Assignment 7

Mohammad Zahid Chowdhury

2025-03-16

Introuduction: For Assignment 7, I have selected three books, each with two authors. Initially, I have written the code for the HTML, XML, and JSON files in a text editor, saved them accordingly, and uploaded them to my GitHub repository. Then, I read the files in HTML, XML, and JSON formats directly from GitHub and obtained the output in a tabular format.

Load the required packages and libraries:

```
install.packages("xml2", repos = "https://cran.rstudio.com/")
```

```
## Installing package into 'C:/Users/zahid/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

##
##   There is a binary version available but the source version is later:
##      binary source needs_compilation
## xml2  1.3.7  1.3.8              TRUE

## installing the source package 'xml2'

## Warning in install.packages("xml2", repos = "https://cran.rstudio.com/"):
## installation of package 'xml2' had non-zero exit status
```

```
library(xml2)      # For xml file
```

```
## Warning: package 'xml2' was built under R version 4.4.3
```

```
library(jsonlite)  # For JSON processing
library(rvest)     # For HTML processing
```

Read the HTML file

```r
html_data <- read_html("https://raw.githubusercontent.com/zahid607/Assignment-7/refs/heads/main/books.h

# Extract the table
books_table <- html_data %>%
  html_node("table") %>%
  html_table(fill = TRUE)

# Convert to a data frame
df_html <- as.data.frame(books_table)

# Print the data frame
print(df_html)
```

```
##                         Title                              Authors Year
## 1         R for Data Science Hadley Wickham, Garrett Grolemund 2017
## 2   The Art of Data Science   Roger D. Peng, Elizabeth Matsui 2015
## 3 Data Science for Business       Foster Provost, Tom Fawcett 2013
##         Publisher
## 1 O'Reilly Media
## 2        Leanpub
## 3 O'Reilly Media
```

Read the xml file:

```r
# Define the raw XML file URL
xml_url <- "https://raw.githubusercontent.com/zahid607/Assignment-7/main/books.xml"


xml_data <- read_xml(xml_url)

# Extract book nodes
books <- xml_find_all(xml_data, "//book")

# Convert to a data frame
df_xml <- data.frame(
  Title = xml_text(xml_find_first(books, "title")),
  Authors = sapply(books, function(book) {
    paste(xml_text(xml_find_all(book, "authors/author")), collapse = ", ")
  }),
  Year = as.numeric(xml_text(xml_find_first(books, "year"))),
  Publisher = xml_text(xml_find_first(books, "publisher")),
  stringsAsFactors = FALSE
)

# Print the data frame
print(df_xml)
```

```
##                         Title                              Authors Year
## 1         R for Data Science Hadley Wickham, Garrett Grolemund 2017
## 2   The Art of Data Science   Roger D. Peng, Elizabeth Matsui 2015
```

```
## 3 Data Science for Business       Foster Provost, Tom Fawcett 2013
##        Publisher
## 1 O'Reilly Media
## 2        Leanpub
## 3 O'Reilly Media
```

Read the Json file:

```
json_data <- fromJSON("https://raw.githubusercontent.com/zahid607/Assignment-7/refs/heads/main/books.js

df_json <- as.data.frame(json_data$books)

# Convert authors list to string
df_json$Authors <- sapply(df_json$authors, paste, collapse = ", ")
df_json$authors <- NULL  # Remove original authors list

print(df_json)
```

```
##                         title year      publisher
## 1        R for Data Science 2017 O'Reilly Media
## 2   The Art of Data Science 2015        Leanpub
## 3 Data Science for Business 2013 O'Reilly Media
##                               Authors
## 1 Hadley Wickham, Garrett Grolemund
## 2   Roger D. Peng, Elizabeth Matsui
## 3       Foster Provost, Tom Fawcett
```

Are the three data frames identical?

```
# Check if df_html and df_xml are identical
identical(df_html, df_xml)
```

```
## [1] FALSE
```

```
# Check if df_html and df_json are identical
identical(df_html, df_json)
```

```
## [1] FALSE
```

```
# Check if df_xml and df_json are identical
identical(df_xml, df_json)
```

```
## [1] FALSE
```

Conclusion: So, we can conclude that all files HTML, XML and JSON files are not identical.