Text analysis workshop: Word clouds (w/ Twitter data mining)

Jessica Couture

Twitter

In this tutorial we'll go through a rough workflow for accessing Twitter data then construct a word cloud from some of the data pulled from the Twitter API.

```
library(tidyverse)
library(tidytext)
library(rtweet)
library(wordcloud)
```

To access Twitter through R, you need the **rtweet** package and some set-up steps through the Twitter Apps website, including applying for a developer account and creating a new app. Here is a more step-by-step tutorial of how to get your account and tokens.

Apply for Twitter developer account

To access the Twitter API, you have to first apply for a developer account through Twitter (providing phone number is required for this account). Once you are approved for an account (which doesn't seem to take much more than verifying your email address), you can 'create a new app', which will provide you with the tokens needed to access the API through R. You'll need internet access to access your tokens and the Twitter API when in R.

I am going to skip a few steps here to keep my tokens hidden but once you get your tokens paste them into code such as:

```
## autheticate via web browser
token <- create_token(
  app = "rtweetToken",
  consumer_key = "Wcxd1...NBY2E",
  consumer_secret = "1sTDb2Xz...ld4jv6jasve3h56bY")
## click 'Authorize' on the pop-up browser window
```

A window will pop-up in your browser to accept access to your account using the specified tokens, click *'Authorize app'*

Search tweets

Now that we have access to Twitter data, we can use `search_tweets()` to search through tweets. This function returns a dataframe with 88 columns of metadata about each tweet (incl. name of account, retweet information, the actual text, etc.). The main arguments of this function allow you to define a search term and number of tweets to return.

This query searched **"#blackhistorymonth"** and limited the response to **500 tweets**.

```
bhm_tweets <- search_tweets(q = "#blackhistorymonth", n = 500)

# bhm_tweets2<-bhm_tweets %>%
```

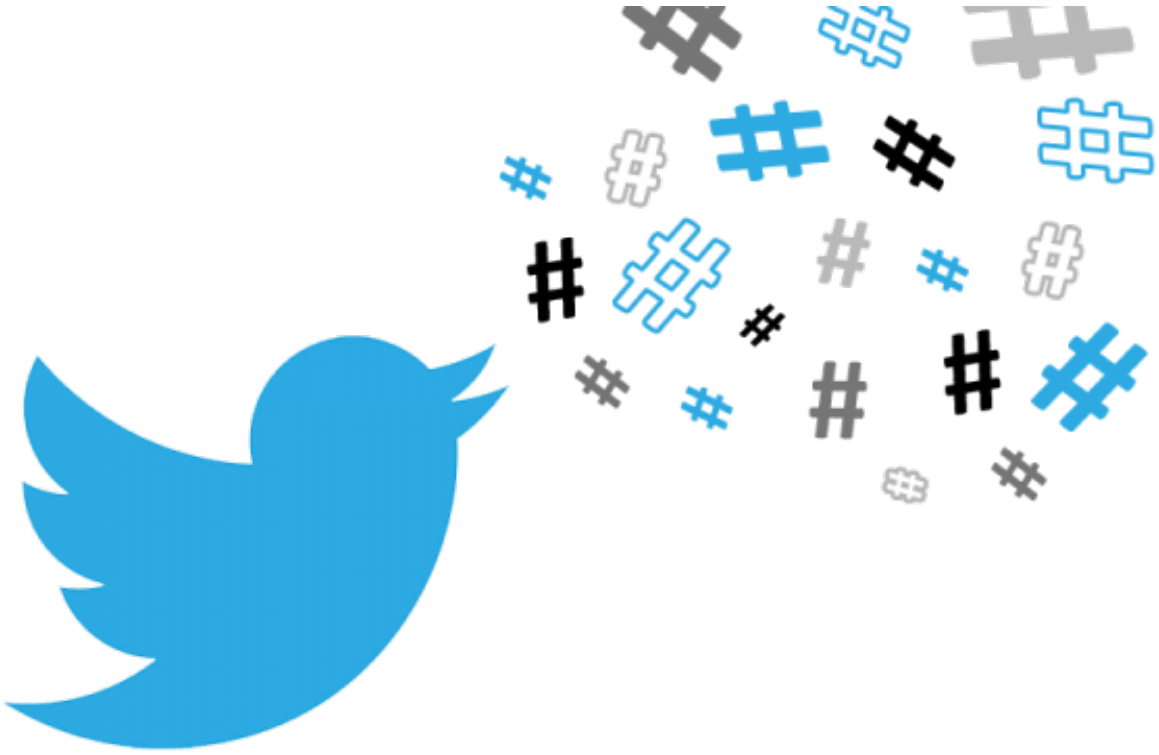


Figure 1:

```
#   select(text) # keep just the tweet text content
#
# write.csv(bhm_tweets2, file="bhmTweets.csv",row.names=F)

colnames(bhm_tweets)
```

```
##  [1] "user_id"                "status_id"
##  [3] "created_at"             "screen_name"
##  [5] "text"                   "source"
##  [7] "display_text_width"     "reply_to_status_id"
##  [9] "reply_to_user_id"       "reply_to_screen_name"
## [11] "is_quote"               "is_retweet"
## [13] "favorite_count"         "retweet_count"
## [15] "hashtags"               "symbols"
## [17] "urls_url"               "urls_t.co"
## [19] "urls_expanded_url"      "media_url"
## [21] "media_t.co"             "media_expanded_url"
## [23] "media_type"             "ext_media_url"
## [25] "ext_media_t.co"         "ext_media_expanded_url"
## [27] "ext_media_type"         "mentions_user_id"
## [29] "mentions_screen_name"   "lang"
## [31] "quoted_status_id"       "quoted_text"
## [33] "quoted_created_at"      "quoted_source"
## [35] "quoted_favorite_count"  "quoted_retweet_count"
## [37] "quoted_user_id"         "quoted_screen_name"
## [39] "quoted_name"            "quoted_followers_count"
```

