파이썬 공공데이터 파싱

공공데이터

- **https://www.data.go.kr**

「공공데이터의 제공 및 이용 활성화에 관한 법률」 제21조(공공데이터포털의 운영)

제21조(공공데이터포털의 운영) ① 행정안전부장관은 공공데이터의 효율적 제공을 위하여 통합제공시스템(이하 "공공데이터포털"이라 한다)을 구축·관리하고 활용을 촉진하여야 한다.
② 행정안전부장관은 공공기관의 장에게 공공데이터포털의 구축과 운영에 필요한 공공데이터의 연계, 제공 등의 협력을 요청할 수 있다. 이 경우 요청을 받은 공공기관의 장은 특별한 사유가 없는 한 이에 따라야 한다.
③ 그 밖에 공공데이터포털의 구축·관리 및 활용촉진 등 필요한 사항은 대통령령으로 정한다.

Come with PNU!

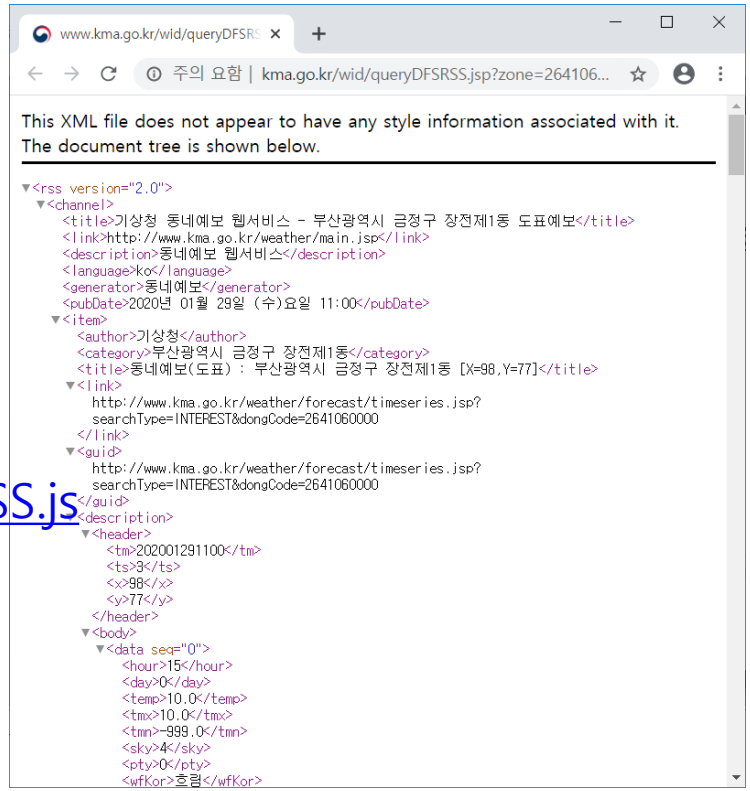
예제 공공데이터

- http://www.kobis.or.kr/kobisopenapi/homepg/apiservice/searchServiceInfo.do


```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <boxOfficeResult>
    <boxofficeType>일별 박스오피스</boxofficeType>
    <showRange>20120101~20120101</showRange>
  - <dailyBoxOfficeList>
    - <dailyBoxOffice>
        <rnum>1</rnum>
        <rank>1</rank>
        <rankInten>0</rankInten>
        <rankOldAndNew>OLD</rankOldAndNew>
        <movieCd>20112207</movieCd>
        <movieNm>미션임파서블:고스트프로토콜</movieNm>
        <openDt>2011-12-15</openDt>
        <salesAmt>2776060500</salesAmt>
        <salesShare>36.3</salesShare>
        <salesInten>-415699000</salesInten>
        <salesChange>-13</salesChange>
        <salesAcc>40541108500</salesAcc>
        <audiCnt>353274</audiCnt>
        <audiInten>-60106</audiInten>
        <audiChange>-14.5</audiChange>
        <audiAcc>5328435</audiAcc>
        <scrnCnt>697</scrnCnt>
        <showCnt>3223</showCnt>
      </dailyBoxOffice>
    - <dailyBoxOffice>
        <rnum>2</rnum>
        <rank>2</rank>
        <rankInten>1</rankInten>
        <rankOldAndNew>OLD</rankOldAndNew>
        <movieCd>20110295</movieCd>
        <movieNm>마이 웨이</movieNm>
        <openDt>2011-12-21</openDt>
        <salesAmt>1189058500</salesAmt>
        <salesShare>15.6</salesShare>
        <salesInten>-105894500</salesInten>
        <salesChange>-8.2</salesChange>
        <salesAcc>13002897500</salesAcc>
        <audiCnt>153501</audiCnt>
        <audiInten>-16465</audiInten>
        <audiChange>-9.7</audiChange>
        <audiAcc>1739543</audiAcc>
        <scrnCnt>588</scrnCnt>
        <showCnt>2321</showCnt>
      </dailyBoxOffice>
    - <dailyBoxOffice>
        <rnum>3</rnum>
        <rank>3</rank>
        <rankInten>-1</rankInten>
        <rankOldAndNew>OLD</rankOldAndNew>
        <movieCd>20112621</movieCd>
        <movieNm>셜록홈즈 : 그림자 게임</movieNm>
        <openDt>2011-12-21</openDt>
