

# 파이썬을 활용한 실전 웹크롤링

2주차 - 요청의 종류와 HTML 문서 이해

2018.07.28

신준호 jjjjooonno@gmail.com



목차 - 이론

1. 서버와 클라이언트와 브라우저의 관계도 이해

2. 웹에 데이터를 요청할 때의 두가지 방식 : GET과 POST



목차 - 실습

1. 웹 문서안의 구조를 제대로 이해하기 : CSS / HTML / JS

2. requests를 활용한 웹사이트 접근



"서버와 클라이언트와 브라우저의 관계도이해"의 목적.

서버, 클라이언트, 브라우저에 대한 이해를 통해 우리가 요청하는 데이터가 어떻게 우 리의 컴퓨터에 오게 되고, 브라우저에 어떻 게 표현되는지 알 수 있습니다.





#### 서버 (웹서버)

서버란 클라이언트(사용자)에게 **네트워크를 통해 서비스를 제공**하는 물리적인 컴퓨터를 의미합니다. 대부분의 인터넷 서비스 업체는 서버를 통해 서비스를 제공합니다. 연말정산 시즌에 국세청 홈페이지가 다운되는 것은 바로 국세청이 이용하는 서버의 가용자원을 넘어선 사용자들의 접근이 원인이 되는 경우가 많습니다. 흔히들 말하는 '서버가 다운됐다'에서 말하는 서버가 웹서버입니다.





#### 클라이언트 (사용자)



클라이언트란 **서버에 접속하기 위한 프로그램**을 의미합니다. 게임을 할때 클라이언트를 사용하여 접근하는데, 이때 게임 클라이언트는 게임 회사의 서버에 접근하여 다른 사람들과 온라인 게임을 즐기는데 사용되는 프로그램 입니다.

물론 웹서버에 접근하는 우리에게 클라이언트 프로그램은 브라우저가 됩니다.

IE(internet explorer) / Chrome / FireFox 등의 다양한 브라우저가 존재하며 각 브라우저는 사용자의 의도에 따라 웹서버에 접근할 수 있게 해줍니다.





#### 웹 브라우저

웹 브라우저란 웹 사용자가 URL등을 사용하여 웹서버에 접근할 때 사용되는 프로그램입니다. 쉽게 설명하자면, **주소를 알려주면 알아서 찾아가 택배를 받아오는** 심부름꾼으로 생각하시면 됩니다.

이러한 웹브라우저는 웹서버에서 사용자가 요청한 정보를 응답받아 실제 사용자가 보고 읽을 수 있는 이미지, 텍스트 등으로 변환해주는 역할도 수행합니다.



브라우저에 응답을 보냄

브라우저에 표현된 응답을 확인



서버에 요청을 전달

정보 요청







### "GET과 POST의 차이에 대한 이해"의 목적.

GET과 POST는 내가 원하는 정보를 한정짓는 변수가 담기는 위치가 다릅니다. 당연하게도 코드도 달라지게 됩니다. 주로 API를 사용할 때 다르게 사용되는 GET과 POST의 차이점을 학습하고 보다 완벽하게 API를 사용해 봅니다.



### GET과 POST의 차이에 대한 이해





#### GET 방식

심부름꾼에게 주소를 알려주면서 주소안에 원하는 정보의 위치 등을 포함시켜 심부름을 보내는 것

#### Ex)

http://comic.naver.com/webtoon/list.nhn?titleId=183559&weekday=mon '신의 탑'이라는 웹툰의 URL 주소입니다. titleId와 weekday라는 정보를 URL에 실어서 보내서 보냅니다.



### GET과 POST의 차이에 대한 이해





POST 방식

POST방식은 이름에서도 알 수 있듯이 무언가를 게시할 때 쓰는 방식입니다. 이 요청은 게시물 작성과 같은 요청을 할 때 수행하는데, 이때 body라는 곳에 정보를 실어서 보내게 됩니다.

Ex) 네이버 카페 글 게시 요청



"웹 문서안의 구조를 제대로 이해하기 : CSS / HTML / JS 학습"의 목적.

