SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

05 PJT

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

웹 크롤링 실습

Confidential

챕터의 포인트

- 목표
- 웹 크롤링 이해하기
- [실습] 웹 크롤링 실습

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

목표

Confidential

프로젝트 파악하기

- 친구들끼리 같이 먹을 음식을 주문하기로 했다.
- 갑자기 궁금해졌다.

  사람들이

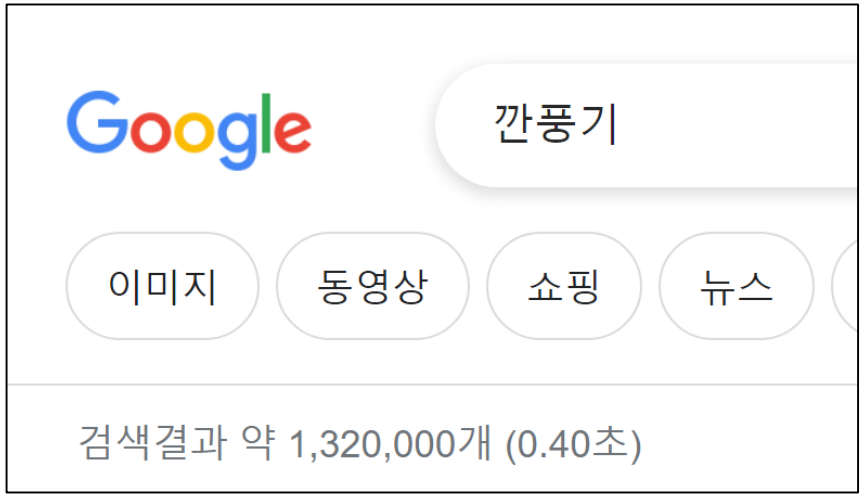
  깐풍기를 더 선호할까?

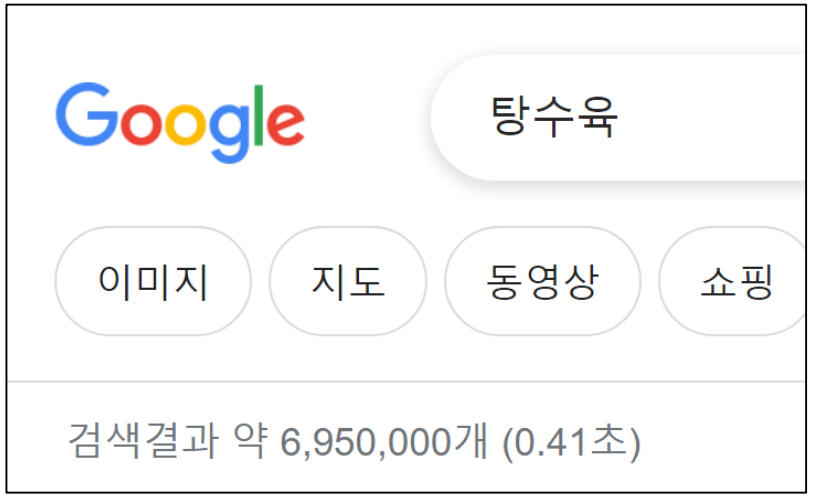
  탕수육을 더 선호할까?

Confidential

프로젝트 파악하기

- 구글에 검색해보고 어떤 메뉴가 더 많이 검색되는 지로 판별해보고자 합니다.





- 어떻게 위와 같은 웹 페이지의 결과를 코드에서 활용할 수 있을까요?

Confidential

파이썬으로 웹 페이지에 있는 정보를 가져오는 방법

- 크게 세 가지 방법으로 가져올 수 있습니다.

1. 누군가 업로드해 둔 데이터를 다운로드 받기 (ex. 캐글)
2. 누군가 만들어 둔 API Server 를 활용하여 정보를 받아오기
    - 아마, 깐풍기와 탕수육 API Server는 아무도 만들어 두지 않았을 거 같다
3. 사람이 검색하는 것처럼 파이썬이 자동으로 검색 후 결과를 수집하는 방법
    - 이러한 기술을 크롤링(Crawling) 이라고 합니다.
    - 이번 프로젝트에서 사용할 기술입니다.

Confidential

Quiz

- 데이터 사이언스에서는 먼저 데이터를 수집하는 것이 중요합니다.

