

## Homework Solution 6

Task 1 <- Pick at least 3 web scraping toolkits and use them to extract the data from a web site of your choice

I tried many tools but few were easier to use then others. The three toolkits I chose to write the report on, are:


- Kimono
- Parse Hub
- Import.io

➔ I chose Goodreads.com to Scrape. My aim was to get the name of the book, the author name, rating and No of editions for the books.


➤ The screen shot below is of Parse hub:

The toolkit is available for free use as well as for Premium and Professional use as well. The tool was kind of difficult to use as one has to adhere to the steps and it takes getting use to. One thing which I would like to bring to attention is it only works with Firefox. Which kind of is a negative as you have to specially download Firefox if you don't use it. One of the positive things it tutorials which help learning a lot easier


Page 1 of about 129640 results (0.48 seconds)




**The Shadow of the Wind (The Cemetery of Forgotten Books, #1)**  
by Carlos Ruiz Zafón, Lucia Graves (Translator)  
★★★★☆ 4.22 avg rating — 250,552 ratings — published 2001 — 239 editions

Want to Read 


Rate this book  
★★★★☆



**The Jungle Books**  
by Rudyard Kipling, Alev Lytle Croutier (Afterword)  
★★★★☆ 3.98 avg rating — 56,928 ratings — published 1895 — 114 editions

Want to Read 


Rate this book  
★★★★☆



A sister's bond.

ADD TO SHELF


sponsored by

 **The Life of a Christmas Tree**

| "names_name of book"   | "names_author"      | "names_rating"                                       | "names_editions" |
|--|---------------------|--|------------------|
| "The Shadow of the Wind (The Cemetery of Forgotten Books, #1)" | "Carlos Ruiz Zafón" | "4.22 of 5 stars\n4.22 avg rating — 250,552 ratings" | "239 editions"   |
| "The Jungle Books"   | "Rudyard Kipling"   | "3.98 of 5 stars\n3.98 avg rating — 56,928 ratings"  | "114 editions"   |
| "The Goose Girl (The Books of Bayern, #1)"                     | "Shannon Hale"      | "4.18 of 5 stars\n4.18 avg rating — 84,839 ratings"  | "29 editions"    |

☒ Sample enabled ⓘ

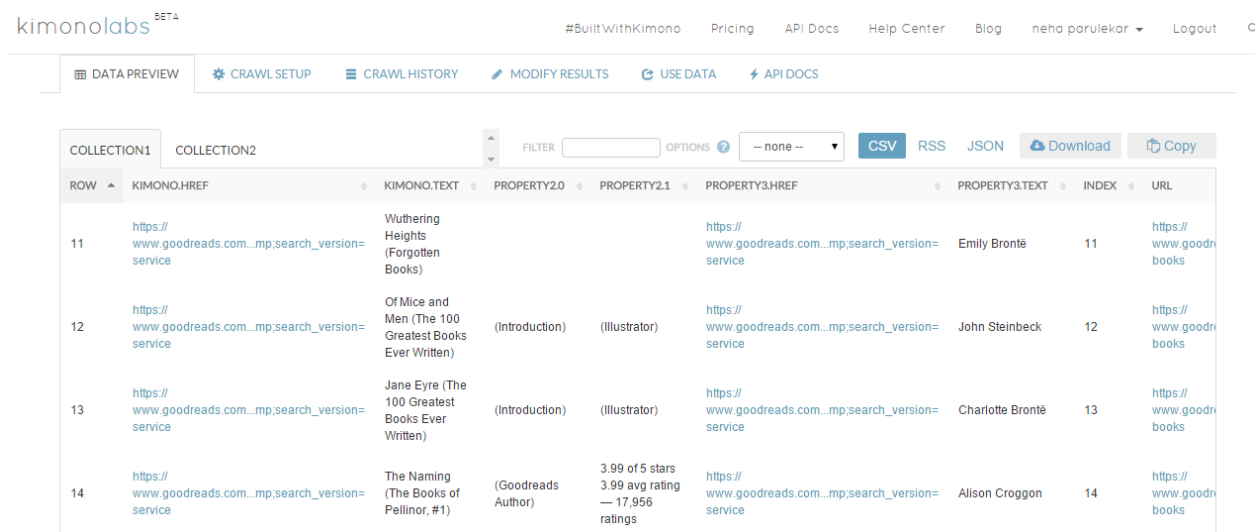
☒ Visuals enabled ⓘ

ParseHub Help 

## Homework Solution 6

➤ The screen shot below is of Kimono:

Free API is available for Kimono. The toolkit is pretty easy to use. The purpose of Kimono is to convert unorganized data to an organized data. Kimono Labs allows you to extract this data either on demand or as a scheduled job. Once you've extracted the data, it then allows you to either download it via a file or extract it via their own API. This is where Kimono really shines—it basically allows you to take any website or data source and turn it into an API or automated export.



The screenshot shows the Kimono Labs interface. At the top, there's a navigation bar with links for #BuiltWithKimono, Pricing, API Docs, Help Center, Blog, and a user profile for neha.parulekar. Below this is a secondary navigation bar with tabs for DATA PREVIEW, CRAWL SETUP, CRAWL HISTORY, MODIFY RESULTS, USE DATA, and API DOCS. The main area displays a table of data from 'COLLECTION2'. The table has columns for ROW, KIMONO.HREF, KIMONO.TEXT, PROPERTY2.0, PROPERTY2.1, PROPERTY3.HREF, PROPERTY3.TEXT, INDEX, and URL. The data rows show book information from Goodreads, including titles like 'Wuthering Heights', 'Of Mice and Men', 'Jane Eyre', and 'The Naming', along with their authors and Goodreads links.

| ROW | KIMONO.HREF   | KIMONO.TEXT   | PROPERTY2.0        | PROPERTY2.1  | PROPERTY3.HREF  | PROPERTY3.TEXT   | INDEX | URL   |
|-----|---|---|--------------------|--|---|------------------|-------|---|
| 11  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Wuthering Heights (Forgotten Books)                   |                    |  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Emily Brontë     | 11    | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> |
| 12  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Of Mice and Men (The 100 Greatest Books Ever Written) | (Introduction)     | (Illustrator)  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | John Steinbeck   | 12    | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> |
| 13  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Jane Eyre (The 100 Greatest Books Ever Written)       | (Introduction)     | (Illustrator)  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Charlotte Brontë | 13    | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> |
| 14  | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | The Naming (The Books of Pellinor, #1)                | (Goodreads Author) | 3.99 of 5 stars<br>3.99 avg rating<br>— 17,956 ratings | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> | Alison Croggon   | 14    | <a href="https://www.goodreads.com...mp;search_version=service">https://www.goodreads.com...mp;search_version=service</a> |

➤ The screen shot below is of Import.io:

The tool kit is pretty user friendly. The setup takes a little time. But the tutorials available help you get the desired results in a very short time. The data that users collect is stored on import.io's cloud servers and can be downloaded as CSV, Excel, Google Sheets or JSON and shared. Users can also generate an API from the data allowing them to easily integrate live web data into their own applications or third party analytics and visualization software.

## Homework Solution 6

Query your API

https://www.goodreads.com/search?utf8=%E2%9C%93&query=books

Run query

Download

| books                               | author            | ratings                                | editions     |
|-------------------------------------|-------------------|--|--------------|
| The Shadow of the Wind (The Ce...   | Carlos Ruiz Zafón | 4.22 of 5 stars 4.22 avg rating — 2... | 239 editions |
| The Jungle Books                    | Rudyard Kipling   | 3.98 of 5 stars 3.98 avg rating — 5... | 114 editions |
| The Goose Girl (The Books of Bay... | Shannon Hale      | 4.18 of 5 stars 4.18 avg rating — 8... | 29 editions  |

### Recommendation:

Though It is not very easy to and takes getting used to I prefer Pasre hub. Once you get a hang of it, the tool is very efficient in extracting data required. Based on the data I could scrape,With the other two I would still need to filter the data. I got the exact results I was aiming for in Parse Hub.

