7 Practice Web Scraping

Juli Furjes

2022-11-06

Clone the repository at https://github.com/Digital-Methods-HASS/WebscrapingPoliceKillings and depending on your familiarity with R, either

1) adapt the web-scraping example to scrape homicide data from FBI site and produce a meaningful report on how homicide trends evolve around US in relation to this urban unrest or
2) use the rvest library to scrape data of your interest (football statistics in Wikipedia?, gender representatives in different governments? global population by country in https://www.worldometers.info/world-population/population-by-country/) or
3) produce data visualisations that shed light on another interesting aspect of the police killing data

Submit both the .rmd and the rendered .html files to your au###### github repository and paste link here.

I chose Exercise 2, and I am using the following link: https://www.nwf.org/Educational-Resources/Wildlife-Guide/Mammals/Raccoon

```
pacman::p_load(robotstxt)

# checking whether scraping is allowed on that website
paths_allowed(paths="https://www.nwf.org/Educational-Resources/Wildlife-Guide/Mammals/Raccoon")
```

```
##  www.nwf.org
```

```
## [1] TRUE
```

```
library(rvest)

# downloading the text from the website
raccoon_html <- read_html("https://www.nwf.org/Educational-Resources/Wildlife-Guide/Mammals/Raccoon")

# finding the name of the animal
name <- raccoon_html %>%
  html_nodes("h2") %>%
  html_text2()
name
```

```
## [1] "Raccoon"
## [2] "Get Involved"
## [3] "What's Trending"
## [4] "Where We Work"
## [5] "You are now leaving\nThe National Wildlife Federation."
```

```
# only keeping the first occurrence from the H2 elements
# (since that's the name of the animal)
name <- name[1]

# finding the latin name of the animal
```

```
latin_name <- raccoon_html %>%
  html_nodes("p.large-subhead") %>%
  html_text2()
latin_name
```

```
## [1] "Procyon lotor"
## [2] "Uniting all Americans to ensure wildlife thrive in a rapidly changing world"
```

```
# only keeping the first occurrence from the subheadings elements
# (since that's the latin name of the animal)
latin_name <- latin_name[1]

# scraping the descriptions based on the class of the paragraph element
all_text <- raccoon_html %>%
  html_nodes("div.bordered-container p") %>%
  html_text2()
all_text
```

```
##  [1] "Description"
##  [2] "A raccoon's face has several markings that help it stand out. The most noticeable marking is tl
##  [3] "Range"
##  [4] "Raccoons live throughout the continental United States in woods, wetlands, suburbs, parks, citi
##  [5] "Diet"
##  [6] "Raccoons are omnivores, meaning they will eat both meat and vegetables. They like grasshoppers
##  [7] "Life History"
##  [8] "Raccoons are solitary, except during the breeding season, which occurs from January to June. Fe
##  [9] "Fun Fact"
## [10] "At the National Wildlife Federation, we love raccoons, especially our mascot Ranger Rick! But i
## [11] "Sources"
## [12] "Animal Diversity Web, University of Michigan Museum of Zoology"
## [13] "Adirondack Ecological Center, College of Environmental Science and Forest, State University of
```

```
# creating empty lists for our two variables
titles <- list()
descriptions <- list()

# separating the titles and the descriptions within 'all_text')
# since they are alternating, we can use a for loop
for (x in 1:length(all_text)) {
  if(x%%2){
    titles <- append(titles, all_text[x])
  } else {
    descriptions <- append(descriptions, all_text[x])
  }
}

titles
```

```
## [[1]]
## [1] "Description"
##
## [[2]]
## [1] "Range"
##
## [[3]]
## [1] "Diet"
```

```
##
## [[4]]
## [1] "Life History"
##
## [[5]]
## [1] "Fun Fact"
##
## [[6]]
## [1] "Sources"
##
## [[7]]
## [1] "Adirondack Ecological Center, College of Environmental Science and Forest, State University of
```

```
descriptions
```

```
## [[1]]
## [1] "A raccoon's face has several markings that help it stand out. The most noticeable marking is th
##
## [[2]]
## [1] "Raccoons live throughout the continental United States in woods, wetlands, suburbs, parks, citi
##
## [[3]]
## [1] "Raccoons are omnivores, meaning they will eat both meat and vegetables. They like grasshoppers,
##
## [[4]]
## [1] "Raccoons are solitary, except during the breeding season, which occurs from January to June. Fe
##
## [[5]]
## [1] "At the National Wildlife Federation, we love raccoons, especially our mascot Ranger Rick! But i
##
## [[6]]
## [1] "Animal Diversity Web, University of Michigan Museum of Zoology"
```

```
# removing the lines which are not part of the actual article
# because they are also mentioned within the same dividers
titles <- titles[-c(6,7)]
descriptions <- descriptions[-6]

titles
```

```
## [[1]]
## [1] "Description"
##
## [[2]]
## [1] "Range"
##
## [[3]]
## [1] "Diet"
##
## [[4]]
## [1] "Life History"
##
## [[5]]
## [1] "Fun Fact"
```

```
descriptions
```

```
## [[1]]
## [1] "A raccoon's face has several markings that help it stand out. The most noticeable marking is th
##
## [[2]]
## [1] "Raccoons live throughout the continental United States in woods, wetlands, suburbs, parks, citi
##
## [[3]]
## [1] "Raccoons are omnivores, meaning they will eat both meat and vegetables. They like grasshoppers,
##
## [[4]]
## [1] "Raccoons are solitary, except during the breeding season, which occurs from January to June. Fe
##
## [[5]]
## [1] "At the National Wildlife Federation, we love raccoons, especially our mascot Ranger Rick! But i
```

```
# this is another way to scrape the titles
# this isdirectly from the page (based on the class of the paragraph element)
title <- raccoon_html %>%
  html_nodes("p.bordered-container-title") %>%
  html_text2()
title
```

```
## [1] "Description"  "Range"        "Diet"         "Life History" "Fun Fact"
## [6] "Sources"
```

```
# creating a dataframe out of the titles and descriptions
raccoon_data = data.frame(unlist(titles),unlist(descriptions))

# naming the columns
names(raccoon_data) = c("titles","descriptions")

# adding the name of the animal to the dataframe
raccoon_data$animal_name <- name
raccoon_data$latin_animal_name <- latin_name
```