Harvest data from Web APIs using the Python Requests library

Amanda Devine
25 July 2019
SI Carpentries Brown Bag

GitHub Repository: https://github.com/amdevine/cbb-python-requests (https://github.com/amdevine/cbb-python-requests)
Detailed Jupyter notebook: https://github.com/amdevine/cbb-python-requests/blob/master/harvest-data-apis-python-requests.ipynb (https://github.com/amdevine/cbb-python-requests/blob/master/harvest-data-apis-python-requests.ipynb)
Presentation slides: https://github.com/amdevine/cbb-python-requests/blob/master/python-requests-slides.pdf (https://github.com/amdevine/cbb-python-requests/blob/master/python-requests-slides.pdf) (run with RISE extension)

Definitions

- **(REST) API**: Application Programming Interface. A special page on a website that provides structured data for other programs and applications.
- **GET Request**: An HTTP command to retrieve code and data from a website.
- **JSON**: JavaScript Object Notation. A common format of structuring data, analogous to a Python dictionary.
- **Base URL**: The "home" website URL for all API data.

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

- **Endpoint**: The specific URL where the API page can be found.

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

- **Parameter**: An additional criterion that is added to the endpoint to filter data returned.

    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.

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

NPS Data API

Official source of data about natural areas managed by the National Park Service

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

NPS API Keys: https://www.nps.gov/subjects/developer/get-started.htm (https://www.nps.gov/subjects/developer/get-started.htm)

NPS Data API documentation: https://www.nps.gov/subjects/developer/api-documentation.htm (https://www.nps.gov/subjects/developer/api-documentation.htm)

Setup

Import the `requests` and `pandas` libraries.

In [1]:
```
import requests
import pandas as pd
import pprint # Prints dictionaries/JSON in a more readable format
```

Save API Key as a constant or read it from a local file.

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

This request returns data on up to 100 parks in Washington DC, Maryland, and Virginia.

In [3]:
```
url = 'https://developer.nps.gov/api/v1/parks'
params = {
    'api_key': API_KEY,
    'stateCode': 'DC,MD,VA', # Per the API documentation, separate multiple values with
 commas
    'fields': 'entranceFees',
    'limit': 100
}
r = requests.get(url, params)
```

`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.

Work with retrieved data

Convert GET request object to dictionary

In [4]:
```
parks_data = r.json()

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

```
First item in 'data':

{'description': 'Over 200,000 African-American soldiers and sailors served in '
                'the U.S. Army and Navy during the Civil War. Their service '
                'helped to end the war and free over four million slaves. The '
                'African American Civil War Memorial honors their service and '
                'sacrifice.',
 'designation': '',
 'directionsInfo': 'The memorial is located at the corner of Vermont Avenue, '
                   '10th St, and U Street NW, near the U '
                   'Street/African-American Civil War Memorial/Cardozo Metro '
                   'Station.',
 'directionsUrl': 'http://www.nps.gov/afam/planyourvisit/directions.htm',
 'entranceFees': [{'cost': '0.0000',
                   'description': 'No Entrance Fee to enter park site.',
                   'title': 'No Entrance Fee'}],
 'fullName': 'African American Civil War Memorial',
 'id': '1A47416F-DAA3-4137-9F30-14AF86B4E547',
 'latLong': 'lat:38.916554, long:-77.025977',
 'name': 'African American Civil War Memorial',
 'parkCode': 'afam',
 'states': 'DC',
 'url': 'https://www.nps.gov/afam/index.htm',
 'weatherInfo': 'Washington DC gets to see all four seasons. Humidity will '
                'make the temps feel hotter in summer and colder in winter.\n'
                '\n'
                'Spring (March - May) Temp: Average high is 65.5 degrees with '
                'a low of 46.5 degrees\n'
                '\n'
                'Summer (June - August) Temp: Average high is 86 degrees with '

                'a low of 68.5 degrees\n'
```

Create a DataFrame

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

In [5]:
```
parks_df = pd.DataFrame(parks_data['data'])
locations_df = parks_df[['parkCode', 'fullName', 'designation', 'states', 'latLong']]
locations_df.head(10)
```

Out[5]:

	parkCode	fullName	designation	
0	afam	African American Civil War Memorial		DC
1	anac	Anacostia Park	Park	DC
2	anti	Antietam National Battlefield	National Battlefield	MD
3	appa	Appalachian National Scenic Trail	National Scenic Trail	CT,GA,MA,MD,ME,NC,NH,NJ,NY,PA,TN,V
4	apco	Appomattox Court House National Historical Park	National Historical Park	VA

	parkCode	fullName	designation	
5	arho	Arlington House, The Robert E. Lee Memorial		VA
6	asis	Assateague Island National Seashore	National Seashore	MD,VA
7	balt	Baltimore National Heritage Area	National Heritage Area	MD
8	bawa	Baltimore-Washington Parkway	Parkway	MD
9	bepa	Belmont-Paul Women's Equality National Monument	National Monument	DC

Restructure/flatten data

Retrieved JSON data for an individual park's multiple entrance fees.

In [6]:
```
pprint.pprint(parks_data['data'][2])
```

```
{'description': '23,000 soldiers were killed, wounded or missing after twelve '
                'hours of savage combat on September 17, 1862. The Battle of '
                "Antietam ended the Confederate Army of Northern Virginia's "
                'first invasion into the North and led Abraham Lincoln to '
                'issue the preliminary Emancipation Proclamation.',
 'designation': 'National Battlefield',
 'directionsInfo': 'Ten miles south of I-70 on Maryland Route 65',
 'directionsUrl': 'http://www.nps.gov/anti/planyourvisit/directions.htm',
 'entranceFees': [{'cost': '7.0000',
                   'description': '3 day pass - $7.00 per bike or motorcycle \n'
                                  'This is the entry fee to the battlefield '
                                  'proper, museum, movie, and ranger programs.',
                   'title': 'Antietam National Battlefield Entrance Fee'},
                  {'cost': '15.0000',
                   'description': '3 day vehicle pass.  This pass covers '
                                  'everyone in a vehicle, ie. family.  The '
                                  'pass covers entry to the battlefield '
                                  'proper, museum, movie, and ranger programs.',
                   'title': 'Antietam National Battlefield Entrance Fee'}],
 'fullName': 'Antietam National Battlefield',
 'id': '8415526C-C932-4236-A634-2D89DF718936',
 'latLong': 'lat:39.46763452, long:-77.73828017',
 'name': 'Antietam',
 'parkCode': 'anti',
 'states': 'MD',
 'url': 'https://www.nps.gov/anti/index.htm',
 'weatherInfo': 'The weather is fairly mild.  Summers can be very warm and '
                'humid and winters cold and snowy.  We have four distinct '
                'seasons with the fall and spring being the best times to '
                'visit the battlefield.'}
```

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 [7]:
```
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': 'National Battlefield',
  'fee_description': '3 day pass - $7.00 per bike or motorcycle \n'
                     'This is the entry fee to the battlefield proper, museum, '
                     'movie, and ranger programs.',
  'fee_type': 'Antietam National Battlefield Entrance Fee',
  'fee_usd': '7.0000',
  'fullName': 'Antietam National Battlefield',
  'parkCode': 'anti'},
 {'designation': 'National Battlefield',
  'fee_description': '3 day vehicle pass.  This pass covers everyone in a '
                     'vehicle, ie. family.  The pass covers entry to the '
                     'battlefield proper, museum, movie, and ranger programs.',
  'fee_type': 'Antietam National Battlefield Entrance Fee',
  'fee_usd': '15.0000',
  'fullName': 'Antietam National Battlefield',
  'parkCode': 'anti'}]
```

Convert `entry_fee_data` to a DataFrame

In [8]:
```
entry_fees_df = pd.DataFrame(entry_fees_data)
entry_fees_df = entry_fees_df[['parkCode', 'fullName', 'designation', 'fee_usd', 'fee_ty
pe']]
entry_fees_df['fee_usd'] = entry_fees_df['fee_usd'].astype(float)
entry_fees_df.head(10)
```

Out[8]:

	parkCode	fullName	designation	fee_usd	fee_type
0	afam	African American Civil War Memorial		0.0	No Entrance Fee
1	anac	Anacostia Park	Park	0.0	Entrance Fees
2	anti	Antietam National Battlefield	National Battlefield	7.0	Antietam National Battlefield Entrance Fee
3	anti	Antietam National Battlefield	National Battlefield	15.0	Antietam National Battlefield Entrance Fee
4	appa	Appalachian National Scenic Trail	National Scenic Trail	0.0	Appalachian National Scenic Trail Entrance Fee

Export data as a tabular file

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 [9]:
```
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/)

List of US Federal Government APIs: https://catalog.data.gov/dataset?res_format=API (https://catalog.data.gov/dataset?res_format=API)

Repository of APIs: https://www.programmableweb.com/ (https://www.programmableweb.com/)