Amanda Devine
21 October 2019
WS61 Biodiversity Informatics 101: Preparing for Biodiversity_Next and Beyond
https://github.com/amdevine/bdn-python-apis (https://github.com/amdevine/bdn-python-apis)

Retrieving data from Web APIs using the Python Requests library

This Jupyter notebook walks through the process of using the Python Requests library, in conjunction with the Pandas library, to download data from website REST API services.

As an example dataset, this notebook uses data on the US National Parks available via the United States National Park Service Data API.

Definitions

- **API**: Application Programming Interface. A special page on a website that provides structured data for other programs and applications.
  - **REST**: Newer API format, easily accessed via URL, returns data in a variety of formats including JSON
  - **SOAP**: Older API format, more complex, accepts queries and returns data in XML only
- **GET Request**: An HTTP command to retrieve code and data from a website. GET requests can be made in a variety of ways; the Requests library offers a very easy way to make GET requests from Python.
- **JSON**: JavaScript Object Notation. A common format of structuring data, analogous to a Python dictionary.
- **Base URL**: The website URL for all API data. A variety of endpoints can be added to the base URL.

  > NPS Base URL: `https://developer.nps.gov/api/v1`

- **Endpoint**: The specific URL where the API page can be found. Each website might have multiple endpoints that return different kinds of data.

  > Parks Endpoint: `https://developer.nps.gov/api/v1/parks`

- **Parameter**: An additional criterion that is added to the endpoint to filter data returned. Parameters are usually added to the endpoint with a `?` character, and are in the format of `field=value`. Multiple parameters can be added to an endpoint, separated by a `&`.

  > parkCode, stateCode, and limit parameters: `https://developer.nps.gov/api/v1/parks?parkCode=yell&stateCode=WY&limit=5`

- **API Key**: A string of characters assigned by the website to identify the user requesting data via the API. For many API services, an API Key is required when making a request.

  > National Parks API Key: `https://developer.nps.gov/api/v1/parks?api_key=1mdaBewB37R0kUA2ZtfA6URe7PeUsig6jLQmSXyx` (not a real key!)

NPS Data API

The National Park Service (NPS) Data API is a service provided by the United States National Park Service to supply data about natural areas managed by the National Park Service. Data available through the API include

- park information
- campground information
- alerts, events, news, and educational resources.

All users of the NPS Data API are required to have an API Key. Keys can be obtained here: https://www.nps.gov/subjects/developer/get-started.htm (https://www.nps.gov/subjects/developer/get-started.htm)

API documentation is available here: https://www.nps.gov/subjects/developer/api-documentation.htm (https://www.nps.gov/subjects/developer/api-documentation.htm)
This documentation shows a list of possible endpoints. Clicking on an endpoint shows the parameters that can be supplied to that endpoint to filter results, as well as an example of the JSON data that will be returned from a GET request.

The NPS Data API GitHub Repository contains examples of using Python and/or PHP to retrieve data. It is available here: https://github.com/nationalparkservice/nps-api-samples (https://github.com/nationalparkservice/nps-api-samples)

Python `requests` library

The Python Requests library provides a simple way to query and retrieve JSON data from API services. It is a wrapper that calls other Python libraries such as `urllib`.

Requests can be download via Terminal/Command Line with `conda install requests` (if you are running Anaconda/Miniconda) or `pip install requests` (if you are running base Python).

Making a GET request to retrieve JSON data with Requests is usually done in the following way:

```
import requests
url = 'https://baseurl.com/endpoint'
params = {
    'field1': 'value1',
    'field2': 'value2',
}
r = requests.get(url, params).json()
```

This creates a dictionary, `r`, from which data can be accessed via key-value pairs.

Parameters supplied to a GET request are automatically encoded to HTML specifications.

> e.g. the parameter `'q': 'César E. Chávez'` is automatically encoded to `q=C%C3%A9sar%20E.%20Ch%C3%A1vez` by Requests

Requests has significantly more functionality and many more options beyond simple GET requests. Quickstart documentation is available here: https://2.python-requests.org/en/master/user/quickstart/ (https://2.python-requests.org/en/master/user/quickstart/)

Setup

Import the `requests` library to retrieve data from the NPS Data API. Import the `pandas` library to work with retrieved data and export as tabular files.

In [1]:
```
import requests
import pandas as pd

import pprint   # Prints dictionaries/JSON in a more human-readable format
```

Assign the API Key to a constant. If the code will be available on GitHub or other public sites, avoid assigning it directly and import it from a local file instead. (And don't add this file to GitHub!)

In [2]:
```
# API_KEY = '1mdaBewB37R0kUA2ZtfA6URe7PeUsig6jLQmSXyx'
with open('api_key_file.txt', 'r') as f:
    API_KEY = f.read().strip()
print("API Key: {}".format("API_KEY")) # Remove quotes to display actual API_KEY
```

```
API Key: API_KEY
```

Make a GET request to the API to retrieve data

Use `requests.get()` to make an **HTTP GET Request** to the API. Any parameters can be provided to requests.get() as a dictionary.

The following request will return data on up to 100 parks in the state of California. `api_key` is a required parameter for all NPS Data API requests. `stateCode` filters parks based on two-letter US state abbreviations. `fields` specifies additional fields to return in addition to the default fields. `limit` specifies the maximum number of results to return.

In [3]:
```
url = 'https://developer.nps.gov/api/v1/parks'
params = {
    'api_key': API_KEY,
    'stateCode': 'CA',
    'fields': 'entranceFees',
    'limit': 100
}
r = requests.get(url, params)
```

`requests.get()` returns a variety of information about the web page retrieved. This info can be useful for troubleshooting.

In [4]:
```
print("The response code is: {}".format(r.status_code))
print("\nThe retrieved URL is: {}".format("r.url")) # Remove quotes to display URL
print("\nThe first 300 characters of the retrieved text are:\n{}".format(r.text[:300]))
```

```
The response code is: 200

The retrieved URL is: r.url

The first 300 characters of the retrieved text are:
{"total":"33","data":[{"states":"CA","entranceFees":[{"cost":"39.9000","description":"A daily scheduled ferry
to Alcatraz has a round-trip fee. Reservations are strongly recommended as we often sell out. See www.alcatra
zcruises.com for reservations, and ticket types and pricing.","title":"Adult Day
```

Work with retrieved data

Convert GET request object to dictionary

When the API data are supplied in the **JSON** format, they can easily be turned into a Python dictionary using the Requests `.json()` method.

In [5]:
```
parks_data = r.json()

print("First item in 'data':\n")
pprint.pprint(parks_data['data'][0])
```

First item in 'data':


```
{'description': 'Alcatraz reveals stories of American incarceration, justice, '
                'and our common humanity. This small island was once a fort, a '
                'military prison, and a maximum security federal penitentiary. '
                'In 1969, the Indians of All Tribes occupied Alcatraz for 19 '
                'months in the name of freedom and Native American civil '
                "rights. We invite you to explore Alcatraz's complex history "
                'and natural beauty.',
 'designation': '',
 'directionsInfo': 'The Alcatraz Ferry Terminal is located on The Embarcadero '
                   'near the intersection of Bay Street at Pier 33.',
 'directionsUrl': 'http://home.nps.gov/alca/planyourvisit/directions.htm',
 'entranceFees': [{'cost': '39.9000',
                   'description': 'A daily scheduled ferry to Alcatraz has a '
                                  'round-trip fee. Reservations are strongly '
                                  'recommended as we often sell out. See '
                                  'www.alcatrazcruises.com for reservations, '
                                  'and ticket types and pricing.',
                   'title': 'Adult Day Ticket (Ferry plus audio tour)'},
                  {'cost': '47.3000',
                   'description': 'A daily scheduled ferry to Alcatraz has a '
                                  'round-trip fee. Reservations are strongly '
                                  'recommended as we often sell out. See '
                                  'www.alcatrazcruises.com for reservations, '
                                  'and ticket types and pricing.',
                   'title': 'Adult Night Ticket (Ferry plus audio tour)'},
                  {'cost': '92.3000',
                   'description': 'A daily scheduled ferry to Alcatraz has a '
                                  'round-trip fee. Reservations are strongly '
                                  'recommended as we often sell out. See '
                                  'www.alcatrazcruises.com for reservations, '
                                  'and ticket types and pricing.',
                   'title': 'Adult Behind the Scenes Tour Ticket (Ferry plus '
                            'audio tour)'}],
 'fullName': 'Alcatraz Island',
 'id': 'C08AD828-98FF-478E-A63C-614E7534274B',
 'latLong': 'lat:37.82676234, long:-122.4230206',
 'name': 'Alcatraz Island',
 'parkCode': 'alca',
 'states': 'CA',
 'url': 'https://www.nps.gov/alca/index.htm',
 'weatherInfo': 'The climate on Alcatraz is unpredictable and can change '
```

```
'suddenly. Cold, foggy mornings may give way to sunny '
'afternoons, which in turn can shift quickly back to more fog '
'and blustery winds. The most pleasant weather usually occurs '
'in spring and fall. Summers tend to be cool and foggy, winter '
'is our rainy season. Temperatures on Alcatraz seldom rise '
'above 75°F (24°C) or fall below 38°'}
```

Create a Pandas DataFrame

Pandas DataFrames make it easy to work with the retrieved data in a tabular format.

This code filters the retrieved data to states and associated lat/long coordinate for each park.

In [6]:
```
parks_df = pd.DataFrame(parks_data['data'])
locations_df = parks_df[['parkCode', 'fullName', 'designation', 'states', 'latLong']]
locations_df.head(10)
```

Out[6]:

	parkCode	fullName	designation	states	latLong
0	alca	Alcatraz Island		CA	lat:37.82676234, long:-122.4230206
1	cabr	Cabrillo National Monument	National Monument	CA	lat:32.6722503, long:-117.2415985
2	cali	California National Historic Trail	National Historic Trail	CA,CO,ID,KS,MO,NE,NV,OR,UT,WY	
3	camo	Castle Mountains National Monument	National Monument	CA	lat:35.29156348, long:-115.0935606
4	cech	César E. Chávez National Monument	National Monument	CA	lat:35.22729389, long:-118.5615781
5	chis	Channel Islands National Park	National Park	CA	lat:33.98680093, long:-119.9112735
6	deva	Death Valley National Park	National Park	CA,NV	lat:36.48753731, long:-117.134395
7	depo	Devils Postpile National Monument	National Monument	CA	lat:37.6152564, long:-119.0873903
8	euon	Eugene O'Neill National Historic Site	National Historic Site	CA	lat:37.82604456, long:-122.0271566
9	fopo	Fort Point National Historic Site	National Historic Site	CA	lat:37.80837439, long:-122.473747

Restructure/flatten data

If the results contain nested data that need to be flattened (e.g. multiple `entranceFees` for each park), or the results could be otherwise restructured in a more "tidy" format, a new list of dictionaries can be created by iterating through the data. This list can then be converted to a DataFrame.

JSON data for Yosemite National Park's multiple entrance fees:

In [7]:
```
pprint.pprint(parks_data['data'][-1])
```

```
{'description': 'Not just a great valley, but a shrine to human foresight, the '
                'strength of granite, the power of glaciers, the persistence '
                'of life, and the tranquility of the High Sierra.\n'
                '\n'
                'First protected in 1864, Yosemite National Park is best known '
                'for its waterfalls, but within its nearly 1,200 square miles, '
                'you can find deep valleys, grand meadows, ancient giant '
                'sequoias, a vast wilderness area, and much more.',
 'designation': 'National Park',
 'directionsInfo': 'You can drive to Yosemite all year and enter via Highways '
                   '41, 140, and 120 from the west. Tioga Pass Entrance (via '
                   'Highway 120 from the east) is closed from around November '
                   'through late May or June. Hetch Hetchy is open all year '
                   'but may close intermittently due to snow.\n'
                   '\n'
                   'Please note that GPS units do not always provide accurate '
                   'directions to or within Yosemite.',
 'directionsUrl': 'http://www.nps.gov/yose/planyourvisit/driving.htm',
 'entranceFees': [{'cost': '35.0000',
                   'description': 'This fee is valid for seven days.',
                   'title': 'Non-commercial car, pickup truck, RV, or van with '
                            '15 or fewer passenger seats'},
                  {'cost': '30.0000',
                   'description': 'The fee is valid for seven days. Cost is '
                                  'per motorcycle (not per person).',
                   'title': 'Motorcycle'},
                  {'cost': '20.0000',
                   'description': 'This fee is valid for seven days. People 15 '
                                  'years and younger are free. Cost is per '
                                  'person.',
                   'title': 'Foot, bicycle, horse, or non-commercial bus or '
                            'van with more than 15 passenger seats'},
                  {'cost': '25.0000',
                   'description': 'The fee is $25 plus $15 per person.',
                   'title': 'Commercial Tour (sedan up to six seats)'},
                  {'cost': '125.0000',
                   'description': '',
                   'title': 'Commercial Tour (van, 7-15 seats, regardless of '
                            'occupancy)'},
                  {'cost': '200.0000',
                   'description': '',
                   'title': 'Commercial Tour (mini bus, 16-25 seats, '
                            'regardless of occupancy)'},
```

```
                {'cost': '300.0000',
                 'description': '',
                 'title': 'Commercial Tour (motor coach, 26 or more seats, '
                          'regardless of occupancy)'}],
 'fullName': 'Yosemite National Park',
 'id': '4324B2B4-D1A3-497F-8E6B-27171FAE4DB2',
 'latLong': 'lat:37.84883288, long:-119.5571873',
 'name': 'Yosemite',
 'parkCode': 'yose',
 'states': 'CA',
 'url': 'https://www.nps.gov/yose/index.htm',
 'weatherInfo': 'Yosemite National Park covers nearly 1,200 square miles '
                '(3,100 square km) in the Sierra Nevada, with elevations '
                'ranging from about 2,000 feet (600 m) to 13,000 ft (4,000 m). '
                'Yosemite receives 95% of its precipitation between October '
                'and May (and over 75% between November and March). Most of '
                'Yosemite is blanketed in snow from about November through '
                'May. (Yosemite Valley can be rainy or snowy in any given '
                'winter storm.)'}
```

For each park in the dataset, and for each entrance fee in that park, add some park and fee values as a dictionary to a new `entry_fee_data` list.

