

Web Scraping For Book Recommendation System

Rasha Shaikh

Department of Computer Engineering, Savitribai Phule University, Pune, Maharashtra, India

ABSTRACT

Article Info Volume9, Issue 2 Page Number: 552-556

Publication Issue March-April-2022

Article History Accepted :03April2022 Published :20April2022 The purpose of a book recommendation system is to predict buyer's interest and recommend books to them accordingly. Personal recommendation systems have been emerged to conduct effective search which mine related books based on user rating and interest. This paper proposed an effective system for recommending books for online users by providing the data which not only counts the ratings but also the users vote for the best books of 2022 along with their genre by using web scraping. Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Rather than using big data, smart data would work much better. The proposed system used Beautiful Soup designed and selenium web drivers for scraping HTML documents. Convenient Pythonic functions for navigating, searching, and modifying a parse tree; a toolkit for decomposing an HTML file and extracting desired information via html parser. The required data was successfully scraped or extracted and saved in csv file. Further a book recommendation model needs to be build using this dataset.

Keywords:

Web scraping, Beautiful Soup, Selenium, Web Drivers, HTML Parser, Data Extraction.

I. INTRODUCTION

Web scraping is a method used to get great amounts of data from websites and then data can be used for any kind of data manipulation and operation on it.

For this technique, we use web browsers. You usually do not have the built-in option to get that data you want. That is why we use Web Scraping to automate the process of getting that data and not having to do it manually. Web Scraping is the technique of automating this process so that instead of manually copying the data from websites.

This is accomplished either manually by a user or automatically by a bot or web crawler. Due to the fact that an enormous amount of heterogeneous data is constantly generated on the WWW, web scraping is widely acknowledged as an efficient and powerful technique for collecting bigdata. To adapt to a variety of scenarios, current web scraping techniques have become customized from smaller ad hoc, humanaided procedures to the utilization of fully automated

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited



systems that are able to convert entire websites into well-organized data set. The purpose of this study is to scrape the best books of 2022 data from the Goodreads website for book recommendation system and convert it into a structured data which can be further used for analysis and building recommendation system.

II. METHODS AND MATERIALS

There are two essential modules of a web scraping program – a module for composing an HTTP request, such as Urllib2 or selenium and another one for parsing and extracting information from raw HTML code, such as Beautiful Soup or Pyquery. Here, the Urllib2 module defines a set of functions to dealing with HTTP requests, such as authentication, redirections, cookies, and so on, while Selenium is a web browser wrapper that builds up a web browser, such as Google Chrome or Internet Explorer, and enables users to automate the process of browsing a website by programming. Regarding data extraction, Beautiful Soup is designed for scraping HTML and other XML documents. It provides convenient Pythonic functions for navigating, searching, and modifying a parse tree; a toolkit for decomposing an HTML file and extracting desired information via lxml or html5lib.In the proposed study, I have used both the methods for retrieving data.

Web data was scrapped utilizing Hypertext Transfer Protocol (HTTP) and through a web browser. The process of scraping data from the Internet can be divided into two sequential steps; acquiring web resources and then extracting desired information from the acquired data. Goodreads website was used to scrape best books of 2022 data along with it's Book Title, Author name, Ratings and Genre.

Important libraries like requests, beautifulsoup and selenium were imported. The program was started by composing a HTTP request from goodreads website. This request was formatted in either a URL containing a GET query. Once the request was successfully received and processed by the goodreads website, beautiful soup was used to parse the text retrieved from the website.Book Title ,Author Name and Ratings were retrieved using find_all function.

A. CHALLENGES ENCOUNTERED

In the pursuit of finding the genre of a book, I stumbled upon a hurdle. My program was flaky while locating the genre element. It passed for some books and for some others it failed. We may classify this under category problem the of 'False Negative'.Although the genre of the book was present, the program result displayed it not to be present, thereby failing.My locator strategy being correct, the problem baffled me at the beginning.I decided to dig down to find the Root Cause.Since the strategy of 'requests' was headless, I could not visualize the real problem.At this juncture, Selenium WebDriver came to the rescue.

B. LITTLE ABOUT SELENIUM WEBDRIVER

Selenium WebDriver provides implementations that can help us visualize the proceedings of the program like I would do manually, also it is more powerful when it comes to parsing the DOM and applying waiting mechanisms.Changing the implementation for the extracting the genre part from 'requests and Beautiful soup' to 'Selenium WebDriver'.

C. BUILDING BLOCKS

I used CHROME as a browser and XPATH as our locator strategy. I also exploited the powers of Fluent Wait which would help us wait for the 'Genre' element to be located for a certain time.

