A Web Crawler

Rajat Mittal

M.Tech. Computer Technology

Student ID - 2020PCT0066

Email: 2020PCT0066@iitjammu.ac.in

September 21, 2021

भारतीय प्रौद्योगिकी संस्थान जम्मू

INDIAN INSTITUTE OF TECHNOLOGY JAMMU

विद्याधनं सर्वधन प्रधानम्

A Text - Crawler

Crawled Website:- https://books.toscrape.com/

In the Text-Crawler, I crawled a bookstore website. This website has around 1000 books in 50 pages. I crawled all of them using Python libraries.

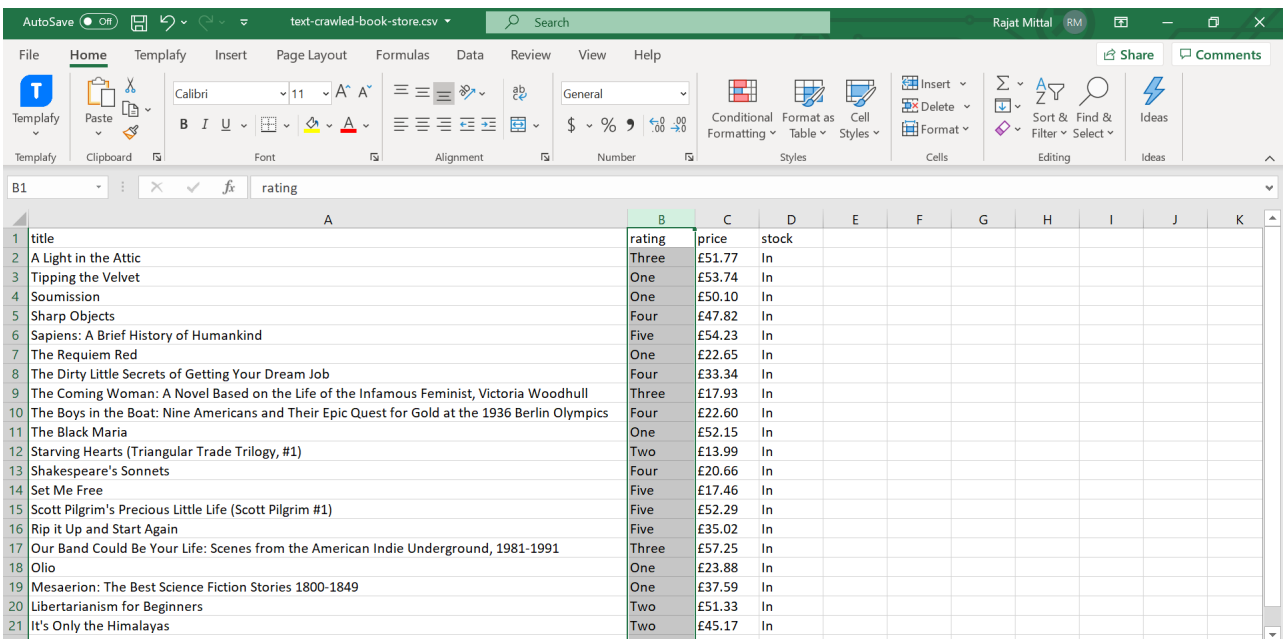
Libraries used: **BeautifulSoup, requests, csv**

Text Crawler Python Code for text-crawling

```
# import necessary libraries
import requests
from bs4 import BeautifulSoup
import csv



# adding given urls
url = "https://books.toscrape.com/"

# get the html content
r = requests.get (url)

html_content = r.content

# making the soup - parsing the HTML
soup = BeautifulSoup (html_content, 'html.parser')

data = []

catalogue = soup.find_all ("ol", {"class":"row"})

for item in catalogue [0].find_all ("li", {"class":"col-xs-6 col-sm-4 col-md-3 col-
  mdata = {}
  mdata ['title'] = item.h3.a.get ("title")
  mdata ['rating'] = item.p.get ("class") [1]
  mdata ['price'] = item.find ("p", {"class":"price_color"}).text
  mdata ['stock'] = item.find ("p", {"class":"instock availability"}).text [15:17]
  data.append (mdata)
```

```
#Save into CSV file

file_csv = 'book-store.csv'

with open (file_csv, 'w', newline='') as f:
    w = csv.DictWriter (f, ['title', 'rating', 'price', 'stock'])
    w.writeheader ()
    for mdata in data:
        w.writerow (mdata)
```

