0]
        df_crawl = pd.read_html(str(table),encoding='utf-8', header=0)[0]
        df_crawl['href'] = [np.where(tag.has_attr('href'),tag.get('href'),"no link") for tag in table.find_all('a')]
        if not df_crawl.empty:
            page += 1
            df = pd.concat([df, df_crawl],sort=False)
        else:
            print(df)
            break
    df_without_index = df.reset_index()
    print(df_without_index)
    df_without_index.to_csv(r'C:/Users/pc/Documents/GitHub/OIA_Text_Wrangling/dataf/'+univ_query+'.csv',index=False,encoding="utf-8")
```

```
def input_text(href):
    empty_lst = []
    review_url = "https://oia.yonsei.ac.kr" + href
    res = requests.get(review_url)
    soup = BeautifulSoup(res.content, 'lxml')
    body_cursor_list = soup.find_all("div", {"class": "exp_txt"})
    # options = webdriver.ChromeOptions()
    # options.add_argument('headless')
    # driver = webdriver.Chrome(r"C:\Users\pc\Desktop\chromedriver", options=options)
    # driver.implicitly_wait(1) # waiting web source for one second implicitly
    # driver.get(review_url)

    # selenium body cursor list
    # body_cursor_list = driver.find_elements_by_class_name("exp_txt")

    text_list = []
    for i in body_cursor_list:
        text_list.append(i.text)

    text_list = text_list[1:]
    return text_list

# input_text("/partner/expReport.asp?id=15129&page=1&bgbn=R")
```



```
def combining_into_csv(file_name):
    initial_df = pd.read_csv(r'C:/Users/pc/Documents/GitHub/OIA_Text_Wrangling/dataf/'+file_name)
    stacked_list = []
    for item in initial_df['href']:
        print(item)
        text_list = input_text(item)
        stacked_list.append(text_list)
    univ_text_df= pd.DataFrame(np.row_stack(stacked_list),
                               columns=["gen_info", "env_info", "food_info", "study_info", "office_info", "facil_info", "mhct_info", "help_i
    print(univ_text_df)
    univ_text_df.to_csv(r'C:/Users/pc/Documents/GitHub/OIA_Text_Wrangling/dataf/'+file_name+'_text_data.csv',index=False,encoding="utf-8")

def make_file(university_query):
    crwl_as_csv(university_query)
    file_name = university_query + ".csv"
    combining_into_csv(file_name)

def give_cde_from_url(href):
    query =urlparse(href)
    query_list = query.query.split('=')
    query_list2 = query_list[1].split('&')
    univ_code = query_list2[0]
    print(univ_code)
    return univ_code
```

한글 데이터 (UTF-8) 변환

1	2	3	4	5
UD IS	경영학	경제학	경제학	경제학
경영학	경제학	경제학	국어국문	기계공학
경영학	경제학	국어국문	사회학	사회환경
경영학	경영학	교육학	교육학	금속시스
UIC	UIC	경영학	경영학	경영학
GSIS	건축학	경영학	경영학	경영학
Informatic	건축공학	경영학	경영학	경영학
UD PSIR	UD CLC	UIC ASD	UIC	UIC
UIC TAD	건축공학	건축학	건축학	경영학
경영학	경영학	경영학	경영학	경영학
경영학	경제학	경제학	경제학	독어독문
경영학	경영학	경영학	경영학	경영학
경영학	경영학	경영학	경영학	경영학
감호학	경영학	경영학	경영학	경영학

정형데이터가 아니라 비정형데이터이기 때문에 학과 이름이 천차만별이었음.

1. 명칭이 축약된 경우

- "정치외교학과", "정외"

2. 명칭이 제각각인 경우

- "언론홍보영상학과", "언론홍보영상학", "언론홍보", "언론홍보학"
- "국어국문학", "국어국문"

3. UIC의 명칭들

- " UIC PSIR ", " UD PSIR ", " 언더우드국제대학 정치외교학과",
"국제대학교 국제대 정치외교학과"

4. 복수전공인 경우에는 두 개의 학과로 나눠서 데이터를 처리함.

- Slash구분: "심리학/사회학", "경제학/응용통계학"
- 쉼표구분: "문헌정보학,응용통계학"

규칙을 세워서 Pandas로 학과 이름 통합 및 데이터 정제

Australian National University	Bond University	Curtin University	La Trobe University	Monash University	University of New South Wales
UD IS	경영학	경영학	경영학	UIC	GSIS
경영학	경제학	경제학	경영학	UIC	건축학
경제학	경제학	국어국문	교육학	경영학	경영학
경제학	국어국문	사회학	교육학	경영학	경영학
경제학	기계공학	사회환경시스템공학	금속시스템공학	경영학	경영학
경제학/정치외교학	도시공학	사회환경시스템공학	물리치료학	경영학	경영학
노어노문	불어불문	생명공학	불어불문	경영학	경영학
대기학	사회복지학	생화학	불어불문	경영학	경영학
문헌정보/경영학	상경계열	신학	생물학	경제학	경제학
문화인류학	상경계열	영어영문	생활디자인	교육학	경제학
생활디자인	상경계열	영어영문	신문방송학	교육학	경제학
생활디자인학/경영학	식품영양학	의공학	신문방송학	독어독문	국어국문
식품영양학	신문방송학	전기전자공학	심리학	문헌정보학	기계공학
식품영양학	신문방송학	주거환경	영어영문	문헌정보학	기계공학

분류를 위해서는 계량데이터가 필요한데, 계량데이터가 없음.

- 각 대학 별로 {학과1: 학생수, 학과2: 학생수 ...} 으로 딕셔너리를 만듦.
- 한 대학 내에서 다른 학과 간의 학생수 차이를 기반으로 분산값을 구함.

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

- 분산값이 클수록 특정 학과에 학생들이 편중됐다는 것을 의미함.
- 분산값이 작을 수록 학과 별로 학생들이 고르게 분포한다는 것을 의미함.

분산값을 계량값으로 구해서, {대학 이름, 분산 값, 학생수}로 dataframe을 만듦.

1	name	variance	no_of_students
2	Aalto University	101.36	72
3	American University	25.69	108
4	Aoyama Gakuin University	0.72	41
5	Arizona State University	10.12	77
6	Australian National University	0.88	27
7	Baylor University	6.39	44
8	Biola University	0.45	17
9	Bocconi University	50.36	48
10	Bond University	0.40	19
11	Brock University	4.50	29
12	Charles University	1.48	36
13	Chinese University of Hong Kong	10.38	91

