Assignment 4: Solution

Karsten Donnay

17.06.2021

Contents


```
library(knitr)

### Global options
options(max.print="150")
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())
```

Assignment 1: Scrape static webpages

Visit the Wikipedia page for the winners of the Eurovision Song Contest since 1956 (https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_winners). Download the table 'Winners by year' and show which two countries are the countries who won the contest the most frequently.

Download

```
url <- "https://en.wikipedia.org/wiki/List_of_Eurovision_Song_Contest_winners"
url_parsed <- xml2::read_html(url)
```

Extract first table and convert it to data.frame

```
library(dplyr)
# Inspect with Google Chrome --> select level you are interested in -->
# Right-Click + Copy XPath
evwinners <- url_parsed %>%
    rvest::html_nodes(xpath = "//*[@id=\"mw-content-text\"]/div/table[1]") %>%
    rvest::html_table(fill = T) %>%
    as.data.frame()
```

Determine most frequent winner country

```
dplyr::top_n(data.frame(table(evwinners$Winner)), 2)  # Ireland and Sweden
```

```
     Var1 Freq
1 Ireland    7
2  Sweden    6
```

Assignment 2: Interact with dynamic pages

Goal: Log in to Twitter.

Assignment: write a script that visits the webpage of the login page of Twitter and tries to log in with the provided credentials.

Optionally: if you are not able to get a RSelenium webdriver to work, at least provide the XPath to the following elements of the website:

- field where you insert your username
- field where you insert your password
- login button

```
url <- "https://www.twitter.com/login"
username <- "testusername"
password <- "testpassword"
```

Navigate to the login page of Twitter.

```
remDr <- RSelenium::remoteDriver(port = 4445L)
remDr$open(silent = TRUE)
remDr$navigate("https://www.twitter.com/login")
```

Extract the two fields where we can put in our username and password (Use e.g. Google Chrome to inspect the page source. Right click on a field and copy the xpath. Paste it here.)

```
xpath_username <- "//*[contains(concat( \" \", @class, \" \" ), concat( \" \", \"r-13qz1uu\", \" \" ))]
xpath_password <- "//*[contains(concat( \" \", @class, \" \" ), concat( \" \", \"r-fdjqy7\", \" \" ))]"

library(yaml)
twitter_login_creds <- yaml::yaml.load_file("~/Documents/Keys/Twitter_login.yaml")

searchElem <- remDr$findElement(using = "xpath", value = xpath_username)
writeFrom <- searchElem$sendKeysToElement(list(twitter_login_creds$username))  # enter username into th

searchElem <- remDr$findElement(using = "xpath", value = xpath_password)
writeFrom <- searchElem$sendKeysToElement(list(twitter_login_creds$password))  # enter password into th
```

Extract the search button and then 'click' it.

```
xpath_loginbutton <- "//*[contains(concat( \" \", @class, \" \" ), concat( \" \", \"r-13qz1uu\", \" \" ]
searchElem <- remDr$findElement(using = "xpath", value = xpath_loginbutton)
resultsPage <- searchElem$clickElement()
```

Now you will likely be blocked and the webdriver will crash if you try to visit a Twitter profile. E.g.:

```
remDr$navigate("https://twitter.com/uzh_ch")
# close connection
remDr$closeServer()
```

Assignment 3: Start an own web scraping attempt

- Write a short script that evaluates data of a website of your choice and transforms it into a data format R can handle.
- Describe a way this data could be used in a potential project and provide some summary statistics of variables of interest.
- Check for alternative ways the website provider offers to access the data (e.g. does the provider have an application programming interface [API; next session]?).
- Check the robots.txt file of the website. What does it allow, what does it not allow ('Dissallow')?
- Check for alternative file dumps. Is there any indication that someone already downloaded contents of the website and provided it publicly?