10조

10조

데이터 분석 실습
미니 프로젝트

소개

윤지원

이창준

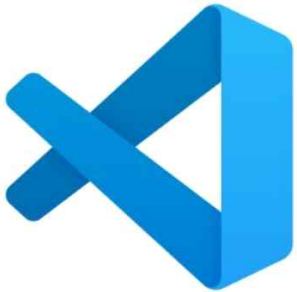
데이터 분석 실습
미니 프로젝트

개발 환경

OS

MacOS

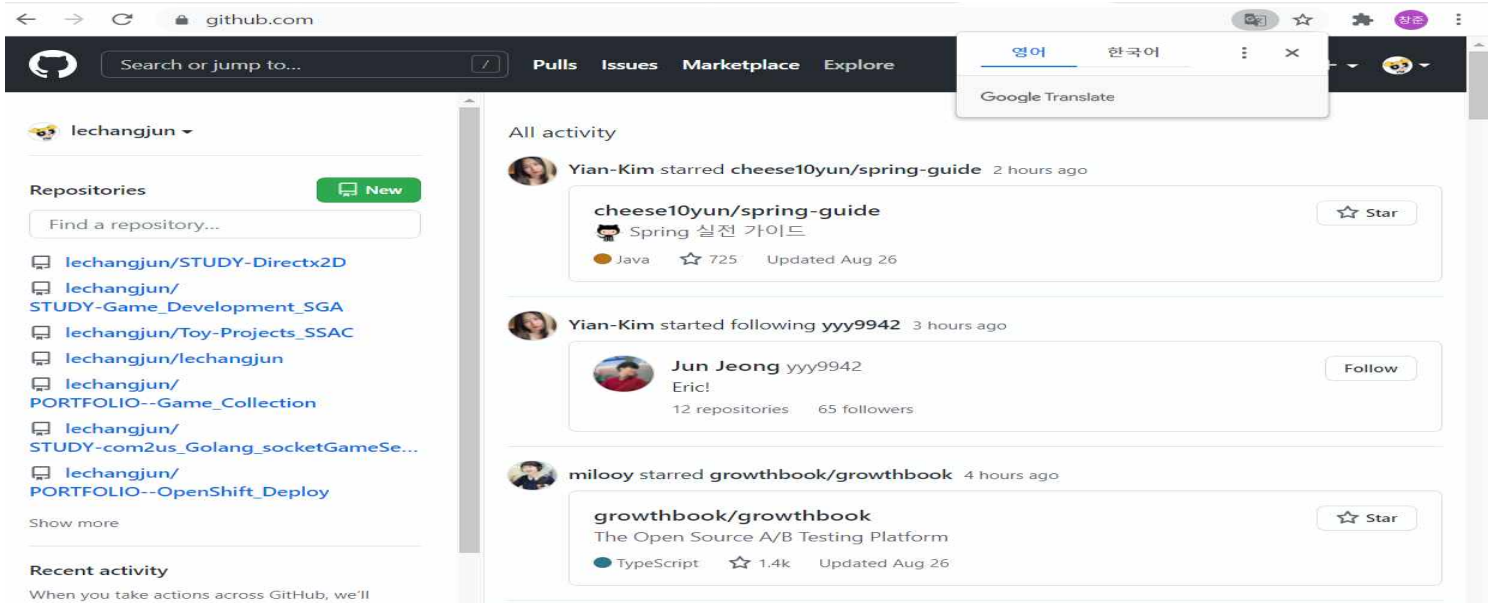
IDE



GitHub

협업 도구 및 버전 관리 도구

원격 저장소로 GitHub(Git 호스팅 사이트) 활용



프로젝트

> 데이터 획득 (크롤링)

> 데이터 저장 [총 6개]

> DB 연동

> 데이터 분석

> 데이터 **시각화**

1. 데이터 획득(크롤링)
• 데이터
  • 당일 지역별 코로나 누적확진자수 및 신규확진자수
  • 최근 7일간 국내발생 및 해외유입 코로나 신규확진자수
  • 데이터 수집처
  • 네이버에서 '코로나 현황'으로 검색
  • 추가 점수 • selenium으로 버튼 클릭 수행

2. 데이터 저장 [총 6개]
• 저장 포맷 2개 -> 주피터 파일 ipynb
  • 엑셀 파일 2개
  • MySQL 테이블 2개

3. 데이터 분석 및 시각화
• 데이터 분석
  • (간단한) 자유 주제
  • 분석 결과 작성
• 시각화
  • (필요하다면) 자유 주제로 시각화
  • 당일 지역별 데이터 ➔ 지도에 표시
  • 최근 7일간 데이터 ➔ 적절한 그래프 활용

• 주의 사항
 • 새 노트에서 수행
 • MySQL에 저장된 데이터 사용

1.데이터 크롤링

• 데이터
  • 당일 지역별 코로나 누적확진자수 및 신규확진자수
  • 최근 7일간 국내발생 및 해외유입 코로나 신규확진자수
oo
  • 데이터 수집처
  • 네이버에서 '코로나 현황'으로 검색