  Kaggle 같은 데이터 공유 플랫폼에서 수집된 데이터들을 쉽게 다운로드 받을 수 있지만

  우리는 XXX 라는 기술을 사용하여 직접 데이터를 수집하고자 합니다.

  이 기술의 이름은 무엇인가요 ?

Confidential

프로젝트 목표

1. Django 없이, 크롤링 하는 방법 학습

2. 구글 검색 수를 크롤링하여 어떤 키워드가 더 많이 검색되는 지 조사하기

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

웹 크롤링 이해하기

Confidential

[복습] 데이터 사이언스 프로세스

- 필요한 정보를 추출하는 5가지 단계

1. 문제 정의 : 해결하고자 하는 문제 정의
2. 데이터 수집 : 문제 해결에 필요한 데이터 수집
3. 데이터 전처리(정제) : 실질적인 분석을 수행하기 위해 데이터를 가공하는 단계
    - 수집한 데이터의 오류 제거(결측치, 이상치), 데이터 형식 변환 등
4. 데이터 분석 : 전처리가 완료된 데이터에서 필요한 정보를 추출하는 단계
5. 결과 해석 및 공유 : 의사 결정에 활용하기 위해 결과를 해석하고 시각화 후 공유하는 단계

Confidential

금융상품비교

영화추천서비스

[복습] 데이터 수집

- 데이터 수집은 다양한 기술과 방법을 활용할 수 있습니다.
  - 웹 스크래핑(Web Scraping): 웹 페이지에서 데이터를 추출하는 기술
  - 웹 크롤링(Web Crawling): 웹 페이지를 자동으로 탐색하고 데이터를 수집하는 기술
  - Open API 활용: 공개된 API 를 통해 데이터를 수집
  - 데이터 공유 플랫폼 활용: 다양한 사용자가 데이터를 공유하고 활용할 수 있는 온라인 플랫폼
    - 종류: 캐글(Kaggle), Data.world , 데이콘(Dacon), 공공데이터포털 등

이미지 출처: https://www.google.com/

Confidential

웹 크롤링이란?

- 여러 웹 페이지를 돌아다니며 원하는 정보를 모으는 기술
- 원하는 정보를 추출하는 스크래핑(Scraping) 과 여러 웹 페이지를 자동으로 탐색하는 크롤링 (Crawling) 의 개념을 합쳐 웹 크롤링이라고 부름
- 즉, 웹 사이트들을 돌아다니며 필요한 데이터를 추출하여 활용할 수 있도록 자동화된 프로세스

Confidential

웹 크롤링 프로세스

- 웹 페이지 다운로드
  - 해당 웹 페이지의 HTML, CSS, JavaScript 등의 코드를 가져오는 단계
- 페이지 파싱
  - 다운로드 받은 코드를 분석하고 필요한 데이터를 추출하는 단계
- 링크 추출 및 다른 페이지 탐색
  - 다른 링크를 추출하고, 다음 단계로 이동하여 원하는 데이터를 추출하는 단계
- 데이터 추출 및 저장
  - 분석 및 시각화에 사용하기 위해 데이터를 처리하고 저장하는 단계

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

웹 크롤링 실습

Confidential

준비 단계

- 실습 및 도전 과제에는 구글 검색 결과 페이지를 크롤링합니다.
- 아래 필수 라이브러리를 설치 후 진행합니다.
  - requests: HTTP 요청을 보내고 응답을 받을 수 있는 모듈
  - BeautifulSoup: HTML 문서에서 원하는 데이터를 추출하는 데 사용되는 파이썬 라이브러리
  - Selenium: 웹 애플리케이션을 테스트하고 자동화하기 위한 파이썬 라이브러리
    - 웹 페이지의 동적인 컨텐츠를 가져오기 위해 사용함 (검색 결과 등)
- `$ pip install requests beautifulsoup4 selenium`

