The rvest Package in R

Amber Brodeur

4/11/2017

Contents

Introduction **1**

Objective **1**

Dataset Description **1**

Variables 2

Goals **2**

Preparation **2**

Tutorial 1 - TripAdvisor **3**

Tutorial 2 **6**

Tutorial 3 **6**

Conclusion **6**

Works Cited **6**

Introduction

Objective

The objective of this project is to master the `rvest` package and the CSS selector tool. This will be accomplished by going through the three tutorials available in the `rvest` package. The `rvest` package is used to scrape data from the Internet and is used in conjuction with the `magrittr` package. The CSS selector tool used in this tutorial is called Selectorgadget. Firstly, "CSS stands for Cascading Style Sheets and is a language used for describing the look and formatting of a document written in a markup language."[1] The Selectorgadget is used to extract specific components from HTML. More specifically, "Selectorgadget is a javascript bookmarklet that allows you to interactively figure out what css selector you need to extract desired components from a page."[2]

Dataset Description

There are three datasets that represent the three webpages scraped in the three tutorials. The first tutorial scrapes reviews on the JW Marriot Indianapolis hotel from the TripAdvisor webpage. The second tutorial scrapes flights from the United webpage. The third tutorial scrapes data from the Zillow webpage.

[1] https://en.wikipedia.org/wiki/Cascading_Style_Sheets

[2] Wickham, Hadley; Selectorgadget

Variables

There is one set of variables for each tutorial with a total of three sets of variables. Below are three tables representing each set.

TripAdvisor Variables

The table below is the data scraped of the reviews on the JW Marriot Indianapolis hotel from the TripAdvisor webpage.

Variable Name	Description	Unit
id	primary key / unique id	id
quote	quote from the hotel review	character
date	date of the review	monnth, day, year
review	hotel review	character

United Variables

The table below is the data scraped on the flights from the United webpage.

Variable Name	Description	Unit

Zillow Variables

The table below is the data scraped from the Zillow webpage.

Variable Name	Description	Unit

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 return documentation about how to download and use the Selector gadget.

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

A tutorial to practice with the Selector Gadget can be found at the following website: http://flukeout.github.io/#.

Install the packages used in this project.

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

Package descriptions:

- The `rvest` package is used to scrape data from the webpage.
- The `xml2` and `httr` packages are downloaded with the `rvest` package. These two packages are used to download and manipulate XML and HTML data.
- The `magrittr` package is used for the pipe operator.
-

Load the libraries into the workspace.

```
library(rvest)
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 *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-chunk below.

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

Next, 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 is used to select nodes from a HTML document. In this case, to select the nodes of the reviews from the Tripadvisor webpage. The Selector gadget is used to find the nodes. 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. The line of code below returns the number of reviews extracted from the TripAdvisor website.

```
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] ""Nice hotel, convenient to downtown "Reviewed yesterday  NEWvia mobileStayed here while in tow
##  [2] ""It was a great stay at the JW Marriott in Indianapolis"Reviewed 2 days ago  NEWI had an aweso
##  [3] ""No way"Reviewed 3 days ago  NEWvia mobileDon't get me wrong, this is a great hotel especially
##  [4] ""Excellent hotel - one of the best I've stayed at in the US"Reviewed 1 week ago  Rooms are spac
##  [5] ""Great hotel in the center of Indy happenings"Reviewed 1 week ago  Just returned from a confer
##  [6] ""Best of the Best in Indy"Reviewed 1 week ago  I cannot say enough about the service we receiv
##  [7] ""Very Poor Customer Service"Reviewed 1 week ago  We were attending a trade show at the attache
##  [8] ""great place to stay"Reviewed 1 week ago  Hotel is beautiful--rooms are great, quiet, big and
##  [9] ""BEWARE if you have physical challenges, Marriott does not care!!!"Reviewed 1 week ago  I atte
## [10] ""Terrific"Reviewed 1 week ago  Can't say enough about this place. From the moment I stepped in
```

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] "rn477556591" "rn477116696" "rn476745109" "rn475414846" "rn475399105"
##  [6] "rn475081150" "rn475007030" "rn474895828" "rn474848581" "rn474334790"
```

