# Social Media

Sarabjit Singh

2022-05-14

```
news0 <- read.csv("OnlineNewsPopularity.csv")
names(news0)
```

```
##  [1] "url"                           "timedelta"
##  [3] "n_tokens_title"                "n_tokens_content"
##  [5] "n_unique_tokens"               "n_non_stop_words"
##  [7] "n_non_stop_unique_tokens"      "num_hrefs"
##  [9] "num_self_hrefs"                "num_imgs"
## [11] "num_videos"                    "average_token_length"
## [13] "num_keywords"                  "data_channel_is_lifestyle"
## [15] "data_channel_is_entertainment" "data_channel_is_bus"
## [17] "data_channel_is_socmed"        "data_channel_is_tech"
## [19] "data_channel_is_world"         "kw_min_min"
## [21] "kw_max_min"                    "kw_avg_min"
## [23] "kw_min_max"                    "kw_max_max"
## [25] "kw_avg_max"                    "kw_min_avg"
## [27] "kw_max_avg"                    "kw_avg_avg"
## [29] "self_reference_min_shares"     "self_reference_max_shares"
## [31] "self_reference_avg_sharess"    "weekday_is_monday"
## [33] "weekday_is_tuesday"            "weekday_is_wednesday"
## [35] "weekday_is_thursday"           "weekday_is_friday"
## [37] "weekday_is_saturday"           "weekday_is_sunday"
## [39] "is_weekend"                    "LDA_00"
## [41] "LDA_01"                        "LDA_02"
## [43] "LDA_03"                        "LDA_04"
## [45] "global_subjectivity"           "global_sentiment_polarity"
## [47] "global_rate_positive_words"    "global_rate_negative_words"
## [49] "rate_positive_words"           "rate_negative_words"
## [51] "avg_positive_polarity"         "min_positive_polarity"
## [53] "max_positive_polarity"         "avg_negative_polarity"
## [55] "min_negative_polarity"         "max_negative_polarity"
## [57] "title_subjectivity"            "title_sentiment_polarity"
## [59] "abs_title_subjectivity"        "abs_title_sentiment_polarity"
## [61] "shares"
```

```
dim(news0)
```

```
## [1] 39644    61
```

```
head(news0)
```

```
##                                                      url timedelta
## 1   http://mashable.com/2013/01/07/amazon-instant-video-browser/       731
## 2    http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/       731
```

```
## 3 http://mashable.com/2013/01/07/apple-40-billion-app-downloads/       731
## 4       http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/       731
## 5              http://mashable.com/2013/01/07/att-u-verse-apps/       731
## 6              http://mashable.com/2013/01/07/beewi-smart-toys/       731
##   n_tokens_title n_tokens_content n_unique_tokens n_non_stop_words
## 1             12              219       0.6635945                1
## 2              9              255       0.6047431                1
## 3              9              211       0.5751295                1
## 4              9              531       0.5037879                1
## 5             13             1072       0.4156456                1
## 6             10              370       0.5598886                1
##   n_non_stop_unique_tokens num_hrefs num_self_hrefs num_imgs num_videos
## 1                0.8153846         4              2        1          0
## 2                0.7919463         3              1        1          0
## 3                0.6638655         3              1        1          0
## 4                0.6656347         9              0        1          0
## 5                0.5408895        19             19       20          0
## 6                0.6981982         2              2        0          0
##   average_token_length num_keywords data_channel_is_lifestyle
## 1             4.680365            5                         0
## 2             4.913725            4                         0
## 3             4.393365            6                         0
## 4             4.404896            7                         0
## 5             4.682836            7                         0
## 6             4.359459            9                         0
##   data_channel_is_entertainment data_channel_is_bus data_channel_is_socmed
## 1                             1                   0                      0
## 2                             0                   1                      0
## 3                             0                   1                      0
## 4                             1                   0                      0
## 5                             0                   0                      0
## 6                             0                   0                      0
##   data_channel_is_tech data_channel_is_world kw_min_min kw_max_min kw_avg_min
## 1                    0                     0         0         0          0
## 2                    0                     0         0         0          0
## 3                    0                     0         0         0          0
## 4                    0                     0         0         0          0
## 5                    1                     0         0         0          0
## 6                    1                     0         0         0          0
##   kw_min_max kw_max_max kw_avg_max kw_min_avg kw_max_avg kw_avg_avg
## 1          0          0          0          0          0          0
## 2          0          0          0          0          0          0
## 3          0          0          0          0          0          0
## 4          0          0          0          0          0          0
## 5          0          0          0          0          0          0
## 6          0          0          0          0          0          0
##   self_reference_min_shares self_reference_max_shares
## 1                       496                       496
## 2                         0                         0
## 3                       918                       918
## 4                         0                         0
## 5                       545                     16000
## 6                      8500                      8500
##   self_reference_avg_sharess weekday_is_monday weekday_is_tuesday
```

```
## 1                   496.000                1                 0
## 2                     0.000                1                 0
## 3                   918.000                1                 0
## 4                     0.000                1                 0
## 5                  3151.158                1                 0
## 6                  8500.000                1                 0
##   weekday_is_wednesday weekday_is_thursday weekday_is_friday
## 1                    0                   0                 0
## 2                    0                   0                 0
## 3                    0                   0                 0
## 4                    0                   0                 0
## 5                    0                   0                 0
## 6                    0                   0                 0
##   weekday_is_saturday weekday_is_sunday is_weekend     LDA_00     LDA_01
## 1                   0                 0          0 0.50033120 0.37827893
## 2                   0                 0          0 0.79975569 0.05004668
## 3                   0                 0          0 0.21779229 0.03333446
## 4                   0                 0          0 0.02857322 0.41929964
## 5                   0                 0          0 0.02863281 0.02879355
## 6                   0                 0          0 0.02224528 0.30671758
##       LDA_02     LDA_03     LDA_04 global_subjectivity
## 1 0.04000468 0.04126265 0.04012254           0.5216171
## 2 0.05009625 0.05010067 0.05000071           0.3412458
## 3 0.03335142 0.03333354 0.68218829           0.7022222
## 4 0.49465083 0.02890472 0.02857160           0.4298497
## 5 0.02857518 0.02857168 0.88542678           0.5135021
## 6 0.02223128 0.02222429 0.62658158           0.4374086
##   global_sentiment_polarity global_rate_positive_words
## 1                0.09256198                 0.04566210
## 2                0.14894781                 0.04313725
## 3                0.32333333                 0.05687204
## 4                0.10070467                 0.04143126
## 5                0.28100348                 0.07462687
## 6                0.07118419                 0.02972973
##   global_rate_negative_words rate_positive_words rate_negative_words
## 1                0.013698630           0.7692308           0.2307692
## 2                0.015686275           0.7333333           0.2666667
## 3                0.009478673           0.8571429           0.1428571
## 4                0.020715631           0.6666667           0.3333333
## 5                0.012126866           0.8602151           0.1397849
## 6                0.027027027           0.5238095           0.4761905
##   avg_positive_polarity min_positive_polarity max_positive_polarity
## 1             0.3786364           0.10000000                   0.7
## 2             0.2869146           0.03333333                   0.7
## 3             0.4958333           0.10000000                   1.0
## 4             0.3859652           0.13636364                   0.8
## 5             0.4111274           0.03333333                   1.0
## 6             0.3506100           0.13636364                   0.6
##   avg_negative_polarity min_negative_polarity max_negative_polarity
## 1            -0.3500000               -0.600           -0.2000000
## 2            -0.1187500               -0.125           -0.1000000
## 3            -0.4666667               -0.800           -0.1333333
## 4            -0.3696970               -0.600           -0.1666667
## 5            -0.2201923               -0.500           -0.0500000
```

```
## 6           -0.1950000                -0.400           -0.1000000
##   title_subjectivity title_sentiment_polarity abs_title_subjectivity
## 1          0.5000000               -0.1875000             0.00000000
## 2          0.0000000                0.0000000             0.50000000
## 3          0.0000000                0.0000000             0.50000000
## 4          0.0000000                0.0000000             0.50000000
## 5          0.4545455                0.1363636             0.04545455
## 6          0.6428571                0.2142857             0.14285714
##   abs_title_sentiment_polarity shares
## 1                    0.1875000    593
## 2                    0.0000000    711
## 3                    0.0000000   1500
## 4                    0.0000000   1200
## 5                    0.1363636    505
## 6                    0.2142857    855
```