Confidential

금융상품비교

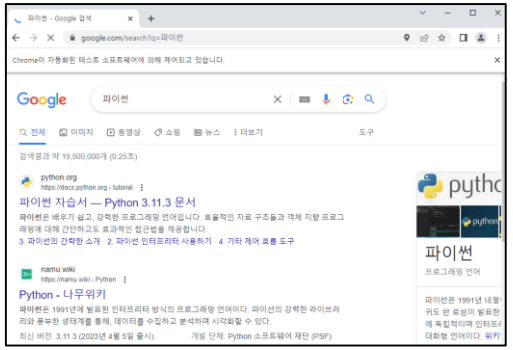
영화추천서비스

기본 예제

- examples/example1.py

```python
from bs4 import BeautifulSoup
from selenium import webdriver

def get_google_data(keyword):
    url = f"https://www.google.com/search?q={keyword}"
    # 크롬 브라우저가 열린다. 이 때, 동적인 내용들이 모두 채워짐
    driver = webdriver.Chrome()
    driver.get(url)

    # 열린 페이지 소스를 받아옴
    html = driver.page_source
    soup = BeautifulSoup(html, "html.parser")

    # 눈으로 보기 좋게 출력
    print(soup.prettify())

    # 파일로 저장하여 확인하기
    with open('soup.txt', 'w', encoding="utf-8") as file:
        file.write(soup.prettify())


# 검색 키워드 설정
keyword = "파이썬"
get_google_data(keyword)
```

- 실행 결과1. 파이썬을 검색 한 구글 창



- 실행 결과2. 엄청 나게 긴 페이지 코드 (분석 불가)

```
ed=1/dg=2/br=1/rs=ACT90oEgzTU-WoZSUdoAV0UvycR8HFVTsw/m=kMFpHd,sy2d,bm51tf?xjs=s3">
  </script>
  <script async="" nonce="" src="/xjs/_/js/k=xjs.s.ko.VNESo4_-d7Q.O/ck=xjs.s.3HFJhKov9JI.L.W.O/
am=CggBIAAgGoRTABtAAPgnDAAAEBAAAAAAAFACYEAgeP8JAQAAQEQMQQwwAJBQAiYFAADg9EMEGACAAGIACgAARQAcNAQq
AAIAAAAgfwDMeQGAgwkLAAAAAAAAAAAIYAmCwQVSKAgAAQAAAAAAAAACAKpm8PCAEAAAC/d=0/excm=A1Sy2b,ABxRVc,AD6
AIb,AOTkuc,CVVp5c,CnT5wd,D1J6He,FXUdw,FmnE6b,FuQWyc,GRJ32c,HfxK9d,JxE93,KrUr5e,LtNDTb,MRb7nf,Mr
kcAd,Mxvwsd,NhUbHc,NmR9jd,NsEUGe,NzGbYd,Oa7Qpb,Ok4XMd,PoJj8d,SKZSKc,SLDaee,TO0csb,U3Ovcc,U6nlJe
,UQpTU,UZNwo,UbcHRb,V9W1ad,WaSRUb,Wx0Z2d,WxJ6g,XHo6qe,XOehOc,XTkmZd,Xk0c,Y0dpFc,Yltq7c,ZrXR8b,Z
udxcb,a0nyD,bXKPzd,bXyZdf,cKV22c,dyUEmd,eTv59e,ee9Gld,f26on,hWJjIf,hfJ9hb,jkRPje,kOSi0d,llagHf,
mL4hG,pMwOEe,pQk1fc,pqUxUc,rL2AR,smKWJb,tzTB5,vJPFse,vPi79c,y25qZb,y6Ihab,yChgtb,yuQBec,zoFt6e/
ed=1/dg=2/br=1/rs=ACT90oEgzTU-WoZSUdoAV0UvycR8HFVTsw/m=syk8,syk9,dt4g2b?xjs=s3">
  </script>
 </body>
</html>
```

