MY-DATA607-Week07-Books-XML-JSON

Michael Y.

October 13, 2019

Contents

Week 7 - Books - Working with XML and JSON in R

Load libraries

```
library(XML)
library(rlist)
library(knitr)
library(kableExtra)
library(jsonlite)
library(RCurl)
library(tidyr)
library(dplyr)
```

Pick three of your favorite books on one of your favorite subjects. At least one of the books should have more than one author. For each book, include the title, authors, and two or three other attributes that you find interesting.

Take the information that you've selected about these three books, and separately create three files which store the book's information in HTML (using an html table), XML, and JSON formats (e.g. "books.html", "books.xml", and "books.json"). To help you better understand the different file structures, I'd prefer that you create each of these files "by hand" unless you're already very comfortable with the file formats.

Write R code, using your packages of choice, to load the information from each of the three sources into separate R data frames.

Your deliverable is the three source files and the R code. If you can, package your assignment solution up into an .Rmd file and publish to rpubs.com. [This will also require finding a way to make your three text files accessible from the web].

Part 1 - HTML

Load the data file in html format

```
# if loading from local file
# my_books_html_file <- 'ThreeBooks.html'
# my_books_html_list <- readHTMLTable(my_books_html_file, stringsAsFactors=F)
```

if loading from github

```
my_books_html_file <- 'https://raw.githubusercontent.com/myampol/MY607/master/ThreeBooks.html'
my_books_html_doc <- getURL(my_books_html_file)

cat(my_books_html_doc)
```

```
## <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
## <html lang="en">
## <head>
## <title>Books - Week07</title>
## </head>
##
## <body>
## <table id="qstable" class="sortable" border="1">
## <thead>
## <tr>
## <th width="20%">Author</th>
## <th width="30%">Title</th>
## <th width="5%">Year</th>
## <th width="15%">Publisher</th>
## <th width="15%">ISBN</th>
## <th width="15%">URL</th>
## </tr>
## </thead>
##
## <tbody>
```

```
##
## <tr id="ProvostFawcett13" class="entry">
##  <td>Provost, Foster &amp;amp; Fawcett, Tom</td>
##  <td>Data Science for Business</td>
##  <td>2013</td>
##  <td>O'Reilly</td>
##  <td>978-1-449-36132-7</td>
##  <td><a href="http://data-science-for-biz.com">http://data-science-for-biz.com/</a></td>
## </tr>
##
## <tr id="Wickham2009Ggplot2" class="entry">
##  <td>Wickham, Hadley</td>
##  <td>ggplot2: Elegant Graphics for Data Analysis</td>
##  <td>2009</td>
##  <td>Springer</td>
##  <td>978-0-387-98140-6</td>
##  <td><a href="http://ggplot2.org/book/">http://ggplot2.org/book/</a></td>
## </tr>
##
## <tr id="Wickham2017R" class="entry">
##  <td>Wickham, Hadley &amp;amp; Grolemund, Garrett</td>
##  <td>R for Data Science: Import, Tidy, Transform, Visualize, and Model Data</td>
##  <td>2017</td>
##  <td>O'Reilly</td>
##  <td>978-1-491-91039-9</td>
##  <td><a href="http://r4ds.had.co.nz/">http://r4ds.had.co.nz/</a></td>
## </tr>
##
##
## </tbody>
## </table>
##
## </body>
## </html>
```

Read the HTML table

```
### the data is read in as a list containing a single element (i.e., the dataframe)
my_books_html_list <- readHTMLTable(my_books_html_doc, stringsAsFactors=F)
str(my_books_html_list)
```

```
## List of 1
##  $ qstable:'data.frame': 3 obs. of  6 variables:
##   ..$ Author   : chr [1:3] "Provost, Foster & Fawcett, Tom" "Wickham, Hadley" "Wickham, Hadley & Grolemund, Garrett"
##   ..$ Title    : chr [1:3] "Data Science for Business" "ggplot2: Elegant Graphics for Data Analysis" "R for Data Science: Import, Tidy,
##   ..$ Year     : chr [1:3] "2013" "2009" "2017"
##   ..$ Publisher: chr [1:3] "O'Reilly" "Springer" "O'Reilly"
##   ..$ ISBN     : chr [1:3] "978-1-449-36132-7" "978-0-387-98140-6" "978-1-491-91039-9"
##   ..$ URL      : chr [1:3] "http://data-science-for-biz.com/" "http://ggplot2.org/book/" "http://r4ds.had.co.nz/"
```

