NLP, Topic Modeling, Term Distribution Analysis

Mohammed Alrashidan

```
# loading all pre-processing functions
source("/Users/mo/Desktop/Desktop/School/USF/Courses/Fall 2021/NLP/Functions/load_NLP_env.R")
path <- "/Users/mo/Desktop/Desktop/School/USF/Courses/Fall 2021/NLP/Functions/"
load_NLP_env(path)
```

```
## [1] "functions loaded:  "
```

```
speech <- read.csv("/Users/mo/Desktop/Desktop/School/USF/Courses/Fall 2021/NLP/datasets/UN_speeches.csv")
speech <- speech[speech$country == 'USA',]
```

```
# text_processing
text <- pre_process_corpus(speech, "text", replace_numbers = T,
                  root_gen = 'lemmatize', output_corpus = T)
```

```
# build tokenizer
tokenizer <- function(x) {unlist(lapply(ngrams(words(x), 1), paste, collapse = " "), use.names = FALSE)}

# set lower bound
lbound <- 2

# create DTM
dtm1 <- as.matrix(DocumentTermMatrix(text, control = list(tokenize = tokenizer, bounds = list(global = c(lbound,
Inf)))))

# create bigram DTM
tokenizer <- function(x) {unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)}
dtm2 <- as.matrix(DocumentTermMatrix(text, control = list(tokenize = tokenizer, bounds = list(global = c(lbound,
Inf)))))

# create trigram DTM
tokenizer <- function(x) {unlist(lapply(ngrams(words(x), 3), paste, collapse = " "), use.names = FALSE)}
dtm3 <- as.matrix(DocumentTermMatrix(text, control = list(tokenize = tokenizer, bounds = list(global = c(lbound,
Inf)))))

dtm <- cbind(dtm1, dtm2)
dtm <- cbind(dtm, dtm3)
dim(dtm)
```

```
## [1]    47 16123
```

1 Frequncy Analysis of all words in the corpus

```
freq_table <- data.frame(term = colnames(dtm), n = colSums(dtm),
                         freq = colSums(dtm)/sum(dtm))

freq_table <- freq_table[order(freq_table$n, decreasing = T),]
head(freq_table)
```

```
##          term    n        freq
## nation nation 1698 0.012300248
## unite   unite 1541 0.011162946
## will     will 1258 0.009112904
## world   world 1068 0.007736552
## state   state  960 0.006954204
## people people  849 0.006150124
```

2, 3 frequency of top countries that were mentioned in the USA speeches

```
library(maptools)
```

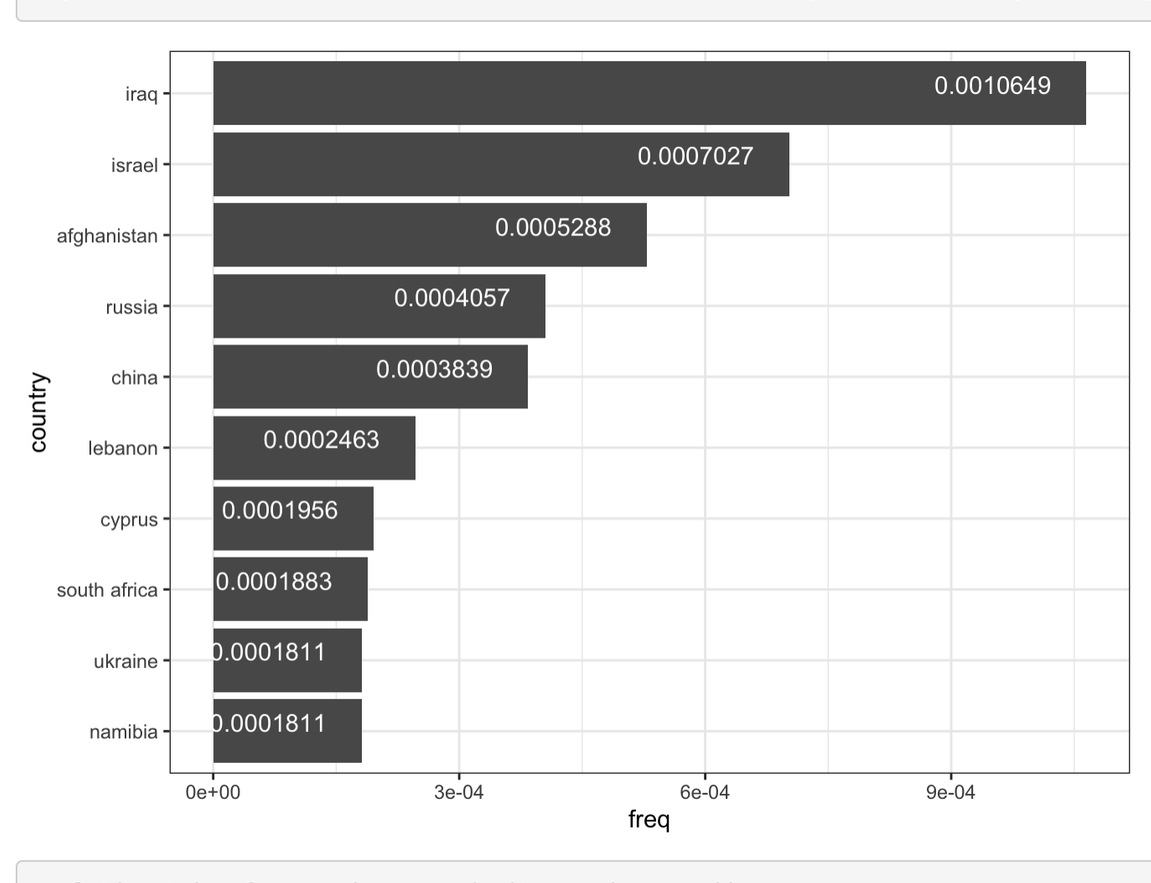
```
## Loading required package: sp
```

```
## Checking rgeos availability: FALSE
##      Note: when rgeos is not available, polygon geometry     computations in maptools depend on gpclib,
##      which has a restricted licence. It is disabled by default;
##      to enable gpclib, type gpclibPermit()
```

