Homework Assignemnt Week7

Author1 :Samriti Malhotra
Author2 :Vishal Arora

3/14/2019

Introduction

In this assignment, data about favorite books are stored in three (3) different file formats –>. HTML, XML, and JSON. these are accessed from github. Book Files Github.

The task was to parse these files and create dataframe.

R packages that are referenced:

```
# Packages for working with HTML, XML & JSON

library(XML)
library(RJSONIO)
library(jsonlite)
library(dplyr)
library(RCurl)
library(kableExtra)
```

XML File

XML is a extensible markup language. It is a data description language used for describing data. *XML File*

```
<books>
 <book copyright='2008' lang='eng' title='breaking dawn' publisher='Brown and Company' isbn='978-0-316-
  <author> Stephenie Meyer.</author>
  </book>

 <book copyright='2004' lang='eng'  title='Guardian of the Horizon' publisher='HarperCollins' isbn='0-0
  <author>Elizabeth Peters.</author>
  </book>
 <book copyright='2015' lang='eng' title='Extreme Ownership' publisher='St.Martins Press' isbn='978-1-2
    <author>
    <author1>Jocko Willink. </author1>
    <author2> Leif Babin.</author2>
    </author>
    </book>
</books>
```

Load XML File

Step 1: Use RCurl package function ::getURL‘ to download the URL for the raw data into R.

```
xml_file <- getURL("https://raw.githubusercontent.com/Vishal0229/Data607/master/Week7/books1.xml")
```

Step 2: Parse xml file , using xmlparse function from XML library. In Next step find the root element of the document and then applying the xpathSApply function to get the attributes of various node element.

```
doc_xml <- xmlParse(xml_file)
root <- xmlRoot(doc_xml)
title <- xpathSApply(doc_xml, "//book", xmlGetAttr, "title")
publisher <- xpathSApply(doc_xml, "//book", xmlGetAttr, "publisher")
isbn <- xpathSApply(doc_xml, "//book", xmlGetAttr, "isbn")
genre <- xpathSApply(doc_xml, "//book", xmlGetAttr, "genre")
copyright <- xpathSApply(doc_xml, "//book", xmlGetAttr, "copyright")
lang <- xpathSApply(doc_xml, "//book", xmlGetAttr, "lang")
```

Step 3 : Convert into dataframe, and also using rbind on root element 3 i.e. Actor to get multiple values inside Actor element.

```
xmldf <- data.frame(lang = unlist(lang),
                    timestamp = unlist(copyright),
                    title = unlist(title),
                    (rbind(xmlSApply(root[[1]], xmlValue),xmlSApply(root[[2]], xmlValue),xmlSApply(root
                    publisher = unlist(publisher),
                    isbn = unlist(isbn),
                    genre = unlist(genre))
```

The Output

lang	timestamp	title	author	publisher	isbn
eng	2008	breaking dawn	Stephenie Meyer.	Brown and Company	98723
eng	2004	Guardian of the Horizon	Elizabeth Peters.	HarperCollins	00662
eng	2015	Extreme Ownership	Jocko Willink. Leif Babin.	St.Martins Press	97812

JSON File

Another standard for data storage and interchange on the Web is the JavaScript Object Notation, abbreviated JSON. JSON is an increasingly popular alternative to XML for data exchange purposes that comes with some preferable features.

JSON File

```
 {
    "copyright": "2008",
    "lang": "eng",
    "title": "breaking dawn",
    "author": "Stephenie Meyer",
    "publisher":"Brown and Company",
    "isbn": "978-0-316-06792-8" ,
    "genre":"Paranormal romance"
},
{
    "copyright": "2004",
    "lang": "eng",
    "title": "Guardian of the Horizon",
    "author": "Elizabeth Peters",
    "publisher":"HarperCollins",
```

```
    "isbn": "0-06-621471-8" ,
    "genre":"Mystery"
},
{
    "copyright": "2004",
    "lang": "eng",
    "title": "Extreme Ownership",
    "author": ["Jocko Willink" ,"Leif Babin"],
    "publisher":"St.Martin'sPress",
    "isbn": "978-1-250-06705" ,
    "genre":"Biography"
}
```

Load JSON File

Step 1: Use getURL to access file from Github

```
json_file <- getURL("https://raw.githubusercontent.com/Vishal0229/Data607/master/Week7/book.json")
```

Step 2: Parse and manipulating the json file. load the json file into R using jsonlite function and then extracting the value of actor name/pair value and then combining all the values back again to form a data frame using c and using paste function concatenate Actor values.

```
doc_json <- jsonlite::fromJSON("book.json")
df <- unlist(doc_json[,4])
doc_json$author <- c(df[1],df[2],paste(df[3],df[4],sep="."))
```

The Output

copyright	lang	title	author	publisher	isbn
2008	eng	breaking dawn	Stephenie Meyer	Brown and Company	9872341
2004	eng	Guardian of the Horizon	Elizabeth Peters	HarperCollins	0066214
2004	eng	Extreme Ownership	Jocko Willink.Leif Babin	St.Martin'sPress	9781250

HTML File

An HTML(Hyper Test Markup Language) file is basically nothing but plain text-it can be opened and edited with any text editor. What makes HTML so powerful is its marked up structure.

HTML File

```
<html>
<head></head>

<body>

<table>

 <tr>

  <td>copyright</td>
```

```
   <td>lang</td>
   <td>title </td>
   <td>author</td>
  <td> publisher</td>
   <td>isbn</td>
   <td> genre</td>
</tr>

<tr>
  <td>2008</td>
   <td>eng</td>
   <td>breaking dawn</td>
   <td> Stephenie Meyer.</td>
  <td> Brown and Company </td>
   <td>978-0-316-06792-8</td>
   <td>Paranormal romance</td>
</tr>
<tr>
  <td>2004</td>
   <td>eng</td>
   <td>Guardian of the Horizon</td>
   <td>Elizabeth Peters.</td>
  <td>HarperCollins</td>
   <td>0-06-621471-8</td>
   <td>Mystery</td>
</tr>
<tr>
```

```
  <td>2015</td>

   <td>eng</td>

   <td>Extreme Ownership</td>

   <td>Jocko Willink. Leif Babin.</td>

  <td> St.Martin'sPress</td>

   <td>978-1-250-06705</td>

   <td>Biography</td>

</tr>

</body>
</html>
```

Load HTML File

Step 1: Use getURL to access file from Github, and using the htmlParse method to load html doc into R memory in tree structure.

```
html_file <- getURL("https://raw.githubusercontent.com/Vishal0229/Data607/master/Week7/books.html")
doc2 <- htmlParse(html_file)
```

Step 2: Parse and manipulating the html file, using getNodeSet method on the parent *table* node and then reading the internal nodes into dataframe onject using readHTMLTable

```
tableNodes <- getNodeSet(doc2, "//table")
myTable <- readHTMLTable(tableNodes[[1]] )
```

The Output

copyright	lang	title	author	publisher	isbn
2008	eng	breaking dawn	Stephenie Meyer.	Brown and Company	987234
2004	eng	Guardian of the Horizon	Elizabeth Peters.	HarperCollins	006621
2015	eng	Extreme Ownership	Jocko Willink. Leif Babin.	St.Martin'sPress	978125

Summary

Using R libraries we can load HTML,XML and JSOn files into R and depending upon the need we can extract the node values from HTML/XML and from JSON we can extract the name/pair values for manipulation and getting the data ready for further analysis.