## test

### 2022-11-06

```
#test4
# df = read.csv("amazon_review_polarity_csv/train.csv", header = FALSE)
set.seed(1)
N = 100000
N t = 0.8*N
reviews_text<-readLines("amazon_review_polarity_csv/train.csv", n = N)</pre>
reviews_text<-data.frame(reviews_text)</pre>
library(tidyr)
reviews_text<-separate(data = reviews_text, col = reviews_text, into = c("Sentiment",
                               "SentimentText"), sep = 4)
# Retaining only alphanumeric values in the sentiment column
reviews_text$Sentiment<-gsub("[^[:alnum:]]","",reviews_text$Sentiment)</pre>
# Retaining only alphanumeric values in the sentiment text
reviews_text$SentimentText<-gsub("[^[:alnum:]]"," ",reviews_text$SentimentText)
# Replacing multiple spaces in the text with single space
reviews\_text\$SentimentText <-gsub("(?<=[\\s])\\s*|^\\s+|\\s+$", "", reviews\_text\$SentimentText,
                                 perl=TRUE)
# Writing the output to a file that can be consumed in other projects
write.table(reviews_text,file = "Sentiment Analysis Dataset.csv",row.names = F,col.names = T,sep=',')
reviews_text<-readLines('amazon_review_polarity_csv/train.csv', n = N)
# Basic EDA to confirm that the data is read correctly
print(class(reviews_text))
## [1] "character"
print(length(reviews_text))
## [1] 100000
# print(head(reviews_text,2))
# Replacing the positive sentiment value 2 with __label__2
reviews_text<-gsub("\\\",",","__label__2 ",reviews_text)</pre>
# Replacing the negative sentiment value 1 with __label__1
reviews_text<-gsub("\\\",","__label__1 ",reviews_text)</pre>
# Removing the unnecessary \" characters
reviews_text<-gsub("\\\""," ",reviews_text)</pre>
# Replacing multiple spaces in the text with single space
```

```
reviews\_text < -gsub("(?<=[\s])\s*|^\s+|\s+|", "", reviews\_text, perl=TRUE)
# Basic EDA post the required processing to confirm input is as desired
print("EDA POST PROCESSING")
## [1] "EDA POST PROCESSING"
print(class(reviews_text))
## [1] "character"
print(length(reviews_text))
## [1] 100000
# print(head(reviews_text,2))
# Writing the revamped file to the directory so we could use it with
# fastText sentiment analyzer project
fileConn<-file("Sentiment Analysis Dataset_ft.txt")</pre>
writeLines(reviews_text, fileConn)
close(fileConn)
BoW approach
library(SnowballC)
library(tm)
## Loading required package: NLP
# Reading the transformed file as a dataframe
text <- read.table(file='Sentiment Analysis Dataset.csv', sep=',', header = TRUE)
# Checking the dataframe to confirm everything is in tact
print(dim(text))
## [1] 100000
# Transforming the text into volatile corpus
train_corp = VCorpus(VectorSource(text$SentimentText))
print(train_corp)
## <<VCorpus>>
## Metadata: corpus specific: 0, document level (indexed): 0
## Content: documents: 100000
# Creating document term matrix
dtm_train <- DocumentTermMatrix(train_corp, control = list( tolower = TRUE,</pre>
  removeNumbers = TRUE, stopwords = TRUE, removePunctuation = TRUE, stemming = TRUE ))
# Basic EDA on dtm
inspect(dtm_train)
```

```
## <<DocumentTermMatrix (documents: 100000, terms: 74760)>>
## Non-/sparse entries: 3399444/7472600556
## Sparsity
                   : 100%
## Maximal term length: 188
## Weighting
               : term frequency (tf)
## Sample
##
        Terms
## Docs
         book get good great just like movi one read time
##
    1250
            0
               0
                    0
                          0
                              0
                                   0
                                        0
                                           0
##
               0
                    0
                          0
                              0
                                   0
                                           0
                                                    0
    56817
            0
##
    63995
            0 2
                  1
                          1
                                   2
                                           2
                                                    1
    6785
            0 7
                              0
                                                0
##
                    1
                          0
                                       0
                                           1
                                                    0
                                   1
            0 0
                  0
                             0
                          0
##
    69262
                                   0
                                          0
                                                0
                                                    0
    73633 1 0 0
                          2 0 3
                                      0 2
                                                    2
##
##
    79144 0 0 0
                          0 0 0
                                      0 0 0
                                                    0
##
    80872
            0 0
                    0
                          0
                             0
                                 0
                                       0 0
                                              0
                                                    0
##
    85894
            0 1
                    0
                          0
                              0 1
                                      0 1
                                             0
                                                    0
##
    87875
# Removing sparse terms
dtm_train = removeSparseTerms(dtm_train, 0.99)
inspect(dtm_train)
## <<DocumentTermMatrix (documents: 100000, terms: 645)>>
## Non-/sparse entries: 2131029/62368971
## Sparsity
## Maximal term length: 10
## Weighting
                   : term frequency (tf)
## Sample
                    :
##
## Docs
         book get good great just like movi one read time
                    0
                          0
                              2
                                           3
##
    34297
               1
                                   5
##
    38984
            6 0
                    0
                          1
                              1
                                   0
                                           1
                                                0
                                                    1
##
    42051
            3 0 1
                          0
                              2
                                   1
                                           3
                                                5
                                                    1
          0 1 0
##
    56269
                          0
                              1
                                   0
                                                0
                                        1
                                           1
                                                    1
                                          1
##
    65117
          0 1 1
                          1
                              1
                                   0
                                               0
                                                    0
    65135 0 0 1
                          2 0 4
##
                                                    0
          0 7 1
##
    6785
                          0 0 1
                                      0 1 0
                                                    0
##
    80366
            0 3
                    1
                          1
                              1
                                  1
                                      0 8 0
                                                    0
##
    87149
            6
                0
                          0
                              0 1
                                     0 2 1
                                                    0
                    1
##
    90397
# Splitting the train and test DTM
dtm_train_train <- dtm_train[1:N_t, ]</pre>
dtm_train_test <- dtm_train[(N_t+1):N, ]</pre>
dtm_train_train_labels <- as.factor(as.character(text[1:N_t, ]$Sentiment))</pre>
dtm_train_test_labels <- as.factor(as.character(text[(N_t+1):N, ]$Sentiment))</pre>
# Convert the cell values with a non-zero value to Y, and in case of a zero convert it to N,
cellconvert<- function(x) { x <- ifelse(x > 0, "Y", "N") }
```

```
# Applying the function to rows in training and test datasets
dtm_train_train <- apply(dtm_train_train, MARGIN = 2,cellconvert)</pre>
dtm_train_test <- apply(dtm_train_test, MARGIN = 2,cellconvert)</pre>
Training model
# Training the naive bayes classifier on the training dtm
library(e1071)
nb_senti_classifier=naiveBayes(dtm_train_train,dtm_train_train_labels)
# Printing the summary of the model created
summary(nb_senti_classifier)
##
            Length Class Mode
## apriori
             2 table numeric
## tables
             645
                   -none- list
## levels
            2
                   -none- character
## isnumeric 645 -none-logical
## call
              3
                   -none- call
# Making predictions on the test data dtm
nb_predicts <- predict(nb_senti_classifier, dtm_train_test,type="class")</pre>
# Computing accuracy of the model
library(rminer)
print(mmetric(nb_predicts, dtm_train_test_labels, c("ACC")))
## [1] 81.19
pretrained word2vec word embedding
# Including the required library
library(softmaxreg)
##
```

```
# Including the required library
library(softmaxreg)

##
## Attaching package: 'softmaxreg'

## The following object is masked from 'package:e1071':
##
## sigmoid

# Importing the word2vec pretrained vector into memory
data(word2vec)
dim(word2vec)
```

