

Week 7 Assignment - Working with XML and JSON in R

Peter Gatica

0320/2021

```
library(XML)
library(RJSONIO)
```

```
filename <- "/Users/Audiorunner13/CUNY MSDS Course Work/DATA607 Spring 2021/Week7/Homework7/Data/Fave_H
fave_books_xml <- xmlParse(filename)
fave_books_xml
```

```
## <?xml version="1.0" encoding="UTF-8"?>
## <books>
##   <book pages="398" lang="eng">
##     <title>Faith of My Fathers</title>
##     <author_1>
##       <first_name>John</first_name>
##       <last_name>McCain</last_name>
##       <educ>US Naval Academy</educ>
##       <career>Republican Senator</career>
##       <misc>Naval Aviator</misc>
##     </author_1>
##     <isbn-13>978-0-3995-9089-4</isbn-13>
##     <publisher>Random House Publishing Group</publisher>
##     <pub_year>1998</pub_year>
##   </book>
##   <book pages="578" lang="eng">
##     <title>Indianapolis</title>
##     <author_1>
##       <first_name>Lynn</first_name>
##       <last_name>Vincent</last_name>
##       <educ>US Navy Veteran</educ>
##       <career>Journalist</career>
##       <misc>Amazon Best of 2018</misc>
##     </author_1>
##     <author_2>
##       <first_name>Sara</first_name>
##       <last_name>Vladic</last_name>
##       <educ>Pepperdine University</educ>
##       <career>Documentary Filmmaker</career>
##       <misc>San Diego Film Award</misc>
##     </author_2>
##     <isbn-13>978-1-5011-3594-1</isbn-13>
##     <publisher>Simon and Schuster Paperbacks</publisher>
##     <pub_year>2019</pub_year>
```

```
## </book>
## <book pages="480" lang="eng">
##   <title>We Were Soldiers Once...and Young</title>
##   <author_1>
##     <first_name>Hal</first_name>
##     <last_name>Moore</last_name>
##     <educ>West Point</educ>
##     <career>Lt General, US Army</career>
##     <misc>Vietnam Veteran</misc>
##   </author_1>
##   <author_2>
##     <first_name>Joseph</first_name>
##     <last_name>Galloway</last_name>
##     <educ>Bronze Star V-Device</educ>
##     <career>Author, Journalist</career>
##     <misc>Decorated non-combatant</misc>
##   </author_2>
##   <isbn-13>978-0-3454-7581-7</isbn-13>
##   <publisher>Random House Publishing Group</publisher>
##   <pub_year>2004</pub_year>
## </book>
## </books>
##
```

```
class(fave_books_xml)
```

```
## [1] "XMLInternalDocument" "XMLAbstractDocument"
```

```
fave_books_xml_1 <- xmlParse(filename,useInternalNodes = FALSE)
fave_books_xml_1
```

```
## $doc
## $file
## [1] "/Users/Audiorunner13/CUNY MSDS Course Work/DATA607 Spring 2021/Week7/Homework7/Data/Fave_Hist_B
##
## $version
## [1] "1.0"
##
## $children
## $children$books
## <books>
##   <book pages="398" lang="eng">
##     <title>Faith of My Fathers</title>
##     <author_1>
##       <first_name>John</first_name>
##       <last_name>McCain</last_name>
##       <educ>US Naval Academy</educ>
##       <career>Republican Senator</career>
##       <misc>Naval Aviator</misc>
##     </author_1>
##     <isbn-13>978-0-3995-9089-4</isbn-13>
##     <publisher>Random House Publishing Group</publisher>
##     <pub_year>1998</pub_year>
```

```

## </book>
## <book pages="578" lang="eng">
##   <title>Indianapolis</title>
##   <author_1>
##     <first_name>Lynn</first_name>
##     <last_name>Vincent</last_name>
##     <educ>US Navy Veteran</educ>
##     <career>Journalist</career>
##     <misc>Amazon Best of 2018</misc>
##   </author_1>
##   <author_2>
##     <first_name>Sara</first_name>
##     <last_name>Vladic</last_name>
##     <educ>Pepperdine University</educ>
##     <career>Documentary Filmmaker</career>
##     <misc>San Diego Film Award</misc>
##   </author_2>
##   <isbn-13>978-1-5011-3594-1</isbn-13>
##   <publisher>Simon and Schuster Paperbacks</publisher>
##   <pub_year>2019</pub_year>
## </book>
## <book pages="480" lang="eng">
##   <title>We Were Soldiers Once...and Young</title>
##   <author_1>
##     <first_name>Hal</first_name>
##     <last_name>Moore</last_name>
##     <educ>West Point</educ>
##     <career>Lt General, US Army</career>
##     <misc>Vietnam Veteran</misc>
##   </author_1>
##   <author_2>
##     <first_name>Joseph</first_name>
##     <last_name>Galloway</last_name>
##     <educ>Bronze Star V-Device</educ>
##     <career>Author, Journalist</career>
##     <misc>Decorated non-combatant</misc>
##   </author_2>
##   <isbn-13>978-0-3454-7581-7</isbn-13>
##   <publisher>Random House Publishing Group</publisher>
##   <pub_year>2004</pub_year>
## </book>
## </books>
##
## attr(,"class")
## [1] "XMLDocumentContent"
##
## $dtd
## $external
## NULL
##
## $internal
## NULL
##

```

```
## attr("class")
## [1] "DTDList"
##
## attr("class")
## [1] "XMLDocument"          "XMLAbstractDocument"
```

```
class(fave_books_xml)
```

```
## [1] "XMLInternalDocument" "XMLAbstractDocument"
```

```
filename <- "/Users/Audiorunner13/CUNY MSDS Course Work/DATA607 Spring 2021/Week7/Homework7/Data/Fave_H_
fave_books_html <- xmlParse(filename,isHTML = TRUE)
```

```
## htmlParseEntityRef: no name
```

```
fave_books_html
```

```
## <!DOCTYPE html>
## <html><body>
##
## <h1>Favorite Books</h1>
## <p>World War II</p>
## <dl>
## <dt>Title: Faith of My Fathers</dt>
##   <dd>Author: John McCain</dd>
##   <dd>Education: US Naval Academy</dd>
##   <dd>Career: Republican Senator</dd>
##   <dd>Miscellaneous: Naval Aviator</dd>
##   <dd>ISBN-13: 978-0-3995-9089-4</dd>
##   <dd>Publisher: Random House Publishing Group</dd>
##   <dd>Publication Year: 1998<dd>
##   <dd>Pages: 398<dd>
##   <dd>Language: English<dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dl>
## <dl>
## <dt>Title: Indianapolis</dt>
##   <dd>Author: Lynn Vincent</dd>
##   <dd>Education: US Navy Veteran</dd>
##   <dd>Career: Journalist</dd>
##   <dd>Miscellaneous: Amazon Best of 2018</dd>
##   <dd>Author: Sara Vladic</dd>
##   <dd>Education: Pepperdine University</dd>
##   <dd>Career: Documentary Filmmaker</dd>
##   <dd>Miscellaneous: San Diego Film Award</dd>
##   <dd>ISBN-13: 978-1-5011-3594-1</dd>
##   <dd>Publisher: Simon & Schuster Paperbacks</dd>
```

```
##      <dd>Publication Year: 2019<dd>
##      <dd>Pages: 578<dd>
##      <dd>Language: English<dd>
##    </dd>
##  </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dl>
## <dl>
## <dt>Title: We Were Soldiers Once...and Young</dt>
##      <dd>Author: Hal Moore</dd>
##      <dd>Education: West Point</dd>
##      <dd>Career: Lt General, US Army</dd>
##      <dd>Miscellaneous: Vietnam Veteran</dd>
##      <dd>Author: Joseph Galloway</dd>
##      <dd>Education: Bronze Star V-Device</dd>
##      <dd>Career: Author, Journalist</dd>
##      <dd>Miscellaneous: Decorated non-combatant</dd>
##      <dd>ISBN-13: 978-0-3454-7581-7</dd>
##      <dd>Publisher: Random House Publishing Group</dd>
##      <dd>Publication Year: 2004</dd>
##      <dd>Pages: 480</dd>
##      <dd>Language: English</dd>
##    </dl>
## </body></html>
##
```

```
class(fave_books_html)
```

```
## [1] "HTMLInternalDocument" "HTMLInternalDocument" "XMLInternalDocument"
## [4] "XMLAbstractDocument"
```

```
fave_books_html_1 <- htmlParse(filename, isHTML = TRUE)
fave_books_html_1
```

```
## <!DOCTYPE html>
## <html><body>
##
## <h1>Favorite Books</h1>
## <p>World War II</p>
## <dl>
## <dt>Title: Faith of My Fathers</dt>
##      <dd>Author: John McCain</dd>
##      <dd>Education: US Naval Academy</dd>
##      <dd>Career: Republican Senator</dd>
##      <dd>Miscellaneous: Naval Aviator</dd>
##      <dd>ISBN-13: 978-0-3995-9089-4</dd>
##      <dd>Publisher: Random House Publishing Group</dd>
##      <dd>Publication Year: 1998<dd>
##      <dd>Pages: 398<dd>
##      <dd>Language: English<dd>
```

```

## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dl>
## <dl>
## <dt>Title: Indianapolis</dt>
## <dd>Author: Lynn Vincent</dd>
## <dd>Education: US Navy Veteran</dd>
## <dd>Career: Journalist</dd>
## <dd>Miscellaneous: Amazon Best of 2018</dd>
## <dd>Author: Sara Vladic</dd>
## <dd>Education: Pepperdine University</dd>
## <dd>Career: Documentary Filmmaker</dd>
## <dd>Miscellaneous: San Diego Film Award</dd>
## <dd>ISBN-13: 978-1-5011-3594-1</dd>
## <dd>Publisher: Simon & Schuster Paperbacks</dd>
## <dd>Publication Year: 2019</dd>
## <dd>Pages: 578</dd>
## <dd>Language: English</dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dd>
## </dl>
## <dl>
## <dt>Title: We Were Soldiers Once...and Young</dt>
## <dd>Author: Hal Moore</dd>
## <dd>Education: West Point</dd>
## <dd>Career: Lt General, US Army</dd>
## <dd>Miscellaneous: Vietnam Veteran</dd>
## <dd>Author: Joseph Galloway</dd>
## <dd>Education: Bronze Star V-Device</dd>
## <dd>Career: Author, Journalist</dd>
## <dd>Miscellaneous: Decorated non-combatant</dd>
## <dd>ISBN-13: 978-0-3454-7581-7</dd>
## <dd>Publisher: Random House Publishing Group</dd>
## <dd>Publication Year: 2004</dd>
## <dd>Pages: 480</dd>
## <dd>Language: English</dd>
## </dl>
## </body></html>
##