### Screenshot of the dataframe of the Pasrehub extracted file in R

|    | names_name.of.book   | names_author        | names_rating   | names_editions |
|----|--|---------------------|--|----------------|
| 1  | The Shadow of the Wind (The Cemetery of Forgotten Books, #1)         | Carlos Ruiz Zafón   | 4.22 of 5 stars\n4.22 avg rating — 250,552 ratings   | 239 editions   |
| 2  | The Jungle Books   | Rudyard Kipling     | 3.98 of 5 stars\n3.98 avg rating — 56,928 ratings    | 114 editions   |
| 3  | The Goose Girl (The Books of Bayern, #1)                             | Shannon Hale        | 4.18 of 5 stars\n4.18 avg rating — 84,839 ratings    | 29 editions    |
| 4  | Matar a un ruiseñor  | Harper Lee          | 4.24 of 5 stars\n4.24 avg rating — 2,525,226 ratings | 384 editions   |
| 5  | The Bachman Books  | Richard Bachman     | 4.06 of 5 stars\n4.06 avg rating — 40,190 ratings    | 22 editions    |
| 6  | The Great Gatsby (100 Greatest Books of All Time)                    | F. Scott Fitzgerald | 3.86 of 5 stars\n3.86 avg rating — 2,192,433 ratings | 1006 editions  |
| 7  | Pride and Prejudice (The 100 Greatest Books Ever Written)            | Jane Austen         | 4.23 of 5 stars\n4.23 avg rating — 1,667,293 ratings | 2316 editions  |
| 8  | The Hobbit Publisher: Houghton Mifflin Books; Later printing edition | J.R.R. Tolkien      | 4.22 of 5 stars\n4.22 avg rating — 1,786,516 ratings | 824 editions   |
| 9  | Romeo And Juliet (Forgotten Books)                                   | William Shakespeare | 3.72 of 5 stars\n3.72 avg rating — 1,388,572 ratings | 1221 editions  |
| 10 | Of Mice and Men (The 100 Greatest Books Ever Written)                | John Steinbeck      | 3.81 of 5 stars\n3.81 avg rating — 1,238,240 ratings | 302 editions   |
| 11 | Jane Eyre (The 100 Greatest Books Ever Written)                      | Charlotte Brontë    | 4.08 of 5 stars\n4.08 avg rating — 1,052,113 ratings | 1628 editions  |
| 12 | Little Women (The 100 Greatest Books Ever Written)                   | Louisa May Alcott   | 4.01 of 5 stars\n4.01 avg rating — 1,082,199 ratings | 1273 editions  |
| 13 | Fahrenheit 451 (Independent Banned Books series #25)                 | Ray Bradbury        | 3.95 of 5 stars\n3.95 avg rating — 942,512 ratings   | 437 editions   |
| 14 | Books of Blood: Volume One (Books of Blood #1)                       | Clive Barker        | 3.98 of 5 stars\n3.98 avg rating — 12,138 ratings    | 42 editions    |
| 15 | The Naming (The Books of Pellinor, #1)                               | Alison Croggon      | 3.99 of 5 stars\n3.99 avg rating — 17,965 ratings    | 35 editions    |
| 16 | Brave New World (The 100 Greatest Books Ever Written)                | Aldous Huxley       | 3.94 of 5 stars\n3.94 avg rating — 858,189 ratings   | 4 editions     |
| 17 | Wuthering Heights (Forgotten Books)                                  | Emily Brontë        | 3.8 of 5 stars\n3.80 avg rating — 805,361 ratings    | 1752 editions  |
| 18 | The Books of Magic   | Neil Gaiman         | 4.07 of 5 stars\n4.07 avg rating — 10,763 ratings    | 17 editions    |
| 19 | Lo que el viento se llevó  | Margaret Mitchell   | 4.26 of 5 stars\n4.26 avg rating — 748,348 ratings   | 318 editions   |
| 20 | Enna Burning (The Books of Bayern, #2)                               | Shannon Hale        | 3.96 of 5 stars\n3.96 avg rating — 29,668 ratings    | 21 editions    |