```
def single_dp_dict(df_column):
    # university_name = df_column[0]
    single_column = df_column[0:]
    single_column_list = single_column.to_list()
    column_lst_without_nan = [x for x in single_column_list if x == x]

    splitted_list = []
    for i in column_lst_without_nan:
        if "/" in i:
            # print(i)
            double_element = i.split("/")
            # print(double_element)
            # splitted_list.remove(i)
            splitted_list += double_element
        elif "," in i:
            double_element = i.split(",")
            splitted_list += double_element
        else:
            splitted_list.append(i)
        pass
    splitted_list

    from collections import defaultdict
    fq= defaultdict( int )
    for w in splitted_list:
        fq[w] += 1
    number_of_departments = len(splitted_list)
    # print(number_of_departments)
    # print(university_name)
    dictionary = dict(fq)
    return dictionary
```

```
def fn_univ_variance(univ_name):
    df_column = df[univ_name]
    single_dict = single_dp_dict(df_column)
    var = variance(single_dict[k] for k in single_dict)
    return var

def no_of_students(univ_name):
    df_column = df[univ_name]
    # print(df_column)
    single_dict = single_dp_dict(df_column)
    # print(single_dict)
    no_of_students = sum(single_dict[k] for k in single_dict)
    return no_of_students

def no_of_departments(univ_name):
    df_column = df[univ_name]
    # print(df_column)
    single_dict = single_dp_dict(df_column)
    no_of_departments = len(single_dict)
    return no_of_departments

def max_department(univ_name):
    df_column = df[univ_name]
    # print(df_column)
    single_dict = single_dp_dict(df_column)
    maximum_dep = max(single_dict.items(), key=operator.itemgetter(1))[0]
    return maximum_dep
```

```
[ ] var_list = []
for univ in header_list:
    var = fn_univ_variance(univ)
    students_no = no_of_students(univ)
    department_no = no_of_departments(univ)
    max_dep = max_department(univ)
    var_dict = {'name':univ,
               'variance':var,
               'no_of_students':students_no,
               'no_of_departments':department_no,
               'maximum_department':max_dep}
    var_list.append(var_dict)

