M8\_L1\_RomilShah

Romil Shah

July 16, 2016

library(RTextTools)

## Warning: package 'RTextTools' was built under R version 3.2.5

## Loading required package: SparseM

##   
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':  
##   
## backsolve

library(tm)

## Warning: package 'tm' was built under R version 3.2.5

## Loading required package: NLP

library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.2.5

## Loading required package: RColorBrewer

seuss <- c("You have brains in your head.",  
 "You have feet in your shoes.",   
 "You can steer yourself any direction you choose.",  
 "You're on your own.",   
 "And you know what you know.",   
 "And YOU are the one who'll decide where to go...",  
 "- Dr. Seuss")  
  
#Term by Document matrix  
myCorpus <- Corpus(VectorSource(seuss))  
myTDM <- t(TermDocumentMatrix(myCorpus))  
inspect(myTDM)

## <<DocumentTermMatrix (documents: 7, terms: 28)>>  
## Non-/sparse entries: 36/160  
## Sparsity : 82%  
## Maximal term length: 9  
## Weighting : term frequency (tf)  
##   
## Terms  
## Docs and any are brains can choose. decide direction dr. feet go... have  
## 1 0 0 0 1 0 0 0 0 0 0 0 1  
## 2 0 0 0 0 0 0 0 0 0 1 0 1  
## 3 0 1 0 0 1 1 0 1 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 0 0 0  
## 5 1 0 0 0 0 0 0 0 0 0 0 0  
## 6 1 0 1 0 0 0 1 0 0 0 1 0  
## 7 0 0 0 0 0 0 0 0 1 0 0 0  
## Terms  
## Docs head. know know. one own. seuss shoes. steer the what where who'll  
## 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 1 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 1 0 0 0 0  
## 4 0 0 0 0 1 0 0 0 0 0 0 0  
## 5 0 1 1 0 0 0 0 0 0 1 0 0  
## 6 0 0 0 1 0 0 0 0 1 0 1 1  
## 7 0 0 0 0 0 1 0 0 0 0 0 0  
## Terms  
## Docs you you're your yourself  
## 1 1 0 1 0  
## 2 1 0 1 0  
## 3 2 0 0 1  
## 4 0 1 1 0  
## 5 2 0 0 0  
## 6 1 0 0 0  
## 7 0 0 0 0

#Assuming each sentence is a separate document, the term by document matrix is formed. We can observe the number of times a word comes in each document.  
  
  
#Calculate td-tdf for three terms  
terms <- DocumentTermMatrix(myCorpus,control=list(weighting = function(x) weightTfIdf(x, normalize = FALSE)))  
mtx <- as.matrix(terms)  
frequency <- sort(colSums(mtx), decreasing = TRUE)  
  
tdidf.brains <- frequency[4]  
tdidf.feet <- frequency[13]  
tdidf.your <- frequency[1]  
tdidf.brains

## you   
## 3.397988

tdidf.feet

## feet   
## 2.807355

tdidf.your

## your   
## 3.667177

#Here the td-idf for 3 terms is calculated using the document term matrix and finding the frequency probability of each. Here I have obtained the td-idf for 'brain', 'feet' and 'you' terms.  
  
  
#Regex for segmenting into separate sentences  
seuss\_sent <- c("You have brains in your head.You have feet in your shoes.You can steer yourself any direction you choose.You're on your own.And you know what you know.And YOU are the one who'll decide where to go...- Dr. Seuss")  
sentences <- strsplit(seuss, "[.]+")  
sentences

## [[1]]  
## [1] "You have brains in your head"  
##   
## [[2]]  
## [1] "You have feet in your shoes"  
##   
## [[3]]  
## [1] "You can steer yourself any direction you choose"  
##   
## [[4]]  
## [1] "You're on your own"  
##   
## [[5]]  
## [1] "And you know what you know"  
##   
## [[6]]  
## [1] "And YOU are the one who'll decide where to go"  
##   
## [[7]]  
## [1] "- Dr" " Seuss"

#Regular expression to separate into sentences. This is done on the basis of splitting when '.' is encountered in the paragraph.  
  
  
#Tokenize quote  
tokens <- strsplit(seuss\_sent, " ")  
tokens

## [[1]]  
## [1] "You" "have" "brains" "in"   
## [5] "your" "head.You" "have" "feet"   
## [9] "in" "your" "shoes.You" "can"   
## [13] "steer" "yourself" "any" "direction"   
## [17] "you" "choose.You're" "on" "your"   
## [21] "own.And" "you" "know" "what"   
## [25] "you" "know.And" "YOU" "are"   
## [29] "the" "one" "who'll" "decide"   
## [33] "where" "to" "go...-" "Dr."   
## [37] "Seuss"

#Tokenization is done by splitting the sentences into different words on based on spaces between them. Thus when a space is found, it is split into the resultant words.  
  
  
#Frequency Signature  
temp <- inspect(myTDM)

## <<DocumentTermMatrix (documents: 7, terms: 28)>>  
## Non-/sparse entries: 36/160  
## Sparsity : 82%  
## Maximal term length: 9  
## Weighting : term frequency (tf)  
##   
## Terms  
## Docs and any are brains can choose. decide direction dr. feet go... have  
## 1 0 0 0 1 0 0 0 0 0 0 0 1  
## 2 0 0 0 0 0 0 0 0 0 1 0 1  
## 3 0 1 0 0 1 1 0 1 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 0 0 0  
## 5 1 0 0 0 0 0 0 0 0 0 0 0  
## 6 1 0 1 0 0 0 1 0 0 0 1 0  
## 7 0 0 0 0 0 0 0 0 1 0 0 0  
## Terms  
## Docs head. know know. one own. seuss shoes. steer the what where who'll  
## 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 1 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 1 0 0 0 0  
## 4 0 0 0 0 1 0 0 0 0 0 0 0  
## 5 0 1 1 0 0 0 0 0 0 1 0 0  
## 6 0 0 0 1 0 0 0 0 1 0 1 1  
## 7 0 0 0 0 0 1 0 0 0 0 0 0  
## Terms  
## Docs you you're your yourself  
## 1 1 0 1 0  
## 2 1 0 1 0  
## 3 2 0 0 1  
## 4 0 1 1 0  
## 5 2 0 0 0  
## 6 1 0 0 0  
## 7 0 0 0 0

freqsign <- data.frame(ST=rownames(temp),Freq = rowSums(temp))  
row.names(freqsign) <- NULL  
freqsign

## ST Freq  
## 1 1 5  
## 2 2 5  
## 3 3 8  
## 4 4 3  
## 5 5 6  
## 6 6 9  
## 7 7 2

#The total frequncy of each word in the documents is obtained and then added together in order to find the frequency signature of the documents for the quote.