Figure 2:

```
## [41] "quoted_friends_count"    "quoted_statuses_count"
## [43] "quoted_location"         "quoted_description"
## [45] "quoted_verified"         "retweet_status_id"
## [47] "retweet_text"            "retweet_created_at"
## [49] "retweet_source"          "retweet_favorite_count"
## [51] "retweet_retweet_count"   "retweet_user_id"
## [53] "retweet_screen_name"     "retweet_name"
## [55] "retweet_followers_count" "retweet_friends_count"
## [57] "retweet_statuses_count"  "retweet_location"
## [59] "retweet_description"     "retweet_verified"
## [61] "place_url"               "place_name"
## [63] "place_full_name"         "place_type"
## [65] "country"                 "country_code"
## [67] "geo_coords"              "coords_coords"
## [69] "bbox_coords"             "status_url"
## [71] "name"                    "location"
## [73] "description"             "url"
## [75] "protected"               "followers_count"
## [77] "friends_count"           "listed_count"
## [79] "statuses_count"          "favourites_count"
## [81] "account_created_at"      "verified"
## [83] "profile_url"             "profile_expanded_url"
## [85] "account_lang"            "profile_banner_url"
## [87] "profile_background_url"  "profile_image_url"
```

Word Cloud

Tidying

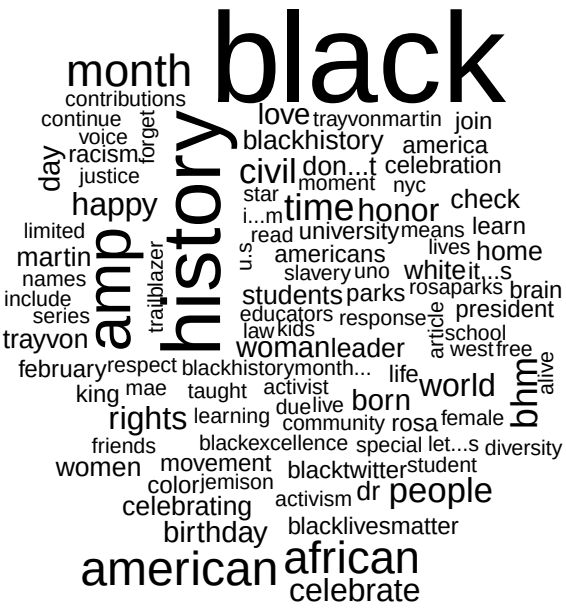
Let's do a bit of tidying to get these data into a format that works with the `wordcloud()` function, which likes a data frame of words with a separate column for frequency. So from the twitter output we only want to look at the "text" column, which contains the actual tweets:

```
data("stop_words")

bhm<-bhm_tweets %>%
  filter(!duplicated(text)) %>%
  unnest_tokens(word,text) %>%
  select(word) %>%
  filter(!word=="#BlackHistoryMonth",!word=="blackhistorymonth",!word=="https",!word=="t.co") %>%
  filter(!str_detect(word, '[0-9]')) %>%
  group_by(word) %>%
  summarize(nWords=n()) %>%
  anti_join(stop_words) %>%
  arrange(-nWords)

head(bhm)
```

```
## # A tibble: 6 x 2
##   word     nWords
##   <chr>     <int>
## 1 black        89
## 2 history      52
## 3 amp          38
## 4 american     25
## 5 african      23
## 6 month        23
```

Constuct your cloud

```
# plot the 100 most common words

wordcloud(bhm$word, bhm$nWords, max.words = 100)
```

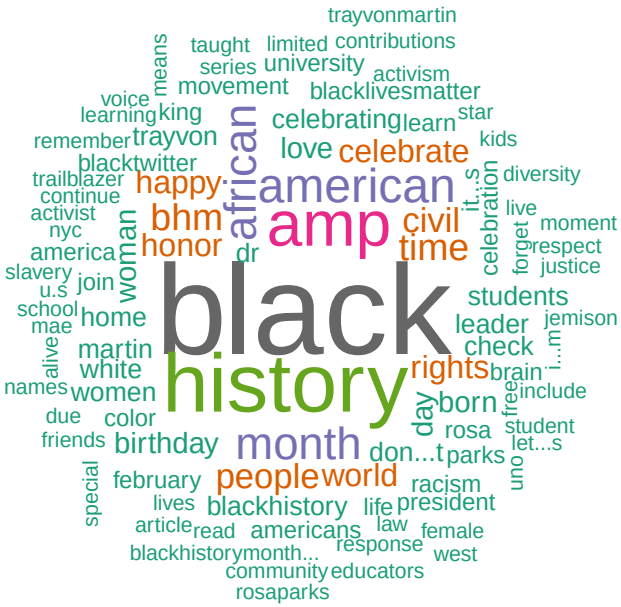


This is a bare-bones word cloud of the words we used, but we can use `random.order=FALSE` to give the cloud more structure:

```
wordcloud(bhm$word, bhm$nWords, random.order = FALSE, max.words = 100)
```

. . . and `colors` to add color:

```
# define a nice color palette
pal <- brewer.pal(9,"Dark2")


wordcloud(bhm$word, bhm$nWords, random.order = FALSE, max.words = 100, colors=pal)
```




```
# bhm %>%
#   with(wordcloud(word, nWords, random.order = FALSE, max.words = 100, colors=pal))
```

```
bhm %>%
  with(wordcloud(word, nWords, random.order = FALSE, max.words = 100, colors=pal,random.color = T))
```