반정형 데이터 수집(JSON)

최수연 교수

mibm400@hanmail.net

학습목표

- JSON에 대해 설명할 수 있다.
- 문자열을 읽고 JSON 형식으로 저장할 수 있다.
- JSON 파일을 읽어와 DataFrame으로 변환할 수 있다.

목차

- JSON의 정의
- 파이썬 객체를 JSON 객체로 변환하기
- JSON 파일 읽어오기

공공데이터 종류

- 정형데이터(structured data)

✓ 미리 정해 놓은 형식과 구조에 따라 저장된 데이터

예) 관계형 데이터베이스의 테이블, 스프레드시트, CSV 등

- 반정형데이터(semi-structured data)

✓ 일정한 규칙의 고정된 필드에 저장되어 있지 않지만

데이터의 구조 정보를 데이터와 함께 제공하는 데이터

예)XML, HTML, JSON, 웹문서, 웹로그 등

JSON 이란?

- JSON(JavaScript Object Notation)

✓ 자바 스크립 언어로 구조화된 문자 기반 표준 포맷

✓ 파이썬의 딕셔너리와 리스트를 중첩한 것과 비슷

JSON 배열 ----▸
JSON 객체 ----▸
JSON 객체 ----▸

```
[
  {
    "휴양림명": "화원자연휴양림",
    "시도명 "  : "대구광역시",
    "휴양림구분": "공유림",
    "휴양림면적": "720000"
  },
  {
    "휴양림명": "옥화자연휴양림",
    "시도명 "  : "충청북도",
    "휴양림구분": "공유림",
    "휴양림면적": "1360000"
  }
]
```

{ key : value } : 객체

[name, Tel] : 배열

{ "휴양림명" : "화원자연휴양림" }

키(key) 값(value)

JSON과 Python 변환

JSON과 Python 변환

- JSON 라이브러리 선언

```
import json
```

- JSON(문자열)과 Python 객체(Dictionary) 변환

JSON 문자열 생성 및 변환

- json.dumps(): 파이썬 객체를 JSON 문자열로 변경

✓ json.dumps(데이터, ensure_ascii=False)

  - 데이터: 파이썬 객체(Dictionary 구조)
  - ensure_ascii = False : 한글이 깨지지 않고 저장된 문자 그대로 출력

- json.loads(): JSON 문자열을 파이썬 객체로 변경

✓ json.loads(데이터)

  - 데이터: json구조를 갖는 문자열

- pd.DataFrame(): 파이썬 객체를 DataFrame 구조로 변경

✓ pd.DataFrame(데이터)

  - 데이터: 파이썬 객체

파이썬 객체 생성하기


```
1 ssudata = {'단과대학':'인문대학', '학과': '사학과', '전화번호':'0380' }
2 print(ssudata)
3 print(ssudata['단과대학'])
4 print(ssudata['학과'])
5 print(ssudata['전화번호'])
6
7 print(type(ssudata))
```

```
{'단과대학': '인문대학', '학과': '사학과', '전화번호': '0380'}
인문대학
사학과
0380
<class 'dict'>
```

JSON 형식으로 변경하기


```
1 import json
```

```
1 jdata = json.dumps(ssudata, ensure_ascii=False)
2 print(jdata)
3 print(type(jdata))
```

```
{"단과대학": "인문대학", "학과": "사학과", "전화번호": "0380"}
<class 'str'>
```

파이썬 객체로 변경하기


```
1 ssudata2 = json.loads(jdata)
2 print(ssudata2)
3 print(type(ssudata2))
```

```
{'단과대학': '인문대학', '학과': '사학과', '전화번호': '0380'}
<class 'dict'>
```

파이썬 객체 생성하기


```
major = [{'단과대학':'인문대학', '학과': '사학과', '전화번호':'0380' },
         {'단과대학':'자연과학대학','학과': '물리학과','전화번호':'0420'},
         {'단과대학':'경영대학', '학과': '회계학과', '전화번호':'0548'}
        ]

major
```

```
[{'단과대학': '인문대학', '학과': '사학과', '전화번호': '0380'},
 {'단과대학': '자연과학대학', '학과': '물리학과', '전화번호': '0420'},
 {'단과대학': '경영대학', '학과': '회계학과', '전화번호': '0548'}]
```

