Week 7 Assignment

Heleine Fouda

2023-10-21

The assignment:

This assignment asks to prepare 3 separate files with the same data in order to practice loading the formats into R, convert them to data frames and compare the differences. The 3 formats are: HTML, JSON and XML. Each file contains data on three of my favorite books and includes the following attributes: the title, the author(s), the theme, the publication year and the number of editions. Each of the files was manually written and then loaded into Github for reference / reproducibility.

Getting Started : Load Libraries

Load the HTML file

```
url1 <-"https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/Books.html"
books_html <- getURL(url1)
# The HTML file loaded as class "character"

class(books_html)
```

```
## [1] "character"
```

Read the HTML file into R:

To read the HTML file into R, I'll need to extract books' information from the html file using regular expressions

```
# Splitting the HTML content into individual book sections
book_sections <- strsplit(books_html, "</ul>")

# Creating empty lists to store extracted data
book_titles <- list()
book_titles_text <- list()
book_info <- list()

# Extract book titles and information
for (section in book_sections) {
  book_title <- gsub("<h2>|</h2>", "", regmatches(section, regexec("<h2>.*?</h2>", section))[[1]])
  book_title_text <- unlist(strsplit(book_title, "\n"))
  book_info_section <- unlist(strsplit(section, "\n"))

  book_titles <- append(book_titles, list(book_title))
  book_titles_text <- append(book_titles_text, list(book_title_text))
  book_info <- append(book_info, list(book_info_section))
}
```

```
# Create a data frame
books_data_html <- data.frame(
  Title = unlist(sapply(book_info, function(x) gsub("<li><strong>Title:</strong> ", "", x[which(grepl("
  Author = unlist(sapply(book_info, function(x) gsub("<li><strong>Author:</strong> ", "", x[which(grepl
  Publisher = unlist(sapply(book_info, function(x) gsub("<li><strong>Publisher:</strong> ", "", x[which
  Publication = unlist(sapply(book_info, function(x) gsub("<li><strong>Publication:</strong> ", "", x[wh
  Theme = unlist(sapply(book_info, function(x) gsub("<li><strong>Theme:</strong> ", "", x[which(grepl("
)

# Print the data frame
books_data_html
```

```
##                                                                              Title
## 1                                                            Farewell to Arms</li>
## 2                                                               War and Peace</li>
## 3         International Law: Laws, Norms, Actors: A Problem-Oriented Approach</li>
##                                                 Author
## 1                                Ernest Hemingway</li>
## 2                                     Leo Tolstoy</li>
## 3         Jeffrey L. Dunoff and Steven R. Ratner</li>
##                                Publisher       Publication
## 1         Charles Scribner\\'s Sons</li>         1929</li>
## 2           Various (Public Domain)</li>         1869</li>
## 3                    Wolters Kluwer</li>         2020</li>
##                            Theme
## 1              War and Love</li>
## 2           Historical Epic</li>
## 3         International Law</li>
```

Load the JSON file

```
library(rjson)
url2 <- "https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/books.json"
books_json <- fromJSON(file = url2)
class(books_json)
```

```
## [1] "list"
```

```
library(tidyverse)

 # Converting the list into a data frame

books_data_json_df <- data.frame(do.call("rbind", books_json))

# Unnesting the data and convert them  back into a data frame

books_data_json_df <- unnest(books_data_json_df, cols = c(Title, Author, Publisher, Publication, Theme)
class(books_data_json_df)
```

```
## [1] "data.frame"
```

```
books_data_json_df
```

```
##                                                                  Title
## 1                                                     Farewell to Arms
## 2                                                        War and Peace
## 3 International Law: Laws, Norms, Actors: A Problem-Oriented Approach
##                                    Author               Publisher Publication
## 1                       Ernest Hemingway Charles Scribner's Sons        1929
## 2                            Leo Tolstoy Various (Public Domain)        1869
## 3 Jeffrey L. Dunoff and Steven R. Ratner          Wolters Kluwer        2020
##               Theme
## 1      War and Love
## 2   Historical Epic
## 3 International Law
```

Load XML file

```
url3 <- "https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/books.xml"

books_xml<- getURL(url3)

class(books_xml)
```

```
## [1] "character"
```

```
 ## read the xml content
books_data_xml <- readLines(url3, warn = FALSE)

# join the lines into a single string
books_data_xml <- paste(books_xml, collapse = "\n")

# parse the XML string
books_data_xml<- xmlParse(books_xml)

# convert to a data frame
books_data_xml <- xmlToDataFrame(books_xml)

# show raw results
books_data_xml
```

```
##                                                                   Title
## 1                                                     Farewell to Arms
## 2                                                        War and Peace
## 3 International Law: Laws, Norms, Actors: A Problem-Oriented Approach
##                                    Author               Publisher Publication
## 1                       Ernest Hemingway Charles Scribner's Sons        1929
## 2                            Leo Tolstoy Various (Public Domain)        1869
## 3 Jeffrey L. Dunoff and Steven R. Ratner          Wolters Kluwer        2020
##               Theme
## 1      War and Love
## 2   Historical Epic
## 3 International Law
```

All the data frames at a glance

HTML FORMAT

```
books_data_html
```

```
##                                                                               Title
## 1                                                             Farewell to Arms</li>
## 2                                                                War and Peace</li>
## 3         International Law: Laws, Norms, Actors: A Problem-Oriented Approach</li>
##                                                 Author
## 1                                Ernest Hemingway</li>
## 2                                     Leo Tolstoy</li>
## 3         Jeffrey L. Dunoff and Steven R. Ratner</li>
##                                Publisher        Publication
## 1        Charles Scribner\\'s Sons</li>          1929</li>
## 2           Various (Public Domain)</li>          1869</li>
## 3                    Wolters Kluwer</li>          2020</li>
##                            Theme
## 1              War and Love</li>
## 2           Historical Epic</li>
## 3         International Law</li>
```

JSON FORMAT

```
books_data_json_df
```

```
##                                                                   Title
## 1                                                      Farewell to Arms
## 2                                                         War and Peace
## 3 International Law: Laws, Norms, Actors: A Problem-Oriented Approach
##                                     Author               Publisher Publication
## 1                         Ernest Hemingway Charles Scribner's Sons        1929
## 2                              Leo Tolstoy Various (Public Domain)        1869
## 3 Jeffrey L. Dunoff and Steven R. Ratner          Wolters Kluwer        2020
##               Theme
## 1      War and Love
## 2   Historical Epic
## 3 International Law
```

XML FORMAT

```
books_data_xml
```

```
##                                                                   Title
## 1                                                      Farewell to Arms
## 2                                                         War and Peace
## 3 International Law: Laws, Norms, Actors: A Problem-Oriented Approach
##                                     Author               Publisher Publication
## 1                         Ernest Hemingway Charles Scribner's Sons        1929
## 2                              Leo Tolstoy Various (Public Domain)        1869
## 3 Jeffrey L. Dunoff and Steven R. Ratner          Wolters Kluwer        2020
##               Theme
## 1      War and Love
## 2   Historical Epic
## 3 International Law
```

Conclusion

Each file format behaves differently when read into R. But, all the three formats rendered the entirety of the information that was stored in them. I had more trouble reading the html format into R despite having loaded the required libraries(data.table, rlist and rvest). Fortunately regular expressions came handy and helped me circumvent the issue I faced.