

데이터베이스 최종 프로젝트

20011815 데이터사이언스 학과 임홍철

세종대학교 및 HiGrad 대상
크롤링 데이터 수집·가공 및
결과 출력

목차

- 1.프로젝트 시행 사유
- 2.프로젝트 시행 경과
- 3.프로젝트 목적
- 4.프로젝트 진행 방법
- 4-1.웹크롤링
- 4-2.데이터 가공
- 4-3.데이터 입력
- 4-4.데이터 출력
- 5.데이터 형식
- 6.프로그램 가동 방법
- 7.아쉬운점
- 8.프로그램 실행 사진

1.프로젝트 시행 사유

최근 3학년 2학기를 지나며 대학원 진학에 관심이 생기기 시작하였다. 그러면서 교수님들의 연구실 홈페이지에 들어가기 시작했는데, 교수님들마다 다른 형태의 홈페이지를 가지고 있었다. 그리고 우리 학교의 경우 연구실 홈페이지 접근성 또한 떨어졌다. 이런 이유로 교수님 연구실 홈페이지들의 각기 다른 형태를 정리하여 낮은 접근성 문제를 해결하고자 이 프로젝트를 시작하였다.

2.프로젝트 시행 경과


우선 우리 학교 인공지능 및 데이터사이언스 학과만으로 프로젝트를 진행을 하였는데 데이터의 수(14개)가 너무 적어 교수님께 충고를 얻어 higrad 홈페이지에서 인공지능 태그를 달고 있는 연구실들을 추가적으로 크롤링을 진행하였다.

3.프로젝트 목적

교수님들마다 다른 연구실 홈페이지 정보들을 웹 크롤링 기술을 활용하여 데이터를 수집하고, 이를 llm을 통하여 비정형데이터를 정형데이터로 가공하여, mysql에 연동하여 파이썬으로 정리된 형태로 보여준다.

4-1.프로젝트 진행 방법 - 웹 크롤링

최근에 유행중인 playwright를 사용하여 프로젝트를 진행하였다. visual code insiders에서 제공해주는 playwright를 사용하여 크롤링을 진행하였다. node.js를 활용하여 코딩을 진행하였으며 세종대연구실crawling.js, higrad인공지능crawling.js 두 개의 파일을 이용하여 크롤링을 진행하였다. 세종대연구실 코드의 경우

'<https://home.sejong.ac.kr/~aidsdpt/12.html>'의 링크에서 html을 분석해본 결과 이런 이모지가 있는 곳에 연구실 하이퍼링크가 있다는 사실을 알게 되었고, 해당 사실을 반영하여 우선 링크를 수집하였다. 앞서 수집한 링크를 바탕으로 접속하여 html에서 navigation에 속하는 메뉴를 찾고 모든 navigation에 접속하며 해당 페이지에 있는 텍스트들을 복사해오는 형식으로 설계하였다. 여기서 텍스트를 복사할 때 1000자로 제한을 두었는데 이는 이후 llm모델을 사용할 때 가공하는데 너무 오랜 시간이 걸리고, 길 경우 로컬 모델이라 중간 내용을 망각하는 현상이 생겨 일부러 제한을 두어 크롤링을 진행하였다. higrad 코드의 경우 '<https://higrad.net/laboratory/organizations?page=6&pagingno=1&keyword=%23%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5&pagesize=10&sortType=FREQ&limit=25&displayType=QNA&siteid=3>' 여기에 해당하는 링크가 인공지능 태그가 된 연구실 모음이라는 사실을 알았다. 그리고 여기서 페이지마다 li.laboratory-list-wrap 이런 html태그가 달린 것이 그 교수님에 관한 higrad 설명 링크인 사실을 알았고, 이 링크를 타고 들어갔을 때 a.button.homepage가 활성화 되어있으면 해당하는 교수님의 개인 연구실 링크로 연결된다는 사실도 알았다. 그래서 이 모든 조건들이 성립했을 경우 위와 같이 navigation을 찾아 모든 navigation링크를 순회하며 데이터를 1000자 제한으로 복사해왔다. 마지막으로 모든 연구실을 한번에 보여주기 어려워 페이지 형식으로 higrad에서 제공하고 있어, 임의로 페이지에 해

당하는 링크만 변경하며 총 6페이지까지 코드를 돌려 크롤링을 진행하였다.