Sample Output Photo



	A	B	C	D
1	title	rating	price	stock
2	A Light in the Attic	Three	£51.77	In
3	Tipping the Velvet	One	£53.74	In
4	Soumission	One	£50.10	In
5	Sharp Objects	Four	£47.82	In
6	Sapiens: A Brief History of Humankind	Five	£54.23	In
7	The Requiem Red	One	£22.65	In
8	The Dirty Little Secrets of Getting Your Dream Job	Four	£33.34	In
9	The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria Woodhull	Three	£17.93	In
10	The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics	Four	£22.60	In
11	The Black Maria	One	£52.15	In
12	Starving Hearts (Triangular Trade Trilogy, #1)	Two	£13.99	In
13	Shakespeare's Sonnets	Four	£20.66	In
14	Set Me Free	Five	£17.46	In
15	Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	Five	£52.29	In
16	Rip it Up and Start Again	Five	£35.02	In
17	Our Band Could Be Your Life: Scenes from the American Indie Underground, 1981-1991	Three	£57.25	In
18	Olio	One	£23.88	In
19	Mesaerion: The Best Science Fiction Stories 1800-1849	One	£37.59	In
20	Libertarianism for Beginners	Two	£51.33	In
21	It's Only the Himalayas	Two	£45.17	In

An Image - Crawler

Crawled Website:- https://books.toscrape.com/

In the Image-Crawler, I crawled a bookstore website. This website has around 1000 books in 50 pages. Also it has image for every book. So, I crawled all images of books using Python libraries.

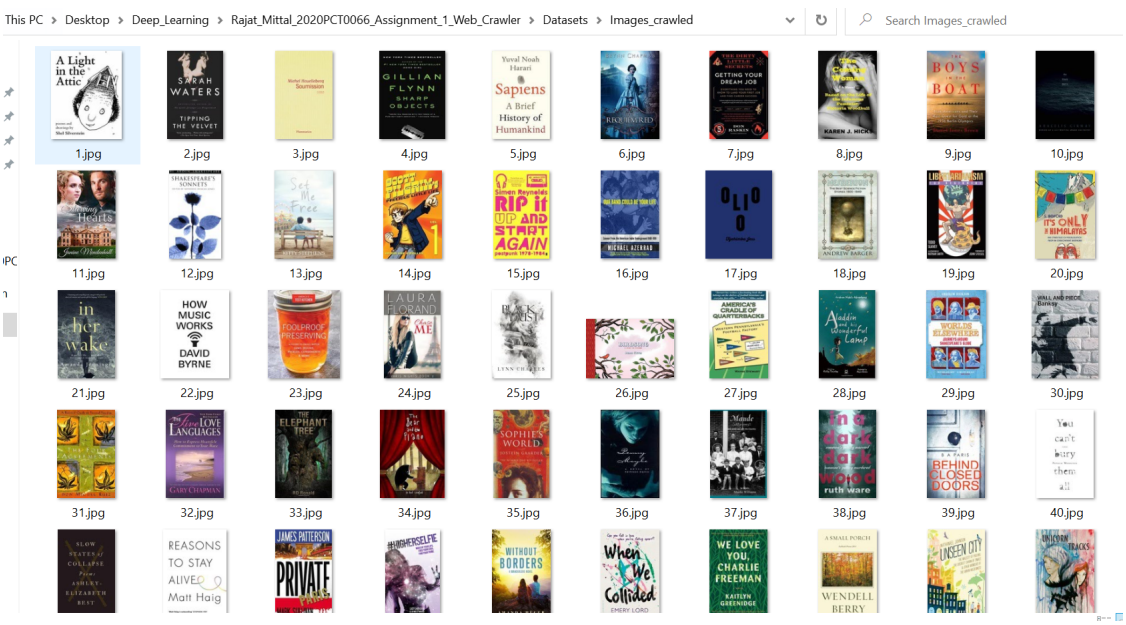
Libraries used: **BeautifulSoup, requests, os**