D. BOOK GENRE EXTRACTION PROGRAM FLOW / MODULES

i. Storing Book URLs: Using Requests, hit the URL of the Good Reads Page. Parse the response HTML using Beautiful Soup 'html parser'. Find the number books present in the list using locators of beautiful soup's findAll method.



Iterate using a For Loop over the number of books to extract the 'anchor tag' thereby pulling the 'href' link of each book. Storing the book URLs in an array. Code Snapshot captured below in Fig. no.1:

Fig no.1

import requests
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui inport WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
<pre>from selenium.webdriver.common.desired_capabilities import DesiredCapabilities</pre>
bestbooks_url='https://www.goodreads.com/list/show/171064.Best_books_of_2022'
response = requests.get(bestbooks_url)
page_contents = response.text
<pre>soup = BeautifulSoup(page_contents, 'html.parser')</pre>
allrows = soup.findAll('tr', itemtype='http://schema.org/Book')
listhref=[]
for i in allrows:
a_tag =1.a
link=a_tag['href']
listhref.append(link)
len(listhref)

ii. Setting Prerequisites for Selenium: up Downloading the required Chrome Driver.Defining the Desired Capabilities and Options for Chrome Driver: Page Strategy as normal. Maximized State of the Chrome Window. Incognito mode. Not loading images. Most important, Headless Mode of Operation. Code Snapshot captured below in Fig. No.2:

Fig no.2

caps = DesiredCapabilities().CHROME caps["pageLoadStrategy"] = "normal"

options = webdriver.ChromeOptions()
prefs = {"profile.managed_default_content_settings.images": 2}
options.add_experimental_option("prefs", prefs)
options.add_argument('--headless')
options.add_argument('--start-maximized')
options.add_argument("--incognito")

iii. Go to Book URLs and extract Genres: Hitting the Book URLs using looping mechanism.Looking whether the Genre Element is Present.If the Genre Element is not present, then find the 'signin' pop up.If the pop up is present, then refresh the page to bypass the pop up.Post refreshing, looking for the Genre Element with the new locator.Conditioning and waiting throughout, either looking out for the presence of pop up, or the two locators of genre elements. Indicating the user if no Genre is present for the book. Capturing the Genre Text corresponding to the Book URL. Repeating the above steps until the Genre of all books are captured. Code Snapshot Depicted Below in Fig.No.3 and Output in Fig.no.4:

Fig no.3

Per	1 in Listbref:				
	driver a webdriver. Chrome/desired capabilitiesscape, antionssoptions,				
	executable_paths'C:\\Users\\Rangl\\PycharmProjects\\pythonProject\\chromet				
	driver.get("https://www.geodreads.com/"+1)				
	print("")				
	print("https://www.goodreads.com/"+1)				
	listgenre = driver.find_elements(By.XPATH, "+//a[Bclass="actionLinkLite_bookPageGenreLink"]")				
	Listpopup + driver.find_elements(By.XPATH,*/html/body/diw[3]/div/div(1]/div/div/butten*)				
	listrefreshpenre + driver.find_elements(By.XP1T#, "//+(Bid+\"mest\")/div[1]/main/div[1]/div[2]/dis				
	if(las(listgenre)==0 and las(listpopup) == 1 and las(listrafrashgenre)=0):				
	driver.refresh()				
	try:				
	HebDriverHait(driver, 100, poll_frequency+5).until(
	<pre>EL.presence_of_element_located((By.XPATH, *//>[@id=*newt*]/div[1]/main/div[1]/div[2]/dix</pre>				
)	"Khatas")				
	genretext = driver.find.element(By.XPATH, "//+[\$10+\"next\"]/61v[1]/#ain/div[1]/div[2]/div				
	print(genretext)				
	except:				
	print(("Genre is not mentioned for this book"))				
	<pre>slif(len(listgenrs)>0 and len(listrefreshgenre)++8):</pre>				
	try:				
	WebDriverWait(driver, 100, poll_frequency=5).until(EC.presence_of_element_located((By.XPATH, 'Khatam')				
	genretext = driver.find_element(By.XP&TH, "#//s[Oclasss'actionLinkLite bookPageGenreLink']")				
	print(genretext)				
	except:				
	print(("Genre is not sectioned for this book"))				
100 1					

Fig no.4



*			
	print("		
Manufacture and the set the bise (631032) and a large of his			
nttps://www.goodreads.com//book/snow/S8341222-reminders-of-him Romance			
	https://www.goodreads.com//book/show/55196813-the-maid		
	Plystery		
	https://www.goodreads.com//book/show/40132775-house-of-sky-and-breath Fantasy		
	https://www.goodreads.com//book/show/57693172-a-flicker-in-the-dark		
	Theilles		
	18F1136F		
	sur 1136.		
	https://www.eoodreads.com//book/show/58590188-book-Invers		