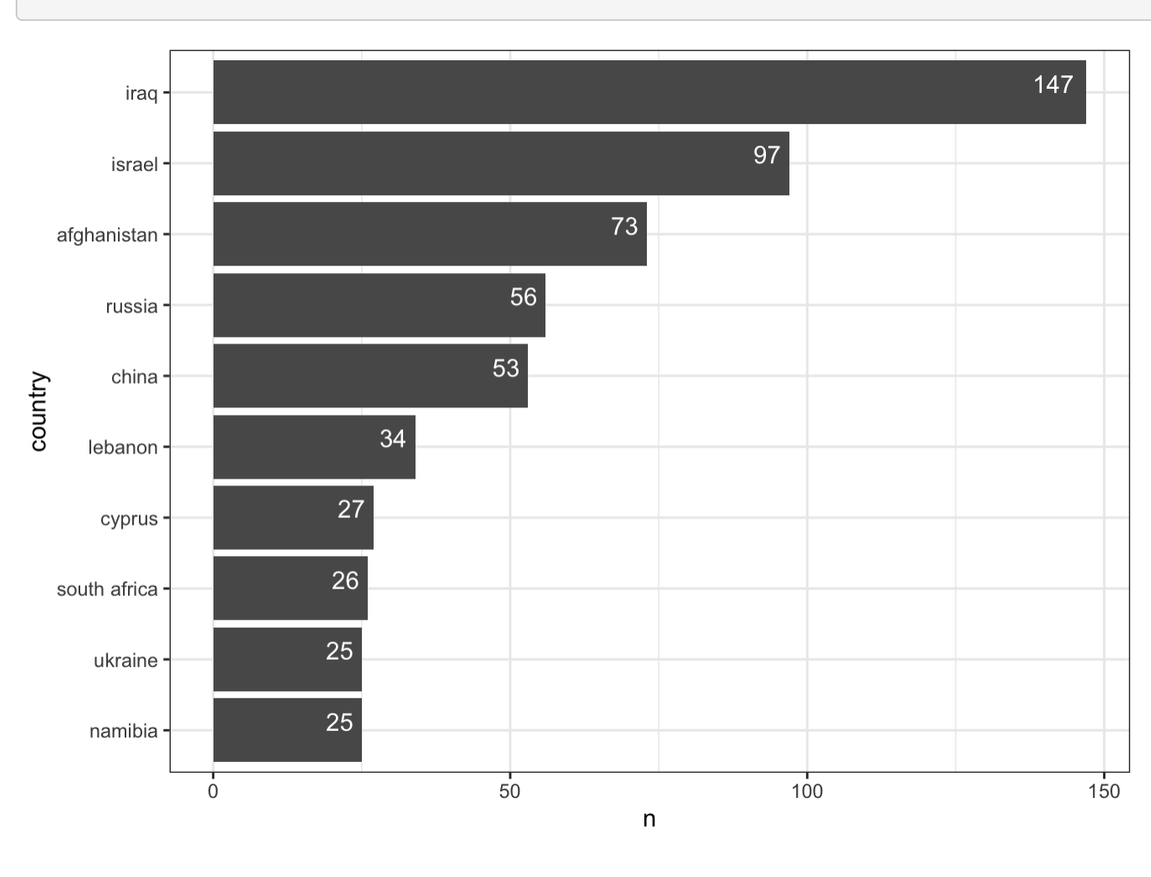
```
data(wrld_simpl)

co <- tolower(wrld_simpl$NAME)

coutries_freq <- freq_table[freq_table$term %in% co,][1:10,]
coutries_freq
```

```
##                     term   n         freq
## iraq                iraq 147 0.0010648624
## israel            israel  97 0.0007026643
## afghanistan  afghanistan  73 0.0005288092
## russia            russia  56 0.0004056619
## china              china  53 0.0003839300
## lebanon          lebanon  34 0.0002462947
## cyprus            cyprus  27 0.0001955870
## south africa south africa  26 0.0001883430
## namibia          namibia  25 0.0001810991
## ukraine          ukraine  25 0.0001810991
```

```
#
ggplot(coutries_freq,aes(x = reorder(term, freq), y = freq)) +
  geom_bar(stat = "identity", show.legend = F) + coord_flip() + xlab("country") + theme_bw() +
  geom_text(aes(label=round(freq,7)), position=position_dodge(width=0.2), vjust=-0, hjust=1.3,colour="white")
```




```
ggplot(coutries_freq,aes(x = reorder(term, n), y = n)) +
  geom_bar(stat = "identity", show.legend = F) + coord_flip() + xlab("country") + theme_bw() +
  geom_text(aes(label=round(n,7)), position=position_dodge(width=0.2), vjust=-0, hjust=1.3,colour="white")
```