depart_var_df = pd.DataFrame(var_list)
depart_var_df
# depart_var_df.to_csv("/content/gdrive/My Drive/_OIA_Project/_department_var_df_mark1.csv",index=False)
```

	name	variance	no_of_students	no_of_departments	maximum_department
0	Australian National University	0.882353	27	17	경제학
1	Bond University	0.401099	19	14	상경계열
2	Curtin University of Technology	0.131868	16	14	사회환경시스템공학
3	La Trobe University	1.897436	22	13	영어영문
4	Monash University	1.747036	40	23	경영학
...
177	Washington College	0.980952	23	15	경영학
178	Washington University in St. Louis	4.576923	25	13	경영학
179	Western Kentucky University	1.435897	20	13	경영학
180	Western Washington University	1.435897	19	13	경영학
181	Westminster College	3.606061	26	12	심리학

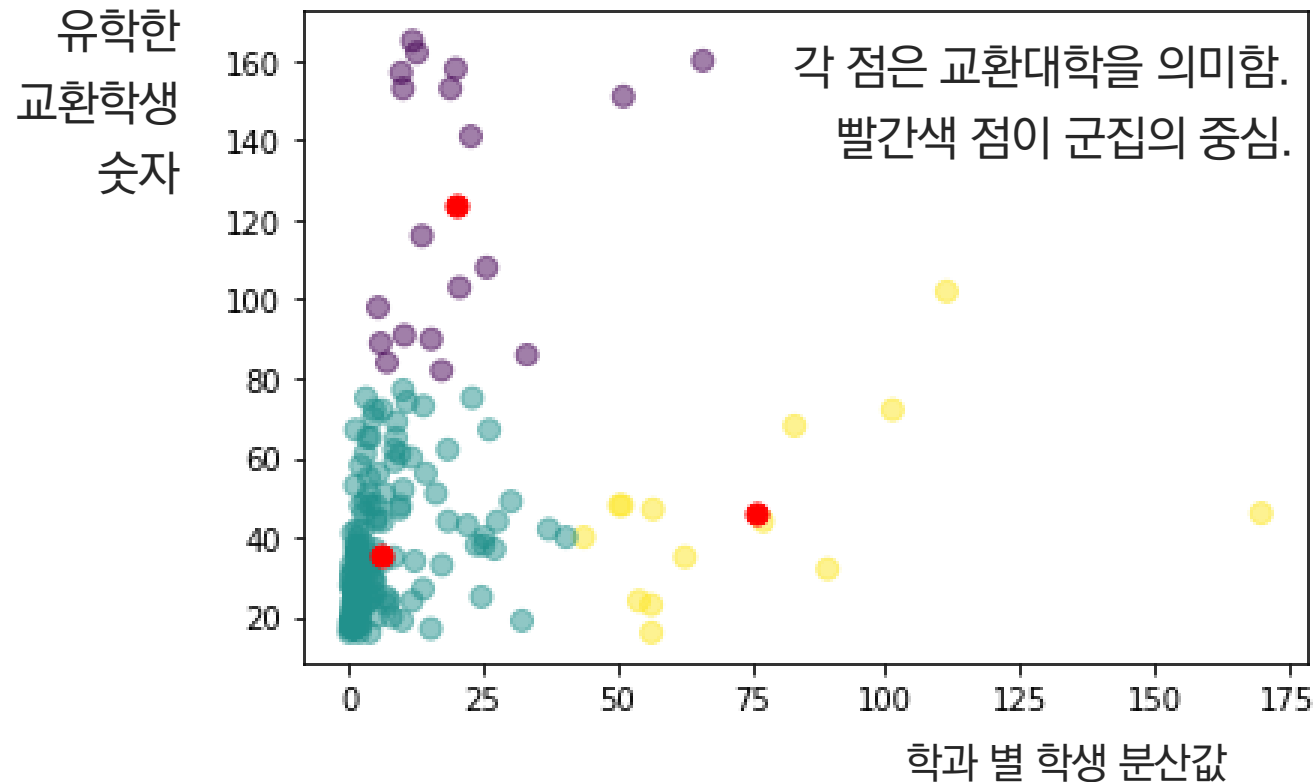
A high-angle, nighttime photograph of a dense urban skyline, likely Hong Kong. The image is filled with numerous skyscrapers and buildings, all of which are brightly lit with various colors of lights (yellow, white, blue, red). The lights create a vibrant, glowing effect against the dark night sky. In the foreground, some buildings are more prominent, showing their architectural details. In the background, the city extends towards a body of water, where some lights are reflected. The overall atmosphere is one of a bustling, modern metropolis.

03 데이터 분석

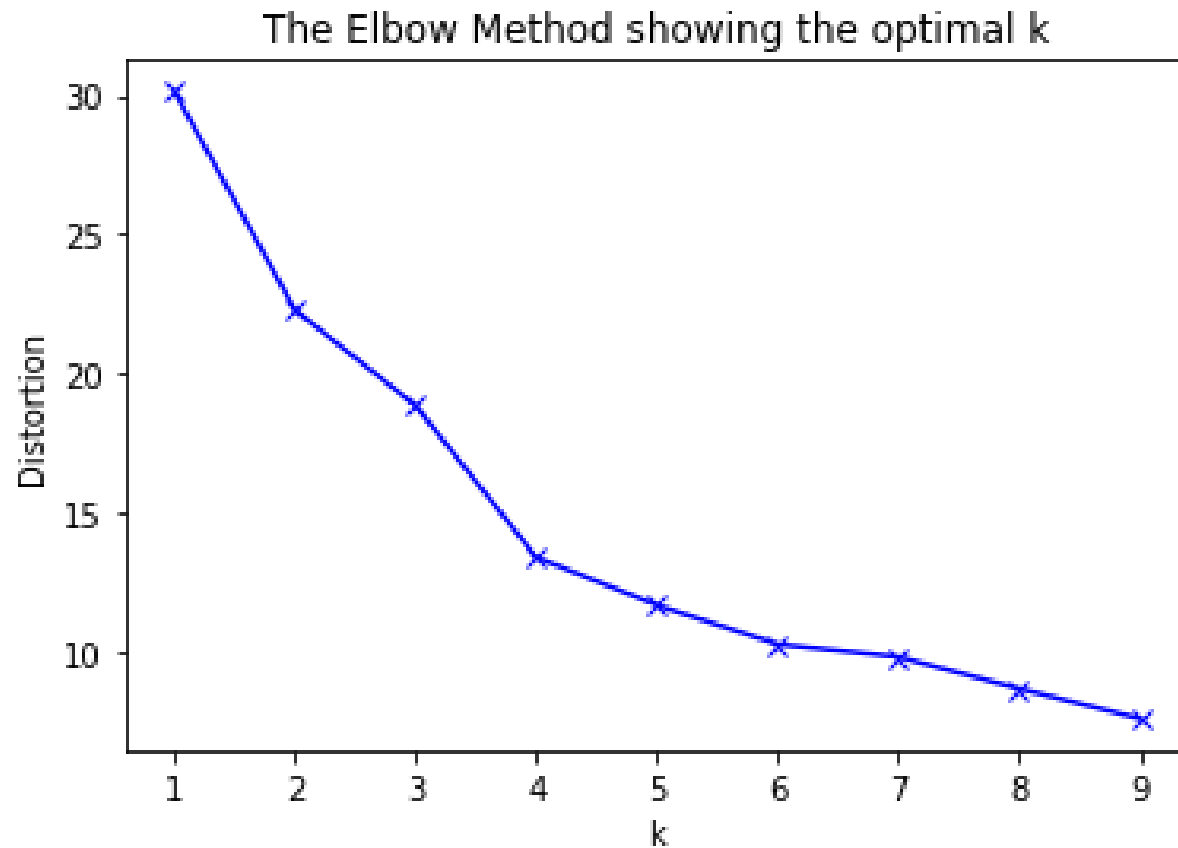
분산값, 재학 학생들 값들이 주어졌으며, 이것만으로 종합대학교와 단과대를 구분해야 함.

1	name	variance	no_of_students
2	Aalto University	101.36	72
3	American University	25.69	108
4	Aoyama Gakuin University	0.72	41
5	Arizona State University	10.12	77
6	Australian National University	0.88	27
7	Baylor University	6.39	44
8	Biola University	0.45	17
9	Bocconi University	50.36	48
10	Bond University	0.40	19
11	Brock University	4.50	29
12	Charles University	1.48	36
13	Chinese University of Hong Kong	10.38	91

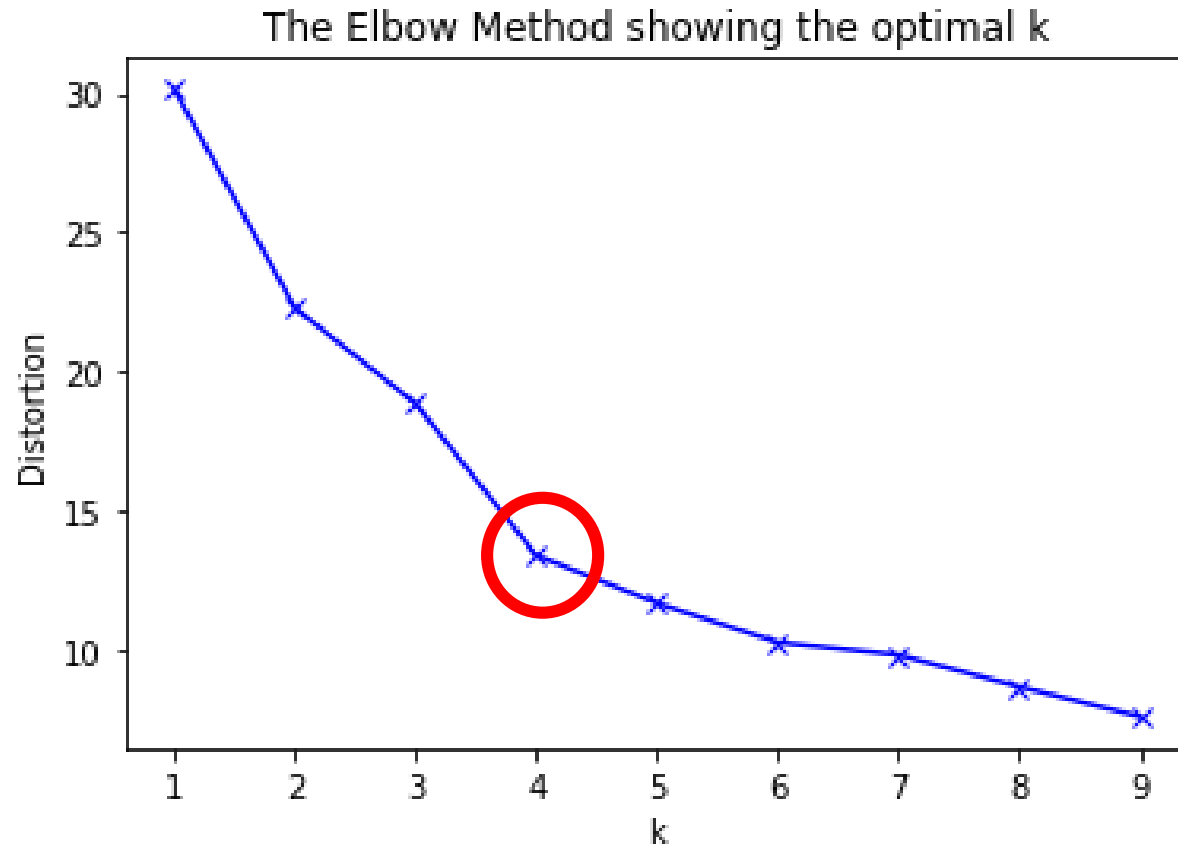
비지도학습기법인 K-means Clustering을 이용해서 컴퓨터로 하여금 분류하게 함.



적합한 K 값을 찾기 위해서 The Elbow Method를 이용함.



기존에 임의로 잡았던 K값인 3 대신에, K값을 4로 선정함.



```
from sklearn.cluster import KMeans
from sklearn import metrics
from scipy.spatial.distance import cdist
import numpy as np
import matplotlib.pyplot as plt

x1 = training_df['var'] #this is x axis
x2 = training_df['size'] #this is y axis

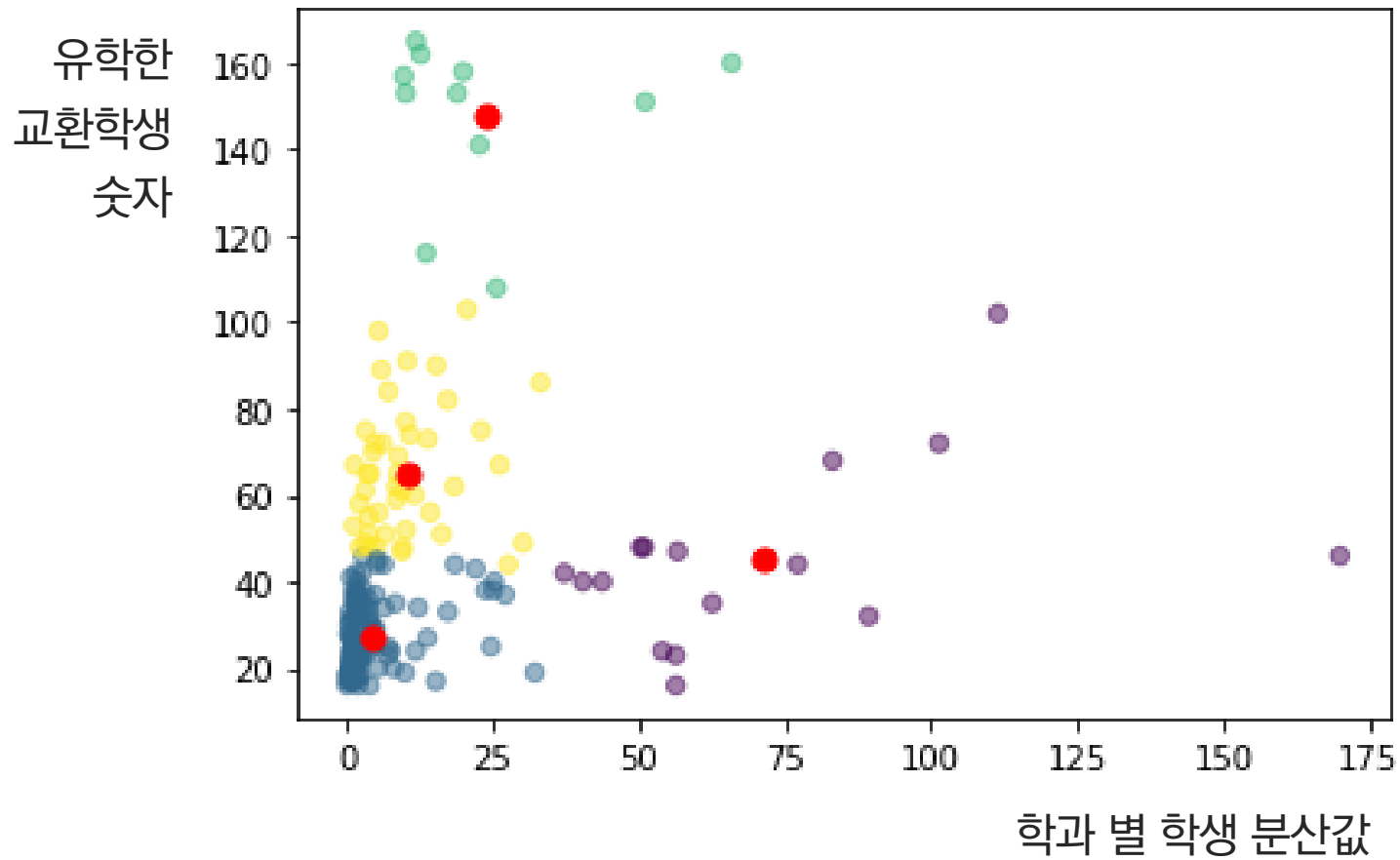
plt.plot()
plt.xlim([-10, 200])
plt.ylim([0, 200])
plt.title('Dataset')
plt.scatter(x1, x2)
plt.show()
```

```
# create new plot and data
plt.plot()
X = np.array(list(zip(x1, x2))).reshape(len(x1), 2)
colors = ['b', 'g', 'r']
markers = ['o', 'v', 's']

# k means determine k
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k).fit(X)
    kmeanModel.fit(X)
    distortions.append(sum(np.min(cdist(X, kmeanModel.cluster_centers_, 'euclidean'), axis=1)) / X.shape[0])

