TM_LAB02_tmPackage

October 25, 2018

0.0.1 tm Package

0.1

01. (tm) ### 02. Corpus() corpus ### 03. ### 04. tm_map ### 05.

0.1.1 01. (tm)

- KoNLP tm .
- tm .
- tm corpus(,) .

```
In [2]: #!install.packages("tm")
        library(tm)
```

```
Loading required package: NLP
```

0.1.2 02. Corpus() corpus

- (: corpus, : corpora) ()
- ,
- tm

```
In [3]: library(tm)
        docs <- c("I am boy", "You are a girl", "I am a student")
        is(docs)
```

1. 'character' 2. 'vector' 3. 'data.frameRowLabels' 4. 'SuperClassMethod'

```
In [4]: VectorSource(docs)
```

```
$encoding
[1] ""

$length
[1] 3

$position
```

```
[1] 0

$reader
function (elem, language, id)
{
    if (!is.null(elem$uri))
        id <- basename(elem$uri)
    PlainTextDocument(elem$content, id = id, language = language)
}
<environment: namespace:tm>

$content
[1] "I am boy"       "You are a girl" "I am a student"

attr(,"class")
[1] "VectorSource" "SimpleSource" "Source"
```

```
In [5]: ##
        print(Corpus(VectorSource(docs)))
        myCorpus <- Corpus(VectorSource(docs))
        is(myCorpus)
```

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

'SimpleCorpus'

```
In [6]: print(myCorpus[1:3])
        print(myCorpus[[1]])  # "I am boy"        ->
        print(myCorpus[[2]])  # "You are a girl"  ->
        print(myCorpus[[3]])  # "I am a student"  ->
```

```
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 3
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14
```

```
In [18]: ## Corpus
         inspect(myCorpus[1:3])
```

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

[1] I am boy       You are a girl I am a student
```

0.1.3 03.

```
In [15]: setwd("C:/Users/WITHJS/Documents/R/00_TM_TextMining")
         textMining = readLines("D:/dataset/textMining/anb-jarena-lee.txt")
         is(textMining)
         print(textMining)
```

```
Warning message in readLines("D:/dataset/textMining/anb-jarena-lee.txt"):
"'D:/dataset/textMining/anb-jarena-lee.txt'    "
```

1. 'character' 2. 'vector' 3. 'data.frameRowLabels' 4. 'SuperClassMethod'

```
 [1] "In 1804, after several months of profound spiritual anxiety, Jarena Lee"
 [2] "moved from New Jersey to Philadelphia. There she labored as a domestic"
 [3] "and worshiped among white congregations of Roman Catholics and mixed"
 [4] "congregations of Methodists. On hearing an inspired sermon by the"
 [5] "Reverend Richard Allen, founder of the Bethel African Methodist"
 [6] "Episcopal Church, Lee joined the Methodists. She was baptized in 1807."
 [7] "Prior to her baptism, she experienced the various physical and emotional"
 [8] "stages of conversion terrifying visions of demons and eternal"
 [9] "perdition; extreme feelings of ecstasy and depression; protracted"
[10] "periods of meditation, fasting, and prayer; ennui and fever; energy and"
[11] "vigor. In 1811 she married Joseph Lee, who pastored an African-American"
[12] "church in Snow Hill, New Jersey. They had six children, four of whom"
[13] "died in infancy."
```

0.1.4 04. tm_map

- tm_map(cor1, stripWhitespace) #
- tm_map(cor1, tolower) #
- tm_map(cor1, removeNumbers) #
- tm_map(cor1, removePunctuation) # , , ,

. tm_map

```
In [16]: myCorpus = Corpus(VectorSource(textMining))
         myCorpus <- tm_map(myCorpus, stripWhitespace)
         myCorpus <- tm_map(myCorpus, tolower)
```

```
          myCorpus <- tm_map(myCorpus, removePunctuation)
          myCorpus <- tm_map(myCorpus, removeNumbers)
          myCorpus <- tm_map(myCorpus, removeWords, stopwords("english"))  #

In [17]: stopword2 <- c(stopwords('en'), "and", "but", "not")  #
          myCorpus <- tm_map(myCorpus, removeWords, stopword2 )
```

. Term-Document Matrix

```
* (document-term matrix) or   (term-document matrix)        .
*       ,

In [18]: str(myCorpus)

