Package ‘streamR’

January 5, 2014

Title Access to Twitter Streaming API via R

Description This package provides a series of functions that allow R users
to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

Version 0.2

Author Pablo Barbera <pablo.barbera@nyu.edu>

Maintainer Pablo Barbera <pablo.barbera@nyu.edu>

Depends R (>= 2.12.0), RCurl, rjson, ROAuth

License GPL-2

Collate 'filterStream.R' 'parseTweets.R' 'sampleStream.R'
'userStream.R' 'streamR-package.R' 'readTweets.R'

R topics documented:

`streamR-package` *Access to Twitter Streaming APIs via R*

Description

This package provides a series of functions that allow R users to access Twitter’s filter, sample, and user streams, and to parse the output into data frames.

Author(s)

Pablo Barbera <pablo.barbera@nyu.edu>

See Also

filterStream, sampleStream, userStream, readTweets, parseTweets

example_tweets *Ten sample tweets published by @twitterapi*

Description

A vector of string characters that contains ten sample tweets in plain text.

Source

http://www.twitter.com/twitterapi

filterStream *Connect to Twitter Streaming API and return public statuses that match one or more filter predicates.*

Description

filterStream opens a connection to Twitter's Streaming API that will return public statuses that match one or more filter predicates. Tweets can be filtered by keywords, users, language, and location. The output can be saved as an object in memory or written to a text file.

Usage

```
filterStream(file.name = NULL, track = NULL,
  follow = NULL, locations = NULL, language = NULL,
  timeout = 0, tweets = NULL, oauth, verbose = TRUE)
```

Arguments

file.name	string, name of the file where tweets will be written. "" indicates output to the console, which can be redirected to an R object (see examples). If the file already exists, tweets will be appended (not overwritten).
track	string or string vector containing keywords to track. See the track parameter information in the Streaming API documentation for details: http://dev.twitter.com/docs/streaming-apis/parameters#track.
follow	string or numeric, vector of Twitter user IDs, indicating the users whose public statuses should be delivered on the stream. See the follow parameter information in the Streaming API documentation for details: http://dev.twitter.com/docs/streaming-apis/parameters#follow.
locations	numeric, a vector of longitude, latitude pairs (with the southwest corner coming first) specifying sets of bounding boxes to filter public statuses by. See the locations parameter information in the Streaming API documentation for details: http://dev.twitter.com/docs/streaming-apis/parameters#locations

`language` string or string vector containing a list of BCP 47 language identifiers. If not `NULL` (default), function will only return tweets that have been detected as being written in the specified languages. Note that this parameter can only be used in combination with any of the other filter parameters. See documentation for details: `https://dev.twitter.com/docs/streaming-apis/parameters#language`

`timeout` numeric, maximum length of time (in seconds) of connection to stream. The connection will be automatically closed after this period. For example, setting `timeout` to 10800 will keep the connection open for 3 hours. The default is 0, which will keep the connection open permanently.

`tweets` numeric, maximum number of tweets to be collected when function is called. After that number of tweets have been captured, function will stop. If set to `NULL` (default), the connection will be open for the number of seconds specified in `timeout` parameter.

`oauth` an object of class `oauth` that contains the access tokens to the user's twitter session. This is currently the only method for authentication. See examples for more details.

`verbose` logical, default is `TRUE`, which generates some output to the R console with information about the capturing process.

Details

`filterStream` provides access to the statuses/filter Twitter stream.

It will return public statuses that match the keywords given in the `track` argument, published by the users specified in the `follow` argument, written in the language specified in the `language` argument, and sent within the location bounding boxes declared in the `locations` argument.

Note that location bounding boxes do not act as filters for other filter parameters. In the fourth example below, we capture all tweets containing the term rstats (even non-geolocated tweets) OR coming from the New York City area. For more information on how the Streaming API request parameters work, check the documentation at: `http://dev.twitter.com/docs/streaming-apis/parameters`.

Also note that the `language` parameter needs to be used in combination with another filter option (either keywords or location).

If any of these arguments is left empty (e.g. no user filter is specified), the function will return all public statuses that match the other filters. At least one predicate parameter must be specified.