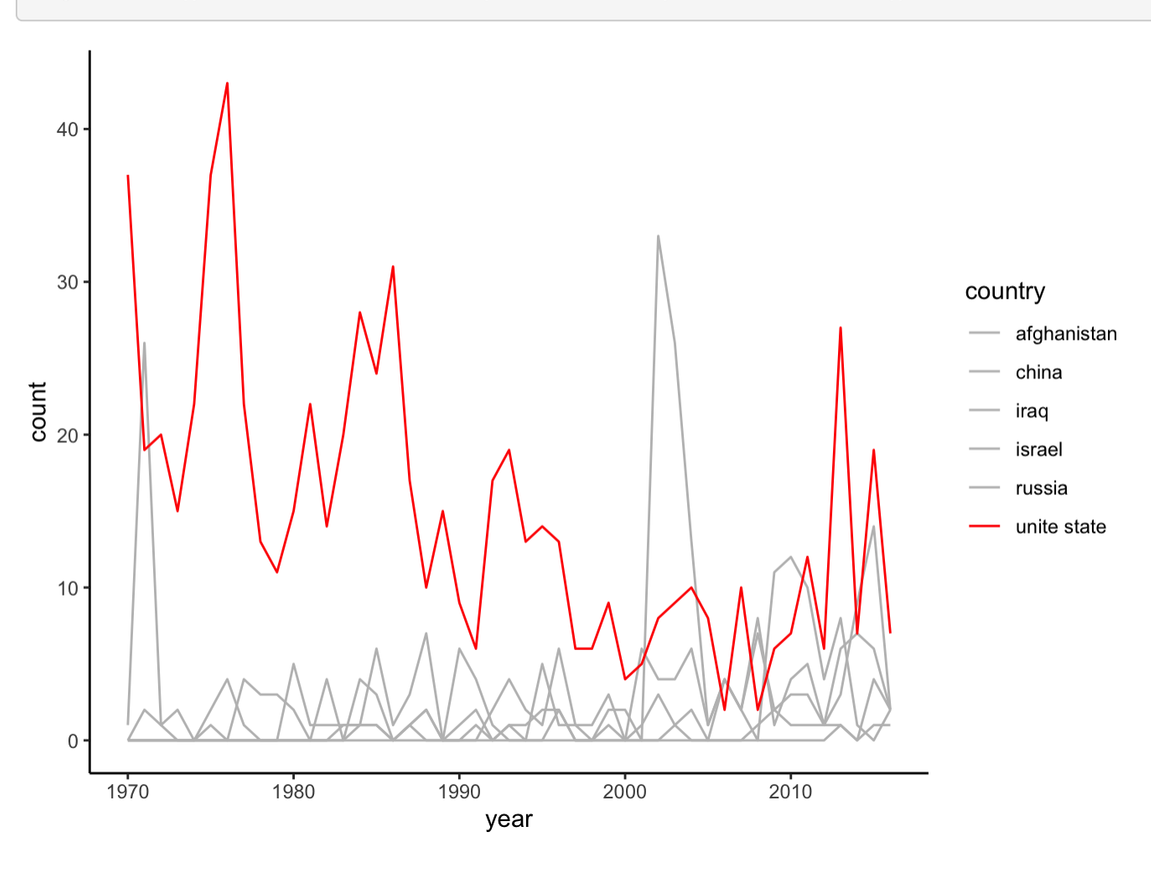


4- Of the 5 countries with the highest term count, compare the counts to the mentions of the United States. Show how these terms counts have changed over time.

```
# adding united states to the comparison
tops <-  c("iraq", "israel", "afghanistan", "russia", "china", "unite state")
term_freq <- dtm[, which(colnames(dtm) %in% sort(tops))]
colnames(term_freq) <- sort(tops)
term_freq <- as.data.frame(term_freq)
term_freq$year <- speech$year

# mellt will convert columns into rows and counting all values of row
df <- melt(term_freq, id.vars = c("year"), variable.name = "country", value.name = "count")
```

```
## Warning in melt(term_freq, id.vars = c("year"), variable.name = "country", :
## The melt generic in data.table has been passed a data.frame and will attempt
## to redirect to the relevant reshape2 method; please note that reshape2 is
## deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(term_freq). In the next
## version, this warning will become an error.
```

```
cols <- c("gray", "gray", "gray", "gray", "gray",  "red")
ggplot(df, aes(x = year, y = count, color= country)) +
  geom_line() + scale_color_manual(values = cols) + theme_classic()
```



5 and 6: Calculate the TF-IDF values for the DTM. In which year's address does the term 'iraq' provide the most semantic contribution? Evaluate the terms with the highest TF-IDF values

```
# compute IDF
number_of_docs <- nrow(dtm)
term_in_docs <- colSums(dtm > 0)
idf <- log(number_of_docs / term_in_docs)

# compute TF/IDF
tf_idf <- t(t(dtm) * idf)
names(tf_idf) <- colnames(dtm)
rownames(tf_idf) <- speech$year

iraq_years <- sort(tf_idf[,"iraq"], decreasing = T)[1:20]
head(iraq_years, 1)
```

```
##     2002
## 17.09212
```

```
iraq_content <- sort(tf_idf["2002",], decreasing = T)[1:20]
as.data.frame(iraq_content)
```

```
##                         iraq_content
## iraqi regime               38.521494
## ninetyone                  34.727005
## hundred ninetyone          34.727005
## nine hundred ninetyone     34.727005
## iraqi                      26.989163
## iraqs                      26.888516
## saddam                     17.925678
## hussein                    17.246973
## saddam hussein             17.246973
## iraq                       17.092122
## kuwait                     14.783119
## ninenine                   14.408717
## inspector                  13.444258
## regime                     12.685586
## six hundred                11.203548
## longrange                  11.006141
## nation inspector           11.006141
## unite nation inspector     11.006141
## renew demand                9.471001
## resolution six              9.471001
```

