```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: 패키지 'tibble'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'tidyr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'dplyr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## Warning: 패키지 'tidymodels'는 R 버전 4.1.3에서 작성되었습니다
```

```
## -- Attaching packages --------------------------------------- tidymodels 0.2.0 --
```

```
## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.1      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0
```

```
## Warning: 패키지 'broom'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'dials'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'infer'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'modeldata'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'parsnip'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'recipes'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'rsample'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'tune'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'workflows'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'workflowsets'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'yardstick'는 R 버전 4.1.3에서 작성되었습니다
```

```
## -- Conflicts --------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```
library(caret)
```

```
## Warning: 패키지 'caret'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: lattice
```

```
##
## 다음의 패키지를 부착합니다: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(tidytext)
```

```
## Warning: 패키지 'tidytext'는 R 버전 4.1.3에서 작성되었습니다
```

```
library(KoNLP)
```

```
## Checking user defined dictionary!
```

```
library(wordcloud)
```

```
## Warning: 패키지 'wordcloud'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: RColorBrewer
```

```
library(wordcloud2)
```

```
## Warning: 패키지 'wordcloud2'는 R 버전 4.1.3에서 작성되었습니다
```

```
df <- read.csv("C:WWUsersWWUserWWDesktopWWdataWWcallmebyyourname.csv")
df$review <- str_replace_all(df$review, '[^가-힣 ]', '')
head(df$review)
```

```
## [1] "이 영화의 아름다움을 영원히 느끼지 못하고 평점테러중인 호모포비아들이 애잔함"
## [2] "우린 서른살이 되기 전에 이미 무너져새로운 사람을 만나도 더이상 보여줄 내가 없어지게 되
지실제로 저런 조언을 해주는 아버지가 몇이나 될것이며 저런 시선으로 바라봐주는 세상이 존재는 할
까영화가 "
## [3] "감정에 솔직하다는건 굉장히 부러운 일이다"
## [4] "소년의 첫사랑 마지막즈음 아버지의 대사가 인상 깊었고 나중에 올리버 시점으로 나왔으면"
## [5] "전화 목소리를 잘 분간하지 못하던 소년은 반년만에 들려온 그의 목소리는 단박에 알아듣는다
오랜만에 인물의 감정 속에 푹 빠져 본 영화였다"
## [6] "관람객      "
```

```
#용어처리
df$review <- str_replace_all(df$review, '^관람객', '')
df$review <- str_replace_all(df$review, '티모시', '티모시 샬라메')
df$review <- str_replace_all(df$review, '티모시샬라메', '티모시 샬라메')
df$review <- str_replace_all(df$review, '알리오', '엘리오')
df$review <- str_replace_all(df$review, '콜마넴', '콜바넴')
df$review <- str_replace_all(df$review, '콜미바이유어네임', '콜바넴')
```

#워드클라우드만들기

```
library(tm)
```

```
## Warning: 패키지 'tm'는 R 버전 4.1.3에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: NLP
```

```
##
## 다음의 패키지를 부착합니다: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
CRPS <- VCorpus(VectorSource(df$review))
DT <- DocumentTermMatrix(CRPS)
```

```
CRPS <- tm_map(CRPS, content_transformer(tolower))
CRPS <- tm_map(CRPS, removeNumbers)
