EDA

Yunting Gu, Shuangyu Zhao, Wenting Liu

2023-03-16

```
# library install
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## 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 dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
## Loading required package: RSQLite
```

```
library(ggplot2)
library(ggcorrplot)
```

Our project focus on two questions: 1. for the word with ˚ train the model to fine whether the speaker misses this pronounciation 2. according to the pronounciation, determine whether this words has ˚ or ‿

```
Rclass <- readRDS("/Users/apple/Desktop/STT811_appl_stat_model/pro/data/RClassifierData_03July2019.Rds")
Tclass <- readRDS("/Users/apple/Desktop/STT811_appl_stat_model/pro/data/TClassifierData_14Nov2019.Rds")
```

see the size of 2 dataframe

```
dim(Rclass)
```

```
## [1] 40614   217
```

```
dim(Tclass)
```

```
## [1] 9888  137
```

1. EDA for question 1

```
dim(Rclass)
```

```
## [1] 40614   217
```

```
sqldf("SELECT DISTINCT Rpresent
       FROM Rclass")
```

```
##   Rpresent
## 1     <NA>
## 2   Absent
## 3  Present
```

```
sqldf("SELECT COUNT(*)
      FROM Rclass
      WHERE Rpresent = 'Absent' ")
```

```
##   COUNT(*)
## 1     4255
```

see the number of data have been coded 'present'(means pronoucing the ˚) and 'absent' (means missing the ˚)

```
sqldf("SELECT COUNT(*)
      FROM Rclass
      WHERE Rpresent = 'Present' ")
```

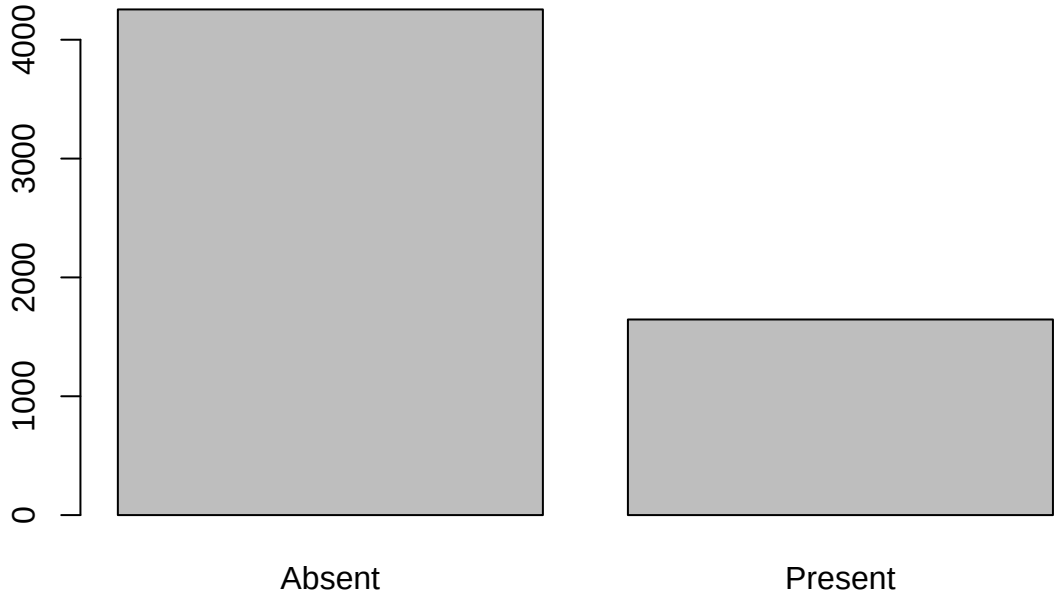
```
##   COUNT(*)
## 1     1646
```

```
sqldf("SELECT COUNT(*)
      FROM Rclass
      WHERE Rpresent = 'Absent' ")
```

```
##   COUNT(*)
## 1     4255
```

```
rdf_pre_ab <- sqldf("SELECT *
                    FROM Rclass
                    WHERE Rpresent = 'Absent' OR Rpresent = 'Present'")
```

```
barplot(table(rdf_pre_ab$Rpresent))
```



we can see that the data are unbalanced.

extract the columns I think useful

```
rdf_useful <- rdf_pre_ab[34:217]
rdf_useful$Rpresent <- rdf_pre_ab$Rpresent
rdf_useful <- na.omit(rdf_useful)
dim(rdf_useful)
```

```
## [1] 5273  185
```

encode absent - 0; present - 1

```
rdf_useful$Rpresent_encode <- ifelse(rdf_useful$Rpresent == "Present", 1, 0)
```

standardize the data

```
# Standardize the data
standardized_rdata <- scale(rdf_useful[, c(1:184)], center = TRUE, scale = TRUE)
rdf_useful[, c(1:184)] <- standardized_rdata
```

the correlation between elements

```
cor_rdf_usefurl <- round(cor(rdf_useful[, c(1:184, 186)]), 2)
write_csv(as.data.frame(cor_rdf_usefurl), file = "rdf_correlation.csv")
```

print all the variables name, whose correlations are higher than 0.9

```
high_corr <- data.frame(var1 = character() , var2 = character(), cor = numeric())
z = 1
for (i in 1:185) {
  for (j in 1:i) {
    if (abs(cor_rdf_usefurl[i, j]) > 0.9 && i != j) {
      high_corr[z, 1] <- colnames(cor_rdf_usefurl)[j]
      high_corr[z, 2] <- colnames(cor_rdf_usefurl)[i]
      high_corr[z, 3] <- cor_rdf_usefurl[i, j]
      z = z +1
    }
  }
}
high_corr
```

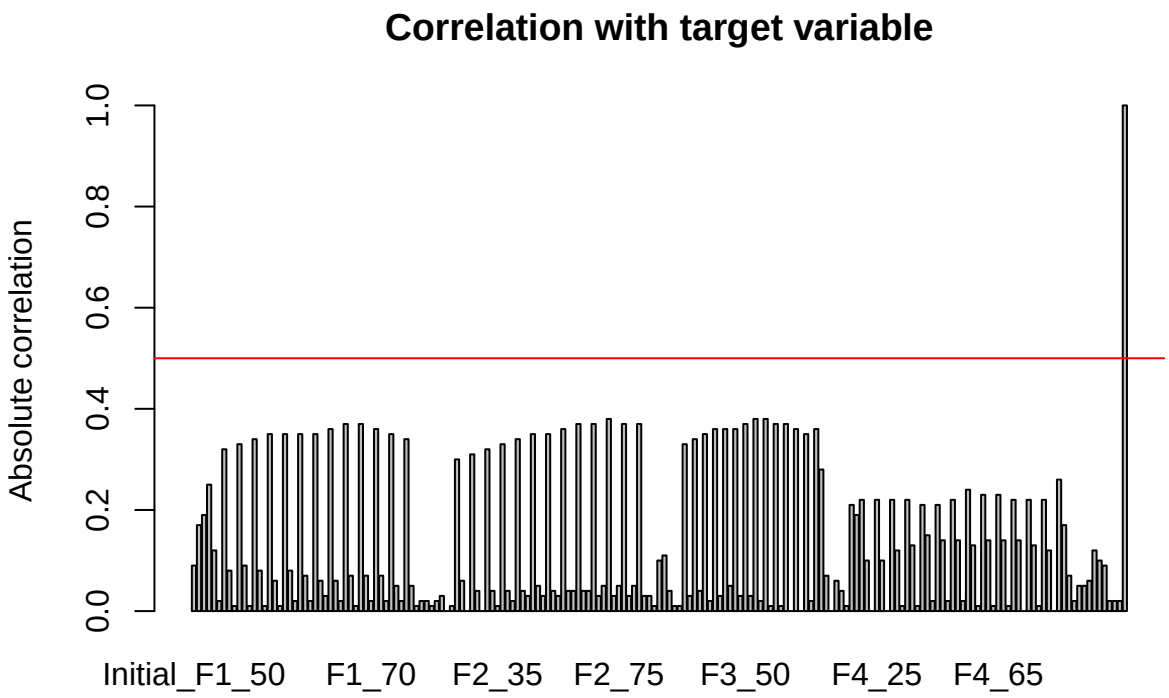
```
##            var1         var2  cor
## 1         F1_20        F1_25 0.92
## 2   diffF3F1_20  diffF3F1_25 0.93
## 3         F1_25        F1_30 0.93
## 4   diffF3F1_25  diffF3F1_30 0.94
## 5   diffF3F1_30  diffF3F1_35 0.94
## 6         F1_35        F1_40 0.93
## 7   diffF3F1_35  diffF3F1_40 0.94
## 8         F1_40        F1_45 0.95
## 9   diffF3F1_40  diffF3F1_45 0.94
## 10        F1_45        F1_50 0.93
## 11  diffF3F1_45  diffF3F1_50 0.92
## 12        F1_50        F1_55 0.93
## 13  diffF3F1_50  diffF3F1_55 0.92
## 14        F1_55        F1_60 0.92
## 15  diffF3F1_55  diffF3F1_60 0.93
## 16        F1_60        F1_65 0.92
## 17  diffF3F1_60  diffF3F1_65 0.93
## 18  diffF3F1_65  diffF3F1_70 0.93
## 19  diffF3F1_70  diffF3F1_75 0.92
## 20        F1_75        F1_80 0.91
## 21  diffF3F1_75  diffF3F1_80 0.92
## 22        F2_20        F2_25 0.92
## 23  diffF3F2_20  diffF3F2_25 0.94
## 24        F2_25        F2_30 0.94
## 25  diffF3F2_25  diffF3F2_30 0.95
## 26        F2_30        F2_35 0.93
## 27  diffF3F2_30  diffF3F2_35 0.95
## 28        F2_35        F2_40 0.94
## 29  diffF3F2_30  diffF3F2_40 0.91
## 30  diffF3F2_35  diffF3F2_40 0.96
## 31        F2_40        F2_45 0.94
## 32  diffF3F2_40  diffF3F2_45 0.95
## 33        F2_45        F2_50 0.94
## 34  diffF3F2_45  diffF3F2_50 0.94
## 35        F2_50        F2_55 0.93
## 36  diffF3F2_45  diffF3F2_55 0.91
## 37  diffF3F2_50  diffF3F2_55 0.94
```

```
## 38        F2_55        F2_60 0.94
## 39  diffF3F2_55  diffF3F2_60 0.95
## 40        F2_60        F2_65 0.93
## 41  diffF3F2_60  diffF3F2_65 0.95
## 42        F2_65        F2_70 0.94
## 43  diffF3F2_65  diffF3F2_70 0.95
## 44        F2_70        F2_75 0.92
## 45  diffF3F2_70  diffF3F2_75 0.94
## 46        F2_75        F2_80 0.93
## 47  diffF3F2_75  diffF3F2_80 0.93
## 48  diffF3F1_20        F3_20 0.94
## 49  diffF3F1_25        F3_25 0.94
## 50        F3_20        F3_25 0.92
## 51  diffF3F1_30        F3_30 0.94
## 52        F3_25        F3_30 0.93
## 53  diffF3F1_35        F3_35 0.94
## 54        F3_30        F3_35 0.93
## 55  diffF3F1_40        F3_40 0.94
## 56        F3_35        F3_40 0.93
## 57  diffF3F1_45        F3_45 0.94
## 58        F3_40        F3_45 0.93
## 59  diffF3F1_50        F3_50 0.94
## 60        F3_45        F3_50 0.91
## 61  diffF3F1_55        F3_55 0.94
## 62        F3_50        F3_55 0.92
## 63  diffF3F1_60        F3_60 0.94
## 64        F3_55        F3_60 0.92
## 65  diffF3F1_65        F3_65 0.94
## 66        F3_60        F3_65 0.91
## 67  diffF3F1_70        F3_70 0.93
## 68        F3_65        F3_70 0.93
## 69  diffF3F1_75        F3_75 0.93
## 70        F3_70        F3_75 0.91
## 71  diffF3F1_80        F3_80 0.92
## 72        F3_75        F3_80 0.91
## 73 intens_F3min intens_F3max 0.92
## 74        F4_35        F4_40 0.91
## 75        F4_40        F4_45 0.91
## 76        F0min        F0max 0.96
```

For the variable listed in this dataframe, we are gonna choose var1 and drop var2, when we feed data into classifer.

for the correlation between variables and target

```
barplot(abs(cor_rdf_usefurl[, 185]), main = "Correlation with target variable", ylab = "Absolute correla
abline(h = 0.5, col = "red")
```



we can see the correlation is not really high.

