TextMining\_1

基于tm和jiebaR以及wordcount作的万科2013年报管理层讨论词云

require(tm)

## Loading required package: tm  
## Loading required package: NLP

require(jiebaR)

## Loading required package: jiebaR  
## Loading required package: jiebaRD

require(wordcloud)

## Loading required package: wordcloud  
## Loading required package: RColorBrewer

#初始化jeibaR的分词器，这里没有使用user自定义词典  
cutter<-worker()  
f1<-scan("~/Downloads/test.txt",encoding = "UTF-8",what=character()) %>%  
 paste(collapse=" ")  
  
#<=符号直接分词，可惜不支持%>% 操作  
f1<-cutter<=f1  
  
#直接定义为dataframeSource语料库，并且去掉数字空白等内容  
f1<-as.data.frame(f1,stringsAsFactors = FALSE) %>%  
 DataframeSource %>%  
 Corpus %>%  
 tm\_map(removeNumbers) %>%  
 tm\_map(removePunctuation) %>%  
 tm\_map(stripWhitespace)   
  
#生成词文矩阵  
tdm <-TermDocumentMatrix(f1)  
  
#找出现频率最高的5个词  
findFreqTerms(tdm,5)

## [1] "万平方米" "新开工" "百分点" "董事会" "进一步"

#调整字体为黑体，防止无法在Mac下输出图形中的汉字  
par(family = "STHeiti")  
  
#作为矩阵并且按照词频排序  
m1<-as.matrix(tdm)  
m1 <- sort(rowSums(m1),decreasing=TRUE)  
m1<-data.frame(word = names(m1),freq=m1)  
  
#输出词云  
wordcloud(m1$word,m1$freq,scale=c(5,0.5), max.words=100,   
 random.order=FALSE,rot.per=0.35, use.r.layout=FALSE,colors=brewer.pal(8,"Dark2"))