Note that when no file name is provided, tweets are written to a temporary file, which is loaded in memory as a string vector when the connection to the stream is closed.

The total number of actual tweets that are captured might be lower than the number of tweets requested because blank lines, deletion notices, and incomplete tweets are included in the count of tweets downloaded.

Author(s)

Pablo Barbera <`pablo.barbera@nyu.edu`>

See Also

`sampleStream`, `userStream`, `parseTweets`

Examples

```
## Not run:

## An example of an authenticated request using the ROAuth package,
## where consumerkey and consumer secret are fictitious.
## You can obtain your own at dev.twitter.com
  library(ROAuth)
  requestURL <- "https://api.twitter.com/oauth/request_token"
  accessURL <- "http://api.twitter.com/oauth/access_token"
  authURL <- "http://api.twitter.com/oauth/authorize"
  consumerKey <- "xxxxxyyyyyzzzzzz"
  consumerSecret <- "xxxxxxyyyyyzzzzzzz111111222222"
  my_oauth <- OAuthFactory$new(consumerKey=consumerKey,
    consumerSecret=consumerSecret, requestURL=requestURL,
    accessURL=accessURL, authURL=authURL)
  my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl"))
  filterStream( file="tweets_rstats.json",
   track="rstats", timeout=3600, oauth=my_oauth )

## capture 10 tweets mentioning the "Rstats" hashtag
  filterStream( file.name="tweets_rstats.json",
     track="rstats", tweets=10, oauth=my_oauth )

## capture tweets published by Twitters official account
  filterStream( file.name="tweets_twitter.json",
     follow="783214", timeout=600, oauth=my_oauth )

## capture tweets sent from New York City in Spanish only, and saving as an object in memory
  tweets <- filterStream( file.name="", language="es",
      locations=c(-74,40,-73,41), timeout=600, oauth=my_oauth )

## capture tweets mentioning the "rstats" hashtag or sent from New York City
  filterStream( file="tweets_rstats.json", track="rstats",
      locations=c(-74,40,-73,41), timeout=600, oauth=my_oauth )


## End(Not run)
```

parseTweets *Converts tweets in JSON format to data frame.*

Description

This function parses tweets downloaded using filterStream, sampleStream or userStream and returns a data frame.

Usage

```
parseTweets(tweets, simplify = FALSE, verbose = TRUE)
```

Arguments

tweets A character string naming the file where tweets are stored or the name of the object in memory where the tweets were saved as strings.

simplify	If TRUE it will return a data frame with only tweet and user fields (i.e., no geographic information or url entities).
verbose	logical, default is TRUE, which will print in the console the number of tweets that have been parsed.

Details

parseTweets parses tweets downloaded using the filterStream, sampleStream or userStream functions and returns a data frame where each row corresponds to one tweet and each column represents a different field for each tweet (id, text, created_at, etc.).

The total number of tweets that are parsed might be lower than the number of lines in the file or object that contains the tweets because blank lines, deletion notices, and incomplete tweets are ignored.

To parse json to a twitter list, see readTweets. That function can be significantly faster for large files, when only a few fields are required.

Author(s)

Pablo Barbera <pablo.barbera@nyu.edu>

See Also

filterStream, sampleStream, userStream

Examples

```
## The dataset example_tweets contains 10 public statuses published
## by @twitterapi in plain text format. The code below converts the object
## into a data frame that can be manipulated by other functions.

data(example_tweets)
tweets.df <- parseTweets(example_tweets, simplify=TRUE)

## Not run:
## A more complete example, that shows how to capture a users home timeline
## for one hour using authentication via OAuth, and then parsing the tweets
## into a data frame.

 library(ROAuth)
 reqURL <- "https://api.twitter.com/oauth/request_token"
 accessURL <- "http://api.twitter.com/oauth/access_token"
 authURL <- "http://api.twitter.com/oauth/authorize"
 consumerKey <- "xxxxxyyyyyzzzzzz"
 consumerSecret <- "xxxxxxyyyyyzzzzzzz111111222222"
 my_oauth <- OAuthFactory$new(consumerKey=consumerKey,
                              consumerSecret=consumerSecret,
                              requestURL=reqURL,
                              accessURL=accessURL,
                              authURL=authURL)
 my_oauth$handshake()
 userStream( file="my_timeline.json", with="followings",
         timeout=3600, oauth=my_oauth )
 tweets.df <- parseTweets("my_timeline.json")

## End(Not run)
```

readTweets *Converts tweets in JSON format to R list.*

Description

This function parses tweets downloaded using `filterStream`, `sampleStream` or `userStream` and returns a list.

