Northeastern University

Collecting Data via Web APIs in R

Martin Schedlbauer, Ph.D., Brinal Pereira, Aniket Ghodke

This tutorial shows you how to use the Twitter Streaming API to get tweet data using R. You can retrieve Twitter tweet data in JSON format through the Twitter Web API external application after registering the application with Twitter.

Prerequistes

Have R and RStudio successfully installed.

Registering your Application on Twitter

To run this script, you need to generate your own consumer key, consumer secret, access token and access token secret by registering your application with Twitter. It is a simple process.

Step 1: Register it here: https://dev.twitter.com/apps. Go to Create New App.

Step 2: Enter a name for your application. e.g "Data Science NEU"
Enter a description for the app e.g. "Getting tweet data using streaming API in R"
For the website enter a placeholder e.g. "http://www.google.com"
The callback URL field is optional
Scroll down the page. Check the license agreement. And click on "Create your Twitter Application".
Your application has been successfully created.

Step 3: Under the Keys and Access Tokens Tab you will see API Key and API Secret. For the access token and access token secret scroll down the page and click on the button "Create my access token". Refresh the page if necessary and you will now see the the access token and the access token secret.

Note: To check the current working directory use the command getwd() in RStudio.
Once you have the twitter credentials ready, enter the keys in the code given below :

Environment Setup

Step 1: Set the working directory to Desktop using the following line of code.

```
setwd("C:/Users/Brinal/Desktop")
```

Step 2: Load the required libraries needed to use Twitter.

```
library(RCurl)
library(ROAuth)
library(streamR)
library(twitteR)
```

NOTE: If you get an error like this " Error in library(RCurl) : there is no package called 'RCurl' ," it means that the package RCurl has not been installed in the R that you have replicated from the server. In the Console window of RStudio type in the command `install.packages("RCurl")` and hit Enter. The package RCurl will be installed. Repeat the procedure for all the libraries you need to load.

Step 3: Download the certificate needed for authentication. This creates a certificate file on the desktop.

```
download.file(url="http://curl.haxx.se/ca/cacert.pem",
destfile="cacert.pem")
```

Step 4: Create a file to collect all the Twitter JSON data received from the API call.

```
outFile <- "tweets_sample.json"
```

Step 5: Set the configuration details to authorize your application to access Twitter data.

```
requestURL       <- "https://api.twitter.com/oauth/request_token"
accessURL        <- "https://api.twitter.com/oauth/access_token"
authURL          <- "https://api.twitter.com/oauth/authorize"
consumerKey      <- "XXXX"
consumerSecret   <- "XXXX"
accessToken      <- "XXXX"
accessTokenSecret<- "XXXX"
```

The requestURL, accessURL and authURL remain the same. For the consumerKey, consumerSecret, accessToken, acessTokenSecret, fill in the information that was provided when you created a developer's account on Twitter (Registering Your Application on Twitter, Step 3).

Step 6: Authenticate user via OAuth handshake and save the OAuth certificate to the local disk for future connections. OAuth is an authentication protocol that enables a third-party application to obtain limited access to an HTTP service without sharing passwords.

```
my_oauth <- OAuthFactory$new( consumerKey=consumerKey,
                              consumerSecret=consumerSecret,
                              requestURL=requestURL,
                              accessURL=accessURL,
                              authURL=authURL)
my_oauth$handshake(cainfo="cacert.pem")
```

Once the above code is executed, you will be given a link to authorize your application to get Twitter feeds. Copy the link in your browser. Click on "Authorize

MyApplication." You will receive a pin number. Copy the pin number and paste it in the console.

Note: If you are getting errors at this point, double check that the values for the consumer key, consumer secret, access token, and access token secret are correct. If you are still getting errors, reinstall all the required packages with in the `install.packages` command then restart R.

Step 7: After your application has been authorized, you will need to register your credentials by setting up the OAuth credentials for a Twitter session.

```
setup_twitter_oauth(consumerKey, consumerSecret, accessToken,
accessTokenSecret)
```

Then press 1 in the console to allow the file to access the credentials.

Getting Tweet Data

Step 1: You can now start getting tweet data. The sampleStream() function in the streamR package opens a connection to Twitter's Streaming API that returns a random sample of public statuses and outputs a JSON file.

```
sampleStream( file=outFile, oauth=my_oauth, tweets=100 )
```

Step 2: The filterStream() function allows for more specific filtering, for example, search for "Boston" or "RedSox" at a certain geolocation, and time out after 5 seconds.

```
follow   <- ""
track    <- "Boston,RedSoxs"
location <- c(23.786699, 60.878590, 37.097000, 77.840813)
filterStream( file.name=outFile, follow=follow, track=track,
     locations=location, oauth=my_oauth, timeout=5)
```

Source Code

Please locate the source code in the file "TwitterAPI.R".