Package 'RedditExtractoR'

January 5, 2019

Type Package

Title Reddit Data Extraction Toolkit

Version 2.1.5

Imports RJSONIO, utils, igraph, grDevices, graphics, magrittr, dplyr, visNetwork, rlang

Depends R (>= 3.2.0)

Date 2019-01-05

Author Ivan Rivera <ivan.s.rivera@gmail.com>

Maintainer Ivan Rivera <ivan.s.rivera@gmail.com>

Description A collection of tools for extracting structured data from <https://www.reddit.com/>.

License GPL-3

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-01-05 17:00:03 UTC

R topics documented:

construct_graph 2
get_reddit 2
RedditExtractoR 3
reddit_content 4
reddit_urls 5
user_network 6

Index **8**

construct_graph *Create a graph file from a single Reddit thread*

Description

Create a graph file from a single Reddit thread

Usage

```
construct_graph(content_data, plot = TRUE, write_to = NA)
```

Arguments

content_data	A data frame produced by reddit_content command
plot	A logical parameter indicating whether a graph is to be plotted or no (TRUE by default). Note that the root node corresponds to the thread itself rather than any of the comments.
write_to	A character string specifying the path and file name where a graph object will be saved. The file must end with an extension such as *.gml. The following formats are allowed: edgelist, pajek, ncol,lgl, graphml, dimacs, gml, dot", leda.

Value

A graph object

Examples

```
## Not run:
my_url = "reddit.com/r/web_design/comments/2wjswo/design_last_reordering_the_web_design_process/"
url_data = reddit_content(my_url)
graph_object = construct_graph(url_data)

## End(Not run)
```

get_reddit *Get all data attributes from search query*

Description

Get all data attributes from search query

Usage

```
get_reddit(search_terms = NA, regex_filter = "", subreddit = NA,
  cn_threshold = 1, page_threshold = 1, sort_by = "comments",
  wait_time = 2)
```

Arguments

search_terms	A string of terms to be searched on Reddit.
regex_filter	An optional regular expression filter that will remove URLs with titles that do not match the condition.
subreddit	An optional character string that will restrict the search to the specified subreddit.
cn_threshold	Comment number threshold that remove URLs with fewer comments that cn_threshold. 0 by default.
page_threshold	Page threshold that controls the number of pages is going to be searched for a given search word. 1 by default.
sort_by	Sorting parameter, either "comments" (default) or "new".
wait_time	wait time in seconds between page requests. 2 by default and it is also the minimum (API rate limit).

Value

A data frame with structure / position of the comment with respect to other comments (structure), ID (id), post / thread date (post_date), comment date (comm_date), number of comments within a post / thread (num_comments), subreddit (subreddit) upvote proportion (upvote_prop), post /thread score (post_score), author of the post / thread (author), user corresponding to the comment (user), comment score (comment_score), controversiality (controversiality), comment (comment), title (title), post / thread text (post_text), URL referenced (link) domain of the references URL (domain)

Examples

```
## Not run:
reddit_data = get_reddit(search_terms = "science",subreddit = "science",cn_threshold=10)

## End(Not run)
```

RedditExtractoR *Reddit Data Extraction Toolkit*

Description

Reddit is an online bulletin board and a social networking website where registered users can submit and discuss content. This package uses Reddit API to retrieve comments together will all corresponding attributes from Reddit threads. Note that at this stage, the extraction produces a data frame with a flat structure, i.e. without preserving the order or heirarchy of individuals comments. This may be addressed in the next version of this package. Also note that due to API limitations, the number of comments available for retrieval is limited to 500 per thread.

Details

Package:	RedditExtractoR
Type:	Package
Version:	2.1.0
Date:	2015-06-14
License:	GPL-3

The package contains a collection of functions for extracting threads of interest and their corresponding comments, as well as functions for analysing the structure of these threads.