### data cleaning

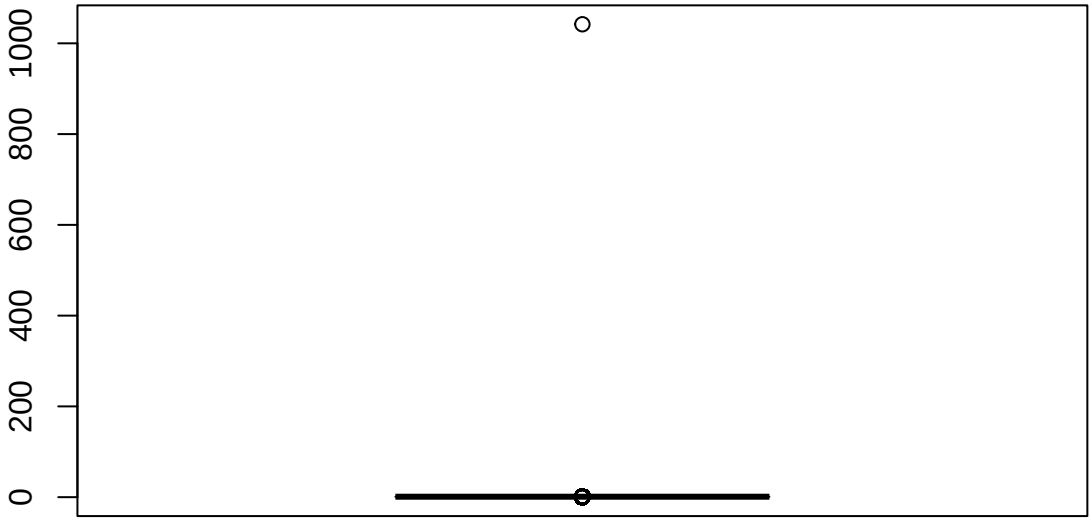
```
#  n_non_stop_words
round(sum(news0$n_non_stop_words==0)/length(news0$n_non_stop_words),2)# 0.97, this variable is almost a
```

```
## [1] 0.03
```

```
summary(news0$n_non_stop_words)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    1.0000    1.0000    0.9965    1.0000 1042.0000
```

```
boxplot(news0$n_non_stop_words,xlab=" n_non_stop_words")
```




```
#install.packages("pastecs")
library(pastecs)
news0n0 <-news0[,-c(14:19)]
names(news0n0)
```

```
##  [1] "url"                          "timedelta"
##  [3] "n_tokens_title"               "n_tokens_content"
##  [5] "n_unique_tokens"              "n_non_stop_words"
##  [7] "n_non_stop_unique_tokens"     "num_hrefs"
##  [9] "num_self_hrefs"               "num_imgs"
## [11] "num_videos"                   "average_token_length"
## [13] "num_keywords"                 "kw_min_min"
## [15] "kw_max_min"                   "kw_avg_min"
## [17] "kw_min_max"                   "kw_max_max"
## [19] "kw_avg_max"                   "kw_min_avg"
## [21] "kw_max_avg"                   "kw_avg_avg"
## [23] "self_reference_min_shares"    "self_reference_max_shares"
## [25] "self_reference_avg_sharess"   "weekday_is_monday"
## [27] "weekday_is_tuesday"           "weekday_is_wednesday"
## [29] "weekday_is_thursday"          "weekday_is_friday"
## [31] "weekday_is_saturday"          "weekday_is_sunday"
## [33] "is_weekend"                   "LDA_00"
## [35] "LDA_01"                       "LDA_02"
## [37] "LDA_03"                       "LDA_04"
## [39] "global_subjectivity"          "global_sentiment_polarity"
## [41] "global_rate_positive_words"   "global_rate_negative_words"
## [43] "rate_positive_words"          "rate_negative_words"
## [45] "avg_positive_polarity"        "min_positive_polarity"
## [47] "max_positive_polarity"        "avg_negative_polarity"
## [49] "min_negative_polarity"        "max_negative_polarity"
## [51] "title_subjectivity"           "title_sentiment_polarity"
## [53] "abs_title_subjectivity"       "abs_title_sentiment_polarity"
## [55] "shares"
```

```
news0n1 <-news0n0[,-c(26:33)]
sumad <- stat.desc(news0n1)

options(scipen = 999,digits=1)
summary_new<-as.data.frame(t(sumad[c(4,5,8,9),c(-1,-2)]))
summary_new
```

```
##                                 min      max    median      mean
## n_tokens_title                  2.0     23.0     10.00     10.40
## n_tokens_content                0.0   8474.0    409.00    546.51
## n_unique_tokens                 0.0    701.0      0.54      0.55
## n_non_stop_words                0.0   1042.0      1.00      1.00
## n_non_stop_unique_tokens        0.0    650.0      0.69      0.69
## num_hrefs                       0.0    304.0      8.00     10.88
## num_self_hrefs                  0.0    116.0      3.00      3.29
## num_imgs                        0.0    128.0      1.00      4.54
## num_videos                      0.0     91.0      0.00      1.25
## average_token_length            0.0      8.0      4.66      4.55
## num_keywords                    1.0     10.0      7.00      7.22
## kw_min_min                     -1.0    377.0     -1.00     26.11
## kw_max_min                      0.0 298400.0    660.00   1153.95
## kw_avg_min                     -1.0  42827.9    235.50    312.37
## kw_min_max                      0.0 843300.0   1400.00  13612.35
## kw_max_max                      0.0 843300.0 843300.00 752324.07
## kw_avg_max                      0.0 843300.0 244572.22 259281.94
## kw_min_avg                     -1.0   3613.0   1023.64   1117.15
```

```
## kw_max_avg                   0.0 298400.0   4355.69   5657.21
## kw_avg_avg                   0.0  43567.7   2870.07   3135.86
## self_reference_min_shares    0.0 843300.0   1200.00   3998.76
## self_reference_max_shares    0.0 843300.0   2800.00  10329.21
## self_reference_avg_sharess   0.0 843300.0   2200.00   6401.70
## LDA_00                       0.0      0.9      0.03      0.18
## LDA_01                       0.0      0.9      0.03      0.14
## LDA_02                       0.0      0.9      0.04      0.22
## LDA_03                       0.0      0.9      0.04      0.22
## LDA_04                       0.0      0.9      0.04      0.23
## global_subjectivity          0.0      1.0      0.45      0.44
## global_sentiment_polarity   -0.4      0.7      0.12      0.12
## global_rate_positive_words   0.0      0.2      0.04      0.04
## global_rate_negative_words   0.0      0.2      0.02      0.02
## rate_positive_words          0.0      1.0      0.71      0.68
## rate_negative_words          0.0      1.0      0.28      0.29
## avg_positive_polarity        0.0      1.0      0.36      0.35
## min_positive_polarity        0.0      1.0      0.10      0.10
## max_positive_polarity        0.0      1.0      0.80      0.76
## avg_negative_polarity       -1.0      0.0     -0.25     -0.26
## min_negative_polarity       -1.0      0.0     -0.50     -0.52
## max_negative_polarity       -1.0      0.0     -0.10     -0.11
## title_subjectivity           0.0      1.0      0.15      0.28
## title_sentiment_polarity    -1.0      1.0      0.00      0.07
## abs_title_subjectivity       0.0      0.5      0.50      0.34
## abs_title_sentiment_polarity 0.0      1.0      0.00      0.16
## shares                       1.0 843300.0   1400.00   3395.38
```

```
# kw_min_min:Worst Keyword (Min. Shares). It is weird to have negative value in this variable, we will
sum(news0$kw_min_min==-1)  #22980 cases
```

```
## [1] 22980
```

```
sum(news0$kw_avg_min==-1) # 694
```

```
## [1] 694
```

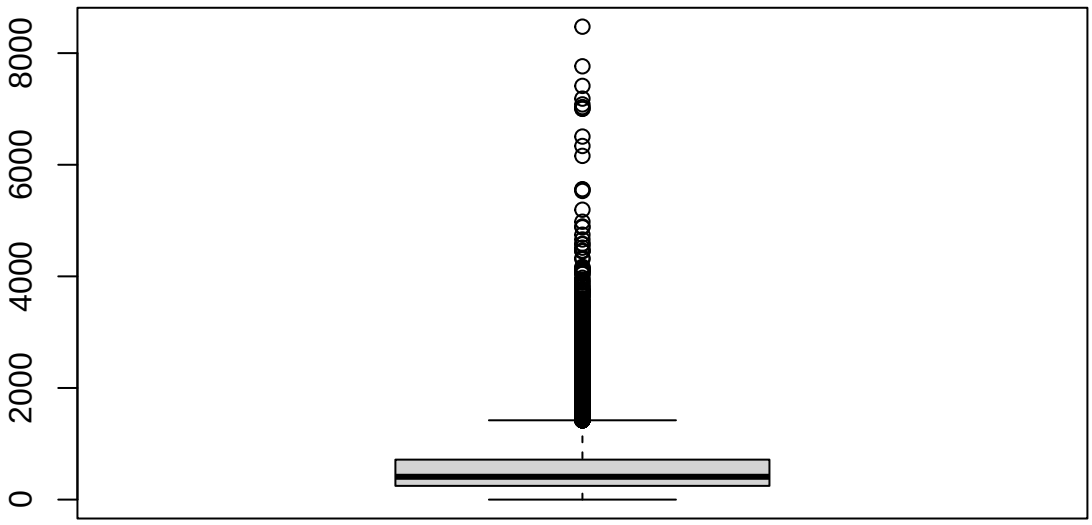
```
sum(news0$kw_avg_max==-1) # 0
```

```
## [1] 0
```

```
cor(news0[20:22]) # since kw_max_min and kw_avg_min are highly related(r=0.94), we will choose kw_avg_m
```

```
##             kw_min_min kw_max_min kw_avg_min
## kw_min_min        1.00       0.01        0.1
## kw_max_min        0.01       1.00        0.9
## kw_avg_min        0.11       0.94        1.0
```

```
variable_name <- c("kw_min_min","kw_avg_min","kw_avg_max")
wrong_value <- c(22980,694,0)

missing_percent <- c(22980/nrow(news0),694/nrow(news0),0)


# n_tokens_content
boxplot(news0$n_tokens_content,xlab="n_tokens_content")
```




n_tokens_content

```
sum(news0$n_tokens_content==0)/length(news0$n_tokens_content)
```

```
## [1] 0.03
```

```
# 0.02979013, 1181 cases
# we will treat 0 as missing value
```

```
# n_unique_tokens is Rate Of Unique Words In The Content, the value range should be (0,1),any value gre
 indt <- which(news0$n_unique_tokens>1)
 wr <- news0[indt,]
wr[,-wr$timedelta] # we can drop this case in our data
```

```
##                                                                 url
## 31038 http://mashable.com/2014/08/18/ukraine-civilian-convoy-attacked/
##       timedelta n_tokens_title n_tokens_content n_unique_tokens
## 31038       142              9             1570              701
##       n_non_stop_words n_non_stop_unique_tokens num_hrefs num_self_hrefs
## 31038             1042                       650        11             10
##       num_imgs num_videos average_token_length num_keywords
## 31038       51          0                    5            7
##       data_channel_is_lifestyle data_channel_is_entertainment
## 31038                         0                             1
##       data_channel_is_bus data_channel_is_socmed data_channel_is_tech
## 31038                   0                      0                    0
##       data_channel_is_world kw_min_min kw_max_min kw_avg_min kw_min_max
## 31038                     0         -1        778        144      23100
##       kw_max_max kw_avg_max kw_min_avg kw_max_avg kw_avg_avg
## 31038     843300     330443       2421       3491       2912
##       self_reference_min_shares self_reference_max_shares
## 31038                       795                         0
##       self_reference_avg_sharess weekday_is_monday weekday_is_tuesday
## 31038                       6924                 0                  1
##       weekday_is_wednesday weekday_is_thursday weekday_is_friday
## 31038                    0                   0                 0
```

```
##       weekday_is_saturday weekday_is_sunday is_weekend LDA_00 LDA_01 LDA_02
## 31038                   0                 0          0      0      0      0
##       LDA_03 LDA_04 global_subjectivity global_sentiment_polarity
## 31038      0      0                   0                         0
##       global_rate_positive_words global_rate_negative_words rate_positive_words
## 31038                          0                          0                   0
##       rate_negative_words avg_positive_polarity min_positive_polarity
## 31038                   0                     0                     0
##       max_positive_polarity avg_negative_polarity min_negative_polarity
## 31038                     0                     0                     0
##       max_negative_polarity title_subjectivity title_sentiment_polarity
## 31038                     0                  0                         0
##       abs_title_subjectivity abs_title_sentiment_polarity shares
## 31038                      0                            0   5900
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract() masks pastecs::extract()
## x dplyr::filter()  masks stats::filter()
## x dplyr::first()   masks pastecs::first()
## x dplyr::lag()     masks stats::lag()
## x dplyr::last()    masks pastecs::last()
```

```
sum(news0$n_unique_tokens>1)
```

```
## [1] 1
```

```
sum(news0$n_non_stop_unique_tokens>1)
```

```
## [1] 1
```

```
news<-news0%>%
  dplyr::filter( n_tokens_content>0 & n_unique_tokens<=1) %>%
 dplyr::select(-kw_min_min,-kw_avg_min,-url,-timedelta,-n_non_stop_words ) # drop non-predictive variab

Missing_percent <- 1-nrow(news)/nrow(news0) #  0.02981536
Missing_case<- nrow(news0)-nrow(news) # 1182 the total cases we delete
Total_case <-nrow(news0)
Model_case <- nrow(news)

Dt <- data.frame(Missing_percent,Missing_case,Total_case,Model_case)
Dt
```

```
##   Missing_percent Missing_case Total_case Model_case
## 1            0.03         1182      39644      38462
```

