1. Web Scraping

March 29, 2022

[1]:
```
#!pip install beautifulsoup4
```

[2]:
```
# Retrieve all text from a wikipedia page
import bs4 as bs
import urllib.request
import re
import nltk

scrapped_data = urllib.request.urlopen('https://en.wikipedia.org/wiki/
 ↪Artificial_intelligence')
article = scrapped_data.read()

parsed_article = bs.BeautifulSoup(article,'lxml')

paragraphs = parsed_article.find_all('p')

article_text = ""

for p in paragraphs:
    article_text += p.text

#article_text
```

[3]:
```
# Retrieve price of an item on a flipkart page
import bs4
import urllib.request as url
path =  "https://www.flipkart.com/
 ↪mi-4a-pro-108-cm-43-inch-full-hd-led-smart-android-tv-google-data-saver/p/
 ↪itmfbzck4mhggxxg"
resp = url.urlopen(path)
print(resp)
# lxml = library xml, parser
page = bs4.BeautifulSoup(resp,'lxml')
# dont print page variable as it will display html content of whole page
# identify the div tag whose value is needed
title1=page.find('div',class_='_30jeq3 _16Jk6d')
title1.text
```

```
#title2=page.find('span',class_='_35KyD6')
#title2.text
# Notice _ after class keyword as it is an html class not a python class
#title3=page.find('div',class_='_1vC4OE _3qQ9m1')
#title3.text
```

```
<http.client.HTTPResponse object at 0x0000026D761227F0>
```

[3]:
```
' 26,999'
```

[4]:
```
# Retrieve prices of items on a flipkart search page
path = "https://www.flipkart.com/search?
 ↪q=tv&sid=ckf%2Cczl&as=on&as-show=on&otracker= \
AS_QueryStore_OrganicAutoSuggest_1_2_na_na_na&otracker1= \
AS_QueryStore_OrganicAutoSuggest_1_2_na_na_na&as-pos=1&as-type= \
RECENT&suggestionId=tv%7CTVs&requestId=474cc2f9-7aad-4de4-b60a-4674b445460c&as-searchtext=tv"
resp = url.urlopen(path)
page = bs4.BeautifulSoup(resp,'lxml')
titles=page.find_all('div',class_='_4rR01T')           # identify the div tag␣
 ↪containing the price
prices=page.find_all('div',class_='_30jeq3 _1_WHN1')   # identify the div tag␣
 ↪containing the title
len(titles)
for i in range(len(titles)):
    print(titles[i].text)
    print(prices[i].text)
```

```
SAMSUNG The Frame 2021 Series 163 cm (65 inch) QLED Ultra HD (4K) Smart TV
 1,19,990
Adsun 98.9 cm (39 inch) HD Ready LED Smart TV
 13,999
LG 80 cm (32 inch) HD Ready LED Smart TV
 17,499
SAMSUNG 80 cm (32 inch) HD Ready LED Smart TV
 16,999
Mi 5X 108 cm (43 inch) Ultra HD (4K) LED Smart Android TV with Dolby Atmos and
Dolby Vision
 31,999
OnePlus Y1 108 cm (43 inch) Full HD LED Smart Android TV
 25,499
OnePlus Y1 100 cm (40 inch) Full HD LED Smart Android TV
 22,999
SAMSUNG The Frame 2021 Series 138 cm (55 inch) QLED Ultra HD (4K) Smart TV
 88,900
OnePlus Y1 80 cm (32 inch) HD Ready LED Smart Android TV
 15,999
MarQ By Flipkart 80 cm (32 inch) HD Ready LED TV
```

```
 8,999
Adsun 80 cm (32 inch) HD Ready LED Smart TV
 9,499
SAMSUNG The Frame 2021 Series 108 cm (43 inch) QLED Ultra HD (4K) Smart TV
 58,990
Vu Premium TV 80 cm (32 inch) HD Ready LED Smart TV with Bezel-Less Frame
 12,999
MarQ By Flipkart 60 cm (24 inch) HD Ready LED TV
 6,999
OnePlus Y1S 80 cm (32 inch) HD Ready LED Smart Android TV
 16,499
SAMSUNG The Frame 2021 Series 125 cm (50 inch) QLED Ultra HD (4K) Smart TV
 73,990
Vu Premium 108 cm (43 inch) Full HD LED Smart Android TV
 22,999
realme 80 cm (32 inch) HD Ready LED Smart Android TV
 15,999
Infinix X3 80 cm (32 inch) HD Ready LED Smart Android TV
 11,999
T-Series 98 cm (40 inch) HD Ready LED Smart Android TV
 19,999
Mi 4A Horizon Edition 108 cm (43 inch) Full HD LED Smart Android TV
 25,999
Mi 4A PRO 80 cm (32 inch) HD Ready LED Smart Android TV
 16,499
Mi 5X 125.7 cm (50 inch) Ultra HD (4K) LED Smart Android TV with Dolby Atmos and
Dolby Vision
 40,999
T-Series 109 cm (43 inch) Full HD LED Smart TV
 26,999
```

[5]:
```
# convert scraped items into a pandas dataset
dataset={"title":[],"price":[]}
for j in range(len(titles)):
    dataset["title"].append(titles[j].text)
    dataset["price"].append(prices[j].text)

import pandas as pd
df=pd.DataFrame(dataset)
df.shape
df.head(10)
```

[5]:

	title	price
0	SAMSUNG The Frame 2021 Series 163 cm (65 inch)…	1,19,990
1	Adsun 98.9 cm (39 inch) HD Ready LED Smart TV	13,999
2	LG 80 cm (32 inch) HD Ready LED Smart TV	17,499
3	SAMSUNG 80 cm (32 inch) HD Ready LED Smart TV	16,999

```
4  Mi 5X 108 cm (43 inch) Ultra HD (4K) LED Smart…    31,999
5  OnePlus Y1 108 cm (43 inch) Full HD LED Smart …    25,499
6  OnePlus Y1 100 cm (40 inch) Full HD LED Smart …    22,999
7  SAMSUNG The Frame 2021 Series 138 cm (55 inch)…    88,900
8  OnePlus Y1 80 cm (32 inch) HD Ready LED Smart …    15,999
9   MarQ By Flipkart 80 cm (32 inch) HD Ready LED TV     8,999
```

[ ]: