Data 607, Working with html, xml, JSON

Mehdi Khan

October 13, 2017

```
library(rvest)
library(stringr)
library(magrittr)
```

read html page:

read_html function was used to load the html. The function returns a list with all the html nodes.

```
htmlData <- read_html("C:\\temp\\book.html")
htmlData
```

```
## {xml_document}
## <html>
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body>\r\n\r\n<h3> My Book List </h3>\n<hr>\n<p>\r\n<b>title:</b> <f ...
```

Since all the attribures of the books were in bold, they were extracted using html_nodes and html_text functions that provided all the texts in between all the and tags:

```
all_attr <- htmlData %>% html_nodes("b") %>% html_text()
all_attr <- str_trim(all_attr)
```

A character vector containing all the unique attributes of a book were produced bu using "unique" function

```
uniq_attr <- unique(all_attr)
```

All the texts of the html data were extracted using html_text function:

```
all_text <- str_replace_all(htmlData %>% html_text(), "[\r\n]", "")
```

Since each of the books starts with its first attribute (in this case 'title') and ends before the first attribute of the next book, the positions of the fisrt attributes were used to seperate all the information of each book as seperate elements of a character vector that has the same length as the number of the books:

```
pos <- str_locate_all(all_text, uniq_attr[1])

bk_text <- rep(NA, nrow(pos[[1]]))

for (i in 1:nrow(pos[[1]])) {
    if (i < 3) {
        bk_text[i] <- str_sub(all_text, pos[[1]][i, 1], pos[[1]][i +
            1, 1] - 1)
    } else {
        bk_text[i] <- str_sub(all_text, pos[[1]][i, 1], str_length(all_text))
    }

}
```

A function was created that would find the positions (start and end) of each of the attributes for each of the books and extract the texts (attribute values) in between those positions and store them as elements of a

character vector:

```
createBKds <- function(x, y) {

    bk_pos <- str_locate(x, y)
    BKDS <- vector("character", nrow(bk_pos))
    for (i in 1:nrow(bk_pos) - 1) {
        BKDS[i] <- str_trim(str_sub(x, bk_pos[i, 2] + 1, bk_pos[i +
            1, 1] - 1), side = "both")
    }
    BKDS[nrow(bk_pos)] <- str_trim(str_sub(x, bk_pos[nrow(bk_pos),
        2] + 1, str_length(x)), side = "both")

    return(BKDS)
}
```

Finally a data frame was created where the column names represent the attributes of the books and each of the rows represents corresponding attribute values of a book:

```
htmlBookDS <- c()
for (i in 1:length(bk_text)) {
    htmlBookDS <- rbind(htmlBookDS, createBKds(bk_text[i], uniq_attr))
}
htmlBookDS <- data.frame(htmlBookDS)
colnames(htmlBookDS) <- str_replace(uniq_attr, ":", "")

htmlBookDS
```

```
##                                                    title
## 1 Happy City: Transforming Our Lives Through Urban Design
## 2                                   Design and Analysis
## 3                                    A New Digital Deal
##                                                                      author
## 1                                                          Charles Montgomery
## 2 Bernard Leupen, Christoph Grafe, Nicola Kornig, Mark Lampe and Peter de Zeeuw
## 3                                                                Bas Boorsma
##                  publisher           category price publish year
## 1 Farrar, Straus and Giroux     Urban Planning 40.00         2013
## 2            010 Publishers       Architecture 35.00         1993
## 3   Rainmaking Publications Information System 35.00         2017
```

read xml file:

Four functions read_xml,xml_children, xml_text, xml_name were used to create the final data frame. xml_children function returned different elements in xml tree and xml_text returned character vectors containing data that were stored in between xml tags.

```
xml_data <- read_xml("C:\\temp\\book.xml")
books <- xml_children(xml_data)

xmlBookDS <- c()
for (i in 1:length(books)) {
    xmlBookDS <- rbind(xmlBookDS, xml_text(xml_children(books[i])))
}
```

```
xmlBookDS <- data.frame(xmlBookDS)

colnames(xmlBookDS) <- xml_name(xml_children(books[1]))

xmlBookDS
```

```
##                                                  title
## 1 Happy City: Transforming Our Lives Through Urban Design
## 2                                  Design and Analysis
## 3                                   A New Digital Deal
##                                                                          author
## 1                                                                Charles Montgomery
## 2 Bernard Leupen, Christoph Grafe, Nicola Kornig, Mark Lampe and Peter de Zeeuw
## 3                                                                     Bas Boorsma
##                   publisher           category price publish_year
## 1 Farrar, Straus and Giroux     Urban Planning 40.00         2013
## 2            010 Publishers       Architecture 35.00         1993
## 3   Rainmaking Publications Information System 25.00         2017
```

read json file:

fromJSON function directly converted data in json format to a data frame in R:

```
library(jsonlite)
```

load file

```
jsonBookDS <- fromJSON("C:\\temp\\book.json")
jsonBookDS
```

```
##                                                  title
## 1 Happy City: Transforming Our Lives Through Urban Design
## 2                                  Design and Analysis
## 3                                   A New Digital Deal
##                                                                          author
## 1                                                                Charles Montgomery
## 2 Bernard Leupen, Christoph Grafe, Nicola Kornig, Mark Lampe and Peter de Zeeuw
## 3                                                                     Bas Boorsma
##                   publisher           category price publish year
## 1 Farrar, Straus and Giroux     Urban Planning 40.00         2013
## 2            010 Publishers       Architecture  35.00         1993
## 3   Rainmaking Publications  Information System 25.00         2017
```

All three data frames are same that were created out of html, xml and JSON file.