Behavior of Twitter Bot in California Fires

Takeshi Oda

2/1/2019

Parameter Settings

```
## Constant Variables ##
threshhold = 0.5
```

Data Loading

```
tweets <- fread("data/CaliforniaFires_Tweet_Stats.csv",
                encoding="UTF-8",
                colClasses = c("numeric",
                               "character",
                               "character",
                               "character",
                               "numeric",   #bot_probability
                               "integer",   #row_num
                               "integer",   #num_words
                               "integer",   #num_question
                               "integer",   #num_exclamation
                               "integer",   #num_digit_screen_name
                               "integer",   #num_political_word
                               "integer",   #num_environmental_word
                               "factor",    #include_retweet
                               "integer"    #num_hashtag
                               ))
```

Exploratory Data Analysis

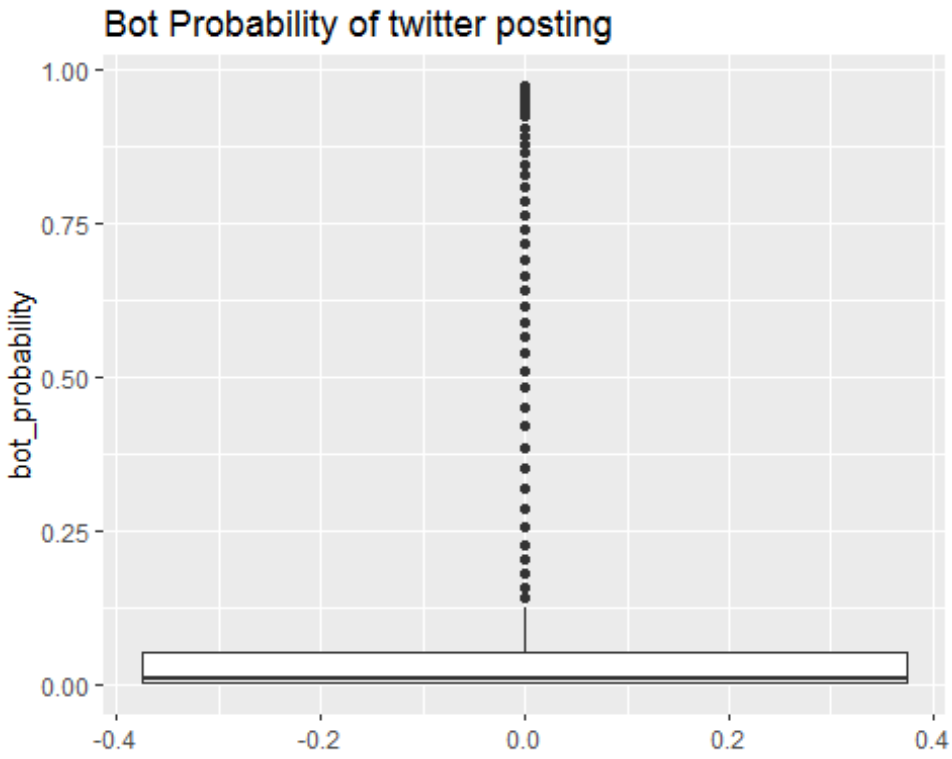
Bot probability was plotted on boxplot and percentiles are identified. According to a prior research, it is estimated between 9% and 15% of active twitter users are bot(Varol, Ferrara, Davis, Menczer & Flammini, 2017). If I apply this percentage into our dataset, the lower boundary of bot probability will be 0.125145748(85% percentile). In this analysis,I chose more conservative threshold 0.5 to decide Bot account.

```
summary(tweets)
```

```
    retweet          user_name          screen_name
 Min.   :    0.0   Length:10955       Length:10955
 1st Qu.:    3.0   Class :character   Class :character
 Median :   38.0   Mode  :character   Mode  :character
```

```
Mean   :  361.9
3rd Qu.:  276.0
Max.   :22481.0
  message          bot_probability       row_num    num_words
Length:10955       Min.   :0.001076   Min.   :0   Min.   :  3.00
Class :character   1st Qu.:0.002857   1st Qu.:0   1st Qu.: 17.00
Mode  :character   Median :0.011259   Median :0   Median : 21.00
                   Mean   :0.069913   Mean   :0   Mean   : 20.84
                   3rd Qu.:0.052723   3rd Qu.:0   3rd Qu.: 24.00
                   Max.   :0.974204   Max.   :0   Max.   :102.00
 num_question     num_exclamation  num_digit_screen_name
Min.   : 0.0000   Min.   :0.0000   Min.   : 0.000
1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.000
Median : 0.0000   Median :0.0000   Median : 0.000
Mean   : 0.1339   Mean   :0.1921   Mean   : 1.029
3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 2.000
Max.   :11.0000   Max.   :9.0000   Max.   :12.000
num_political_word num_environmental_word include_retweet
Min.   :0.00000    Min.   :0.00000        0:1951
1st Qu.:0.00000    1st Qu.:0.00000        1:9004
Median :0.00000    Median :0.00000
Mean   :0.02556    Mean   :0.04874
3rd Qu.:0.00000    3rd Qu.:0.00000
Max.   :2.00000    Max.   :2.00000
 num_hashtag
Min.   : 0.000
1st Qu.: 1.000
Median : 1.000
Mean   : 2.345
3rd Qu.: 3.000
Max.   :23.000
```

```
gf_boxplot(~bot_probability, data=tweets, title="Bot Probability of twitt
er posting")
```



