United Nations General Debates Text Analysis: Isreal vs. Palestine

Load required packages

```
#loading the packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tokenizers)
```

```
## Warning: package 'tokenizers' was built under R version 4.0.5
```

```
library(quanteda)
```

```
## Warning: package 'quanteda' was built under R version 4.0.5
```

```
## Package version: 3.0.0
## Unicode version: 10.0
## ICU version: 61.1
```

```
## Parallel computing: 8 of 8 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(quanteda.textplots)
```

```
## Warning: package 'quanteda.textplots' was built under R version 4.0.5
```

```
#install.packages("stm")
library(stm)
```

```
## Warning: package 'stm' was built under R version 4.0.5
```

```
## stm v1.3.6 successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com
```

```
#install.packages("seededlda")
library(seededlda)
```

```
## Warning: package 'seededlda' was built under R version 4.0.5
```

```
##
## Attaching package: 'seededlda'
```

```
## The following object is masked from 'package:stats':
##
##     terms
```

Load the dataset

Load the United Nations General Debates dataset, take a peak of its top 5 rows.

```
metadata <- read_csv("UNGDspeeches.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   doc_id = col_character(),
##   text = col_character(),
##   country = col_character(),
##   session = col_double(),
##   year = col_double()
## )
```

```
head(metadata)
```

```
## # A tibble: 6 x 5
##   doc_id      text                                          country session  year
##   <chr>       <chr>                                         <chr>     <dbl> <dbl>
## 1 ALB_25_197~ "33: May I first convey to our President th~ ALB          25  1970
## 2 ARG_25_197~ "177.\t : It is a fortunate coincidence tha~ ARG          25  1970
## 3 AUS_25_197~ "100.\t  It is a pleasure for me to extend ~ AUS          25  1970
## 4 AUT_25_197~ "155.\t  May I begin by expressing to Ambas~ AUT          25  1970
## 5 BEL_25_197~ "176. No doubt each of us, before coming up~ BEL          25  1970
## 6 BLR_25_197~ "\n71.\t. We are today mourning the untimel~ BLR          25  1970
```

Data exploration is done in another notebook (python) on distribution of speeches/document over years, and speeches by Israel and Palestine.

However, we will seperate Israel by the date 1998, because it is the date when Palestine first joined the United Nations.

```
for (i in 1:nrow(metadata)){
  #if the country is israel
  if (metadata[i, ]$country == 'ISR'){
    if (metadata[i, ]$year < 1998) {
      metadata[i, 'country'] = 'ISR_prev_1998'
    } else {
      metadata[i, 'country'] = 'ISR_post_1998'
    }
  }
}
```

Create document frequency matrix

```
#use quanteda to turn the data into a corpus
corpus_un <- corpus(metadata, text_field = "text")
toks_un <- tokens(corpus_un)
dfm_un <- dfm(toks_un)
dfm_un
```

```
## Document-feature matrix of: 8,093 documents, 76,792 features (98.80% sparse) and 3 docvars.
##                  features
## docs              33 : may  i first convey  to our president the
##   ALB_25_1970.txt  1 6   5  1     4      1 240   9         1 872
##   ARG_25_1970.txt  0 5   5  6     7      0 165  24         3 443
##   AUS_25_1970.txt  0 1   7 23     7      0 169  22         2 444
##   AUT_25_1970.txt  0 2  10 22     4      0 165  23         5 412
##   BEL_25_1970.txt  1 7   3 13     7      2 131  31         4 345
##   BLR_25_1970.txt  0 3   2  2     3      0 140  10         5 710
## [ reached max_ndoc ... 8,087 more documents, reached max_nfeat ... 76,782 more features ]
```

Corpus `corpus_un` consisting of 8,093 documents and 3 docvars;

Tokens `toks_un` consisting of 8,093 documents and 3 docvars.

dfm_un is a Document-feature matrix of: 8,093 documents, 51,006 features (98.65% sparse) and 3 docvars.

Words such as "I", "to" should not be included: we need to retokenize the corpus to have punctuation, numbers, stemwords and stopwords removed:

```
toks_un <- tokens(corpus_un, remove_punct = TRUE, remove_numbers=TRUE)
toks_un <- tokens_wordstem(toks_un)
toks_un <- tokens_select(toks_un,  stopwords("en"), selection = "remove")
dfm_un <- dfm(toks_un)
dfm_un
```

```
## Document-feature matrix of: 8,093 documents, 51,006 features (98.65% sparse) and 3 docvars.
##                  features
```

```
## docs              may first convey presid congratul albanian deleg elect
##   ALB_25_1970.txt   5     4      1      3         1        9     3     1
##   ARG_25_1970.txt   5     7      0      4         1        0     2     1
##   AUS_25_1970.txt   7     7      0      4         2        0     6     2
##   AUT_25_1970.txt  10     4      2      8         0        0     2     2
##   BEL_25_1970.txt   3     7      2      5         1        0     2     0
##   BLR_25_1970.txt   2     3      0      5         1        0     5     1
##                  features
## docs              twenty-fifth session
##   ALB_25_1970.txt            3       5
##   ARG_25_1970.txt            1       6
##   AUS_25_1970.txt            4       7
##   AUT_25_1970.txt            4       7
##   BEL_25_1970.txt            0       1
##   BLR_25_1970.txt            8       5
## [ reached max_ndoc ... 8,087 more documents, reached max_nfeat ... 50,996 more features ]
```

51006 features are too many for the analysis: reduce the number to 5% of it. Calling method dfm_trim from the quanteda package, and obtain a new document frequency matrix with 2471 features.

```
dfm_trimmed <- dfm_trim(dfm_un, min_docfreq = 0.05, docfreq_type = "prop")
dfm_trimmed
```