        <salesAmt>1176022500</salesAmt>
        <salesShare>15.4</salesShare>
        <salesInten>-210328500</salesInten>
        <salesChange>-15.2</salesChange>
        <salesAcc>10678327500</salesAcc>
        <audiCnt>153004</audiCnt>
        <audiInten>-31283</audiInten>
        <audiChange>-17</audiChange>
```

예제 공공데이터

- https://www.weather.go.kr/weather/lifeindustry/sevice_rss.jsp



http://www.kma.go.kr/wid/queryDFSRSS.jsp?zone=2641060000

XML 데이터 크롤링

- **BeautifulSoup을 이용한 파싱**

```
from bs4 import BeautifulSoup
def dailyBoxOfficeList(data) :
    bs = BeautifulSoup(data, 'lxml-xml')
    dailyBoxOffices = bs.find_all("dailyBoxOffice")

    return dailyBoxOffices
```

- **ElementTree 을 이용한 파싱**

```
import xml.etree.ElementTree as ET
def dailyBoxOfficeList(data) :
    tree = ET.fromstring(data) # xml 문자를 이용한 파싱
    dailyBoxOffices = tree.find("dailyBoxOfficeList").findall("dailyBoxOffice")
    #dailyBoxOffices = tree.findall("dailyBoxOfficeList/dailyBoxOffice")

    return dailyBoxOffices
```

실습


```
▼<boxOfficeResult>
    <boxofficeType>일별 박스오피스</boxofficeType>
    <showRange>20200120~20200120</showRange>
  ▼<dailyBoxOfficeList>
    ▼<dailyBoxOffice>
        <rnum>1</rnum>
        <rank>1</rank>
        <rankInten>0</rankInten>
        <rankOldAndNew>OLD</rankOldAndNew>
        <movieCd>20183002</movieCd>
        <movieNm>해치지않아</movieNm>
        <openDt>2020-01-15</openDt>
        <salesAmt>541763240</salesAmt>
        <salesShare>35.5</salesShare>
        <salesInten>-1386039100</salesInten>
        <salesChange>-71.9</salesChange>
        <salesAcc>7444487020</salesAcc>
        <audiCnt>67840</audiCnt>
        <audiInten>-153978</audiInten>
        <audiChange>-69.4</audiChange>
        <audiAcc>881242</audiAcc>
        <scrnCnt>1118</scrnCnt>
        <showCnt>5281</showCnt>
      </dailyBoxOffice>
    ▼<dailyBoxOffice>
        <rnum>2</rnum>
        <rank>2</rank>
        <rankInten>0</rankInten>
        <rankOldAndNew>OLD</rankOldAndNew>
        <movieCd>20196670</movieCd>
        <movieNm>나쁜 녀석들: 포에버</movieNm>
        <openDt>2020-01-15</openDt>
        <salesAmt>298107730</salesAmt>
        <salesShare>19.5</salesShare>
        <salesInten>-702046880</salesInten>
        <salesChange>-70.2</salesChange>
        <salesAcc>3952522900</salesAcc>
```


