用定性数据分析包 RQDA tm 进行文本挖掘

Written by Benson Ye (bensonye@189.cn)
Revised by Ronggui Huang (ronggui.huang@gmail.com)
2010-07-22

在对访谈内容或剧本、小说部分内容进行文本挖掘时，如果用不断的剪粘保存的方法非常繁琐而且容易漏掉一些内容。好在黄荣贵开发的 RQDA 包可以进行文档管理和内容编码及提取，大大方便了利用 tm 包进行文本挖掘，既提高了效率又提高了准确性，下面举一个小例子：

对（人民网 >> 时政 >> 时政专题 >> 网友进言）中的公安部回应进行分析

相关链接：http://politics.people.com.cn/GB/8198/138817/index.html

1、安装 RQDA 包、tm 包和中文分词软件；

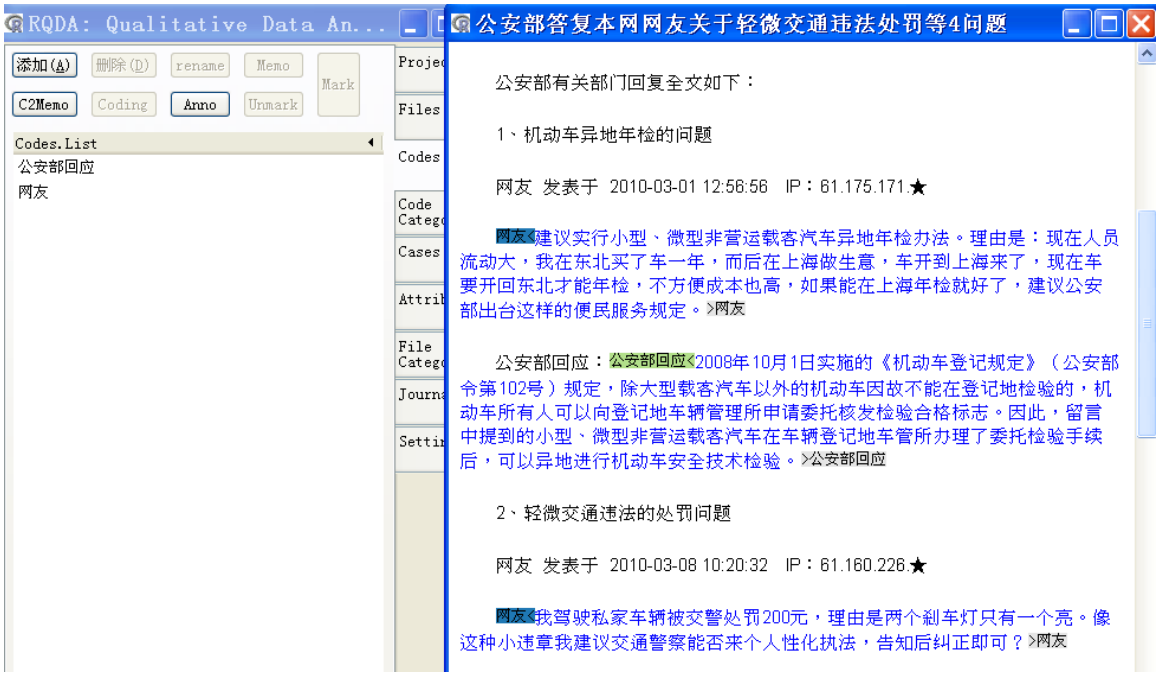
```
> install.packages(c("rJava","tm", "gsubfn"))
> install.packages(c("RQDA","RQDAtm"),repos=" http://R-Forge.R-project.org",type='source')
```

2、装载 RQDA 包并建立一个新的工程项目；

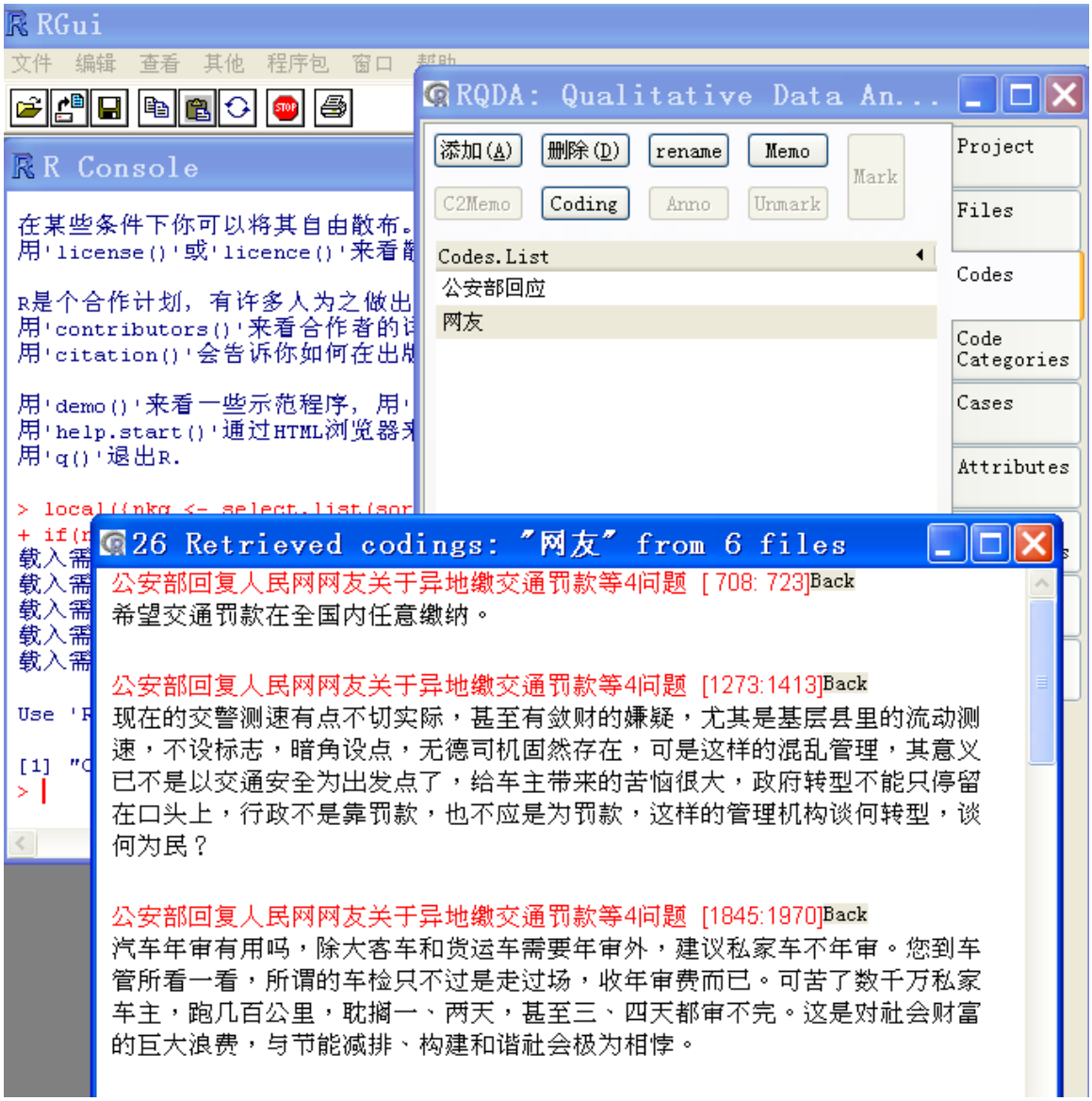
```
> library(RQDAtm)
```

3、输入相关文本文件；

4、进行编码和作标记；



5、双击想要提取的编码即可提取相关文本；



6、运行下面下载的程序进行文本提取、转换、分词、文本挖掘工作。
(以上步骤的结果为 RQDA2tm_example.rqda)，可直接打开该文件继续如下步骤。

```
> gg <- RQDA2tm("公安部回应" )
> summary(gg)
A corpus with 26 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
   create_date creator
Available variables in the data frame are:
   MetaID cid fid selfirst selend fname
> inspect(gg)
----------------------------------------------
> ##    去掉多余空格    ####
```

```
> reuters <- tm_map(gg, stripWhitespace)
> reuters[[3]]
```
公安部规定，县级公安机关交通管理部门车辆管理所可以办理本行政辖区内初次申领和增加准驾车型为低速载货汽车、三轮汽车、普通三轮摩托车、普通二轮摩托车、轻便摩托车的机动车驾驶证业务，具体业务范围和办理条件由省级公安机关交通管理部门确定。目前，全国仅有个别县级车辆管理所受条件限制无法开展增加准驾车型为摩托车的考试业务。

```
------------------------------------------------
> ##    全文搜索      ##
> searchFullText(gg[[1]], "是临[时]?改")
[1] FALSE
------------------------------------------------
> ###    查找以某字开头、结尾等的词条    ###
> stemCompletion(gg, c("机", "交", "证"))
                                                    机
                                "机动车驾驶证申领和使用规定"
                                                    交
                                  "交通管理服务群众十项措施"
                                                    证
"证件所有人不应该为自己没有从事的行为承担法律责任"
------------------------------------------------
> ###    中文分词    ###
> txt <- prescindMeta(gg,c("ID"))
> re <- list()
> for (i in 1:nrow(txt)) {
+         re[[i]]<-    CWS(PlainTextDocument(reuters)[[i]],TRUE) ##  包括停用词
+         }
> ###    生成新的文集    ###
> reuters <- Corpus(VectorSource(re))
------------------------------------------------
> ###    元数据管理    ###
> DublinCore(reuters[[2]], "title") <- "建国 60 周年"
> meta(reuters[[2]])
Available meta data pairs are:
   Author           :
   DateTimeStamp: 2010-07-22 01:03:57
   Description    :
   Heading          :  建国 60 周年
   ID                  : 2
   Language        : eng
   Origin           :

------------------------------------------------
> ###    创建词条-文件矩阵
```

```
> dtm <- DocumentTermMatrix(reuters,control = list(minWordLength=2))##最短词两个字
> dtm
A document-term matrix (26 documents, 778 terms)

Non-/sparse entries: 1521/18707
Sparsity           : 92%
Maximal term length: 7
Weighting          : term frequency (tf)
> inspect(dtm[1:2, 3:6])   ## 结果有一定随机性
A document-term matrix (2 documents, 4 terms)

Non-/sparse entries: 3/5
Sparsity           : 62%
Maximal term length: 5
Weighting          : term frequency (tf)

