

## R Programming for Data Analytics Assignment #4

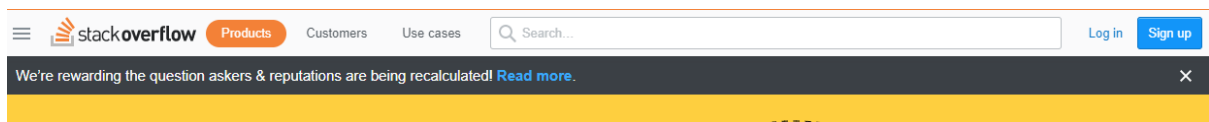
- Web Data Scraping -

Submit three files

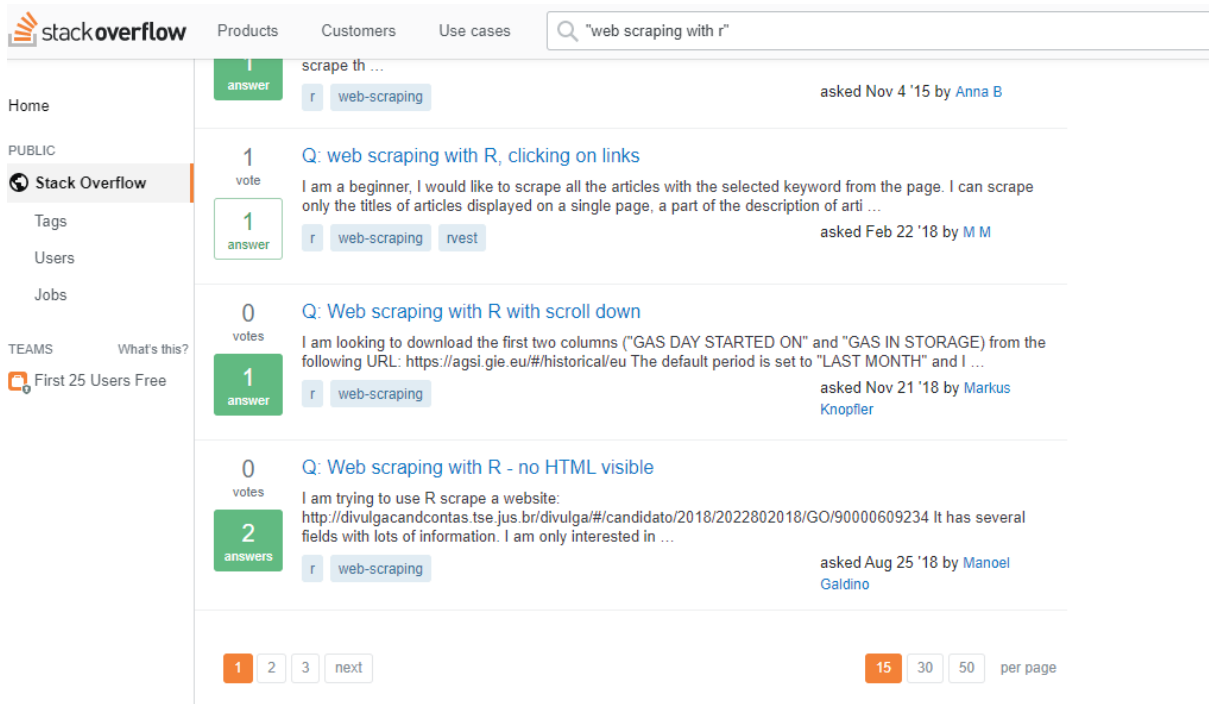
- R script for data scraping
- Rmd file for the plots
- HTML files generated from the Rmd file

[1] Visit the Stackoverflow website: <https://stackoverflow.com/>

[2] Use your own query that returns the result at least 10 pages



For example, the query “web scraping with r” returns the results in only 3 pages (not acceptable query)



However, the query “web scraping” returns the results in 33 pages (acceptable query)

The screenshot shows the Stack Overflow search results for the query "web scraping". The search bar at the top contains the text "web scraping". The results are displayed in a list format. The first result is titled "Q: Web scraping image inside canvas" and has 5 votes and 1 answer. The second result is titled "Q: How to web-scraping from a (javascript?) website?" and has -1 votes and 1 answer. The third result is titled "Q: Web scraping in dialogflow" and has 0 votes and 2 answers. The left sidebar shows the Stack Overflow logo and navigation links. The bottom of the page shows pagination controls indicating 33 pages of results.

[Note: the following instructions are based on the query example “web scraping.” Students must use their own query phrase to scrape the data]

[3] Scrape the following information from the result in the first 10 pages (not from all pages)

- Title of the question (Title)

## Web scraping with Java

Asked 9 years, 4 months ago Active 4 months ago Viewed 126k times

71 I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.

Are there ways other than web scraping?

★ java web-scraping frameworks

45

share improve this question

edited Feb 15 at 8:25

5377037

8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38

NoneType

867 ● 1 ● 7 ● 10

add a comment

- The date the question is asked (Date)
  - (Note) The date must be stored as "2010-07-08T09:38:26" as shown in the HTML source code instead of "9 years, 4 months ago" as shown in the web page.

