Data 607 Assignment 7

Joao De Oliveira

2025-10-08

Overview

This report intends to upload three different format files (HTML, XML, and JSON) with the same information to understand the similarities and differences between these three formats. I will normalize and clean the data in order to properly compare if all methods preserve the same data.

Load Libraries

```
library(dplyr)
library(rvest)
library(xml2)
library(jsonlite)
```

Load Data

```
raw_html <- "https://raw.githubusercontent.com/JDO-MSDS/DATA-607/refs/heads/main/Assignment7/books.html"
raw_xml <- "https://raw.githubusercontent.com/JDO-MSDS/DATA-607/refs/heads/main/Assignment7/books.xml"
raw_json <- "https://raw.githubusercontent.com/JDO-MSDS/DATA-607/refs/heads/main/Assignment7/books.json"
```

Data Frames

```
# read html and transform into data frame
doc_html <- read_html(raw_html)
tbl_html <- doc_html %>%
  html_element("table") %>%
  html_table()

# normalization for comparison
df_html <- tbl_html %>%
  mutate(
    title = trimws(as.character(title)),
    authors = trimws(as.character(authors)),
    year = suppressWarnings(as.integer(year)),
    publisher = trimws(as.character(publisher)),
    isbn = trimws(as.character(isbn))
  )
```

```
df_html
```

```
## # A tibble: 3 x 5
##   title                                    authors       year publisher isbn
##   <chr>                                    <chr>        <int> <chr>     <chr>
## 1 Futebol - Observacao, Analise e Intervencao Antonio Bar~  2022 Prime Bo~ 9789~
## 2 Mourinho Anatomy of a Winner             Patrick Bar~  2006 Orion Pu~ 9780~
## 3 The Young Soccer Player                  Gary Lineker  1994 Dorling ~ 978-~
```

```
# read xml and transform into data fram
doc_xml <- read_xml(raw_xml)
books_node <- xml_find_all(doc_xml, ".//book")

rows <- lapply(books_node, function(node){
  data.frame(
    title = xml_text(xml_find_first(node, "./title")),
    authors = paste(xml_text(xml_find_all(node, "./authors/author")), collapse = "; "),
    year = as.integer(xml_text(xml_find_first(node, "./year"))),
    publisher = xml_text(xml_find_first(node, "./publisher")),
    isbn = xml_text(xml_find_first(node, "./isbn")),
    stringsAsFactors = FALSE
  )
})

df_xml <- do.call(rbind, rows)

df_xml
```

```
##                                       title                     authors year
## 1 Futebol - Observacao, Analise e Intervencao Antonio Barbosa; Rui Resende 2022
## 2               Mourinho Anatomy of a Winner             Patrick Barclay 2006
## 3                    The Young Soccer Player                Gary Lineker 1994
##                          publisher           isbn
## 1                      Prime Books  9789896554620
## 2   Orion Publishing Group, Limited  9780752877655
## 3 Dorling Kindersley Publishers Ltd 978-0751351651
```

```
# read json and transform into data frame
j <- fromJSON(raw_json, simplifyVector = FALSE)

df_json <- data.frame(
  title = vapply(j, '[[', "", "title"),
  authors = vapply(j, function(x) paste(x$authors, collapse = "; "), ""),
  year = as.integer(vapply(j, '[[', 0, "year")),
  publisher = vapply(j, '[[', "", "publisher"),
  isbn = vapply(j, '[[', "", "isbn"),
  stringsAsFactors = FALSE
)

df_json
```

```
##                                       title                     authors year
```

```
## 1 Futebol - Observacao, Analise e Intervencao Antonio Barbosa; Rui Resende 2022
## 2                Mourinho Anatomy of a Winner              Patrick Barclay 2006
## 3                     The Young Soccer Player                 Gary Lineker 1994
##                           publisher           isbn
## 1                       Prime Books  9789896554620
## 2   Orion Publishing Group, Limited  9780752877655
## 3 Dorling Kindersley Publishers Ltd 978-0751351651
```

```
desired_cols <- c("title", "authors", "year", "publisher", "isbn")

norm_types <- function(df) {
  df %>%
    mutate(
      title = as.character(title),
      authors = as.character(authors),
      year = as.integer(year),
      publisher = as.character(publisher),
      isbn = as.character(isbn)
    ) %>%
    select(all_of(desired_cols)) %>%
    arrange(title, isbn)
}

df_html_n <- norm_types(df_html)
df_xml_n <- norm_types(df_xml)
df_json_n <- norm_types(df_json)

df_html_n
```

```
## # A tibble: 3 x 5
##   title                                    authors       year publisher isbn
##   <chr>                                    <chr>        <int> <chr>     <chr>
## 1 Futebol - Observacao, Analise e Intervencao Antonio Bar~  2022 Prime Bo~ 9789~
## 2 Mourinho Anatomy of a Winner             Patrick Bar~  2006 Orion Pu~ 9780~
## 3 The Young Soccer Player                  Gary Lineker  1994 Dorling ~ 978-~
```

```
df_xml_n
```

```
##                                       title                    authors year
## 1 Futebol - Observacao, Analise e Intervencao Antonio Barbosa; Rui Resende 2022
## 2                Mourinho Anatomy of a Winner              Patrick Barclay 2006
## 3                     The Young Soccer Player                 Gary Lineker 1994
##                           publisher           isbn
## 1                       Prime Books  9789896554620
## 2   Orion Publishing Group, Limited  9780752877655
## 3 Dorling Kindersley Publishers Ltd 978-0751351651
```

```
df_json_n
```

```
##                                       title                    authors year
## 1 Futebol - Observacao, Analise e Intervencao Antonio Barbosa; Rui Resende 2022
## 2                Mourinho Anatomy of a Winner              Patrick Barclay 2006
```

```
## 3                     The Young Soccer Player                 Gary Lineker 1994
##                           publisher           isbn
## 1                       Prime Books  9789896554620
## 2   Orion Publishing Group, Limited  9780752877655
## 3 Dorling Kindersley Publishers Ltd 978-0751351651
```

```
cmp_html_xml <- all.equal(df_html_n, df_xml_n, check.attributes = FALSE)
cmp_html_json <- all.equal(df_html_n, df_json_n, check.attributes = FALSE)
cmp_json_xml <- all.equal(df_json_n, df_xml_n, check.attributes = FALSE)

cat("html vs xml: ", cmp_html_xml, "\n")
```

```
## html vs xml:  TRUE
```

```
cat("html vs json: ", cmp_html_json, "\n")
```

```
## html vs json:  TRUE
```

```
cat("json vs xml: ", cmp_json_xml, "\n")
```

```
## json vs xml:  TRUE
```

Conclusion

The datasets were all successfully read into R and transformed into data frames with a consistent column structure. After data normalization, I got TRUE for all three comparisons between the normalized data frames with the HTML, XML, and JSON files. Despite the different syntax and document structure before being uploaded, their content was still the same.