# Plotting the elow
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

K = 4일 때 K-Means Clustering



```
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

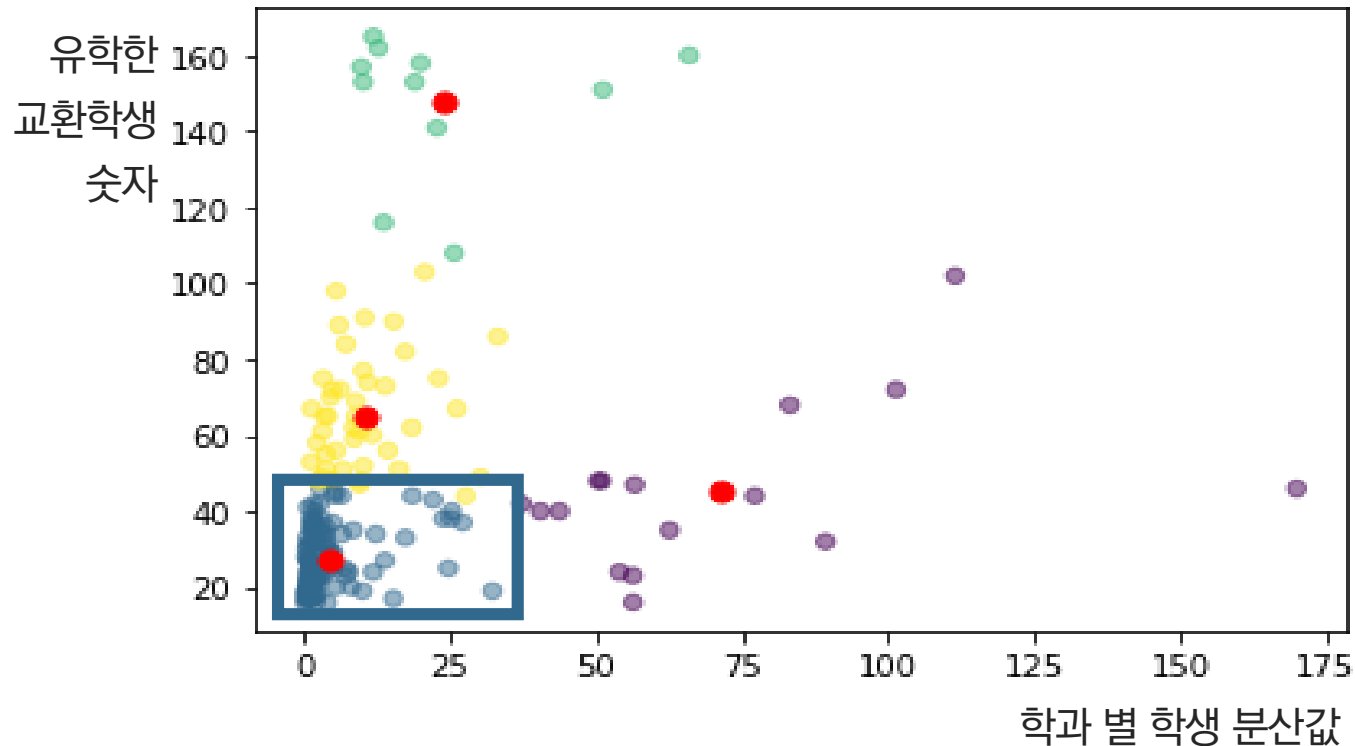
#variance as x axis
#number of students as y axis
training_df = pd.concat([depart_var_df['variance'], depart_var_df['no_of_students']], axis=1, keys=['var', 'size'])

kmeans = KMeans(n_clusters=4).fit(training_df)
centroids = kmeans.cluster_centers_
print(centroids)

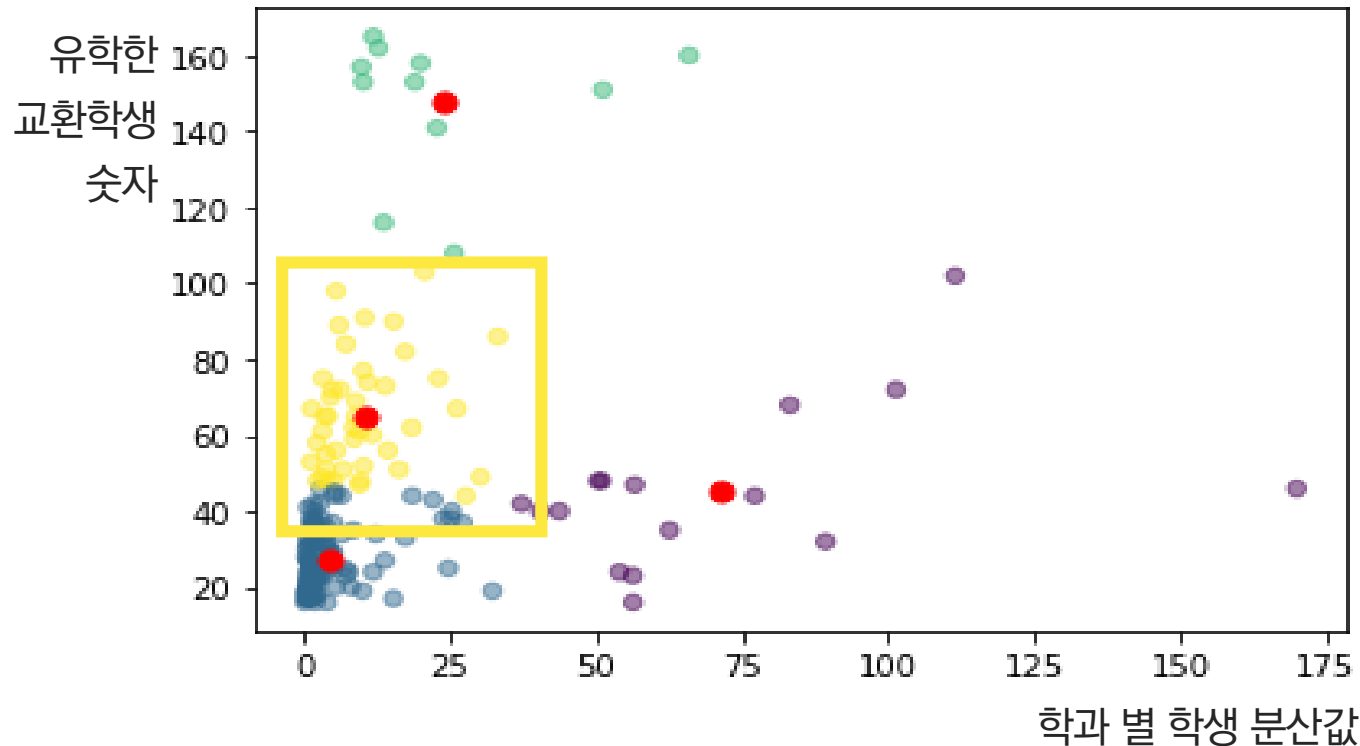
plt.scatter(training_df['var'], training_df['size'], c= kmeans.labels_.astype(float), s=30, alpha=0.5)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=50)
```


