Module8L1

Neha Parulekar

March 21, 2016

# Additional packages needed

install.packages("tm");  
install.packages("wordcloud"); install.packages("RTextTools");

require(tm)

## Loading required package: tm

## Loading required package: NLP

require(wordcloud)

## Loading required package: wordcloud

## Warning: package 'wordcloud' was built under R version 3.2.4

## Loading required package: RColorBrewer

require(RTextTools)

## Loading required package: RTextTools

## Warning: package 'RTextTools' was built under R version 3.2.4

## Loading required package: SparseM

##   
## Attaching package: 'SparseM'

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

# Data

You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own, and you know what you know. And you are the guy who'll decide where to go.  
  
- Dr. Seuss

**Perform the following tasks (either by hand or in R):**

***Create a term by document matrix for the Dr. Seuss quote. Assume each sentence is a new documment.*****Calculate the td-idf for three terms in the text. Assume each sentence is a new documment.**

calculating the td-idf, we get respective td-idf for *brains = 1.160964* can = 0.580482 \*youre = 0.773976

SeussQuote <- c(D1 = "You have brains in your head.",  
 D2 = "You have feet in your shoes.",   
 D3 = "You can steer yourself in any direction you choose.",  
 D4 = "You're on your own, and you know what you know.",  
 D5 = "And you are the guy who'll decide where to go.")  
  
# convert these documents into a corpus.  
SeussQuoteCor <- Corpus(VectorSource(SeussQuote))  
# Getting the summary of the data  
summary(SeussQuoteCor)

## Length Class Mode  
## 1 2 PlainTextDocument list  
## 2 2 PlainTextDocument list  
## 3 2 PlainTextDocument list  
## 4 2 PlainTextDocument list  
## 5 2 PlainTextDocument list

# cleaning things up a bit.   
# Remove the white space  
SeussQuoteCor <- tm\_map(SeussQuoteCor, stripWhitespace)  
# First we convert all of the text to lowercase  
SeussQuoteCor <- tm\_map(SeussQuoteCor,content\_transformer(tolower))  
# then remove punctuation  
SeussQuoteCor <- tm\_map(SeussQuoteCor, removePunctuation)  
# remove common English stopwords.(if any)  
SeussQuoteCor <- tm\_map(SeussQuoteCor, removeWords, stopwords("english"))  
  
# have a look at what's left by inspecting.  
inspect(SeussQuoteCor)

## <<VCorpus>>  
## Metadata: corpus specific: 0, document level (indexed): 0  
## Content: documents: 5  
##   
## $D1  
## <<PlainTextDocument>>  
## Metadata: 7  
## Content: chars: 15  
##   
## $D2  
## <<PlainTextDocument>>  
## Metadata: 7  
## Content: chars: 14  
##   
## $D3  
## <<PlainTextDocument>>  
## Metadata: 7  
## Content: chars: 31  
##   
## $D4  
## <<PlainTextDocument>>  
## Metadata: 7  
## Content: chars: 22  
##   
## $D5  
## <<PlainTextDocument>>  
## Metadata: 7  
## Content: chars: 25

# Building a Document-Term Matrix.  
TermDocumentMatrix <- TermDocumentMatrix(SeussQuoteCor)  
inspect(TermDocumentMatrix[, 1:5])

## <<TermDocumentMatrix (terms: 13, documents: 5)>>  
## Non-/sparse entries: 13/52  
## Sparsity : 80%  
## Maximal term length: 9  
## Weighting : term frequency (tf)  
##   
## Docs  
## Terms 1 2 3 4 5  
## brains 1 0 0 0 0  
## can 0 0 1 0 0  
## choose 0 0 1 0 0  
## decide 0 0 0 0 1  
## direction 0 0 1 0 0  
## feet 0 1 0 0 0  
## guy 0 0 0 0 1  
## head 1 0 0 0 0  
## know 0 0 0 2 0  
## shoes 0 1 0 0 0  
## steer 0 0 1 0 0  
## wholl 0 0 0 0 1  
## youre 0 0 0 1 0

DocumentTermMatrix <- DocumentTermMatrix(SeussQuoteCor, control = list(weighting = weightTfIdf))  
as.matrix(DocumentTermMatrix)

## Terms  
## Docs brains can choose decide direction feet guy  
## 1 1.160964 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000  
## 2 0.000000 0.000000 0.000000 0.000000 0.000000 1.160964 0.000000  
## 3 0.000000 0.580482 0.580482 0.000000 0.580482 0.000000 0.000000  
## 4 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000  
## 5 0.000000 0.000000 0.000000 0.773976 0.000000 0.000000 0.773976  
## Terms  
## Docs head know shoes steer wholl youre  
## 1 1.160964 0.000000 0.000000 0.000000 0.000000 0.000000  
## 2 0.000000 0.000000 1.160964 0.000000 0.000000 0.000000  
## 3 0.000000 0.000000 0.000000 0.580482 0.000000 0.000000  
## 4 0.000000 1.547952 0.000000 0.000000 0.000000 0.773976  
## 5 0.000000 0.000000 0.000000 0.000000 0.773976 0.000000

\***Write a regular expression to segment the Dr. Seuss quote in to seperate sentences.**

Quote <- c("You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own,and you know what you know. And you are the guy who'll decide where to go")  
  
# Printing the quote  
print(Quote)

## [1] "You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own,and you know what you know. And you are the guy who'll decide where to go"

# segmenting the Dr. Seuss quote in to seperate sentences  
strsplit(Quote, "\\. ")

## [[1]]  
## [1] "You have brains in your head"   
## [2] "You have feet in your shoes"   
## [3] "You can steer yourself in any direction you choose"  
## [4] "You're on your own,and you know what you know"   
## [5] "And you are the guy who'll decide where to go"

\***Write a regular expression to tokenize the Dr. Seuss quote.**

print(Quote)

## [1] "You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own,and you know what you know. And you are the guy who'll decide where to go"

# tokenize the Dr. Seuss quote  
strsplit(Quote, " ")

## [[1]]  
## [1] "You" "have" "brains" "in" "your"   
## [6] "head." "You" "have" "feet" "in"   
## [11] "your" "shoes." "You" "can" "steer"   
## [16] "yourself" "in" "any" "direction" "you"   
## [21] "choose." "You're" "on" "your" "own,and"   
## [26] "you" "know" "what" "you" "know."   
## [31] "And" "you" "are" "the" "guy"   
## [36] "who'll" "decide" "where" "to" "go"

\***Create a frequency signature for the Dr. Seuss quote. Assume each sentence is a new documment.**

Assuming each sentence is a new document. Looking at the pair of words, "brains & Feet", we can see that there is no occurance of brains and feet each without each other and no co-occurance. Also the total no of word count is 14 per tdm. Heance we can say that the frequency signatue would be [0,1,1,14]