Author(s)

Ivan Rivera

Maintainer: Ivan Rivera <ivan.s.rivera@gmail.com>

References

https://www.reddit.com/dev/api

See Also

www.reddit.com

Examples

```
example_urls = reddit_urls(search_terms="science")
## Not run:
example_attr = reddit_content(URL="reddit.com/r/gifs/comments/39tzsy/whale_watching")
example_data = get_reddit(search_terms="economy")
## End(Not run)
```

---

reddit_content *Extract data attributes*

Description

Extract data attributes

Usage

```
reddit_content(URL, wait_time = 2)
```

Arguments

URL	a string or a vector of strings with the URL address of pages of interest
wait_time	wait time in seconds between page requests. 2 by default and it is also the minimum (API rate limit).

Value

A data frame with structure / position of the comment with respect to other comments (structure), ID (id), post / thread date (post_date), comment date (comm_date), number of comments within a post / thread (num_comments), subreddit (subreddit) upvote proportion (upvote_prop), post /thread score (post_score), author of the post / thread (author), user corresponding to the comment (user), comment score (comment_score), controversiality (controversiality), comment (comment), title (title), post / thread text (post_text), URL referenced (link) domain of the references URL (domain)

Examples

```
## Not run:
example_attr = reddit_content(URL="reddit.com/r/gifs/comments/39tzsy/whale_watching")

## End(Not run)
```

`reddit_urls` *Returns relevant reddit URLs*

Description

Returns relevant reddit URLs

Usage

```
reddit_urls(search_terms = NA, regex_filter = "", subreddit = NA,
  cn_threshold = 0, page_threshold = 1, sort_by = "relevance",
  wait_time = 2)
```

Arguments

`search_terms`	A character string to be searched on Reddit.
`regex_filter`	An optional regular expression filter that will remove URLs with titles that do not match the condition.
`subreddit`	An optional character string that will restrict the search to the specified subreddits (separated by space).
`cn_threshold`	Comment number threshold that remove URLs with fewer comments that cn_threshold. 0 by default.
`page_threshold`	Page threshold that controls the number of pages is going to be searched for a given search word. 1 by default.
`sort_by`	Sorting parameter, either "comments" (default) or "new".
`wait_time`	wait time in seconds between page requests. 2 by default and it is also the minimum (API rate limit).

Value

A data frame with URLs (links), number of comments (num_comments), title (title),date (date) and subreddit (subreddit).

Examples

```
## Not run:
example_urls = reddit_urls(search_terms="science")

## End(Not run)
```

`user_network` *User relationship network*

Description

User relationship network

Usage

```
user_network(thread_df, include_author = TRUE, agg = FALSE)
```

Arguments

`thread_df`	a data frame with columns structure, user, author – designed for the output of the reddit_content() or get_reddit() functions
`include_author`	a TRUE/FALSE indicator for whether or not the author should be considered in the resulting network, if TRUE, comments at the top of the tree will be treated as relies to the original posts, otherwise these comments will be disregarded
`agg`	a TRUE/FALSE indicator that allows you to aggregate results, if FALSE, the results will remain disaggregated

Value

a list with df (effectively an edge list without IDs), node_df (node list), edge_df (edge list), igrpah (igraph object), plot (plot object)

Examples

```
## Not run:
# load libraries
library(dplyr)
library(RedditExtractoR)
target_urls <- reddit_urls(search_terms="cats", subreddit="Art", cn_threshold=50)
target_df <- target_urls %>%
filter(num_comments==min(target_urls$num_comments)) %$%
URL %>% reddit_content # get the contents of a small thread
```

```
network_list <- target_df %>% user_network(include_author=FALSE, agg=TRUE) # extract the network
network_list$plot # explore the plot
str(network_list$df) # check out the contents

## End(Not run)
```

Index

∗Topic **reddit**
RedditExtractoR, 3

construct_graph, 2

get_reddit, 2

reddit_content, 4
reddit_urls, 5
RedditExtractoR, 3

user_network, 6