Lab One - Text Analysis

Derek Wales

23JAN20

Question One:

Creating a token using the creat token function

```
# Commmented out do to knitting issues
# create_token(app = "udemysucks",consumer_key, consumer_secret)
```

Question Two:

Utilizing the get_timeline() function, collect the most recent 250 tweets from @espn. Inspect the contents of those tweets.

```
espn_tweets = get_timeline("espn", n = 250, max_id = NULL, home = FALSE, parse = TRUE,
                          check = TRUE, token = NULL)

head(espn_tweets)
```

```
## # A tibble: 6 x 90
##   user_id status_id created_at          screen_name text  source
##   <chr>   <chr>     <dttm>              <chr>       <chr> <chr>
## 1 2557521 12205528~ 2020-01-24 03:44:08 espn        "ALL~ Khoros
## 2 2557521 12205485~ 2020-01-24 03:27:11 espn        Sere~ Tweet~
## 3 2557521 12205482~ 2020-01-24 03:26:01 espn        "Bro~ Khoros
## 4 2557521 12205387~ 2020-01-24 02:48:09 espn        "Aft~ Tweet~
## 5 2557521 12205332~ 2020-01-24 02:26:29 espn        ".@D~ Khoros
## 6 2557521 12205262~ 2020-01-24 01:58:28 espn        "No.~ Khoros
## # ... with 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <lgl>, reply_to_user_id <lgl>,
## #   reply_to_screen_name <lgl>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>,
## #   ext_media_expanded_url <list>, ext_media_type <chr>,
## #   mentions_user_id <list>, mentions_screen_name <list>, lang <chr>,
## #   quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>,
## #   quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>,
## #   quoted_screen_name <chr>, quoted_name <chr>,
## #   quoted_followers_count <int>, quoted_friends_count <int>,
## #   quoted_statuses_count <int>, quoted_location <chr>,
## #   quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>,
## #   retweet_created_at <dttm>, retweet_source <chr>,
## #   retweet_favorite_count <int>, retweet_retweet_count <int>,
## #   retweet_user_id <chr>, retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>,
## #   country <chr>, country_code <chr>, geo_coords <list>,
## #   coords_coords <list>, bbox_coords <list>, status_url <chr>,
## #   name <chr>, location <chr>, description <chr>, url <chr>,
## #   protected <lgl>, followers_count <int>, friends_count <int>,
## #   listed_count <int>, statuses_count <int>, favourites_count <int>,
## #   account_created_at <dttm>, verified <lgl>, profile_url <chr>,
## #   profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>
```

```
#names(espn_tweets)
```

Question Three:

The case for most social media is very few individuals that use a platform actually directly create content, in this case 'tweet'. Therefore many users will use other functions such as favorite or retweet other user's content. As such you may want to see what other ways people are participating in a platform. Utilizing the get_favorites() function, gather @PatrickMahomes favorites/likes.

Note: limited N = 500

```
pat_mahomes = get_favorites("PatrickMahomes", n = 500, since_id = NULL, max_id = NULL,
                            parse = TRUE, token = NULL)

head(pat_mahomes)
```

```
## # A tibble: 6 x 91
##   user_id status_id created_at          screen_name text  source
##   <chr>   <chr>     <dttm>              <chr>       <chr> <chr>
## 1 312419~ 12197636~ 2020-01-21 23:28:05 GehrigDiet~ I ne~ Twitt~
## 2 312419~ 12027421~ 2019-12-06 00:10:59 GehrigDiet~ @Pat~ Twitt~
## 3 312419~ 11157391~ 2019-04-09 22:11:47 GehrigDiet~ Not ~ Twitt~
## 4 312419~ 11548017~ 2019-07-26 17:12:42 GehrigDiet~ @Pat~ Twitt~
## 5 312419~ 11215458~ 2019-04-25 22:45:34 GehrigDiet~ Cong~ Twitt~
## 6 312419~ 11454010~ 2019-06-30 18:37:49 GehrigDiet~ @Pat~ Twitt~
## # ... with 85 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>,
## #   ext_media_expanded_url <list>, ext_media_type <chr>,
## #   mentions_user_id <list>, mentions_screen_name <list>, lang <chr>,
## #   quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>,
## #   quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>,
## #   quoted_screen_name <chr>, quoted_name <chr>,
## #   quoted_followers_count <int>, quoted_friends_count <int>,
## #   quoted_statuses_count <int>, quoted_location <chr>,
## #   quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>,
## #   retweet_created_at <dttm>, retweet_source <chr>,
## #   retweet_favorite_count <int>, retweet_retweet_count <int>,
## #   retweet_user_id <chr>, retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>,
## #   country <chr>, country_code <chr>, geo_coords <list>,
## #   coords_coords <list>, bbox_coords <list>, status_url <chr>,
## #   name <chr>, location <chr>, description <chr>, url <chr>,
## #   protected <lgl>, followers_count <int>, friends_count <int>,
## #   listed_count <int>, statuses_count <int>, favourites_count <int>,
## #   account_created_at <dttm>, verified <lgl>, profile_url <chr>,
## #   profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>, favorited_by <chr>
```

