

Analysing Twitter for Ubisoft

Ryan Greenup

May 4, 2020

Contents

8.1 Analysing the Relationship Between Friends and Followers for Twitter Users	1
8.1.1 Retrieve the posts from Twitter	1
8.2.2 Count of Followers and Friends	1
8.1.3 Summary Statistics	1
8.1.4 Above Average Followers	2
8.1.5 Bootstrap confidence intervals	2
.1 a/b.) Generate a bootstrap distribution	2
.2 c.) Estimate a Confidence Interval for the population mean Follower Counts . . .	6
.3 d.) Estimate a Confidence Interval for the population mean Friend Counts	7
FIXME 8.1.6 Estimate a 97% Confidence Interval for the High Friend Count Proportion	7
8.1.7 Is the Number of Friends Independent to the Number of Followers	10
.1 Bin the Follower and Friend Categories	10
.2 Find the Group frequency	11
.3 Find the Expected Counts under each group and test for independence	11
.4 FIXME Conclusion	14
8.2 Finding Themes in tweets	14
8.2.8 Find Users with Above Average Friend Counts	14
8.2.9 Find Users with Below Average Friend Counts	15
8.2.10 Find the <i>Tweets</i> corresponding to users with high or low friend counts	15
8.2.11 Clean the tweets	15
.1 Create a Corpus Object	15
.2 Process the tweets	17
8.2.12 Display the first two tweets before/after processing	19
8.2.13 Create a Term Document Matrix	19
.1 Apply Weighting Manually	19
.2 Apply Weighting with Built in Function	20
.3 Remove Empty Documents	20
8.2.14 How many Clusters are there	21
.1 Use Cosine Distance	21
.2 Project into Euclidean Space	22
.3 Measure Within Cluster Variance	22

Appendix	24
Users with High Friend Count	24
Users with Low Friend Count	24
TF-IDF Matrix	24
Relevant XKCD	24

8.1 Analysing the Relationship Between Friends and Followers for Twitter Users

8.1.1 Retrieve the posts from Twitter

relevant posts can be retrieved from twitter by utilising the `rtweet` package, packages can be loaded for use in **R** thusly:

```

1  # Load Packages
   ↪ -----
2  setwd("~/Dropbox/Notes/DataSci/Social_Web_Analytics/SWA-Project/scripts")
   ↪ "/"
3
4  if (require("pacman")) {
5    library(pacman)
6  } else{
7    install.packages("pacman")
8    library(pacman)
9  }
10
11  pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2,
12                ggmap, parallel, dplyr, plotly, tidyverse,
13                reticulate, UsingR, Rmpfr, swirl, corrplot,
14                gridExtra, mise, latex2exp, tree, rpart,
15                lattice, coin, primes, epitools, maps, clipr,
16                ggmap, twitterR, ROAuth, tm, rtweet, base64enc,
17                httpuv, SnowballC, RColorBrewer, wordcloud,
18                ggwordcloud, tidyverse, boot)

```

Listing 1: Load the Packages for **R**

The `rtweet` API will search for tweets that contain all the words of a query regardless of uppercase or lowercase usage [5].

In order to leverage the *Twitter* API it is necessary to use tokens provided through a *Twitter* developer account:

and hence all tweets containing a mention of *Ubisoft* can be returned and saved to disk as shown in listing 3:

```

1  # Set up Tokens
   ↪ =====
2
3  options(RCurlOptions = list(
4    verbose = FALSE,
5    capath = system.file("CurlSSL", "cacert.pem", package = "RCurl"),
6    ssl.verifypeer = FALSE
7  ))
8
9  setup_twitter_oauth(
10     consumer_key = "*****",
11     consumer_secret =
12     ↪ "*****",
13     access_token = "*****",
14     access_secret = "*****"
15   )
16
17  # rtweet
   ↪ =====
18  tk <- rtweet::create_token(
19     app = "SWA",
20     consumer_key = "*****",
21     consumer_secret =
22     ↪ "*****",
23     access_token =
24     ↪ "*****",
25     access_secret = "*****",
26     set_renv = FALSE

```

Listing 2: Import the twitter tokens (redacted)

```

1  n <- 1000
2  tweets.company <- search_tweets(q = 'ubisoft', n = n, token = tk,
3                                include_rts = FALSE)
4  save(tweets.company[,], file = "resources/Download_1.Rdata")

```

Listing 3: Save the Tweets to the HDD as an rdata file

8.2.2 Count of Followers and Friends

In order to identify the number of users that are contained in the *tweets* the `unique()` function can be used to return a vector of names which can then be passed as an index to the vector of counts as shown in listing 4, this provides that 81.7% of the tweets are by unique users.

```
1 (users <- unique(tweets.company$name)) %>% length()
2 x <- tweets.company$followers_count[duplicated(tweets.company$name)]
3 y <- tweets.company$friends_count[duplicated(tweets.company$name)]
4
5 ## > [1] 817
```

Listing 4: Return follower count of twitter posts

8.1.3 Summary Statistics

The average number of friends and followers from users who posted tweets mentioning *Ubisoft* can be returned using the `mean()` as shown in listing 5 this provides that on average each user has 586 friends and 63,620 followers.

```
1 x<- rnorm(090)
2 y<- rnorm(090)
3 (xbar <- mean(x))
4 (ybar <- mean(y))
5
6 ## > [1] 4295.195
7 ## > [1] 435.9449
```

Listing 5: Determine the average number of friends and followers

8.1.4 Above Average Followers

Each user can be compared to the average number of followers, by using a logical operator on the vector (e.g. `y > ybar`), this will return an output of logical values. *R* will coerce logicals into 1/0 values meaning that the mean value will return the proportion of TRUE responses as shown in listing 6. This provides that:

- 2.4% of the have identified have an above average **number of followers**.
- 20.6% of the users identified have an above average **number of friends**.

```

1 (px_hat <- mean(x>xbar))
2 (py_hat <- mean(y>ybar))
3
4 ## > [1] 0.0244798
5 ## > [1] 0.2729498

```

Listing 6: Calculate the proportion of users with above average follower counts

8.1.5 Bootstrap confidence intervals

a/b.) Generate a bootstrap distribution

A bootstrap assumes that the population is an infinitely large repetition of the sample and may be produced with respect to follower counts by resampling with replacement/repetition and plotted using the ggplot2 library as demonstrated in listings 7 and 8 and shown in figure 1.

This shows that the population follower counts is a non-normal skew-right distribution, which is expected because the number of friends is an integer value bound by zero [6].

```

1 ## Resample the Data
2 (bt_pop <- sample(x, size = 10^6, replace = TRUE)) %>% head()
3
4 ## > [1] 7 515 262 309 186 166

```

Listing 7: Bootstrapping a population from the sample.

```

1 ## Make the Population
2 bt_pop_data <- tibble("Followers" = bt_pop)
3 ggplot(data = bt_pop_data, aes(x = Followers)) +
4   geom_histogram(aes(y = ..density..), fill = "lightblue", bins = 35,
5     ↪ col = "pink") +
6   geom_density(col = "violetred2") +
7   scale_x_continuous(limits = c(1, 800)) +
8   theme_bw() +
9   labs(x = "Number of Followers", y = "Density",
10     title = "Bootstrapped population of Follower Numbers")

```

c.) Estimate a Confidence Interval for the population mean Follower Counts

In order to perform a bootstrap for the population mean value of follower counts it is necessary to:

1. Resample the data with replacement

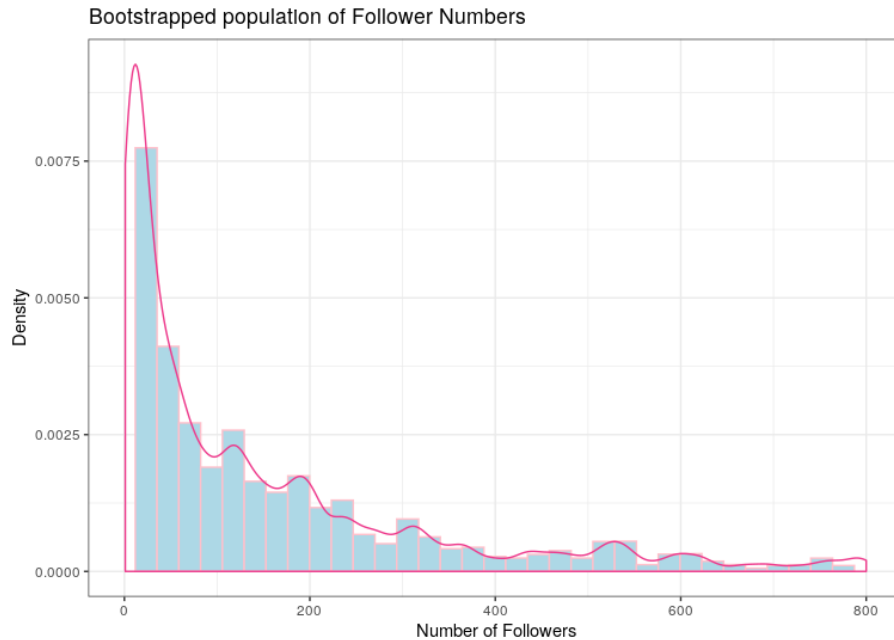


Figure 1: Histogram of the bootstrapped population of follower counts

- i.e. randomly select values from the sample allowing for repetition
2. Measure the statistic of concern
 3. Replicate this a sufficient number of times
 - i.e. Greater than or equal to 1000 times [2, Ch. 5]

This is equivalent to drawing a sample from a population that is infinitely large and constructed of repetitions of the sample. This can be performed in **R** as shown in listing 8.

```

1 xbar_boot_loop <- replicate(10^3, {
2   s <- sample(x, replace = TRUE)
3   mean(s)
4 })
5 quantile(xbar_boot_loop, c((1-0.97)/2, (1+0.97)/2))
6
7 ##          1.5%          98.5%
8 ## 588.4189 10228.7352

```

Listing 8: Confidence Interval of Mean Follower Count in Population

A 97% probability interval is such that a sample drawn from a population will contain the population mean in that interval 97% of the time, this means that it may be concluded with a high degree of certainty that the true population mean lies between 588 and 10228.

1. Alternative Approaches If this data was normally distributed it may have been appropriate to consider bootstrapping the standard error and using a t distribution, however it is more appropriate to use a percentile interval for skewed data such as this, in saying that however this method is not considered to be very accurate in the literature and is often too narrow. [3, Section 4.1]

- It's worth noting that the normal t value bootstrap offers no advantage over using a t distribution (other than being illustrative of bootstrapping generally) [3, Section 4.1]

The `boot` package is a bootstrapping library common among authors in the data science sphere [4, p. 295] [8, p. 237] that implements confidence intervals consistent with work by Davison and Hinkley [7] in their textbook *Bootstrap Methods and their Application*. In this work it is provided that the BC_a method of constructing confidence intervals is superior to mere percentile methods in terms of accuracy [2, Ch. 5], a sentiment echoed in the literature. [1, 2, Ch. 5]

Such methods can be implemented in **R** by passing a function to the `boot` call as shown in listing 9. This provides a broader interval, providing that the true confidence interval could lie between 1079 and 16227 followers.

```
1 xbar_boot <- boot(data = x, statistic = mean_val, R = 10^3)
2 boot.ci(xbar_boot, conf = 0.97, type = "bca", index = 1)
3
4 ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
5 ## Based on 1000 bootstrap replicates
6 ##
7 ## CALL :
8 ## boot.ci(boot.out = xbar_boot, conf = 0.97, type = "bca", index = 1)
9 ##
10 ## Intervals :
11 ## Level      BCa
12 ## 97%      ( 1079, 16227 )
13 ## Calculations and Intervals on Original Scale
14 ## Warning : BCa Intervals used Extreme Quantiles
15 ## Some BCa intervals may be unstable
16 ## Warning message:
17 ## In norm.inter(t, adj.alpha) : extreme order statistics used as
   ↪ endpoints
```

Listing 9: Bootstrap of population mean follower count implementing the BC_a method

references

d.) Estimate a Confidence Interval for the population mean Friend Counts

A Confidence interval for the population mean friend counts may be constructed in a like wise fashion as shown in listings 10. This provides that the 97% confidence interval for the population mean friend count is between 384 and 502 (or 387 and 496 if the BC_a method used, they're quite close and so the more conservative percentile method will be accepted).

```

1  # d.) Estimate a Confidence Interval for the populattion mean Friend
   ↪ Count ===
2  # Using a Percentile Method
   ↪ #####
3  ybar_boot_loop <- replicate(10^3, {
4    s <- sample(y, replace = TRUE)
5    mean(s)
6  })
7  quantile(ybar_boot_loop, c(0.015, 0.985))
8
9  # Using BCA Method
   ↪ #####
10 mean_val <- function(data, index) {
11   X = data[index]
12   return(mean(X))
13 }
14
15 xbar_boot <- boot(data = y, statistic = mean_val, R = 10^3)
16 boot.ci(xbar_boot, conf = 0.97, type = "bca", index = 1)
17
18
19 ##      1.5%    98.5%
20 ## 383.7619 501.5903
21 ##
22 ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
23 ## Based on 1000 bootstrap replicates
24 ##
25 ## CALL :
26 ## boot.ci(boot.out = xbar_boot, conf = 0.97, type = "bca", index = 1)
27 ##
28 ## Intervals :
29 ## Level      BCa
30 ## 97%      (386.8, 496.7 )
31 ## Calculations and Intervals on Original Scale
32 ## Some BCa intervals may be unstable

```

Listing 10: Bootstrap of population mean follower count

FIXME 8.1.6 Estimate a 97% Confidence Interval for the High Friend Count Proportion

In order to bootstrap a confidence interval for the proportion of users with above average follower counts, repeatedly draw random samples from an infinitely large population composed entirely of the sample, and record the sampled proportion. this can be achieved by resampling the observations of above and below as shown in listing 11.

This provides that:

- The 97% confidence interval for the population proportion of users that have an above average number of friends is between 0.24 and 0.31.
 - i.e. The probability of any given sample containing the population mean within this interval would be 97%, although that doesn't however mean that there is a 97% probability that this interval contains the value, merely that we may be 97% *confident*

8.1.7 Is the Number of Friends Independent to the Number of Followers

One method to determine whether or not the number of followers is independent of the number of friends is to bin the counts and determine whether or not the distribution of users across those counts is consistent with the hypothesis of independence.

Bin the Follower and Friend Categories

The counts may be binned by performing a logical interval test as shown in listing 12.

Find the Group frequency

These values may be tabulated in order to count the occurrence of users among these categories as shown in listing 13 and table 1.

Table 1: Table of Binned Friend and Follower counts, transposed relative to code.

	<i>Followers</i>	<i>Friends</i>
<i>Tens</i>	421	262
<i>Hundreds</i>	317	476
<i>1 - Thousands</i>	39	47
<i>2 - Thousands</i>	11	15
<i>3 - Thousands</i>	9	6
<i>4 - Thousands</i>	2	9
<i>5 Thousand or More</i>	18	2

Find the Expected Counts under each group and test for independence

The expected count of each cell, under the assumption that the two metrics are independent, will be the proportion users per bracket multiplied by the number of users in that status group. This implies that any cell will be:

```

1  # 8.1.6 High Friend Count Proportion
   ↪ -----
2  prop <- factor(c("Below", "Above"))
3  ## 1 is above average, 2 is below
4  py_hat_bt <- replicate(10^3, {
5      rs      <- sample(c("Below", "Above"),
6                        size = length(y),
7                        prob = c(py_hat, 1-py_hat),
8                        replace = TRUE)
9      isabove <- rs == "Above"
10     mean(isabove)
11 })
12 quantile(py_hat_bt, c(0.015, 0.985))
13
14
15 ##      1.5%      98.5%
16 ## 0.2399021 0.3072215
17 ## > > > . + > > >
18 ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
19 ## Based on 1000 bootstrap replicates
20 ##
21 ## CALL :
22 ## boot.ci(boot.out = py_hat_boot, conf = 0.97, type = "bca")
23 ##
24 ## Intervals :
25 ## Level      BCa
26 ## 97%      ( 0.2399,  0.3072 )
27 ## Calculations and Intervals on Original Scale

```

Listing 11: Bootstrap of Proportion of Friends above average

```

1  ## Assign Categories
2  x_df <- data.frame(x)
3  x_df$cat[0      <= x_df$x & x_df$x < 100] <- "Tens"
4  x_df$cat[100    <= x_df$x & x_df$x < 1000] <- "Hundreds"
5  x_df$cat[1000   <= x_df$x & x_df$x < 2000] <- "1Thousands"
6  x_df$cat[2000   <= x_df$x & x_df$x < 3000] <- "2Thousands"
7  x_df$cat[3000   <= x_df$x & x_df$x < 4000] <- "3Thousands"
8  x_df$cat[4000   <= x_df$x & x_df$x < 5000] <- "4Thousands"
9  x_df$cat[5000   <= x_df$x & x_df$x < Inf]  <- "5ThousandOrMore"
10
11 ### Make a factor
12 x_df$cat <- factor(x_df$cat, levels = var_levels, ordered = TRUE)
13
14 ### Determine Frequencies
15 (x_freq <- table(x_df$cat) %>% as.matrix())
16
17 ## ** b) Find the Friend Count Frequency
18   ↪ =====
19 ## Assign Categories
20 y_df <- data.frame(y)
21 y_df$cat[0      <= y_df$y & y_df$y < 100] <- "Tens"
22 y_df$cat[100    <= y_df$y & y_df$y < 1000] <- "Hundreds"
23 y_df$cat[1000   <= y_df$y & y_df$y < 2000] <- "1Thousands"
24 y_df$cat[2000   <= y_df$y & y_df$y < 3000] <- "2Thousands"
25 y_df$cat[3000   <= y_df$y & y_df$y < 4000] <- "3Thousands"
26 y_df$cat[4000   <= y_df$y & y_df$y < 5000] <- "4Thousands"
27 y_df$cat[5000   <= y_df$y & y_df$y < Inf]  <- "5ThousandOrMore"
28
29 ### Make a factor
30 y_df$cat <- factor(y_df$cat, levels = var_levels, ordered = TRUE)
31
32 ### Determine Frequencies
33 (y_freq <- table(y_df$cat) %>% as.matrix())

```

Listing 12: Use Logical Test to Assign observations into bins

```

1 vals <- t(cbind(x_freq, y_freq))
2 rownames(vals) <- c("Followers.x", "followers.y")
3 vals
4
5 ##           Tens Hundreds 1Thousands 2Thousands 3Thousands 4Thousands
6 ## Followers.x  421       317        39         11         9         2
7 ## followers.y  262       476        47         15         6         9
8 ##           5ThousandOrMore
9 ## Followers.x           18
10 ## followers.y           2

```

Listing 13: Tabulate the binned counts for the distribution of users among amount and status.

- the product of the row sum, multiplied by the column sum divided by the number of counts.

This can be equivalently expressed as an outer product as shown in equation (1), in **R** this operation is denoted by the `%o%` operator, which is shorthand for the `outer()` function, this and other summary statistics may be evaluated as shown in listing 14.

The outer product is such that:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}.$$

This means the matrix of expected frequencies can be expressed as an outer product thusly:

$$\tilde{\mathbf{e}} = \frac{1}{n} \times \begin{bmatrix} \sum_{j=1}^n [o_{1j}] \\ \sum_{j=1}^n [o_{2j}] \\ \sum_{j=1}^n [o_{3j}] \\ \sum_{j=1}^n [o_{4j}] \\ \vdots \\ \sum_{j=1}^n [o_{nj}] \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n [o_{i1}] \\ \sum_{j=1}^n [o_{i2}] \\ \sum_{j=1}^n [o_{i3}] \\ \dots \\ \sum_{j=1}^n [o_{in}] \end{bmatrix}^T \quad (1)$$

```

1 ## ***** Calculate Summary Stats
2 n <- sum(vals)
3 bracket_prop <- colSums(vals) / n
4 metric_prop <- rowSums(vals) / n
5 o <- vals
6 e <- rowSums(vals) %o% colSums(vals) / n
7 chi_obs <- sum((e-o)^2/e)

```

Listing 14: Calculate Expected frequency of values under the assumption of independence.

1. Testing Independence In order to test whether or not the distribution of users among brackets is independent of being a follower or friend a χ^2 test may be used, this can be evaluated from a model or simulated, in **R**, the simulated test is shown in listing 15, this provides a p -value < 0.0005 , which means that the hypothesis of independence may be rejected with a high degree of certainty.

```
1  chisq.test(vals, simulate.p.value = TRUE)
2
3
4  ## ~IPearson's Chi-squared test with simulated p-value (based on 2000
5  ## ~Ireplicates)
6  ##
7  ## data:  vals
8  ## X-squared = 88.109, df = NA, p-value = 0.0004998
```

Listing 15: Chi-Square testing for independence between friend and follower bin categories.

- (a) From First Principles The χ^2 statistic may be performed from first principles by randomly sampling the values at the rate at which they occurred, tabulating those counts, measuring the χ^2 -value and then repeating this many times.

Because the samples are random they must be independent and average number of positives is hence an estimate for the *FPR*, which is in turn an estimate for the p -value. This technique is demonstrated in listing 16, the p -value being returned as 0.0004, this value is consistent with the value produced by **R**'s built in `chisq.test` function and so is accepted.

FIXME Conclusion

The p -value measures the probability of rejecting the null hypothesis when it is true, i.e. the probability of detecting a *false positive*, a very small p -value is hence good evidence that the null hypothesis should be rejected (because doing so would unlikely to be a mistake).

In saying that however the p -value is distinct from the *power* statistic, which is a measure of */the probability of accepting the alternative hypothesis* when it is true, a low p -value is not a measurement of the probability of being correct.

Hence we can conclude, with a high degree of certainty, that the follower and friend counts are not independent of one another.

8.2 Finding Themes in tweets

8.2.8 Find Users with Above Average Friend Counts

Users with Above average Friend Counts can be identified by filtering the tweets data frame for two conditions:

1. non-duplicated user-id

```

1  ## ***** Create Vectors of factor levels
2  brackets <- unique(x_df$cat)
3  metrics <- c("follower", "friend")
4
5  ## ***** Simulate the data Assuming H_0
6  ## I.e. assuming that the null hypothesis is true in that
7  ## the brackets assigned to followers are independent of the friends
8  ## (this is a symmetric relation)
9
10 s <- replicate(10^4,{
11   ## Sample the set of Metrics
12   m <- sample(metrics, size = n, replace = TRUE, prob = metric_prop)
13
14   ## Sample the set of Brackets (i.e. which performance bracket the
15   ↪ user falls in)
16   b <- sample(brackets, size = n, replace = TRUE, prob = bracket_prop)
17
18   ## Make a table of results
19   o <- table(m, b)
20
21   ## Find What the expected value would be
22   e_sim <- t(colSums(o) %o% rowSums(o) / n)
23
24   ## Calculate the Chi Stat
25   chi_sim <- sum((e_sim-o)^2/e_sim)
26   chi_sim
27
28   ## Is this more extreme, i.e. would we reject null hypothesis?
29   chi_sim > chi_obs
30
31 })
32
33 mean(s)
34
35 ## [1] 4e-04

```

Listing 16: Performing a χ^2 statistic from first principles

2. friend_count greater than average

This can be achieved easily using the dplyr package as shown in 17, the top 20 of these users are shown in table 2 of the appendix

```
1 select <- dplyr::select
2 filter <- dplyr::filter
3 interested_vars <- c("user_id", "friends_count")
4 (friend_counts <- tweets.company %>%
5   select(interested_vars) %>%
6   filter(!duplicated(user_id)))
7
8 (high_friends <- friend_counts %>%
9   filter(friends_count > mean(friends_count, na.rm = TRUE)))
10
11 ## Export Friends List
12 write.csv(high_friends[order(
13   high_friends$friends_count,
14   decreasing = TRUE),], file = "/tmp/highfriend.csv")
```

Listing 17: Use dplyr to Filter for Users with a high Friend Count

8.2.9 Find Users with Below Average Friend Counts

Users with high friends may be determined by a similar method (or by taking the complement of the high friends) as shown in listing 18, the lowest 20 of these users are shown in table 3 of the appendix.

```
1 (low_friends <- friend_counts %>%
2   filter(friends_count <= mean(friends_count, na.rm = TRUE)))
3
4 low_friends <- low_friends[order(
5   low_friends$friends_count,
6   decreasing = TRUE),]
7
8 ## Export Users
9 write.csv(low_friends[order(
10   low_friends$friends_count,
11   decreasing = FALSE),], file = "/tmp/lowfriend.csv")
```

Listing 18: Use dplyr to Filter for Users with a low Friend Count

8.2.10 Find the *Tweets* corresponding to users with high or low friend counts

The tweets corresponding to users with high and low friend counts can be identified by filtering the dataframe based on the friend count and using that to the index the tweets from the data frame ¹, alternatively it is possible to test whether or not the ID of a user appears in the high or low vector set using the `%in%` operator as shown in listing 19.

```
1  ## Method 1
2  friend_test <- tweets.company$friends_count >
   ↪ mean(tweets.company$friends_count)
3  tweets_high <- tweets.company$text[friend_test]
4  tweets_low <- tweets.company$text[!friend_test]
5
6  ## Method 2
   ↪      :15b5a74:
7  ## top part is the high friends and the low part is the low friends
8
9  tweets_high <- tweets.company$text[tweets.company$user_id %in%
   ↪ high_friends$user_id]
10 tweets_low  <- tweets.company$text[tweets.company$user_id %in%
   ↪ low_friends$user_id]
11 tweets <- c(tweets_high, tweets_low)
```

Listing 19: Identify tweets corresponding to users with high and low friend counts

8.2.11 Clean the tweets

Create a Corpus Object

In order to clean the tweets it is necessary to create a corpus object as shown in listing 20, it is possible to pass a dataframe source in order to include the user ID, this isn't strictly necessary however because the `tm` package preserves order when performing transformations.

Next it is necessary to choose an encoding, a primary consideration of this is whether or not the use of *emoji* characters will influence the model performance. There is research to suggest that Emoji's can be used as predictive features [lecompte2017] and that they can improve *sentiment analysis* models [shiha2017] that implement a *bag of words* approach. For these reasons *emoji* characters will be preserved and `UTF-8` implemented.

In order to encode the data as `UTF-8`, the `iconv` function can be used as shown in listing 21.

