Mission_to_Mars_Midlomarie

October 21, 2019

0.1 UR Data Analytics Homework #12: Web Scraping

0.1.1 Introduction to Beautiful Soup, Splinter, and MongoDB

Submitted by MidloMarie, October, 2019

0.2 *Mission to Mars*

In this assignment, we build a web application that scrapes various websites for data related to NASA's Mission to Mars and displays the information in a single HTML page.

[1]:
```
# Dependencies and Setup
from bs4 import BeautifulSoup
from splinter import Browser
import pandas as pd
```

[2]:
```
# Set Executable Path & Initialize Chrome Browser to view and control desired␣
 ↪Web pages
executable_path = {"executable_path": "./chromedriver.exe"}
browser = Browser("chrome", **executable_path)
```

[3]:
```
# Visit the NASA Mars News Site
url = "https://mars.nasa.gov/news/"
browser.visit(url)
```

0.2.1 What's going on with NASA Mars missions?

Let's look at the News on the mars.nasa.gov web page and get the most recent article > From "inspection" of activated NASA Mars Web page and using Devtools, we note that the Title > and "teaser" body of each article are found under the > ul class="item list" li class=slide > div class="content title" > div class="article_teaser_body"

[6]:
```
# Parse Results HTML with BeautifulSoup
#   <ul class="item_list">
#     <li class="slide">

html = browser.html
NASAnews_soup = BeautifulSoup(html, "html.parser")
grid_element = NASAnews_soup.select_one("ul.item_list li.slide")
```

[7]:
```
print(grid_element.prettify())
```

```
<li class="slide">
 <div class="image_and_description_container">
  <a href="/news/8531/mars-2020-unwrapped-and-ready-for-more-testing/"
target="_self">
   <div class="rollover_description">
    <div class="rollover_description_inner">
     In time-lapse video, bunny-suited engineers remove the inner layer of
protective foil on NASA's Mars 2020 rover after it was relocated for testing.
    </div>
    <div class="overlay_arrow">
     <img alt="More" src="/assets/overlay-arrow.png"/>
    </div>
   </div>
   <div class="list_image">
    <img alt="Mars 2020 Unwrapped and Ready for Testing: In time-lapse video
bunny-suited engineers remove the inner layer of protective foil on NASA's Mars
2020 rover after it was moved to a different building at JPL for testing."
src="/system/news_items/list_view_images/8531_PIA23467-320x240.gif"/>
   </div>
   <div class="bottom_gradient">
    <div>
     <h3>
      Mars 2020 Unwrapped and Ready for More Testing
     </h3>
    </div>
   </div>
  </a>
  <div class="list_text">
   <div class="list_date">
    October 18, 2019
   </div>
   <div class="content_title">
    <a href="/news/8531/mars-2020-unwrapped-and-ready-for-more-testing/"
target="_self">
     Mars 2020 Unwrapped and Ready for More Testing
    </a>
   </div>
   <div class="article_teaser_body">
    In time-lapse video, bunny-suited engineers remove the inner layer of
protective foil on NASA's Mars 2020 rover after it was relocated for testing.
   </div>
  </div>
 </div>
</li>
```

[8]:
```
# Now find just the title of the latest article (first one in the list) and␣
 ↪article text
news_date = grid_element.find("div", class_="list_date").get_text()
news_title = grid_element.find("div", class_="content_title").get_text()
news_teaser = grid_element.find("div", class_="article_teaser_body").get_text()

print(f"From mars.nasa.gov on {news_date} we learn that: \n\t'{news_title}'")
print(f"\t{news_teaser}")
```

```
From mars.nasa.gov on October 18, 2019 we learn that:
        'Mars 2020 Unwrapped and Ready for More Testing'
        In time-lapse video, bunny-suited engineers remove the inner layer of
protective foil on NASA's Mars 2020 rover after it was relocated for testing.
```

0.2.2 What does Mars look like? Any featured images associated with our article?

Let's look at the images on the jpl.nasa.gov web page and look at recent images > From "inspection" of activated NASA JPL Web page and using Devtools, we find the > featured image at the top of the page is id'd as a "full_image". > The featured image may not be the same as the latest news article on the NASA news page.

