**R Code**

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

**################################# ANALYSIS ########################################**

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

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

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

**library(tm)**

**library(slam)**

**library(wordcloud)**

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

**## loading the data (see Chapter 4)**

**load("C:\\Johannes Ledolter\\2020March01Book\\Chapter10WEB\\PrelimData.RData")**

**## here we use all speeches**

**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))**

**dim(corp.dtm)**

**## BI-GRAMS**

**BigramTokenizer <- function(x)**

 **unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)**

**bi2.dtm <- DocumentTermMatrix(corpus5, control = list(tokenize = BigramTokenizer))**

**dim(bi2.dtm)**

**Bbi2.dtm=weightBin(bi2.dtm)**

**nn=dim(Bbi2.dtm)[2]**

**string=c(1:nn)**

**stringind=dim(nn)**

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

**stringind[i]=FALSE**

**}**

**stringind[string[Terms(Bbi2.dtm)=="senfrom connecticut"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom delaware"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom illinois"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom indiana"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom iowa"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom kansas"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom kentucky"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom maine"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom maryland"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom massachusetts"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom michigan"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom minnesota"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom missouri"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom nevada"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom newhampshire"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom newjersey"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom newyork"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom ohio"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom oregon"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom pennsylvania"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom rhdisland"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom vermont"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom westvirginia"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom wisconsin"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom california"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom tennessee"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom virginia"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom florida"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom louisiana"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom mississippi"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom ncarolina"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom scarolina"]]=TRUE**

**stringind[string[Terms(Bbi2.dtm)=="senfrom texas"]]=TRUE**

**table(stringind)**

**## matrix 1**

**Bbi2red.dtm=Bbi2.dtm[,stringind]**

**dim(Bbi2red.dtm)**

**Terms(Bbi2red.dtm)**

**senators=c("foot","wade","sumner","trumbull","collamer","foster","harlan","chandler","doolittle","clark","anthony","saulsbury","grimes","howe","mcdougall","lane","nesmith","cowan","sherman","pomeroy(S)","howard","ramsey","sprague","conness","hendricks","buckalew","willey","vanwinkle","brown","riddle","stewart","nye","cragin","yates","fessenden","norton","guthrie","creswell","stockton","poland","kirkwood","edmunds","fowler","fogg","cattell","henderson(S)","wilson(S)","johnson(S)","davis(S)","harris(S)","williams(S)","dixon(S)","morgan(S)","wright(S)","ross(S)","patterson(S)","morrill(S)","frelinghuysen")**

**senators**

**state=c("vermont","ohio","massachusetts","illinois","vermont","connecticutt","iowa",**

**"michigan","wisconsin","new hampshire","rhode island","delaware","iowa","wisconsin",**

**"california","indiana/kansas","oregon","pennsylvania","ohio","kansas","michigan",**

**"minnesota","rhode island","california","indiana", "pennsylvania","west virginia","west virginia",**

**"missouri","delaware","nevada","nevada","new hampshire","illinois","maine",**

**"minnesota","kentucky","maryland","new jersey","vermont","iowa","vermont",**

**"tennessee","new hampshire","new jersey","missouri","massachusetts","maryland","kentucky",**

**"new york","oregon","connecticutt","new york", "new jersey","kansas","tennessee","maine","new jersey")**

**table(state)**

**data.frame(senators,state)**

**## matrix 2**

**n=length(senators)**

**nr=dim(Bbi2red.dtm)[2]**

**nu=dim(n)**

**out=matrix(nrow=n,ncol=nr)**

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

**ff=Bbi2red.dtm[meta2==senators[i],]**

**nu[i]=dim(ff)[1]**

**out[i,]=colSums(as.matrix(ff))**

**}**

**dd=data.frame(senators,state,nu,out)**

**dim(dd)**

**f1=Terms(Bbi2red.dtm)**

**f2=gsub("senfrom ","",f1)**

**f2**

**colnames(dd)=c("senator","state","#speeches",f2)**

**dd**

**dd=dd[order(dd[,3],decreasing=TRUE),]**

**dd**

**## table of number of references made by senator (row) to senators of states (columns)**

**ddcount=dd**

**ddcount**

**## total number of references to the various states (most made to senator of mass)**

**refsen=colSums(dd[,4:(nr+3)])**

**refsen**

**numentions=refsen[order(refsen,decreasing=TRUE)]**

**numentions**

**library(wordcloud)**

**freq=dim(length(numentions))**

**for (i in 1:length(numentions)) {**

**freq[i]=numentions[[i]]**

**}**

**freq**

**## displaying word clouds**

**set.seed(142)**

**dark2 <- brewer.pal(6,"Dark2")**

**wordcloud(labels(numentions),freq,max.words=50,rot.per=0.2,colors=dark2)**

**## table of row proportions**

**nrow=dim(dd)[1]**

**ncol=dim(dd)[2]**

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

**sum=0**

**for (j in 4:ncol) {**

**sum=sum+dd[i,j]**

**}**

**for (j in 4:ncol) {**

**dd[i,j]=round(dd[i,j]/sum,3)**

**}**

**}**

**ddratio=dd ## ratios**

**ddratio**

**## determining the most frequent state referral**

**corr=dim(nrow)**

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

**terms=f2**

**gg=dd[i,4:ncol]**

**corr[i]=terms[gg==max(gg)]**

**}**

**referral=data.frame(dd[,1:3],corr)**

**colnames(referral)=c("senator","state","#speeches","most frequent referral")**

**referral**

**## for simplifying the tables**

**limitr=25 ## senator mentioning senfrom at least 25 times**

**indicatorr=c(rowSums(dd[,4:(nr+3)]))>limitr**

**indicatorr**

**limitc=300 ## state mentioned at least 300 times**

**indicatorc=c(TRUE,TRUE,TRUE,colSums(dd[,4:(nr+3)])>limitc)**

**indicatorc**

**## table of row ratios**

**## ratios only shown for senators who address more than limitr senators**

**## ratios only shown for states that are mentioned more than limitc times**

**ddratiored=dd[indicatorr,]**

**ddratiored=ddratiored[,indicatorc]**

**ddratiored**

**Discussion**

The speeches include references to other speakers, and it is of interest to learn who addresses whom. When formatting the text we had inserted the keyword "senfrom" in front of a state whenever a speaker addressed a Senator from a particular state. For example, "senfrom Iowa" indicates that the speech makes reference to a senator from Iowa. We can use this information to learn which states are referred to most often; this tells us how important senators from a certain state are in the discourse. Here we only look at speeches in the Senate and stratify this information according to each Senator. This tells us how each Senator from a given state engages with his Senate colleagues from the other states.

The enclosed R program processes this information. We use the document-term matrix for bigrams to learn whether a speech includes terms such as "senfrom iowa", "senfrom massachusetts", … . This leads us to a very large table of 0/1 indicators; the row of the table indicates the speech, the column indicates the state, and the indicator expresses whether a particular speech referred to (the Senator of) a particular state. We then summarize the information for each Senate speaker. The summary matrix displays the number of speeches given by each Senator that refer to (Senators of) each state. The column sums of this matrix express the number of Senate speeches that refer to (Senators of) the various states.

Below we list the number of times senators from a particular state are mentioned in Senate speeches. Senators from the state of Massachusetts, Ohio and Maine are mentioned most often. The large number of references to Senators from these states reflects their importance in the Senate discourse. The word cloud of states, with their font size selected proportional to their frequencies, visualizes this information.

**massachusetts ohio maine illinois iowa indiana**

 **937 895 763 564 440 431**

 **wisconsin maryland vermont michigan california missouri**

 **423 387 357 334 306 275**

 **kentucky oregon pennsylvania nevada kansas newhampshire**

 **258 251 220 198 185 153**

 **newyork newjersey connecticut rhdisland delaware westvirginia**

 **150 146 130 130 113 93**

 **minnesota tennessee virginia texas florida louisiana**

 **71 24 4 2 1 1**

 **mississippi ncarolina scarolina**

 **1 1 1**



The table of row proportions shown below expresses how the speeches of each Senator that refer to another Senator are distributed across the different states. For example, 16.7 percent of the speeches of the Maine Senator Fessenden that include a reference to another Senator are directed to (a Senator from) Massachussetts. [Note that while Senator Fessenden has given 4,213 speeches, not all his speeches make a reference to a Senate colleague]. Senator's Fessenden most frequent reference is to Senators from the state of Massachusetts. The following table indicates this information for the other Senators. We don't show all senators (only senators with more than references are shown) and we don't show all states (only states that are mentioned more than 300 times are shown).

