Python 데이터 분석과 이미지 처리

나동빈

웹 크롤링

Web Crawler

- 웹 크롤러란 자동화된 방법으로 웹(Web)에서 다양한 정보를 수집하는 소프트웨어입니다.
- 원하는 서비스에서 원하는 정보를 편하게 얻어올 수 있습니다.
- 언어를 막론하고 구현할 수 있지만, 주로 Python을 이용합니다.

웹 크롤링

특정 웹 사이트 HTML 코드 추출 ①

```
import requests

# 특정 URL에 접속하는 요청(Request) 객체를 생성합니다.
request = requests.get('http://www.dowellcomputer.com/main.jsp')

# 접속한 이후의 웹 사이트 소스코드를 추출합니다.
html = request.text

print(html)
```

웹 크롤링

특정 웹 사이트 HTML 코드 추출 ②

```
import requests
from bs4 import BeautifulSoup

# 특정 URL에 접속하는 요청(Request) 객체를 생성합니다.
request = requests.get('http://www.dowellcomputer.com/main.jsp')
# 접속한 이후의 웹 사이트 소스코드를 추출합니다.
html = request.text
# HTML 소스코드를 파이썬 BeatifulSoup 객체로 변환합니다.
soup = BeautifulSoup(html, 'html.parser')

# <a> 태그를 포함하는 요소를 추출합니다.
links = soup.select('td > a')

# 모든 링크에 하나씩 접근합니다.
for link in links:
    # 링크가 href 속성을 가지고 있다면
    if link.has_attr('href'):
        # href 속성의 값으로 notice라는 문자가 포함되어 있다면
        if link.get('href').find('notice') != -1:
            print(link.text)
```