Text Ming

by Yang Chen

```
install.packages(“tm”)
install.packages(“SnowballC”)
install.packages(“wordcloud”)

library(tm)
library(SnowballC)
library(wordcloud)
```

```
data <-VCorpus(DirSource("C:\\Users\\Desktop\\Corpus"))
```

The main structure for managing documents in tm is a so-called Corpus, representing a collection of text documents. A corpus is an abstract concept.

VCorpus corpora are R objects held fully in memory

PCorpus the documents are physically stored outside of R (e.g., in a database)

```
data <-VCorpus(DirSource("C:\\Users\\Desktop\\Corpus"))
```

Within the corpus constructor, x must be a Source object which abstracts the input location. tm provides a set of predefined sources, e.g., DirSource, VectorSource, or DataframeSource

Data Export

writeCorpus(data, path = “ ”, filename = “ ”)

Transformation

Transformations are done via the tm_map() function which applies (maps) a function to all elements of the corpus.

Transformation

```
data <- tm_map(data, content_transformer(tolower))
```

content_transformer()-----tm 包装函数

```
data <- tm_map(data, removeNumbers)
```

```
data <- tm_map(data, removeWords, stopwords("english"))
```

Transformation

```
data <- tm_map(data, removePunctuation)
removePunctuation(“hello...world”)

replacePunctuation <- function(x) {
  gsub(":[[:punct:]]+", " ", x)
}

data <- tm_map(data, replacePunctuation)
```

Transformation

```
library(SnowballC)

data <- tm_map(data, stemDocument)

data <- tm_map(data, stripWhitespace)
```

Term Document Matrix
& Document Term Matrix

```
data <- data <- tm_map(data, PlainTextDocument)

tdm <- TermDocumentMatrix(data)

dtm <- DocumentTermMatrix(data)
```

wordcloud

wordcloud(words, freq, scale = c(4, .5), min.freq = 3, max.words =
Inf, random.order = TRUE, random.color = FAlSE, rot.per = .1,
colors = “black”, ordered.colors = FALSE, use.r.layout =
FALSE, fixed.asp = TRUE, ...)

wordcloud

word ----关键词列表

freq ---- 关键词对应的词频列表

scale----显示字体大小的范围

min.freq----最小词频

max.words----显示的最大词数量

random.order----词在图上的排列顺序 T：随机排列 F:词按频数从图中心位置往外降序排列

random.color----控制词的字体颜色 T：颜色随机分配 F：根据频数分配字体颜色

wordcloud(data, min.freq = 50)