Confidential

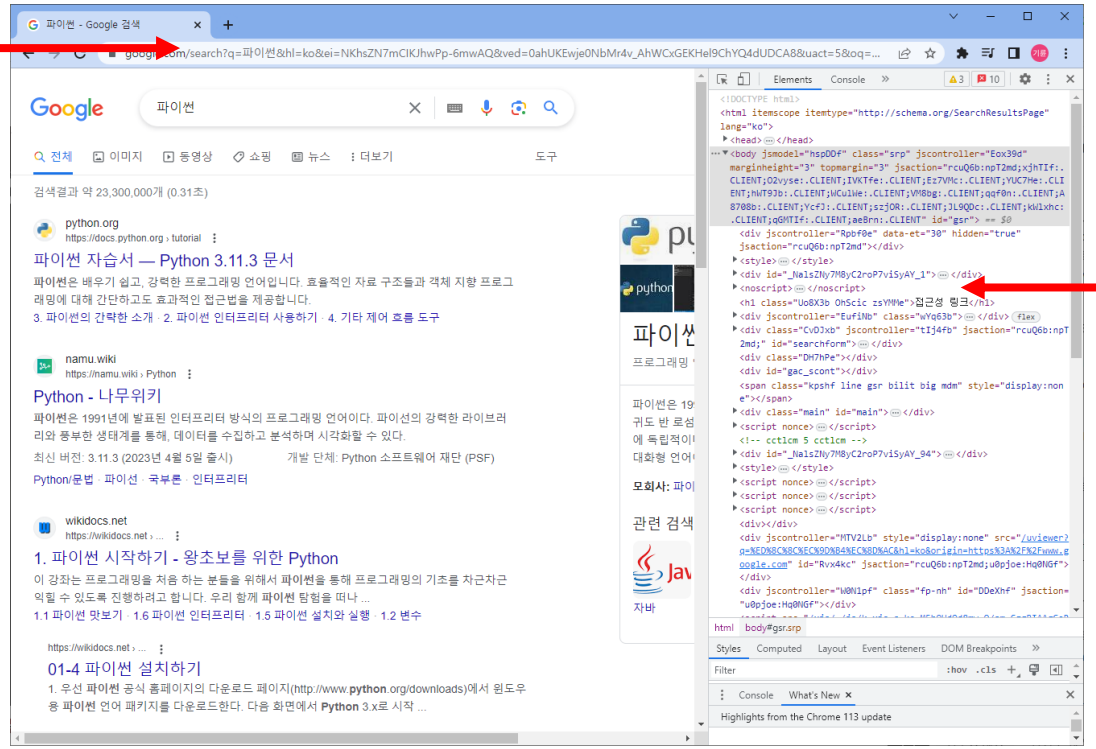
금융상품비교

영화추천서비스

구글 검색 결과 분석하기(1/4)

- “F12” 혹은 “우측 클릭 - 검사” 로 크롬 개발자 도구를 열어 활용합니다.

q=파이썬



HTML, CSS, JavaScript 코드

- id 와 class 이름이 이상하다!
- id 는 새로고침마다 변한다.

사람이 정하는 것이 아니라,

프로그래밍 되어 있다.

즉, id 값이 아닌 class 와 태그를

기준으로 정보를 추출해야 한다.

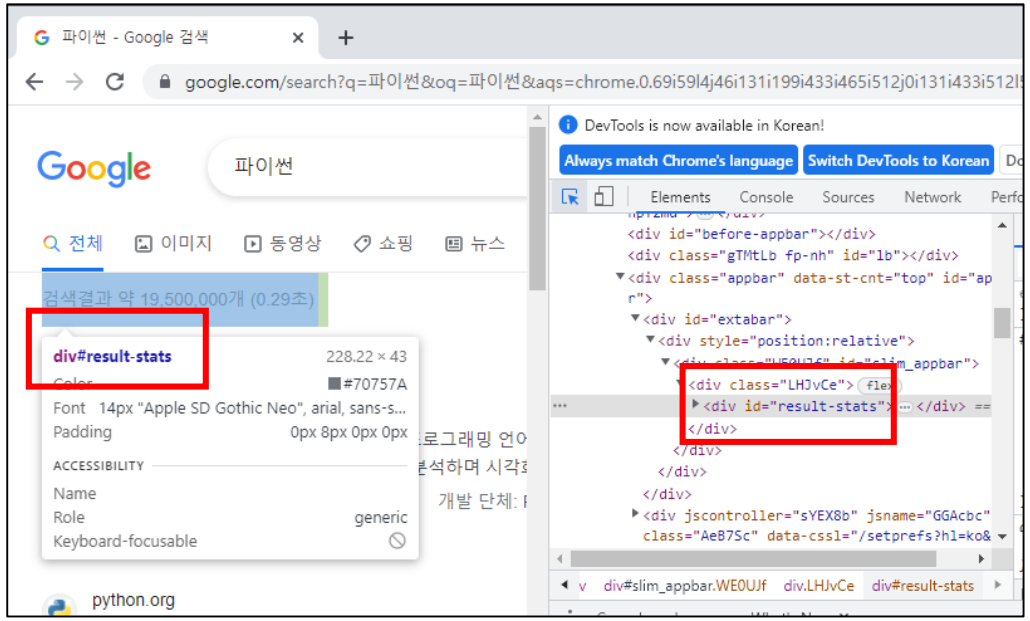
Confidential

금융상품비교

영화추천서비스

구글 검색 결과 분석하기(2/4)

