Text mining

Prabin Kharel

Making a Corpus to perform text mining

```
#install.packages("tm")
#Uncomment above line to install "tm" package if not already installed. This package is need for text m
library(tm)
```

```
## Loading required package: NLP
```

```
#provide a directory to a variable in which the text documents to be mined are stored
file <- DirSource('txt/')

#making corpus
fileCorpus <- Corpus(file)

#inspecting a corpus. The number inside the [] is the file which we intend to inspect.
inspect(fileCorpus[3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## \nThe continuous updating of data under their management creates a dynamic bank, whose rules are auto
```

Pre-processing of the data involves cleaning or tidying the data. We get rid of the punctuation, numbers, white spaces, typos and stopwords that don't require analysis. Also since R is case-sensitive, it will assume "Read" and "read" to be separate words. To remove this redundancy of words, all the words need to be made in lower case. Also we will stem the words from Corpus.

```

fileCorpus <- tm_map(fileCorpus, stripWhitespace)
inspect(fileCorpus[1])
```

Removing Whitespaces

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## Introduction to artificial neural networks Introduction rhythms arise from stochastic, nonlinear biol
```

```r
fileCorpus <- tm_map(fileCorpus,removePunctuation)
inspect(fileCorpus[1])
```

Removing Punctuation

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## Introduction to artificial neural networks Introduction rhythms arise from stochastic nonlinear biolo
```

```r
fileCorpus <- tm_map(fileCorpus, removeNumbers)
inspect(fileCorpus[1])
```

Removing Numbers

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## Introduction to artificial neural networks Introduction rhythms arise from stochastic nonlinear biolo
```

```r
fileCorpus <- tm_map(fileCorpus, tolower)
inspect(fileCorpus[1])
```

Changing text to lower case

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## introduction to artificial neural networks introduction rhythms arise from stochastic nonlinear biolo
```

```r
stopwords("english")
```

Removing Stopwords

```
##   [1] "i"          "me"         "my"         "myself"     "we"
##   [6] "our"        "ours"       "ourselves"  "you"        "your"
##  [11] "yours"      "yourself"   "yourselves" "he"         "him"
##  [16] "his"        "himself"    "she"        "her"        "hers"
##  [21] "herself"    "it"         "its"        "itself"     "they"
##  [26] "them"       "their"      "theirs"     "themselves" "what"
##  [31] "which"      "who"        "whom"       "this"       "that"
##  [36] "these"      "those"      "am"         "is"         "are"
##  [41] "was"        "were"       "be"         "been"       "being"
##  [46] "have"       "has"        "had"        "having"     "do"
##  [51] "does"       "did"        "doing"      "would"      "should"
##  [56] "could"      "ought"      "i'm"        "you're"     "he's"
##  [61] "she's"      "it's"       "we're"      "they're"    "i've"
##  [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
##  [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
##  [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
##  [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
##  [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
##  [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
##  [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
## [101] "who's"      "what's"     "here's"     "there's"    "when's"
## [106] "where's"    "why's"      "how's"      "a"          "an"
## [111] "the"        "and"        "but"        "if"         "or"
## [116] "because"    "as"         "until"      "while"      "of"
## [121] "at"         "by"         "for"        "with"       "about"
## [126] "against"    "between"    "into"       "through"    "during"
## [131] "before"     "after"      "above"      "below"      "to"
## [136] "from"       "up"         "down"       "in"         "out"
## [141] "on"         "off"        "over"       "under"      "again"
## [146] "further"    "then"       "once"       "here"       "there"
## [151] "when"       "where"      "why"        "how"        "all"
## [156] "any"        "both"       "each"       "few"        "more"
## [161] "most"       "other"      "some"       "such"       "no"
## [166] "nor"        "not"        "only"       "own"        "same"
## [171] "so"         "than"       "too"        "very"
```

```
fileStopwords <- c(stopwords("english"))
fileCorpus <- tm_map(fileCorpus, removeWords, fileStopwords)
inspect(fileCorpus[1])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## introduction  artificial neural networks introduction rhythms arise  stochastic nonlinear biological
```

Stemming Stemming a word means to reduce the word to it's base form or the root form. Since our goal is to get information from the text, we do not need the words repeating in different forms. We would rather prefer to use only 'go' for, 'go','went' and 'gone'.

```
library(SnowballC)
fileCorpusCopy <- fileCorpus
tm_map(fileCorpus, stemDocument)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
```

```
inspect(fileCorpus[1])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## introduction  artificial neural networks introduction rhythms arise  stochastic nonlinear biological
```

