Assignment 2

Karsten Donnay

15.06.2021

Contents

```
library(knitr)

### Global options
options(max.print="75")
opts_chunk$set(echo=FALSE,
                 cache=FALSE,
               prompt=FALSE,
               tidy=TRUE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
rm(list = ls())
```

Preparation

Install the 'stringr', 'xml2', and the 'jsonlite' package and load them.

Assignment 1: Data extraction

Data type

In the Github repository (and on OLAT) under "assignment2/data/", you will find two files: file_a and file_b. Use a text editor of your choice, inspect each file, and determine what type of data each file includes.

File A

Use the appropriate package to load the **file_a** into R. Try the functions associated with the package you used to load the data to get a feeling for the dataset and extract the following information with it.

- Extract a list of the IDs of the books.
- What is the name of the author of the 4th book?

```
rm(list = ls())
```

File B

Use the appropriate package to load **file_b** into R.

- What is the email of Mr. Bea?

```
rm(list = ls())
```

Assignment 2: Extract dates with Regular Expressions

You have the URL to an article of the Guardian. Use regular expressions to extract the publication date of the linked article. Your procedure should be generally applicable to the articles of The Guardian and not only work for this one link. Use the package **stringr** and its function **str_extract** or optionally **str_replace** to solve this task. (Hint: Slashes and Backslashes must be escaped with two preceding backslashes; e.g. ‘ /’ or ’ ’. The package lubridate and its function as_date() are helpful when trying to transform text to dates.)

```
url <- "https://www.theguardian.com/society/2018/oct/11/new-law-employers-reveal-race-pay-gap-figures"
# Structure: https://www.theguardian.com/category/year/month/day/headlinetext
result <- lubridate::as_date("2018-10-11")  # The result should look like this.
```

Assignment 3: Translation

Take the content of file_b used in the first assignment. Transform it into an XML-File format by hand. You can use https://www.xmlvalidation.com to validate your attempts.