library(tm)

library(SnowballC)

library(wordcloud)

library(RColorBrewer)

sou2010 = paste(scan(url("http://textuploader.com/a5vq4/raw"), what="character"),collapse=" ")

sou2010=iconv(sou2010, "latin1", "ASCII", "")

sou2010

sou2011 = paste(scan(url("http://textuploader.com/a5vm0/raw"), what="character"),collapse=" ")

sou2011=iconv(sou2011, "latin1", "ASCII", "")

sou2012 = paste(scan(url("http://textuploader.com/a5vmp/raw"), what="character"),collapse=" ")

sou2012=iconv(sou2012, "latin1", "ASCII", "")

sou2013 = paste(scan(url("http://textuploader.com/a5vh0/raw"), what="character"),collapse=" ")

sou2013=iconv(sou2013, "latin1", "ASCII", "")

sou2014 = paste(scan(url("http://textuploader.com/a5vhp/raw"), what="character"),collapse=" ")

sou2014=iconv(sou2014, "latin1", "ASCII", "")

sou2015 = paste(scan(url("http://textuploader.com/a5vhb/raw"), what="character"),collapse=" ")

sou2015=iconv(sou2015, "latin1", "ASCII", "")

getwd()

write.table(sou2010, "c:/Users/clesmeister/chap12/text/sou2010.txt")

write.table(sou2011, "c:/Users/clesmeister/chap12/text/sou2011.txt")

write.table(sou2012, "c:/Users/clesmeister/chap12/text/sou2012.txt")

write.table(sou2013, "c:/Users/clesmeister/chap12/text/sou2013.txt")

write.table(sou2014, "c:/Users/clesmeister/chap12/text/sou2014.txt")

write.table(sou2015, "c:/Users/clesmeister/chap12/text/sou2015.txt")

name = file.path("C:/Users/clesmeister/chap12/text")

length(dir(name))

dir(name)

docs = Corpus(DirSource(name))

docs

docs = tm\_map(docs, tolower)

docs = tm\_map(docs, removeNumbers)

docs = tm\_map(docs, removePunctuation)

docs = tm\_map(docs, removeWords, stopwords("english"))

docs = tm\_map(docs, stripWhitespace)

docs = tm\_map(docs, stemDocument)

docs = tm\_map(docs, removeWords, c("applaus","can","cant","will","that","weve",

"dont","wont"))

docs <- tm\_map(docs, PlainTextDocument)

dtm <- DocumentTermMatrix(docs)

dim(dtm)

dtm = removeSparseTerms(dtm, 0.51)

dim(dtm)

rownames(dtm) = c("2010","2011","2012","2013","2014","2015")

inspect(dtm[1:6, 1:5])

freq = colSums(as.matrix(dtm))

ord = order(-freq) #order the frequency

freq[head(ord)]

freq[tail(ord)]

head(table(freq))

tail(table(freq))

findFreqTerms(dtm, 100)

findAssocs(dtm, "busi", corlimit=0.9)

findAssocs(dtm, "job", corlimit=0.9)

wordcloud(names(freq), freq, min.freq=50,scale=c(3, .5), colors=brewer.pal(6, "Dark2"))

wordcloud(names(freq), freq, max.words=30)

freq = sort(colSums(as.matrix(dtm)), decreasing=TRUE)

wf = data.frame(word=names(freq), freq=freq)

wf = wf[1:10,]

barplot(wf$freq, names=wf$word, main="Word Frequency",

xlab="Words", ylab="Counts", ylim=c(0,250))

#topic models

library(topicmodels)

set.seed(123)

lda3 = LDA(dtm, k=3, method="Gibbs")

topics(lda3)

set.seed(456)

lda4 = LDA(dtm, k=4, method="Gibbs")

topics(lda4)

terms(lda3,20)

library(qdap)

state15 = gsub("(Applause.)", "", sou2015)

speech15 = data.frame(speech=state15)

sent15 = sentSplit(speech15, "speech")

sent15$year = "2015"

state10 = gsub("(Applause.)", "", sou2010)

speech10 = data.frame(speech=state10)

sent10 = sentSplit(speech10, "speech")

sent10$year = "2010"

sentences = rbind(sent10, sent15)

which.min(sentences$pol)

pol = polarity(text.var=sentences$speech,grouping.var=sentences$year)

pol

plot(pol)

pol.df = pol$all

which.min(pol.df$polarity)

pol.df$text.var[12]

ari = automated\_readability\_index(text.var=sentences$speech,

grouping.var=sentences$year)

ari$Readability

form = formality(sentences$speech, sentences$year)

form

form$form.prop.by

plot(form)

div = diversity(sentences$speech, sentences$year)

div

dispersion\_plot(sentences$speech,

grouping.var=sentences$year,

c("economy","jobs","families"),

color="black", bg.color="white")

freq2010 = freq\_terms(sent10$speech, top=10, stopwords=Top200Words)

freq2010

freq2015 = freq\_terms(sent15$speech, top=50, stopwords=Top200Words)

freq2015