## [1] 12853

21

```
# Function to get word vector for each review
docVectors = function(x) { wordEmbed(x, word2vec, meanVec = TRUE) }
text = read.csv(file='Sentiment Analysis Dataset.csv', header = TRUE)
# Applying the docVector function on each of the reviews
# Storing the matrix of word vectors as temp
temp=t(sapply(text$SentimentText, docVectors))
dim(temp)
## [1] 100000
                  20
# Splitting the dataset into train and test
temp_train=temp[1:N_t,]
temp_test=temp[(N_t+1):N,]
labels_train=as.factor(as.character(text[1:N_t,]$Sentiment))
labels_test=as.factor(as.character(text[(N_t+1):N,]$Sentiment))
library(randomForest)
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
# Training a model using random forest classifier with training dataset
# Observe that we are using 20 trees to create the model
rf_senti_classifier=randomForest(temp_train, labels_train,ntree=20)
print(rf_senti_classifier)
##
## randomForest(x = temp_train, y = labels_train, ntree = 20)
##
                  Type of random forest: classification
##
                        Number of trees: 20
## No. of variables tried at each split: 4
           OOB estimate of error rate: 39.82%
## Confusion matrix:
        1
              2 class.error
## 1 23699 15284 0.3920683
## 2 16570 24441 0.4040379
# Making predictions on the dataset
rf_predicts<-predict(rf_senti_classifier, temp_test)</pre>
library(rminer)
print(mmetric(rf_predicts, labels_test, c("ACC")))
```

GloVe word embedding

## [1] 62.875

```
# Including the required library
library(text2vec)
##
## Attaching package: 'text2vec'
## The following object is masked from 'package:rminer':
##
##
       fit
# Reading the dataset
text = read.csv(file='Sentiment Analysis Dataset.csv', header = TRUE)
# Subsetting only the review text so as to create Glove word embedding
wiki = as.character(text$SentimentText)
# Create iterator over tokens
tokens = space_tokenizer(wiki)
# Create vocabulary. Terms will be uniquams (simple words).
it = itoken(tokens, progressbar = FALSE)
vocab = create_vocabulary(it)
# Consider a term in the vocabulary if and only if the term has appeared at least
# three times in the dataset
vocab = prune vocabulary(vocab, term count min = 3L)
# Use the filtered vocabulary
vectorizer = vocab_vectorizer(vocab)
# Use window of 5 for context words and create a term co-occurance matrix
tcm = create_tcm(it, vectorizer, skip_grams_window = 5L)
# Create the glove embedding for each in the vocab and
# the dimension of the word embedding should set to 50
# x_max is the maximum number of co-occurrences to use in the weighting function
glove = GlobalVectors$new(rank = 50, x_max = 100)
wv_main = glove$fit_transform(tcm, n_iter = 10, convergence_tol = 0.01)
## INFO [14:06:03.185] epoch 1, loss 0.0502
## INFO [14:06:08.706] epoch 2, loss 0.0317
## INFO [14:06:14.236] epoch 3, loss 0.0266
## INFO [14:06:19.757] epoch 4, loss 0.0239
## INFO [14:06:25.376] epoch 5, loss 0.0221
## INFO [14:06:30.890] epoch 6, loss 0.0209
## INFO [14:06:36.293] epoch 7, loss 0.0199
## INFO [14:06:41.729] epoch 8, loss 0.0191
## INFO [14:06:47.231] epoch 9, loss 0.0185
## INFO [14:06:52.696] epoch 10, loss 0.0179
# Glove model learns two sets of word vectors - main and context.
# Both matrices may be added to get the combined word vector
wv_context = glove$components
word vectors = wv main + t(wv context)
# Converting the word_vector to a dataframe for visualization
word vectors = data.frame(word vectors)
# The word for each embedding is set as row name by default
# Using the tibble library rownames_to_column function, the rownames is copied
```

```
# as first column of the dataframe
# We also name the first column of the dataframe as words
library(tibble)
word_vectors=rownames_to_column(word_vectors, var = "words")
library(softmaxreg)
docVectors = function(x) { wordEmbed(x, word_vectors, meanVec = TRUE) }
# Applying the function docVectors function on the entire reviews dataset
# This will result in word embedding representation of the entire reviews dataset
temp = t(sapply(text$SentimentText, docVectors))
# Splitting the dataset into train and test portions
temp_train=temp[1:N_t,]
temp_test=temp[(N_t+1):N,]
labels_train=as.factor(as.character(text[1:N_t,]$Sentiment))
labels_test=as.factor(as.character(text[(N_t+1):N,]$Sentiment))
# Using randomforest to build a model on train data
library(randomForest)
rf_senti_classifier=randomForest(temp_train, labels_train, ntree=20)
print(rf_senti_classifier)
##
## Call:
## randomForest(x = temp_train, y = labels_train, ntree = 20)
##
                  Type of random forest: classification
                        Number of trees: 20
## No. of variables tried at each split: 7
##
           OOB estimate of error rate: 31.77%
## Confusion matrix:
              2 class.error
        1
## 1 26911 12070 0.3096380
## 2 13340 27671 0.3252786
# Predicting labels using the randomforest model created
rf_predicts<-predict(rf_senti_classifier, temp_test)</pre>
# Estimating the accuracy from the predictions
library(rminer)
print(mmetric(rf_predicts, labels_test, c("ACC")))
## [1] 71.995
fastText word embedding
library(fastTextR)
# Input reviews file
text = readLines("Sentiment Analysis Dataset_ft.txt")
```

```
# Dividing the reviews into training and test
temp_train=text[1:N_t]
temp test=text[(N t+1):N]
# Creating txt file for train and test dataset
fileConn<-file("train.ft.txt")</pre>
writeLines(temp_train, fileConn)
close(fileConn)
fileConn<-file("test.ft.txt")</pre>
writeLines(temp_test, fileConn)
close(fileConn)
# Creating a test file with no labels
temp_test_nolabel<- gsub("__label__1", "", temp_test, perl=TRUE)</pre>
temp_test_nolabel<- gsub("__label__2", "", temp_test_nolabel, perl=TRUE)</pre>
fileConn<-file("test_nolabel.ft.txt")</pre>
writeLines(temp_test_nolabel, fileConn)
close(fileConn)
# training a supervised classification model with training dataset file
model<-ft_train("train.ft.txt", method = "supervised", control = ft_control(nthreads = 3L, seed = 1))</pre>
# Obtain all the words from a previously trained model
words<-ft_words(model)</pre>
# Obtain word vectors from a previously trained model.
word_vec<-ft_word_vectors(model, words)</pre>
# Predicting the labels for the reviews in the no labels test dataset
# Getting the predictions into a dataframe so as to compute performance measurement
ft_preds<-ft_predict(model, newdata = temp_test_nolabel)</pre>
# Reading the test file to extract the actual labels
reviewstestfile<- readLines("test.ft.txt")</pre>
# Extracting just the labels frm each line
library(stringi)
actlabels<-stri_extract_first(reviewstestfile, regex="\\w+")</pre>
# Converting the actual labels and predicted labels into factors
actlabels<-as.factor(as.character(actlabels))</pre>
ft_preds<-as.factor(as.character(ft_preds$label))</pre>
# Getting the estimate of the accuracy
library(rminer)
print(mmetric(actlabels, ft_preds, c("ACC")))
```