## data exploration

### data discription

```
names(news)
```

```
##  [1] "n_tokens_title"               "n_tokens_content"
##  [3] "n_unique_tokens"              "n_non_stop_unique_tokens"
##  [5] "num_hrefs"                    "num_self_hrefs"
##  [7] "num_imgs"                     "num_videos"
##  [9] "average_token_length"         "num_keywords"
## [11] "data_channel_is_lifestyle"    "data_channel_is_entertainment"
## [13] "data_channel_is_bus"          "data_channel_is_socmed"
## [15] "data_channel_is_tech"         "data_channel_is_world"
## [17] "kw_max_min"                   "kw_min_max"
## [19] "kw_max_max"                   "kw_avg_max"
## [21] "kw_min_avg"                   "kw_max_avg"
## [23] "kw_avg_avg"                   "self_reference_min_shares"
## [25] "self_reference_max_shares"    "self_reference_avg_sharess"
## [27] "weekday_is_monday"            "weekday_is_tuesday"
## [29] "weekday_is_wednesday"         "weekday_is_thursday"
## [31] "weekday_is_friday"            "weekday_is_saturday"
## [33] "weekday_is_sunday"            "is_weekend"
## [35] "LDA_00"                       "LDA_01"
## [37] "LDA_02"                       "LDA_03"
## [39] "LDA_04"                       "global_subjectivity"
## [41] "global_sentiment_polarity"    "global_rate_positive_words"
## [43] "global_rate_negative_words"   "rate_positive_words"
## [45] "rate_negative_words"          "avg_positive_polarity"
## [47] "min_positive_polarity"        "max_positive_polarity"
## [49] "avg_negative_polarity"        "min_negative_polarity"
## [51] "max_negative_polarity"        "title_subjectivity"
## [53] "title_sentiment_polarity"     "abs_title_subjectivity"
## [55] "abs_title_sentiment_polarity" "shares"
```

```
summary(news)
```

```
##  n_tokens_title n_tokens_content n_unique_tokens n_non_stop_unique_tokens
##  Min.   : 2     Min.   :  18     Min.   :0.1     Min.   :0.1
##  1st Qu.: 9     1st Qu.: 259     1st Qu.:0.5     1st Qu.:0.6
##  Median :10     Median : 423     Median :0.5     Median :0.7
##  Mean   :10     Mean   : 563     Mean   :0.5     Mean   :0.7
##  3rd Qu.:12     3rd Qu.: 729     3rd Qu.:0.6     3rd Qu.:0.8
##  Max.   :23     Max.   :8474     Max.   :1.0     Max.   :1.0
##    num_hrefs   num_self_hrefs    num_imgs     num_videos average_token_length
##  Min.   :  0   Min.   :  0    Min.   :  0   Min.   : 0   Min.   :4
##  1st Qu.:  5   1st Qu.:  1    1st Qu.:  1   1st Qu.: 0   1st Qu.:4
##  Median :  8   Median :  3    Median :  1   Median : 0   Median :5
##  Mean   : 11   Mean   :  3    Mean   :  5   Mean   : 1   Mean   :5
##  3rd Qu.: 14   3rd Qu.:  4    3rd Qu.:  4   3rd Qu.: 1   3rd Qu.:5
##  Max.   :304   Max.   :116    Max.   :128   Max.   :91   Max.   :8
##   num_keywords data_channel_is_lifestyle data_channel_is_entertainment
##  Min.   : 1    Min.   :0.0              Min.   :0.0
##  1st Qu.: 6    1st Qu.:0.0              1st Qu.:0.0
##  Median : 7    Median :0.0              Median :0.0
```

```
##  Mean   : 7    Mean   :0.1              Mean   :0.2
##  3rd Qu.: 9    3rd Qu.:0.0              3rd Qu.:0.0
##  Max.   :10    Max.   :1.0              Max.   :1.0
##  data_channel_is_bus data_channel_is_socmed data_channel_is_tech
##  Min.   :0.0         Min.   :0.0            Min.   :0.0
##  1st Qu.:0.0         1st Qu.:0.0            1st Qu.:0.0
##  Median :0.0         Median :0.0            Median :0.0
##  Mean   :0.2         Mean   :0.1            Mean   :0.2
##  3rd Qu.:0.0         3rd Qu.:0.0            3rd Qu.:0.0
##  Max.   :1.0         Max.   :1.0            Max.   :1.0
##  data_channel_is_world   kw_max_min       kw_min_max       kw_max_max
##  Min.   :0.0           Min.   :     0   Min.   :     0   Min.   :     0
##  1st Qu.:0.0           1st Qu.:   445   1st Qu.:     0   1st Qu.:843300
##  Median :0.0           Median :   660   Median :  1400   Median :843300
##  Mean   :0.2           Mean   :  1152   Mean   : 13182   Mean   :750315
##  3rd Qu.:0.0           3rd Qu.:  1000   3rd Qu.:  7700   3rd Qu.:843300
##  Max.   :1.0           Max.   :298400   Max.   :843300   Max.   :843300
##    kw_avg_max       kw_min_avg     kw_max_avg       kw_avg_avg
##  Min.   :     0   Min.   : -1   Min.   :     0   Min.   :    0
##  1st Qu.:171300   1st Qu.:   0   1st Qu.:  3549   1st Qu.: 2374
##  Median :242080   Median :1009   Median :  4312   Median : 2851
##  Mean   :255213   Mean   :1102   Mean   :  5604   Mean   : 3103
##  3rd Qu.:326864   3rd Qu.:2031   3rd Qu.:  5962   3rd Qu.: 3551
##  Max.   :843300   Max.   :3613   Max.   :298400   Max.   :43568
##  self_reference_min_shares self_reference_max_shares self_reference_avg_sharess
##  Min.   :     0            Min.   :     0            Min.   :     0
##  1st Qu.:   703            1st Qu.:  1200            1st Qu.:  1100
##  Median :  1200            Median :  3000            Median :  2300
##  Mean   :  4122            Mean   : 10647            Mean   :  6598
##  3rd Qu.:  2700            3rd Qu.:  8200            3rd Qu.:  5300
##  Max.   :843300            Max.   :843300            Max.   :843300
##  weekday_is_monday weekday_is_tuesday weekday_is_wednesday weekday_is_thursday
##  Min.   :0.0       Min.   :0.0        Min.   :0.0          Min.   :0.0
##  1st Qu.:0.0       1st Qu.:0.0        1st Qu.:0.0          1st Qu.:0.0
##  Median :0.0       Median :0.0        Median :0.0          Median :0.0
##  Mean   :0.2       Mean   :0.2        Mean   :0.2          Mean   :0.2
##  3rd Qu.:0.0       3rd Qu.:0.0        3rd Qu.:0.0          3rd Qu.:0.0
##  Max.   :1.0       Max.   :1.0        Max.   :1.0          Max.   :1.0
##  weekday_is_friday weekday_is_saturday weekday_is_sunday   is_weekend
##  Min.   :0.0       Min.   :0.0         Min.   :0.0       Min.   :0.0
##  1st Qu.:0.0       1st Qu.:0.0         1st Qu.:0.0       1st Qu.:0.0
##  Median :0.0       Median :0.0         Median :0.0       Median :0.0
##  Mean   :0.1       Mean   :0.1         Mean   :0.1       Mean   :0.1
##  3rd Qu.:0.0       3rd Qu.:0.0         3rd Qu.:0.0       3rd Qu.:0.0
##  Max.   :1.0       Max.   :1.0         Max.   :1.0       Max.   :1.0
##      LDA_00        LDA_01        LDA_02        LDA_03        LDA_04
##  Min.   :0.0   Min.   :0.0   Min.   :0.0   Min.   :0.0   Min.   :0.0
##  1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.0
##  Median :0.0   Median :0.0   Median :0.0   Median :0.0   Median :0.1
##  Mean   :0.2   Mean   :0.1   Mean   :0.2   Mean   :0.2   Mean   :0.2
##  3rd Qu.:0.3   3rd Qu.:0.2   3rd Qu.:0.3   3rd Qu.:0.3   3rd Qu.:0.4
##  Max.   :0.9   Max.   :0.9   Max.   :0.9   Max.   :0.9   Max.   :0.9
##  global_subjectivity global_sentiment_polarity global_rate_positive_words
##  Min.   :0.0         Min.   :-0.4              Min.   :0.00
```

```
##  1st Qu.:0.4         1st Qu.: 0.1              1st Qu.:0.03
##  Median :0.5         Median : 0.1              Median :0.04
##  Mean   :0.5         Mean   : 0.1              Mean   :0.04
##  3rd Qu.:0.5         3rd Qu.: 0.2              3rd Qu.:0.05
##  Max.   :1.0         Max.   : 0.7              Max.   :0.16
##  global_rate_negative_words rate_positive_words rate_negative_words
##  Min.   :0.00               Min.   :0.0         Min.   :0.0
##  1st Qu.:0.01               1st Qu.:0.6         1st Qu.:0.2
##  Median :0.02               Median :0.7         Median :0.3
##  Mean   :0.02               Mean   :0.7         Mean   :0.3
##  3rd Qu.:0.02               3rd Qu.:0.8         3rd Qu.:0.4
##  Max.   :0.18               Max.   :1.0         Max.   :1.0
##  avg_positive_polarity min_positive_polarity max_positive_polarity
##  Min.   :0.0           Min.   :0.0           Min.   :0.0
##  1st Qu.:0.3           1st Qu.:0.0           1st Qu.:0.6
##  Median :0.4           Median :0.1           Median :0.8
##  Mean   :0.4           Mean   :0.1           Mean   :0.8
##  3rd Qu.:0.4           3rd Qu.:0.1           3rd Qu.:1.0
##  Max.   :1.0           Max.   :1.0           Max.   :1.0
##  avg_negative_polarity min_negative_polarity max_negative_polarity
##  Min.   :-1.0          Min.   :-1.0          Min.   :-1.0
##  1st Qu.:-0.3          1st Qu.:-0.7          1st Qu.:-0.1
##  Median :-0.3          Median :-0.5          Median :-0.1
##  Mean   :-0.3          Mean   :-0.5          Mean   :-0.1
##  3rd Qu.:-0.2          3rd Qu.:-0.3          3rd Qu.: 0.0
##  Max.   : 0.0          Max.   : 0.0          Max.   : 0.0
##  title_subjectivity title_sentiment_polarity abs_title_subjectivity
##  Min.   :0.0        Min.   :-1.0             Min.   :0.0
##  1st Qu.:0.0        1st Qu.: 0.0             1st Qu.:0.2
##  Median :0.1        Median : 0.0             Median :0.5
##  Mean   :0.3        Mean   : 0.1             Mean   :0.3
##  3rd Qu.:0.5        3rd Qu.: 0.1             3rd Qu.:0.5
##  Max.   :1.0        Max.   : 1.0             Max.   :0.5
##  abs_title_sentiment_polarity     shares
##  Min.   :0.0                  Min.   :     1
##  1st Qu.:0.0                  1st Qu.:   945
##  Median :0.0                  Median :  1400
##  Mean   :0.2                  Mean   :  3355
##  3rd Qu.:0.2                  3rd Qu.:  2700
##  Max.   :1.0                  Max.   :843300
```

```
indtm <- which(news$shares==843300)
wrm <- news[indtm,]
wrm
```

```
##      n_tokens_title n_tokens_content n_unique_tokens n_non_stop_unique_tokens
## 9292             12              688             0.5                       0.6
##      num_hrefs num_self_hrefs num_imgs num_videos average_token_length
## 9292        28              3       15          1                    5
##      num_keywords data_channel_is_lifestyle data_channel_is_entertainment
## 9292            6                         0                             0
##      data_channel_is_bus data_channel_is_socmed data_channel_is_tech
## 9292                   0                      0                    0
##      data_channel_is_world kw_max_min kw_min_max kw_max_max kw_avg_max
## 9292                     0       1100       2800     690400     430717
```

```
