Package ‘htm2txt’

October 19, 2017

Title Convert Html into Text

Version 2.1.1

Description Convert a html document to simple plain texts by removing all html tags. This package utilizes regular expressions to strip off html tags. It also offers gettxt() and browse() function, which enables you to get or browse texts at a certain web page.

Depends R (>= 3.0.0)

License GPL (>= 2)

URL https://github.com/sangchulpark

BugReports https://github.com/sangchulpark/htm2txt/issues

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Author Sangchul Park [aut, cre]

Maintainer Sangchul Park <mail@sangchul.com>

Repository CRAN

Date/Publication 2017-10-19 20:34:11 UTC

R topics documented:

browse ................................................................. 2
gettext ............................................................ 2
htm2txt ............................................................ 3

Index 4
**browse**  
*Display plain texts in a web page at a certain URL*

**Description**
Display plain texts in a web page at a certain URL.

**Usage**
browse(URL, ...)

**Arguments**
- `URL` A character indicating the URL of a web page.
- `...` Other htm2txt::htm2txt arguments.

**Value**
None (invisible NULL).

**Examples**
browse("https://CRAN.R-project.org/package=htm2txt")

---

**gettext**  
*Extract plain texts from a web page at a certain URL*

**Description**
Extract plain texts from a web page at a certain URL.

**Usage**
ggettext(URL, encoding = "UTF-8", ...

**Arguments**
- `URL` A character indicating the URL of a web page.
- `encoding` Encoding method (e.g., "UTF-8", "latin1", "bytes", "unknown", etc.).
- `...` Other htm2txt::htm2txt arguments.

**Value**
A character containing plain texts converted from the htm document at the URL.

**Examples**
text = gettext("https://CRAN.R-project.org/package=htm2txt")
**htm2txt**  
Convert a html document to simple plain texts by removing all html tags

---

**Description**

Convert a html document to simple plain texts by removing all html tags

**Usage**

```r
htm2txt(htm, list = "\n&#8226; ", pagebreak = "\n\n-------\n\n")
```

**Arguments**

- **htm**: A character vector, containing a html document, to be converted into plain texts (other objects are coerced into character vectors).
- **list**: A character that replaces a `<li>`...`<li>` tag (referring to a numbering or bullet for lists).
- **pagebreak**: A character that replaces a `<hr>` tag (referring to a thematic change in the content or a page break).

**Value**

A character vector containing plain texts converted from the html document.

**Examples**

```r
text = htm2txt("<html><body>html texts</body></html>")
text = htm2txt(c("Hello<p>World", "Goodbye<br>Friends"))
text = htm2txt("<p>Menu:<ul><li>Coffee</li><li>Tea</li></ul>", list = "\n- ")
text = htm2txt("Page 1<hr>Page 2", pagebreak = "\n\n[NEW PAGE]\n\n")
```
Index

browse, 2
gettxt, 2
htm2txt, 3