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

**###################### ANALYSIS (SECTION 8.2: 39th Congress) ###########################**

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

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

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

**library(tm)**

**library(slam)**

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

**load("C:\\Johannes Ledolter\\2020March01Book\\Chapter8WEB\\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"**

**## 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**

**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]]**

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

**## cutting out short letters**

**## working with / simplifying corpus**

**dim(corp.dtm)**

**## as.matrix(corp.dtm)**

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

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

**dim(corp.dtm)**

**corpStripped.dtm=removeSparseTerms(corp.dtm,0.98)**

**## terms that show up in at least 2 percent of documents**

**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]]**

**dim(corpStripped.dtm)**

**length(meta2)**

**sort(table(meta2))**

**meta2orig=meta2**

**## CASE 1: comparing two speakers (stevens vs johnson)**

**## data set-up**

**meta2=meta2orig**

**ind1=meta2=="stevens" ## liberal**

**ind2=meta2=="johnson(S)" ## conservative**

**indcomb=ind1|ind2 ## logical or**

**table(ind1)**

**table(ind2)**

**table(indcomb)**

**cor1=corpStripped.dtm[indcomb,]**

**dim(cor1)**

**ndoc=dim(cor1)[1]**

**meta2=meta2[indcomb]**

**length(meta2)**

**## considering occurrences**

**Bcor1=weightBin(cor1) ## for 0/1 occurrences**

**Bxx=as.matrix(Bcor1)**

**dim(Bxx)**

**## considering occurrences**

**## considering frequencies**

**Fxx=as.matrix(cor1)**

**dim(Fxx)**

**## considering frequencies**

**## determining test and evaluation documents**

**nt=round(ndoc\*0.80)**

**ndoc**

**nt**

**set.seed(1) ## to make the calculations reproducible in repeated runs**

**training <- sample(1:ndoc,nt)**

**## the evaluation carried out below uses just one simulation**

**## we study results of a single data split (80/20) into an estimation and an evaluation data set**

**## you may want to use repeated simulations for the 80/20 split**

**## nearest neighbor method**

**library(class)**

**set.seed(1) ## as ties broken at random**

**nearest1 <- knn(train=Bxx[training,],test=Bxx[-training,],cl=meta2[training],k=1)**

**nearest5 <- knn(train=Bxx[training,],test=Bxx[-training,],cl=meta2[training],k=5)**

**nearest10 <- knn(train=Bxx[training,],test=Bxx[-training,],cl=meta2[training],k=10)**

**nearest25 <- knn(train=Bxx[training,],test=Bxx[-training,],cl=meta2[training],k=25)**

**data.frame(meta2[-training],nearest1,nearest5,nearest10,nearest25)[1:50,]**

**## calculate the proportion of correct classifications on this one training set**

**pcorrn1=100\*sum(meta2[-training]==nearest1)/(ndoc-nt)**

**pcorrn5=100\*sum(meta2[-training]==nearest5)/(ndoc-nt)**

**pcorrn10=100\*sum(meta2[-training]==nearest10)/(ndoc-nt)**

**pcorrn25=100\*sum(meta2[-training]==nearest25)/(ndoc-nt)**

**mpcorrn1=100-pcorrn1**

**mpcorrn5=100-pcorrn5**

**mpcorrn10=100-pcorrn10**

**mpcorrn25=100-pcorrn25**

**pcorrn1**

**pcorrn5**

**pcorrn10**

**pcorrn25**

**mpcorrn1**

**mpcorrn5**

**mpcorrn10**

**mpcorrn25**

**## NAÏVE BAYES Bernouilli (detailed program)**

**mm=dim(Bxx)[2]**

**mm**

**pastdata=Bxx[training,]**

**pastcl=meta2[training]**

**testdata=Bxx[-training,]**

**testcl=meta2[-training]**

**pastdatag1=pastdata[pastcl=="stevens",]**

**pastdatag2=pastdata[pastcl=="johnson(S)",]**

**pastclg1=pastcl[pastcl=="stevens"]**

**pastclg2=pastcl[pastcl=="johnson(S)"]**

**propprior=length(pastclg1)/length(pastcl)**

**propprior**

**ratioprior=(1-propprior)/propprior**

**ratioprior**

**alpha=1 ## with Laplace smoothing**

**prop1=(colSums(pastdatag1)+alpha)/(length(pastclg1)+(2\*alpha)) ## Laplace smoothing**

**prop2=(colSums(pastdatag2)+alpha)/(length(pastclg2)+(2\*alpha)) ## Laplace smoothing**

**factor1=dim(mm)**

**factor2=dim(mm)**

**ratio=dim(mm)**

**clpred=dim(length(testcl))**

**prop=dim(length(testcl))**

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

