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

**## Johannes Ledolter, University of Iowa**

**## Program to read the raw data (speeches of 43 Congress)**

**## Matt Gentzkow, Jesse M. Shapiro, and Matt Taddy: https://data.stanford.edu/congress\_text**

**## Merging the text with meta variables (on speaker and time)**

**## Three methods**

**## Method 1**

**## only considers speeches with clearly-identified speakers. Stored in speechesWithMeta1\_xxx.txt**

**## uses the file xxx\_SpeakerMap.txt for matching**

**## advantage: corrected speaker information**

**## disadvantage: fewer speeches**

**## Method 2**

**## uses all speeches. Stored in speechesWithMeta2\_xxx.txt**

**## uses the file descr\_xxx.txt file**

**## speakercorr contains corrected speaker names (marked as "unknown" if no correction**

**## can be found); the names in speakercorr are certainly better than those in speakerorig**

**## includes the date of the speech and speaker information such chamber, state (even though state ## is often missing)**

**## advantage: contains all speeches, but speaker meta variable not as clean as in**

**## speechesWithMeta1\_xxx.txt**

**## Method 3**

**## Method 3 == Method 1 (plus the meta variable date)**

**## only considers speeches with clearly-identified speakers. Stored in speechesWithMeta3\_xxx.txt**

**## advantage: corrected speaker information**

**## disadvantage: fewer speeches**

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

**## merging with meta variables**

**## Method 1**

**## only considers speeches with clearly-identified speakers. Stored in speechesWithMeta1\_xxx.txt**

**## uses the file xxx\_SpeakerMap.txt for matching**

**## advantage: corrected speaker information**

**## disadvantage: fewer speeches**

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

**library(stringi)**

**library(tm)**

**## read from txt file**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speeches\_043.txt')**

**## text file**

**datatext=dim(length(test))**

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

**datatext[i]=test[i]**

**}**

**datatext[1:10]**

**length(datatext)**

**speech=dim(length(datatext))**

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

**txt=datatext[i]**

**temp=strsplit(txt," ")[[1]]**

**speech[i]=stri\_sub(temp[1],1,9)**

**temp[1]=stri\_sub(temp[1],11,nchar(temp[1]))**

**datatext[i]=toString(temp)**

**datatext[i]=gsub(",","",datatext[i])**

**}**

**speech[1:10]**

**datatext[1:10]**

**number=dim(length(datatext))**

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

**number[i]=i**

**}**

**number**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\043\_SpeakerMap.txt') ## speaker file**

**dataaut=dim(length(test))**

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

**dataaut[i]=test[i]**

**}**

**dataaut[1:10]**

**length(dataaut)**

**aut=matrix(nrow=length(dataaut),ncol=8)**

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

**txt=dataaut[i]**

**txt=gsub("[|]"," ",txt)**

**txt=strsplit(txt," ")[[1]]**

**for (j in 1:8) {**

**aut[i,j]=txt[j]**

**}**

**}**

**aut[1:10,1:8]**

**ind=dim(length(datatext))**

**text=dim(length(dataaut))**

**text[1]="text"**

**for (i in 2:length(dataaut)) {**

**ind=aut[i,2]==speech**

**text[i]=datatext[number[ind]]**

**}**

**rrr=cbind(aut[,2],text,aut[,1],aut[,3],aut[,4],aut[,5],aut[,6],aut[,7],aut[,8])**

**dim(rrr)**

**rrr[1:10,]**

**## starting the analysis**

**lastname=aut[,3]**

**lastname=lastname[-1] ## stripping off the label**

**sort(table(lastname)) ## illustrates that speaker names are fairly clean**

**text=text[-1] ## stripping off the label**

**## creating corpus for speeches associated with a well-identified speaker**

**corpus <- VCorpus(VectorSource(text),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)**

**## writing the output file speechesWithMeta1\_043.tx**

**write.table(rrr,**

**file='C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta1\_043.txt',**

**sep=" ",** **row.names = FALSE, col.names = FALSE)**

**## reading from the output file speechesWithMeta1\_043.tx**

**new=read.table('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta1\_043.txt',**

**header=T)**

**## Method 2**

**## uses all speeches. Stored in speechesWithMeta2\_xxx.txt**

**## uses the file descr\_xxx.txt file**

**## speakercorr contains corrected speaker names (marked as "unknown" if no correction**

**## can be found); the names in speakercorr are certainly better than those in speakerorig**

**## includes the date of the speech and speaker information such chamber, state (even though state ## is often missing)**

