Package 'ralger'

February 18, 2020

Type Package

Title Easy Web Scraping

Version 1.1.0

Author Mohamed El Fodil Ihaddaden

Maintainer Mohamed El Fodil Ihaddaden <ihaddaden.fodeil@gmail.com>

Description The goal of 'ralger' is to facilitate web scraping in R.
The user has the ability to extract a vector with scrap(), a tidy dataframe using tidy_scrap(), a table with table_scrap() and web links with weblink_scrap().

License MIT + file LICENSE

Encoding UTF-8

LazyData true

URL https://github.com/feddelegrand7/ralger

BugReports https://github.com/feddelegrand7/ralger/issues

VignetteBuilder knitr

Imports magrittr, rvest, xml2, testthat, tidyr, dplyr, stringr,
robotstxt, crayon, purrr

Suggests knitr, rmarkdown

RoxygenNote 7.0.2

NeedsCompilation no

Repository CRAN

Date/Publication 2020-02-18 15:50:05 UTC

R topics documented:

scrap 2
table_scrap 2
tidy_scrap 3
weblink_scrap 4

Index **5**

scrap *Simple website scraping*

Description

This function is used to scrap one element from a website.

Usage

```
scrap(link, node, clean = FALSE, askRobot = FALSE)
```

Arguments

link	the link of the webpage to scrap
node	the HTML or CSS element to consider, the SelectorGadget tool is highly recommended
clean	logical. Should the function clean the extracted vector or not ? Default is FALSE.
askRobot	logical. Should the function ask the robots.txt if we're allowed or not to scrap the web page ? Default is FALSE.

Value

a character vector

Examples

```
# Extracting imdb top 250 movie titles

link <- "https://www.imdb.com/chart/top/"
node <- ".titleColumn a"

scrap(link, node)
```

table_scrap *HTML table scraping*

Description

This function is used to scrap an html table from a website.

Usage

```
table_scrap(link, askRobot = FALSE)
```

Arguments

link	the link of the webpage containing the table to scrap
askRobot	logical. Should the function ask the robots.txt if we're allowed or not to scrap the web page ? Default is FALSE.

Value

a tidy dataframe.

Examples

```
# Extracting premier ligue 2019/2020 top scorers

link     <- "https://www.topscorersfootball.com/premier-league"
table_scrap(link)
```

---

tidy_scrap *Website Tidy scraping*

Description

This function is used to scrap a tibble from a website.

Usage

```
tidy_scrap(link, nodes, colnames, clean = FALSE, askRobot = FALSE)
```

Arguments

link	the link of the webpage to scrap
nodes	the vector of HTML or CSS elements to consider, the SelectorGadget tool is highly recommended.
colnames	the names of the expected columns.
clean	logical. Should the function clean the extracted tibble or not ? Default is FALSE.
askRobot	logical. Should the function ask the robots.txt if we're allowed or not to scrap the web page ? Default is FALSE.

Value

a tidy dataframe.

Examples

```
# Extracting imdb movie titles and rating

link     <- "https://www.imdb.com/chart/top/"
my_nodes <- c(".titleColumn a", "strong")
names    <- c("title", "rating")

tidy_scrap(link, my_nodes, names)
```

`weblink_scrap` *Website Tidy scraping*

Description

This function is used to scrap a tibble from a website.

Usage

```
weblink_scrap(link, contain = NULL, askRobot = FALSE)
```

Arguments

`link`	the link of the webpage to scrap
`contain`	filter the web links according the character string provided. Particulary useful when extracting PDF or xlsx links (works also with regex)
`askRobot`	logical. Should the function ask the robots.txt if we're allowed or not to scrap the web page ? Default is FALSE.

Value

a character vector.

Examples

```
# Extracting the web links within the World Bank research and publications page

link     <- "https://www.worldbank.org/en/research"

weblink_scrap(link)
```

Index

scrap, 2

table_scrap, 2
tidy_scrap, 3

weblink_scrap, 4