WebTech-HTML,XML,JSON

Chunhui Zhu

October 11, 2017

Assignment - Working with 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. "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].

R enviroment

1.HTML

Read in .html file

```
setwd("C://Users/Ivy/Desktop/607/W7")
htmldata<-htmlTreeParse('Books.html', useInternalNodes = T)
htmldata
```

```
## <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose
## <html>
## <head></head>
## <body><ul>
## <p>
##         </p>
## <li><b>Book: R for Data Science</b></li>
##         <li><i>Author: Hadley Wickham, Garrett Grolemund</i></li>
##          <li>ISBN : 1491910399  </li>
##      <li>Year : 2015 </li>
##      <li>Publisher: John Wiley and Sons,Inc </li>
##
##         <p>
##         </p>
## <li><b>Book: Practical Statistics for Data Scientists: 50 Essential Concepts</b></li>
##         <li><i>Author: Peter Bruce, Andrew Bruce </i></li>
##          <li>ISBN : 1491952962 </li>
##      <li>Year : 2017 </li>
##      <li>Publisher: O'Reilly Media,Inc </li>
##
##
```

```
##         <p>
##         </p>
## <li><b>Book: An Introduction to Statistical Learning: with Applications in R </b></li>
##      <li><i>Author: Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani</i></li>
##          <li>ISBN : 1461471370  </li>
##      <li>Year : 2013 </li>
##      <li>Publisher: Springer </li>
##
##
##         </ul></body>
## </html>
##
```

GetNodeSet will get value between "ul" and "li" sets. then store value in a list

```
htmldf<-getNodeSet(htmldata,"//ul//li")
htmldf<-sapply(htmldf,xmlValue)
htmldf
```

```
##  [1] "Book: R for Data Science"
##  [2] "Author: Hadley Wickham, Garrett Grolemund"
##  [3] "ISBN : 1491910399  "
##  [4] "Year : 2015 "
##  [5] "Publisher: John Wiley and Sons,Inc "
##  [6] "Book: Practical Statistics for Data Scientists: 50 Essential Concepts"
##  [7] "Author: Peter Bruce, Andrew Bruce "
##  [8] "ISBN : 1491952962 "
##  [9] "Year : 2017 "
## [10] "Publisher: O'Reilly Media,Inc "
## [11] "Book: An Introduction to Statistical Learning: with Applications in R "
## [12] "Author: Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani"
## [13] "ISBN : 1461471370  "
## [14] "Year : 2013 "
## [15] "Publisher: Springer "
```

Strsplit returns a list; use sapply to get the 2nd obs of each list element

```
htmldf1<- sapply(strsplit(htmldf,"\\:"), `[`, 2)
htmldf1
```

```
##  [1] " R for Data Science"
##  [2] " Hadley Wickham, Garrett Grolemund"
##  [3] " 1491910399  "
##  [4] " 2015 "
##  [5] " John Wiley and Sons,Inc "
##  [6] " Practical Statistics for Data Scientists"
##  [7] " Peter Bruce, Andrew Bruce "
##  [8] " 1491952962 "
##  [9] " 2017 "
## [10] " O'Reilly Media,Inc "
## [11] " An Introduction to Statistical Learning"
## [12] " Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani"
## [13] " 1461471370  "
## [14] " 2013 "
## [15] " Springer "
```

Transform data frame

```
htmldf1<-as.data.frame(matrix (htmldf1,nrow=5))
htmldf1<-as.data.frame(t(htmldf1))
colnames(htmldf1)<-c("Book","Author","ISBN","Year","Publisher")
htmldf1
```

```
##                                       Book
## V1                       R for Data Science
## V2  Practical Statistics for Data Scientists
## V3   An Introduction to Statistical Learning
##                                                                Author
## V1                              Hadley Wickham, Garrett Grolemund
## V2                                     Peter Bruce, Andrew Bruce
## V3  Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani
##             ISBN   Year                 Publisher
## V1  1491910399    2015   John Wiley and Sons,Inc
## V2   1491952962   2017        O'Reilly Media,Inc
## V3  1461471370    2013                  Springer
```