# Drug Data

## [1] 86.47

```
set.seed(1)
N_Drug = 146942
reviews_text_Drug<-readLines("Drug Train.csv", n = N_Drug)
reviews_text_Drug<-data.frame(reviews_text_Drug)</pre>
```

```
library(tidyr)
reviews_text_Drug<-separate(data = reviews_text_Drug, col = reviews_text_Drug, into = c("Sentiment", "S
reviews_text_Drug<-reviews_text_Drug[-1,]</pre>
N_Drug = N_Drug - 1
# Retaining only alphanumeric values in the sentiment column
reviews_text_Drug$Sentiment<-gsub("[^[:alnum:]]","",reviews_text_Drug$Sentiment)
# Retaining only alphanumeric values in the sentiment text
reviews_text_Drug$SentimentText<-gsub("[^[:alnum:]]"," ",reviews_text_Drug$SentimentText)
# Replacing multiple spaces in the text with single space
reviews_text_Drug$SentimentText<-gsub("(?<=[\\s])\\s*|^\\s+$", "", reviews_text_Drug$SentimentText
# Balance our data
minlabel <- names (which (table (reviews_text_Drug$Sentiment) == min(table (reviews_text_Drug$Sentiment))))
maxlabel<-names(which(table(reviews_text_Drug$Sentiment)==max(table(reviews_text_Drug$Sentiment))))
n_maxlabel<-min(table(reviews_text_Drug$Sentiment))</pre>
minlabelid<-c(1:N_Drug)[reviews_text_Drug$Sentiment==minlabel]</pre>
maxlabelid <- sample (c(1:N_Drug) [reviews_text_Drug$Sentiment==maxlabel], n_maxlabel)
balanceid<-sample(c(minlabelid,maxlabelid))</pre>
reviews_text_Drug<-reviews_text_Drug[balanceid,]</pre>
N Drug = nrow(reviews text Drug)
N_train_Drug = round(0.8*N_Drug)
# Writing the output to a file that can be consumed in other projects
write.table(reviews_text_Drug,file = "Sentiment Analysis Dataset_Drug.csv",row.names = F,col.names = T,
# reading the first 1000 reviews from the dataset
reviews_text_Drug<-readLines("Drug Train.csv", n = 146942)</pre>
reviews_text_Drug<-reviews_text_Drug[-1]</pre>
reviews_text_Drug<-reviews_text_Drug[balanceid]</pre>
# basic EDA to confirm that the data is read correctly
print(class(reviews_text_Drug))
## [1] "character"
print(length(reviews_text_Drug))
## [1] 80150
print(head(reviews_text_Drug,2))
## [1] "\"2\",\"Clarithromycin Sinusitis Works well very tired after course of clarithyromycin\""
## [2] "\"2\",\"Mesalamine Ulcerative Colitis Active I was diagnosed with ulcerative colitis a year ago
# replacing the positive sentiment value 2 with __label__2
reviews_text_Drug<-gsub("\\\",",","__label__2 ",reviews_text_Drug)</pre>
# replacing the negative sentiment value 1 with __label__1
```

reviews\_text\_Drug<-gsub("\\\",",","\_label\_1 ",reviews\_text\_Drug)</pre>

```
# removing the unnecessary \" characters
reviews_text_Drug<-gsub("\\\""," ",reviews_text_Drug)</pre>
# replacing multiple spaces in the text with single space
reviews_text_Drug<-gsub("(?<=[\\s])\\s*|^\\s+\\, "", reviews_text_Drug, perl=TRUE)
# Basic EDA post the required processing to confirm input is as desired
print("EDA POST PROCESSING")
## [1] "EDA POST PROCESSING"
print(class(reviews_text_Drug))
## [1] "character"
print(length(reviews_text_Drug))
## [1] 80150
print(head(reviews_text_Drug,2))
## [1] "__label__2 Clarithromycin Sinusitis Works well very tired after course of clarithyromycin"
## [2] " label 2 Mesalamine Ulcerative Colitis Active I was diagnosed with ulcerative colitis a year
# writing the revamped file to the directory so we could use it with
# fastText sentiment analyzer project
fileConn<-file("Sentiment Analysis Dataset_ft_Drug.txt")</pre>
writeLines(reviews_text_Drug, fileConn)
close(fileConn)
BoW approach
# including the required libraries
library(SnowballC)
library(tm)
# setting the working directory where the text reviews dataset is located
# reading the transformed file as a dataframe
text_Drug <- read.table(file='Sentiment Analysis Dataset_Drug.csv', sep=',', header = TRUE)</pre>
# checking the dataframe to confirm everything is in tact
print(dim(text_Drug))
```

```
## [1] 80150 2

# View(text)

# transforming the text into volatile corpus
train_corp_Drug = VCorpus(VectorSource(text_Drug$SentimentText))
print(train_corp_Drug)
```

```
## <<VCorpus>>
## Metadata: corpus specific: 0, document level (indexed): 0
## Content: documents: 80150
# creating document term matrix
dtm_train_Drug <- DocumentTermMatrix(train_corp_Drug, control = list( tolower = TRUE,removeNumbers = TR</pre>
# Basic EDA on dtm
inspect(dtm_train_Drug)
## <<DocumentTermMatrix (documents: 80150, terms: 44610)>>
## Non-/sparse entries: 2919201/3572572299
## Sparsity
                   : 100%
## Maximal term length: 95
## Weighting
                   : term frequency (tf)
## Sample
                     :
##
         Terms
          day effect get month pain start take week work year
## Docs
##
                  7
                      0
                                                0
    14443
          0
                            0
                                0
                                       1
                                           1
    21739 3
##
                  4
                      3
                                10
                                       4
                                           3
                                                5
                                                     6
                                                          1
                            1
    32948 9
                                                         5
##
                                1
                                       3
                                           5
                                                1
                                                     1
                  1
                      1
                            1
          7
                  7
                                                4
##
    35157
                      4
                            1
                                 3
                                      1
                                           6
                                                         0
##
    39889 0
                  2
                      2
                                 3
                                           1
                                                0
                            4
                                                         1
##
    4810
           0
                  7
                      0
                            0
                                 0
                                      1
                                           1
                                                0
                                                         0
##
    48674 7
                  7
                                3
                      4
                            1
                                           6
                                                4
                                                    4
                                                         0
                                         9
                                                6
##
    50714 7
                  5
                    3
                          2
                                 0
                                                    2
                                                         1
                  2 2
##
    56489 0
                                 3
                                                         1
    79862
                  0
                                                         2
##
          6
                            0
# Removing sparse terms
dtm_train_Drug = removeSparseTerms(dtm_train_Drug, 0.99)
inspect(dtm_train_Drug)
## <<DocumentTermMatrix (documents: 80150, terms: 645)>>
## Non-/sparse entries: 2177799/49518951
## Sparsity
                   : 96%
## Maximal term length: 14
## Weighting
                   : term frequency (tf)
## Sample
         Terms
##
## Docs
          day effect get month pain start take week work year
##
                  3
                      4
                            0
                                           3
                                                3
                                                     1
    1194
            2
                                 0
                                       1
##
    21739 3
                  4
                      3
                            1
                                10
                                           3
                                                5
                                                         1
          9
    32948
                      1
                                           5
                                                1
                                                     1
                                                         5
##
                  1
                            1
                                1
                                      3
##
    35157
          7
                  7
                      4
                            1
                                 3
                                      1
                                           6
                                                4
                                                     4
                                                         0
##
    35179
          2
                  3
                      4
                            0
                                 0
                                           3
                                                         0
##
    39889 0
                  2
                      2
                            4
                                 3
                                      4
                                           1
                                                0
                                                     2
                                                         1
##
    48674 7
                  7
                                3
                                           6
                                                         0
                      4
                            1
                                                4
                 5 3
                                         9
##
    50714 7
                           2
                                0
                                                6
                                                    2
                                                         1
                  2 2
                           4 3
##
    56489
                                      4 1
                                                         1
```