위에서 크롤링 할 때 원래 각 홈페이지에 올려져 있는 이미지 링크도 같이 가져와 가공하고 싶었으나, 이미지 링크의 경우 잘 수집되지 않아 해당하는 부분은 버리기로 했다.

이렇게 크롤링을 진행하여 총 7개(세종대1개+hi grad6개)의 xlsx 엑셀 파일이 나오게 되었고, 132개의 연구실 홈페이지와 1424개의 연구실 홈페이지의 페이지들에 대해 조사를 진행하게 되었다.

4-2.프로젝트 진행 방법 - 데이터 가공

세종대연구실crawling.js 파일을 통해 나온 결과물은 우선 한글이 깨진 형태로 나왔다. 그리고 줄바꿈, 공백등 여러 내용이 깨져 있어 이를 보기 쉽게 data_refine.ipynb를 통하여 가공하였다. 우선 한글이 깨지는 문제는 utf-8로 읽어 들여 해결 하였고 저장할때도 utf-8-sig를 통하여 깨지지 않게 만들었다. 또한 re 라이브러리를 사용하여 직접적으로 이상한 줄바꿈이나 필요없는 부분을 제거하였다. 이후 행한줄로 이어진 파일을 코드를 통해 줄을 바꿔 내가 원하는 형태로 변경하였다. 이 후 데이터들을 보니 연구실마다 묶여 있기는한데 교수님들마다 연구원들을 부르는 방식이 매우 다른 문제가 발생하였다. 어떤 교수님은 people로 연구원을 표현하기도 했지만 researcher, member, members등 같은 것을 지칭하는데 다르게 표현하는 문제가 있었다. 이를 해결하기 위하여 dlama를 사용하여 최신 llm 모델인 llm3: 12b을 이용하여 각 페이지에 해당하는 내용과 그 페이지를 교수님이 어떻게 지정했는지 타이틀 정보 두 개를 주고 categories = ['professor', 'members', 'research', 'publications', 'lectures', 'projects', 'events', 'recruit', 'others'] 다음의 카테고리 안에서 가장 적합한 것으로 선택하게 만들었다. 이후 이를 기반으로 다시 한번 llm에 위에서 그룹 지어진 categories와 페이지 내용을 아래와 같은 프롬프트에 같이 넣어 반정형 형태로 가공하였다.

1차변환 - 페이지내열정리 열 생성

```
instructions = {
    'professor': "이 교수님의 대표 논문과 주요 연구 내용을 한 문장으로 요약해 주세요.",
    'members': (
        "연구실 구성원(학생) 정보를 아래 항목을 기준으로 정리해 주세요.\n"
        "학생 이름 / 근속 년수(있으면) / 연구분야(있으면) / 연구주제(있으면) / 기타사항(논문, 수상 등)\n"
        "가능하면 리스트 형태로 주세요."
    ),
    'research': "이 연구실이 수행 중인 연구 주제를 한 문장으로 요약해 주세요.",
    'publications': "연구실의 주요 논문 또는 출판물을 리스트 형식으로 정리해 주세요.",
    'lectures': "교수님이 가르친 과목들을 리스트 형식으로 정리해 주세요.",
    'projects': (
        "연구실에서 수행한 프로젝트를 정리해 주세요. 각 항목마다 아래 형식으로 작성해 주세요:\n"
```

```

"프로젝트명: 설명 / 어떤 일을 수행했는지\n"
"각 프로젝트는 줄바꿈으로 구분해 주세요."
),
'recruit': "연구실에서 모집하는 학생의 조건이나 방향을 한 문장으로 정리해 주
세요.",
'events': "행사의 종류와 주요 내용(언제, 무슨 행사)을 간단히 정리해 주세요."
}