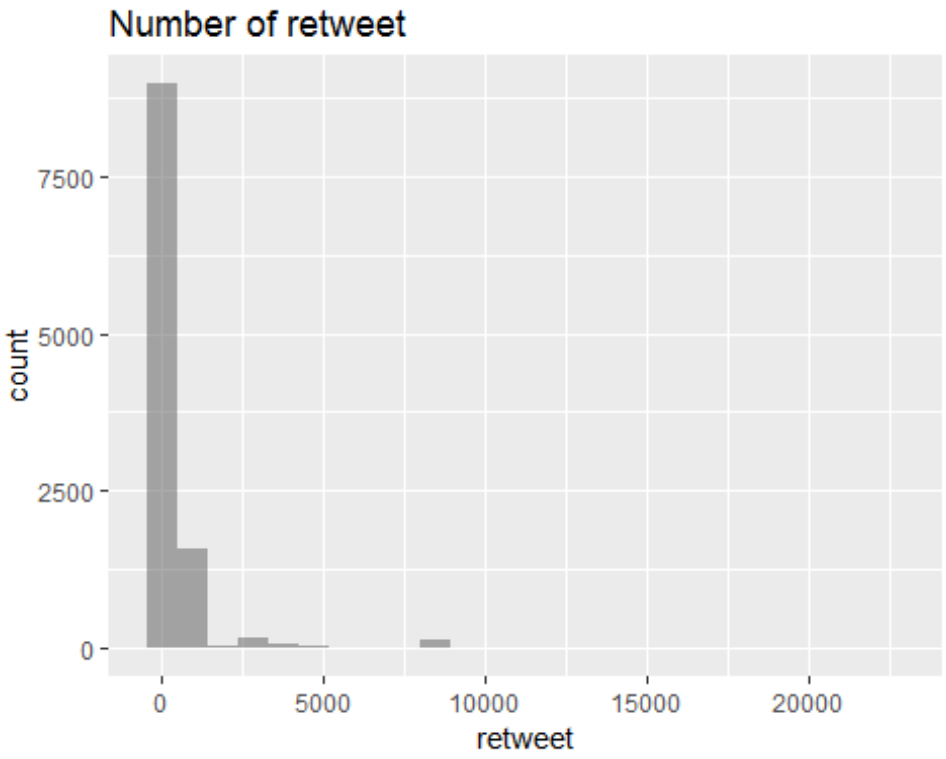

```
quantile(tweets$bot_probability, c(0,0.1,0.25,0.5,0.75,0.85,0.9,1))
```

```
         0%         10%         25%         50%         75%         85%
0.001076330 0.001725458 0.002856940 0.011259347 0.052723484 0.125145748
        90%        100%
0.203165259 0.974204075
```

```
tweets <- tweets %>% mutate(is_bot = ifelse(bot_probability > threshhold,
 "Bot", "Non-Bot"))
```

```
Warning: package 'bindrcpp' was built under R version 3.4.4
```

