Errata for 'Automated Data Collection with R'

Last update: 2015-10-08 10:36:19

```
library(stringr)
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.2.2
```

```
library(XML)
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 3.2.2
```

page 2

Credit: Suryapratim Sarkar (2015-06-25)

Wikipedia changed its server communication from HTTP to HTTPS. As a result, the following lines on page 2 return an error:

```
heritage_parsed <- htmlParse("http://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger",
                             encoding = "UTF-8")
```

```
## Error: failed to load external entity "http://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger
```

There are at least two solutions to the problem:

1. Use `getURL()` and specify the location of CA signatures (see Section 9.1.7 of our book).
2. Use Hadley Wickham's `rvest` package, which came out after our book was published. It facilitates scraping with R considerably, in particular in such scenarios. In this specific example, use the following code instead:

```
library(rvest) # the new package, version 0.3.0
heritage_parsed <- read_html("http://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger", encoding =
tables <- html_table(heritage_parsed, fill = TRUE) # html_table() from the revest package, which replac
```

From thereon, the rest of the chapter code should work. If you want to learn more about the rvest package, have a look here. We are planning to cover it extensively in the next edition of our book.

page 35

Credit: Jüri Kuusik (2015-01-22)

typo: change *"parsed_doc"* to *"parsed_fortunes$children"*

page 136

Credit: Laurent Franckx (2015-02-18)

Due to (supposedly) a bug in the RCurl package (version 1.95-4.5, bug has been reported) the following lines on page 136 give an error:

```
handle <- getCurlHandle(customrequest = "HEAD")
res <- getURL(url = url, curl = handle, header = TRUE)
```

There are two workaronds at the moment.

(1) To make the code simply run through you might give up the HTTP HEAD method used in the code above and use `customrequest = "GET"` instead:

```
require(RCurl)
require(stringr)

url <- "http://www.r-datacollection.com/materials/http/helloworld.html"
res <- getURL(url = url, header = TRUE)
cat(str_split(res, "\r")[[1]])

handle <- getCurlHandle(customrequest = "GET")
res <- getURL(url = url, curl = handle, header = TRUE)
```

(2) If you want to have a working example involving HTTP HEAD method you might switch to the httr package like this:

```
require(httr)

url <- "http://www.r-datacollection.com/materials/http/helloworld.html"
res <- HEAD(url)

res
```

```
## Response [http://www.r-datacollection.com/materials/http/helloworld.html]
##   Date: 2015-10-08 08:36
##   Status: 200
##   Content-Type: text/html
## <EMPTY BODY>
```

```
res$request
```

```
## <request>
## HEAD http://www.r-datacollection.com/materials/http/helloworld.html
## Output: write_memory
## Options:
## * useragent: libcurl/7.39.0 r-curl/0.9 httr/1.0.0
## * cainfo: C:/Users/Simon Munzert/Documents/R/win-library/3.2/httr/cacert.pem
## * nobody: TRUE
## * customrequest: HEAD
## Headers:
## * Accept: application/json, text/xml, application/xml, */*
```

```
res$headers[1:3]
```

```
## $date
## [1] "Thu, 08 Oct 2015 08:36:43 GMT"
##
## $server
## [1] "Apache"
##
## $vary
## [1] "Accept-Encoding"
```

page 194

Reported by: Laurent Franckx (2015-05-11)

The URL on page 194 to the parlgov SQLite database has changed and does not work anymore. The new URL is:

http://www.parlgov.org/static/stable/2014/parlgov-stable.db

page 249

Reported by: Laurent Franckx (2015-06-08)

