Assignment Week 7

Umais Siddiqui

October 15, 2017

Rpubs:http://rpubs.com/umais/data607_week7_assignment

Github:https://github.com/umais/DATA-607/tree/master/Week7_Assignment

Assignment Week 7

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. "books.html", "books.xml", and "books.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. Write R code, using your packages of choice, to load the information from each of the three sources into separate R data frames. Are the three data frames identical? Your deliverable is the three source files and the R code. If you can, package your assignment solution up into an .Rmd file and publish to rpubs.com. [This will also require finding a way to make your three text files accessible from the web].

Reading Data from an HTML File

```
url <- "books.html"

tbls_xml <- readHTMLTable(url)

htmlData=tbls_xml[[1]]
names(htmlData)<-c("title","ISBN","copyright","author1","author2")

htmlData
```

```
##                                          title           ISBN copyright
## 1 Agile Principles, Patterns, and Practices in C# 978-0131857254      2007
## 2                                       Dracula                      2005
## 3                    The Island of Doctor Moreau   978155111327      2009
##               author1      author2
## 1    Robert C. Martin Micah Martin
## 2 William Shakespeare         <NA>
## 3          H.G. Wells         <NA>
```

Reading Data from XML file

```
data <- xmlParse("books.xml")
rootNode <- xmlRoot(data)
rootNode[1]
```

```
## $book
## <book>
##   <title>Agile Principles, Patterns, and Practices in C#</title>
##   <ISBN>978-0131857254</ISBN>
##   <copyright>2007</copyright>
##   <author1>Robert C. Martin</author1>
##   <author2>Micah Martin</author2>
## </book>
##
## attr(,"class")
## [1] "XMLInternalNodeList" "XMLNodeList"
```

```
data <- xmlSApply(rootNode,function(x) xmlSApply(x, xmlValue))
booksXML <- data.frame(t(data),row.names=NULL)

booksXML
```

```
##                                          title           ISBN copyright
## 1 Agile Principles, Patterns, and Practices in C# 978-0131857254      2007
## 2                                      Dracula  9780316014816      2005
## 3                    The Island of Doctor Moreau   978155111327      2009
##               author1      author2
## 1    Robert C. Martin Micah Martin
## 2 William Shakespeare
## 3          H.G. Wells
```

Reading data from JSON File

```
library(jsonlite)

url <- 'books.json'

# read url and convert to data.frame
document <- fromJSON(txt=url)

document
```

```
##                                          title         ISBN13 copyright
## 1 Agile Principles, Patterns, and Practices in C# 978-0131857254      2007
## 2                                      Dracula  9780316014816      2005
## 3                    The Island of Doctor Moreau   978155111327      2009
##               author1      author2
## 1    Robert C. Martin Micah Martin
## 2 William Shakespeare
## 3          H.G. Wells
```

Observations

The three data sets are almost identical. Even though the data is read from the files using different menthods the XML and JSON data frames are identical. However, when data is read from HTML the columns are named V1 , V2 etc. The columns had to be renamed.