AmrithRavindraHW3.R

Chanti

Sat Apr 8 21:54:21 2017

```
x= getwd()
setwd(x)
library(rpart)
library(rpart.plot)

#######
hw3 <- read.csv("hw3.csv", stringsAsFactors = FALSE)
str(hw3)

## 'data.frame':    3882 obs. of  7 variables:
##  $ X.1     : chr  "1" "2" "3" "4" ...
##  $ X       : chr  "1" "2" "3" "4" ...
##  $ Vandal  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minor   : int  1 1 0 1 1 0 0 0 1 0 ...
##  $ Loggedin: int  1 1 1 0 1 1 1 1 1 0 ...
##  $ Added   : chr  "  represent psycholinguisticspsycholinguistics orthogra
phyorthography help text all actions through human ethnologue relationsh"| __
truncated__ " website external links" " " " afghanistan used iran mostly that
farsiis is countries some xmlspacepreservepersian parts tajikestan region" ..
.
##  $ Removed : chr  " " " talklanguagetalk" " regarded as technologytechnolo
gies human first" "  represent psycholinguisticspsycholinguistics orthography
orthography help all actions through ethnologue relationships linguis"| __tru
ncated__ ...

table(hw3$Vandal) #This tells us that there were 1815 recorded cases of vanda
lism

##
##    0    1
## 2061 1815

#######
library(tm)

## Loading required package: NLP

addcorpus <- Corpus(VectorSource(hw3$Added))
addcorpus <- tm_map(addcorpus, removeWords, stopwords("english"))
addcorpus <- tm_map(addcorpus, stemDocument)
adddoc <- DocumentTermMatrix(addcorpus)
adddoc #This tells us that our document term matrix contains 3882 documents a
nd 6675 terms
```

```
## <<DocumentTermMatrix (documents: 3882, terms: 6675)>>
## Non-/sparse entries: 15368/25896982
## Sparsity           : 100%
## Maximal term length: 784
## Weighting          : term frequency (tf)
```

```
sparseAdded <- removeSparseTerms(adddoc, 0.3)
sparseAdded
```

```
## <<DocumentTermMatrix (documents: 3882, terms: 0)>>
## Non-/sparse entries: 0/0
## Sparsity           : 100%
## Maximal term length: 0
## Weighting          : term frequency (tf)
```

```
wordsAdded <- as.data.frame(as.matrix(sparseAdded))

#Repeating all the steps again
removecorpus<- Corpus(VectorSource(hw3$Removed))
removecorpus <- tm_map(removecorpus, removeWords, stopwords("english"))
removecorpus <- tm_map(removecorpus, stemDocument)
removedoc <- DocumentTermMatrix(removecorpus)
sparseRemoved <- removeSparseTerms(removedoc, 0.3)
wordsRemoved <- as.data.frame(as.matrix(sparseRemoved))

#Combining both the dataframes
wikiWords <- cbind(wordsAdded, wordsRemoved)

#Adding the vandal column
wikiWords$Vandal <- hw3$Vandal
library(caTools)

#Splitting the data into testing and training sets
set.seed(123)
split <- sample.split(wikiWords$Vandal, SplitRatio = 0.7)
train <- subset(wikiWords, split == TRUE)
test <- subset(wikiWords, split == FALSE)
table(test$Vandal)
```

```
## 
##   0   1 
## 618 545
```

```
#Building the CART Model

#CART <- rpart(Vandal~.,data = train,method = "class", parms = list(split="gi
ni"))
```

However,when I use the following code I am getting different results.

You have to copy and paste the code as it is and run it in R to see what I mean. Only when I use 0,99% am I able to obtain 15 terms from the document term matrix which are not sparse. If you run the following code you will understand what I mean.

```
x = getwd()

setwd(x)

library(rpart)

library(rpart.plot)


#Code to read data and count number of cases of vandalism detected


vdata = read.csv(file = "hw3.csv", header = T, check.names = T, na.strings = "", strip.white = T)

colnames(vdata)

vcount <- subset(vdata, vdata$Vandal == 1)

nrow(vcount) #This tells us there were 1815 counts of vandalism detected


#Preprocessing of text data and creating a corpus from the 'Added' column

library(tm)

library(NLP)

library(SnowballC)


added = vdata[,c(6)]

added  = as.data.frame(added)

addedNONA = as.data.frame(added[complete.cases(added),])

myCorpus<- Corpus(DataframeSource(addedNONA))

getTransformations()
```

```
myCorpus = tm_map(myCorpus, tolower)
myCorpus = tm_map(myCorpus, removeNumbers)
myCorpus = tm_map(myCorpus, removePunctuation)
myCorpus = tm_map(myCorpus, removeWords, stopwords("english"))
myCorpus = tm_map(myCorpus, stemDocument)
myCorpus = tm_map(myCorpus, stripWhitespace)
myCorpus = tm_map(myCorpus, PlainTextDocument)

test = myCorpus
length(test) #This tells us that 2395 documents were finally added to the corpus after preprocessing

#Creating a Document Term Matrix and filtering out sparse terms

tdm <- DocumentTermMatrix(test)
inspect(tdm) #This tells us there are 2395 documents and 6336 terms in the document term matrix
tm <- as.matrix(tdm)
length(tm)

notSparse = removeSparseTerms(tdm, 0.99) #Here I realized that chosing a value less than 0.99 always leaves me with no terms to inspect
inspect(notSparse) #This tells us there are 15 terms in 2395 documents which are not sparse
sparseAdded <- as.data.frame(as.matrix(notSparse))
View(sparseAdded)
wordsAdded <- as.data.frame(as.matrix(sparseAdded))
```

```
#### Repeating the steps again ####


removecorpus <- Corpus(DataframeSource(addedNONA))

removecorpus <- tm_map(removecorpus, removeWords, stopwords("english"))

removecorpus <- tm_map(removecorpus, stemDocument)

removedoc <- DocumentTermMatrix(removecorpus)

sparseRemoved <- removeSparseTerms(removedoc, 0.99)

wordsRemoved <- as.data.frame(as.matrix(sparseRemoved))


View(wordsAdded)

View(wordsRemoved)


#Creating wikiWords

wikiWords <- cbind(wordsAdded, wordsRemoved)


#Adding the vandal column

wikiWords2 <- cbind(wordsAdded, wordsRemoved, hw3$Vandal)

wikiWords$Vandal <- vdata$Vandal


library(caTools)


#Splitting the data into testing and training sets

set.seed(123)

split <- sample.split(vdata$Vandal, SplitRatio = 0.7)

train <- subset(wikiWords, split == TRUE)

test <- subset(wikiWords, split == FALSE)

table(test$Vandal)
```

```
#Building the CART Model


#CART <- rpart(Vandal~.,data = train,method = "class", parms = list(split="gini")
```

I also experienced a few errors and was unable to solve it completely but I did give it a hard try and I’m still working on it hoping to crack it completely. Meanwhile I am submitting this version just to make sure I don’t miss the deadline.