# Opgave 4 Analytics 2019

Naam student: *graag hier invullen*

Data voor deze opgave: twts.csv (documentatie: twts.txt).

## Inleveren

Inleveren UITSLUITEND via Blackboard. Uiterste datum van inleveren: 24-3-2019.

Inleveren UITSLUITEND in formaat .PDF, .DOC of .DOCX.

Als je met RStudio werkt kan je de .Rmd file als basis gebruiken. Anders mag je de resultaten in dit document opnemen. Vergeet niet je naam in het document te zetten.

## Inleiding

De opgave gaat over het analyseren van tweets. De vraag is om vast te stellen of een tweet een positieve dan wel een negatieve lading heeft.

Bij deze opgave begin je met ruwe data. In de data staan een aantal variabelen die je waarschijnlijk beter niet kan gebruiken. Het kan zij dat een aantal tweets meerdere malen voorkomt. (Misschien is het beter om deze teksten eerst te ontdubbelen, misschien maakt dat ook niet zo veel uit) Het is aan jou om deze data zodanig te behandelen dat je de analyse goed kan uitvoeren. De afhankelijke variabele is “Sentiment”.

De vraag is om een zo goed mogelijk CART-tree model te maken om uit de tekst van een tweet op te maken of het een positieve of negatieve lading heeft. De methode (text analytics) is uitgelegd op het college en is ook te vinden in de video’s “Turning Tweets into Knowledge”.

Bij deze opgave wordt van je verwacht dat je zelf de juiste stappen zet om tot het gewenste eindresultaat te komen. Probeer zo goed mogelijk uit te leggen WAAROM je bepaalde stappen zet.

## Vraag 1

Maak een zo optimaal mogelijk CART-tree model om uit de tekst van een tweet het (positieve/negatieve) sentiment te halen.

Let op: de Sentiment-variabele heeft DRIE verschillende waarden. Maak hiervan eerst een factor met slechts TWEE categorieen (positief/negatief). (Om te zorgen dat het een factor wordt kan je de functie as.factor gebruiken).

Gebruik daarvoor de tekst-analyse methodiek uit het college en/of de video’s. Zorg er ook voor dat je de data splitst in een trainingsset en een testset, zodat je de validatie goed kan uitvoeren.

In je antwoorddocument verwacht ik in elk geval:

* Alle R-commando’s
* Relevant commentaar zodat het duidelijk is wat je doet en waarom je het doet
* Alle relevante R-uitvoer (bijvoorbeeld confusion matrices)
* Alle relevante R-plots
* Niet alleen het opstellen van een model (op basis training data) maar ook validatie van het model (op basis van testdata). Validatie bestaat in elk geval uit confusion matrix + accuracy + sensitivity + specificity en een ROC-curve, plus een argumentatie over hoe goed het model de betreffende data kan voorspellen.
* Een goed onderbouwde conclusie

> setwd("C:/Users/wiebe/Documents/analytics/opgaven\_data")

> getwd

> tweets <- read.csv(file ='twts.csv', stringAsFactors = FALSE)

> View(tweets)

> tweets$X <- NULL

Tweets$negative = as.factor(tweets$AVG<= -1)

Table(tweets$negative)

library(tm)

library(SnowballC)

corpus= Corpus(vectorSource(tweets$tweet))

corpus = tm\_map(corpus, tolower)

corpus = tm\_map(corpus, removePunctuation)

stopwords(“dutch”)[1:5]

corpus = tm\_map(corpus, removeWords, c(“apple”,stopwords(“dutch”)))

corpus = tm\_map(corpus,PlainTextDocument)

corpus = tm\_map(corpus, StemDocument)

frequencies = DocumentTermMatrix(corpus)

str(frequencies)

inspect(frequencies[1000:1005,505:515])

sparse = removeSparseTerms(frequencies, 0.995)

tweetsSparse = as.data.frame(as.matrix(sparse))

colnames(tweetsSparse) = make.names(colnames(tweetsSparse))

library(caTools)

set.seed(9022)

split = sample.split(tweetsSparse$negative, SplitRatio = 0.7)

trainSparse = subset(tweetsSparse, split==TRUE)

testSparse = subset(tweetsSparse, split==FALSE)

> library(rpart)

> library(rpart.plot)

tweetCart = rpart(Negative~.,data=trainSparse,method=”class”)

prp(tweetCart)

## Vraag 2

Bij de eerste vraag heb je de 3 waarden voor Sentiment omgezet in 2 waarden voor de tekst-analyse. Je zou ook een CART-tree kunnen maken met daarbij alle drie waarden van Sentiment. Doe dit en geef het resultaat.

Gebruik cross-validation om de best mogelijke cp-parameter vast te stellen hiervoor.

De confusion-matrix is nu 3 bij 3. Geef aan wat dat voor consequenties heeft voor je validatie.

In je antwoorddocument verwacht ik in elk geval:

* Alle R-commando’s
* Relevant commentaar zodat het duidelijk is wat je doet en waarom je het doet
* Alle relevante R-uitvoer (bijvoorbeeld confusion matrices)
* Alle relevante R-plots
* Je antwoord op de vraag in de laatste zin van vraag 2 (over de 3x3 matrix)
* Een goed onderbouwde conclusie

> model=glm(Churn~IP+VMP+NVM+TDCH+TEM+TNM+TIC+TICH+CSC,data=tweets,family=binomial)

> summary(model)

predictTweet=predict(model,type="response",newdata=telcotest)

library(ROCR)

ROCRperfi=performance(predictTweet,"tpr","fpr")

plot(ROCRperfi,colorize=true,print.cutoffs.at=seq(0,1,0.1))

abline(0,1)

AUC=as.numeric(performance(predictTelco,”AUC”)@y.values)

AUC

> library(rpart)

> library(rpart.plot)

tree=rpart(Churn~.,data=telcoTest,method=”class”,control=rpart.control(minbucket=25))

rpr(tree)

confusion <-function(arg1,arg2,arg3){

+ table = table(arg1,arg2>arg3)

+ specificity <-table[1,1]/(table[1,1]+table[1,2])

+ sensitivity <-table[2,2]/(table[2,1]+table[2,2])

+ accuracy <-(table[1,1]+table[2,2])/(table[1,1]+table[1,2]+table[2,1]+table[2,2])

+ return(data.frame(Specificity=specificity,Sensitivity=sensitivity,Accuracy=accuracy))}

> confusion(telcoTest$Churn,predictTest,0.1)

library(caTools)

set.seed(9022)

split=sample.split(data$Churn,SplitRatio=0.65)

trainset=subset(data,split=TRUE)

testset=subset(data,split=FALSE)

library(caret)

library(e1071)

fitcontrol=trainControl(method=”cv”,number=10)

cartGrid=expand.grid(.cp=(1:50)\*0.01)

cv=telcoTest(Churn~.,data=telcoTest,method=”rpart”,trControl=fitControl,tuneGrid=cartGrid)

tweets =

## Eindbeoordeling Analytics

Je wordt uitgenodigd om de resultaten van deze opgave in een assessment te bespreken met de docent. (Het assessment duurt ongeveer 8 minuten). Het eindcijfer van het vak Analytics wordt daarna bepaald uit het gemiddelde cijfer voor de vier opgaven en het cijfer voor het assessment. Voor een voldoende resultaat moeten minimaal drie opgaven zijn ingeleverd en het assessment moet met een voldoende zijn afgesloten.