2025-1 데이터 마이닝 기말 프로젝트

자연어처리를 활용한 강의평가 데이터 분석 교수님 추천 서비스

ProfPick팀 : 임나리 장재훈 최윤서 한서희

https://github.com/hanseohee031/ProfPick

HALLYM UNIVERSITY

프로젝트 소개

ProfPick Public

Pin | Watch 0 | Fork 0 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

About

2025-1 데이터마이닝 프로젝트 <자연어 처리를 활용한 강의평가 데이터 교수님 추천 서비스>

닫기

hanseohee031 Update README.md ded526c · 2 days ago 26 Commits

core	Initial commit	5 days ago
data	나리 완료	
data_preprocessing	Add files via upload	
image	Add files via upload	
professorpick	모델 올릴 준비 끝	
recommend	나리 완료	
templates	나리 완료	
users	일단 수정	
.gitignore	Initial commit	
README.md	Update README.md	
manage.py	Initial commit	
yunseo_reulst.html	윤서 끝	

README

ProfPick (Professor Pick)

자연어처리를 활용한 강의평가 데이터 분석 교수님 추천 서비스

- 2025-1 데이터마이닝 기말 프로젝트

127.0.0.1:8000

ProfPick

회원가입 로그인

교수님 추천 서비스에 오신 걸 환영합니다!

회원가입 후 프로필에서 원하는 교수님을 입력하면, 맞춤 추천을 받을 수 있습니다 :)

자연어처리를 활용한 강의 평가 데이터 분석 교수님 추천 시스템

2025-1 데이터마이닝 프로젝트 | 김유섭 교수님

팀원: 임나리 · 장재훈 · 최윤서 · 한서희

서비스명: ProfPick

프로젝트 소개

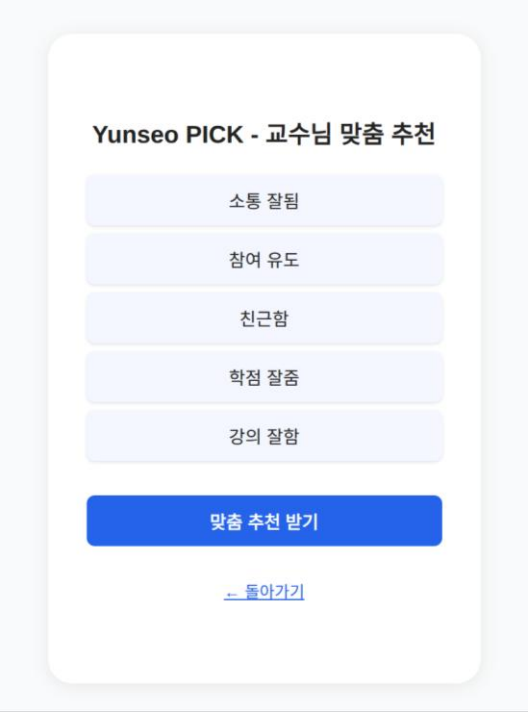
소프트웨어학과 수업을 듣는 학생들을 대상으로

직접 설문 조사한 데이터와 **에브리타임 강의평가 데이터**를 활용하여,

교수님들에 대한 정보가 없는 신입생 및 재학생을 위해

학생 개개인에게 적합한 교수님을 추천하는 시스템입니다.

(설문조사는 네이버폼으로 진행하였으며 소프트웨어학과 수업을 듣는 총 50명의 학생이 참여하였습니다)



우선순위 설정 후
맞춤 추천 받기

프로젝트 소개

실습 수업에서는 항상 이미 준비된 데이터를 사용하지만,
저희는 직접 데이터를 구해보고 **원시(raw) 데이터를 정리·가공**하는 전 과정을 경험해 보고 싶었습니다.
따라서 단순히 주어진 데이터를 다루는 데 그치지 않고,
실제로 데이터를 수집하고 전처리하는 과정에 집중하였습니다.
이 과정을 통해 **빅데이터 시대에 '데이터'의 가치**를 몸소 느낄 수 있었습니다.

어떤 데이터를 확보하느냐에 따라,
그리고 어떤 형식으로 정리·전처리하느냐에 따라 분석 결과가 크게 달라진다는 것을 직접 체감했습니다.
기본적인 전처리(결측값 제거, 형식 통일)와 파일 통합을 완료한 뒤에는,
동일한 데이터를 가지고 **팀원 각각이 다른 방식으로 추가 전처리**를 적용해 보았습니다.
그 결과, 데이터 활용 방식에 따라 결과가 얼마나 다양하게 나타나는지를 확인할 수 있었고,
이는 데이터 처리 전략의 중요성을 다시 한번 깨닫게 해 주었습니다.

데이터

(1) 직접 설문조사

종료

데이터마이닝 기말프로젝트 설문조사: 교수님을 평가해봅시다

2025. 05. 13. 오후 06:48 ~ 2025. 05. 29. 오전 01:06 • 총 참여 50

종합 결과 참여자별 결과 일자별 참여수

전체

총 질문 6

1. 김유섭 교수님

아래의 키워드 중 3가지 이상을 선택해서 자유롭게 답변해주세요~

교수님 인상과 말투는 어떠신지 (ex. 친근하다. 무섭다. 말이 느리시다. 딕션이 좋으시다.)

강의 자료가 좋은지 (ex. 강의자료가 영어라 싫다. 강의자료만 봐도 이해가 간다. 오타가 많다.)

수업이 이해하기 쉬운지 (ex. 듣다보면 졸리다. 설명을 진짜 잘해주신다. 듣고만 있어도 쏙쏙 들어온다.)

과제량이 만족스러운지 (ex. 과제가 좀 많은 것 같다. 과제가 별로 없어서 좋다. 과제 변별력이 너무 없다.)

과제 난이도는 어떠한지 (ex. 몇시간이면 한다. 하루 종일 이것만 해야된다.)

학생 수업 참여 유도를 하시는 지 (ex. 학생들에게 질문하신다. 수업시간에 대답하는 학생들이 많다. 학생들이 자도 방치한다.)

시험 난이도는 만족스러운 지 (ex. 이정도면 괜찮은 것 같다. 시험이 쉽다. 외우라고 한것만 외워가면 문제 없다.)

수업에 대한 열정은 어떠신지 (ex. 적극적이다. 학생들이 질문하면 좋아하신다.)

시험 성적 입력이 빠르신지 (ex. 성적 입력을 바로바로 해주신다. 시험 치는 당일에 바로 나온다. 엄청 느리다)

출결 방식, 그리고 출결 관리가 철저한지 (ex. 지각해도 상관없다. 마지막에 출석 부를때만 있으면 된다. 전자 출석 이용하신다.)

참여자의 답변 입력란 (최대 2000자)

데이터

(2) 에브리타임 리뷰

get_everytime_review.ipynb

	professor	lecture_id	lecture_name	year	semester	text	rate	posvote
0	고영웅	2338758	소프트웨어캡스톤디자인	2024	1	점수도 잘나오고 기부니가 좋다 열심히만 하면 되는듯 교수님이 말씀하신거 다 수용하고	5	0
1	고영웅	2338758	소프트웨어캡스톤디자인	2024	1	이번에 링크사업단으로 넘어가면서 교수님이 잘 모르시는게 많았음 그리고 학생들끼리하는...	2	0
2	고영웅	2338758	소프트웨어캡스톤디자인	2023	2	캡스톤 졸업만 하고싶다는 생각이면 다른 분반을 추천합니다 졸업만 하자라는 생각으로 ...	5	1
3	고영웅	2338758	소프트웨어캡스톤디자인	2023	2	이미 질리도록 들으셨겠지만 그 만큼 캡스톤 같이 할 팀원 잘 찾는게 좋습니다... ...	5	1
4	고영웅	2338758	소프트웨어캡스톤디자인	2023	1	소융대 학생들에 수준을 알 수 있는 수업 4년동안 무엇을 배웠는지 모르는 사람들이 ...	5	1
...	...	...	...	...	...	...	...	...
4465	신범주	2700057	파이썬과학프로그래밍기초	2024	1	교수님 열정도 좋으시고 질문하면 이해할때까지 다 햐주시고 본인도 직접 하셔서 해결할...	5	0
4466	신범주	2700045	머신러닝	2024	1	교수님이 열심히는 가르쳐 주시는데 너무 설명을 어렵게 해주신다고 느꼈습니다. 그래도...	3	0
4467	신범주	2700045	머신러닝	2024	1	어려움 처음 배우면서 이걸 들으니 못 따라감 저학년은 듣지 마세요	3	0
4468	신범주	2700045	머신러닝	2024	1	머신러닝의 알고리즘을 수학적 수식으로 설명하는 수업 선수 학습내용(거의 필수)...	5	0
4469	신범주	2700045	머신러닝	2024	1	이론을 수학으로 알려주십니다. 머신러닝이 수학으로 만들어진거다보니 수학으로 수업을 ...	5	0