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

**factor1[j]=testdata[i,j]\*prop1[j]+(1-testdata[i,j])\*(1-prop1[j])**

**factor2[j]=testdata[i,j]\*prop2[j]+(1-testdata[i,j])\*(1-prop2[j])**

**ratio[j]=factor2[j]/factor1[j]**

**}**

**sus=0**

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

**sus=sus+log(ratio[j])**

**}**

**fac=exp(sus)**

**prop[i]=1/(1+fac\*ratioprior)**

**if(prop[i]>=0.5) clpred[i]="stevens" ## "slavestate"**

**if(prop[i]<=0.5) clpred[i]="johnson(S)" ## "nonslavestate"**

**}**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**

**round(prop,dig=3)**

**library(lattice)**

**hist(prop)**

**boxplot(prop~testcl)**

**mean(prop[testcl=="stevens"]) ## "slavestate"**

**mean(prop[testcl=="johnson(S)"]) ## "nonslavestate"**

**## NAÏVE BAYES Bernouilli (Package: naivebayes)**

**library(naivebayes)**

**pastdata=Bxx[training,]**

**pastcl=meta2[training]**

**testdata=Bxx[-training,]**

**testcl=meta2[-training]**

**bnb <- bernoulli\_naive\_bayes(x=pastdata,y=pastcl,laplace=1)**

**bnb**

**head(predict(bnb,newdata=testdata,type="class"))**

**head(predict(bnb,newdata=testdata,type="prob"))**

**clpred=predict(bnb,newdata=testdata,type="class")**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**

**## NAÏVE BAYES Binomial (Package: naivebayes)**

**pastdata=Fxx[training,]**

**pastcl=meta2[training]**

**testdata=Fxx[-training,]**

**testcl=meta2[-training]**

**binomnb=multinomial\_naive\_bayes(pastdata,pastcl,prior=NULL,laplace=1)**

**binomnb**

**head(predict(binomnb,newdata=testdata,type="class"))**

**head(predict(binomnb,newdata=testdata,type="prob"))**

**clpred=predict(binomnb,newdata=testdata,type="class")**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**

**## CASE 2: combining groups of speakers**

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

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

**meta2=meta2orig**

**length(meta2)**

**dim(corpStripped.dtm)**

**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)**

**indcomb=indslave|indnonslave ## logical or**

**table(indcomb)**

**indslave=indslave[indcomb]**

**indnonslave=indnonslave[indcomb]**

**meta2=meta2[indcomb]**

**ndoc=length(meta2)**

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

**if (indslave[i]==TRUE) meta2[i]="slavestate"**

**if (indnonslave[i]==TRUE) meta2[i]="nonslavestate"**

**}**

**cor1=corpStripped.dtm[indcomb,]**

**ndoc=dim(cor1)[1]**

**xx=as.matrix(cor1)**

**## considering occurrences**

**Bcor1=weightBin(cor1) ## to make binary**

**Bxx=as.matrix(Bcor1)**

**dim(Bxx)**

**## considering occurrences**

**## considering frequencies**

**Fxx=as.matrix(cor1)**

**dim(Fxx)**

**## considering frequencies**

**## determining test and evaluation documents**

**nt=round(ndoc\*0.80)**

**ndoc**

**nt**

**set.seed(1) ## to make the calculations reproducible in repeated runs**

**training <- sample(1:ndoc,nt)**

**## the evaluation carried out below uses just one simulation**

**## we study results of a single data split (80/20) into an estimation and an evaluation data set**

**## you may want to use repeated simulations for the 80/20 split**

**## nearest neighbor method**

**library(class)**

**set.seed(1) ## as ties broken at random**

**nearest1 <- knn(train=xx[training,],test=xx[-training,],cl=meta2[training],k=1)**

**nearest5 <- knn(train=xx[training,],test=xx[-training,],cl=meta2[training],k=5)**

**nearest10 <- knn(train=xx[training,],test=xx[-training,],cl=meta2[training],k=10)**

**nearest25 <- knn(train=xx[training,],test=xx[-training,],cl=meta2[training],k=25)**

**## data.frame(meta2[-training],nearest1,nearest5,nearest10,nearest25)[1:50,]**

**## calculate the proportion of correct classifications on this one training set**

**pcorrn1=100\*sum(meta2[-training]==nearest1)/(ndoc-nt)**

**pcorrn5=100\*sum(meta2[-training]==nearest5)/(ndoc-nt)**

**pcorrn10=100\*sum(meta2[-training]==nearest10)/(ndoc-nt)**

**pcorrn25=100\*sum(meta2[-training]==nearest25)/(ndoc-nt)**

**mpcorrn1=100-pcorrn1**

**mpcorrn5=100-pcorrn5**

**mpcorrn10=100-pcorrn10**

**mpcorrn25=100-pcorrn25**

**pcorrn1**

**pcorrn5**

**pcorrn10**

**pcorrn25**

**mpcorrn1**

**mpcorrn5**

**mpcorrn10**

**mpcorrn25**

**## NAÏVE BAYES Bernouilli (detailed program)**

**mm=dim(Bxx)[2]**

**mm**

**pastdata=Bxx[training,]**

**pastcl=meta2[training]**

**testdata=Bxx[-training,]**

**testcl=meta2[-training]**

**pastdatag1=pastdata[pastcl=="slavestate",] ## "slavestate"**

**pastdatag2=pastdata[pastcl=="nonslavestate",] ## "nonslavestate"**

**pastclg1=pastcl[pastcl=="slavestate"] ## "slavestate"**

**pastclg2=pastcl[pastcl=="nonslavestate"] ## "nonslavestate"**

**propprior=length(pastclg1)/length(pastcl)**

**propprior**

**ratioprior=(1-propprior)/propprior**

**ratioprior**

**alpha=1 ## with Laplace smoothing**

**prop1=(colSums(pastdatag1)+alpha)/(length(pastclg1)+(2\*alpha)) ## Laplace smoothing**

**prop2=(colSums(pastdatag2)+alpha)/(length(pastclg2)+(2\*alpha)) ## Laplace smoothing**

**factor1=dim(mm)**

**factor2=dim(mm)**

**ratio=dim(mm)**

**clpred=dim(length(testcl))**

**prop=dim(length(testcl))**

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

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

**factor1[j]=testdata[i,j]\*prop1[j]+(1-testdata[i,j])\*(1-prop1[j])**

**factor2[j]=testdata[i,j]\*prop2[j]+(1-testdata[i,j])\*(1-prop2[j])**

**ratio[j]=factor2[j]/factor1[j]**

**}**

**sus=0**

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

**sus=sus+log(ratio[j])**

**}**

**fac=exp(sus)**

**prop[i]=1/(1+fac\*ratioprior)**

**if(prop[i]>=0.5) clpred[i]="slavestate" ## "slavestate"**

**if(prop[i]<=0.5) clpred[i]="nonslavestate" ## "nonslavestate"**

**}**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**

**round(prop,dig=3)**

**library(lattice)**

**hist(prop)**

**boxplot(prop~testcl)**

**mean(prop[testcl=="slavestate"]) ## "slavestate"**

**mean(prop[testcl=="nonslavestate"]) ## "nonslavestate"**

**## NAÏVE BAYES Bernouilli (Package: naivebayes)**

**library(naivebayes)**

**pastdata=Bxx[training,]**

**pastcl=meta2[training]**

**testdata=Bxx[-training,]**

**testcl=meta2[-training]**

**bnb <- bernoulli\_naive\_bayes(x=pastdata,y=pastcl,laplace=1)**

**bnb**

**head(predict(bnb,newdata=testdata,type="class"))**

**head(predict(bnb,newdata=testdata,type="prob"))**

**clpred=predict(bnb,newdata=testdata,type="class")**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**

**## NAÏVE BAYES Binomial (Package: naivebayes)**

**library(naivebayes)**

**pastdata=Fxx[training,]**

**pastcl=meta2[training]**

**testdata=Fxx[-training,]**

**testcl=meta2[-training]**

**binomnb=multinomial\_naive\_bayes(pastdata,pastcl,prior=NULL,laplace=1)**

**binomnb**

**head(predict(binomnb,newdata=testdata,type="class"))**

**head(predict(binomnb,newdata=testdata,type="prob"))**

**clpred=predict(binomnb,newdata=testdata,type="class")**

**tt=table(testcl,clpred)**

**tt**

**misprop=(tt[1,2]+tt[2,1])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**misprop**