CRPS <- tm_map(CRPS, removeWords, stopwords('English'))
CRPS <- tm_map(CRPS, removePunctuation)
CRPS <- tm_map(CRPS, stripWhitespace)
DT <- DocumentTermMatrix(CRPS)
```

```
DTM <- as.matrix(DT)
wrdfrq <- sort(colSums(DTM), decreasing=TRUE)
WRDFRQ <- data.frame(wrd=names(wrdfrq), n=wrdfrq)
head(WRDFRQ)
```

```
##              wrd  n
## 아름다운 아름다운 49
## 여운이     여운이 36
## 엘리오의 엘리오의 32
## 마지막     마지막 31
## 엘리오     엘리오 31
## 영화를     영화를 24
```

```
library(RColorBrewer)
pal <- brewer.pal(11,'Spectral')
wordcloud(WRDFRQ$wrd, WRDFRQ$n, min.freq=1, max.words=100,
 random.order=FALSE, rot.per=0.5, colors=pal)
```

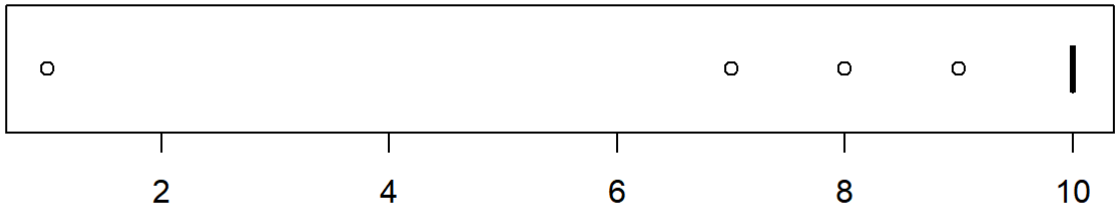
#

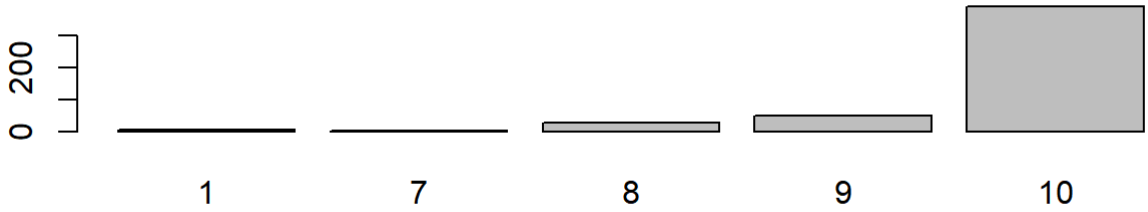
평점예측

```
df$rvw <- sapply(df$review, function(txt){paste(KoNLP::extractNoun(txt), collapse=' ')},USE.NAMES=FALSE)
head(df$rvw)
```

```
## [1] "영화 아름다움 평점 테러 중인 호모포비아들이 애잔 함"
## [2] "우 서른 살 전 무너져새로운 사람 나 되지실제로 조언 아버지 몇 것 시선 세상 존재 할까영화"
## [3] "감정 솔직하다는건 일"
## [4] "소년 첫사랑 마지막 즈음 아버지 대사 인상 나중 올리버 시점 나왔으 면"
## [5] "전화 목소리 분간 소년 반년 만 그 목소리 오랜만 인물 감정 속 영화였"
## [6] "      "
```

```
#타겟정의
par(mfrow=c(2,1))
boxplot(df$rating, horizontal=TRUE)
barplot(table(df$rating))
```






```
par(mfrow=c(1,1))
summary(df$rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000  10.000  10.000   9.625  10.000  10.000
```

```
#y정의
library(dplyr)
library(RmecabKo)
```

```
## Warning: 패키지 'RmecabKo'는 R 버전 4.1.3에서 작성되었습니다
```

```
##
## 다음의 패키지를 부착합니다: 'RmecabKo'
```

```
## The following object is masked from 'package:NLP':
##
##     words
```

```
df <- df %>% mutate(y=factor(ifelse(df$rating>=9, 'pos','neg')))
table(df$y)
```

```
##
## neg pos
##  39 441
```

#분할

```
set.seed(20180178)
ISP <- initial_split(df, prop=3/4, strata=y)
TRDF <- training(ISP)
TSDF <- testing(ISP)
rbind(dim(TRDF), dim(TSDF))
```

```
##      [,1] [,2]
## [1,]  360    8
## [2,]  120    8
```

```
#불용어처리
mystopwords <- c('가는', '그는', '그를', '그의', '그리고','이건','이게','이런')


CRPS <- CRPS %>% tm_map(content_transformer(tolower)) %>% tm_map(removeNumbers) %>%tm_map(remov
eWords, stopwords('english')) %>% tm_map(removeWords, mystopwords) %>% tm_map(removePunctuatio
n) %>%tm_map(stripWhitespace)
```

```
#TRCRPS 생성
TRCRPS <- VCorpus(VectorSource(TRDF$rvw))
TRCRPS <- tm_map(TRCRPS, removeWords, mystopwords)
TRCRPS #요약정보
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 360
```

```
TRDT <- DocumentTermMatrix(TRCRPS,control=list(wordLengths=c(2,10), weighting=weightTf))
TRDT <- removeSparseTerms(TRDT, sparse=0.995)
c(nDocs(TRDT), nTerms(TRDT))
```

```
## [1] 360 240
```

```
head(Terms(TRDT))
```

```
## [1] "가슴" "각오" "간만" "간직" "감각" "감독"
```

```
TR <- as.data.frame(as.matrix(TRDT))
TR <- TR %>% mutate(y=TRDF$y)
TR[1:3, 1:5]
```

```
##   가슴 각오 간만 간직 감각
## 1    0    0    0    0    0
## 2    0    0    0    0    0
## 3    0    0    0    0    0
```

```
#TSCRPS 생성
TSCRPS <- VCorpus(VectorSource(TSDF$rvw))
mystopwords <- c('가는', '그는', '그를', '그의', '그리고','이건','이게','이런')
TSCRPS <- tm_map(TSCRPS, removeWords, mystopwords)
TSCRPS
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 120
```

```