extract the dataframe from the list

```
my_books_html_df <- my_books_html_list[[1]]
my_books_html_df
```

```
##                                Author
## 1       Provost, Foster & Fawcett, Tom
## 2                      Wickham, Hadley
## 3 Wickham, Hadley & Grolemund, Garrett
##                                                                    Title
## 1                                                Data Science for Business
## 2                              ggplot2: Elegant Graphics for Data Analysis
## 3 R for Data Science: Import, Tidy, Transform, Visualize, and Model Data
##   Year Publisher              ISBN                              URL
## 1 2013  O'Reilly 978-1-449-36132-7 http://data-science-for-biz.com/
## 2 2009  Springer 978-0-387-98140-6         http://ggplot2.org/book/
## 3 2017  O'Reilly 978-1-491-91039-9           http://r4ds.had.co.nz/
```

Improve the display, using kable

```
my_books_html_df %>%
  kable() %>%
  column_spec(.,column = c(1,2), width = "10em") %>% kable_styling(c("striped", "bordered"))
```

Author	Title	Year	Publisher	ISBN	URL
Provost, Foster & Fawcett, Tom	Data Science for Business	2013	O'Reilly	978-1-449-36132-7	http://data-science-for-biz.com/
Wickham, Hadley	ggplot2: Elegant Graphics for Data Analysis	2009	Springer	978-0-387-98140-6	http://ggplot2.org/book/
Wickham, Hadley & Grolemund, Garrett	R for Data Science: Import, Tidy, Transform, Visualize, and Model Data	2017	O'Reilly	978-1-491-91039-9	http://r4ds.had.co.nz/

Separate out the multiple authors into individual columns

```
my_books_html_df <- separate(data = my_books_html_df,
                             col = Author,
                             into =c("Author1","Author2"),
                             fill="right",
                             sep = " & ",
                             remove = T)

my_books_html_df %>%
  kable() %>% column_spec(.,column = 3, width = "10em") %>% kable_styling(c("striped", "bordered"))
```

Author1	Author2	Title	Year	Publisher	ISBN	URL
Provost, Foster	Fawcett, Tom	Data Science for Business	2013	O'Reilly	978-1-449-36132-7	http://data-science-for-biz.com/
Wickham, Hadley	NA	ggplot2: Elegant Graphics for Data Analysis	2009	Springer	978-0-387-98140-6	http://ggplot2.org/book/
Wickham, Hadley	Grolemund, Garrett	R for Data Science: Import, Tidy, Transform, Visualize, and Model Data	2017	O'Reilly	978-1-491-91039-9	http://r4ds.had.co.nz/

Part 2 - XML

Select the file to load

```
# if loading from local file
# my_books_xml_file <- 'ThreeBooks.xml'
# my_books_xml <- xmlParse(my_books_xml_file)
# my_books_xml
```

If loading from github

```
my_books_xml <- getURL('https://raw.githubusercontent.com/myampol/MY607/master/ThreeBooks.xml')
cat(my_books_xml)
```

```
## <?xml version="1.0" encoding="UTF-8"?>
## <MY_Booklist>
##  <Book id='ProvostFawcett13'>
##      <Author1>Provost, Foster</Author1>
##      <Author2>Fawcett, Tom</Author2>
##      <Title>Data Science for Business</Title>
##      <Year>2013</Year>
##      <Publisher>O'Reilly</Publisher>
##      <ISBN>978-1-449-36132-7</ISBN>
##      <URL>http://data-science-for-biz.com/</URL>
##  </Book>
##  <Book id='Wickham2009Ggplot2'>
##      <Author1>Wickham, Hadley</Author1>
##      <Title>ggplot2: Elegant Graphics for Data Analysis</Title>
##      <Year>2009</Year>
##      <Publisher>Springer</Publisher>
##      <ISBN>978-0-387-98140-6</ISBN>
##      <URL>http://ggplot2.org/book/</URL>
##  </Book>
##  <Book id='Wickham2017R'>
##      <Author1>Wickham, Hadley</Author1>
```

```
##       <Author2>Grolemund, Garrett</Author2>
##       <Title>R for Data Science: Import, Tidy, Transform, Visualize, and Model Data</Title>
##       <Year>2017</Year>
##       <Publisher>O'Reilly</Publisher>
##       <ISBN>978-1-491-91039-9</ISBN>
##       <URL>http://r4ds.had.co.nz/</URL>
##   </Book>
## </MY_Booklist>
```