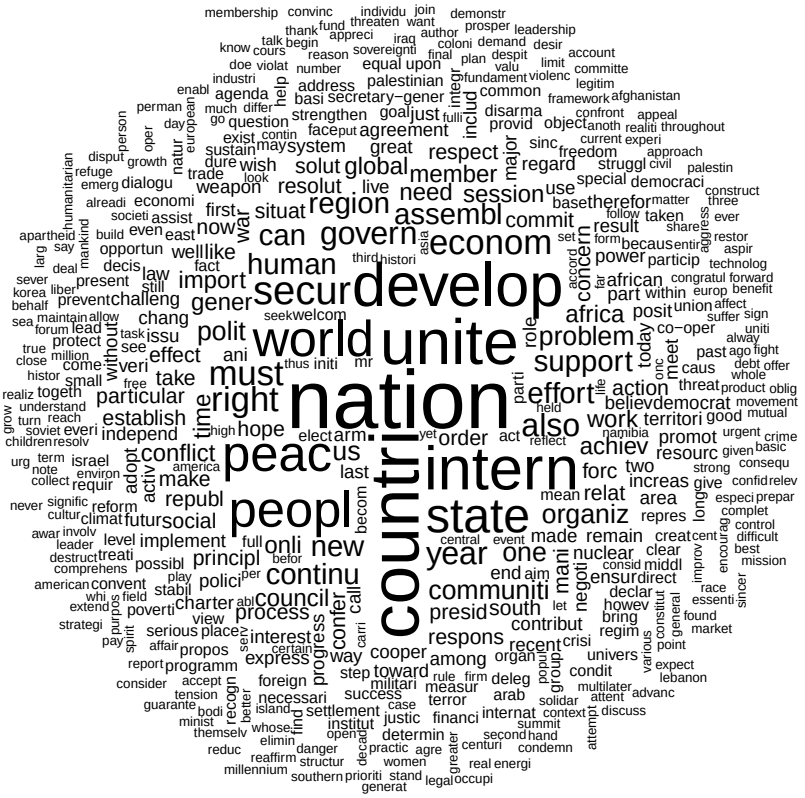
```
## Document-feature matrix of: 8,093 documents, 2,471 features (76.26% sparse) and 3 docvars.
##                  features
## docs              may first convey presid congratul deleg elect session gener
##   ALB_25_1970.txt   5     4      1      3         1     3     1       5     5
##   ARG_25_1970.txt   5     7      0      4         1     2     1       6    12
##   AUS_25_1970.txt   7     7      0      4         2     6     2       7     5
##   AUT_25_1970.txt  10     4      2      8         0     2     2       7    13
##   BEL_25_1970.txt   3     7      2      5         1     2     0       1     6
##   BLR_25_1970.txt   2     3      0      5         1     5     1       5     5
##                  features
## docs              assembl
##   ALB_25_1970.txt       7
##   ARG_25_1970.txt      14
##   AUS_25_1970.txt      12
##   AUT_25_1970.txt      14
##   BEL_25_1970.txt       6
##   BLR_25_1970.txt       6
## [ reached max_ndoc ... 8,087 more documents, reached max_nfeat ... 2,461 more features ]
```

```
#2,471 features
```

Most frequent word: visualization

Generate a word cloud of all features that we selected, based on their word frequency.

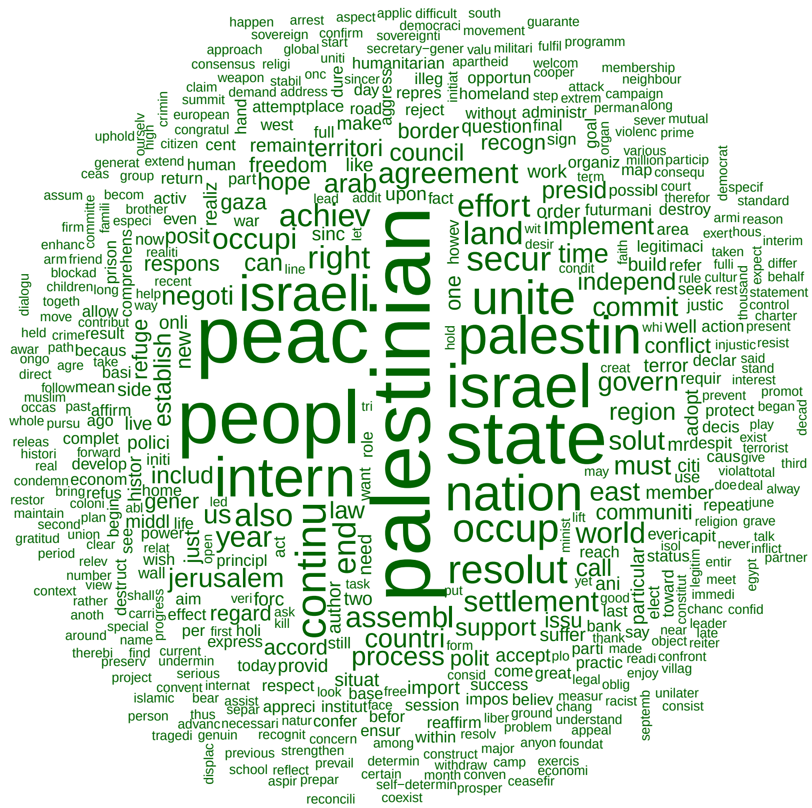
```
#all word based on their word frequency.
textplot_wordcloud(dfm_trimmed, col="black")
```




```
#Subset dfm and metadata to speech made by the Israel before 1998, after 1998, and Palestine.

dfm_trimmed <- dfm_trimmed[metadata$country%in%c("PSE", "ISR_prev_1998", "ISR_post_1998"),]
metadata <- metadata[metadata$country%in%c("PSE", "ISR_prev_1998", "ISR_post_1998"),]
```

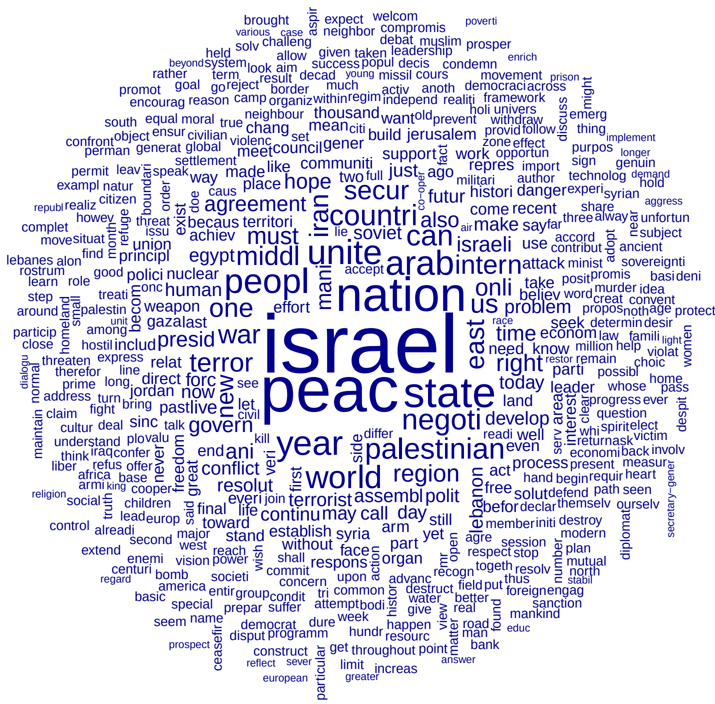
Word Cloud of only Palestine.

```
textplot_wordcloud(dfm_trimmed[metadata$country == "PSE",], col="darkgreen")
```