List of 13
 $ 1 :List of 2
  ..$ content: chr "   several months  profound spiritual anxiety jarena lee"
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "1"
  .. ..$ language     : chr "en"
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 $ 2 :List of 2
  ..$ content: chr "moved  new jersey  philadelphia   labored   domestic"
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "2"
  .. ..$ language     : chr "en"
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 $ 3 :List of 2
  ..$ content: chr " worshiped among white congregations  roman catholics  mixed"
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "3"
  .. ..$ language     : chr "en"
```

```
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 4 :List of 2
 ..$ content: chr "congregations  methodists  hearing  inspired sermon  "
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "4"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 5 :List of 2
 ..$ content: chr "reverend richard allen founder   bethel african methodist"
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "5"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 6 :List of 2
 ..$ content: chr "episcopal church lee joined  methodists   baptized  "
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "6"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 7 :List of 2
 ..$ content: chr "prior   baptism  experienced  various physical  emotional"
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "7"
 .. ..$ language     : chr "en"
```

```
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 8 :List of 2
 ..$ content: chr "stages  conversion terrifying visions  demons  eternal"
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "8"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 9 :List of 2
 ..$ content: chr "perdition extreme feelings  ecstasy  depression protracted"
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "9"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 10:List of 2
 ..$ content: chr "periods  meditation fasting  prayer ennui  fever energy "
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "10"
 .. ..$ language     : chr "en"
 .. ..$ origin       : chr(0)
 .. ..- attr(*, "class")= chr "TextDocumentMeta"
 ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
$ 11:List of 2
 ..$ content: chr "vigor    married joseph lee  pastored  africanamerican"
 ..$ meta   :List of 7
 .. ..$ author       : chr(0)
 .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
 .. ..$ description  : chr(0)
 .. ..$ heading      : chr(0)
 .. ..$ id           : chr "11"
 .. ..$ language     : chr "en"
```

```
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 $ 12:List of 2
  ..$ content: chr "church  snow hill new jersey   six children four  "
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "12"
  .. ..$ language     : chr "en"
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 $ 13:List of 2
  ..$ content: chr "died  infancy"
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2018-10-24 15:15:24"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "13"
  .. ..$ language     : chr "en"
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 - attr(*, "class")= chr [1:2] "SimpleCorpus" "Corpus"
```

```
In [21]: tdm <- TermDocumentMatrix(myCorpus)
         tdm
         # terms  72 , documents:13  ( 13 .)
         # sparsit 92% tdm 92% 0 .
```

```
<<TermDocumentMatrix (terms: 72, documents: 13)>>
Non-/sparse entries: 79/857
Sparsity           : 92%
Maximal term length: 15
Weighting          : term frequency (tf)
```