III. RESULTS AND DISCUSSION

A dataframe was created with attributes as Book Title, Author Name, Ratings and Genre using pandas libraries. This dataframe was then converted into csv file with hundred book details or hundred rows.Csv file snapshot is shown in Fig.no.4:

Fig no.4

File 1 80 7 Re 3 Th 4 "H 5 A 6 80 7 Th	Edt Vew Language ok Title,Author Name,Ratings,Gerre minders of Him,Colleen Hoover, * 4.56 avg rating - 254,742 ratings",Romance #Vaid,Mita Proce,* 3.89 avg rating - 116,121 ratings",Mystery oute of Sky and Breath (Crescent City, a2)",Sarah 3. Nams,* 4.55 avg rating - 102,285 ratings",Fantasy Filcker in the Dark,Stacy Hillingham,* 4.09 avg rating - 58,659 ratings",Thriller ok lovers,Lmily,Honry,* 4.44 avg rating - 71,355 ratings",Mystery e Paris Apartment,Lucy Foley,* 3.71 avg rating - 59,691 ratings",Romance g Diso Breaming,Kohiti González,* 4.09 avg rating - 4,269 ratings",Fiction a of Tranquility,Emily 51. John Mandel,* 4.25 avg rating - 24,269 ratings",Fiction					
1 80 7 8e 9 Th 5 A 6 Bo 7 Th	ok Title, Author Name, Ratings, Genre minders of Him, Colleen Hoover, * 4.56 avg rating - 254,742 ratings*, Romance e Maid, Mita Prose, * 3.89 avg rating - 116,121 ratings*, Mystery ouse of Sky and Breath (Crescent City, z2)*, Sarah J. Naus, * 4.55 avg ruting - 102,286 ratings*, Fantasy Filcker in the Ouex, Stacy Willingham. * 4.03 avg rating - 58,650 ratings*, Thriller ok Lovers, fmily Henry, * 4.48 avg rating - 71,355 ratings*, Mystery B Paris Apartment, Lowy Foley, * 3.71 avg rating - 56,901 ratings*, Romance g Dies Oneaming, Modell Guides, * 4.03 avg rating - 24,260 ratings*, Fiction a of Tranquility, fmily St. John Mandel, * 4.25 avg rating - 24,260 ratings*, Fiction					
1 80 7 Re 9 Th 5 A 6 Bo 7 Th	So lites, autors Amarings, vere sinders of Hiss, Audor, Harlings, Vere # Kaid, Mita Brose, " 3.48 avg rating - 116, 121 ratings", Mystery ouse of Sky and Breath (Crescent City, 421', Scanh, Nuas," 4.55 avg rating - 102, 285 ratings", Fantasy Elicker in the Oark, Stacy Willingham, 4.40 avg rating - 56, 565 ratings", Theiller ok Iovers, Enily Henry, - 4.48 avg rating - 71, 355 ratings", Mystery e Paris Aparteent, Jucy Foley, " 3.71 avg rating - 56, 901 ratings", Romance g Dies Dreaming, Michiel Bonziles," 4.48 avg rating - 56, 901 ratings", Fiction a of Tranquility, Enily 51. John Mandel, " 4.25 avg rating - 24, 260 ratings", Fiction a of Ioranguility, Enily 51. John Mandel, " 4.25 avg rating - 24, 260 ratings", Fiction					
) Th 4 "H 5 A 6 Bo 7 Th	Mankes of num_collect nover, 4.50 mg rating - 55,42 rating , howeve Maidwita Power, 5.80 mg rating - 116,212 ratings", Mytery oute of Sky and Breath (Crescent City, 42)", Sarch J. Nuas, * 4.55 any rating - 102,285 ratings", Fantasy Elcker in the Dark, Stacy Willingham, * 4.01 ang rating - 58,650 ratings", Thriller ok Lovers, Isily Henry, * 4.40 ang rating - 71,355 ratings", Mystery B Paris Agartment, Lucy Foley, * 3.71 ang rating - 56,300 ratings", Remark g Diss Dreaming, Michiel Consider, * 4.03 ang rating - 14,280 ratings", Fiction a of Tranquility, Kmily St. John Mandel, * 4.25 ang rating - 24,280 ratings", Fiction a folder Caroline Gener Medrice, * 4.04 ang rating - 44,280 ratings", Fiction					
5 A 6 Bo 7 Th	e Main, artose, J. aso Wg reting - 10,121 raings, hystery oue of Sky and Breach (Crescut City, 21); Sarah J. Naus, "4.55 avg rating - 102,286 ratings", Fantasy Filcker in the Dark, Stacy Willingham," 4.00 avg rating - 58,659 ratings", Thriller ek lovers, faily Henry, "4.48 avg rating - 71,355 ratings", Mystery e Paris Agarteent, Lucy Folgy, "3.71 avg rating - 56,901 ratings", Romace ga Dies Dreaming, Köchitl González, "4.03 avg rating - 14,935 ratings", Fiction a of Tranquility, fmily St. John Mandel, "4.25 avg rating - 24,280 ratings", Fiction is Order Gener Medricht, "4.08 avg rating - 40,810 ratings", Fiction					
5 A 6 Bo 7 Th	Sole or Sky and Freen (created triy, at) ,safan J. Namas, 4.55 ang Fating - 104,200 ratings ,rantasy Ficker in the Dark,Stacy Willingham, 4.40 ang rating - 58,650 ratings",Theiler ok (Dowrs,Ently Henry, 4.40 ang rating - 71,355 ratings",Mystery e Paris Apartment,Locy Foley, 7.3,71 ang rating - 96,901 ratings",Romance ga Dies Dreaming,Michiel Gonziles, "4.00 ang rating - 14,955 ratings",Fiction a of Tranquility,Emily St. John Mandel, "4.25 ang rating - 24,200 ratings",Fiction is Orden Courde Gonze Hendrick - 4.00 ang rating - 40,955 ratings",Fiction					
