Twitter Analysis

Tony

Sun Oct 01 16:13:43 2017

Introduction

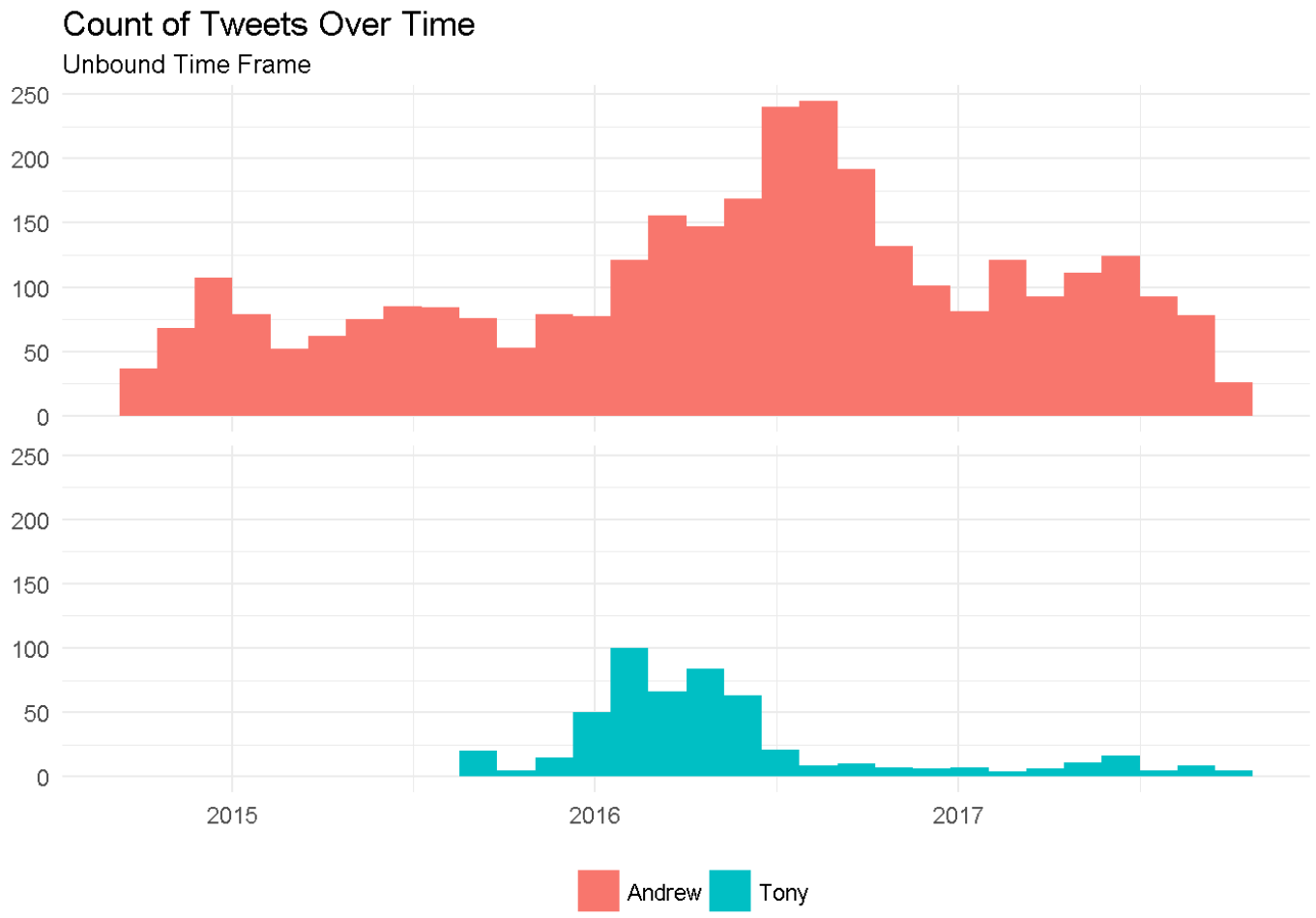
This report compares the volume, behavior, and content of the tweets made by Andrew and Tony.

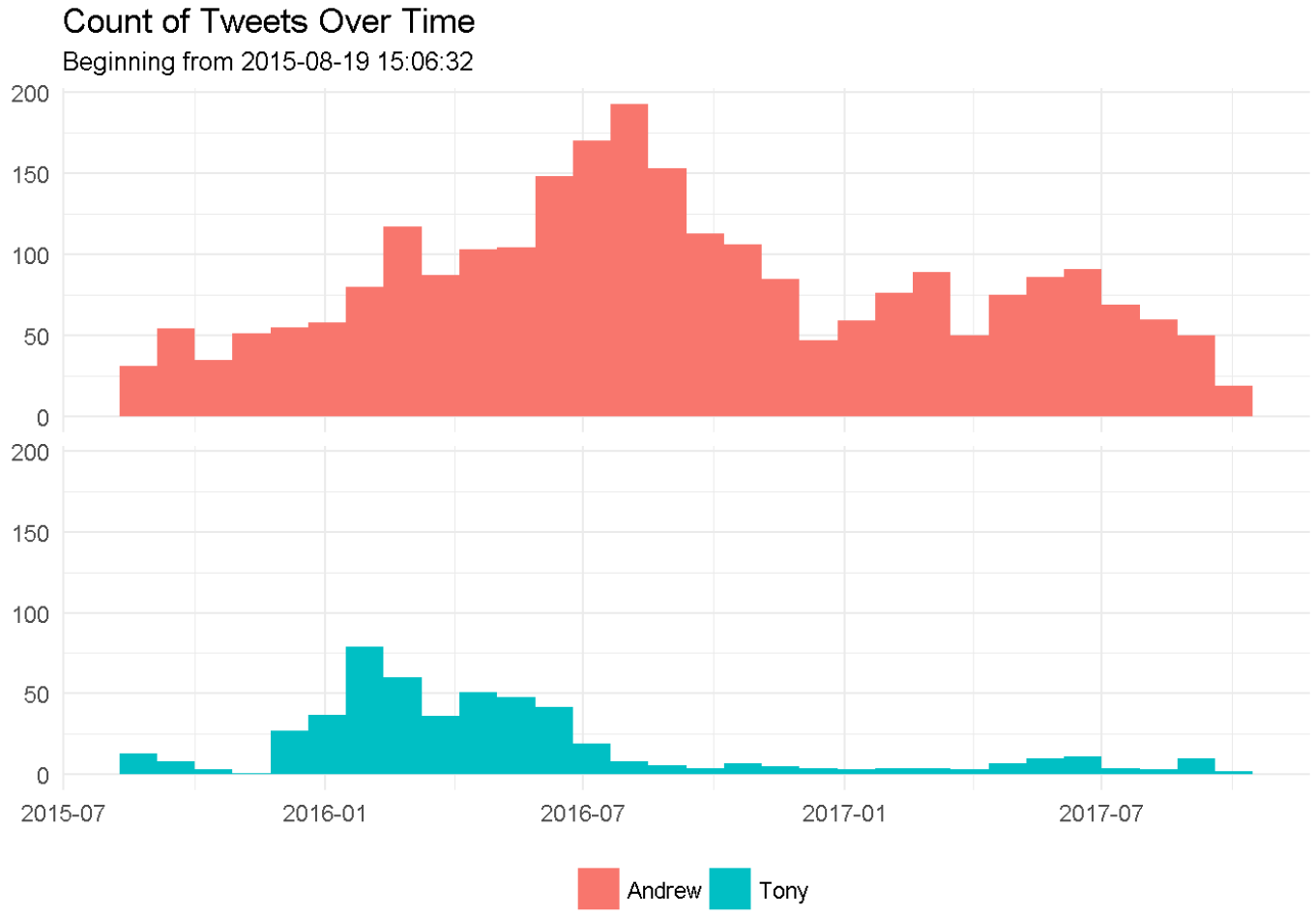
```
## # A tibble: 2 x 2
##   person     n
##    <chr> <int>
## 1 Andrew  3164
## 2   Tony   519
## # A tibble: 2 x 2
##   person           timestamp
##    <chr>              <dttm>
## 1 Andrew 2014-09-23 17:23:37
## 2   Tony 2015-08-19 15:06:32
```

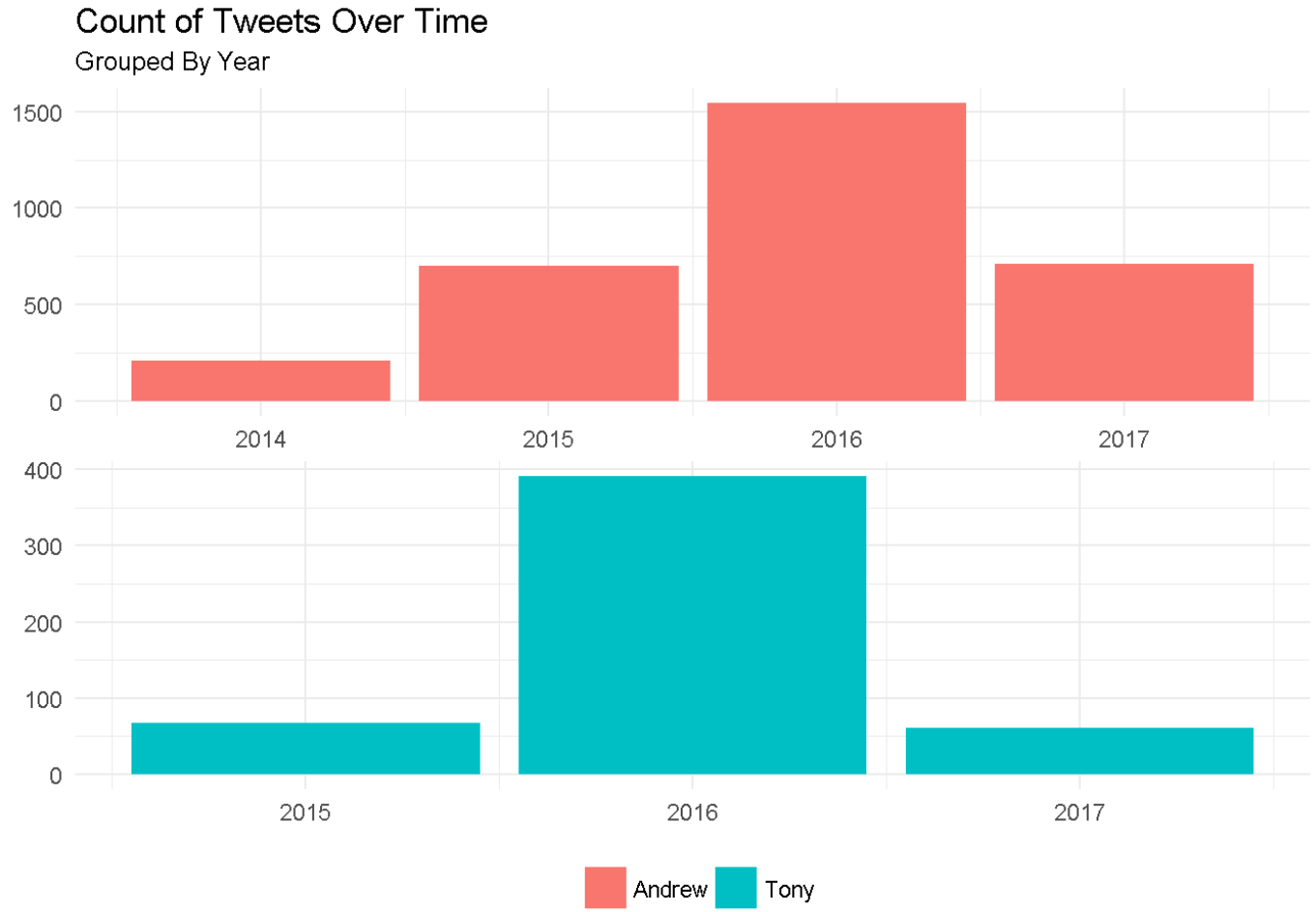
3164 have been collected for Andrew. 519 have been collected for Tony. Note that the oldest tweet made by Andrew (in the collected data) is from 2014-09-23 17:23:37 and the oldest tweet from Tony is from 2015-08-19 15:06:32.