```
In [22]: # tdm(Term-Document Matrix) tm    .
         #       Matrix .
         m <- as.matrix(tdm)
         m
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
anxiety	1	0	0	0	0	0	0	0	0	0	0	0	0
jarena	1	0	0	0	0	0	0	0	0	0	0	0	0
lee	1	0	0	0	0	1	0	0	0	0	1	0	0
months	1	0	0	0	0	0	0	0	0	0	0	0	0
profound	1	0	0	0	0	0	0	0	0	0	0	0	0
several	1	0	0	0	0	0	0	0	0	0	0	0	0
spiritual	1	0	0	0	0	0	0	0	0	0	0	0	0
domestic	0	1	0	0	0	0	0	0	0	0	0	0	0
jersey	0	1	0	0	0	0	0	0	0	0	0	1	0
labored	0	1	0	0	0	0	0	0	0	0	0	0	0
moved	0	1	0	0	0	0	0	0	0	0	0	0	0
new	0	1	0	0	0	0	0	0	0	0	0	1	0
philadelphia	0	1	0	0	0	0	0	0	0	0	0	0	0
among	0	0	1	0	0	0	0	0	0	0	0	0	0
catholics	0	0	1	0	0	0	0	0	0	0	0	0	0
congregations	0	0	1	1	0	0	0	0	0	0	0	0	0
mixed	0	0	1	0	0	0	0	0	0	0	0	0	0
roman	0	0	1	0	0	0	0	0	0	0	0	0	0
white	0	0	1	0	0	0	0	0	0	0	0	0	0
worshiped	0	0	1	0	0	0	0	0	0	0	0	0	0
hearing	0	0	0	1	0	0	0	0	0	0	0	0	0
inspired	0	0	0	1	0	0	0	0	0	0	0	0	0
methodists	0	0	0	1	0	1	0	0	0	0	0	0	0
sermon	0	0	0	1	0	0	0	0	0	0	0	0	0
african	0	0	0	0	1	0	0	0	0	0	0	0	0
allen	0	0	0	0	1	0	0	0	0	0	0	0	0
bethel	0	0	0	0	1	0	0	0	0	0	0	0	0
founder	0	0	0	0	1	0	0	0	0	0	0	0	0
methodist	0	0	0	0	1	0	0	0	0	0	0	0	0
reverend	0	0	0	0	1	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
demons	0	0	0	0	0	0	0	1	0	0	0	0	0
eternal	0	0	0	0	0	0	0	1	0	0	0	0	0
stages	0	0	0	0	0	0	0	1	0	0	0	0	0
terrifying	0	0	0	0	0	0	0	1	0	0	0	0	0
visions	0	0	0	0	0	0	0	1	0	0	0	0	0
depression	0	0	0	0	0	0	0	0	1	0	0	0	0
ecstasy	0	0	0	0	0	0	0	0	1	0	0	0	0
extreme	0	0	0	0	0	0	0	0	1	0	0	0	0
feelings	0	0	0	0	0	0	0	0	1	0	0	0	0
perdition	0	0	0	0	0	0	0	0	1	0	0	0	0
protracted	0	0	0	0	0	0	0	0	1	0	0	0	0
energy	0	0	0	0	0	0	0	0	0	1	0	0	0
ennui	0	0	0	0	0	0	0	0	0	1	0	0	0
fasting	0	0	0	0	0	0	0	0	0	1	0	0	0
fever	0	0	0	0	0	0	0	0	0	1	0	0	0
meditation	0	0	0	0	0	0	0	0	0	1	0	0	0
periods	0	0	0	0	0	0	0	0	0	1	0	0	0
prayer	0	0	0	0	0	0	0	0	0	1	0	0	0
africanamerican	0	0	0	0	0	0	0	0	0	0	1	0	0
joseph	0	0	0	0	0	0	0	0	0	0	1	0	0
married	0	0	0	0	0	0	0	0	0	0	1	0	0

```
In [23]: print(textMining)
```

```
 [1] "In 1804, after several months of profound spiritual anxiety, Jarena Lee"
 [2] "moved from New Jersey to Philadelphia. There she labored as a domestic"
 [3] "and worshiped among white congregations of Roman Catholics and mixed"
 [4] "congregations of Methodists. On hearing an inspired sermon by the"
 [5] "Reverend Richard Allen, founder of the Bethel African Methodist"
 [6] "Episcopal Church, Lee joined the Methodists. She was baptized in 1807."
 [7] "Prior to her baptism, she experienced the various physical and emotional"
 [8] "stages of conversion terrifying visions of demons and eternal"
 [9] "perdition; extreme feelings of ecstasy and depression; protracted"
[10] "periods of meditation, fasting, and prayer; ennui and fever; energy and"
[11] "vigor. In 1811 she married Joseph Lee, who pastored an African-American"
[12] "church in Snow Hill, New Jersey. They had six children, four of whom"
[13] "died in infancy."
```

0.1.5

- lee 1 6 .

```
In [25]: ###      .
         stopword2 <- c(stopwords('en'), "new", "among", "ennui")  #
         myCorpus <- tm_map(myCorpus, removeWords, stopword2 )
         tdm2 <- TermDocumentMatrix(myCorpus)
         tdm2
```

```
<<TermDocumentMatrix (terms: 69, documents: 13)>>
Non-/sparse entries: 75/822
Sparsity           : 92%
Maximal term length: 15
Weighting          : term frequency (tf)
```

```
In [26]: ##
         m2 <- as.matrix(tdm2)
         m2
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
anxiety	1	0	0	0	0	0	0	0	0	0	0	0	0
jarena	1	0	0	0	0	0	0	0	0	0	0	0	0
lee	1	0	0	0	0	1	0	0	0	0	1	0	0
months	1	0	0	0	0	0	0	0	0	0	0	0	0
profound	1	0	0	0	0	0	0	0	0	0	0	0	0
several	1	0	0	0	0	0	0	0	0	0	0	0	0
spiritual	1	0	0	0	0	0	0	0	0	0	0	0	0
domestic	0	1	0	0	0	0	0	0	0	0	0	0	0
jersey	0	1	0	0	0	0	0	0	0	0	0	1	0
labored	0	1	0	0	0	0	0	0	0	0	0	0	0
moved	0	1	0	0	0	0	0	0	0	0	0	0	0
philadelphia	0	1	0	0	0	0	0	0	0	0	0	0	0
catholics	0	0	1	0	0	0	0	0	0	0	0	0	0
congregations	0	0	1	1	0	0	0	0	0	0	0	0	0
mixed	0	0	1	0	0	0	0	0	0	0	0	0	0
roman	0	0	1	0	0	0	0	0	0	0	0	0	0
white	0	0	1	0	0	0	0	0	0	0	0	0	0
worshiped	0	0	1	0	0	0	0	0	0	0	0	0	0
hearing	0	0	0	1	0	0	0	0	0	0	0	0	0
inspired	0	0	0	1	0	0	0	0	0	0	0	0	0
methodists	0	0	0	1	0	1	0	0	0	0	0	0	0
sermon	0	0	0	1	0	0	0	0	0	0	0	0	0
african	0	0	0	0	1	0	0	0	0	0	0	0	0
allen	0	0	0	0	1	0	0	0	0	0	0	0	0
bethel	0	0	0	0	1	0	0	0	0	0	0	0	0
founder	0	0	0	0	1	0	0	0	0	0	0	0	0
methodist	0	0	0	0	1	0	0	0	0	0	0	0	0
reverend	0	0	0	0	1	0	0	0	0	0	0	0	0
richard	0	0	0	0	1	0	0	0	0	0	0	0	0
baptized	0	0	0	0	0	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
conversion	0	0	0	0	0	0	0	1	0	0	0	0	0
demons	0	0	0	0	0	0	0	1	0	0	0	0	0
eternal	0	0	0	0	0	0	0	1	0	0	0	0	0
stages	0	0	0	0	0	0	0	1	0	0	0	0	0
terrifying	0	0	0	0	0	0	0	1	0	0	0	0	0
visions	0	0	0	0	0	0	0	1	0	0	0	0	0
depression	0	0	0	0	0	0	0	0	1	0	0	0	0
ecstasy	0	0	0	0	0	0	0	0	1	0	0	0	0
extreme	0	0	0	0	0	0	0	0	1	0	0	0	0
feelings	0	0	0	0	0	0	0	0	1	0	0	0	0
perdition	0	0	0	0	0	0	0	0	1	0	0	0	0
protracted	0	0	0	0	0	0	0	0	1	0	0	0	0
energy	0	0	0	0	0	0	0	0	0	1	0	0	0
fasting	0	0	0	0	0	0	0	0	0	1	0	0	0
fever	0	0	0	0	0	0	0	0	0	1	0	0	0
meditation	0	0	0	0	0	0	0	0	0	1	0	0	0
periods	0	0	0	0	0	0	0	0	0	1	0	0	0
prayer	0	0	0	0	0	0	0	0	0	1	0	0	0
africanamerican	0	0	0	0	0	0	0	0	0	0	1	0	0
joseph	0	0	0	0	0	0	0	0	0	0	1	0	0
married	0	0	0	0	0	0	0	0	0	0	1	0	0

0.1.6

```
In [46]: library(RColorBrewer)
         wordFreq <- sort(rowSums(m2), decreasing=TRUE)
         head(wordFreq, 20)
```

lee 3 **jersey** 2 **congregations** 2 **methodists** 2 **church** 2 **anxiety** 1 **jarena** 1 **months** 1 **profound** 1 **several** 1 **spiritual** 1 **domestic** 1 **labored** 1 **moved** 1 **philadelphia** 1 **catholics** 1 **mixed** 1 **roman** 1 **white** 1 **worshiped** 1

```
In [47]: ##
         wordFreq2 <- sort(colSums(m2), decreasing=TRUE)
         wordFreq2
```

1 7 **5** 7 **12** 7 **3** 6 **6** 6 **7** 6 **8** 6 **9** 6 **10** 6 **11** 6 **2** 5 **4** 5 **13** 2

0.1.7

```
# findFreqTerms(x, lowfreq, highfreq)
# x: term-document
# lowfreq :
# highfreq :
```

```
In [48]: findFreqTerms(tdm2, lowfreq=2, highfreq=Inf)
```

1. 'lee' 2. 'jersey' 3. 'congregations' 4. 'methodists' 5. 'church'

0.1.8

- findAssocs() :
-
- findAssocs(x, term, corlimit)
- x : term-document
- term :
- corlimit :

```
In [44]: findAssocs(tdm2, "jersey", 0.2)   # jersey 0.2
```

$jersey = **domestic** 0.68 **labored** 0.68 **moved** 0.68 **philadelphia** 0.68 **children** 0.68 **four** 0.68 **hill** 0.68 **six** 0.68 **snow** 0.68 **church** 0.41

0.1.9 05.

```
In [39]: library(RColorBrewer)
         library(wordcloud)
         search()
```

1. '.GlobalEnv' 2. 'package:wordcloud' 3. 'package:RColorBrewer' 4. 'package:tm' 5. 'package:NLP' 6. 'jupyter:irkernel' 7. 'package:RevoUtils' 8. 'package:stats' 9. 'package:graphics' 10. 'package:grDevices' 11. 'package:utils' 12. 'package:datasets' 13. 'package:RevoUtilsMath' 14. 'package:methods' 15. 'Autoloads' 16. 'package:base'

```
In [45]: names(wordFreq)
         wordFreq
```

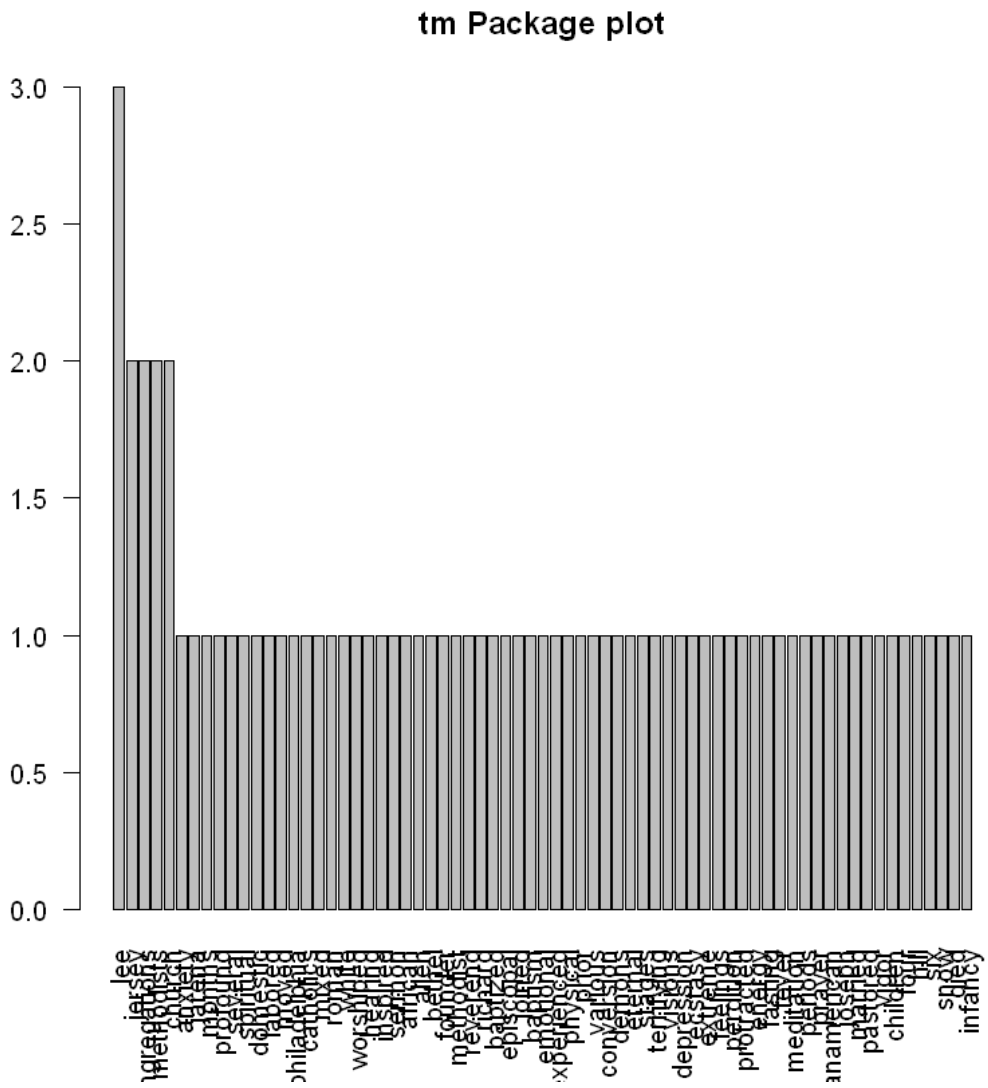
1. '1' 2. '5' 3. '12' 4. '3' 5. '6' 6. '7' 7. '8' 8. '9' 9. '10' 10. '11' 11. '2' 12. '4' 13. '13'

1 7 **5** 7 **12** 7 **3** 6 **6** 6 **7** 6 **8** 6 **9** 6 **10** 6 **11** 6 **2** 5 **4** 5 **13** 2