## Web scraping with Java

Asked 9 years, 4 months ago Active 4 months ago Viewed 126k times

71

I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.



Are there ways other than web scraping?



java web-scraping frameworks

45

share improve this question

edited Feb 15 at 8:25



5377037

8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38



NoneType

867 ● 1 ● 7 ● 10

add a comment

```
<div class="grid fw-wrap pb8 mb16 bb bc-black-2" style="user-select: auto;">
  <div class="grid--cell ws-nowrap mr16 mb8" title="2010-07-08 09:38:26Z" style="
    user-select: auto;">
    <span class="fc-light mr2" style="user-select: auto;">Asked</span>
    <time itemprop="dateCreated" datetime="2010-07-08T09:38:26" style="user-select:
    auto;">9 years, 4 months ago</time> == $0
  </div>
  <div class="grid--cell ws-nowrap mr16 mb8" style="user-select: auto;">...</div>
  <div class="grid--cell ws-nowrap mb8" title="Viewed 126,487 times" style="user-
  select: auto;">...</div>
</div>
```

- The number of views (Views)
  - (Note) The number of views must be stored as "126487" as shown in the HTML source code instead of "126k" as shown in the web page.

## Web scraping with Java

Asked 9 years, 4 months ago   Active 4 months ago   Viewed **126k times**

71

I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.



Are there ways other than web scraping?



java   web-scraping   frameworks

45

share   improve this question

edited Feb 15 at 8:25



5377037

8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38



NoneType

867 ● 1 ● 7 ● 10

add a comment

```

<div class="grid fw-wrap pb8 mb16 bb bc-black-2" style="user-select: auto;">
  <div class="grid--cell ws-nowrap mr16 mb8" title="2010-07-08 09:38:26Z" style="
    user-select: auto;">
    <span class="fc-light mr2" style="user-select: auto;">Asked</span>
    <time itemprop="dateCreated" datetime="2010-07-08T09:38:26" style="user-select:
    auto;">9 years, 4 months ago</time> == $0
  </div>
  <div class="grid--cell ws-nowrap mr16 mb8" style="user-select: auto;">...</div>
  <div class="grid--cell ws-nowrap mb8" title="Viewed 126,487 times" style="user-
  select: auto;">...</div>
</div>

```


- Main body of the question (Question)

## Web scraping with Java


Asked 9 years, 4 months ago Active 4 months ago Viewed 126k times

  
71

I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.




Are there ways other than web scraping?

 java web-scraping frameworks

45

share improve this question

edited Feb 15 at 8:25  
 5377037  
8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38  
 NoneType  
867 ● 1 ● 7 ● 10

add a comment

- Tags of the question (Tags)
  - (Note) Store multiple tags as a single string
    - ◆ Example: “java web-scraping frameworks” for the tags below

## Web scraping with Java

Asked 9 years, 4 months ago Active 4 months ago Viewed 126k times

  
71

I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.




Are there ways other than web scraping?

 java web-scraping frameworks

45

share improve this question

edited Feb 15 at 8:25  
 5377037  
8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38  
 NoneType  
867 ● 1 ● 7 ● 10

add a comment

- Comments of the question (Q\_comments)
  - If there are more than one comments, paste them as a single string with the collapse symbol "OOO"
    - ◆ Example:
      - Comment 1: Thanks
      - Comment 2: Excellent library
      - Comment 3: Great
    - ◆ Comment variable stored in the dataframe: "Thanks OOO Excellent library OOO Great"
  - If there is no comment, skip this variable

## What is the difference between web-crawling and web-scraping

Asked 8 years, 11 months ago   Active 1 year, 11 months ago   Viewed 58k times

85

**This question already has answers here:**  
[crawler vs scraper](#) (4 answers)  
 Closed last year.

★ Is there a difference between Crawling and Web-scraping?

26 If there's a difference, what's the best method to use in order to collect some web data to supply a database for later use in a customised search engine?

search-engine   web-scraping   web-crawler

share   improve this question

edited Sep 6 '13 at 16:53

asked Dec 1 '10 at 17:54

**wassimans**

6,446 ● 9 ● 39 ● 56

14 Scraping means pulling content from a page. Crawling means following links to reach numerous pages. Crawlers have to scrape, and that's for two reasons: one is that useful crawlers don't just traverse pages for nothing; they collect info (e.g. indexing words to build a search index for a search engine). Secondly, they have to discover links to other pages. – Kaz Oct 10 '13 at 23:50

add a comment

- Main body of each answer (Answer)

▲  
108



Crawling would be essentially what Google, Yahoo, MSN, etc. do, looking for ANY information. Scraping is generally targeted at certain websites, for specific data, e.g. for price comparison, so are coded quite differently.