An aerial night view of a city skyline, likely Hong Kong, with numerous illuminated skyscrapers and a body of water in the background. A large yellow rectangular box is centered over the image, containing the text '04 데이터 해석'.

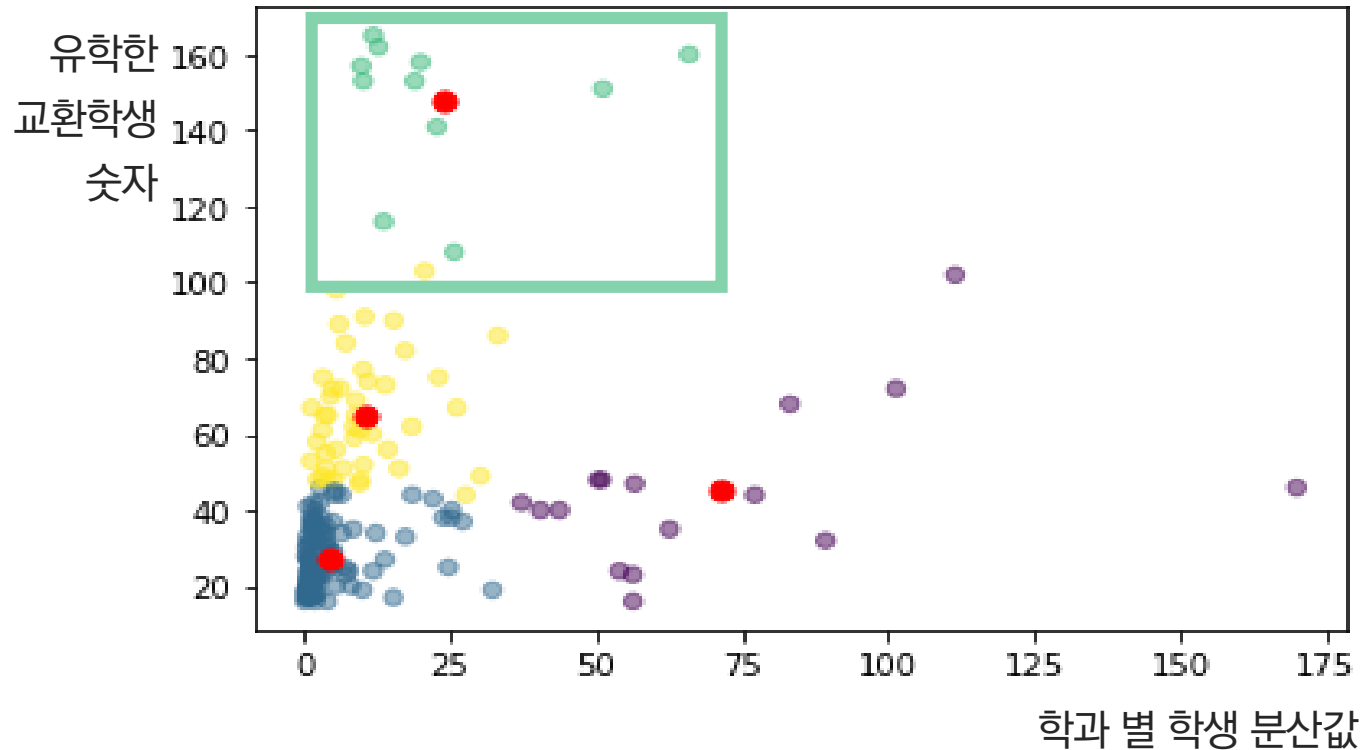
04 데이터 해석



파란색 군집은 학과들이 분포된 종합대학이지만, 많은 학생들이 가지는 않는 대학들.
Meiji Gakuin University 같이 많은 정원을 뽑지 않는 종합대학이 해당.