우리의 요청에 따라 응답받게 되는 HTML문서의 구조를 학습하여 우리가 찾게 될 정보가 담긴 형식을 이해합니다.



#### **HTML**

HyperText Markup Language의 줄임말입니다.

웹 브라우저가 서버에서 받는 정보는 HTML이라는 웹 문서로 이것을 브라우저가 해석하여 우리가 보는 실제 웹사이트로 구성합니다.

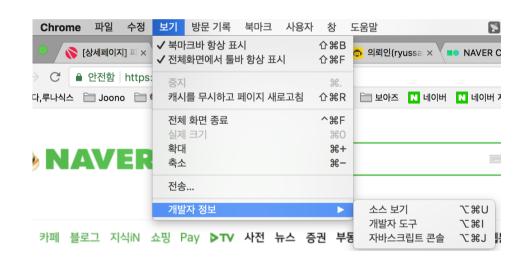
이러한 HTML 웹문서는 대부분 안에 CSS라는 언어와 JavaScript를 포함하고 있습니다. HTML은 우리가 보는 웹사이트의 제목, 본문, 버튼, 입력칸 등의 모든 **엘리먼트(element)**를 **태그(Tag)와 속성(attribute)으로 표현**하고, 그 **위치나 크기, 색상 등을 CSS언어**로 표현합니다.

또한 사용자가 URL의 변화 없이, 즉 다른 웹문서를 불러오지 않고 **같은 문서 안에서 다른 정보를** 사용자의 입력에 따라 표현하기 위해 JavaScript를 사용하기도 합니다.



```
<!doctype html>
                                             - 기본 구조
<html>
<head>
                                             <태그 속성 = "속성값">
 <title>Hello HTML</title>
</head>
<body>
                                             텍스트 내용
 Hello World!
<a href="http://example.org"><img src="image.gif"
                                             </태그>
alt="alternative text" width="50" height="50"></a>
</body>
</html>
```





#### 크롬의 개발자 도구

현재 웹사이트의 HTML 웹문서를 확인하기 위한 도구

HTML 문서의 엘리먼트들의 태그와 속성값을 확인 가능하고, 문서의 구조가 어떻게 짜여져 있는지 확인이 가능합니다.

