oncf.R

nisrin

Wed Jan 09 23:42:32 2019

#Extract database from Twitter, make text analytics on tweets  
  
echo=TRUE  
eval=TRUE

#install.packages("tm")  
library(tm)

## Warning: package 'tm' was built under R version 3.5.2

## Loading required package: NLP

#install.packages("SnowballC")  
library(SnowballC)

## Warning: package 'SnowballC' was built under R version 3.5.2

#install.packages("caTools")  
library(caTools)  
#install.packages("rpart")  
library(rpart)  
#install.packages("rpart.plot")  
library(rpart.plot)

#export data  
oncf = read.csv("ONCF3.csv", header = TRUE, sep = ";" )  
  
#view data   
str(oncf)

## 'data.frame': 105 obs. of 13 variables:  
## $ Number\_of\_Records : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Created\_Time : Factor w/ 102 levels "16/10/2018 11:48",..: 102 101 100 99 98 97 96 95 94 93 ...  
## $ Filter : Factor w/ 12 levels "Aden","Clarendon",..: 2 9 9 5 9 2 9 9 9 9 ...  
## $ Image\_URL : Factor w/ 105 levels "https://scontent.cdninstagram.com/vp/0382098c077234d5e4ef523eec2c5f3b/5C7DC701/t51.2885-15/e35/s320x320/4307934"| \_\_truncated\_\_,..: 9 103 11 101 96 75 38 88 13 35 ...  
## $ latitude : Factor w/ 12 levels "0","22,59372606",..: 1 1 3 1 1 12 6 5 1 1 ...  
## $ Number\_of\_likes : int 23 49 36 47 5685 27 10 13 9 2 ...  
## $ Link : Factor w/ 105 levels "https://www.instagram.com/\_ibrahiim\_naji/p/BpCoflGnd8Z/",..: 103 104 102 101 100 99 98 97 96 94 ...  
## $ Location : Factor w/ 12 levels "","Bou Knadel, Rabat-Sale, Morocco",..: 1 1 8 1 1 12 9 7 1 1 ...  
## $ longitude : Factor w/ 12 levels "-5,0032","-5,8195",..: 11 11 9 11 11 2 3 8 11 11 ...  
## $ number\_of\_Comments: int 4 0 3 3 365 0 0 5 0 0 ...  
## $ Tags : Factor w/ 96 levels "","???,morocco,????,oncf,rabatagdal,dosaindia",..: 36 32 67 2 1 79 54 60 64 57 ...  
## $ Text : Factor w/ 102 levels "!?????? ???? ????? ??????? ??? ??? ???? 7 ????? ????? ?? 80 ??????\nUn train déraille au Maroc. Dernier bilan: "| \_\_truncated\_\_,..: 101 13 21 102 34 84 98 65 95 29 ...  
## $ Username : Factor w/ 95 levels "\_ibrahiim\_naji",..: 17 29 77 28 82 88 11 90 73 16 ...

summary(oncf)