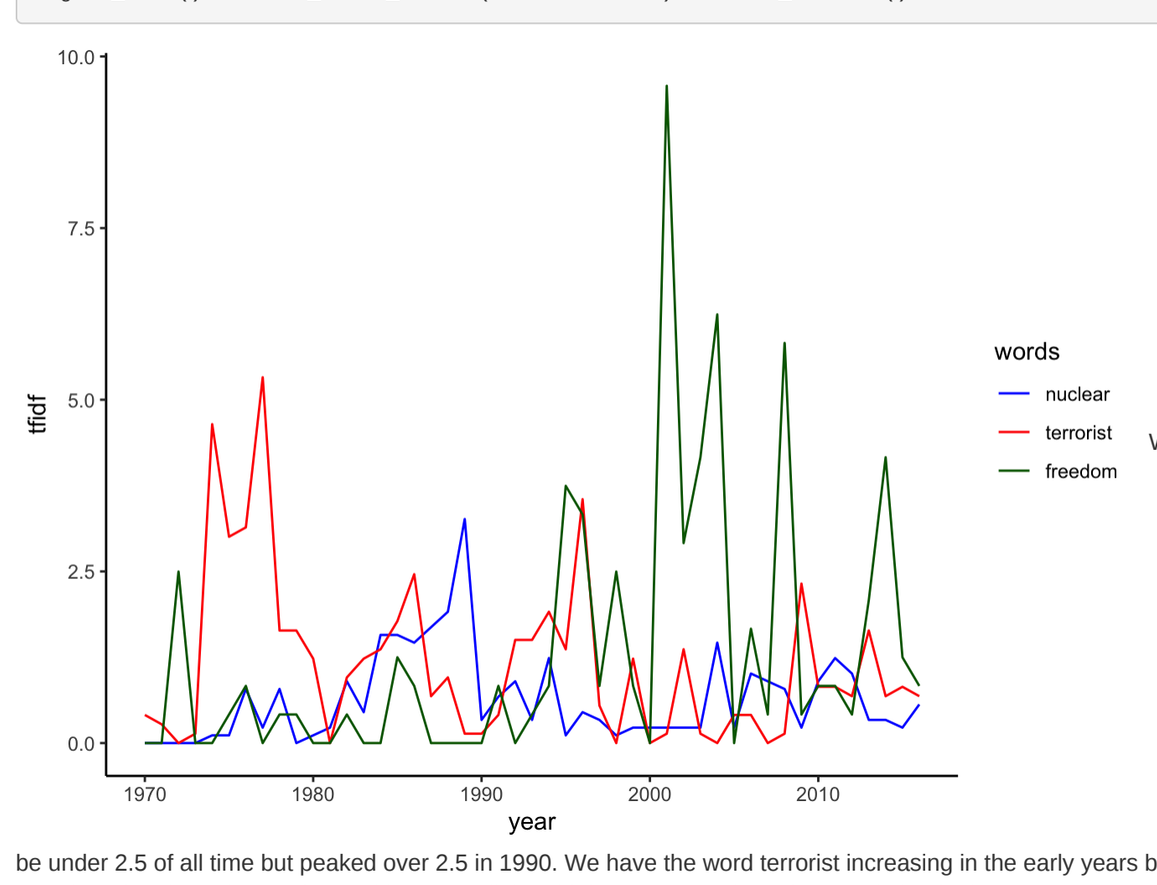
For 6, it seems that the context based on the tf/idf values we can infer that the speech was about nations, and renew demand and discussion of iraq regime

7- Perform a time series analysis on the terms 'nuclear,' 'terrorist,' and 'freedom' using TF-IDF values. How does the use of these terms change over time? Make a visual and describe your results.

```
words <- c("nuclear", "terrorist", "freedom")
tf_idf_words <- tf_idf[, which(colnames(tf_idf) %in% words)]
colnames(tf_idf_words) <- words
tf_idf_words <- data.frame(year = speech$year, as.data.frame(tf_idf_words))
tf_idf_words$year_n <- 1:nrow(tf_idf_words)

# create visual
tf_idf_words <- melt(tf_idf_words, id.vars = c("year_n", "year"), variable.name = "words", value.name = "tfidf")
```

```
## Warning in melt(tf_idf_words, id.vars = c("year_n", "year"), variable.name =
## "words", : The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(tf_idf_words). In the next
## version, this warning will become an error.
```

```
df <- tf_idf_words[order(tf_idf_words$tfidf, decreasing = T),]

cols <- c("blue", "red", "dark green")
ggplot(df, aes(x = year, y = tfidf, color = words)) +
  geom_line() + scale_color_manual(values = cols) + theme_classic()
```



We can see the changes in nuclear to be under 2.5 of all time but peaked over 2.5 in 1990. We have the word terrorist increasing in the early years but decreased after 2010. Also, "freedom" has been increasing and become all time high of compared to other words beginning of 2000s.

Defining 'Topic' in Text Analysis

Latent Dirichlet Allocation (LDA)

Using Topic Modeling for Unsupervised Learning

Tuning Topic Distributions

```
g8 <- unique(speech$country)
g8 <- c("CAN","FRA","GBR", "ITA","JPN","USA", "RUS", "DEU")
```

```
text <- pre_process_corpus(speech, "text", replace_numbers = T,
                           extra_stopwords = g8,
                           root_gen = "lemmatize")
speech$review_preprocessed <- text
```

```
it <- itoken(text, tokenizer = word_tokenizer)
vocab_full <- create_vocabulary(it, ngram = c(1, 3))

# set lower bound
lbound <- 2

vocab <- vocab_full[vocab_full$doc_count > lbound,]

vectorizer <- vocab_vectorizer(vocab)
dtm <- create_dtm(it, vectorizer)
dim(dtm)
```

```
## [1]   47 7600
```

```
library(Matrix)
sparse_corpus <- Matrix(dtm, sparse = T)
```

```
topic_content <- as.data.frame(t(exp(topic_model$beta$logbeta[[1]])))
colSums(topic_content)
```

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12
##   1   1   1   1   1   1   1   1   1   1   1   1
```

```
apply(topic_content, 2, function(x) {topic_model$vocab[order(x, decreasing = T)[1:10]]})
```

```
##       V1             V2        V3             V4             V5
##  [1,] "will"         "will"    "nation"       "nation"       "people"
##  [2,] "one"          "people"  "unite"        "unite"        "will"
##  [3,] "can"          "world"   "unite_nation" "will"         "can"
##  [4,] "world"        "unite"   "world"        "weapon"       "nation"
##  [5,] "us"           "state"   "will"         "world"        "world"
##  [6,] "nuclear"      "america" "state"        "unite_nation" "us"
##  [7,] "good"         "must"    "must"         "one"          "unite"
##  [8,] "peace"        "s"       "s"            "nuclear"      "make"
##  [9,] "unite_state"  "can"     "us"           "state"        "good"
## [10,] "state"        "nation"  "new"          "terrorist"    "country"
##       V6             V7             V8              V9       V10
##  [1,] "unite"        "unite"        "nation"        "will"   "nation"
##  [2,] "soviet"       "state"        "will"          "people" "unite"
##  [3,] "state"        "nation"       "unite"         "nation" "unite_nation"
##  [4,] "nation"       "unite_nation" "state"         "world"  "must"
##  [5,] "unite_state"  "unite_state"  "international" "peace"  "people"
##  [6,] "world"        "will"         "must"          "must"   "world"
##  [7,] "peace"        "world"        "unite_state"   "can"    "state"
##  [8,] "will"         "one"          "world"         "unite"  "will"
##  [9,] "union"        "peace"        "peace"         "state"  "good"
## [10,] "soviet_union" "year"         "can"           "us"     "right"
##       V11             V12
##  [1,] "unite"         "nation"
##  [2,] "one"           "unite"
##  [3,] "state"         "world"
##  [4,] "country"       "people"
##  [5,] "development"   "will"
##  [6,] "develop"       "unite_nation"
##  [7,] "international" "one"
##  [8,] "will"          "peace"
##  [9,] "nation"        "state"
## [10,] "unite_state"   "freedom"
```

