# 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("D:/GitHub/School/P2.3/Analytics/opgaven\_data")

> data=read.csv("twts.csv", stringsAsFactors = FALSE)

> View(data)

> str(data)

'data.frame': 498 obs. of 6 variables:

$ Sentiment: int 4 4 4 4 4 4 0 4 4 4 ...

$ ID : int 3 4 5 6 7 8 9 10 11 12 ...

$ Date : chr "Mon May 11 03:17:40 UTC 2009" "Mon May 11 03:18:03 UTC 2009" "Mon May 11 03:18:54 UTC 2009" "Mon May 11 03:19:04 UTC 2009" ...

$ Query : chr "kindle2" "kindle2" "kindle2" "kindle2" ...

$ User : chr "tpryan" "vcu451" "chadfu" "SIX15" ...

$ Tweet : chr "@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right." "Reading my kindle2... Love it... Lee childs is good read." "Ok, first assesment of the #kindle2 ...it fucking rocks!!!" "@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is "| \_\_truncated\_\_ ...

> data$ID <- NULL

> data$User <- NULL

> data$Query <- NULL

> data$Date <- NULL

> data$Negative <- as.factor(data$Sentiment <= 1)

> str(data)

'data.frame': 498 obs. of 3 variables:

$ Sentiment: int 4 4 4 4 4 4 0 4 4 4 ...

$ Tweet : chr "@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right." "Reading my kindle2... Love it... Lee childs is good read." "Ok, first assesment of the #kindle2 ...it fucking rocks!!!" "@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is "| \_\_truncated\_\_ ...

$ Negative : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 2 1 1 1 ...

> table(data$Negative)

FALSE TRUE

321 177

Check op dubbele waardes, deze zijn er niet

> duplicated(data)

[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[20] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[39] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[58] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[77] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[96] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[115] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[134] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[153] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[172] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[191] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[248] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[267] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[286] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[305] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[324] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[343] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[362] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[381] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[400] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[419] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[438] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[476] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[495] FALSE FALSE FALSE FALSE

> library(tm)

> library(snowballC)

Corpus aanmaken en deze filteren en perfectioneren

> corpus = Corpus(VectorSource(data$Tweet))

> corpus

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 498

> corpus = tm\_map(corpus, tolower)

Warning message:

In tm\_map.SimpleCorpus(corpus, tolower) : transformation drops documents

> corpus = tm\_map(corpus, removePunctuation)

Warning message:

In tm\_map.SimpleCorpus(corpus, removePunctuation) :

transformation drops documents

> corpus = tm\_map(corpus, removewords, stopwords("English"))

Error in tm\_map.SimpleCorpus(corpus, removewords, stopwords("English")) :

object 'removewords' not found

> corpus = tm\_map(corpus, stemDocument)

Warning message:

In tm\_map.SimpleCorpus(corpus, stemDocument) :

transformation drops documents

> corpus

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 498

Met behulp van frequencies probeer ik te zien hoevaak bepaalde woorden voorkomen.

> frequencies = DocumentTermMatrix(corpus)

> str(frequencies)

List of 6

$ i : int [1:4969] 1 1 1 1 1 1 1 1 1 1 ...

$ j : int [1:4969] 1 2 3 4 5 6 7 8 9 10 ...

$ v : num [1:4969] 1 1 1 1 1 1 1 1 1 1 ...

$ nrow : int 498

$ ncol : int 1943

$ dimnames:List of 2

..$ Docs : chr [1:498] "1" "2" "3" "4" ...

..$ Terms: chr [1:1943] "but" "cool" "fantast" "kindle2" ...

- attr(\*, "class")= chr [1:2] "DocumentTermMatrix" "simple\_triplet\_matrix"

- attr(\*, "weighting")= chr [1:2] "term frequency" "tf"

> inspect(frequencies)

<<DocumentTermMatrix (documents: 498, terms: 1943)>>

Non-/sparse entries: 4969/962645

Sparsity : 99%

Maximal term length: 46

Weighting : term frequency (tf)

Sample :

Terms

Docs and for have just love night the time with you

165 0 0 0 0 0 0 1 0 0 0

206 0 1 0 0 0 0 3 1 0 0

240 0 0 0 0 1 0 3 0 0 0

30 1 1 0 0 1 0 4 0 0 0

381 1 3 0 0 0 0 1 0 0 0

382 1 1 0 0 0 0 3 0 0 0

4 1 2 0 0 1 0 1 0 0 0

40 0 2 0 0 0 0 0 0 1 0

455 0 1 0 0 0 0 2 0 0 0

61 0 0 0 0 0 0 0 0 2 0

> frequencies

<<DocumentTermMatrix (documents: 498, terms: 1943)>>

Non-/sparse entries: 4969/962645

Sparsity : 99%

Maximal term length: 46

Weighting : term frequency (tf)

Met behulp van de Sparse methodes verwijder ik termen die nauwelijks voorkomen

> sparse=removeSparseTerms(frequencies, 0.995)

> sparse

<<DocumentTermMatrix (documents: 498, terms: 374)>>

Non-/sparse entries: 3158/183094

Sparsity : 98%

Maximal term length: 27

Weighting : term frequency (tf)

> sparseData = as.data.frame(as.matrix(sparse))

> str(sparseData)

'data.frame': 498 obs. of 374 variables:

$ but : num 1 0 0 0 1 0 0 0 0 0 ...