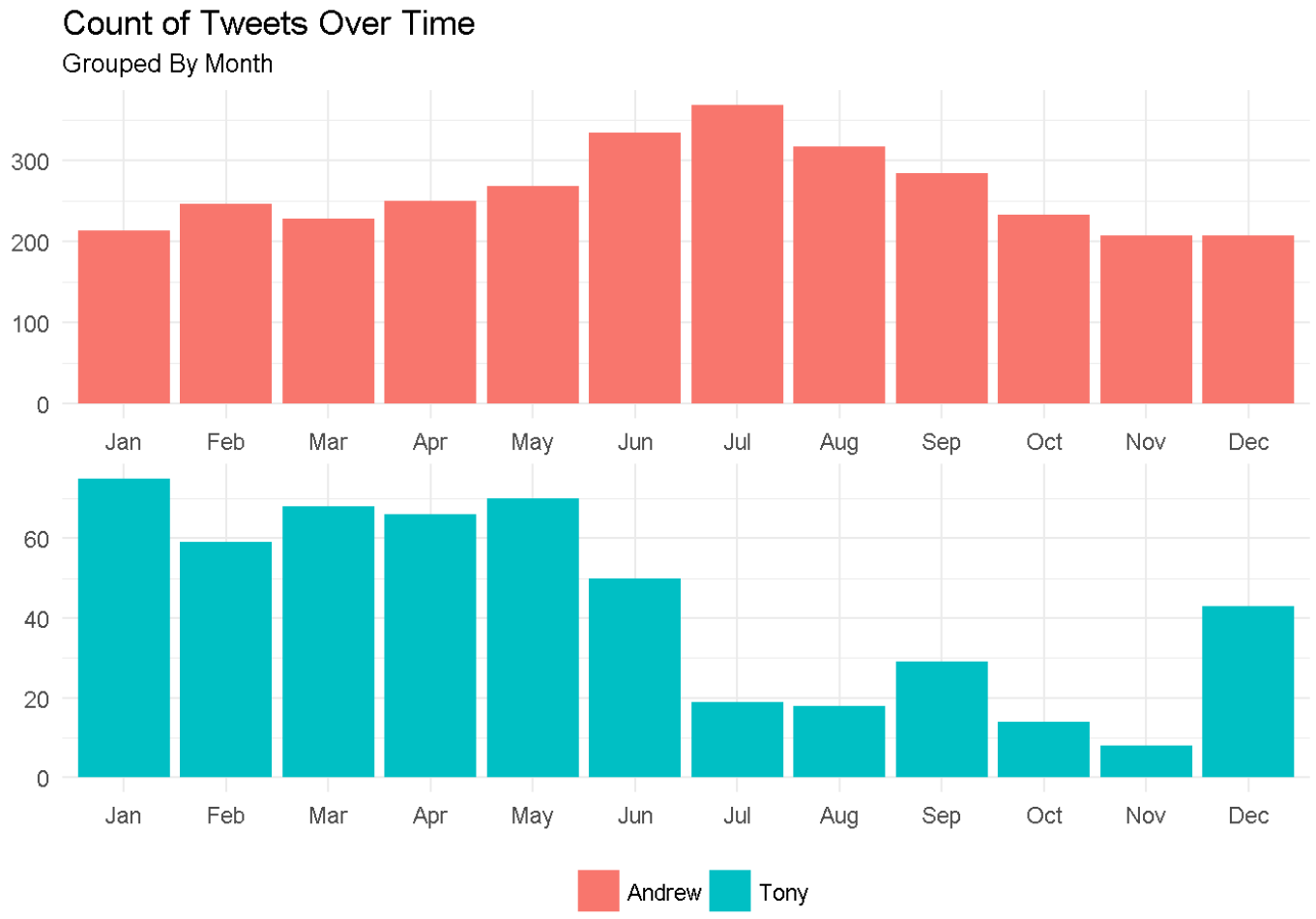
Tweet Volume

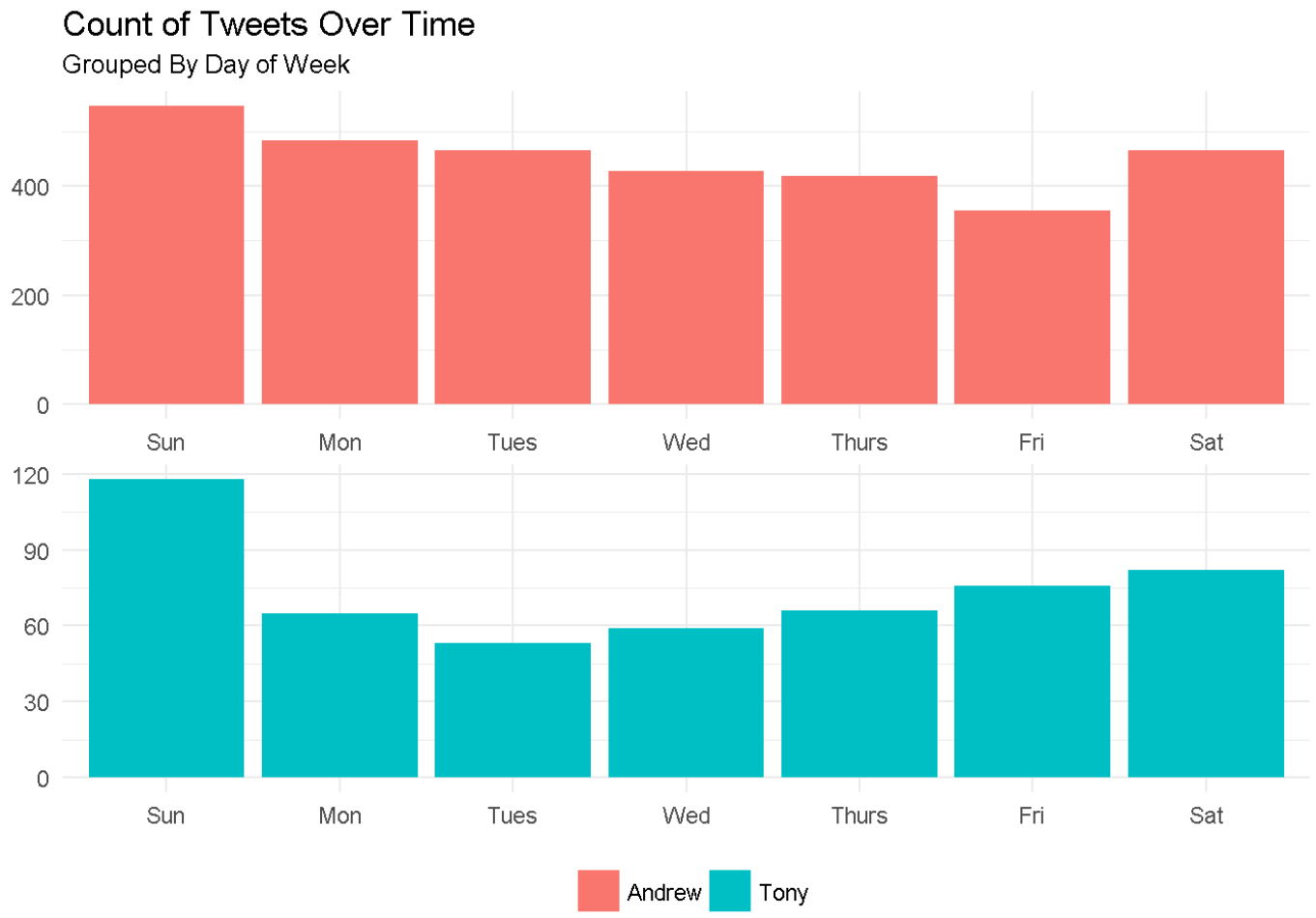
How often do Andrew and Tony tweet? Does the volume of tweets look different for temporal periods (e.g. year, month, etc.)?

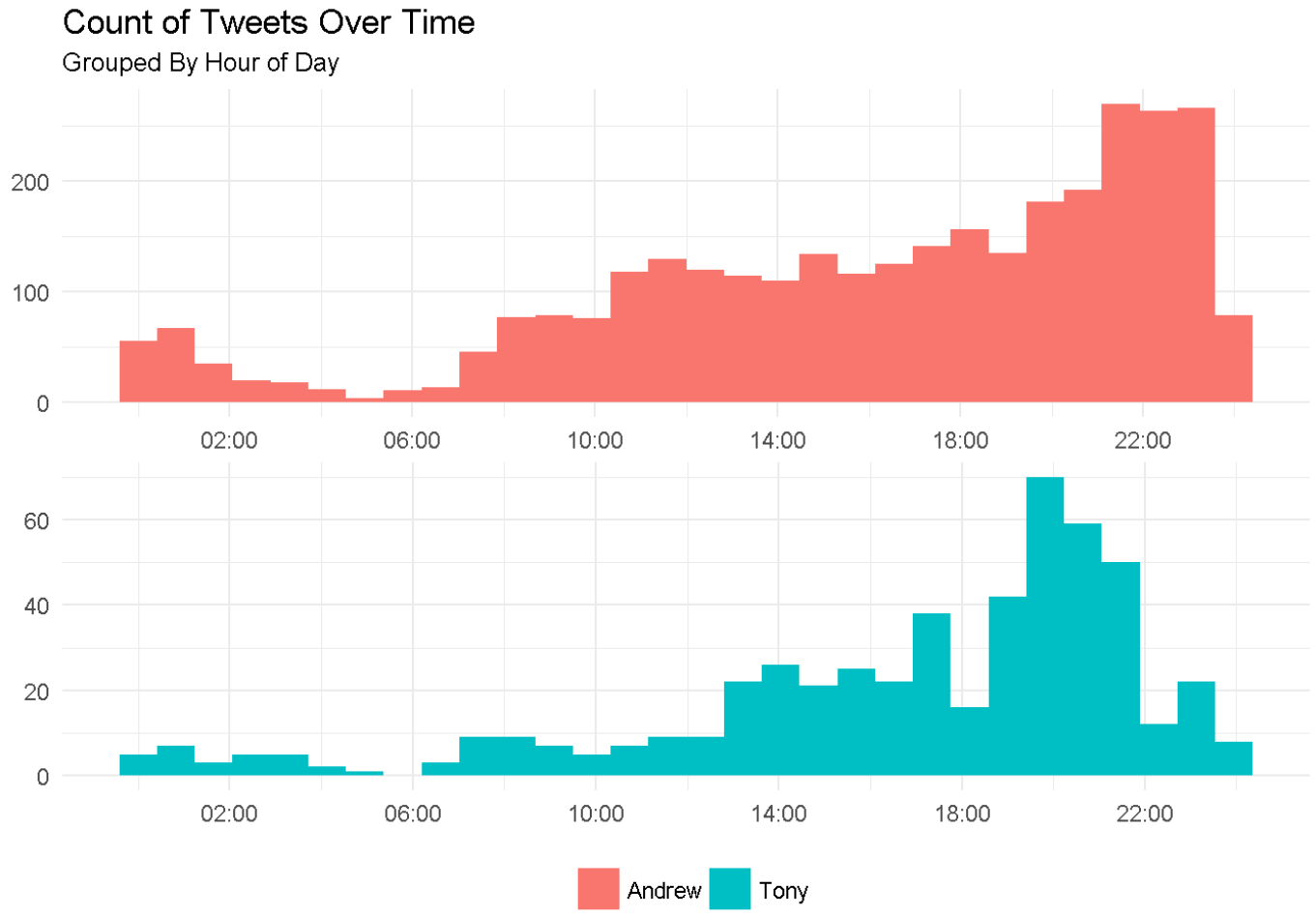

Count of Tweets Over Time
Unbound Time Frame
250
200
150
100
50
0
250
200
150
100
50
0
2015
2016
2017
Andrew
Tony


Count of Tweets Over Time
Beginning from 2015-08-19 15:06:32
200
150
100
50
0
200
150
100
50
0
2015-07
2016-01
2016-07
2017-01
2017-07
Andrew
Tony


Count of Tweets Over Time
Grouped By Year
1500
1000
500
0
2014
2015
2016
2017
400
300
200
100
0
2015
2016
2017
Andrew
Tony


Count of Tweets Over Time
Grouped By Month
300
200
100
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
60
40
20
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Andrew
Tony





Is the distribution of our volume of tweets given a certain temporal period statistically significant? Here, I use the Chi-Squared Test. If the p-value is calculated to be less thatn some threshold value (e.g. 0.05), then I

can deduce that the the null hypothes (that the distribution is uniform) is invalid. In fact, it appears that our tweet volume does differ depending on the month and day of the week.

```
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 116.75, df = 11, p-value < 2.2e-16
##
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 154.32, df = 11, p-value < 2.2e-16
##
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 48.031, df = 6, p-value = 1.165e-08
##
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 37.965, df = 6, p-value = 1.141e-06
##
## [1] 0.9859181
## [1] 0.6603261
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 47.579, df = 6, p-value = 1.434e-08
##
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 13.008, df = 6, p-value = 0.04292
```

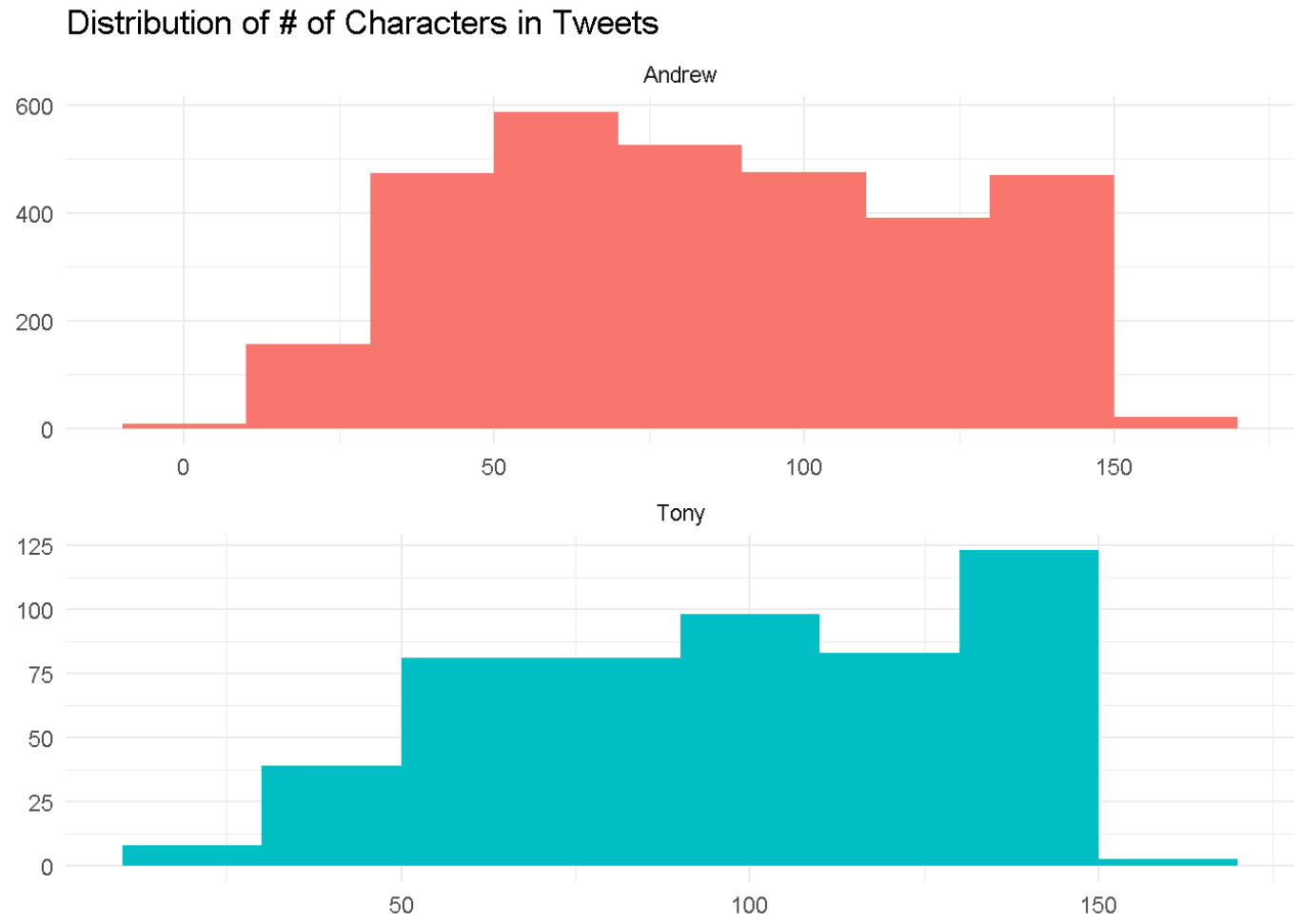
Tweet Behavior

How often do we use hashtags, RT, and reply?

```
## # A tibble: 24 x 4
##    person     type response  value
##  *  <chr>    <chr>    <chr>  <dbl>
##  1 Andrew  hashtag      yes 0.0553
##  2   Tony  hashtag      yes 0.0539
##  3 Andrew hashtag2      yes 0.0531
##  4   Tony hashtag2      yes 0.0539
##  5 Andrew  hashtag       no 0.9447
##  6   Tony  hashtag       no 0.9461
##  7 Andrew hashtag2       no 0.9469
##  8   Tony hashtag2       no 0.9461
##  9 Andrew     link      yes 0.3666
## 10   Tony     link      yes 0.4586
## # ... with 14 more rows
```

Tweet Content

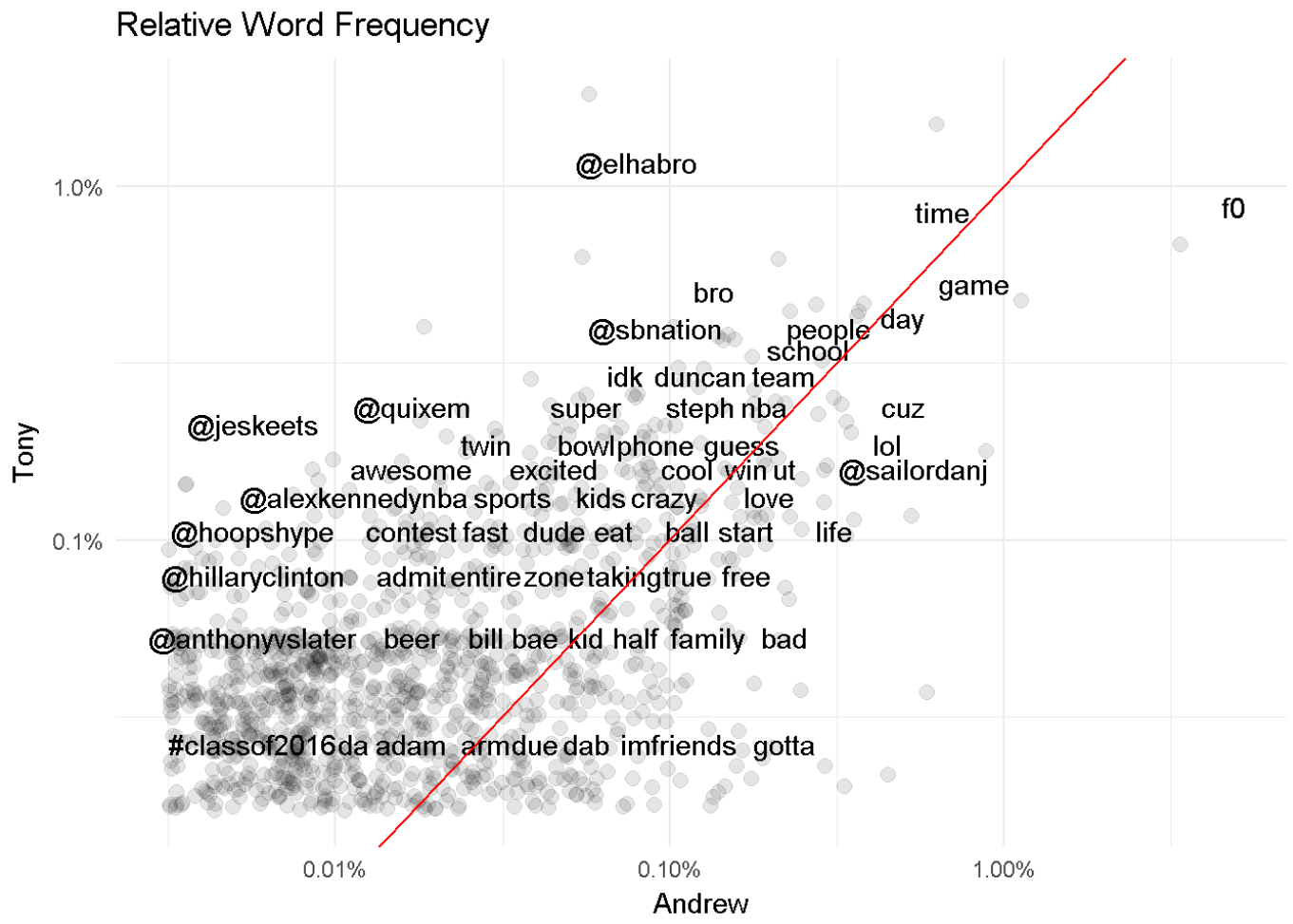
How Long are our tweets/




```
## # A tibble: 1 x 3
##   char_count_count char_count_avg char_count_max
##              <int>          <dbl>          <dbl>
## 1               60         344.85           3542
```

