Text Processing - Word Cloud and Sentiment Analysis

Monday, May 11, 2015

#### Text documents (.txt files) can be analyzed for language usage and emotion through several tools including word clouds and sentiment analysis.

##### Four separate documents are reviewed, all are papers on various subjects within the field of Human Resources: Compensation and Benefits, Industrial Relations, Recruitment, and Strategic Human Resources.

library(tm) ## text management package  
library(wordcloud) ## word cloud addendum  
library(plyr) ## data manipulation  
library(ggplot2) ## plotting  
library(RColorBrewer) ## plot colors used in wordcloud  
library(SnowballC) ## needed for wordcloud  
library(syuzhet) ## new package available for sentiment analysis  
library(reshape2) ## matrix management  
library(mpa) ## co-words process analysis

#### Section 1: Load needed libraries for processing.

mydata <- Corpus(DirSource("C:/wordanalysis"))  
mydata <- tm\_map(mydata,stripWhitespace)  
mydata <- tm\_map(mydata, removePunctuation)  
mydata <- tm\_map(mydata, tolower) ## note: tolower doesn't return a Text Doc  
mydata <- tm\_map(mydata,removeWords,stopwords("english"))  
mydata <- tm\_map(mydata,stemDocument)  
mydata <- tm\_map(mydata, PlainTextDocument) ## need to do this before process  
  
filelist <- list.files("C:/wordanalysis")  
  
  
##coword\_mpa <- tm\_map(mydata,leer.mpa)  
##coword\_matrix <- tm\_map(coword\_mpa,matriz.mpa)  
  
##cowords <- TermDocumentMatrix(mydata)  
##terms <- findFreqTerms(cowords,lowfreq=3,highfreq=Inf)

#### Section 2: Process the documents.

##### 1. Load all files in the directory "wordanalysis" into a Corpus (cluster)

##### 2. Strip all whitespace

##### 3. Remove punctuation so as not to double count "win" and "win,"

##### 4. Convert all words to lowercase, to avoid count differences between "Word" and "word"

##### 5. Remove the majority of common non-impact words ("the", "and", etc.)

##### 6. "Stem" words: put 'example' and 'examples' together as one

##### 7. Process into wordcloud

wordcloud(mydata, scale=c(5,0.5),   
 max.words=100,  
 random.order=FALSE,  
 rot.per=0.35,  
 use.r.layout=FALSE,  
 colors=brewer.pal(8,"Dark2"))



## Wordcloud of 4 Human Resources texts

#### Section 3. Generate sentiment analysis.

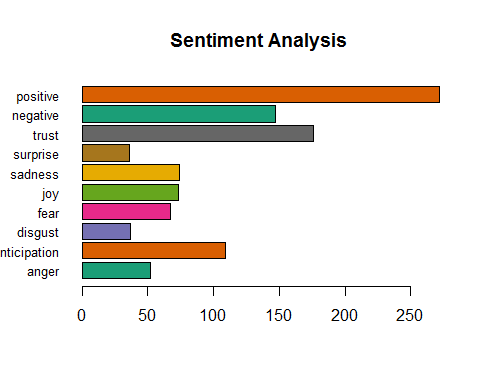
##### 1. Convert the word data to a character vector

##### 2. Process for sentiments (value words) via syuzhet.

##### Note: Previous R package "sentiment" has been removed to archives.

##### 3. Reshape the matrix into tidy data ready for plot via ggplot2.

mycharv <- as.character(mydata)  
mysent <- get\_nrc\_sentiment(mycharv)  
x <- mysent[1,]  
plotdf <- melt(x)  
barplot(plotdf$value,   
 col=brewer.pal(8,"Dark2"),  
 names.arg = plotdf$variable,  
 cex.names = 0.80,  
 horiz = TRUE,  
 main = "Sentiment Analysis",  
 mar = c(5,6,4,1) + 0.3,  
 las = 1)



## 4 Human Resources texts