Homework 7

OMER OZEREN

March 17, 2019

Table of Contents

Assignment

Working with HTML, XML and JSON in R

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.

Take the information that you've selected about these three books, and separately create three files which store the book's information in HTML (using an html table), XML, and JSON formats (e.g. "book.html", "book.xml", and "book.json"). To help you better understand the different file structures, I'd prefer that you create each of these files "by hand" unless you're already very comfortable with the file formats.

References to the books.html, books.xml, and books.json.

books.html:
https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.html

books.xml:
https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.xml

books.jason:
https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.json

Load libraries:

```
library(RCurl)
library(XML)
library(jsonlite)
library(data.table)
```

HTML

Import HTML file

```
html_url  <-
"https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.html
"
html_file <- getURL(html_url)
```

Parsing the data

```
html_file <- htmlParse(html_file)
```

Converting into a data.frame

```
html_df <- as.data.frame(readHTMLTable(html_file))
html_df

##              X.Top.three.books..Book.Title X.Top.three.books..Year
## 1                            Deep Learning                    2015
## 2 Pattern Recognition and Machine Learning                    2017
## 3                    Financial Time Series                    2008
##   X.Top.three.books..ISBN X.Top.three.books..Author
## 1          978-0739435571            Ian Goodfellow
## 2          978-0060883287               Mark Bishop
## 3          978-0131103627                 Jack Tsay
##     X.Top.three.books..Publisher
## 1                      Mit Press
## 2     Cambridge University Press
## 3 World Scientific Publishing Co
```

XML

Import XML file

```
xml_url <-
"https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.xml"
xml_file <- getURL(xml_url)
```

Parsing the data

```
xml_file <- xmlParse(xml_file)
```

Converting into a data.frame

```
part_1 <- xmlRoot(xml_file)
xml_df <- xmlToDataFrame(part_1)
xml_df

##                                       title year           isbn
## 1                             Deep Learning 2015 978-0739435571
## 2 Pattern Recognition and Machine Learning 2017 978-0060883287
## 3                     Financial Time Series 2008 978-0131103627
##            author                      publisher
## 1 Ian Goodfellow                      MIT Press
## 2    Mark Bishop     Cambridge University Press
## 3      Jack Tsay World Scientific Publishing Co
```

JSON

Import JSON file

```
json_url <-
"https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_7/books.json
"
json_file <- getURL(json_url)
```

Parsing the data

```
json_file <- fromJSON(json_file)
```

Converting into a data.frame

```
json_file_df <- as.data.frame(json_file)
json_file_df

##                        sergios_books.title sergios_books.year
## 1                            Deep Learning               2015
## 2 attern Recognition and Machine Learning               2017
## 3                    Financial Time Series               2008
##   sergios_books.isbn sergios_books.author        sergios_books.publisher
## 1     978-0739435571       Ian Goodfellow                      MIT Press
## 2     978-0060883287          Mark Bishop     Cambridge University Press
## 3     978-0131103627            Jack Tsay World Scientific Publishing Co
```

Are the three data frames identical?

We can look at the structures of each data.frame to see if they are identical.

```
str(html_df)

## 'data.frame':    3 obs. of  5 variables:
##  $ X.Top.three.books..Book.Title: Factor w/ 3 levels "Deep Learning",..: 1
3 2
##  $ X.Top.three.books..Year      : Factor w/ 3 levels "2008","2015",..: 2 3
```

```
1
##  $ X.Top.three.books..ISBN      : Factor w/ 3 levels "978-0060883287",..:
3 1 2
##  $ X.Top.three.books..Author    : Factor w/ 3 levels "Ian Goodfellow",..:
1 3 2
##  $ X.Top.three.books..Publisher : Factor w/ 3 levels "Cambridge University
Press",..: 2 1 3
```

```
str(xml_df)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ title    : Factor w/ 3 levels "Deep Learning",..: 1 3 2
##  $ year     : Factor w/ 3 levels "2008","2015",..: 2 3 1
##  $ isbn     : Factor w/ 3 levels "978-0060883287",..: 3 1 2
##  $ author   : Factor w/ 3 levels "Ian Goodfellow",..: 1 3 2
##  $ publisher: Factor w/ 3 levels "Cambridge University Press",..: 2 1 3
```

```
str(json_file_df)
```

```
## 'data.frame':    3 obs. of  5 variables:
##  $ sergios_books.title    : chr  "Deep Learning" "attern Recognition and
Machine Learning" "Financial Time Series"
##  $ sergios_books.year     : chr  "2015" "2017" "2008"
##  $ sergios_books.isbn     : chr  "978-0739435571" "978-0060883287" "978-
0131103627"
##  $ sergios_books.author   : chr  "Ian Goodfellow" "Mark Bishop" "Jack
Tsay"
##  $ sergios_books.publisher: chr  "MIT Press" "Cambridge University Press"
"World Scientific Publishing Co"
```

After looking at the structures, the HTML and XML data.frame looks identical to each other (other than the names of the headers i.e. Null.Book.Name vs. Book_Name). They are all listed as factors with multiple levels, whereas, the JSON data frame uses chr, int in their data.frame.