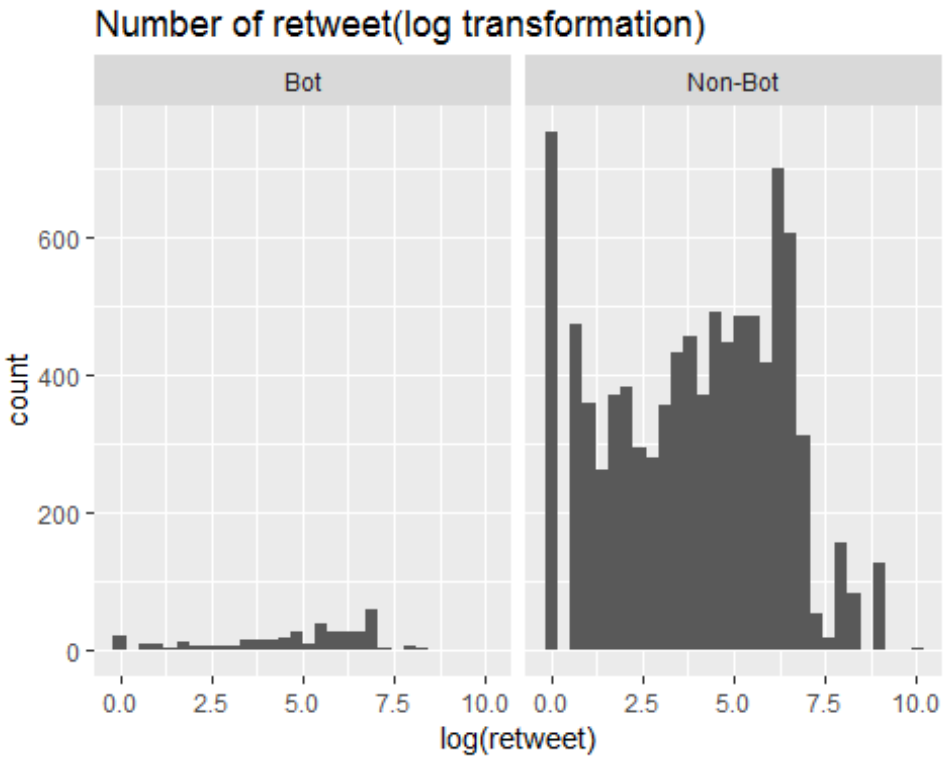
```
#Retweet
gf_histogram(~retweet, data=tweets, title="Number of retweet")
```