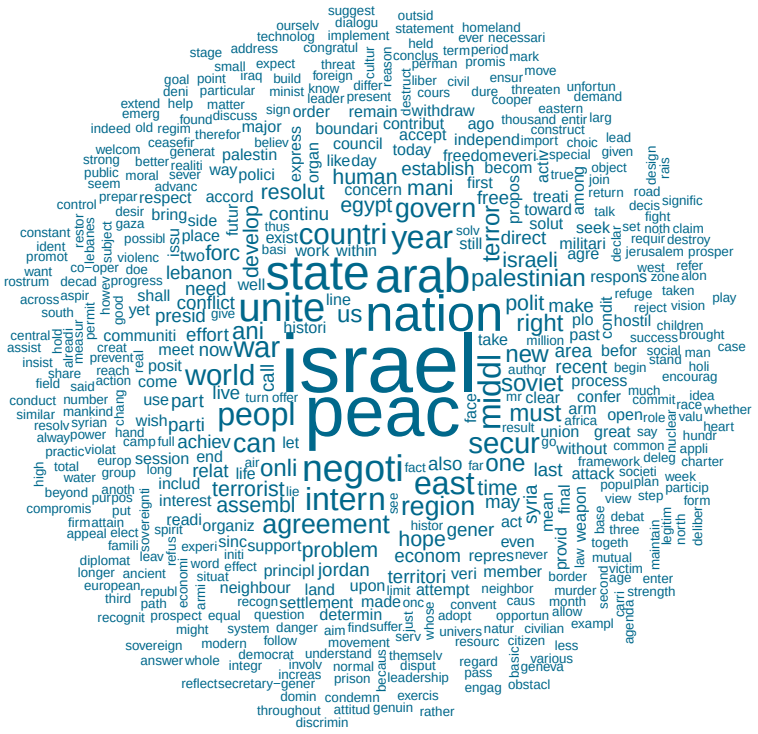
Word Cloud of only Israel

```r
textplot_wordcloud(dfm_trimmed[metadata$country%in%c("ISR_prev_1998", "ISR_post_1998"),])
```



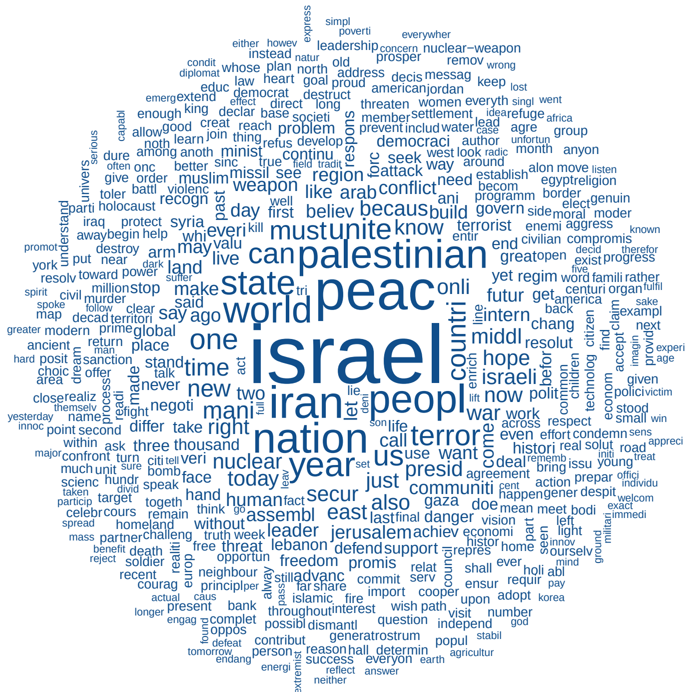
Word Cloud of only Israel before 1998, when Palestine was not in the United nations.

```r
textplot_wordcloud(dfm_trimmed[metadata$country=="ISR_prev_1998",],
                   col = 'deepskyblue4')
```



Palestine joined the UN in 1998, print Word Cloud of only Israel after 1998:

```
textplot_wordcloud(dfm_trimmed[metadata$country=="ISR_post_1998",],
                   col = 'dodgerblue4')
```



By directly observing the word cloud, we can see that the word "Iran" was more frequent after 1998.

Find distinctive words

```r
#DSC161 codes: Fightin' words
clusterFightinWords <- function(dfm, clust.vect, alpha.0=100) {
  # we need to get the overall corpus word distribution and the cluster-specific words dists
  # y_{kw} in Monroe et al.
  overall.terms <- colSums(dfm)
  # n and n_k in Monroe et al.
  n <- sum(overall.terms)
  # alpha_{kw} in Monroe et al.
  prior.terms <- overall.terms / n * alpha.0
  # y_{kw}(i) in Monroe et al.
  cluster.terms <- colSums(dfm[clust.vect, ])
  # n_k(i) in Monroe et al.
  cluster.n <- sum(cluster.terms)

  cluster.term.odds <-
    (cluster.terms + prior.terms) /
    (cluster.n + alpha.0 - cluster.terms - prior.terms)
  overall.term.odds <-
    (overall.terms + prior.terms) /
    (n + alpha.0 - overall.terms - prior.terms)
```

```
  log.odds <- log(cluster.term.odds) - log(overall.term.odds)

  variance <- 1/(cluster.terms + prior.terms) + 1/(overall.terms + prior.terms)

  # return the variance weighted log-odds for each term
  output <- log.odds / sqrt(variance)
  names(output) <- colnames(dfm)
  return(output)
}
```

```
#Find words that are distinctive of Israel before 1998, after 1998, and Palestine

#terms <- clusterFightinWords(dfm_trimmed, metadata$country=="ISR")
#sort(terms, decreasing=T)[1:10]

terms <- clusterFightinWords(dfm_trimmed,
                             metadata$country=="ISR_prev_1998")
sort(terms, decreasing=T)[1:10]
```

```
##     arab   soviet   negoti    egypt   propos    middl boundari neighbor
## 7.083559 7.065167 5.255596 5.130174 4.453895 4.413279 4.116798 3.849411
##   jordan      war
## 3.695982 3.668016
```

The 10 most distinctive words for Israel's speech in before 1998 is:

- arab
- soviet
- negoti(ate)
- egypt
- propos(e)
- middl(e)
- boundari(y)
- neighbor
- jordan
- war

```
terms <- clusterFightinWords(dfm_trimmed,
                             metadata$country=="ISR_post_1998")
sort(terms, decreasing=T)[1:10]
```

```
##      iran   nuclear      know    terror    israel    becaus       get    global
## 16.802166  7.845091  6.808745  6.106231  6.091778  5.819943  5.678615  5.509139
##     world      want
##  5.421542  5.292286
```

The 10 most distinctive words for Israel's speech in after 1998 is:

- iran
- nuclear
- know

- terror
- israel
- becaus(e)
- get
- global
- world
- want

```
#Find words that are distinctive of PSE

terms <- clusterFightinWords(dfm_trimmed,
                             metadata$country=="PSE")
sort(terms, decreasing=T)[1:10]
```

```
##    palestin       occup palestinian      occupi     israeli       peopl
##   13.133044   12.124380   11.669960    9.434397    8.754563    8.362793
##     continu      intern   implement     resolut
##    8.277747    7.771255    7.351733    7.095903
```

The 10 most distinctive words for Israel's speech in after 1998 is:

- palestin
- occup(y)
- palestinian
- occupi(y)
- israeli
- peopl(e)
- continu(e)
- intern
- implement
- resolut(ion)

```
dfm_trimmed
```

```
## Document-feature matrix of: 70 documents, 2,471 features (76.15% sparse) and 3 docvars.
##                  features
## docs              may first convey presid congratul deleg elect session gener
##   ISR_25_1970.txt   6     5      0      4         0     3     0       3     9
##   ISR_26_1971.txt  11     8      0      6         0     0     0       2     3
##   ISR_27_1972.txt   4     2      0      8         1     3     1       3     9
##   ISR_28_1973.txt   4     2      1      6         0     1     0       1     4
##   ISR_29_1974.txt   2     7      0      2         1     1     0       2     3
##   ISR_30_1975.txt   5     1      0      3         1     0     1       5    10
##                  features
## docs              assembl
##   ISR_25_1970.txt       8
##   ISR_26_1971.txt       4
##   ISR_27_1972.txt       8
##   ISR_28_1973.txt       4
##   ISR_29_1974.txt       3
##   ISR_30_1975.txt      13
## [ reached max_ndoc ... 64 more documents, reached max_nfeat ... 2,461 more features ]
```

