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

## top\_three\_books.title top\_three\_books.year  
## 1 Deep Learning 2015  
## 2 attern Recognition and Machine Learning 2017  
## 3 Financial Time Series 2008  
## top\_three\_books.isbn top\_three\_books.author  
## 1 978-0739435571 Ian Goodfellow  
## 2 978-0060883287 Mark Bishop  
## 3 978-0131103627 Jack Tsay  
## top\_three\_books.publisher  
## 1 MIT Press  
## 2 Cambridge University Press  
## 3 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:  
## $ top\_three\_books.title : chr "Deep Learning" "attern Recognition and Machine Learning" "Financial Time Series"  
## $ top\_three\_books.year : chr "2015" "2017" "2008"  
## $ top\_three\_books.isbn : chr "978-0739435571" "978-0060883287" "978-0131103627"  
## $ top\_three\_books.author : chr "Ian Goodfellow" "Mark Bishop" "Jack Tsay"  
## $ top\_three\_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.