DATA 607 Assignment Week 7

Samuel I Kigamba

October 13, 2019

R Markdown

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 youve selected about these three books, and separately create three files which store the books 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, Id prefer that you create each of these files by hand unless youre 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].

Install required packages

Make sure to install all the required packages, I have commented them out since they are already installed in my case.

```r
#install.packages('RCurl')
#install.packages('rjson')
#install.packages('XML')
#install.packages('selectr')
#install.packages('ROAuth')
#install.packages('httr')
#install.packages('rvest')
#install.packages('stringr')
#install.packages('JSONIO')
#install.packages('jsonlite')
```

Load all the required libraries, set message = FALSE and warning = FALSE inside the R-code to prevent Rmd from printing out the contents of the load library.

```r
library(jsonlite)
library(knitr)
library(RJSONIO)
library(tidyverse)
library(plyr)
library(dplyr)
library(XML)
library(xml2)
#gc() #Clean up the memory
```

HTML file parsing

Following the insturction above, i have created a html file and uploaded it to Github for ease of accessibility.

Lets load the file into R and parse through it to create a data frame and display its contents.

```
#Follow this Github link to view the raw contents of the file.
books_html <- "https://raw.githubusercontent.com/igukusamuel/DATA-607-Week-7-Assignment/master/MyFavour

download.file(books_html, destfile = "~/MyFavouriteBooks.html")

books_html <- file.path("MyFavouriteBooks.html")

#Lets use htmlParse() for parsing the file.

books_html <- htmlParse(books_html)

#We then use readHTMLTable to read through the html file and create a table

books_html_tbl <- readHTMLTable(books_html, stringAsFactors = FALSE)

dframe.html <- books_html_tbl[[1]] %>% tbl_df()

dframe.html
```

```
## # A tibble: 3 x 4
##   BookTitle                        AuthorS                Edition    Year
##   <fct>                            <fct>                  <fct>      <fct>
## 1 Options, Futures, and Other Der~ John C. Hull           10th Edit~ 2018
## 2 Fixed Income Securities          Bruce Tuckman, Angel S~ 3rd Editi~ 2012
## 3 Doing Bayesian Data Analysis     John K. Kruschke       2nd Editi~ 2015
```

JSON file parsing

Following the insturction above, i have created a json file and uploaded it to Github for ease of accessibility.

Use isValidJSON() to check whether the JSON file is valid before parsing to avoid brakages in the process.

```
isValidJSON("https://raw.githubusercontent.com/igukusamuel/DATA-607-Week-7-Assignment/master/MyFavourit
```

```
## [1] TRUE
```

Lets load the file into R and parse through it to create a data frame and display its contents.

```
#Follow this Github link to view the raw contents of the file.
MyBooks_json <- "https://raw.githubusercontent.com/igukusamuel/DATA-607-Week-7-Assignment/master/MyFavou

download.file(MyBooks_json, destfile = "~/MyFavouriteBooks.json")
MyBooks_json <- file.path("MyFavouriteBooks.json")
MyBooks_JSON <- fromJSON(content = MyBooks_json)

#Save the data into a data frame

MyBooks_JSON_df = as.data.frame(MyBooks_JSON)

MyBooks_JSON_df
```

```
##                         My_Books.BookTitle My_Books.AuthorS
## 1 Options, Futures, and Other Derivatives     John C. Hull
##   My_Books.edition My_Books.Year    My_Books.BookTitle.1
## 1     10th Edition          2018 Fixed Income Securities
##             My_Books.AuthorS.1 My_Books.edition.1 My_Books.Year.1
## 1 Bruce Tuckman, Angel Serrat        3rd Edition            2012
##            My_Books.BookTitle.2 My_Books.AuthorS.2 My_Books.edition.2
## 1 Doing Bayesian Data Analysis   John K. Kruschke        2nd Edition
##   My_Books.Year.2
## 1            2015
```

XML file parsing

Following the insturction above, i have created a XML file and uploaded it to Github for ease of accessibility.

Lets load the file into R and parse through it to create a data frame and display its contents.

```
#Follow this Github link to view the raw contents of the file.
MyBooks_XML <- "https://raw.githubusercontent.com/igukusamuel/DATA-607-Week-7-Assignment/master/MyFavou

download.file(MyBooks_XML, destfile = "~/MyFavouriteBooks.xml")
MyBooks_XML <- file.path("MyFavouriteBooks.xml")
MyBooks_XML <- xmlParse(MyBooks_XML)

MyBooks_XML <- xmlRoot(MyBooks_XML)

dframe.xml <- xmlToDataFrame(MyBooks_XML, stringsAsFactors = F) %>% tbl_df()
dframe.xml
```

```
## # A tibble: 3 x 4
##   BookTitle                Authors                          Edition   Year
##   <chr>                    <chr>                            <chr>     <chr>
## 1 Options, Futures, and O~ " Main_Author=\"John C. Hull\" " 10th Edi~ 2018
## 2 Fixed Income Securities  " Main_Author=\"Bruce Tuckman\"~ 3rd Edit~ 2012
## 3 Doing Bayesian Data Ana~ " Main_Author=\"John K. Kruschk~ 2nd Edit~ 2015
```