for the variable having highest correlation with target, we will observe the distribution of it.

```
cor_rdf_usefurl_df <- as.data.frame(cor_rdf_usefurl)
head(cor_rdf_usefurl_df)
```

```
##               Initial_F1_50 Initial_F2_50 Initial_F3_50 Initial_F4_50 TokenDur
## Initial_F1_50          1.00          0.30          0.25          0.47     0.02
## Initial_F2_50          0.30          1.00          0.82          0.67     0.06
## Initial_F3_50          0.25          0.82          1.00          0.74     0.04
## Initial_F4_50          0.47          0.67          0.74          1.00    -0.01
## TokenDur               0.02          0.06          0.04         -0.01     1.00
## F1_20                  0.29          0.14          0.12          0.20     0.17
##               F1_20 diffF3F1_20 BW1_20 F1_25 diffF3F1_25 BW1_25 F1_30
## Initial_F1_50  0.29        0.03   0.00  0.30        0.03  -0.01  0.30
## Initial_F2_50  0.14        0.22   0.09  0.14        0.22   0.07  0.13
## Initial_F3_50  0.12        0.33   0.09  0.12        0.32   0.08  0.12
## Initial_F4_50  0.20        0.23   0.10  0.20        0.23   0.08  0.20
## TokenDur       0.17       -0.07   0.02  0.16       -0.06   0.04  0.18
## F1_20          1.00       -0.22   0.13  0.92       -0.22   0.16  0.85
##               diffF3F1_30 BW1_30 F1_35 diffF3F1_35 BW1_35 F1_40 diffF3F1_40
## Initial_F1_50        0.04   0.00  0.31        0.05   0.00  0.32        0.05
## Initial_F2_50        0.23   0.07  0.13        0.24   0.05  0.13        0.25
## Initial_F3_50        0.33   0.08  0.11        0.33   0.07  0.11        0.34
## Initial_F4_50        0.23   0.09  0.20        0.24   0.08  0.20        0.24
```

```
## TokenDur            -0.07   0.03  0.22       -0.08   0.04  0.22       -0.09
## F1_20               -0.21   0.16  0.78       -0.19   0.19  0.73       -0.18
##               BW1_40 F1_45 diffF3F1_45 BW1_45 F1_50 diffF3F1_50 BW1_50 F1_55
## Initial_F1_50   0.01  0.31        0.06  -0.01  0.29        0.06   0.00  0.29
## Initial_F2_50   0.06  0.12        0.25   0.04  0.12        0.25   0.04  0.13
## Initial_F3_50   0.07  0.11        0.33   0.05  0.10        0.33   0.04  0.11
## Initial_F4_50   0.08  0.19        0.24   0.05  0.19        0.25   0.05  0.19
## TokenDur        0.05  0.23       -0.10   0.05  0.19       -0.11   0.06  0.22
## F1_20           0.16  0.71       -0.18   0.15  0.69       -0.17   0.13  0.68
##               diffF3F1_55 BW1_55 F1_60 diffF3F1_60 BW1_60 F1_65 diffF3F1_65
## Initial_F1_50        0.07   0.00  0.29        0.08   0.01  0.28        0.10
## Initial_F2_50        0.26   0.05  0.12        0.27   0.06  0.11        0.28
## Initial_F3_50        0.34   0.05  0.11        0.35   0.07  0.10        0.36
## Initial_F4_50        0.27   0.05  0.18        0.29   0.06  0.17        0.30
## TokenDur            -0.11   0.05  0.22       -0.11   0.12  0.21       -0.12
## F1_20               -0.16   0.13  0.64       -0.17   0.14  0.60       -0.15
##               BW1_65 F1_70 diffF3F1_70 BW1_70 F1_75 diffF3F1_75 BW1_75 F1_80
## Initial_F1_50   0.00  0.27        0.10   0.01  0.25        0.11   0.02  0.24
## Initial_F2_50   0.06  0.09        0.29   0.07  0.07        0.29   0.07  0.06
## Initial_F3_50   0.08  0.09        0.37   0.08  0.07        0.37   0.08  0.06
## Initial_F4_50   0.07  0.15        0.31   0.06  0.13        0.33   0.06  0.11
## TokenDur        0.06  0.22       -0.12   0.13  0.20       -0.12   0.08  0.20
## F1_20           0.14  0.57       -0.14   0.15  0.53       -0.13   0.15  0.50
##               diffF3F1_80 BW1_80 F1min F1max time_F1min time_F1max F1range
## Initial_F1_50        0.13   0.00  0.29  0.16       0.01      -0.01   -0.01
## Initial_F2_50        0.31   0.05  0.03  0.11       0.00      -0.02    0.08
## Initial_F3_50        0.39   0.07  0.01  0.11      -0.02       0.00    0.10
## Initial_F4_50        0.35   0.05  0.10  0.12       0.01      -0.03    0.06
## TokenDur            -0.12   0.09 -0.05  0.37      -0.04       0.13    0.37
## F1_20               -0.11   0.13  0.53  0.63       0.12      -0.17    0.29
##               time_F1range slopeF1 F2_20 diffF3F2_20 BW2_20 F2_25 diffF3F2_25
## Initial_F1_50         0.01   -0.02  0.02        0.11   0.09  0.02        0.11
## Initial_F2_50        -0.02    0.01  0.14        0.12   0.09  0.14        0.12
## Initial_F3_50        -0.04    0.03  0.09        0.26   0.13  0.09        0.25
## Initial_F4_50        -0.02    0.00  0.05        0.23   0.21  0.05        0.23
## TokenDur             -0.28    0.41  0.01       -0.02  -0.02  0.01       -0.02
## F1_20                -0.07    0.11  0.18       -0.04   0.02  0.14       -0.04
##               BW2_25 F2_30 diffF3F2_30 BW2_30 F2_35 diffF3F2_35 BW2_35 F2_40
## Initial_F1_50   0.08  0.02        0.11   0.09  0.03        0.11   0.08  0.03
## Initial_F2_50   0.10  0.15        0.11   0.11  0.15        0.13   0.11  0.15
## Initial_F3_50   0.15  0.09        0.25   0.15  0.08        0.26   0.15  0.09
## Initial_F4_50   0.21  0.05        0.22   0.21  0.05        0.23   0.21  0.05
## TokenDur       -0.04  0.01       -0.02  -0.04  0.02       -0.02  -0.02  0.01
## F1_20           0.02  0.12       -0.03   0.01  0.09       -0.02   0.00  0.08
##               diffF3F2_40 BW2_40 F2_45 diffF3F2_45 BW2_45 F2_50 diffF3F2_50
## Initial_F1_50        0.12   0.09  0.03        0.11   0.10  0.04        0.11
## Initial_F2_50        0.13   0.09  0.15        0.13   0.10  0.14        0.13
## Initial_F3_50        0.26   0.14  0.09        0.25   0.14  0.08        0.26
## Initial_F4_50        0.24   0.20  0.06        0.23   0.20  0.06        0.24
## TokenDur            -0.03  -0.04  0.02       -0.04  -0.04  0.00       -0.04
## F1_20               -0.02  -0.01  0.08       -0.02  -0.01  0.08       -0.01
##               BW2_50 F2_55 diffF3F2_55 BW2_55 F2_60 diffF3F2_60 BW2_60 F2_65
## Initial_F1_50   0.09  0.05        0.12   0.11  0.05        0.12   0.10  0.06
## Initial_F2_50   0.10  0.16        0.13   0.10  0.16        0.14   0.10  0.17
```

```
## Initial_F3_50   0.15  0.09        0.27   0.15  0.09        0.27   0.14  0.10
## Initial_F4_50   0.20  0.06        0.25   0.20  0.07        0.25   0.19  0.08
## TokenDur       -0.04  0.01       -0.04  -0.03  0.01       -0.04  -0.05  0.02
## F1_20          -0.01  0.09       -0.01  -0.01  0.09       -0.02  -0.02  0.08
##               diffF3F2_65 BW2_65 F2_70 diffF3F2_70 BW2_70 F2_75 diffF3F2_75
## Initial_F1_50        0.13   0.10  0.06        0.13   0.10  0.06        0.14
## Initial_F2_50        0.15   0.10  0.16        0.15   0.10  0.16        0.16
## Initial_F3_50        0.28   0.15  0.10        0.28   0.15  0.10        0.29
## Initial_F4_50        0.26   0.21  0.08        0.26   0.21  0.09        0.27
## TokenDur            -0.06  -0.05  0.01       -0.05  -0.07  0.00       -0.05
## F1_20               -0.01   0.00  0.08       -0.01   0.01  0.07        0.01
##               BW2_75 F2_80 diffF3F2_80 BW2_80 F2min F2max time_F2min time_F2max
## Initial_F1_50   0.08  0.06        0.15   0.08 -0.07  0.10       0.02       0.00
## Initial_F2_50   0.09  0.17        0.17   0.09  0.05  0.21       0.02       0.00
## Initial_F3_50   0.14  0.11        0.30   0.14 -0.03  0.19       0.01       0.00
## Initial_F4_50   0.19  0.10        0.29   0.19 -0.11  0.19       0.01      -0.01
## TokenDur       -0.05  0.01       -0.04  -0.03 -0.17  0.21       0.00       0.00
## F1_20           0.00  0.07        0.02   0.01  0.03  0.16       0.03      -0.04
##               F2range time_F2range slopeF2 F3_20 BW3_20 F3_25 BW3_25 F3_30
## Initial_F1_50    0.15        -0.06    0.09  0.14   0.14  0.14   0.13  0.14
## Initial_F2_50    0.15        -0.08    0.08  0.28   0.08  0.27   0.09  0.28
## Initial_F3_50    0.19        -0.08    0.10  0.37   0.12  0.37   0.12  0.37
## Initial_F4_50    0.26        -0.10    0.13  0.31   0.16  0.30   0.14  0.30
## TokenDur         0.33        -0.20    0.38 -0.01  -0.02 -0.01  -0.02 -0.01
## F1_20            0.11        -0.08    0.08  0.14   0.04  0.10   0.03  0.08
##               BW3_30 F3_35 BW3_35 F3_40 BW3_40 F3_45 BW3_45 F3_50 BW3_50 F3_55
## Initial_F1_50   0.13  0.15   0.12  0.16   0.13  0.16   0.13  0.16   0.13  0.17
## Initial_F2_50   0.09  0.28   0.08  0.30   0.10  0.29   0.08  0.29   0.08  0.30
## Initial_F3_50   0.11  0.38   0.11  0.38   0.13  0.37   0.12  0.37   0.11  0.38
## Initial_F4_50   0.15  0.31   0.14  0.32   0.16  0.31   0.15  0.32   0.14  0.34
## TokenDur       -0.01 -0.01  -0.02 -0.02  -0.01 -0.02   0.00 -0.05   0.00 -0.03
## F1_20           0.02  0.07   0.01  0.06   0.03  0.06   0.03  0.07   0.02  0.07
##               BW3_55 F3_60 BW3_60 F3_65 BW3_65 F3_70 BW3_70 F3_75 BW3_75 F3_80
## Initial_F1_50   0.15  0.18   0.16  0.20   0.15  0.20   0.15  0.21   0.15  0.22
## Initial_F2_50   0.10  0.31   0.10  0.32   0.10  0.32   0.09  0.32   0.09  0.34
## Initial_F3_50   0.11  0.39   0.12  0.40   0.12  0.41   0.12  0.40   0.13  0.41
## Initial_F4_50   0.15  0.35   0.16  0.36   0.16  0.37   0.17  0.38   0.19  0.40
## TokenDur       -0.01 -0.04   0.00 -0.04   0.03 -0.04   0.02 -0.05   0.01 -0.04
## F1_20           0.03  0.06   0.04  0.06   0.03  0.07   0.03  0.07   0.02  0.08
##               BW3_80 F3min F3max time_F3min time_F3max F3range time_F3range
## Initial_F1_50   0.16  0.05  0.25      -0.03       0.05    0.18        -0.08
## Initial_F2_50   0.09  0.18  0.35      -0.02       0.04    0.16        -0.08
## Initial_F3_50   0.13  0.25  0.45      -0.03       0.03    0.19        -0.08
## Initial_F4_50   0.18  0.19  0.43      -0.05       0.06    0.23        -0.08
## TokenDur        0.02 -0.24  0.19       0.06      -0.02    0.38        -0.23
## F1_20           0.05 -0.01  0.17       0.03      -0.03    0.16        -0.08
##               slopeF3 intens_F3min intens_F3max F4_20 diffF4F3_20 BW4_20 F4_25
## Initial_F1_50    0.11        -0.10        -0.12  0.17        0.05   0.14  0.17
## Initial_F2_50    0.09        -0.10        -0.13  0.30        0.06   0.10  0.30
## Initial_F3_50    0.10        -0.14        -0.16  0.35        0.01   0.05  0.35
## Initial_F4_50    0.11        -0.15        -0.18  0.48        0.23   0.13  0.48
## TokenDur         0.44         0.02        -0.03 -0.06       -0.06  -0.01 -0.05
## F1_20            0.12         0.02        -0.02  0.22        0.11   0.03  0.18
##               diffF4F3_25 BW4_25 F4_30 diffF4F3_30 BW4_30 F4_35 diffF4F3_35
```

```
## Initial_F1_50        0.06   0.15  0.18        0.05   0.15  0.18        0.05
## Initial_F2_50        0.06   0.10  0.30        0.05   0.10  0.30        0.04
## Initial_F3_50        0.01   0.06  0.34        0.01   0.06  0.34        0.00
## Initial_F4_50        0.24   0.14  0.47        0.23   0.14  0.47        0.22
## TokenDur            -0.06  -0.01 -0.06       -0.05   0.00 -0.04       -0.04
## F1_20                0.10   0.04  0.15        0.09   0.04  0.13        0.08
##                BW4_35 F4_40 diffF4F3_40 BW4_40 F4_45 diffF4F3_45 BW4_45 F4_50
## Initial_F1_50    0.15  0.18        0.04   0.15  0.18        0.04   0.15  0.18
## Initial_F2_50    0.11  0.30        0.03   0.11  0.30        0.04   0.12  0.29
## Initial_F3_50    0.06  0.34       -0.01   0.06  0.33       -0.01   0.05  0.32
## Initial_F4_50    0.14  0.48        0.22   0.15  0.47        0.22   0.13  0.46
## TokenDur         0.00 -0.05       -0.04  -0.01 -0.06       -0.04   0.00 -0.07
## F1_20            0.05  0.12        0.09   0.05  0.13        0.09   0.06  0.12
##                diffF4F3_50 BW4_50 F4_55 diffF4F3_55 BW4_55 F4_60 diffF4F3_60
## Initial_F1_50        0.03   0.14  0.19        0.04   0.14  0.20        0.03
## Initial_F2_50        0.03   0.11  0.30        0.03   0.12  0.32        0.04
## Initial_F3_50       -0.02   0.05  0.33       -0.02   0.05  0.35       -0.01
## Initial_F4_50        0.20   0.13  0.48        0.20   0.14  0.49        0.21
## TokenDur            -0.03   0.00 -0.06       -0.03   0.00 -0.04       -0.01
## F1_20                0.07   0.06  0.12        0.07   0.06  0.11        0.07
##                BW4_60 F4_65 diffF4F3_65 BW4_65 F4_70 diffF4F3_70 BW4_70 F4_75
## Initial_F1_50    0.16  0.20        0.03   0.17  0.22        0.04   0.15  0.22
## Initial_F2_50    0.13  0.32        0.03   0.12  0.31        0.03   0.11  0.31
## Initial_F3_50    0.07  0.35       -0.02   0.05  0.35       -0.03   0.04  0.34
## Initial_F4_50    0.16  0.49        0.20   0.15  0.49        0.19   0.14  0.49
## TokenDur         0.03 -0.05       -0.01   0.02 -0.04       -0.01   0.02 -0.06
## F1_20            0.07  0.12        0.07   0.07  0.12        0.08   0.05  0.12
##                diffF4F3_75 BW4_75 F4_80 diffF4F3_80 BW4_80 F4min F4max
## Initial_F1_50        0.04   0.14  0.23        0.03   0.16  0.09  0.27
## Initial_F2_50        0.03   0.11  0.32        0.02   0.13  0.22  0.36
## Initial_F3_50       -0.03   0.05  0.35       -0.03   0.06  0.28  0.37
## Initial_F4_50        0.18   0.13  0.50        0.17   0.15  0.37  0.54
## TokenDur            -0.02   0.04 -0.04        0.00   0.03 -0.28  0.16
## F1_20                0.07   0.04  0.12        0.06   0.04  0.04  0.22
##                time_F4min time_F4max F4range time_F4range slopeF4 F0min F0max
## Initial_F1_50       -0.03       0.01    0.16        -0.06    0.09  0.34  0.34
## Initial_F2_50       -0.03       0.01    0.15        -0.06    0.09  0.45  0.46
## Initial_F3_50       -0.02       0.00    0.10        -0.06    0.06  0.43  0.44
## Initial_F4_50       -0.03       0.00    0.18        -0.08    0.09  0.61  0.60
## TokenDur             0.02       0.01    0.37        -0.26    0.41 -0.07  0.07
## F1_20                0.04      -0.06    0.16        -0.08    0.10  0.17  0.19
##                F0rangeST time_F0min time_F0max absSlopeF0 Rpresent_encode
## Initial_F1_50       0.06      -0.08       0.04       0.17           -0.09
## Initial_F2_50       0.14       0.01      -0.06       0.23           -0.17
## Initial_F3_50       0.11       0.01      -0.03       0.25           -0.19
## Initial_F4_50       0.06       0.00       0.00       0.17           -0.25
## TokenDur            0.54      -0.17       0.08      -0.05            0.12
## F1_20               0.08      -0.05       0.04       0.03           -0.02
```

