DATA607_HW7

Sean Connin

Overview: Working with XML and JSON in R

For this assignment we created individual xml, JSON, and html files to store information related to books we have read and enjoyed. After uploading these files to the internet, we developed R scripts to retrieve the files and then the information into separate dataframes.

This exercise facilitates our understanding of xml, JSON, and html file structures as well as our ablity to manipulate this information.

```
library(xml2)
library(flatxml)
library(methods)
library(htmltab)
library(kableExtra)
library(tidyverse)
library(magrittr)
library(jsonlite)
```

HTML

Read html table and save into dataframe. In this case, I used the htmltab function.

```
#Read in html table from github and save as dataframe

url <- "https://raw.githubusercontent.com/sconnin/DATA607_HW7/main/HW7_html_table.html?token=APC5QM2CW2P3H5W6W6QXZSTAK6PXI"

df_books<-htmltab(doc=url)

#Print table without assigned row numbers

row.names(df_books) <- NULL

df_books%>%kbl%>%kable_material(c("striped"))
```

ID	Title	Author	ISBN	Publisher	Pages
1	Their Eyes Were Watching God	Zora Neale Hurston	9780060916503	HARPERCOLLINS PUBLISHERS	207
2	Sand County Almanac With Essays on Conservation from Round River	Aldo Leopold	9780345345059	PENGUIN RANDOM HOUSE	320
3	Desert Solitaire	Edward Abbey	9780671695880	POCKET BOOKS	288

XML

Read xml file from Github and convert to dataframe. I used the flatxml library for this purpose.

```
# Convert the input xml file to a data frame.

df_xml <- fxml_importXMLFlat("https://raw.githubusercontent.com/sconnin/DATA607_HW7/main/books.xml")


df_xml%>%kbl%>%kable_material(c("striped"))
```

elem.	elemid.	attr.	value.	level1	level2	level3
Books	1	NA	NA	Books	NA	NA
Book	2	NA	NA	Books	Book	NA
ID	3	NA	1	Books	Book	ID
Title	4	NA	Their Eyes Were Watching God	Books	Book	Title
Author	5	NA	Zora Neale Hurston	Books	Book	Author
ISBN	6	NA	9780060916503	Books	Book	ISBN
Publisher	7	NA	HARPERCOLLINS PUBLISHERS	Books	Book	Publisher
Pages	8	NA	207	Books	Book	Pages
Copyright	9	NA	1990	Books	Book	Copyright
Book	10	NA	NA	Books	Book	NA

elem.	elemid.	attr.	value.	level1	level2	level3
ID	11	NA	2	Books	Book	ID
Title	12	NA	Sand County Almanac With Essays on Conservation from Round River	Books	Book	Title
Author	13	NA	Aldo Leopold	Books	Book	Author
ISBN	14	NA	9780345345059	Books	Book	ISBN
Publisher	15	NA	PENGUIN RANDOM HOUSE	Books	Book	Publisher
Pages	16	NA	320	Books	Book	Pages
Copyright	17	NA	1966	Books	Book	Copyright
Book	18	NA	NA	Books	Book	NA
ID	19	NA	3	Books	Book	ID
Title	20	NA	Desert Solitaire	Books	Book	Title
Author	21	NA	Edward Abbey	Books	Book	Author
ISBN	22	NA	9780671695880	Books	Book	ISBN
Publisher	23	NA	POCKET BOOKS	Books	Book	Publisher
Pages	24	NA	288	Books	Book	Pages
Copyright	25	NA	1990	Books	Book	Copyright

JSON

Read Json file from Github and convert to dataframe. I used the JSONlite library for this purpose.

```
url <- "https://raw.githubusercontent.com/sconnin/DATA607_HW7/main/HW7_JSON.json"

books <- jsonlite::fromJSON(url, simplifyVector = TRUE)

df_bks <- data.frame(books)

row.names(df_bks)
```

```
## [1] "1" "2" "3"
```

```
df_bks%>%kbl%>%kable_material(c("striped"))
```

ID	Title	Author	ISBN	Pages	Copyright.Year
1	Their Eyes Were Watching God	Zora Neale Hurston	9780060916503	207	1990
2	Sand County Almanac With Essays on Conservation from Round River	Aldo Leopold	9780345345059	320	1996
3	Desert Solitaire	Edward Abbey	9780671695880	288	1990

Assessment

The dataframes produced from html and JSON file formats are essentially identical. The dataframe produced from the xml file format has been flattened but still retains the hierarchical relationships (e.g., root, path) inherent to xml. Additional manipulation is required to remove these features.