6 Bo	Ficker in the Gara, Stacky Millinghum, 4.03 and Facing - 36,000 rating, InFilter of Lowers, Enly Henry, 4.43 and seg rating - 71,355 ratings", Mystery e Paris Apartement, Lucy Foley, " 3.71 and rating - 56,000 ratings", Romance ga Dies Dreaming, Nochill González, " 4.03 and rating - 14,935 ratings", Fiction a of Tranquility, Enly 51. John Mandel, " 4.25 and rating - 24,260 ratings", Fiction a folder Garani Gara Handfring - 4.04 and rating - 40,810 ratings", Fiction					
7 Th	ok Lovers, mily menry, a ve ang raing - 71,355 raing, rhystery Paris Aparteent Lycv Folgy, 3.71 aug raing = 55,001 raings, Romance ga Dies Dreaming, Nochill González, 4.83 aug raing = 14,935 raings, Fiction a of Tranquility, fimily 51. John Mandel, 4.25 aug raing - 24,260 raings, Fiction e foldes (camel Gener Modelring = 4.04 aug raing = 4.04 aug raing)					
	e Varis Apartment, Very Yozy, S./I ang Tating - 20, 200 rating, Industry go Dos Dreaming, Modrid Loczialer, 4.03 ang rating - 14,950 ratings, Fiction a of Tranquility, Emily 51. John Mandel, 4.25 ang rating - 24,260 ratings, Fiction a Colden Canadi Gener Modelrie, 4.04 and rating - 40.810 relient. Theiller					
	ge Uses Urewaing,Johnsti Gonzalez, 4.03 avg reting - 14,355 ratings, ristion a of Tranquility,Imily St. John Mandel, 4.25 avg rating - 24,260 ratings",Fiction a Golden Conduct Scare Menderick - 4.04 avg rating - 40.810 ratings", Fiction					
01	a of Franquility,tmily St. John Handel, 4.25 avg rating - 24,260 ratings ,Fiction					
7 20	A DATINGS LINES A STREET BOOMSTICKS. ALL AND AND THE STREET STREET					
10 10	The Golden Couple,Green Hendricks, 4.04 aug rating - 50,819 ratings , Inriller					
11 10	How High we so in the Dark, sequoia magamarsu, 3.94 avg rating - 10,055 ratings , Science Fiction					
11 75	ine Last Mouse on the Street,Dame Chamberlain, 4.28 avg rating - 18,722 ratings Mistorical					
14 81	The Final Flaw, Michael R. Sullivan, 4.37 avg rating - 43 ratings, Science Fiction					
11 04	bloomsbury uiris, matalle semer, 4.13 avg rating - 940 ratings, mistorical					
15 00	the Italian Summer, Rebecta serie, 3.73 avg rating = 31,335 ratings, Fiction					
10 10	ine utamono tyy,kate Quini, 4.55 avg rating - 17,526 ratings ,Historical					
10 10	ine overnight Guest, Heather Gueenkait, 4.00 avg rating - 11,752 ratings, inviller					
	ine Hourgiass innone (ine land sequence, as) ,k.u. towards, 4.73 avg rating - dos ratings, rantasy					
10 10	Dessons in Cremistry, Bonnie Garmus, 4.40 avg rating - 22,002 ratings ,Fiction					
	the broaded diris of (wayward children, w/) seeman neutres, will any neutral - 7,573 ratings (randay					
22 40	Search of a strategiout Statut, 4.20 and rating - 51 failing humans					
11 75	a Book of Cold Civian Classa Cf. Jamas " 3 01 was entire - 35 101 entires" Montany					
14 76	e buck of cole cases,simone sc. sames, stat avg rating - ssiter ratings involvely a Violia Coerciente Receive Clarumb * 8 30 sus esting - 8 104 estings Montage					
24 IB	 Violan Comparenzami Successor, m. Au ang ratang - 6,224 ratangs prysicity Choise Affolio Blins de Compare 7: 25 per enting - 37 201 estimates "Bilinear" 					
35 84	Packlass Cials Puckal Hunkles * 2 52 non matian					
11 76	The Marid Creat Cius Tara Tashalla Burton 1 Cl aut enting - 510 rations Eistion					
21 84	In me worst cannot sive; rang isabella burton; 3:33 ang taling = 950 falings; Fiction By Equations to Vill Mar & Noval Parkant Namion & 36 mum ration - 31 rations Historical Ficking					
20 44	 ry workings, reads to vale me, a myrespector memory 4.20 Mg failing = 31 failing (1500 2011 failing) Making J, Manis ef Complete Randing County of Charl Yandan. These 7 75 and rates and and an official statement of the second se					
10 51	ster Mother Mareion Vanessa Riley, 4-35 aug rating - 17 ratings Historical					
11 Te	11 Ne the Truth Kiersten Rodelin." 3.91 avg rating = 2.592 ratings" Audiobook					

