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]:

'<!DOCTYPE html>₩n<html>₩n₩t<head>₩n₩t<title> 나의 웹 페이지 </title>₩n₩t</head>₩t₩n₩t₩n₩t<body>₩t₩n₩t 나의 웹 페이지 내용1 ₩n₩t 나의 웹 페이지 내용2 ₩n₩t 나의 웹 페이지 내용3 ₩n₩t웹 페이지 내용 4
₩n₩t웹 페이지 내용 5
₩n₩t네이버 페이지 연결하기
₩n₩tyes24
₩n₩t<a>다음 페이지 연결하기
₩n₩t<a>카카오 페이지 연결하기
₩n₩t ₩n₩t₩t<img src="OIP1.jpg" alt="이미지가 없어요." width="400" height="400">₩n₩t₩n₩t</body>₩n</html>'

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]:

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

Out[5]:

```
<!DOCTYPE html>

<html>
<head>
<title> 나의 웹 페이지 </title>
</head>
<body>
<p> 나의 웹 페이지 내용1 </p>
<p> 나의 웹 페이지 내용2 </p>
<p> 나의 웹 페이지 내용3 </p>
        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/>
<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&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&amp;
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=608012845
0816915 27&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="OIP1.jpg" width="400"/>
</a>
</body>
</html>
```

In [6]:

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

```
<!DOCTYPE html>
<html>
 <head>
  <title>
   나의 웹 페이지
  </title>
 </head>
 <body>
  <p>
   나의 웹 페이지 내용1
  </p>
  <p>
   나의 웹 페이지 내용2
  </p>
  <p>
   나의 웹 페이지 내용3
  </p>
  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/>
  <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
=D3CBADA9496B772E91DC06B4A289C313A4B10E2B&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=6080128
45081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
   <img alt="이미지가 없어요." height="400" src="OIP1.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>
 <p> 나의 웹 페이지 내용1 </p>
 <p> 나의 웹 페이지 내용2 </p>
 <p> 나의 웹 페이지 내용3 </p>
        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/>
 <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=
D3CBADA9496B772E91DC06B4A289C313A4B10E2B&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=6080128
45081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
 <img alt="이미지가 없어요." height="400" src="OIP1.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>
<p> 나의 웹 페이지 내용1 </p>
<p> 나의 웹 페이지 내용2 </p>
<p> 나의 웹 페이지 내용3 </p>
        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/>
<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&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&amp;
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=608012845
081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="OIP1.jpg" width="400"/>
</a>
</body>
</html>
```

In [12]:

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

Out[12]:

```
['₩n',
 <head>
 <title> 나의 웹 페이지 </title>
 </head>,
 '₩n',
 <body>
 <p> 나의 웹 페이지 내용1 </p>
 <p> 나의 웹 페이지 내용2 </p>
 <p> 나의 웹 페이지 내용3 </p>
        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/>
 <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=
D3CBADA9496B772E91DC06B4A289C313A4B10E2B&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=6080128
45081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
 <img alt="이미지가 없어요." height="400" src="OIP1.jpg" width="400"/>
 </a>
 </body>,
 '₩n']
```

body 부분 정보 얻기

In [13]:

```
Content_Body = soup.body
Content_Body
```

Out[13]:

```
<body>
<p> 나의 웹 페이지 내용1 </p>
<p> 나의 웹 페이지 내용2 </p>
<p> 나의 웹 페이지 내용3 </p>
        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/>
<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&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&amp;
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=608012845
081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="OIP1.jpg" width="400"/>
</a>
</body>
```

In [14]:

```
Content_Body = list(tmp.children)[3]
Content_Body
```

Out[14]:

```
<body>
<p> 나의 웹 페이지 내용1 </p>
<p> 나의 웹 페이지 내용2 </p>
<p> 나의 웹 페이지 내용3 </p>
        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/>
<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&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&amp;
mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-tele
charger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=608012845
081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+filte
rui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
<img alt="이미지가 없어요." height="400" src="OIP1.jpg" width="400"/>
</a>
</body>
```

1-2 find_all, find을 이용한 태그를 이용한 찾기

In [15]:

```
soup.find_all('p')
```

Out[15]:

```
[<p> 나의 웹 페이지 내용1 </p>, <p> 나의 웹 페이지 내용2 </p>, <p> 나의 웹 페이지 내
용3 </p>]
```

In [16]:

```
soup.find('p')
```

Out[16]:

```
<p> 나의 웹 페이지 내용1 </p>
```

In [17]:

```
soup.find('div')
```

In [18]:

```
soup.find_all('p', class_='p3')
```

Out[18]:

```
[]
```

In [19]:

```
soup.find_all(id='p4')
```

Out[19]:

```
[]
```

1-3 p 태그의 문자들만 가져오기

In [20]:

```
for ptag in soup.find_all('p'):
    print(ptag)
```

```
<p> 나의 웹 페이지 내용1 </p>
<p> 나의 웹 페이지 내용2 </p>
<p> 나의 웹 페이지 내용3 </p>
```

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=
D3CBADA9496B772E91DC06B4A289C313A4B10E2B&amp;thid=OIP.bwogiAQLFQcauPBShl8rMgHaEo&am
p;mediaurl=https%3a%2f%2fstileex.xyz%2fwp-content%2fuploads%2f2019%2f01%2fimage-a-te
lecharger-gratuitement.jpg&amp;exph=1137&amp;expw=1820&amp;q=image&amp;simid=6080128
45081691527&amp;ck=6E87BD0A98028EE049D14662D202C61F&amp;selectedIndex=5&amp;qft=+fil
terui%3alicense-L2_L3_L4_L5_L6_L7&amp;FORM=IRPRST&amp;ajaxhist=0">
 <img alt="이미지가 없어요." height="400" src="OIP1.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)
yes24
```

In [ ]:

In [ ]: