RDataMining

Week 11 Lab in R: Text Mining

Dr. Liang (Leon) Chen

Instructions

- R script for Text Mining is provided in the next slide, please follow it and complete the lab in R.
- You do not need to type notes (starting at #), but it’s a good manner to have them in you script.
- In order to see codes and notes clearly, I show the script in RStudio.
- Advanced materials are available to you at the end.

```r
#define and choose the dataset: this is same with Step 2 in RM import configuration wizard
Lab11Data<-read.csv(file.choose(),header=FALSE, sep="\t", quote="", stringsAsFactors=FALSE)
#check the dimension of your data
dim(Lab11Data)
#assign column names for your data
colnames(Lab11Data) <- c("type", "text")
#select your first column as your factor
Lab11Data$type <- factor(Lab11Data$type)
#check the structure of your data
str(Lab11Data)
#install and then use the package tm for text mining
install.packages("tm")
library(tm)
#Load the data as a corpus and rename it
Lab11Doc<-Corpus(VectorSource(Lab11Data$text))
#Inspect the content of the document using inspect(Lab11Doc); too long, so not included here
#Data cleaning,remove stopwords,numbers, etc.
Lab11Doc1<-tm_map(Lab11Doc,removePunctuation)
Lab11Doc1<-tm_map(Lab11Doc1,tolower)
Lab11Doc1<-tm_map(Lab11Doc1,removeWords, stopwords(kind="en"))
Lab11Doc1<-tm_map(Lab11Doc1,stripWhitespace)
Lab11Doc1<-tm_map(Lab11Doc1,removeNumbers)
# Build a term-document matrix
Lab11Doc2<-TermDocumentMatrix(Lab11Doc1)
#find words that occur at least 100 times using findFreqTerms()
findFreqTerms(Lab11Doc2, lowfreq = 100)
#Analyze the association between frequent terms (i.e. terms which correlate) using findAssocs() function
findAssocs(Lab11Doc2, terms = "call", corlimit = 0.1)
#generate a matrix
Lab11Matrix<-as.matrix(Lab11Doc2)
#generate a frequency matrix based on the number frquency of each word
Lab11_freqs = sort(rowSums(Lab11Matrix), decreasing=TRUE)
Lab11DF = data.frame(word=names(Lab11_freqs), freq=Lab11_freqs)
#view the first 20 most frequent terms
head(Lab11DF,20)
#visualize the first 20 most frequent terms
library(ggplot2)
Lab11DF_top<-Lab11DF[1:20,]
ggplot(Lab11DF_top, aes(x = word, y = freq))+geom_bar(stat = "identity") + coord_flip() + labs(title = "Most Frequent words")
#install and use wordcloud library for visualizaiton: bag of words
install.packages("wordcloud")
library(wordcloud)
set.seed(123)
#generate bag of words to show the top 200 words.
wordcloud(word=Lab11DF$word, freq=Lab11DF$freq, max.words=200, random.order=FALSE, colors=brewer.pal(8, "Dark2"))
```

You can use stopwords(kind = "en") to see all the stopwords in R

Deliverable 1: Take a screenshot of the associations with term “call”.

Deliverable 2: Take a screenshot of the top 20 most frequent words. Are the top five words the same with those in RapidMiner (sort the wordlist by total occurrences in Deliverable 4 of RapidMiner lab)

Deliverable 3: Take a screenshot of the bag of words you generate.

Advanced Materials (Optional)

- You can follow this site to conduct Naive Bayes Classification for our dataset
- Please find the instruction for ngram package here.
- **Text Mining with R *A Tidy Approach***
- **Text mining and word cloud fundamentals in R**
- **S**imple **W**ordcloud
- Text Mining and N-Grams