파이썬 객체 생성하기


```
1 print(type(major))
2
3 print(major[1]['단과대학'])
4 print(major[1]['학과'])
5 print(major[1]['전화번호'])
```

```
<class 'list'>
자연과학대학
물리학과
0420
```

JSON 형식으로 변경하기


```
1 jdata = json.dumps(major, ensure_ascii=False)
2 print(type(jdata))
3 jdata
```

```
<class 'str'>
'[{"단과대학": "인문대학", "학과": "사학과", "전화번호": "0380"}, {"단과대학": "자연과학대학", "학과": "물리학과", "전화번호": "0420"}, {"단과대학": "경영대학", "학과": "회계학과", "전화번호": "0548"}]'
```

파이썬 객체로 변경하기


```
1 jdata = json.loads(jdata)
2 jdata
```

```
[{'단과대학': '인문대학', '학과': '사학과', '전화번호': '0380'},
 {'단과대학': '자연과학대학', '학과': '물리학과', '전화번호': '0420'},
 {'단과대학': '경영대학', '학과': '회계학과', '전화번호': '0548'}]
```

파이썬 객체 생성


```
1 import pandas as pd
2 data = pd.DataFrame(jdata)
3 data
```

	단과대학	학과	전화번호
0	인문대학	사학과	0380
1	자연과학대학	물리학과	0420
2	경영대학	회계학과	0548

JSON 파일 읽어오기

Python 객체(dictionary)를 JSON 화일로 변환

❖ with open('화일명', 'w') as 파일객체 : 파일을 열기 위한 명령문

✓ 화일명 : JSON 경로명과 화일명(*.json)

✓ 'w' : 파일을 쓰기 전용으로 open

✓ 파일객체: 파일을 관리하는 객체

- json.dump(데이터, 파일객체): 파이썬 객체를 JSON 화일로 변경

✓ 데이터: 파이썬 객체(Dictionary 구조)

✓ 파일객체: 파일을 관리하는 객체

JSON 파일과 Python 객체(화일) 변환

파이썬 객체 <화일>

json.load()

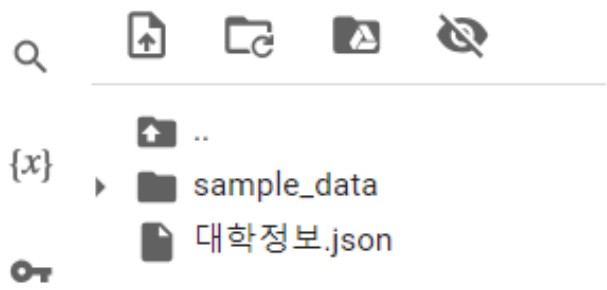
json.dump()

JSON <화일>

```
1 major = [{'단과대학':'인문대학', '학과': '사학과', '전화번호':'0380'},
2          {'단과대학':'자연과학대학','학과': '물리학과','전화번호':'0420'},
3          {'단과대학':'경영대학','학과': '회계학과','전화번호':'0548'}]
4 major
```

