Working with JSON, XML, and HTML Files

Jose Rodriguez

2022-10-17

Overview

In this document we explore way to load three different types of files into R; JSON, XML, and HTML. In addtion to loading the files we will tidy the information within these files to portray a flat file. In other words, we will transform them into data frames.

For each file type we first load the data from the github url. We then parse the information. Lastly, we transform it into a data frame class type.

```
url_json = "https://raw.githubusercontent.com/jlixander/DATA607/main/Assignment5/Assignment5.json"
json_df = jsonlite::fromJSON((url_json), simplifyDataFrame = TRUE) %>%
  as.data.frame()


knitr::kable(json_df, "pipe")
```

Books.TopBooks.Title	Books.TopBooks.Authors	Books.TopBooks.ISBN.10	Books.TopBooks.ISBN.13	Books.TopBooks.Publisher	Books.TopBooks.Language
Beautiful Creatures	Kami Garcia , Margaret Stohl	0316077038	978-0316077033	Little, Brown Books for Young Readers	English
The Five People You Meet in Heaven	Mitch Albom, NA	9781401308582	978-1401308582	Hachette Books	English
The Alchemist	Paulo Coelho, NA	0062315005	978-0062315007	HarperOne	English

```
url_xml = GET("https://raw.githubusercontent.com/jlixander/DATA607/main/Assignment5/assignment5.xml") #
xml_doc <- xmlParse(url_xml) #parse xml file
xml_df <- xmlToDataFrame(xml_doc, nodes=getNodeSet(xml_doc, "//TopBooks")) #Select which level to unlis

knitr::kable(xml_df, "pipe")
```

Title	Authors	ISBN-10	ISBN-13	Publisher	Language
Beautiful Creatures	Kami GarciaMargaret Stohl	0316077038	978-0316077033	Little, Brown Books for Young Readers	English
The Five People You Meet in Heaven	Mitch AlbomNA	9781401308582	978-1401308582	Hachette Books	English

Title	Authors	ISBN-10	ISBN-13	Publisher	Language
The Alchemist	Paulo CoelhoNA	0062315005	978-0062315007	HarperOne	English

```
html_url <- "https://raw.githubusercontent.com/jlixander/DATA607/main/Assignment5/Assignment5.html"
html_doc <- xml2::read_html(html_url)
tables = html_doc |> html_table()
table_one = tables[[1]]
html_df <- head(table_one, - 1) #Remove extra blank row.

knitr::kable(html_df, "pipe")
```

Title	Authors	ISBN-10	ISBN-13	Publisher	Language
Beautiful Creatures	Kami Garcia|Margaret Stohl	0316077038	978-0316077033	Little, Brown Books for Young Readers	English
The Five People You Meet in Heaven	Mitch Albom|NA	9781401308582	978-1401308582	Hachette Books	English
The Alchemist	Paulo Coelho|NA	0062315005	978-0062315007	HarperOne	English

Differences

There are some distinguishable differences when looking at all three tables. On table one we can see that this parser created the column names by taking the full hierarchy of where the information is found. In a sense we can say that its given us the XML directory of the objects. Another difference is that the lists of authors were placed together in the same cell, delimited by a comma.If we take a closer look at our dataframe in our environment, we will find that we have data frames within a data frame.

On table 2 we have clean column names. One major difference is that the lists of authors were concatenated together in one single string.

On table 3 we also have clean column names. The difference in this table is that we have the list of author names delimited by pipes.