Convert to Data Frame

```
my_books_xml_df <- xmlToDataFrame(my_books_xml, stringsAsFactors = F)

my_books_xml_df %>%
  kable() %>% column_spec(.,column = 3, width = "10em") %>% kable_styling(c("striped", "bordered"))
```

Author1	Author2	Title	Year	Publisher	ISBN	URL
Provost, Foster	Fawcett, Tom	Data Science for Business	2013	O'Reilly	978-1-449-36132-7	http://data-science-for-biz.com/
Wickham, Hadley	NA	ggplot2: Elegant Graphics for Data Analysis	2009	Springer	978-0-387-98140-6	http://ggplot2.org/book/
Wickham, Hadley	Grolemund, Garrett	R for Data Science: Import, Tidy, Transform, Visualize, and Model Data	2017	O'Reilly	978-1-491-91039-9	http://r4ds.had.co.nz/

Are the above two data frames (from HTML and from XML) identical?

```
identical(my_books_html_df, my_books_xml_df )
```

```
## [1] TRUE
```

Part 3 - JSON

Select the file containing the books in JSON

```
# if loading from local file
# my_books_json_file <- 'ThreeBooks.json'
```

if loading from github

```
my_books_json_file <- 'https://raw.githubusercontent.com/myampol/MY607/master/ThreeBooks.json'
my_books_json_doc <- getURL(my_books_json_file)
cat(my_books_json_doc)
```

```
## [
##     {
##         "Author1": "Provost, Foster",
##         "Author2": "Fawcett, Tom",
##         "Title": "Data Science for Business",
##         "Year": "2013",
##         "Publisher": "O'Reilly",
##         "ISBN": "978-1-449-36132-7",
##         "URL": "http://data-science-for-biz.com/"
##     },
##     {
##         "Author1": "Wickham, Hadley",
##         "Title": "ggplot2: Elegant Graphics for Data Analysis",
##         "Year": "2009",
##         "Publisher": "Springer",
##         "ISBN": "978-0-387-98140-6",
##         "URL": "http://ggplot2.org/book/"
##     },
##     {
##         "Author1": "Wickham, Hadley",
##         "Author2": "Grolemund, Garrett",
##         "Title": "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data",
```

```
##         "Year": "2017",
##         "Publisher": "O'Reilly",
##         "ISBN": "978-1-491-91039-9",
##         "URL": "http://r4ds.had.co.nz/"
##     }
## ]
##
```

load the JSON file using jsonlite

```
### This automatically loads it into a data frame, unless simplifyDataFrame=FALSE has been specified

my_books_json_df <- fromJSON(txt = my_books_json_file )
my_books_json_df %>%
  kable() %>% column_spec(.,column = 3, width = "10em") %>% kable_styling(c("striped", "bordered"))
```

Author1	Author2	Title	Year	Publisher	ISBN	URL
Provost, Foster	Fawcett, Tom	Data Science for Business	2013	O'Reilly	978-1-449-36132-7	http://data-science-for-biz.com/
Wickham, Hadley	NA	ggplot2: Elegant Graphics for Data Analysis	2009	Springer	978-0-387-98140-6	http://ggplot2.org/book/
Wickham, Hadley	Grolemund, Garrett	R for Data Science: Import, Tidy, Transform, Visualize, and Model Data	2017	O'Reilly	978-1-491-91039-9	http://r4ds.had.co.nz/

Conclusion

Are the three data frames identical?

```
## HTML data frame vs. JSON data frame
identical(my_books_html_df, my_books_json_df)
```

```
## [1] TRUE
```

```
## XML data frame vs. JSON data frame
identical(my_books_xml_df, my_books_json_df)
```

```
## [1] TRUE
```

All three data frames are identical.