607 Fall 2021 HW6 Json_HTML_XML

Mark Schmalfeld

10/10/2021

Load library

```
library(RCurl,curl)
library (rvest)
library(jsonlite, rjson)
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:jsonlite':
##
##     flatten
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::complete()       masks RCurl::complete()
## x dplyr::filter()         masks stats::filter()
## x purrr::flatten()        masks jsonlite::flatten()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
```

```
library(xml2)
library(XML)
library (methods)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

HTML loaded in this case directly into R as HTML formated text to extract the table.

Create book list in HTML

```
HTML_df <- as.data.frame(read_html("<html>
<body>

 <h1>Books Catalog</h1>
   <table style= "width:100%">
        <tr>
            <th>Book Title</th>
            <th>Author1</th>
           <th>Author2</th>
            <th>Attribute1</th>
            <th>Attribute2</th>
            <th>Attribute3</th>
            <th>Attribute4</th>
         </tr>
         <tr>

            <td>Odyssey</td>
            <td>Homer</td>
           <td>NA</td>
            <td>Adventure</td>
            <td>Heroic</td>
            <td>Greek</td>
            <td>Trust</td>
         </tr>
         <tr>

            <td>The Federalist Papers</td>
            <td>Alexander Hamilton</td>
          <td>John Jay</td>
            <td>States Rights</td>
            <td>Taxiation</td>
            <td>Liberty</td>
            <td>Union</td>
         </tr>
         <tr>

           <td>Practical Cats</td>
            <td>TS Eliot</td>
           <td>NA</td>
            <td>Society</td>
            <td>Humor</td>
            <td>Life</td>
            <td>Cats as humans</td>
         </tr>
      </table>
```

```
   </body>
</html>") %>% html_table(fill=TRUE))
HTML_df
```

```
##              Book.Title            Author1  Author2    Attribute1 Attribute2
## 1               Odyssey              Homer     <NA>     Adventure     Heroic
## 2 The Federalist Papers Alexander Hamilton John Jay States Rights  Taxiation
## 3        Practical Cats           TS Eliot     <NA>       Society      Humor
##   Attribute3     Attribute4
## 1      Greek          Trust
## 2    Liberty          Union
## 3       Life Cats as humans
```

```
strHTML<-str(HTML_df)
```

```
## 'data.frame':    3 obs. of  7 variables:
##  $ Book.Title: chr  "Odyssey" "The Federalist Papers" "Practical Cats"
##  $ Author1   : chr  "Homer" "Alexander Hamilton" "TS Eliot"
##  $ Author2   : chr  NA "John Jay" NA
##  $ Attribute1: chr  "Adventure" "States Rights" "Society"
##  $ Attribute2: chr  "Heroic" "Taxiation" "Humor"
##  $ Attribute3: chr  "Greek" "Liberty" "Life"
##  $ Attribute4: chr  "Trust" "Union" "Cats as humans"
```

Download the Booklist XML file from the github and convert into a dataframe Evaluate for comparison using the str comparison.

```
download.file("https://raw.githubusercontent.com/schmalmr/607-Fall-2021-HW6-Jason/main/Booklist3.xml", d

booksXML_File <- xmlParse("BooksList.xml")
booksXML_df <- xmlToDataFrame(booksXML_File)
booksXML_df
```

```
##               Booktitle            Author1  Author2    Attribute1 Attribute2
## 1               Odyssey              Homer       na     Adventure     Heroic
## 2 The Federalist Papers Alexander Hamilton John Jay States Rights  Taxiation
## 3        Practical Cats           TS Eliot       na       Society      Humor
##   Attribute3     Attribute4
## 1      Greek          Trust
## 2    Liberty          Union
## 3       Life Cats as humans
```