¹This works because the `tm` package preserves the order of the data, this can be confirmed by using a dataframe source as opposed to a vector source (e.g. in listing 20) and comparing the ID's before/after transformation.


```

1 tweet_source <- tm::VectorSource(doc_id = tweets$user_id, text =
  ↳ tweets$text)
2 tweet_corpus <- tm::Corpus(x = tweet_source)

```

Listing 20: Create a Corpus from the tweets

```

1 encode <- function(x) {
2   iconv(x, to = "UTF-8")
3   #   iconv(x, to = "latin1")
4   #   iconv(x, to = "ASCII")
5 }
6
7 tweet_corpus <- tm_map(x = tweet_corpus, FUN = encode)
8 tweet_corpus_raw <- tweet_corpus

```

Listing 21: Encode the Data as UTF-8

Process the tweets

Before analysis the tweets should be modified to remove characters that may interfere with categorising words, this is referred to as cleaning, in particular the following should be implemented:

1. Remove URL's
2. Remove Usernames
3. remove numbers
4. remove punctuation
5. remove whitespace
6. case fold all characters to lower case
7. remove a set of stop words
8. reduce each word to its stem

In particular it is important to reduce words to lower case before removing stop words otherwise an unorthodox use of capitalisation may prevent the word from being removed throughout.

The stop word `ubisoft` will also be used, this was the query term so it's expected to turn up at a very high frequency, the words `can` and `'s` also occurred quite frequently and so were removed.

The cleaning can be implemented by mapping functions over the corpus, which is fundamentally a list, this can be performed via the `tm_map` function as shown in listing 22.

```

1  mystop <- c(stopwords(), "s", "can", "ubisoft", "@ubisoft",
  ↪ "#ubisoft")# <<stphere>>
2
3  clean_corp <- function(corpus) {
4    ## Remove URL's
5    corpus <- tm_map(corpus, content_transformer(function(x)
  ↪ gsub("(f|ht)tp(s?)://\\S+", "", x)))
6    ## Remove Usernames
7    corpus <- tm_map(corpus, content_transformer(function(x)
  ↪ gsub("@\\w+", "", x)))
8    ## Misc
9    corpus <- tm_map(corpus, FUN = removeNumbers)
10   corpus <- tm_map(corpus, FUN = removePunctuation)
11   corpus <- tm_map(corpus, FUN = stripWhitespace)
12   corpus <- tm_map(corpus, FUN = tolower)
13   corpus <- tm_map(corpus, FUN = removeWords, mystop)
14   ## stopwords() returns characters and is fead as second argument
15   corpus <- tm_map(corpus, FUN = stemDocument)
16   return(corpus)
17 }
18
19 tweet_corpus_clean <- clean_corp(tweet_corpus)

```

Listing 22: Use the tm_map function to clean the tweets

8.2.12 Display the first two tweets before/after processing

The tweets can be viewed from inside the corpus by selecting with the `[` function ² as demonstrated in listing 23, the first *tweet* was rendered empty by the processing and the following two tweets were:

- **Pre-Processing**

- “Today was the first time in over a month that I have gone 24 hours without checking the coronavirus death toll. Thanks @Ubisoft.”
- “@btwimskrank @TheDivisionGame @UbiMassive @Ubisoft @jgerighty @hamishbode @Tideman92 @janeyo_jane @slimjd Very odd... I'll even post a video about it.”

- **Post-Processing**

- “today first time month gone hour without check coronavirus death toll thank”
- “odd ill even post video”

```
1 tweet_corpus_raw[[1]]$content
2 tweet_corpus_clean[[1]]$content
3 tweet_corpus_raw[[2]]$content
4 tweet_corpus_clean[[2]]$content
5 tweet_corpus_raw[[3]]$content
6 tweet_corpus_clean[[3]]$content
```

Listing 23: Load the Packages for *R*

8.2.13 Create a Term Document Matrix

Apply Weighting Manually

1. Create Term Document Matrix A term Document matrix (and it's transpose) can be constructed from a corpus using the `tm:TermDocumentMatrix` function as shown in listing 24.
2. Apply TF-IDF Weighting Weighted term frequency is defined as shown in equation (2), where:
 - $f_{d,t}$ is the frequency of a given term t in the a document d
 - $w_{d,t}$ is the weight of a given term t in the a document d
 - N is the number of documents

²The `[` function is actually shorthand for `Extract()`, most things in *R* are functions, this is similar to *LISP* and has to do with the origins of the language, e.g. `sum(1:10) == (sum (1:10))`, also relevant see the relevant xkcd in figure 3.

```

1 tweet_matrix_tdm <- TermDocumentMatrix(tweet_corpus_clean)
2 tweet_matrix_dtm <- DocumentTermMatrix(tweet_corpus_clean)

```

Listing 24: Load the Packages for *R*

- f_t is the number of documents containing t

$$\begin{aligned}
 w_{d,t} &= \text{TF}_t \times \text{IDF}_{d,t} \\
 &= \log_e(f_{d,t} + 1) \log_e\left(\frac{N}{f_t}\right)
 \end{aligned}
 \tag{2}$$

This would require multiplying each term of each row of the TF_t matrix by the corresponding vector element of $\text{IDF}_{d,t}$, this can be implemented by taking the matrix product of a diagonalised matrix, this is shown in listing 25.

```

1 N <- nrow(as.matrix(tweet_matrix_dtm)) # Number of Documents
2 ft=colSums(as.matrix(tweet_matrix_dtm) > 0) #in how many documents term
   ↪ t appeared in,
3
4 TF <- log(as.matrix(tweet_matrix_dtm) + 1) # built in uses log2()
5 IDF <- log(N/ft)
6
7 tweet_weighted <- TF %*% diag(IDF)
8 colnames(tweet_weighted) <- colnames(tweet_matrix_dtm)

```

Listing 25: Apply TF-IDF Weigting

There is however a function built in to the *tm* package that will weight term document matrices and this will instead be implented to analyse the data because it will produce more maintainable code.

