

# Package ‘htm2txt’

May 8, 2026

**Title** Convert Html into Text

**Version** 2.2.2

**Author** Sangchul Park [aut, cre]

**Maintainer** Sangchul Park <mail@sangchul.com>

**Description** Convert a html document to plain texts by stripping off all html tags.

**License** GPL (>= 2)

**URL** <https://github.com/replicable/htm2txt>

**BugReports** <https://github.com/replicable/htm2txt/issues>

**Encoding** UTF-8

**RoxygenNote** 7.2.0

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2022-06-12 15:50:09 UTC

## Contents

browse . . . . .	1
gettxt . . . . .	2
htm2txt . . . . .	3

<b>Index</b>	<b>4</b>
--------------	----------

---

browse	<i>Display simple plain texts in a web page at a certain URL</i>
--------	--

---

## Description

Display simple 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 <a href="#">gettxt</a> arguments.

**Value**

None (invisible NULL).

**Examples**

```
browse("https://www.wikipedia.org/")
```

---

gettxt

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

---

**Description**

Extract simple plain texts from a web page at a certain URL

**Usage**

```
gettxt(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 <a href="#">htm2txt</a> arguments.

**Value**

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

**Examples**

```
text = gettxt("https://www.wikipedia.org/")
```

---

htm2txt	<i>Convert a html document to plain texts by stripping off all html tags</i>
---------	--

---

**Description**

Convert a html document to plain texts by stripping off all html tags

**Usage**

```
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 "li" tags (referring to a numbering or bullet for lists). The default is a line change followed by a bullet character and a space.
pagebreak	A character that replaces "hr" tags (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**

```
text = htm2txt("<html><body>html texts</body></html>")
text = htm2txt(c("Hello<p>World", "Goodbye<br>Friends"))
text = htm2txt("<p>Menu:</p><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`, 1

`gettext`, 2, 2

`htm2txt`, 2, 3