The page structure had changed and code did not work anymore.

```
# define urls
search_url <- "www.biblio.com/search.php?keyisbn=data"
cart_url   <- "www.biblio.com/cart.php"

# download and parse page
search_page <- htmlParse(getURL(url = search_url, curl = handle))

# identify form fields
xpathApply(search_page, "//div[@class='row-fixed'][position()<2]/form")
```

```
## [[1]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[2]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[3]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[4]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[5]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
```

```
## [[6]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[7]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[8]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[9]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[10]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[11]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[12]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[13]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[14]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[15]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[16]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[17]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[18]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[19]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## [[20]]
## <form class="ob-add-form" action="http://www.biblio.com/cart.php" method="get"><input type="hidden" n
##
## attr(,"class")
## [1] "XMLNodeSet"
```

```
# extract book ids
xpath <- "//div[@class='row-fixed'][position()<4]/form/input[@name='bid']/@value"
bids <- unlist(xpathApply(search_page, xpath, as.numeric))
bids
```

```
##  [1] 755785471 745162375 433842671 724847744 745443134 724420313 232036641
```

```
##  [8] 724419579 755662729 231125829 564631291 756720938 244413814 405284703
## [15] 249002914 832933198 231125839 755663238 815848891 815848914
```

```
# add items to shopping cart
for(i in seq_along(bids))  {
    res <- getForm(uri = cart_url,
                   curl = handle,
                   bid = bids[i],
                   add = 1,
                   int = "keyword_search")
}

# inspect shopping cart
cart  <- htmlParse(getURL(url=cart_url, curl=handle))
clean <- function(x)  str_replace_all(xmlValue(x),"(\t)|(\n)"," ")
xpathSApply(cart, "//h3/a", clean)
```

```
##  [1] "Introduction to Statistics and Data Analysis (4th Hardcover Edition) by Jay L. Devore, Roxy Pe
##  [2] "Doing Data Analysis with SPSS: Version 18.0 (5th US Edition) by Robert Carver and Jane Gradwoh
##  [3] "Nonparametric Tests for Complete Data by Vilijandas Bagdonavi?us and Julius Kruopis"
##  [4] "Mining Graph Data by Diane J. Cook and Lawrence B. Holder"
##  [5] "SPSS survival manual: a step by step guide to data analysis using IBM SPSS (5th US Edition) by
##  [6] "Anabaptist Families from Langnau, Switzerland, 1749-1875 by Peden, Monty C"
##  [7] "Both Sides of the Ocean, Amish-Mennonites from Switzerland to America by Miller, J Virgil"
##  [8] "Agricultural and Federal Census Schedules, 1850-1880, Brecknock Twp by Frey, James E"
##  [9] "The Pennsylvania-Kentucky Rifle by Kauffman, Henry J"
## [10] "Christ is Our Cornerstone: 100 Years at Lititz Mennonite Church by Lapp, Alice Weber"
## [11] "Genealogical Data Relating to the German Settlers of Pennsylvania and Adjacent Territory by Ho
## [12] "Source-Data Collection Form: QuickSheet: Your Stripped-Bare Guide to Citing Sources by Mills, 
## [13] "Data Structures and Algorithm Analysis in Java (3rd Edition) by Mark Allen Weiss"
## [14] "Multivariate Data Analysis (7th Edition) by Hair Jr, Joseph F.; Black, William C.; Babin, Barr
## [15] "Data Structures and Algorithms in C++ by DROZDEK ADAM"
## [16] "Data Structures and Other Objects Using Java (4th Edition) by Main, Michael"
## [17] "Skillful Inquiry/Data Team by Nancy Love"
## [18] "Data Literacy for Teachers by Nancy Love"
## [19] "Tales from Shakespeare (Illustrated by Norman M. Price) by Lamb, Charles; Lamb, Mary"
## [20] "Data Management: Database and Organizations by Watson, Richard T.; Bostrom, Robert P"
```

```
# request header
cat(str_split(info$value()["headerOut"],"\r")[[1]][1:13])
```

```
## GET /search.php?keyisbn=data HTTP/1.1
## Host: www.biblio.com
## Accept: */*
## from: eddie@r-datacollection.com
## user-agent: R version 3.2.1 (2015-06-18), x86_64-w64-mingw32
##
## GET /cart.php?bid=755785471&add=1&int=keyword_search HTTP/1.1
## Host: www.biblio.com
## Accept: */*
## Cookie: variation=res_a; vis=language%3Ade%7Ccountry%3A6%7Ccurrency%3A9%7Cvisitor%3A5cb7Uqna0GrYIB5T
## from: eddie@r-datacollection.com
## user-agent: R version 3.2.1 (2015-06-18), x86_64-w64-mingw32
```

```
# response header
cat(str_split(info$value()["headerIn"],"\r")[[1]][1:14])
```

```
## HTTP/1.1 200 OK
## Server: nginx
## Date: Thu, 08 Oct 2015 08:36:43 GMT
## Content-Type: text/html; charset=UTF-8
## Content-Length: 108349
## Connection: keep-alive
## Keep-Alive: timeout=60
## Set-Cookie: vis=language%3Ade%7Ccountry%3A6%7Ccurrency%3A9%7Cvisitor%3A5cb7UqnaOGrYIB5TzftI0HFP3ORrI
## Set-Cookie: variation=res_a; expires=Fri, 09-Oct-2015 08:36:42 GMT; path=/; domain=.biblio.com; http
## X-Mod-Pagespeed: 1.9.32.3-4448
## Access-Control-Allow-Credentials: true
## Vary: User-Agent,Accept-Encoding
## Expires: Fri, 09 Oct 2015 08:36:43 GMT
## Cache-Control: max-age=86400
```

page 254

Reported by: Laurent Franckx (2015-06-10)

There has been a change to the install_github function of the devtools package. To install Rwebdriver use:

```
library(devtools)

install_github("crubba/Rwebdriver")
```

page 299–310

The website holding the UK government press releases has been altered slightly. To get the date and organisation you need to change the XPaths here...

```
library(XML)

organisation <- xpathSApply(tmp, "//dl[@data-trackposition='top']//a[@class='organisation-link']", xmlV
publication <- xpathSApply(tmp, "//dl[@class='primary-metadata']//abbr[@class='date']", xmlValue)
```

... and here...

```
for(i in 2:length(list.files("Press_Releases/"))){
    tmp <- readLines(str_c("Press_Releases/", i, ".html"))
    tmp <- str_c(tmp, collapse = "")
    tmp <- htmlParse(tmp)
    release <- xpathSApply(tmp, "//div[@class='block-4']", xmlValue)
    organisation <- xpathSApply(tmp, "//dl[@data-trackposition='top']//a[@class='organisation-link']",
    publication <- xpathSApply(tmp, "//dl[@class='primary-metadata']//abbr[@class='date']", xmlValue)
    if(length(release) != 0){
        n <- n + 1
```

```
        tmp_corpus <- Corpus(VectorSource(release))
        release_corpus <- c(release_corpus, tmp_corpus)
        meta(release_corpus[[n]], "organisation") <- organisation[1]
        meta(release_corpus[[n]], "publication") <- publication
    }
}
```

The prescindMeta() function is defunct as of version 0.6 of the tm package. The meta data can now be gathered with the meta() function.

```
meta_organisation <- meta(release_corpus, type = "local", tag = "organisation")
meta_publication <- meta(release_corpus, type = "local", tag = "publication")

meta_data <- data.frame(
    organisation = unlist(meta_organisation),
    publication = unlist(meta_publication)
)
```

The sFilter() function is also defunct. You can filter the corpus using meta().

```
release_corpus <- release_corpus[
    meta(release_corpus, tag = "organisation") == "Department for Business, Innovation & Skills" |
    meta(release_corpus, tag = "organisation") == "Department for Communities and Local Government" |
    meta(release_corpus, tag = "organisation") == "Department for Environment, Food & Rural Affairs" |
    meta(release_corpus, tag = "organisation") == "Foreign & Commonwealth Office" |
    meta(release_corpus, tag = "organisation") == "Ministry of Defence" |
    meta(release_corpus, tag = "organisation") == "Wales Office"
]
```

The *stringr* package also produces a hick-up with the updated version of the tm package, thus we switch to base R.

```
tm_filter(release_corpus, FUN = function(x) any(grep("Afghanistan", content(x))))
```

We need to wrap the replace function with the new content_transformer()...

```
release_corpus <- tm_map(
    release_corpus,
    content_transformer(
        function(x, pattern){
            gsub(
                pattern = "[[:punct:]]",
                replacement = " ",
                x
            )
        )
    )
)
```

Moreover, the tolower() function needs to be wrapped with the content_transformer()...

```
release_corpus <- tm_map(release_corpus, content_transformer(tolower))
```

The prescindMeta() function is also defunct on page 310...

```
org_labels <- unlist(meta(release_corpus, "organisation"))
```

page 315

Since the sFilter() and prescindMeta() functions are defunct as of version 0.6 of the tm package, you need to change the code on page 315 to filter the corpus.

```
short_corpus <- release_corpus[c(
    which(
        meta(
            release_corpus, tag = "organisation"
        ) == "Department for Business, Innovation & Skills"
    )[1:20],
    which(
        meta(
            release_corpus, tag = "organisation"
        ) == "Wales Office"
    )[1:20],
    which(
        meta(
            release_corpus, tag = "organisation"
        ) == "Department for Environment, Food & Rural Affairs"
    )[1:20]
)]

table(unlist(meta(short_corpus, "organisation")))
```