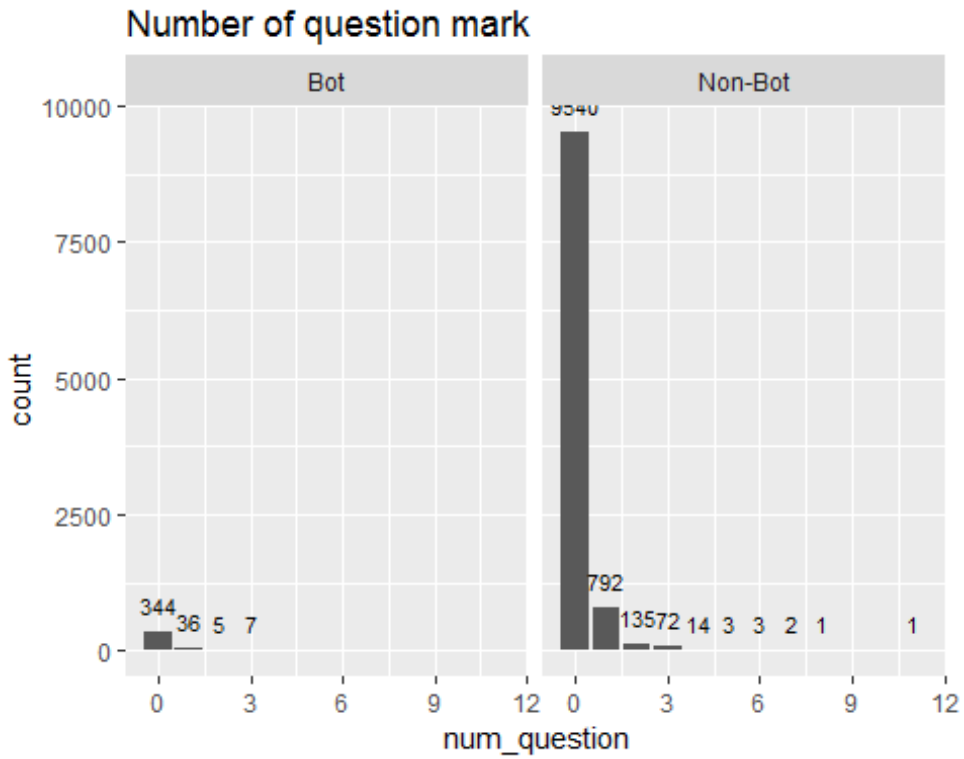

```
ggplot(data=tweets, aes(x=log(retweet))) + facet_grid(. ~is_bot) + geom_h
istogram() + labs(title="Number of retweet(log transformation)")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

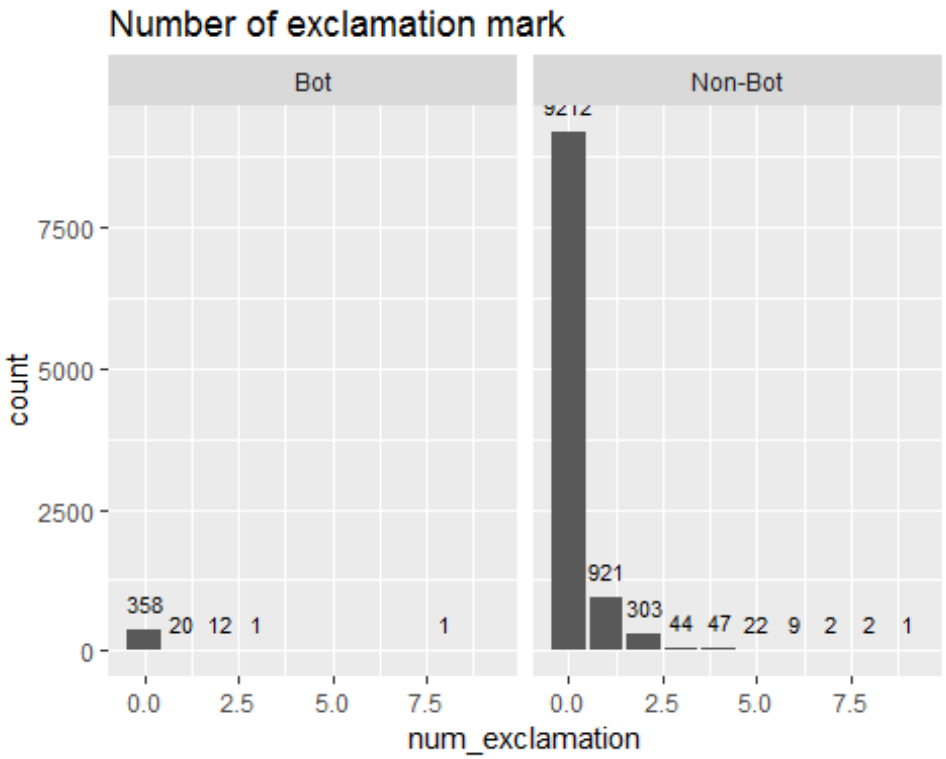
```
Warning: Removed 1413 rows containing non-finite values (stat_bin).
```



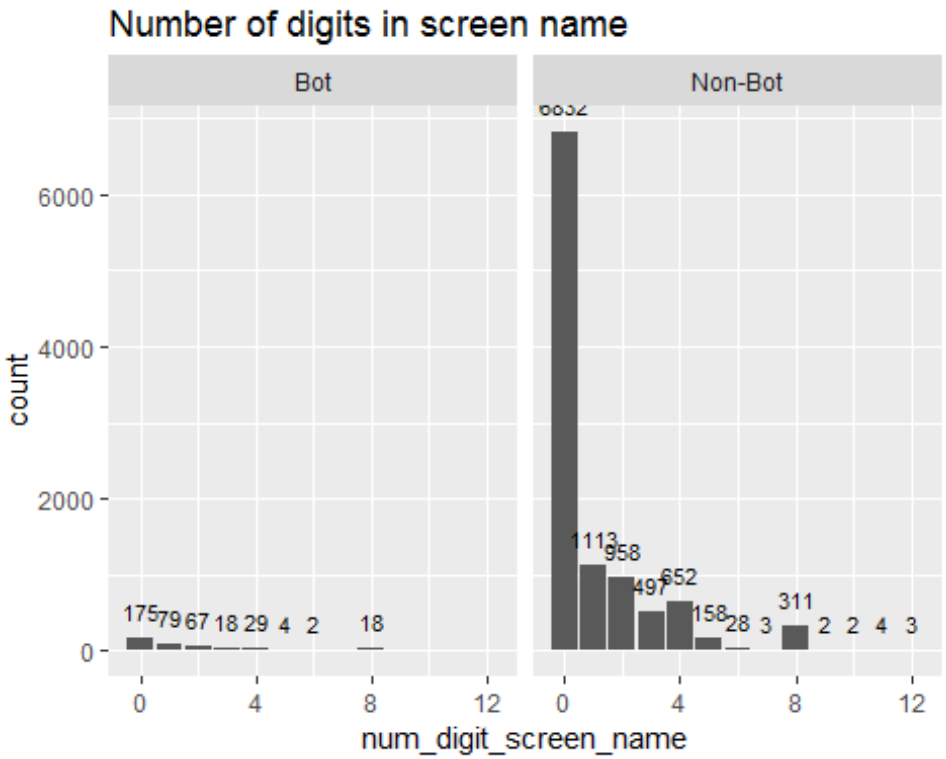