[9]:
```
## Now we look for space imagery from NASA Jet Propulsion Laboratory Featured␣
 ↪Space Image site
executable_path = {"executable_path": "./chromedriver.exe"}
browser = Browser("chrome", **executable_path)
url = "https://www.jpl.nasa.gov/spaceimages/?search=&category=Mars"
browser.visit(url)
```

[10]:
```
# Use Splinter to find the featured image by its id='full_image' in the HTML␣
 ↪code
# <button class="full_image">Full Image</button>
full_image_button = browser.find_by_id("full_image")
full_image_button.click()
```

[11]:
```
# Find "More Info" Button and Click It
browser.is_element_present_by_text("more info", wait_time=1)
more_info_element = browser.find_link_by_partial_text("more info")
more_info_element.click()
```

[21]:
```
# Parse Results HTML with BeautifulSoup
html = browser.html
image_soup = BeautifulSoup(html, "html.parser")

img_url = image_soup.select_one("figure.lede a img").get("src")
img_url = f"https://www.jpl.nasa.gov{img_url}"
img_url
```

[21]:
```
'https://www.jpl.nasa.gov/spaceimages/images/largesize/PIA07137_hires.jpg'
```

0.2.3 Now let's find out about Martian weather from the Mars Twitter account

[31]:
```
# Set Executable Path & Initialize Chrome Browser to view and control desired␣
 ↪Web pages
executable_path = {"executable_path": "./chromedriver.exe"}
browser = Browser("chrome", **executable_path)
url = "https://twitter.com/marswxreport?lang=en"
browser.visit(url)
```

[32]:
```
# Parse Results HTML with BeautifulSoup
html = browser.html
weather_soup = BeautifulSoup(html, "html.parser")
# print(weather_soup.prettify())
```

[33]:
```
# Find a Tweet with the data-name `Mars Weather`
mars_weather_tweet = weather_soup.find("div",
                                       attrs={
                                           "class": "tweet",
                                            "data-name": "Mars Weather"
                                        })
# print(mars_weather_tweet.prettify())
```

[34]:
```
# Search Within Tweet for <p> Tag Containing Tweet Text
mars_weather = mars_weather_tweet.find("p", "tweet-text").get_text()
print(mars_weather)
```

```
InSight sol 319 (2019-10-19) low -101.5žC (-150.7žF) high -25.5žC (-13.9žF)
winds from the SSE at 4.6 m/s (10.4 mph) gusting to 18.4 m/s (41.2 mph)
pressure at 7.10 hPapic.twitter.com/gdBUdujdVM
```

[ ]:
```
## Now look at Mars Facts site to scrape the table for data about the planet␣
 ↪including size, mass.
* Use Pandas to convert the data to an HTML table string
```

[35]:
```
mars_df = pd.read_html("https://space-facts.com/mars/")[0]
print(mars_df)
mars_df.columns=["Description", "Mars", "Earth"]
# mars_df
mars_facts_df=mars_df.drop(columns=["Earth"])
mars_facts_df.set_index("Description",inplace=True)
mars_facts_df
```

	Mars - Earth Comparison	Mars	Earth
0	Diameter:	6,779 km	12,742 km
1	Mass:	6.39 ℇ 10^23 kg	5.97 ℇ 10^24 kg
2	Moons:	2	1
3	Distance from Sun:	227,943,824 km	149,598,262 km
4	Length of Year:	687 Earth days	365.24 days
5	Temperature:	-153 to 20 řC	-88 to 58řC

[35]:
```
                                Mars
Description
Diameter:                    6,779 km
Mass:                6.39 ℰ 10^23 kg
Moons:                             2
Distance from Sun:     227,943,824 km
Length of Year:        687 Earth days
Temperature:            -153 to 20 řC
```