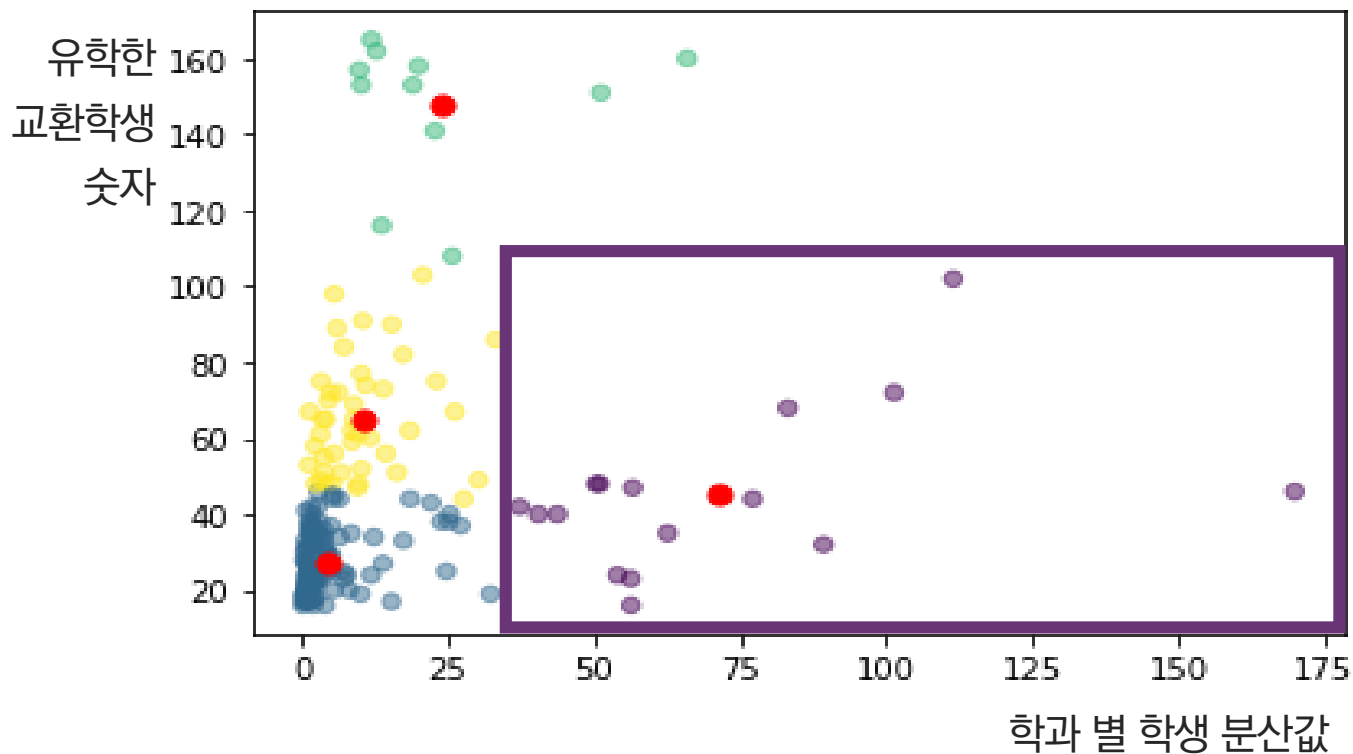


노란색 군집은 학과들에 따라서 학생들이 고르게 분포한 종합대학들.
진학하는 학생들이 많고, 대학 정원이 큰 경우. UCLA 같은 경우가 이 경우에 해당.



초록색 군집에 해당하는 대학들을 확인해보니, “학생 수가 많은 종합대”였음.

- 초록색 점들은 분산값이 이전 그룹들보다 높음. 비록 특정학과에 편중되어 있어도, 학생수가 충분히 많으면 종합대로 분류가 되는 것임.
- 분산값이 제일 높은 Maastricht University에 교환을 간 100명 중에 68명이 경영학, 경제학이었음. 그러나 교환학생들의 학과 종류 수는 21개였음.



보라색 군집이 단과대학교들: 독어독문/노어노문, 경영학 단과대 등

A high-angle, nighttime photograph of a dense urban skyline, likely Hong Kong. The image is filled with numerous skyscrapers and buildings, all illuminated with various lights, creating a vibrant and complex pattern of light against the dark night sky. The lights from the buildings reflect on the water in the foreground, which appears to be a harbor or bay. The overall scene conveys a sense of a bustling, modern metropolis.

05 성능 평가

"특정 학과 적합한 대학교(보라색 Cluster)"로 분류할 확률 정확도는 93%

대학 이름	X Axis (분산)	Y axis (학생 수)	특징	Cluster
Freie Universitat Berlin	37.23	42	독어독문 특화	보라색
European Business School	40.42	40	경영학 특화	보라색
KEDGE Business School	43.74	40	경영학 특화	보라색
Bocconi University	50.36	48	상경대 특화	보라색
Jonkoping International Business School	50.90	48	상경대 특화	보라색
SUNY at Albany	51.11	151	종합대	초록
ESADE	54.00	24	경영학 특화	보라색
Moscow State University	56.30	23	노어노문 특화	보라색
New York University - Stern School of Business	56.33	16	경영학 특화	보라색
University of Manitoba	56.62	47	상경대 특화	보라색
York University: Schulich School of Business	62.57	35	상경대 특화	보라색
University of Washington	65.89	160	종합대	초록
Singapore Management University	77.16	44	경영학 특화	초록
CUNY - Baruch College	83.15	68	종합대	보라색
Saxion University	89.29	32	경영학 특화	보라색
Aalto University	101.36	72	경영학 특화	보라색
Maastricht University	111.43	102	상경대 특화	보라색
St. Petersburg State University	169.93	46	노어노문 특화	보라색

A high-angle, nighttime photograph of a dense urban skyline, likely Hong Kong. The image is filled with numerous skyscrapers and buildings, all illuminated with various lights, creating a vibrant and complex pattern of light against the dark night sky. The lights from the buildings reflect on the water in the foreground. The overall scene conveys a sense of a bustling, modern metropolis.

06

대학 특징 시각화

코펜하겐대학교 워드클라우드 빈도수로 단순 시각화 (불용어처리 O)



45

University of Minnesota Morris 워드클라우드 빈도수로 시각화 (불용어처리 O)



지리적으로 미국 중서부에 속
월초에도 눈을 봤습니다. 춥긴
정도라고 생각하시면 됩니다
무척 추웠습니다. 제가 도착
기보다 조금 작은 것 같습니다

미네소타가 겁나게 춥긴 춥나봐
요

겨울 얘기 밖에 없어요...

ㅋㅋㅋㅋㅋㅋㅋㅋ

9:30 AM




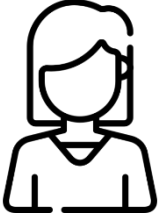

Muzi showing off money

마자 거기 진짜

못살아

9:30 AM

불용어사전을 제작하고, Mecab 사전을 이용하여 워드클라우드로 시각화함.
밑에 세 가지 자료들을 참고하여 데이터를 시각화 함.