1위 (0) 해치지않아
2위 (0) 나쁜 녀석들: 포에버
3위 (0) 닥터 두리틀
4위 (0) 백두산
5위 (▲2) 타오르는 여인의 초상
6위 (0) 천문: 하늘에 묻는다
7위 (0) 아내를 죽였다
8위 (▼3) 스타워즈: 라이즈 오브 스카이워커
9위 (0) 남산의 부장들
10위 (0) 시동

파이썬 datetime 모듈

- 날짜 및 시간 계산을 지원
  - 날짜만 저장 : datetime.date
  - 시간만 저장 : datetime.time
  - 날짜와 시간을 저장 : datetime.datetime
  - 시간 구간 정보 : datetime.timedelta
    - 예) dspDate = dt + timedelta(days=2)
      dt날짜에서 2일을 더함
  - 두 날짜, 시간 또는 날짜 시간의 인스턴스 객체 간 차이를 마이크로 초 해상도로 나타내는 기간 : datetime.tzinfo
  - tzinfo 추상 기본 클래스를 UTC의 고정 offset으로 구현하는 클래스 : datetime.timezone

파이썬 datetime.datetime

- 속성
  - year: 연도
  - month: 월
  - day: 일
  - hour: 시
  - minute: 분
  - second: 초
  - microsecond: 마이크로초(micro seconds, 백만분의 일초)
- 메소드
  - weekday(): 요일 반환 (0:월, 1:화, 2:수, 3:목, 4:금, 5:토, 6:일)
  - strftime(): 문자열 반환
    - %Y(4자리 연도 숫자), %m(2자리 월 숫자), %d(2자리 일 숫자)
  - now() : 컴퓨터의 현재 시각을 datetime.datetime 클래스 반환
  - date(): 날짜 정보만 가지는 datetime.date 클래스 객체 반환
  - time(): 시간 정보만 가지는 datetime.time 클래스 객체 반환

실습: 자동으로 어제 날짜 조회


```
▼<boxOfficeResult>
  <boxofficeType>일별 박스오피스</boxofficeType>
  <showRange>20200120~20200120</showRange>
 ▼<dailyBoxOfficeList>
  ▼<dailyBoxOffice>
    <rnum>1</rnum>
    <rank>1</rank>
    <rankInten>0</rankInten>
    <rankOldAndNew>OLD</rankOldAndNew>
    <movieCd>20183002</movieCd>
    <movieNm>해치지않아</movieNm>
    <openDt>2020-01-15</openDt>
    <salesAmt>541763240</salesAmt>
    <salesShare>35.5</salesShare>
    <salesInten>-1386039100</salesInten>
    <salesChange>-71.9</salesChange>
    <salesAcc>7444487020</salesAcc>
    <audiCnt>67840</audiCnt>
    <audiInten>-153978</audiInten>
    <audiChange>-69.4</audiChange>
    <audiAcc>881242</audiAcc>
    <scrnCnt>1118</scrnCnt>
    <showCnt>5281</showCnt>
  </dailyBoxOffice>
  ▼<dailyBoxOffice>
    <rnum>2</rnum>
    <rank>2</rank>
    <rankInten>0</rankInten>
    <rankOldAndNew>OLD</rankOldAndNew>
    <movieCd>20196670</movieCd>
    <movieNm>나쁜 녀석들: 포에버</movieNm>
    <openDt>2020-01-15</openDt>
    <salesAmt>298107730</salesAmt>
    <salesShare>19.5</salesShare>
    <salesInten>-702046880</salesInten>
    <salesChange>-70.2</salesChange>
    <salesAcc>3952522900</salesAcc>
```


```
1위 (0) 해치지않아
2위 (0) 나쁜 녀석들: 포에버
3위 (0) 닥터 두리틀
4위 (0) 백두산
5위 (▲2) 타오르는 여인의 초상
6위 (0) 천문: 하늘에 묻는다
7위 (0) 아내를 죽였다
8위 (▼3) 스타워즈: 라이즈 오브 스카이워커
9위 (0) 남산의 부장들
10위 (0) 시동
```

```python
from datetime import datetime, timedelta, date

dt = datetime.now() - timedelta(days=1)
dt = datetime.strftime(dt, "%Y%m%d")
```