__ BeautifulSoup을 이용한 정보 찾기
__ HTML 정보 찾기 ① - 태그 속성 활용
__ HTML 정보 찾기 ② - 상위 구조 활용

  • 추가 점수[• selenium으로 버튼 클릭 수행]
    html -> dom
    _ 멜론 노래 순위 정보 크롤링
    _ selenium을 활용한 크롤링

코딩 거니 강의 시청

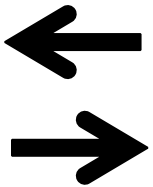
책 이름: 이것이 MySQL이다

selenium 이용해서 페이지 넘기기 >> driver.find_element_by_link_text('텍스트')¶

[당일 지역별 코로나 누적확진자수 / 신규 확진자수]
지역별 표 화면> driver.find_element_by_link_text('지역별 표').click()
지역별 표 다음 페이지> driver.find_element_by_link_text('다음').click()

프로젝트

NAVER



프로젝트 주제

> 당일 지역별 코로나 누적확진자수 및 신규확진자수
> 최근 7일간 국내발생 및 해외유입 코로나 신규확진자수

10조

윤지원

당일 지역별 코로나 누적확진자수 및 신규확진자수

이창준

최근 7일간 국내발생 및 해외유입 코로나 신규확진자수

Openshy
4j

데이터 크롤링

Listly

데이터 수집 코드

자료수집

```
from bs4 import BeautifulSoup
import requests
from selenium import webdriver
import pandas as pd
```

```
browser = webdriver.Chrome('./chromedriver.exe')
url = 'https://search.naver.com/search.naver?query=%EC%BD%94%EB%A1%9C%EB%82%98+%ED%98%84%ED%99%A9&ie=utf8&sm=whl_nht'
browser.get(url)
html = browser.page_source
soup = BeautifulSoup(html, 'html.parser')
```

```
section = soup.select('tbody > tr')
section = list(section)
```

```
data_list = [[sub_section.select('span')[0].text.strip(),
              sub_section.select('span')[1].text.replace(",",""),
              sub_section.select('span.confirmed_case')[0].text
              ] for num, sub_section in enumerate(section)]

data_list
```

데이터 수집 도전

2. 최근 7일간 국내발생 및 해외유입 코로나 신규확진자수

자료수집

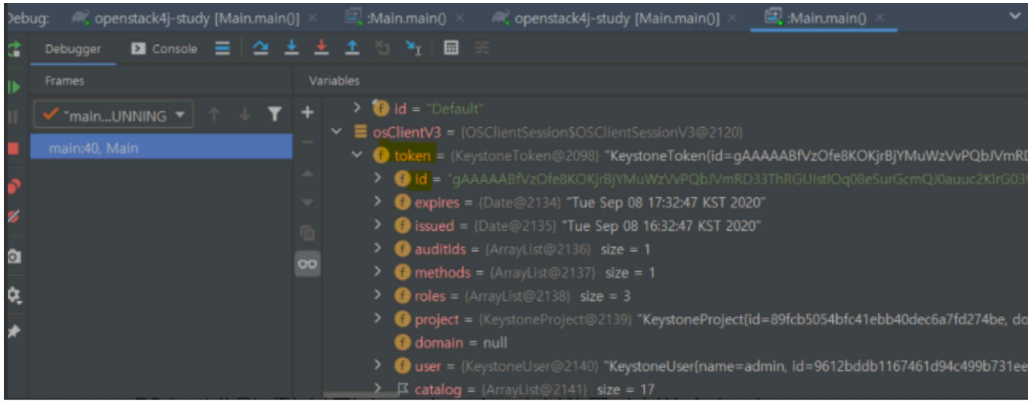
- 1차: get 함수 사용 --> 실패
- 2차: get 함수 사용 --> 실패
- 3차: Listly - Data Scraper 도구 사용

Listly

> 확장 프로그램 추가

> 확장 프로그램 관리

> 화면 지정



DB

데이터 저장

엑셀

```
df_data_list.to_excel('./covid19_status.xlsx', index = False)
```

MySQL

DB 생성

```
import pymysql
```

```
connect = pymysql.connect(host="127.0.0.1", user="root", password="1234",
                          charset='utf8',cursorclass=pymysql.cursors.DictCursor)
try:
    cur = connect.cursor()

    cur.execute("DROP DATABASE IF EXISTS projectDB")
    cur.execute("CREATE DATABASE projectDB")

    cur.execute("SHOW DATABASES")

    db_list = cur.fetchall()

    for db in db_list:
        print(db)

except Exception as e:
    print("Exeception occured:{}".format(e))

# finally:
#     connect.close()
```