## Number\_of\_Records Created\_Time Filter   
## Min. :1 17/10/2018 06:54: 2 Normal :78   
## 1st Qu.:1 17/10/2018 19:41: 2 Clarendon:12   
## Median :1 18/10/2018 15:18: 2 Gingham : 2   
## Mean :1 16/10/2018 11:48: 1 Hefe : 2   
## 3rd Qu.:1 17/10/2018 01:24: 1 Lark : 2   
## Max. :1 17/10/2018 06:42: 1 Ludwig : 2   
## (Other) :96 (Other) : 7   
## Image\_URL   
## https://scontent.cdninstagram.com/vp/0382098c077234d5e4ef523eec2c5f3b/5C7DC701/t51.2885-15/e35/s320x320/43079346\_310841449695923\_8385799250393832596\_n.jpg : 1   
## https://scontent.cdninstagram.com/vp/038312cc1713bb5fda24d48602b49d57/5C67D503/t51.2885-15/e35/s320x320/43377468\_989903237884294\_8739503876806183684\_n.jpg : 1   
## https://scontent.cdninstagram.com/vp/06e31e5b54106e60ac9015781e6c944b/5C87E04B/t51.2885-15/e15/s320x320/43143666\_175804796662247\_1631101817997212826\_n.jpg : 1   
## https://scontent.cdninstagram.com/vp/096fcc69a8dc754b5187b82d9f9dbd92/5C4F117F/t51.2885-15/e35/s320x320/42869662\_1927369654009983\_2279538585656329928\_n.jpg: 1   
## https://scontent.cdninstagram.com/vp/0f5862b89273a680009a79a19114c4f9/5C7CF68E/t51.2885-15/e35/s320x320/43913187\_265795407343000\_3192559832287794618\_n.jpg : 1   
## https://scontent.cdninstagram.com/vp/0f6e003f418d2f5d56a470af06a113d6/5C811D4A/t51.2885-15/e35/s320x320/43462016\_346110679551033\_6178669420682691266\_n.jpg : 1   
## (Other) :99   
## latitude Number\_of\_likes   
## 0 :87 Min. : 1.0   
## 32 : 3 1st Qu.: 12.0   
## 35,7614 : 3 Median : 28.0   
## 31,6333 : 2 Mean : 193.6   
## 33,595538: 2 3rd Qu.: 58.0   
## 34,0435 : 2 Max. :5685.0   
## (Other) : 6   
## Link   
## https://www.instagram.com/\_ibrahiim\_naji/p/BpCoflGnd8Z/: 1   
## https://www.instagram.com/e\_joesgallery/p/BpJ1FNTlkTT/ : 1   
## https://www.instagram.com/e\_joesgallery/p/BpKZLzUFF39/ : 1   
## https://www.instagram.com/p/Bo\_i3qen0V-/ : 1   
## https://www.instagram.com/p/BpB-p2DlMPD/ : 1   
## https://www.instagram.com/p/BpB1EfShRjp/ : 1   
## (Other) :99   
## Location longitude number\_of\_Comments  
## :87 0 :87 Min. : 0.000   
## Morocco : 3 -5,8195 : 3 1st Qu.: 0.000   
## Tangier, Morocco : 3 -6 : 3 Median : 1.000   
## Casablanca, Morocco: 2 -5,0032 : 2 Mean : 7.067   
## Fez, Morocco : 2 -7,61446: 2 3rd Qu.: 3.000   
## Marrakech : 2 -8 : 2 Max. :365.000   
## (Other) : 6 (Other) : 6   
## Tags   
## oncf : 7   
## : 2   
## ???\_????\_??????\_?????\_????????,?????\_????\_?\_????\_??????,???\_?????\_????\_????\_??????,oncf : 2   
## oncf,port : 2   
## ???,morocco,????,oncf,rabatagdal,dosaindia : 1   
## ????,??????,?????\_??????,?????,oncfs,????,????????,?????,???????,?????????,rabat,?????,?????,?????,????,???????,??????,??????,????\_???????,?????,oncf,kenitra: 1   
## (Other) :90   
## Text   
## #???\_?????\_????\_????\_?????? #???\_????\_??????\_?????\_???????? #oncf\n#?????\_????\_?\_????\_?????? : 3   
## #???\_?????\_????\_????\_?????? #???\_????\_??????\_?????\_???????? #oncf\n#?????\_????\_?\_????\_?????? \n@the\_moroccan\_darck\_sarcasm .\n.\n.\n#saadlamjarred1 #summer #soulking #snapchat?? #samirasaid #salmarachid #douzi #duniabtma #dubai #amrdiab #agadir #asmaalamnawar #algerie #noamanebeliachi #zouhirbahaoui #ihabamir #ibtissamtiskat #halaturk #hatimammor #hatimammor #morocco???? #mohamedramadan #le\_roi\_mohamed\_6 #?????\_?????? #????? #???? #2018: 2   
## !?????? ???? ????? ??????? ??? ??? ???? 7 ????? ????? ?? 80 ??????\nUn train déraille au Maroc. Dernier bilan: 7 morts et plus de 80 blessés ! ?\n\n\n\n\n#accident #ferroviaire #oncf #maroc #morocco #deraillement #train #wagon #bouknadel #tragedie #breakingnews #actualite #Wada7 : 1   
## " ???? ???????? ????? ??? ??????? ????? ????? ??? ??? ???? ???? ?????? " ??\n????? ????? ??????? ??????? ????? ???? ?????? ?????? ??? ?????? ????????? ??????? ?????? " ??????? " ?????? ?????? ?????? ????? ???????? ??\n#??\_???\_????\_????\_?????? ?? #ONCF ?? : 1   
## "?? ??????? ?????????????? ?????????? ????? ??????????? ????????? ??????? ?????? ?????? ?? ?????? ???????? ??????????". #savemorocco #saveyouth #morocco #rabat #bouknadel #protection #oncf : 1   
## #???? #??????? ??? ????? ???? \n#??????\n#???? \n#oncf : 1   
## (Other) :96   
## Username   
## morocco\_news1 : 3   
## sara\_habraoui : 3   
## e\_joesgallery : 2   
## hamza\_zahir : 2   
## karim\_hadri : 2   
## maghribinachet: 2   
## (Other) :91

attach(oncf)  
summary(Number\_of\_likes)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 12.0 28.0 193.6 58.0 5685.0

table(Number\_of\_likes)