```
In [49]: set.seed(1234)
         wordF = findFreqTerms(tdm2, lowfreq=1, highfreq=Inf)
         pal = brewer.pal(8, "Dark2")
```

```
In [59]: wordcloud(words=names(wordFreq),
                   freq=wordFreq,
                   scale=c(4, 1),
                   min.freq=2, colors=pal, random.order=F, random.color=T)
         legend(0.3, 1 ,"tm Package Test", cex=1, fill=NA, border=NA, bg='white', text.col='re
```


```
In [62]: barplot(wordFreq, main='tm Package plot', las=2)
```




```
In [65]: testText <- readLines("D:/dataset/textMining/___.txt")  #  text
         print(testText)
         str(testText)
```

```
Warning message in readLines("D:/dataset/textMining/wikiTextming.txt"):
"'D:/dataset/textMining/wikiTextming.txt'    "
```

```
  [1] "From Wikipedia, the free encyclopedia"
  [2] "Text mining, also referred to as text data mining, roughly equivalent to text analytics
  [3] ""
  [4] "Text analysis involves information retrieval, lexical analysis to study word frequency
  [5] ""
  [6] "A typical application is to scan a set of documents written in a natural language and e
```

```
 [7] ""
 [8] "Contents "
 [9] "1\tText analytics"
[10] "2\tFuture"
[11] "3\tText analysis processes"
[12] "4\tApplications"
[13] "4.1\tSecurity applications"
[14] "4.2\tBiomedical applications"
[15] "4.3\tSoftware applications"
[16] "4.4\tOnline media applications"
[17] "4.5\tBusiness and marketing applications"
[18] "4.6\tSentiment analysis"
[19] "4.7\tAcademic applications"
[20] "4.8\tDigital humanities and computational sociology"
[21] "5\tSoftware"
[22] "6\tIntellectual property law"
[23] "6.1\tSituation in Europe"
[24] "6.2\tSituation in the United States"
[25] "7\tImplications"
[26] "8\tSee also"
[27] "9\tReferences"
[28] "9.1\tCitations"
[29] "9.2\tSources"
[30] "10\tExternal links"
[31] "Text analytics"
[32] "The term text analytics describes a set of linguistic, statistical, and machine learning
[33] ""
[34] "The term text analytics also describes that application of text analytics to respond to
[35] ""
[36] "Future"
[37] "Increasing interest is being paid to multilingual data mining: the ability to gain infor
[38] ""
[39] "The challenge of exploiting the large proportion of enterprise information that originat
[40] ""
[41] "\"...utilize data-processing machines for auto-abstracting and auto-encoding of document
[42] ""
[43] "Yet as management information systems developed starting in the 1960s, and as BI emerged
[44] ""
[45] "For almost a decade the computational linguistics community has viewed large text colle
[46] ""
[47] "Hearst's 1999 statement of need fairly well describes the state of text analytics techno
[48] ""
[49] "Text analysis processes"
[50] "Subtasks\xe2\u0080omponents of a larger text-analytics effort\xe2\u0080ypically include
[51] ""
[52] "Information retrieval or identification of a corpus is a preparatory step: collecting or
[53] "Although some text analytics systems apply exclusively advanced statistical methods, mar
[54] "Named entity recognition is the use of gazetteers or statistical techniques to identify
```