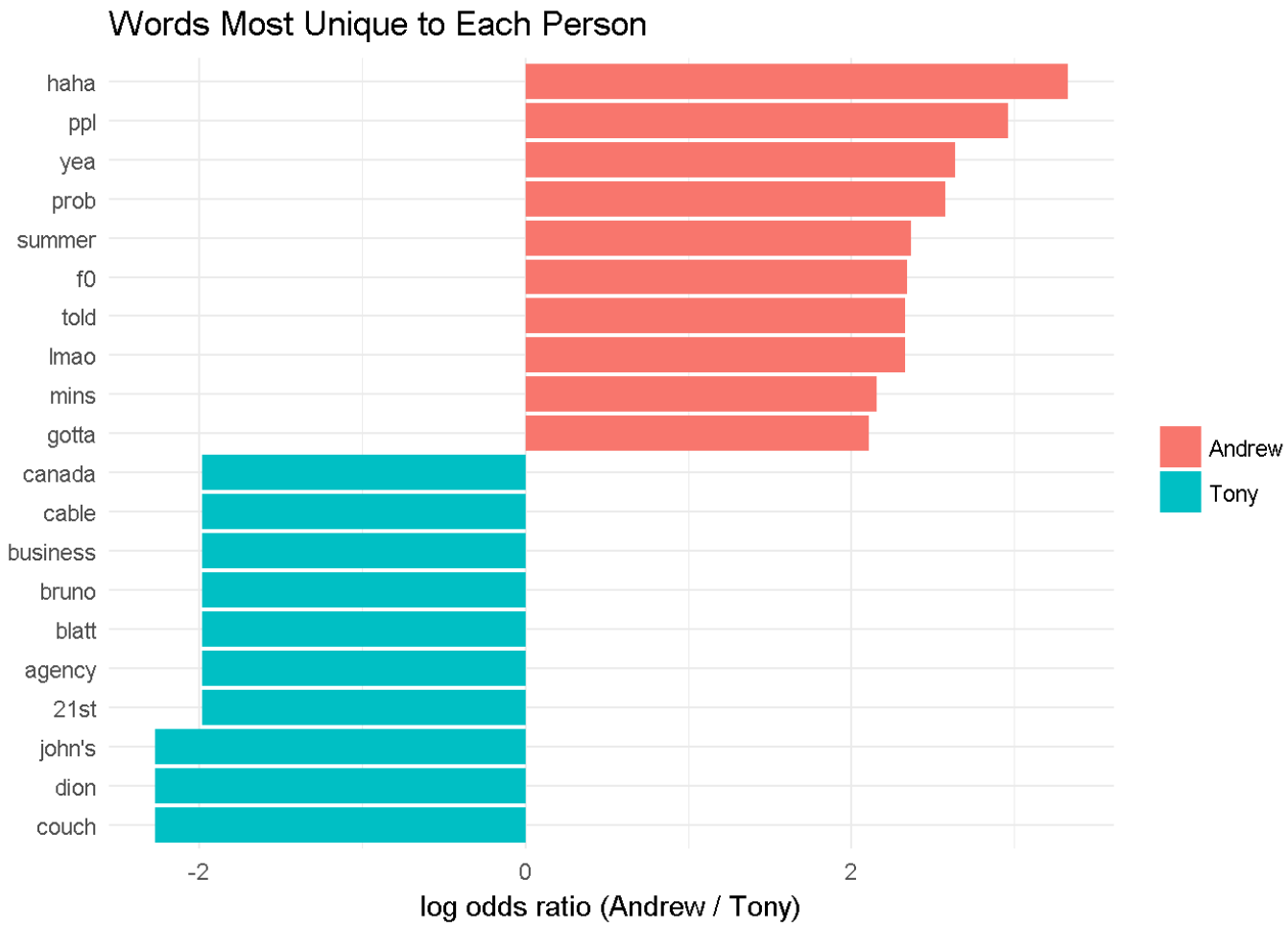
Word Frequency and Usage

Which words are we most likely to use?



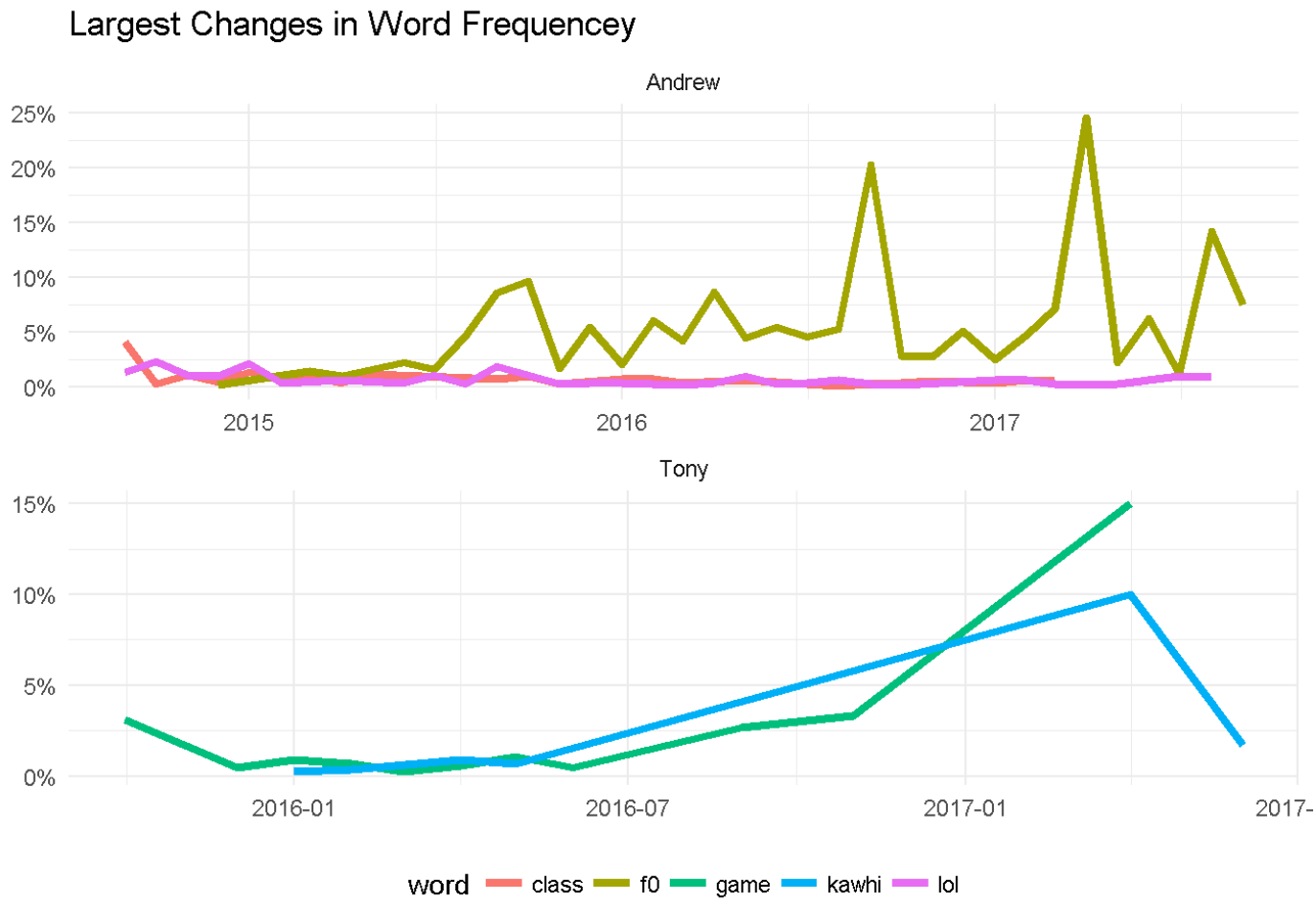
```
## # A tibble: 1,538 x 3
##    screen_name  word          created_at
##          <chr> <chr>               <chr>
##  1     elhabro  00a8 2016-04-01 14:56:28
##  2  TonyElHabr  00a8 2015-12-22 03:19:00
##  3  TonyElHabr  00a8 2015-12-22 03:19:00
##  4     elhabro  00a5 2016-04-01 14:56:28
##  5     elhabro  00a5 2016-04-03 02:56:22
##  6     elhabro  00a5 2016-04-03 02:56:22
##  7     elhabro  00a5 2016-04-03 02:56:22
##  8     elhabro  00a5 2016-04-03 02:56:22
##  9     elhabro  00a5 2016-04-03 18:23:27
## 10     elhabro  00a5 2016-04-06 23:48:42
## # ... with 1,528 more rows
## # A tibble: 2 x 2
##   person total
##    <chr> <int>
## 1 Andrew 17701
## 2   Tony  3283
## # A tibble: 8,121 x 5
## # Groups:   person [2]
##    person            word     n total        freq
##     <chr>           <chr> <int> <int>       <dbl>
##  1 Andrew              f0   862 17701 0.048697814
##  2 Andrew            game   145 17701 0.008191628
##  3 Andrew            time   116 17701 0.006553302
##  4 Andrew     @sailordanj    95 17701 0.005366928
##  5 Andrew             cuz    89 17701 0.005027965
##  6 Andrew             day    88 17701 0.004971471
##  7 Andrew             lol    79 17701 0.004463025
##  8 Andrew @dragonflyjonez    73 17701 0.004124061
##  9 Andrew           gonna    73 17701 0.004124061
## 10 Andrew            haha    67 17701 0.003785097
## # ... with 8,111 more rows
## # A tibble: 6,922 x 3
##               word       Andrew         Tony
##              <chr>        <dbl>        <dbl>
##  1    #classof2016 5.649398e-05 0.0003045995
##  2      #finalfour 5.649398e-05 0.0003045995
##  3     #graduation 5.649398e-05 0.0003045995
##  4   #marchmadness 5.649398e-05 0.0003045995
##  5         #rstats 5.649398e-05 0.0003045995
##  6 @alisongriswold 5.649398e-05 0.0003045995
##  7 @basketballtalk 5.649398e-05 0.0003045995
##  8  @bill_easterly 5.649398e-05 0.0003045995
##  9 @chldishricardo 5.649398e-05 0.0003045995
## 10 @coliegestudent 5.649398e-05 0.0003045995
## # ... with 6,912 more rows
```