또한 각 엘리먼트들의 CSS 스타일도 확인이 가능합니다.





```
🖫 🛅 Elements Console Sources Network Performance Memory Application Security >> 🕲 1
                                                        Styles Computed Event Listeners DOM Breakpoints
 <html lang="ko" data-useragent="Mozilla/5.0
(Macintosh; Intel Mac OS X 10_13_4) AppleWebKit/537.36 Filter
                                                                                                   :hov .cls +
(KHTML, like Gecko) Chrome/65.0.3325.181 Safari/
537.36" class="gr__news_naver_com">
                                                        element.style {
 ▶#shadow-root (open)
 ▶ <head>...</head>
                                                        .main_content_new .main_component {
                                                                                                    common.css:605
 ▼ <body class="chrome" data-gr-c-s-loaded="true">
                                                          border-bottom: ▶ 1px solid #dedede
   ▼<div id="wrap">
                                                          margin: ▶ 0 0 15px 0:
       <div id="da_base"></div>
                                                          padding-top: 9px;
      <div id="da stake"></div>
     ▶ <div id="header" class="header">...</div>
                                                        .main component {
                                                                                                    common.css:479
     ▶ <div class="home timelate">...</div>
                                                          position: relative;
     ▶ <script type="text/javascript">...</script>
                                                           margin-top: 16px;
      <div class="main space"></div>
                                                          background: ▶ □ #fff;
     ▼<div id="container" class="main">
                                                        body, div, p, h1, h2, h3, h4, h5, h6, ul,
                                                                                                    common.css:2
       ▼<div id="main_content" class="main_content
                                                        ol. li. dl. dt. dd. table. th. td. form.
                                                         fieldset, legend, input, textarea, button {
         ▶ <div class="main comp
                                                          margin: ▶ 0;
         "today main news">...</div> == $0
                                                          padding: ▶ 0;
         ▶ <div class="main_component droppable" id=
         "section politics">...</div>
         ▶ <div class="main_component droppable" id=
                                                                                            user agent stylesheet
         "section_economy">...</div>
                                                          display: block;
         ▶ <div class="main_component droppable" id=
         "section_society">...</div>
         ▶<div class="main component droppable" id=
                                                        Inherited from div#container.main
         "section life">...</div>
                                                        #container {
                                                                                                    common.css:165
         ▶ <div class="main_component droppable" id=
                                                          display: table:
         "section_world">...</div>
                                                          table-layout: fixed;
         ▶ <div class="main_component droppable" id=
                                                          width: 1080px;
         "section_it">...</div>
                                                          margin: ▶ 0 auto;
          <div id="home space" class="space" style=
                                                          text-align: left;
           "display:none"></div>
        ▶ <div id="home_phantom" class="home_move"
        style="left:500px; top:220px; display:none;"
                                                        Inherited from div#wrap
         ...</div>
                                                        #wrap {
                                                                                                    common.css:161
        ▶ <div class="airs">...</div>
                                                          margin: ▶ 0 auto:
         ::after
                                                          text-align: center;
         </div>
        <hr>
       ▶ <div class="main_aside">...</div>
                                                        Inherited from body, chrome
      </div>
                                                        body {
                                                                                                      common.css:4
      <hr>
                                                          color: ##000:
     ▶ <div class="index">...</div>
                                                          font-size: 12px;
      <hr>
                                                           text-align: center;
     ▶ <div id="footer">...</div>
                                                          background: ▶ □ #fff;
     ▶ <script type="text/javascript">...</script>
    </div>
                                                        body, table, table td, input, select,
                                                                                                      common.css:3
   </body>
                                                        textarea {
  </html>
                                                          font-size: 12px:
                                                           font-family: "Helvetica Neue", "Apple SD Gothic