```

2차변환 - 2차정제 열 생성

위에서 만들어진 반정형 데이터를 바로 mysql에서 이용할 수 없어 gemma3:12b에 위에서 반정형 형태로 가공된 내용과 카테고리를 아래와 같이 형식을 지정해주는 프롬프트랑 같이 넣어 정형 데이터로 만들었다. 변환 결과 몇 개의 특이 케이스가 있지만 이는 data_refie.ipynb에 있는 'I'을 기반으로 데이터를 찾는 코드를 사용하면 쉽게 처리가 가능하다.

```

PROMPT_TEMPLATES = {
    "members": """다음은 "members" 섹션에 있는 텍스트야. 이 안에서 학생 이름, 근속
    년수, 연구분야, 연구주제, 기타사항을 정리해서 아래와 같은 표로 변환해줘:

    형식:
    | 학생 이름 | 근속 년수 | 연구분야 | 연구주제 | 기타사항 |
    |-----|-----|-----|-----|-----|

    출력은 오직 위 표 형식으로만 해줘. 다른 설명은 하지 마. 텍스트는 다음과 같아:

    ---
    {content}
    """,
    "publications": """다음은 "publications" 섹션의 텍스트야. 아래 형식으로 논문 리스
    트를 표로 변환해줘:

    형식:
    | 논문 이름 | 저널 | 저자 | 연도 |
    |-----|-----|-----|-----|

    출력은 위 표로만 해줘. 불필요한 설명 없이.

    텍스트:
    ---
    {content}
    """,
    "lectures": """다음은 "lectures" 섹션 텍스트야. 여기에 언급된 강의 과목명을 리스트

```

형태로 변환해줘:

형식:

["과목1", "과목2", "과목3", ...]

텍스트:

{content}

"""

"projects": """다음은 "projects" 섹션이야. 아래와 같은 테이블 형식으로 변환해줘:

| 프로젝트명 | 프로젝트 설명 |

|-----|-----|

출력은 테이블 형태로만 해줘. 텍스트:

{content}

"""

}

3차변환 - 3차정제(교수님 열 생성)

1차 정제에서 정제된 데이터를 보니 교수님관련 데이터는 너무 적어 정리 할 수 없다는 output이 많이 나온 사실을 확인하였다. 따라서 3차로 아래 프롬프트를 넣어 추가적인 분류를 진행하였다. 이후 3차정제에서 나온 열을 2차정제에 합치는 과정을 진행하였다.

PROFESSOR_PROMPT = """아래 텍스트에서 교수 정보를 표 형식으로 추출해줘:

[교수이름,이메일주소,소속대학교,교수전화번호,연구실위치,기타사항]

형식:

| 교수이름 | 이메일주소 | 소속대학교 | 교수전화번호 | 연구실위치 | 기타사항 |

|-----|-----|-----|-----|

출력은 위 표로만 해줘. 불필요한 설명 없이. 교수이름은 영어로 되어있으면 영어로 출력해줘. 이메일주소는 이메일 형식으로, 전화번호는 숫자만 포함된 형식으로 출력해줘.

텍스트:

{content}

"""

4-3.프로젝트 진행 방법 - 데이터 입력

attach_mysql.ipynb파일을 이용하여 위에서 가공한 연구실_8개_테이블_정리.xlsx 파일을 읽어 mysql에 집어넣었다.

4-4.프로젝트 진행 방법 - 데이터 출력

교수님께서 수업 시간에 제공해주신 회원 관리 시스템 통합 코드를 최대한 활용하여 조회 프로그램을 만들어보고자 한다. 프로그램은 파이썬을 이용하여 작성되었으며 sql을 사용하기위해 pymysql 그리고 데이터를 가시성있게 출력하기 위해 tabulate 라이브러리를 사용하였다. 프로그램은 아래와 같은 다이어그램으로 기능하며 페이지를 통해 가시성을 높이고, 필터를 통해 찾고자 하는 것을 쉽게 찾을 수 있게 하였다. 참고로 page_sizes의 크기를 변경하면 한 페이지에서 보여주는 데이터 개수를 변경할 수 있다.

[프로그램 시작]



[메인 메뉴]

- └─ (1) Raw 데이터 조회
 - └─ [테이블 선택] → [데이터 조회(페이지)]
- └─ (2) 데이터 분석
 - └─ (1) 전체 열 분석
 - └─ (2) 교수별 정보 요약(페이지,필터)
 - └─ (3) 연구실별 내용 요약(필터, 연구실 상세 보기)
 - └─ [연구원/프로젝트/논문 세부 조회]
- └─ (3) 종료

5.데이터 형식

데이터의 가장 뼈대가 되는 ‘연구실_정보_추출결과2.xlsx’ 엑셀파일은 아래와 같은 열을 가진다.

