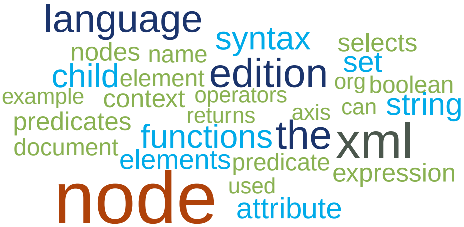
Introduction to Web Scraping with R

Scraping Multiple Pages



Simon Munzert | IPSDS

An advanced scraping scenario

Motivation

- until now, the toy examples were limited to single HTML pages
- often, we want to scrape data from multiple pages
- in such scenarios, automating the scraping process becomes **really** powerful
- also, the principles of polite scraping are more relevant

The scenario

Goal: examine download statistics of articles of the Journal of Statistical Software

- download HTML pages
- extract bibliometrical information

Tasks:

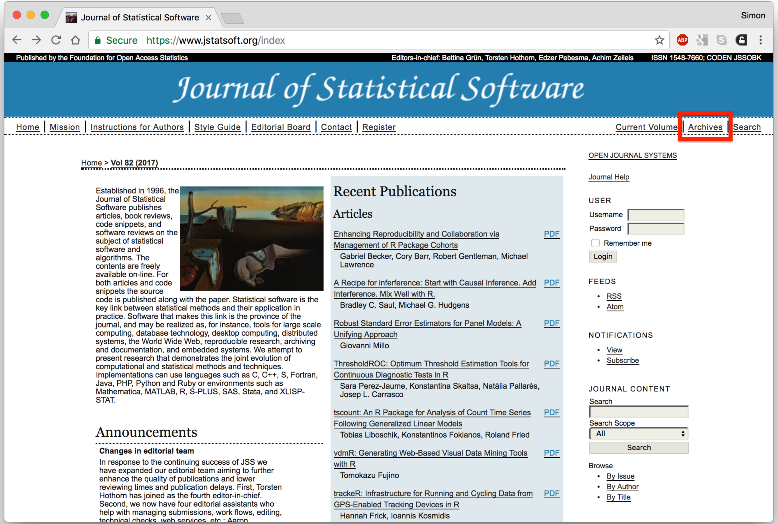
- identify relevant resources on http://www.jstatsoft.org/
- download HTML pages
- import them into R
- extract information via XPath

Scraping multiple pages with R

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"



Published by the Foundation for Open Access Statistics

Editors-in-chief: Bettina Grün, Torsten Hothorn, Edzer Pebesma, Achim Zeileis

ISSN 1548-7660; CODEN JSSOBK

Journal of Statistical Software

Home > Vol 82 (2017)

Established in 1996, the Journal of Statistical Software publishes articles, book reviews, code snippets, and software reviews on the subject of statistical software and algorithms. The contents are freely available on-line. For both articles and code snippets the source code is published along with the paper. Statistical software is the key link between statistical methods and their application in practice. Software that makes this link is the province of the journal, and may be realized as, for instance, tools for large scale computing, database technology, desktop computing, distributed systems, the World Wide Web, reproducible research, archiving and documentation, and embedded systems. We attempt to present research that demonstrates the joint evolution of computational and statistical methods and techniques. Implementations can use languages such as C, C++, S, Fortran, Java, PHP, Python and Ruby or environments such as Mathematica, MATLAB, R, S-PLUS, SAS, Stata, and XLISP-STAT.