```
#Number of question mark
ggplot(data=tweets, aes(x=num_question)) + facet_grid(. ~is_bot) + geom_b
ar() + geom_text(stat='count', aes(label=..count..), vjust=-1, size=3) +
labs(title="Number of question mark")
```



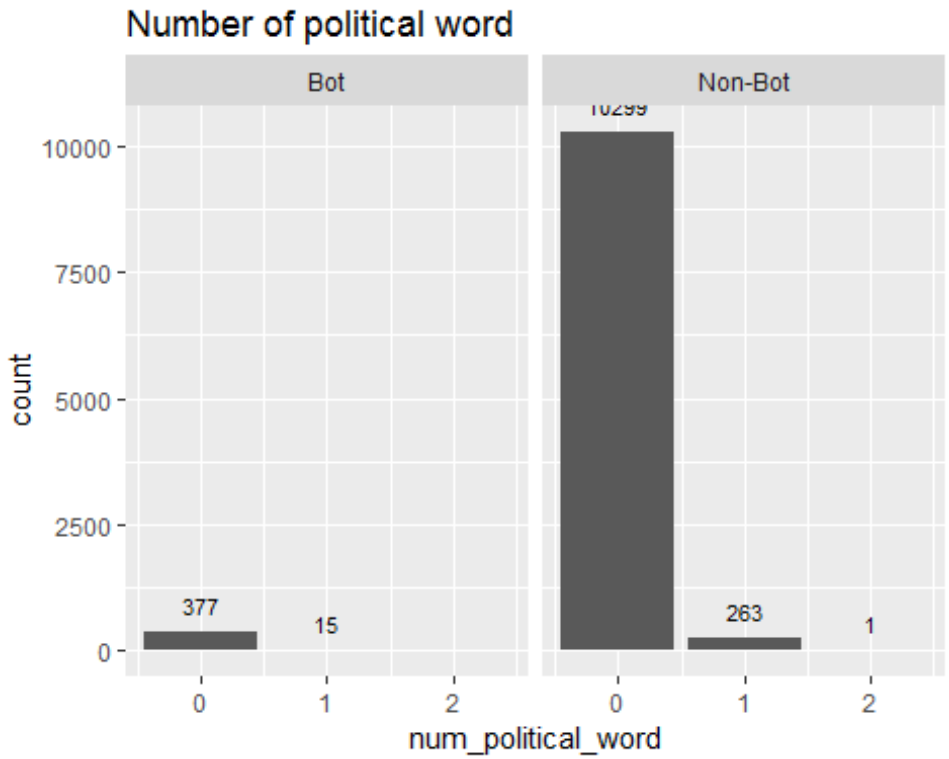