4252 rows × 8 columns

데이터

(1) 설문조사 데이터 survey_reviews_prep.ipynb

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Professor names
professor_names = ['김유섭', '김은주', '이정근', '양은샘', '신미영', '김선정']
```

"모르겠다", "수업을 안들어서 모른다", "기억 안남" 등의 응답을 NaN 처리하고 결측값 제거하기

1. **문장 유사도 기반 NaN 처리**
   - `SentenceTransformer`를 이용해 기준 문장("모르겠다", "들어본 적 없음", "기억 안난다" 등)을 먼저 임베딩하여 저장
   - 각 리뷰 문장에 대해 동일한 임베딩 모델로 벡터화하고, 기준 임베딩들과의 코사인 유사도를 계산
   - 유사도가 임계치(예: 0.85) 이상인 경우 `np.nan`으로 반환하여 "모름" 응답을 결측값으로 표시
2. **결측값 제거**
   - Pandas의 `dropna()` 또는 `DataFrame.dropna(axis=0, how='any')` 메서드를 사용해 NaN이 된 행(응답)들을 제거
   - 이렇게 하면 실제 수업을 들은 경험에 기반한 리뷰만 남겨, 이후 분석의 왜곡을 막습니다

Generate embeddings

In [ ]:
```
from sentence_transformers import SentenceTransformer, util
import numpy as np

# 1) 임베딩 모델 로드 (가볍고 빠른 모델 추천)
model = SentenceTransformer('snunlp/KR-SBERT-V40K-klueNLI-augSTS')

# 2) 기준 문장 리스트 (수업 안 들음/모름 의미)
reference_texts = [
    '수업을 들은 적이 없다',
    '수강한 적 없음',
    '기억이 안 난다',
    '모름',
    '머름',
    '_'
]

# 기준 문장 임베딩
ref_embeddings = model.encode(reference_texts, convert_to_tensor=True)

def to_nan_if_similar(text, threshold=0.55):
    emb = model.encode(text, convert_to_tensor=True)
    cosine_scores = util.pytorch_cos_sim(emb, ref_embeddings)
    max_score = cosine_scores.max().item()
    if max_score > threshold:
        print(f'NaN 처리됨: "{text}", 유사도: {max_score}')
        return np.nan
    return text

# professor_cols는 교수님별 컬럼 리스트
df['review'] = df['review'].apply(to_nan_if_similar)
```

In [ ]:
```
df.dropna(inplace=True)
```

In [ ]:
```
# 결측값 처리 후
df['professor'].value_counts()
```

전처리 과정

(2) 에브리타임 데이터 everytime_reviews_prep.ipynb

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Professor names
professor_names = ['김유섭', '김은주', '이정근', '양은샘', '신미영', '김선정']

# Everytime session
everytime_token = ''
```

동명이인 및 전공 외 과목 처리

In [ ]:
```
df = df[df['professor'].isin(professor_names)]
df = df[~df['name'].isin([
    '음성학과발음연습', '영어문법',
    # '오디세이세미나1', '오디세이세미나2(리더십과 기업가정신)', '오디세이세미나3', '오디세이세미나4',
    # '글로벌취업전략', '직무및기업탐색', '취업성공전략', '취업설계', '진로설계', '여대생커리어개발과취업전략', '해외취업및인턴준비.
])]
```

In [ ]:
```
df['professor'].value_counts()
```

In [ ]:
```
df['professor'] = pd.Categorical(
    df['professor'],
    categories=professor_names,
    ordered=True
)

# Sort and reset index
df = df.sort_values('professor').reset_index(drop=True)
```

In [ ]:
```
df.head()
```

전처리 과정

데이터 통합

Merge data

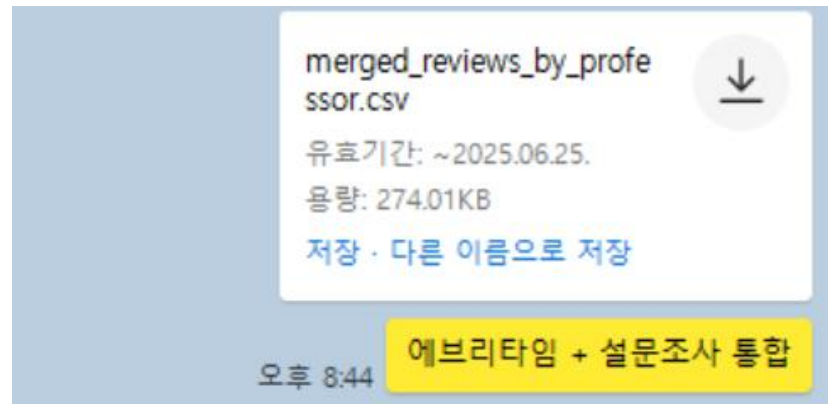
In [ ]:
```
everytime_reviews = pd.read_csv('everytime_reviews.csv', encoding='utf-8-sig')
survey_reviews = pd.read_csv('survey_reviews.csv', encoding='utf-8-sig')

everytime_reviews = everytime_reviews[['professor', 'review']]
survey_reviews = survey_reviews[['professor', 'review']]

df = pd.concat([everytime_reviews, survey_reviews], ignore_index=True)
```

In [ ]:
```
everytime_reviews['professor'].value_counts()
```

In [ ]:
```
survey_reviews['professor'].value_counts()
```

In [ ]:
```
df['professor'].value_counts()
```

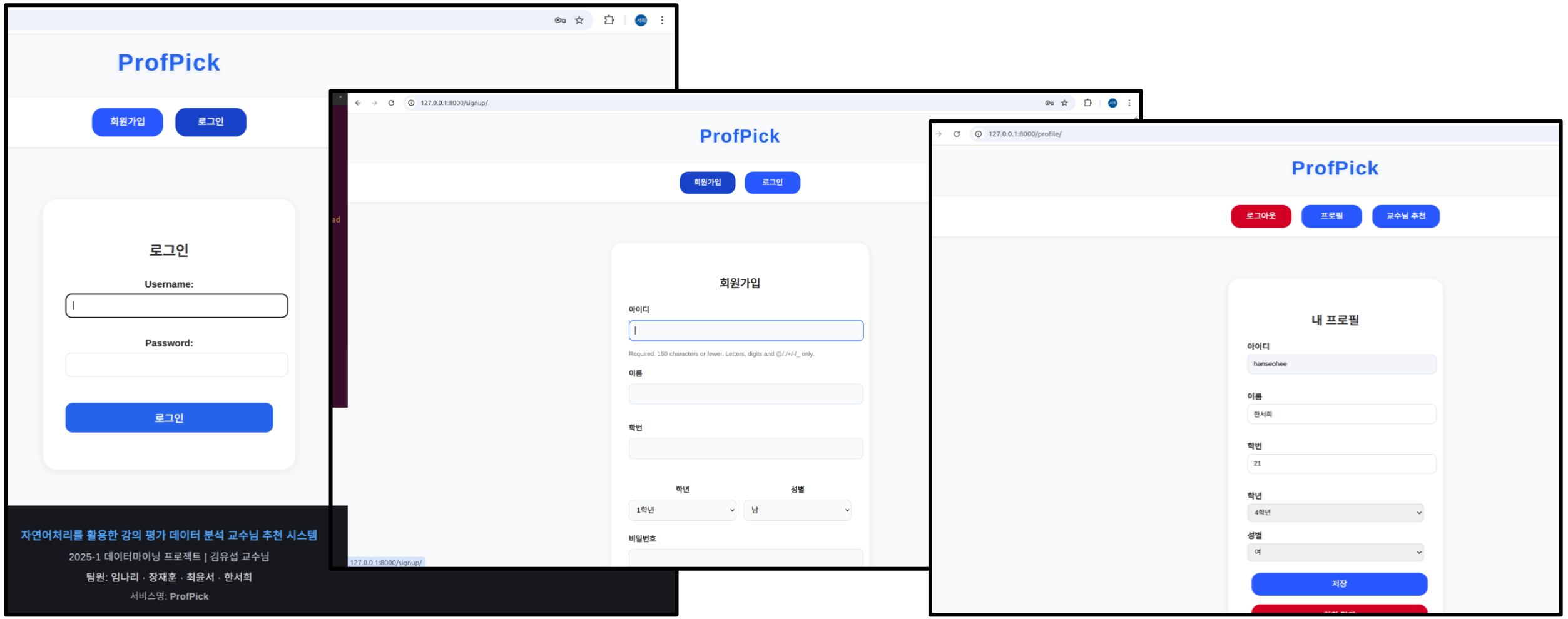
In [ ]:
```
df.to_csv('merged_reviews_by_professor.csv', index=False, encoding='utf-8-sig')
```