```
topic_content <- as.data.frame(t(exp(topic_model$beta$logbeta[[1]])))

apply(topic_content, 2, function(x) {topic_model$vocab[order(x, decreasing = T)[1:10]]})
```

```
##       V1                V2            V3             V4
##  [1,] "development"     "agreement"   "nuclear"      "iraq"
##  [2,] "develop"         "economic"    "community"    "regime"
##  [3,] "economic"        "development" "cooperation"  "hundred"
##  [4,] "peacekeeping"    "issue"       "conference"   "iraqi"
##  [5,] "growth"          "now"         "hundred"      "council"
##  [6,] "develop_country" "negotiation" "organization" "security_council"
##  [7,] "need"            "energy"      "agreement"    "nine"
##  [8,] "trade"           "develop"     "step"         "democracy"
##  [9,] "strengthen"      "resolution"  "towards"      "challenge"
## [10,] "system"          "problem"     "economic"     "one_thousand"
##       V5             V6          V7             V8          V9
##  [1,] "soviet"       "freedom"   "nuclear"      "terrorist" "development"
##  [2,] "freedom"      "now"       "arm"          "terror"    "develop"
##  [3,] "hope"         "terrorism" "soviet"       "s"         "problem"
##  [4,] "soviet_union" "common"    "union"        "iraq"      "four"
##  [5,] "union"        "terror"    "soviet_union" "free"      "hope"
##  [6,] "say"          "believe"   "negotiation"  "freedom"   "need"
##  [7,] "free"         "chemical"  "obligation"   "america"   "american"
##  [8,] "economic"     "free"      "control"      "democracy" "general"
##  [9,] "now"          "s"         "arm_control"  "region"    "session"
## [10,] "nuclear"      "even"      "peaceful"     "stand"     "concern"
##       V10           V11         V12
##  [1,] "america"     "s"         "nuclear"
##  [2,] "s"           "nuclear"   "economic"
##  [3,] "believe"     "child"     "problem"
##  [4,] "conflict"    "now"       "let"
##  [5,] "power"       "democracy" "political"
##  [6,] "live"        "world_s"   "conflict"
##  [7,] "nuclear"     "threat"    "food"
##  [8,] "now"         "century"   "challenge"
##  [9,] "child"       "let"       "now"
## [10,] "palestinian" "treaty"    "history"
```

```
topic_prevalence <- as.data.frame(topic_model$theta)
```

```
top1 <- findThoughts(topic_model, topics = 8, texts = speech$text, n = 1)
substr(top1$docs[[1]], 1, 1000)
```

```
## [1] "Thank you for the honour of addressing the General Assembly. The American people respect the idealism tha
t gave life to this Organization. And we respect the men and women of the United Nations, who stand for peace and
human rights in every part of the world. Welcome to New York City, and welcome to the United States of America. D
uring the past three years, I have addressed the General Assembly in a time of tragedy for my country, and in tim
es of decision for all of us. Now we gather at a time of tremendous opportunity for the United Nations and for al
l peaceful nations. For decades, the circle of liberty, security and development has been expanding in our world.
This progress has brought unity to Europe, self-government to Latin America and Asia and new hope to Africa. Now
we have the historic 8 chance to widen the circle even further, to fight radicalism and terror with justice and d
ignity and to achieve a true peace, founded on human freedom. The United Nations and my country share "
```