## Number\_of\_likes  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 1 2 1 1 2 7 2 1 3 2 4 2 1 2 1   
## 16 17 18 19 20 22 23 24 25 26 27 28 29 30 31   
## 3 2 1 1 1 1 2 2 3 3 1 1 2 2 2   
## 32 34 36 38 39 40 44 45 47 49 50 52 57 58 59   
## 1 2 3 1 1 2 2 1 1 2 1 1 1 1 1   
## 60 61 76 86 90 100 108 137 152 176 200 208 233 247 279   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 303 328 346 391 524 928 1505 1778 4508 5685   
## 1 1 1 1 1 1 1 1 1 1

corpus = VCorpus(VectorSource(oncf$Text))   
corpus

## <<VCorpus>>  
## Metadata: corpus specific: 0, document level (indexed): 0  
## Content: documents: 105

corpus[[1]]$content

## [1] "The finest picture with the best friend and his colleagues at work ????? #boungissa #hawmiaw #oncf #bestfriend #bestpeople #smile #livingeasy #coffee #petitpain ??????"

# Convert to lower-case  
corpus = tm\_map(corpus, content\_transformer(tolower))  
# Remove punctuation  
corpus = tm\_map(corpus, removePunctuation)  
corpus[[3]]$content

## [1] "morocco train oncf marrakesh red gare voyageurs northafrica africa for the first time journey traveldairies wildcraft indian kerala arab connection"

# Remove stopwords   
corpus = tm\_map(corpus, removeWords, c(stopwords("english")))  
# Stem document   
corpus = tm\_map(corpus, stemDocument)

frequencies = DocumentTermMatrix(corpus)  
frequencies

## <<DocumentTermMatrix (documents: 105, terms: 898)>>  
## Non-/sparse entries: 1337/92953  
## Sparsity : 99%  
## Maximal term length: 27  
## Weighting : term frequency (tf)

# Check for sparsity  
findFreqTerms(frequencies, lowfreq=20)

## [1] "maroc" "morocco" "oncf" "train"

# Remove sparse terms  
sparse = removeSparseTerms(frequencies, 0.995)  
sparse

## <<DocumentTermMatrix (documents: 105, terms: 898)>>  
## Non-/sparse entries: 1337/92953  
## Sparsity : 99%  
## Maximal term length: 27  
## Weighting : term frequency (tf)

oncfSparse = as.data.frame(as.matrix(sparse))  
colnames(oncfSparse) = make.names(colnames(oncfSparse))

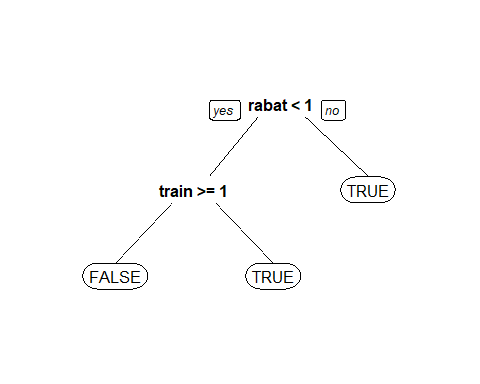
oncf$Negative = as.factor(oncf$Number\_of\_likes <= 40)  
table(oncf$Negative)

##   
## FALSE TRUE   
## 36 69

oncfSparse$Negative = oncf$Negative

set.seed(12)  
split = sample.split(oncfSparse$Negative, SplitRatio = 0.6)  
trainSparse = subset(oncfSparse, split==TRUE)  
testSparse = subset(oncfSparse, split==FALSE)

oncfCART = rpart(Negative ~ ., data=trainSparse, method="class")  
prp(oncfCART)



# Evaluate the performance of the model  
predictCART = predict(oncfCART, newdata=testSparse, type="class")  
table(testSparse$Negative, predictCART)

## predictCART  
## FALSE TRUE  
## FALSE 1 13  
## TRUE 5 23

#model's accuracy  
table(testSparse$Negative)

##   
## FALSE TRUE   
## 14 28