2.XML

Read in .xml file

```
xmldata<-xmlParse(file="Books.xml")
xmldata
```

```
## <?xml version="1.0" encoding="utf-8"?>
## <Books>
##   <Info>
##     <Book>
##       <b>R for Data Science</b>
##     </Book>
##     <Author>
##       <i>Hadley Wickham, Garrett Grolemund </i>
##     </Author>
##     <ISBN>1491910399 </ISBN>
##     <Year>2015 </Year>
##     <Publisher>John Wiley and Sons,Inc </Publisher>
##   </Info>
##   <Info>
##     <Book>
##       <b>Practical Statistics for Data Scientists: 50 Essential Concepts</b>
##     </Book>
##     <Author>
##       <i>Peter Bruce, Andrew Bruce </i>
##     </Author>
##     <ISBN>1491952962 </ISBN>
##     <Year>2017 </Year>
##     <Publisher>O'Reilly Media,Inc </Publisher>
##   </Info>
##   <Info>
##     <Book>
##       <b>An Introduction to Statistical Learning: with Applications in R </b>
##     </Book>
```

```
##     <Author>
##       <i>Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani</i>
##     </Author>
##     <ISBN>1461471370  </ISBN>
##     <Year>2013 </Year>
##     <Publisher>Springer </Publisher>
##   </Info>
## </Books>
##
```

Xml has nice data frame formate when read into r.

```
xmldf<-xmlRoot(xmldata)
xmldf1<-xmlToDataFrame(xmldf)
xmldf1
```

```
##                                                                  Book
## 1                                                 R for Data Science
## 2  Practical Statistics for Data Scientists: 50 Essential Concepts
## 3 An Introduction to Statistical Learning: with Applications in R
##                                                            Author         ISBN
## 1                          Hadley Wickham, Garrett Grolemund   1491910399
## 2                                  Peter Bruce, Andrew Bruce   1491952962
## 3 Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani 1461471370
##    Year                Publisher
## 1 2015  John Wiley and Sons,Inc
## 2 2017       O'Reilly Media,Inc
## 3 2013                 Springer
```

3.JSON

Read in .json file, and value automatically store in a list.

```
jsondata<-fromJSON(file="Books.json")
jsondata
```

```
## [[1]]
## [[1]]$Book
## [1] "R for Data Science"
##
## [[1]]$Author
## [1] "Hadley Wickham, Garrett Grolemund"
##
## [[1]]$ISBN
## [1] 1491910399
##
## [[1]]$Year
## [1] 2015
##
## [[1]]$Publisher
## [1] "John Wiley and Sons,Inc"
##
##
## [[2]]
```

```
## [[2]]$Book
## [1] "Practical Statistics for Data Scientists: 50 Essential Concepts"
##
## [[2]]$Author
## [1] "Peter Bruce, Andrew Bruce"
##
## [[2]]$ISBN
## [1] 1491952962
##
## [[2]]$Year
## [1] 2017
##
## [[2]]$Publisher
## [1] "John Wiley and Sons,Inc"
##
##
## [[3]]
## [[3]]$Book
## [1] "An Introduction to Statistical Learning: with Applications in R"
##
## [[3]]$Author
## [1] "Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani"
##
## [[3]]$ISBN
## [1] 1461471370
##
## [[3]]$Year
## [1] 2013
##
## [[3]]$Publisher
## [1] "Springer"
```

lapply returns a list of jsondata length as X, use do.call to constructs and executes "rbind" call from jsondf of arguments to be passed to it.

```
jsondf<- lapply(jsondata, function(x) {unlist(x)})
as.data.frame( do.call("rbind", jsondf))
```

```
##                                                                  Book
## 1                                            R for Data Science
## 2 Practical Statistics for Data Scientists: 50 Essential Concepts
## 3 An Introduction to Statistical Learning: with Applications in R
##                                                          Author       ISBN
## 1                          Hadley Wickham, Garrett Grolemund 1491910399
## 2                                 Peter Bruce, Andrew Bruce 1491952962
## 3 Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani 1461471370
##   Year               Publisher
## 1 2015 John Wiley and Sons,Inc
## 2 2017 John Wiley and Sons,Inc
## 3 2013                Springer
```