```
#Number of exclamation mark
ggplot(data=tweets, aes(x=num_exclamation)) + facet_grid(. ~is_bot) + geo
m_bar() + geom_text(stat='count', aes(label=..count..), vjust=-1, size=3)
 + labs(title="Number of exclamation mark")
```



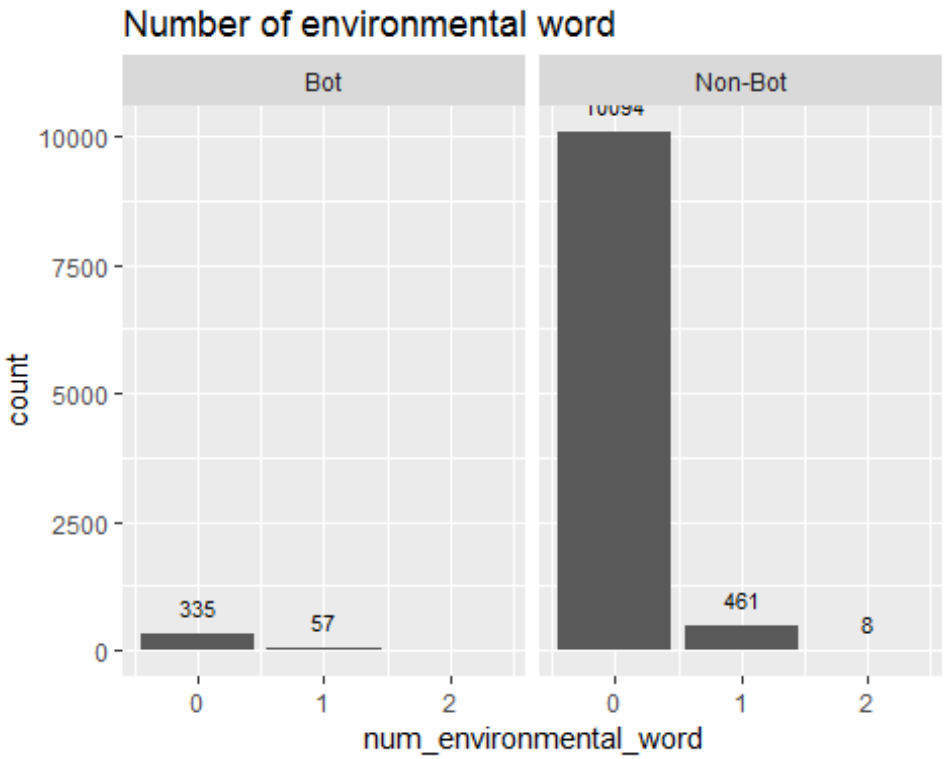

```
#Number of digits in screen name
ggplot(data=tweets, aes(x=num_digit_screen_name)) + facet_grid(. ~is_bot)
 + geom_bar() + geom_text(stat='count', aes(label=..count..), vjust=-1, s
ize=3) + labs(title="Number of digits in screen name")
```



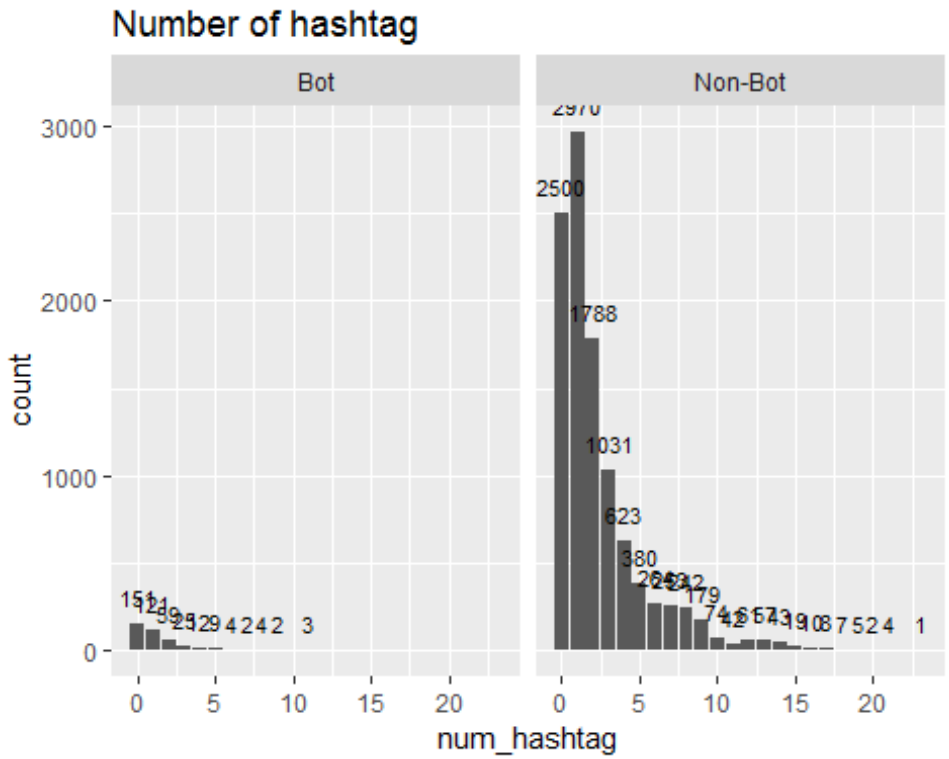

