데이터 크롤링

Crawling

- requests
- BeautifulSoup
  - find / find_all
  - select_one / select

■ URL

● Uniform Resource Locator

① ② ③ ④ ⑤ ⑥

● https://news.naver.com:80/main/read.naver

⑦

?mode=LSD&mid=shm&sid1=105&oid=001&aid=0009847211

⑧

#da_727145

① http:// or https:// - Protocol

② news - Sub Domain

③ naver.com - Domain

④ :80 - Port

⑤ /main/ - Path

⑥ read.nhn - Page

⑦ ?mode=LSD&mid=shm&sid1=105&oid=001&aid=0009847211 - Query String (Parameter)

⑧ #da_727145 - Fragment

■ Crawling

- 조직적 / 자동화 된 방법으로 데이터를 탐색하거나 수집하는 것

- 데이터 수집 절차
  - 원하는 URL에 request를 보내고 결과를 받아온다
  - 받은 결과물 (HTML / JSON / XML)을 파싱(Parsing)한다
  - 필요한 정보만 추출한다

- 파이썬에서 크롤링을 하기 위해 필요한 라이브러리
  - 데이터 통신 : requests / urllib
  - 데이터 추출 : bs4 (BeautifulSoup)

※ selenium

■ 웹 페이지의 종류

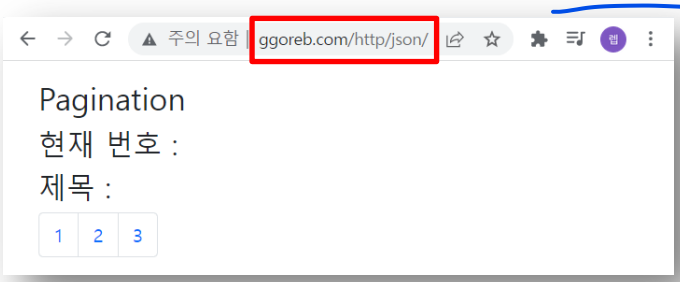
● 정적 페이지

- http://ggoreb.com/http/get/

- http://ggoreb.com/http/post/

● 동적 페이지

- http://ggoreb.com/http/json/



■ requests vs selenium

● requests

- 받아오는 문자열에 따라 두가지 방법으로 구분
- html / xml : requests + BeautifulSoup
- json : requests

● selenium

- 브라우저를 직접 열어서 데이터를 받는 방법
- 웹 드라이버를 통해 브라우저를 제어

※ 속도 순위

requests > requests + BeautifulSoup > selenium

json　　html / xml　　브라우저 제어

■ requests

● 모듈 가져오기

```
import requests
```

● URL 호출 후 응답코드 확인

```
result = requests.get('http://ggoreb.com/http')
print(result.status_code)
```

```
200
```

● 응답결과 확인

```
print(result.text)
```

```
"<html lang='ko'>\n<head>\n    <meta charset='utf-8' >\n</head>\n<body>\n
<h3>\n        <a href='#'>í\x97\x88ë\x8b\x88ë¹\x84</a>\n
```

● Encoding 처리 (한글 깨짐)

```
result.encoding = 'utf-8'
print(result.text)
```

```
"<html lang='ko'>\n<head>\n    <meta charset='utf-8' >\n</head>\n<body>\n
<h3>\n        <a href='#'>허니비</a>\n    </h3>\n</body>\n</html>"
```

■ requests

● 기본 사용법

```
import requests
res = requests.get('http://ggoreb.com/python/request.jsp')
print(res.status_code)
print(res.text)
```

```
200



method : GET<br>
query string<br>
<br><br>
header<br>
key : accept, value : */*<br>
key : Accept-Encoding, value : gzip, deflate<br>
key : connection, value : close<br>
key : host, value : ggoreb.com<br>
key : HOSTING_CONTINENT_CODE, value : AS<br>
key : HOSTING_COUNTRY_CODE, value : KR<br>
key : HOSTING_WHITE_IP, value : false<br>
key : user-agent, value : python-requests/2.25.0<br>
key : X-Forwarded-Proto, value : http<br>
key : X-SERVER_PORT, value : 80<br>
key : X-SERVER_PROTOCOL, value : HTTP/1.1<br>
key : X-SIMPLEXI, value : 110.70.51.5<br>
key : content-length, value : 0<br>
```

■ requests

● Parameter 사용 (GET - **params**)

```
import requests
param = { 'page': 1, 'search': '검색어' }
res = requests.get(
    'http://ggoreb.com/python/request.jsp', params=param)
print(res.text)
```

```
method : GET<br>
query string<br>
key : search, value : 검색어<br>
key : page, value : 1<br>
<br><br>
header<br>
key : accept, value : */*<br>
key : Accept-Encoding, value : gzip, deflate, br<br>
key : connection, value : close<br>
key : host, value : ggoreb.com<br>
```

■ requests

● Parameter 사용 (POST - **data**)

```
import requests
param = { 'page': 1, 'search': '검색어' }
res = requests.post(
    'http://ggoreb.com/python/request.jsp', data=param)
print(res.text)
```

```
method : POST<br>
query string<br>
key : search, value : 검색어<br>
key : page, value : 1<br>
<br><br>
header<br>
key : accept, value : */*<br>
key : Accept-Encoding, value : gzip, deflate, br<br>
key : connection, value : close<br>
key : content-length, value : 41<br>
```

