The rvest Package in R

Amber Brodeur

4/11/2017

Contents

Introduction

Objective

Dataset Description

Variables

Goals

Preparation

In this tutorial we used the CSS selector tool (Selector gadget). To get started with webscraping, we will first take a look at how to use the Selector gadget. To grasp a basic understanding of how to use it in R, use the command `vignette("selectorgadget")` to learn about the Selector gadget.

```
vignette("selectorgadget")
```

Install the packages used in this project.

```
install.packages('rvest')
install.packages('magrittr')
```

Package descriptions:

- The `rvest` package....

- The 'magrittr' package is used for the pipe operator
-

Load the libraries into the workspace.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(magrittr)
```

We will explore the `rvest` package by reviewing demonstrations available in the `rvest` package. The command below lists the demonstrations in the `rvest` package.

```
demo(package = "rvest")
```

There are three demonstrations available for the `rvest` package: tripadvisor, united, and zillow.

Tutorial 1 - TripAdvisor

The command below explores the tripadvisor demonstration from the `rvest` package. When the `ask` setting is set to true, the demonstration pauses between pages. The results are not shown for the code.

```
demo(tripadvisor, package = "rvest", ask = FALSE)
```

The Tripadvisor url is saved as an object in the R environment.

```
url <- "http://www.tripadvisor.com/Hotel_Review-g37209-d1762915-Reviews-JW_Marriott_Indianapolis-Indiana
```

The `read_html` command is used to read the the Tripadvisor url. The pipe operator is used to forward the result of this expression to the `html_nodes` command. The `html_nodes` command, to specify the nodes of the reviews from the Tripadvisor webpage. The nodes are found by using the Selector gadget. The result is assigned to `reviews`.

```
reviews <- url %>%
   read_html() %>%
   html_nodes("#REVIEWS .innerBubble")
```

The command `length` is a base R command. This command returns the length of an object.

```
length(reviews)
```

```
## [1] 10
```

There were ten reviews extracted.

The `html_text` command extracts text from HTML. The code below extracts the text of the ten Tripadvisor reviews.

```
html_text(reviews)
```

```
##  [1] ""Terrific"Reviewed yesterday  NEWCan't say enough about this place. From the moment I stepped i
##  [2] ""Comfortable space and terrific room service food!"Reviewed yesterday  NEWvia mobileThe JW Mari
##  [3] ""My preferred hotel for business"Reviewed yesterday  NEWvia mobileI love almost everything abou
##  [4] ""business trip"Reviewed yesterday  NEWhelpful, courteous staff. walking to the convention cente
##  [5] ""Fantastic Hotel"Reviewed yesterday  NEWThis is my third trip to Indianapolis and my third stay
##  [6] ""A Beautiful View of Indianapolis"Reviewed yesterday  NEWThe JW Marriott is a perfect choice fo
##  [7] ""Lovely stay, likely cheaper options"Reviewed 3 days ago  NEWvia mobileI have no complaints abo
##  [8] ""Beautiful hotel!"Reviewed 4 days ago  NEWvia mobileThis is a beautiful hotel. It is definitely
```

```
##  [9] ""Comfortable Stay"Reviewed 6 days ago  NEWWarmly greeted upon arriving, we found the staff to
## [10] ""Executive Lounge Going Downhill"Reviewed 1 week ago  I have Marriott Platinum status, and have
```

The `reviews` are piped to the`html_node` command. The

With the Selector gadget, click on the title/quote of the review. When hovering over this selection, in small print you will see an `a` and `span`. These are two nodes for the quote. The `a` node is used to extract the id from the reviews. The `span` node will be used later. The `html_attr` command extracts attributes from HTML.

```
id <- reviews %>%
   html_node(".quote a") %>%
   html_attr("id")
```

```
length(id)
```

```
## [1] 10
```

The id attribute is returned below.

```
id
```

```
##  [1] "rn474334790" "rn474333652" "rn474331857" "rn474327331" "rn474225776"
##  [6] "rn474173318" "rn473694850" "rn473376421" "rn472983546" "rn472638625"
```

The `span` node mentioned above is used here. The `span` node extracts the quote from the Tripadvisor review. The `html_text` command extracts text from HTML.

```
quote <- reviews %>%
   html_node(".quote span") %>%
   html_text()
```

```
quote
```

```
##  [1] "Terrific"
##  [2] "Comfortable space and terrific room service food!"
##  [3] "My preferred hotel for business"
##  [4] "business trip"
##  [5] "Fantastic Hotel"
##  [6] "A Beautiful View of Indianapolis"
##  [7] "Lovely stay, likely cheaper options"
##  [8] "Beautiful hotel!"
##  [9] "Comfortable Stay"
## [10] "Executive Lounge Going Downhill"
```

The rating is extracted from the reviews object. The nodes are found using the css Selector gadget. The `html_attr` command extracts the attribute from HTML.

```
rating <- reviews %>%
   html_node(".rating .rating_s_fill") %>%
   html_attr("alt") %>%
   gsub(" of 5 stars", "", .) %>%
   as.integer()
```

The rating is not correct in the tutorial. All values returned NA.

```
rating
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA
```

To Be Continued...

Tutorial 2

The command below explores the `united` demonstration from the `rvest` package. When the `ask` setting is set to true, the demonstration pauses between pages.

```
demo(united, package = "rvest", ask = FALSE)
```

Tutorial 3

The command below explores the `zillow` demonstration from the `rvest` package. When the `ask` setting is set to true, the demonstration pauses between pages.

```
demo(zillow, package = "rvest", ask = FALSE)
```