```

```
class(fave_books_html_1)
```

```

## [1] "HTMLInternalDocument" "HTMLInternalDocument" "XMLInternalDocument"
## [4] "XMLAbstractDocument"

```



```

##      <dd>
##      Pages: 578
##      <dd>
##      <dd>
##      Language: English
##      <dd/>
##      </dd>
##      </dd>
##      </dd>
##      </dd>
##      </dl>
##      <dl>
##      <dt>Title: We Were Soldiers Once...and Young</dt>
##      <dd>Author: Hal Moore</dd>
##      <dd>Education: West Point</dd>
##      <dd>Career: Lt General, US Army</dd>
##      <dd>Miscellaneous: Vietnam Veteran</dd>
##      <dd>Author: Joseph Galloway</dd>
##      <dd>Education: Bronze Star V-Device</dd>
##      <dd>Career: Author, Journalist</dd>
##      <dd>Miscellaneous: Decorated non-combatant</dd>
##      <dd>ISBN-13: 978-0-3454-7581-7</dd>
##      <dd>Publisher: Random House Publishing Group</dd>
##      <dd>Publication Year: 2004</dd>
##      <dd>Pages: 480</dd>
##      <dd>Language: English</dd>
##      </dl>
##      </body>
##      </html>
##
##      attr(,"class")
##      [1] "XMLDocumentContent"

```

```
class(fave_books_html_2)
```

```
## [1] "XMLDocumentContent"
```

```

filename <- "/Users/Audiorunner13/CUNY MSDS Course Work/DATA607 Spring 2021/Week7/Homework7/Data/Fave_H
(fave_books_json = fromJSON(filename))

```

```

## $book1
##              Title              Author
##      "Faith of My Fathers"      "John McCain"
##              Education          Career
##      "US Naval Academy"        "Republican Senator"
##              Author             Miscellaneous
##      "John McCain"             "Naval Aviator"
##              ISBN-13            Publisher
##      "978-0-3995-9089-4" "Random House Publishing Group"
##              Publication Year    Pages
##      "1998"                   "398"

```



```

##          Language
##          "English"
##
## $book2
##          Title          Author
##          "Indianapolis"  "Lynn Vincent"
##          Education       Career
##          "US Navy Veteran" "Journalist"
##          Miscellaneous    Author
##          "Amazon Best of 2018" "Sara Vladic"
##          Education       Career
##          "Pepperdine University" "Documentary Filmmaker"
##          Miscellaneous    ISBN-13
##          "San Diego Film Award" "978-1-5011-3594-1"
##          Publisher        Publication Year
##          "Simon & Schuster Paperbacks" "2019"
##          Pages           Language
##          "578"           "English"
##
## $book3
##          Title          Author
##          "We Were Soldiers Once...and Young" "Hal Moore"
##          Education       Career
##          "West Point"    "Lt General, US Army"
##          Miscellaneous    Author
##          "Vietname Veteran" "Joseph Galloway"
##          Education       Career
##          "Bronze Star V-Device" "Author, Journalist"
##          Miscellaneous    ISBN-13
##          "Decorated non-combatant" "978-0-3454-7581-7"
##          Publisher        Publication Year
##          "Random House Publishing Group" "2004"
##          Pages           Language
##          "480"           "English"

```

```
class(fave_books_json)
```

```
## [1] "list"
```