IV. CONCLUSION

While this project may not be as sophisticated as web scrapers made by large corporations, there is enough scope in this application to make a decent impact in the world of book recommendation. By first scraping and then utilizing a set of information like genre and ratings, users may be recommended books based on collaborative andcontentbased recommendation techniques, that would help both the users and the business. Users' search time for the right book may be significantly reduced, thereby the saved time may be invested in reading the recommended book.

V. REFERENCES

 Saurkar, Anand V., Kedar G. Pathare and Shweta A. Gode, An Overview On Web ScrapingTechniques And Tools, International Journal on Future Revolution in Computer Science&Communication Engineering, pages 363-367, 2018.

- [2]. Liu B., Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, pages 627- 666, 2010.
- [3]. PratikshaAshiwal, S.R. Tandan, Priyanka Tripathi and Rohit Miri, Web Information Retrieval UsingPython and BeautifulSoup, International Journal for Research in Applied Science &EngineeringTechnology (IJRASET), pages 335-339, 2016.
- [4]. Rahul Dhawani, Marudav Shukla, Priyanka Puvar, Bhagirath Prajapati, A Novel Approach toWebScraping Technology, International Journal of Advanced Research in Computer Science and SoftwareEngineering, Volume 5, Issue 5, 2015.

PROBLEM	IMPLEMENTED SOLUTION
The Good Reads page throws the registration pop-up, once the user has accessed around 10 books.	Using Selenium, I checked for the presence of a sign-in pop-up on the page. If a pop up was encountered, page refresh function of selenium was invoked, resulting in the page being displayed in its normal state, without the sign-in pop up bothering us
Selector of the "Genre Element" changing on page refresh	When the Good Reads page was refreshed, to deal with the 'sign-in' pop-up, the program could no longer identify that 'Genre Element' which it was seamlessly able to find, pre-refresh. On Printing the HTML code in the 'except' snippet, I observed that the locator of the 'Genre Element' has changed. I handled this using programming conditional statements on 'locators' - pre and post refresh.
Some Good Read books were in languages other than English, and did not have a Genre associated with it.	This is a special and rare occurrence on the Good Reads page for a book not to have a Genre. This was handled, using conditions statements.
Program was trying to find the Genre Element before the entire page loads.	Selenium's Desired Capabilities allows to set the 'page- strategy' to 'normal' which will allow the program to run only after the page has fully loaded.
Program execution was slow	To improve the speed of the program, Selenium helped us to have some options with Chrome Browser, in which I chose not to load images of the page, because Image takes more time to load than text. Secondly, Selenium also gives us the options to run the Chrome Browser in Headless mode. Both these helped us to fasten the execution by 50% with regards to execution time.

TABLE NO.1. Chronological Order: Discovery of problem navigating through them successfully