Which words are most likely to be shared/different between us?


Words Most Unique to Each Person
haha
ppl
yea
prob
summer
f0
told
lmao
mins
gotta
canada
cable
business
bruno
blatt
agency
21st
john's
dion
couch
-2
0
2
log odds ratio (Andrew / Tony)
Andrew
Tony

```
## # A tibble: 5,858 x 4
##      word       Andrew         Tony logratio
##     <chr>        <dbl>        <dbl>    <dbl>
##  1   haha 0.0031969911 0.0001141944 3.332063
##  2    ppl 0.0022096850 0.0001141944 2.962703
##  3    yea 0.0015984955 0.0001141944 2.638916
##  4   prob 0.0015044664 0.0001141944 2.578292
##  5 summer 0.0012223789 0.0001141944 2.370652
##  6     f0 0.0405735778 0.0038826082 2.346610
##  7   lmao 0.0023507287 0.0002283887 2.331432
##  8   told 0.0011753644 0.0001141944 2.331432
##  9   mins 0.0009873061 0.0001141944 2.157078
## 10  gotta 0.0018805830 0.0002283887 2.108288
## # ... with 5,848 more rows
## # A tibble: 5,858 x 4
##          word       Andrew         Tony    logratio
##         <chr>        <dbl>        <dbl>       <dbl>
##  1     career 0.0005641749 0.0005709718 -0.01197551
##  2        job 0.0005641749 0.0005709718 -0.01197551
##  3     lakers 0.0005641749 0.0005709718 -0.01197551
##  4 #nbafinals 0.0002350729 0.0002283887  0.02884648
##  5        1st 0.0002350729 0.0002283887  0.02884648
##  6    android 0.0002350729 0.0002283887  0.02884648
##  7      bring 0.0002350729 0.0002283887  0.02884648
##  8        bus 0.0002350729 0.0002283887  0.02884648
##  9     buying 0.0002350729 0.0002283887  0.02884648
## 10     caught 0.0002350729 0.0002283887  0.02884648
## # ... with 5,848 more rows
## # A tibble: 5,858 x 4
##      word       Andrew         Tony  logratio
##     <chr>        <dbl>        <dbl>     <dbl>
##  1   haha 3.196991e-03 0.0001141944  3.332063
##  2    ppl 2.209685e-03 0.0001141944  2.962703
##  3    yea 1.598496e-03 0.0001141944  2.638916
##  4   prob 1.504466e-03 0.0001141944  2.578292
##  5 summer 1.222379e-03 0.0001141944  2.370652
##  6     f0 4.057358e-02 0.0038826082  2.346610
##  7   lmao 2.350729e-03 0.0002283887  2.331432
##  8   told 1.175364e-03 0.0001141944  2.331432
##  9  couch 4.701457e-05 0.0004567774 -2.273739
## 10   dion 4.701457e-05 0.0004567774 -2.273739
## # ... with 5,848 more rows
```

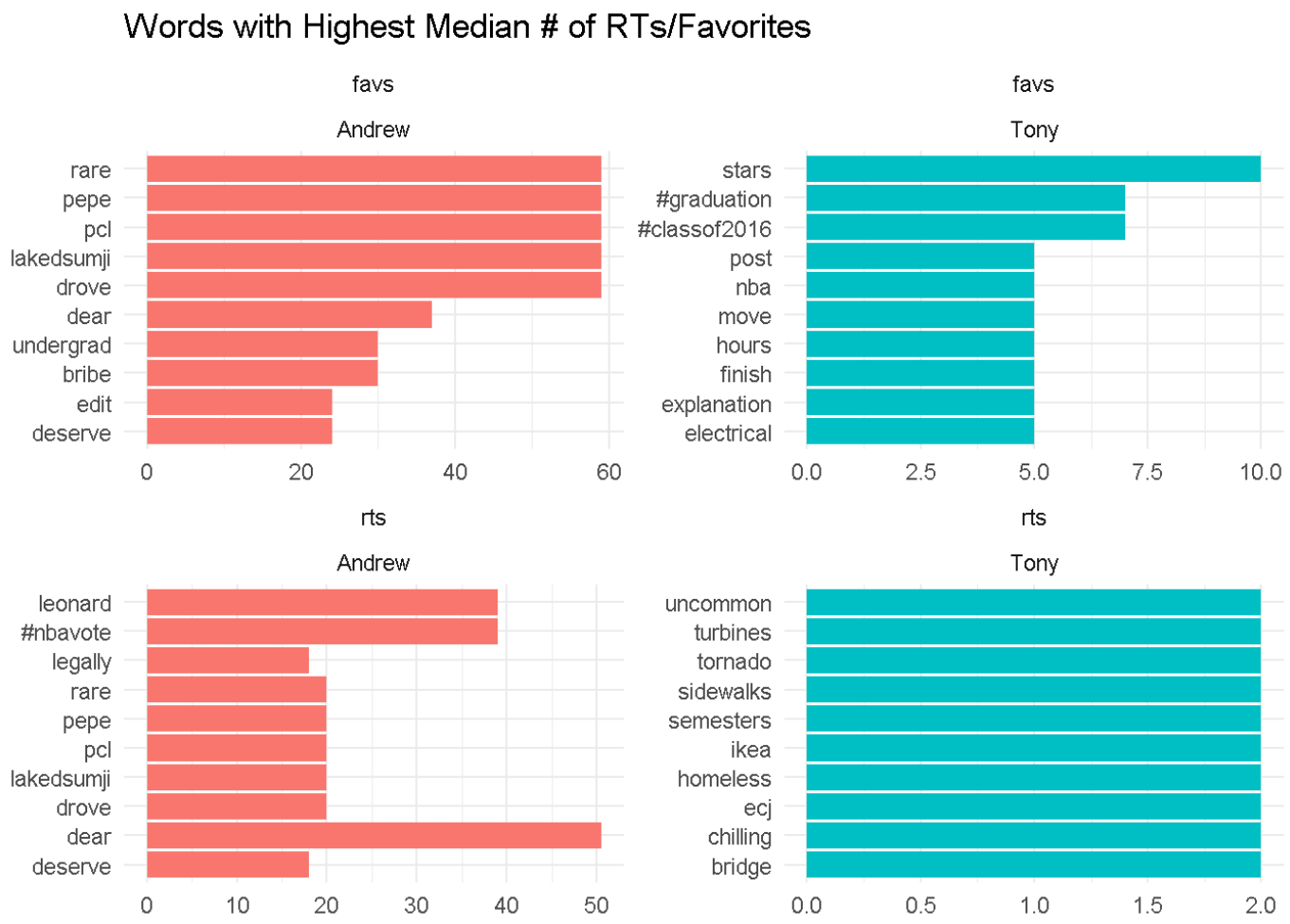
Which words have we used more/less frequently over time?