                                                             Neo", "Malgun Gothic", "맑은 고딕", Dotum, "돋움", sans-
    div#today_main_news.main_component.droppable
```



```
Elements Console Sources Network Performance Memory Application Security » 01 : X
                                                         Styles Computed Event Listeners DOM Breakpoints >>>
 <html lang="ko" data-useragent="Mozilla/5.0
(Macintosh; Intel Mac OS X 10_13_4) AppleWebKit/537.36 Filter
(KHTML, like Gecko) Chrome/65.0.3325.181 Safari/
537.36" class="gr_news_naver_com">
                                                        element.stvle {
 ▶#shadow-root (open)
 ▶ <head>...</head>
                                                         .main content new .main component {
                                                                                                   common.css:605
 ▼ <body class="chrome" data-gr-c-s-loaded="true">
                                                          border-bottom: ▶ 1px solid □#dedede;
   ▼<div id="wrap">
                                                           margin: ▶ 0 0 15px 0:
       <div id="da_base"></div>
                                                          padding-top: 9px:
      <div id="da stake"></div>
     ▶ <div id="header" class="header">...</div>
                                                        .main_component {
                                                                                                   common.css:479
     ▶ <div class="home timelate">...</div>
                                                          position: relative:
     ▶ <script type="text/javascript">...</script>
                                                           margin-top: 16px:
      <div class="main space"></div>
                                                           background: ▶ □ #fff;
     ▼<div id="container" class="main">
                                                        body, div, p, h1, h2, h3, h4, h5, h6, ul,
      ▼ <div id="main_content" class="main_content
                                                         ol, li, dl, dt, dd, table, th, td, form,
       main content new">
                                                         ieldset, legend, input, textarea, button {
        "today_main_news">...</div> == $0
                                                          margin: ▶ 0:
         ▶ <div class="main_component droppable" id=
                                                          padding: ▶ 0;
         "section_politics">...</div>
         ▶ <div class="main_component droppable" id=
                                                                                            user agent stylesheet
         "section economy">...</div>
                                                          display: block;
        ▶ <div class="main_component droppable" id=
         "section society">...</div>
         ▶ <div class="main component droppable" id=
                                                         Inherited from div#container.main
         "section life">...</div>
                                                                                                   common.css:165
         ▶ <div class="main_component droppable" id=
                                                          display: table:
         "section_world">...</div>
                                                           table-layout: fixed;
         ▶ <div class="main_component droppable" id=
                                                          width: 1080px;
         "section_it">...</div>
                                                          margin: ▶ 0 auto;
          <div id="home_space" class="space" style=
                                                          text-align: left:
          "display:none"></div>
         ▶ <div id="home phantom" class="home move"
        style="left:500px; top:220px; display:none;":
                                                        Inherited from div#wrap
         ...</div>
                                                        #wrap {
                                                                                                   common.css:161
        ▶ <div class="airs">...</div>
          ::after
                                                           text-align: center;
        </div>
                                                        Inherited from body.chrome
       ▶ <div class="main aside">...</div>
      </div>
                                                        body {
                                                                                                      common.css:4
      <hr>
                                                          color: ■#000;
     ▶ <div class="index">...</div>
                                                           font-size: 12px;
                                                           text-align: center:
    ▶ <div id="footer">,,,</div>
                                                          background: ▶ ☐ #fff;
     ▶ <script type="text/javascript">...</script>
    </div>
                                                       body, table, table td, input, select,
                                                                                                      common.css:3
   </body>
                                                       textarea {
 </html>
                                                           font-size: 12px:
                                                           font-family: "Helvetica Neue", "Apple SD Gothic
                                                             Neo", "Malgun Gothic", "맑은 고딕", Dotum, "돋움", sans-
    div#today_main_news.main_component.droppable
```