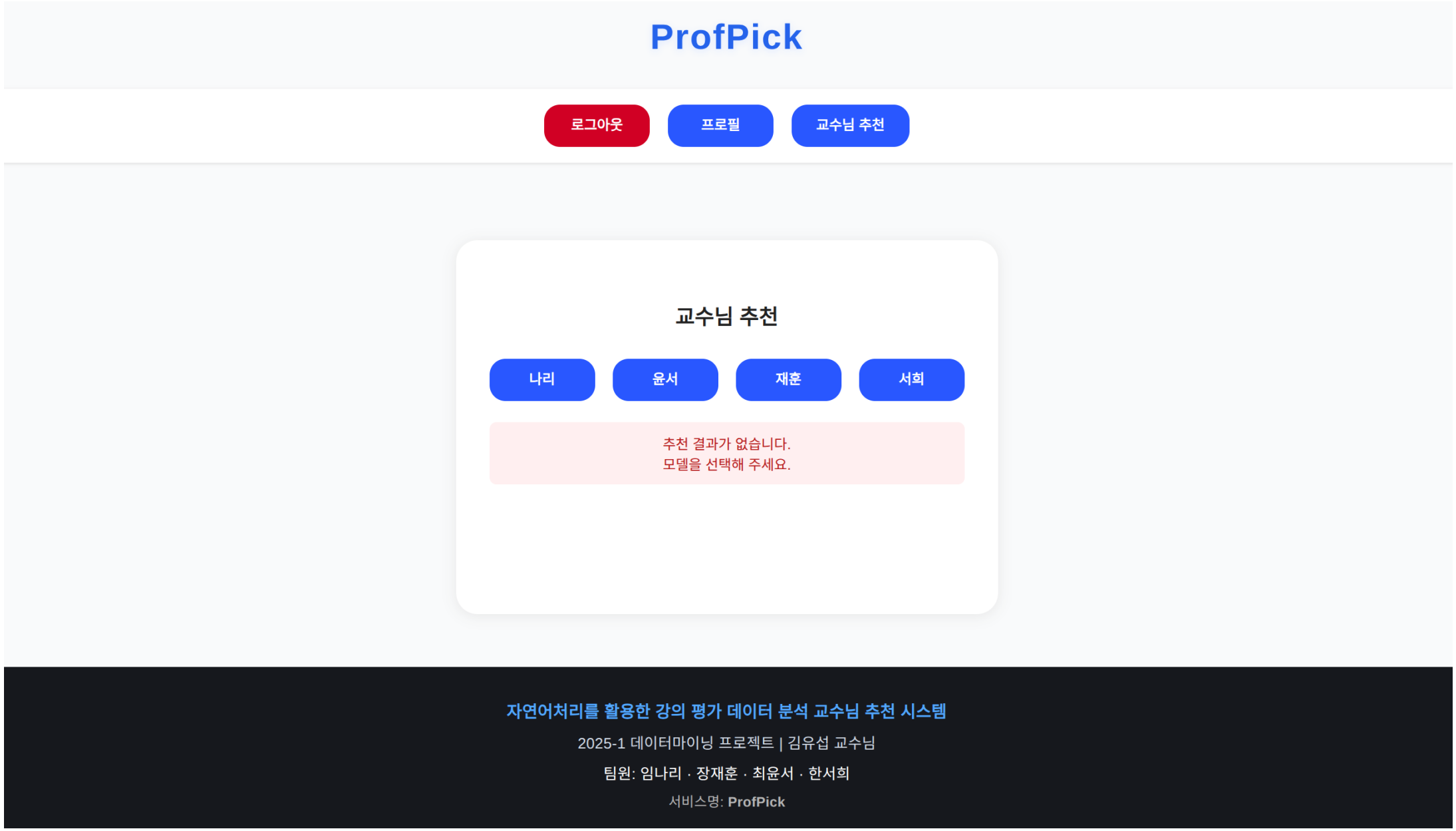
교수님 추천 서비스 기본 구현

환경 정보

> 이 프로젝트는 Ubuntu 24.04.2 LTS (코드네임: noble) 기반 리눅스 환경에서 개발 및 테스트되었습니다.
> Windows, MacOS에서도 실행 가능하지만, 일부 명령어는 다를 수 있으니 참고해 주세요.