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] "Nice hotel, convenient to downtown "
##  [2] "It was a great stay at the JW Marriott in Indianapolis"
##  [3] "No way"
##  [4] "Excellent hotel - one of the best I've stayed at in the US"
##  [5] "Great hotel in the center of Indy happenings"
##  [6] "Best of the Best in Indy"
##  [7] "Very Poor Customer Service"
##  [8] "great place to stay"
##  [9] "BEWARE if you have physical challenges, Marriott does not care!!!"
## [10] "Terrific"
```

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
```

The date is extracted from the reviews object. The nodes are found using the css Selector gadget. The `html_attr` command extracts the attribute from HTML. The `strptime` command converts character vector to class “POSIXlt” and each input string is processed for the specified format. “POSIXlt” represents calendar dates and times and is a list of vectors that represents: seconds, minutes, hours, day of the month, year, day of the week, day of the year, daylight saving time, and time zone.

```
date <- reviews %>%
   html_node(".rating .ratingDate") %>%
   html_attr("title") %>%
   strptime("%b %d, %Y") %>%
   as.POSIXct()
```

```
date
```

```
##  [1] "2017-04-20 EDT" "2017-04-19 EDT" "2017-04-18 EDT" "2017-04-13 EDT"
##  [5] "2017-04-13 EDT" "2017-04-12 EDT" "2017-04-12 EDT" "2017-04-12 EDT"
##  [9] "2017-04-11 EDT" "2017-04-10 EDT"
```

There are 10 dates with the time zone in `date`.

The reviews are extracted from the reviews object. The nodes are found using the css Selector gadget. The `html_text` command extracts text from HTML.

```
review <- reviews %>%
   html_node(".entry .partial_entry") %>%
   html_text()
```

```
review
```

```
##  [1] "Stayed here while in town for a meeting at the convention center. The hotel is a full service
##  [2] "I had an awesome visit at the JW Marriott in Indianapolis. I stayed on the 33rd floor and the
##  [3] "Don't get me wrong, this is a great hotel especially for not a big city like Indianapolis. But
##  [4] "Rooms are spacious and clean. Great views of the city and canal. Walking distance to Lucas Oil
##  [5] "Just returned from a conference in Indianapolis and had the pleasure to stay at JW Marriott. T
##  [6] "I cannot say enough about the service we receive at the Indianapolis JW Marriott each time we
##  [7] "We were attending a trade show at the attached convention center. We had reserved 6 rooms mont
##  [8] "Hotel is beautiful--rooms are great, quiet, big and clean. Staff are really nice and say hello
##  [9] "I attended the Mideast Qualifier and reserved my room in November 2016. At that time, I took t
## [10] "Can't say enough about this place. From the moment I stepped into the facility until the moment
```

The `id`, `quote`, `date`, and `review` are combined to form a dataframe named `ta`.

```
ta <- data.frame(id, quote, date, review, stringsAsFactors = FALSE)
```

Review the structure of the `ta` dataframe.

```
str(ta)
```

```
## 'data.frame':    10 obs. of  4 variables:
##  $ id    : chr  "rn477556591" "rn477116696" "rn476745109" "rn475414846" ...
##  $ quote : chr  "Nice hotel, convenient to downtown " "It was a great stay at the JW Marriott in Ind
##  $ date  : POSIXct, format: "2017-04-20" "2017-04-19" ...
##  $ review: chr  "Stayed here while in town for a meeting at the convention center. The hotel is a ful
```

The dataframe has 10 observations and four variables. Conveniently, the data looks good and does not need any clean-up.

In order to save the data set as a csv file, we set the working directory and then used the write.csv command to export the data to the file path. Below, the csv file was saved on the desktop.

```
setwd("/Users/amberbrodeur/Desktop/Web Scrape")
write.csv(ta, 'tripadvisor.csv', row.names=FALSE)
```

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)
```

To Be Continued…

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)
```

Conclusion

Works Cited