```
# name topics for 6 most likely terms from each
topic_names <- apply(topic_content, 2, function(x) {paste(topic_model$vocab[order(x,
                                     decreasing = T)[1:6]], collapse = " ")})
topic_names
```

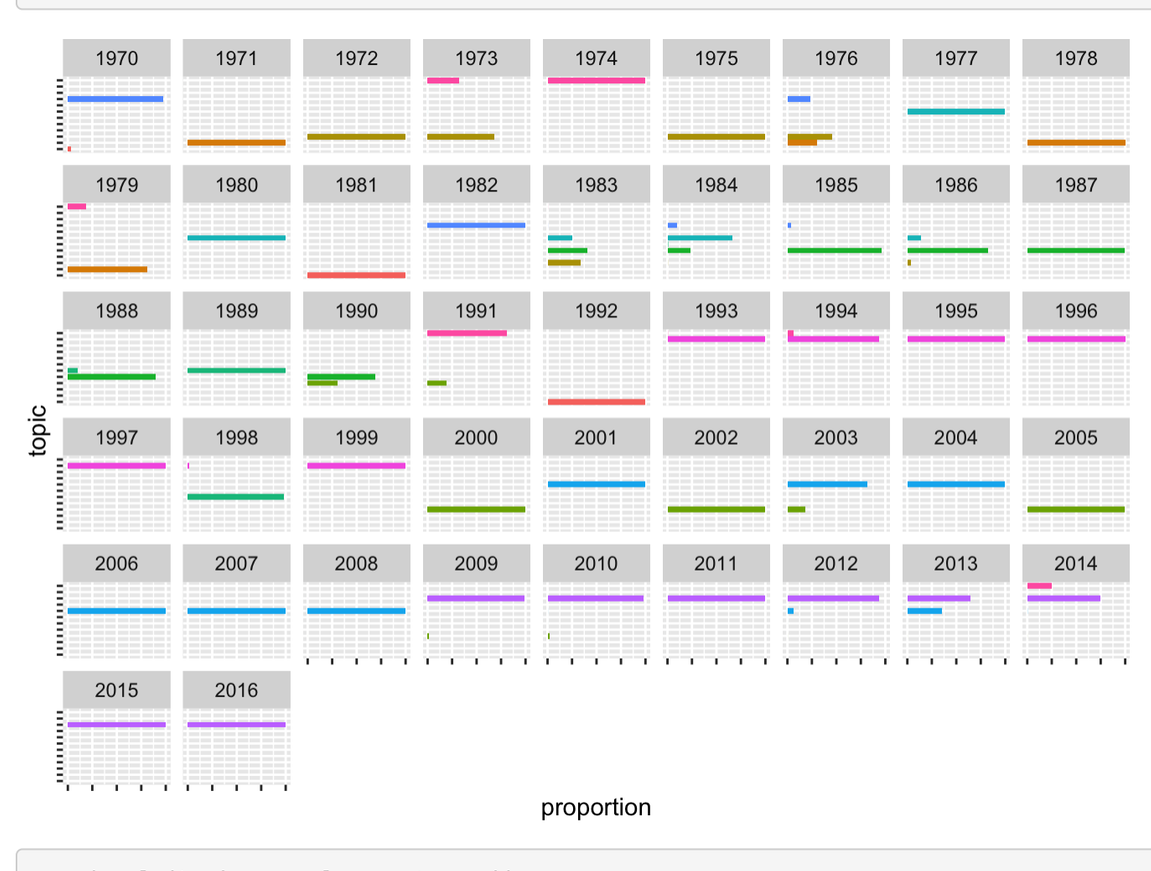
```
##                                                                 V1
## "development develop economic peacekeeping growth develop_country"
##                                                                 V2
##             "agreement economic development issue now negotiation"
##                                                                 V3
##    "nuclear community cooperation conference hundred organization"
##                                                                 V4
##               "iraq regime hundred iraqi council security_council"
##                                                                 V5
##                      "soviet freedom hope soviet_union union say"
##                                                                 V6
##                     "freedom now terrorism common terror believe"
##                                                                 V7
##                "nuclear arm soviet union soviet_union negotiation"
##                                                                 V8
##                           "terrorist terror s iraq free freedom"
##                                                                 V9
##                      "development develop problem four hope need"
##                                                                V10
##                          "america s believe conflict power live"
##                                                                V11
##                          "s nuclear child now democracy world_s"
##                                                                V12
##                 "nuclear economic problem let political conflict"
```

Analyze how the G8 countries align on topics and how that alignment has changed over time.

```
df <- topic_prevalence
colnames(df) <- topic_names
df$year <- as.character(speech$year)
df <- melt(df, id.vars = 'year', value.name = 'proportion', variable.name = 'topic')
```

```
## Warning in melt(df, id.vars = "year", value.name = "proportion", variable.name
## = "topic"): The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(df). In the next version, this
## warning will become an error.
```

```
ggplot(df, aes(x = topic, y = proportion, fill = topic)) + geom_bar(stat = 'identity') +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank(), legend.position = "none") +
  coord_flip() + facet_wrap(~ year, ncol = 9)
```




```
mean(apply(topic_prevalence, 1, max))
```

```
## [1] 0.9186626
```

```
topic_prevalence <- as.data.frame(topic_model2$theta)
mean(apply(topic_prevalence, 1, max))
```

```
## [1] 0.8774324
```

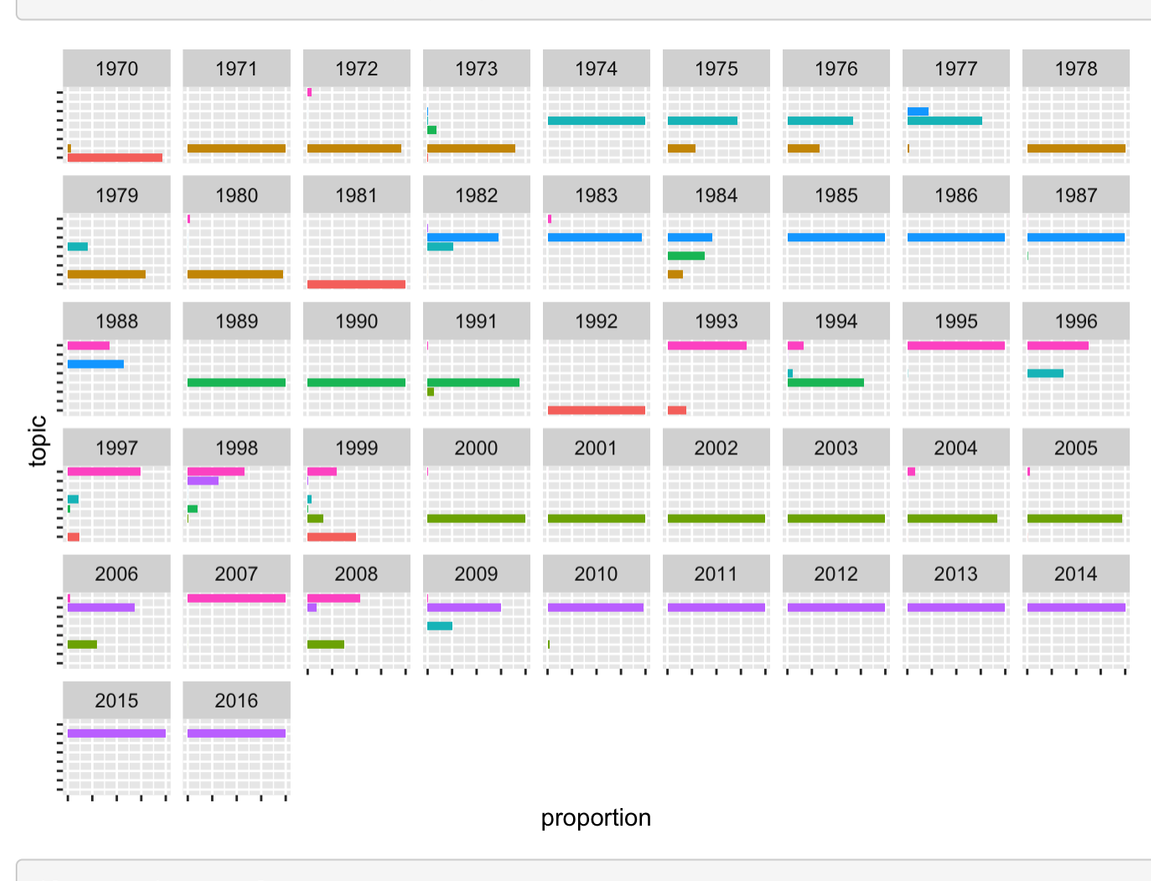
```
topic_content <- as.data.frame(t(exp(topic_model2$beta$logbeta[[1]])))
topic_names <- apply(topic_content, 2, function(x) {paste(topic_model2$vocab[order(x,
                                     decreasing = T)[1:6]], collapse = " ")})
topic_names
```

