-BeautifulSoup를 이용한 파이썬 웹 크롤링

안녕하세요. 소프트웨어학과 조희진입니다.

이번 튜토리얼에서는 BeautifulSoup를 이용한 파이썬 웹 크롤링에 대해 알아보도록 하겠습니다.

1. 크롤링의 정의

웹 크롤링 혹은 웹 스크랩핑이란 컴퓨터 소프트웨어 기술로 웹 사이트들에서 원하는 정보를 추축하는 것을 의미합니다.

2. 사용할 오픈소스 소개

BeautifulSoup는 HTML과 XML 파일로부터 데이터를 뽑아내기 위한 파이썬 오픈소스 라이브러리입니다.

공식사이트 URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

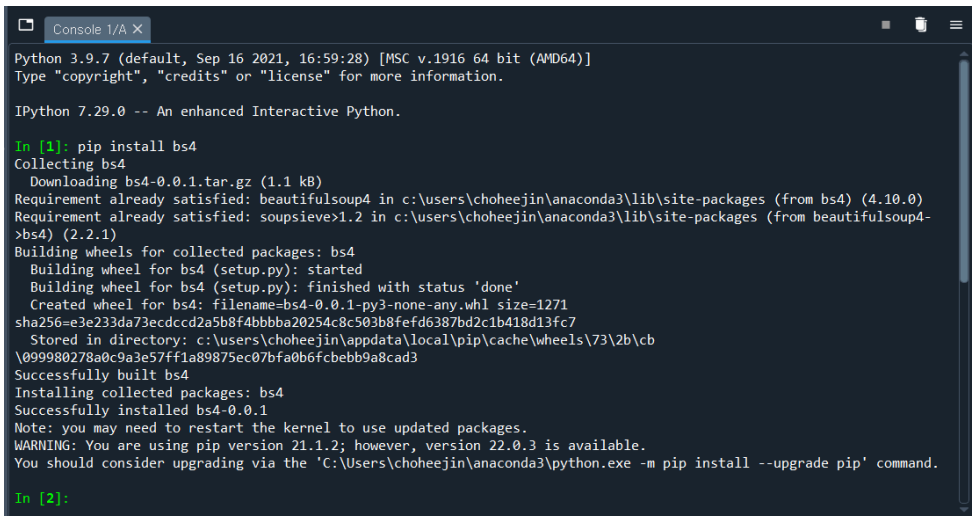
3. 설치

1) 아나콘다(Anaconda) 설치

2) Syder 열기

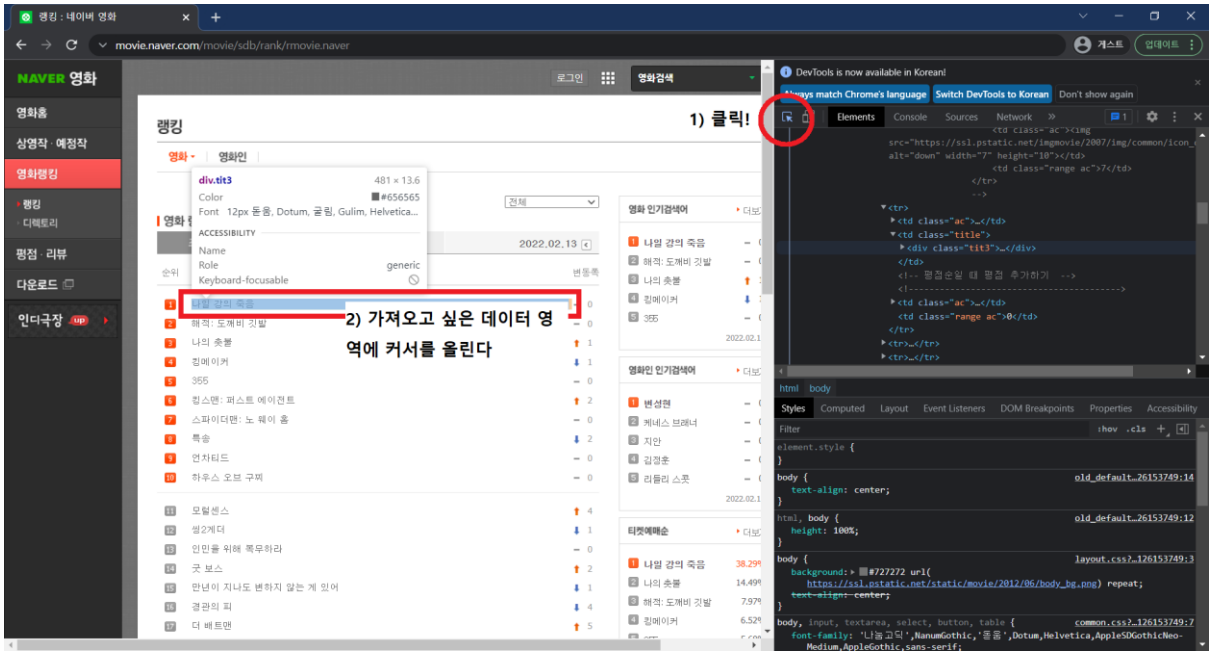
3) BeautifulSoup install

콘솔창에 **pip install bs4** 입력하여 BeautifulSoups 설치하기



4. 크롤링하고자 하는 웹 페이지의 구조 파악하기

