Package 'wikiScraper'

September 4, 2019

Title Scraping and formatting information from Wikipedia pages.

Version 0.0.0.9002

Description A series of wrapper functions around the rvest package
to retrieve data from Wikipedia pages.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Imports rvest,
xml2,
tidyverse,
sjmisc,
magrittr

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

`ws_get_card` *Get Data from Wikipedia Card*

Description

A function to extract data from the "infocards" some Wikipedia pages provide.

Usage

```
ws_get_card(page, format = "long", delay = 1)
```

Arguments

format	Either 'long' or 'wide'. 'long' returns an output with two columns (header and data), 'wide' returns an output with a column for each data entry.
delay	Rate at which to throttle calls. There is no delay if the function is passed an HTML object (e.g. from ws_get_page). Defaults to 1, can be turned off by setting to 0. Time between calls is determined by multiplying the value of this parameter with the response time by the server.
html	Either a url to a wikipediea page, or an object that contains the body of a wikipedia page (e.g. from ws_scrape_page).

Value

Returns a data_frame (tibble) that contains the data from a table with the class "infobox".

Examples

```
ws_get_card("wiki/New_York_City")

# OR
# get page THEN get card

page <- ws_get_page("New_York_City") # get page
ws_get_card(page) # then get data from the card
```

---

ws_get_geometry *Get Lat/Long Coordinates from Wikipedia Page*

Description

A helper function to parse Wikipedia-style geographic coordinates.

Usage

```
ws_get_geometry(data, delete_coords = T, coordinates)
```

Arguments

data	A data frame that contains a column of coordinates from Wikipedia.
delete_coords	Boolean value that indicates whether to remove the original coordinates column.
coordinates	A character object that contains the name of the column that contains the coordiantes to parse. If missing, the function will look for columns named coords or coordinates.

Value

Returns a the dataframe passed in the argument data, with columns lat and lon appended. By default, the original coordinates column is deleted.

Examples

```
ws_get_table("List_of_metro_systems") %>% ws_tidy_names()
```

---

ws_get_page *Get HTML from Wikipedia Page*

Description

Returns a list containing the contents of the requested webpage. Useful in conjunction with [ws_get_table()], [ws_get_card()], and [ws_get_section()]. This function is a wrapper around the function [xml2::get_html()].

Usage

```
ws_get_page(page, url = "https://wikipedia.org/wiki/", delay = 1)
```

Arguments

page	Extension of the page, e.g. "New_York_City"
url	The base url of the site to visit, defaults to "https://wikipedia.org/wiki/"
delay	Rate at which to throttle calls. Defaults to 1, can be turned off by setting to 0. Time between calls is determined by multiplying the value of this parameter with the response time by the server.

Value

An object that contains the HTML content of the requested URL.

Examples

```
ws_get_page("New York City")
# is equivelant to
ws_get_page("https://wikipedia.org/wiki/New_York_City")
# and
#xml2::read_html("https://wikipedia.org/wiki/New_York_City")
```

---

ws_get_section *Get a Section from a Wikipedia Page*

Description

Returns an HTML object that contains the requested section of the web page. Useful for getting a section of a long Wikipedia page to pass to [ws_get_table()].

Usage

```
ws_get_section(page, section, delay = 1)
```

Arguments

page	Either a url to a wikipediea page, or an object that contains the body of a wikipedia page (e.g. from ws_scrape_page).
section	The header or css id of the section to retrieve
delay	Rate at which to throttle calls. There is no delay if the function is passed an HTML object (e.g. from ws_get_page). Defaults to 1, delay can be turned off by setting this value to 0. Time between calls is determined by multiplying the value of this parameter with the server's response time.

Value

Returns an HTML object that contains the requested section of the web page.

Examples

```
ws_get_page("List_of_metro_systems")
# is equivelant to
ws_get_page("https://wikipedia.org/wiki/List_of_metro_systems")
```

---

ws_get_table *Scrape a Table from Wikipedia Page*

Description

A function to extract a table from a Wikipedia page.

Usage

```
ws_get_table(page, table = 1, skip = 0, header_length = 1,
  col_names = NULL, exclude_brackets = TRUE, exclude_parens = FALSE,
  format = NULL, delay = 1)
```

Arguments

page	Either the url of a Wikipedia, or an object that contains a Wikipedia page
table	An integer that specifies which table to get data from. Defaults to 1, which retrieves the first table element in the HTML object or web page passed.
skip	The number of rows to skip before collecting data. This is useful for omitting full-width "title" cells. Takes an integery and defaults to 0.
header_length	Set to a number greater than one to deal with multi-row headers. Takes an integer and defaults to 1.
col_names	Optional argument that takes a character vector to name columns in the output table.
exclude_brackets	Whether to exclude brackets and their contents in output. Takes a boolean and defaults to TRUE.
exclude_parens	Whether to exclude parenthesis and their contents in output. Takes a boolean and defaults to FALSE

delay	Rate at which to throttle calls. There is no delay if the function is passed an HTML object (e.g. from ws_get_page). Defaults to 1, can be turned off by setting to 0. Time between calls is determined by multiplying the value of this parameter with the response time by the server.

Value

Returns a dataframe (tibble) that contains the data from the table specified by the table argument.

Examples

```
ws_get_table("https://wikipedia.org/wiki/List_of_metro_systems")
ws_get_table("List_of_metro_systems")
```

---

ws_tidy_names *Tidy column names in tables*

Description

A helper function to clean up column names.

Usage

```
ws_tidy_names(data, replace_all = c(""), remove = NULL,
  rename = NULL, lowercase = TRUE)
```

Arguments

data	The data frame or tibble to tidy.
remove	A regular expression or vector of regular expressions to remove in all header names. For example, the argument c("aa","b") would delete ALL occurances of the characters "aa" and "b".
rename	A vector of the form c("old_name1", "new_name1", "old_name2", "new_name2"). In this case, "old_name1" or "old_name2" could be replaced with an integer indicating the column index in which the new name should be inserted.
lowercase	A boolean indicating whether the output names should be forced to lowercase. Defaults to TRUE.
replace	A vector of regular expressions and character string. The first element in the vector is replaced by the second, the third is replaced by the fourth, and so on. If a vector with an odd length is passed, the last element will be ignored will replace the first value with the second. In other words c("foo","bar") in the title would replace all occurances of "foo" with "bar" in the new menu.

Note

Order of Operations: 1) replace punctuation and spaces with underscores, 2) rename, 3) replace_all, 4) remove, 5) to lowercase. This could impact the results of the function.

Examples

```
ws_get_table("List_of_metro_systems") %>% ws_tidy_names()
```

Index