Package 'htmltab'

July 22, 2015

Title Assemble Data Frames from HTML Tables

Version 0.6.0

Description HTML tables are a valuable data source but extracting and recasting these data into a useful format can be tedious. htmltab is a package for extracting structured information from HTML tables. It is similar to readHTMLTable() of the XML package but provides two major advantages. First, the function automatically expands row and column spans in the header and body cells. Second, users are given more control over the identification of header and body rows which will end up in the R table. Additionally, the function preprocesses table code, removes unneeded parts and so helps to alleviate the need for tedious post-processing.

Depends R (>= 3.0.0)

Imports XML (>= 3.98.1.3), httr (>= 1.0.0)

License MIT + file LICENSE

LazyData true

Suggests testthat, knitr, magrittr (>= 1.5), tidyr

URL https://github.com/crubba/htmltab

BugReports https://github.com/crubba/htmltab/issues

VignetteBuilder knitr

NeedsCompilation no

Author Christian Rubba [aut, cre]

Maintainer Christian Rubba <christian.rubba@gmail.com>

Repository CRAN

Date/Publication 2015-07-22 12:21:09

R topics documented:

htmltab *Assemble a data frame from HTML table data*

Description

Robust and flexible methods for extracting structured information out of HTML tables

Usage

```
htmltab(doc, which = NULL, header = NULL, headerFun = function(node)
  XML::xmlValue(node), headerSep = " >> ", body = NULL,
  bodyFun = function(node) XML::xmlValue(node), complementary = T,
  fillNA = NA, rm_superscript = T, rm_escape = " ", rm_footnotes = T,
  rm_nodata_cols = T, rm_invisible = T, rm_whitespace = T,
  colNames = NULL, ...)
```

Arguments

doc	the HTML document which can be a file name or a URL or an already parsed document (by XML's parsing functions)
which	a vector of length one for identification of the table in the document. Either a numeric vector for the tables' rank or a character vector that describes an XPath for the table
header	the header formula, see details for specifics
headerFun	a function that is executed over the header cell nodes
headerSep	a character vector that is used as a seperator in the construction of the table's variable names (default ' » ')
body	a vector that specifies which table rows should be used as body information. A numeric vector can be specified where each element corresponds to a table row. A character vector may be specified that describes an XPath for the body rows. If left unspecified, htmltab tries to use semantic information from the HTML code
bodyFun	a function that is executed over the body cell nodes
complementary	logical, should htmltab ensure complementarity of header, inbody header and body elements (default TRUE)?
fillNA	character vector of symbols that are replaced by NA (default c(''))
rm_superscript	logical, should superscript information be removed from header and body cells (default TRUE)?
rm_escape	a character vector that, if specified, is used to replace escape sequences in header and body cells (default ' ')
rm_footnotes	logical, should semantic footer information be removed (default TRUE)?
rm_nodata_cols	logical, should columns that have no alphanumeric data be removed (default TRUE)?

rm_invisible	logical, should nodes that are not visible be removed (default TRUE)?
rm_whitespace	logical, should leading/trailing whitespace be removed from cell values (default TRUE)?
colNames	a character vector of column names, or a function that can be used to replace specific column names (default NULL)
...	additional arguments passed to HTML parsers

Details

The header formula has the following format: level1 + level2 + level3 + ... level1 specifies the main header dimension (column names). This information must be for rows. level2 and deeper signify header dimensions that appear throughout the body. Those information muste be for cell elements, not rows. Header information may be one of the following types:

- the NULL value (default). No information passed, htmltab will try to identify header elements through heuristics (heuristics only work for the main header)
- A numeric vector that retrieves rows in the respective position
- A character string of an XPath expression
- A function that when evaluated produces a numeric or character vector
- 0, when the process of finding the main header should be skipped (only works for main header)

Value

An R data frame

Author(s)

Christian Rubba <http://www.christianrubba.com>

References

https://github.com/crubba/htmltab

Examples

```
## Not run:
# When no spans are present, htmltab produces output identical to XML's readHTMLTable()

 url <- "http://en.wikipedia.org/wiki/World_population"
 xp <- "//caption[starts-with(text(),'World historical')]/ancestor::table"
 htmltab(doc = url, which = xp)

 popFun <- function(node) {
   x <- XML::xmlValue(node)
   gsub(',', '', x)
 }

 htmltab(doc = url, which = xp, bodyFun = popFun)
```

```
#This table lacks header information. We provide them through colNames.
#We also need to set header = 0 to indicate that no header is present.
doc <- "http://en.wikipedia.org/wiki/FC_Bayern_Munich"
xp2 <- "//td[text() = 'Head coach']/ancestor::table"
htmltab(doc = doc, which = xp2, header = 0, encoding = "UTF-8", colNames = c("name", "role"))

#htmltab recognizes column spans and produces a one-dimension vector of variable information,
#also removes automatically superscript information since these are usually not of use.

 doc <- "http://en.wikipedia.org/wiki/Usage_share_of_web_browsers"
 xp3 <-  "//table[7]"
 bFun <- function(node) {
   x <- XML::xmlValue(node)
   gsub('%$', '', x)
 }

 htmltab(doc = doc, which = xp3, bodyFun = bFun)


#When header information appear throughout the body, you can specify their
#position in the header formula

htmltab("https://en.wikipedia.org/wiki/Arjen_Robben", which = 3,
header = 1:2 + "//tr/th[@colspan='3' and not(contains(text(), 'Club'))]")

## End(Not run)
```

---

num_xpath *num_xpath: Generate numeric XPath expression*

Description

Generate numeric XPath expression

Usage

```
num_xpath(data)
```

Arguments

data	the header XPath

Index