

Data-607 Week-7 Assignment

Student Name : Sachid Deshmukh

Date : 10/09/2018

- GitHub Location for rmd file
 - GitHub Location for pdf file
 - RPub's location of published file
 - HTML File
 - XML File
 - JSON File
-

1] Library Initialization

```
library(XML)
library(RJSONIO)
library(XML)
```

2] Read HTML File

Read HTML file for Books catalogue

```
books.html <- readHTMLTable("D:/MSDS/MSDSQ1/Data607/Week7/Books.html")
```

Preview R dataframe constructed from HTML file

```
head(books.html)
```

```
## $`NULL`
##
##              Title
## 1          Fluent in 3 Months
## 2 How to Learn Almost Anything in 48 Hours
## 3      CompTIA A+ Complete Study Guide
##
##              Authors      Type Price
## 1          Benny Lewis Audio Book    15
## 2              Tansel Ali    E Book    12
## 3 Quentin Docter, Emmett Dulaney, Toby Skandier    Book    33
##
##      Publisher
## 1 HarperCollins
## 2   Adams Media
## 3        Wiley
```

3] Read XML File

Read XML file for Books catalogue

```
books.xml <- xmlToDataFrame("D:/MSDS/MSDSQ1/Data607/Week7/Books.xml")
```

Preview R dataframe constructed from XML file

```
head(books.xml)
```

```
##                               Title
## 1                        Fluent in 3 Months
## 2 How to Learn Almost Anything in 48 Hours
## 3           CompTIA A+ Complete Study Guide
##                               Authors      Type Price
## 1                        Benny Lewis Audio Book    15
## 2                        Tansel Ali    E Book    12
## 3 Quentin Docter, Emmett Dulaney, Toby Skandier    Book    33
##                               Publisher
## 1 HarperCollins
## 2 Adams Media
## 3 Wiley
```

4] Read JSON File

Read JSON file for Books catalogue

```
books.json <- fromJSON("D:/MSDS/MSDSQ1/Data607/Week7/Books.json")
books.json <- lapply(books.json, function(x) {
  x[sapply(x, is.null)] <- NA
  unlist(x)
})
books.json<-as.data.frame(do.call("cbind", books.json))
```

Preview R dataframe constructed from JSON file

```
head(books.json)
```

```
##                               Title
## Book1                        Fluent in 3 Months
## Book2 How to Learn Almost Anything in 48 Hours
## Book3           CompTIA A+ Complete Study Guide
##                               Authors      Type Price
## Book1                        Benny Lewis Audio Book    15
## Book2                        Tansel Ali    E Book    12
## Book3 Quentin Docter, Emmett Dulaney, Toby Skandier    Book    33
##                               Publisher
## Book1 HarperCollins
## Book2 Adams Media
## Book3 Wiley
```

5] Are the three data frames identical?

Check the column data types of three data frame

```
str(books.html)
```

```
## List of 1
## $ NULL:'data.frame': 3 obs. of 5 variables:
## ..$ Title : Factor w/ 3 levels "CompTIA A+ Complete Study Guide",...: 2 3 1
## ..$ Authors : Factor w/ 3 levels "Benny Lewis",...: 1 3 2
## ..$ Type : Factor w/ 3 levels "Audio Book","Book",...: 1 3 2
## ..$ Price : Factor w/ 3 levels "12","15","33": 2 1 3
## ..$ Publisher: Factor w/ 3 levels "Adams Media",...: 2 1 3
```

```
str(books.xml)
```

```
## 'data.frame': 3 obs. of 5 variables:
## $ Title : Factor w/ 3 levels "CompTIA A+ Complete Study Guide",...: 2 3 1
## $ Authors : Factor w/ 3 levels "Benny Lewis",...: 1 3 2
## $ Type : Factor w/ 3 levels "Audio Book","Book",...: 1 3 2
## $ Price : Factor w/ 3 levels "12","15","33": 2 1 3
## $ Publisher: Factor w/ 3 levels "Adams Media",...: 2 1 3
```

```
str(books.json)
```

```
## 'data.frame': 3 obs. of 5 variables:
## $ Title : Factor w/ 3 levels "CompTIA A+ Complete Study Guide",...: 2 3 1
## ..- attr(*, "names")= chr "Book1" "Book2" "Book3"
## $ Authors : Factor w/ 3 levels "Benny Lewis",...: 1 3 2
## ..- attr(*, "names")= chr "Book1" "Book2" "Book3"
## $ Type : Factor w/ 3 levels "Audio Book","Book",...: 1 3 2
## ..- attr(*, "names")= chr "Book1" "Book2" "Book3"
## $ Price : Factor w/ 3 levels "12","15","33": 2 1 3
## ..- attr(*, "names")= chr "Book1" "Book2" "Book3"
## $ Publisher: Factor w/ 3 levels "Adams Media",...: 2 1 3
## ..- attr(*, "names")= chr "Book1" "Book2" "Book3"
```

1] We can see that column data types of all three data frame are same

2] We can see that no of columns and column orders for all three data frame are same

3] We can see that data frame generated from json file have row names as Book1, Book2 and Book3. The data frames generated from HTML and XML files have numeric value assigned for row names.