```
##                                                            V1
##    "development develop need economic growth peacekeeping"
##                                                            V2
## "agreement economic negotiation development issue problem"
##                                                            V3
##            "iraq terrorist regime terror council hundred"
##                                                            V4
##                "freedom now free democracy century hope"
##                                                            V5
##           "nuclear interest common economic conflict now"
##                                                            V6
##               "soviet union soviet_union freedom arm hope"
##                                                            V7
##                 "s america believe live palestinian power"
##                                                            V8
##                      "s world_s child now threat nuclear"
```

```
df <- topic_prevalence
colnames(df) <- topic_names
df$year <- as.character(speech$year)
df <- melt(df, id.vars = 'year', value.name = 'proportion', variable.name = 'topic')
```

```
## Warning in melt(df, id.vars = "year", value.name = "proportion", variable.name
## = "topic"): The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(df). In the next version, this
## warning will become an error.
```

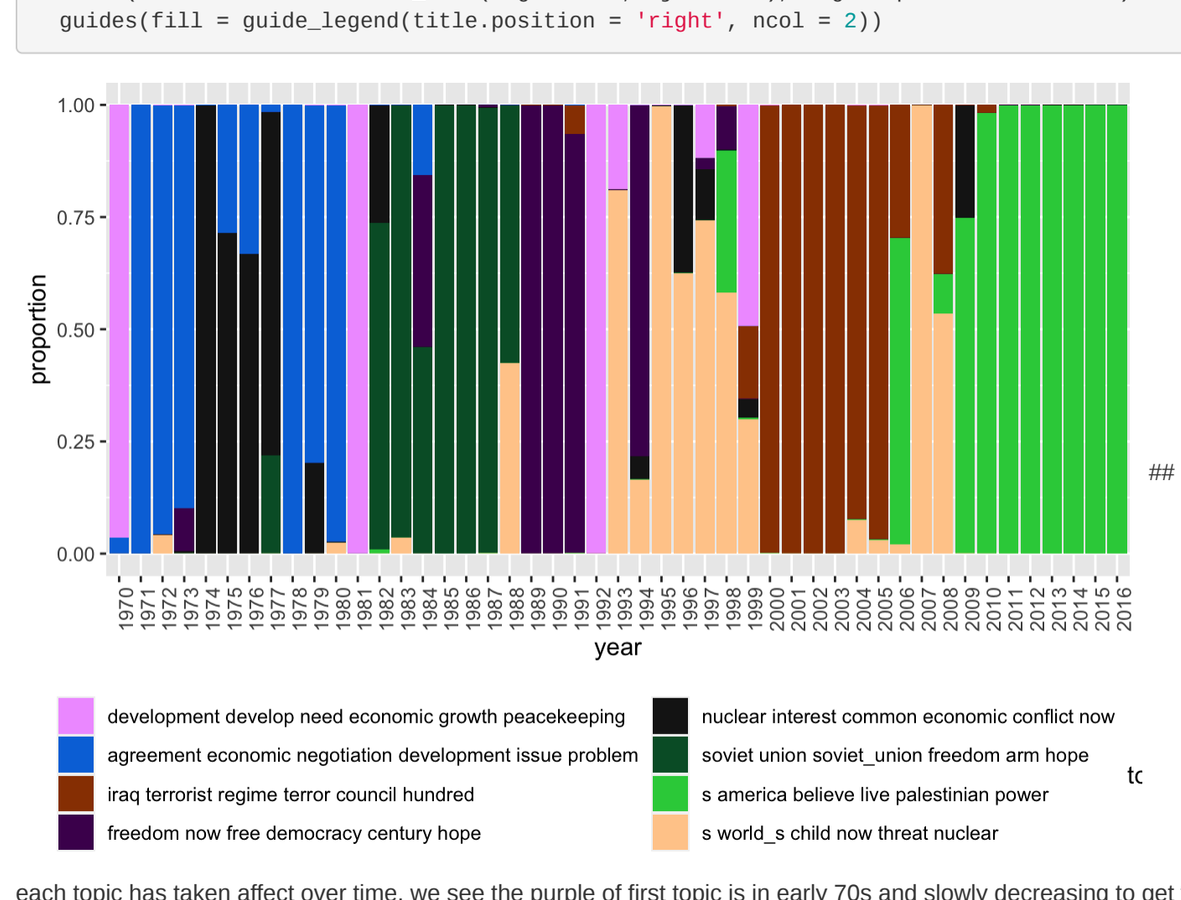
```
ggplot(df, aes(x = topic, y = proportion, fill = topic)) + geom_bar(stat = 'identity') +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank(), legend.position = "none") +
  coord_flip() + facet_wrap(~ year, ncol = 9)
```