Usage

```
readTweets(tweets, verbose = TRUE)
```

Arguments

`tweets`	A character string naming the file where tweets are stored or the name of the object in memory where the tweets were saved as strings.
`verbose`	logical, default is `TRUE`, which will print in the console the number of tweets that have been parsed.

Details

This function is the first step in the `parseTweets` function and is provided now as an independent function for convenience purposes. In cases where only one field is needed, it can be faster to extract it directly from the JSON data read in R as a list. It can also be useful to extract fields that are not parsed by `parseTweets`, such as hashtags or mentions.

The total number of tweets that are parsed might be lower than the number of lines in the file or object that contains the tweets because blank lines, deletion notices, and incomplete tweets are ignored.

Author(s)

Pablo Barbera <`pablo.barbera@nyu.edu`>

See Also

`parseTweets`.

Examples

```
## The dataset example_tweets contains 10 public statuses published
## by @twitterapi in plain text format. The code below converts the object
## into a list and extracts only the text.

data(example_tweets)
tweets.list <- readTweets(example_tweets)
only.text <- unlist(lapply(tweets.list, [[, text))
## it can be done with an explicit loop:
only.text <- c()
for (i in 1:length(tweets.list)){
   only.text[i] <- tweets.list[[i]][text]
}
print(unlist(only.text))
```

sampleStream — *Connect to Twitter Streaming API and return a small random sample of all public statuses.*

Description

sampleStream opens a connection to Twitter's Streaming API that will return a small random sample of public statuses, around 1% at any given time.

Usage

```
sampleStream(file.name, timeout = 0, tweets = NULL,
  oauth = NULL, verbose = TRUE)
```

Arguments

file.name — string, name of the file where tweets will be written. "" indicates output to the console, which can be redirected to an R object. If the file already exists, tweets will be appended (not overwritten).

timeout — numeric, maximum length of time (in seconds) of connection to stream. The connection will be automatically closed after this period. For example, setting timeout to 10800 will keep the connection open for 3 hours. The default is 0, which will keep the connection open permanently.

tweets — numeric, maximum number of tweets to be collected when function is called. After that number of tweets have been captured, function will stop. If set to NULL (default), the connection will be open for the number of seconds specified in timeout parameter.

oauth — an object of class oauth that contains the access tokens to the user's twitter session. This is currently the only method for authentication. See examples for more details.

verbose — logical, default is TRUE, which generates some output to the R console with information about the capturing process.

Details

For more information, check the documentation at: https://dev.twitter.com/docs/api/1.1/get/statuses/sample

Note that when no file name is provided, tweets are written to a temporary file, which is loaded in memory as a string vector when the connection to the stream is closed.

The total number of actual tweets that are captured might be lower than the number of tweets requested because blank lines, deletion notices, and incomplete tweets are included in the count of tweets downloaded.

Author(s)

Pablo Barbera <pablo.barbera@nyu.edu>

See Also

filterStream, userStream, parseTweets

Examples

```
## Not run:
## capture a random sample of tweets
sampleStream( file.name="tweets_sample.json", user=FOO, password=BAR )

## An example of an authenticated request using the ROAuth package,
## where consumerkey and consumer secret are fictitious.
## You can obtain your own at dev.twitter.com
 library(ROAuth)
 reqURL <- "https://api.twitter.com/oauth/request_token"
 accessURL <- "http://api.twitter.com/oauth/access_token"
 authURL <- "http://api.twitter.com/oauth/authorize"
 consumerKey <- "xxxxxyyyyyzzzzzz"
 consumerSecret <- "xxxxxxyyyyyzzzzzzz111111222222"
  my_oauth <- OAuthFactory$new(consumerKey=consumerKey,
    consumerSecret=consumerSecret, requestURL=requestURL,
    accessURL=accessURL, authURL=authURL)
 my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl"))
 sampleStream( file.name="tweets_sample.json", oauth=my_oauth )


## End(Not run)
```

---

userStream — *Connect to Twitter Streaming API and return messages for a single user.*

Description

userStream opens a connection to Twitter's Streaming API that will return statuses specific to the authenticated user. The output can be saved as an object in memory or written to a text file.