```
#Number of political word
ggplot(data=tweets, aes(x=num_political_word)) + facet_grid(. ~is_bot) +
geom_bar() + geom_text(stat='count', aes(label=..count..), vjust=-1, size
=3) + labs(title="Number of political word")
```




```
#Number of environmental word
ggplot(data=tweets, aes(x=num_environmental_word)) + facet_grid(. ~is_bo
t) + geom_bar() + geom_text(stat='count', aes(label=..count..), vjust=-1,
 size=3) + labs(title="Number of environmental word")
```




```
#Number of hashtag
ggplot(data=tweets, aes(x=num_hashtag)) + facet_grid(. ~is_bot) + geom_ba
r() + geom_text(stat='count', aes(label=..count..), vjust=-1, size=3) + l
abs(title="Number of hashtag")
```

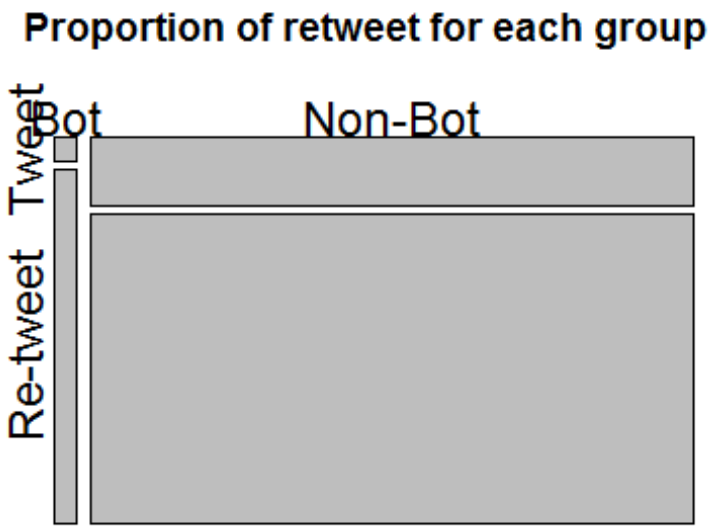



```
bot <- tweets %>% filter(is_bot =="Bot")
no_bot <- tweets %>% filter(is_bot =="Non-Bot")

tbl_ret <- table(tweets$is_bot, tweets$include_retweet)
rownames(tbl_ret) <- c("Bot", "Non-Bot")
colnames(tbl_ret) <- c("Tweet", "Re-tweet")
print(tbl_ret)
```

```
          Tweet Re-tweet
  Bot        26      366
  Non-Bot  1925     8638
```

```
plot(tbl_ret, cex=1.5, main="Proportion of retweet for each group")
```



Statistical Testing

Statistical significance between two groups, i.e. Bots and Non-Bots were tested. Average of seven numerical features were compared through two-sample t-test. One categorical feature was compared through chi-squared goodness-of-fit test.
Since I apply eight testing on single sample, I use adjusted rejection region .00625 (.05 /8).

```
#statistical testing
#Number of retweet
t1 <- t.test(log(bot$retweet+1), log(no_bot$retweet+1))

#Number of question mark
t2 <- t.test(bot$num_question, no_bot$num_question)

#Number of exclamation mark
t3 <- t.test(bot$num_exclamation, no_bot$num_exclamation)

#Number of digit in screen name
t4 <- t.test(bot$num_digit_screen_name, no_bot$num_digit_screen_name)

#Number of political word
t5 <- t.test(bot$num_political_word, no_bot$num_political_word)
```

```r
#Number of environmental word
t6 <- t.test(bot$num_environmental_word, no_bot$num_environmental_word)

#Number of hashtag
t7 <- t.test(bot$num_hashtag, no_bot$num_hashtag)

#Include retweet
t8 <- chisq.test(tbl_ret)

reject_region <- 0.00625

#Create table for variables and their p-values
summary <- tweets %>% group_by(is_bot)
summary <- summary %>% summarize(Avg_Retweet = mean(retweet),
                                 Avg_Question = mean(num_question),
                                 Avg_Exclamation = mean(num_exclamation),
                                 Avg_Digit = mean(num_digit_screen_name),
                                 Avg_Political_Word = mean(num_political_wor
d),
                                 Avg_Environmental_Word = mean(num_environme
ntal_word),
                                 Avg_Hashtag = mean(num_hashtag)
                                 )

