

# 웹 데이터 수집 - BeautifulSoup 실습

## 학습 목표

- 실제 웹 사이트의 정보를 가져오는 것을 실습해 본다.

## 학습 내용

- 간단한 웹 페이지의 정보를 가져와보기

## 목차

01. 웹 페이지의 정보 가져오기
02. 여러개의 정보를 가져오기
03. 연결된 링크(URL) 정보를 가져오기


## 01. 웹 페이지의 정보 가져오기

목차로 이동하기

- <https://pythonstart.github.io/web/>

```
In [1]: from urllib.request import urlopen
        from bs4 import BeautifulSoup
```

## 웹 페이지 화면

 No description has been provided for this image

## 정보 가져오기 단계

- 웹의 URL 지정
- urlopen() 함수를 이용하여 URL로 HTTP 요청을 보내고 서버의 응답을 받는다.
- page는 서버의 응답 객체(HTML 내용 포함)
- BeautifulSoup()은 HTML이나 XML을 파싱하는 Python 라이브러리.
- 정리 :웹의 정보를 가져와서 이를 구조화하고, 여기에서 title을 가져온다.

```
In [4]: ## 정보 가져오기
        url = "https://pythonstart.github.io/web/"
        page = urlopen(url)
        soup = BeautifulSoup(page, 'lxml')
        soup.title
```

Out[4]: <title>나의 웹 페이지</title>

## 02. 여러개의 정보를 가져오기

목차로 이동하기

```
In [5]: content = []
link_content = soup.find_all("a")
for one in link_content:
    # print(one)
    content.append(one.text)

print("가져온 내용 : ", content)
```

가져온 내용 : [' 01. 제목 가져오기(title) ', ' 02. 텍스트 가져오기(p) ', ' 03. 링크 가져오기(a) ', ' 04. 이미지 정보 가져오기(img) ', ' 05. 리스트 정보 가져오기(ul,ol) ', ' 06. id를 활용한 정보 획득 ', ' 07. class를 활용한 정보 획득 ', ' 08. 하나의 이미지 다운로드 ', ' 09. 여러개의 이미지 다운로드 ', ' 10. 랭킹 정보 가져오기(웹 크롤링) ', ' 11. 네이버 뉴스 가져오기(1) ', ' 12. 네이버 뉴스 가져오기(1) ']

(추가 실습) 공백을 없애서 가져오기

## 03. 연결된 링크(URL) 정보를 가져오기

목차로 이동하기

```
In [13]: content = []
link_inf = []
link_content = soup.find_all("a")
for one in link_content:
    # print(one)
    content.append(one.text)
    link_inf.append(one['href'])

print("가져온 내용 : ", content)
print("가져온 링크 정보 : ", link_inf)
```

가져온 내용 : [' 01. 제목 가져오기(title) ', ' 02. 텍스트 가져오기(p) ', ' 03. 링크 가져오기(a) ', ' 04. 이미지 정보 가져오기(img) ', ' 05. 리스트 정보 가져오기(ul,ol) ', ' 06. id를 활용한 정보 획득 ', ' 07. class를 활용한 정보 획득 ', ' 08. 하나의 이미지 다운로드 ', ' 09. 여러개의 이미지 다운로드 ', ' 10. 랭킹 정보 가져오기(웹 크롤링) ']

가져온 링크 정보 : ['./01.html', './02.html', './03.html', 'https://pythonstart.github.io/web/04.html', './05.html', './06.html', './07.html', './08.html', 'https://pythonstart.github.io/web/09.html', './10.html']

```
In [15]: content = []
link_inf = []
link_content = soup.find_all("a")
for one in link_content:
    # print(one)
    content.append(one.text)
    link_one = one['href'].replace("./", "") # 특수 문자 제거
    link_all = url + "/" + link_one
    link_inf.append(link_all)
```

```
print("가져온 내용 : ", content)
print("가져온 링크 정보 : ", link_inf)
```

가져온 내용 : [' 01. 제목 가져오기(title) ', ' 02. 텍스트 가져오기(p) ', ' 03. 링크 가져오기(a) ', ' 04. 이미지 정보 가져오기(img) ', ' 05. 리스트 정보 가져오기(ul,ol) ', ' 06. id를 활용한 정보 획득 ', ' 07. class를 활용한 정보 획득 ', ' 08. 하나의 이미지 다운로드 ', ' 09. 여러개의 이미지 다운로드 ', ' 10. 랭킹 정보 가져오기(웹 크롤링) ']

가져온 링크 정보 : ['https://pythonstart.github.io/web//01.html', 'https://pythonstart.github.io/web//02.html', 'https://pythonstart.github.io/web//03.html', 'https://pythonstart.github.io/web//https://pythonstart.github.io/web/04.html', 'https://pythonstart.github.io/web//05.html', 'https://pythonstart.github.io/web//06.html', 'https://pythonstart.github.io/web//07.html', 'https://pythonstart.github.io/web//08.html', 'https://pythonstart.github.io/web//https://pythonstart.github.io/web/09.html', 'https://pythonstart.github.io/web//10.html']

## 실습 과제

- 현재 정보 가져오기에서 추가적으로 링크를 타고 들어가서 정보가져오기