Usage

```
userStream(file.name = NULL, with = "followings",
  replies = NULL, track = NULL, locations = NULL,
  timeout = 0, tweets = NULL, oauth, verbose = TRUE)
```

Arguments

file.name
: string, name of the file where tweets will be written. "" indicates output to the console, which can be redirected to an R object. If the file already exists, tweets will be appended (not overwritten).

with
: string, detault is "followings", which will stream messages from accounts the authenticated user follow. If set to "user", will only stream messages from authenticated user.

  See the with parameter information in the Streaming API documentation for details: https://dev.twitter.com/docs/streaming-apis/parameters#with

replies
: string, default is NULL, which will only stream replies sent by a different user if the authenticated user follows the receiver of the reply. All replies to users that the authenticated user follows will be included if this argument is set to "all".

  See the replies parameter information in the Streaming API documentation for details: https://dev.twitter.com/docs/streaming-apis/parameters#replies `track` string or string vector containing keywords to track. See the track parameter information in the Streaming API documentation for details: `http://dev.twitter.com/docs/streaming-apis/parameters#track`.

`locations` numeric, a vector of longitude, latitude pairs (with the southwest corner coming first) specifying sets of bounding boxes to filter statuses by. See the locations parameter information in the Streaming API documentation for details: `http://dev.twitter.com/docs/streaming-apis/parameters#locations`

`timeout` numeric, maximum length of time (in seconds) of connection to stream. The connection will be automatically closed after this period. For example, setting `timeout` to 10800 will keep the connection open for 3 hours. The default is 0, which will keep the connection open permanently.

`tweets` numeric, maximum number of tweets to be collected when function is called. After that number of tweets have been captured, function will stop. If set to `NULL` (default), the connection will be open for the number of seconds specified in `timeout` parameter.

`oauth` an object of class `oauth` that contains the access tokens to the user's twitter session. This is the only method for authentication available for user streams. See examples for more details.

`verbose` logical, default is `TRUE`, which generates some output to the R console with information about the capturing process.

Details

This function provides access to messages for a single user.

The set of messages to be returned can include the user's tweets and/or replies, and public statuses published by the accounts the user follows, as well to replies to those accounts.

Tweets can also be filtered by keywords and location, using the `track` and `locations` arguments.

The total number of actual tweets that are captured might be lower than the number of tweets requested because blank lines, deletion notices, and incomplete tweets are included in the count of tweets downloaded.

Note that when no file name is provided, tweets are written to a temporary file, which is loaded in memory as a string vector when the connection to the stream is closed.

Author(s)

Pablo Barbera <`pablo.barbera@nyu.edu`>

See Also

`filterStream`, `sampleStream`, `parseTweets`

Examples

```
## Not run:
## The following example shows how to capture a users home timeline
## with the Streaming API and using authentication via the ROAuth
## package, with fictitious consumerkey and consumer secret.
## You can obtain your own at dev.twitter.com
 library(ROAuth)
 requestURL <- "https://api.twitter.com/oauth/request_token"
 accessURL <- "http://api.twitter.com/oauth/access_token"
 authURL <- "http://api.twitter.com/oauth/authorize"
 consumerKey <- "xxxxxyyyyyzzzzzz"
 consumerSecret <- "xxxxxxyyyyyzzzzzzz111111222222"
 my_oauth <- OAuthFactory$new(consumerKey=consumerKey,
    consumerSecret=consumerSecret, requestURL=requestURL,
    accessURL=accessURL, authURL=authURL)
 my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl"))
 save(my_oauth, file="my_oauth")
## Capturing 10 tweets from a users timeline
 userStream( file.name="my_timeline.json", with="followings",
     tweets=10, oauth=my_oauth )

## End(Not run)
```

Index