Web Scraping with JSON and Python

Overview

Web scraping is the process of extracting data from the internet.

Common web data formats:

- HTML
- JSON
- CSV/TSV

Basic Web Request

Pulling information from the web:

```
import requests
response =
requests.get("https://en.wikipedia.org/robots.txt
")
txt = response.text
print(txt)
```

HTML Overview

See Web site HTML by using View Source

Different objects enclosed in tags

- Open tag <>
- Close tag </>

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Basic HTML page

```
<!DOCTYPE html>
<html>
<body>

<h1> Heading 1 </h1>
<h2> Heading 2 </h2>

<p> Paragraph 1 </p>
<p> Here's our example HTML page.</p>

<ul>
<li> item 1 </li>
<li> item 2 </li>
</ul>


</body>
</html>
```

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Scraping HTML– install packages

Install Python Packages:

- beauitifulsoup4
- requests

For more information on HTML format:
https://www.w3schools.com/html/

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Scraping HTML

```
import requests
from bs4 import BeautifulSoup

url = "https://wiki.python.org/moin/IntroductoryBooks"

response = requests.get(url)

content = response.content

soup = BeautifulSoup(content, "lxml")

all_a = soup.find_all("a")

all_a_https = soup.find_all("a", "https")

for x in all_a_https:
    print(x)
```

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Scraping JSON

Scraping data from the JSON format is even easier than parsing raw HTML.

Scraping CSV/TSV format

CSV and TSV files are some of the most commonly used formats in data science.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Scraping CSV Data

Your Turn

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

CSV to Pandas Dataframe

Convert the following web page into a pandas dataframe and add meaningful column headers:

Mammographic Masses Database

Solution

```
import pandas as pd

# Mammographic Masses URL
url = http://archive.ics.uci.edu/ml/machine-learning-
databases/mammographic-masses/mammographic_masses.data

# use pandas to convert csv into a dataframe
Mamm = pd.read_csv(url, header=None)

# add a list of column headers
Mamm.columns = ["BI-RADS", "Age", "Shape", "Margin",
"Density", "Severity"]
```

Summary

>We learned the basic function for web scraping

>Basic HTML structure

>Applied web scraping HTML, JSON and CSV formats

Web Scraping