Topic Modelling

LDA:

```
#LDA
######
#Run LDA using quanteda
lda <- textmodel_lda(dfm_trimmed, k = 10)

#Most likely term for each topic
lda.terms <- terms(lda, 10)
lda.terms
```

```
##       topic1      topic2    topic3      topic4        topic5      topic6
##  [1,] "negoti"    "new"     "peac"      "terror"      "arab"      "israel"
##  [2,] "agreement" "peopl"   "nation"    "world"       "soviet"    "nation"
##  [3,] "secur"     "can"     "also"      "human"       "war"       "countri"
##  [4,] "govern"    "east"    "process"   "war"         "intern"    "lebanon"
##  [5,] "israeli"   "may"     "unite"     "nation"      "middl"     "govern"
##  [6,] "propos"    "middl"   "presid"    "terrorist"   "terrorist" "continu"
##  [7,] "side"      "nation"  "jerusalem" "must"        "problem"   "develop"
##  [8,] "parti"     "econom"  "secur"     "right"       "human"     "peopl"
##  [9,] "boundari"  "economi" "agreement" "palestinian" "area"      "world"
## [10,] "line"      "land"    "posit"     "also"        "east"      "ani"
##       topic7    topic8        topic9        topic10
##  [1,] "israel"  "palestinian" "peac"        "iran"
##  [2,] "state"   "peopl"       "region"      "israel"
##  [3,] "unite"   "palestin"    "palestinian" "peopl"
##  [4,] "peac"    "intern"      "negoti"      "nuclear"
##  [5,] "nation"  "state"       "must"        "presid"
##  [6,] "year"    "occup"       "peopl"       "year"
##  [7,] "arab"    "israeli"     "year"        "palestinian"
##  [8,] "one"     "continu"     "us"          "want"
##  [9,] "can"     "land"        "time"        "countri"
## [10,] "assembl" "settlement"  "futur"       "weapon"
```

```
#Topical content matrix
mu <- lda$phi
dim(mu) #10 topics, 5923 words
```

```
## [1]   10 2471
```

```
mu[1:10,1:20]
```

```
##                  may        first       convey       presid    congratul
## topic1  1.560769e-05 4.853990e-03 1.560769e-05 1.576376e-03 1.560769e-05
## topic2  1.278896e-02 2.703030e-03 1.344791e-05 2.434072e-03 1.223760e-03
## topic3  8.628053e-06 8.628053e-06 8.628053e-06 1.510772e-02 3.804971e-03
## topic4  5.141631e-03 1.484460e-04 1.349509e-05 1.349509e-05 2.833970e-04
## topic5  1.162102e-05 1.162102e-05 1.278312e-04 1.278312e-04 1.278312e-04
## topic6  1.526927e-05 3.527202e-03 1.526927e-05 1.526927e-05 3.206547e-04
```

```
## topic7  3.174335e-03 1.906714e-03 5.281755e-06 5.281755e-06 5.281755e-06
## topic8  9.249753e-06 3.792399e-04 9.342250e-04 9.249753e-06 1.942448e-04
## topic9  7.018480e-06 7.018480e-06 7.018480e-06 7.018480e-06 7.018480e-06
## topic10 1.386559e-05 6.392036e-03 1.386559e-05 1.457273e-02 1.386559e-05
##               deleg        elect      session        gener      assembl
## topic1  1.560769e-05 1.560769e-05 1.560769e-05 1.560769e-05 1.560769e-05
## topic2  1.358239e-03 1.627197e-03 1.344791e-05 5.513643e-04 1.344791e-05
## topic3  8.628053e-06 5.099179e-03 5.012899e-03 8.377840e-03 8.291559e-03
## topic4  1.349509e-05 1.767857e-03 1.349509e-05 1.349509e-05 2.172710e-03
## topic5  2.800665e-03 1.162102e-05 1.987194e-03 3.381715e-03 1.162102e-05
## topic6  1.526927e-05 1.526927e-05 5.512208e-03 1.526927e-05 1.526927e-05
## topic7  5.281755e-06 5.281755e-06 5.281755e-06 4.177868e-03 1.004062e-02
## topic8  1.017473e-04 9.249753e-06 7.492300e-04 9.249753e-06 2.876673e-03
## topic9  7.018480e-06 7.018480e-06 7.018480e-06 7.018480e-06 7.018480e-06
## topic10 5.684891e-04 1.386559e-05 1.386559e-05 1.386559e-05 1.386559e-05
##                take         work       agenda        fifth         held
## topic1  1.560769e-05 4.073606e-03 1.560769e-05 1.560769e-05 1.716845e-04
## topic2  1.344791e-05 1.344791e-05 6.858434e-04 6.858434e-04 1.344791e-05
## topic3  8.628053e-06 7.256193e-03 2.674697e-04 2.674697e-04 8.628053e-06
## topic4  6.221239e-03 1.349509e-05 1.349509e-05 2.833970e-04 1.349509e-05
## topic5  1.162102e-05 1.162102e-05 1.173723e-03 1.162102e-05 1.162102e-05
## topic6  1.526927e-05 1.526927e-05 1.526927e-05 1.526927e-05 1.526927e-05
## topic7  5.075767e-03 5.281755e-06 5.281755e-06 5.281755e-06 2.487707e-03
## topic8  9.249753e-06 9.249753e-06 9.249753e-06 9.249753e-06 9.249753e-06
## topic9  1.410714e-03 1.129975e-03 1.761638e-03 7.018480e-06 3.579425e-04
## topic10 1.525215e-04 5.144133e-03 1.386559e-05 1.386559e-05 1.386559e-05
##          anniversari         come         forc      charter        unite
## topic1  1.560769e-05 3.293222e-03 4.229683e-03 1.560769e-05 1.560769e-05
## topic2  1.344791e-05 1.089281e-03 1.344791e-05 1.344791e-05 1.344791e-05
## topic3  3.537502e-04 8.628053e-06 8.628053e-06 1.906800e-03 1.709217e-02
## topic4  1.349509e-05 2.037759e-03 1.484460e-04 1.349509e-05 1.349509e-05
## topic5  1.162102e-05 1.162102e-05 4.427607e-03 2.219614e-03 1.162102e-05
## topic6  1.236811e-03 1.526927e-05 1.679620e-04 1.526927e-05 6.428364e-03
## topic7  5.281755e-06 5.281755e-06 8.139185e-03 5.281755e-06 2.963593e-02
## topic8  2.867423e-04 9.249753e-06 9.249753e-06 1.942448e-04 9.249753e-06
## topic9  4.983121e-04 7.165868e-03 7.018480e-06 7.018480e-06 7.018480e-06
## topic10 1.386559e-05 4.728165e-03 1.386559e-05 4.298332e-04 1.386559e-05
```

```
#Most representative words in Topic 1
mu[1,][order(mu[1,], decreasing=T)][1:10]
```

```
##     negoti  agreement      secur     govern    israeli     propos       side
## 0.03185529 0.02342714 0.02155421 0.01890091 0.01437468 0.01109706 0.01109706
##      parti   boundari       line
## 0.01109706 0.01094099 0.01078491
```

```
#Topical prevalence matrix
pi <- lda$theta
dim(pi) #number of docs by number of topics
```

```
## [1] 70 10
```

```r
#Most representative documents in Topic 1
metadata[order(pi[1,],decreasing=T),]
```

```
## # A tibble: 10 x 5
##    doc_id      text                                      country   session  year
##    <chr>       <chr>                                     <chr>       <dbl> <dbl>
##  1 ISR_29_197~ "At the outset of my remarks, I wish to ~ ISR_prev~      29  1974
##  2 ISR_31_197~ "Permit me to join the representatives w~ ISR_prev~      31  1976
##  3 ISR_25_197~ "93.\t: Mr. President, your country, Nor~ ISR_prev~      25  1970
##  4 ISR_33_197~ "70.\tMr. President, permit me to join t~ ISR_prev~      33  1978
##  5 ISR_26_197~ "60.\t Mr. President, you come to the le~ ISR_prev~      26  1971
##  6 ISR_27_197~ "Mr. President, I congratulate you on yo~ ISR_prev~      27  1972
##  7 ISR_30_197~ "33.\t Mr. President, it is with pleasur~ ISR_prev~      30  1975
##  8 ISR_34_197~ "161.\tMr, President, please accept my c~ ISR_prev~      34  1979
##  9 ISR_28_197~ "52.\t Mr. President, those of us who kn~ ISR_prev~      28  1973
## 10 ISR_32_197~ "154.\tMr. President, permit me to join ~ ISR_prev~      32  1977
```

