훈련 및 테스트 데이터셋 생성

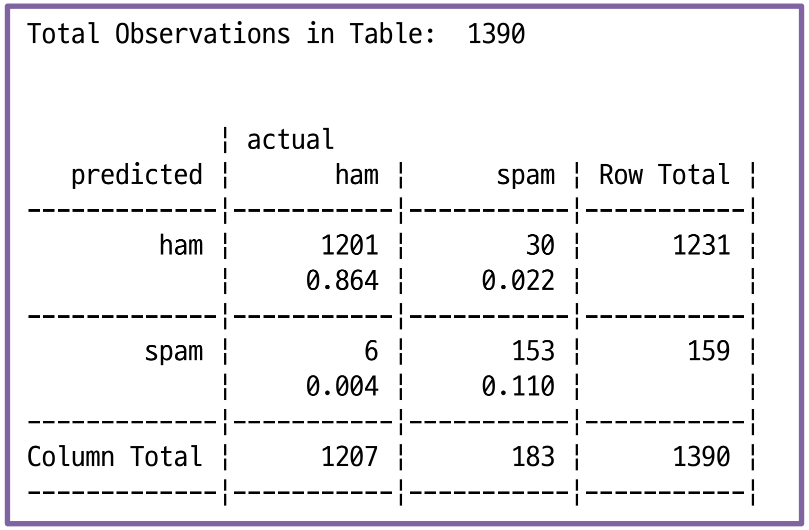
```
#훈련데이터와 테스트데이터 생성
sms_dtm_train <- sms_dtm[1:4169,] (75%)
sms_dtm_test <- sms_dtm[4170:5559,] (25%)

  #레이블 생성
  sms_train_labels <- sms_raw[1:4169,]$type
  sms_test_labels <- sms_raw[4170:5559,]$type

  #spam 비율 비교
  prop.table(table(sms_train_labels))
  prop.table(table(sms_test_labels))
```

훈련 및 테스트 데이터셋 생성

```
#자주 나타나는 단어 저장
sms_freq_words <- findFreqTerms(sms_dtm_train, 5)

#자주 나타나는 단어들로 구성된 DTM만들기
sms_dtm_freq_train <- sms_dtm_train[ , sms_freq_words]
sms_dtm_freq_test <- sms_dtm_test[ , sms_freq_words]

#sparse matrix 값을 범주형으로 변환
convert_counts <- function(x){
  x <- ifelse(x>0, "yes", "no")
}
sms_train <- apply(sms_dtm_freq_train, MARGIN =2, convert_counts)
sms_test <- apply(sms_dtm_freq_test, MARGIN =2, convert_counts)
```

모델 훈련 및 성능 평가

```
# 모델 훈련
sms_classifier <- naiveBayes(sms_train,sms_train_labels)

#모델 성능 평가
sms_test_pred <- predict(sms_classifier,sms_test)
library(gmodels)
CrossTable(sms_test_pred,sms_test_labels,prop.chisq = FALSE, prop.c = FALSE,
prop.r = FALSE, dnn=c('predicted', 'actual'))
```



```
Total Observations in Table:  1390


             | actual
   predicted |       ham |      spam | Row Total |
-------------|-----------|-----------|-----------|
         ham |      1201 |        30 |      1231 |
             |     0.864 |     0.022 |           |
-------------|-----------|-----------|-----------|
        spam |         6 |       153 |       159 |
             |     0.004 |     0.110 |           |
-------------|-----------|-----------|-----------|
Column Total |      1207 |       183 |      1390 |
-------------|-----------|-----------|-----------|
```

Accuracy=0.974

Error=0.026

모델 성능 개선

```r
#모델 성능 개선
sms_classifier3 <- naiveBayes(sms_train, sms_train_labels, laplace = 1)
sms_test_pred3 <- predict(sms_classifier3,sms_test)
CrossTable(sms_test_pred3, sms_test_labels,
                    prop.chisq = FALSE, prop.c=FALSE, prop.r = FALSE,
                    dnn=c('predicted','actual'))
```

```
Total Observations in Table:  1390


             ¦ actual
   predicted ¦       ham ¦      spam ¦ Row Total ¦
-------------¦-----------¦-----------¦-----------¦
         ham ¦      1189 ¦        16 ¦      1205 ¦
             ¦     0.855 ¦     0.012 ¦           ¦
-------------¦-----------¦-----------¦-----------¦
        spam ¦        18 ¦       167 ¦       185 ¦
             ¦     0.013 ¦     0.120 ¦           ¦
-------------¦-----------¦-----------¦-----------¦
Column Total ¦      1207 ¦       183 ¦      1390 ¦
-------------¦-----------¦-----------¦-----------¦
```

Accuracy=0.976

Error=0.024