##

0 2

0 7

```
# splitting the train and test DTM
dtm_train_train_Drug <- dtm_train_Drug[1:N_train_Drug, ]</pre>
dtm_train_test_Drug <- dtm_train_Drug[(N_train_Drug+1):N_Drug, ]</pre>
dim(dtm train Drug)
## [1] 80150
            645
dtm_train_train_Drug_labels <- as.factor(as.character(text_Drug[1:N_train_Drug, ]$Sentiment))
dtm_train_test_Drug_labels <- as.factor(as.character(text_Drug[(N_train_Drug+1):N_Drug, ]$Sentiment))
cellconvert<- function(x) { x <- ifelse(x > 0, "Y", "N") }
# applying the function to rows in training and test datasets
dtm_train_train_Drug <- apply(dtm_train_train_Drug, MARGIN = 2,cellconvert)</pre>
dtm_train_test_Drug <- apply(dtm_train_test_Drug, MARGIN = 2,cellconvert)</pre>
# inspecting the train dtm to confirm all is in tact
# View(dtm_train_train)
training model
# training the naive bayes classifier on the training dtm
library(e1071)
nb_senti_classifier_Drug=naiveBayes(dtm_train_train_Drug,dtm_train_train_Drug_labels)
# printing the summary of the model created
summary(nb_senti_classifier_Drug)
##
          Length Class Mode
## apriori
           2 table numeric
## tables
          645
               -none- list
          2
## levels
                -none- character
## isnumeric 645 -none- logical
## call 3 -none- call
# making predictions on the test data dtm
nb_predicts_Drug <- predict(nb_senti_classifier_Drug, dtm_train_test_Drug,type="class")</pre>
# printing the predictions from the model
print(nb predicts Drug)
##
     ##
     ##
     [73] \ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 1
##
    [145] 2 2 2 2 1 2 2 1 2 2 1 2 2 1 1 1 2 2 1 2 2 2 2 2 2 2 1 1 1 2 1 2 2 1 1 1
##
    ##
##
    ##
    [253] 2 1 1 2 2 2 1 2 1 1 1 2 2 2 2 2 2 2 1 1 2 2 1 1 1 2 2 2 2 2 2 1 1 1 1
    [289] 2 1 2 2 2 1 2 2 2 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 2 1 2 1 1 1
##
##
    [361] 1 1 1 1 2 2 2 2 2 2 2 1 1 1 2 2 1 1 1 2 2 1 2 1 2 1 1 1 2 2 2 2 2 2 2 1 1 2 2 1
##
```

 $[2377] \ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1$ ## ## ## ## [2521] 2 1 1 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 2 1 2 1 ## [2557] 2 2 2 1 1 2 2 1 1 2 1 1 2 2 1 2 2 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 ## [2593] 1 2 2 1 2 1 2 2 2 2 2 2 2 1 2 1 1 2 1 2 2 2 2 2 2 1 2 1 2 2 1 2 1 1 ## ## ## ## ## ## [2809] 2 2 2 2 1 2 1 1 2 2 1 2 2 1 1 2 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 1 2 1 1 1 ##  $[2845] \ 1\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 2\ 2$ ## ##  $[2917] \ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 1$ ## ## [2953] 2 2 2 2 2 1 1 2 2 2 2 1 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 1 1 1 2 2 ## ## ##  $[3061] \ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 1$ ## ## [3169] 1 2 1 1 1 1 2 2 2 1 2 1 2 2 2 2 2 2 1 1 1 1 1 2 2 2 1 2 1 1 2 1 1 2 1 1 ## ## ## ## ##  $[3349] \ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 2$ ## ## ## ## ##  $[3529] \ 1\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1$ ## ## [3565] 2 2 1 1 2 2 1 1 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 1 1 1 1 1 2 1 2 2 1 2 1 2 [3601] 1 2 2 2 2 1 2 2 2 1 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 ## ## ## ##  $[3745] \ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1$ ## ## ## [3853] 1 1 1 2 2 1 1 2 2 1 1 2 2 2 1 1 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 1 2 1 2 1 1 2 2 2 1 1 2 ## ## ##  $[3925] \ 2\ 2\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 1$ ## ## [3997] 2 1 2 1 2 2 1 2 2 1 1 2 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 2 2 2 2 1 1 2 2 2 [4033] 1 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 2 2 1 1 1 2 1 2 1 2 1 1 1 2 2 2 1 1 1 2 ## ## ##  $[4141] \ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2$ ## ## ## [4213] 2 1 2 1 2 2 1 2 2 2 1 1 2 1 1 2 1 1 2 2 1 1 2 1 2 1 2 1 2 2 2 1 2 2 2 1 2 2 1 ## 

