

Python Programming

웹 데이터 활용(웹크롤링)



권기웅



2. 웹크롤링

2.1 모듈 설치

2.1.1 간단한 웹서버에 요청 모듈

명령프롬프트에서 **pip install requests**

```
C:\> 명령 프롬프트
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\Ki-ung>pip install requests
Collecting requests
  Using cached requests-2.18.4-py2.py3-none-any.whl
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\ki-ung\appdata\local\programs\python\python36\lib\site-packages (from requests)
Requirement already satisfied: idna<2.7,>=2.5 in c:\users\ki-ung\appdata\local\programs\python\python36\lib\site-packages (from requests)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\ki-ung\appdata\local\programs\python\python36\lib\site-packages (from requests)
Requirement already satisfied: urllib3<1.23,>=1.21.1 in c:\users\ki-ung\appdata\local\programs\python\python36\lib\site-packages (from requests)
Installing collected packages: requests
Successfully installed requests-2.18.4

C:\Users\Ki-ung>
```

2.1.2 HTML Parser 모듈

명령프롬프트에서 **pip install bs4**

```
C:\> 명령 프롬프트
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\Ki-ung>pip install bs4
Collecting bs4
  Using cached bs4-0.0.1.tar.gz
Collecting beautifulsoup4 (from bs4)
  Using cached beautifulsoup4-4.6.0-py3-none-any.whl
Installing collected packages: beautifulsoup4, bs4
  Running setup.py install for bs4 ... done
Successfully installed beautifulsoup4-4.6.0 bs4-0.0.1

C:\Users\Ki-ung>
```



2.2 HTML 다운받기

```
Python 3.6.3 Shell
File Edit Shell Debug Options Window Help

<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
<title>멜론차트>멜론 TOP 100</title>
<meta name="keywords" content="멜론, 다운로드, 아티스트폴러스, 최향을 아는 똑똑한 추천 라디오, 내
<meta name="description"
<meta name="naver-site-verification"
<meta property="fb:app_id"
<meta property="og:title"
<meta property="og:image"

webcrawling.py - D:/workspacePython/webcrawling.py (3.6.3)
File Edit Format Run Options Window Help

import requests

url = "http://www.melon.com/chart/index.htm"
res = requests.get(url)
print(res.text)
```

2.3 HTML Parsing

```
Python 3.6.3 Shell
File Edit Shell Debug Options Window Help

===== RESTART: D:/workspacePython/webcrawling.py =====
1위[순위 동일] Beautiful(1-1=0 (NOTHING WITHOUT YOU)) : Wanna One (워너원)
2위[순위 동일] 좋아(2017 월간 윤종신 11월호) : 민서
3위[1단계 상승] 갖고 싶어(1-1=0 (NOTHING WITHOUT YOU)) : Wanna One (워너원)
4위[1단계 하락] 선물(Moonlight) :
5위[1단계 상승] 에너제틱 (Energetic) :
6위[2단계 상승] Twilight(1-1=0 (NOTHING WITHOUT YOU)) :
7위[2단계 하락] 피카부 (Peek-A-Boo) :
8위[1단계 상승] Nothing Without You :
9위[2단계 하락] 그때의 나, 그때의 너 :
10위[3단계 상승] To Be One (Outro) :
11위[4단계 상승] 활활 (Burn It Up) :
12위[2단계 하락] 좋니(LISTEN 010) :
13위[1단계 상승] DNA(LOVE YOURSELF) :
14위[2단계 하락] 그리워하다(Brother) :
15위[2단계 상승] Wanna Be (My Baby) :
16위[5단계 하락] LIKEY(twice) :
17위[2단계 상승] 에너제틱 (Energetic) :
18위[3단계 상승] 활활 (Burn It Up) :
19위[3단계 하락] 연애소설 (Feat. O.J. Borg) :
20위[4단계 상승] 봄날(YOU NEVER KNOW) :

webcrawling.py - D:/workspacePython/webcrawling.py (3.6.3)
File Edit Format Run Options Window Help

import requests
from bs4 import BeautifulSoup

url = "http://www.melon.com/chart/index.htm"
res = requests.get(url)
soup = BeautifulSoup(res.text, "html.parser")
lst50 = soup.select("tr.lst50") # soup.select("css선택자")
for i, music in enumerate(lst50):
    rank = i + 1
    updown = music.select("span.rank_wrap")[0].get("title") #.get("속성명")
    title = music.select("div.rank01.ellipsis > span > a")[0].string # 내용
    artist = music.select("div.rank02.ellipsis > span > a")[0].string
    album = music.select("div.rank03.ellipsis > a")[0].string
    musicInfo = "%d위[%s] %s(%s) : %s" % (rank, updown, title, album, artist)
    print(musicInfo)
```

