

# Week 8

*Dan Brooks*

*March 15, 2016*

```
library(XML)
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(httr)
library(RJSONIO)
library(plyr)
library(stringr)
```

```
url <- "https://raw.githubusercontent.com/DanielBrooks39/IS607/master/Week%208/books.xml"
doc.xml <- xmlParse(rawToChar(GET(url)$content))
doc.xml
```

```
## <?xml version="1.0" encoding="ISO-8859-1"?>
## <Textbooks>
##   <book id="1">
##     <title>Differential Equations with Boundary Value Problems</title>
##     <author_one>John Porking</author_one>
##     <author_two>Al Boggess</author_two>
##     <author_three>David Arnold</author_three>
##     <ISBN>0131862367</ISBN>
##     <publisher>Prentice Hall</publisher>
##     <year>2006</year>
##     <book_type>Hardback</book_type>
##     <used_price>$150.60</used_price>
##     <new_price>$158.40</new_price>
##     <website><![CDATA[Textbooks.com]]></website>
##   </book>
##   <book id="2">
##     <title>Fourier Series</title>
##     <author_one>Rajendra Bhatia</author_one>
##     <ISBN>0883857405</ISBN>
##     <publisher>Mathematical Association of America</publisher>
##     <year>2004</year>
##     <book_type>Hardback</book_type>
##     <used_price>$12.21</used_price>
##     <new_price>$35.00</new_price>
##     <website><![CDATA[Amazon.com]]></website>
##   </book>
##   <book id="3">
##     <title>Linear Algebra</title>
##     <author_one>Stephen H. Freidberg</author_one>
##     <author_two>Arnold J. Insel</author_two>
##     <author_three>Lawrence E. Spence</author_three>
##     <ISBN>0130084514</ISBN>
```

```
##      <publisher>Prentice Hall</publisher>
##      <year>2003</year>
##      <book_type>Hardback</book_type>
##      <used_price>$115.21</used_price>
##      <new_price>$163.48</new_price>
##      <website><![CDATA[Textbooks.com]]></website>
##    </book>
## </Textbooks>
##
```

```
root <- xmlRoot(doc.xml)
```

```
xml.df <- xmlToDataFrame(root)
```

```
names(xml.df) <- c("Title", "Author One", "Author Two", "Author Three", "ISBN", "Publisher", "Year", "B", "Price", "New Price", "Website")
xml.df
```

```
##                                     Title                Author One
## 1 Differential Equations with Boundary Value Problems      John Porking
## 2                                     Fourier Series      Rajendra Bhatia
## 3                                     Linear Algebra Stephen H. Freidberg
##      Author Two      Author Three      ISBN
## 1      Al Boggess      David Arnold 0131862367
## 2      <NA>            <NA> 0883857405
## 3 Arnold J. Insel Lawrence E. Spence 0130084514
##                                     Publisher Year Book Type Used Price New Price
## 1                                     Prentice Hall 2006  Hardback   $150.60  $158.40
## 2 Mathematical Association of America 2004  Hardback    $12.21   $35.00
## 3                                     Prentice Hall 2003  Hardback   $115.21  $163.48
##      Website
## 1 Textbooks.com
## 2      Amazon.com
## 3 Textbooks.com
```

```
url <- "https://raw.githubusercontent.com/DanielBrooks39/IS607/master/Week%208/books.json"
doc.json <- fromJSON(url)
doc.json
```

```
## $textbooks
## $textbooks[[1]]
## $textbooks[[1]]$rank
## [1] 1
##
## $textbooks[[1]]$title
## [1] "Differential Equations with Boundary Value Problems"
##
## $textbooks[[1]]$authors
##      one      two      three
## "John Porking" "Al Boggess" "David Arnold"
##
## $textbooks[[1]]$ISBN
## [1] "0131862367"
##
## $textbooks[[1]]$publisher
```

```

## [1] "Prentice Hall"
##
## $textbooks[[1]]$Year
## [1] 2006
##
## $textbooks[[1]]$Book_Type
## [1] "Harback"
##
## $textbooks[[1]]$Used_Price
## [1] 105.6
##
## $textbooks[[1]]$New_Price
## [1] 158.4
##
## $textbooks[[1]]$website
## [1] "Textbooks.com"
##
##
## $textbooks[[2]]
## $textbooks[[2]]$rank
## [1] 2
##
## $textbooks[[2]]$title
## [1] "Fourier Series"
##
## $textbooks[[2]]$authors
##          one
## "Rajendra Bhatia"
##
## $textbooks[[2]]$ISBN
## [1] "0883857405"
##
## $textbooks[[2]]$publisher
## [1] "Mathematical Association of America"
##
## $textbooks[[2]]$Year
## [1] 2004
##
## $textbooks[[2]]$Book_Type
## [1] "Harback"
##
## $textbooks[[2]]$Used_Price
## [1] 12.21
##
## $textbooks[[2]]$New_Price
## [1] 35
##
## $textbooks[[2]]$website
## [1] "Amazon.com"
##
##
## $textbooks[[3]]
## $textbooks[[3]]$rank
## [1] 3

