# **RVest Package**

By Gregory Albarian

# **Installation**

install.packages("rvest")

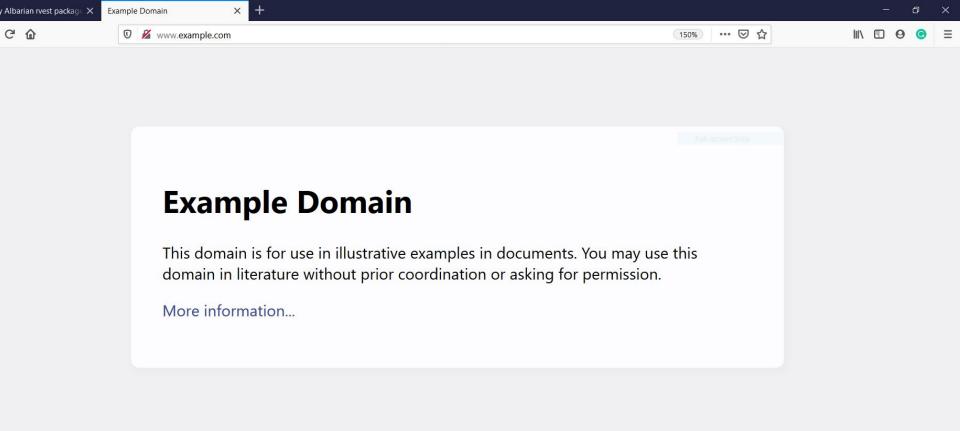
library(rvest)

# 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>
   "Paragraphs"
</body>
</html>
```

## Example.com body example



#### 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

```
2,2
2,3
1,1
         1,2
          3,1
 1,3
          3,2
3,3
2,1
```

# 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"))

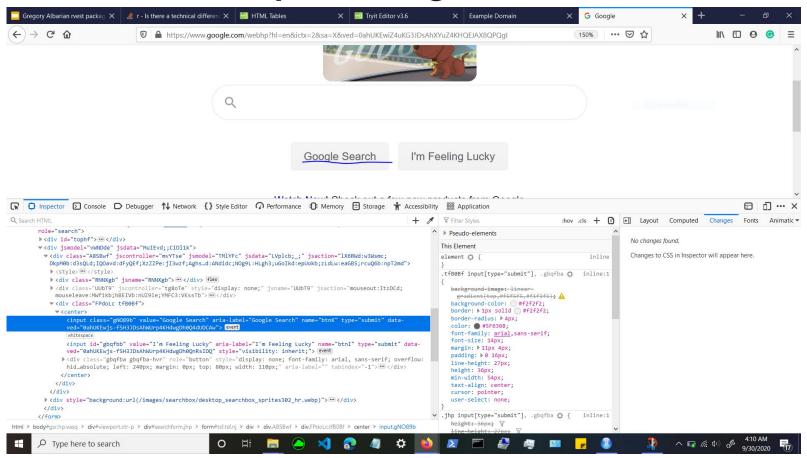
## Output

- 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



#### Code:

url <-

"https://www.google.com/webhp?hl=en&sa=X&ved=0ahUKEwiAquTZ8IzsAhUOrJ4 KHaesD7EQPAgI"

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