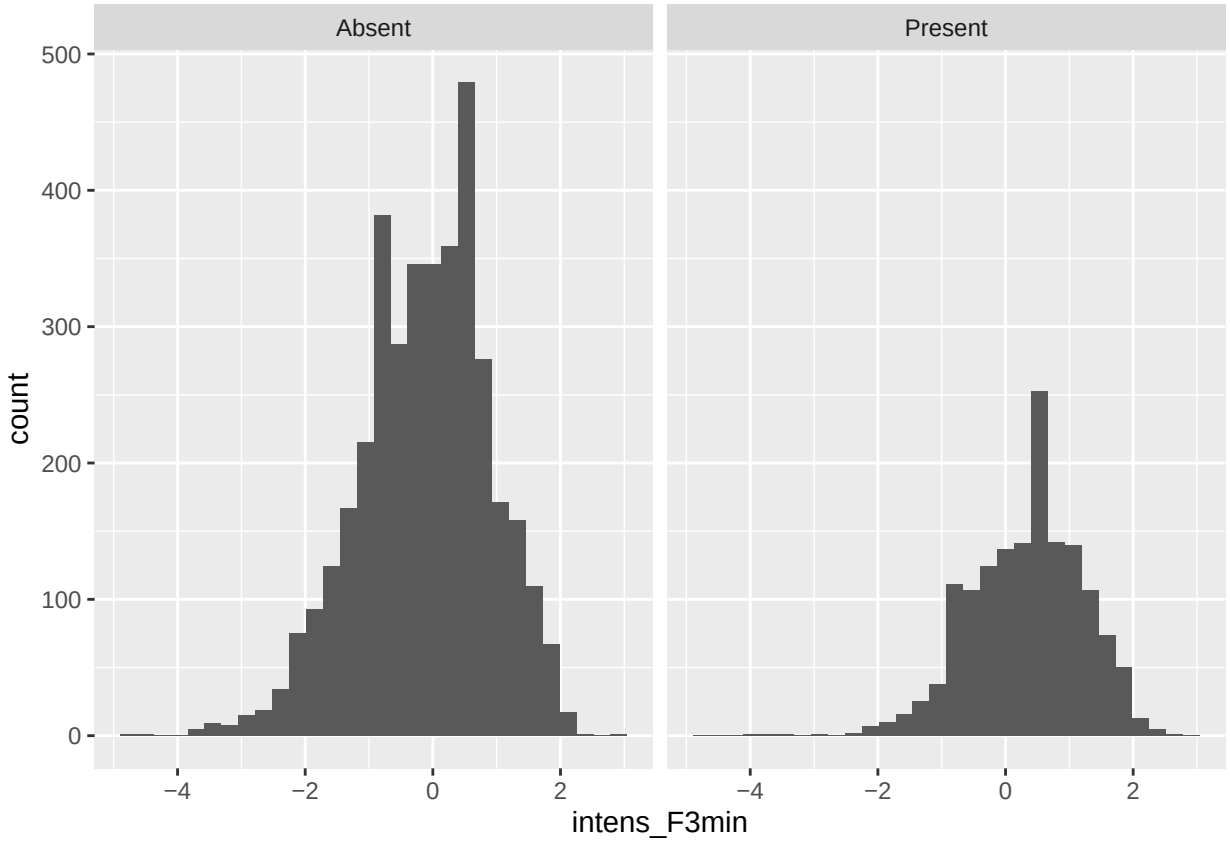
```
cor_rdf_usefurl_df$row <- colnames(cor_rdf_usefurl_df)
df <- sqldf("SELECT row, Rpresent_encode
             FROM cor_rdf_usefurl_df
             ORDER BY Rpresent_encode DESC
             LIMIT 4")
```

```
df
```

```
##                row Rpresent_encode
## 1 Rpresent_encode            1.00
## 2    intens_F3min            0.21
## 3    intens_F3max            0.19
## 4     diffF4F3_40            0.15
```

```
ggplot(data = rdf_useful, aes(x = intens_F3min)) + geom_histogram() + facet_grid(.~Rpresent)
```

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



2. EDA for question 2

```
Rclass_token <- Rclass
Rclass_token$token <- 'r'
```

set column names to the lower case

```
colnames(Rclass_token) <- tolower(colnames(Rclass_token))
colnames(Tclass) <- tolower(colnames(Tclass))
```

list the all same column in two files

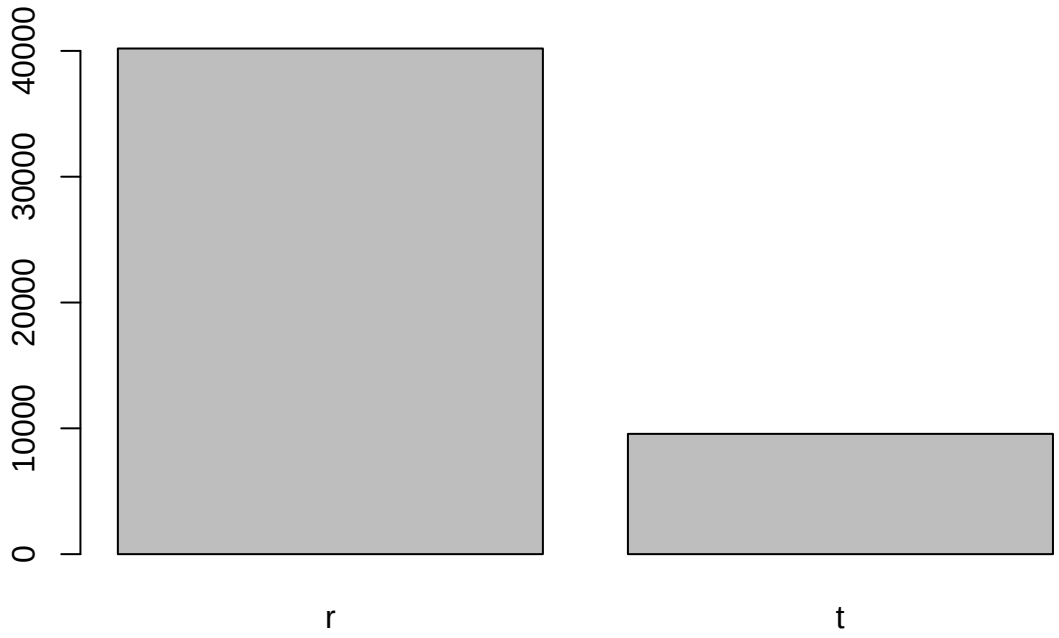
```
same <- Reduce(intersect, list(colnames(Rclass_token), colnames(Tclass)))
same
```

```
##  [1] "tokennum"   "speaker"    "gender"     "matchid"    "stress"
##  [6] "celexfreq"  "syllable"   "syllstart"  "syllend"    "word"
## [11] "wordstart"  "wordend"    "tokenstart" "tokenend"   "corpusfreq"
## [16] "howcoded"   "tokendur"   "bw1_20"     "bw1_25"     "bw1_30"
## [21] "bw1_35"     "bw1_40"     "bw1_45"     "bw1_50"     "bw1_55"
## [26] "bw1_60"     "bw1_65"     "bw1_70"     "bw1_75"     "bw1_80"
## [31] "bw2_20"     "bw2_25"     "bw2_30"     "bw2_35"     "bw2_40"
## [36] "bw2_45"     "bw2_50"     "bw2_55"     "bw2_60"     "bw2_65"
## [41] "bw2_70"     "bw2_75"     "bw2_80"     "bw3_20"     "bw3_25"
## [46] "bw3_30"     "bw3_35"     "bw3_40"     "bw3_45"     "bw3_50"
## [51] "bw3_55"     "bw3_60"     "bw3_65"     "bw3_70"     "bw3_75"
## [56] "bw3_80"     "token"
```

merge these two dataframe with same colums

```
merge_df <- rbind(Rclass_token[, same], Tclass[, same])
merge_df <- na.omit(merge_df)
```

```
barplot(table(merge_df$token))
```



the label are unbalanced

```
head(merge_df)
```

```
## # A tibble: 6 x 57
##   tokennum speaker  gender matchid  stress celex~1 sylla~2 sylls~3 syllend word
##      <int> <fct>    <fct>  <fct>    <fct>    <int> <fct>     <dbl>   <dbl> <fct>
## 1        1 Speaker1 Male   g_243;e~ 0          522 t@         0.18    0.28 Word~
## 2        2 Speaker1 Male   g_243;e~ ’         1634 ’b$n      61.7    61.9  Word~
## 3        3 Speaker1 Male   g_243;e~ 0          725 b@        65.2    65.4  Word~
## 4        4 Speaker1 Male   g_243;e~ ’           11 ’g$       66.6    67.0  Word~
## 5        5 Speaker1 Male   g_243;e~ ’        21107 ’w3       67.7    67.8  Word~
## 6        6 Speaker1 Male   g_243;e~ ’         1383 ’T3       68.2    68.4  Word~
## # ... with 47 more variables: wordstart <dbl>, wordend <dbl>, tokenstart <dbl>,
## #   tokenend <dbl>, corpusfreq <int>, howcoded <fct>, tokendur <dbl>,
## #   bw1_20 <dbl>, bw1_25 <dbl>, bw1_30 <dbl>, bw1_35 <dbl>, bw1_40 <dbl>,
## #   bw1_45 <dbl>, bw1_50 <dbl>, bw1_55 <dbl>, bw1_60 <dbl>, bw1_65 <dbl>,
## #   bw1_70 <dbl>, bw1_75 <dbl>, bw1_80 <dbl>, bw2_20 <dbl>, bw2_25 <dbl>,
## #   bw2_30 <dbl>, bw2_35 <dbl>, bw2_40 <dbl>, bw2_45 <dbl>, bw2_50 <dbl>,
## #   bw2_55 <dbl>, bw2_60 <dbl>, bw2_65 <dbl>, bw2_70 <dbl>, bw2_75 <dbl>, ...
## # i Use ‘colnames()‘ to see all variable names
```

```
num_token <- select_if(merge_df, is.numeric)
num_token$token <- merge_df$token
head(num_token)
```

```
## # A tibble: 6 x 50
##   tokennum celexfreq syllstart syllend wordstart wordend token~1 token~2 corpu~3
##      <int>     <int>     <dbl>   <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <int>
## 1        1       522      0.18    0.28      0.12    0.62    0.24    0.28      41
## 2        2      1634     61.7    61.9      61.7    61.9    61.7    61.9      117
## 3        3       725     65.2    65.4      64.7    65.4    65.2    65.4       13
## 4        4        11     66.6    67.0      66.6    67.0    66.7    67.0       95
## 5        5     21107     67.7    67.8      67.7    67.8    67.8    67.8     2442
## 6        6      1383     68.2    68.4      68.2    68.5    68.3    68.4      103
## # ... with 41 more variables: tokendur <dbl>, bw1_20 <dbl>, bw1_25 <dbl>,
## #   bw1_30 <dbl>, bw1_35 <dbl>, bw1_40 <dbl>, bw1_45 <dbl>, bw1_50 <dbl>,
## #   bw1_55 <dbl>, bw1_60 <dbl>, bw1_65 <dbl>, bw1_70 <dbl>, bw1_75 <dbl>,
## #   bw1_80 <dbl>, bw2_20 <dbl>, bw2_25 <dbl>, bw2_30 <dbl>, bw2_35 <dbl>,
## #   bw2_40 <dbl>, bw2_45 <dbl>, bw2_50 <dbl>, bw2_55 <dbl>, bw2_60 <dbl>,
## #   bw2_65 <dbl>, bw2_70 <dbl>, bw2_75 <dbl>, bw2_80 <dbl>, bw3_20 <dbl>,
## #   bw3_25 <dbl>, bw3_30 <dbl>, bw3_35 <dbl>, bw3_40 <dbl>, bw3_45 <dbl>, ...
## # i Use ‘colnames()‘ to see all variable names
```

