wrangling_json_filetypes

November 3, 2019

1 Wrangling JSON Filetypes

1.1 Data-Wrangling techniques using a World Bank JSON file

1.1.1 Phase I. Business Understanding.

In this overview, I will venture into the lightweight data-interchange format JavaScript Object Notation (JSON) data structure.

JSON is effortless for humans to read and write while also being accessible for machines to parse from and generate. JSON is a lightweight data-interchange format working between clients, servers, and embedded in programming languages. It is a subset of the JavaScript object and distinct from any other programming language. JSON is a subset of the JavaScript Programming Language Standard: ECMA-262 3rd Edition - December 1999. It has a text format that is completely language independent but utilizes analogs that are accustomed to programmers of the "C-family" of languages: C, C++, C#, Java, JavaScript, Perl, Python, and many others[1].

JSON's human-readable data structures are efficient for more extensive data exports that contain a hierarchically structured format.

The basic structure of the JSON object is as follows:

i. Data is by name/value pairs with colons.

ii. Commas split data-objects.

iii. Curly braces {} hold Data-Objects.

iv. Square brackets [] can be utilized to indicate a list of objects in a group.

v. Data-Element enclosed with quotes for characters, and without quotes for a numeric value.

During the following example, I will demonstrate how to use Data-Wrangling techniques on a ***World Bank JSON file***. Additionally, I will answer three questions (set objectives) that emulate a real-world example of what a Data Scientist may encounter in an organization attempting to gain insight from a JSON file!

Set objectives:

1. What are the top 10 countries with the most projects?
2. What are the top 10 major project themes (using column 'mjtheme_namecode') you see in the data?

3. Finally, create a DataFrame with the missing names filled in.

Initial assessment of resources and tools:

1. **Resourse:** Data Scientist, Alfred Hull
2. **Resourse:** The data in this overview was provided at the World Bank and is 2.8 Megabytes (MB). The data contains a 500 by 50 matrix: five hundred (500) rows by fifty (50) fields.
3. **Tool:** The computing resources used for this example were the MS Surface Book 2 for Business - 15" Display /256 GB / Intel Core i7. High-speed Intel processors, (quad-core available),NVIDIA GeForce GTX graphics, 17 hours of battery life, and running Windows 10 Pro.
4. **Tool:** The software used for this examples was Windows 10, Anaconda, IPython, and Jupyter Notebook

Reference:

1. Crawford, D. (1999, December). Introducing JSON. Retrieved November 2, 2019, from https://www.json.org/.
2. Creation of normalized dataframes (tables) from nested json string, ver. 025.3. (2019, October 31). Retrieved November 2, 2019, from https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html.
3. Github Repository: https://github.com/ahull002/wrangling_json.git
4. Data File Location: http://jsonstudio.jsonar.com/world_bank.html

[ ]:

1.1.2 Phase II. Data Understanding.

Load Libraries & Assign Variables, Load Data, Transform Data, Describe Data, and Explore Data

[2]:
```
# Load Libraries: Required files and variables

%matplotlib inline
import json
import requests
import pandas as pd
from pandas.io.json import json_normalize
from IPython.core.display import HTML


pd.set_option('display.max_rows', 2000)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
```

```
# Load Data: JSON data as a string to preview data
file = 'data/world_bank_projects.json'

with open(file) as f:
    raw = json.load(f)


# Transform Data: JSON file into pandas DataFrame. look at first 20␣
 ↪observations/records

json_df = pd.read_json(file)
json_df.head()
```

