# Errata for 'Automated Data Collection with R'

Last update: 2015-05-11 22:37:28

### 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: 2015-05-11 22:37
##   Status: 200
##   Content-Type: text/html
## <EMPTY BODY>
```

```
res$request
```

```
## $handle
## Host: http://www.r-datacollection.com/ <0x00000000085647d0>
##
## $writer
## <write_memory>
##
## $method
## [1] "HEAD"
##
## $opts
## Config:
## List of 9
##  $ followlocation:TRUE
##  $ maxredirs     :10
##  $ encoding      :"gzip"
##  $ useragent     :"curl/7.39.0 Rcurl/1.95.4.5 httr/0.6.1"
##  $ cainfo        :"C:/Users/Peter/Documents/R/win-library/3.1/httr/cacert.pem"
##  $ httpheader    :"application/json, text/xml, application/xml, */*"
##   ..- attr(*, "names")="Accept"
##  $ nobody        :TRUE
##  $ customrequest :"HEAD"
##  $ url           :"http://www.r-datacollection.com/materials/http/helloworld.html"
##
## $body
## NULL
```

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

```
## $date
## [1] "Mon, 11 May 2015 20:37:29 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 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']", xmlVa
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']", x
    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")))
```