

10: Brief Notes on Text Mining

John H Maindonald

December 18, 2018

```
doFigs <- FALSE
```

Load the *DAAGviz* package:

```
library(DAAGviz, quietly=TRUE)
```

Set up a path to the directory where text (.txt) files are stored:

```
## Create paths to the files
```

```
txdir <- system.file(package='DAAGviz', "texts")
```

```
print(dir(txdir, pattern=".txt$"))
```

```
[1] "data6-7.txt"      "graphics8-9.txt"
```

```
[3] "prelims1-5.txt"
```

```
txfiles <- dir(txdir, pattern=".txt$", full.names=TRUE)
```

Choose a color palette:

```
pal <- RColorBrewer::brewer.pal(6, "Dark2")
```

```
dirSource <- tm::DirSource(directory=txdir,  
                           pattern=".txt$")
```

```
txcorp <- tm::Corpus(dirSource)
```

```
txcorp <- tm::tm_map(txcorp,
```

```
  tm::content_transformer(  
    function(x) iconv(x, to="UTF-8",  
                      sub = "byte")))
```

```

ctl <- list(stopwords = c(tm::stopwords(), "[1]"),
            removePunctuation = list(preserve_intra_word_dashes = FALSE),
            removeNumbers = TRUE, stopwords=c(tm::stopwords(), "[1]"),
            minDocFreq = 2)
tx.tdm <- tm::TermDocumentMatrix(txcorp, control=ctl)

```

```

figset14 <- function(){
  if(!requireNamespace('tm', quietly = TRUE))stop('tm must be installed')
  if(!requireNamespace('wordcloud', quietly=TRUE))stop('wordcloud must be installed')
}

```

```
figset14()
```

```

doTDM <- if(exists("tx.tdm")) FALSE else TRUE
doCorp <- if(doTDM & !exists("tx.corp")) TRUE else FALSE
if(doCorp){
  dirSource <- tm::DirSource(directory=txdir,
                             pattern=".txt$")
  txcorp <- tm::Corpus(dirSource)
  txcorp <- tm::tm_map(txcorp,
                      tm::content_transformer(
                        function(x) iconv(x, to="UTF-8",
                                           sub = "byte")))
}
if(doTDM){
  ctl <- list(stopwords = c(tm::stopwords(), "[1]"),
              removePunctuation = list(preserve_intra_word_dashes = FALSE),
              removeNumbers = TRUE, stopwords=c(tm::stopwords(), "[1]"),
              minDocFreq = 2)
  tx.tdm <- tm::TermDocumentMatrix(txcorp, control=ctl)
}
fig10.1A <- function(){
  fnam1 <- as.matrix(tx.tdm)[,1]
  wordcloud::wordcloud(names(fnam1), fnam1, max.words=80, colors=pal[-1],
                       random.order=FALSE, scale=c(10.5,.5))
}
fig10.1B <- function(){
  fnam2 <- as.matrix(tx.tdm)[,2]
  wordcloud::wordcloud(names(fnam2), fnam2, max.words=80, colors=pal[-1],
                       random.order=FALSE, scale=c(5,.5))
}

```

```
fig10.1A()  
fig10.1B()  
fig10.1C()
```

Figure 1: Wordcloud plots are A: for the words in Chapters 1 - 5; B: 6 - 7; and C: 8 - 9.

```
fig10.1C <- function(){  
  fnam3 <- as.matrix(tx.tdm)[,3]  
  wordcloud::wordcloud(names(fnam3), fnam3, max.words=80, colors=pal[-1],  
    random.order=FALSE, scale=c(6.5,.5))  
}
```

```
fig10.1 <- function(){  
  cat("\n",  
    "Run the separate functions fig10.1A(), fig10.1B() & fig10.1C()",  
    "\n")  
}
```