Web Scraping - News

Lisa Halmschlager (1902224)

11/12/2019

Task: Download news data from any websites for a given keyword:
- The outcome should be a dataframe
- Apply your function to at least 3 page
- Write function, use lapply and rbindlist

R Setup

```
library(rvest)
library(stringr)
library(data.table)

# Have your SelectorGadget on Google Chrome: https://selectorgadget.com/
```

Write function to scrape website

```
# my_url <- "https://www.diepresse.com/suche?s=Rauchverbot"

scrape_diepresse <- function(my_url) {

  page_html <- read_html(my_url, encoding="ISO-8859-1") # reads html document from URL

  page_titles <- page_html %>%
    html_nodes('.card__link') %>%
    html_text() %>%
    trimws()

  page_summary <- page_html %>%
    html_nodes(".card__text") %>%
    html_text() %>%
    trimws()

  page_links <- page_html %>%
    html_nodes('.card__link') %>%
    html_attr("href")

  page_kicker <- page_html %>%
    html_nodes(".card__kicker") %>%
    html_text() %>%
    sapply(function(txt){
      return(str_sub(strsplit(txt,"\n", fixed =T)[[1]][3],start = -10))
      }) %>%
    as.Date("%d.%m.%Y")

  page_df <- data.frame(title = page_titles,
                        teaser = page_summary,
                        links = page_links,
```

```r
                        date = page_kicker,
                        stringsAsFactors = FALSE)
  return(page_df)
}
```

Scrape website (first 5 pages)

```r
# links for first 5 pages
presse_links <- paste0('https://www.diepresse.com/suche?s=Rauchverbot&p=', 1:5)

# Scrape website
listofdf_presse <- lapply(presse_links, scrape_diepresse)
presse_df <- rbindlist(listofdf_presse)

# Save to file
write.table(presse_df,file="presse_df.csv")

# print head
head(presse_df)
```

```
##                                                                          title
## 1:                                       Die Betriebe brauchen Luft zum Atmen
## 2:                     Rauchverbot: Einbu<U+00C3><U+009F>en bis zu 15 Prozent
## 3: Diese f<U+00C3><U+00BC>nf Wirtschaftsgeschichten sollten Sie gelesen haben
## 4:              Stilvoll dilettantisch: Adam Green brummelt sich ins Paradies
## 5:                                                  Rauchverbot belastet Casinos
## 6:                  Voodoo J<U+00C3><U+00BC>rgens: Singsang aus der Unterwelt
##
## 1:                                                                                                 Wl
## 2:
## 3:                                         Der Bund macht <U+00C3><U+009C>bersch<U+00C3><U+00BC>sse,
## 4: Adam Green, der gro<U+00C3><U+009F>e Naive der US-Popmusik, stand einmal auf gro<U+00C3><U+009F>en
## 5:
## 6:
##
## 1:                       https://www.diepresse.com/5720877/Interview_Die-Betriebe-brauchen-Luft-zum-A
## 2:                  https://www.diepresse.com/5721163/Gastronomie_Rauchverbot_Einbussen-bis-zu-15-Pro
## 3:         https://www.diepresse.com/5718980/Diese-fuenf-Wirtschaftsgeschichten-sollten-Sie-gelesen-h
## 4: https://www.diepresse.com/5717943/Konzert_Stilvoll-dilettantisch_Adam-Green-brummelt-sich-ins-Para
## 5:                            https://www.diepresse.com/5717878/Gluecksspiel_Rauchverbot-belastet-Ca
## 6:                 https://www.diepresse.com/5717826/Gespraech_Voodoo-Juergens_Singsang-aus-der-Unte
##          date
## 1: 2019-11-12
## 2: 2019-11-12
## 3: 2019-11-08
## 4: 2019-11-06
## 5: 2019-11-06
## 6: 2019-11-06
```