scale the numeric variable

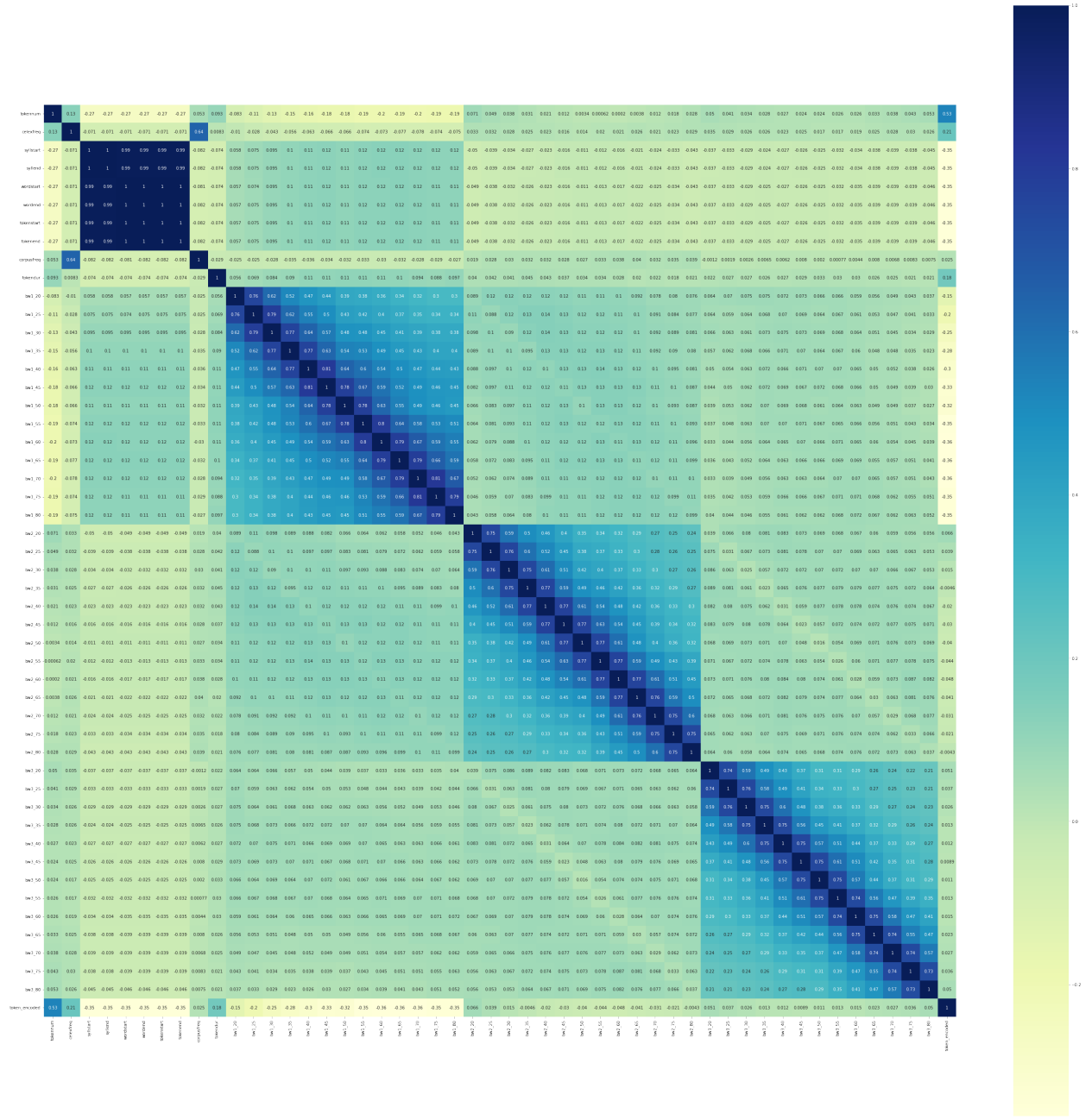
```
# Standardize the data
standardized_mergedata <- scale(num_token[, c(1:49)], center = TRUE, scale = TRUE)
num_token[, c(1:49)] <- standardized_mergedata
```

encode the token. r-1, t-0

```
num_token$token_encoded <- ifelse(num_token$token == 'r', 1, 0)
merge_cor <- cor(num_token[, c(1:49, 51)])
```

correlation map between variables(use python to draw it)

```
write_csv(as.data.frame(num_token[, c(1:49, 51)]), file = "merge_df.csv")
```



