Scraping Scientific American

Veronika Palotai

11/24/2019

Description

The code below consists of a function, `get_one_page_news` which gets the title, the teaser, the date and the author of an article published on the *Computing* section of *Scientific American*. All together the latest 10 pages of articles are scraped and the resulting list of dataframes is converted into a single dataframe.

Workflow

Importing the necessary libraries

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

Scraping the website

```
# function to get news from one page

get_one_page_news <- function(my_url) {
  scientific_american_html <- read_html(my_url)

  write_html(scientific_american_html, 'scientific_american_html.html')

  title <- scientific_american_html%>%
    html_nodes('.t_listing-title a') %>%
    html_text()

  teaser <- scientific_american_html %>%
    html_nodes('.listing-wide__inner__desc') %>%
    html_text()

  date_and_author <- scientific_american_html %>%
    html_nodes('#sa_body .t_meta') %>%
    html_text()

  news_df <- data.frame('news' = title, 'teaser' = teaser, 'date and author' = date_and_author)
  return(news_df)
}

urls <- paste0("https://www.scientificamerican.com/computing/?page=", seq(from=1, to=10, by=1))

news_by_page <- lapply(urls, get_one_page_news)

all_news <- rbindlist(news_by_page)

all_news <- data.frame(all_news)

write.table(all_news, 'news.csv', sep = ",", dec = ".", col.names = TRUE)
```