08/18/2020

Ankita Ghosh

**Reference:** <https://first-web-scraper.readthedocs.io/en/latest/>

**Data:** <https://report.boonecountymo.org/mrcjava/servlet/RMS01_MP.I00030s>

Data will be extracted from the roster of inmates at the Boone County Jail in Missouri from the above website and save it as comma-delimited text ready for analysis.

**Scope:**

* Analyze the crime rate according by gender, state, city race and state
* Try to classify according to the degree and crimes

1. Install python 3.8 , set environment path

<https://projects.raspberrypi.org/en/projects/using-pip-on-windows/4>

1. Install or check for pip (package management system used to manage, and handle software packages written in Python)

<https://projects.raspberrypi.org/en/projects/using-pip-on-windows/5>

1. Install beautifulsoup4

**Command** ~ *pip install beautifulsoup4*

1. Install Request library

**Command** ~ *python -m pip install requests*

1. Create a new directory > Scrape.py file > try to import request library and download HTML from the desired site

**Command**

*import requests*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response=requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*print html*

The **User**-**Agent** request **header** is a characteristic string that lets servers and network peers identify the application, operating system, vendor, and/or version of the requesting **user agent**

Why Mozilla?

<https://stackoverflow.com/questions/1114254/why-do-all-browsers-user-agents-start-with-mozilla>

1. To import BeautifulSoup

**Command~**

*import requests*

*from bs4 import BeautifulSoup*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response=requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*print(html)*

*soup = BeautifulSoup(html)*

*print(soup.prettify())*

Run ~ python Scarpe.py

1. We do not want the whole page. We want that particular table, so we will need to take any table id, with attributes to identify from where the table will be taken.

**Command:**

*import requests*

*from BeautifulSoup import BeautifulSoup*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response = requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*soup = BeautifulSoup(html)*

*table = soup.find('tbody', attrs={'class': 'stripe'})*

*print table.prettify()*

1. Now rows of the table needs to be converted to lists, because we have to loop through and grab all the data.

**Command:**

*import requests*

*from BeautifulSoup import BeautifulSoup*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response = requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*soup = BeautifulSoup(html)*

*table = soup.find('tbody', attrs={'class': 'stripe'})*

*for row in table.findAll('tr'):*

*for cell in row.findAll('td'):*

*print cell.text*

1. We got the data from the web page, now we need to structure the data in .csv file format. So, we will add each cell in a row to a new python list

*import requests*

*from bs4 import BeautifulSoup*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response=requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*print(html)*

*soup = BeautifulSoup(html)*

*table = soup.find('tbody', attrs={'class': 'stripe'})*

*#print(soup.prettify())*

*#print(table.prettify())*

*#to convert the table rows to list so that data can be traversed*

*for row in table.findAll('tr'):*

*list\_of\_cells = []*

*for cell in row.findAll('td'):*

*text = cell.text.replace('&nbsp;', '') #to replace &nbsp (non breaking space)*

*list\_of\_cells.append(text)*

*print(list\_of\_cells)*

1. Those lists can now be lumped together into one big list of lists, which, when you think about it, isn’t all that different from how a spreadsheet is structured.

*import requests*

*from BeautifulSoup import BeautifulSoup*

*url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s'*

*response = requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})*

*html = response.content*

*soup = BeautifulSoup(html)*

*table = soup.find('tbody', attrs={'class': 'stripe'})*

*list\_of\_rows = []*

*for row in table.findAll('tr'):*

*list\_of\_cells = []*

*for cell in row.findAll('td'):*

*text = cell.text.replace('&nbsp;', '')*

*list\_of\_cells.append(text)*

*list\_of\_rows.append(list\_of\_cells)*

*print list\_of\_rows*

1. To write the list out of CSV type, we need to import csv and at the end add the below code

**Command ~**

import csv

import requests

from bs4 import BeautifulSoup

url = 'https://report.boonecountymo.org/mrcjava/servlet/SH01\_MP.I00290s?max\_rows=500'

response=requests.get(url, headers={'User-Agent': 'Mozilla/5.0'})

html = response.content

#print(html)

soup = BeautifulSoup(html)

table = soup.find('tbody', attrs={'class': 'stripe'})

#print(soup.prettify())

#print(table.prettify())

list\_of\_rows = []

#to convert the table rows to list so that data can be traversed

for row in table.findAll('tr'):

list\_of\_cells = []

for cell in row.findAll('td'):

text = cell.text.replace('&nbsp;', '') #to replace &nbsp (non breaking space)

list\_of\_cells.append(text)

list\_of\_rows.append(list\_of\_cells)

#print(list\_of\_rows)

outfile = open("./try.csv", "w")

writer = csv.writer(outfile)

writer.writerow(["Last", "First", "Middle", "Gender", "Race", "Age", "City", "State"])

writer.writerows(list\_of\_rows) #handy tool it has called writerows to dump out our list of lists