Apply Weighting with Built in Function

In order to create a term document matrix (and its transpose) with TF-IDF weighted values, the weighting argument may be specified as `weightTfIdf` by passing an appropriate list to the `control` argument of the `TermDocumentMatrix`, as shown in listing 26

Remove Empty Documents

Empty Documents may be removed from the matrix by a logical test as shown in listing 27³ this provides that 328 documents were empty following the processing. A summary of the first rows and columns of

³It is important not to filter based the logic of an empty vector, because otherwise an empty vector will returned, hence the if statement in listing 27.

```

1 tweet_weighted_tdm <- tm::TermDocumentMatrix(x = tweet_corpus_clean,
  ↪ control = list(weighting = weightTfIdf))
2 tweet_weighted_dtm <- as.DocumentTermMatrix() %>%
3     as.matrix()

```

Listing 26: Create a Document Term Matrix by transforming a Term Document Matrix

this matrix, following the removal of empty documents, is provided in table 4 of the appendix.

```

1 null = which(rowSums(as.matrix(tweet_weighted_dtm)) == 0)
2 rowSums(as.matrix(tweet_weighted_dtm)==0)
3
4 if(length(null)!=0){
5     tweet_weighted_dtm = tweet_weighted_dtm[-null,]
6 }
7
8 length(null)
9
10 ## [1] 328

```

Listing 27: Load the Packages for *R*

8.2.14 How many Clusters are there

Use Cosine Distance

In order to consider clustering, it can be more effective to consider the distance between the weighted documents in terms of cosine distance, the cosine distance can be calculated from the euclidean distance using the identity shown in (11), and this can be performed in *R* by taking the matrix product of a diagonalised matrix as shown in listing 28.

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\| \quad (3)$$

$$= \sqrt{\sum_{i=1}^n [(x_i - y_i)^2]} \quad (4)$$

$$\text{dist}(\mathbf{X}, \mathbf{Y})^2 = \sum_{i=1}^n [(x_i - y_i)^2] \quad (5)$$

$$= \sum_{i=1}^n (x_i^2) + \sum_{i=1}^n (y_i^2) + 2 \sum_{i=1}^n (x_i y_i) \quad (6)$$

$$= 1 + 1 + 2 \times \frac{\sum_{i=1}^n (x_i y_i)}{(1)} \quad (7)$$

$$= 2 + 2 \times \frac{\sum_{i=1}^n (x_i y_i)}{\|\mathbf{X}\| \times \|\mathbf{Y}\|} \quad (8)$$

$$= 2 + 2 \cos(\mathbf{X}, \mathbf{Y}) \quad (9)$$

$$(10)$$

$$\Rightarrow (1 - \cos(\mathbf{X}, \mathbf{Y})) = \frac{\text{dist}(\mathbf{X}, \mathbf{Y})^2}{2} \quad (11)$$

```

1 norm.tweet_weighted_dtm = diag(
2     1/sqrt(rowSums(tweet_weighted_dtm^2))
3     ) %*% tweet_weighted_dtm
4 D =dist(norm.tweet_weighted_dtm, method = "euclidean")^2/2

```

Listing 28: Load the Packages for *R*

Project into Euclidean Space

The cosine distance however is not appropriate to perform clustering on and so instead should be projected back into euclidean space, this can be achieved using *Multi-Dimensional Scaling* via the `cmdscale` function as shown in listing 29. The distance is a measure of between document distance so the number of dimensions should correspond to the number of documents, however, if there are zero-value eigenvalues, these dimensions won't help explain the data in the projection, hence the number of eigenvalues has been used as the dimension of projection in this case.

Measure Within Cluster Variance

In order to determine the appropriate number of clusters, the within cluster variance can be measured, the number of clusters at which this value ceases to decrease is indicative of a potentially appropriate number of clusters. This is implemented in listing 30 and shown in figure 2.

Figure 2 Indicates a sudden stop of decrease in variance at 7 clusters and following that the within cluster variance begins to decrease at a slightly slower rate. For this reason 7 could be an appropriate candidate for the number of clusters, however the minimal amount of change in the within-cluster variance indicates that the data is most likely not clustered at all.

```

1 l <- min(nrow(tweet_weighted_dtm),
2         ncol(tweet_weighted_dtm))
3 ev <- eigen(tweet_weighted_dtm[1:l, 1:l])
4 k <- (ev$values != 0) %>% sum()
5
6 mds.tweet_weighted_dtm <- cmdscale(D, k=k) #TODO What should K be? see
  ↪ issue #10

```

Listing 29: Load the Packages for *R*

```

1 set.seed(271)
2 n = 15 # Assume it bends at 7 clusters
3 SSW = rep(0, n)
4 for (a in 1:n) {
5   K = kmeans(mds.tweet_weighted_dtm, a, nstart = 20)
6   SSW[a] = K$tot.withinss
7   paste(a*100/n, "%") %>% print()
8 }
9 SSW
10
11
12 SSW_tb <- tibble::enframe(SSW)
13 ggplot(SSW_tb, aes(x = name, y = value)) +
14   geom_point(col = "#Cd5b45", size = 5) +
15   geom_line(col = "#Da70d6") +
16   geom_vline(xintercept = 7, lty = 3, col = "blue") +
17   theme_bw() +
18   labs(x = "Number of Clusters",
19        y = "Within Cluster Sum of Square Distance",
20        title = "Within Cluster Variance across Clusters")

```

Listing 30: Use a loop to evaluate the performace of various cluster models, plot this with *ggplot2*