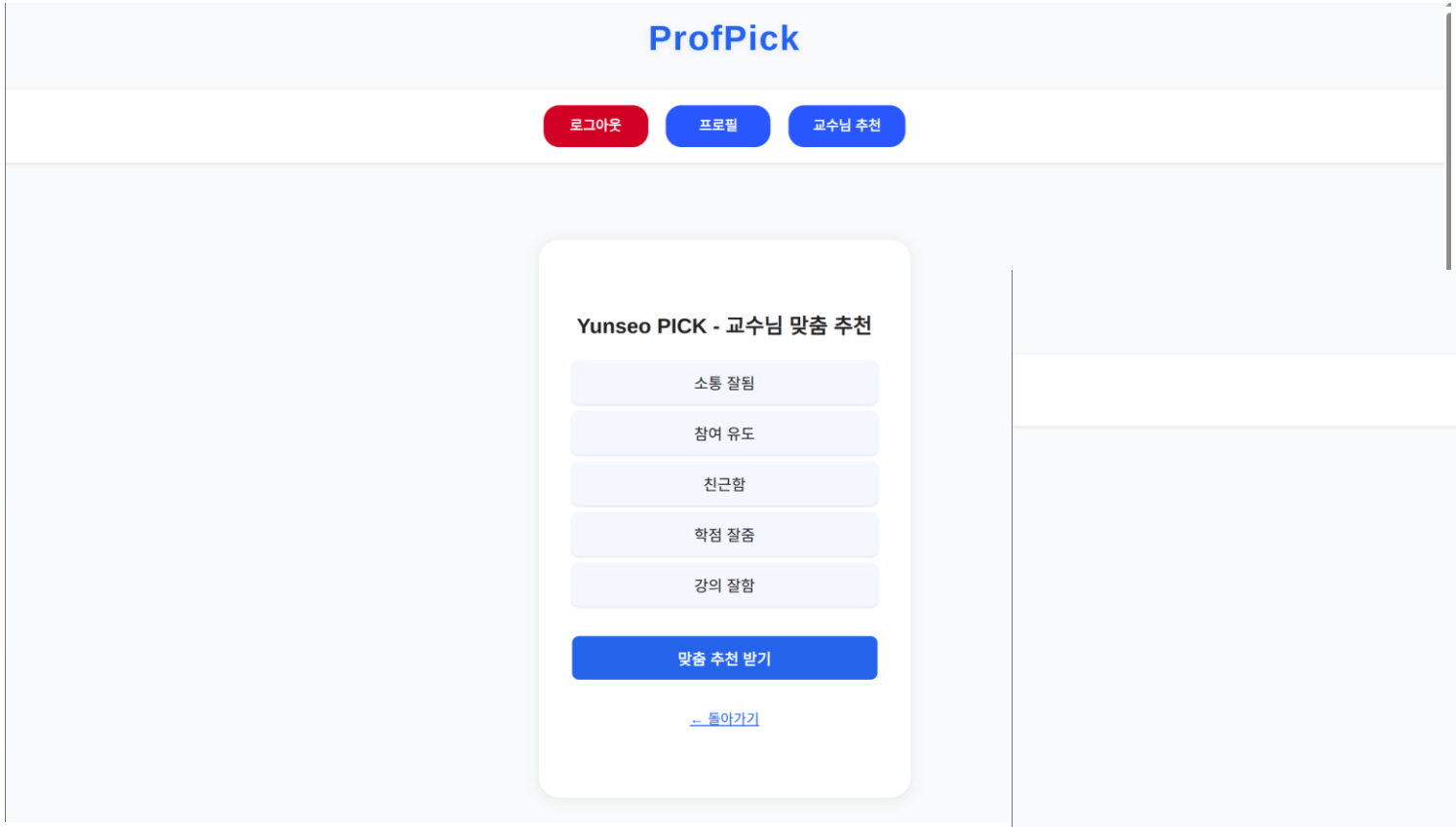


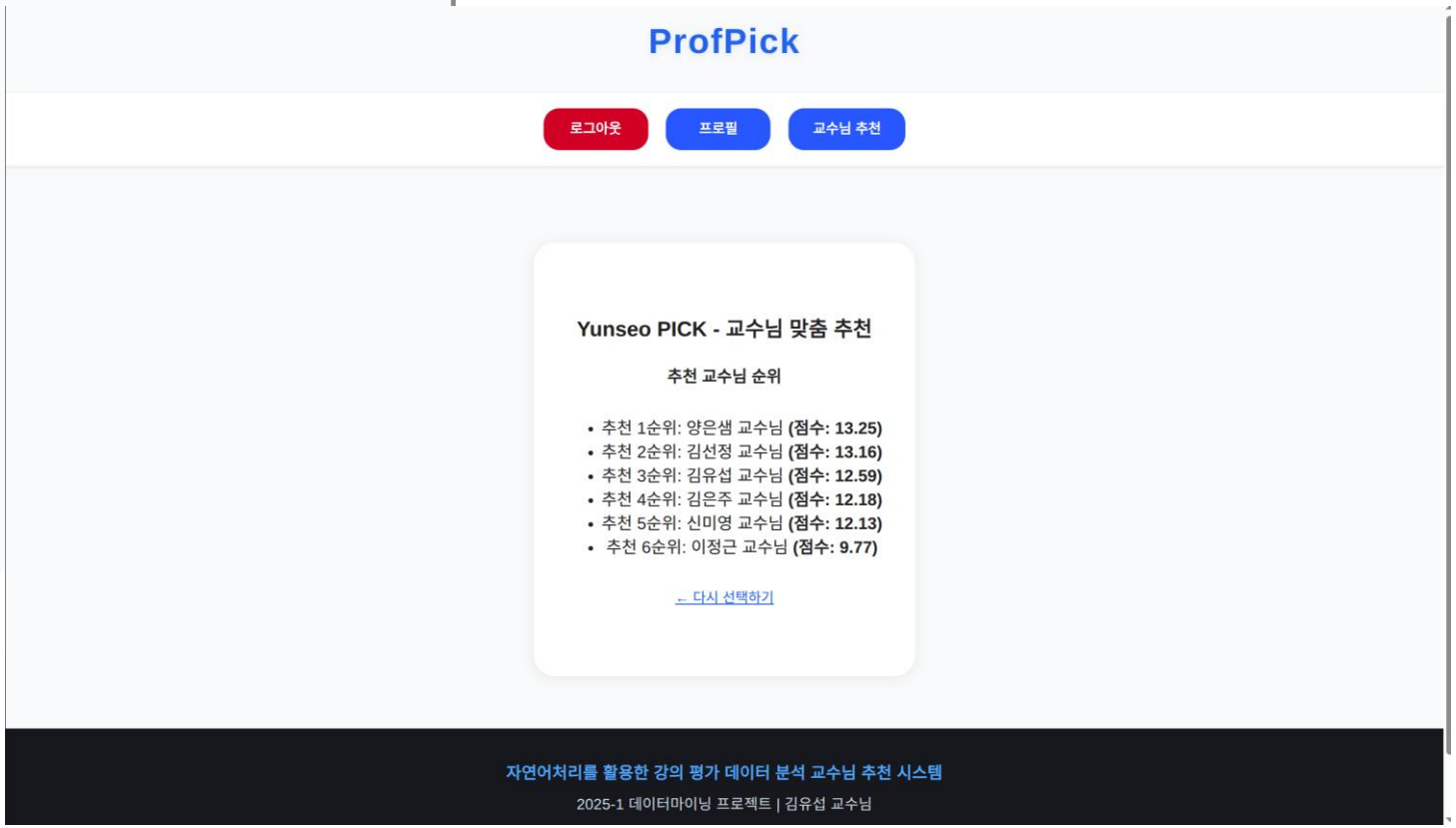
교수님 추천 서비스 기본 구현



교수님 추천 결과 및 설명

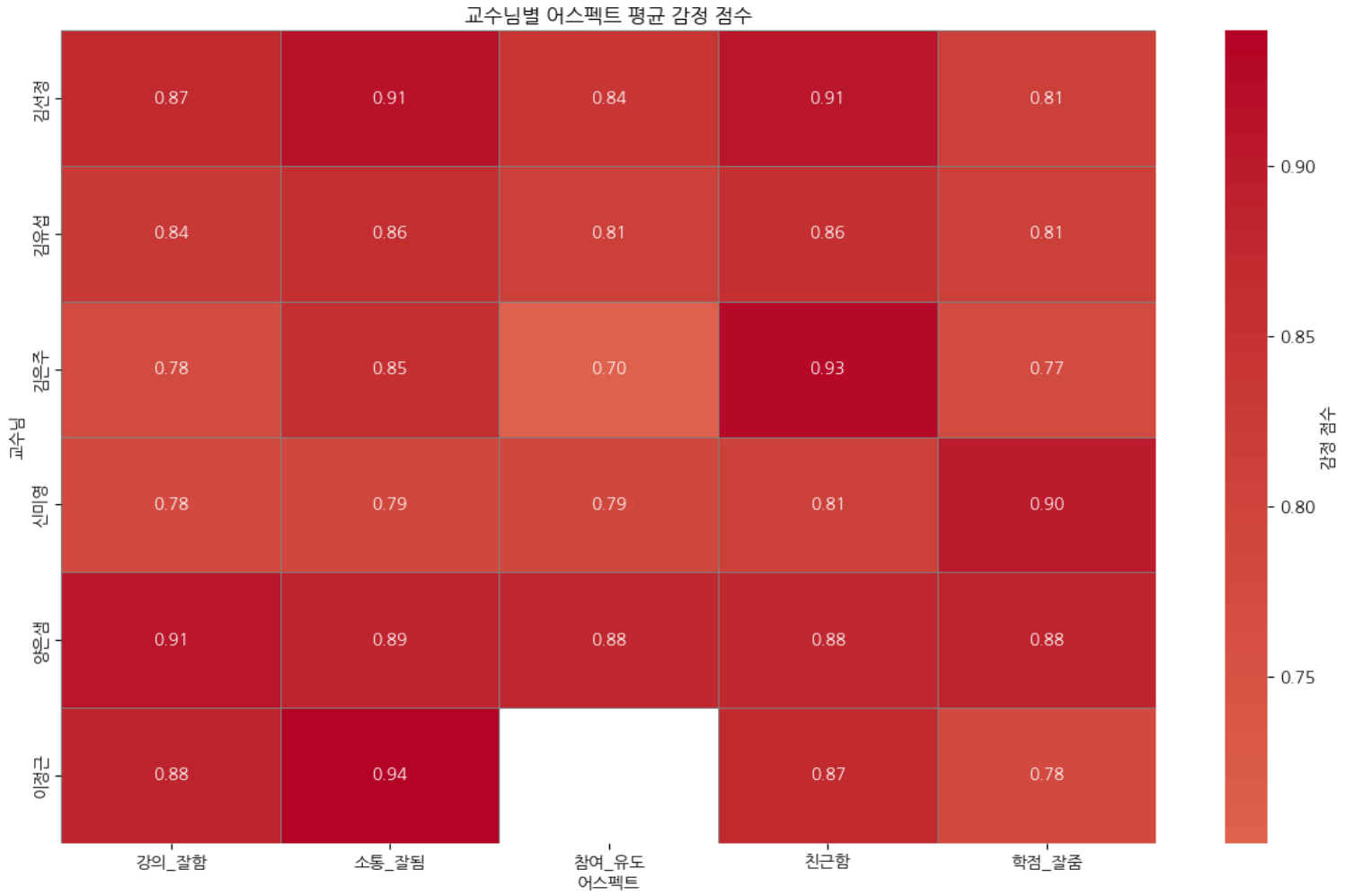
윤서 Pick　윤서.ipynb





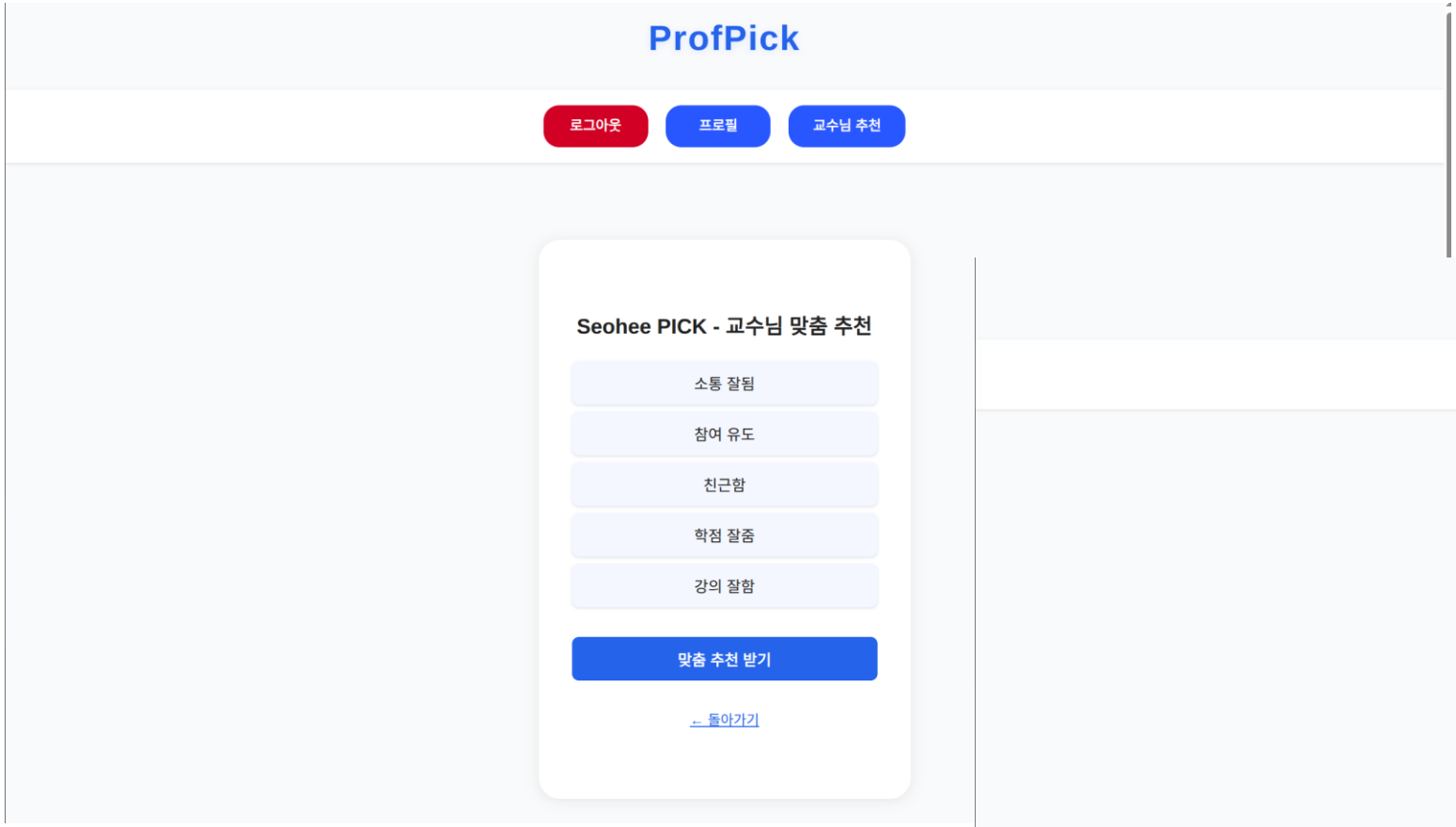
교수님 추천 결과 및 설명

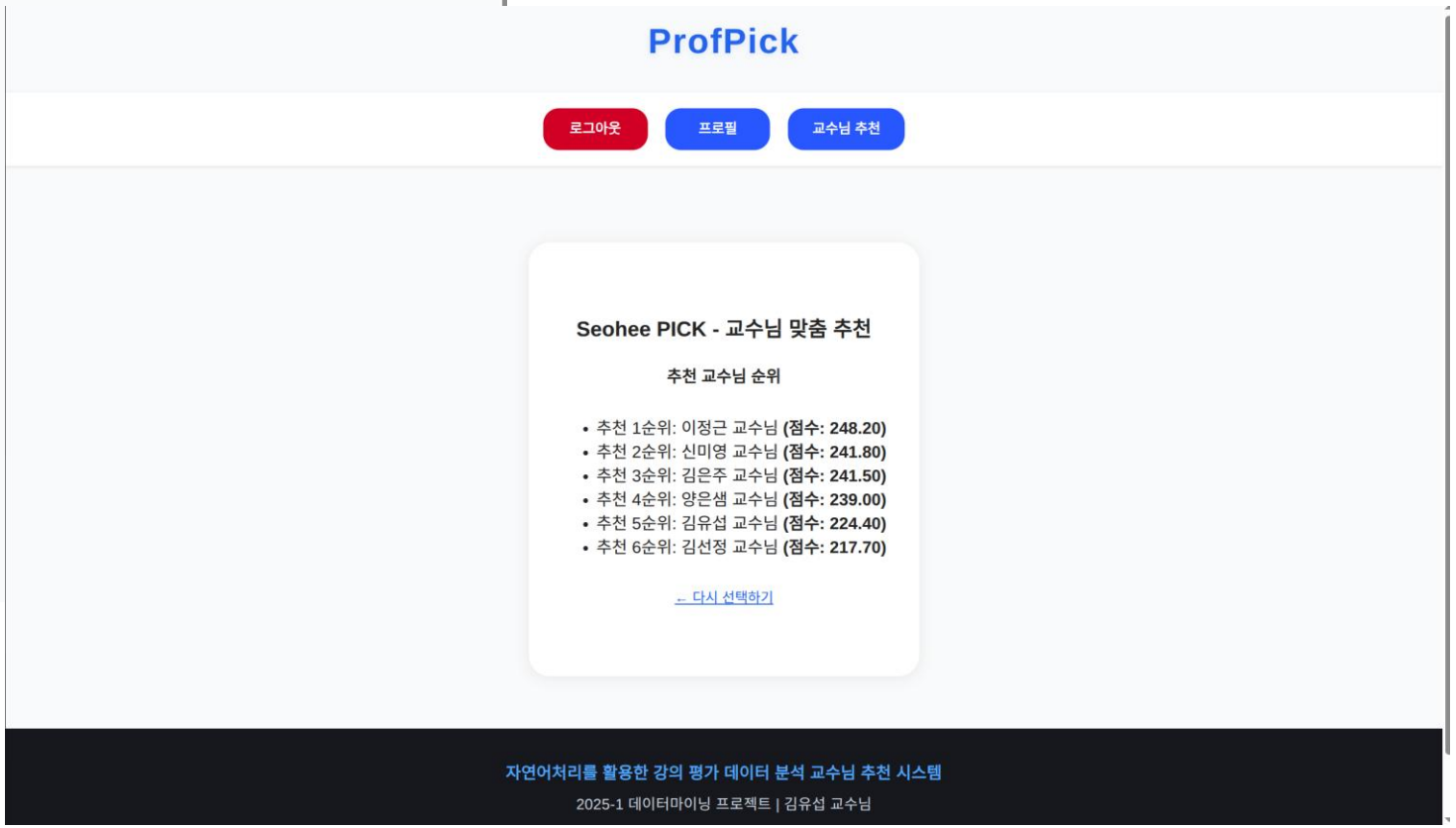
윤서 Pick



교수님 추천 결과 및 설명

서희 Pick　서희.ipynb

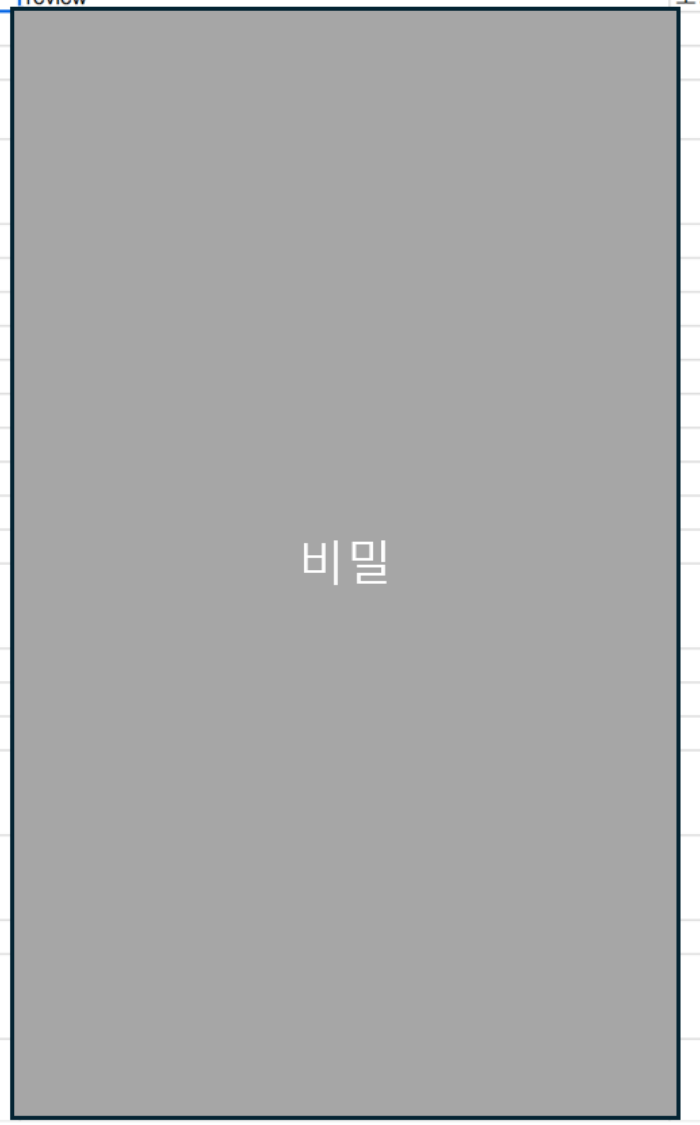




교수님 추천 결과 및 설명

서희 Pick

	A	B	C	D	E	F	G	H
1	professor	review	소통_잘됨	참여_유도	학점_잘줌	친근함	강의_잘함	
2	김유섭		0	1	0	1	0	
3	김유섭		0	0	0	1	0	
4	김유섭		0	0	0	1	2	
5	김유섭		0	2	0	1	0	
6	김유섭		0	1	0	0	1	
7	김유섭		0	0	0	1	0	
8	김유섭		0	3	0	0	0	
9	김유섭		0	0	0	0	1	
10	김유섭		0	0	0	1	0	
11	김유섭		2	2	0	1	0	
12	김유섭		0	0	0	0	0	
13	김유섭		0	0	0	0	0	
14	김유섭		0	0	0	0	0	
15	김유섭		0	0	0	1	0	
16	김유섭		0	0	0	1	0	
17	김유섭		0	0	0	0	0	
18	김유섭		0	0	0	0	0	
19	김유섭		0	0	0	0	0	
20	김유섭		0	1	0	1	1	
21	김유섭		0	0	0	0	0	
22	김유섭		0	1	0	0	1	
23	김유섭		0	0	0	0	0	
24	김유섭		0	0	0	0	0	