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ not : num 1 0 0 0 0 0 0 0 0 1 ...

$ own : num 1 0 0 0 0 0 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ that : num 1 0 0 0 0 0 0 0 0 0 ...

$ the : num 2 0 1 1 1 1 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ and : num 0 0 0 1 1 0 1 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ few : num 0 0 0 1 0 0 0 0 0 0 ...

$ for : num 0 0 0 2 0 0 0 0 0 0 ...

$ had : num 0 0 0 1 0 0 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ your : num 0 0 0 1 0 0 0 0 0 0 ...

$ have : num 0 0 0 0 1 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ too : num 0 0 0 0 0 1 0 0 0 0 ...

$ with : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ this : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ about : num 0 0 0 0 0 0 0 0 0 1 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ how : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ you : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ out : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ they : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ was : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ all : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ into : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ our : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ are : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ there : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ from : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> colnames(sparseData) = make.names(colnames(sparseData))

> str(sparseData)

'data.frame': 498 obs. of 374 variables:

$ but : num 1 0 0 0 1 0 0 0 0 0 ...

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ not : num 1 0 0 0 0 0 0 0 0 1 ...

$ own : num 1 0 0 0 0 0 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ that : num 1 0 0 0 0 0 0 0 0 0 ...

$ the : num 2 0 1 1 1 1 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ and : num 0 0 0 1 1 0 1 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ few : num 0 0 0 1 0 0 0 0 0 0 ...

$ for. : num 0 0 0 2 0 0 0 0 0 0 ...

$ had : num 0 0 0 1 0 0 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ your : num 0 0 0 1 0 0 0 0 0 0 ...

$ have : num 0 0 0 0 1 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ too : num 0 0 0 0 0 1 0 0 0 0 ...

$ with : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ this : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ about : num 0 0 0 0 0 0 0 0 0 1 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ how : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ you : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ out : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ they : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ was : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ all : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ into : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ our : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ are : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ there : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ from : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> sparseData$Negative = data$Negative

> str(sparseData)

'data.frame': 498 obs. of 375 variables:

$ but : num 1 0 0 0 1 0 0 0 0 0 ...

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ not : num 1 0 0 0 0 0 0 0 0 1 ...

$ own : num 1 0 0 0 0 0 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ that : num 1 0 0 0 0 0 0 0 0 0 ...

$ the : num 2 0 1 1 1 1 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ and : num 0 0 0 1 1 0 1 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ few : num 0 0 0 1 0 0 0 0 0 0 ...

$ for. : num 0 0 0 2 0 0 0 0 0 0 ...

$ had : num 0 0 0 1 0 0 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ your : num 0 0 0 1 0 0 0 0 0 0 ...

$ have : num 0 0 0 0 1 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ too : num 0 0 0 0 0 1 0 0 0 0 ...

$ with : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ this : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ about : num 0 0 0 0 0 0 0 0 0 1 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ how : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ you : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ out : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ they : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ was : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ all : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ into : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ our : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ are : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ there : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ from : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> install.packages("caTools")

Installing package into ‘C:/Users/kisha/Documents/R/win-library/3.5’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/caTools\_1.17.1.2.zip'

Content type 'application/zip' length 329608 bytes (321 KB)

downloaded 321 KB

package ‘caTools’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\kisha\AppData\Local\Temp\Rtmp6VwGyD\downloaded\_packages

> library(caTools)

Warning message:

package ‘caTools’ was built under R version 3.5.3

Na het afronden van de voorbereiding ga ik nu de data splitsen en een model

maken dmv cross validation

> set.seed(8602)

> split = sample.split(sparseData$Negative, SplitRatio = 0.7)

> trainData = subset(sparseData, split==TRUE)

> testData = subset(sparseData, split==FALSE)

> library(caret)

Loading required package: lattice

Loading required package: ggplot2

Attaching package: ‘ggplot2’

The following object is masked from ‘package:NLP’:

annotate

Warning messages:

1: package ‘caret’ was built under R version 3.5.3

2: package ‘ggplot2’ was built under R version 3.5.3

> library(e1071)

Warning message:

package ‘e1071’ was built under R version 3.5.3

> fitControl=trainControl(method="cv", number = 10)

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

> cv=train(Negative~.,data=trainData,method="rpart", trControl=fitControl,tuneGrid=cartGrid)

> cv

CART

349 samples

374 predictors

2 classes: 'FALSE', 'TRUE'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 315, 315, 314, 314, 313, 313, ...

Resampling results across tuning parameters:

cp Accuracy Kappa

0.01 0.7214659 0.29891599

0.02 0.7329038 0.31980235

0.03 0.7299627 0.30777380

0.04 0.7330672 0.30581616

0.05 0.7247339 0.27921912

0.06 0.7162465 0.24899128

0.07 0.7105275 0.23123397

0.08 0.7105275 0.22844289

0.09 0.7133847 0.23441385

0.10 0.7133847 0.23441385

0.11 0.7133847 0.23441385

0.12 0.7133847 0.23441385

0.13 0.7133847 0.23441385

0.14 0.7133847 0.23441385

0.15 0.7133847 0.23441385

0.16 0.7133847 0.23441385

0.17 0.7133847 0.23441385

0.18 0.7019561 0.19475867

0.19 0.6736881 0.10069037

0.20 0.6504949 0.02014975

0.21 0.6447759 0.00000000

0.22 0.6447759 0.00000000

0.23 0.6447759 0.00000000

0.24 0.6447759 0.00000000

0.25 0.6447759 0.00000000

0.26 0.6447759 0.00000000

0.27 0.6447759 0.00000000

0.28 0.6447759 0.00000000

0.29 0.6447759 0.00000000

0.30 0.6447759 0.00000000

0.31 0.6447759 0.00000000

0.32 0.6447759 0.00000000

0.33 0.6447759 0.00000000

0.34 0.6447759 0.00000000

0.35 0.6447759 0.00000000

0.36 0.6447759 0.00000000

0.37 0.6447759 0.00000000

0.38 0.6447759 0.00000000

0.39 0.6447759 0.00000000

0.40 0.6447759 0.00000000

0.41 0.6447759 0.00000000

0.42 0.6447759 0.00000000

0.43 0.6447759 0.00000000

0.44 0.6447759 0.00000000

0.45 0.6447759 0.00000000

0.46 0.6447759 0.00000000

0.47 0.6447759 0.00000000

0.48 0.6447759 0.00000000

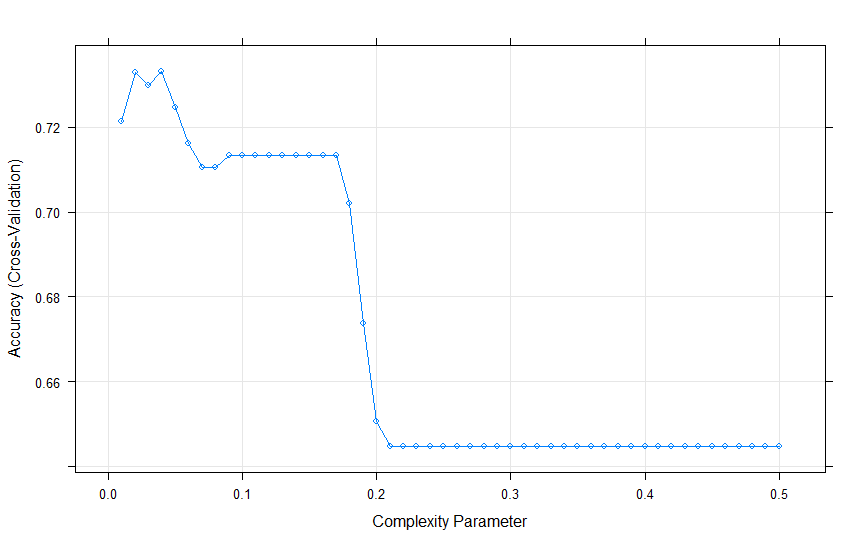
0.49 0.6447759 0.00000000

0.50 0.6447759 0.00000000

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.04.

> plot(cv)

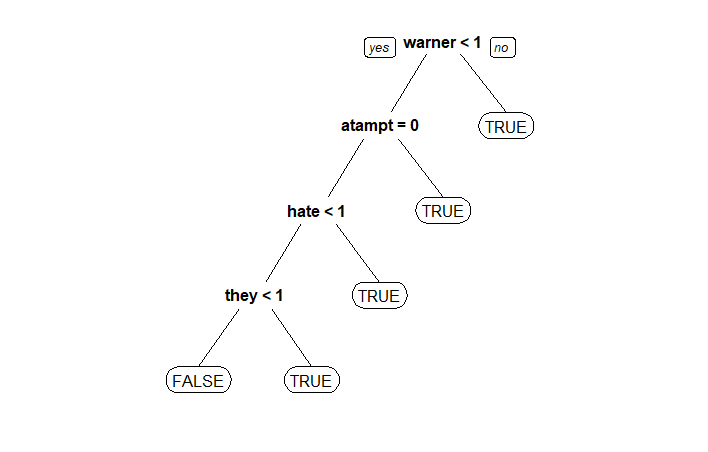


> library(rpart)

> tree=rpart(Negative~.,data=trainData,method="class",control=rpart.control(cp=0.04))

> library(rpart.plot)

> prp(tree)



> library(ROCR)

Loading required package: gplots

Attaching package: ‘gplots’

The following object is masked from ‘package:stats’:

lowess

Warning message:

package ‘ROCR’ was built under R version 3.5.3

Nu maken we een roc curve hier van

> predictTreeROC=predict(tree,newdata=testData)

> treeROC=prediction(predictTreeROC[,2], testData$Negative)

> treePerf=performance(treeROC,"tpr","fpr")

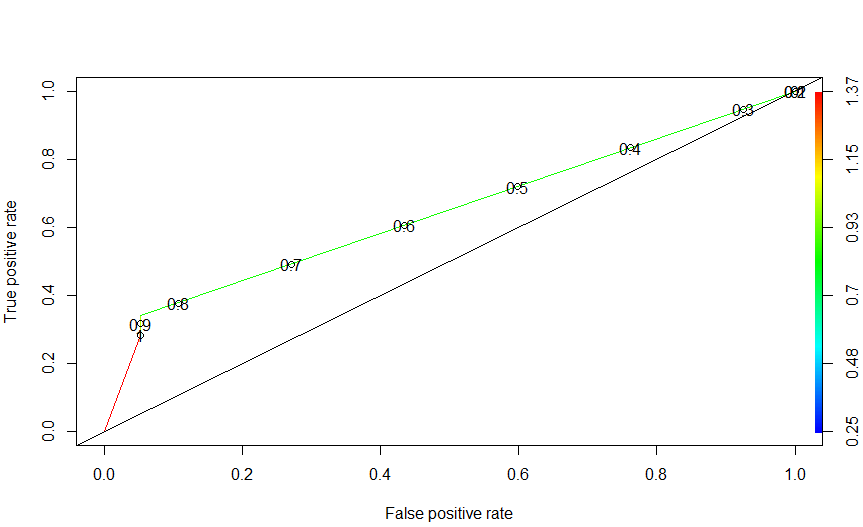
> plot(treePerf, colorize=TRUE,print.cutoffs.at=seq(0,1,0.1))

> abline(0,1)

> AUC=as.numeric(performance(treeROC, "auc")@y.values)

> AUC

[1] 0.6422956



> source('D:/GitHub/School/P2.3/Analytics/opgaven\_data/conf.R')

> conf(testData$Negative, 0.5)

n FALSE TRUE

FALSE 91 5

TRUE 35 18

[1] "Sensitivity = 0.782608695652174"

[1] "Specificify = 0.722222222222222"

[1] "Accuracy = 0.731543624161074"

Door de hoge Sensitivity voorspelt het model de meeste(84%) van de negatieve tweets als negatieve tweets. Dit vindt ik goed vooral gezien de meeste online forums al hele tijd negatieve reviews en comments proberen te verbannen van hun sites en het altijd word gezien al bijna onmogelijk om negatieve berichten te blokkeren.

De Specificify is helaas wat lager (71%) Dit vindt ik als ik zelf een social media forum had niet acceptabel. Doordat de Specificify 71% is zou dit niet handig zijn op een revieuw site of zelfs een forum voor hobbyisten. Wanneer meer dan een kwart van de goede berichten moeten worden nagekeken of geblocked kom je snel in een situatie waar dit veel te veel werk word en mensen zich gecensureerd voelen.

Ik zie echter wel specifieke onderdelen van een webpagina waar dit handig kan wezen. Neem bijvoorbeeld de facebook feed van een ziek kind. Door de hoge Sensitivity worden eventuele negatieve berichten gefilterd en hopelijk helpt dit de stress te verhelpen bij het kind.

Ter conclusie: Dit model is handig voor specifieke situaties zoals boven eerder al aangegeven maar is naar mijn mening niet goed genoeg om globaal ingezet te worden op forums/social media. De specificity is op dit moment te laag om negatieve reacties te voorkomen zonder een hoog hinder percentage.

## 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

## 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.

> data=read.csv("twts.csv", stringsAsFactors = FALSE)

> str(data)

'data.frame': 498 obs. of 6 variables:

$ Sentiment: int 4 4 4 4 4 4 0 4 4 4 ...

$ ID : int 3 4 5 6 7 8 9 10 11 12 ...

$ Date : chr "Mon May 11 03:17:40 UTC 2009" "Mon May 11 03:18:03 UTC 2009" "Mon May 11 03:18:54 UTC 2009" "Mon May 11 03:19:04 UTC 2009" ...

$ Query : chr "kindle2" "kindle2" "kindle2" "kindle2" ...

$ User : chr "tpryan" "vcu451" "chadfu" "SIX15" ...

$ Tweet : chr "@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right." "Reading my kindle2... Love it... Lee childs is good read." "Ok, first assesment of the #kindle2 ...it fucking rocks!!!" "@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is "| \_\_truncated\_\_ ...

> View(data)

> data$ID <- NULL

> data$User <- NULL

> data$Query <- NULL

> data$Date <- NULL

> View(data)

> table(data$Sentiment)

0 2 4

177 139 182

> corpus = Corpus(VectorSource(data$Tweet))

> corpus = tm\_map(corpus, tolower)

Warning message:

In tm\_map.SimpleCorpus(corpus, tolower) : transformation drops documents

> corpus = tm\_map(corpus, removePunctuation)

Warning message:

In tm\_map.SimpleCorpus(corpus, removePunctuation) :

transformation drops documents

> corpus = tm\_map(corpus, removeWords, stopwords("English"))

Warning message:

In tm\_map.SimpleCorpus(corpus, removeWords, stopwords("English")) :

transformation drops documents

> corpus = tm\_map(corpus, stemDocument)

Warning message:

In tm\_map.SimpleCorpus(corpus, stemDocument) :

transformation drops documents

> corpus

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 498

> frequencies = DocumentTermMatrix(corpus)

> str(frequencies)

List of 6

$ i : int [1:3992] 1 1 1 1 1 1 2 2 2 2 ...

$ j : int [1:3992] 1 2 3 4 5 6 7 8 3 9 ...

$ v : num [1:3992] 1 1 1 1 1 1 1 1 1 1 ...

$ nrow : int 498

$ ncol : int 1872

$ dimnames:List of 2

..$ Docs : chr [1:498] "1" "2" "3" "4" ...

..$ Terms: chr [1:1872] "cool" "fantast" "kindle2" "loooooooovvvvvvee" ...

