ds607_week7_JagdishChhabria

Jagdish Chhabria

March 16, 2019

```
# This is the code chunk for reading in the HTML file into R as a data frame
library(XML)
library(RCurl)
```

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

```
library(rvest)
```

```
## Loading required package: xml2
```

```
## 
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:XML':
## 
##     xml
```

```
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
#fileurl<-"C:\\Jagdish\\Masters Programs\\CUNY\\Data 607 Data Acquisition and Management\\Week7\\books.
fileurl<-getURL("https://raw.githubusercontent.com/Jagdish16/jagdish_r_repo/master/DATA607/Week7/books.

#books1<-readHTMLTable(doc=fileurl, trim=T, as.data.frame=T, header=T, Encoding="windows-1252")
#books2<-htmlParse(file=fileurl, encoding = "windows-1252", as.data.frame=T)
#books2<-read_html(fileurl, encoding = "UTF-8")
#tables<-html_nodes(books2, "table")
# table1 <- html_table(tables, fill = TRUE)

booksdf <- as.data.frame(read_html(fileurl) %>% html_table(fill=TRUE))
booksdf
```

```
##   ID                                  Title           ISBN Edition
## 1  1 Options, Futures and other Derivatives 978-0136015864 seventh
## 2  2     Advanced Financial Risk Management 978-1118278543  second
## 3  3 Introduction to Fixed Income Analytics 978-0470572139  second
##                  Author1        Author2     Author3     Publisher
## 1           John C. Hull                            Prentice Hall
## 2 Donald R. Van Deventer     Kenji Imai Mark Mesler         Wiley
## 3       Frank J. Fabozzi Steven V. Mann                     Wiley
##   Latest_Publish_Year
## 1                2008
## 2                2013
## 3                2010
```

```r
# This is the code chunk for reading in the XML file into R as a data frame
library(XML)
library(RCurl)
library(httr)
```

```
## Warning: package 'httr' was built under R version 3.5.3
```

```r
#fileurl<-"C:\\Jagdish\\Masters Programs\\CUNY\\Data 607 Data Acquisition and Management\\Week7\\books1

fileurl<-getURL("https://raw.githubusercontent.com/Jagdish16/jagdish_r_repo/master/DATA607/Week7/books1

books<-xmlParse(file=fileurl)
books_df<-xmlToDataFrame(books, stringsAsFactors = FALSE)
books_df
```

```
##                                    title           isbn edition
## 1 Options, Futures and other Derivatives 978-0136015864 seventh
## 2     Advanced Financial Risk Management 978-1118278543  second
## 3 Introduction to Fixed Income Analytics 978-0470572139  second
##                    author1     publisher latest_publish_year
## 1            John C. Hull  Prentice Hall                2008
## 2  Donald R. Van Deventer          Wiley                2013
## 3        Frank J. Fabozzi          Wiley                2010
##            author2       author3
## 1             <NA>          <NA>
## 2      Kenji Imai   Mark Mesler
## 3  Steven V. Mann           <NA>
```

```r
# This is the code chunk for reading in the JSON file into R as a data frame
library(jsonlite)
library(dplyr)
library(RCurl)

#fileurl<-"C:\\Jagdish\\Masters Programs\\CUNY\\Data 607 Data Acquisition and Management\\Week7\\books5
fileurl<-getURL("https://raw.githubusercontent.com/Jagdish16/jagdish_r_repo/master/DATA607/Week7/books5

books.df<-fromJSON(fileurl) %>% as.data.frame
books.df
```

```
##   finance_books.book.id               finance_books.book.title
## 1                     1 Options, Futures and other Derivatives
## 2                     2     Advanced Financial Risk Management
## 3                     3 Introduction to Fixed Income Analytics
##   finance_books.book.isbn finance_books.book.edition
## 1          978-0136015864                    seventh
## 2          978-1118278543                     second
## 3          978-0470572139                     second
##   finance_books.book.author1 finance_books.book.author2
## 1               John C. Hull
## 2     Donald R. Van Deventer                  Kenji Ima
## 3           Frank J. Fabozzi             Steven V. Mann
##   finance_books.book.author3 finance_books.book.publisher
## 1                                           Prentice Hall
## 2                Mark Mesler                        Wiley
## 3                                                   Wiley
##   finance_books.book.latest_publish_year
## 1                                   2008
## 2                                   2013
## 3                                   2010
```

The 3 dataframes are not completely identical. There are minor differences such as: 1) the “id” column missing in the dataframe converted from the html file; 2) capitalization of column names is different across the datafarmes created from html versus xml, while the column names for the dataframe created from the json file are pre-fixed by the name of the first object in the json file. Presumably these differences are on account of how the different handling functions work under the hood.