Assignment7_ZhouxinShi

Introduction

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.

```
knitr::opts_chunk$set(warning=FALSE,
                      message=FALSE,
                      tidy=F,
                      #comment = "",
                      dev="png",
                      dev.args=list(type="cairo"))


#create vector with all needed packages
load_packages <- c("RCurl","prettydoc", "stringr", "dplyr", "knitr", "janitor", "XML", "tidyr", "RJSONI


t(t(sapply(load_packages, require, character.only = TRUE, quietly = TRUE,  warn.conflicts = FALSE)))
```

```
##           [,1]
## RCurl     TRUE
## prettydoc TRUE
## stringr   TRUE
## dplyr     TRUE
## knitr     TRUE
## janitor   TRUE
## XML       TRUE
## tidyr     TRUE
## RJSONIO   TRUE
```

Step 1. HTML Table Parsing Load data, look at the HTML table & then its data frame

```
url_html <- getURLContent("https://raw.githubusercontent.com/szx868/data607/master/Assignment7/books2.h


writeLines(url_html)
```

```
## <table>
## <tr> <th>Book Title</th> <th>Author</th> <th>Cover Type</th> <th>Subject</th>    <th>Pages</th>  </t
## <tr> <td>Introduction to Algorithms</td> <td>Thomas H.Cormen</td>    <td>Hard</td>   <td>Computer Sc
## <tr> <td>Introduction to Algorithms</td> <td>Charles E.Leiserson</td>    <td>Hard</td>   <td>Compute
## <tr> <td>Introduction to Algorithms</td> <td>Ronald L.Rivest</td>    <td>Hard</td>   <td>Computer Sc
## <tr> <td>Introduction to Algorithms</td> <td>Clifford Stein</td> <td>Hard</td>   <td>Computer Scienc
```

```
## <tr> <td>Harry Potter and the Sorcerer's Stone</td>  <td>J. K. Rowling</td>  <td>Hard</td>   <td>Fic
## <tr> <td>New Moon: The Twilight Saga</td>    <td>Stephenie Meyer</td>    <td>Soft</td>   <td>Fiction
## </table>
```

```
my_html_df <- url_html %>%
  readHTMLTable(header=TRUE, as.data.frame = TRUE) %>%
  data.frame(stringsAsFactors = FALSE) %>%
  clean_names()

colnames(my_html_df) <- str_replace(colnames(my_html_df),"null_", "")
my_html_df <- my_html_df %>% arrange(desc(book_title))
kable(my_html_df)
```

book_title	author	cover_type	subject	pages
New Moon: The Twilight Saga	Stephenie Meyer	Soft	Fiction	512
Introduction to Algorithms	Thomas H.Cormen	Hard	Computer Science	1292
Introduction to Algorithms	Charles E.Leiserson	Hard	Computer Science	1292
Introduction to Algorithms	Ronald L.Rivest	Hard	Computer Science	1292
Introduction to Algorithms	Clifford Stein	Hard	Computer Science	1292
Harry Potter and the Sorcerer's Stone	J. K. Rowling	Hard	Fiction	256

Step 2. JSON Parsing Load data, look at the JSON & then its data frame

```
url_json <- getURLContent("https://raw.githubusercontent.com/szx868/data607/master/Assignment7/books.js

print("Is my JSON valid?")
```

```
## [1] "Is my JSON valid?"
```

```
isValidJSON("https://raw.githubusercontent.com/szx868/data607/master/Assignment7/books.json")
```

```
## [1] TRUE
```

```
writeLines(url_json)
```

```
## {"favorite recent books":[
##  {
##  "Book Title": "Harry Potter and the Sorcerer's Stone",
##  "Authors": "J. K. Rowling",
##  "Cover Type": "Hard",
##  "Subject": "Fiction",
##  "Pages": 256
##  },
##  {
##  "Book Title": "New Moon: The Twilight Saga",
##  "Authors": "Stephenie Meyer",
##  "Cover Type": "Soft",
##  "Subject": "Fiction",
##  "Pages": 512
```

```
##  },
##      {
##  "Book Title": "Introduction to Algorithms",
##  "Authors": ["Thomas H.Cormen", "Charles E.Leiserson", "Ronald L.Rivest", "Clifford Stein"],
##  "Cover Type": "Hard",
##  "Subject": "Computer Science",
##  "Pages": 1292
##  }]
## }
```

```
my_json_df <- fromJSON(url_json)

my_json_df <- do.call("rbind", lapply(my_json_df$`favorite recent books`, data.frame, stringsAsFactors =

my_json_df <- my_json_df %>%
  clean_names() %>%
  arrange(desc(book_title))

kable(my_json_df, caption = "data frame looks the same as the HTML one.")
```

Table 2: data frame looks the same as the HTML one.