## [6229] 2 2 2 2 2 2 1 2 1 2 1 1 2 2 2 1 2 2 1 2 2 1 1 1 2 1 2 2 2 1 2 1 1 ## ##  $[6337] \ 2\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2$ ## ##  $[6373] \ 1\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2$  $[6409] \ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 1$ ## [6445] 1 2 1 1 1 2 2 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 1 1 2 2 1 2 2 1 2 2 1 ## ## ## ## ## ## ## ## ## ##  $[6769] \ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1$  $[6805] \ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 1$ ## ##  $[6841] \ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 2$  $[6877] \ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 1$ ## ## ## ## ## [7057] 1 2 2 2 2 1 2 1 1 2 1 2 1 2 2 2 2 2 1 1 2 1 1 1 2 2 2 1 2 1 2 1 2 1 2 1 ## ## ## ## ## ## ## ## ## ##  $[7381] \ 2\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 1$ ## ## [7453] 2 1 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 1 2 2 2 1 2 1 2 2 2 1 1 1 2 2 2 1 2 2 ## ## ## ## ## ##  $[7669] \ 2\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1$ ## ## [7741] 2 1 2 1 2 2 2 1 1 1 2 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 1 2 1 2 1 2 1 2 2 1 1 1 2 ## ## ##  $[7885] \ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2$ ## ## ## ## ## ## ##  $[8137] \ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1$ ##

