RVest Package

By Gregory Albarian

Installation

```
install.packages("rvest")

library(rvest)

# used for web scraping but can also do some web scraping
```

Basic HTML Syntax

```
<!DOCTYPE html>
<html>
<head>
    <meta charset="UTF-8">
    <title>blank site</title>
    <!--comments-->
</head>
<body>
    <p>”Paragraphs”</p>
</body>
</html>
```

Example.com body example

```
</head>

<body>
<div>
    <h1>Example Domain</h1>
    <p>This domain is for use in illustrative examples in documents. You may use this
    domain in literature without prior coordination or asking for permission.</p>
    <p><a href="https://www.iana.org/domains/example">More information...</a></p>
</div>
</body>
</html>
```

y Albarian rvest packag

Example Domain

www.example.com

150%

Full-screen Snip

Example Domain

This domain is for use in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission.

More information...

Scraping Example.com

```
html_code = read_html("http://www.example.com/")

# %>% is like a pipe it says we are still referring to the same data in the next line

html_code %>%

html_nodes("p") %>% #use node to select one and nodes to select all

html_text()

#one side note you can find tables but putting into data frames nad parsing them well needs other libraries
```

The form of a table - goes in the body

```
<table style="width:50%">
 <tr>
      <th>1,1</th>
      <th>1,2</th>
      <th>1,3</th>
 </tr>
 <tr>
      <td>2,1</td>
      <td>2,2</td>
      <td>2,3</td>
 </tr>
 <tr>
      <td>3,1</td>
      <td>3,2</td>
      <td>3,3</td>
 </tr>
</table>
```

Display from last slide

1,1	**1,2**	**1,3**
2,1	2,2	2,3
3,1	3,2	3,3

Different ways to scrape tables

Practical Example scraping the stock information for the DOW Jones on Yahoo! Finance

One way: - use the html_node() or html_nodes() functions like we have been doing

Second way: - use a special builtin function - can store as list in a variable

html_table(read_html("https://finance.yahoo.com/quote/%5EDJI?p=^DJI&.tsrc=fin-srch"))

Output

- The output is stored as a list
- The list is full of String the numbers need to be converted to numeric values
- Can be used for more practical data analysis

Scraping Forms - finding user input on sites

- Again you can just use the get_nodes() function
- There is a special function
- Later example https://google.com/

Practical example - Google Search

Gregory Albarian rvest packag | r - Is there a technical differen | HTML Tables | Tryit Editor v3.6 | Example Domain | Google

https://www.google.com/webhp?hl=en&ictx=2&sa=X&ved=0ahUKEwiZ4uKG3JDsAhXYuZ4KHQEJAX8QPQgI 150%

Google Search | I'm Feeling Lucky

Inspector | Console | Debugger | Network | Style Editor | Performance | Memory | Storage | Accessibility | Application

```
role="search">
  ▸ <div id="tophf">…</div>
  ▾ <div jsmodel="vWNDde" jsdata="MuIEvd;;CiDl1k">
    ▾ <div class="A8SBwf" jscontroller="mvYTse" jsmodel="TMlYFc" jsdata="LVplcb;_;" jsaction="lX6RWd:w3Wsmc;
      DkpM0b:d3sQLd;IQOavd:dFyQEf;XzZZPe:jI3wzf;Aghs…d:ANdidc;NOg9L:HLgh3;uGoIkd:epUokb;zLdLw:eaGBS;rcuQ6b:npT2md">
      ▸ <style>…</style>
      ▸ <div class="RNNXgb" jsname="RNNXgb">…</div> flex
      ▸ <div class="UUbT9" jscontroller="tg8oTe" style="display: none;" jsname="UUbT9" jsaction="mouseout:ItzDCd;
        mouseleave:MWfikb;hBEIVb:nUZ9le;YMFC3:VKssTb">…</div>
      ▾ <div class="FPdoLc tfB0Bf">
        ▾ <center>
            <input class="gNO89b" value="Google Search" aria-label="Google Search" name="btnK" type="submit" data-
            ved="0ahUKEwjs-fSH3JDsAhWUrp4KHdwgDh0Q4dUDCAw"> event
            whitespace
            <input id="gbqfbb" value="I'm Feeling Lucky" aria-label="I'm Feeling Lucky" name="btnI" type="submit" data-
            ved="0ahUKEwjs-fSH3JDsAhWUrp4KHdwgDh0QnRsIDQ" style="visibility: inherit;"> event
          ▸ <div class="gbqfba gbqfba-hvr" role="button" style="display: none; font-family: arial, sans-serif; overflow:
            hid…absolute; left: 240px; margin: 0px; top: 80px; width: 110px;" aria-label="" tabindex="-1">…</div>
        </center>
      </div>
    </div>
  ▸ <div style="background:url(/images/searchbox/desktop_searchbox_sprites302_hr.webp)">…</div>
  </div>
</form>
```

html > body#gsr.hp.vasq > div#viewport.ctr-p > div#searchform.jhp > form#tsf.tsf.nj > div > div.A8SBwf > div.FPdoLc.tfB0Bf > center > input.gNO89b

Filter Styles :hov .cls

Pseudo-elements

This Element

```
element {                                          inline
}
.tfB0Bf input[type="submit"], .gbqfba              inline:1
{
  background-image: linear-gradient(top,#f5f5f5,#f1f1f1);
  background-color: #f2f2f2;
  border: 1px solid #f2f2f2;
  border-radius: 4px;
  color: #5F6368;
  font-family: arial,sans-serif;
  font-size: 14px;
  margin: 11px 4px;
  padding: 0 16px;
  line-height: 27px;
  height: 36px;
  min-width: 54px;
  text-align: center;
  cursor: pointer;
  user-select: none;
}
.jhp input[type="submit"], .gbqfba {               inline:1
  height: 36px;
  line-height: 27px;
```

Layout | Computed | Changes | Fonts | Animatic

No changes found.

Changes to CSS in Inspector will appear here.

Type here to search

4:10 AM
9/30/2020

Code:

```
url <-
"https://www.google.com/webhp?hl=en&sa=X&ved=0ahUKEwiAquTZ8IzsAhUOrJ4KHaesD7EQPAgI"

html_form(read_html(url))

# the output should show the different user inputs on the page

# allows us to see the Google Search button in the code
```

Sources:

- https://rdrr.io/cran/rvest/man/html_nodes.html#heading-3
- https://github.com/tidyverse/rvest
- http://rvest.tidyverse.org/
- https://www.dataquest.io/blog/web-scraping-in-r-rvest/
- https://cran.r-project.org/web/packages/rvest/rvest.pdf