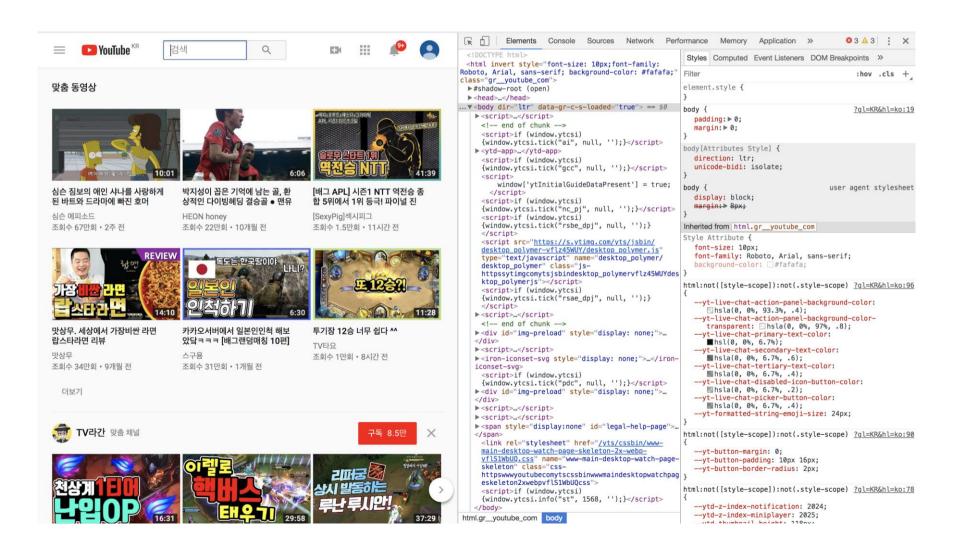
태그

속성

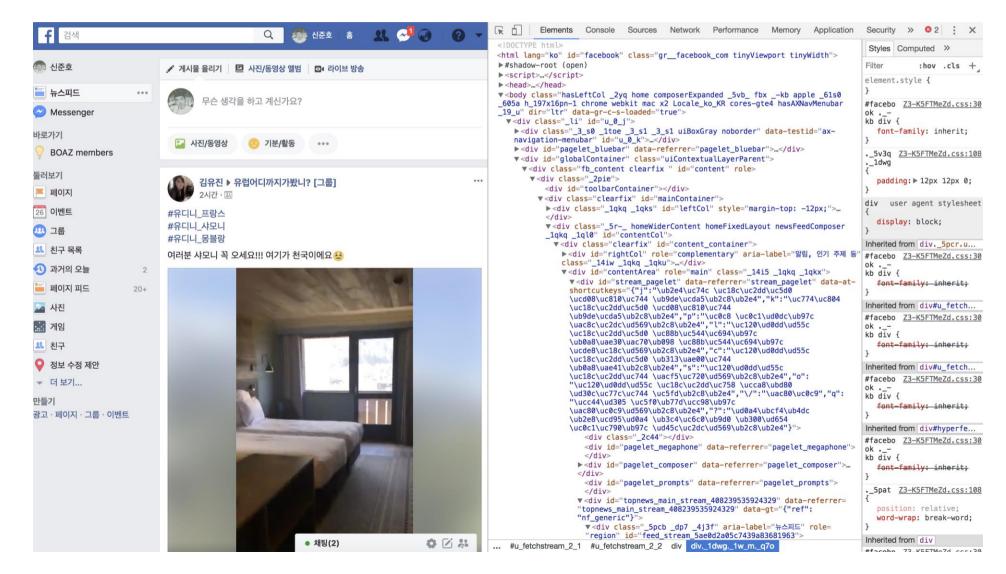
"속성값"

CSS 스타일











#### 크롤링을 할때 주로 접근하는 태그

div: 공간을 의미합니다. 엘리먼트를 다른 것과 구분하기 위해 사용합니다.

p: 단락을 의미합니다. 대부분 기사의 본문이나 댓글의 본문 등에 사용됩니다.

a : <mark>링크</mark>를 의미합니다. 대부분 **href라는 속성**을 가지고 있는데 이 속성이 링크와 연결된 사이트의 주소를 가지고 있습니다.

td: 여러개의 게시판 목록처럼 테이블과 같은 것들의 하나의 셀을 의미합니다. 주로 웹페이지의 엘리먼트를 공간을 구분지어 구성해 놓은 사이트에서 많이 사용합니다.

img : 말그대로 이미지를 나타내는 태그입니다. 이때 **src라는 속성으로 이미지의 주소**를 표현해 놓습니다. 주소를 통해 이미지에 접근이 가능합니다.

span : 그 자체로는 무의미 하지만, 내부의 속성값을 사용하여 박스공간 등을 표현할 때 사용합니다.



#### **CSS**

```
Cascading Style Sheets의 줄임말입니다.
HTML 문서의 디자인을 담당하는 언어입니다.
웹페이지의 배경색을 정하거나, 텍스트의 폰트, 폰트색 등을 정할때 사용합니다.
 font-size: 110%;
 font-family: garamond, sans-serif;
h2{
 color: red;
 background: white;
위와 같이 CSS는 중괄호로 표현하며 속성: 속성값; 으로 표현합니다.
```



#### **JavaScript**

Java와는 다른 언어입니다.

<script>라는 태그에 쌓여서 HTML 문서 안에서 작동하는 코드입니다.

HTML 웹 문서는 정적인 문서(하나의 URL에는 정해진 컨텐츠가 있어서 어떤 경우에도 바뀌지 않고 표현됨)와 동적인 문서(같은 URL로 접근해도 URL안에서 다른 정보를 표현할 수 있는 문서)가 있습니다.

동적인 문서는 웹 브라우저에서 사용자의 액션에 반응하여 이미지 / 컨텐츠를 변경하거나 액션에 맞는 반응을 표현하는 문서입니다. 이러한 동적인 문서를 구성하기 위해 사용되는 언어이며, 수업에서는 JavaScript 코드를 짜기보다는 읽고, 이해하여 웹문서의 동적인 부분을 이해하는데 초점을 맞춥니다.



"requests를 활용한 웹사이트 접근"의 목적.

실제로 Python 코드를 사용하여 웹사이트에 접근 해 응답을 Python 객체로 받아 적용시켜 보며, 우리가 원하는 element를 어떻게 찾을 수 있는지 학습합니다.



#### Requests 패키지

파이썬의 기본 내장 패키지. get방식과 post 방식 둘다 사용가능하고, 따로 웹드라이버가 필요없습니다. 하지만 실제 브라우저를 통한 접근이 아니기 때문에 크롤링bot을 막기 위한 제어장치를 가지고 있는 웹에 접근할때는 여러가지 추가 코딩이 필요하다는 단점이 있습니다.