      Terms
Docs 0.016 10 102 105
   1       0   1     1     0
   2       0   2     0     0
----------------------------------------------------------------------
> ##   操作词条-文件矩阵   ##
> ##   1、找出最少出现过 10 次的词条   ##
> findFreqTerms(dtm, 10)
[1] "汽车"    "驾驶"    "部门"    "居民"    "身份证" "使用"    "安全"    "检验"
[9] "公民"
----------------------------------------------------------------
> #   2、找出与"应该"相关度到少达 0.9 的词条   ###
> findAssocs(dtm, "应该", 0.9)
  保密   必须   便捷   表面   参考   常识   承担   读取   负有   复印 复印件
  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
  公众   过程   核对   核实   经营   快速   留存   切实   权益   确认   确实
  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
  十分   实践   司法 同一性   外观   伪造   文字   无误 行为人   行业   一致
  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
  义务   意识   应该   有损   责任   真伪   职能   只能   作用   法律   社会
  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   0.97   0.97
  证件   事务   相应   从事   使用   相关
  0.96   0.95   0.95   0.94   0.92   0.91

> ###   去掉较少词频（保留 80%以上）的词条后   ####
> inspect(removeSparseTerms(dtm, 0.8))
> ##  结果省略
----------------------------------------------------------------
```

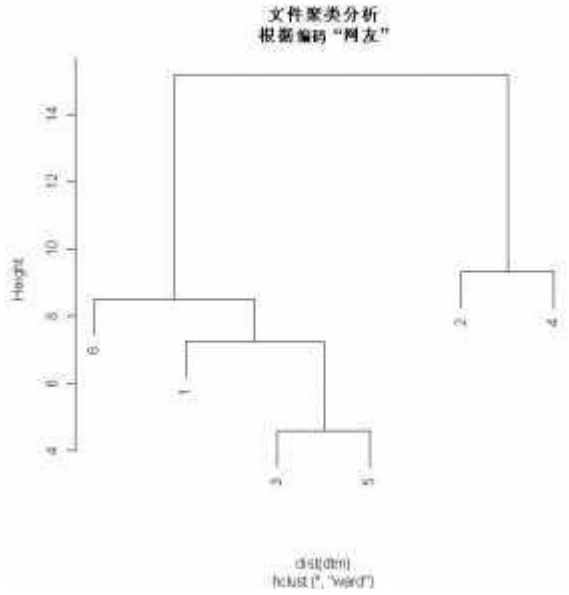
```
> ###   词典   ###   它通常用来表示文本挖掘有关词条
> (d <- Dictionary(c("车辆", "驾驶证")))
[1] "车辆"     "驾驶证"
attr(,"class")
[1] "Dictionary" "character"
> inspect(DocumentTermMatrix(reuters, list(dictionary = d)))
A document-term matrix (26 documents, 1 terms)

Non-/sparse entries: 7/19
Sparsity           : 73%
Maximal term length: 3
Weighting          : term frequency (tf)

      Terms
Docs  驾驶证
   1        0
   2        0
   3        1
   4        0
   5        4
   6        6
   7        4
   8        0
   9        0
   10       3
   11       0
   12       1
   13       0
   14       0
   15       0
   16       4
   17       0
   18       0
   19       0
   20       0
   21       0
   22       0
   23       0
   24       0
   25       0
   26       0

------------------------------------------------------------
> ##   根据词条频率对文件进行聚类分析   ##
```

```
> gg <- RQDA2tm("公安部回应" ,byFile = TRUE)
> reuters <- tm_map(gg, stripWhitespace)
> txt <- prescindMeta(gg,c("ID"))
> re <-    list()
> for (i in 1:nrow(txt)) {
+           re[[i]]<-    CWS(PlainTextDocument(reuters)[[i]],TRUE)
+         }
> reuters <- Corpus(VectorSource(re))
> dtm <- DocumentTermMatrix(reuters,control = list(minWordLength=2))
> reHClust <- hclust(dist(dtm), method = "ward")
> plot(reHClust,main ="文件聚类分析")
> ##  图形省略
> head(txt)
```

