## 01. 간단한 정보 가져오기 실습

```
In [3]:
from bs4 import BeautifulSoup

In [4]:

page = open("mypage.html", 'r', encoding="utf-8").read()
page
```

### Out [4]:

### 1-1 html.parser : HTML/XHTML을 텍스트 파일을 구문분석하기 위한 html.parser 클래스

https://docs.python.org/3/library/html.parser.html (https://docs.python.org/3/library/html.parser.html)

In [5]: H

```
soup = BeautifulSoup(page, 'html.parser')
```

```
Out [5]:
<!DOCTYPE html>
<html>
<head>
<title> 나의 웹 페이지 </title>
</head>
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a><br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=D")</pre>
3CBADA9496B772E91DC06B4A289C313A4B10E2B& thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=608012845
081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+filte
rui%3alicense-L2 L3 L4 L5 L6 L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
</body>
</html>
```

In [6]:

```
print(soup.prettify())
```

```
<!DOCTYPE html>
<html>
<head>
 <title>
  나의 웹 페이지
 </title>
</head>
<body>
 >
  나의 웹 페이지 내용1
 >
  나의 웹 페이지 내용2
 >
  나의 웹 페이지 내용3
 </n>
 웹 페이지 내용 4
 \langle br/ \rangle
 웹 페이지 내용 5
 <a href="https://www.naver.com/" target=" blank">
  네이버 페이지 연결하기
 </a>
 \langle br/ \rangle
 <a href="http://www.yes24.com/Main/default.aspx" target="_blank">
  yes24
 </a>
 \langle br/ \rangle
 <a>
  다음 페이지 연결하기
 </a>
 <br/>
 <a>
  카카오 페이지 연결하기
 </a>
 <br/>
 =D3CBADA9496B772E91DC06B4A289C313A4B10E2B&thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&exph=1137&expw=1820&q=image&simid=6080128
45081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&gft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
  <img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
 </a>
</body>
</html>
```

In [7]: ▶

soup.children

#### Out [7]:

<list\_iterator at 0x250bc71c730>

In [8]: ▶

list(soup.children)

### Out[8]:

```
['html',
 '₩n',
<html>
<head>
<title> 나의 웹 페이지 </title>
</head>
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a></br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a><br/>br/>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=</pre>
D3CBADA9496B772E91DC06B4A289C313A4B10E2B&thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=6080128
45081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
</body>
</html>
```

In [11]:

```
tmp = list(soup.children)[2]
tmp
```

#### Out[11]:

```
<html>
<head>
<title> 나의 웹 페이지 </title>
</head>
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a><br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=D</pre>
3CBADA9496B772E91DC06B4A289C313A4B10E2B& thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&
mediaur I=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=608012845
081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
</body>
</html>
```

In [12]: ▶

```
list(tmp.children)
```

```
Out[12]:
```

```
['₩n'.
<head>
<title> 나의 웹 페이지 </title>
</head>,
 '₩n',
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
  나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a><br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a><br/>br/>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=</pre>
D3CBADA9496B772E91DC06B4A289C313A4B10E2B& thid=0IP.bwogiAQLFQcauPBSh18rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=6080128
45081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
</body>,
 '₩n'l
```

## body 부분 정보 얻기

```
In [13]:
                                                                                      H
Content_Body = soup.body
Content_Body
Out [13]:
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a><br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a><br/>br/>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=D</pre>
3CBADA9496B772E91DC06B4A289C313A4B10E2B& thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&
mediaur I=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=608012845
081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
</body>
In [14]:
                                                                                      H
Content_Body = list(tmp.children)[3]
Content_Body
Out [14]:
<body>
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3 
       웹 페이지 내용 4<br/>
       웹 페이지 내용 5<br/>
<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a><br/>
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a>
<a>다음 페이지 연결하기</a><br/>
<a>카카오 페이지 연결하기</a><br/>
<a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=D
3CBADA9496B772E91DC06B4A289C313A4B10E2B& thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&exph=1137&expw=1820&q=image&simid=608012845
081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&qft=+filte
rui%3alicense-L2 L3 L4 L5 L6 L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="0IP1.jpg" width="400"/>
</a>
```

# 1-2 find\_all, find을 이용한 태그를 이용한 찾기

</body>

```
H
In [15]:
soup.find_all('p')
Out[15]:
[ 나의 웹 페이지 내용1 ,  나의 웹 페이지 내용2 ,  나의 웹 페이지 내용2 
용3 ]
In [16]:
                                                                             M
soup.find('p')
Out[16]:
 나의 웹 페이지 내용1 
In [17]:
                                                                             H
soup.find('div')
In [18]:
                                                                             H
soup.find_all('p', class_='p3')
Out[18]:
[]
In [19]:
                                                                             M
soup.find_all(id='p4')
Out[19]:
[]
1-3 p 태그의 문자들만 가져오기
In [20]:
                                                                             H
for ptag in soup.find_all('p'):
   print(ptag)
 나의 웹 페이지 내용1 
 나의 웹 페이지 내용2 
 나의 웹 페이지 내용3
```

```
In [21]:
for ptag in soup.find_all('p'):
   print(ptag.get_text())
 나의 웹 페이지 내용1
 나의 웹 페이지 내용2
 나의 웹 페이지 내용3
1-4 링크 가져오기(link)
In [22]:
soup.find_all('a')
Out [22]:
[<a href="https://www.naver.com/" target="_blank">네이버 페이지 연결하기</a>,
<a href="http://www.yes24.com/Main/default.aspx" target="_blank">yes24</a>,
<a>다음 페이지 연결하기</a>,
<a>카카오 페이지 연결하기</a>.
 <a href="https://www.bing.com/images/search?view=detailV2&amp;ccid=bwogiAQL&amp;id=</pre>
D3CBADA9496B772E91DC06B4A289C313A4B10E2B&thid=01P.bwogiAQLFQcauPBSh18rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&exph=1137&expw=1820&g=image&simid=6080128
45081691527&ck=6E87BD0A98028EE049D14662D202C61F&selectedIndex=5&gft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&FORM=IRPRST&ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="이P1.jpg" width="400"/>
</a>]
In [23]:
links = soup.find_all('a')
print(links[1]['href'])
print(links[1].string)
http://www.yes24.com/Main/default.aspx (http://www.yes24.com/Main/default.aspx)
ves24
In [ ]:
                                                                                          H
In [ ]:
```