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

**#################### ANALYSIS (SECTION 9.3.2 (39th Congress) ###########################**

**################### Reading the data from file PrelimData.RData #########################**

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

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

**library(tm)**

**library(slam)**

**library(stm)**

**## LOADING DATA (39th Congress)**

**load("C:\\Johannes Ledolter\\2020March01Book\\Chapter9WEB\\PrelimData.RData")**

**## creating corpus**

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

**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"))**

**corp.dtm <- DocumentTermMatrix(corpus5,control=list(stemming=FALSE)) ## no stemming is the default**

**dim(corp.dtm)**

**## omitting a certain term from the corpus such as "globe" or "unitedstates"**

**## as these terms either come up very often ("unitedstates")**

**## or have little to do with discourse ("globe")**

**should not have to take out senfrom gentfrom …. ???**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="globe"]**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="unitedstates"]**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="gentfrom"]**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="senfrom"]**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="friendfrom"]**

**dim(corp.dtm)**

**corp.dtm=corp.dtm[,!labels(corp.dtm)$Terms=="sensreps"]**

**dim(corp.dtm)**

**quantile(len)**

**## cutting out short letters**

**## could do that, but the same letters may get cut when stripping words**

**indexlen=len>=32 ## can pick a length to omit short letters**

**nnn=dim(corp.dtm)[1]**

**number=dim(nnn)**

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

**number[i]=i**

**}**

**corp.dtm=corp.dtm[number[indexlen],]**

**dim(corp.dtm)**

**meta2=meta2[number[indexlen]]**

**meta3=meta3[number[indexlen]]**

**meta4=meta4[number[indexlen]]**

**meta5=meta5[number[indexlen]]**

**meta6=meta6[number[indexlen]]**

**group=group[number[indexlen]]**

**len=len[number[indexlen]]**

**## cutting out short letters**

**## working with / simplifying corpus**

**dim(corp.dtm)**

**findFreqTerms(corp.dtm,10)**

**findFreqTerms(corp.dtm,2000)**

**dim(corp.dtm)**

**corpStripped.dtm=removeSparseTerms(corp.dtm,0.995) ## experiment with sparcity**

**## terms that show up in at least 0.5 percent of documents (50,000 x 0.005 = 250)**

**dim(corpStripped.dtm)**

**## may want to include this**

**## Check for documents with zero words**

**dim(corpStripped.dtm)**

**index=row\_sums(corpStripped.dtm)==0**

**nnn=dim(corpStripped.dtm)[1]**

**number=dim(nnn)**

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

**number[i]=i**

**}**

**number[index]**

**## do following only if there are empty documents**

**corpStripped.dtm=corpStripped.dtm[-number[index],]**

**dim(corpStripped.dtm)**

**meta2=meta2[-number[index]]**

**meta3=meta3[-number[index]]**

**meta4=meta4[-number[index]]**

**meta5=meta5[-number[index]]**

**meta6=meta6[-number[index]]**

**group=group[-number[index]]**

**len=len[-number[index]]**

**labels(corpStripped.dtm)**

**## TOPIC MODELS WITHOUT METAVARIABLES**

**length(meta2)**

**dim(corpStripped.dtm)**

**yearnum=as.numeric(meta5)**

**yearfac=factor(yearnum)**

**table(yearfac)**

**plot(yearfac)**

**group ## different months in ascending order**

**table(group)**

**plot(group)**

**################ STM topic model WITHOUT metavariables ###################**

**processed <- readCorpus(corpStripped.dtm, type = 'slam')**

**out <- prepDocuments(processed$documents, processed$vocab)**

**out**

**nutopics=20**

**results=stm(out$documents,out$vocab,K=nutopics,seed = 1234)**

**labelTopics(results,c(1:20), n=20)**

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

**plot(results)**

**plot(results,type="summary",labeltype="prob",n=10)**

**plot(results,type="summary",labeltype="frex",n=10)**

**plot(results,type="summary",labeltype="lift",n=10)**

**plot(results,type="summary",labeltype="score",n=10)**

**summary(results)**

**cloud(results,max.words=100)**

**## displaying overall (average) topic proportions**

**average=dim(nutopics)**

**for (j in 1:nutopics) {**

**uu=results$theta[,j]**

**average[j]=mean(uu)**

**}**

**plot(average)**

**average**

**top=c("Topic 1","Topic 2","Topic 3","Topic 4","Topic 5","Topic 6","Topic 7","Topic 8","Topic 9","Topic 10","Topic 11","Topic 12","Topic 13","Topic 14","Topic 15","Topic 16","Topic 17","Topic 18","Topic 19","Topic 20")**