 **senator state #speeches california illinois indiana iowa maine maryland massachusetts michigan ohio vermont wisconsin**

**35 fessenden maine 4213 0.031 0.053 0.042 0.068 0.020 0.062 0.167 0.044 0.134 0.046 0.046**

**4 trumbull illinois 2425 0.027 0.009 0.089 0.041 0.090 0.068 0.090 0.030 0.098 0.043 0.065**

**19 sherman ohio 2033 0.032 0.072 0.058 0.075 0.130 0.026 0.167 0.037 0.014 0.061 0.058**

**13 grimes iowa 1845 0.047 0.047 0.038 0.007 0.083 0.033 0.135 0.071 0.121 0.045 0.052**

**3 sumner massachusetts 1781 0.036 0.090 0.055 0.083 0.168 0.088 0.021 0.012 0.135 0.081 0.040**

**24 conness california 1731 0.011 0.034 0.063 0.069 0.117 0.020 0.094 0.037 0.109 0.034 0.051**

**25 hendricks indiana 1677 0.050 0.095 0.008 0.050 0.085 0.037 0.091 0.056 0.131 0.019 0.035**

**48 johnson(S) maryland 1668 0.047 0.054 0.034 0.041 0.162 0.034 0.128 0.061 0.095 0.034 0.061**

**47 wilson(S) massachusetts 1558 0.050 0.083 0.062 0.067 0.100 0.054 0.025 0.012 0.083 0.038 0.054**

**2 wade ohio 1204 0.035 0.042 0.049 0.042 0.125 0.042 0.188 0.035 0.035 0.076 0.076**

**46 henderson(S) missouri 1175 0.021 0.121 0.047 0.057 0.078 0.069 0.097 0.047 0.116 0.038 0.040**

**9 doolittle wisconsin 1151 0.038 0.091 0.061 0.045 0.053 0.008 0.170 0.049 0.193 0.034 0.008**

**14 howe wisconsin 1058 0.013 0.058 0.045 0.042 0.127 0.127 0.097 0.032 0.091 0.068 0.013**

**21 howard michigan 1040 0.054 0.051 0.069 0.036 0.058 0.047 0.090 0.000 0.101 0.101 0.079**

**10 clark new hampshire 1034 0.039 0.066 0.135 0.023 0.054 0.031 0.097 0.066 0.081 0.035 0.042**

**57 morrill(S) maine 999 0.013 0.066 0.060 0.066 0.013 0.093 0.093 0.046 0.079 0.086 0.040**

**20 pomeroy(S) kansas 867 0.100 0.060 0.033 0.067 0.067 0.007 0.113 0.047 0.100 0.047 0.087**

**8 chandler michigan 786 0.018 0.027 0.036 0.089 0.277 0.009 0.062 0.027 0.125 0.045 0.027**

**16 lane indiana/kansas 771 0.000 0.069 0.026 0.164 0.052 0.017 0.207 0.000 0.121 0.043 0.034**

**51 williams(S) oregon 768 0.053 0.047 0.073 0.027 0.053 0.047 0.067 0.067 0.100 0.073 0.107**

**42 edmunds vermont 706 0.032 0.056 0.024 0.048 0.161 0.065 0.073 0.073 0.145 0.040 0.065**

**49 davis(S) kentucky 633 0.062 0.207 0.019 0.034 0.043 0.067 0.111 0.048 0.062 0.029 0.053**

**31 stewart nevada 616 0.076 0.065 0.054 0.033 0.054 0.022 0.174 0.043 0.120 0.011 0.087**

**11 anthony rhode island 572 0.045 0.067 0.034 0.112 0.090 0.022 0.101 0.034 0.067 0.000 0.056**

**18 cowan pennsylvania 530 0.014 0.092 0.028 0.028 0.078 0.000 0.191 0.078 0.128 0.028 0.057**

**15 mcdougall california 452 0.070 0.070 0.051 0.045 0.032 0.051 0.255 0.032 0.115 0.019 0.051**

**22 ramsey minnesota 432 0.059 0.039 0.059 0.098 0.039 0.000 0.157 0.078 0.118 0.000 0.039**

**40 poland vermont 428 0.026 0.105 0.066 0.026 0.171 0.026 0.118 0.066 0.079 0.039 0.026**

**29 brown missouri 377 0.043 0.065 0.022 0.152 0.109 0.043 0.065 0.000 0.109 0.043 0.022**

**32 nye nevada 368 0.065 0.057 0.057 0.098 0.049 0.041 0.033 0.016 0.138 0.016 0.122**

**23 sprague rhode island 353 0.157 0.043 0.014 0.086 0.214 0.000 0.071 0.071 0.071 0.014 0.043**

**50 harris(S) new york 347 0.038 0.113 0.038 0.038 0.189 0.019 0.094 0.075 0.057 0.057 0.132**

**38 creswell maryland 331 0.000 0.000 0.065 0.043 0.217 0.000 0.217 0.000 0.174 0.065 0.065**

**41 kirkwood iowa 330 0.032 0.048 0.081 0.000 0.097 0.065 0.145 0.048 0.048 0.048 0.048**

**26 buckalew pennsylvania 317 0.044 0.067 0.056 0.011 0.000 0.056 0.133 0.033 0.289 