주소	제목	메인주소	타이틀	본문내용	분류	페이지내용정리	2차정제
----	----	------	-----	------	----	---------	------

주소는 각 페이지의 상세 주소이다. 예를들어 메인주소가 www.tedlab.com가 있으면 www.tedlab.com/people 등 상세 주소를 가르킨다.

제목은 각 페이지에서 홈페이지 저자가 설정한 홈페이지 이름이다.

메인주소는 위에서 언급한 주소의 기본뼈대가 되는 주소이다.

타이틀은 html상에서 navigator에 저장되어있는 해당 페이지의 이름이다.

본문내용은 해당 페이지에 있는 내용 1000자만 복사해온 내용이다.

분류는 해당 페이지에 해당하는 본문내용과 타이틀을 gemma3:12b에 넣고 분류시켜 얻은 결과물이다.

페이지내용정리는 본문내용과 분류를 gemma3:12b에 넣고 프롬프트로 반정형 형태로 만든 내용이다.

2차정제는 페이지내용정리를 위에서 형식을 지정해주는 프롬프트와 함께 gemma3:12b에 넣

어 나온 형태이다.

데이터는 8개의 테이블로 나누어진다.

모든 테이블은 정규화를 거쳤고 메인주소와 분류 2개를 합쳐서 primary key로 설정하여 추후 join을 통해 연결할 것이다. 각 연구실에서 크롤링을 해왔기 때문에 메인주소, 분류만 필수적으로 있어야 할 데이터이고, 나머지는 없을 수도 있다. 교수님의 경우 lecture와 professor이 각각 다른 상세링크로 이어지기 때문에 어쩔 수 없이 분리하였다.

1테이블 메인주소, 분류, 타이틀, 주소

2테이블 메인주소, 분류, 연구원 이름, 근속 연수, 연구분야, 연구주제, 연구원기타사항

3테이블 메인주소, 분류, 논문 이름, 저널, 저자, 연도

4테이블 메인주소, 분류, 교수님이 가르치시는 과목들

5테이블 메인주소, 분류, 프로젝트 이름, 프로젝트 설명

6테이블 메인주소, 분류, recruit(모집 정보)

7테이블 메인주소, 분류, 연구주제

8테이블 메인주소, 분류, 교수님 이름, 교수님 설명

mysql를 통해 아래와 같이 table을 생성하였다.

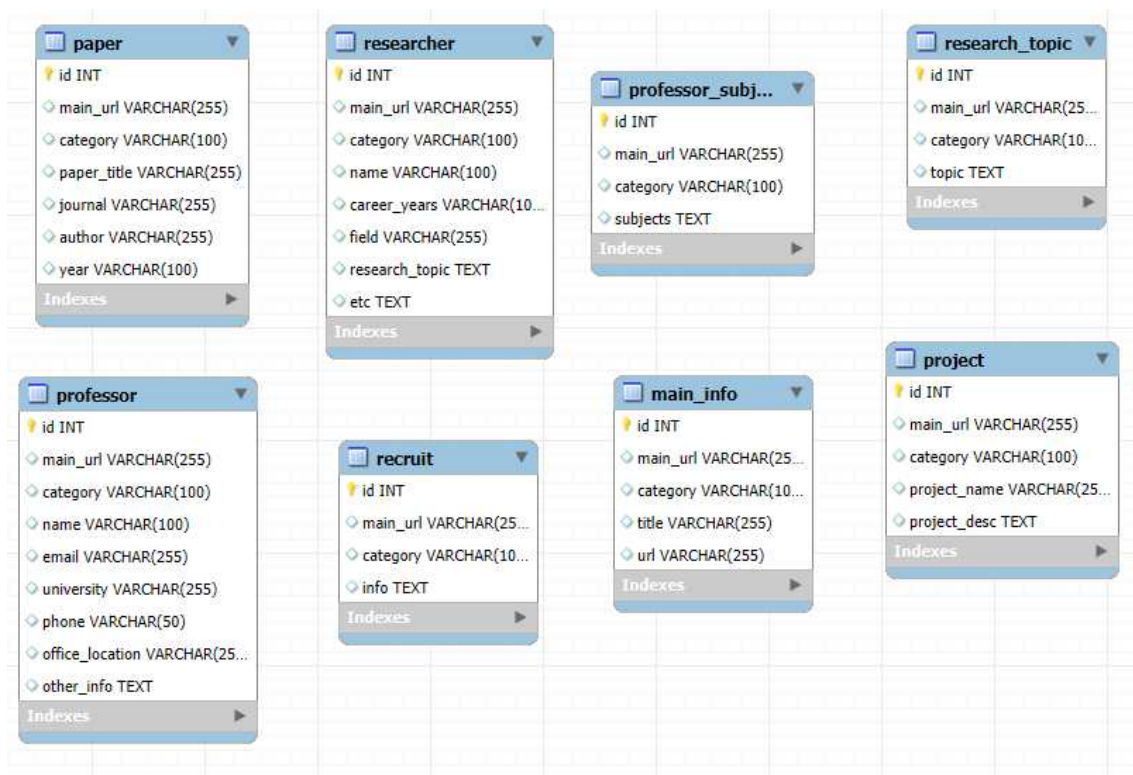
```
CREATE TABLE IF NOT EXISTS main_info (  
  id INT AUTO_INCREMENT PRIMARY KEY,  
  main_url VARCHAR(255),  
  category VARCHAR(100),  
  title VARCHAR(255),  
  url VARCHAR(255)  
);
```

```
CREATE TABLE IF NOT EXISTS researcher (  
  id INT AUTO_INCREMENT PRIMARY KEY,  
  main_url VARCHAR(255),  
  category VARCHAR(100),  
  name VARCHAR(100),  
  career_years VARCHAR(100),  
  field VARCHAR(255),  
  research_topic TEXT,  
  etc TEXT  
);
```

```
CREATE TABLE IF NOT EXISTS paper (  
  id INT AUTO_INCREMENT PRIMARY KEY,  
  main_url VARCHAR(255),  
  category VARCHAR(100),  
  paper_title VARCHAR(255),  
  journal VARCHAR(255),  
  author VARCHAR(255),
```