STM

```r
#STM
#Process the data to put it in STM format.Textprocessor() automatically does pre-processing
temp <- textProcessor(documents=metadata$text,metadata=metadata)
```

```
## Building corpus...
## Converting to Lower Case...
## Removing punctuation...
## Removing stopwords...
## Removing numbers...
## Stemming...
## Creating Output...
```

```r
#prepDocuments() removes words/docs that are now empty after pre-processing
out <- prepDocuments(temp$documents, temp$vocab, temp$meta)
```

```
## Removing 2825 of 7143 terms (2825 of 51399 tokens) due to frequency
## Your corpus now has 70 documents, 4318 terms and 48574 tokens.
```

```r
#Let's try to distinguish between topics

#number of topic
num_topic = 5
model.stm <- stm(out$documents, out$vocab, K = num_topic, prevalence = ~country + s(year),
                 data = out$meta, max.em.its = 10)
```

```
## Beginning Spectral Initialization
##   Calculating the gram matrix...
##   Finding anchor words...
##      .....
##   Recovering initialization...
```

```
##      ......................................
## Initialization complete.
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 1 (approx. per word bound = -7.091)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 2 (approx. per word bound = -6.954, relative change = 1.938e-02)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 3 (approx. per word bound = -6.930, relative change = 3.337e-03)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 4 (approx. per word bound = -6.926, relative change = 6.930e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 5 (approx. per word bound = -6.924, relative change = 2.595e-04)
## Topic 1: palestinian, peac, state, peopl, will
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, nation, state, peac
##  Topic 5: peac, israel, nation, negoti, will
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 6 (approx. per word bound = -6.923, relative change = 1.495e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 7 (approx. per word bound = -6.922, relative change = 1.072e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 8 (approx. per word bound = -6.921, relative change = 8.387e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 9 (approx. per word bound = -6.921, relative change = 6.833e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Model Terminated Before Convergence Reached
```

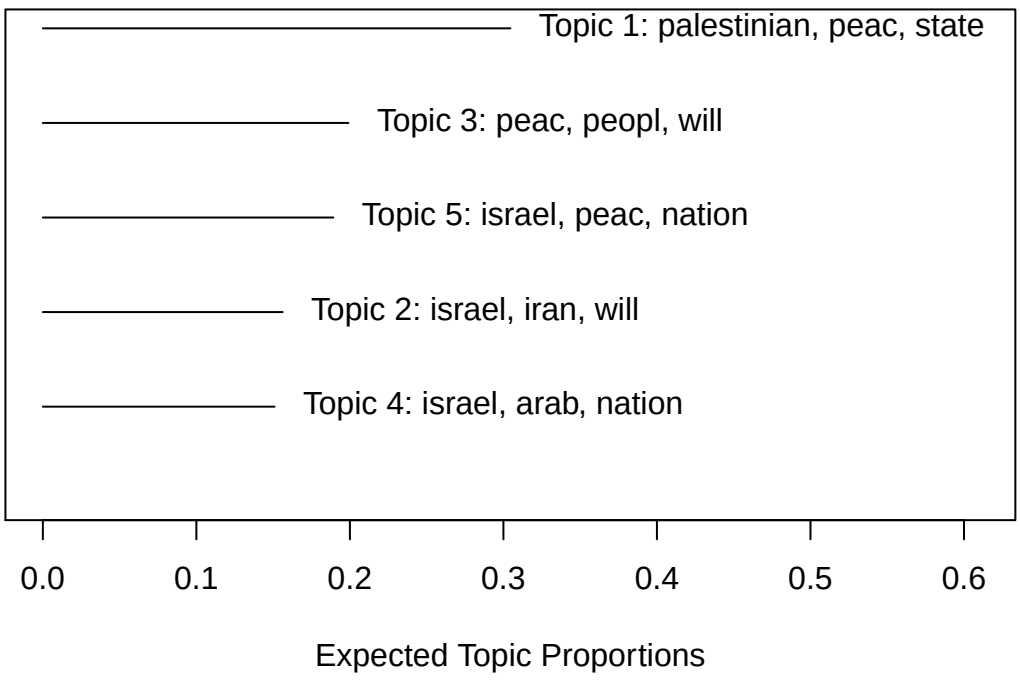
```
#Find most probable words in each topic
labelTopics(model.stm)
```

```
## Topic 1 Top Words:
##       Highest Prob: palestinian, peac, state, peopl, will, intern, israel
##       FREX: occup, occupi, palestin, implement, strip, two-stat, settler
```

```
##       Lift: accompli, advisori, agreed-upon, albright, anger, ascens, barack
##       Score: canton, occup, sieg, settler, occupi, al-nakba, two-stat
## Topic 2 Top Words:
##       Highest Prob: israel, iran, will, peac, year, peopl, nation
##       FREX: iran', iran, rouhani, hama, iranian, milit, nuclear
##       Lift: big, "death, al-assad', amazon, america", anchor, anytim
##       Score: iran', vow, rouhani, milit, uranium, isi, iran
## Topic 3 Top Words:
##       Highest Prob: peac, peopl, will, nation, new, palestinian, can
##       FREX: longer, economi, democraci, promis, global, valu, choic
##       Lift: wheel, borderless, boutros-ghali, children', east-west, forty-eighth, genesi
##       Score: wheel, forty-eighth, fresh, laden, saddam, scienc, wealth
## Topic 4 Top Words:
##       Highest Prob: israel, arab, nation, peac, state, unit, will
##       FREX: ceasefir, neighbor, arab, soviet, hijack, sovereignti, detent
##       Lift: amiti, arabisrael, blackmail, chad, congeni, coward, delic
##       Score: ehud, neighbor, aircraft, detent, propos, geneva, plo
## Topic 5 Top Words:
##       Highest Prob: israel, peac, nation, will, negoti, state, unit
##       FREX: soviet, framework, withdraw, treati, jar, boundari, relat
##       Lift: color, fifty-first, inviol, ismail, linchpin, marri, moratorium
##       Score: moratorium, canal, jar, judaea, neighbor, propos, suez
```

```
#This takes a while to run!
```

```
#And most common topics
plot(model.stm)
```

