**Assignment 8**

**performing sentiment analysis on a corpus of product reviews that has already been supplied.**

Step 1: Read the excel filename and load the column containing product reviews, install package pander, syuzhet

Step 2: Tokenize and stem each review from Northeastern\_class.csv file

Step 3: Compare each stemmed review with the stemmed positive/negative dictionary to compute the percentage of positive and negative words in each review

Step 4: Visualize an individual review's sentiment.

Step 5: Compute overall sentiment of the corpus by averaging and also the overall sentiment

**Problem 1 : each review sentiment**

> library(devtools)

> #install.packages("pander")

> library(pander)

**> #read file browse on my laptop northeastrern\_class.csv**

> product\_review<-read.csv(file.choose(),header = T)

> **#tokenizing, removingnon-essential characters such as punctuation, numbers**

> product\_review$Body = gsub("[[:punct:]]", "", product\_review$Body)

> product\_review$Body = gsub("[[:digit:]]", "", product\_review$Body)

> product\_review$Body = gsub("http\\w+", "", product\_review$Body)

> product\_review$Body = gsub("[ \t]{2,}", "", product\_review$Body)

> product\_review$Body = gsub("^\\s+|\\s+$", "", product\_review$Body)

> #chcek the rows number

> nrow(product\_review)

[1] 500

> #first set the positive in sentiment column

> product\_review["Sentiment"]<-"positive"

> #loop all rows to match with dictionary uisng pander package

> for(i in 1:nrow(product\_review)) {

+ row <- product\_review[i,]

+ nrc\_data <- get\_nrc\_sentiment(as.String(product\_review$Body[i]))

+ pander::pandoc.table(nrc\_data[, 1:8], split.table = Inf)

+ pander::pandoc.table(nrc\_data[, 9:10])

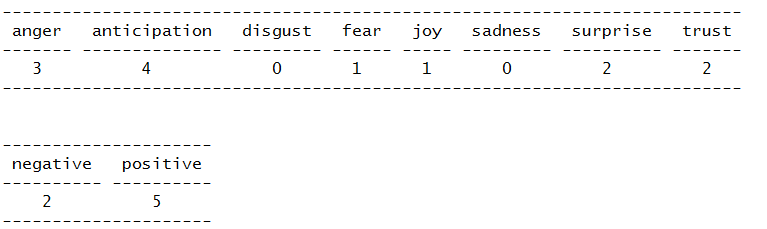
+ valence <- (nrc\_data[, 9]\*-1) + nrc\_data[, 10]#compare with dictionary

+ if(valence<0)#define the negtive words

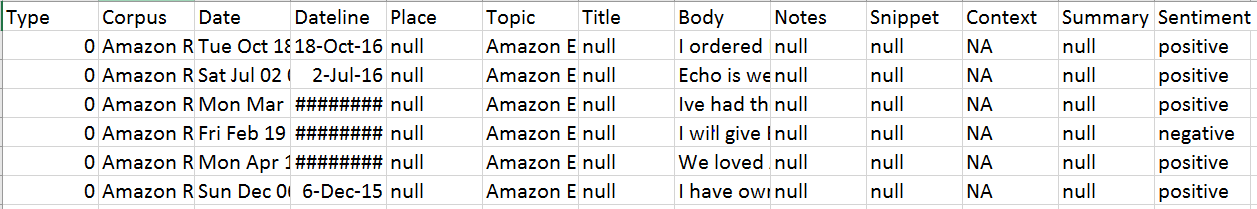
+ { product\_review[i,c('Sentiment')]="negative"}

}

**Result for every row:**



Write out the csv file with sentiment result(positive & negative)



> #Visualization of individual review enter the number of row we want to analyze

> n <- readline(prompt="Enter an integer between 1 and 500: ")

Enter an integer between 1 and 500: 300

> n<-as.integer(n)

> nrc\_data <- get\_nrc\_sentiment(as.String(product\_review$Body[n]))

> pander::pandoc.table(nrc\_data[, 1:8], split.table = Inf)