실습


```
▼<rss version="2.0">
  ▼<channel>
     <title>기상청 동네예보 웹서비스 - 부산광역시 금정구 장전제1동 도표예보</title>
     <link>http://www.kma.go.kr/weather/main.jsp</link>
     <description>동네예보 웹서비스</description>
     <language>ko</language>
     <generator>동네예보</generator>
     <pubDate>2020년 01월 30일 (목)요일 08:00</pubDate>
    ▼<item>
       <author>기상청</author>
       <category>부산광역시 금정구 장전제1동</category>
       <title>동네예보(도표) : 부산광역시 금정구 장전제1동 [X=98,Y=77]</title>
      ▼<link>
         http://www.kma.go.kr/weather/forecast/timeseries.jsp?searchType=INTEREST&dongCode=2641060000
       </link>
      ▼<guid>
         http://www.kma.go.kr/weather/forecast/timeseries.jsp?searchType=INTEREST&dongCode=2641060000
       </guid>
      ▼<description>
        ▼<header>
           <tm>202001300800</tm>
           <ts>2</ts>
           <x>98</x>
           <y>77</y>
         </header>
        ▼<body>
          ▼<data seq="0">
             <hour>12</hour>
             <day>0</day>
             <temp>9.0</temp>
             <tmx>10.0</tmx>
             <tmn>-999.0</tmn>
             <sky>3</sky>
             <pty>0</pty>
             <wfKor>구름 많음</wfKor>
             <wfEn>Mostly Cloudy</wfEn>
             <pop>20</pop>
             <r12>0.0</r12>
             <s12>0.0</s12>
             <ws>7.5</ws>
             <wd>0</wd>
             <wdKor>북</wdKor>
             <wdEn>N</wdEn>
             <reh>50</reh>
             <r06>0.0</r06>
             <s06>0.0</s06>
           </data>
```


```
2020-01-30 11:00:00 일자
2020 년 01 월 30 일  :  15 시 ( 구름 많음 )
2020 년 01 월 30 일  :  18 시 ( 구름 많음 )
2020 년 01 월 30 일  :  21 시 ( 구름 많음 )
2020 년 01 월 30 일  :  24 시 ( 구름 많음 )
2020 년 01 월 31 일  :  3 시 ( 구름 많음 )
2020 년 01 월 31 일  :  6 시 ( 구름 많음 )
2020 년 01 월 31 일  :  9 시 ( 구름 많음 )
2020 년 01 월 31 일  :  12 시 ( 구름 많음 )
2020 년 01 월 31 일  :  15 시 ( 구름 많음 )
2020 년 01 월 31 일  :  18 시 ( 구름 많음 )
2020 년 01 월 31 일  :  21 시 ( 맑음 )
2020 년 01 월 31 일  :  24 시 ( 맑음 )
2020 년 02 월 01 일  :  3 시 ( 맑음 )
2020 년 02 월 01 일  :  6 시 ( 맑음 )
2020 년 02 월 01 일  :  9 시 ( 맑음 )
2020 년 02 월 01 일  :  12 시 ( 맑음 )
2020 년 02 월 01 일  :  15 시 ( 맑음 )
2020 년 02 월 01 일  :  18 시 ( 맑음 )
2020 년 02 월 01 일  :  21 시 ( 맑음 )
2020 년 02 월 01 일  :  24 시 ( 맑음 )
```