TSDT <- DocumentTermMatrix(TSCRPS,control=list(dictionary=Terms(TRDT)))
TSDT
```

```
## <<DocumentTermMatrix (documents: 120, terms: 240)>>
## Non-/sparse entries: 101/28699
## Sparsity           : 100%
## Maximal term length: 7
## Weighting          : term frequency (tf)
```

```
as.matrix(TSDT)[1:3, 1:5]
```

```
##     Terms
## Docs 가슴 각오 간만 간직 감각
##    1    0    0    0    0    0
##    2    0    0    0    0    0
##    3    0    0    0    0    0
```

```
TS <- as.data.frame(as.matrix(TSDT))
TS <- TS %>% mutate(y=TSDF$y)
TS[1:3, 1:5]
```

```
##   가슴 각오 간만 간직 감각
## 1    0    0    0    0    0
## 2    0    0    0    0    0
## 3    0    0    0    0    0
```

```
Mg <- glm(y~., data=TR, family=binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
sort(coef(Mg))[1:10]
```

```
##       유어       처럼       간직       절절       나이       풍경     연출력       보고
## -330.0147 -295.2396 -281.3112 -260.6019 -239.1911 -237.7401 -237.1751 -188.5096
##     사람들       이유
## -182.0045 -180.3376
```

```
#긍정단어
sort(coef(Mg), decreasing=TRUE)[1:10]
```

```
##       관객       어린 인상깊었다       그것       얼굴     사운드       이거
##   336.8996   321.6811   311.6716   220.9998   217.2676   210.8095   208.9785
## 엘리오처럼       맨살       인물
##   204.8795   198.3074   180.7822
```

```
yhg <- factor(ifelse(predict(Mg, newdata=TR, type='response')>=0.5, 'pos', 'neg'))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
head(yhg)
```

```
##   1   2   3   4   5   6
## pos pos pos pos pos pos
## Levels: neg pos
```

```
#TR평가 적중률 0.9889(과적합?)
confusionMatrix(yhg, TR$y, positive='pos')
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction neg pos
##        neg  23   0
##        pos   4 333
##
##               Accuracy : 0.9889
##                 95% CI : (0.9718, 0.997)
##    No Information Rate : 0.925
##    P-Value [Acc > NIR] : 2.221e-08
##
##                  Kappa : 0.9141
##
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 1.0000
##             Specificity : 0.8519
##          Pos Pred Value : 0.9881
##          Neg Pred Value : 1.0000
##              Prevalence : 0.9250
##          Detection Rate : 0.9250
##    Detection Prevalence : 0.9361
##       Balanced Accuracy : 0.9259
##
##        'Positive' Class : pos
##
```

```
#TS에서 평가
#적중률0.7417
yhg <- factor(ifelse(predict(Mg, newdata=TS, type='response')>=0.5, 'pos', 'neg'))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
confusionMatrix(yhg, TS$y, positive='pos')
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction neg pos
##        neg   1  20
##        pos  11  88
##
##               Accuracy : 0.7417
##                 95% CI : (0.6538, 0.8172)
##    No Information Rate : 0.9
##    P-Value [Acc > NIR] : 1.0000
##
##                  Kappa : -0.0764
##
##  Mcnemar's Test P-Value : 0.1508
##
##             Sensitivity : 0.81481
##             Specificity : 0.08333
##          Pos Pred Value : 0.88889
##          Neg Pred Value : 0.04762
##              Prevalence : 0.90000
##          Detection Rate : 0.73333
##    Detection Prevalence : 0.82500
##       Balanced Accuracy : 0.44907
##
##        'Positive' Class : pos
##
```