- 예시1. 검색 결과 개수 출력
- div 태그 이면서 id 가 “result-stats” 이다



- example2.py

```python
from bs4 import BeautifulSoup
from selenium import webdriver

def get_google_data(keyword):
    url = f"https://www.google.com/search?q={keyword}"
    # 크롬 브라우저가 열린다. 이 때, 동적인 내용들이 모두 채워짐
    driver = webdriver.Chrome()
    driver.get(url)

    # 열린 페이지 소스를 받아옴
    html = driver.page_source
    soup = BeautifulSoup(html, "html.parser")

    # div 태그 중 id 가 result-stats 인 요소 검색
    result_stats = soup.select_one("div#result-stats")
    print(result_stats)

# 검색 키워드 설정
keyword = "파이썬"
get_google_data(keyword)
```

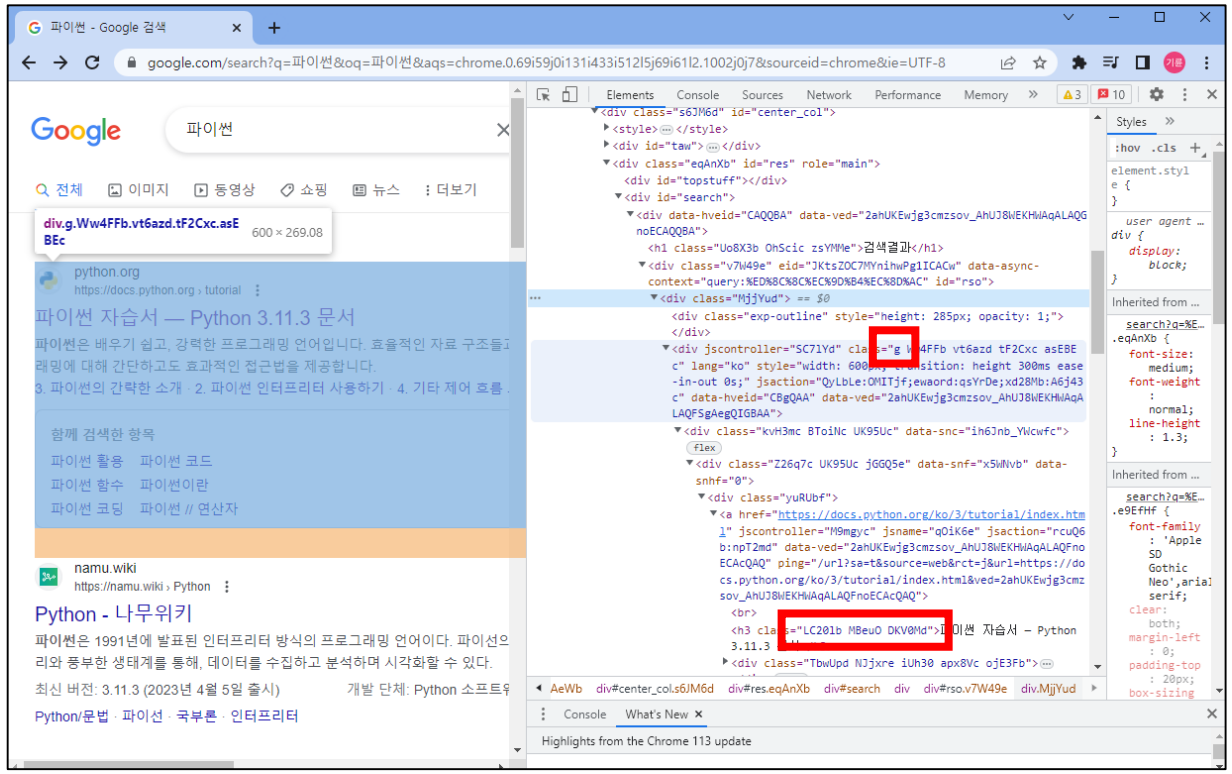
Confidential

금융상품비교

영화추천서비스

구글 검색 결과 분석하기(3/4)

- 예시2. 검색 결과 페이지들의 제목 가져오기



공통적으로

결과를 감싸는 div 에는 “g” 클래스

제목에는 “LC20lb MBeuO DKV0Md”

클래스를 가지고 있습니다.