print all the variables name, whose correlations are higher than 0.9

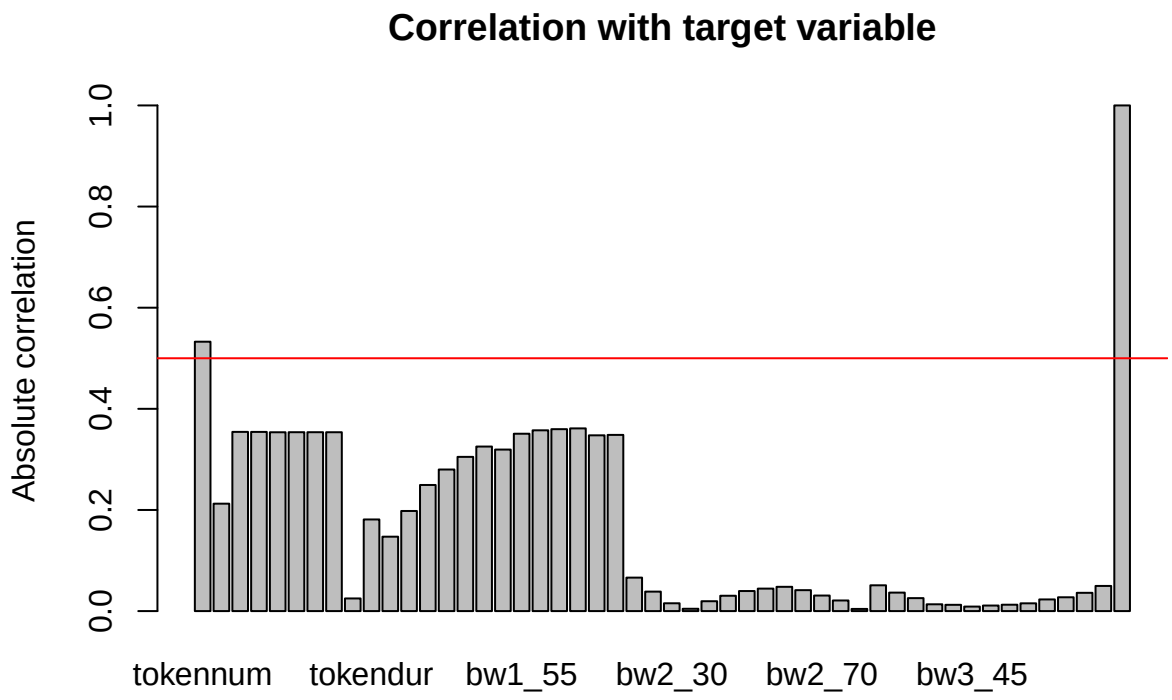
```
high_corr2 <- data.frame(var1 = character() , var2 = character(), cor = numeric())
z = 1
for (i in 1:50) {
```

```
  for (j in 1:i) {
    if (abs(merge_cor[i, j]) > 0.9 && i != j) {
      high_corr2[z, 1] <- colnames(merge_cor)[j]
      high_corr2[z, 2] <- colnames(merge_cor)[i]
      high_corr2[z, 3] <- merge_cor[i, j]
      z = z +1
    }
  }
}
high_corr2
```

```
##          var1       var2       cor
## 1   syllstart    syllend 0.9999999
## 2   syllstart  wordstart 0.9907643
## 3     syllend  wordstart 0.9907647
## 4   syllstart    wordend 0.9907647
## 5     syllend    wordend 0.9907653
## 6   wordstart    wordend 0.9999998
## 7   syllstart tokenstart 0.9907649
## 8     syllend tokenstart 0.9907654
## 9   wordstart tokenstart 0.9999999
## 10    wordend tokenstart 0.9999999
## 11  syllstart   tokenend 0.9907645
## 12    syllend   tokenend 0.9907651
## 13  wordstart   tokenend 0.9999998
## 14    wordend   tokenend 0.9999999
## 15 tokenstart   tokenend 0.9999999
```

for the correlation between variables and target

```
barplot(abs(merge_cor[, 50]), main = "Correlation with target variable", ylab = "Absolute correlation")
abline(h = 0.5, col = "red")
```




```
merge_cor_df <- as.data.frame(merge_cor)
head(merge_cor_df)
```

```
##             tokennum   celexfreq   syllstart     syllend   wordstart
## tokennum   1.0000000  0.12622317 -0.26666716 -0.26664952 -0.26636383
## celexfreq  0.1262232  1.00000000 -0.07123087 -0.07132948 -0.07110637
## syllstart -0.2666672 -0.07123087  1.00000000  0.99999985  0.99076432
## syllend   -0.2666495 -0.07132948  0.99999985  1.00000000  0.99076471
## wordstart -0.2663638 -0.07110637  0.99076432  0.99076471  1.00000000
## wordend   -0.2664148 -0.07137406  0.99076470  0.99076533  0.99999976
##               wordend  tokenstart    tokenend  corpusfreq     tokendur
## tokennum  -0.26641479 -0.26640901 -0.26637849  0.05308553  0.092704090
## celexfreq -0.07137406 -0.07123759 -0.07125554  0.64001510  0.008308352
## syllstart  0.99076470  0.99076494  0.99076450 -0.08180775 -0.074272734
## syllend    0.99076533  0.99076536  0.99076509 -0.08192476 -0.074079564
## wordstart  0.99999976  0.99999991  0.99999984 -0.08137903 -0.074125266
## wordend    1.00000000  0.99999986  0.99999992 -0.08164840 -0.073956194
##                bw1_20      bw1_25      bw1_30      bw1_35      bw1_40
## tokennum  -0.08301178 -0.10774826 -0.13368482 -0.15371226 -0.16470995
## celexfreq -0.01000265 -0.02846127 -0.04304287 -0.05602254 -0.06287913
## syllstart  0.05774585  0.07529451  0.09480663  0.10399363  0.11430654
## syllend    0.05775126  0.07530717  0.09482029  0.10400774  0.11432411
## wordstart  0.05709911  0.07447045  0.09453451  0.10340941  0.11462350
## wordend    0.05715179  0.07453856  0.09461156  0.10349239  0.11471399
##                bw1_45      bw1_50      bw1_55      bw1_60      bw1_65
```

```
## tokennum  -0.17776892 -0.17582381 -0.19407216 -0.19619835 -0.19370642
## celexfreq -0.06595692 -0.06572713 -0.07419592 -0.07265706 -0.07683835
## syllstart  0.11896894  0.11041069  0.12058766  0.12261137  0.12332538
## syllend    0.11898905  0.11043143  0.12060815  0.12263650  0.12334830
## wordstart  0.11849624  0.11025820  0.12057172  0.12225102  0.12292031
## wordend    0.11859308  0.11035335  0.12067062  0.12235429  0.12302323
##                bw1_70      bw1_75     bw1_80      bw2_20      bw2_25
## tokennum  -0.19794399 -0.19086765 -0.1899234  0.07137890  0.04895939
## celexfreq -0.07752644 -0.07447837 -0.0754227  0.03322598  0.03230431
## syllstart  0.11913874  0.11525554  0.1154944 -0.04982556 -0.03926407
## syllend    0.11915983  0.11527866  0.1155216 -0.04982175 -0.03926346
## wordstart  0.11870656  0.11427354  0.1148882 -0.04937254 -0.03807990
## wordend    0.11881136  0.11437924  0.1149968 -0.04937753 -0.03808427
##                bw2_30      bw2_35      bw2_40      bw2_45      bw2_50
## tokennum   0.03787788  0.03097908  0.02102057  0.01226560  0.00341216
## celexfreq  0.02770229  0.02545353  0.02264716  0.01646396  0.01398996
## syllstart -0.03378832 -0.02675691 -0.02280291 -0.01559208 -0.01081496
## syllend   -0.03379208 -0.02676122 -0.02280826 -0.01559930 -0.01082360
## wordstart -0.03248183 -0.02620927 -0.02267662 -0.01576303 -0.01113131
## wordend   -0.03248622 -0.02621072 -0.02267832 -0.01576207 -0.01113157
##                bw2_55        bw2_60       bw2_65      bw2_70      bw2_75
## tokennum   0.000615153  0.0002017475  0.003752653  0.01247119  0.01804350
## celexfreq  0.019539451  0.0214781281  0.026130384  0.02125110  0.02277750
## syllstart -0.012132562 -0.0164749616 -0.021418263 -0.02427605 -0.03331663
## syllend   -0.012142356 -0.0164905565 -0.021431175 -0.02428911 -0.03333028
## wordstart -0.012519019 -0.0169438934 -0.021588905 -0.02462119 -0.03363747
## wordend   -0.012518298 -0.0169475795 -0.021591170 -0.02462387 -0.03364069
##                bw2_80      bw3_20      bw3_25      bw3_30      bw3_35
## tokennum   0.02817396  0.04969308  0.04073739  0.03369148  0.02765857
## celexfreq  0.02870046  0.03479764  0.02914559  0.02605637  0.02588928
## syllstart -0.04253146 -0.03701013 -0.03297883 -0.02864128 -0.02445693
## syllend   -0.04254336 -0.03701442 -0.03297994 -0.02864034 -0.02445886
## wordstart -0.04284923 -0.03691591 -0.03310146 -0.02888397 -0.02526114
## wordend   -0.04285422 -0.03691907 -0.03309975 -0.02887842 -0.02525751
##                bw3_40      bw3_45      bw3_50      bw3_55      bw3_60
## tokennum   0.02666640  0.02358924  0.02419332  0.02602049  0.02552752
## celexfreq  0.02330098  0.02498030  0.01711503  0.01673983  0.01918719
## syllstart -0.02651951 -0.02579327 -0.02506965 -0.03168192 -0.03431759
## syllend   -0.02652062 -0.02579854 -0.02507134 -0.03168385 -0.03431868
## wordstart -0.02712898 -0.02637458 -0.02531881 -0.03172231 -0.03454167
## wordend   -0.02712228 -0.02637027 -0.02531231 -0.03171646 -0.03453523
##                bw3_65      bw3_70      bw3_75      bw3_80 token_encoded
## tokennum   0.03270611  0.03768054  0.04304405  0.05306184     0.5327371
## celexfreq  0.02520056  0.02765112  0.03000720  0.02603427     0.2124238
## syllstart -0.03814874 -0.03865564 -0.03843187 -0.04525579    -0.3543960
## syllend   -0.03815194 -0.03865708 -0.03843278 -0.04525398    -0.3543411
## wordstart -0.03860257 -0.03926068 -0.03907648 -0.04601838    -0.3536606
## wordend   -0.03859950 -0.03925697 -0.03907452 -0.04601602    -0.3537470
```

