

Automatic Sampling and Analysis of YouTube Data

Excursus Screenscraping

Julian Kohne
Johannes Breuer
M. Rohangis Mohseni

2021-02-24

Screenscraping

There are two packages for screenscraping in R

- **rvest**: Sufficient for scraping **static** websites
- **R Selenium**: Can also deal with **dynamic** websites

Dynamic websites are pages that *dynamically* load content from the database without changing the URL

Example: When you click on "show more" on the comment replies of a *YouTube* video, new content is loaded from the database but not the whole website is reloaded (this is done with **Ajax**).

Screenscraping with rvest

```
# installing and loading the package
if ("rvest" %in% installed.packages() != TRUE) {
  install.packages("rvest");library(rvest)

# defining website and XPath from inspect function in browser
page <- "https://www.youtube.com/watch?v=1aheRpmurAo&"
Xp <- "/html/body/div[2]/div[4]/div/div[5]
/div[2]/div[2]/div/div[2]/meta[2]"

# getting page
Website <- read_html(page)

# getting node containing the description
Description <- html_nodes(Website, xpath = Xp)

# printing description
html_attr(Description, name = "content")
```

"John Oliver discusses the census, why it matters, and the consequences of an undercount. Connect with Last Week Tonight online. Subscribe to the Last Week ..."

Screenscraping with RSeLenium

```
# We first have to configure Docker and open a Docker container:  
# https://callumgtaylor.github.io/blog/2018/02/01/using-rselenium-a  
  
# installing package  
if ("RSeLenium" %in% installed.packages()) != TRUE) {  
  install.packages("RSeLenium")  
}  
  
# loading package  
library(RSeLenium)  
  
# opening Docker container from system  
check <- system2("docker", args = "ps", stdout = TRUE)
```


Screenscraping with RSeLenium

```
# Assigning Google Chrome Docker session
remDr <- RSelenium::remoteDriver(remoteServerAddr = "localhost",
                                  port = 4445L,
                                  browserName = "chrome")

# Waiting for 5 seconds to finish initialization of Docker session
Sys.sleep(5)
```

Screenscraping with RSeLenium

- We can now navigate to a website and print a screenshot

```
# Open remote connection
remDr$open()

# Navigate to website
remDr$navigate("https://www.youtube.com/watch?v=1aheRpmurAo&")

# Wait for 2 seconds for the website to load
Sys.sleep(2)

# Scroll down a bit
webElem <- remDr$findElement("css", "body")
for (i in 20){
  webElem$sendKeysToElement(list(key = "down_arrow"))
}

# Take screenshot
remDr$screenshot(file = 'Images/R Selenium Screenshot.png')
```

Screenshot

Screenscraping with RSeLenuim

- We can then navigate to the "show more" button, and click it

```
# Xpath of "show more" button (using inspect element in browser)
xp <- '//*[@id="more"]/yt-formatted-string'

# navigating to button element
element <- remDr$findElement(using = 'xpath', xp)

# click on button
element$clickElement()

# scrolling down a bit
webElem <- remDr$findElement("css", "body")
for (i in 20){
  webElem$sendKeysToElement(list(key = "down_arrow"))
}

# take screenshot (we can see that the description box is now expanded)
remDr$screenshot(file = 'Images/R SeleniumScreenshot2.png')
```

Screenshot





Screenscraping with RSeLenium

- We can also extract the contents of the expanded description box

```
# navigate to description element
xp2 <- '//*[@id="description"]/yt-formatted-string'
element2 <- remDr$findElement(using = 'xpath', xp2)

# get element text
unlist(element2$getElementText())
```

"John Oliver discusses the census, why it matters, and the consequences of an undercount.\n\nConnect with Last Week Tonight online... \n\nSubscribe to the Last Week Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/lastweektonight \n\nFind Last Week Tonight on Facebook like your mom would: www.facebook.com/lastweektonight \n\nFollow us on Twitter for news about jokes and jokes about news: www.twitter.com/lastweektonight \n\nVisit our official site for all that other stuff at once: www.hbo.com/lastweektonight"

Exercise time    

Solutions