Largest Changes in Word Frequencey
Andrew
25%
20%
15%
10%
5%
0%
2015
2016
2017
Tony
15%
10%
5%
0%
2016-01
2016-07
2017-01
2017-
word
class
f0
game
kawhi
lol

```
## # A tibble: 1,382 x 6
##    time_floor person    word count time_total word_total
##        <dttm>  <chr>   <chr> <int>      <int>      <int>
##  1 2014-09-01 Andrew     bad     1         73         41
##  2 2014-09-01 Andrew   class     3         73         59
##  3 2014-09-01 Andrew   gonna     1         73         77
##  4 2014-09-01 Andrew   guess     2         73         36
##  5 2014-09-01 Andrew   hours     1         73         40
##  6 2014-09-01 Andrew     lol     1         73         86
##  7 2014-09-01 Andrew   start     1         73         34
##  8 2014-09-01 Andrew twitter     1         73         49
##  9 2014-09-01 Andrew    week     1         73         50
## 10 2014-10-01 Andrew     bro     1        350         43
## # ... with 1,372 more rows
## # A tibble: 102 x 3
##    person    word              data
##     <chr>   <chr>            <list>
##  1 Andrew     bad <tibble [27 x 4]>
##  2 Andrew   class <tibble [24 x 4]>
##  3 Andrew   gonna <tibble [32 x 4]>
##  4 Andrew   guess <tibble [22 x 4]>
##  5 Andrew   hours <tibble [22 x 4]>
##  6 Andrew     lol <tibble [29 x 4]>
##  7 Andrew   start <tibble [22 x 4]>
##  8 Andrew twitter <tibble [20 x 4]>
##  9 Andrew    week <tibble [26 x 4]>
## 10 Andrew     bro <tibble [14 x 4]>
## # ... with 92 more rows
## # A tibble: 102 x 4
##    person    word              data    models
##     <chr>   <chr>            <list>    <list>
##  1 Andrew     bad <tibble [27 x 4]> <S3: glm>
##  2 Andrew   class <tibble [24 x 4]> <S3: glm>
##  3 Andrew   gonna <tibble [32 x 4]> <S3: glm>
##  4 Andrew   guess <tibble [22 x 4]> <S3: glm>
##  5 Andrew   hours <tibble [22 x 4]> <S3: glm>
##  6 Andrew     lol <tibble [29 x 4]> <S3: glm>
##  7 Andrew   start <tibble [22 x 4]> <S3: glm>
##  8 Andrew twitter <tibble [20 x 4]> <S3: glm>
##  9 Andrew    week <tibble [26 x 4]> <S3: glm>
## 10 Andrew     bro <tibble [14 x 4]> <S3: glm>
## # ... with 92 more rows
## # A tibble: 5 x 8
##   person  word       term      estimate    std.error statistic
##    <chr> <chr>      <chr>         <dbl>        <dbl>     <dbl>
## 1 Andrew    f0 time_floor  1.643854e-08 1.730128e-09  9.501347
## 2 Andrew   lol time_floor -1.625020e-08 4.422427e-09 -3.674498
## 3   Tony  game time_floor  7.449339e-08 2.036544e-08  3.657834
## 4   Tony kawhi time_floor  5.225945e-08 1.808690e-08  2.889354
## 5 Andrew class time_floor -1.561842e-08 5.672361e-09 -2.753424
## # ... with 2 more variables: p.value <dbl>, adjusted_p_value <dbl>
```

Tweet Popularity

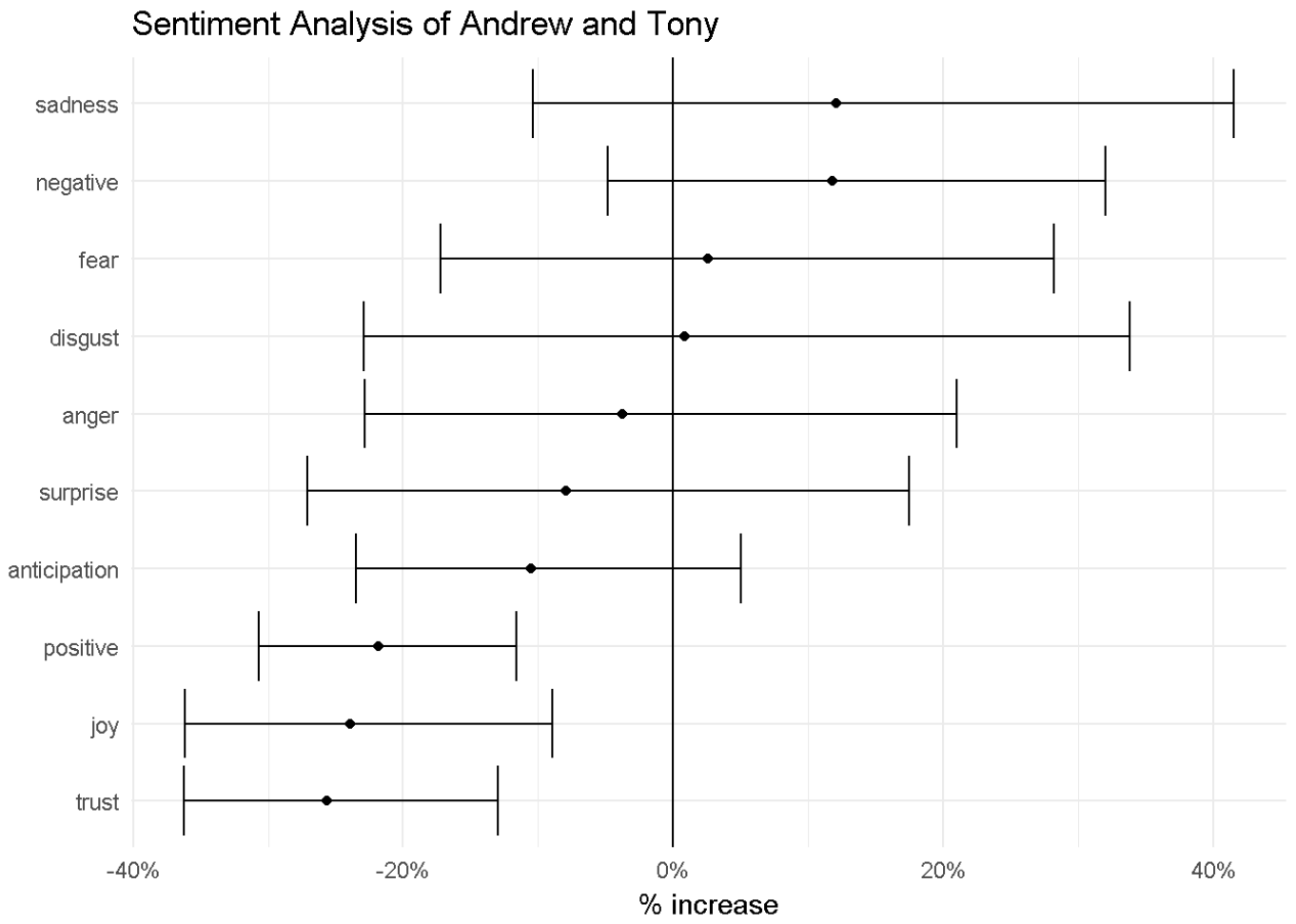
How often do our tweets get liked/favorited/retweeted?

