

Reading data from the web

Jeffrey Leek

May 18, 2016

Webscraping

Webscraping: Programatically extracting data from the HTML code of websites.


- ▶ It can be a great way to get data How Netflix reverse engineered Hollywood
- ▶ Many websites have information you may want to programatically read
- ▶ In some cases this is against the terms of service for the website
- ▶ Attempting to read too many pages too quickly can get your IP address blocked

http://en.wikipedia.org/wiki/Web_scraping

Example: Google scholar

Click to edit [Jeff Leek - Google Scholar](#) [scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en&oi=ao](#) [jtleek@gmail.com](#)

Web Images More...


 **Jeff Leek** [Edit](#)
Assistant Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health [Edit](#)
[Statistics](#) - [Computing](#) - [Genomics](#) - [Personalized Medicine](#) - [Scientific Communication](#) [Edit](#)
Verified email at [jhsp.h.edu](#) [Edit](#)
My profile is public [Edit](#) [Link](#) [Homepage](#) [Edit](#)

[Change photo](#)

Citation indices

| | All | Since 2008 |
|-----------|------|------------|
| Citations | 1285 | 1146 |
| h-index | 10 | 10 |
| i10-index | 11 | 11 |

Citations to my articles



Select: [All](#), [None](#) [Actions](#) Show: [20](#) [5](#) [1-20](#) [Next >](#)

| Title / Author | Cited by | Year |
|--|----------|------|
| <input type="checkbox"/> Significance analysis of time course microarray experiments JD Storey, W Xiao, JT Leek, RG Tompkins, RW Davis Proceedings of the National Academy of Sciences of the United States of ... | 338 | 2005 |
| <input type="checkbox"/> Capturing heterogeneity in gene expression studies by surrogate variable analysis JT Leek, JD Storey PLoS Genetics 3 (9), e161 | 171 | 2007 |
| <input type="checkbox"/> EDGE: extraction and analysis of differential gene expression JT Leek, E Monsen, AR Dabney, JD Storey Bioinformatics 22 (4), 507-508 | 140 | 2006 |
| <input type="checkbox"/> Tackling the widespread and critical impact of batch effects in high-throughput data JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, D German, K ... Nature Reviews Genetics 11 (10), 733-739 | 133 | 2010 |
| <input type="checkbox"/> The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments JD Storey, JY Dai, JT Leek UW Biostatistics Working Paper Series, 260 | 107 | 2005 |
| Systems-level dynamic analyses of fate change in murine embryonic stem | | |

Google scholar

[Search Authors](#)

My Citations - Help

Follow this author

5 Followers

[Follow new articles](#)
[Follow new citations](#)

Add co-authors

[Add -](#) John D. Storey
[Add -](#) Rafael A Irizarry
[Add -](#) Ben Langmead
[Add -](#) Hector Corrada Br...
[Add -](#) wenzhong xiao
[Add -](#) W. Evan Johnson
[Add -](#) Alexander Lachm...
[Add -](#) Olga Troyanskaya
[Add -](#) Avi Ma'ayan
[Add -](#) Edoardo M Airoidi
[View all co-authors](#)

Co-authors

No co-authors

☐ Inviting co-author
[Send invitation](#)

`http://scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en`

Getting data off webpages - readLines()

```
con = url("http://scholar.google.com/citations?user=HI-I6C0")
htmlCode = readLines(con)
```

```
## Warning in readLines(con): incomplete final line found on file
## scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en'
```

```
close(con)
htmlCode
```

```
## Warning in readLines(con): incomplete final line found on file
## scholar.google.com/citations?user=HI-I6C0AAAAJ&hl=en'
```

Parsing with XML

```
library(XML)
```

```
## Loading required package: methods
```

```
url <- "http://scholar.google.com/citations?user=HI-I6C0AA"
```

```
html <- htmlTreeParse(url, useInternalNodes=T)
```

```
xpathSApply(html, "//title", xmlValue)
```

```
## [1] "Jeff Leek - Google Scholar Citations"
```

```
xpathSApply(html, "//td[@id='col-citedby']", xmlValue)
```

```
## list()
```

GET from the httr package

```
library(httr); html2 = GET(url)
content2 = content(html2,as="text")
parsedHtml = htmlParse(content2,asText=TRUE)
xpathSApply(parsedHtml, "//title", xmlValue)
```

```
## [1] "Jeff Leek - Google Scholar Citations"
```

Accessing websites with passwords

```
pg1 = GET("http://httpbin.org/basic-auth/user/passwd")  
pg1
```

```
## Response [http://httpbin.org/basic-auth/user/passwd]  
##   Date: 2016-05-18 14:57  
##   Status: 401  
##   Content-Type: <unknown>  
## <EMPTY BODY>
```

```
http://cran.r-project.org/web/packages/httr/httr.pdf
```

Accessing websites with passwords

```
pg2 = GET("http://httpbin.org/basic-auth/user/passwd",  
          authenticate("user", "passwd"))  
pg2
```

```
## Response [http://httpbin.org/basic-auth/user/passwd]  
##   Date: 2016-05-18 14:57  
##   Status: 200  
##   Content-Type: application/json  
##   Size: 47 B  
  
## No encoding supplied: defaulting to UTF-8.  
  
## {  
##   "authenticated": true,  
##   "user": "user"  
## }
```

```
names(pg2)
```


Using handles

```
google = handle("http://google.com")  
pg1 = GET(handle=google,path="/")  
pg2 = GET(handle=google,path="search")
```

<http://cran.r-project.org/web/packages/httr/httr.pdf>

Notes and further resources

- ▶ R Bloggers has a number of examples of web scraping
`http://www.r-bloggers.com/?s=Web+Scraping`
- ▶ The httr help file has useful examples `http://cran.r-project.org/web/packages/httr/httr.pdf`
- ▶ See later lectures on APIs