

DSO 545: Statistical Computing and Data Visualization






Web Scrapping in Python

1. Use HTML to create the following webpage:

Totally Normal Gifts

Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is hand-curated by well-paid, free-range Tibetan monks.

We haven't figured out how to make online shopping carts yet, but you can send us a check.

Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with super-colorful bell peppers!</i>	\$15.00	
Russian Nesting Dolls	Hand-painted by trained monkeys, these exquisite dolls are priceless! And by "priceless," we mean "extremely expensive"! <i>8 entire dolls per set! Octuple the presents!</i>	\$10,000.52	
Fish Painting	If something seems fishy about this painting, it's because it's a fish! <i>Also hand-painted by trained monkeys!</i>	\$10,005.00	
Dead Parrot	This is an ex-parrot! <i>Or maybe he's only resting?</i>	\$0.50	
Mystery Box	If you love surprises, this mystery box is for you! Do not place on light-colored surfaces. May cause oil staining. <i>Keep your friends guessing!</i>	\$1.50	

2. Go to <http://www.pythonscraping.com/pages/warandpeace.html> and extract all green text.

```
import re
from urllib.request import urlopen
from bs4 import BeautifulSoup

html = urlopen("http://www.pythonscraping.com/pages/warandpeace.html")
bsObj = BeautifulSoup(html)

### find_all(tag_name, tagAttributes)

namelist = bsObj.find_all("span", {"class": "green"})
namelist

# strips all tags and returns a list of strings
```

```
for name in namelist:
    print(name.get_text())
```

3. Import the list of presidents of United States from this wikipedia page:

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

```
import pandas as pd
import requests
from bs4 import BeautifulSoup

url = "https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States"
page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')

tbl = soup.find("table", {"class": "wikitable"})

for link in tbl.find_all('b'):
    name = link.find("a")
    # print(name)
    print(name.get_text('title'))

# this could be shortened using list comprehension

[link.find('a').get_text('title') for link in tbl.find_all('b')]
```

4. Import the full table into csv table from :

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

```
URL = "<https://en.wikipedia.org/wiki/List_of_largest_manufacturing_companies_by_revenue"
response = requests.get(URL)
soup = BeautifulSoup(response.content, "html.parser") # response.text

# the table is in table > tbody

table = soup.find("table", {"class": "wikitable sortable plainrowheads"}).tbody

# find all rows

rows = table.find_all('tr')
rows[0].text for var in rows[0].find_all('th')
cols = [var.text.replace('\n', '') for var in rows[0].find_all('th')]

df = pd.DataFrame()

# \[var.text.replace("\n", "") for var in rows[1].find_all('td')]

for i in range(1, len(rows)):
```

```
# tds = rows[i].find_all('td')
# print(len(tds))
values = [td.text.replace("\n", "").replace('\xa0', ' ') for td in rows[i].find_
df = df.append(pd.Series(values), ignore_index=True)

df.columns = cols
df.head()

# for the electronics industry, find the total revenue:

electronics = df.loc[df.Industry == "Electronics", :]
electronics.head()

revenue = pd.to_numeric(electronics.iloc[:, 3].str.replace(',', ''))
revenue.sum()

df.to_csv("/Users/abbassalsharif/Documents/GitHub/DS0545/02_Sessions/Week_09/webscr
```