Announcements

Changes in editorial team

In response to the continuing success of JSS we have expanded our editorial team aiming to further enhance the quality of publications and lower reviewing times and publication delays. First, Torsten Hothorn has joined as the fourth editor-in-chief. Second, we now have four editorial assistants who help with managing submissions, work flows, editing, technical checks, web services, etc.: Aaron

Recent Publications

Articles

Enhancing Reproducibility and Collaboration via Management of R Package Cohorts PDF
Gabriel Becker, Cory Barr, Robert Gentleman, Michael Lawrence

A Recipe for inferference: Start with Causal Inference. Add Interference. Mix Well with R. PDF
Bradley C. Saul, Michael G. Hudgens

Robust Standard Error Estimators for Panel Models: A Unifying Approach PDF
Giovanni Millo

ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R PDF
Sara Perez-Jaume, Konstantina Skaltsa, Natàlia Pallarès, Josep L. Carrasco

tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models PDF
Tobias Liboschik, Konstantinos Fokianos, Roland Fried

vdmR: Generating Web-Based Visual Data Mining Tools with R PDF
Tomokazu Fujino

trackeR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R PDF
Hannah Frick, Ioannis Kosmidis

OPEN JOURNAL SYSTEMS

Journal Help

USER

Username

Password

□ Remember me

Login

FEEDS

- RSS
- Atom

NOTIFICATIONS

- View
- Subscribe

JOURNAL CONTENT

Search

Search Scope

All

Search

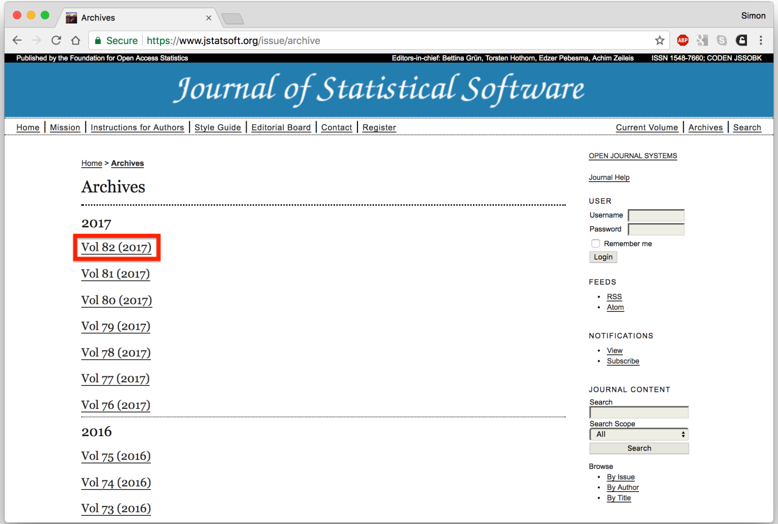
Browse

- By Issue
- By Author
- By Title

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"
- inspect the most recent volume



Published by the Foundation for Open Access Statistics

Editors-in-chief: Bettina Grün, Torsten Hothorn, Edzer Pebesma, Achim Zeileis

ISSN 1548-7660; CODEN JSSOBK

Journal of Statistical Software

Home | Mission | Instructions for Authors | Style Guide | Editorial Board | Contact | Register
Current Volume | Archives | Search

Home > **Archives**

Archives

2017

Vol 82 (2017)

Vol 81 (2017)

Vol 80 (2017)

Vol 79 (2017)

Vol 78 (2017)

Vol 77 (2017)

Vol 76 (2017)

2016

Vol 75 (2016)

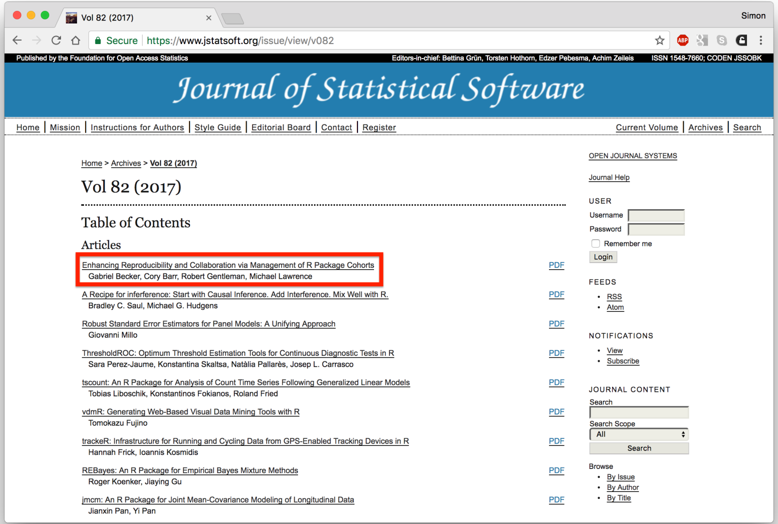
Vol 74 (2016)