##  $[8173] \ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1$ ## ##  $[8281] \ 2\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 1$ ## ##  $[8353] \ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2$ ## [8389] 1 1 2 2 2 1 2 2 1 1 1 1 1 2 2 2 1 1 2 1 2 1 2 2 2 1 1 2 1 2 2 2 1 1 2 1 1 1 1 ##  $[8425] \ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2$ ## ## [8461] 2 1 2 2 2 2 1 1 1 1 1 1 2 1 2 2 1 2 2 1 2 1 1 1 1 2 1 2 1 2 2 2 1 1 1 1 ## ## ## ## ## ##  $[8677] \ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1$ ##  $[8713] \ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 1\ 2\ 1\ 1\ 1\ 2$ ## ## ## ##  $[8857] \ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2$ ## ## ## ## ## ## ## ## ## ## [9217] 1 2 1 2 2 2 1 1 2 2 2 1 1 2 1 1 1 2 2 1 1 1 1 1 2 2 1 2 1 2 2 2 2 2 2 2 ## ##  $[9289] \ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2$ ## ## ## [9433] 1 2 2 1 1 1 2 2 2 1 1 1 2 1 1 1 2 2 2 2 2 1 1 2 2 1 1 2 1 1 2 1 1 2 2 1 ## ## ##  $[9505] \ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1$  $[9541] \ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 1$ ##  $[9577] \ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 2$ ## ## ## ## [9685] 2 2 1 2 1 2 1 2 1 2 1 2 2 1 2 1 2 2 1 2 1 1 1 1 1 2 2 1 2 2 1 2 2 1 1 1 1 1 2 ## ## ## ## ## ## ##  $\begin{smallmatrix} 9973 \end{smallmatrix} ] \ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 2\ 2$ ## ## [10045] 2 2 2 2 2 2 1 2 2 1 2 1 1 2 1 1 2 2 2 2 1 2 2 1 1 1 1 2 2 2 1 1 1 1 2 2 2 1 2 1 ## [10081] 2 2 2 1 2 2 2 2 1 1 2 1 2 2 2 1 1 2 1 1 2 2 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 2 2 2

```
## [10117] 1 1 2 2 1 1 1 2 2 1 2 2 1 2 2 1 1 1 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 2
## [10225] 2 1 2 1 2 2 1 2 2 2 2 1 1 1 1 1 2 1 1 1 1 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2
## [10261] 1 2 1 2 1 1 2 2 1 2 2 1 1 2 2 1 1 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 1 1 2 2
## [10333] 2 1 2 2 2 2 2 1 2 1 1 1 2 1 1 1 2 2 1 1 1 1 2 2 1 1 1 2 2 1 1 1 2 1 1 1 2 1 1 1
## [10405] 1 1 1 1 1 2 1 1 2 2 1 1 1 2 1 1 1 2 1 2 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 2 1 1
## [10441] 2 2 2 2 1 2 2 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 1 1 1 1 1 2 1 1 1
## [10477] 1 2 1 2 2 1 1 1 2 1 2 1 2 1 2 2 2 1 1 2 1 2 1 2 1 2 1 1 1 1 1 2 1 2 2 2 1 2 2
## [10513] 1 1 2 1 2 1 1 1 1 2 1 2 1 2 1 1 1 2 2 2 1 1 2 1 2 2 1 1 2 1 2 1 2 2 2
## [10549] 2 2 2 2 2 2 2 1 2 1 1 2 2 1 1 2 1 1 2 2 2 1 1 2 2 1 1 2 2 1 2 1 2 1 2 2 2 2 2 2
## [10585] 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 2 1 1 2 1 2 1 1 1 1 2
## [10729] 2 1 2 2 2 1 1 2 2 1 2 2 2 1 1 1 2 2 1 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 2 2 1 2
## [10801] 2 1 2 1 2 2 1 2 1 2 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 1
## [10873] 2 1 1 1 2 2 1 1 2 2 2 2 2 1 1 2 1 1 2 1 1 2 2 1 1 1 2 1 2 2 1 2 2 1 2 2 1 1 2
## [10909] 2 1 2 2 2 2 2 2 1 2 1 2 1 1 2 1 1 2 2 1 1 2 2 2 1 2 1 2 1 2 1 2 1 2 2 2
## [10981] 1 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 2 2 2 2 2 1 1 2 2 1 1 2 1 1 2 2 1 2 2 1 1 1
## [11053] 2 1 1 2 2 2 1 1 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1 2 2 1
## [11089] 1 1 2 2 2 1 2 1 2 1 1 2 2 1 1 1 2 2 1 1 1 2 2 1 2 2 2 2 2 2 2 2 1 2 1 2 2 2
## [11125] 2 2 1 1 1 2 1 2 2 2 2 1 2 1 2 1 2 2 2 2 1 1 1 1 2 2 2 2 1 2 2 2 2 1
## [11161] 1 1 2 2 1 1 1 2 2 2 2 2 1 2 1 1 1 1 2 1 1 2 2 1 1 1 1 2 2 2 2 1 2 1 1 1 1
## [11233] 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1 1 2
## [11269] 1 2 1 2 1 2 1 1 1 1 1 1 2 1 2 1 2 2 2 1 2 1 2 2 2 2 1 1 2 1 2 1 2 1 2 2 2
## [11377] 2 1 1 1 2 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2 1 1 2 2 1 1 2 1 1 2 2 1 1 1 2 1 1 2 1
## [11449] 1 2 2 1 2 2 2 1 1 1 2 1 1 2 2 2 1 1 1 1 2 1 1 1 2 2 1 2 1 1 1 1 1 2 2 1 2 1 2 1 1 1 1 1
## [11485] 2 2 1 2 1 2 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 2 2 2 1 2 1 1 2 2 2 2 2 1 2 1 2 1
## [11701] 1 1 1 2 1 1 2 2 2 1 1 1 2 1 1 2 2 1 2 1 2 1 2 1 2 1 1 2 1 1 2 1 1 2 1 2 2 1 2
## [11845] 1 2 2 1 2 1 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 1 1 1 1 2 2 1 2 1 1 1 2 2 2 2 2 2 2 1 1
## [11917] 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 2 1 2 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 2
## [11953] 1 1 1 2 1 2 2 1 1 1 1 1 2 2 1 1 2 2 2 1 1 2 2 2 1 1 2 2 1 2 1 2 1 2 2 2 1 1 2 2
## [11989] 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 2 2 1 2 2 1 2 1 1 1 1 2 1 1 1 1 1 2 1 1 2 2
## [12025] 2 2 2 2 2 2 1 2 2 2 2 1 1 1 1 1 2 2 1 2 2 2 2 1 1 2 1 2 1 2 2 1
```

```
## [12061] 1 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 1 2 1 1 2 2 1 2 1 1
## [12097] 1 2 1 1 1 2 2 1 1 1 2 2 2 1 1 1 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 1 1 2 1 2 2 2 1 1
## [12133] 2 1 2 2 2 2 2 1 1 2 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 1 1 1
## [12169] 1 1 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 2 1 2 1
## [12349] 2 1 1 2 1 1 1 2 1 2 1 1 2 2 1 2 2 1 2 2 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2
## [12457] 1 2 2 1 1 1 1 1 2 2 2 1 1 2 1 2 1 2 2 2 1 2 2 1 2 2 1 2 2 2 2 2 1 1 1 2
## [12493] 1 2 2 1 1 2 1 1 1 1 2 1 2 2 1 2 2 1 2 1 2 2 2 2 2 1 1 1 2 1 2 1 2 2 2 2
## [12529] 2 1 1 1 1 2 1 2 1 1 1 1 2 2 2 1 2 2 1 2 1 2 1 2 1 2 1 1 1 2 2 1 1 2 2 1 2
## [12673] 1 1 1 1 2 2 1 1 2 2 1 1 2 1 2 2 1 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 2 2 1 2
## [12709] 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 1 1 2 2 1 2 1 2 1 2 1 2 1 2 2 2 2 1 2 1 2 2 2 2
## [12745] 2 2 2 2 2 1 1 2 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 2 1
## [12817] 2 1 1 2 2 2 1 1 1 1 2 1 1 1 1 2 2 1 1 2 2 2 2 2 1 2 1 2 1 1 1 2 1 1 1 1
## [12853] 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 1 1 1 1 1 1 1 2 2 1 2 1 2 1 1 1 2 2 1 2 1 2 1 2 1 2 1 2 1
## [12889] 1 1 1 1 1 1 2 1 2 2 1 1 2 1 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 1 1 2 2 1 1 1 2
## [12925] 2 1 1 2 1 1 2 2 1 2 2 1 1 2 2 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 1
## [12961] 1 1 1 2 2 2 1 2 2 2 2 2 1 1 1 2 2 2 1 2 1 1 1 2 2 2 2 2 2 2 2 2 2 1
## [12997] 2 1 2 2 2 2 1 1 1 1 1 2 1 2 2 1 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2
## [13105] 2 2 1 2 1 1 2 1 1 1 1 2 2 1 2 2 1 1 1 1 2 2 1 2 1 1 2 1 2 2 2 2 2 2 2 2 1
## [13141] 2 1 1 2 1 1 1 1 2 1 2 2 2 1 2 2 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 2 2 1
## [13177] 2 2 2 1 1 2 1 2 1 2 1 2 2 1 2 1 2 2 2 1 1 2 1 2 2 2 2 2 1 2 1 