Data 607 Assignment Week 7

Leticia Salazar

2021-10-09

Assignment - Working with XML and JSON in R

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", "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?

```
#Install Packages
#install.packages("XML")
#install.packages("rjson")
```

```
#Import Libraries
library(tidyverse)
```

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.]

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RCurl)
```

```
##
## Attaching package: 'RCurl'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(XML)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
```

```
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(methods)
library(dplyr)
```

To create these files, I used Visual Studio Code which is a great tool for Macs, Windows, or Linux, etc. Upon creating a new file it gives you an option to select what type of file you want to create, in this case HTML, XML and JSON. Once the files were created, I imported them to my GitHub and the results are below:

HTML

```
#Read HTML table
HTML_File <- getURL("https://raw.githubusercontent.com/letisalba/Data607_Assignment_Week7/main/Books.ht

#Set file as data frame
books_html <- as.data.frame(readHTMLTable(HTML_File, stringAsFactors = FALSE))
books_html
```

```
##                 NULL.Title               NULL.Authors             NULL.Type
## 1             The Talisman Stephen King, Peter Straub             Paperback
## 2 And Then There Were None            Agatha Christie Mass Market Paperback
## 3              Night Shift               Stephen King Mass Market Paperback
##   NULL.Publisher   NULL.ISBN.13 NULL.Price
## 1  Gallery Books 978-1501192272      13.50
## 2 William Morrow 978-0062073488       7.99
## 3         Anchor 978-0307743640       8.99
```

XML

```
#Read XML File
XML_File <-  getURL("https://raw.githubusercontent.com/letisalba/Data607_Assignment_Week7/main/Books2.xm

#Parse File
books_xml <-  xmlParse(XML_File)

#Set file as data frame
books_xml2 <-  xmlToDataFrame(books_xml)
books_xml2
```

```
##                       Title                    Authors                Type
## 1             The Talisman Stephen King, Peter Straub             Paperback
## 2 And Then There Were None            Agatha Christie Mass Market Paperback
## 3              Night Shift               Stephen King Mass Market Paperback
##        Publisher        ISBN-13 Price
## 1  Gallery Books 978-1501192272 13.50
## 2 William Morrow 978-0062073488  7.99
## 3         Anchor 978-0307743640  8.99
```

JSON

```
#Load JSON file
JSON_File <- fromJSON("https://raw.githubusercontent.com/letisalba/Data607_Assignment_Week7/main/Books2

#Set file as data frame
books_json <- as.data.frame(JSON_File)
books_json
```

```
##                Books.Title               Books.Author            Books.Type
## 1             The Talisman Stephen King, Peter Straub             Paperback
## 2 And Then There Were None            Agatha Christie Mass Market Paperback
## 3              Night Shift               Stephen King Mass Market Paperback
##   Books.Publisher  Books.ISBN.13 Books.Price
## 1   Gallery Books 978-1501192272       13.50
## 2  William Morrow 978-0062073488        7.99
## 3          Anchor 978-0307743640        8.99
```

Are the three data frames identical?

At first glance it seems like the data frames are identical. The more I analyzed it I noticed:

For the HTML, XML, and JSON there are 6 columns and 3 rows, all named very similar. With some tidying, I'd rename the columns, drop the NULL in HTML column names and remove "Books." in JSON column names.

```
#Glimpse of HTML
glimpse(books_html)
```

For all except for "Price" in JSON, it's characterized as a character, where as "Price" in the JSON file is characterize as numeric even though all three were written the same way when creating each file.

```
## Rows: 3
## Columns: 6
## $ NULL.Title     <chr> "The Talisman", "And Then There Were None", "Night Shif~
## $ NULL.Authors   <chr> "Stephen King, Peter Straub", "Agatha Christie", "Steph~
## $ NULL.Type      <chr> "Paperback", "Mass Market Paperback", "Mass Market Pape~
## $ NULL.Publisher <chr> "Gallery Books", "William Morrow", "Anchor"
## $ NULL.ISBN.13   <chr> "978-1501192272", "978-0062073488", "978-0307743640"
## $ NULL.Price     <chr> "13.50", "7.99", "8.99"
```

```
#Glimpse of XML
glimpse(books_xml2)
```

```
## Rows: 3
## Columns: 6
## $ Title     <chr> "The Talisman", "And Then There Were None", "Night Shift"
## $ Authors   <chr> "Stephen King, Peter Straub", "Agatha Christie", "Stephen Ki~
## $ Type      <chr> "Paperback", "Mass Market Paperback", "Mass Market Paperback"
## $ Publisher <chr> "Gallery Books", "William Morrow", "Anchor"
## $ `ISBN-13` <chr> "978-1501192272", "978-0062073488", "978-0307743640"
## $ Price     <chr> "13.50", "7.99", "8.99"
```

```
#Glimpse of JSON
glimpse(books_json)
```

```
## Rows: 3
## Columns: 6
## $ Books.Title     <chr> "The Talisman", "And Then There Were None", "Night Shi~
## $ Books.Author    <chr> "Stephen King, Peter Straub", "Agatha Christie", "Step~
## $ Books.Type      <chr> "Paperback", "Mass Market Paperback", "Mass Market Pap~
## $ Books.Publisher <chr> "Gallery Books", "William Morrow", "Anchor"
## $ Books.ISBN.13   <chr> "978-1501192272", "978-0062073488", "978-0307743640"
## $ Books.Price     <dbl> 13.50, 7.99, 8.99
```