

# Package ‘scrapeR’

February 22, 2024

**Type** Package

**Title** These Functions Fetch and Extract Text Content from Specified Web Pages

**Version** 0.1.8

**Description** The 'scrapeR' package utilizes functions that fetch and extract text content from specified web pages. It handles HTTP errors and parses HTML efficiently. The package can handle hundreds of websites at a time using the `scrapeR_in_batches()` command.

**License** MIT + file LICENSE

**Encoding** UTF-8

**Imports** htr, rvest, utils, magrittr, stringr

**NeedsCompilation** no

**Author** Mathieu Dubeau [aut, cre, cph]

**Maintainer** Mathieu Dubeau <mdubeau20@gmail.com>

**Repository** CRAN

**Date/Publication** 2024-02-22 00:40:02 UTC

## R topics documented:

scrapeR . . . . .	1
scrapeR_in_batches . . . . .	3

<b>Index</b>	<b>5</b>
--------------	----------

---

scrapeR *Web Page Content Scraper*

---

## Description

The `scrapeR` function fetches and extracts text content from the specified web page. It handles HTTP errors and parses HTML efficiently.

**Usage**

```
scrapeR(url)
```

**Arguments**

`url` A character string specifying the URL of the web page to be scraped.

**Details**

The function uses `tryCatch` to handle potential web scraping errors. It fetches the webpage content, checks for HTTP errors, and then parses the HTML content to extract text. The text from different HTML nodes like headings and paragraphs is combined into a single string.

**Value**

A character string containing the combined text from the specified HTML nodes of the web page. Returns NA if an error occurs or if the page content is not accessible.

**Note**

This function requires the **httr** and **rvest** packages. Ensure that these dependencies are installed and loaded in your R environment.

**Author(s)**

Mathieu Dubeau, Ph.D.

**References**

Refer to the [rvest package documentation](#) for underlying HTML parsing and extraction methods.

**See Also**

[GET](#), [read\\_html](#), [html\\_nodes](#), [html\\_text](#)

**Examples**

```
url <- "http://www.example.com"  
scraped_text <- scrapeR(url)
```

---

scrapeR\_in\_batches      *Batch Web Page Content Scraper*

---

### Description

The `scrapeR_in_batches` function processes a dataframe in batches, scraping web content from URLs in a specified column and writing the scraped content to a column in `df`.

### Usage

```
scrapeR_in_batches(df, url_column, extract_contacts)
```

### Arguments

<code>df</code>	A dataframe containing the URLs to be scraped.
<code>url_column</code>	The name of the column in <code>df</code> that contains the URLs.
<code>extract_contacts</code>	A function that searches scraped content for emails and phone numbers, defaults to <code>FALSE</code> .

### Details

This function divides the input dataframe into batches of a fixed size (default: 100). For each batch, it extracts the combined text content from the web pages of the URLs in the specified column. The results are appended to the `df`. The function also includes a throttling mechanism to pause between batch processing, reducing the load on the server being scraped.

### Value

The values are returned to `content` column and optionally to an `email` and `phone_number` column if `extract_contacts` is `TRUE`.

### Note

Ensure that the **httr**, **rvest**, and **stringr** packages are installed and loaded. Also, handle large datasets and output files with care to avoid memory issues.

### Author(s)

Mathieu Dubeau Ph.D

### References

Refer to [rvest package documentation](#) and [httr package documentation](#) for underlying web scraping methods.

**See Also**

[GET](#), [read\\_html](#), [html\\_nodes](#), [html\\_text](#), [write.table](#)

**Examples**

```
mock_scrapeR <- function(url) {
  return(paste("Scraped content from", url))
}

df <- data.frame(url = c("http://site1.com", "http://site2.com"), stringsAsFactors = FALSE)

## Not run:
  scrapeR_in_batches(df, url_column = "url", extract_contacts = FALSE)

## End(Not run)
```

# Index

- \* **HTML parsing**
    - scrapeR, 1
  - \* **URL processing**
    - scrapeR\_in\_batches, 3
  - \* **batch processing**
    - scrapeR\_in\_batches, 3
  - \* **data extraction**
    - scrapeR\_in\_batches, 3
  - \* **text extraction**
    - scrapeR, 1
  - \* **web content extraction**
    - scrapeR, 1
  - \* **web scraping**
    - scrapeR, 1
    - scrapeR\_in\_batches, 3
- GET, 2, 4
- html\_nodes, 2, 4
- html\_text, 2, 4
- read\_html, 2, 4
- scrapeR, 1
- scrapeR\_in\_batches, 3
- write.table, 4