학습 정리

팀	이진조	구성원	이규진

일정	발제자	주제
5/30 (목)	이규진	requests 모듈 기초, 웹 브라우저 없는 스크랩핑 및 파싱 실습
주요 내용 요약		

Section 3 - 파이썬 고급 스크랩핑

Http 통신 기초

• 오늘 내용 정리

1. request, response 간단 개념 알아보기
2. cookie, session 개념 알아보기

http 통신

1. 비연결지향
2. 상태정보 유지안함

=> 만약 쿠기와 세션정보가 없다면 페이지 이동할 때마다 로그인해야 할 수 있음.

- **쿠키**: 자동로그인, 오늘 하루는 이 창을 열지 않겠다!, 쇼핑몰 장바구니 정보
- **세션**: 서버

파이썬 requests 모듈 기초

사용법(1)

• 오늘 내용 정리

1. requests 모듈 사용법(1) 및 장점 - urllib
2. Json 데이터 핸들링
3. requests 모듈 테스트 실습

requests 모듈 사용

```
s = requests.Session() # Session을 열어주는 명령어,

# r = s.get("https://www.naver.com") # PUT(FETCH), DELETE, GET, POST

# print('1', r.text)


# r = s.get('http://httpbin.org/cookies', cookies={'from':'myName'})

# print(r.text)


url = "https://httpbin.org/get"

headers = {'user-agent': 'myPythonApp_1.0.0'}


# r = s.get(url,headers=headers)

# print(r.text)


s.close() # 반드시 리소스 낭비를 위해 닫아줘야함


with requests.Session() as s:

  r = s.get('https://www.naver.com')

  print(r.text)
```

Json 데이터 핸들링

```
# Response 상태 코드
```

```
s = requests.Session()
r = s.get("http://httpbin.org/get")
# print(r.status_code)
# print(r.ok)


# https://jsonplaceholder.typicode.com


r = s.get('https://jsonplaceholder.typicode.com/posts/1')
# print(r.text)
print(r.json()) # 데이터를 json형태로 컨버트해줌
print(r.json().keys())
print(r.json().values())
print(r.encoding)
print(r.content) # b가 붙음으로 바이너리 형태로 줄바꿈을 문자를 포함하여 가져옴
print(r.raw)
= requests.Session()


r = s.get('http://httpbin.org/stream/20', stream=True)
# print(r.text)
# print(r.encoding) # 위 사이트는 encoding형태가 none으로 되있음
# print(r.json())


if r.encoding == None:
  r.encoding = 'utf-8'


for line in r.iter_lines(decode_unicode=True):
  # print(line)
  b = json.loads(line) # dict형태
  # print(type(b))
```

```
  # print(b['origin'])

  for e in b.keys():

      print("key:", e, "vlaues:",b[e])
```

파이썬 requests 모듈 기초 사용법(2)

• 오늘 내용 정리

1. requests 모듈 사용법(2)
2. requests 모듈 Rest API 실습

실습(과제): https://www.apistore.co.kr/api/apiList.do 사용해보기

```
r = requests.get("https://api.github.com/events")

r.raise_for_status() # requests에서 에러가 발생했을 때 예외를 발생 시켜줌

# 페이지가 없으면 404에러를 발생시킴

print(r.text)


jar = requests.cookies.RequestsCookieJar()


jar.set('name','kim',domain='httpbin.org',path='/cookies')


r = requests.get('http://httpbin.org/cookies',cookies=jar)

r.raise_for_status()

print(r.text)


r = requests.get('https://github.com',timeout=3)

print(r.text)


r = requests.post('http://httpbin.org/post', data={'name':'kim'}, cookies=jar)
```

```
print(r.text)


payload1 = {'key1':'value1', 'key2':'value2'}

payload2 = (('key2','value2'),('key3','value3'))

payload3 = {'some':'nice'}


r = requests.post("http://httpbin.org/post", data=payload1) # form데이터로 요청

print(r.text)


r = requests.post("http://httpbin.org/post", data=json.dumps(payload3)) # json데이터로 요청

print(r.text)

# Rest Api = POST(보내기), GET(가져오기), PUT(FETCH)(수정), DELETE


sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')

sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')


payload1 = {'key1':'value1', 'key2':'value2'}

payload2 = (('key2','value2'),('key3','value3'))

payload3 = {'some':'nice'}


r = requests.put("http://httpbin.org./put", data=payload1)

print(r.text)


r = requests.put('https://jsonplaceholder.typicode.com/posts/1', data=payload1)

print(r.text)


r = requests.delete('https://jsonplaceholder.typicode.com/posts/1')

print(r.text) # 실제로 삭제된 것은 아님
```

파이썬 requests 통신 실습 고급(1)

• 오늘 내용 정리

1. 루리웹(Ruliweb) 사이트 로그인 처리 후 게시판 글 가져오기
2. 인프런(inflearn) 사이트 로그인 처리 후 개인정보 가져오기

1. 루리웹

```
import requests

from bs4 import BeautifulSoup

import sys

import io



sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')

sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')


# 로그인 유저정보

LOGIN_INFO = {

  'user_id': 'ymbrdii2002',

  'user_pw': 'rbwlsl91'

}


# Session 생성, with 구문안에서 유지

with requests.Session() as s:

  login_req = s.post("https://user.ruliweb.com/member/login_proc",data=LOGIN_INFO)

  # HTML 소스 확인
```

```
    # print('login_req',login_req.text)

    # Header 확인

    # print('headers',login_req.headers)

    if login_req.status_code == 200 and login_req.ok:

        post_one =
s.get('https://market.ruliweb.com/read.htm?table=market_ps&page=1&num=4457079&find=&ftext='
)

        post_one.raise_for_status()

        soup = BeautifulSoup(post_one.text, 'html.parser')

        # print(soup.prettify())



        article = soup.select_one("table:nth-child(1)").find_all('p')

        # print(article)

        for i in article:

            if i.string is not None:

                print(i.string)
```

2.인프런

```
import requests

from bs4 import BeautifulSoup

import urllib.parse as rep

import urllib.request as req

import sys

import io

import os


sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')

sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')
```

```
# 로그인 유저정보
LOGIN_INFO = {
  'email': '본인 아이디(인프런)',
  'password': '본안 비번(인프런)'
}

# Session 생성, with 구문 안에서 유지
with requests.Session() as s:

  login_req = s.post('https://www.inflearn.com/api/signin', data=LOGIN_INFO)
  # HTML 소스 확인
  # print('login_req'.format(login_req.text))
  # HTTP Header 확인
  # print('login_req'.format(login_req.headers))

  # Response 정상 확인
  if login_req.status_code == 200 and login_req.ok:
    # 인프런 개인 대시보드 정보 확인하기

    # URL 확인
    # URL 부분의 숫자 부분은 본인의 대시보드 URL 확인 후 수정 한다.
    # 예) https://www.inflearn.com/users/40769/dashboard
    dash_info = s.get('https://www.inflearn.com/users/숫자/dashboard')

    # 수신 에러시 예외 발생
    dash_info.raise_for_status()

    # 수신 확인
    # print(dash_info.text)
```

```
    #BeautifulSoup 선언
    soup = BeautifulSoup(dash_info.text,'html.parser')

    # 수신 HTML 정리
    # print(soup.prettify())

    statistics = soup.select("div.box.statistics > div.box_content > div > div")

    # 확인
    # print(statistics)

    for v in statistics:
        print()
        # 레이블 명
        lable = v.find('div', class_="status_label").text.strip()
        # 통계
        status = v.find('div', class_="status_value").text.strip()
        # 출력
        print('{} : {}'.format(lable, status))
```