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

**####### COMBINING DTM FOR WORDS, BIGRAMS AND TRIGRAMS ########**

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

**## start with document-term matrices for bi-grams and tri-grams**

**## DTMs come from the same corpus, hence equal row dimensions**

**dim(bi2.dtm)**

**labels(bi2.dtm)$Terms**

**dim(bi3.dtm)**

**labels(bi3.dtm)$Terms**

**as.matrix(bi2.dtm[1:2,1:20]) ## looking at first 20 bi-grams**

**n1=dim(bi3.dtm)[2]-19**

**n2=dim(bi3.dtm)[2]**

**as.matrix(bi3.dtm[1:2,n1:n2]) ## looking at last 20 tri-grams**

**## combining the two .dtm (bi-gram and tri-gram)**

**comb.dtm=as.DocumentTermMatrix(cbind(bi2.dtm,bi3.dtm),weighting=weightTf)**

**comb.dtm**

**dim(comb.dtm)**

**labels(comb.dtm)$Terms**

**findFreqTerms(comb.dtm,100)**

**as.matrix(comb.dtm[1:2,1:20]) ## looking at first 20 terms of the combined matrix**

**n1=dim(comb.dtm)[2]-19**

**n2=dim(comb.dtm)[2]**

**as.matrix(comb.dtm[1:2,n1:n2]) ## looking at last 20 terms of the combined matrix**

**## reducing terms by requiring a certain coverage**

**dim(bi2.dtm)**

**bi2corpStripped.dtm=removeSparseTerms(bi2.dtm,0.95)**

**dim(bi2corpStripped.dtm)**

**dim(bi3.dtm)**

**bi3corpStripped.dtm=removeSparseTerms(bi3.dtm,0.95)**

**dim(bi3corpStripped.dtm)**

**dim(comb.dtm)**

**combcorpStripped.dtm=removeSparseTerms(comb.dtm,0.95)**

**dim(combcorpStripped.dtm)**

**## checking results for a given tri-gram (bi3.dtm and comb.dtm lead to the same result)**

**indterm=as.matrix(bi3.dtm[,"republican form government"]) ## frequencies for the searched term**

**indterm**

**indterm=indterm>0 ## indicator whether searched term is present in a document**

**indterm**

**## meta ## vector containing the outcomes of a meta variable (such as speaker)**

**## meta[indterm] ## outcomes of meta variable for documents containing the searched term**

**indterm=as.matrix(comb.dtm[,"republican form government"]) ## frequencies for the searched term**

**indterm**

**indterm=indterm>0 ## indicator whether searched term is present**

**indterm**

**## meta ## vector containing the outcomes of a meta variable (such as speaker)**

**## meta[indterm] ## outcomes of meta variable for documents containing the searched term**