Confidential

금융상품비교

영화추천서비스

구글 검색 결과 분석하기(4/4)

- 예시2. 검색 결과 페이지들의 제목 가져오기

- example3.py

```python
from bs4 import BeautifulSoup
from selenium import webdriver

def get_google_data(keyword):
    url = f"https://www.google.com/search?q={keyword}"
    # 크롬 브라우저가 열린다. 이 때, 동적인 내용들이 모두 채워짐
    driver = webdriver.Chrome()
    driver.get(url)

    # 열린 페이지 소스를 받아옴
    html = driver.page_source
    soup = BeautifulSoup(html, "html.parser")

    # div 태그 중 g 클래스를 가진 모든 요소 선택
    g_list = soup.select("div.g")
    # 해당 요소를 반복하며
    for g in g_list:
        # 요소 안에 LC20lb MBeuO DKV0Md 클래스를 가진 특정 요소 선택
        title = g.select_one(".LC20lb.MBeuO.DKV0Md")
        # 요소가 존재 한다면
        if title is not None:
            title_text = title.text
            print('제목 = ', title_text)

# 검색 키워드 설정
keyword = "파이썬"
get_google_data(keyword)
```

- 출력 결과

```
제목 =  Python - 나무위키
제목 =  파이썬 자습서 — Python 3.11.3 문서
제목 =  1. 파이썬 시작하기 - 왕초보를 위한 Python
제목 =  [Python] Python이란? - Maker's VAP - 티스토리
제목 =  Python란 무엇인가요? - Python 언어 설명 - Amazon AWS
제목 =  파이썬 - 위키백과, 우리 모두의 백과사전
제목 =  Python란 무엇인가요? - Python 언어 설명 - Amazon AWS
제목 =  최신 파이썬 코딩 무료 강의 - 5시간만 투자하면 개발자가 됩니다
제목 =  1 장 파이썬(Python) 입문 | 파이썬 프로그래밍 기초
제목 =  1) 파이썬 개요 - 코딩의 시작, TCP School
제목 =  파이썬 코딩을 시작하기 좋은 쉬운 아이디어들 - freeCodeCamp
```

Confidential

[참고] BeautifulSoup4 요소 선택 메서드 종류

- find()
  - 태그를 사용하여 요소를 검색. 첫 번째로 일치하는 요소를 반환
- find_all()
  - 태그를 사용하여 요소를 검색. 모든 일치하는 요소를 리스트로 반환
- select()
  - CSS 선택자를 사용하여 요소를 검색. 모든 일치하는 요소를 리스트로 반환
- select_one()
  - CSS 선택자를 사용하여 요소를 검색. 첫 번째로 일치하는 요소를 반환
- find_parent() / find_next_sibling() / find_previous_sibling()
  - 태그를 사용하여 요소를 검색. 각각 일치하는 요소의 부모/다음 형제 요소/이전 형제 요소를 반환
- 공식문서 참고

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

도전 과제

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

금융 상품 비교 앱
PJT 05

Confidential

관통 Ver1 - PJT05 도전 과제

- 프로젝트명: 키워드 검색량 분석을 위한 데이터 수집
- 목표
  - 크롤링을 통한 데이터 수집
  - 수집한 데이터를 DB 에 저장하고, 저장한 데이터 활용하기
- 특징
  - 데이터 사이언스 패키지 사용
  - 수집한 데이터를 저장하고 활용할 수 있도록 DB 설계

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

영화 추천 서비스
PJT 05

관통 Ver2 - PJT05 도전 과제

- 프로젝트명: DB를 활용한 웹 페이지 구현
- 목표
  - 영화, 회원, 댓글 간의 모델 관계가 형성된 애플리케이션 완성
- 특징
  - 1:N 관계 이해 및 응용