파이썬 웹 크롤링2

파이썬 웹 크롤링 응용하기

네이버 웹툰 크롤링

```
from bs4 import BeautifulSoup
import requests
```

> 필요한 모듈 임포트

```
# 웹페이지를 열고 소스코드를 읽어오는 작업
html = requests.get("https://comic.naver.com/webtoon/weekday.nhn")
soup = BeautifulSoup(html.text, 'html.parser')
html.close()

data1_list = soup.findAll('div', {'class':'col_inner'})
print(data1_list)
```

> 웹페이지를 열고 소스코드를 읽어오는 작업

```
# 요일병 웹툰 영역중 제목과 썸네일 영역을 하나의 리스트로
li_list = []

for data1 in data1_list:
    # 제목+썸네일 영역 추출
    # 해당 부분을 찾아 li_list와 병합
    li_list.extend(data1.findAll('li'))

print(li_list)
```

> 요일별 웹툰 영역중 제목과 썸네일 영역을 하나의 리스트로

```
# 각각의 요소 중 <img> 태그의 제목과 썸네일(~.jpg)만 추출하기
for li in li_list:
    img = li.find('img')
    title = img['title']
    img_src = img['src']
    print(title, img_src)
```

> 각각의 요소 중 태그의 제목과 썸네일(~.jpg)만 추출하기

썸네일 추출 및 다운로드

- **다운로드 하기**
  이미지 또는 동영상 링크가 있다면 다운로드 하는 방법은 쉽다.
  from urllib.request import urlretrieve 를 추가한 뒤,
  urlretrieve 호출 시에 링크와 저장 할 파일명을 넣으면 된다.

- **특수문자 처리**
  도중에 에러가 난 부분을 보면 파일명에 특수문자가 있는 경우,
  따라서 추출한 제목에서 특수문자는 다른 문자로 변경해주거나 삭제.

  변경은 replace를 하면 되는데,
  여기서는 정규식 표현을 이용한 re모듈을 사용하여 삭제.
  따라서 re모듈을 import

- **저장폴더 생성**
  여기서는 os 모듈을 참조
  os.path.isdir : 이미 디렉토리가 있는지 검사
  os.path.join : 현재 경로를 계산하여 입력을 ㅗ들어온 텍스트를 합하여 새로운 경로를 만듦

  os.makedirs : 입력으로 들어온 경로로 폴더를 생성

  모듈 참조와 아래 urlretrieve 부분도 변경

```
import errno
from bs4 import BeautifulSoup

import requests, re, os
from urllib.request import urlretrieve # 추가
```

> 필요한 모듈 import

웹페이지를 열고 소스 코드를 읽어오는 작업

```
# 웹페이지를 열고 소스 코드를 읽어오는 작업
html = requests.get("https://comic.naver.com/webtoon/weekday.nhn")
soup = BeautifulSoup(html.text, 'html.parser')
html.close()
```

요일별 웹툰 영역 추출하기

```
# 요일별 웹툰 영역 추출하기
data1_list = soup.findAll('div',{'class':'col_inner'})
print(data1_list)
```

전체 웹툰 리스트

```
# 전체 웹툰 리스트
li_list = []

for data1 in data1_list:
    # 제목+썸네일 영역 추출
    # 해당 부분을 찾아 li_list와 병합
    li_list.extend(data1.findAll('li'))

print(li_list)
```

각각 썸네일과 제목 추출하기

```
# 각각 썸네일과 제목 추출하기
for li in li_list:
    img = li.find('img')
    title = img['title']
    img_src = img['src']

    # 해당 영역의 글자가 아닌것은 ''로 치환시킨다.
    title = re.sub('[^0-9a-zA-Zㄱ-힗]', '', title)
```

```
# 주소, 파일경로 + 파일명 + 확장자
urlretrieve(img_src, './image/' + title + '.jpg')
```