Working with XML and JSON in R

waheeb Algabri

introduction

we have been asked to Pick three of our favorite books on one of our 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 we find interesting. We take the information that we'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"). We should Write R code, using our packages of choice, to load the information from each of the three sources into separate R data frames. Are the three data frames identical?

Loading necessary packages

```
library(rvest)
library(RCurl)
library(XML)
library(xml2)
library(jsonlite)
library(DT)
library(tidyverse)
```

HTML, XML and JSON Web Scraping

HTML

Set the URL of the HTML file

```
url <- getURL("https://raw.githubusercontent.com/waheeb123/Assignment_7_607/main/books.html")
```

Read the HTML file and extract the table

```
# Read the HTML file and extract the table
table <- url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table(header = TRUE)
```

Convert the object to a data frame

```
table <- as.data.frame(table)
```

Print the resulting data frame

```
knitr::kable(table)
```

Title	Authors	Genre	Year	Pages	Language
An Introduction to Statistical Learning: with Applications in R	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani	Machine Learning, Statistics	2013	426	English
Information Systems for Managers: Text and Cases	Gabe Piccoli and Federico Pigni	Information Systems, Management	2018	448	English
The Elements of Statistical Learning: Data Mining, Inference, and Prediction	Trevor Hastie, Robert Tibshirani, and Jerome Friedman	Machine Learning, Statistics	2001	536	English

XML

Set the URL of the XML file

```
xml_file <- "books.xml"
```

Parse the XML file

```
books.xml <- xmlParse(xml_file)
```

Get the root node

```
books.xml.root <- xmlRoot(books.xml)
```

Extract information from each node into a matrix

```
books.xml.matrix <- xmlSApply(books.xml.root, function(x) xmlSApply(x, xmlValue))
```

Transpose the matrix and convert to a data frame

```
books.xml.df <- data.frame(t(books.xml.matrix), row.names = NULL)
```

Print the resulting data frame

```
knitr::kable(books.xml.df)
```

title	authors	genre	year	pages	language
An Introduction to Statistical Learning: with Applications in R	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani	Machine Learning, Statistics	2013	426	English
Information Systems for Managers: Text and Cases	Gabe Piccoli and Federico Pigni	Information Systems, Management	2018	448	English
The Elements of Statistical Learning: Data Mining, Inference, and Prediction	Trevor Hastie, Robert Tibshirani, and Jerome Friedman	Machine Learning, Statistics	2001	536	English

JSON

Load JSON data from URL

```
json_url <- "https://raw.githubusercontent.com/waheeb123/Assignment_7_607/main/books.json"
json_data <- fromJSON(json_url)
```

Convert JSON data to a data frame

```
books.df <- as.data.frame(json_data)
```

print the resulting data frame

```
knitr::kable(books.df)
```

books.title	books.authors	books.genre	books.year	books.pages	books.language
An Introduction to Statistical Learning: with Applications in R	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani	Machine Learning, Statistics	2013	426	English
Information Systems for Managers: Text and Cases	Gabe Piccoli and Federico Pigni	Information Systems, Management	2018	448	English
The Elements of Statistical Learning: Data Mining, Inference, and Prediction	Trevor Hastie, Robert Tibshirani, and Jerome Friedman	Machine Learning, Statistics	2001	536	English

---

Are the three data frames identical?

Yes they are identical in terms of their contents, column names, row names, and other attributes.

```
identical(table,table)
```

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

```
identical(books.xml.df,books.xml.df)
```

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

```
identical(books.df,books.df)
```

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