[36]:
```
# Output table in HTML format
mars_facts_df.to_html(open('mars_facts.html', 'w'))
```

0.3 Look for images of Mars Hemispheres

The two hemispheres of Mars are dramatically different from each other—a characteristic not seen on any other planet in our solar system. Non-volcanic, flat lowlands characterize the northern hemisphere, while highlands punctuated by countless volcanoes extend across the southern hemisphere.Jan 29, 2015 https://www.futurity.org/mars-hemispheres-846802/

[37]:
```
# Visit the USGS Astrogeology Science Center Site
executable_path = {"executable_path": "./chromedriver.exe"}
browser = Browser("chrome", **executable_path, headless=False)
url = "https://astrogeology.usgs.gov/search/results?
 ↪q=hemisphere+enhanced&k1=target&v1=Mars"
browser.visit(url)
```

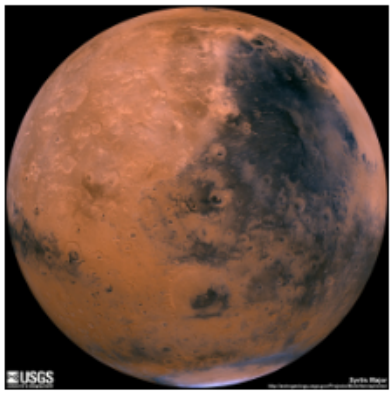
[38]:
```
# Find all of the level 3 header information for hemisphere products.  Loop␣
 ↪through images based on number of products.
hemisphere_image_urls = []

products = browser.find_by_css("a.product-item h3")

for i in range(len(products)):
    # initialize hemisphere dictionary
    hemisphere = {}
    # click on each product link to get to actual image
    browser.find_by_css("a.product-item h3")[i].click()

    # get url (href) for the "Sample" image option since full-res images are␣
 ↪very large
    sample_product = browser.find_link_by_text("Sample").first
    hemisphere["img_url"] = sample_product["href"]

    # Get Hemisphere Title
    hemisphere["title"] = browser.find_by_css("h2.title").text

    # Append Hemisphere Object to List
```

```
    hemisphere_image_urls.append(hemisphere)

    # Go back to product screen to move to next product on the page
    browser.back()

# print out the hemisphere urls
print(hemisphere_image_urls)
```

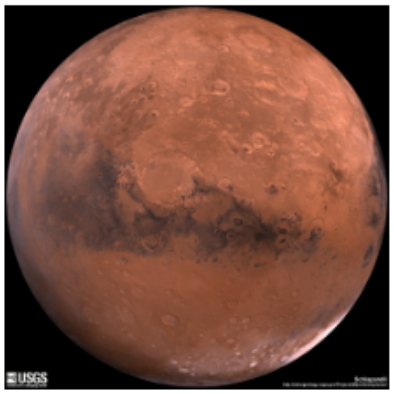
```
[{'img_url': 'http://astropedia.astrogeology.usgs.gov/download/Mars/Viking/cerberus_enhanced.tif/full.jpg', 'title': 'Cerberus Hemisphere Enhanced'},
{'img_url': 'http://astropedia.astrogeology.usgs.gov/download/Mars/Viking/schiaparelli_enhanced.tif/full.jpg', 'title': 'Schiaparelli Hemisphere Enhanced'},
{'img_url': 'http://astropedia.astrogeology.usgs.gov/download/Mars/Viking/syrtis_major_enhanced.tif/full.jpg', 'title': 'Syrtis Major Hemisphere Enhanced'},
{'img_url': 'http://astropedia.astrogeology.usgs.gov/download/Mars/Viking/valles_marineris_enhanced.tif/full.jpg', 'title': 'Valles Marineris Hemisphere Enhanced'}]
```

Cerberus

Syrtis



Schiaparelli



Valles Marineris