


Project 4: Web scraping in Python + REGEX

In this project we use BS- BeautifulSoup and REGEX to find some whose last name starts with the letter 'A'

WE'LL COVER THE FOLLOWING ^

- Solution

Web scraping or web data extraction is data scraping used for extracting data from websites. In this project, we will extract tabular data from the Boone County Sheriff's Dept [website](#) of criminal records (see the image below) and then find all the people whose name start with the word "A".



Boone County Sheriff's Department
Boone County Sheriff's Department
2121 County Dr.
Columbia, MO 65202

Office (573) 875-1111
Fax (573) 874-8953

Home Menu Help

Current Detainees of Boone County Jail

Search By Name
Last Name:
First Name:

Run Date: 08/20/17

232 Total Items | Page Size << < Page 1 of 5 > >>

Last Name ▲	First Name ▲	Middle Name	Sex	Race	Age	City	State	
ACTON	ANTHONY	SEAN	M	B	25	COLUMBIA	MO	<input type="button" value="Details"/>
ADAM	OMER	SIRAJ	M	B	29	COLUMBIA	MO	<input type="button" value="Details"/>
ALEXANDER	CHARLES	CODY	M	W	23	COLUMBIA	MO	<input type="button" value="Details"/>
ALLEN	WILLIAM	LAMAR	M	B	55	COLUMBIA	MO	<input type="button" value="Details"/>
AVALOS-AVALOS	JOSE		M	H	19	STANN	MO	<input type="button" value="Details"/>
BAIMA	TERESA	MAE	F	W	46	COLUMBIA	MO	<input type="button" value="Details"/>

Solution

The problem solution uses [BeautifulSoup](#). A detailed explanation of the code is out-of-scope for this course (hint: read the BS docs). In this code, we first extract HTML data and format/convert into BS's table using BS's

`BeautifulSoup()` function, then find and extract the table from the HTML

code.

```
import re
from pprint import pprint
import csv
import requests
from bs4 import BeautifulSoup

url = 'http://www.showmeboone.com/sheriff/JailResidents/JailResidents.asp'
response = requests.get(url)
html = response.content

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

list_of_rows = []

for row in table.findAll('tr')[0:]:
    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)

for line in list_of_rows:
    row = '\t'.join(str(i) for i in line) # python 2
    s=row[0:5] # Select only the Last names (1st column)
    m = re.search(r"^[A]",s,re.I)
    if m:
        print row
```

Expected output (Surnames started with the letter “A”):

```
ACTON    ANTHONY SEAN    M    B    25    COLUMBIA    MO    Details
ADAM     OMER    SIRAJ    M    B    29    COLUMBIA    MO    Details
ALEXANDER    CHARLES CODY    M    W    23    COLUMBIA    MO    Details
ALLEN    WILLIAM LAMAR    M    B    55    COLUMBIA    MO    Details
AVALOS-AVALOS    JOSE    M    H    19    ST.ANN    MO    Details
```

Easy!

This examples has been **adopted and extended** from the Python-BeautifulSoup’s [first web scraper](#), originally developed by Chase Davis, Jackie Kazil, Sisi Wei and Matt Wynn for bootcamps held by Investigative Reporters and Editors at the University of Missouri in Columbia, Missouri.