book_title	authors	cover_type	subject	pages
New Moon: The Twilight Saga	Stephenie Meyer	Soft	Fiction	512
Introduction to Algorithms	Thomas H.Cormen	Hard	Computer Science	1292
Introduction to Algorithms	Charles E.Leiserson	Hard	Computer Science	1292
Introduction to Algorithms	Ronald L.Rivest	Hard	Computer Science	1292
Introduction to Algorithms	Clifford Stein	Hard	Computer Science	1292
Harry Potter and the Sorcerer's Stone	J. K. Rowling	Hard	Fiction	256

Step 3. XML Parsing Load data, look at the XML & then its data frame

```
url_XML <- getURLContent("https://raw.githubusercontent.com/szx868/data607/master/Assignment7/books.xml

writeLines(url_XML)
```

```
## <?xml version="1.0" encoding="UTF-8" ?>
## <!--These are some of my recent favorite books-->
## <books>
##  <book>
##      <book_title>Introduction to Algorithms</book_title>
##      <authors>
##          <author ID="1">Thomas H.Cormen</author>
##          <author ID="2">Charles E.Leiserson</author>
##          <author ID="3">Ronald L.Rivest</author>
##          <author ID="4">Clifford Stein</author>
##      </authors>
##      <cover_type>Hard</cover_type>
##      <subject>Computer Science</subject>
##      <pages>1292</pages>
##  </book>
```

```
##  <book>
##      <book_title>Harry Potter and the Sorcerer's Stone</book_title>
##      <authors>
##          <author ID="1">J. K. Rowling</author>
##      </authors>
##      <cover_type>Hard</cover_type>
##      <subject>Fiction</subject>
##      <pages>256</pages>
##  </book>
##  <book>
##      <book_title>New Moon: The Twilight Saga</book_title>
##      <authors>
##          <author ID="1">Stephenie Meyer</author>
##      </authors>
##      <cover_type>Soft</cover_type>
##      <subject>Fiction</subject>
##      <pages>512</pages>
##  </book>
## </books>
```

```
my_XML_df <- url_XML %>%
  xmlParse() %>%
  xmlToDataFrame(stringsAsFactors = FALSE)

kable(my_XML_df, caption = "This does not look the same as the first 2 data frames. For the book with m
```

Table 3: This does not look the same as the first 2 data frames. For the book with more than one author, the function concatenated all of them into a single cell. Let's do a little bit of surgery to get the same result.

book_title	authors	cover_type	subject	pages
Introduction to Algorithms	Thomas H.CormenCharles E.LeisersonRonald L.RivestClifford Stein	Hard	Computer Science	1292
Harry Potter and the Sorcerer's Stone	J. K. Rowling	Hard	Fiction	256
New Moon: The Twilight Saga	Stephenie Meyer	Soft	Fiction	512

```
my_XML_df2 <- my_XML_df %>%
  mutate(authors = paste(str_replace_all(authors, "([a-z])([A-Z])", "\\1,\\2"))) %>%
  separate(authors, c(paste0("author_", 1:4)), sep = ",") %>%
  gather(author_num, author, author_1:author_4, na.rm = T) %>%
  select(book_title, author, everything(), -author_num) %>%
  arrange(desc(book_title))

kable(my_XML_df2, caption = "And now it looks like the other 2 data frames")
```

Table 4: And now it looks like the other 2 data frames

book_title	author	cover_type	subject	pages
New Moon: The Twilight Saga	Stephenie Meyer	Soft	Fiction	512
Introduction to Algorithms	Thomas H.Cormen	Hard	Computer Science	1292
Introduction to Algorithms	Charles E.Leiserson	Hard	Computer Science	1292
Introduction to Algorithms	Ronald L.Rivest	Hard	Computer Science	1292
Introduction to Algorithms	Clifford Stein	Hard	Computer Science	1292
Harry Potter and the Sorcerer's Stone	J. K. Rowling	Hard	Fiction	256

```
my_json_df == my_html_df
```

```
##      book_title authors cover_type subject pages
## [1,]       TRUE    TRUE       TRUE    TRUE  TRUE
## [2,]       TRUE    TRUE       TRUE    TRUE  TRUE
## [3,]       TRUE    TRUE       TRUE    TRUE  TRUE
## [4,]       TRUE    TRUE       TRUE    TRUE  TRUE
## [5,]       TRUE    TRUE       TRUE    TRUE  TRUE
## [6,]       TRUE    TRUE       TRUE    TRUE  TRUE
```

```
my_html_df == my_XML_df2
```

```
##      book_title author cover_type subject pages
## [1,]       TRUE   TRUE       TRUE    TRUE  TRUE
## [2,]       TRUE   TRUE       TRUE    TRUE  TRUE
## [3,]       TRUE   TRUE       TRUE    TRUE  TRUE
## [4,]       TRUE   TRUE       TRUE    TRUE  TRUE
## [5,]       TRUE   TRUE       TRUE    TRUE  TRUE
## [6,]       TRUE   TRUE       TRUE    TRUE  TRUE
```