**###############################################################**

**########### CHAPTER 5: TERRITORIAL PAPERS #######################**

**###############################################################**

**rm(list = ls())**

**library(stm)**

**library(tm)**

**library(slam)**

**###############################################################**

**######################### TERRITORIAL PAPERS ###################**

**################## Reading from file datacomb.csv ##################**

**############ authors and recipients already cleaned up ###############**

**###############################################################**

**data <- read.csv("C:\\Johannes Ledolter\\2020March01Book\\Chapter5WEB\\datacomb.csv",stringsAsFactors=F)**

**dim(data)**

**data[1:10,1:5]**

**ids1=data$indc**

**ids2=data$autr**

**ids3=data$recr**

**ids4=data$yearc**

**ids1**

**length(ids1)**

**ids2**

**length(ids2)**

**ids3**

**length(ids3)**

**ids4**

**length(ids4)**

**corpus <- VCorpus(VectorSource(data$text),readerControl = list(reader = readPlain))**

**## this is how to create corpus**

**corpus1 <- tm\_map(corpus, stripWhitespace)**

**corpus2 <- tm\_map(corpus1, content\_transformer(tolower))**

**corpus3 <- tm\_map(corpus2, removePunctuation)**

**corpus4 <- tm\_map(corpus3, removeNumbers)**

**corpus5 <- tm\_map(corpus4, removeWords, stopwords("english"))**

**corpus.dtm <- DocumentTermMatrix(corpus5,control=list(stemming=FALSE))**

**## no stemming as default**

**dim(corpus.dtm)**

**####################################################################**

**#### display frequency of most-common words according to time / vol ########**

**####################################################################**

**findFreqTerms(corpus.dtm,3000,10000000)**

**findFreqTerms(corpus.dtm,3000,10000000)[1]**

**high=length(findFreqTerms(corpus.dtm,3000,10000000))**

**high**

**ids=ids4 ## year**

**ids=as.numeric(ids)**

**## plotting 9 top words**

**## allows us to exclude some stopwords**

**toplot=dim(9)**

**toplot[1]=1**

**toplot[2]=4**

**toplot[3]=5**

**toplot[4]=6**

**toplot[5]=12**

**toplot[6]=13**

**toplot[7]=16**

**toplot[8]=21**

**toplot[9]=22**

**par(mfrow=c(3,3))**

**for (i in 1:9) {**

**ind=findFreqTerms(corpus.dtm,3000,10000000)[toplot[i]]**

**nuterms=row\_sums(as.matrix(corpus.dtm[,ind]))**

**indicator<-function(condition) ifelse(condition,1,0)**

**nutermsind=indicator(nuterms)**

**t50=tapply(nutermsind,ids,FUN=mean,na.rm=TRUE)**

**t51=t50**

**t51[colSums(table(nutermsind,ids))<=9]=NA**

**barplot(t51,ylim=c(0,1),main=ind,xlab="Year",ylab="Percentage of letters")**

**}**

**ids=ids1 ## volume**

**ids=as.numeric(ids)**

**## plotting 9 top words**

**## allows us to exclude some stopwords**

**toplot=dim(9)**

**toplot[1]=1**

**toplot[2]=4**

**toplot[3]=5**

**toplot[4]=6**

**toplot[5]=12**

**toplot[6]=13**

**toplot[7]=16**

**toplot[8]=21**

**toplot[9]=22**

**par(mfrow=c(3,3))**

**for (i in 1:9) {**

**ind=findFreqTerms(corpus.dtm,3000,10000000)[toplot[i]]**

**nuterms=row\_sums(as.matrix(corpus.dtm[,ind]))**

**indicator<-function(condition) ifelse(condition,1,0)**

**nutermsind=indicator(nuterms)**

**t50=tapply(nutermsind,ids,FUN=mean,na.rm=TRUE)**

**t51=t50**

**t51[colSums(table(nutermsind,ids))<=9]=NA**

**barplot(t51,ylim=c(0,1),main=ind,xlab="Corpus",ylab="Percentage of letters")**

**}**