Vol 73 (2016)

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"
- inspect the most recent volume
- inspect the first article



Vol 82 (2017)
Simon
Secure https://www.jstatsoft.org/issue/view/v082
Home | Mission | Instructions for Authors | Style Guide | Editorial Board | Contact | Register
Current Volume | Archives | Search

Published by the Foundation for Open Access Statistics

Editors-in-chief: Bettina Grün, Torsten Hothorn, Edzer Pebesma, Achim Zeileis

ISSN 1548-7660; CODEN JSSOBK

Journal of Statistical Software

Home > Archives > Vol 82 (2017)

Vol 82 (2017)

Table of Contents

Articles

Enhancing Reproducibility and Collaboration via Management of R Package Cohorts — PDF
Gabriel Becker, Cory Barr, Robert Gentleman, Michael Lawrence

A Recipe for inferference: Start with Causal Inference. Add Interference. Mix Well with R. — PDF
Bradley C. Saul, Michael G. Hudgens

Robust Standard Error Estimators for Panel Models: A Unifying Approach — PDF
Giovanni Millo

ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R — PDF
Sara Perez-Jaume, Konstantina Skaltsa, Natàlia Pallarès, Josep L. Carrasco

tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models — PDF
Tobias Liboschik, Konstantinos Fokianos, Roland Fried

vdmR: Generating Web-Based Visual Data Mining Tools with R — PDF
Tomokazu Fujino

trackeR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R — PDF
Hannah Frick, Ioannis Kosmidis

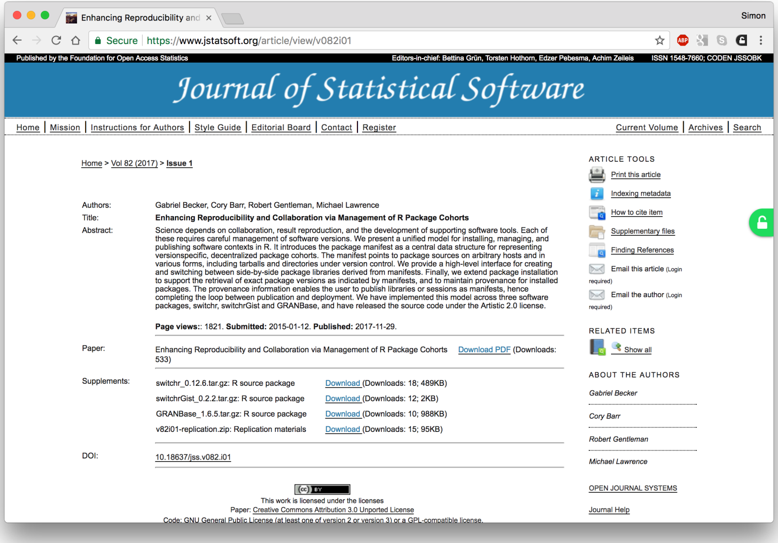
REBayes: An R Package for Empirical Bayes Mixture Methods — PDF
Roger Koenker, Jiaying Gu

jmcm: An R Package for Joint Mean-Covariance Modeling of Longitudinal Data — PDF
Jianxin Pan, Yi Pan

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"
- inspect the most recent volume
- inspect the first article


Enhancing Reproducibility and ×
Simon
Secure https://www.jstatsoft.org/article/view/v082i01

Published by the Foundation for Open Access Statistics Editors-in-chief: Bettina Grün, Torsten Hothorn, Edzer Pebesma, Achim Zeileis ISSN 1548-7660; CODEN JSSOBK