##      kw_min_avg kw_max_avg kw_avg_avg self_reference_min_shares
## 9292       1700       5117       3753                      2100
##      self_reference_max_shares self_reference_avg_sharess weekday_is_monday
## 9292                     40000                      17067                 0
##      weekday_is_tuesday weekday_is_wednesday weekday_is_thursday
## 9292                  0                    1                   0
##      weekday_is_friday weekday_is_saturday weekday_is_sunday is_weekend LDA_00
## 9292                 0                   0                 0          0   0.03
##      LDA_01 LDA_02 LDA_03 LDA_04 global_subjectivity global_sentiment_polarity
## 9292   0.03   0.03    0.7    0.2                 0.5                        0.2
##      global_rate_positive_words global_rate_negative_words rate_positive_words
## 9292                       0.06                       0.02                0.8
##      rate_negative_words avg_positive_polarity min_positive_polarity
## 9292                 0.2                   0.3                  0.05
##      max_positive_polarity avg_negative_polarity min_negative_polarity
## 9292                     1                  -0.2                  -0.4
##      max_negative_polarity title_subjectivity title_sentiment_polarity
## 9292                 -0.05                0.1                      -0.3
##      abs_title_subjectivity abs_title_sentiment_polarity shares
## 9292                    0.4                         0.3 843300
```

```
indts <- which(news$self_reference_max_shares==843300)
wrs <- news[indts,]
wrs
```

```
##       n_tokens_title n_tokens_content n_unique_tokens n_non_stop_unique_tokens
## 10181             11              486             0.6                       0.8
## 10217              9              721             0.5                       0.7
## 10256              9              854             0.5                       0.6
## 16235             11              716             0.5                       0.6
## 17978              9             1322             0.5                       0.6
## 18064             11              924             0.6                       0.7
## 18178              8              593             0.6                       0.7
##       num_hrefs num_self_hrefs num_imgs num_videos average_token_length
## 10181        17              3        0          0                    5
## 10217        28              9       12          0                    5
## 10256        36              6       15          0                    5
## 16235        29              4       15          1                    5
## 17978       143             51        0         91                    5
## 18064        89             41        0         42                    5
## 18178        83             41        0         42                    5
##       num_keywords data_channel_is_lifestyle data_channel_is_entertainment
## 10181            7                         0                             0
## 10217            9                         0                             0
## 10256            7                         0                             0
## 16235            9                         0                             0
## 17978            7                         0                             0
## 18064            7                         0                             0
## 18178            6                         0                             0
##       data_channel_is_bus data_channel_is_socmed data_channel_is_tech
## 10181                   0                     0                    0
## 10217                   0                     0                    0
## 10256                   0                     0                    0
## 16235                   0                     0                    0
## 17978                   0                     0                    0
```

```
## 18064                   0                     0                   0
## 18178                   0                     0                   0
##       data_channel_is_world kw_max_min kw_min_max kw_max_max kw_avg_max
## 10181                     0        586          0     843300     269186
## 10217                     0        733          0     843300     236089
## 10256                     0        205          0     843300     278743
## 16235                     0        826       1200     843300     475878
## 17978                     0       2100       4800     843300     591286
## 18064                     0        932          0     843300     570290
## 18178                     0        932      19500     843300     576333
##       kw_min_avg kw_max_avg kw_avg_avg self_reference_min_shares
## 10181          0       7459       3628                   843300
## 10217          0       5558       3481                     2900
## 10256          0       5259       3248                     2100
## 16235       1041      65065      11063                     2100
## 17978       3450       8201       5421                        2
## 18064          0       6861       3629                     1500
## 18178       3513      10216       6683                     1500
##       self_reference_max_shares self_reference_avg_sharess weekday_is_monday
## 10181                    843300                    843300                 0
## 10217                    843300                    423100                 1
## 10256                    843300                    295133                 1
## 16235                    843300                    295133                 0
## 17978                    843300                     83223                 0
## 18064                    843300                     80505                 0
## 18178                    843300                     80505                 0
##       weekday_is_tuesday weekday_is_wednesday weekday_is_thursday
## 10181                  0                    0                   0
## 10217                  0                    0                   0
## 10256                  0                    0                   0
## 16235                  1                    0                   0
## 17978                  0                    0                   0
## 18064                  1                    0                   0
## 18178                  0                    0                   0
##       weekday_is_friday weekday_is_saturday weekday_is_sunday is_weekend LDA_00
## 10181                 0                   0                 1          1   0.03
## 10217                 0                   0                 0          0   0.02
## 10256                 0                   0                 0          0   0.03
## 16235                 0                   0                 0          0   0.02
## 17978                 1                   0                 0          0   0.03
## 18064                 0                   0                 0          0   0.03
## 18178                 1                   0                 0          0   0.03
##       LDA_01 LDA_02 LDA_03 LDA_04 global_subjectivity global_sentiment_polarity
## 10181   0.03   0.17    0.7   0.03                 0.6                    -0.006
## 10217   0.02   0.02    0.6   0.30                 0.6                     0.265
## 10256   0.03   0.03    0.9   0.03                 0.5                     0.220
## 16235   0.02   0.02    0.7   0.25                 0.5                     0.216
## 17978   0.03   0.03    0.9   0.03                 0.5                     0.013
## 18064   0.03   0.17    0.7   0.03                 0.5                     0.008
## 18178   0.03   0.03    0.9   0.03                 0.5                    -0.091
##       global_rate_positive_words global_rate_negative_words rate_positive_words
## 10181                       0.04                      0.03                 0.6
## 10217                       0.05                      0.01                 0.8
## 10256                       0.06                      0.01                 0.8
```

```
## 16235                      0.07                      0.02                0.8
## 17978                      0.03                      0.02                0.5
## 18064                      0.03                      0.02                0.6
## 18178                      0.03                      0.03                0.4
##       rate_negative_words avg_positive_polarity min_positive_polarity
## 10181                 0.4                   0.4                 0.05
## 10217                 0.2                   0.5                 0.10
## 10256                 0.2                   0.4                 0.05
## 16235                 0.2                   0.4                 0.05
## 17978                 0.5                   0.4                 0.10
## 18064                 0.4                   0.4                 0.10
## 18178                 0.6                   0.4                 0.10
##       max_positive_polarity avg_negative_polarity min_negative_polarity
## 10181                   1.0                  -0.4                 -1.0
## 10217                   1.0                  -0.2                 -0.6
## 10256                   1.0                  -0.2                 -0.4
## 16235                   1.0                  -0.2                 -0.5
## 17978                   0.8                  -0.4                 -0.8
## 18064                   0.8                  -0.4                 -0.7
## 18178                   0.8                  -0.4                 -1.0
##       max_negative_polarity title_subjectivity title_sentiment_polarity
## 10181                 -0.10                0.0                     0.0
## 10217                 -0.17                0.0                     0.0
## 10256                 -0.05                0.5                     0.4
## 16235                 -0.05                0.7                     0.3
## 17978                 -0.10                0.5                     0.5
## 18064                 -0.05                0.0                     0.0
## 18178                 -0.16                0.0                     0.0
##       abs_title_subjectivity abs_title_sentiment_polarity shares
## 10181                    0.5                         0.0    728
## 10217                    0.5                         0.0   8000
## 10256                    0.0                         0.4   1200
## 16235                    0.2                         0.3   1100
## 17978                    0.0                         0.5  13600
## 18064                    0.5                         0.0   3900
## 18178                    0.5                         0.0   3800
```

```
indtkwm <- which(news$kw_max_max==843300)
wrkwm <- news[indtkwm,]
nrow(wrkwm)/nrow(news)
```

```
## [1] 0.8
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

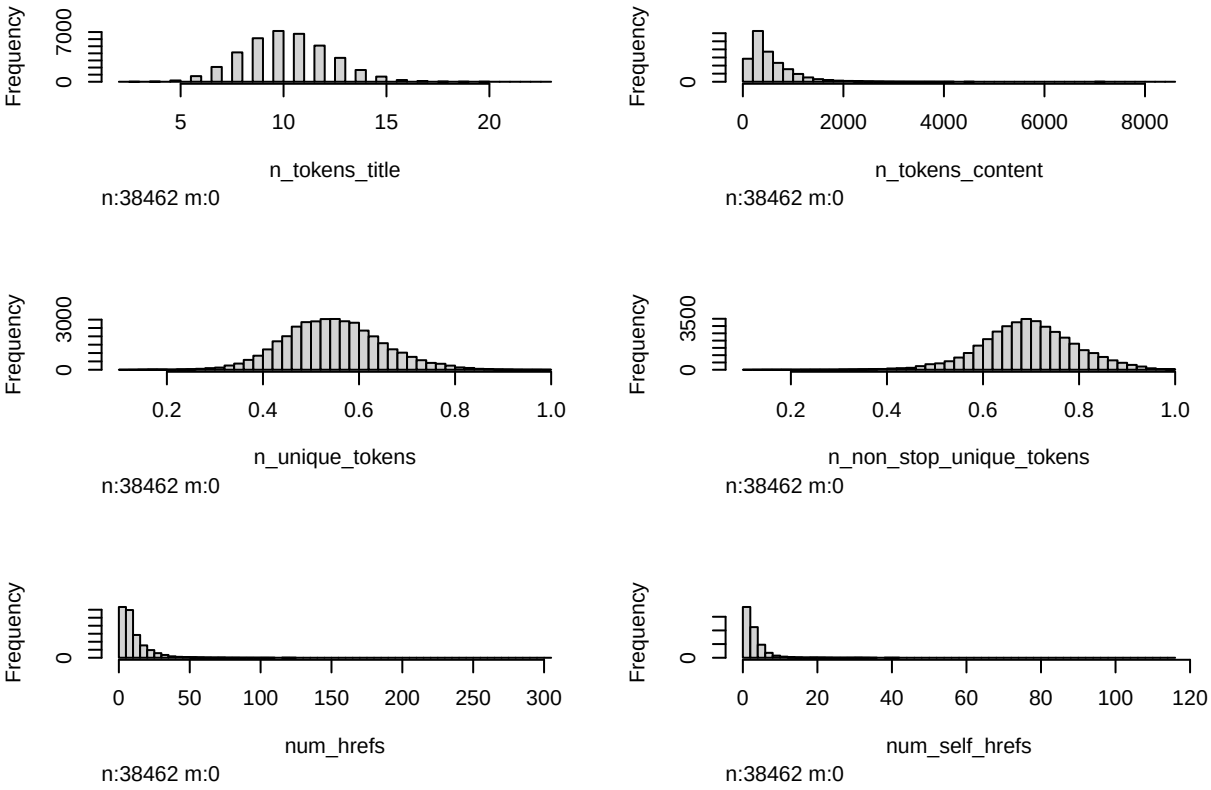
```
## Loading required package: Formula
```

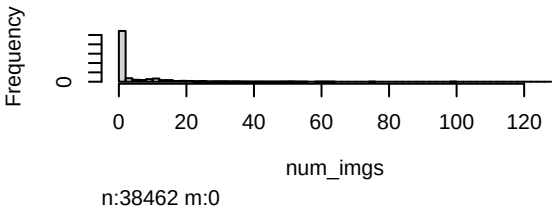
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

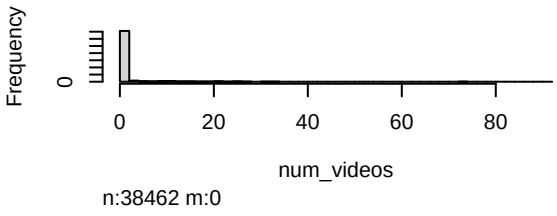
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
par(mfrow = c(3, 2))
hist.data.frame(news) # obviously left or right screwed, a lot of transformations would be needed to bo
```

