

Web Scraping 101

WORKING WITH WEB DATA IN R



Charlotte Wickham

Instructor

Selectors

- Little browser extensions
- Identify the specific bit(s) you want
- Give you a unique ID to grab them with
- Not used in this course (but worth grabbing after)

rvest

- `rvest` is a dedicated web scraping package
- Makes things shockingly easy
- Read HTML page with `read_html(url = ___)`

Parsing HTML

- `read_html()` returns an XML document
- Use `html_node()` to extract contents with XPATHs

Parsing HTML

```
wiki_r <- read_html(  
  "https://en.wikipedia.org/wiki/R_(programming_language)"  
)  
wiki_r
```

```
{xml_document}  
<html class="client-nojs" lang="en" dir="ltr">  
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; c ...  
[2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ...
```

```
wiki_r <- read_html(  
  "https://en.wikipedia.org/wiki/R_(programming_language)"  
)  
wiki_r
```

```
{xml_document}  
<html class="client-nojs" lang="en" dir="ltr">  
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; c ...  
[2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ...
```

```
html_node(wiki_r, xpath = "//ul")
```

```
{xml_node}  
<ul>  
[1] <li><a href="/wiki/Common_Lisp" title="Common Lisp">Common Li ...  
[2] <li><a href="/wiki/S_(programming_language)" title="S (progra ...  
[3] <li>\n<a href="/wiki/Scheme_(programming_language)" title="Sc ...  
[4] <li><a href="/wiki/XLispStat" title="XLispStat">XLispStat</a> ...
```

Let's practice!
WORKING WITH WEB DATA IN R

HTML Structure

WORKING WITH WEB DATA IN R



Oliver Keyes
Instructor

Tags

- HTML is content within tags
- Like XML
- `<p> this is a test </p>`

Attributes

```
<a href = "https://en.wikipedia.org/"> this is a test </a>
```

Extracting information

- `html_text(x = ___)` - get text contents
- `html_attr(x = ___, name = ___)` - get specific attribute
- `html_name(x = ___)` - get tag name

Let's practice!
WORKING WITH WEB DATA IN R

Reformatting Data

WORKING WITH WEB DATA IN R



Charlotte Wickham

Instructor

HTML tables

- HTML tables are dedicated structures: `<table>...</table>`
- They can be turned into data.frames with `html_table()`
- Use `colnames(table) <- c("name", "second_name")` to name the columns

Turning things into data.frames

- Non-tables can also become data.frames
- Use `data.frame()` , with the vectors of text or names or attributes

Let's practice!
WORKING WITH WEB DATA IN R