Week 7_Assignment_Chunjie_Nan

Chunjie Nan

10/7/2021

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?

HTML

```
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
```

```
library(XML)
htmlbook <- readLines("https://raw.githubusercontent.com/nancunjie4560/Data607/master/HTML") %>%
  htmlParse() %>%
  readHTMLTable()%>%
  data.frame()

colnames(htmlbook)<-c('Title','Author','Release Year','Price','Amazon Rating','Language')
htmlbook
```

```
##                                Title                      Author Release Year
## 1       Practical Data Science with R       Nina Zumel,John Mount     Dec 2019
## 2                Linear Models with R            Julian J.Faraway     Jul 2014
## 3 Text Mining with R: A Tidy Approach Julia Silge, David Robinson     Jul 2017
##    Price Amazon Rating Language
## 1 $34.99         4.7/5  English
## 2 $46.11         4.7/5  English
## 3 $21.49         4.5/5  English
```

XML

```
xmlbook<-readLines("https://raw.githubusercontent.com/nancunjie4560/Data607/master/xml")%>%
  xmlParse()%>%
  xmlToDataFrame()

xmlbook
```

```
##                                 title                       author     year
## 1       Practical Data Science with R       Nina Zumel,John Mount Dec 2019
## 2                Linear Models with R            Julian J.Faraway Jul 2014
## 3 Text Mining with R: A Tidy Approach Julia Silge, David Robinson Jul 2017
##    price rating language
## 1 $34.99 4.75/5  English
## 2 $46.11 4.75/5  English
## 3 $21.49  4.5/5  English
```

JSON

```
library(rjson)
library(RCurl)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
```

```
## The following objects are masked from 'package:rjson':
##
##     fromJSON, toJSON
```

```
json<-readLines("https://raw.githubusercontent.com/nancunjie4560/Data607/master/json")
jsonbook<-paste(json, collapse = '')%>%
  fromJSON()
jsonbook
```

```
##                                 title                       author     year
## 1      Practical Data Science with R      Nina Zumel, John Mount Dec 2019
## 2               Linear Models with R            Julian J.Faraway Jul 2014
## 3 Text Mining with R:A Tidy Approach Julia Silge, David Robinson Jul 2017
##    price rating language
## 1 $34.99 4.75/5  English
## 2 $46.11 4.75/5  English
## 3 $21.49  4.5/5  English
```

In conclusion, the dataframes are identical for HTML, XML, and JSON. For HTML, I had to rename the variables, and the XML is very straight forward. However, for running JSON() function, need to load the jsonlite(), not the JSONIO(). Because JSONIO() converts a list to a list; but we want a data frame, not the list. Therefore, run fromJSON() with jsonlite() library is the correct option.