#Convert gplyr summary into matrix
m <- t(as.matrix(summary[,-1])) #Remove grouping column 'is_bot'
m <- cbind(m, rep(0,7)) #Column for p-value
m <- cbind(m, rep(0,7)) #Column for Significance

colnames(m) <- c("Bot", "Non-Bot", "P-value", "Difference significant")
#Add p-values into the matrix
m[1,3] <- t1$p.value
m[2,3] <- t2$p.value
m[3,3] <- t3$p.value
m[4,3] <- t4$p.value
m[5,3] <- t5$p.value
m[6,3] <- t6$p.value
m[7,3] <- t7$p.value

#Round value to two decimal places
m <- round(m, 2)

#Add significance into the matrix
m[1,4] <- ifelse(t1$p.value < reject_region, "Significant", "Not Signific
ant")
```

```r
m[2,4] <- ifelse(t2$p.value < reject_region, "Significant", "Not Signific
ant")
m[3,4] <- ifelse(t3$p.value < reject_region, "Significant", "Not Signific
ant")
m[4,4] <- ifelse(t4$p.value < reject_region, "Significant", "Not Signific
ant")
m[5,4] <- ifelse(t5$p.value < reject_region, "Significant", "Not Signific
ant")
m[6,4] <- ifelse(t6$p.value < reject_region, "Significant", "Not Signific
ant")
m[7,4] <- ifelse(t7$p.value < reject_region, "Significant", "Not Signific
ant")


#Show matrix
print(m[1:7,1:4])
```

```
                      Bot      Non-Bot  P-value Difference significant
Avg_Retweet           "422.81" "359.62" "0"     "Significant"
Avg_Question          "0.17"   "0.13"   "0.15"  "Not Significant"
Avg_Exclamation       "0.14"   "0.19"   "0.07"  "Not Significant"
Avg_Digit             "1.43"   "1.01"   "0"     "Significant"
Avg_Political_Word    "0.04"   "0.03"   "0.18"  "Not Significant"
Avg_Environmental_Word "0.15"  "0.05"   "0"     "Significant"
Avg_Hashtag           "1.35"   "2.38"   "0"     "Significant"
```

```r
#Percentage of retweets in bot tweets
retweet_pct_bot <- round(tbl_ret[1,2] / (tbl_ret[1,1] + tbl_ret[1,2]), 2)
#Percentage of retweets in non bot tweets
retweet_pct_notbot <- round(tbl_ret[2,2] / (tbl_ret[2,1] + tbl_ret[2,2]),
 2)

n <- matrix(
        c(
                retweet_pct_notbot,
                retweet_pct_bot,
                round(t8$p.value,2),
                ifelse(t8$p.value < reject_region, "Significant", "Not Si
gnificant")
                ),
        nrow=1,
        dimnames=list(c("Percent of retweet"), c("Not Bot", "Bot", "P-Val
ue", "Proportion Significant")))

print(n)
```

```
                    Not Bot Bot    P-Value Proportion Significant
Percent of retweet "0.82"  "0.93" "0"     "Significant"

# ---- R Output from hypothesis test ---------------------------------
-
#                         Bot      Non-Bot  P-value Difference significant
#Avg_Retweet             "422.81" "359.62" "0"     "Significant"
#Avg_Question            "0.17"   "0.13"   "0.15"  "Not Significant"
#Avg_Exclamation         "0.14"   "0.19"   "0.07"  "Not Significant"
#Avg_Digit               "1.43"   "1.01"   "0"     "Significant"
#Avg_Political_Word      "0.04"   "0.03"   "0.18"  "Not Significant"
#Avg_Environmental_Word "0.15"   "0.05"   "0"     "Significant"
#Avg_Hashtag             "1.35"   "2.38"   "0"     "Significant"
#                    Not Bot Bot    P-Value Proportion Significant
#Percent of retweet "0.82"  "0.93" "0"     "Significant"
#
# ------------------------------------------------------------------------
-
```

Conclusion

We found statistical significance in below behaviors in tweets about California wild fire.

Tweets by Bot are more likely to be retweeted
Screen name of Bot account is more likely to have digits (Same user might have many user accounts for professional purpose)
Tweets by Bot tend to comment about environmental aspects of California fire
Normal tweets are more likely to embed hashtags in their tweets
Tweets by Bot are more likely to be retweets of other tweets

This result might imply that Bot tweets are aimed at propagating specific opinion to the public by retweeting on multiple user accounts.

Reference

Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. Accepted paper for ICWSM'17.