교수님 추천 결과 및 설명

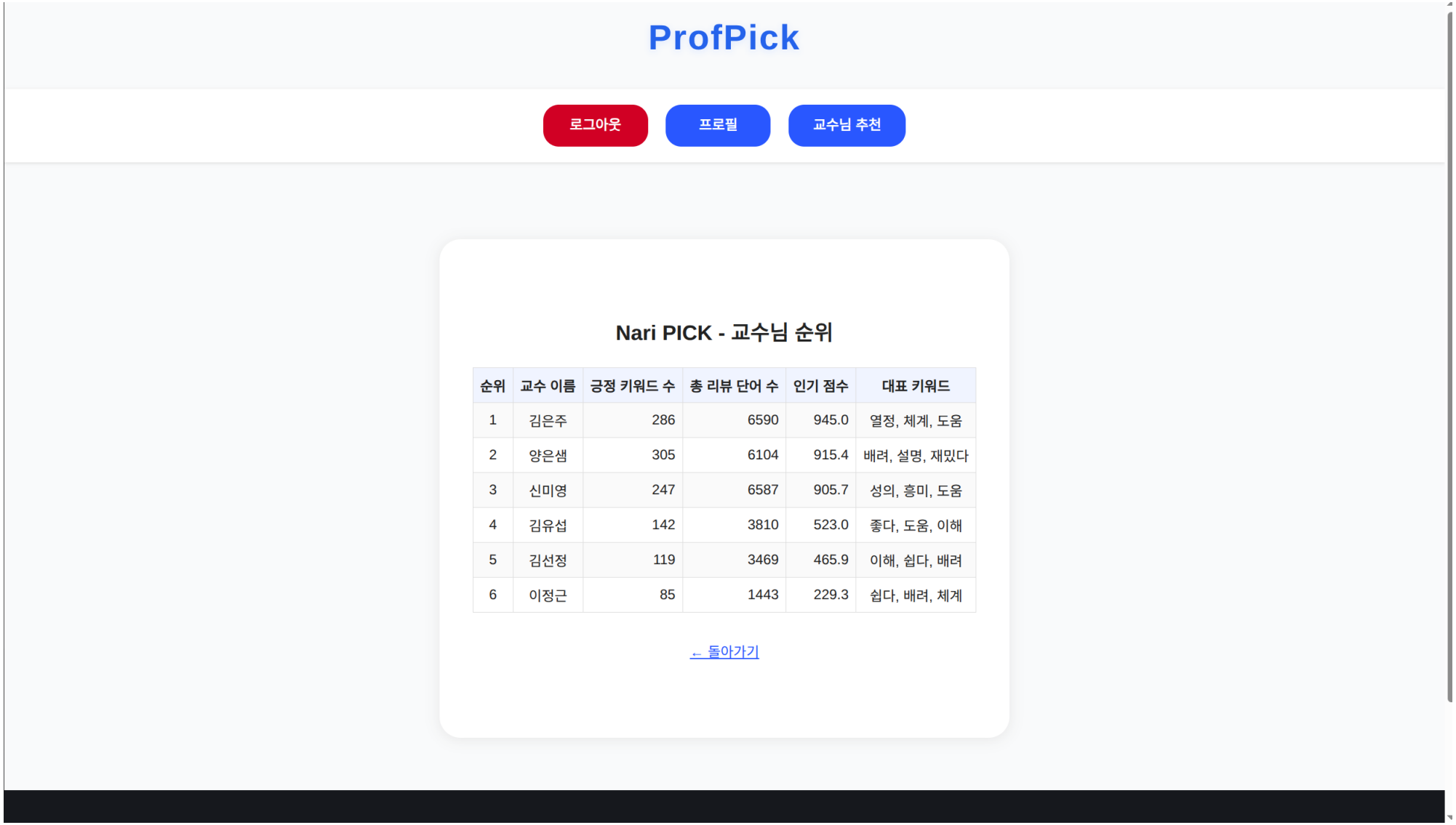
서희 Pick

	A	B	C	D	E	F
1	professor	소통 잘됨	참여 유도	학점 잘줌	친근함	강의 잘함
2	김유섭	15	25	54	27	74
3	양은샘	56	43	56	36	155
4	신미영	41	43	51	47	127
5	김선정	13	14	29	19	62
6	이정근	16	15	15	27	47
7	김은주	24	59	50	28	107

	professor	소통 잘됨	참여 유도	학점 잘줌	친근함	강의 잘함
0	김유섭	7.7	12.8	27.7	13.8	37.9
1	양은샘	16.2	12.4	16.2	10.4	44.8
2	신미영	13.3	13.9	16.5	15.2	41.1
3	김선정	9.5	10.2	21.2	13.9	45.3
4	이정근	13.3	12.5	12.5	22.5	39.2
5	김은주	9.0	22.0	18.7	10.4	39.9

교수님 추천 결과 및 설명

나리 Pick 나리.ipynb

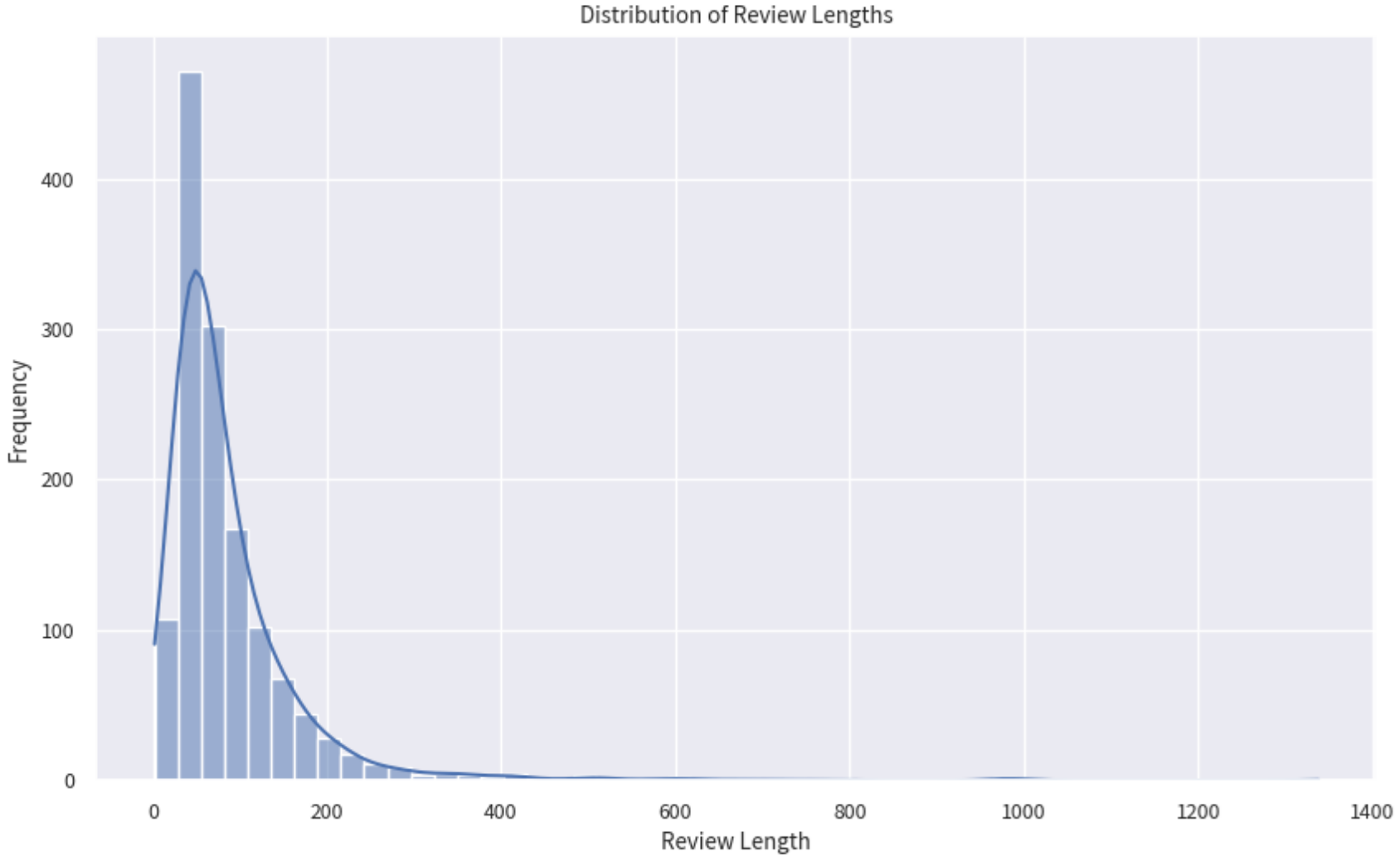


순위	교수 이름	긍정 키워드 수	총 리뷰 단어 수	인기 점수	대표 키워드
1	김은주	286	6590	945.0	열정, 체계, 도움
2	양은샘	305	6104	915.4	배려, 설명, 재밌다
3	신미영	247	6587	905.7	성의, 흥미, 도움
4	김유섭	142	3810	523.0	좋다, 도움, 이해
5	김선정	119	3469	465.9	이해, 쉽다, 배려
6	이정근	85	1443	229.3	쉽다, 배려, 체계