JSON(JavaScript Object Notation)

- JavaScript 문법에 영향을 받아 개발된 Lightweight한 데이타 표현 방식
- 데이타를 교환하는 한 포맷으로서 그 단순함과 유연함 때문에 널리 사용
- 키(Key)-값(Value)의 쌍으로 이루어짐

- Python JSON 표준 라이브러리 : json
  - JSON 인코딩 : Python 타입의 Object를 JSON 문자열로 변경
    - .dumps(딕션너리)
    - - indent 옵션 : JSON 문자열을 읽기 쉽게 작성
    - - ensure_ascii 옵션 : 한글 처리 ensure_ascii=False
  - JSON 디코딩 : JSON 문자열을 다시 Python 타입으로 변환
    - .loads(문자열)

Come with PNU!

딕션너리와 JSON 다루기

```python
import json

dic = {"boxofficeType":"일별 박스오피스","showRange":"20200129~20200129"}

print("-"*20)
for item in dic.items() :
    print(item)

print("-"*20)
for key in dic.keys() :
    print(key)

print("-"*20)
for value in dic.values() :
    print(value)

print("-"*20)
for key, value in dic.items() :
    print(key, value)

print("-"*20)
print(dic.get("boxofficeType"))
print(dic["boxofficeType"])

#딕션너리 -> JSON 문자열
print("-"*20)
jdic = json.dumps(dic, indent=4, ensure_ascii=False)
print(jdic)
print(type(jdic))

#JSON 문자열 -> 딕션너리
print("-"*20)
dic2 = json.loads(jdic)
print(dic2)
print(type(dic2))
```

```
--------------------
('boxofficeType', '일별 박스오피스')
('showRange', '20200129~20200129')
--------------------
boxofficeType
showRange
--------------------
일별 박스오피스
20200129~20200129
--------------------
boxofficeType 일별 박스오피스
showRange 20200129~20200129
--------------------
일별 박스오피스
일별 박스오피스
--------------------
{
    "boxofficeType": "일별 박스오피스",
    "showRange": "20200129~20200129"
}
<class 'str'>
--------------------
{'boxofficeType': '일별 박스오피스', 'showRange': '20200129~20200129'}
<class 'dict'>
```