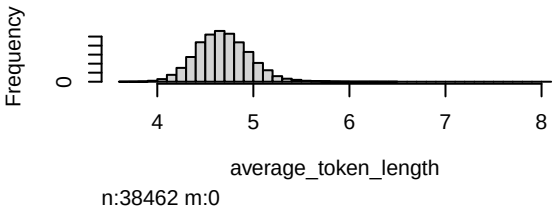




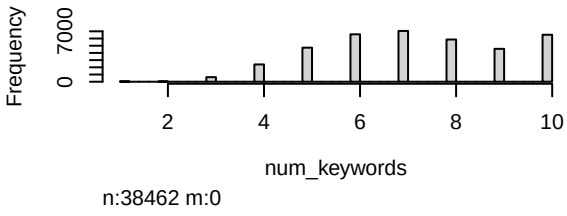

n:38462 m:0



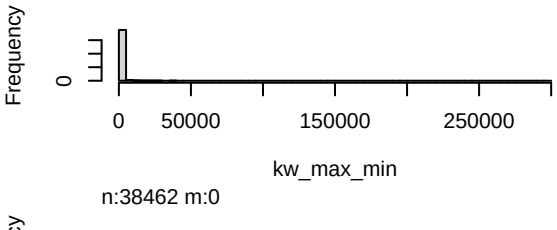

n:38462 m:0



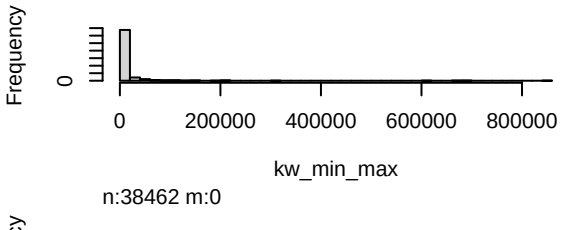

n:38462 m:0



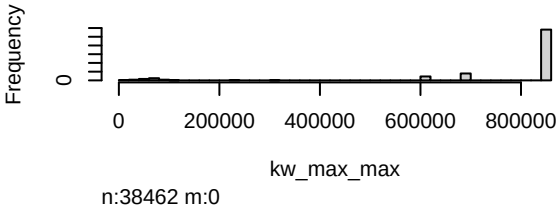

n:38462 m:0



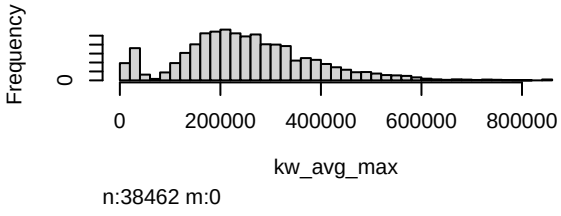

n:38462 m:0



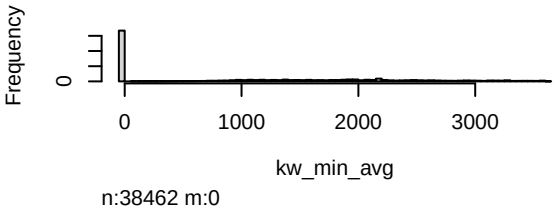

n:38462 m:0



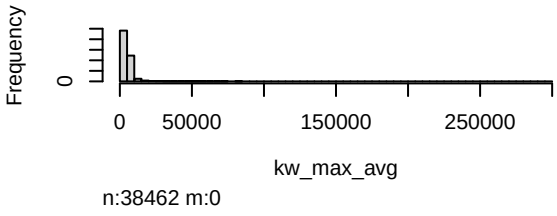

n:38462 m:0



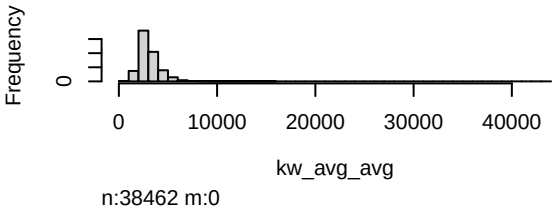

n:38462 m:0



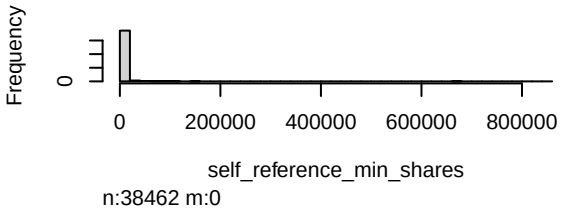

n:38462 m:0



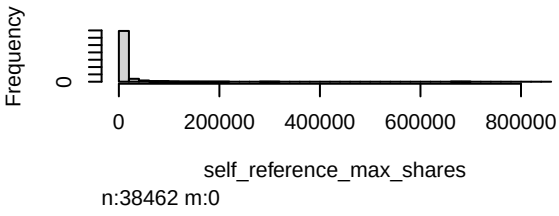

n:38462 m:0



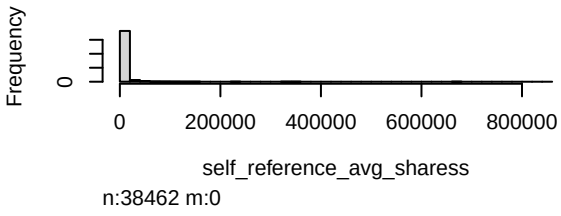

n:38462 m:0



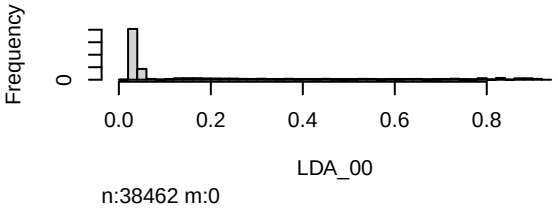

n:38462 m:0



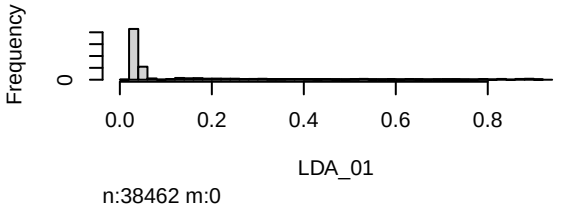
n:38462 m:0



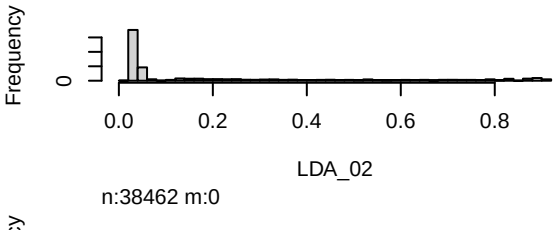
n:38462 m:0



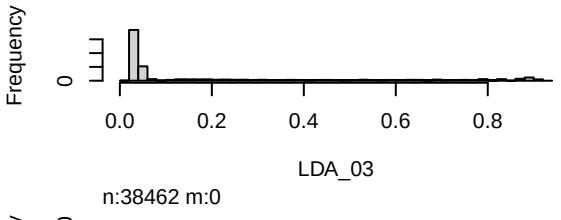
n:38462 m:0



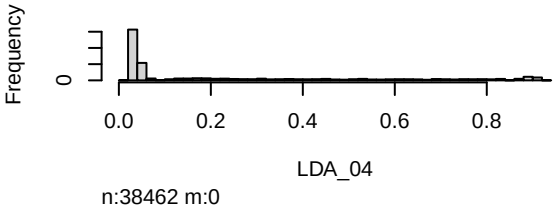
n:38462 m:0



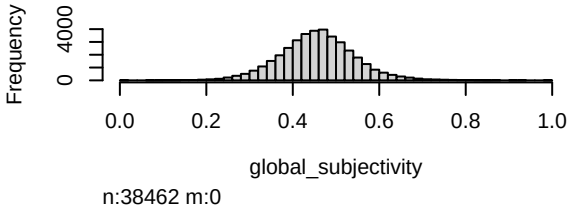
n:38462 m:0



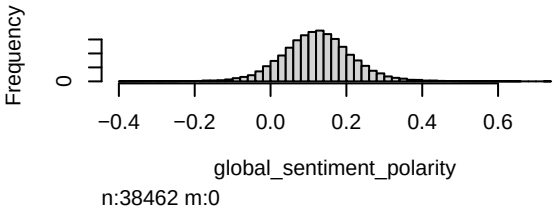
n:38462 m:0



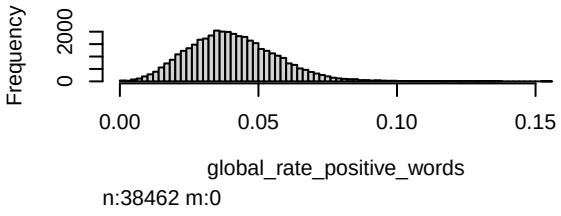
n:38462 m:0



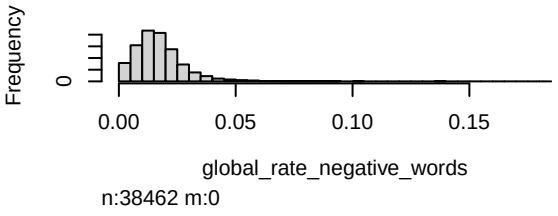
