Assignment 8 - Working with XML and JSON in R

Jawaid Hakim

2022-10-13

Contents

1 Load libraries

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(jsonlite)
```

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

```
library(xml2)
library(XML)
library(htmltab)
```

2 Assignment

> 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]

3 Data

Each book entry in the data set contains 5 variables: title, author, medium, pages, and isbn. For example:

> "title": "Thinking Fast and Slow", "author": "Daniel Kahneman", "medium": "Paperback", "pages": 499, "isbn": "0374533555"

4 JSON

First load the JSON file. As expected, the resulting data frame contains 3 observatios of 5 variables. Note, the *pages* variable is of *int* type.

```
d <- jsonlite::fromJSON(txt = "https://raw.githubusercontent.com/himalayahall/DATA607/main/Assignment%2
df_json <- as.data.frame(d)
str(df_json)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ book.title : chr  "Thinking Fast and Slow" "Wolf Hall" "Man-Eaters of Kumaon"
##  $ book.author: chr  "Daniel Kahneman" "Hilary Mantel" "Jim Corbett"
##  $ book.medium: chr  "Paperback" "Paperback" "Paperback"
##  $ book.pages : int  499 640 184
##  $ book.isbn  : chr  "0374533555" "1250806712" "8129140365"
```

```
df_json
```

```
##               book.title     book.author book.medium book.pages  book.isbn
## 1 Thinking Fast and Slow Daniel Kahneman   Paperback        499 0374533555
## 2              Wolf Hall   Hilary Mantel   Paperback        640 1250806712
## 3   Man-Eaters of Kumaon     Jim Corbett   Paperback        184 8129140365
```

5 XML

Next, we load the same data in XML format. Note, both JSON and XML data frames are very similar. However, unlike the JSON version, the data type for *pages* is *chr* (string)

The correct way to fix the issue of type safety in XML documents is by through the use of **XML Schema** (DTD or XSD). For example, here is a basic XSD for this assignment.

```
book_data_xml <- xml2::read_xml("https://raw.githubusercontent.com/himalayahall/DATA607/main/Assignment
book_xml <- XML::xmlParse(book_data_xml)
df_xml <- xmlToDataFrame(nodes=getNodeSet(book_xml, "//book"))
str(df_xml)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ title : chr  "Thinking Fast and Slow" "Wolf Hall" "Man-Eaters of Kumaon"
##  $ author: chr  "Daniel Kahneman" "Hilary Mantel" "Jim Corbett"
##  $ medium: chr  "Paperback" "Paperback" "Paperback"
##  $ pages : chr  "499" "640" "184"
##  $ isbn  : chr  "0374533555" "1250806712" "8129140365"
```

```
df_xml
```

```
##                    title          author    medium pages       isbn
## 1 Thinking Fast and Slow Daniel Kahneman Paperback   499 0374533555
## 2              Wolf Hall   Hilary Mantel Paperback   640 1250806712
## 3   Man-Eaters of Kumaon     Jim Corbett Paperback   184 8129140365
```

6 HTML

Finally, the same data is loaded in HTML format. Here, the data frame is identical to XML.

```
df_html <- htmltab::htmltab(doc = "https://raw.githubusercontent.com/himalayahall/DATA607/main/Assignmen
```

```
## Argument 'which' was left unspecified. Choosing first table.
```

```
str(df_html)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ title : chr  "Thinking Fast and Slow" "Wolf Hall" "Man-Eaters of Kumaon"
##  $ author: chr  "Daniel Kahneman" "Hilary Mantel" "Jim Corbett"
##  $ medium: chr  "Paperback" "Paperback" "Paperback"
##  $ pages : chr  "499" "640" "184"
##  $ isbn  : chr  "0374533555" "1250806712" "8129140365"
```

```
df_html
```

```
##                    title          author    medium pages       isbn
## 2 Thinking Fast and Slow Daniel Kahneman Paperback   499 0374533555
## 3              Wolf Hall   Hilary Mantel Paperback   640 1250806712
## 4   Man-Eaters of Kumaon     Jim Corbett Paperback   184 8129140365
```