**## advantage: contains all speeches, but speaker meta variable not as clean as in**

**## speechesWithMeta1\_xxx.txt**

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

**library(stringi)**

**library(tm)**

**## read from txt file**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speeches\_043.txt')**

**## text file**

**datatext=dim(length(test))**

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

**datatext[i]=test[i]**

**}**

**datatext[1:10]**

**length(datatext)**

**speech=dim(length(datatext))**

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

**txt=datatext[i]**

**temp=strsplit(txt," ")[[1]]**

**speech[i]=stri\_sub(temp[1],1,9)**

**temp[1]=stri\_sub(temp[1],11,nchar(temp[1]))**

**datatext[i]=toString(temp)**

**datatext[i]=gsub(",","",datatext[i])**

**}**

**speech[1:10]**

**datatext[1:10]**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\descr\_043.txt')**

**## speaker file**

**dataaut=dim(length(test))**

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

**dataaut[i]=test[i]**

**}**

**dataaut[1:10]**

**length(dataaut)**

**## strip blanks**

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

**txt=dataaut[i]**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**dataaut[i]=txt**

**}**

**## create speaker list: speakerorig (full information); speakercorr (last name/unknown)**

**speech=dim(length(dataaut))**

**chamber=dim(length(dataaut))**

**date=dim(length(dataaut))**

**speakerorig=dim(length(dataaut))**

**speakercorr=dim(length(dataaut))**

**state=dim(length(dataaut))**

**gender=dim(length(dataaut))**

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

**txt=dataaut[i]**

**txt=gsub("[|]"," ",txt)**

**txt=strsplit(txt," ")[[1]]**

**speech[i]=txt[1]**

**chamber[i]=txt[2]**

**date[i]=txt[3]**

**speakerorig[i]=txt[5]**

**speakercorr[i]=txt[7]**

**state[i]=txt[8]**

**gender[i]=txt[9]**

**}**

**rrr=cbind(speech,chamber,date,speakerorig,speakercorr,state,gender,datatext)**

**dim(rrr)**

**rrr[1:10,]**

**## starting the analysis**

**speakercorr=speakercorr[-1] ## stripping off the label**

**sort(table(speakercorr)) ## shows that not all speaker names have been corrected**

**datatext=datatext[-1] ## stripping off the label**

**## creating corpus for all speeches**

**corpus <- VCorpus(VectorSource(datatext),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)**

**## writing the output file speechesWithMeta2\_043.tx**

**write.table(rrr,**

**file='C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta2\_043.txt',**

**sep=" ",** **row.names = FALSE, col.names = FALSE)**

**## reading from the output file speechesWithMeta2\_043.tx**

**new=read.table('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta2\_043.txt',**

**header=T)**

**## Method 3**

**## Method 3 == Method 1 (plus the meta variable date)**

**## only considers speeches with clearly-identified speakers. Stored in speechesWithMeta3\_xxx.txt**

**## advantage: corrected speaker information**

**## disadvantage: fewer speeches**

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

**library(stringi)**

**library(tm)**

**## read from txt file**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speeches\_043.txt')**

**## text file**

**datatext=dim(length(test))**

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

**datatext[i]=test[i]**

**}**

**datatext[1:10]**

**length(datatext)**

**speech=dim(length(datatext))**

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

**txt=datatext[i]**

**temp=strsplit(txt," ")[[1]]**

**speech[i]=stri\_sub(temp[1],1,9)**

**temp[1]=stri\_sub(temp[1],11,nchar(temp[1]))**

**datatext[i]=toString(temp)**

**datatext[i]=gsub(",","",datatext[i])**

**}**

**speech[1:10]**

**datatext[1:10]**

**number=dim(length(datatext))**

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

**number[i]=i**

**}**

**number**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\descr\_043.txt') ## speaker file**

**dataaut=dim(length(test))**

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

**dataaut[i]=test[i]**

**}**

**dataaut[1:10]**

**length(dataaut)**

**## strip blanks**

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

**txt=dataaut[i]**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**txt=gsub(" ","", ignore.case = TRUE,txt)**

**dataaut[i]=txt**

**}**

**## create speaker list: speakerorig (full information); speakercorr (last name/unknown)**

**speech=dim(length(dataaut))**

**chamber=dim(length(dataaut))**

**date=dim(length(dataaut))**

**speakerorig=dim(length(dataaut))**

**speakercorr=dim(length(dataaut))**

**state=dim(length(dataaut))**

**gender=dim(length(dataaut))**

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

**txt=dataaut[i]**

**txt=gsub("[|]"," ",txt)**

**txt=strsplit(txt," ")[[1]]**

**speech[i]=txt[1]**

**chamber[i]=txt[2]**

**date[i]=txt[3]**

**speakerorig[i]=txt[5]**

**speakercorr[i]=txt[7]**

**state[i]=txt[8]**

**gender[i]=txt[9]**

**}**

**test=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\043\_SpeakerMap.txt') ## speaker file**

**dataaut=dim(length(test))**

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

**dataaut[i]=test[i]**

**}**

**dataaut[1:10]**

**length(dataaut)**

**aut=matrix(nrow=length(dataaut),ncol=8)**

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

**txt=dataaut[i]**

**txt=gsub("[|]"," ",txt)**

**txt=strsplit(txt," ")[[1]]**

**for (j in 1:8) {**

**aut[i,j]=txt[j]**

**}**

**}**

**aut[1:10,1:8]**

**ind=dim(length(datatext))**

**text=dim(length(dataaut))**

**dateshort=dim(length(dataaut))**

**dateshort[1]="date"**

**text[1]="text"**

**for (i in 2:length(dataaut)) {**

**ind=aut[i,2]==speech**

**text[i]=datatext[number[ind]]**

**dateshort[i]=date[number[i]]**

**}**

**rrr=cbind(aut[,2],dateshort,text,aut[,1],aut[,3],aut[,4],aut[,5],aut[,6],aut[,7],aut[,8])**

**dim(rrr)**

**rrr[1:10,]**

**## starting the analysis**

**text=text[-1] ## stripping off the label**

**## creating corpus for speeches associated with a well-identified speaker**

**corpus <- VCorpus(VectorSource(text),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)**

**## writing the output file speechesWithMeta1\_043.tx**

**write.table(rrr,**

**file='C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta3\_043.txt',**

**sep=" ",** **row.names = FALSE, col.names = FALSE)**

**## reading from the output file speechesWithMeta3\_043.tx**

**new=read.table('C:\\Johannes Ledolter\\2020March01Book\\Chapter2WEB\\speechesWithMeta3\_043.txt',**

**header=T)**