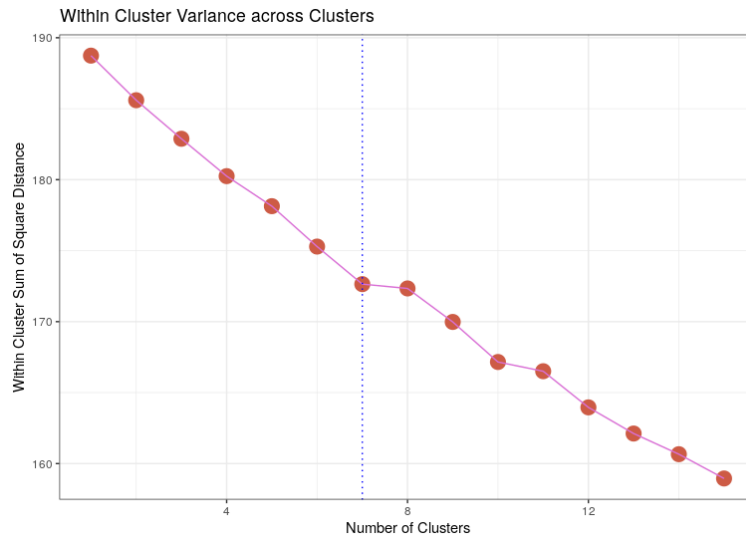


Figure 2: Plot of the Within Cluster Variance of the tweets using cosine distance (projected into Euclidean Space)

Appendix

Users with High Friend Count

Users with Low Friend Count

TF-IDF Matrix

Relevant XKCD

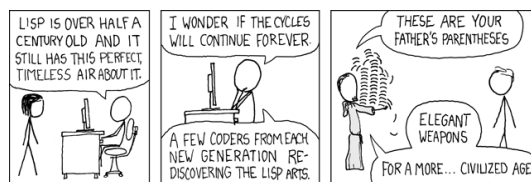


Figure 3: *xkcd* # 297

Table 2: User ID and Friend Count of users with above highest friend count in sample

<i>User ID</i>	<i>Friend Count</i>
274488119	8752
743771665	5002
1036014247	4999
2281452613	4992
1554453560	4958
981233818408570880	4944
931765564388921344	4836
807405140	4710
1112579152970842112	4514
2441577446	4322
552692862	4229
956297007127252992	3976
22493896	3675
255922782	3500
1067409881332936709	3312
27998570	3210
715118521555017728	3099
2356170174	2885
2372688230	2880
1868357425	2719

Table 3: User ID and Friend Count of users with above highest friend count in sample

<i>User ID</i>	<i>Friend Count</i>
1254280995592966145	0
875126772978913280	0
1254256124217319425	0
1250219450210480128	0
1214921087328411648	0
1254115699628421120	0
1217600080376520704	0
1253480062453600257	0
1254178435502571521	0
1251955545092718592	0
1106864828700712960	0
1160744587620524032	0
1254256536710504448	1
1129040408384868352	1
1254121201871589376	1
1248687797755658243	2
1210265263867932675	2
3380784928	3
1177274165239275520	3
54645521	3

Table 4: Document Term Matrix of first 6 documents and first 6 Words

check	coronavirus	death	first	gone	hour
0.615	0.747	0.830	0.596	0.747	0.698
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

References

- [1] James Carpenter and John Bithell. "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians". en. In: *Statistics in Medicine* 19.9 (2000), pp. 1141–1164. ISSN: 1097-0258. DOI: [10.1002/\(SICI\)1097-0258\(20000515\)19:9<1141::AID-SIM479>3.0.CO;2-F](https://doi-org.ezproxy.uws.edu.au/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F). URL: [https://doi-org.ezproxy.uws.edu.au/10.1002/\(SICI\)1097-0258\(20000515\)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F](https://doi-org.ezproxy.uws.edu.au/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F) (visited on 04/27/2020) (cit. on p. 6).
- [2] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge ; New York, NY, USA: Cambridge University Press, 1997. ISBN: 978-0-521-57391-7 978-0-521-57471-6 (cit. on p. 6).
- [3] Tim C. Hesterberg. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum". In: *The American Statistician* 69.4 (Oct. 2015), pp. 371–386. ISSN: 0003-1305. DOI: [10.1080/00031305.2015.1089789](https://doi.org/10.1080/00031305.2015.1089789). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/> (visited on 04/26/2020) (cit. on p. 6).
- [4] Gareth James et al., eds. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. OCLC: ocn828488009. New York: Springer, 2013. ISBN: 978-1-4614-7137-0 (cit. on p. 6).
- [5] Michael Kearney. *Get Tweets Data on Statuses Identified via Search Query*. *Search_tweets*. en. Manual. 2019. URL: https://rtweet.info/reference/search_tweets.html (visited on 04/26/2020) (cit. on p. 1).
- [6] NIST. *1.3.3.14.6. Histogram Interpretation: Skewed (Non-Normal) Right*. Oct. 2013. URL: <https://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm> (visited on 04/26/2020) (cit. on p. 2).
- [7] Brian Ripley. *Boot.Ci Function | R Documentation*. Apr. 2020. URL: <https://www.rdocumentation.org/packages/boot/versions/1.3-25/topics/boot.ci> (visited on 04/27/2020) (cit. on p. 6).
- [8] Matt Wiley and Joshua Wiley. *Advanced R Statistical Programming and Data Models*. New York, NY: Springer Berlin Heidelberg, 2019. ISBN: 978-1-4842-2871-5 (cit. on p. 6).