Assignment 4

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.

Assignment 2: Interact with dynamic pages

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"
```

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?