Image Crawler Python Code for images-crawling

```
import requests
from bs4 import BeautifulSoup
import os

data = []
images = []
image_url = "https://books.toscrape.com"
# add your url
for i in range(1,51):
    url = "https://books.toscrape.com/catalogue/page-" + str(i) + ".html"
    #print(url)
    r = requests.get (url)
    html_content = r.content
    soup = BeautifulSoup (html_content, 'html.parser')
    catalogue = soup.find_all("img")


    for image in catalogue:
      images.append(image['src'])

for i in range(len(images)):
  temp = images[i][2:]
  images[i] = image_url+temp
  #print(image)


os.mkdir('Rajat_photos')
i = 1
```

```
for index, img_link in enumerate(images):
    if i <= len(images):
        img_data = requests.get(img_link).content
        with open("my_pics/"+str(index+1)+'.jpg', 'wb+') as f:
            f.write(img_data)
        i += 1
    else:
        f.close()
        break
```

Sample Output Photo



A Video - Crawler

Crawled Website:- https://sample-videos.com/

In the Video-Crawler, I crawled a sample video website.In a one video page of that webiste, it has around 80+ videos. I crawled all of them using Python libraries. It has .mp4, .3gp and .flv videos.

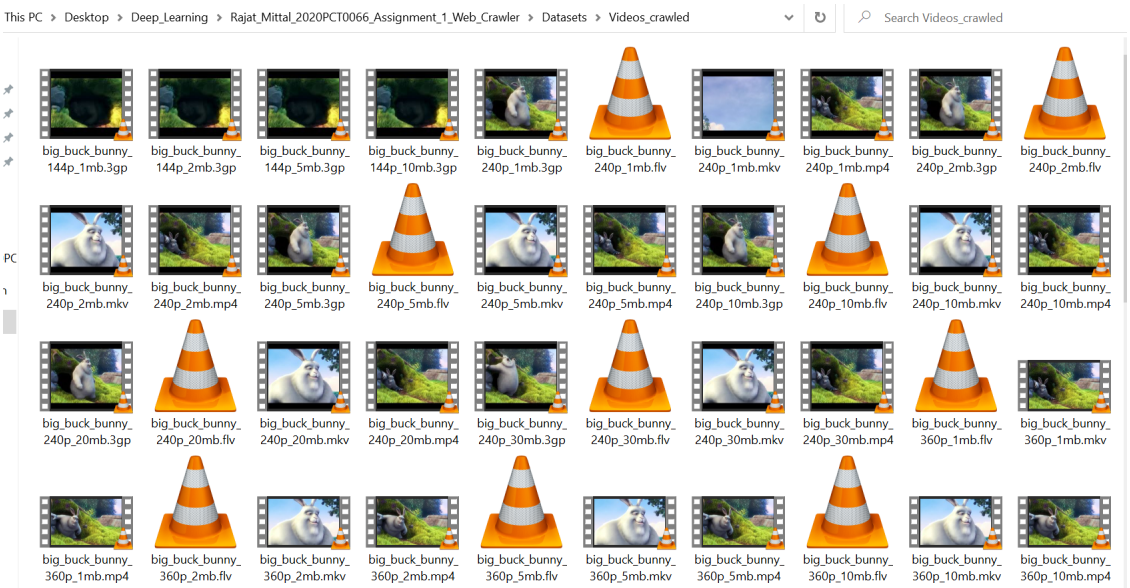
Libraries used: **BeautifulSoup, requests**

Videos Crawler Python Code for videos-crawling

```
import requests
from bs4 import BeautifulSoup

links = []
videos = []

videos_url = "https://sample-videos.com/"

url = "https://sample-videos.com/index.php#sample-mp4-video"

r = requests.get (url)
html_content = r.content
soup = BeautifulSoup (html_content, 'html.parser')

for a in soup.find_all('a', href=True):
    videos.append(a['href'])

videos = videos[26:len(videos)-1]


for i in range(len(videos)):
  links.append(videos_url + videos[i])

def download_video_series(video_links):
        for link in video_links:
                file_name = link.split('/')[-1]
                r = requests.get(link, stream = True)
                with open(file_name, 'wb') as f:
                        for chunk in r.iter_content(chunk_size = 1024*1024):
```

```
                    if chunk:
                        f.write(chunk)
        return
#temp = links[0:2]

download_video_series(links)
```

Sample Output Photo



Thank You!