Term Document Matrix

A term document matrix gives us the frequency of the terms that occurs in the corpus. We are going to make a term document matrix with words having lengths between 5 and 10.

```
fileTdm <- TermDocumentMatrix(fileCorpus, control =list(wordLengths=c(5,10)))
inspect(fileTdm)
```

```
## <<TermDocumentMatrix (terms: 821, documents: 3)>>
## Non-/sparse entries: 1092/1371
## Sparsity           : 56%
## Maximal term length: 10
## Weighting          : term frequency (tf)
## Sample             :
##             Docs
## Terms         net1.txt net2.txt net3.txt
##   artificial         8        3        5
##   individual         0       14        0
##   input             11        4        3
##   networks          11        8        7
##   neural            11        7        7
##   number             2        8        4
##   problem            2        5       17
##   rules              7        2        4
##   training           1        3       10
##   variables          1       19       19
```

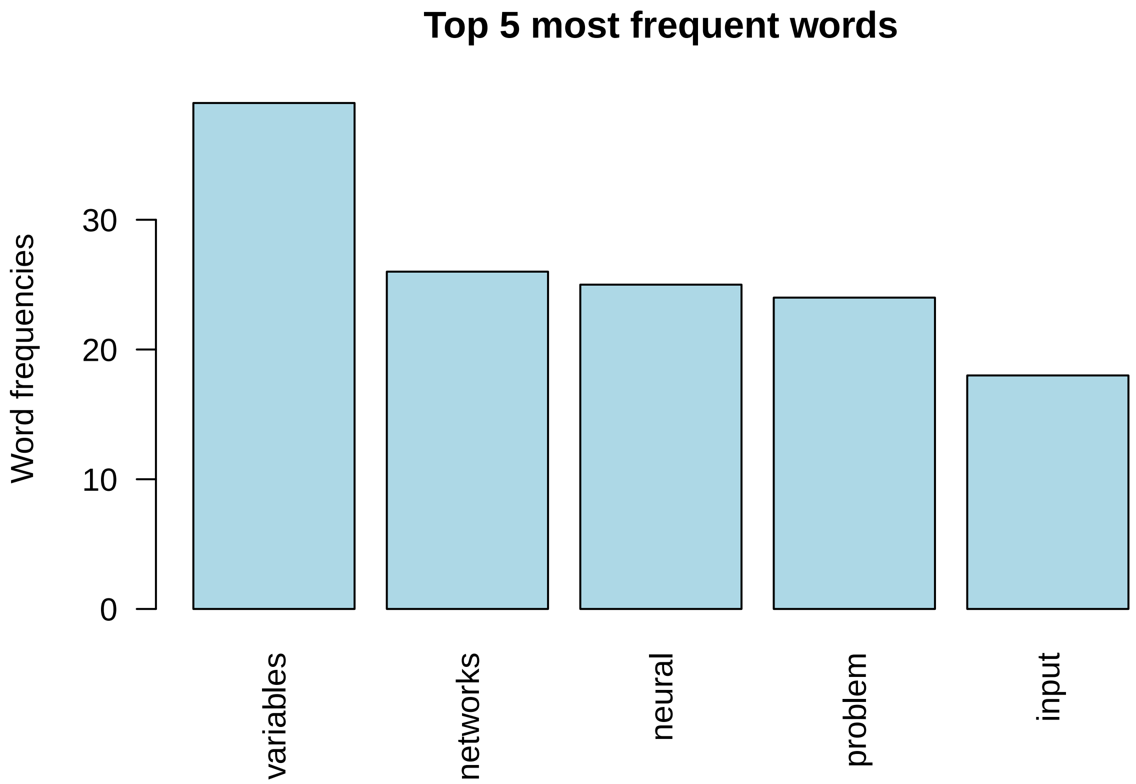
```

Tdm_m <- as.matrix(fileTdm)
# Sort by decreasing frequency
Tdm_f <- sort(rowSums(Tdm_m),decreasing=TRUE)
Tdm_n <- data.frame(word = names(Tdm_f),freq=Tdm_f)
# Inspect the top 5 most frequent words
head(Tdm_n, 5)
```

Plotting top 5 most frequent words

```
##                word freq
## variables variables   39
## networks   networks   26
## neural       neural   25
## problem     problem   24
## input         input   18
```

```
# Plot the most frequent words
barplot(Tdm_n[1:5,]$freq, las = 2, names.arg = Tdm_n[1:5,]$word,
        col ="lightblue", main ="Top 5 most frequent words",
        ylab = "Word frequencies")
```



```r
findAssocs(fileTdm, terms = c("artificial"), corlimit = 0.8)
```