1 2 1 2 2 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 
## [13213] 2 1 2 1 1 1 2 2 1 1 2 2 2 1 2 1 2 2 1 2 1 2 1 2 1 1 1 1 2 1 2 2 2 2 1 1 2 1 1
## [13249] 1 2 2 2 1 1 2 2 1 2 2 2 1 1 2 2 2 1 1 2 2 2 1 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 1 1 2 2 2 1
## [13285] 2 1 2 2 1 1 1 2 1 2 2 2 2 1 2 1 1 2 1 2 2 2 2 1 1 1 2 2 2 2 1 1 1 2 2 2 2 2
## [13321] 2 2 2 1 2 2 2 2 2 2 2 2 1 2 1 2 2 2 1 1 1 2 1 1 1 1 1 2 1 1 1 2 2 1 2 1 1 2
## [13393] 1 2 2 1 2 2 1 1 1 1 1 1 1 1 2 1 2 2 1 1 1 1 2 2 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 2 1 1 2 1
## [13465] 1 1 1 1 2 2 1 2 2 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 2 2 1 1 2 1 2 2 2
## [13501] 1 2 1 2 1 2 1 1 2 1 1 2 2 2 2 2 1 2 2 2 2 2 2 1 2 1 2 1 2 1 2 1 2 2 2 2 1 2
## [13573] 1 2 1 1 2 1 2 2 1 2 2 2 1 2 2 2 1 2 2 2 1 2 1 2 1 1 1 2 1 2 1 1 2 1 2 1 2 2 1 2
## [13609] 1 2 2 2 1 1 2 2 1 1 2 1 2 2 2 2 2 1 1 1 1 1 2 2 2 1 1 1 1 2 2 2 2 1 2 1 2 1 2
## [13717] 2 1 2 2 1 1 2 2 2 1 2 1 1 1 2 1 1 2 2 2 1 2 2 1 2 2 1 2 2 2 2 2 2 1 1 1 1 1
## [13753] 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 1 1
## [13789] 2 2 1 2 1 2 2 1 2 1 1 2 2 2 2 2 2 2 1 1 1 2 1 1 1 2 2 2 1 2 1 2 1 2 1 2 1
## [13825] 1 2 1 1 2 1 2 2 2 1 2 2 1 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 1 2 1 2 2 2 2
## [13933] 2 2 2 2 1 1 1 2 1 2 2 1 2 1 2 1 2 2 2 2 1 2 2 1 1 2 1 1 1 1 1 1 2 1 2 2 2 2
```

```
## [14005] 1 1 1 1 2 1 2 2 1 2 1 1 2 2 2 1 1 2 2 2 2 2 2 1 2 1 1 2 2 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## [14077] 2 2 2 2 1 1 1 2 2 2 2 1 1 1 1 2 1 1 2 2 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2
## [14149] 1 1 1 1 2 1 2 2 1 2 1 1 2 2 2 1 1 2 2 2 2 1 1 1 1 1 1 1 1 2 1 2 2 1 1 1
## [14185] 2 1 2 2 2 2 1 1 2 2 2 2 1 2 2 2 2 1 2 1 1 1 2 2 1 2 1 1 1 1 2 2 2 2 2 2 1
## [14221] 1 1 1 2 1 1 2 1 2 1 2 1 2 2 2 2 1 1 2 1 1 1 1 2 2 2 2 2 2 1 2 1 1 2 2
## [14257] 2 2 1 2 1 2 1 2 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 2 1 1 1 1 2
## [14329] 1 2 1 2 1 1 2 2 2 1 2 2 1 1 1 2 2 1 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1
## [14401] 1 1 1 1 2 1 1 1 2 1 2 1 2 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 2 2 2 2 2 2 2 2
## [14509] 1 1 2 2 2 1 1 1 2 1 1 1 2 2 2 2 2 1 1 1 2 2 2 2 2 2 1 1 1 2 1 2 2 2 2 2 1 1 1 2 2 2 2 2 1
## [14689] 1 1 1 1 2 1 2 1 1 2 2 1 1 2 1 2 1 1 1 2 2 2 2 1 2 1 1 1 1 2 1 1 2 2 1 1
## [14761] 1 1 1 1 2 1 1 2 2 2 2 2 2 2 1 2 1 1 1 2 2 2 2 2 1 1 1 1 2 2 2 2 1 1 2 2 2
## [14797] 2 2 2 1 2 2 1 2 1 2 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 2 1 2 2 1 2 2
## [14833] 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 2 1 2 2 2 1 1 2 1 2 2 2 1 2 2 2 1 1 2 1 2 2 2 1 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 
## [14869] 1 2 2 2 1 2 1 1 2 1 2 1 1 1 1 1 2 1 1 1 1 1 2 2 2 2 2 1 1 1 2 1 1 1 2 1 1 1
## [14905] 2 1 2 1 2 1 1 1 2 1 2 2 1 1 1 2 2 1 1 1 1 2 2 1 1 2 2 1 2 2 2 2 2 1
## [14977] 2 2 1 1 2 1 2 1 2 2 2 1 2 1 2 2 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 2 2 1 1 2 1 1 2 1 1 2 1 1 1 2
## [15049] 2 1 1 2 1 2 1 2 2 1 2 2 1 1 1 2 1 1 2 1 1 2 1 2 2 2 2 2 2 1 1 1 1
## [15085] 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 1 2 1 1 1 1 2 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1
## [15193] 2 1 2 1 2 1 2 2 2 2 2 1 1 2 1 1 2 2 2 2 2 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 2 2
## [15229] 1 2 2 1 2 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2
## [15409] 2 2 2 1 1 2 1 2 2 2 2 2 1 1 2 2 1 1 1 2 2 2 2 2 2 2 1 2 2 1 2 1 1 1 2 2
## [15445] 2 2 1 1 2 2 2 1 2 2 1 2 1 1 1 1 2 2 2 2 1 2 2 2 1 1 2 1 1 2 2 2 2 1
## [15517] 2 2 1 2 1 2 1 2 2 2 1 2 2 1 2 2 2 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 2 2 2 1
## [15553] 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 1 1 2 2 1 1 1 2 2 2 1 1 2 1 2 2 2 2 1 1 2 1 2 2 2 2 1 1 2 1 1 2
## [15589] 1 2 1 2 2 1 1 2 1 2 2 2 2 1 1 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1
## [15733] 2 2 1 2 2 2 1 2 2 2 2 2 1 1 2 1 2 1 1 1 1 1 2 2 2 2 1 1 1 1 2 1 2 2 2 2 2
## [15805] 1 2 1 1 1 1 2 1 2 1 2 1 1 1 1 2 2 2 1 1 2 1 1 2 2 1 2 1 2 1 1 1 1 2 2 1 1
## [15877] 1 2 1 2 1 2 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 1 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 1 2 2 2 2 1 1 1 2
```

```
## [16021] 1 1 2 1 1 2 1 2 2 2
## Levels: 1 2
# computing accuracy of the model
library(rminer)
print(mmetric(nb_predicts_Drug, dtm_train_test_Drug_labels, c("ACC")))
## [1] 74.7723
pretrained word2vec word embedding
# including the required library
# install.packages("https://cran.r-project.org/src/contrib/Archive/softmaxreg/softmaxreg_1.2.tar.gz",re
library(softmaxreg)
# importing the word2vec pretrained vector into memory
data(word2vec)
# View(word2vec)
dim(word2vec)
## [1] 12853
              21
# function to get word vector for each review
docVectors = function(x) { wordEmbed(x, word2vec, meanVec = TRUE) }
text_Drug = read.csv(file='Sentiment Analysis Dataset_Drug.csv', header = TRUE)
# applying the docVector function on each of the reviews
# storing the matrix of word vectors as temp
temp_Drug=t(sapply(text_Drug$SentimentText, docVectors))
# visualizing the word vectors output
# View(temp)
dim(temp_Drug)
## [1] 80150
              20
# splitting the dataset into train and test
temp_train_Drug=temp_Drug[1:N_train_Drug,]
temp_test_Drug=temp_Drug[(N_train_Drug+1):N_Drug,]
labels_train_Drug=as.factor(as.character(text_Drug[1:N_train_Drug,]$Sentiment))
labels_test_Drug=as.factor(as.character(text_Drug[(N_train_Drug+1):N_Drug,]$Sentiment))
# including the random forest library
library(randomForest)
# training a model using random forest classifier with training dataset
# observe that we are using 20 trees to create the model
rf_senti_classifier_Drug=randomForest(temp_train_Drug, labels_train_Drug,ntree=20)
print(rf_senti_classifier_Drug)
##
## Call:
## randomForest(x = temp_train_Drug, y = labels_train_Drug, ntree = 20)
```

```
##
                  Type of random forest: classification
                        Number of trees: 20
##
## No. of variables tried at each split: 4
##
          OOB estimate of error rate: 31.92%
## Confusion matrix:
              2 class.error
## 1 23000 8985
                 0.2809129
## 2 11479 20649
                 0.3572896
# making predictions on the dataset
rf_predicts_Drug<-predict(rf_senti_classifier_Drug, temp_test_Drug)
library(rminer)
print(mmetric(rf_predicts_Drug, labels_test_Drug, c("ACC")))
```

