

Web Scraping Project

Instructions

This project will be done in python so make sure you have it installed. If it is not installed, you can go to <https://www.python.org/> and install it.

You will also need two python packages: requests and BeautifulSoup4. To install these, you can type a simple command in command line for each module: "pip install _____" (ex: pip install BeautifulSoup4)

If you are on Linux or Mac, you may have to get python3-pip or create a virtual environment first.

Now that everything is set up, let's get started!

For this project, you will be scraping the information card data off Wikipedia's Penn State page (shown on the right).

https://en.wikipedia.org/wiki/Pennsylvania_State_University

You will scrape the data from 'Motto' to 'Website' on the information card on the right ('Website' not shown here).

Your code should store this data in some sort of file (json, csv, txt, etc) in an organized format. For example, you can use a table header (th) to table data (td) format (Motto, Making Life Better). 'Motto' is a th and 'Making Life Better' is a td.

Refer to the example files on page 3 if you want a visual example of an output file. Your csv or json output file can differ from the ones shown.

Clean up the data as much as you can (try to get rid of the [1], [2], etc.)

Submission

If you use Github, upload your files to a public GitHub repo and turn in that link (recommended).

Otherwise, submit a python file.

Getting started

The Pennsylvania State University



Motto	Making Life Better
Type	Public state-related land-grant flagship research university
Established	1855; 166 years ago
Academic affiliations	AAU · APLU · BTAA · CDIO · CRL · ORAU · UCAR · URA · Sea-grant · Space-grant · Sun-grant
Endowment	\$4.55 billion (2019) ^[1] • \$2.2 billion (University Park) ^[2]
President	Eric J. Barron ^[3]
Provost	Nicholas P. Jones ^[4]
Academic staff	8,864 ^[5]
Students	96,408 ^[6] • 46,723 (University Park)
Undergraduates	81,080 ^[6] • 40,639 (University Park)
Postgraduates	15,328 ^[6] • 6,084 (University Park)
Location	University Park,

Right click on the page and select 'Inspect' to view the HTML code.

Select the entire infobox card and click it or locate it on your own (this should be a table tag).

Figure out which parent/sibling/child relationships the table rows have.

Determine what data is not needed and what differentiates these from the data you need (for example, the table row with the image has no table header).

Determine what tags are unnecessary (if you use the `get_text()` function, all text will be retrieved).

Resources:

Refer to <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> or click the links below for specific details and examples

[find\(\)](#)

- Finds the first descendant (children and children of those children and so on) and returns that tag

[find_all\(\)](#)

- Finds all the descendants and returns them as a list

[decompose\(\)](#)

- Removes a tag and everything inside that tag

[get_text\(\)](#)

- Gets all the text in between a tag (and text between children tags too)

[prettify\(\)](#)

- Will create a nicely formatted and easy to read version of the HTML code

Starter Code (if needed):

```
import requests
from bs4 import BeautifulSoup
# import csv
# import json

url = "https://en.wikipedia.org/wiki/Pennsylvania_State_University"

response = requests.get(url)
# print(response.status_code)

soup = BeautifulSoup(response.content, 'html.parser')
```

Issues

It is recommended to try to solve issues on your own first. Software engineers will have to read the documentation for tools they use and debug their code using online resources. Stack Overflow and the [Beautiful Soup documentation](#) are great resources that you can use for this project. However, feel free to send us a Canvas mail if you are having any issues.



Example csv file (structured data):

	data.csv
1	University,The Pennsylvania State University
2	Motto,Making Life Better
3	Type,Public state-related land-grant flagship research university
4	Established,1855
5	Academic affiliations,AAUAPLUBTAACDIOCRLOAUUCARURASea-grantSpace-grantSun-grant
6	Endowment,\$4.55 billion (2019)• \$2.2 billion (University Park)
7	President,Eric J. Barron
8	Provost,Nicholas P. Jones
9	Academic staff,"8,864"
10	Students,"96,408• 46,723 (University Park)"
11	Undergraduates,"81,080• 40,639 (University Park)"
12	Postgraduates,"15,328• 6,084 (University Park)"
13	Location,"University Park, Pennsylvania, United States"
14	Campus,"Rural/College town, 7,343 acres (29.72km)"
15	Newspaper,The Daily Collegian
16	Colors,Nittany Navy and White Out
17	Nickname,Nittany Lions
18	Sporting affiliations,NCAA Division I - Big Ten; MAISA
19	Mascot,Nittany Lion
20	Website,www.psu.edu
21	

Csv file imported to Excel:

	A	B
1	Column1	Column2
2	University	The Pennsylvania State University
3	Motto	Making Life Better
4	Type	Public state-related land-grant flagship research university
5	Established	1855
6	Academic affiliations	AAUAPLUBTAACDIOCRLOAUUCARURASea-grantSpace-grantSun-grant
7	Endowment	\$4.55 billion (2019)• \$2.2 billion (University Park)
8	President	Eric J. Barron
9	Provost	Nicholas P. Jones
10	Academic staff	8,864
11	Students	96,408• 46,723 (University Park)
12	Undergraduates	81,080• 40,639 (University Park)
13	Postgraduates	15,328• 6,084 (University Park)
14	Location	University Park, Pennsylvania, United States
15	Campus	Rural/College town, 7,343 acres (29.72km)
16	Newspaper	The Daily Collegian
17	Colors	Nittany Navy and White Out
18	Nickname	Nittany Lions
19	Sporting affiliations	NCAA Division I – Big Ten; MAISA
20	Mascot	Nittany Lion
21	Website	www.psu.edu

Example json file (semi structured data):

```
{
  "The Pennsylvania State University": {
    "Academic affiliations": "AAUAPLUBTAACDIOCRLOAUUCARURASea-grantSpace-grantSun-grant",
    "Academic staff": "8,864",
    "Campus": "Rural/College town, 7,343 acres (29.72km)",
    "Colors": "Nittany Navy and White Out",
    "Endowment": "$4.55 billion (2019); $2.2 billion (University Park)",
    "Established": "1855",
    "Location": "University Park, Pennsylvania, United States",
    "Mascot": "Nittany Lion",
    "Motto": "Making Life Better",
    "Newspaper": "The Daily Collegian",
    "Nickname": "Nittany Lions",
    "Postgraduates": "15,328; 6,084 (University Park)",
    "President": "Eric J. Barron",
    "Provost": "Nicholas P. Jones",
    "Sporting affiliations": "NCAA Division I - Big Ten; MAISA",
    "Students": "96,408; 46,723 (University Park)",
    "Type": "Public state-related land-grant flagship research university",
    "Undergraduates": "81,080; 40,639 (University Park)",
    "Website": "www.psu.edu"
  }
}
```

Challenge (Optional)

Start with https://en.wikipedia.org/wiki/Big_Ten_Conference and get the infobox card data from every single university in the Member Schools table and store it in a file.

An example is shown below in json file format, but any type of file is fine.

```
129 "Postgraduates": "14,385 (Columbus)14,400 (all campuses)",
130 "President": "Kristina M. Johnson",
131 "Sporting affiliations": "NCAA Division IBig Ten Conference",
132 "Students": "61,369 (Columbus)67,957 (all campuses)",
133 "Type": "Public flagship land-grant research university",
134 "Undergraduates": "46,984 (Columbus)53,557 (all campuses)",
135 "University": "The Ohio State University",
136 "Website": "osu.edu"
137 },
138 "The Pennsylvania State University": {
139 "Academic affiliations": "AAUAPLUBTAACDIOCRLOAUUCARURASea-grantSpace-grantSun-grant",
140 "Academic staff": "8,864",
141 "Campus": "Rural/College town, 7,343 acres (29.72km)",
142 "Colors": "Nittany Navy and White Out",
143 "Endowment": "$4.55 billion (2019); $2.2 billion (University Park)",
144 "Established": "1855",
145 "Location": "University Park, Pennsylvania, United States",
146 "Mascot": "Nittany Lion",
147 "Motto": "Making Life Better",
148 "Newspaper": "The Daily Collegian",
149 "Nickname": "Nittany Lions",
150 "Postgraduates": "15,328; 6,084 (University Park)",
151 "President": "Eric J. Barron",
152 "Provost": "Nicholas P. Jones",
153 "Sporting affiliations": "NCAA Division I - Big Ten; MAISA",
154 "Students": "96,408; 46,723 (University Park)",
155 "Type": "Public state-related land-grant flagship research university",
156 "Undergraduates": "81,080; 40,639 (University Park)",
157 "University": "The Pennsylvania State University",
158 "Website": "www.psu.edu"
159 },
160 "The University of Iowa": {
161 "Academic affiliations": "AAUAPLUBTAUAURASpace-grant",
162 "Administrative staff": "2,296",
163 "Athletics": "NCAA Division I - Big Ten",
164 "Campus": "Urban1,880 acres",
165 "Colors": "Black and Gold",
166 "Endowment": "$1.58 billion (2019)",
167 "Established": "1847",
168 "Location": "Iowa City, Iowa, United States",
169 "Mascot": "Herky the Hawk",
170 "Nickname": "Hawkeyes",
171 "Postgraduates": "10,958"
```