**Extracting documents that mention certain words**

There are times when researchers want to limit their focus to documents that mention certain words. Instead of sifting through all documents of the corpus, they want to read only those documents that mention a specified word (or a word from a specified group of words) that describes their interest. For example, among all 96,000 speeches of the 39th Congress, researchers may want to study only those documents that relate to indentured Chinese labor; and they are able to specify a list of words that describe their search, such as china, chinamen, coolie, … . They want an automatic and fast way of selecting those documents that mention at least one word from the specified group of words, and they want those documents written to a file for more detailed study. In the appendix, we list the R-code that executes such queries (in the file ExtractDocuments.docx). Also note that the code can be changed easily to the situation when one wants to select (even fewer) documents that contain at least two (or more) words from the specified list.

**###** **ExtractDocuments.docx: Extracting and saving documents that contain specified query words ### ### Program to extract all speeches that mention at least one word from a list of several words ###**

**### Example: any of the following words: "china","chinamen","chinese" ###**

**### Resulting documents are written out in full to a CSV file ###**

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

**comb=c("china","chinamen","coolie") ## words to combine**

**comb**

**data=readLines('C:\\Johannes Ledolter\\2020March01Book\\Chapter5WEB\\combined39Jan272020.txt') ## speeches of 39th Congress**

**## from file combined39.txt (with line return separating speeches)**

**## creating corpus**

**library(tm)**

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

**## creating the 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"))**

**corp.dtm <- DocumentTermMatrix(corpus5,control=list(stemming=FALSE))**

**dim(corp.dtm)**

**Bcorp.dtm=weightBin(corp.dtm) ## occurrences only**

**dim(Bcorp.dtm)**

**xx=dim(dim(Bcorp.dtm)[1])**

**for (i in 1: dim(Bcorp.dtm)[1]) {**

**xx[i]=0**

**}**

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

**ind=labels(Bcorp.dtm)$Terms==comb[i]**

**if (sum(ind)!=0) xx=xx+as.matrix(Bcorp.dtm[,ind])**

**}**

**index=xx>0**

**table(index)**

**## index=xx>1 ## if you select on any two of the words from the list of words**

**## table(index)**

**datareduced=data[index]**

**length(datareduced)**

**datareduced**

**## creating .csv output file**

**write.csv(datareduced,'C:\\Johannes Ledolter\\2020March01Book\\Chapter5WEB\\CorpusReduced.csv')**

**## reading from the .csv output file**

**data = read.csv('C:\\Johannes Ledolter\\2020March01Book\\Chapter5WEB\\CorpusReduced.csv', header = TRUE, stringsAsFactors = FALSE)**

**data[1,1:2]**

**text=data[,2]**

**text[1]**

**text**