# R语言实现文本聚类分类

###### 读入100篇文本数据：

csv = read.csv("train.csv", header = T, sep='\t', stringsAsFactors = F, encoding = 'UTF-8')

###### 读入停用词表：

mystopwords <- unlist(read.table("StopWords.txt", stringsAsFactors = F))

###### 预处理+分词：

removeNumbers = function(x) {ret = gsub("[0-9０１２３４５６７８９]", "", x)}

removeStopWords = function(x, words) {

ret = character(0)

index <- 1

it\_max <- length(x)

while (index <= it\_max) {

if (length(words[words == x[index]]) < 1)ret <- c(ret, x[index])

index <= index + 1

}

ret

}

sample.words <- lapply(csv$text, removeNumbers)

sample.words <- lapply(sample.words, segmentCN)

sample.words <- lapply(sample.words, removeStopWords, mystopwords)

###### 生成文本库格式：

corpus = Corpus(VectorSource(sample.words))

meta(corpus, "cluster") <- csv$type

unique\_type = unique(csv$type)

###### 得到文档词频矩阵

sample.dtm <- DocumentTermMatrix(corpus, control = list(wordLengths = c(2, Inf)))

###### 按照类目绘制词云图

n<-nrow(csv)

zz1 = 1:n

cluster\_matrix<-sapply(unique\_type, function(type){apply(tdm.matrix[,zz1[csv$type==type]], 1, sum)})

png("sample\_cluster\_comparision.png", width = 800, height = 800)

comparison.cloud(family='STKaiti',cluster\_matrix)

title(main="smaple cluster comparison")

dev.off()



###### 文本聚类：

sample\_matrix = as.matrix(sample.dtm)

rownames(sample\_matrix) <- csv$type

k <- length(unique(csv$type))

sample\_KMeans <- kmeans(sample\_matrix, k)

library(clue)

cl\_agreement(sample\_KMeans, as.cl\_partition(csv$type), "diag")

输出一致度：0.28

###### 文本分类（1-NN）：

sample\_knnCl <- knn(train, test, trainC1)

trueC1 <- as.factor(rownames(test))

(nnTable <-table("1-NN" = sample\_knnCl, sample = trueC1))

sum(diag(nnTable)) / nrow(test)

输出准确率：0.1