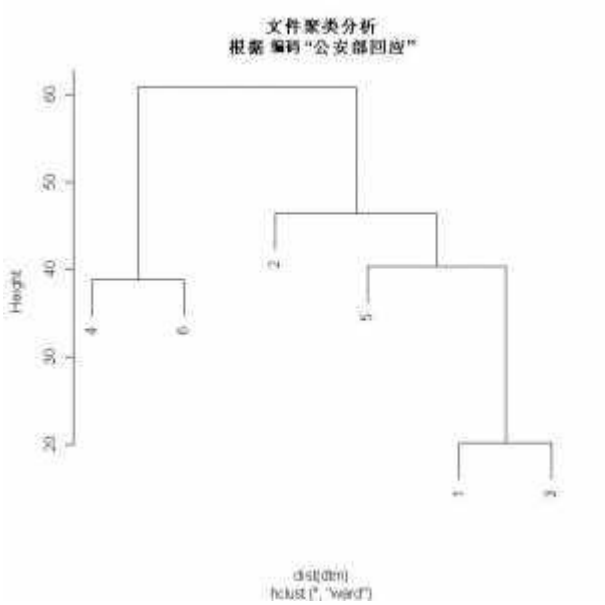
	MetaID	fname	fid	ID
1	0	公安部答复本网网友关于轻微交通违法处罚等 4 问题	1	1
2	0	公安部答复本网网友关于驾龄计算、异地购车上牌、老人驾车等 8 问题	2	2
3	0	公安部答复本网网友关于如何转回农业户口等 3 问题	3	3
4	0	公安部回复本网网友关于驾驶证年检被注销等 3 问题	4	4
5	0	公安部回复人民网网友关于异地缴交通罚款等 4 问题	5	5
6	0	公安部回复人民网网友关于身份证重号错号等 4 问题	6	6

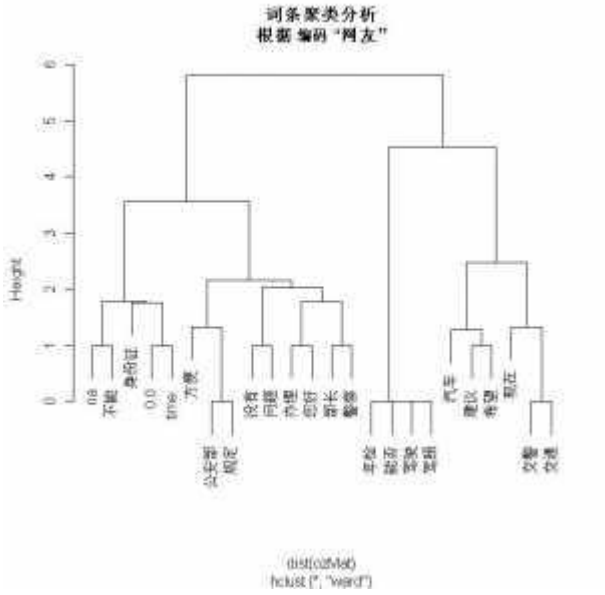
```
> ##   对词条进行分类   ###
> kmeans(dtm, 3)
##   结果省略
```