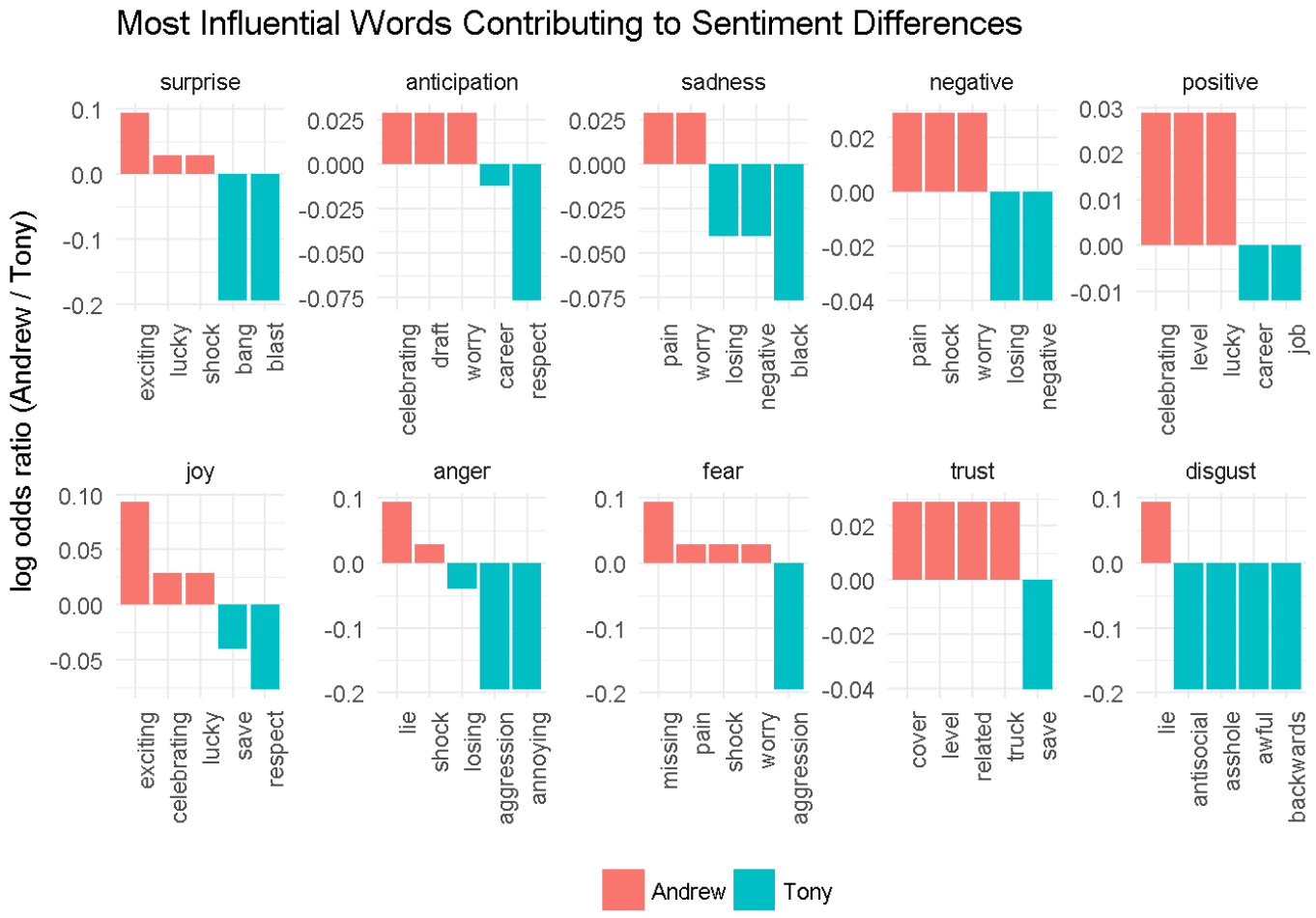


```
## # A tibble: 2 x 10
##   person  uses rts_total favs_total rts_max favs_max rts_avg favs_avg
##    <chr> <int>     <int>      <int>   <dbl>    <dbl>   <dbl>    <dbl>
## 1 Andrew  1148      4414      30412     909      737    3.84    26.49
## 2   Tony   184       131       1025      22       80    0.71     5.57
## # ... with 2 more variables: rts_median <dbl>, favs_median <dbl>
## # A tibble: 2,856 x 4
##    person          word rts_median favs_median
##     <chr>         <chr>      <dbl>       <dbl>
##  1 Andrew           '16        0.0          20
##  2 Andrew           '93        0.0          15
##  3 Andrew            'd        0.0           6
##  4 Andrew           'em        0.5           3
##  5 Andrew       'murica        1.0           5
##  6 Andrew         'twas        3.0          11
##  7 Andrew            #1        0.0           5
##  8 Andrew   #aparnahive        0.0           3
##  9 Andrew   #ballislife        0.0           2
## 10 Andrew #battleof3009        0.0           1
## # ... with 2,846 more rows
## # A tibble: 5,712 x 5
##    person          word  type   calc value
##  *  <chr>         <chr> <chr>  <chr> <dbl>
##  1 Andrew           '16   rts median   0.0
##  2 Andrew           '93   rts median   0.0
##  3 Andrew            'd   rts median   0.0
##  4 Andrew           'em   rts median   0.5
##  5 Andrew       'murica   rts median   1.0
##  6 Andrew         'twas   rts median   3.0
##  7 Andrew            #1   rts median   0.0
##  8 Andrew   #aparnahive   rts median   0.0
##  9 Andrew   #ballislife   rts median   0.0
## 10 Andrew #battleof3009   rts median   0.0
## # ... with 5,702 more rows
```

Sentiment Analysis

What is the sentiment (i.e. “tone”) of our tweets?






```
## # A tibble: 3,642 x 3
##       status_id person total_words
##           <dbl> <chr>       <int>
```

```
##            <dbl>  <chr>        <int>
##  1 5.159100e+17 Andrew       17701
##  2 7.594746e+17 Andrew       17701
##  3 8.284065e+17 Andrew       17701
##  4 8.853525e+17 Andrew       17701
##  5 6.700127e+17 Andrew       17701
##  6 7.972158e+17 Andrew       17701
##  7 5.526847e+17 Andrew       17701
##  8 6.892333e+17 Andrew       17701
##  9 7.172034e+17   Tony        3283
## 10 7.114075e+17 Andrew       17701
## # ... with 3,632 more rows
## # A tibble: 20 x 4
##    person    sentiment total_words words
##     <chr>        <chr>       <int> <dbl>
##  1 Andrew        anger       17701   498
##  2 Andrew anticipation       17701   931
##  3 Andrew      disgust       17701   348
##  4 Andrew         fear       17701   553
##  5 Andrew          joy       17701   652
##  6 Andrew     negative       17701  1049
##  7 Andrew     positive       17701  1404
##  8 Andrew      sadness       17701   550
##  9 Andrew     surprise       17701   427
## 10 Andrew        trust       17701   830
## 11   Tony        anger        3283    96
## 12   Tony anticipation        3283   193
## 13   Tony      disgust        3283    64
## 14   Tony         fear        3283   100
## 15   Tony          joy        3283   159
## 16   Tony     negative        3283   174
## 17   Tony     positive        3283   333
## 18   Tony      sadness        3283    91
## 19   Tony     surprise        3283    86
## 20   Tony        trust        3283   207
## # A tibble: 10 x 4
##       sentiment Andrew   Tony sentiment_diff
##           <chr>  <dbl>  <dbl>          <dbl>
##  1     positive 0.0793 0.1014        -0.0221
##  2        trust 0.0469 0.0631        -0.0162
##  3          joy 0.0368 0.0484        -0.0116
##  4 anticipation 0.0526 0.0588        -0.0062
##  5     surprise 0.0241 0.0262        -0.0021
##  6        anger 0.0281 0.0292        -0.0011
##  7      disgust 0.0197 0.0195         0.0002
##  8         fear 0.0312 0.0305         0.0007
##  9      sadness 0.0311 0.0277         0.0034
## 10     negative 0.0593 0.0530         0.0063
## # A tibble: 10 x 9
##       sentiment  estimate statistic      p.value parameter  conf.low
##           <chr>     <dbl>     <dbl>        <dbl>     <dbl>     <dbl>
##  1        anger 0.9621243       498 7.346559e-01  501.0672 0.7719001
##  2 anticipation 0.8946753       931 1.626770e-01  948.1474 0.7654059
##  3      disgust 1.0084918       348 1.000000e+00  347.5416 0.7706078
##  4         fear 1.0256477       553 8.716172e-01  550.8365 0.8275475
##  5          joy 0.7605426       652 2.685413e-03  684.1170 0.6385506
##  6     negative 1.1181481      1049 1.808206e-01 1031.6585 0.9515492
```

```
##  7     positive 0.7819806      1404 8.331172e-05 1465.2419 0.6934323
##  8      sadness 1.1209710       550 3.281363e-01  540.7139 0.8965403
##  9     surprise 0.9208789       427 4.663807e-01  432.7398 0.7289193
## 10        trust 0.7436710       830 1.972226e-04  874.7587 0.6378809
## # ... with 3 more variables: conf.high <dbl>, method <fctr>,
## #   alternative <fctr>
```

Conclusion

That's it!