Journal of Statistical Software

Home | Mission | Instructions for Authors | Style Guide | Editorial Board | Contact | Register
Current Volume | Archives | Search

Home > Vol 82 (2017) > **Issue 1**

Authors: Gabriel Becker, Cory Barr, Robert Gentleman, Michael Lawrence

Title: **Enhancing Reproducibility and Collaboration via Management of R Package Cohorts**

Abstract: Science depends on collaboration, result reproduction, and the development of supporting software tools. Each of these requires careful management of software versions. We present a unified model for installing, managing, and publishing software contexts in R. It introduces the package manifest as a central data structure for representing versionspecific, decentralized package cohorts. The manifest points to package sources on arbitrary hosts and in various forms, including tarballs and directories under version control. We provide a high-level interface for creating and switching between side-by-side package libraries derived from manifests. Finally, we extend package installation to support the retrieval of exact package versions as indicated by manifests, and to maintain provenance for installed packages. The provenance information enables the user to publish libraries or sessions as manifests, hence completing the loop between publication and deployment. We have implemented this model across three software packages, switchr, switchrGist and GRANBase, and have released the source code under the Artistic 2.0 license.

Page views:: 1821. **Submitted:** 2015-01-12. **Published:** 2017-11-29.

Paper: Enhancing Reproducibility and Collaboration via Management of R Package Cohorts Download PDF (Downloads: 533)

Supplements:

switchr_0.12.6.tar.gz: R source package	Download (Downloads: 18; 489KB)
switchrGist_0.2.2.tar.gz: R source package	Download (Downloads: 12; 2KB)
GRANBase_1.6.5.tar.gz: R source package	Download (Downloads: 10; 988KB)
v82i01-replication.zip: Replication materials	Download (Downloads: 15; 95KB)

DOI: 10.18637/jss.v082.i01

This work is licensed under the licenses
Paper: Creative Commons Attribution 3.0 Unported License
Code: GNU General Public License (at least one of version 2 or version 3) or a GPL-compatible license.

ARTICLE TOOLS

- Print this article
- Indexing metadata
- How to cite item
- Supplementary files
- Finding References
- Email this article (Login required)
- Email the author (Login required)

RELATED ITEMS

- Show all

ABOUT THE AUTHORS

Gabriel Becker

Cory Barr

Robert Gentleman

Michael Lawrence

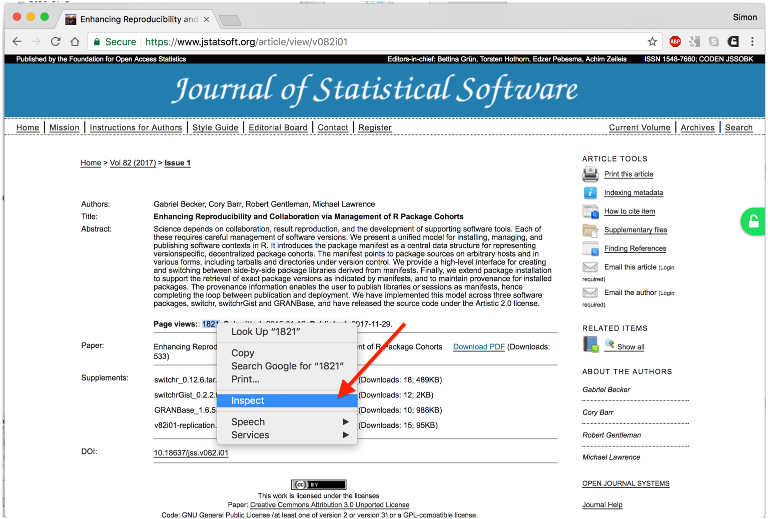
OPEN JOURNAL SYSTEMS

Journal Help

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"
- inspect the most recent volume
- inspect the first article
- inspect the page views element


Enhancing Reproducibility and ×
Simon
Secure
https://www.jstatsoft.org/article/view/v082i01