교수님 추천 결과 및 설명

재훈 Pick　재훈.ipynb

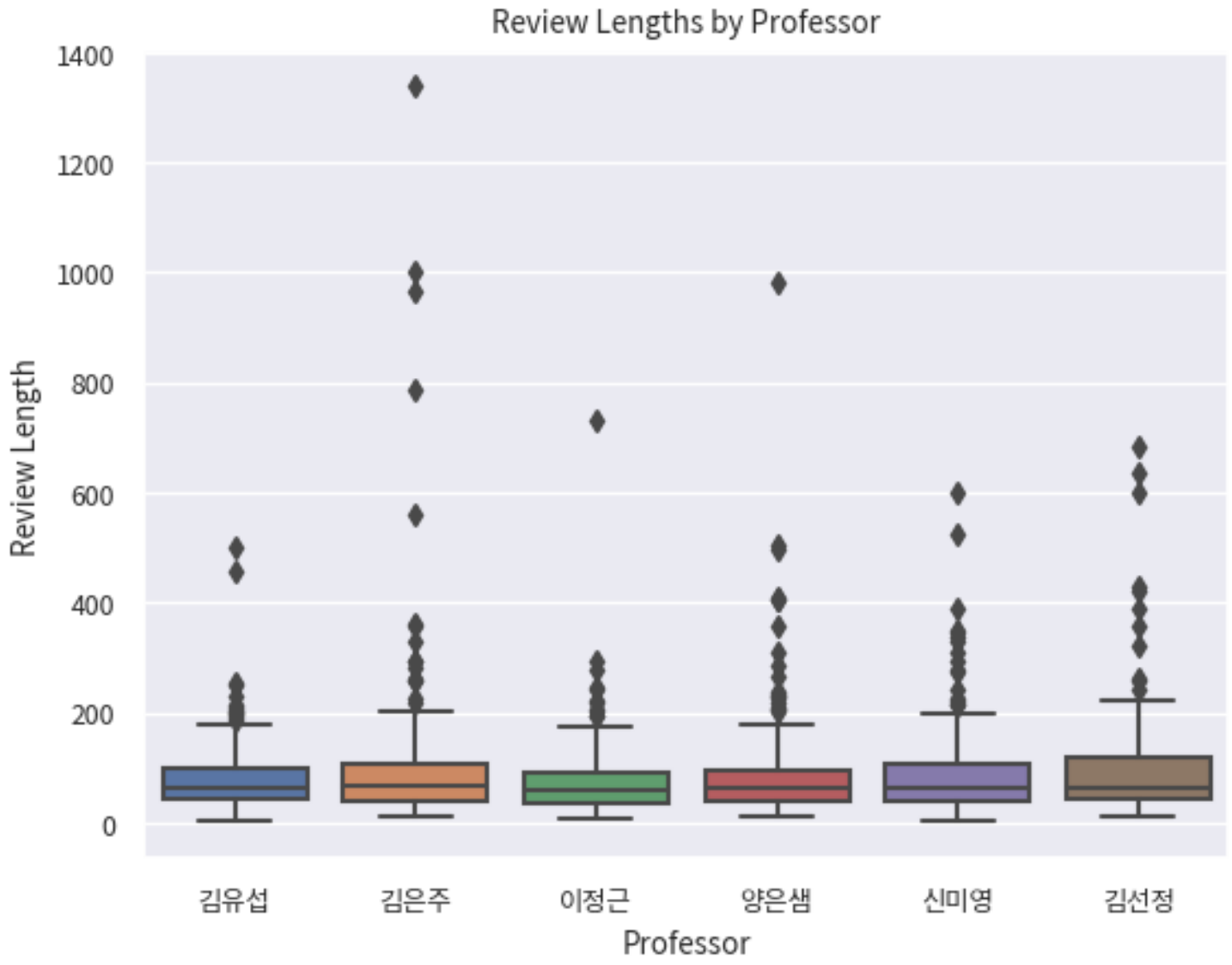
EDA



교수님 추천 결과 및 설명

재훈 Pick

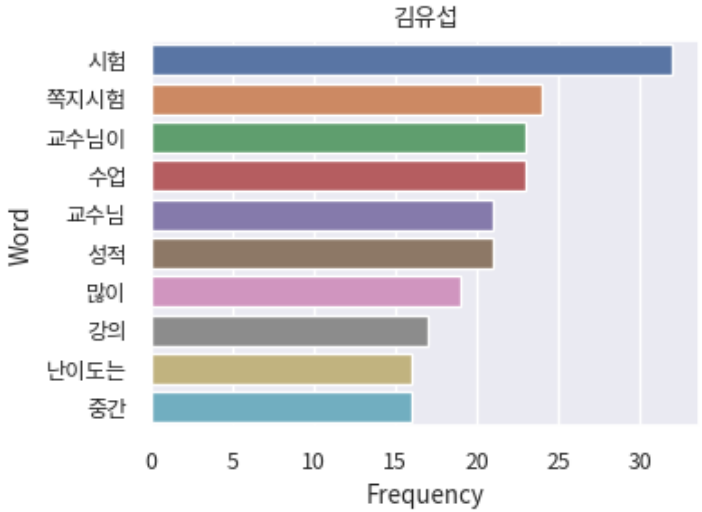
EDA

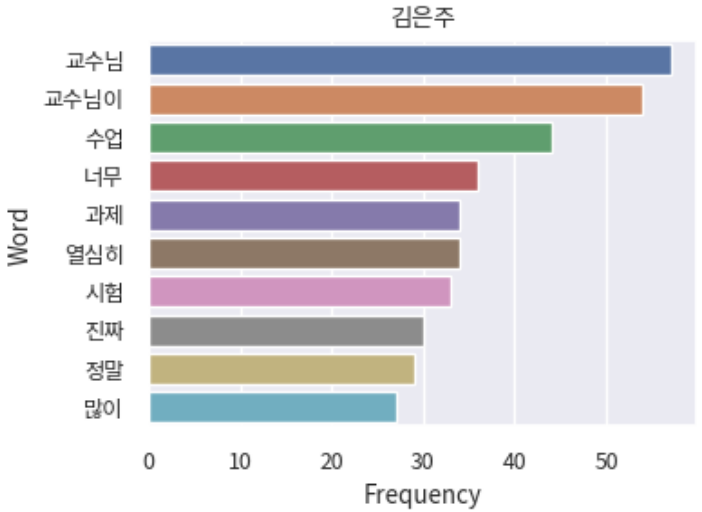


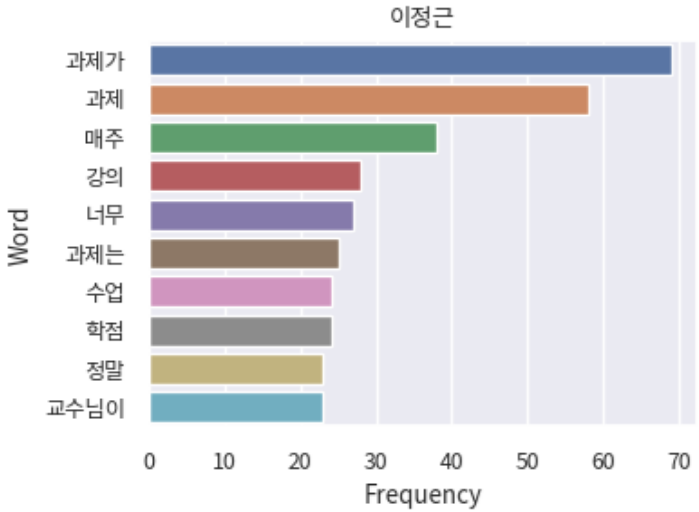
교수님 추천 결과 및 설명

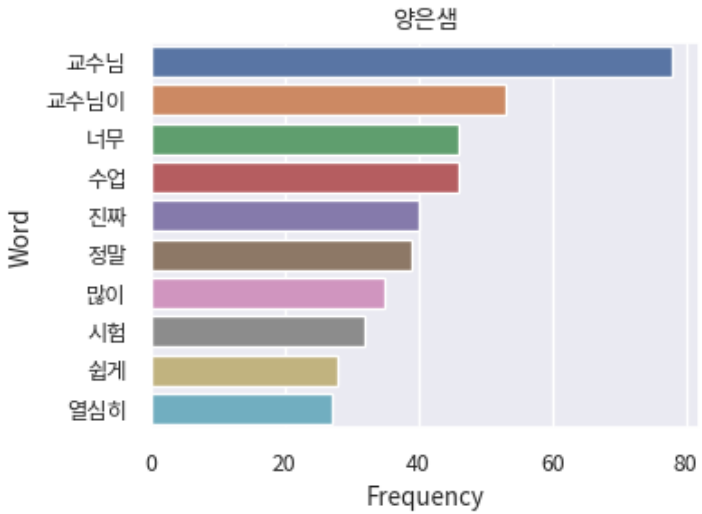
재훈 Pick

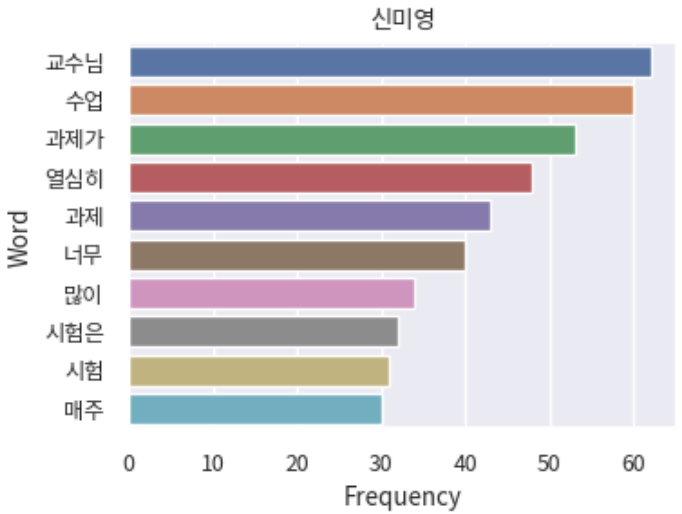
EDA

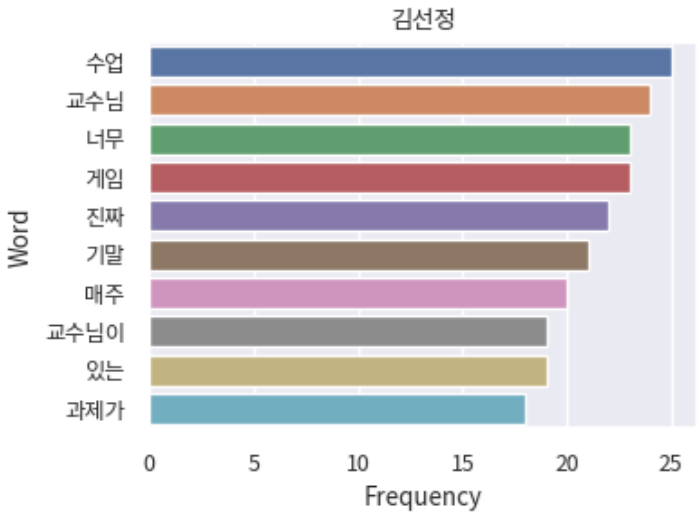












교수님 추천 결과 및 설명

재훈 Pick

EDA

김유섭

김은주

이정근

양은샘

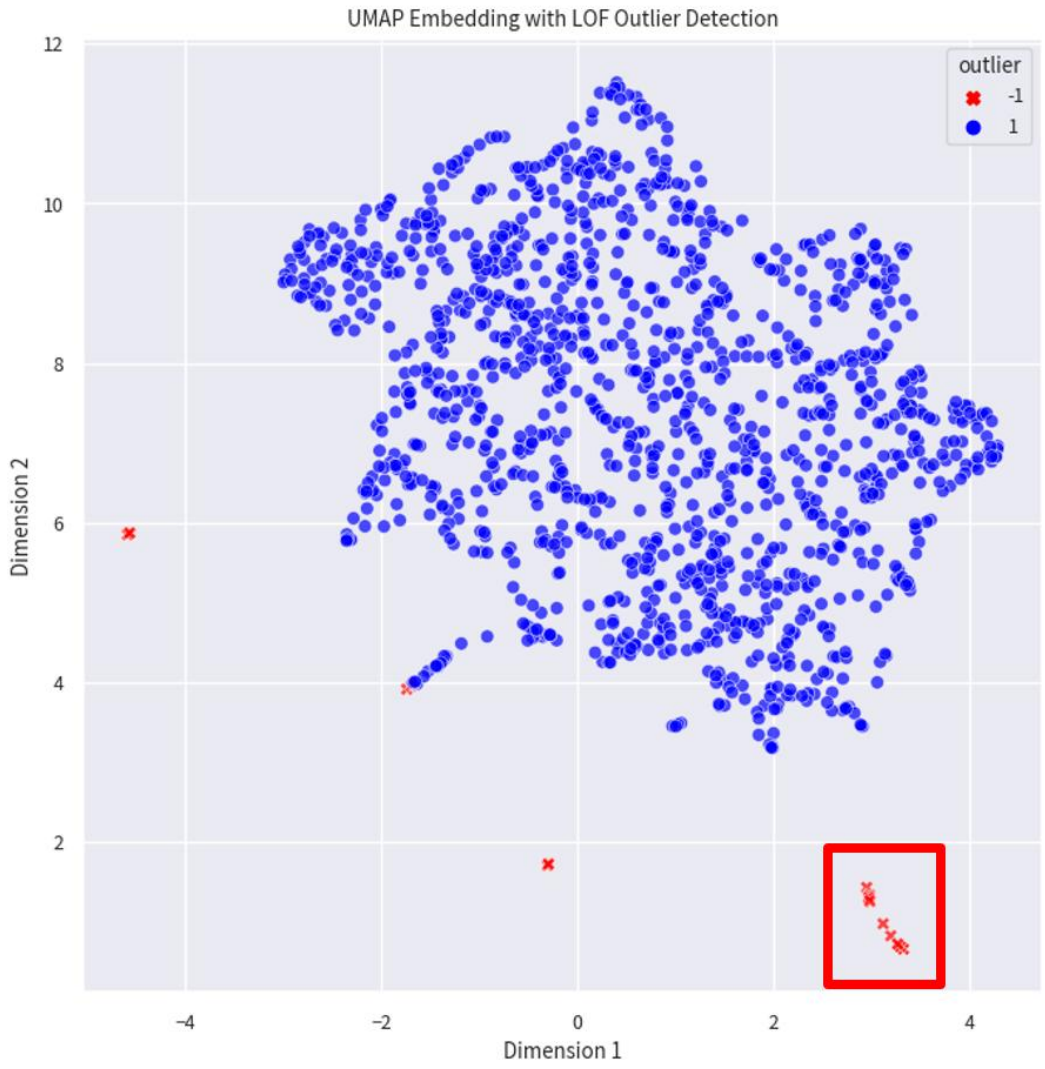
신미영

김선정

교수님 추천 결과 및 설명

재훈 Pick

EDA



친근하다, 몇시간이면 한다, 이정도면 괜찮은 것 같다

친근하다,적극적이다,좋다

설명을 잘해주신다. 수업도 열심히 진행하신다. 적극적이시다.

친근하다,적극적이다,좋다

적극적이다

친근하다
적극적이다
수업 설명을 잘해주신다

친근하다,적극적이다,좋다

친근하다. 듣고만 있어도 쏙쏙 들어온다. 과제는 몇 시간이면 한다.