- attr(\*, "class")= chr [1:2] "DocumentTermMatrix" "simple\_triplet\_matrix"

- attr(\*, "weighting")= chr [1:2] "term frequency" "tf"

> inspect(frequencies)

<<DocumentTermMatrix (documents: 498, terms: 1872)>>

Non-/sparse entries: 3992/928264

Sparsity : 100%

Maximal term length: 46

Weighting : term frequency (tf)

Sample :

Terms

Docs good just love museum new night nike see time

124 0 0 0 0 0 0 0 0 0

150 0 0 0 0 0 0 0 0 1

165 0 0 0 0 0 0 0 0 0

4 0 0 1 0 1 0 0 0 0

470 1 0 0 0 0 0 0 0 0

480 0 0 0 0 0 1 0 0 0

61 0 0 0 0 1 0 0 0 0

77 0 0 0 0 2 0 0 0 0

81 0 0 0 0 0 0 0 0 0

95 0 1 0 0 0 0 0 0 0

Terms

Docs warner

124 0

150 1

165 0

4 0

470 0

480 0

61 0

77 0

81 0

95 0

> frequencies

<<DocumentTermMatrix (documents: 498, terms: 1872)>>

Non-/sparse entries: 3992/928264

Sparsity : 100%

Maximal term length: 46

Weighting : term frequency (tf)

> sparse=removeSparseTerms(frequencies, 0.995)

> sparse

<<DocumentTermMatrix (documents: 498, terms: 316)>>

Non-/sparse entries: 2200/155168

Sparsity : 99%

Maximal term length: 27

Weighting : term frequency (tf)

> sparseData = as.data.frame(as.matrix(sparse))

> str(sparseData)

'data.frame': 498 obs. of 316 variables:

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

$ say : num 0 0 0 0 0 0 0 0 0 0 ...

$ sick : num 0 0 0 0 0 0 0 0 0 0 ...

$ visa : num 0 0 0 0 0 0 0 0 0 0 ...

$ long : num 0 0 0 0 0 0 0 0 0 0 ...

$ weekend : num 0 0 0 0 0 0 0 0 0 0 ...

$ allen : num 0 0 0 0 0 0 0 0 0 0 ...

$ bad : num 0 0 0 0 0 0 0 0 0 0 ...

$ booz : num 0 0 0 0 0 0 0 0 0 0 ...

$ hamilton : num 0 0 0 0 0 0 0 0 0 0 ...

$ way : num 0 0 0 0 0 0 0 0 0 0 ...

$ custom : num 0 0 0 0 0 0 0 0 0 0 ...

$ canon : num 0 0 0 0 0 0 0 0 0 0 ...

$ use : num 0 0 0 0 0 0 0 0 0 0 ...

$ 40d : num 0 0 0 0 0 0 0 0 0 0 ...

$ suggest : num 0 0 0 0 0 0 0 0 0 0 ...

$ googl : num 0 0 0 0 0 0 0 0 0 0 ...

$ show : num 0 0 0 0 0 0 0 0 0 0 ...

$ dad : num 0 0 0 0 0 0 0 0 0 0 ...

$ place : num 0 0 0 0 0 0 0 0 0 0 ...

$ work : num 0 0 0 0 0 0 0 0 0 0 ...

$ android : num 0 0 0 0 0 0 0 0 0 0 ...

$ fast : num 0 0 0 0 0 0 0 0 0 0 ...

$ phone : num 0 0 0 0 0 0 0 0 0 0 ...

$ play : num 0 0 0 0 0 0 0 0 0 0 ...

$ slide : num 0 0 0 0 0 0 0 0 0 0 ...

$ bay : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> colnames(sparseData) = make.names(colnames(sparseData))

> str(sparseData)

'data.frame': 498 obs. of 316 variables:

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

$ say : num 0 0 0 0 0 0 0 0 0 0 ...

$ sick : num 0 0 0 0 0 0 0 0 0 0 ...

$ visa : num 0 0 0 0 0 0 0 0 0 0 ...

$ long : num 0 0 0 0 0 0 0 0 0 0 ...

$ weekend : num 0 0 0 0 0 0 0 0 0 0 ...

$ allen : num 0 0 0 0 0 0 0 0 0 0 ...

$ bad : num 0 0 0 0 0 0 0 0 0 0 ...

$ booz : num 0 0 0 0 0 0 0 0 0 0 ...

$ hamilton : num 0 0 0 0 0 0 0 0 0 0 ...

$ way : num 0 0 0 0 0 0 0 0 0 0 ...

$ custom : num 0 0 0 0 0 0 0 0 0 0 ...

$ canon : num 0 0 0 0 0 0 0 0 0 0 ...

$ use : num 0 0 0 0 0 0 0 0 0 0 ...

$ X40d : num 0 0 0 0 0 0 0 0 0 0 ...

$ suggest : num 0 0 0 0 0 0 0 0 0 0 ...

$ googl : num 0 0 0 0 0 0 0 0 0 0 ...

$ show : num 0 0 0 0 0 0 0 0 0 0 ...

$ dad : num 0 0 0 0 0 0 0 0 0 0 ...

$ place : num 0 0 0 0 0 0 0 0 0 0 ...

$ work : num 0 0 0 0 0 0 0 0 0 0 ...

$ android : num 0 0 0 0 0 0 0 0 0 0 ...

$ fast : num 0 0 0 0 0 0 0 0 0 0 ...

$ phone : num 0 0 0 0 0 0 0 0 0 0 ...