Published by the Foundation for Open Access Statistics   Editors-in-chief: Bettina Grün, Torsten Hothorn, Edzer Pebesma, Achim Zeileis   ISSN 1548-7660; CODEN JSSOBK

Journal of Statistical Software

Home | Mission | Instructions for Authors | Style Guide | Editorial Board | Contact | Register   Current Volume | Archives | Search

Home > Vol 82 (2017) > Issue 1

Authors: Gabriel Becker, Cory Barr, Robert Gentleman, Michael Lawrence

Title: **Enhancing Reproducibility and Collaboration via Management of R Package Cohorts**

Abstract: Science depends on collaboration, result reproduction, and the development of supporting software tools. Each of these requires careful management of software versions. We present a unified model for installing, managing, and publishing software contexts in R. It introduces the package manifest as a central data structure for representing versionspecific, decentralized package cohorts. The manifest points to package sources on arbitrary hosts and in various forms, including tarballs and directories under version control. We provide a high-level interface for creating and switching between side-by-side package libraries derived from manifests. Finally, we extend package installation to support the retrieval of exact package versions as indicated by manifests, and to maintain provenance for installed packages. The provenance information enables the user to publish libraries or sessions as manifests, hence completing the loop between publication and deployment. We have implemented this model across three software packages, switchr, switchrGist and GRANBase, and have released the source code under the Artistic 2.0 license.

Page views:: 1821 017-11-29.

Paper: Enhancing Reprod nt of R Package Cohorts Download PDF (Downloads: 533)

Supplements: switchr_0.12.6.tar. (Downloads: 18; 489KB)
switchrGist_0.2.2. (Downloads: 12; 2KB)
GRANBase_1.6.5 (Downloads: 10; 988KB)
v82i01-replication. (Downloads: 15; 95KB)

DOI: 10.18637/jss.v082.i01

This work is licensed under the licenses
Paper: Creative Commons Attribution 3.0 Unported License
Code: GNU General Public License (at least one of version 2 or version 3) or a GPL-compatible license.

ARTICLE TOOLS
Print this article
Indexing metadata
How to cite item
Supplementary files
Finding References
Email this article (Login required)
Email the author (Login required)

RELATED ITEMS
Show all

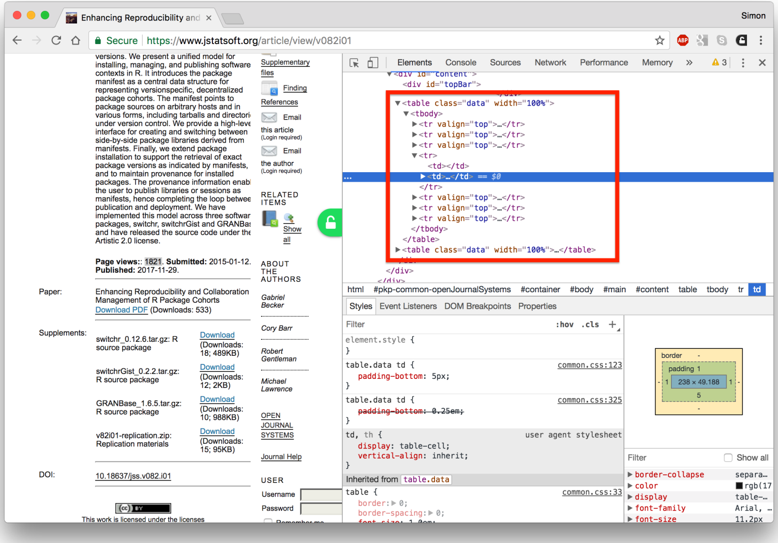
ABOUT THE AUTHORS
Gabriel Becker
Cory Barr
Robert Gentleman
Michael Lawrence

OPEN JOURNAL SYSTEMS
Journal Help

Step 1: Inspect the source

Procedure

- source: http://www.jstatsoft.org/
- go to "Archives"
- inspect the most recent volume
- inspect the first article
- inspect the page views element
- it's in a table!