Finding Association with words

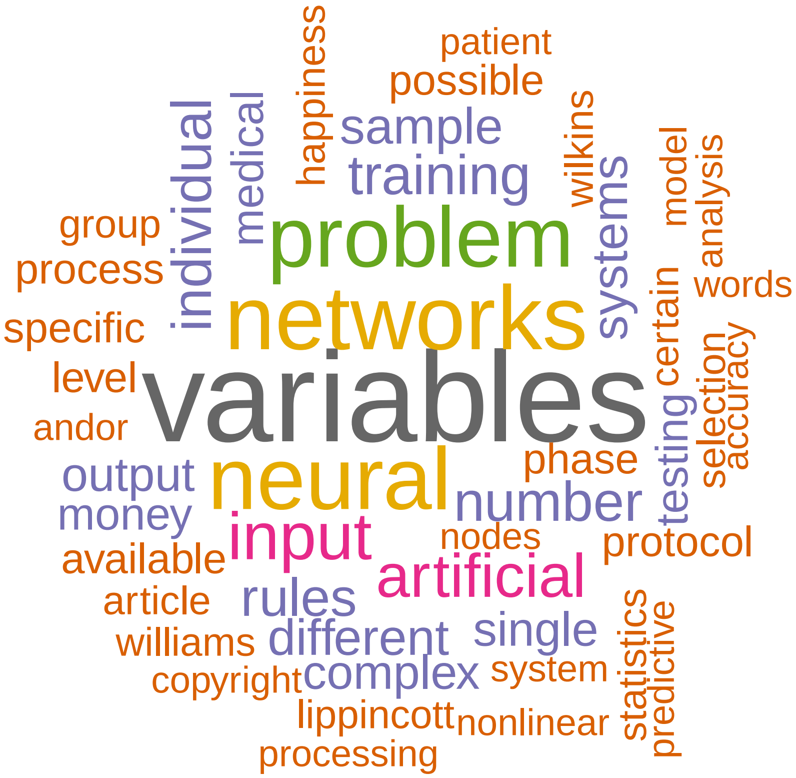
```
## $artificial
##      called      result       rules  understand  connection  represents    problems
##        1.00        1.00        1.00        1.00        0.99        0.99        0.95
##   abilities    accepted  accordance       adapt      adjust    advanced   algorithm
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##       ann's       appro       arise      august       axons       basic      binary
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
## biological       black      blocks      bodily      bracco       brain     briefly
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##       built   calculate     capture    carnegie   cartesian      centre   character
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##       chart    classify      clemen    combined      common       commu   companies
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
## complexity      connec     connect   connected   currently   dendrites  department
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##   depending  determined     diagram     discuss     element    elements       email
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##    emerging    ensemble      entire   equations    everyday  excitatory     excited
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##      facing     finally    flexible       folli     forward   forwarded    function
##        0.92        0.92        0.92        0.92        0.92        0.92        0.92
##     gastro     global     govern     health    imaging  including increasing
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##   indicate  inhibited inhibitory     inside   integral   internal intestinal
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##   involved     issues      italy   keywords     kluwer  knowledge      layer
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##     layers   linkages     living     manner mechanisms      milan     milano
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##   modifies     modify   negative     neural neuralware     neuron    neurons
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##  nications  nonlinear   normally    obtains  organized   packages      pairs
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##    paradox    pattern   patterns     person   positive     priate  processes
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##   progress properties   provides   purposes    receive   received   receives
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##  recompose    related   remained    rhythms     robust    science scientific
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##    semeion     simple   simplify    skilled  solutions    solving   somewhat
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##      sorts     starts stochastic  structure structures   subtypes  technical
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##     things      tions      today      total transforms   trigoria ubiquitous
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##   violates   weighted widespread     within    without    wolters     world’
##       0.92       0.92       0.92       0.92       0.92       0.92       0.92
##    written       ðcxþ        ðxþ       ‘law      ‘real      ‘see’      input
##       0.92       0.92       0.92       0.92       0.92       0.92       0.87
##     output      nodes     amount  addressed      apply   approach      basis
##       0.86       0.84       0.83       0.80       0.80       0.80       0.80
##   behavior capability conditions    dynamic     figure  governing     handle
##       0.80       0.80       0.80       0.80       0.80       0.80       0.80
##     hidden      human      later     learns    limited      makes     nature
##       0.80       0.80       0.80       0.80       0.80       0.80       0.80
##     needed      offer    popular   quantity  relations   strength      terms
##       0.80       0.80       0.80       0.80       0.80       0.80       0.80
##      times   valuable     weight    whether
##       0.80       0.80       0.80       0.80
```