[2]:
```
                                    _id  approvalfy board_approval_month
boardapprovaldate                                    borrower           closingdate
country_namecode countrycode                                  countryname
countryshortname                                                  docty
envassesmentcategorycode  grantamt  ibrdcommamt       id  idacommamt
impagency                  lendinginstr lendinginstrtype  lendprojectcost
majorsector_percent                                    mjsector_namecode
mjtheme                                  mjtheme_namecode mjthemecode prodline
prodlinetext productlinetype                                    project_abstract
project_name                                          projectdocs
projectfinancialtype projectstatusdisplay  \
0  {'$oid': '52b213b38594d8a2be17c780'}        1999             November
2013-11-12T00:00:00Z  FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA
2018-07-07T00:00:00Z  Federal Democratic Republic of Ethiopia!$!ET          ET
Federal Democratic Republic of Ethiopia            Ethiopia  Project Information
Document,Indigenous People...                        C         0            0
P129828   130000000                     MINISTRY OF EDUCATION  Investment
Project Financing               IN        550000000  [{'Percent': 46, 'Name':
'Education'}, {'Perce...  [{'code': 'EX', 'name': 'Education'}, {'code':...
[Human development]  [{'code': '8', 'name': 'Human development'}, {...
8,11       PE                       IBRD/IDA               L  {'cdata': 'The
development objective of the Se...  Ethiopia General Education Quality
Improvement...  [{'DocDate': '28-AUG-2013', 'EntityID': '09022...
IDA               Active
1  {'$oid': '52b213b38594d8a2be17c781'}        2015             November
2013-11-04T00:00:00Z                    GOVERNMENT OF TUNISIA
NaN                     Republic of Tunisia!$!TN          TN
Republic of Tunisia             Tunisia  Project Information Document,Integrated
Safegu...                         C   4700000            0  P144674           0
MINISTRY OF FINANCE      Specific Investment Loan               IN
5700000  [{'Percent': 70, 'Name': 'Public Administratio...  [{'code': 'BX',
'name': 'Public Administration...  [Economic management, Social protection and
ri...  [{'code': '1', 'name': 'Economic management'},...         1,6       RE
Recipient Executed Activities               L
```

```
NaN          TN: DTF Social Protection Reforms Support  [{'DocDate':
'29-MAR-2013', 'EntityID': '00033...                OTHER               Active
2  {'$oid': '52b213b38594d8a2be17c782'}        2014             November
2013-11-01T00:00:00Z   MINISTRY OF FINANCE AND ECONOMIC DEVEL
NaN                                  Tuvalu!$!TV          TV
Tuvalu             Tuvalu  Resettlement Plan,Environmental Assessment,Int...
B         0            0  P145310     6060000  MINISTRY OF TRANSPORT AND
COMMUNICATIONS  Investment Project Financing               IN          6060000
[{'Percent': 100, 'Name': 'Transportation'}]         [{'code': 'TX', 'name':
'Transportation'}]  [Trade and integration, Public sector governan...  [{'code':
'5', 'name': 'Trade and integration'...    5,2,11,6       PE
IBRD/IDA               L                                                NaN
Tuvalu Aviation Investment Project - Additiona...  [{'DocDate': '21-OCT-2013',
'EntityID': '00033...                   IDA               Active
3  {'$oid': '52b213b38594d8a2be17c783'}        2014              October
2013-10-31T00:00:00Z   MIN. OF PLANNING AND INT'L COOPERATION
NaN                          Republic of Yemen!$!RY          RY
Republic of Yemen  Yemen, Republic of  Procurement Plan,Project Information
Document,...                          C   1500000            0  P144665
0  LABOR INTENSIVE PUBLIC WORKS PROJECT PMU     Technical Assistance Loan
IN          1500000  [{'Percent': 100, 'Name': 'Health and other so...
[{'code': 'JX', 'name': 'Health and other soci...  [Social dev/gender/inclusion,
Social dev/gende...  [{'code': '7', 'name': 'Social dev/gender/incl...