versions. We present a unified model for installing, managing, and publishing software contexts in R. It introduces the package manifest as a central data structure for representing versionspecific, decentralized package cohorts. The manifest points to package sources on arbitrary hosts and in various forms, including tarballs and directori under version control. We provide a high-leve interface for creating and switching between side-by-side package libraries derived from manifests. Finally, we extend package installation to support the retrieval of exact package versions as indicated by manifests, and to maintain provenance for installed packages. The provenance information enabl the user to publish libraries or sessions as manifests, hence completing the loop betwee publication and deployment. We have implemented this model across three softwar packages, switchr, switchrGist and GRANBas and have released the source code under the Artistic 2.0 license.

Page views:: 1821. **Submitted:** 2015-01-12. **Published:** 2017-11-29.

Paper:	Enhancing Reproducibility and Collaboration Management of R Package Cohorts Download PDF (Downloads: 533)	
Supplements:	switchr_0.12.6.tar.gz: R source package	Download (Downloads: 18; 489KB)
	switchrGist_0.2.2.tar.gz: R source package	Download (Downloads: 12; 2KB)
	GRANBase_1.6.5.tar.gz: R source package	Download (Downloads: 10; 988KB)
	v82i01-replication.zip: Replication materials	Download (Downloads: 15; 95KB)
DOI:	10.18637/jss.v082.i01	

(cc) BY

This work is licensed under the licenses

Gabriel Becker

Cory Barr

Robert Gentleman

Michael Lawrence

```
▼<div id="content">
   <div id="topBar">
   </div>
  ▼<table class="data" width="100%">
    ▼<tbody>
      ▶<tr valign="top">…</tr>
      ▶<tr valign="top">…</tr>
      ▶<tr valign="top">…</tr>
      ▼<tr>
          <td></td>
...     ▶<td>…</td> == $0
        </tr>
      ▶<tr valign="top">…</tr>
      ▶<tr valign="top">…</tr>
      ▶<tr valign="top">…</tr>
      </tbody>
    </table>
  ▶<table class="data" width="100%">…</table>
  </div>
 </div>
</div>
```

html #pkp-common-openJournalSystems #container #body #main #content table tbody tr td

```
element.style {
}
table.data td {                              common.css:123
   padding-bottom: 5px;
}
table.data td {                              common.css:325
   ~~padding-bottom: 0.25em;~~
}
td, th {                         user agent stylesheet
   display: table-cell;
   vertical-align: inherit;
}
Inherited from table.data
table {                                       common.css:33
   border:▶ 0;
   border-spacing:▶ 0;
   font-size: 1.0em;
```

```
border          -
padding         1
-  1  238 × 49.188  1  -
                5
                -
```

```
▶ border-collapse     separa…
▶ color               ■ rgb(17
▶ display             table-…
▶ font-family         Arial, …
▶ font-size           11.2px
```

Step 2: Develop a scraping strategy

Observations

- getting the information out of the table will be straightforward
- this applies to all articles (check other articles on a sample basis)
- what we need is the set of **URLs leading to all articles**

Step 2: Develop a scraping strategy

Observations

- getting the information out of the table will be straightforward
- this applies to all articles (check other articles on a sample basis)
- what we need is the set of **URLs leading to all articles**

Inspecting the URLs

- the URL of the selected article looks as follows: https://www.jstatsoft.org/article/view/v082i01
- we find out that the final part, `v082i01`, always follows the same pattern: `v<volume number>i<issue number>`