■ requests

● Header 사용

```
import requests
header = { 'user-agent': 'android', 'accept-language': 'en' }
res = requests.get(
    'http://ggoreb.com/python/request.jsp', headers=header)
print(res.text)
```

```
method : GET<br>
query string<br>
<br><br>
header<br>
key : accept, value : */*<br>
key : Accept-Encoding, value : gzip, deflate<br>
key : accept-language, value : en<br>
key : connection, value : close<br>
key : host, value : ggoreb.com<br>
key : HOSTING_CONTINENT_CODE, value : AS<br>
key : HOSTING_COUNTRY_CODE, value : KR<br>
key : HOSTING_WHITE_IP, value : false<br>
key : user-agent, value : android<br>
```

■ requests

● 요청 메소드 (GET / POST / PUT / DELETE)

```
import requests
res = requests.get(
    'http://ggoreb.com/http/method.jsp')
print(res.text)
```

```
<h1>Method => <span style="color:blue;">GET</span></h1>
```

```
import requests
res = requests.post(
    'http://ggoreb.com/http/method.jsp')
print(res.text)
```

```
<h1>Method => <span style="color:red;">POST</span></h1>
```

```
import requests
res = requests.put(
    'http://ggoreb.com/http/method.jsp')
print(res.text)
```

```
<h1>Method => <span style="color:green;">PUT</span></h1>
```

```
import requests
res = requests.delete(
    'http://ggoreb.com/http/method.jsp')
print(res.text)
```

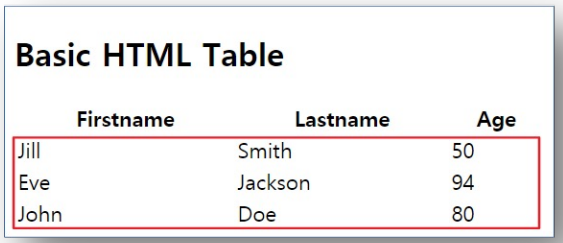
```
<h1>Method => <span style="color:yellow;">DELETE</span></h1>
```

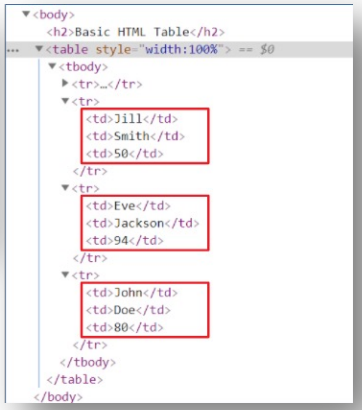
■ requests

● 데이터 추출하기

```
import requests
result = requests.get(
    'http://ggoreb.com/python/html/data1.html').text

s_idx = 0
e_idx = 0
while True:
    s_idx = result.find('<td>', e_idx)
    if s_idx == -1:
        break
    e_idx = result.find('</td>', s_idx)
    print(result[s_idx + 4 : e_idx])
```





```
Jill
Smith
50
Eve
Jackson
94
John
Doe
80
```

■ 연습문제

● 결과와 같이 span의 숫자 내용만 출력하기

주의 요함 | view-source:ggoreb.com/python/html/number.html

```
<body>
    <div class="win_result">
        <h4><strong>1000회</strong> 당첨결과</h4>
        <p class="desc">(2022년 01월 29일 추첨)</p>
        <div class="nums">
            <div class="num win">
                <strong>당첨번호</strong>
                <p>
                    <span class="ball_645 lrg ball1">2</span>
                    <span class="ball_645 lrg ball1">8</span>
                    <span class="ball_645 lrg ball2">19</span>
                    <span class="ball_645 lrg ball3">22</span>
                    <span class="ball_645 lrg ball4">32</span>
                    <span class="ball_645 lrg ball5">42</span>
                </p>
            </div>
            <div class="num bonus">
                <strong>보너스</strong>
                <p><span class="ball_645 lrg ball4">39</span></p>
            </div>
        </div>
    </div>
</body>
```


```
import requests
address = 'http://ggoreb.com/python/html/number.html'
res = requests.get(address)
res.encoding = None  # 한글 깨짐 처리
```

방법 1) <span 문자열 이후에 등장하는 "> 문자열을 찾아서 시작 인덱스로

방법 2) 정규식

■ 연습문제

● 결과와 같이 img의 src 속성 값 출력하기

주의 요함 | view-source:ggoreb.com/python/html/attribute.html

```
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8">
<title>Insert title here</title>
</head>
<body>
    <div class="info">
        <img src="/images/luffy.jpg" width="10%" title="루피">
        <img src="/images/nami.jpg" width="10%" title="나미">
        <img src="/images/robin.jpg" width="10%" title="로빈">
        <img src="/images/zoro.jpg" width="10%" title="조로">
    </div>
</body>
</html>
```

주의 요함 | ggoreb.com/python/html/attribute.html


```
import requests
address = 'http://ggoreb.com/python/html/attribute.html'
res = requests.get(address)
res.encoding = None  # 한글 깨짐 처리
```