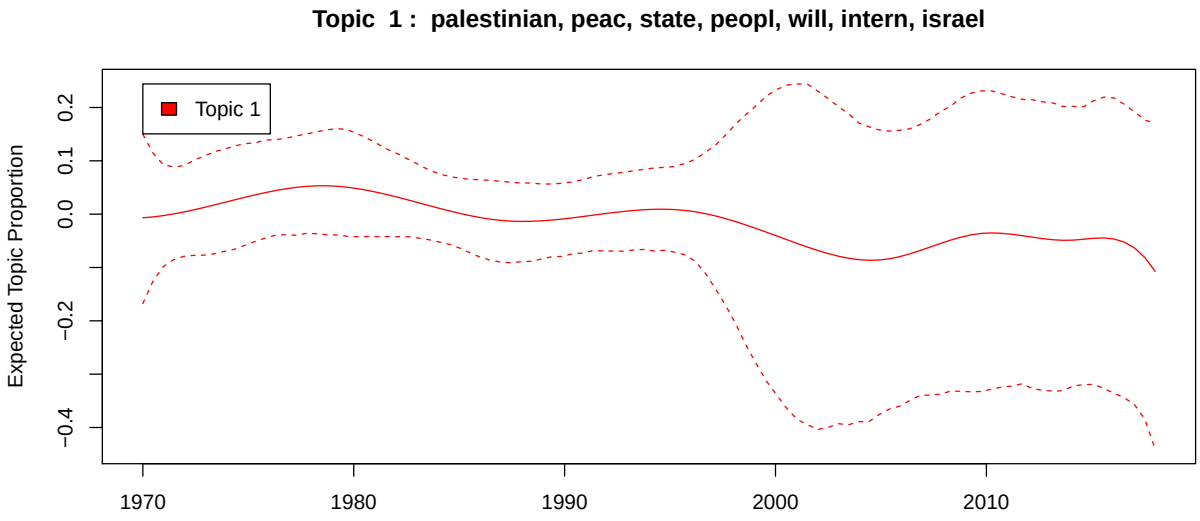


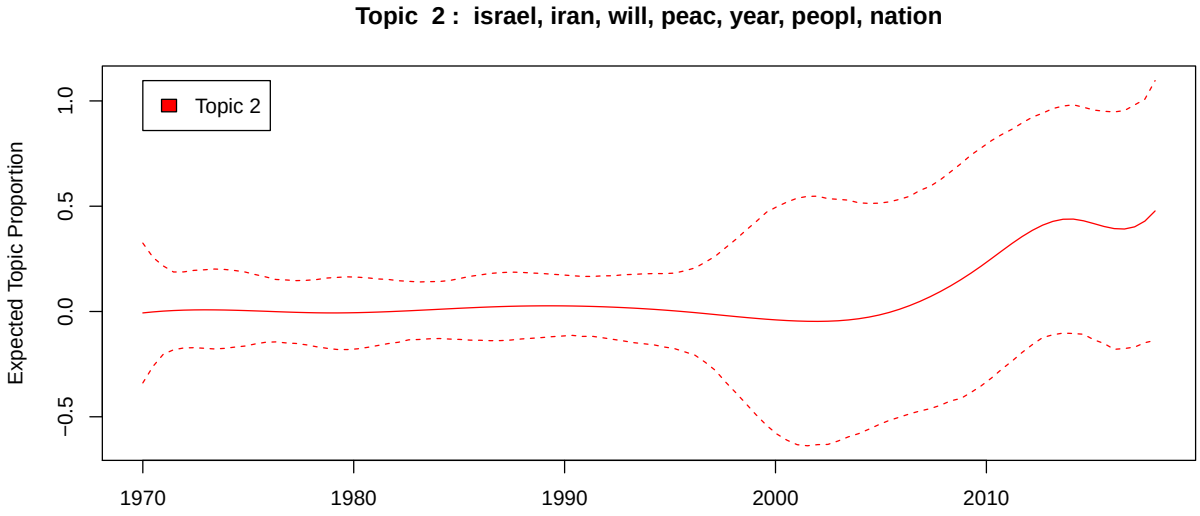

```
topic_words =
  c("palestinian, peac, state, peopl, will, intern, israel",
      "israel, iran, will, peac, year, peopl, nation",
      "peac, peopl, will, nation, new, palestinian, can ",
      "israel, arab, nation, peac, state, unit, will",
      "israel, peac, nation, will, negoti, state, unit"
  )
```

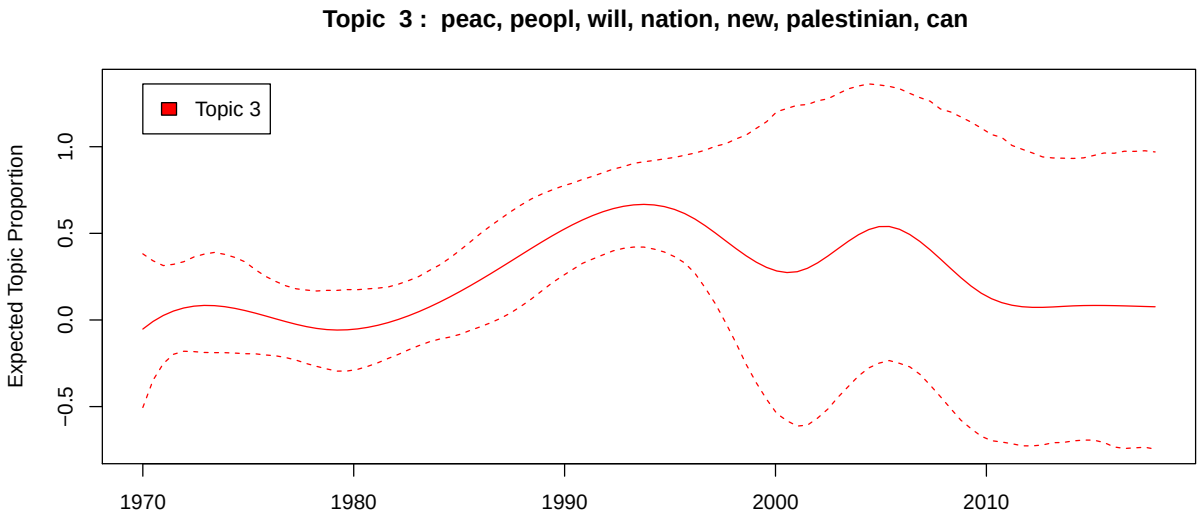
Plot each topic vs. countries, and effect of the topic over the years.