Step 2: Develop a scraping strategy

Let’s try to construct the list of URLs from scratch

R code

```r
baseurl <- "http://www.jstatsoft.org/article/view/v"
volurl <- paste0("0", seq(1, 78, 1))
volurl[1:9] <- paste0("00", seq(1, 9, 1))
brurl <- paste0("0", seq(1, 9, 1))
urls_list <- paste0(baseurl, volurl)
urls_list <- paste0(rep(urls_list, each = 9), "i", brurl)
urls_list[1:5]
[1] "http://www.jstatsoft.org/article/view/v001i01"
[2] "http://www.jstatsoft.org/article/view/v001i02"
[3] "http://www.jstatsoft.org/article/view/v001i03"
[4] "http://www.jstatsoft.org/article/view/v001i04"
[5] "http://www.jstatsoft.org/article/view/v001i05"
names <- paste0(rep(volurl, each = 9), "_", brurl, ".html")
names[1:5]
[1] "001_01.html" "001_02.html" "001_03.html" "001_04.html" "001_05.html"
```

end

Step 3: Download the files

Set working directory

R code

```r
tempwd <- ("data/jstatsoftStats")
dir.create(tempwd)
setwd(tempwd)
```

end

Download pages

R code

```r
folder <- "html_articles/"
dir.create(folder)
for (i in 1:length(urls_list)) {
    if (!file.exists(paste0(folder, names[i]))) {
        download.file(urls_list[i], destfile = paste0(folder, names[i]))
        Sys.sleep(runif(1, 0, 1))
    }
}
```

end

Step 3: Download the files

Check success

R code

```r
list_files <- list.files(folder, pattern = "0.*")
list_files_path <- list.files(folder, pattern = "0.*", full.names = TRUE)
length(list_files)
```

```
[1] 666
```

end

Step 4: Import files and parse out information

Build loop

R code

```r
authors <- character()
title <- character()
statistics <- character()
numViews <- numeric()
datePublish <- character()
for (i in 1:length(list_files_path)) {
    html_out <- read_html(list_files_path[i])
    table_out <- html_table(html_out, fill = TRUE)[[6]]
    authors[i] <- table_out[1, 2]
    title[i] <- table_out[2, 2]
    statistics[i] <- table_out[4, 2]
    numViews[i] <- statistics[i] %>% str_extract("[[:digit:]]+") %>% as.numeric()
    datePublish[i] <- statistics[i] %>% str_extract("[[:digit:]]{4}-[[:digit:]]{2}-[[:digit:]]{2}.$") %>%
        str_replace("\\.", "")
}
```

end

Step 4: Import files and parse out information

Inspect parsed data

R code

```
authors[1:3]
```

```
[1] "Ronald Barry"   "Jason Bond, George Michailides"   "Thomas Lumley"
```

```
title[1:2]
```

```
[1] "A Diagnostic to Assess the Fit of a Variogram Model to Spatial Data"
[2] "Homogeneity Analysis in Xlisp-Stat"
```

```
numViews[1:3]
```

```
[1] 5835 3939 4379
```

end

Construct data frame

R code

```
dat <- data.frame(authors = authors, title = title, numViews = numViews, datePublish =
datePublish)
dim(dat)
```

```
[1] 666   4
```

end

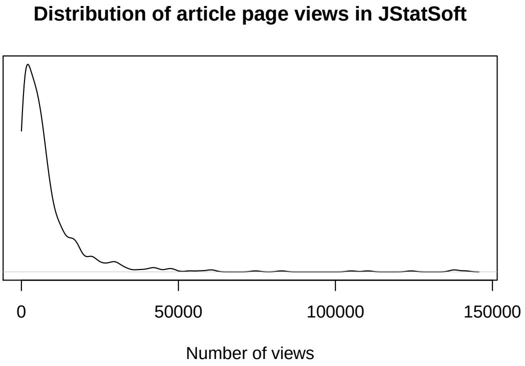
Step 5: Visualize data

Density plot of download statistics

R code

```r
plot(density(dat$numViews, from = 0), yaxt="n", ylab="", xlab="Number of views", main="
Distribution of article page views in JStatSoft")
```

end



Summary

Summary

- scraping data from multiple pages is no problem in R
- most of the brain work often goes into developing a scraping strategy and tidying the data, not into the actual downloading/scraping part
- scraping is also possible in even more complex scenarios, e.g., when HTML forms are involved or you have to take care of cookies or authentication
- this is beyond the scope of this course → check out the book for more applications

Source: Horia Varlan