Twitter client for R

Jeff Gentry

December 30, 2014

1 Introduction

Twitter is a popular service that allows users to broadcast short messages ('*tweets*') for others to read. Over the years this has become a valuable tool not just for standard social media purposes but also for data mining experiments such as sentiment analysis. The *twitteR* package is intended to provide access to the Twitter API within R, allowing users to grab interesting subsets of Twitter data for their analyses.

This document is not intended to be exhaustive nor comprehensive but rather a brief introduction to some of the more common bits of functionality and some basic examples of how they can be used. In the last section I've included a variety of links to people using *twitteR* to solve real world problems.

2 Some Initial Notes

2.1 Support mailing list

While this package doesn't generate a huge volume of emails to me, I have found that the same questions tends to come up repeatedly (often when something has been broken!). I also field requests for advice on practical application of this package which is an area that I'm far from expert at. I've set up a mailing list to better manage emails from users as this way, with the idea being that there'll now be a searchable archive and perhaps other users might be able to chime in. The URL for this mailing list is `http://lists.hexdump.org/listinfo.cgi/twitter-users-hexdump.org`

3 Authentication with OAuth

As of March 2013 OAuth authentication is *required* for all Twitter transactions. You will need to follow these instructions to continue.

OAuth is an authentication mechanism gaining popularity which allows applications to provide client functionality to a web service without granting an end user's credentials to the client itself. This causes a few wrinkles for cases

like ours, where we're accessing Twitter programatically. *twitteR* uses the *httr* package under the hood to manage this.

The first step is to create a Twitter application for yourself. Go to `https://twitter.com/apps/new` and log in. After filling in the basic info, go to the "Settings" tab and select "Read, Write and Access direct messages". Make sure to click on the save button after doing this. In the "Details" tab, take note of your consumer key and consumer secret.

In your R session, you'll want to do the following with the appropriate values from the web page:

```
>    setup_twitter_oauth("API key", "API secret")
```

This will authenticate via *httr*, I recommend looking at that package's `Token` man page for more information regarding how to manage the authentication and caching processes.

If you are in a headless environment or otherwise don't want to deal with the browser based authentication dance, you can get your access token and secret from your apps webpage and call `setup_twitter_oauth` a bit differently:

```
>   setup_twitter_oauth("API key", "API secret", "Access token", "Access secret")
```

4 Getting Started

This document is intended to demonstrate basic techniques rather than an exhaustive tour of the functionality. For more in depth examples I recommend exploring the mailing list or StackOverflow. I've also included some links to examples of `twitteR` being used in the real word at the end.

```
> library(twitteR)

> setup_twitter_oauth("API key", "API secret")
```

5 Exploring Twitter

5.1 Searching Twitter

The `searchTwitter` function can be used to search for tweets that match a desired term. Example searches are such things as hashtags, basic boolean logic such as AND and OR. It is worth looking at `https://dev.twitter.com/docs/using-search` for an example of what can and can not be done here. The `n` argument can be used to specify the number of tweets to return, defaulting to 25. Note that while `searchTwitter` will wrap an arbitrary number of actual search calls to provide the number of tweets requested, the Twitter API has limitations on just how much it will actually return. In general you can only go back a handful of days worth of tweets.

```
>    tweets = searchTwitter('#rstats', n=50)
>    head(tweets)

[[1]]
[1] "hadleywickham: RT @quominus: Scraping the IETF with rvest: http://t.co/KVFMOo1JYR #rsta

[[2]]
[1] "hughparsonage: RT @sharon000: \"Most likely, R became the top statistics package used d

[[3]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 3 - http://t.co/q1CZhfgJT9 via @sjmga

[[4]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 2 - http://t.co/IZGfwrfwtk via @sjmga

[[5]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 1 - http://t.co/CkfIzvRifY via @sjmga

[[6]]
[1] "JonathanAFrye: RT @sharon000: \"Most likely, R became the top statistics package used d
```

There's a handy utility method, `strip_retweets` which attempts to do exactly what it describes. It takes a list of *status* objects (e.g. from `searchTwitter`) and by default will remove official API-based retweets from Twitter, i.e. cases where the retweet button was pressed instead of the manual *RT @soandso .....* method. However there are two arguments, `strip_manual` and `strip_mt`, corresponding to the manual retweets described above and modified retweets (MT). If either of these are `TRUE`, the appropriate type of tweets will also be removed, but leaving everything to the left of the RT/MT.

Note that this example may or may not do anything depending on the data available when this vignette was compiled.

```
>   head(strip_retweets(tweets, strip_manual=TRUE, strip_mt=TRUE))

[[1]]
[1] "JBYoder: Vectorize!\nLet no for() loops evade your eyes!\nRemember why #R gave you all

[[2]]
[1] "ayeimanol: \"ItâĂŹs not just the confidence and drive to act. ItâĂŹs having engraved in

[[3]]
[1] "Rbloggers: Plot with ggplot2 and plotly within knitr reports http://t.co/Xisq4jAkUZ #rs

[[4]]
[1] "ayeimanol: \"SAS is #1âĂęIn Plans to Discontinue Use\" #rstats @myen http://t.co/5uWMNc

[[5]]
```

```
[1] "LearnRinaDay: SAS is #1âĂęIn Plans to Discontinue Use: http://t.co/1KafbigJeO\n#SAS #pr

[[6]]
[1] "daroczig: Only 5 days left to submit your tutorial proposal for @user2015aalborg at #us
```

5.2 Looking at users

To take a closer look at a Twitter user (including yourself!), run the command `getUser`. This will only work correctly with users who have their profiles public, or if you're authenticated and granted access. You can also see things such as a user's followers, who they follow, retweets, and more. The `getUser` function returns a *user* object, which can then be polled for further information.

```
>    crantastic = getUser('crantastic')
>    crantastic$getDescription()

[1] "I like some things, and I dislike everything else."

>    crantastic$getFollowersCount()

[1] 30

>    crantastic$getFriends(n=5)

$`2657140026`
[1] "SilasUniversity"

$`2319087912`
[1] "HeyCarmilla"

$`2797069390`
[1] "Elise3aum"

$`132345262`
[1] "natvanlis"

$`2612008842`
[1] "Laura2theLetter"

>    crantastic$getFavorites(n=5)

[[1]]
[1] "ohnikkers: @keithkurson naw they did two shows in london and i went to both.  i made ha

[[2]]
[1] "mrcartaire: @karmytho Absolutely not.  Amy is definitely not straight."

[[3]]
```

```
[1] "mrcartaire: @saraGG14 @thekatiestevens @greggsulkin @therealritavolk Amy definitely isn

[[4]]
[1] "AnnaKendrick47: Guys I won't be able to see the fireworks from where I am today. Can so

[[5]]
[1] "SmashKarenC: Don't get me wrong, I think it's great that Ivy sleeps with her directors
```

5.3 Conversion to data.frames

There are times when it is convenient to display the object lists as an `data.frame` structure. To do this, every class has a reference method `toDataFrame` as well as a corresponding S4 method `as.data.frame` that works in the traditional sense. Converting a single object will typically not be particularly useful by itself but there is a convenience method to convert an entire list, `twListToDF` which takes a list of objects from a single *twitteR* class:

```
>    df = twListToDF(tweets)
>    head(df)
```

```
1                                                                  RT @quominus: Scraping the IETF
2 RT @sharon000: "Most likely, R became the top statistics package used during summer of thi
3                             RT @moorejh: #rstats vs #Python Round 3 - http://t.co/q1CZhfgJT
4                             RT @moorejh: #rstats vs #Python Round 2 - http://t.co/IZGfwrfwt
5                             RT @moorejh: #rstats vs #Python Round 1 - http://t.co/CkfIzvRif
6 RT @sharon000: "Most likely, R became the top statistics package used during summer of thi
  favorited favoriteCount replyToSN             created truncated replyToSID
1     FALSE             0      <NA> 2014-12-30 23:33:44     FALSE         NA
2     FALSE             0      <NA> 2014-12-30 23:30:35     FALSE         NA
3     FALSE             0      <NA> 2014-12-30 23:30:34     FALSE         NA
4     FALSE             0      <NA> 2014-12-30 23:30:32     FALSE         NA
5     FALSE             0      <NA> 2014-12-30 23:30:30     FALSE         NA
6     FALSE             0      <NA> 2014-12-30 23:27:56     FALSE         NA
                  id replyToUID
1 550072274947043328       <NA>
2 550071480587800576       <NA>
3 550071478142529536       <NA>
4 550071468344635392       <NA>
5 550071459284918272       <NA>
6 550070814239121409       <NA>
                                                                                statusSource
1                        <a href="http://www.echofon.com/" rel="nofollow">Echofon</a>
2  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
3  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
4  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
```

```
5  <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
6 <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
     screenName retweetCount isRetweet retweeted longitude latitude
1 hadleywickham            1      TRUE     FALSE        NA       NA
2 hughparsonage           17      TRUE     FALSE        NA       NA
3    liketree36            8      TRUE     FALSE        NA       NA
4    liketree36            5      TRUE     FALSE        NA       NA
5    liketree36            6      TRUE     FALSE        NA       NA
6 JonathanAFrye           17      TRUE     FALSE        NA       NA
```

5.4 Database Persistence

A question that I'm often asked is how to retrieve data from the past, generally people are doing a study on some major event that has already happened (e.g. Arab Spring, an election, etc). Using the Twitter API this is impossible as you can only go back a small amount. However, if you have the ability to look ahead, it is easy to enable a prospective study by collecting data and automatically persisting it to a database. This will then allow you to load everything into a later R session, including using tools such as *dplyr*. There's a full writeup of this functionality at `http://geoffjentry.blogspot.com/2014/02/twitter-now-supports-database.html`.

Here's a brief example:

```
>   sql_lite_file = tempfile()
>   register_sqlite_backend(sql_lite_file)
>   store_tweets_db(tweets)

[1] TRUE

>   from_db = load_tweets_db()
>   head(from_db)

[[1]]
[1] "hadleywickham: RT @quominus: Scraping the IETF with rvest: http://t.co/KVFMOo1JYR #rsta

[[2]]
[1] "hughparsonage: RT @sharon000: \"Most likely, R became the top statistics package used d

[[3]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 3 - http://t.co/q1CZhfgJT9 via @sjmga

[[4]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 2 - http://t.co/IZGfwrfwtk via @sjmga

[[5]]
[1] "liketree36: RT @moorejh: #rstats vs #Python Round 1 - http://t.co/CkfIzvRifY via @sjmga
```

```
[[6]]
[1] "JonathanAFrye: RT @sharon000: \"Most likely, R became the top statistics package used d
```

5.5 Timelines

A Twitter *timeline* is simply a stream of tweets. We support two timelines, the *user timeline* and the *home timeline*. The former provides the most recent tweets of a specified user while the latter is used to display your own most recent tweets. These both return a list of *status* objects.

To look at a particular user's timeline that user must either have a public account or you must have access to their account. You can either pass in the user's name or an object of class *user* (more on this later). For this example, let's use the user *cranatic*.

```
>    cran_tweets = userTimeline('cranatic')
>    cran_tweets[1:5]

[[1]]
[1] "cranatic: Update: Bchron, BoolNet, caribou, CePa, fmri, HTSCluster, isa2, lessR, lgcp,

[[2]]
[1] "cranatic: New: extrafont, extrafontdb, Rttf2pt1, x12GUI. http://t.co/skyrajMA #rstats"

[[3]]
[1] "cranatic: Update: drc, RcmdrPlugin.survival, rrcov, spls. http://t.co/eEoXNifB #rstats"

[[4]]
[1] "cranatic: New: hzar. http://t.co/eEoXNifB #rstats"

[[5]]
[1] "cranatic: Update: directlabels, forensim, gdata, gWidgetstcltk, gWidgetsWWW, harvestr,
```

By default this command returns the 20 most recent tweet. As with most (but not all) of the functions, it also provides a mechanism to retrieve an arbitrarily large number of tweets up to limits set by the Twitter API, which vary based on the specific type of request.

```
>    cran_tweets_large = userTimeline('cranatic', n=100)
>    length(cran_tweets_large)

[1] 100
```

The `homeTimeline` function works nearly identically except you do not pass in a user, it uses your own timeline.

5.6 Trends

Twitter keeps track of topics that are popular at any given point of time, and allows one to extract that data. The `getTrends` function is used to pull current trend information from a given location, which is specified using a WOEID (see `http://developer.yahoo.com/geo/geoplanet/`). Luckily there are two other functions to help you identify WOEIDs that you might be interested in. The `availableTrendLocations` function will return a `data.frame` with a location in each row and the `woeid` giving that location's WOEID. Similarly the `closestTrendLocations` function is passed a latitude and longitude and will return the same style `data.frame`.