Usually a scraper will be bespoke to the websites it is supposed to be scraping, and would be doing things a (good) crawler wouldn't do, i.e.:

- Have no regard for robots.txt
- Identify itself as a browser
- Submit forms with data
- Execute Javascript (if required to act like a user)

share improve this answer

edited Jun 13 '13 at 20:44

answered Dec 1 '10 at 18:07



the Tin Man

141k ● 27 ● 182 ● 262



Ben

3,536 ● 1 ● 18 ● 20

- 1 @Ben Do you know where I can find out more about how a web scraper identifies itself as a browser? Wikipedia says "implementing low-level Hypertext Transfer Protocol (HTTP)" but I'd like to really know more how it works. – Abdul Jul 13 '15 at 18:28
- 1 @Abdul in HTTP requests, you can specify a "User-Agent" property to identify yourself. If you for instance set this to "Mozilla/5.0 ... Chrome" or something that Chrome uses, your scraper would look like a browser to the server. – Amani Kilumanga Mar 16 '16 at 0:17

add a comment

▲  
58



Yes, they are different. In practice, you may need to use both.

(I have to jump in because, so far, the other answers don't get to the essence of it. They use examples but don't make the distinctions clear. Granted, they are from 2010!)

**Web scraping**, to use a minimal definition, is the process of processing a web document and extracting information out of it. You can do web scraping without doing web crawling.

**Web crawling**, to use a minimal definition, is the process of iteratively finding and fetching web links starting from a list of seed URL's. Strictly speaking, to do web crawling, you have to do some degree of web scraping (to extract the URL's.)

To clear up some concepts mentioned in the other answers:

- `robots.txt` is intended to apply to any automated process that accesses a web page. So it applies to both crawlers and scrapers.
- 'Proper' crawlers and scrapers, both, should identify themselves accurately.

Some references:

- [Wikipedia on web scraping](#)
- [Wikipedia on web crawlers](#)
- [Wikipedia on robots.txt](#)

share improve this answer

answered Jun 21 '12 at 17:08



David J.

25.6k ● 16 ● 96 ● 150

add a comment

- Comments for each answer: the same rule for the comments of the question is applied (A Comments)

## Web scraping with Java

Asked 9 years, 4 months ago   Active 4 months ago   Viewed 126k times

71 I'm not able to find any good web scraping Java based API. The site which I need to scrape does not provide any API as well; I want to iterate over all web pages using some `pageID` and extract the HTML titles / other stuff in their DOM trees.

Are there ways other than web scraping?



java   web-scraping   frameworks

45

share   improve this question

edited Feb 15 at 8:25



5377037  
8,796 ● 12 ● 33 ● 66

asked Jul 8 '10 at 9:38



NoneType  
867 ● 1 ● 7 ● 10

add a comment

### 9 Answers

active   oldest   votes

#### jsoup

94

Extracting the title is not difficult, and you have many options, search here on Stack Overflow for "Java HTML parsers". One of them is [Jsoup](#).



You can navigate the page using DOM if you know the page structure, see <http://jsoup.org/cookbook/extracting-data/dom-navigation>

It's a good library and I've used it in my last projects.

share   improve this answer

edited Jun 21 '18 at 0:45



Basil Bourque  
146k ● 38 ● 473 ● 659

answered Jul 8 '10 at 9:44



Wajdy Essam  
3,667 ● 21 ● 30

2 Thanks, it's a nice library with no dependencies so it's quite lightweight. Also, it's headless so it doesn't need a browser (I've had problems with Selenium opening Chrome and I couldn't use HtmlUnit at all). Selenium must be more realistic but this library might serve the purpose in most scraping cases and it's really easy to setup: add the dependency and you're good to go. – [Ferran Maylinch](#) May 31 '14 at 17:13

Excellent library indeed. Easy setup and powerful regex support. `doc.select("li[id^=cosid_]")`. Cool. – [EMM](#) Jul 19 '16 at 15:21

I have recently open sourced my web scraping framework that not only allows you to parse the documents with Jsoup and HtmlUnit, but also handles the parallelization for you and can manage a large pool of proxy servers if required: [github.com/subes/invesdwin-webproxy](https://github.com/subes/invesdwin-webproxy) – [subes](#) Jun 9 '17 at 18:57

@subes can your framework be used for web analytics testing ? – [vikramvi](#) Nov 11 '17 at 10:53

My requirement is to do "Web Analytics" automation, is Jsoup can do all the testing activities ? – [vikramvi](#) Nov 11 '17 at 10:54

Well you can automate users coming from various parts of the world by creating proxy enabled bots (just web scrapers that navigate your analytics enabled website). Though be aware that some analytics packages filter users from public proxies. So better use a service for this or your own servers with proxies installed. Or you can navigate your analytics website itself and collect data for a dashboard there. So yes, but JSoup might not be enough for this since HtmlUnit provides cookies, JS and other essential support for this. – [subes](#) Nov 12 '17 at 11:17

add a comment



[4] Store the result to the dataframe named "Stackoverflow\_QA"

(Dataframe format)

Title	Date	Views	Question	Tags	Q_comments	Answer	A_Comments

[5] Draw at least five different plots based on the collected data

- You can create additional variables from the above dataframe to draw plots
- Some examples
  - Histogram on the "Views" variable
  - Scatter plot (x-axis: number of characters in the title, y-axis: number of characters in the question)
  - Box plot of "Views" with regard to the number of Tags
  - Etc.