Question Four:

As with favorites, many users retweet, essentially sharing content other users have made utilizing the get_retweet() get_retweeters() functions to gather the retweets of this Patrick Mahomes' tweet (status: 857944539029483520).

```
pat_mahomes_retweet = get_retweeters("857944539029483520", n = 1000, parse = TRUE,
                              token = NULL)

#dim(pat_mahomes_retweet)
head(pat_mahomes_retweet)
```

```
##               user_id
## 1 1204431973990780930
## 2 1207031863048429568
## 3           199909841
## 4          1695342242
## 5  806263468819152896
## 6           385657504
```

Question Five:

Social media is very good at giving insight on how users interaction between themselves. Later in the course we will touch on social network analysis, but there are a number of functions that can help us build a network once we aquire those skills. Part of understanding a network is to look at who follows who, essentially who is being given content from a user directly. Utilizing the get_followers() and get_friends() functions, collect @PatrickMahome followers, and who he follows respectively.

```
pat_mahomes_followers = get_followers("PatrickMahomes", n = 5000)
pat_mahomes_friends = get_friends("PatrickMahomes", n = 5000)

head(pat_mahomes_followers)
```

```
## # A tibble: 6 x 1
##   user_id
##   <chr>
## 1 1220558231216566274
## 2 1220557474811478017
## 3 843108080934117376
## 4 1220559264290430977
## 5 43744453
## 6 878490965383589888
```

```
head(pat_mahomes_friends)
```

```
## # A tibble: 6 x 2
##   user           user_id
##   <chr>          <chr>
## 1 PatrickMahomes 39440317
## 2 PatrickMahomes 537186173
## 3 PatrickMahomes 757775576
## 4 PatrickMahomes 15955515
## 5 PatrickMahomes 55307193
## 6 PatrickMahomes 2640922334
```

Question Six:

Sometimes when using social media you're not exactly sure what you're looking for directly. Having specific users names, or content IDs is very helpful, but not always directly available without some prior research. Twitter specifically has functions that allow you to look up what topic are popular in certain regions, which can lead to more specific information about a specific item you're looking for. Utilizing the get_trends() function, identify trends that are happening in North Carolina.

```
us_trends = get_trends(woeid = "23424977")
head(us_trends)
```

```
## # A tibble: 6 x 9
##   trend url   promoted_content query tweet_volume place  woeid
##   <chr> <chr> <lgl>            <chr>        <int> <chr>  <int>
## 1 #Rig~ http~ NA               %23R~           NA Unit~ 2.34e7
## 2 #YAN~ http~ NA               %23Y~        19076 Unit~ 2.34e7
## 3 #iubb http~ NA               %23i~           NA Unit~ 2.34e7
## 4 #Imp~ http~ NA               %23I~        13296 Unit~ 2.34e7
## 5 #Gre~ http~ NA               %23G~        15307 Unit~ 2.34e7
## 6 Jare~ http~ NA               %22J~           NA Unit~ 2.34e7
## # ... with 2 more variables: as_of <dttm>, created_at <dttm>
```

Question Seven:

If trends or stream don't work in content collection, you can search tweets using key words or hashtags. With our basic account, we are only allow to search tweets made in the last 6-9 days, and 18,000 total at a time. Utilizing the search_tweets() function, collect 5000 tweets containing the keywords 'superbowl' and 'sbliv'

```
super_bowl_tweets = search_tweets("superbowl OR sbliv", n =5000)
head(super_bowl_tweets)
```

```
## # A tibble: 6 x 90
##   user_id status_id created_at          screen_name text  source
##   <chr>   <chr>     <dttm>              <chr>       <chr> <chr>
## 1 154232~ 12205594~ 2020-01-24 04:10:29 Trevor_ebo~ if t~ Twitt~
## 2 350760~ 12205594~ 2020-01-24 04:10:28 Moana42     "Wor~ Twitt~
## 3 369822~ 12205594~ 2020-01-24 04:10:21 johnavalos~ @kai~ Twitt~
## 4 231625~ 12205594~ 2020-01-24 04:10:21 A8DULRHMAN  "<U+062D><U+064A><U+0627>~ Twitt~
## 5 116107~ 12205594~ 2020-01-24 04:10:18 CarlosOla5  "Com~ Twitt~
## 6 599284~ 12205594~ 2020-01-24 04:10:14 MNIMN_      SUPE~ Twitt~
## # ... with 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>,
## #   ext_media_expanded_url <list>, ext_media_type <chr>,
## #   mentions_user_id <list>, mentions_screen_name <list>, lang <chr>,
## #   quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>,
## #   quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>,
## #   quoted_screen_name <chr>, quoted_name <chr>,
## #   quoted_followers_count <int>, quoted_friends_count <int>,
## #   quoted_statuses_count <int>, quoted_location <chr>,
## #   quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>,
## #   retweet_created_at <dttm>, retweet_source <chr>,
## #   retweet_favorite_count <int>, retweet_retweet_count <int>,
## #   retweet_user_id <chr>, retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>,
## #   country <chr>, country_code <chr>, geo_coords <list>,
## #   coords_coords <list>, bbox_coords <list>, status_url <chr>,
## #   name <chr>, location <chr>, description <chr>, url <chr>,
## #   protected <lgl>, followers_count <int>, friends_count <int>,
## #   listed_count <int>, statuses_count <int>, favourites_count <int>,
## #   account_created_at <dttm>, verified <lgl>, profile_url <chr>,
## #   profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>
```

Question Eight:

Similarly, you can search straight from the stream as posts are being created. You can choose a random sampling or narrow it down by user_id or key word. Without narrowing down the stream, a lot of random content, much without context will be found. Utlizing the stream_tweets() function, collect 2 minutes worth of tweets.

```
streamed_tweets_23JAN20 = stream_tweets(q="", timeout = 120)
```

```
## Streaming tweets for 120 seconds...
```

```
## Finished streaming tweets!
```

```
head(streamed_tweets_23JAN20)
```

```
## # A tibble: 6 x 90
##   user_id status_id created_at          screen_name text  source
##   <chr>   <chr>     <dttm>              <chr>       <chr> <chr>
## 1 965929~ 12205595~ 2020-01-24 04:10:57 Octis_Rega~ @Kos~ Twitt~
## 2 911866~ 12205595~ 2020-01-24 04:10:57 inamap2the~ <U+30D1><U+30E9><U+30A2><U+30B9>~ Twitt~
## 3 114188~ 12205595~ 2020-01-24 04:10:58 ayachan124~ "<U+5143><U+604B><U+4EBA>~ Twitt~
## 4 138863~ 12205595~ 2020-01-24 04:10:58 ospococo    meid~ Twitt~
## 5 997060~ 12205595~ 2020-01-24 04:10:58 pei0811     "@ma~ Twitt~
## 6 143203~ 12205595~ 2020-01-24 04:10:58 williamsdi~ Hay ~ Twitt~
## # ... with 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>,
## #   ext_media_expanded_url <list>, ext_media_type <chr>,
## #   mentions_user_id <list>, mentions_screen_name <list>, lang <chr>,
## #   quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>,
## #   quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>,
## #   quoted_screen_name <chr>, quoted_name <chr>,
## #   quoted_followers_count <int>, quoted_friends_count <int>,
## #   quoted_statuses_count <int>, quoted_location <chr>,
## #   quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>,
## #   retweet_created_at <dttm>, retweet_source <chr>,
## #   retweet_favorite_count <int>, retweet_retweet_count <int>,
## #   retweet_user_id <chr>, retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>,
## #   country <chr>, country_code <chr>, geo_coords <list>,
## #   coords_coords <list>, bbox_coords <list>, status_url <chr>,
## #   name <chr>, location <chr>, description <chr>, url <chr>,
## #   protected <lgl>, followers_count <int>, friends_count <int>,
## #   listed_count <int>, statuses_count <int>, favourites_count <int>,
## #   account_created_at <dttm>, verified <lgl>, profile_url <chr>,
## #   profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>
```