	<p>“청와대 국민청원 데이터 분석” - 박조은 (2019.01)</p> <ul style="list-style-type: none">• 불용어 사전 제작, 빈도수 높은 명사 추출, 워드클라우드 시각화 자료 참고• KONLPY 사용법 및 Mecab 사전 사용법 숙지
	<p>“한국어와 NLTK, Gensim의 만남” - 박은정(2015)</p> <ul style="list-style-type: none">• 명사 Tokenizing 자료 참고.• KONLPY 사용법 숙지.
	<p>“한국어 임베딩” - 이기창 (2019.09)</p> <ul style="list-style-type: none">• TF-IDF 알고리즘 학습 및 명사 Tokenizing 자료 참고.

```
def noun_bow(df):
    bow = []
    for element in df:
        nouns = mecab.nouns(element)
        print(nouns)
        print(len(nouns))
        bow = bow + nouns
        print(len(bow))
    return bow

from collections import Counter
counted_nouns = Counter(noun_bow(df['gen_info']))
counted_nouns
tags = counted_nouns.most_common(100)

import pytagcloud
taglist = pytagcloud.make_tags(tags,maxsize=100)
tag_list = []

stopwords_kr = ['미네소타', '캠퍼스', '덴마크', '코펜하겐 대학교', '여름', '겨울', '날씨', '대학', '모리스', '미국',
                '때문', '정도', '경우', '학생', '교환', '학교', '문화', '충격', '기숙사', '한국', '국제', '교육부', '하
                '그럼', '이런', '저런', '합니다', '많은', '많이', '정말', '너무', '여기', '이곳', '우리', '학부', '사

for n, item in tags:
    if n not in stopwords_kr:
        tag_list.append(n)

displayWordCloud(' '.join(tag_list))
```

```
# =====
# -- TF-IDF function
# =====
def f(t, d):
    # d is document == tokens
    return d.count(t)

def tf(t, d):
    # d is document == tokens
    return 0.5 + 0.5*f(t,d)/max([f(w,d) for w in d])

def idf(t, D):
    # D is documents == document list
    numerator = len(D)
    denominator = 1 + len([ True for d in D if t in d])
    return log10(numerator/denominator)

def tfidf(t, d, D):
    return tf(t,d)*idf(t, D)

def tokenizer(d):
    # return [ t for t in d.split() if len(t) > 1 ]
    return d.split()

def tfidfScorer(D):
    tokenized_D = [tokenizer(d) for d in D]
    result = []
    for d in tokenized_D:
        result.append([(t, tfidf(t, d, tokenized_D)) for t in d])
    return result

if __name__ == '__main__':
    corpus = df['gen_info'].tolist()

    for i, doc in enumerate(tfidfScorer(corpus)):
        print('===== document[%d] =====' % i)
        print(doc)
```



07 토의 사항

난관

- Pandas 라이브러리를 이용해서 비정형 텍스트 데이터를 전처리하는 것이 까다로웠음.

프로젝트 진행 시 학습한 것

- 지금껏 소외되어 왔던 교환학생 데이터를 corpus로 제작하면서 Pandas 사용법을 숙지함.
- K-Means Clustering을 Python3으로 실습을 진행함.
- 교환학생 후기 corpus를 바탕으로 Mecab, Twitter 등의 사전을 이용해서 명사 Tokenizing 방법을 학습함.
- 교환학생 후기 corpus를 바탕으로 TF-IDF를 Python3으로 실습함.

한계점

- 교환 생활에 유용한 정보인 “대학의 위치”를 시각화하지 못했음.
- 아직 모든 대학의 워드클라우드를 제작하지 못했음.

모든 대학의 워드클라우드를 대학의 위치와 함께 시각화해서 1월 10일 이전에 배포할 예정

End of Document

교환학생 후기 데이터 분석

해외로 교환학생을 가는 연세대 학생을 위해서 전공 학점인정이 가능한 대학을 분류하고, 해외 생활정보를 시각화함.

문제 인식	<ul style="list-style-type: none"> 교환학생을 준비하는 연세대학교 학생들은 전공 학점 인정, 해외 생활 환경에 관심이 많음. 하지만 연세대학교 후기 게시판에서 국가 별로 / 대학 별로 후기들을 열람하기는 번거로움.
연구 설계와 연구 질문	<ul style="list-style-type: none"> 종합대학과 단과대학의 분류: 해외 파견대학에서 "전공 학점을 채우기 힘든 상황"을 방지하기 위해서 전공 학과가 고르게 존재하는 종합대학과, 특정 학과만이 존재하는 단과대학을 분류함. 해외 생활정보 시각화: 물가, 시내와의 근접성, 도심/교외, 기후 등을 각 대학별로 워드클라우드로 시각화함.
데이터 선정	<ul style="list-style-type: none"> 해외로 교환학생을 다녀온 연세대학교 학생들은 7주 이내에 필수적으로 후기를 작성해야 함. 후기 데이터는 비정형 자연어(문어체)이며, 기후/주변환경/식사 등 9개 항목들로 분류되어 있음.
데이터 수집 및 처리	<p>종합대학과 단과대학의 분류</p> <ul style="list-style-type: none"> BeautifulSoup4와 Selenium으로 후기 데이터 8326개 크롤링. Pandas, statistic을 이용해서 파견 인원 수와 학과 별 인원의 분산값으로 2차원 df를 제작함. Scikit-learn, matplotlib을 이용해서, 각 파견 대학들을 K-Means로 4가지 군집으로 분류함. <p>해외 생활정보 시각화</p> <ul style="list-style-type: none"> KONLPY를 이용해서 문서에서 자주 등장하는 명사를 워드클라우드로 제작함. TF-IDF와 KONLPY를 이용해서 같은 대학의 9개 항목 문서들 간에 공통적으로 등장하는 명사는 비중을 낮추고, 각 문서들에서 자주 등장하는 명사들을 갖고 워드클라우드를 제작함.
결과 분석	<ul style="list-style-type: none"> 종합대학 / 단과대학을 분류하는 정확도(accuracy)는 93%로 나옴. 워드클라우드는 각 파견대학의 차이점들을 시각화한 것에서 의의가 있음.
토의	<ul style="list-style-type: none"> Pandas 라이브러리를 이용해서 데이터를 전처리하는 것이 까다로웠음. 지금껏 소외된 교환학생 데이터를 corpus로 만들고, 분류와 시각화를 한 것에서 의의가 있음.

교환학생 후기 데이터 분석

해외로 교환학생을 가는 연세대 학생을 위해서 전공 학점인정이 가능한 대학을 분류하고, 해외 생활정보를 시각화함.

- 연세대학교 교환후기 데이터 수집 코드

https://github.com/snoop2head/OIA_Text_Wrangling/blob/master/data_collect.py

- 교환후기 Corpus (CSV Format, 364 files, 96MB)

<https://drive.google.com/drive/u/0/folders/1iSXS9UVp0d0J-s2hRVKZGUTWfYOI4-f0>

- 해외 교환대학을 종합대학, 단과대학으로 분류하기 (Jupyter Notebook)

https://colab.research.google.com/drive/1-ALy9nUHQ9ioJYkDHdbd_0BOTiadhbpm

- 대학 별 특징 워드클라우드로 시각화하기 (Jupyter Notebook)

https://colab.research.google.com/drive/1_DYkc6sJqcUF7DRRZHUrOjsMan0uD5zi