### ## [1] 71.01684

### GloVe word embedding

```
# including the required library
library(text2vec)
# reading the dataset
text_Drug = read.csv(file='Sentiment Analysis Dataset_Drug.csv', header = TRUE)
# subsetting only the review text so as to create Glove word embedding
wiki_Drug = as.character(text_Drug$SentimentText)
# Create iterator over tokens
tokens Drug = space tokenizer(wiki Drug)
# Create vocabulary. Terms will be unigrams (simple words).
it_Drug = itoken(tokens_Drug, progressbar = FALSE)
vocab_Drug = create_vocabulary(it_Drug)
# consider a term in the vocabulary if and only if the term has appeared at least three times in the da
vocab_Drug = prune_vocabulary(vocab_Drug, term_count_min = 3L)
# Use the filtered vocabulary
vectorizer_Drug = vocab_vectorizer(vocab_Drug)
# use window of 5 for context words and create a term co-occurance matrix
tcm_Drug = create_tcm(it_Drug, vectorizer_Drug, skip_grams_window = 5L)
# create the glove embedding for each in the vocab and
# the dimension of the word embedding should set to 50
# x_max is the maximum number of co-occurrences to use in the weighting
# function
# note that training the word embedding is time consuming - be patient
glove = GlobalVectors$new(rank = 50, x_max = 100)
wv_main_Drug = glove$fit_transform(tcm_Drug, n_iter = 10, convergence_tol = 0.01)
## INFO [14:16:42.664] epoch 1, loss 0.0758
## INFO [14:16:44.940] epoch 2, loss 0.0489
## INFO [14:16:47.121] epoch 3, loss 0.0402
## INFO [14:16:49.328] epoch 4, loss 0.0355
## INFO [14:16:51.581] epoch 5, loss 0.0324
## INFO [14:16:53.825] epoch 6, loss 0.0302
```

## INFO [14:16:55.996] epoch 7, loss 0.0285

```
## INFO [14:16:58.241] epoch 8, loss 0.0272
## INFO [14:17:00.432] epoch 9, loss 0.0262
## INFO [14:17:02.657] epoch 10, loss 0.0253
# Glove model learns two sets of word vectors - main and context.
# both matrices may be added to get the combined word vector
wv_context = glove$components
word_vectors_Drug = wv_main_Drug + t(wv_context)
# converting the word_vector to a dataframe for visualization
word_vectors_Drug=data.frame(word_vectors_Drug)
# the word for each embedding is set as row name by default
# using the tibble library rownames_to_column function, the rownames is copied as first column of the d
# we also name the first column of the dataframe as words
library(tibble)
word_vectors_Drug=rownames_to_column(word_vectors_Drug, var = "words")
# View(word_vectors)
library(softmaxreg)
docVectors_Drug = function(x) { wordEmbed(x, word_vectors_Drug, meanVec = TRUE) }
# applying the function docVectors function on the entire reviews dataset
# this will result in word embedding representation of the entire reviews
# dataset
temp_Drug=t(sapply(text_Drug$SentimentText, docVectors_Drug))
# View(temp)
# splitting the dataset into train and test portions
temp_train_Drug=temp_Drug[1:N_train_Drug,]
temp_test_Drug=temp_Drug[(N_train_Drug+1):N_Drug,]
labels_train_Drug=as.factor(as.character(text_Drug[1:N_train_Drug,]$Sentiment))
labels_test_Drug=as.factor(as.character(text_Drug[(N_train_Drug+1):N_Drug,]$Sentiment))
# using randomforest to build a model on train data library(randomForest)
rf_senti_classifier_Drug=randomForest(temp_train_Drug, labels_train_Drug, ntree=20)
print(rf_senti_classifier_Drug)
##
## Call:
## randomForest(x = temp_train_Drug, y = labels_train_Drug, ntree = 20)
                  Type of random forest: classification
##
##
                        Number of trees: 20
## No. of variables tried at each split: 7
##
##
           OOB estimate of error rate: 29%
## Confusion matrix:
##
              2 class.error
        1
## 1 23384 8600 0.2688844
## 2 9994 22133
                 0.3110779
# predicting labels using the randomforest model created
rf_predicts_Drug<-predict(rf_senti_classifier_Drug, temp_test_Drug)
# estimating the accuracy from the predictions
library(rminer)
print(mmetric(rf_predicts_Drug, labels_test_Drug, c("ACC")))
```

## [1] 74.82845

#### fastText word embedding

```
# loading the required libary
library(fastTextR)
# reading the input reviews file
# recollect that fastText needs the file in a specific format and we created one compatiable file in
# "Understanding the Amazon Reviews Dataset" section of this chapter
text_Drug = readLines("Sentiment Analysis Dataset_ft_Drug.txt")
# Viewing the text vector for conformation
# View(text)
# dividing the reviews into training and test
temp_train_Drug=text_Drug[1:N_train_Drug]
temp_test_Drug=text_Drug[(N_train_Drug+1):N_Drug]
# Viewing the train datasets for confirmation
# View(temp_train)
# creating txt file for train and test dataset
# the fasttext function expects files to be passed for training and testing
fileConn<-file("train_Drug.ft.txt")</pre>
writeLines(temp_train_Drug, fileConn)
close(fileConn)
fileConn<-file("test_Drug.ft.txt")</pre>
writeLines(temp_test_Drug, fileConn)
close(fileConn)
# creating a test file with no labels
# recollect the original test dataset has labels in it
# as the dataset is just a subset obtained from full dataset
temp_test_Drug_nolabel<- gsub("__label__1", "", temp_test_Drug, perl=TRUE)</pre>
temp_test_Drug_nolabel<- gsub("__label__2", "", temp_test_Drug_nolabel, perl=TRUE)</pre>
# View(temp_test_nolabel)
fileConn<-file("test_Drug_nolabel.ft.txt")</pre>
writeLines(temp_test_Drug_nolabel, fileConn)
close(fileConn)
# training a supervised classification model with training dataset file
model_Drug<-ft_train("train_Drug.ft.txt", method = "supervised", control = ft_control(nthreads = 3L, se</pre>
# Obtain all the words from a previously trained model=
words_Drug<-ft_words(model_Drug)</pre>
# viewing the words for confirmation. These are the set of words present in our training data
# View(words)
# Obtain word vectors from a previously trained model.
word_vec_Drug<-ft_word_vectors(model_Drug, words_Drug)</pre>
# Viewing the word vectors for each word in our training dataset
# observe that the word embedding dimension is 5
# View(word_vec)
# predicting the labels for the reviews in the no labels test dataset
# getting the predictions into a dataframe so as to compute performance measurement
ft_preds_Drug<-ft_predict(model_Drug, newdata = temp_test_Drug_nolabel)</pre>
# reading the test file to extract the actual labels
```

```
reviewstestfile_Drug<- readLines("test_Drug.ft.txt")
# extracting just the labels frm each line
library(stringi)
actlabels_Drug<-stri_extract_first(reviewstestfile_Drug, regex="\\w+")
# converting the actual labels and predicted labels into factors
actlabels_Drug<-as.factor(as.character(actlabels_Drug))
ft_preds_Drug<-as.factor(as.character(ft_preds_Drug$label))
# getting the estimate of the accuracy
library(rminer)
print(mmetric(actlabels_Drug, ft_preds_Drug, c("ACC")))</pre>
```

## [1] 78.67748