실습 : JSON 데이터 크롤링


```
{"boxOfficeResult":{"boxofficeType":"일별 박스오피스","showRange":"20200130~20200130","dailyBoxOfficeList":
[{"rnum":"1","rank":"1","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20180989","movieNm":"남산의 부장
들","openDt":"2020-01-
22","salesAmt":"945133180","salesShare":"44.0","salesInten":"-355938620","salesChange":"-27.4","salesAcc":"32547
182270","audiCnt":"118084","audiInten":"-93873","audiChange":"-44.3","audiAcc":"3743586","scrnCnt":"1383","showC
nt":"7262"},{"rnum":"2","rank":"2","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20192101","movieNm":"히트
맨","openDt":"2020-01-
22","salesAmt":"602166980","salesShare":"28.1","salesInten":"-228841510","salesChange":"-27.5","salesAcc":"15390
499240","audiCnt":"74508","audiInten":"-62668","audiChange":"-45.7","audiAcc":"1789019","scrnCnt":"1027","showCn
t":"4745"},{"rnum":"3","rank":"3","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20188421","movieNm":"미스터
주: 사라진 VIP","openDt":"2020-01-
22","salesAmt":"92604980","salesShare":"4.3","salesInten":"-57051680","salesChange":"-38.1","salesAcc":"45814210
60","audiCnt":"12298","audiInten":"-12367","audiChange":"-50.1","audiAcc":"543798","scrnCnt":"596","showCnt":"15
70"},{"rnum":"4","rank":"4","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20192300","movieNm":"스파이 지니어
스 ","openDt":"2020-01-
22","salesAmt":"88064900","salesShare":"4.1","salesInten":"-37017120","salesChange":"-29.6","salesAcc":"30754843
00","audiCnt":"11730","audiInten":"-8596","audiChange":"-42.3","audiAcc":"374662","scrnCnt":"574","showCnt":"116
9"},{"rnum":"5","rank":"5","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20100312","movieNm":"인셉
션","openDt":"2010-07-
21","salesAmt":"67568700","salesShare":"3.1","salesInten":"-21178300","salesChange":"-23.9","salesAcc":"43645889
200","audiCnt":"8346","audiInten":"-6634","audiChange":"-44.3","audiAcc":"5859100","scrnCnt":"204","showCnt":"60
5"},{"rnum":"6","rank":"6","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20183002","movieNm":"해치지않
아","openDt":"2020-01-
15","salesAmt":"47241640","salesShare":"2.2","salesInten":"-34285520","salesChange":"-42.1","salesAcc":"10049037
830","audiCnt":"6106","audiInten":"-7222","audiChange":"-54.2","audiAcc":"1195437","scrnCnt":"471","showCnt":"89
5"},{"rnum":"7","rank":"7","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20198676","movieNm":"타오르는 여인의
초상","openDt":"2020-01-
16","salesAmt":"35329360","salesShare":"1.6","salesInten":"-7492020","salesChange":"-17.5","salesAcc":"791441040
","audiCnt":"4480","audiInten":"-2579","audiChange":"-36.5","audiAcc":"95398","scrnCnt":"142","showCnt":"275"},
{"rnum":"8","rank":"8","rankInten":"13","rankOldAndNew":"OLD","movieCd":"20193321","movieNm":"정직한 후
보","openDt":"2020-02-
12","salesAmt":"29833000","salesShare":"1.4","salesInten":"25136000","salesChange":"535.2","salesAcc":"72029000"
,"audiCnt":"3943","audiInten":"3420","audiChange":"653.9","audiAcc":"10143","scrnCnt":"21","showCnt":"22"},
{"rnum":"9","rank":"9","rankInten":"0","rankOldAndNew":"OLD","movieCd":"20196670","movieNm":"나쁜 녀석들: 포에
버","openDt":"2020-01-
15","salesAmt":"18913800","salesShare":"0.9","salesInten":"-8837500","salesChange":"-31.8","salesAcc":"476391630
0","audiCnt":"2765","audiInten":"-1533","audiChange":"-35.7","audiAcc":"531933","scrnCnt":"204","showCnt":"297"}
,{"rnum":"10","rank":"10","rankInten":"0","rankOldAndNew":"NEW","movieCd":"20193836","movieNm":"핑크퐁 시네마 콘
서트 : 우주대탐험","openDt":"2020-01-
30","salesAmt":"20010500","salesShare":"0.9","salesInten":"20010500","salesChange":"100","salesAcc":"36416500","
audiCnt":"2693","audiInten":"2693","audiChange":"100","audiAcc":"5235","scrnCnt":"176","showCnt":"341"}]}}
```


1위 (0) 남산의 부장들
2위 (0) 히트맨
3위 (0) 미스터 주: 사라진 VIP
4위 (0) 스파이 지니어스
5위 (0) 인셉션
6위 (0) 해치지않아
7위 (0) 타오르는 여인의 초상
8위 (▲13) 정직한 후보
9위 (0) 나쁜 녀석들: 포에버
10위 (0) 핑크퐁 시네마 콘서트 : 우주대탐험