n:38462 m:0



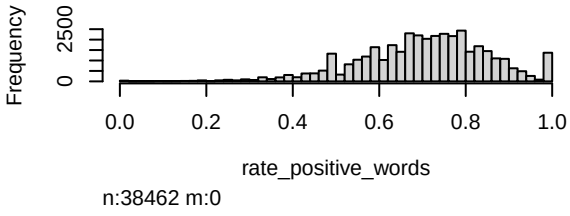
n:38462 m:0



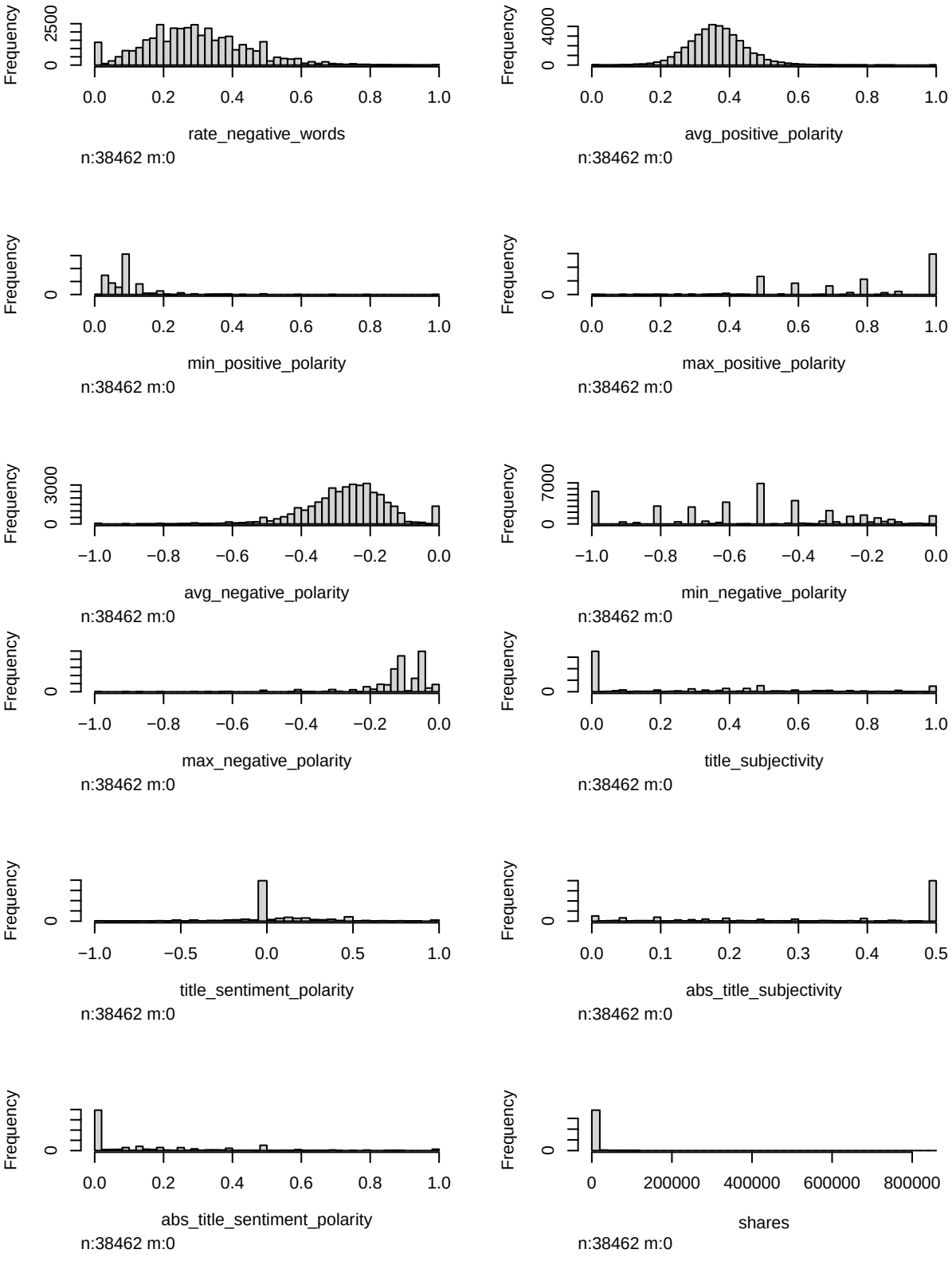
n:38462 m:0



n:38462 m:0



n:38462 m:0

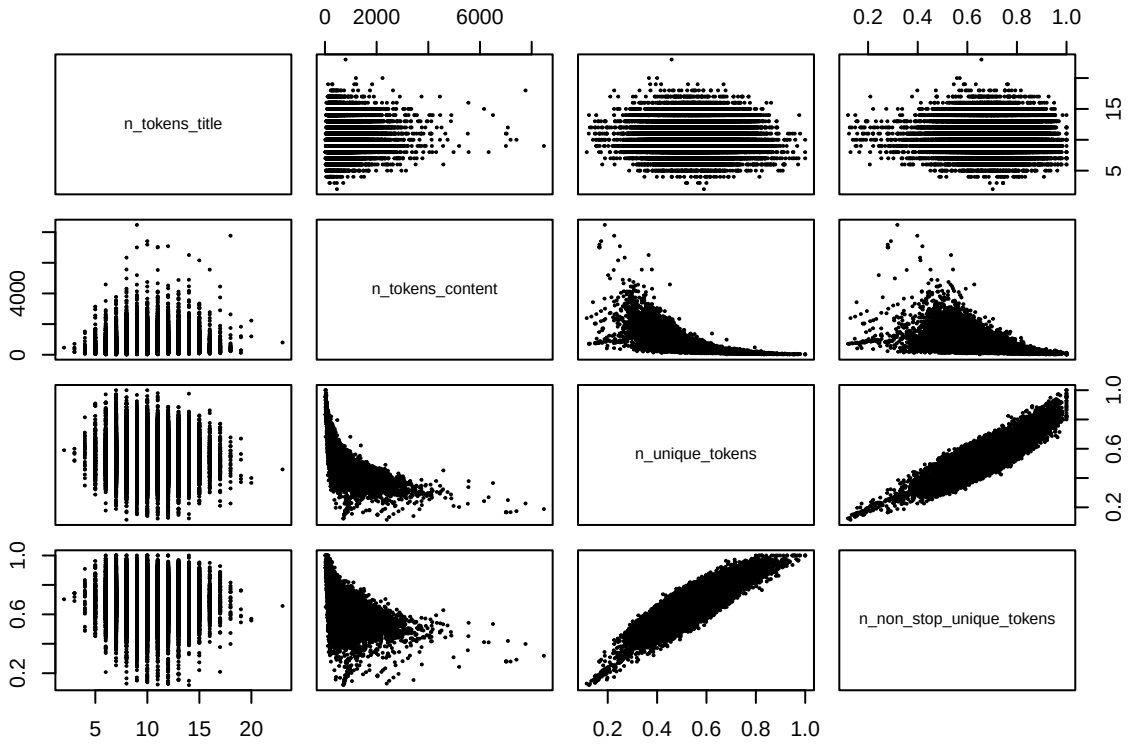

Frequency
0
2500
0.0
0.2
0.4
0.6
0.8
1.0
rate_negative_words
n:38462 m:0
Frequency
0
4000
0.0
0.2
0.4
0.6
0.8
1.0
avg_positive_polarity
n:38462 m:0
Frequency
0
0.0
0.2
0.4
0.6
0.8
1.0
min_positive_polarity
n:38462 m:0
Frequency
0
0.0
0.2
0.4
0.6
0.8
1.0
max_positive_polarity
n:38462 m:0
Frequency
0
3000
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
avg_negative_polarity
n:38462 m:0
Frequency
0
7000
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
min_negative_polarity
n:38462 m:0
Frequency
0
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
max_negative_polarity
n:38462 m:0
Frequency
0
0.0
0.2
0.4
0.6
0.8
1.0
title_subjectivity
n:38462 m:0
Frequency
0
−1.0
−0.5
0.0
0.5
1.0
title_sentiment_polarity
n:38462 m:0
Frequency
0
0.0
0.1
0.2
0.3
0.4
0.5
abs_title_subjectivity
n:38462 m:0
Frequency
0
0.0
0.2
0.4
0.6
0.8
1.0
abs_title_sentiment_polarity
n:38462 m:0
Frequency
0
0
200000
400000
600000
800000
shares
n:38462 m:0