--------------------------------------------------------------------------

anger anticipation disgust fear joy sadness surprise trust

------- -------------- --------- ------ ----- --------- ---------- -------

1 3 0 1 6 1 4 4

--------------------------------------------------------------------------

> pander::pandoc.table(nrc\_data[, 9:10])

---------------------

negative positive

---------- ----------

1 11

---------------------

> #nrc\_data[,9]

> valence <- (nrc\_data[, 9]\*-1) + nrc\_data[, 10]

> valence

[1] 10

> #fOR Visualization

> barplot(

+ sort(colSums(prop.table(nrc\_data[, 1:8]))),

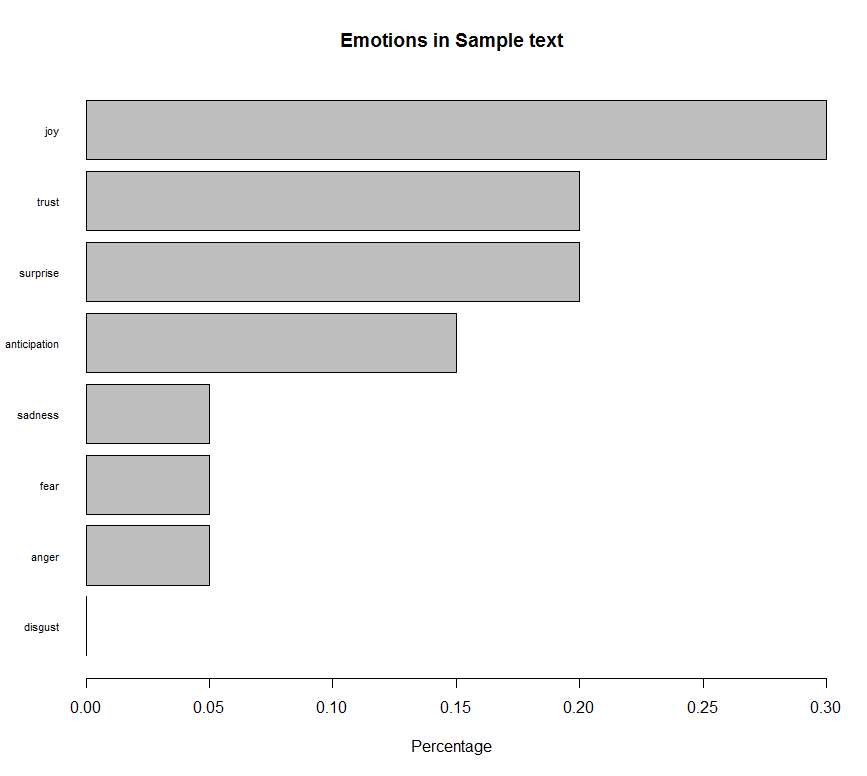
+ horiz = TRUE,

+ cex.names = 0.7,

+ las = 1,

+ main = "Emotions in Sample text", xlab="Percentage"

+ )



**Problem 2 : Overall sentiment**

> syuzhet\_vector <- get\_sentiment(as.String(product\_review$Body), method="syuzhet")

> head(syuzhet\_vector)

[1] 104

> nrc\_data <- get\_nrc\_sentiment(as.String(product\_review$Body))

> pander::pandoc.table(nrc\_data[, 1:8], split.table = Inf)

--------------------------------------------------------------------------

anger anticipation disgust fear joy sadness surprise trust

------- -------------- --------- ------ ----- --------- ---------- -------

184 267 124 232 233 209 165 341

--------------------------------------------------------------------------

> pander::pandoc.table(nrc\_data[, 9:10])

---------------------

negative positive

---------- ----------

498 668

---------------------

> barplot(

+ sort(colSums(prop.table(nrc\_data[, 1:8]))),

+ horiz = TRUE,

+ cex.names = 0.7,

+ las = 1,

+ main = "Emotions in Sample text", xlab="Percentage"

+ )

