DATA607_w7 - Working with XML and JSON in R

David Simbandumwe

2021-10-10

Introduction

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. Are the three data frames identical?

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].

HTML

```
# setup files
html_url <- "https://raw.githubusercontent.com/dsimband/DATA607/main/Homework/w7_files/books.html"
html_file <- tempfile()


# download file to temp directory
curl_download(url = html_url,
              destfile = html_file
              )


# read file as html
html_df <- readHTMLTable(html_file,
                         header=TRUE,
                          which = 1,
                          as.data.frame = TRUE,
                          colClasses = c("character",
                                         "character",
                                         "character",
                                         "numeric",
```

```
                                        "numeric",
                                        "numeric")
                          )

html_df
```

```
##                         title                           sub_title
## 1    This Time is Different Eight Centuries of Financial Filly
## 2                   Capital        in the Twenty_First Century
## 3 Market Neutral Strategies
##                                  author year_published length  cost
## 1 Carmen M. Reinhart & Kenneth S Rogoff           2009    463 35.00
## 2                        Thomas Piketty           2014    685 39.95
## 3     Bruce I. Jacobs & Kenneth N. Levy           2005    284 80.00
```

XML

```
# setup files
xml_url <- "https://raw.githubusercontent.com/dsimband/DATA607/main/Homework/w7_files/books.xml"
xml_file <- tempfile()


# download file to temp directory
curl_download(url = xml_url,
              destfile = xml_file
              )
# load file as xml
xml_df <- xmlToDataFrame(xml_file) %>%
        transform(year_published = as.numeric(year_published),
              length = as.numeric(length),
              price = as.numeric(price)
              )

xml_df
```

```
##                         title                           sub_title
## 1    This Time is Different Eight Centuries of Financial Folly
## 2                   Capital        in the Twenty_First Century
## 3 Market Neutral Strategies
##                                  author year_published length price
## 1 Carmen M. Reinhart & Kenneth S Rogoff           2009    463 35.00
## 2                        Thomas Piketty           2014    685 39.95
## 3     Bruce I. Jacobs & Kenneth N. Levy           2005    284 80.00
```

JSON

```r
json_url <- "https://raw.githubusercontent.com/dsimband/DATA607/main/Homework/w7_files/books.json"

json_df = fromJSON(json_url,
                   simplifyDataFrame = TRUE
                   )$books$book %>%
    transform(year_published = as.numeric(year_published),
              length = as.numeric(length),
              price = as.numeric(price)
              )

json_df
```

```
##                       title                           sub_title
## 1    This Time is Different Eight Centuries of Financial Folly
## 2                   Capital        in the Twenty_First Century
## 3 Market Neutral Strategies
##                                 author year_published length price
## 1 Carmen M. Reinhart & Kenneth S Rogoff           2009    463 35.00
## 2                        Thomas Piketty           2014    685 39.95
## 3     Bruce I. Jacobs & Kenneth N. Levy           2005    284 80.00
```

Conclusions

The data frames are identical with the exception of special characters such as & for the HTML and the XML versions of the files we had to use & to represent &. Working with JSON is definitely cleaner from a separation of data and presentation logic perspective. HTML and XML still contain formating data and tend to be verbose for big datasets. It is also easier because you can read the file contents directly from git hub without having to create a intermediate file