### data relationship exploration

```
# since they are heterogeneous set of features, we will explore the relationship among the predictor va
names(news)
```

```
##  [1] "n_tokens_title"                "n_tokens_content"
##  [3] "n_unique_tokens"               "n_non_stop_unique_tokens"
##  [5] "num_hrefs"                     "num_self_hrefs"
##  [7] "num_imgs"                      "num_videos"
##  [9] "average_token_length"          "num_keywords"
## [11] "data_channel_is_lifestyle"     "data_channel_is_entertainment"
## [13] "data_channel_is_bus"           "data_channel_is_socmed"
## [15] "data_channel_is_tech"          "data_channel_is_world"
## [17] "kw_max_min"                    "kw_min_max"
## [19] "kw_max_max"                    "kw_avg_max"
## [21] "kw_min_avg"                    "kw_max_avg"
## [23] "kw_avg_avg"                    "self_reference_min_shares"
## [25] "self_reference_max_shares"     "self_reference_avg_sharess"
## [27] "weekday_is_monday"             "weekday_is_tuesday"
## [29] "weekday_is_wednesday"          "weekday_is_thursday"
## [31] "weekday_is_friday"             "weekday_is_saturday"
## [33] "weekday_is_sunday"             "is_weekend"
## [35] "LDA_00"                        "LDA_01"
## [37] "LDA_02"                        "LDA_03"
## [39] "LDA_04"                        "global_subjectivity"
## [41] "global_sentiment_polarity"     "global_rate_positive_words"
## [43] "global_rate_negative_words"    "rate_positive_words"
## [45] "rate_negative_words"           "avg_positive_polarity"
## [47] "min_positive_polarity"         "max_positive_polarity"
## [49] "avg_negative_polarity"         "min_negative_polarity"
## [51] "max_negative_polarity"         "title_subjectivity"
## [53] "title_sentiment_polarity"      "abs_title_subjectivity"
## [55] "abs_title_sentiment_polarity"  "shares"
```

```
# Word features
pairs(news[1:4],cex=0.2)
```



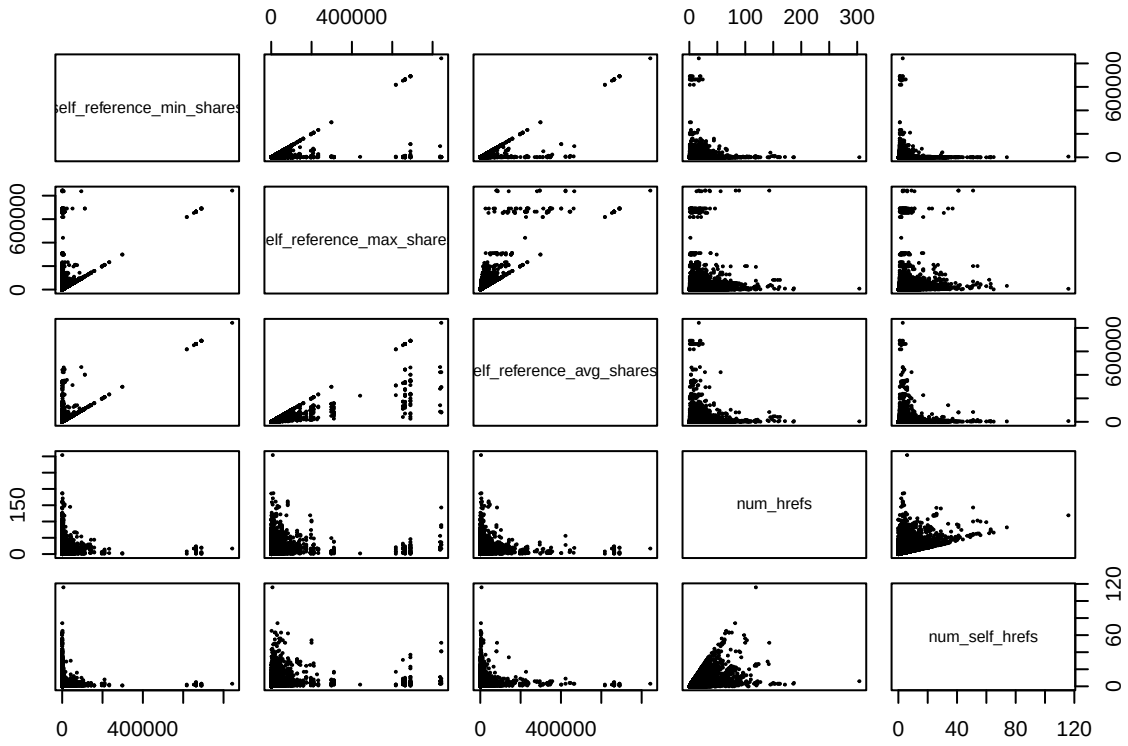

```
cor(news[3:4])
```

```
##                            n_unique_tokens n_non_stop_unique_tokens
## n_unique_tokens                        1.0                     0.9
## n_non_stop_unique_tokens               0.9                     1.0
```

```
cor(news$n_unique_tokens , news$n_non_stop_unique_tokens) # 0.885152 we will choose n_non_stop_unique_t
```

```
## [1] 0.9
```

```
# Links & References features
#"self_reference_min_shares",self_reference_max_shares" and "self_reference_avg_sharess", "num_hrefs","
pairs(~self_reference_min_shares+self_reference_max_shares+self_reference_avg_sharess+num_hrefs+num_seli
```



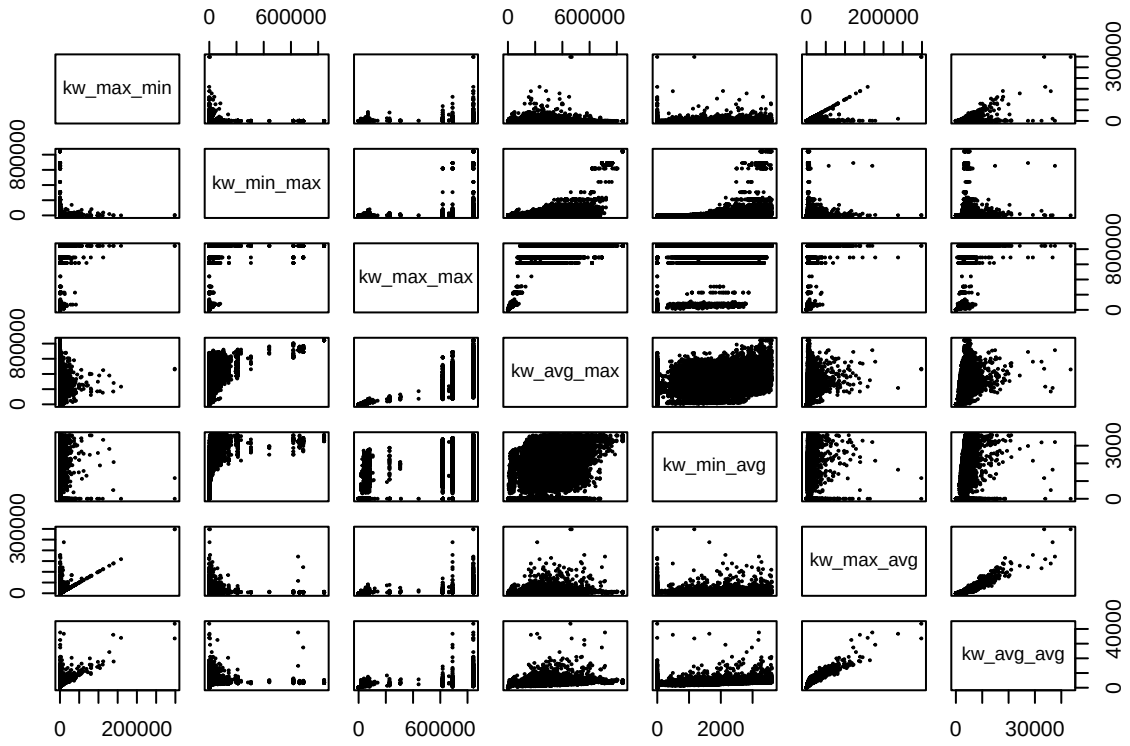

```
cor(news[24:26]) #  in this case we will keep "self_reference_avg_sharess" and drop "self_reference_min
```

```
##                            self_reference_min_shares self_reference_max_shares
## self_reference_min_shares                        1.0                       0.5
## self_reference_max_shares                        0.5                       1.0
## self_reference_avg_sharess                       0.8                       0.9
##                            self_reference_avg_sharess
## self_reference_min_shares                         0.8
## self_reference_max_shares                         0.9
## self_reference_avg_sharess                        1.0
```

```
cor(news[5:6]) # we will keep these 2 variables
```

```
##                num_hrefs num_self_hrefs
## num_hrefs            1.0            0.4
## num_self_hrefs       0.4            1.0
```

```
# Keyword features
pairs(news[17:23],cex=0.2) # for variables kw_max_min,kw_min_max,kw_max_max,kw_avg_max,kw_min_avg, kw_m
```