이번에 크롤링하고자 하는 웹 페이지는 네이버 영화의 영화 랭킹입니다. 크롬 브라우저로 https://movie.naver.com/movie/sdb/rank/rmovie.naver 에 접속하여 F12 키를 눌러 개발자 도구를 열어줍니다.

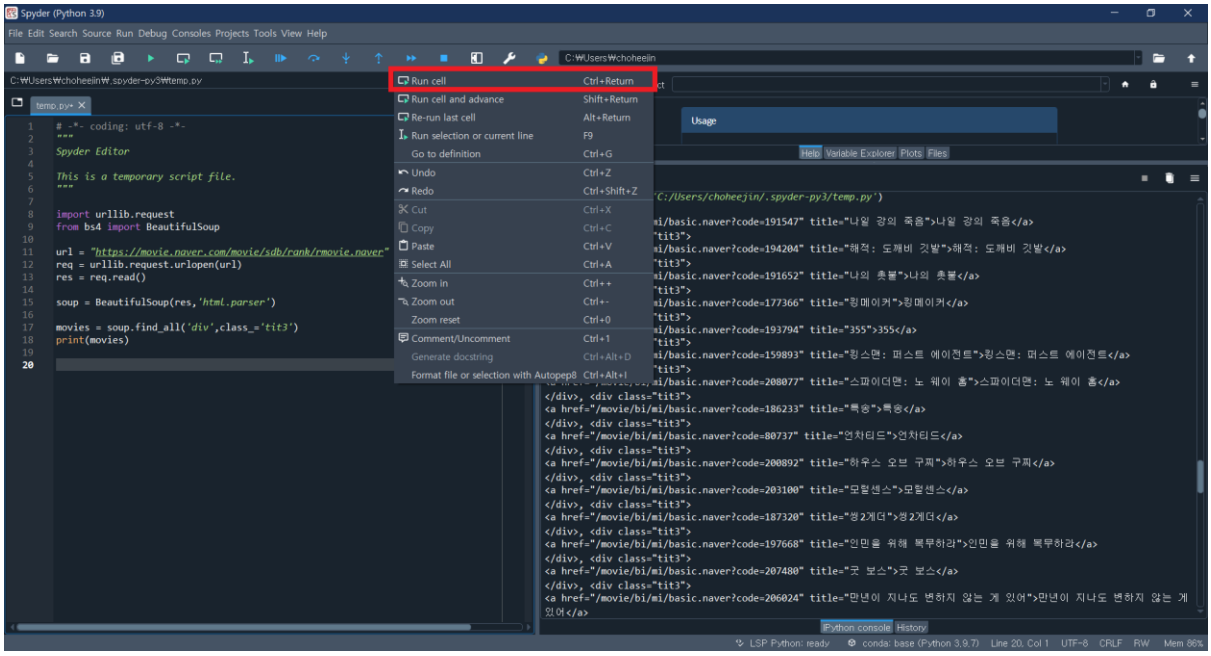


첫 번째 사진대로 따라하시면, 영화 조회순 랭킹이 div 태그에 tit3 클래스를 사용하는 것을 확인할 수 있습니다.

5. 실습

```python
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""

import urllib.request
from bs4 import BeautifulSoup

url = "https://movie.naver.com/movie/sdb/rank/rmovie.naver"
req = urllib.request.urlopen(url)
res = req.read()

soup = BeautifulSoup(res,'html.parser')

movies = soup.find_all('div',class_='tit3')
print(movies)
```

url 변수에 네이버 영화 랭킹 주소를 넣고, 해당 조건에 맞는 모든 태그들을 가져오는 **find_all()** 함수를 이용하여, 클래스 이름이 tit3인 div 태그를 모두 가져옵니다.



