Python Class *5*

BeautifulSoup

이번 수업에서는 `BeautifulSoup` 을 다루어 보도록 하겠습니다.

BeautifulSoup은 html 코드를 Python이 이해하는 객체 구조로 변환하는 Parsing을 맡고 있고, 이 라이브러리를 이용해 우리는 제대로 된 '의미있는' 정보를 추출해 낼 수 있습니다. 간단히 말하해서 BeautifulSoup이란 HTML과 XML 파일들로 부터 데이터를 끄집어 내는 파이썬 라이브러리입니다.

BeautifulSoup을 직접 import하는 방법 보단 `bs4` 라는 wrapper를 통해 import하는 방식이 더 쉽고 안전합니다. 따라서 아래와 같은 코드를 통해 환경설정을 합니다.

```
import urllib.request
from bs4 import BeautifulSoup
```

urllib.request.urlopen()

다루게 될 html페이지를 아래 코드처럼 지정해 줍니다.

```
html = urllib.request.urlopen("http://movie.naver.com/movie/sdb/rank/rmovie.nhn")
```

BeautifulSoup(html, 'html.parser')

아래 코드는 우리가 설정한 html페이지를 BeautifulSoup의 html 파서로 파싱하여 soup라는 객체에 할당합니다.

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

파싱된 내용의 일부를 살펴보도록 하겠습니다.

```
soup.head()
```

```
 [<meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>,
  <meta content="IE=edge" http-equiv="X-UA-Compatible"/>,
  <meta content="http://imgmovie.naver.com/today/naverme/naverme_profile.jpg"
property="me2:image">
  <meta content="네이버영화 " property="me2:post_tag">
  <meta content="네이버영화" property="me2:category1"/>
  <meta content="" property="me2:category2"/>
  <meta content="랭킹 : 네이버 영화" property="og:title"/>
  <meta content="영화, 영화인, 예매, 박스오피스 랭킹 정보 제공" property="og:description"/>
  <meta content="article" property="og:type"/>
  <meta content="http://movie.naver.com/movie/sdb/rank/rmovie.nhn"
```

```
property="og:url"/>
 <meta content="http://static.naver.net/m/movie/icons/OG_270_270.png"
property="og:image"/><!-- http://static.naver.net/m/movie/im/navermovie.jpg -->
 <meta content="http://imgmovie.naver.com/today/naverme/naverme_profile.jpg"
property="og:article:thumbnailUrl"/>
 <meta content="네이버 영화" property="og:article:author"/>
 <meta content="http://movie.naver.com/" property="og:article:author:url"/>
 <link href="http://static.naver.net/m/movie/icons/naver_movie_favicon.ico"
rel="shortcut icon" type="image/x-icon"/>
 <title>랭킹 : 네이버 영화</title>
 <link href="/common/css/movie_tablet.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/common.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/layout.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_default.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_layout.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_common.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_super_db.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <script src="/common/js/default-min.js" type="text/javascript"></script>
 </meta></meta>,
 <meta content="네이버영화 " property="me2:post_tag">
 <meta content="네이버영화" property="me2:category1"/>
 <meta content="" property="me2:category2"/>
 <meta content="랭킹 : 네이버 영화" property="og:title"/>
 <meta content="영화, 영화인, 예매, 박스오피스 랭킹 정보 제공" property="og:description"/>
 <meta content="article" property="og:type"/>
 <meta content="http://movie.naver.com/movie/sdb/rank/rmovie.nhn"
property="og:url"/>
 <meta content="http://static.naver.net/m/movie/icons/OG_270_270.png"
property="og:image"/><!-- http://static.naver.net/m/movie/im/navermovie.jpg -->
 <meta content="http://imgmovie.naver.com/today/naverme/naverme_profile.jpg"
property="og:article:thumbnailUrl"/>
 <meta content="네이버 영화" property="og:article:author"/>
 <meta content="http://movie.naver.com/" property="og:article:author:url"/>
 <link href="http://static.naver.net/m/movie/icons/naver_movie_favicon.ico"
rel="shortcut icon" type="image/x-icon"/>
 <title>랭킹 : 네이버 영화</title>
 <link href="/common/css/movie_tablet.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/common.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/layout.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_default.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_layout.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_common.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <link href="/common/css/old_super_db.css?20170925160128" rel="stylesheet"
type="text/css"/>
 <script src="/common/js/default-min.js" type="text/javascript"></script>
```

```
 </meta>,
 <meta content="네이버영화" property="me2:category1"/>,
 <meta content="" property="me2:category2"/>,
 <meta content="랭킹 : 네이버 영화" property="og:title"/>,
 <meta content="영화, 영화인, 예매, 박스오피스 랭킹 정보 제공" property="og:description"/>,
 <meta content="article" property="og:type"/>,
 <meta content="http://movie.naver.com/movie/sdb/rank/rmovie.nhn"
property="og:url"/>,
 <meta content="http://static.naver.net/m/movie/icons/OG_270_270.png"
property="og:image"/>,
 <meta content="http://imgmovie.naver.com/today/naverme/naverme_profile.jpg"
property="og:article:thumbnailUrl"/>,
 <meta content="네이버 영화" property="og:article:author"/>,
 <meta content="http://movie.naver.com/" property="og:article:author:url"/>,
 <link href="http://static.naver.net/m/movie/icons/naver_movie_favicon.ico"
rel="shortcut icon" type="image/x-icon"/>,
 <title>랭킹 : 네이버 영화</title>,
 <link href="/common/css/movie_tablet.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/common.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/layout.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/old_default.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/old_layout.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/old_common.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <link href="/common/css/old_super_db.css?20170925160128" rel="stylesheet"
type="text/css"/>,
 <script src="/common/js/default-min.js" type="text/javascript"></script>]
```

BeautifulSoup.findAll()

위 데이터 중에 우리가 찾기를 원하는 데이터들만 뽑아 보겠습니다.