```

```
##
## $textbooks[[3]]$title
## [1] "Linear Algebra"
##
## $textbooks[[3]]$authors
##           one           two           three
## "Stephen H. Freidberg" "Arnold J. Insel" "Lawrence E. Spence"
##
## $textbooks[[3]]$ISBN
## [1] "0130084514"
##
## $textbooks[[3]]$publisher
## [1] "Prentice Hall"
##
## $textbooks[[3]]$Year
## [1] 2003
##
## $textbooks[[3]]$Book_Type
## [1] "Harback"
##
## $textbooks[[3]]$Used_Price
## [1] 115.21
##
## $textbooks[[3]]$New_Price
## [1] 163.48
##
## $textbooks[[3]]$website
## [1] "Textbooks.com"
```

```
unlist.doc <- sapply(doc.json[[1]], unlist)
json.df <- do.call("rbind.fill", lapply(lapply(unlist.doc,t),data.frame, stringAsFactors = FALSE))
names(json.df) <- c("Title", "Author One", "Author Two", "Author Three", "ISBN", "Publisher", "Year", "Book Type", "Used Price", "New Price", "Website")
json.df
```

```
## Title Author One
## 1 1 Differential Equations with Boundary Value Problems
## 2 2 Fourier Series
## 3 3 Linear Algebra
## Author Two Author Three ISBN Publisher
## 1 John Porking Al Boggess David Arnold 0131862367
## 2 Rajendra Bhatia <NA> <NA> 0883857405
## 3 Stephen H. Freidberg Arnold J. Insel Lawrence E. Spence 0130084514
## Year Book Type Used Price New Price
## 1 Prentice Hall 2006 Harback 105.6
## 2 Mathematical Association of America 2004 Harback 12.21
## 3 Prentice Hall 2003 Harback 115.21
## Website NA NA
## 1 158.4 Textbooks.com FALSE
## 2 35 Amazon.com FALSE
## 3 163.48 Textbooks.com FALSE
```

```
url <- "https://raw.githubusercontent.com/DanielBrooks39/IS607/master/Week%208/books.html"
doc.html <-htmlParse(rawToChar(GET(url)$content))
doc.html
```

```
## <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose
## <html><body>
##     <table>
## <tr>
## <th>Rank</th> <th> Title </th> <th> Author 1 </th> <th> Author 2 </th> <th> Author 3 </th> <th> ISBN
## <tr>
## <td>1</td> <td> Differential Equations With Boundry Value Problems </td> <td> John Porking </td> <td>
## <tr>
## <td>2</td> <td> Fourier Series </td> <td> Rajendra Bhatia </td> <td> </td> <td> </td> <td> 0883857
## <tr>
## <td>3</td> <td> Linear Algebra </td> <td> Stephen H. Freidberg </td> <td> Arnold J. Insel </td> <td>
## </table>
## </body></html>
##
```

```
html.df <- as.data.frame(readHTMLTable(doc.html,header = TRUE))
names(html.df) <- c("Title", "Author One", "Author Two", "Author Three", "ISBN", "Publisher", "Year", "
html.df
```

```
## Title Author One
## 1 1 Differential Equations With Boundry Value Problems
## 2 2 Fourier Series
## 3 3 Linear Algebra
## Author Two Author Three ISBN Publisher
## 1 John Porking Al Boggess David Arnold 0131862367
## 2 Rajendra Bhatia 0883857405
## 3 Stephen H. Freidberg Arnold J. Insel Lawrence E. Spence 0130084514
## Year Book Type Used Price New Price
## 1 Prentince Hall 2006 Hardback $ 105.60
## 2 Mathematical Association of America 2004 Hardback $ 12.21
## 3 Prentince Hall 2003 Hardback $ 115.21
## Website NA
## 1 $ 158.40 Textbooks.com
## 2 $ 35.00 Amazon.com
## 3 $ 163.48 Textbooks.com
```