**top**

**average=round(average,3)**

**hh=data.frame(top,average)**

**print(hh)**

**## stratify topic proportions according to year, frequent speakers, speakers from slave/nonslave states**

**## stratify according to year**

**ave=matrix(nrow=3,ncol=nutopics)**

**relave=matrix(nrow=3,ncol=nutopics)**

**for (j in 1:nutopics) {**

**uu=results$theta[,j]**

**ave[1,j]=mean(uu[yearnum==1865])**

**ave[2,j]=mean(uu[yearnum==1866])**

**ave[3,j]=mean(uu[yearnum==1867])**

**}**

**max=max(ave)**

**TOPIC=c(1:20)**

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

**barplot(ave[1,]~TOPIC,ylim=c(0,max),main="1865")**

**barplot(ave[2,]~TOPIC,ylim=c(0,max),main="1866")**

**barplot(ave[3,]~TOPIC,ylim=c(0,max),main="1867")**

**## compared to average for each topic**

**for (j in 1:nutopics) {**

**average=mean(ave[,j])**

**for (i in 1:3) {**

**relave[i,j]=ave[i,j]/average**

**}**

**}**

**max=max(relave)**

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

**barplot(relave[1,]~TOPIC,ylim=c(0,max),main="1865")**

**abline(h=1)**

**barplot(relave[2,]~TOPIC,ylim=c(0,max),main="1866")**

**abline(h=1)**

**barplot(relave[3,]~TOPIC,ylim=c(0,max),main="1867")**

**abline(h=1)**

**## Topic 16 (on slavery and race) is more important in early years (1865)**

**## while Topics 8 and 15 (on commerce/business) is more important in later years (1867)**

**## stratify according to the kkk most frequent speakers**

**kkk=40**

**frequent=dim(kkk)**

**frequency=dim(kkk)**

**tt=sort(table(meta2),decreasing=TRUE)**

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

**frequent[i]=labels(tt[i])**

**frequency[i]=tt[i]**

**}**

**frequent**

**frequency**

**avetop=matrix(nrow=kkk,ncol=nutopics)**

**relavetop=matrix(nrow=kkk,ncol=nutopics)**

**for (j in 1:nutopics) {**

**uu=results$theta[,j]**

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

**avetop[i,j]=mean(uu[meta2==frequent[i]])**

**}**

**}**

**## for selected topics**

**hh=data.frame(frequent,frequency,round(avetop[,6],3),round(avetop[,15],3),round(avetop[,16],3),round(avetop[,17],3))**

**print(hh)**

**for (j in 1:nutopics) {**

**average=mean(avetop[,j])**

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

**relavetop[i,j]=avetop[i,j]/average ## compared to average for each topic**

**}**

**}**

**max=max(relavetop)**

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

**barplot(relavetop[17,]~TOPIC,ylim=c(0,max),main=frequent[17])**

**abline(h=1)**

**barplot(relavetop[24,]~TOPIC,ylim=c(0,max),main=frequent[24])**

**abline(h=1)**

**barplot(relavetop[27,]~TOPIC,ylim=c(0,max),main=frequent[27])**

**abline(h=1)**

**barplot(relavetop[28,]~TOPIC,ylim=c(0,max),main=frequent[28])**

**abline(h=1)**

**barplot(relavetop[38,]~TOPIC,ylim=c(0,max),main=frequent[38])**

**abline(h=1)**

**## topic 6 (indian affairs): mentioned often by DOOLITTLE, WILSON(S) and SCHENCK**

**## topic 15 (money): mentioned often by POMEROY**

**## topic 16 (slavery/race): mentioned often by BINGHAM**

**## topic 17 (infrastructure/railroads): mentioned often by POMEROY**

**## stratify according to speakers from**

**## slave states (Delaware, Kentucky, Missouri without Henderson, Maryland) vs**

**## nonslave states (Ohio, Illinois, Iowa)**

**indss1=meta2=="saulsbury"**

**table(indss1)**

**indss2=meta2=="riddle"**

**table(indss2)**

**indss3=meta2=="davis(S)"**

**table(indss3)**

**indss4=meta2=="guthrie"**

**table(indss4)**

indss5**=meta2=="brown"**

**table(indss5)**

**indss6=meta2=="johnson(S)"**

**table(indss6)**

**indss7=meta2=="creswell"**

**table(indss7)**

**indslave=indss1|indss2|indss3|indss4|indss5|indss6|indss7 ## logical or**

**table(indslave)**

**indns1=meta2=="wade"**

**table(indns1)**

**indns2=meta2=="sherman"**

**table(indns2)**

**indns3=meta2=="trumbull"**

**table(indns3)**

**indns4=meta2=="yates"**

**table(indns4)**

**indns5=meta2=="grimes"**

**table(indns5)**

**indns6=meta2=="kirkwood"**

**table(indns6)**

**indnonslave=indns1|indns2|indns3|indns4|indns5|indns6 ## logical or**

**table(indnonslave)**

**aveslave=matrix(nrow=2,ncol=nutopics)**

**relaveslave=matrix(nrow=2,ncol=nutopics)**

**for (j in 1:nutopics) {**

**uu=results$theta[,j]**

**aveslave[1,j]=mean(uu[indslave])**

**aveslave[2,j]=mean(uu[indnonslave])**

**}**

**legend=c("slave","nonslave")**

**aveslave[,]=round(aveslave[,],3)**

**hh=data.frame(legend,aveslave[,6],aveslave[,15],aveslave[,16],aveslave[,17])**

**print(hh)**

**max=max(aveslave)**

**par(mfrow=c(2,1))**

**barplot(aveslave[1,],ylim=c(0,max),main="slave")**

**barplot(aveslave[2,],ylim=c(0,max),main="non-slave")**

**## compared to average for each topic**

**for (j in 1:nutopics) {**

**average=mean(aveslave[,j])**

**for (i in 1:2) {**

**relaveslave[i,j]=aveslave[i,j]/average**

**}**

**}**

**max=max(relaveslave)**

**par(mfrow=c(2,1))**

**barplot(relaveslave[1,],ylim=c(0,max),main="slave")**

**abline(h=1)**

**barplot(relaveslave[2,],ylim=c(0,max),main="non-slave")**

**abline(h=1)**

**## topic 16 (slavery/race): mentioned more often by former slave states**