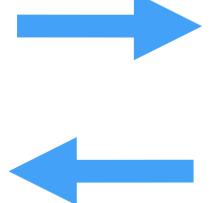
#### Selenium 패키지

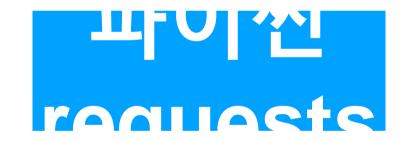
웹 개발자들이 만들어진 웹페이지를 테스트 하기위해 만들어진 패키지입니다. 웹을 코드로 작동시킬수 있기 때문에 크롤링에 사용하기에 매우 편리합니다. 하지만 post방식을 사용할때 매우 코딩이 어렵다는 단점이 있습니다. 실제 브라우저를 통해 접근하기 때문에 크롤링 bot을 막는 제어장치를 피할수 있고, 크롤링 도중에 에러가 난 경우 브라우저를 통해 에러의 원인을 살펴보고 수정하기 용이하다는 장점이 있습니다.



받은 응답을 파이썬으로 전달







정보 요청을 웹서버로 보냄



# 기본적인 문법

import requests

#GET 방식

```
response_html = response.text

#POST 방식
Data = "{key값 : value값}"
Response = requests.post('웹사이트 주소', data = data)
response_value = Response.text
# POST로 받는 것이 json이라면 Response.json()으로 받음.
```

#statue\_code는 응답이 알맞게 왔는지 확인하는 코드를 뽑아줍니다.

# https://developer.mozilla.org/ko/docs/Web/HTTP/Status

# 200은 알맞게 온 것이고, 404는 요청한 자원을 찾지 못했다는 의미입니다.

# 위 사이트에는 각종 응답코드에 대한 설명이 있고, 또한 사설 API를 사용하는 경우

response = requests.get('웹사이트 주소')

Response code = response.statue code

# API마다의 응답 코드가 존재할 수 있습니다.



```
import requests
# requests 패키지를 불러옵니다.
response = requests.get('https://www.naver.com')
# get을 사용해 naver홈페이지 정보를 요청합니다.
#이때 웹서버의 응답은 response라는 변수에 저장됩니다!
response_code = response.status_code
# 웹서버로부터 도착한 response에 status_code를 사용해
# 요청에 대한 응답이 잘 왔는지 확인합니다.
if response_code == 200:
 print(response.text)
else:
 print('error code : '+str(response_code))
```



import requests

url = "https://openapi.naver.com/v1/datalab/search"

client id = "X9S83UuJsvBu1zKJsdep"

client secret = "Vph796WJWB"

```
# naver api를 사용하기 위해 아이디와 비밀번호를 입력합니다.
body = "{\"startDate\":\"2017-01-01\",\"endDate\":\"2017-04-30\",
\"timeUnit\":\"month\",\"keywordGroups\":[{\"groupName\":\"배그\",
\"keywords\":[\"배틀그라운드\",\"pubg\",\"배그\",\"battleground\",\"카배\"]}],
\"device\":\"pc\",\"ages\":[\"1\",\"2\",\"3\",\"4\"],\"gender\":\"f\"}"
header = {"X-Naver-Client-Id": client id,
"X-Naver-Client-Secret": client secret, "Content-Type": "application/json"}
# POST요청을 보내기 위해 택배상자 안에 넣을 정보를 따로 기입해 놓습니다.
r = requests.post(url, data = body.encode(encoding='utf-8'), headers = header)
# post를 사용해 data라는 변수에 우리가 준비한 택배 내용물을 넣고 요청을 보냅니다.
rcod = r.status code
if(rcod==200):
 dc = r.ison()
 for i in dc['results'][0]['data']:
   print(i['period'], ':', i['ratio'])
else:
 print("Error Code:" + str(rcod))
# 위의 네이버 API는 키워드에 대한 검색 비율을 ison으로 응답해주는 API입니다. 따라서 응답을 .json()으로 해석하여 뽑아냅니다.
```

# 감사합니다.