$ play : num 0 0 0 0 0 0 0 0 0 0 ...

$ slide : num 0 0 0 0 0 0 0 0 0 0 ...

$ bay : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> sparseData$Sentiment = data$Sentiment

> str(sparseData)

'data.frame': 498 obs. of 317 variables:

$ cool : num 1 0 0 0 0 0 0 0 0 0 ...

$ kindle2 : num 1 1 1 1 1 1 0 0 0 0 ...

$ right : num 1 0 0 0 0 0 0 0 0 0 ...

$ good : num 0 1 0 0 0 0 0 0 0 0 ...

$ love : num 0 1 0 1 0 0 0 0 1 1 ...

$ read : num 0 2 0 0 0 0 0 0 0 0 ...

$ first : num 0 0 1 0 0 0 0 0 0 0 ...

$ fuck : num 0 0 1 0 0 0 1 0 0 0 ...

$ rock : num 0 0 1 0 0 0 0 0 0 0 ...

$ back : num 0 0 0 1 0 0 0 0 0 0 ...

$ big : num 0 0 0 1 0 1 0 0 0 0 ...

$ ive : num 0 0 0 1 0 0 0 0 0 0 ...

$ look : num 0 0 0 1 0 0 0 0 0 0 ...

$ need : num 0 0 0 1 0 0 0 0 0 0 ...

$ never : num 0 0 0 1 0 0 0 0 0 0 ...

$ new : num 0 0 0 1 0 0 0 1 0 0 ...

$ one : num 0 0 0 1 0 0 0 0 0 0 ...

$ think : num 0 0 0 0 1 0 0 0 0 0 ...

$ happi : num 0 0 0 0 0 1 0 0 0 0 ...

$ quit : num 0 0 0 0 0 1 0 0 0 0 ...

$ aig : num 0 0 0 0 0 0 1 0 0 0 ...

$ ass : num 0 0 0 0 0 0 1 0 0 0 ...

$ hate : num 0 0 0 0 0 0 1 0 0 0 ...

$ best : num 0 0 0 0 0 0 0 1 0 0 ...

$ jqueri : num 0 0 0 0 0 0 0 1 0 0 ...

$ twitter : num 0 0 0 0 0 0 0 0 1 0 ...

$ can : num 0 0 0 0 0 0 0 0 0 1 ...

$ joke : num 0 0 0 0 0 0 0 0 0 1 ...

$ make : num 0 0 0 0 0 0 0 0 0 1 ...

$ obama : num 0 0 0 0 0 0 0 0 0 1 ...

$ check : num 0 0 0 0 0 0 0 0 0 0 ...

$ dinner : num 0 0 0 0 0 0 0 0 0 0 ...

$ hous : num 0 0 0 0 0 0 0 0 0 0 ...

$ presid : num 0 0 0 0 0 0 0 0 0 0 ...

$ video : num 0 0 0 0 0 0 0 0 0 0 ...

$ white : num 0 0 0 0 0 0 0 0 0 0 ...

$ slogan : num 0 0 0 0 0 0 0 0 0 0 ...

$ want : num 0 0 0 0 0 0 0 0 0 0 ...

$ amp : num 0 0 0 0 0 0 0 0 0 0 ...

$ got : num 0 0 0 0 0 0 0 0 0 0 ...

$ last : num 0 0 0 0 0 0 0 0 0 0 ...

$ night : num 0 0 0 0 0 0 0 0 0 0 ...

$ went : num 0 0 0 0 0 0 0 0 0 0 ...

$ nike : num 0 0 0 0 0 0 0 0 0 0 ...

$ seen : num 0 0 0 0 0 0 0 0 0 0 ...

$ shit : num 0 0 0 0 0 0 0 0 0 0 ...

$ stop : num 0 0 0 0 0 0 0 0 0 0 ...

$ wast : num 0 0 0 0 0 0 0 0 0 0 ...

$ basketbal : num 0 0 0 0 0 0 0 0 0 0 ...

$ dont : num 0 0 0 0 0 0 0 0 0 0 ...

$ get : num 0 0 0 0 0 0 0 0 0 0 ...

$ lebron : num 0 0 0 0 0 0 0 0 0 0 ...

$ time : num 0 0 0 0 0 0 0 0 0 0 ...

$ also : num 0 0 0 0 0 0 0 0 0 0 ...

$ die : num 0 0 0 0 0 0 0 0 0 0 ...

$ guy : num 0 0 0 0 0 0 0 0 0 0 ...

$ talk : num 0 0 0 0 0 0 0 0 0 0 ...

$ tell : num 0 0 0 0 0 0 0 0 0 0 ...

$ still : num 0 0 0 0 0 0 0 0 0 0 ...

$ laker : num 0 0 0 0 0 0 0 0 0 0 ...

$ let : num 0 0 0 0 0 0 0 0 0 0 ...

$ lol : num 0 0 0 0 0 0 0 0 0 0 ...

$ awesom : num 0 0 0 0 0 0 0 0 0 0 ...

$ come : num 0 0 0 0 0 0 0 0 0 0 ...

$ even : num 0 0 0 0 0 0 0 0 0 0 ...

$ nba : num 0 0 0 0 0 0 0 0 0 0 ...

$ app : num 0 0 0 0 0 0 0 0 0 0 ...

$ fun : num 0 0 0 0 0 0 0 0 0 0 ...