마우스 우클릭을 하여 [Run cell]을 눌러 오른쪽 콘솔창에서 결과를 확인합니다. 결과를 보시면, html 태그들이 리스트 형태로 반환된 것을 확인하실 수 있습니다. 여기서, 원하는 데이터인 영화 이름들만 리스트로 뽑아내 보도록 하겠습니다.

```
7
8   import urllib.request
9   from bs4 import BeautifulSoup
10
11  url = "https://movie.naver.com/movie/sdb/rank/rmovie.naver"
12  req = urllib.request.urlopen(url)
13  res = req.read()
14
15  soup = BeautifulSoup(res,'html.parser')
16  movies = soup.find_all('div',class_='tit3')
17  result = []
18
19  for movie in movies:
20      result.append(movie.get_text().strip())
21
22  print(result)
```

현재 HTML 문서의 모든 텍스트를 추출할 수 있는 **get_text()** 함수를 통해 영화 이름들을 모두 뽑아낸 뒤, 문자열의 양쪽 끝에 있는 공백을 제거해주는 **strip()** 함수를 사용하여, 깔끔하게 데이터를 정리합니다.

- strip() 함수 사용

```
In [7]: runcell(0, 'C:/Users/choheejin/.spyder-py3/temp.py')
['나일 강의 죽음', '해적: 도깨비 깃발', '나의 촛불', '킹메이커', '355', '킹스맨: 퍼스트 에이전트', '스파이더맨: 노 웨이 홈', '특송', '언차티드', '하우스 오브 구찌', '모럴센스', '씽2게더', '인민을 위해 복무하라', '굿 보스', '만년이 지나도 변하지 않는 게 있어', '경관의 피', '더 배트맨', '극장판 주술회전 0', '듄', '미싱타는 여자들', '셰터드', '극장판 안녕 자두야: 제주도의 비밀', '리코리쉬 피자', '어나더 라운드', '드라이브 마이 카', '안테벨룸', '가슴이 떨리는 건 너 때문', '이상한 나라의 수학자', '대한민국 대통령', '스크림', '더 마더', '비틀즈 겟 백: 루프탑 콘서트', '나이트메어 앨리', '오리엔트 특급 살인', '장르만 로맨스', '문폴', '리프레쉬', '프랑스', '코로나', '피그', '시크릿 카운터', '하드코어 로맨스', '해탄적일천', '장민호 드라마 최종회', '강릉', '코만도', '저수지의 피크닉', '신들의 분노', '쥬라기 월드: 도미니언', '킹스맨 : 시크릿 에이전트']
```

- strip() 함수 미사용

```
In [8]: runcell(0, 'C:/Users/choheejin/.spyder-py3/temp.py')
['\n나일 강의 죽음\n', '\n해적: 도깨비 깃발\n', '\n나의 촛불\n', '\n킹메이커\n', '\n355\n', '\n킹스맨: 퍼스트 에이전트\n', '\n스파이더맨: 노 웨이 홈\n', '\n특송\n', '\n언차티드\n', '\n하우스 오브 구찌\n', '\n모럴센스\n', '\n씽2게더\n', '\n인민을 위해 복무하라\n', '\n굿 보스\n', '\n만년이 지나도 변하지 않는 게 있어\n', '\n경관의 피\n', '\n더 배트맨\n', '\n극장판 주술회전 0\n', '\n듄\n', '\n미싱타는 여자들\n', '\n셰터드\n', '\n극장판 안녕 자두야: 제주도의 비밀\n', '\n리코리쉬 피자\n', '\n어나더 라운드\n', '\n드라이브 마이 카\n', '\n안테벨룸\n', '\n가슴이 떨리는 건 너 때문\n', '\n이상한 나라의 수학자\n', '\n대한민국 대통령\n', '\n스크림\n', '\n더 마더\n', '\n비틀즈 겟 백: 루프탑 콘서트\n', '\n나이트메어 앨리\n', '\n오리엔트 특급 살인\n', '\n장르만 로맨스\n', '\n문폴\n', '\n리프레쉬\n', '\n프랑스\n', '\n코로나\n', '\n피그\n', '\n시크릿 카운터\n', '\n하드코어 로맨스\n', '\n해탄적일천\n', '\n장민호 드라마 최종회\n', '\n강릉\n', '\n코만도\n', '\n저수지의 피크닉\n', '\n신들의 분노\n', '\n쥬라기 월드: 도미니언\n', '\n킹스맨 : 시크릿 에이전트\n']
```

이렇게 조회수가 1위~50위인 영화를 크롤링해봤습니다. 위와 같은 방법으로 조회수뿐만 아니라, 평점순으로 영화 순위를 크롤링 해보시길 바랍니다.

이상으로, BeautifulSoup를 이용한 파이썬 웹 크롤링 튜토리얼을 마치도록 하겠습니다.

감사합니다.