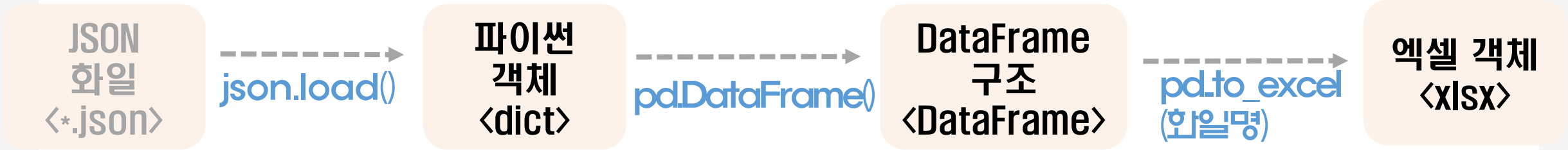
```
[{'단과대학': '인문대학', '학과': '사학과', '전화번호': '0380'},
 {'단과대학': '자연과학대학', '학과': '물리학과', '전화번호': '0420'},
 {'단과대학': '경영대학', '학과': '회계학과', '전화번호': '0548'}]
```

```
1 with open('대학정보.json', 'w') as fp:
2   json.dump(major, fp)
```




```
4  [
5    {
6      "단과대학": "인문대학",
7      "학과": "사학과",
8      "전화번호": "0380"
9    },
10   {
11     "단과대학": "자연과학대학",
12     "학과": "물리학과",
13     "전화번호": "0420"
14   },
15   {
16     "단과대학": "경영대학",
17     "학과": "회계학과",
18     "전화번호": "0548"
19   }
20 ]
```

JSON 파일 읽어오기와 excel로 저장하기



❖ with open('화일명', 'r') as 파일객체: 파일을 열기 위한 명령문

✓ 화일명 : JSON 경로명과 화일명(*.json)