```
 [55] "Disambiguation\xe2\u0080he use of contextual clues\xe2\u0080ay be required to decide whe
 [56] "Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail a
 [57] "Coreference: identification of noun phrases and other terms that refer to the same objec
 [58] "Relationship, fact, and event Extraction: identification of associations among entities
 [59] "Sentiment analysis involves discerning subjective (as opposed to factual) material and e
 [60] "Quantitative text analysis is a set of techniques stemming from the social sciences wher
 [61] "Applications"
 [62] "The technology is now broadly applied for a wide variety of government, research, and bu
 [63] ""
 [64] "Enterprise Business Intelligence/Data Mining, Competitive Intelligence"
 [65] "E-Discovery, Records Management"
 [66] "National Security/Intelligence"
 [67] "Scientific discovery, especially Life Sciences"
 [68] "Sentiment Analysis Tools, Listening Platforms"
 [69] "Natural Language/Semantic Toolkit or Service"
 [70] "Publishing"
 [71] "Automated ad placement"
 [72] "Search/Information Access"
 [73] "Social media monitoring"
 [74] "Security applications"
 [75] "Many text mining software packages are marketed for security applications, especially mo
 [76] ""
 [77] "Biomedical applications"
 [78] "Main article: Biomedical text mining"
 [79] "A range of text mining applications in the biomedical literature has been described.[11]
 [80] ""
 [81] "One online text mining application in the biomedical literature is PubGene that combines
 [82] ""
 [83] "GoPubMed is a knowledge-based search engine for biomedical texts."
 [84] ""
 [85] "Software applications"
 [86] "Text mining methods and software is also being researched and developed by major firms,
 [87] ""
 [88] "Online media applications"
 [89] "Text mining is being used by large media companies, such as the Tribune Company, to clar
 [90] ""
 [91] "Business and marketing applications"
 [92] "Text mining is starting to be used in marketing as well, more specifically in analytical
 [93] ""
 [94] "Sentiment analysis"
 [95] "Sentiment analysis may involve analysis of movie reviews for estimating how favorable a
 [96] ""
 [97] "Text has been used to detect emotions in the related area of affective computing.[22] Te
 [98] ""
 [99] "Academic applications"
[100] "The issue of text mining is of importance to publishers who hold large databases of info
[101] ""
[102] "Academic institutions have also become involved in the text mining initiative:"
```

```
[103] ""
[104] "The National Centre for Text Mining (NaCTeM), is the first publicly funded text mining c
[105] "In the United States, the School of Information at University of California, Berkeley is
[106] "The Text Analysis Portal for Research (TAPoR), currently housed at the University of All
[107] "Digital humanities and computational sociology"
[108] "The automatic analysis of vast textual corpora has created the possibility for scholars
[109] ""
[110] ""
[111] "Narrative network of US Elections 2012[26]"
[112] "The automatic parsing of textual corpora has enabled the extraction of actors and their
[113] ""
[114] "Content analysis has been a traditional part of social sciences and media studies for a
[115] ""
[116] "Software"
[117] "Text mining computer programs are available from many commercial and open source compani
[118] ""
[119] "Intellectual property law"
[120] "Situation in Europe"
[121] "File:FixCopyright- Copyright & Research - Text & Data Mining (TDM) Explained.webm"
[122] "Video by Fix Copyright campaign explaining TDM and its copyright issues in the EU, 2016
[123] "Because of a lack of flexibilities in European copyright and database law, the mining of
[124] ""
[125] "The European Commission facilitated stakeholder discussion on text and data mining in 20
[126] ""
[127] "Situation in the United States"
[128] "By contrast to Europe, the flexible nature of US copyright law, and in particular fair u
[129] ""
[130] "Implications"
[131] "Until recently, websites most often used text-based searches, which only found documents
[132] ""
[133] "See also"
[134] "Concept mining"
[135] "Document processing"
[136] "Full text search"
[137] "List of text mining software"
[138] "Market sentiment"
[139] "Name resolution (semantics and text extraction)"
[140] "Named entity recognition"
[141] "News analytics"
[142] "Record linkage"
[143] "Sequential pattern mining (string and sequence mining)"
[144] "w-shingling"
[145] "Web mining, a task that may involve text mining (e.g. first find appropriate web pages b
 chr [1:145] "From Wikipedia, the free encyclopedia" ...
```

0.1.10 2. , Corpus , .