0.067 0.033**

**52 dixon(S) connecticutt 304 0.017 0.149 0.000 0.033 0.149 0.008 0.207 0.050 0.116 0.008 0.058**

**12 saulsbury delaware 295 0.021 0.115 0.021 0.021 0.021 0.042 0.177 0.021 0.104 0.021 0.042**

**28 vanwinkle west virginia 287 0.000 0.050 0.025 0.050 0.100 0.025 0.100 0.100 0.250 0.050 0.025**

**27 willey west virginia 249 0.044 0.059 0.059 0.059 0.132 0.044 0.103 0.015 0.074 0.044 0.029**

**17 nesmith oregon 188 0.074 0.056 0.037 0.111 0.185 0.000 0.167 0.037 0.037 0.000 0.074**

**34 yates illinois 130 0.017 0.017 0.150 0.000 0.117 0.083 0.183 0.000 0.117 0.017 0.033**

**6 foster connecticutt 123 0.016 0.000 0.164 0.180 0.082 0.000 0.098 0.066 0.033 0.000 0.016**

The maximum of each vector of row proportions indicates the state that is referred to most often. It is the state of Massachusetts for Senator Fessenden of Maine, and it is the state of Maine for the Senator Sherman of Ohio.

 **senator state #speeches most frequent referral**

**35 fessenden maine 4213 massachusetts**

**4 trumbull illinois 2425 ohio**

**19 sherman ohio 2033 massachusetts**

**13 grimes iowa 1845 massachusetts**

**3 sumner massachusetts 1781 maine**

**24 conness california 1731 maine**

**25 hendricks indiana 1677 ohio**

**48 johnson(S) maryland 1668 maine**

**47 wilson(S) massachusetts 1558 maine**

**2 wade ohio 1204 massachusetts**

**46 henderson(S) missouri 1175 illinois**

**9 doolittle wisconsin 1151 ohio**

**14 howe wisconsin 1058 maine**

**21 howard michigan 1040 ohio**

**10 clark new hampshire 1034 indiana**

**57 morrill(S) maine 999 maryland**

**20 pomeroy(S) kansas 867 massachusetts**

**8 chandler michigan 786 maine**

**16 lane indiana/kansas 771 massachusetts**

**51 williams(S) oregon 768 wisconsin**

**42 edmunds vermont 706 maine**

**49 davis(S) kentucky 633 illinois**

**31 stewart nevada 616 massachusetts**

**11 anthony rhode island 572 iowa**

**18 cowan pennsylvania 530 massachusetts**

**15 mcdougall california 452 massachusetts**

**22 ramsey minnesota 432 massachusetts**

**40 poland vermont 428 maine**

**29 brown missouri 377 iowa**

**32 nye nevada 368 ohio**

**23 sprague rhode island 353 maine**

**50 harris(S) new york 347 maine**

**38 creswell maryland 331 maine**

**41 kirkwood iowa 330 missouri**

**26 buckalew pennsylvania 317 ohio**

**52 dixon(S) connecticutt 304 massachusetts**

**12 saulsbury delaware 295 massachusetts**

**28 vanwinkle west virginia 287 ohio**

**27 willey west virginia 249 maine**

**17 nesmith oregon 188 maine**

**53 morgan(S) new york 145 ohio**

**34 yates illinois 130 massachusetts**

**37 guthrie kentucky 128 massachusetts**

**6 foster connecticutt 123 iowa**

**33 cragin new hampshire 92 indiana**

**58 frelinghuysen new jersey 90 ohio**

**36 norton minnesota 82 wisconsin**

**45 cattell new jersey 71 missouri**

**30 riddle delaware 69 iowa**

**1 foot vermont 61 maine**

**55 ross(S) kansas 56 illinois**

**39 stockton new jersey 46 maine**

**43 fowler tennessee 39 missouri**

**56 patterson(S) tennessee 32 <NA>**

**44 fogg new hampshire 24 connecticut**

**54 wright(S) new jersey 19 maine**

**7 harlan iowa 3 illinois**

**5 collamer vermont 2 <NA>**

This analysis illustrates how the text and the information from meta variables can be used to address specific interesting hypotheses (like in this case, who refers to whom). As with all other analyses, it takes problem solving skills and computing experience to get this done.