607 Assignment 7

Lu Beyer

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(XML)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

Overview

I want to read the same data in three different file formats: HTML, XML, and JSON. Hopefully this will give me a better idea of which formats are best or easiest to use in R for webscraping/APIs.

```
html <-  read_html('books.html')

html_df <- html %>%
```

```
  html_node("table") %>%
  html_table()

html_df
```

```
## # A tibble: 3 x 3
##   Series      Volume Author
##   <chr>        <int> <chr>
## 1 Nana             1 Ai Yazawa
## 2 Death Note       3 Tsugumi Ohba and Takeshi Obata
## 3 Sailor Moon      1 Naoko Takeuchi
```

```
json_df <- fromJSON("books.JSON")
json_df
```

```
##        Series Volume                         Author
## 1        Nana      1                      Ai Yazawa
## 2  Death Note      3 Tsugumi Ohba and Takeshi Obata
## 3 Sailor Moon      1                 Naoko Takeuchi
```

```
xml_data <- xmlParse("books.xml")
xml_df <- xmlToDataFrame(xml_data)
xml_df
```

```
##        Series Volume                         Author
## 1        Nana      1                      Ai Yazawa
## 2  Death Note      3 Tsugumi Ohba and Takeshi Obata
## 3 Sailor Moon      1                 Naoko Takeuchi
```

Conclusion

With the same information across three different file formats, each was read and loaded into R differently. While HTML and XML files appear very similar, they read numbers differently, with HTML being read numerically, and XML being read as character. JSON has the most confusing structure to read, but took the least amount of steps to read and present. HTML and XML are very intuitive for how they're formatted, lending to easier visual parsing.