Package 'newscatcheR'

April 30, 2022

Title Programmatically Collect Normalized News from (Almost) Any Website

Version 0.1.1

Description Programmatically collect normalized news from (almost) any website. An 'R' clone of the <https://github.com/kotartemiy/newscatcher> 'Python' module.

License MIT + file LICENSE

URL https://github.com/discindo/newscatcheR/

BugReports https://github.com/discindo/newscatcheR/issues/

Depends R (>= 2.10)

Imports tidyRSS (>= 2.0.2), utils

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Language en-US

NeedsCompilation no

Author Novica Nakov [aut, cre],
Teofil Nakov [ctb],
Artem Bugara [ctb],
Discindo [cph]

Maintainer Novica Nakov <nnovica@gmail.com>

Repository CRAN

Date/Publication 2022-04-29 22:50:08 UTC

R topics documented:

`check_url` *Check URL A helper function to verify user input before fetching the feed.*

Description

Check URL A helper function to verify user input before fetching the feed.

Usage

```
check_url(website = "ycombinator.com", rss_table = package_rss)
```

Arguments

`website`	a url of a new source in the format "news.ycombinator.com"
`rss_table`	a dataframe with urls and rss feeds in case you need to construct your own out of websites not in the included database. Be sure to have the same format as the included data. See ‘R/package_rss.R‘ for details.

`describe_url` *Describe URL*

Description

Describe URL

Usage

```
describe_url(website = "ycombinator.com", rss_table = package_rss)
```

Arguments

website	a url of a new source in the format "news.ycombinator.com"
rss_table	a dataframe with urls and rss feeds in case you need to construct your own out of websites not in the included database. Be sure to have the same format as the included data. See package_rss.R for details.

Value

A character vector with topics.

Examples

```
describe_url(website = "ycombinator.com", rss_table = package_rss)
```

---

filter_urls *Filter URLs in the provided database based on topic, country and language*

Description

Filter URLs in the provided database based on topic, country and language

Usage

```
filter_urls(
  topic = NULL,
  country = NULL,
  language = NULL,
  rss_table = package_rss
)
```

Arguments

topic	the topic of the feed see show_topics() for more info.
country	the country of origin of the feed using two capital letters, for example "US". See show_countries() for more info.
language	the language of the content of the feed using two lowercase letters, for example "en". See show_languages() for more info.
rss_table	a dataframe with urls and rss feeds in case you need to construct your own out of websites not in the included database. Be sure to have the same format as the included data. See package_rss.R for details.

Value

a tibble filtered according to the given parameters

Examples

```
filter_urls(topic = "tech", country = "US", language = "en")
```

`get_headlines` *Get headlines A helper function to get just the headlines of the feed*

Description

Get headlines A helper function to get just the headlines of the feed

Usage

```
get_headlines(
  website = "ycombinator.com",
  topic = NULL,
  rss_table = package_rss
)
```

Arguments

`website`	a url of a new source in the format "news.ycombinator.com"
`topic`	the topic of the feed, by default it is NULL which means it will fetch the "main" feed. topics are 'tech', 'news', 'business', 'science', 'finance', 'food', 'politics', 'economics', 'travel', 'entertainment', 'music', 'sport', 'world', but not all site have all topics. use `describe_url("website")` to check for available feeds.
`rss_table`	a dataframe with urls and rss feeds in case you need to construct your own out of websites not in the included database. Be sure to have the same format as the included data. See `package_rss` for details.

Value

a tibble containing the headlines contained in the feed

Examples

```
## Not run:
Sys.sleep(3) # adding a small time delay to avoid
# simultaneous posts to the API
get_headlines(website = "ycombinator.com", rss_table = package_rss)

## End(Not run)
```

get_news *Get news Get the contents of a rss feed*

Description

Get news Get the contents of a rss feed

Usage

```
get_news(website = "ycombinator.com", topic = NULL, rss_table = package_rss)
```

Arguments

website	a url of a new source in the format "news.ycombinator.com"
topic	the topic of the feed, by default it is NULL which means it will fetch the "main" feed. topics are 'tech', 'news', 'business', 'science', 'finance', 'food', 'politics', 'economics', 'travel', 'entertainment', 'music', 'sport', 'world', but not all site have all topics. use `describe_url("website")` to check for available feeds.
rss_table	a dataframe with urls and rss feeds in case you need to construct your own out of websites not in the included database. Be sure to have the same format as the included data. See `?package_rss` for details.

Value

a tibble containing the contents of the rss feed

Examples

```
## Not run:
Sys.sleep(3) # adding a small time delay to avoid
# simultaneous posts to the API
get_news(website = "ycombinator.com", rss_table = package_rss)

## End(Not run)
```

---

newscatcheR *newscatcheR: Programmatically collect normalized news from (almost) any website using R*

Description

Two functions that work as a wrapper around tidyRSS can be used to fetch the feed from a given website. Two additional functions can be used to conveniently browse the websites dataset.

newscatcheR functions are

get_news(website) returns the contents of a rss feed of a website. get_headlines(website) returns just the headlines of the website's rss feed. describe_url(website) returns the topics of a given web site. filter_urls(topic, country, language) can be used to browse the dataset by topic, country or language. See more in the vignette.

`package_rss` *RSS table from python package newscatcher*

Description

A dataset containing sample medical data.

Usage

```
package_rss
```

Format

A data frame with 4505 rows and 7 variables:

clean_url url of news website

language the language of the website

topic_unified the topic of the website

main main

clean_country clean_country

rss_url location of feed

GlobalRank rank of website

Source

https://github.com/kotartemiy/newscatcher

show_countries *Show countries Show all countries in the database.*

Description

Show countries Show all countries in the database.

Usage

```
show_countries(rss_table = package_rss)
```

Arguments

rss_table	a dataframe with urls and rss feeds in case you #need to construct your own out of websites not in the included database. #Be sure to have the same format as the included data. See 'R/package_rss.R' #for details.

Value

a character vector of available countries

show_languages *Show languages Show all languages in the database.*

Description

Show languages Show all languages in the database.

Usage

```
show_languages(rss_table = package_rss)
```

Arguments

rss_table	a dataframe with urls and rss feeds in case you #need to construct your own out of websites not in the included database.#' #Be sure to have the same format as the included data. See 'R/package_rss.R' #for details.

Value

a character vector of available languages

show_topics *Show topics Show all topics in the database.*

Description

Show topics Show all topics in the database.

Usage

```
show_topics(rss_table = package_rss)
```

Arguments

rss_table	a dataframe with urls and rss feeds in case you #need to construct your own out of websites not in the included database. #Be sure to have the same format as the included data. See 'R/package_rss.R' #for details.

Value

a character vector of available topics

Index