DB

Data 불러오기

In [49]:

```
data1, data2, data3 == "","",""
row=None


conn = pymysql.connect(host='127.0.0.1', user='root', password='1234',db= 'projectDB', charset='utf8')
cur = conn.cursor()
cur.execute("SELECT * FROM covid19")

print("지역₩t₩t누적확진자₩t₩t신규확진자")

while 1:
    row = cur.fetchone()
    if row == None:
        break
    data1 = row[0]
    data2 = row[1]
    data3 = row[2]

ex =[]
    ex += [row[0], row[1], row[2]]

    print("{}₩t₩t{}₩t₩t{}".format(data1, data2, data3))
print(ex)
```

```
지역	누적확진자		신규확진자
서울	76814	569
경기	68671	513
대구	13287	95
인천	11406	102
부산	11077	78
경남	9683	68
경북	6866	64
충남	6639	92
```

데이터 랭킹

3. 랭킹 데이터 시각화하기

```
# 라이브러리 추가하기
import pandas as pd
import matplotlib.pyplot as plt
```

```
# 그래프에서 한글을 표기하기 위한 글꼴 변경
from matplotlib import font_manager, rc
import platform
if platform.system() == 'Windows':
    path = 'c:/Windows/Fonts/malgun.ttf'
    font_name = font_manager.FontProperties(fname = path).get_name()
    rc('font', family = font_name)
elif platform.system() == 'Darwin':
    rc('font', family = 'AppleGothic')
else:
    print('Check your OS system')
```

```
# 엑셀 파일 불러오기
df = pd.read_excel('./files/LISTLY_TRIAL_20210826.xlsx')
df.head()
```

[26]:

	LABEL-1	LABEL-2
0	8.18	2,152
1	8.19	2,050
2	8.20	1,879
3	8.21	1,626
4	8.22	1,417

데이터 시각화 준비

시각화

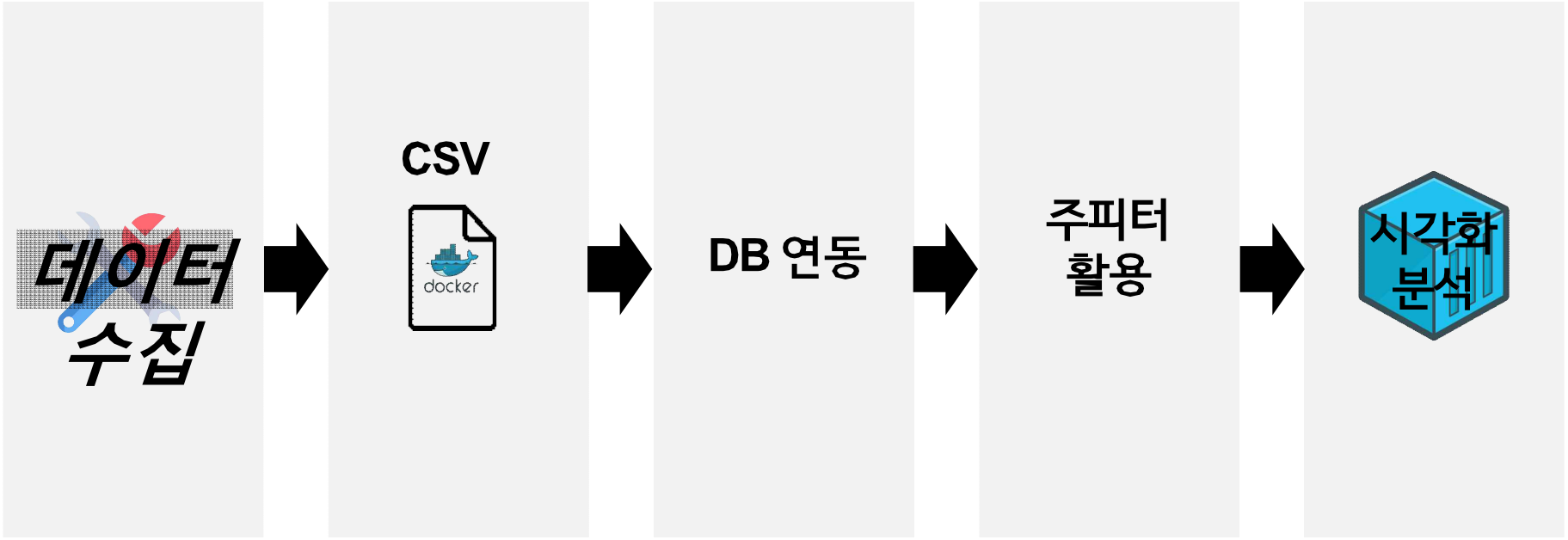
```
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df.head()
```

```
df = pd.read_excel('./files/LISTLY_TRIAL_20210826 (1).xlsx')
plt.plot(df_filter['LABEL-1'], df_filter['LABEL-2'])
plt.show()
```

```
df_pivot = df_filter.pivot_table(index = 'LABEL-1', columns = 'LABEL-2')
df_pivot
```

데이터 분석 Workflow



qna

Thank
yoU