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

**############# 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)**

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

**#### HISTOGRAM / BOXPLOTS: START ###############################**

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

**yearc=data$yearc**

**indc=data$indc**

**library(lattice)**

**par(mfrow=c(1,1))**

**hist(as.numeric(yearc),xlab="Year",main="Histogram Year")**

**boxplot(as.numeric(yearc)~indc,xlab="Year",ylab="Corpus",horizontal = TRUE,main="Boxplots Year")**

**histogram(~as.numeric(yearc)|indc,xlab="Year",layout=c(1,7))**

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

**#### HISTOGRAM / BOXPLOTS: END #################################**

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

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

**#### WRITE OUT AUTHORS / RECIPIENTS: START ######################**

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

**autr=data$autr**

**recr=data$recr**

**sort(table(c(autr,recr)),decreasing=TRUE)[1:50]**

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

**#### WRITE OUT AUTHORS / RECIPIENTS: END ########################**

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

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

**################ NETWORK MODELS: START ###################**

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

**autr=data$autr**

**recr=data$recr**

**length(autr)**

**length(recr)**

**table(autr)**

**f10=rownames(table(autr))**

**f10**

**f20=rownames(table(recr))**

**f20**

**f10[1:10]**

**f20[1:10]**

**xx=table(autr,recr)**

**dim(xx)**

**xx**

**f1=rownames(xx)**

**f2=colnames(xx)**

**sum(xx)**

**f3=rownames(table(c(f1,f2))) ## also could use union but union doesn't order alphabetically (overall results the same)**

**f1**

**f2**

**f3**

**length(f1)**

**length(f2)**

**length(f3)**

**h3=c(autr,f3)**

**h4=c(recr,f3)**

**h3**

**h4**

**length(h3)**

**length(h4)**

**adj=table(h3,h4)**

**adj**

**dim(adj)**

**n=dim(adj)[1]**

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

**adj[i,i]=0 ## we take out self-references**

**}**

**sum(adj)**

**adj**

**## not exactly the same as sum(xx)**

**## reason is that adj subtracts out self-references (where to/from the same); some proclamations were coded that way; also governor to governor**

**## above is checked to get the correct adjacency matrix**

**library(igraph) ## load the package**

**## with names for nodes**

**letters <- graph.adjacency(adj,mode="directed",weighted=TRUE,diag=FALSE,add.colnames=NULL,add.rownames=NA)**

**letters**

**letters=delete.vertices(letters,which(degree(letters,mode="total")<51))**

**letters**

**V(letters)**

**E(letters)**

**degree(letters)**

**set.seed(1234)**

**plot(letters)**

**set.seed(1234)**

**plot.igraph(letters,vertex.label=V(letters)$name,layout=layout\_with\_fr,edge.color="black",**

**edge.width=E(letters)$weight/10, edge.arrow.size=1.5,edge.curved=TRUE)**

 **## other graph layouts when placing nodes**

 **## random placement**

 **## plot.igraph(letters,vertex.label=V(letters)$name,layout=layout\_randomly,edge.color="black",edge.width=E(letters)$weight/10, edge.arrow.size=1.5,edge.curved=TRUE)**

 **## circle placement**

 **## plot.igraph(letters,vertex.label=V(letters)$name,layout=layout\_in\_circle,edge.color="black",edge.width=E(letters)$weight/10, edge.arrow.size=1.5,edge.curved=TRUE)**

**### About omitting nodes**

**### Above have omitted on total degrees (in and out)**

**### Another possibility is to omit on the number of letters written / received (we omit self-referrals)**

**m=dim(adj)[1]**

**m**

**sumletter=dim(m)**

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

**sumletter[i]=sum(adj[i,])+sum(adj[,i]) ##total number of letters written or received**

**}**

**mean(sumletter)**

**par(mfrow=c(3,1))**

**hist(sumletter)**

**boxplot(sumletter)**

**plot(sumletter)**

**par(mfrow=c(1,1))**

**letters <- graph.adjacency(adj,mode="directed",weighted=TRUE,diag=FALSE,add.colnames=NULL,add.rownames=NA)**

**letters**

**letters=delete.vertices(letters,which(sumletter<251))**

**letters**

**V(letters)**

**E(letters)**

**set.seed(1234)**

**plot(letters)**

**set.seed(1234)**

**plot.igraph(letters,vertex.label=V(letters)$name,layout=layout\_with\_fr, edge.color="black",**

**edge.width=E(letters)$weight/10, edge.arrow.size=1.5,edge.curved=TRUE)**

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

**################ NETWORK MODELS: END #####################**

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