Homework 6 - DATA 607

Kenan Sooklall

3/18/2021

`Assignment - Working with XML and JSON in R`

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.

Topic Computer Science

```
books = c('Hello World', 'The signal and the noise', 'Why we sleep')
authors = c('Hannah Fry', 'Nate Silver', 'Matthew Walker')
cover_color = c('Green', 'Yellow', 'Black')

df <- cbind(books, authors, cover_color)
df
```

```
##      books                       authors          cover_color
## [1,] "Hello World"               "Hannah Fry"     "Green"
## [2,] "The signal and the noise" "Nate Silver"    "Yellow"
## [3,] "Why we sleep"              "Matthew Walker" "Black"
```

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.

XML data

```
xurl <- 'https://raw.githubusercontent.com/ksooklall/CUNY-SPS-Masters-DS/main/DATA_607/homework/homework

xdf <- getURI(xurl) %>% xmlParse %>% xmlToDataFrame
xdf
```

```
##                        Book         Author Cover_color
## 1               Hello World     Hannah Fry       Green
## 2 The signal and the noise    Nate Silver      Yellow
## 3              Why we sleep Matthew Walker       Black
```

JSON data

```
jurl <- 'https://raw.githubusercontent.com/ksooklall/CUNY-SPS-Masters-DS/main/DATA_607/homework/homeworl

jdf <- fromJSON(jurl)$Books_Table
jdf
```

```
##                        Book         Author Cover_color
## 1             Hellow World     Hannah Fry       Green
## 2 The signal and the noise    Nate Silver      Yellow
## 3             Why we sleep Matthew Walker       Black
```

HTML data

```
hurl <- 'https://raw.githubusercontent.com/ksooklall/CUNY-SPS-Masters-DS/main/DATA_607/homework/homeworl

hdf <- as.data.frame(read_html(hurl) %>% html_table(fill=TRUE))
hdf
```

```
##                        Book         Author Cover_color
## 1              Hello World     Hannah Fry       Green
## 2 The signal and the noise    Nate Silver      Yellow
## 3             Why we sleep Matthew Walker       Black
```

Are the three data frames identical?

The data isn't read in the same, but after some parsing and cleaning their all identical now.