Generating Wordcloud

```
#install.packages("wordcloud")
library(wordcloud)
```

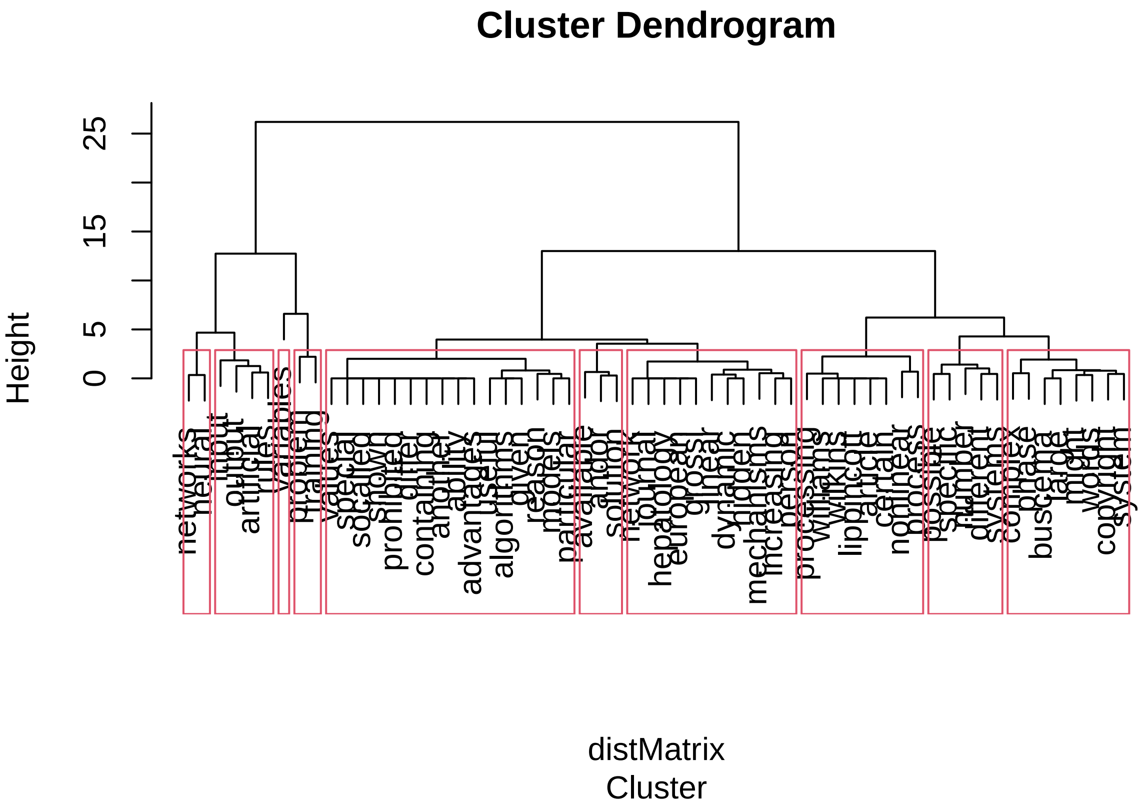
```
## Loading required package: RColorBrewer
```

```
set.seed(1234)
wordcloud(words = names(Tdm_f), freq = Tdm_f, min.freq = 7, random.order=F, rot.per= 0.3,max.words = 100
```



Clustering of Words

```
# remove sparse terms
fileTdm2 <- removeSparseTerms(fileTdm,sparse = 0.3)
m2 <- as.matrix(fileTdm2)
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward.D")
plot(fit, sub = "Cluster")
# cut tree into 10 clusters
rect.hclust(fit, k=10)
```


Cluster Dendrogram
Height
0
5
15
25
networks
neural
input
output
artificial
rules
variables
problem
training
values
special
socalled
shown
prohibited
filter
containing
another
ability
advantages
useful
algorithms
given
reason
models
particular
available
author
solution
network
journal
hepatology
european
glossar
linear
dynamic
hidden
mechanisms
increasing
person
processing
williams
wilkins
lippincott
article
certain
nonlinear
process
possible
specific
number
different
systems
complex
phase
buscema
target
might
words
copyright
system
distMatrix
Cluster