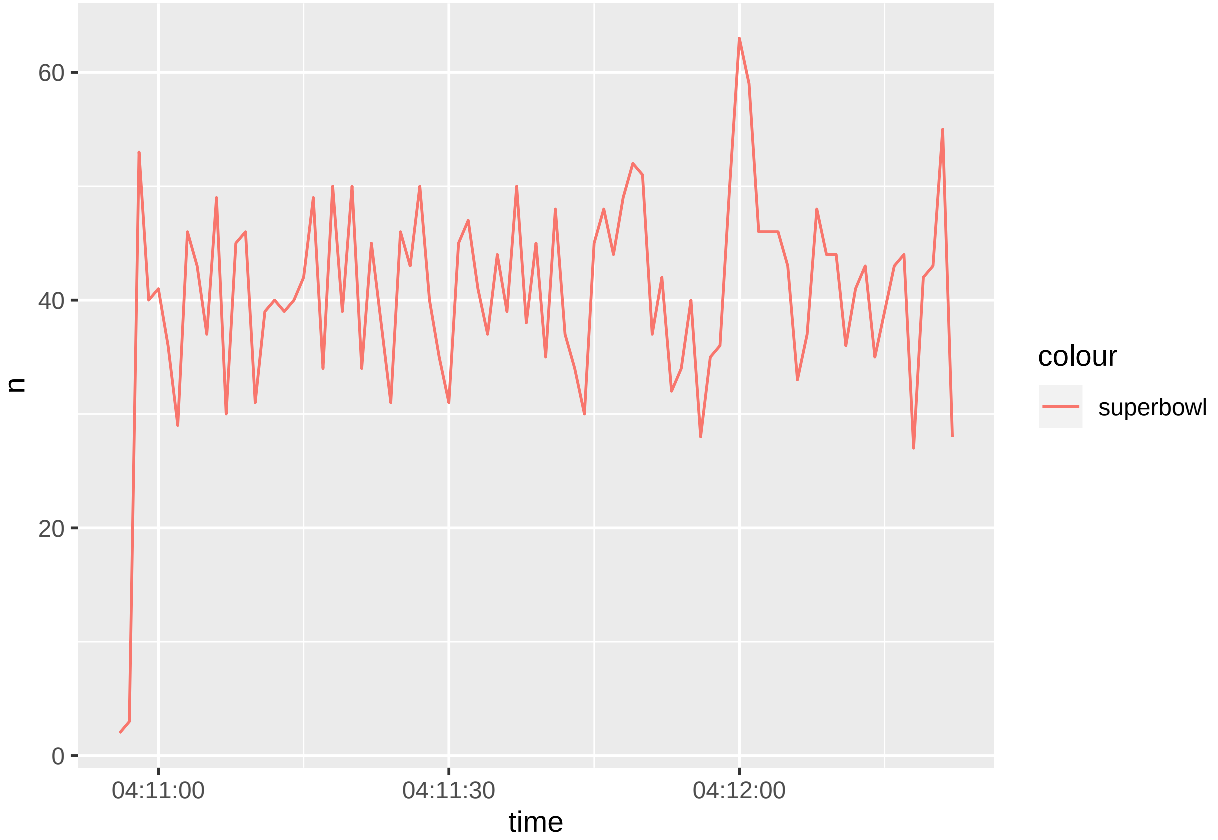
Question Nine:

Because tweets are timestamped, we can plot content to determine patterns or relevancy of post. Utilizng the ts_data() and ts_plot() functions, plot the content of from question 7 concening the super bowl.

```
streamed_data = ts_data(streamed_tweets_23JAN20, by = "secs")
streamed_tweets_23JAN20 %>%
  dplyr::group_by("superbowl") %>%
  ts_plot("secs")
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```



### Question Ten: Another interesting function is geotagging, since most tweets are made from cellular phones, their data can contain positiona data that can be tracked. Typically this opted out, but about 10% of posts are geotagged. Utilizng the Map package in R, plot the geotagged tweets from question 7.

```
geocoded <- lat_lng(streamed_tweets_23JAN20)
par(mar = c(0, 0, 0, 0))
maps::map("state", lwd = .25)
with(geocoded, points(lng, lat, pch = 20, cex = .75, col = rgb(0, .3, .7, .75)))
```