7,7       RE  Recipient Executed Activities               L
NaN   Gov't and Civil Society Organization Partnership  [{'DocDate':
'15-MAY-2013', 'EntityID': '00035...                OTHER               Active
4  {'$oid': '52b213b38594d8a2be17c784'}        2014              October
2013-10-31T00:00:00Z                       MINISTRY OF FINANCE
2019-04-30T00:00:00Z                        Kingdom of Lesotho!$!LS          LS
Kingdom of Lesotho             Lesotho  Project Information Document,Integrated
Safegu...                           B         0            0  P144933    13100000
MINISTRY OF TRADE AND INDUSTRY  Investment Project Financing               IN
15000000  [{'Percent': 50, 'Name': 'Industry and trade'}...  [{'code': 'YX',
'name': 'Industry and trade'},...  [Trade and integration, Financial and private
...  [{'code': '5', 'name': 'Trade and integration'...         5,4       PE
IBRD/IDA               L  {'cdata': 'The development objective of the Se...
Second Private Sector Competitiveness and Econ...  [{'DocDate': '06-SEP-2013',
'EntityID': '09022...                   IDA               Active

                     regionname
sector                                          sector1
sector2                                          sector3
sector4                                    sector_namecode   sectorcode source
status supplementprojectflg                                              theme1
theme_namecode    themecode   totalamt  totalcommamt
url
0                        Africa  [{'Name': 'Primary education'}, {'Name':
```

```
'Seco...       {'Percent': 46, 'Name': 'Primary education'}     {'Percent': 26,
'Name': 'Secondary education'}  {'Percent': 16, 'Name': 'Public
administration...  {'Percent': 12, 'Name': 'Tertiary education'}  [{'code':
'EP', 'name': 'Primary education'}, ...  ET,BS,ES,EP   IBRD  Active
N      {'Percent': 100, 'Name': 'Education for all'}      [{'code': '65',
'name': 'Education for all'}]           65  130000000     130000000
http://www.worldbank.org/projects/P129828/ethi...
1  Middle East and North Africa  [{'Name': 'Public administration- Other
social...  {'Percent': 70, 'Name': 'Public administration...  {'Percent': 30,
'Name': 'General public admini...
NaN                                            NaN  [{'code': 'BS', 'name':
'Public administration...        BZ,BS   IBRD  Active                    N
{'Percent': 30, 'Name': 'Other economic manage...  [{'code': '24', 'name':
'Other economic manage...        54,24          0       4700000
http://www.worldbank.org/projects/P144674?lang=en
2         East Asia and Pacific  [{'Name': 'Rural and Inter-Urban Roads and
Hig...  {'Percent': 100, 'Name': 'Rural and Inter-Urba...
NaN                                                NaN
NaN  [{'code': 'TI', 'name': 'Rural and Inter-Urban...           TI   IBRD
Active                    Y    {'Percent': 46, 'Name': 'Regional integration'}
[{'code': '47', 'name': 'Regional integration'...  52,81,25,47    6060000
6060000  http://www.worldbank.org/projects/P145310?lang=en
3  Middle East and North Africa                [{'Name': 'Other social
services'}]  {'Percent': 100, 'Name': 'Other social services'}
NaN                                                NaN
NaN  [{'code': 'JB', 'name': 'Other social services'}]           JB   IBRD
Active                    N  {'Percent': 50, 'Name': 'Participation and civ...
[{'code': '57', 'name': 'Participation and civ...        59,57          0
1500000  http://www.worldbank.org/projects/P144665?lang=en
4                        Africa  [{'Name': 'General industry and trade
sector'}...  {'Percent': 50, 'Name': 'General industry and ...
{'Percent': 40, 'Name': 'Other industry'}            {'Percent': 10, 'Name':
'SME Finance'}                                          NaN  [{'code': 'YZ',
'name': 'General industry and ...     FH,YW,YZ   IBRD  Active
N  {'Percent': 30, 'Name': 'Export development an...  [{'code': '45', 'name':
'Export development an...        41,45   13100000      13100000
http://www.worldbank.org/projects/P144933/seco...
```

