Web Data Collection with R

JavaScript Case Study

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

The Queens Christmas Speeches

The Queens Christmas Speeches

Teaser

- god ssave the queen

some preparations

```r
require(stringr)
require(RSelenium)
require(rvest)
require(wordcloud)
```

some preparations

```r
url1 <- "http://www.royal.gov.uk/ImagesandBroadcasts"
url2 <- "/TheQueensChristmasBroadcasts/ChristmasBroadcasts"
url3 <- "/ChristmasBroadcast"
base_url <- str_c(url1, url2, url3)

# make directory to save files in
dir.create("Queens_Speeches", showWarnings = F)
```

starting up engines

```
checkForServer() # search for and download Selenium Server
# linux: https://christopher.su/2015/selenium-chromedriver-
# windows: make sure to use Java 64bit for R 64Bit !!!
```

starting up engines

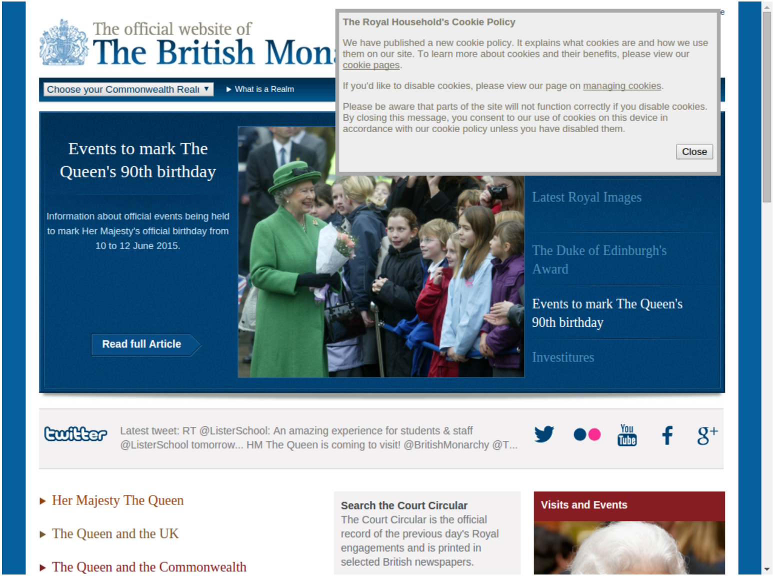
```r
startServer() # run Selenium Server binary
remDr <- remoteDriver(browserName="chrome", port=4444)
remDr$open()
remDr$navigate("http://www.royal.gov.uk/")

?remoteDriver
```

screen shot

```
remDr$screenshot(display = F,
                 useViewer=F,
                 file="screen1.png")
```

screen shot



The Royal Household's Cookie Policy

We have published a new cookie policy. It explains what cookies are and how we use them on our site. To learn more about cookies and their benefits, please view our cookie pages.

If you'd like to disable cookies, please view our page on managing cookies.

Please be aware that parts of the site will not function correctly if you disable cookies. By closing this message, you consent to our use of cookies on this device in accordance with our cookie policy unless you have disabled them.

Events to mark The Queen's 90th birthday

Information about official events being held to mark Her Majesty's official birthday from 10 to 12 June 2015.

Latest tweet: RT @ListerSchool: An amazing experience for students & staff @ListerSchool tomorrow... HM The Queen is coming to visit! @BritishMonarchy @T...

Search the Court Circular
The Court Circular is the official record of the previous day's Royal engagements and is printed in selected British newspapers.

getting speeches

```
for(i in 1953:2015){
  url <- str_c(base_url, i, ".aspx")
  remDr$navigate(url)

  html   <- remDr$getPageSource()[[1]]
  parsed <- htmlParse(html)
  speech <- xpathSApply(parsed,
                        '//div[@id="content"]/p',
                        xmlValue)
  speech_short <- speech[2:length(speech)]
  write(speech_short, str_c("Queens_Speeches/", i, ".txt"))
}
```

text

```r
text  <- str_c(readLines("Queens_Speeches/2014.txt"),
               collapse="\n")
substring(text, 1, 150)
```

```
## [1] "In the ruins of the old Coventry Cathedral is a scu
```

some cleansing

```r
library(tm)
corpus <- tolower(text)
corpus <- removeNumbers(corpus)
corpus <- removeWords(corpus, stopwords('english'))
corpus <- stemDocument(corpus)
```

word frequencies

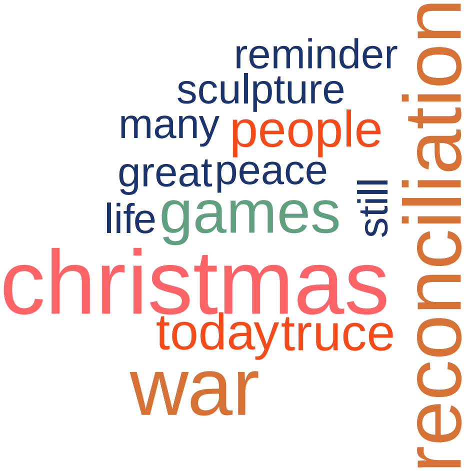
```
words <- unlist(str_split(corpus, "\\W"))
words <- words[words!="" & words!=" "]
tail(sort(table(words)),10)
```

```
## words
##       sculpture          still             us          pe
##               3              3              3
##           truce          games reconciliation
##               4              5              7
```

wordclouds

```
colors <- c("#CD8B34", "#89B151", "#01ABE9",
            "#1B346C", "#F54B1A", "#4B574D",
            "#609F80", "#AF420A", "#D67236",
            "#FD6467")
wordcloud(words, colors=colors)
```

wordclouds


reminder
sculpture
many
people
great
peace
still
life
games
christmas
today
truce
war
reconciliation

RSelenium alternative

- rdom / phantom.js
- https://github.com/cpsievert/rdom

Live Clicking // Remote Building