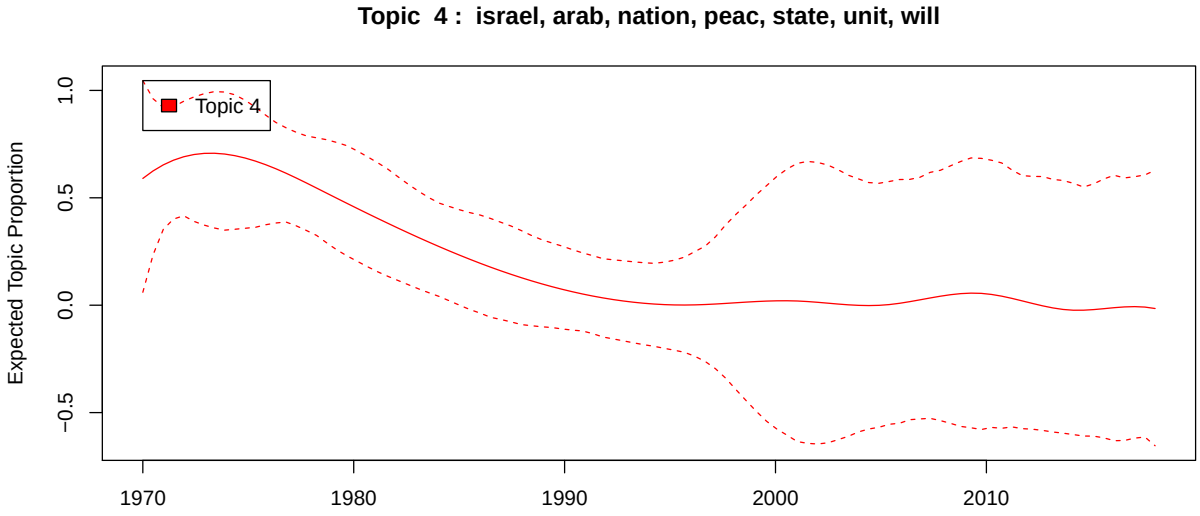
```
model.stm.ee <- estimateEffect(1:num_topic ~ country + s(year), model.stm, meta = out$meta)

dev.new(width=100, height=50, unit="in")
plot(model.stm.ee, "country", main="Topic num vs. Countries")
```

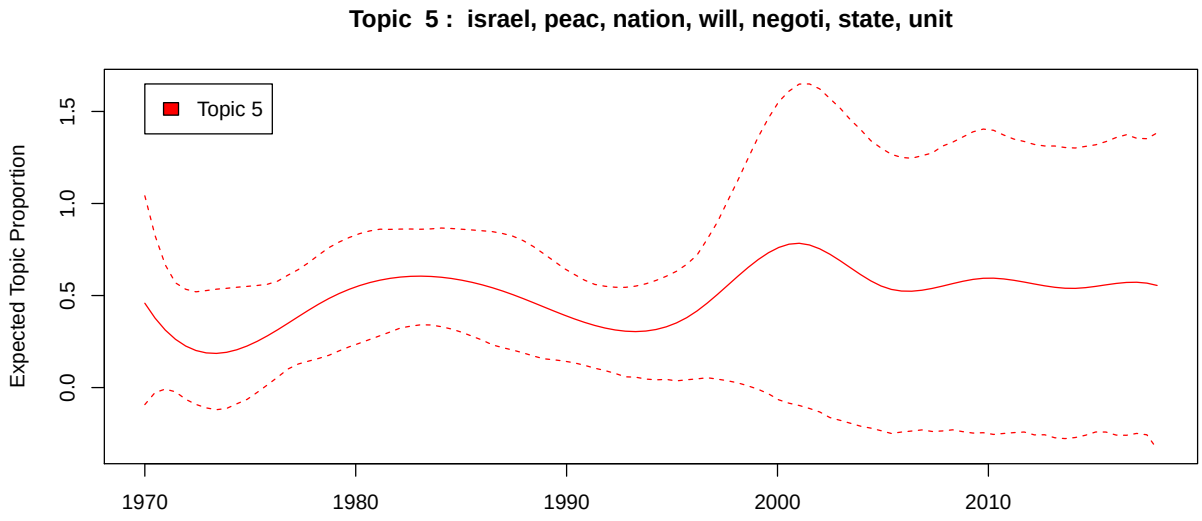
```
for (i in 1:num_topic){
  #plot(model.stm.ee, "country")
  plot(model.stm.ee, "year", method="continuous", topics=i, main = paste("Topic ", i, ": ", topic_words
}
```


Topic 1 : palestinian, peac, state, peopl, will, intern, israel
Topic 1
Expected Topic Proportion
0.2
0.1
0.0
−0.2
−0.4
1970
1980
1990
2000
2010


Topic 2 : israel, iran, will, peac, year, peopl, nation
Topic 2
Expected Topic Proportion
1.0
0.5
0.0
−0.5
1970
1980
1990
2000
2010


Topic 3 : peac, peopl, will, nation, new, palestinian, can
Topic 3
Expected Topic Proportion
−0.5
0.0
0.5
1.0
1970
1980
1990
2000
2010


Topic 4 : israel, arab, nation, peac, state, unit, will
Topic 4
Expected Topic Proportion
−0.5
0.0
0.5
1.0
1970
1980
1990
2000
2010



Get representative document

```
findThoughts(model.stm, texts=out$meta$year, topics=1, n=3)$docs
```

```
## $`Topic 1`
## [1] 2010 2009 2011
```

```
findThoughts(model.stm, texts=out$meta$country, topics=1, n=3)$docs
```

```
## $`Topic 1`
## [1] "PSE" "PSE" "PSE"
```

```
#findThoughts(model.stm, texts=out$meta$text, topics=i, n=1)$docs[1]
```

We can save the output document to a dataframe.

```
df = data.frame(matrix(vector(), 0, 4,
                dimnames=list(c(), c("Topic", "Year", "Country", "Text"))),
                stringsAsFactors=T)
```

```
for (i in 1:num_topic){

  df[(i-1) * 10 + 1: (i * 10), "Topic"] = list(rep(i,10))

  df[(i-1) * 10 + 1: (i * 10), "Year"] = findThoughts(model.stm, texts=out$meta$year, topics=i, n=10)$d

  df[(i-1) * 10 + 1: (i * 10), "Country"] = findThoughts(model.stm, texts=out$meta$country, topics=i, n

  df[(i-1) * 10 + 1: (i * 10), "Text"] = findThoughts(model.stm, texts=out$meta$text, topics=i, n=10)$d

}
```

Generate output dataframe with topics and document

```
head(df, 2)
```

```
##   Topic Year Country
## 1     1 2010     PSE
## 2     1 2009     PSE
##
## 1 It gives \nme pleasure to congratulate Mr. Joseph Deiss on his \nelection to the presidency at this
## 2
```

```
write.csv(df,'topic_model_output.csv', row.names = FALSE)
```