**כריית טקסט**

***מיני תרגיל מספר 2 – לא להגשה***

**מטרת התרגיל:**

בחינת ההבדלים בין commonality cloud ו comparison cloud

**שאלות:**

1. טען שני קבצים, company1.csv, company2.csv
2. בנה קורפוס המכיל מסמך אחד עם כל המילים המופיעות בקובץ של company1 ומסמך שני המכיל את כל המילים המופיעות בקובץ של company2.
3. צור commonality cloud ו comparison cloud עבור הקורפוס הנ"ל.
4. צור טבלת שכיחויות המכילה את מספר הפעמים שכל מילה מופיעה בכל מסמך.
5. האם קיימות מילים המופיעות בשני העננים שיצרת בסעיף הקודם?
6. צור טבלה המכילה רק מילים המופיעות בשני הקבצים (ע"י subset).
7. חשב את סכום המופעים בשני הקבצים ומיין בסדר יורד. מיהם ארבעת המונחים הראשונים ברשימה?
8. שנה את פרמטר max.words = 4 ב commonality cloud. אילו מילים מופיעות בענן?

**פתרון:**

setwd("E:/Text mining course/2")

library(tm)

library(wordcloud)

####load data

cmp1 <- read.csv('company1.csv')

cmp2 <- read.csv('company2.csv')

cmp1.vec <- paste(cmp1$text, collapse=" ")

cmp2.vec <- paste(cmp2$text, collapse=" ")

all <- c(cmp1.vec, cmp2.vec)

length(all)

corpus <- VCorpus(VectorSource(all))

tdm <- TermDocumentMatrix(corpus)

tdm.m <- as.matrix(tdm) #convert to a matrix

colnames(tdm.m) = c("cmp1", "cmp2")

#choose color palette

pal <- brewer.pal(9, "Purples")

pal <- pal[-(1:4)]

set.seed(1) # for reproducibility

commonality.cloud(tdm.m, max.words = 4, random.order=FALSE,colors=pal)

comparison.cloud(tdm.m, max.words = 30, random.order=FALSE,title.size = 1,

colors=c('black','darkred'))

#only terms that appear in both documents

common.words <- subset(tdm.m, tdm.m[,1] > 0 & tdm.m[,2] > 0)

common.words

sum <- rowSums(common.words)

sort(sum, decreasing = T)