```
strXML<-str(booksXML_df)
```

```
## 'data.frame':    3 obs. of  7 variables:
##  $ Booktitle : chr  "Odyssey" "The Federalist Papers" "Practical Cats"
##  $ Author1   : chr  "Homer" "Alexander Hamilton" "TS Eliot"
##  $ Author2   : chr  "na" "John Jay" "na"
##  $ Attribute1: chr  "Adventure" "States Rights" "Society"
##  $ Attribute2: chr  "Heroic" "Taxiation" "Humor"
##  $ Attribute3: chr  "Greek" "Liberty" "Life"
##  $ Attribute4: chr  "Trust" "Union" "Cats as humans"
```

Load the JSON code for the book list into the system and convert to data frame. Data frame conversion was a wide data frame.

Additonal tidying is needed to convert it to a tidy data frame.

Create the string comparison file.

```
#url<-"https://raw.githubusercontent.com/schmalmr/607-Fall-2021-HW6-Jason/main/Booklisttext.json"

json <-
'[
  {"Book" : "Odyssey", "Author1" : "Homer", "Author2" : "NA" , "Attribute1" : "Adventure", "Attribute2"
  {"Book" : "The Federlist Papers", "Author1" : "Alexander Hamilton", "Author2" : "John Jay", "Attribut
{"Book" : "Practical Cats", "Author1" : "T.S. Eliot" , "Author2" : "NA", "Attribute1": "Society","Attri
]'
Json_df <- fromJSON(json)
Json_df
```

```
##                   Book            Author1  Author2    Attribute1 Attribute2
## 1              Odyssey              Homer       NA     Adventure     Heroic
## 2 The Federlist Papers Alexander Hamilton John Jay States Rights  Taxiation
## 3       Practical Cats         T.S. Eliot       NA       Society      Humor
##   Attribute3     Attribute4
## 1      Greek          Trust
## 2    Liberty          Union
## 3       Life Cats as Humans
```

```
df_Json <- as.data.frame(Json_df)
df_Json
```

```
##                   Book            Author1  Author2    Attribute1 Attribute2
## 1              Odyssey              Homer       NA     Adventure     Heroic
## 2 The Federlist Papers Alexander Hamilton John Jay States Rights  Taxiation
## 3       Practical Cats         T.S. Eliot       NA       Society      Humor
##   Attribute3     Attribute4
## 1      Greek          Trust
## 2    Liberty          Union
## 3       Life Cats as Humans
```

```
Json_str <-str(df_Json)
```

```
## 'data.frame':    3 obs. of  7 variables:
##  $ Book      : chr  "Odyssey" "The Federlist Papers" "Practical Cats"
##  $ Author1   : chr  "Homer" "Alexander Hamilton" "T.S. Eliot"
##  $ Author2   : chr  "NA" "John Jay" "NA"
##  $ Attribute1: chr  "Adventure" "States Rights" "Society"
##  $ Attribute2: chr  "Heroic" "Taxiation" "Humor"
##  $ Attribute3: chr  "Greek" "Liberty" "Life"
##  $ Attribute4: chr  "Trust" "Union" "Cats as Humans"
```

Compare HTML and XML formats

```
all_equal(HTML_df,booksXML_df)
```

```
## [1] "not compatible: \n- Cols in y but not x: ‘Booktitle‘.\n- Cols in x but not y: ‘Book.Title‘.\n"
```

Compare the HTML and the XML with the Jason table. (already know they are not the same pending further work to gather into header columns instead of the wide format create)

```
all.equal(df_Json,HTML_df)
```

```
## [1] "Names: 1 string mismatch"
## [2] "Component 1: 1 string mismatch"
## [3] "Component \"Author1\": 1 string mismatch"
## [4] "Component \"Author2\": 'is.NA' value mismatch: 2 in current 0 in target"
## [5] "Component \"Attribute4\": 1 string mismatch"
```

```
all.equal(df_Json,booksXML_df)
```

```
## [1] "Names: 1 string mismatch"
## [2] "Component 1: 1 string mismatch"
## [3] "Component \"Author1\": 1 string mismatch"
## [4] "Component \"Author2\": 2 string mismatches"
## [5] "Component \"Attribute4\": 1 string mismatch"
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.