

Web Data Collection with R

HTML Forms GET Case Study

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT


Wikipedia general search

Wikipedia general search

HTML forms

http:

//en.wikipedia.org/w/index.php?title=Special%3ASearch



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)

Special page

Search

Search

[Content pages](#) [Multimedia](#) [Everything](#) [Advanced](#)

Not logged in

inspecting the search bar

```
<form id="search" method="get" action="/w/index.php">  
  <input type="hidden" value="Special:Search" name="title">  
  <input type="hidden" value="default" name="profile">  
  <input id="searchText" size="50" class="mw-ui-input mw-ui-input-text">  
  <input type="hidden" value="Search" name="fulltext">  
  <input class="mw-ui-button mw-ui-progressive" type="submit" value="Search"/>  
</form>
```

<https://en.wikipedia.org/w/index.php?title=Special%3ASearch&profile=default&search=franken&fulltext=Search>

inspecting the search bar

```
require(rvest)
url <- "http://en.wikipedia.org/w/index.php?title=Special%3F"
html <- read_html(url)
```

inspecting the serach bar

ADCR: page 236

```
attr_inspector <- function(parsed_html, xpath){  
  x <- html_nodes(parsed_html, xpath=xpath)  
  x <- html_attr(x)  
  x <- lapply(x, function(x) as.data.frame(t(x)) )  
  do.call(plyr::rbind.fill, x)  
}  
  
attr_inspector(html, "//form")
```

| ## | | id | method | action |
|------|------------|----|--------|--------------|
| ## 1 | search | | get | /w/index.php |
| ## 2 | searchform | | <NA> | /w/index.php |

inspecting the search bar

```
attr_inspector(html, "//form[1]//input")[,1:5]
```

| ## | type | value | name | id | size |
|------|--------|----------------|----------|-----------------|------|
| ## 1 | hidden | Special:Search | title | <NA> | <NA> |
| ## 2 | hidden | default | profile | <NA> | <NA> |
| ## 3 | search | <NA> | search | searchText | 50 |
| ## 4 | hidden | Search | fulltext | <NA> | <NA> |
| ## 5 | submit | Search | <NA> | <NA> | <NA> |
| ## 6 | search | <NA> | search | searchInput | <NA> |
| ## 7 | hidden | Special:Search | title | <NA> | <NA> |
| ## 8 | submit | Search | fulltext | mw-searchButton | <NA> |
| ## 9 | submit | Go | go | searchButton | <NA> |

filling out forms

```
require(stringr)
url1 <- str_c(url, "&search=Peter")
url2 <- str_c(url, "&search=Peter", "&fulltext=search")
url3 <- "http://en.wikipedia.org/w/index.php?search=Peter&f"
## browseURL(url1)
## browseURL(url2)
## browseURL(url3)
```

filling out forms - more elegant

```
require(httr)
url <- "http://en.wikipedia.org/w/index.php"
resp <-
  GET(url,
      query = list(
        title    = "Special:Search",
        profile  = "default",
        search   = "Bamberg",
        fulltext = "search"
      )
  )
```

filling out forms - more elegant

```
xpath = "//*[@class='mw-search-result-heading']/a"
results <-
html_attr(
  html_nodes(
    content(resp, "parsed"),
    xpath=xpath
  ), "title" )
```