$ iphon : num 0 0 0 0 0 0 0 0 0 0 ...

$ just : num 0 0 0 0 0 0 0 0 0 0 ...

$ much : num 0 0 0 0 0 0 0 0 0 0 ...

$ call : num 0 0 0 0 0 0 0 0 0 0 ...

$ offic : num 0 0 0 0 0 0 0 0 0 0 ...

$ say : num 0 0 0 0 0 0 0 0 0 0 ...

$ sick : num 0 0 0 0 0 0 0 0 0 0 ...

$ visa : num 0 0 0 0 0 0 0 0 0 0 ...

$ long : num 0 0 0 0 0 0 0 0 0 0 ...

$ weekend : num 0 0 0 0 0 0 0 0 0 0 ...

$ allen : num 0 0 0 0 0 0 0 0 0 0 ...

$ bad : num 0 0 0 0 0 0 0 0 0 0 ...

$ booz : num 0 0 0 0 0 0 0 0 0 0 ...

$ hamilton : num 0 0 0 0 0 0 0 0 0 0 ...

$ way : num 0 0 0 0 0 0 0 0 0 0 ...

$ custom : num 0 0 0 0 0 0 0 0 0 0 ...

$ canon : num 0 0 0 0 0 0 0 0 0 0 ...

$ use : num 0 0 0 0 0 0 0 0 0 0 ...

$ X40d : num 0 0 0 0 0 0 0 0 0 0 ...

$ suggest : num 0 0 0 0 0 0 0 0 0 0 ...

$ googl : num 0 0 0 0 0 0 0 0 0 0 ...

$ show : num 0 0 0 0 0 0 0 0 0 0 ...

$ dad : num 0 0 0 0 0 0 0 0 0 0 ...

$ place : num 0 0 0 0 0 0 0 0 0 0 ...

$ work : num 0 0 0 0 0 0 0 0 0 0 ...

$ android : num 0 0 0 0 0 0 0 0 0 0 ...

$ fast : num 0 0 0 0 0 0 0 0 0 0 ...

$ phone : num 0 0 0 0 0 0 0 0 0 0 ...

$ play : num 0 0 0 0 0 0 0 0 0 0 ...

$ slide : num 0 0 0 0 0 0 0 0 0 0 ...

$ bay : num 0 0 0 0 0 0 0 0 0 0 ...

[list output truncated]

> set.seed(8602)

> split = sample.split(sparseData$Sentiment, SplitRatio = 0.7)

> trainData = subset(sparseData, split==TRUE)

> testData = subset(sparseData, split==FALSE)

> fitControl=trainControl(method="cv", number = 10)

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

> cv=train(Sentiment~.,data=trainData,method="rpart", trControl=fitControl,tuneGrid=cartGrid)

Warning message:

In nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :

There were missing values in resampled performance measures.

> cv

CART

348 samples

316 predictors

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 314, 314, 313, 313, 313, 313, ...

Resampling results across tuning parameters:

cp RMSE Rsquared MAE

0.01 1.514380 0.23083713 1.150484

0.02 1.522140 0.21929906 1.155059

0.03 1.558450 0.17530604 1.211974

0.04 1.640264 0.08923005 1.320585

0.05 1.636071 0.09148276 1.339852

0.06 1.631342 0.09159490 1.350609

0.07 1.623562 0.09329692 1.358239

0.08 1.649885 0.07707145 1.394392

0.09 1.691380 0.04160281 1.446293

0.10 1.698691 NaN 1.448335

0.11 1.698691 NaN 1.448335

0.12 1.698691 NaN 1.448335

0.13 1.698691 NaN 1.448335

0.14 1.698691 NaN 1.448335

0.15 1.698691 NaN 1.448335

0.16 1.698691 NaN 1.448335

0.17 1.698691 NaN 1.448335

0.18 1.698691 NaN 1.448335

0.19 1.698691 NaN 1.448335

0.20 1.698691 NaN 1.448335

0.21 1.698691 NaN 1.448335

0.22 1.698691 NaN 1.448335

0.23 1.698691 NaN 1.448335

0.24 1.698691 NaN 1.448335

0.25 1.698691 NaN 1.448335

0.26 1.698691 NaN 1.448335

0.27 1.698691 NaN 1.448335

0.28 1.698691 NaN 1.448335

0.29 1.698691 NaN 1.448335

0.30 1.698691 NaN 1.448335

0.31 1.698691 NaN 1.448335

0.32 1.698691 NaN 1.448335

0.33 1.698691 NaN 1.448335

0.34 1.698691 NaN 1.448335

0.35 1.698691 NaN 1.448335

0.36 1.698691 NaN 1.448335

0.37 1.698691 NaN 1.448335

0.38 1.698691 NaN 1.448335

0.39 1.698691 NaN 1.448335

0.40 1.698691 NaN 1.448335

0.41 1.698691 NaN 1.448335

0.42 1.698691 NaN 1.448335

0.43 1.698691 NaN 1.448335

0.44 1.698691 NaN 1.448335

0.45 1.698691 NaN 1.448335

0.46 1.698691 NaN 1.448335

0.47 1.698691 NaN 1.448335

0.48 1.698691 NaN 1.448335

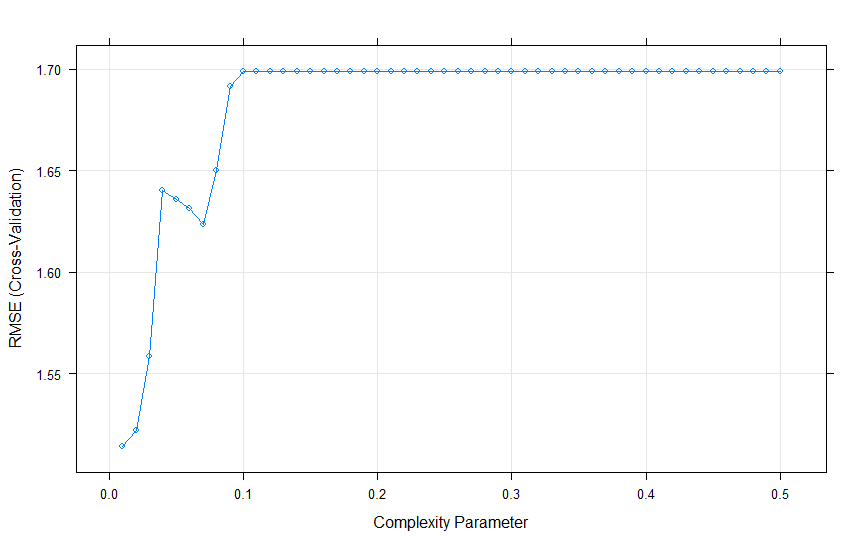
0.49 1.698691 NaN 1.448335

0.50 1.698691 NaN 1.448335

RMSE was used to select the optimal model using

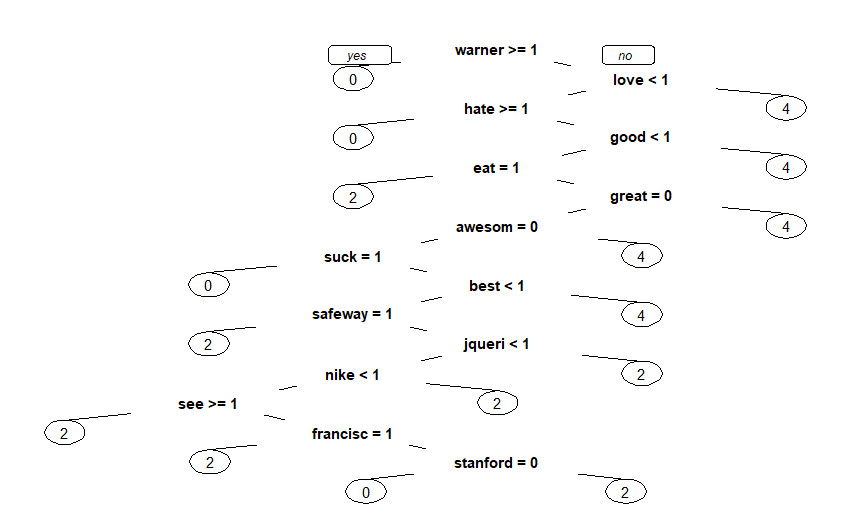
the smallest value.

The final value used for the model was cp = 0.01.

> plot(cv) 

> tree=rpart(Sentiment~.,data=trainData,method="class",control=rpart.control(cp=0.01))

> prp(tree)



> testData$Sentiment <- as.factor(testData$Sentiment)

conf <- function(n, d)

{

predicttest= predict (tree, newdata=testData, type="class")

t=table(n, predicttest)

T1=t[1]

F1a=t[2]

F1b=t[3]

F2a=t[4]

T2=t[5]

F2b=t[6]

F3a=t[7]

F3b=t[8]

T3=t[9]

sensitivity0 <- T1 / (T1 + F1a + F1b)

specificity0 <- (T2 + T3) / (T2 + T3 + F2a + F3a)

sensitivity2 <- T2 / (T2 + F2a + F2b)

specificity2 <- (T1 + T3) / (T1 + T3 + F1a + F3a)

sensitivity4 <- T3 / (T3 + F3a + F3b)

specificity4 <- (T2 + T1) / (T2 + T1 + F2a + F1a)

accuracy <- (T1 + T2 + T3) / (T1 + T2 + T3 + F1a + F1b + F2a + F2b + F3a + F3b)

print(t)

print(paste("Sensitivity 0 = ", sensitivity0))

print(paste("Specificify 0 = ", specificity0))

print(paste("Sensitivity 2 = ", sensitivity2))

print(paste("Specificify 2 = ", specificity2))

print(paste("Sensitivity 4 = ", sensitivity4))

print(paste("Specificify 4 = ", specificity4))

print(paste("Accuracy = ", accuracy))

}

> conf(testData$Sentiment, 0.0)

predicttest

n 0 2 4

0 47 3 3

2 24 17 1

4 29 6 20

[1] "Sensitivity 0 = 0.47"

[1] "Specificify 0 = 0.86046511627907"

[1] "Sensitivity 2 = 0.653846153846154"

[1] "Specificify 2 = 0.712765957446808"

[1] "Sensitivity 4 = 0.833333333333333"

[1] "Specificify 4 = 0.703296703296703"

[1] "Accuracy = 0.56"

Nu de confusion matrix 3x3 is maakt dit het een stuk lastiger om te lezen, en te weten wat wat betekent. Hiervoor moets het script ook opnieuw geschreven worden.