year VARCHAR(100));
CREATE TABLE IF NOT EXISTS professor_subject (id INT AUTO_INCREMENT PRIMARY KEY, main_url VARCHAR(255), category VARCHAR(100), subjects TEXT);
CREATE TABLE IF NOT EXISTS project (id INT AUTO_INCREMENT PRIMARY KEY, main_url VARCHAR(255), category VARCHAR(100), project_name VARCHAR(255), project_desc TEXT);
CREATE TABLE IF NOT EXISTS recruit (id INT AUTO_INCREMENT PRIMARY KEY, main_url VARCHAR(255), category VARCHAR(100), info TEXT);
CREATE TABLE IF NOT EXISTS research_topic (id INT AUTO_INCREMENT PRIMARY KEY, main_url VARCHAR(255), category VARCHAR(100), topic TEXT);
CREATE TABLE IF NOT EXISTS professor (id INT AUTO_INCREMENT PRIMARY KEY, main_url VARCHAR(255), category VARCHAR(100), name varchar(100), email VARCHAR(255), university VARCHAR(255), phone VARCHAR(50), office_location VARCHAR(255), other_info TEXT);

아래 사진은 이를 토대로 만들어진 다이어그램입니다.



[전체 열 및 데이터 개수 분석]

테이블명	열 목록	데이터 개수
main_info	id, main_url, category, title, url	1415
researcher	id, main_url, category, name, career_years, field, research_topic, etc	721
paper	id, main_url, category, paper_title, journal, author, year	702
professor_subject	id, main_url, category, subjects	38
project	id, main_url, category, project_name, project_desc	158
recruit	id, main_url, category, info	131
research_topic	id, main_url, category, topic	383
professor	id, main_url, category, name, email, university, phone, office_location, other_info	73

위 사진은 university_show_program.ipynb 전체 열 분석을 통해 출력한 값으로 전체 데이터의 개수를 보여줍니다.

6. 프로그램 가동 방법

- 1.mysql을 커서 new schema를 만든다음에 university_df schema를 생성한다.
- 2.table_creator.sql를 전체 실행하여 8개의 테이블을 생성한다.
- 3.attach_mysql.ipynb를 실행하여 생성된 테이블에 '연구실_8개_테이블_정리.xlsx'파일 내용을 넣는다.
- 4.university_show_program.ipynb 또는 university_show_program.py을 열어
conn = pymysql.connect(host="127.0.0.1", user="root", password="1234",

```
db="university_df", charset="utf8")
```

이 쪽 코드를 환경에 맞게 설정한다.