✓ 'r' : 파일을 읽기 전용으로 open

✓ 파일객체: 파일을 관리하는 객체

❖ 변수명 = json.load(파일객체): JSON 파일의 binary를 파이썬 객체로 변경

✓ 데이터: json구조를 갖는 binary 데이터

❖ pd.to_excel(화일명): DataFrame에 저장된 데이터를 excel 파일로 저장

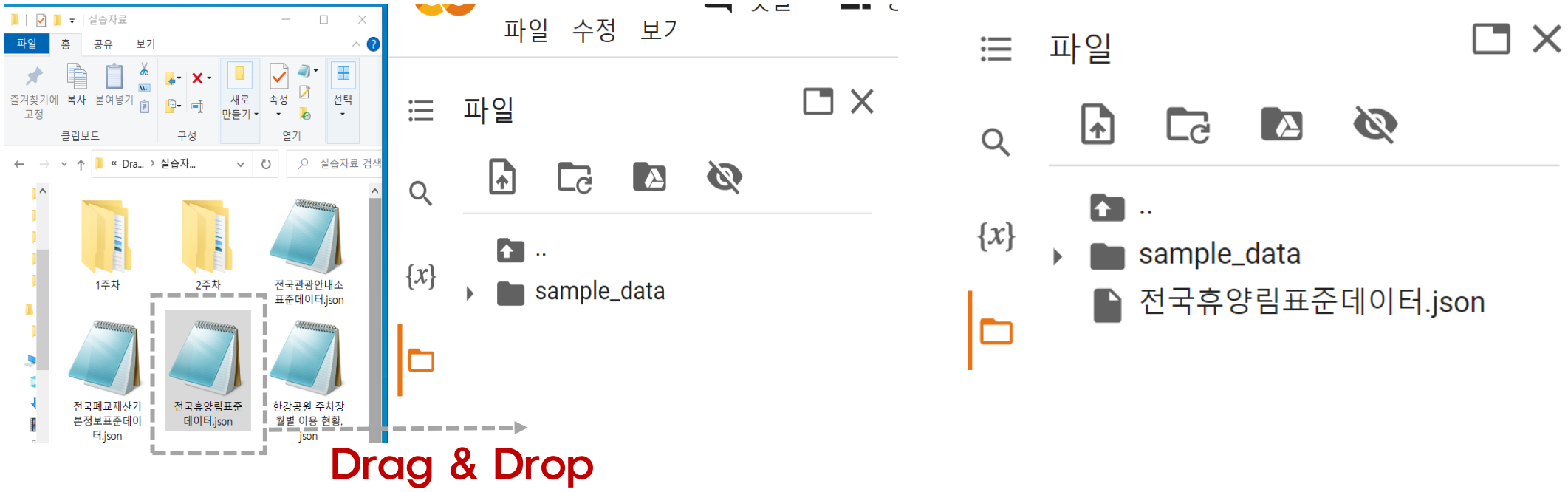
✓ 화일명 : excel 화일명

전국휴양림표준데이터 JSON 파일 다운로드

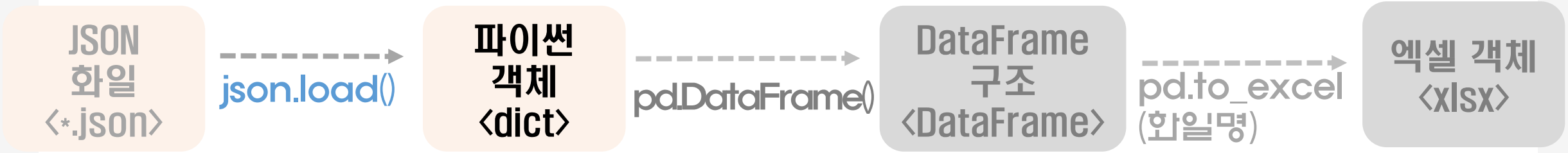
- URL: https://www.data.go.kr/data/15013111/standard.do

JSON 파일 Colab 으로 가져오기

- 다운로드 받은 JSON 파일을 Colab 으로 Drag&Drop



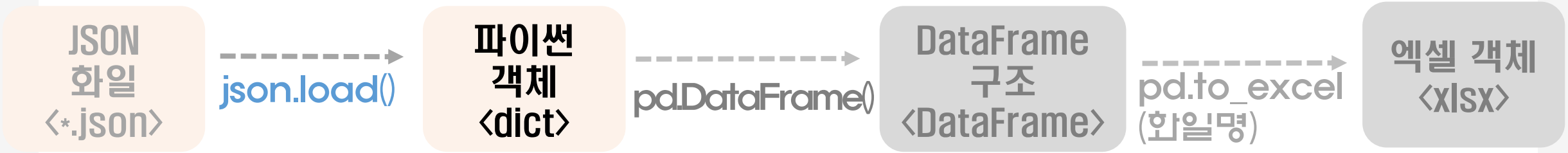
JSO 파일 읽어오고 구조 확인하기




```
1 import json
2 with open('전국휴양림표준데이터.json','r') as files:
3     temp = json.load(files)
4 temp
```

```
{'fields': [{'id': '휴양림명'},
  {'id': '시도명'},
  {'id': '휴양림구분'},
  {'id': '휴양림면적'},
  {'id': '수용인원수'},
  {'id': '입장료'},
  {'id': '숙박가능여부'},
  {'id': '주요시설명'},
  {'id': '소재지도로명주소'},
  {'id': '관리기관명'},
  {'id': '휴양림전화번호'},
  {'id': '홈페이지주소'},
  {'id': '위도'},
  {'id': '경도'},
  {'id': '데이터기준일자'},
  {'id': '제공기관코드'},
```

JSON 데이터 구조 확인하고 필요부분 가져오기



fields
- id
- id
- …

records
- 휴양림명
- ^시도명
- 휴양림구분
- …

```
1 data = []
2 for d in temp['records']:
3   data.append(d)
4 data
```

pd.DataFrame으로 변환하기

JSON 화일 <*.json> → json.load() → 파이썬 객체 <dict> → pd.DataFrame() → DataFrame 구조 <DataFrame> → pd.to_excel (화일명) → 엑셀 객체 <xlsx>