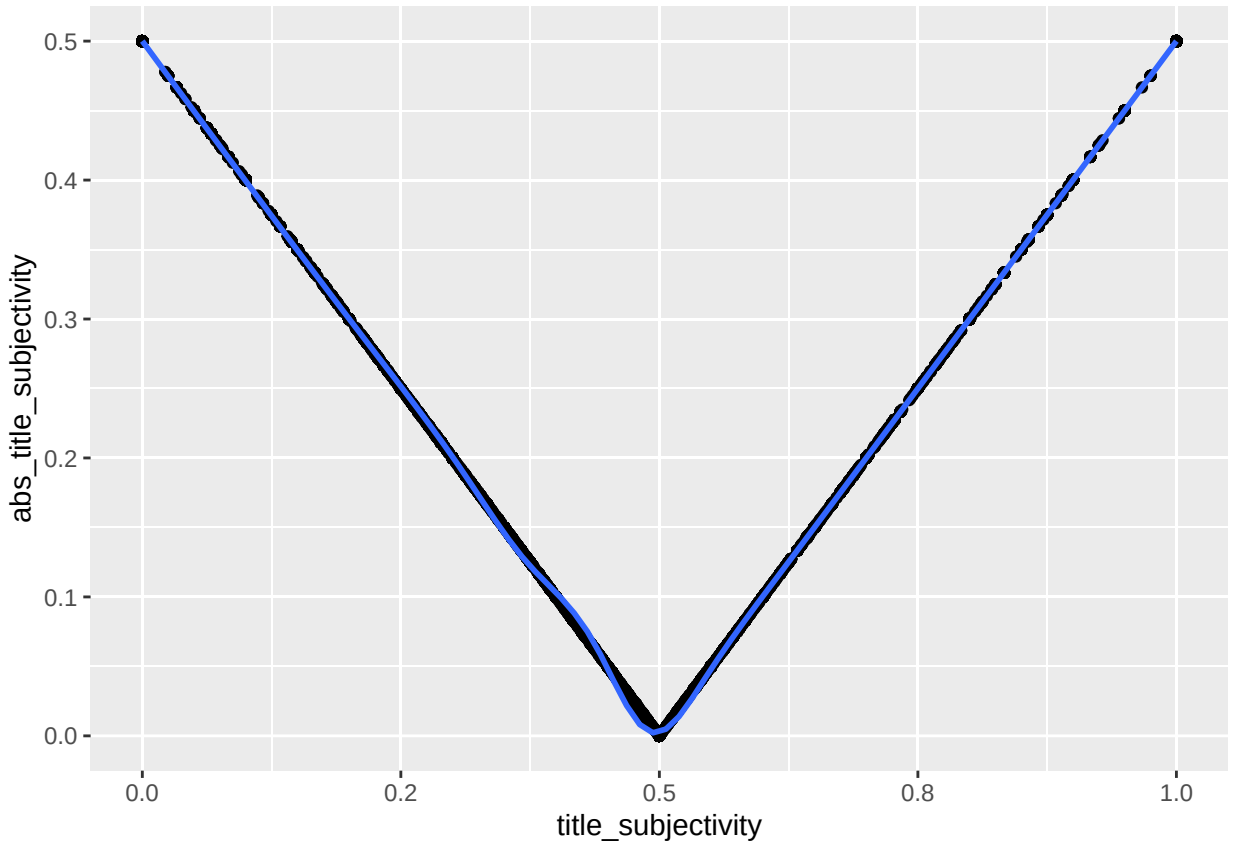



```
cor(news[17:23]) #we find kw_max_avg and kw_avg_avg are high correlated (0.8164474),we will choose kw_a
```

```
##             kw_max_min kw_min_max kw_max_max kw_avg_max kw_min_avg kw_max_avg
## kw_max_min       1.000      -0.04     -0.005      -0.04       0.006       0.59
## kw_min_max      -0.036       1.00      0.065       0.41       0.350       0.03
## kw_max_max      -0.005       0.06      1.000       0.57       0.157       0.09
## kw_avg_max      -0.036       0.41      0.571       1.00       0.396       0.13
## kw_min_avg       0.006       0.35      0.157       0.40       1.000       0.09
## kw_max_avg       0.594       0.03      0.091       0.13       0.089       1.00
## kw_avg_avg       0.418       0.16      0.223       0.40       0.443       0.82
##             kw_avg_avg
## kw_max_min         0.4
## kw_min_max         0.2
## kw_max_max         0.2
## kw_avg_max         0.4
## kw_min_avg         0.4
## kw_max_avg         0.8
## kw_avg_avg         1.0
```

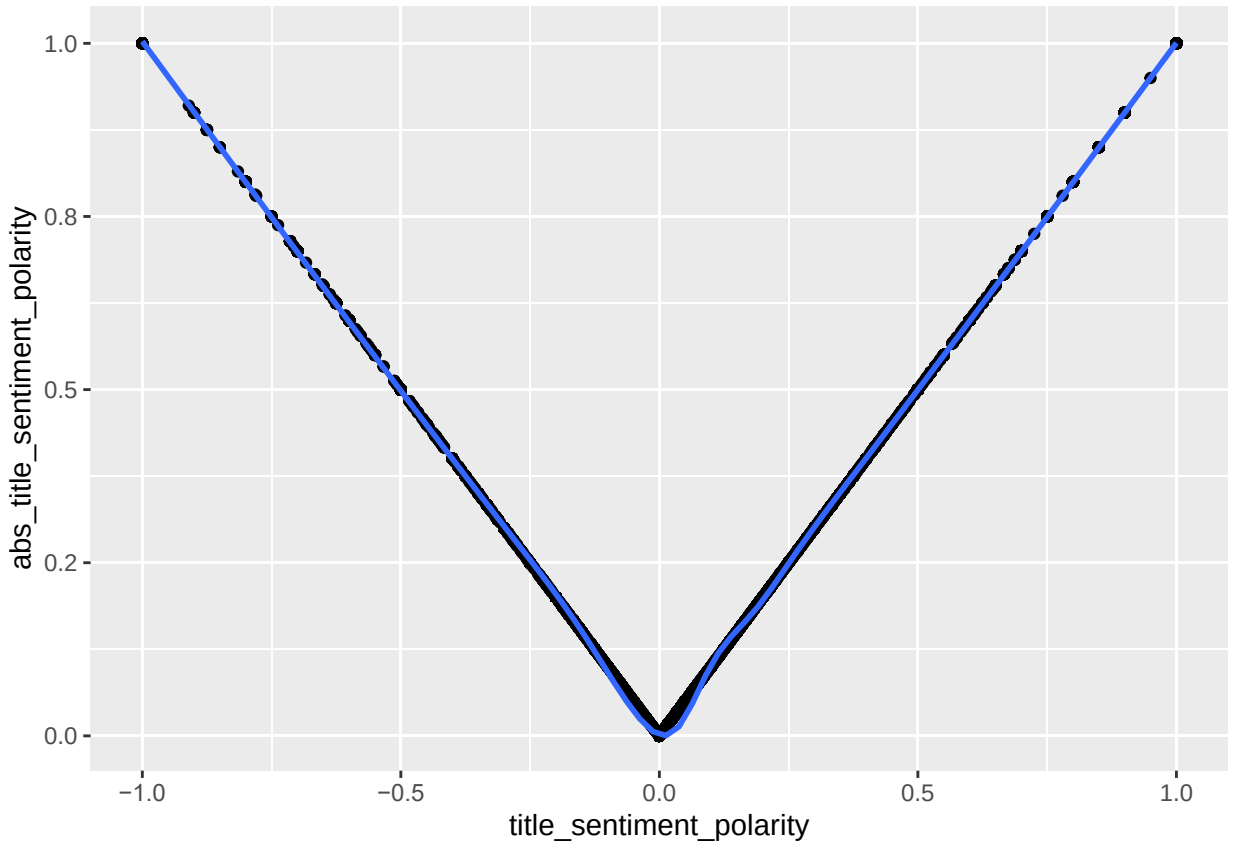
```
# NLP features

ggplot(aes(title_subjectivity,abs_title_subjectivity) ,data=news)+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(aes(title_sentiment_polarity,abs_title_sentiment_polarity) ,data=news)+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

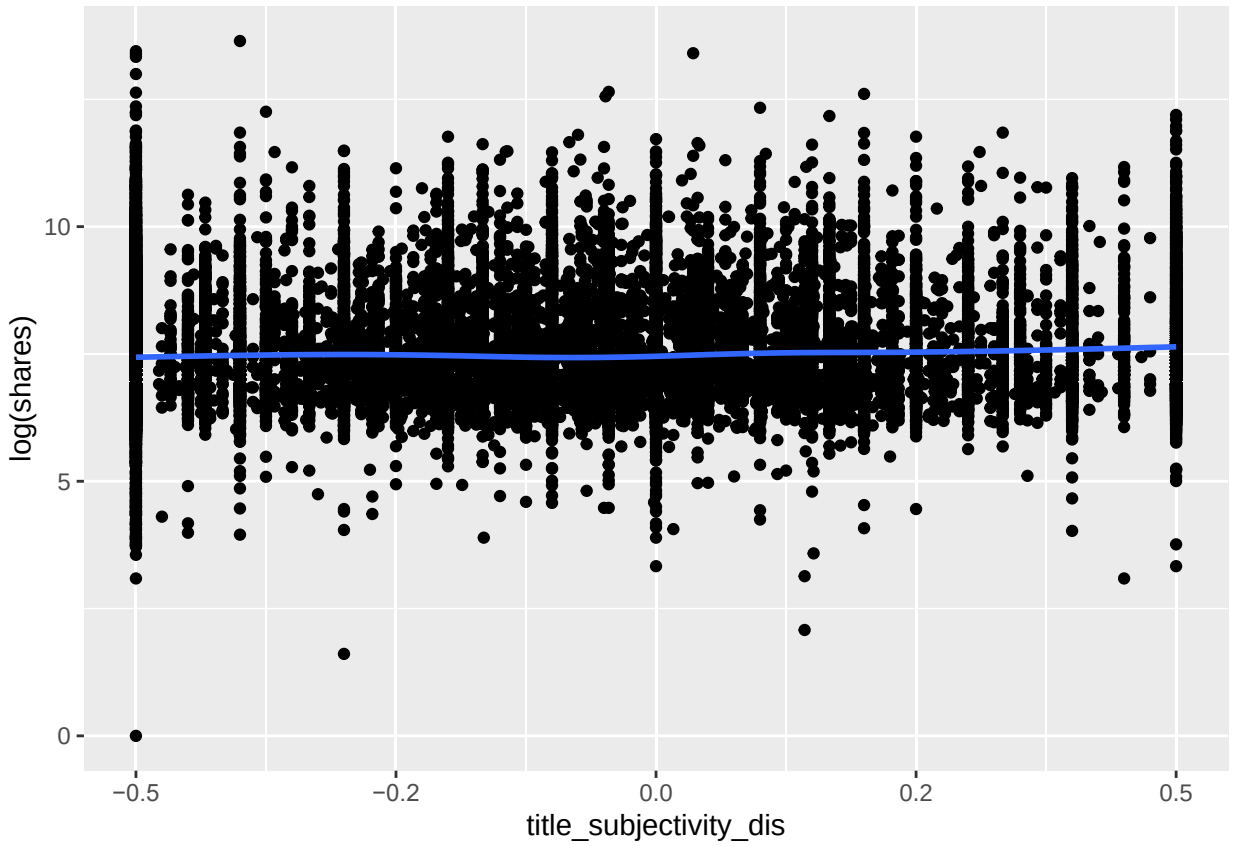



```
sum(news$title_subjectivity < 0.5)/nrow(news) # title_subjectivity is Article text subjectivity score
```

```
## [1] 0.7
```

```
# abs_title_subjectivity is title_subjectivity absolute difference to 0.5

# in order not to loose information about the direction of this original value, we create a new variabl

news<- mutate(news, title_subjectivity_dis= ifelse( title_subjectivity>=0.5, abs_title_subjectivity, -al


# in this step we will do same process to abs_title_sentiment_polarity and title_sentiment_polarity
# abs_title_sentiment_polarity is title_sentiment_polarity absolute difference to 0(not 0.5 ), the defi
news<- mutate(news, title_sentiment_polarity_dis= ifelse( title_sentiment_polarity>=0, abs_title_sentime

# we will drop abs_title_sentiment_polarity and title_sentiment_polarity, abs_title_subjectivity and ti

ggplot(aes(title_subjectivity_dis,log(shares)) ,data=news)+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(aes(title_sentiment_polarity_dis,log(shares)) ,data=news)+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```