5.앞에서 수정한 내용을 저장한 후 university_show_program.ipynb 또는 university_show_program.py을 실행하면 완성된 코드가 실행됩니다.

참고 할점

추가적으로 mysql_dump에 제 schema를 직접 담았습니다.

university_show_program.py의 경우 앞에서 연구실 정보를 잘 보여주지 못한 것을 개선한 버전으로 이것으로 실행해주시길 바랍니다.

연구실에서 각 연구실의 멤버목록, 논문 목록, 프로젝트 목록을 보기 위해서는 2[데이터분석],3[연구실별 내용 요약],s[연구실 선택],1~10[보기 원하는 연구실 번호 입력] 순으로 가시면 됩니다.

7.아쉬운점

node.js의 경우 잘몰라 vibe coding으로 진행하였다. 그래서 우선 코드를 돌릴 때 에러 사항이 많았고, 데이터를 일관된 형식으로 뽑을 수 없었다. 덕분에 파이썬으로 해당되는 내용을 고칠 필요가 있었다.

또한 로컬 컴퓨터의 한계로 인해 글자수를 1000자로 제한하여, 잘린 데이터가 많아 상세한 분석을 하지 못한점이 아쉽다.

웹사이트 크롤링 및 llm을 사용하여 데이터를 수집하고 가공하여 noise가 조금 섞여 있는 점도 아쉽다.

비정형 데이터를 정형 데이터로 만드는 과정에서 llm을 사용하였는데 논문 같은 경우 실제로 논문이 아닌데도 불구하고 출력하는 경우가 있었고(조선일보,책같은경우), 연도 같이 숫자를 다루는 부분에 있어서 사소한 오류가 조금씩 발견 되었다.

8.프로그램 실행 사진

교수님들 모음

연구실들 모임

연구실 학생 모임 - 최우석 교수님 연구실

장학 프로그램 - python C:\Users\god1\Desktop\202501\img\test\withcloud\재출출리딩\university_show_program.py

[연구실 정보] (페이지 1/1)

name	career_years	field	research_topic	etc

정보 없음	정보 없음	정보 없음	정보 없음	정보 없음

Hen Gyeol Song	미상	미상	Data-driven analysis for air quality	Email: thdksrnf0@gmail.com

Kyu Bo Jeon	미상	미상	Recommender system for climate prediction	Email: ksang1977@naver.com

Sungjoon Sohn	미상	미상	Station-temporal variation of O3	Email: sohnbo@naver.com

Ye Eun Kwon	미상	미상	Numerical weather prediction	Email: aba0327@naver.com

Chae Min Lim	미상	미상	Anomaly prediction	Email: lca1207@naver.com

Yi Kyung Kim	미상	미상	Energy prediction using time series analysis	Email: ykk1231@naver.com

Soo Beom Yun	미상	미상	Early detecting technology of wildfire	Email: scarlet0411@naver.com

Younghun Jeon	미상	미상	Wind power forecasting	Email: youngh1108@naver.com

Na Eun Kang	미상	미상	Indoor air quality analysis	Email: j_ameon@naver.com

[n] 다음 | [p] 이전 | [r] 필터 | [a] 종료

연구실 논문 모음 - 최우석 교수님 연구실

장학 프로그램 - python C:\Users\god1\Desktop\202501\img\test\withcloud\재출출리딩\university_show_program.py

[논문] - <https://sites.google.com/view/wschol-sejong> (페이지 1/1)

paper_title	year	journal	author

Regional classification based on spatiotemporal variability of ozone concentrations in the Seoul Metropolitan Area and diagnosis of key mechanisms	2025	to be submitted	Sohn, S., and H. Choi*

Particulate matter characteristics by forecast regions in Korea using the fusion of explainable artificial intelligence and clustering		SQA, to be submitted	Kwon, Y., H.-J. Song, and H. Choi*

Impact of anthropogenic factors on wildfire occurrence prediction using deep learning in Gangwon state, South Korea	2025	Engineering Applications of Artificial Intelligence, to be submitted	Jeon, G., and H. Choi*

Analysis of summertime green-roof effects on temperature and O2-concentration reduction in Seoul	2025	S (게시된 내용이 없습니다.)	Lim, C., and H. Choi*

[n] 다음 | [p] 이전 | [a] 종료