Collecting Twitter Data and Analyzing it as Text

Lars Hinrichs

2020-04-01

Contents

Introduction

Collecting and analyzing tweets, all within R, has been getting much, much easier in the last couple of years. We now have the package `rtweet` and it makes all the difference. I won’t go in detail here about all the improvements this package offers compared to earlier packages, but its authors mention a few on the package website.

Let’s get to it!

Collecting tweets

Before we get started

You will need your own personal **Twitter account with a user name and password**.

Load packages

Let’s start with some packages that we’ll need.‘

```
if (!require(pacman)) library(pacman)
p_load(rtweet, tidyverse, rio, janitor)
```

Collect tweets ("harvesting")

The core function of `rtweet` is `search_tweets()`. All of its arguments are described in the help manual and on this page. Using that information, let us clobber together an initial search. First, we'll define a keyword to search for.

```
q = "virus"
```

We can now use this keyword as an argument to `search_tweets()`. Notice that I am setting the `include_rts` argument to `FALSE` because I want only original tweets, no retweets, among my results. In addition, I am piping the results I am getting into a call to `select()` so that I'll see the text of each tweet and the writer's screen name, but none of the other information that also gets collected.

```
search_tweets(
  q,
  n = 70,
  include_rts = FALSE
  ) %>%
  select(text, screen_name)
```

This was just a search to see how `rtweet` works - we did not store any data in a variable yet.

Let us try to set up a serious search. Why don't we compare the way that "virus" is being talked about in the US and in the UK? (Our machinery is tuned to English-language material, that is why I am sticking with English-speaking locales.)

And we'll also re-use `q`, our query, which is "virus". Running our search then, we should be sure to give the resulting data a variable name so that we can analyze it later. - I will ask for a total of 10,000 tweets.

```
mytweets <- search_tweets(
  q,
  n = 10000,
  include_rts = FALSE
  )
```

Let's ask how many rows are in our resulting dataset - in other words, how many tweets we got:

```
mytweets %>% nrow()
```

```
## [1] 7997
```

So we do in fact get (about) 10,000 tweets. By using the geo-information contained in the data (it's in the format of an R dataframe), we can see how many tweets we got from the US and how many from the UK.

```
mytweets %>%
  tabyl(country) %>%
  select(1, 2) %>%
  arrange(-n) %>%
  head(10)
```

```
##           country    n
## 1                  7772
## 2   United States   48
## 3           India   32
## 4  United Kingdom   25
## 5       Indonesia   18
## 6          Brazil   15
## 7    South Africa   11
## 8         Nigeria    8
## 9           Italy    7
## 10          Spain    7
```

It looks like we are getting a fair number of tweets from the US and the UK. However, if we wanted a larger corpus of tweets for our study, we'd have to collect a lot more data. I won't do this here, as I am only showing how this method works, but I encourage you to put in the time to collect, let's say, 50,000 tweets. Here is how you would collect that many. Since there is a limit to how many you can download, we need to use the `retryonratelimit` argument (setting it to `TRUE`).

```
mytweets <- search_tweets(
  q,
  n = 50000,
  retryonratelimit = TRUE,
  include_rts = FALSE
)
```

Once you're done collecting the larger amount of data, apply a few filters to be sure that all your data is relevant:

- keep only tweets that were written in the UK and the US,
- keep only tweets in English, and
- remove duplicates, since some tweets will have been harvested twice.

```
df <- mytweets %>%
  filter(country %in% c("United Kingdom", "United States"),
         lang == "en") %>%
  unique()
df %>%
```

```
  select(text, country) %>%
  head()
```

```
##
## 1                    @techdino @MarkDice Viruses are spread through bodily fluids.  A virus cannot li
## 2
## 3                      The possible exposure to #COVID19 for those not present for patient care?  If
## 4 Hey y'all, I am a pharmacy technician and we are on the front end of the virus! If you could sign t
## 5
## 6                                 @RealDealAxelrod Omg I may have to check if my son's Doctor gave hir
##          country
## 1  United States
## 2 United Kingdom
## 3  United States
## 4  United States
## 5 United Kingdom
## 6  United States
```

Since # Analyzing tweets

Tidy format

Wordclouds

Sentiment analysis