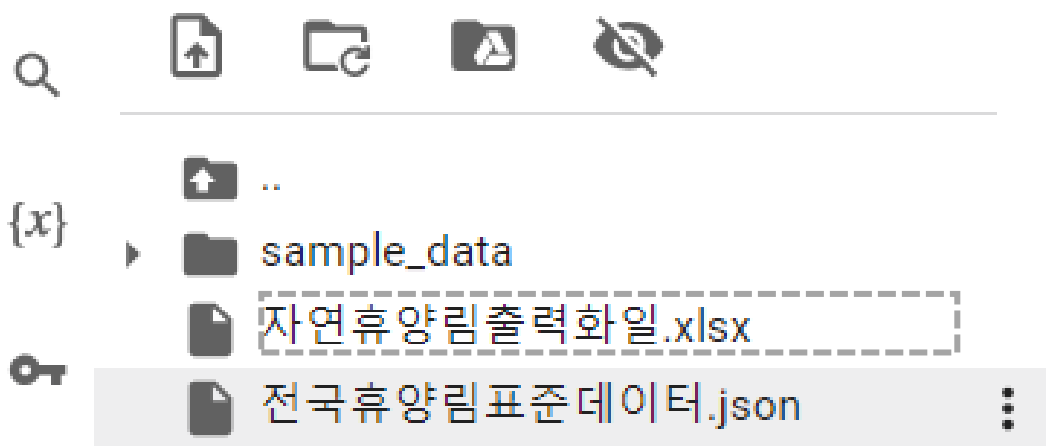
```
1 import pandas as pd
2 pdData = pd.DataFrame(data)
3 pdData
```

	휴양림명	시도명	휴양림구분	휴양림면적	수용인원수	입장료	구박가능여부	주요시설명	소재지도로명주소	관리기관명	휴양림전화번호	홈페이지주소	위도	경도	데이터기준일자	제공기관코드	제공기관명
0	화원자연휴양림	대구광역시	공유림	720000	115	없음	Y	숲속의집+산림문화휴양관+산림욕장+전망데크등	대구광역시 달성군 화원읍 화원휴양림길 126	대구광역시 달성군시설관리공단	053-659-4455	http://hwawon.dssiseol.or.kr/hwawon/html/main....	35.77179173	128.5367385	2022-06-10	3480000	대구광역시 달성군
1	옥화자연휴양림	충청북도	공유림	1360000	373	어른(개인 1000원+단체 500원)+청소년(개인 500원+단체 300원)+어린이(...	Y	숲속의집+산림휴양관+국민여가오토캠핑장+물놀이장 등	충청북도 청주시 상당구 미원면 운암옥화길 140	청주시 푸른도시사업본부	043-270-7384	www.foresttrip.go.kr/indvz	36.5989501183	127.6943715683	2022-06-15	5710000	충청북도 청주시
2	광치자연휴양림	강원도	공유림	51	160	3000원	Y	숲속의집, 휴양관, 숲속카페,무장애나눔길	강원도 양구군 남면 광치령로 1794번길 265	강원도 양구군청	033-482-3115	http://www.foresttrip.go.kr	38.14199686	128.0710755	2022-08-31	4320000	강원도 양구군
3	금원산자연휴양림(거창)	경상남도	공유림	1300000	685	개인 어른1000원+개인 군인청소년 600원+개인 어린이 300원+단체 어른800원...	Y	숲속의집+휴양관	경상남도 거창군 위천면 금원산길 471-27	경상남도청	055-254-3971	https://www.foresttrip.go.kr/indvz/main.do?hmp...	35.72617571	127.795727	2022-11-29	6480000	경상남도
4	항노화자연휴양림(항노화힐링랜드)	경상남도	공유림	479276	100	일반 3000원(만 7세 이상 만 65세 미만)	Y	숲속의집+휴양관	경상남도 거창군 가조면 의상봉길 834	경상남도 거창군청	055-940-7933	https://www.foresttrip.go.kr/indvz/main.do?hmp...	35.73662049	128.0408983	2022-11-29	6480000	경상남도

pd.DataFrame을 excel 파일로 저장하기

JSON 화일 <*.json> → json.load() → 파이썬 객체 <dict> → pd.DataFrame() → DataFrame 구조 <DataFrame> → pd.to_excel (화일명) → 엑셀 객체 <xlsx>

```
1 outputFile = '자연휴양림출력화일.xlsx'
2 pdData.to_excel(outputFile)
```



수고하셨습니다.