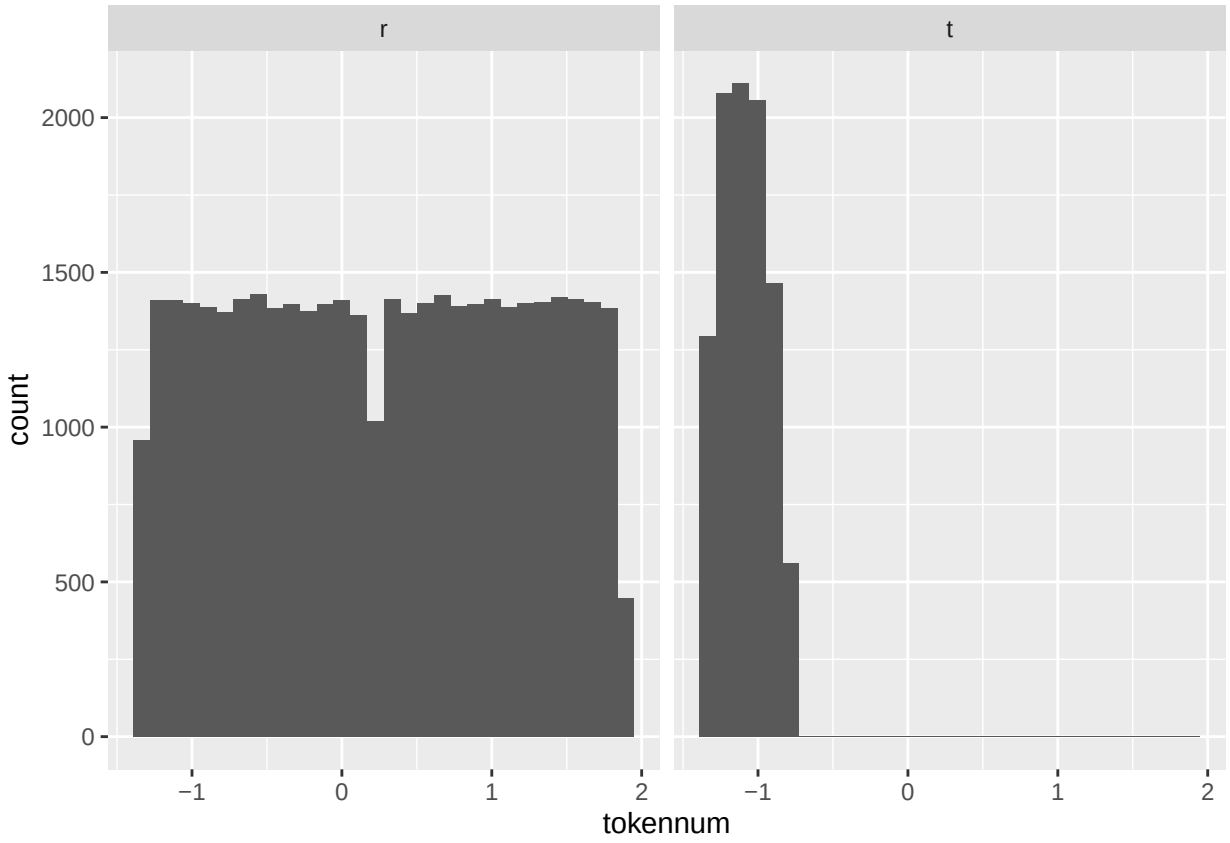
```
merge_cor_df$row <- colnames(merge_cor_df)
sqldf("SELECT row, token_encoded
      FROM merge_cor_df
      ORDER BY token_encoded DESC
      LIMIT 4")
```

```
##             row token_encoded
## 1 token_encoded     1.0000000
## 2      tokennum     0.5327371
## 3     celexfreq     0.2124238
## 4      tokendur     0.1811820
```

the distribution of variables with relatively high correlation with token

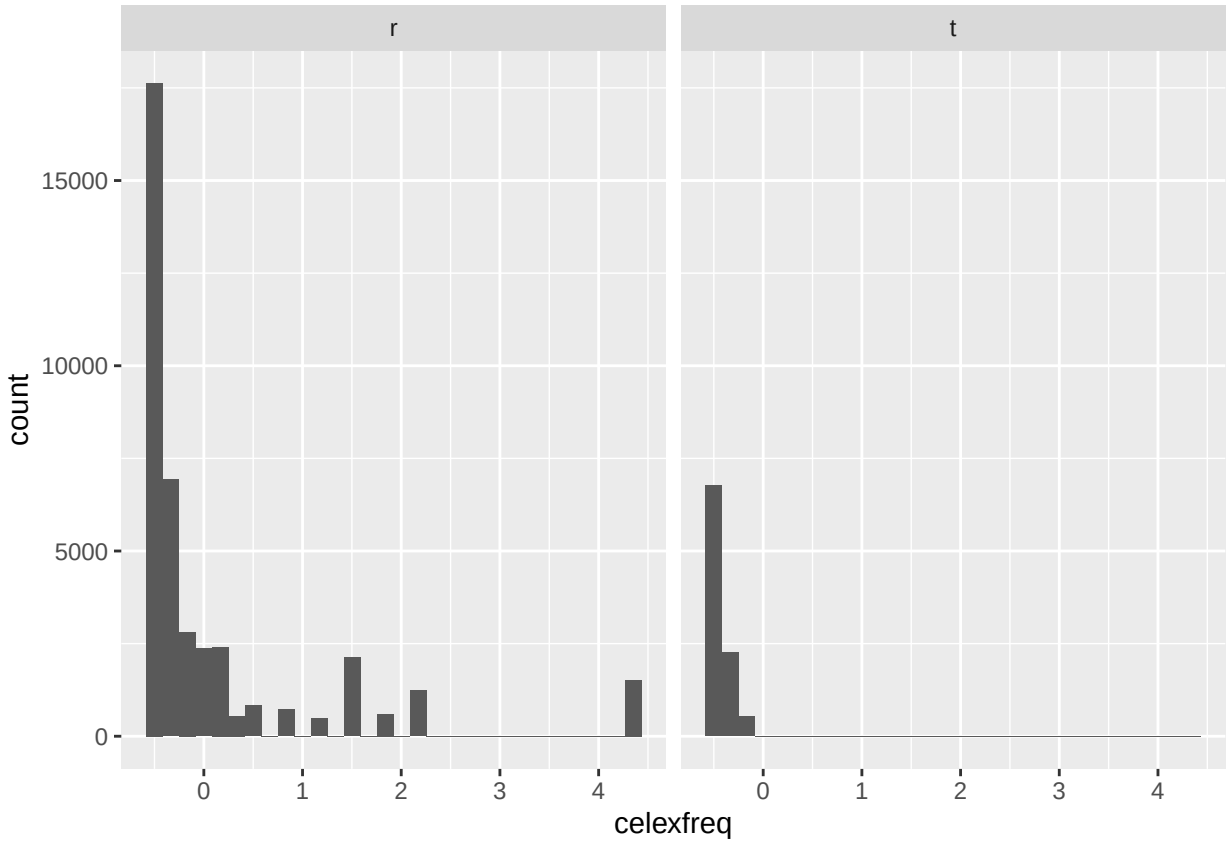
```
ggplot(data = num_token, aes(x = tokennum)) + geom_histogram() + facet_grid(.~token)
```

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




```
ggplot(data = num_token, aes(x = celexfreq)) + geom_histogram() + facet_grid(.~token)
```

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


r
t
15000
10000
5000
0
count
0
1
2
3
4
celexfreq