Package: scrapeR (via r-universe)

August 20, 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 httr, rvest, utils, magrittr, stringr

NeedsCompilation no

Author Mathieu Dubeau [aut, cre, cph]

Maintainer Mathieu Dubeau <mdubeau20@gmail.com>

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

Repository https://mdubeau20.r-universe.dev

RemoteUrl https://github.com/cran/scrapeR

RemoteRef HEAD

RemoteSha 7444f28615c960a2f93e4c75812173ff5e45c34b

Contents

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

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

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