Web Scraping Using R

Cesar Augusto JImenez Sanchez

1/22/2022

Reasoning

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Getting to know the web page's structure

Limitations

You can also embed plots, for example:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Defining a strategy

Hands-on

Creating a list of neighborhoods

```
# Libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
library(rvest)
```

```
##
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```r
library(stringi)

# Creating empty data frame
df <- data.frame(province = character(),
                 id = character(),
                 description = character())

# There are 31 provinces in the Dominican Republic, but supercasas.com considers
# municipalities as provinces, depending on its importance. The province ID does
# not increases orderly (some numerals have no province assigned to it).
for (province in 1:50) {
    # URL to iterate looking for all neighborhoods in each province
    base_url <- "https://www.supercasas.com/assets/js/autocomplete-search-location-sectors.js?text=&val=
    list_url <- paste0(base_url,
                       province)

    # Extracting HTML
    list_html <- read_html(list_url)

    # Verifying whether the URL contains a list of neighborhoods, the loop will
    # continue only if it contains a list
    empty_list <- stri_detect_fixed(html_text(list_html), "Selecciona")
    if(empty_list){
        next
    }

    sector <- 1

    repeat{
        links <- html_nodes(list_html,
                            paste0("body > div > ul > li:nth-child(",
                                   sector,
                                   ")"))
        len <- length(links)
        if(len == 0){
            break
        } else {
            df <- rbind(df, data.frame(province = province,
                                       id = html_attrs(links)[[1]][3],
                                       description = html_attrs(links)[[1]][4]))
            sector = sector + 1
        }
    }
```

```
}

# Removing the option "All neighborhoods" from the data set
df <- df %>%
    filter(description != "Todos los sectores") %>%
    distinct(id, .keep_all = TRUE)

# write.csv(df, "./1_data/0_raw/neighborhoods.csv",  row.names = FALSE)
```