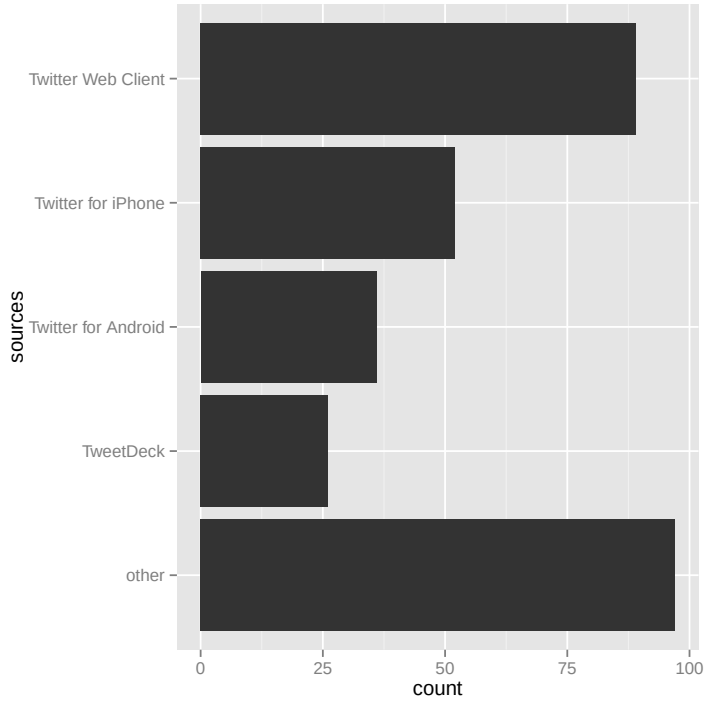
```
>   avail_trends = availableTrendLocations()
>   head(avail_trends)

       name country woeid
1 Worldwide             1
2  Winnipeg  Canada  2972
3    Ottawa  Canada  3369
4    Quebec  Canada  3444
5  Montreal  Canada  3534
6   Toronto  Canada  4118

>   close_trends = closestTrendLocations(-42.8, -71.1)
>   head(close_trends)

        name country  woeid
1 Concepcion   Chile 349860

>   trends = getTrends(2367105)
>   head(trends)

                     name                                                    url
1 #SubscribeToTylerOakley http://twitter.com/search?q=%23SubscribeToTylerOakley
2               New Years           http://twitter.com/search?q=%22New+Years%22
3                #GoIrish                http://twitter.com/search?q=%23GoIrish
4              Notre Dame          http://twitter.com/search?q=%22Notre+Dame%22
5                 Walmart                   http://twitter.com/search?q=Walmart
6           #LeelahAlcorn           http://twitter.com/search?q=%23LeelahAlcorn
                      query   woeid
1 %23SubscribeToTylerOakley 2367105
2           %22New+Years%22 2367105
3                %23GoIrish 2367105
4          %22Notre+Dame%22 2367105
5                   Walmart 2367105
6           %23LeelahAlcorn 2367105
```

5.7 A simple example

Just a quick example of how one can interact with actual data. Here we will pull the most recent results from the public timeline and see the clients that were used to post those statuses. We can look at a plot of the most common clients, as well as seeing hos many others were used.

Note that sources which are not the standard web interface will be presented as an anchored URL string (<A>...</A>). There are more efficient means to rip out the anchor string than how it is done below, but this is a bit more robust for the purposes of this vignette due to issues with character encoding, locales, etc.

```
>    library(ggplot2)
>    r_tweets <- searchTwitter("#rstats", n=300)
>    sources <- sapply(r_tweets, function(x) x$getStatusSource())
>    sources <- gsub("</a>", "", sources)
>    sources <- strsplit(sources, ">")
>    sources <- sapply(sources, function(x) ifelse(length(x) > 1, x[2], x[1]))
>    source_table = table(sources)
>    filtered_sources = names(source_table[source_table < quantile(source_table, 0.9)])
>    sources[sources %in% filtered_sources] = "other"
>    source_df = as.data.frame(sources)
>    ggplot(source_df, aes(sources)) + geom_bar() + coord_flip()
```



6 Examples Of twitteR In The Wild

I've found some examples around the web of people using this package for various purposes, hopefully some of these can give you good ideas on how to do things. Unfortunately I didn't give the package the most easily searched name! If you know of a good example please let me know.

NB: Many of these predate the changes to using the *httr* package, so specifics might have changed, rather view these as examples of things you could do.

- Jeffrey Stanton's free book on data science discusses *twitteR*: `http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf`
- Rare bird siting twitter bot: `https://twitter.com/crd_rare_bird`
- Jeffrey Breen's sentiment analysis example: `http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment`
- Mapping your followers: `http://simplystatistics.org/2011/12/21/an-r-function-to-map-your-twitter-followers/`
- Yangchao Zhao's book on data mining w/ R `http://www.amazon.com/Data-Mining-Examples-Case-Studies/dp/0123969638`
- Gary Miner et al's book on data mining `http://www.amazon.com/Practical-Statistical-Analysis-N dp/012386979X`
- Mining Twitter with R `https://sites.google.com/site/miningtwitter/home`
- Organization or conversation in Twitter: A case study of chatterboxing `https://www.asis.org/asist2012/proceedings/Submissions/185.pdf`

7 Session Information

The version number of R and packages loaded for generating the vignette were:

```
R version 3.1.2 (2014-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggplot2_1.0.0 RSQLite_1.0.0 DBI_0.3.1     twitteR_1.1.8
```

```
loaded via a namespace (and not attached):
 [1] bit_1.1-12       bit64_0.9-4      bitops_1.0-6     colorspace_1.2-4
 [5] digest_0.6.6     grid_3.1.2       gtable_0.1.2     httr_0.6.0
 [9] labeling_0.3     MASS_7.3-35      munsell_0.4.2    plyr_1.8.1
[13] proto_0.3-10     R6_2.0.1         Rcpp_0.11.3      RCurl_1.95-4.5
[17] reshape2_1.4.1   rjson_0.2.15     scales_0.2.4     stringr_0.6.2
[21] tools_3.1.2
```