In [8]:
```
entry_fees_data = []

for park in parks_data['data']:
    for fee in park['entranceFees']:
        entry_fees_data.append({
            'parkCode': park['parkCode'],
            'fullName': park['fullName'],
            'designation': park['designation'],
            'fee_usd': fee['cost'],
            'fee_type': fee['title'],
            'fee_description': fee['description']
        })

pprint.pprint(entry_fees_data[2:4])
```

```
[{'designation': '',
  'fee_description': 'A daily scheduled ferry to Alcatraz has a round-trip '
                     'fee. Reservations are strongly recommended as we often '
                     'sell out. See www.alcatrazcruises.com for reservations, '
                     'and ticket types and pricing.',
  'fee_type': 'Adult Behind the Scenes Tour Ticket (Ferry plus audio tour)',
  'fee_usd': '92.3000',
  'fullName': 'Alcatraz Island',
  'parkCode': 'alca'},
 {'designation': 'National Monument',
  'fee_description': 'The pass is valid for seven full days.',
  'fee_type': 'Cabrillo Entrance Fee - Per non commercial vehicle',
  'fee_usd': '20.0000',
  'fullName': 'Cabrillo National Monument',
  'parkCode': 'cabr'}]
```

Convert `entry_fee_data` to a DataFrame

In [9]:
```
entry_fees_df = pd.DataFrame(entry_fees_data)
entry_fees_df = entry_fees_df[['parkCode', 'fullName', 'designation', 'fee_usd', 'fee_type']]
entry_fees_df['fee_usd'] = entry_fees_df['fee_usd'].astype(float)
entry_fees_df.head(10)
```

Out[9]:

	parkCode	fullName	designation	fee_usd	fee_type
0	alca	Alcatraz Island		39.9	Adult Day Ticket (Ferry plus audio tour)
1	alca	Alcatraz Island		47.3	Adult Night Ticket (Ferry plus audio tour)
2	alca	Alcatraz Island		92.3	Adult Behind the Scenes Tour Ticket (Ferry plu...
3	cabr	Cabrillo National Monument	National Monument	20.0	Cabrillo Entrance Fee - Per non commercial veh...
4	cabr	Cabrillo National Monument	National Monument	15.0	Cabrillo Entrance Fee - per motorcycle
5	cabr	Cabrillo National Monument	National Monument	10.0	Cabrillo Entrance Fee - per walker or bicyclist
6	cabr	Cabrillo National Monument	National Monument	30.0	Cabrillo Entrance Fee - per commercical vehicl...
7	cabr	Cabrillo National Monument	National Monument	45.0	Cabrillo Entrance Fee - per commercial vehicle...
8	cabr	Cabrillo National Monument	National Monument	100.0	Cabrillo Entrance Fee - per commercial vehicle...
9	cali	California National Historic Trail	National Historic Trail	0.0	Entrance fees vary site by site

Export data as a tabular file

Pandas DataFrames have a method, `.to_csv()`, that allows them to be exported as a CSV or TSV file. This file can be imported into another program for further analysis.

CSV file: `df_name.to_csv('output_file_name.csv', index=False)`

TSV file: `df_name.to_csv('output_file_name.tsv', sep='\t', index=False)`

In [10]:
```
locations_df.to_csv('parks_data.tsv', sep='\t', index=False)
entry_fees_df.to_csv('parks_entry_fees.tsv', sep='\t', index=False)
```

Additional API Resources

Full **Requests documentation**: https://2.python-requests.org/en/master (https://2.python-requests.org/en/master)

data.gov is a repository of data produced by US Federal and State Government departments. In addition to data files, data.gov also provides a list of API services for Federal data (https://catalog.data.gov/dataset?res_format=API (https://catalog.data.gov/dataset?res_format=API)).

Programmable Web (https://www.programmableweb.com/ (https://www.programmableweb.com/)) is a website that aggregates information on APIs provided by governmental and private organizations. This can be a good resource for locating APIs of interest.