아래 코드는 div 태그의 class 속성이 'tit3'이라는 값을 가진 태그들을 모두 찾아 tags라는 변수에 할당하였습니다.

```
tags = soup.findAll('div', attrs={'class':'tit3'})
print(tags)
```

```
[<div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=161242" title="범죄도시">범죄도시</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=129095" title="지오스톰">지오스톰</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=154353" title="대장 김창수">대장 김창수</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=150637" title="남한산성">남한산성</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=157178" title="나는 내일, 어제의 너와 만난다">나는
```

```
내일, 어제의 너와 만난다</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=88227" title="블레이드 러너 2049">블레이드 러너
2049</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=152650" title="마더!">마더!</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=140696" title="희생부활자">희생부활자</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=149747" title="킹스맨: 골든 서클">킹스맨: 골든 서클
</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=134898" title="토르: 라그나로크">토르: 라그나로크
</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=159830" title="너의 췌장을 먹고 싶어">너의 췌장을 먹
고 싶어</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=137890" title="살인자의 기억법">살인자의 기억법</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=159741" title="잠깐만 회사 좀 관두고 올게">잠깐만 회
사 좀 관두고 올게</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=161850" title="아이 캔 스피크">아이 캔 스피크</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=158649" title="잇 컴스 앳 나잇">잇 컴스 앳 나잇
</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=140806" title="사랑은 없다">사랑은 없다</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=149517" title="유리정원">유리정원</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=146407" title="다크타워: 희망의 탑">다크타워: 희망의
탑</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=153642" title="침묵">침묵</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=85579" title="신과함께">신과함께</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=70627" title="잃어버린 도시 Z">잃어버린 도시 Z</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=164968" title="노 게임 노 라이프 -제로-">노 게임 노
라이프 -제로-</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=138601" title="여배우는 오늘도">여배우는 오늘도</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=162420" title="미스 프레지던트">미스 프레지던트</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=162932" title="채비">채비</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=37732" title="루터">루터</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=161963" title="아이 앰 히스 레저">아이 앰 히스 레저
</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=160749" title="부라더">부라더</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=137945" title="아메리칸 메이드">아메리칸 메이드</a>
```

```
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=154573" title="다시 태어나도 우리">다시 태어나도 우
리</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=137326" title="블랙 팬서">블랙 팬서</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=142317" title="미옥">미옥</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=144215" title="아토믹 블론드">아토믹 블론드</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=116866" title="저스티스 리그">저스티스 리그</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=152268" title="직쏘">직쏘</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=121052" title="넛잡 2">넛잡 2</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=163503" title="용의 치과의사">용의 치과의사</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=160009" title="내 친구 정일우">내 친구 정일우</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=152267" title="22년째의 고백 - 내가 살인범이다 -
">22년째의 고백 - 내가 살인범이다 -</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=152385" title="꾼">꾼</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=143469" title="어메이징 메리">어메이징 메리</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=106069" title="히든 아이덴티티">히든 아이덴티티</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=140622" title="내 사랑 왕가흔">내 사랑 왕가흔</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=131440" title="테이킹">테이킹</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=162415" title="가을 우체국">가을 우체국</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=151559" title="새벽의 저주: 좀비랜드">새벽의 저주:
좀비랜드</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=155715" title="7호실">7호실</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=161129" title="스코어: 영화음악의 모든 것">스코어:
영화음악의 모든 것</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=155665" title="강철비">강철비</a>
 </div>, <div class="tit3">
 <a href="/movie/bi/mi/basic.nhn?code=143390" title="우리의 20세기">우리의 20세기</a>
 </div>]
```

영화제목 뽑아내기

아직까진 사람이 보기에는 너무 불필요한 정보들이 섞여 있습니다. 유의미한 데이터들만 추출하기 위해 위 tags 중에 영화 제목만을 출력하도록 해보겠습니다.

```
for tag in tags:
```

```
    print(tag.a['title'])
```

```
범죄도시
지오스톰
대장 김창수
남한산성
블레이드 러너 2049
희생부활자
나는 내일, 어제의 너와 만난다
킹스맨: 골든 서클
마더!
너의 췌장을 먹고 싶어
아이 캔 스피크
토르: 라그나로크
잠깐만 회사 좀 관두고 올게
살인자의 기억법
사랑은 없다
잇 컴스 앳 나잇
다크타워: 희망의 탑
신과함께
유리정원
침묵
루터
잃어버린 도시 Z
여배우는 오늘도
노 게임 노 라이프 -제로-
미스 프레지던트
아메리칸 메이드
채비
아토믹 블론드
부라더
아이 앰 히스 레저
다시 태어나도 우리
미옥
블랙 팬서
저스티스 리그
용의 치과의사
넛잡 2
어메이징 메리
직쏘
7호실
가을 우체국
꾼
히든 아이덴티티
22년째의 고백 - 내가 살인범이다 -
기억의 밤
강철비
내 친구 정일우
스코어: 영화음악의 모든 것
우리의 20세기
해피 데스데이
블레이드 러너
```