DA607 Week 07 Assignment

Dan Fanelli

March 16, 2016

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].

```
library(RCurl)
library(XML)
library(knitr)
library(jsonlite)
library(plyr)
library(stringr)
```

Html Table to Data Frame:

```
#html.theurl <- "https://raw.githubusercontent.com/fandang/DA607/master/Wk08/DanF_DA607_Week_08_Books.h
html.theurl <- "DanF_DA607_Week_08_Books.html"
html.tables <- readHTMLTable(html.theurl)
html.n.rows <- unlist(lapply(html.tables, function(t) dim(t)[1]))
html.data <- html.tables[[which.max(html.n.rows)]]
html.df <- as.data.frame(html.data, stringsAsFactors=TRUE)
colnames(html.df) <- c("title", "author1", "author2", "amazon_stars",   "num_pages",    "summary")
html.df[is.na(html.df)] <- ''
html.df$authors <- paste(html.df$author1,html.df$author2,sep=", ")
html.df <- html.df[,c(1,7,4,5,6)]
kable(html.df)
```

title	authors	amazon_star
The Big Short: Inside the Doomsday Machine	Michael Lewis,	4.4
The One Year Uncommon Life Daily Challenge	Tony Dungy, Nathan Whitaker	4.8
The Signal and the Noise: Why So Many Predictions Fail-but Some Don't	Nate Silver,	4.3

XML to Data Frame:

```
#xml.theurl <- "https://raw.githubusercontent.com/fandang/DA607/master/Wk08/DanF_DA607_Week_08_Books.xm
xml.theurl <- "DanF_DA607_Week_08_Books.xml"
xml.doc <- xmlParse(xml.theurl)
xml.df <- xmlToDataFrame(nodes = xmlChildren(xmlRoot(xml.doc)[["books"]]), stringsAsFactors=TRUE)
xml.df <- xml.df[, c(1,4,6,2,3,5)]
xml.df[is.na(xml.df)] <- ''
xml.df$authors <- paste(xml.df$author1,xml.df$author2,sep=", ")
xml.df <- xml.df[,c(1,7,4,5,6)]

kable(xml.df)
```

title	authors	amazon_star
The Big Short: Inside the Doomsday Machine	Michael Lewis,	4.4
The One Year Uncommon Life Daily Challenge	Tony Dungy, Nathan Whitaker	4.8
The Signal and the Noise: Why So Many Predictions Fail-but Some Don't	Nate Silver,	4.3

JSON to Data Frame:

```
#json.theurl <- "https://raw.githubusercontent.com/fandang/DA607/master/Wk08/DanF_DA607_Week_08_Books.j
json.theurl <- "DanF_DA607_Week_08_Books.json"
json.document <- fromJSON(json.theurl, simplifyVector = TRUE)
json.df <- as.data.frame(json.document)
json.df <- json.df[, c(1,4,2,3,5)]
colnames(json.df) <- c("title", "authors",  "amazon_stars", "num_pages",    "summary")

kable(json.df)
```

title	authors	amazon_star
The Big Short: Inside the Doomsday Machine	Michael Lewis	4.4
The One Year Uncommon Life Daily Challenge	Tony Dungy, Nathan Whitaker	4.8
The Signal and the Noise: Why So Many Predictions Fail-but Some Don't	Nate Silver	4.3

Are they Equal? Not exactly, but probably with a little munging and string trimming they would all be...

```
html.df == xml.df
```

```
##      title authors amazon_stars num_pages summary
## [1,]  TRUE    TRUE         TRUE      TRUE    TRUE
## [2,]  TRUE    TRUE         TRUE      TRUE    TRUE
## [3,]  TRUE    TRUE         TRUE      TRUE    TRUE
```

```
html.df == json.df
```

```
##      title authors amazon_stars num_pages summary
## [1,]  TRUE   FALSE         TRUE      TRUE    TRUE
## [2,]  TRUE   FALSE         TRUE      TRUE    TRUE
## [3,]  TRUE   FALSE         TRUE      TRUE    TRUE
```