친근하다, 몇시간이면 한다, 설명을 진짜 잘해주신다

친근하다,적극적이다,좋다

친근하다
적극적이다
수업 설명을 잘하주신다

친근하다,적극적이다,좋다

친근하다, 설명을 진짜 잘해주신다, 적극적이다

교수님 추천 결과 및 설명

재훈 Pick

모델링

SVM Test Accuracy: 0.4615

	precision	recall	f1-score	support
김선정	0.57	0.19	0.29	21
김유섭	0.53	0.56	0.55	32
김은주	0.34	0.23	0.27	53
신미영	0.49	0.61	0.54	57
양은샘	0.39	0.46	0.42	63
이정근	0.55	0.60	0.57	47
accuracy			0.46	273
macro avg	0.48	0.44	0.44	273
weighted avg	0.46	0.46	0.45	273

Random Forest Test Accuracy: 0.4066

	precision	recall	f1-score	support
김선정	0.67	0.19	0.30	21
김유섭	0.46	0.38	0.41	32
김은주	0.27	0.15	0.19	53
신미영	0.38	0.61	0.47	57
양은샘	0.41	0.48	0.44	63
이정근	0.49	0.47	0.48	47
accuracy			0.41	273
macro avg	0.45	0.38	0.38	273
weighted avg	0.41	0.41	0.39	273

감사합니다

HALLYM
UNIVERSITY