```r
# Week 5 - Recitation


# Video 2

# Load the dataset

emails = read.csv("energy_bids.csv", stringsAsFactors=FALSE)

str(emails)

# Look at emails

emails$email[1]
emails$responsive[1]

emails$email[2]
emails$responsive[2]

# Responsive emails

table(emails$responsive)



# Video 3


# Load tm package

library(tm)


# Create corpus

corpus = Corpus(VectorSource(emails$email))

corpus[[1]]


# Pre-process data
corpus <- tm_map(corpus, tolower)

corpus <- tm_map(corpus, removePunctuation)

corpus <- tm_map(corpus, removeWords, stopwords("english"))

corpus <- tm_map(corpus, stemDocument)
```

```r
# Look at first email
corpus[[1]]



# Video 4

# Create matrix

dtm = DocumentTermMatrix(corpus)
dtm

# Remove sparse terms
dtm = removeSparseTerms(dtm, 0.97)
dtm

# Create data frame
labeledTerms = as.data.frame(as.matrix(dtm))

# Add in the outcome variable
labeledTerms$responsive = emails$responsive

str(labeledTerms)



# Video 5


# Split the data

library(caTools)

set.seed(144)

spl = sample.split(labeledTerms$responsive, 0.7)

train = subset(labeledTerms, spl == TRUE)
test = subset(labeledTerms, spl == FALSE)

# Build a CART model

library(rpart)
library(rpart.plot)

emailCART = rpart(responsive~., data=train, method="class")
```

```
prp(emailCART)



# Video 6

# Make predictions on the test set

pred = predict(emailCART, newdata=test)
pred[1:10,]
pred.prob = pred[,2]

# Compute accuracy

table(test$responsive, pred.prob >= 0.5)

(195+25)/(195+25+17+20)

# Baseline model accuracy

table(test$responsive)
215/(215+42)



# Video 7

# ROC curve

library(ROCR)

predROCR = prediction(pred.prob, test$responsive)

perfROCR = performance(predROCR, "tpr", "fpr")

plot(perfROCR, colorize=TRUE)

# Compute AUC

performance(predROCR, "auc")@y.values
```