# 파이썬 웹 크롤링

BeautifulSoup - 파이썬 웹 크롤링 라이브러리
BeautifulSoup은 파이썬 웹 크롤링에 가장 널리 사용되는 라이브러리이자 툴 입니다.

웹 크롤링(Web crawling) 또는 스크래핑(Scrapping)은
웹 페이지들을 긁어와서 데이터를 추출하는 것을 말합니다.

웹 크롤러는 자동화된 방식으로 웹 페이지들을 탐색하는 컴퓨터 프로그램입니다.

파이썬과 BeautifulSoup 라이브러리를 이용하면
프로그래밍에 익숙하지 않은 비전공자나 입문자도 쉽게 크롤링을 할 수 있습니다.

BeautifulSoup 크롤링 예제에서 Requests와 BeautifulSoup 라이브러리를 사용하는데,
기본적으로 아나콘다 통합 패키지에 포함되어 있지만
설치되어 있지 않다면 설치를 진행합니다.

```
# Requests 설치
> pip install requests

# BeutifulSoup 설치
> pip install beautifulsoup4
```

## 네이버 날씨 미세먼지 가져오기

```
# 웹페이지 가져오기
# 'https://search.naver.com/search.naver?query=날씨'

from bs4 import BeautifulSoup as bs
import requests

html = requests.get('https://search.naver.com/search.naver?query=날씨')
print(html.text)
```

파싱

```
soup = bs(html.text, 'html.parser')
print(soup)
```

요소 1개 찾기 (find)
미세먼지 정보가 있는 div 요소만 추출

```
# 요소 1개 찾기 (find)
# 미세먼지 정보가 있는 div 요소만 추출
data1 = soup.find('div', {'class':'detail_box'})
print(data1)
```

**요소 모두 찾기**

find와 사용방법이 똑같으나
find는 처음 매칭된 1개만,
fildAll은 매칭된 모든 것을 리스트로 반환

```
data2 = data1.findAll('dd')
print(data2)
```

내부 텍스트만 골라내도록 .text를 이용

```
fine_dust = data2[0].find('span',{'class':'num'}).text
print(fine_dust)
```

내부 텍스트 추출
span태그에 속성과 class = :num

```
fine_dust = data2[0].find('span',{'class':'num'})
print(fine_dust)
```

내부 텍스트만 골라내도록 .text를 이용

```
# 내부 텍스트만 골라내도록 .text를 이용
fine_dust = data2[0].find('span',{'class':'num'}).text
print(fine_dust)
```

- 초미세먼지 추출

```
ultra_fine_dust = data2[1].find('span',{'class':'num'}).text
print(ultra_fine_dust)
```

> data2 변수에서
> 미세먼지는 0번 인덱스
> 초미세먼지는 1번 인덱스

- 오존지수 추출

```
ozone = data2[2].find('span',{'class':'num'}).text
print(ozone)
```

> 오존지수는 2번 인덱스