Package 'scraply'

December 15, 2012

Maintainer Brian Abelson <brianabelson@gmail.com>

Author Brian Abelson <brianabelson@gmail.com>

Version 1.0

License MIT

Title

Description Tools for making scraping easier in R

Depends R (>= 2.15.1), RCurl, XML, plyr

Suggests stringr

Collate 'ahref.R' 'scraply.R' 'tree2node.R' 'url2tree.R'

R topics documented:

`ahref` *extract link and xmlValue from an xml node representing an "a" tag*

Description

extract link and xmlValue from an xml node representing an "a" tag

Usage

```
ahref(node, colnames = c("value", "link"))
```

Arguments

`node`	an XML node representing an "a" tag.
`colnames`	what to call the value and link extracted (in that order)

Value

a data.frame consisting of the link path and the text associated with the "a" tag.

Examples

```
# see example in the README
```

`scraply` *Scrape urls with llply, handling errors*

Description

This function works like `ldply`, but specifically for page scraping. Like `ldply`, it applies a function over a list (in this case a list of urls) and returns a data.frame. The difference is that scraply includes error handling and logging automagically This saves you a ton of time when you want to quickly write and deploy a page scraper. Happy scraplying!

Usage

```
scraply(ids, fx, sleep = 0)
```

Arguments

`ids`	A character vector of ids/urls to feed to a scraping function
`fx`	The scraping function to apply across the ids/urls
`sleep`	Seconds to sleep between iterations.

Value

A data.frame created by the scraping function (fx), with an added "error" column. Urls that dont return data will have scraped fields filled with NAs.

Examples

```
# see example in the README
```

`tree2node` *extract nodes from a html tree without having to write xpath!*

Description

extract nodes from a html tree without having to write xpath!

Usage

```
tree2node(tree, select, children = NULL)
```

Arguments

tree	a html tree returned from url2tree
select	a css selector to navigate to, e.g. 'class="keyword"'. This can also be a node attribute, e.g. 'cellpadding="1"'. Make sure to include double quotes around the value associated with the class/id.
children	OPTIONAL subsequent children nodes to navigate to. e.g. "a", "ul/li/a" etc.

Value

a list of nodes matching the constructed xpath expression

Examples

```
# see example in the README
```

---

url2tree *download a url and convert the html into a parseable tree in one step.*

Description

this function combines getURL and htmlTreeParse

Usage

```
url2tree(url)
```

Arguments

url	A url to download and convert into a html tree

Value

an html tree to parse with tree2node

Examples

```
# see example in the README
```

Index