```
df <- topic_prevalence
colnames(df) <- topic_names
df$year <- as.character(speech$year)
df <- melt(df, id.vars = 'year', value.name = 'proportion', variable.name = 'topic')
```

```
## Warning in melt(df, id.vars = "year", value.name = "proportion", variable.name
## = "topic"): The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(df). In the next version, this
## warning will become an error.
```

```
library(pals)
ggplot(df, aes(x = year, y = proportion, fill = topic)) + geom_bar(stat = 'identity') +
  scale_fill_manual(values = paste0(alphabet(20), "FF"), name = "topic") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position="bottom") +
  guides(fill = guide_legend(title.position = 'right', ncol = 2))
```



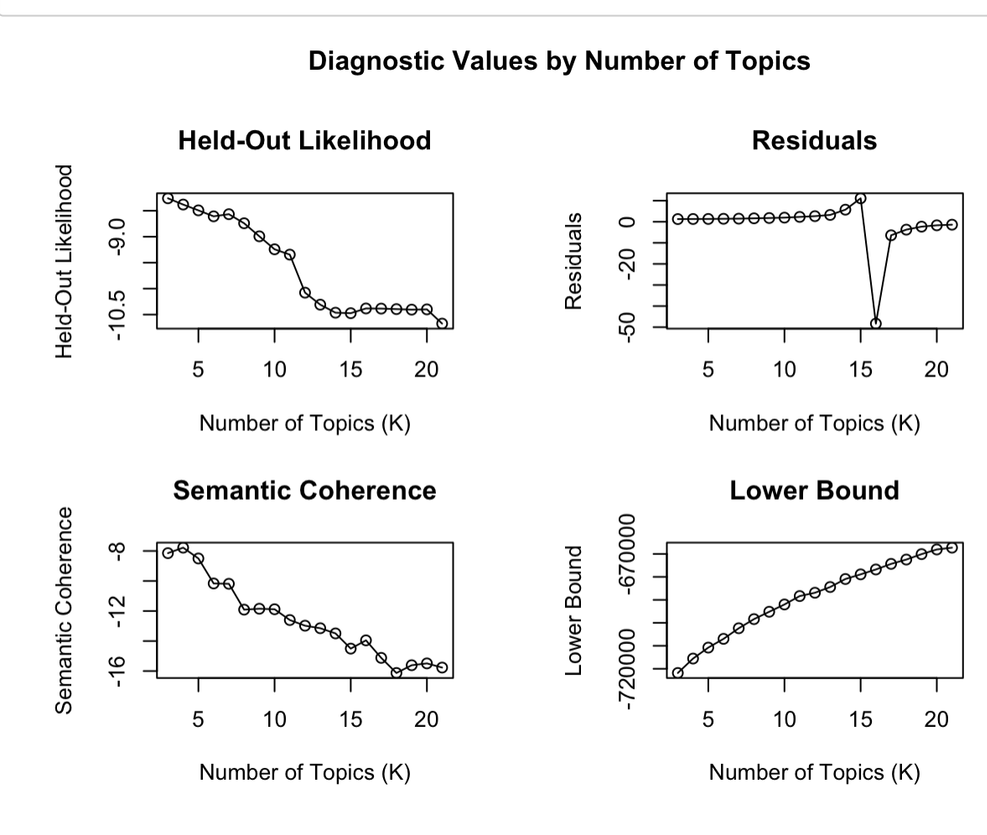
This is a comprehensive visual of

to

each topic has taken affect over time. we see the purple of first topic is in early 70s and slowly decreasing to get to grean which is topics about prenventions and diplomacy

```
ksearch
```

```
## $results
##     K   exclus    semcoh   heldout  residual     bound    lbound em.its
## 1   3 8.020834 -8.144812 -8.259909  1.276254 -721838.2 -721836.4     31
## 2   4 8.228381 -7.778511 -8.380555  1.314681 -715556.2   -715553     34
## 3   5 8.320627  -8.49905 -8.492595  1.340772 -710780.5 -710775.7     29
## 4   6 8.318788 -10.16403 -8.607401  1.412801 -706983.9 -706977.4     22
## 5   7 8.401406   -10.187 -8.566305  1.477005 -702366.5   -702358     28
## 6   8 8.493556 -11.90788 -8.740941  1.633326 -698391.1 -698380.5     27
## 7   9  8.53164 -11.84834 -8.992541  1.780874   -695191 -695178.2     25
## 8  10 8.586126 -11.88252 -9.242476  1.965839 -692014.4 -691999.3     36
## 9  11 8.695476 -12.60053 -9.345184  2.306203 -688357.2 -688339.7     20
## 10 12 8.767495 -12.97785 -10.07767  2.661024 -686922.1 -686902.1     19
## 11 13 8.812171 -13.14665 -10.30923   3.24492 -684430.9 -684408.3     33
## 12 14 8.877194 -13.49236 -10.46399  5.738061 -680940.9 -680915.7     50
## 13 15 8.930664 -14.50541 -10.47326  11.14244 -678917.6 -678889.7     42
## 14 16 8.960784 -13.95488 -10.38226 -48.33192 -676790.1 -676759.4     42
## 15 17 8.992802 -15.12183 -10.38492 -6.416407 -674361.6 -674328.1     28
## 16 18 9.037252 -16.12895 -10.39485  -3.73503 -672477.8 -672441.4     27
## 17 19 9.059132 -15.61238 -10.40483 -2.313853 -670120.4   -670081     28
## 18 20 9.083703 -15.48962 -10.39974 -1.683086 -668093.7 -668051.4     34
## 19 21 9.110316 -15.76781 -10.67427 -1.386643 -667309.2 -667263.8     32
##
## $call
## searchK(documents = docs, vocab = colnames(dtm), K = c(3:21))
##
## attr(,"class")
## [1] "searchK"
```



from search k we can create range of numbers that we desire for k and the function will estimate different outcomes of K. we can notice that semantice cogerence are performing better at 15 number of topics but we see a slightly increase in residuals since we might increase the number of k.