[ ]:

[3]:
```
# Describe Data 1.0: Get a concise summary of the dataframe

json_df.shape
```

[3]:
```
(500, 50)
```

[ ]:

[4]:
```
# Describe Data 1.1: Get a concise summary of the dataframe

json_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 50 columns):
_id                         500 non-null object
approvalfy                  500 non-null int64
board_approval_month        500 non-null object
boardapprovaldate           500 non-null object
borrower                    485 non-null object
closingdate                 370 non-null object
country_namecode            500 non-null object
countrycode                 500 non-null object
countryname                 500 non-null object
countryshortname            500 non-null object
docty                       446 non-null object
envassesmentcategorycode    430 non-null object
grantamt                    500 non-null int64
ibrdcommamt                 500 non-null int64
id                          500 non-null object
idacommamt                  500 non-null int64
impagency                   472 non-null object
lendinginstr                495 non-null object
lendinginstrtype            495 non-null object
lendprojectcost             500 non-null int64
majorsector_percent         500 non-null object
mjsector_namecode           500 non-null object
mjtheme                     491 non-null object
mjtheme_namecode            500 non-null object
mjthemecode                 500 non-null object
prodline                    500 non-null object
prodlinetext                500 non-null object
productlinetype             500 non-null object
project_abstract            362 non-null object
project_name                500 non-null object
projectdocs                 446 non-null object
projectfinancialtype        500 non-null object
projectstatusdisplay        500 non-null object
regionname                  500 non-null object
sector                      500 non-null object
sector1                     500 non-null object
sector2                     380 non-null object
sector3                     265 non-null object
sector4                     174 non-null object
sector_namecode             500 non-null object
```

```
sectorcode                  500 non-null object
source                      500 non-null object
status                      500 non-null object
supplementprojectflg        498 non-null object
theme1                      500 non-null object
theme_namecode              491 non-null object
themecode                   491 non-null object
totalamt                    500 non-null int64
totalcommamt                500 non-null int64
url                         500 non-null object
dtypes: int64(7), object(43)
memory usage: 195.4+ KB
```

[ ]:

1.1.3 Answering Set Objectives.

Set objectives:

1. What are the top 10 countries with the most projects?

[7]:
```
# Defining themes while normalizing for the one to many relationship between: 'code', 'name', and 'countryname'
# Top 10 Major World Bank Project Themes

themes_df = json_normalize(raw, 'mjtheme_namecode', ['countryname'])
countries_t10_df = themes_df['countryname'].value_counts().head(10)
countries_t10_df
```

[7]:
```
Republic of Indonesia             56
Republic of India                 51
Socialist Republic of Vietnam     43
Federative Republic of Brazil     41
People's Republic of Bangladesh   41
People's Republic of China        40
Africa                            39
Republic of Yemen                 34
Kingdom of Morocco                32
Republic of Mozambique            31
Name: countryname, dtype: int64
```

[8]:
```
# Grouping on 'countryname' to analyze unique project themes count

ucountries_t10_df = themes_df.groupby('countryname').code.nunique().sort_values(ascending=False).head(10)
ucountries_t10_df
```

[8]:
```
countryname
Republic of Kenya                 10
Republic of Indonesia             10
```

```
Republic of Mozambique              10
Federative Republic of Brazil        9
People's Republic of China           9
Islamic State of Afghanistan         9
Nepal                                9
United Republic of Tanzania          9
Burkina Faso                         9
People's Republic of Bangladesh      8
Name: code, dtype: int64
```

[ ]:

2. What are the top 10 major project themes (using column 'mjtheme_namecode') you see in the data?

[9]:
```
# Top 10 Major World Bank Project Themes:'
project_t10_df = themes_df['name'].value_counts().head(10)
project_t10_df
```

[9]:
```
Environment and natural resources management    223
Rural development                               202
Human development                               197
Public sector governance                        184
Social protection and risk management           158
Financial and private sector development        130
                                                122
Social dev/gender/inclusion                     119
Trade and integration                            72
Urban development                                47
Name: name, dtype: int64
```

[ ]:

3. Finally, create a DataFrame with the missing names filled in.

[11]:
```
# Creating a dictionary to map code values with project theme names

project_dict = {}
for row in themes_df.itertuples():
    if row[2] != '':
        project_dict[row[1]] = row[2]

project_dict
```

[11]:
```
{'8': 'Human development',
 '1': 'Economic management',
 '6': 'Social protection and risk management',
 '5': 'Trade and integration',
 '2': 'Public sector governance',
 '11': 'Environment and natural resources management',
 '7': 'Social dev/gender/inclusion',
 '4': 'Financial and private sector development',
```