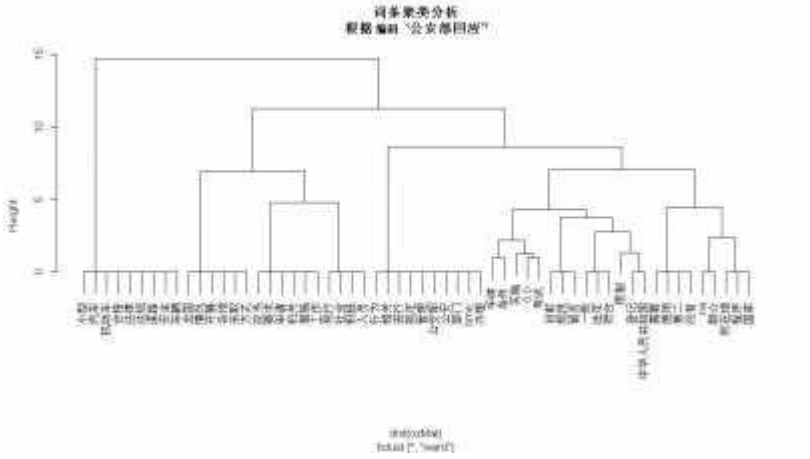
下面是按照以上方法对文档对不同编码进行聚类分析所绘树图：



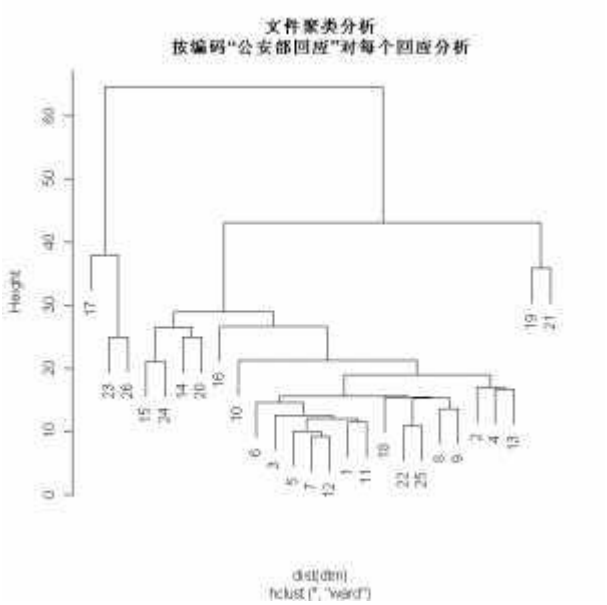
这是用编码“网友”提取相关文档进行分类的结果。其中，数字是指文件 ID（fname ID），后面几个图是对去掉较少词频的词条后的结果。







对上面的数据改为将每条回应为研究对象进行文档聚类分析，结果如下：



综合上面两种聚类分析可以判断：公安部负责对人民网网友进行回应的工作人员有两名，因为每个人的写作用词习惯是比较固定的。

```
> ###     主成分分析   ###
> ozMat <- TermDocumentMatrix(makeChunks(reuters, 50),
+           list(weighting = weightBin,minWordLength=2))
> ##将文档按 50 个字为单位分开形成多个文件，保留最短词两个字
>
> k <- princomp(as.matrix(ozMat), features = 2)
> windows()
> screeplot(k,npcs=6,type='lines')
> windows()
> biplot(k)
```

```
> ###     对词条进行聚类分析   ####
> ozHClust <- hclust(dist(ozMat), method = "ward")
> windows()
> plot(ozHClust,main="词条聚类分析")
> (x <- identify(ozHClust))
> memb <- cutree(ozHClust, k = 5)   #按 5 分类砍树
> memb
> cutree(ozHClust, h = 20)   #按 20 高度砍树
```

其他看上面的链接中的内容，其实生成词条-文件矩阵后还有许多工作可以做，比如用支持向量机进行文件分类、话题分类、根据话题用词频率分析作者所熟悉的行业等等……

运行环境：

```
> sessionInfo()
R version 2.11.0 (2010-04-22)
i386-pc-mingw32

locale:
[1] LC_COLLATE=Chinese (Simplified)_People's Republic of China.936
[2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936
[3] LC_MONETARY=Chinese (Simplified)_People's Republic of China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_People's Republic of China.936

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] RQDAtm_0.1-0              rJava_0.8-4               gsubfn_0.5-3
 [4] proto_0.3-8               tm_0.5-3                  RQDA_0.2-0
 [7] gWidgetsRGtk2_0.0-65 gWidgets_0.0-41           RSQLite_0.9-0
[10] DBI_0.2-5

loaded via a namespace (and not attached):
[1] RGtk2_2.12.18 slam_0.1-13   tools_2.11.0
```