3.17.10 R code we learned for text mining. May 9,2018

3.17.10.1

3.17.10.2 library(tm) #text mining.

3.17.10.3 docs <- **VCorpus**(DirSource(cname))

a. corp <- **Corpus**(ZipSource("AutoElectronics.zip", recursive = T)). #similar, different file source

3.17.10.4 Corp <- **tm_map**(Corp, removeNumbers)

a. • "removePunctuation" "removeWords" tolower

3.17.10.5 **writeLines**(as.character(Corp[1])) # Check to see if it worked.

3.17.10.6 Don't forget: Look at samples of RAW data

a. Original source, in the wild
b. Before it's read into R
c. When first read into R
d. After cleaning

3.17.10.7 Tip: Add some variables

3.17.10.8 **str**(data). #work with any data

3.17.10.9 **rep**(1, 1000) #repeat 1000 times

3.17.10.10

3.17.10.11 training <- **sample**(c(1:2000), 0.6*2000). #take 1200 samples from 1:2000

3.17.10.12 tdm <- **TermDocumentMatrix**(corp)

3.17.10.13 tfidf <- **weightTfIdf(**tdm)

3.17.10.14 Sparse2 <- **removeSparseTerms**(Tdm, .98) #keep only 2% approximately

3.17.10.15 **head**(Sparse2$dimnames$Terms, 50). #first 50 elements.

a. # could also use sample

b. **sample**(Sparse2$dimnames$Terms, 20, replace=TRUE)

3.17.10.16 lsa.tfidf <- **lsa**(tfidf, dim = 20) #Doing LSA analysis. Careful, can take minutes

3.17.10.17 lsa.tfidf$dk # extracts variable (column) called 'dk'

3.17.10.18 words.df <- **as.data.frame**(as.matrix(lsa.tfidf$dk))

3.17.10.19 reg <- **glm**(label ~ ., data = trainData, family = 'binomial')

3.17.10.20 pred <- **predict**(reg, newdata = validData, type = "response")