United Nations General Debates Text Analysis: Isreal vs. Palestine

Selected Corpus

UN General Debate Corpus, collected by Baturo, Dasandi, and Mikhaylov, containing all 8093 UN General Debate statements presented from 1970 to 2018, and their corresponding metadata. This corpus is great for our research questions because the nature of UNGD could allow smaller states to raise issues that they believe are important but received less attention. While other states could use GD as a way to influence international perceptions of their states and other states. Thus, we can observe the fairly accurate policy preferences of Israel and Palestine.

Research Question

How did the policy preferences in UNGD of Israel change overtime? Is there a dramatic shift after Palestine joined the UNGD in 1998? Could the General Debates reflect the conflicts between Israel and Palestine?

Research Question Significance

About two weeks ago, a serious armed conflict involving airstrikes and missile attacks broke out between Israel and Palestine. The enduring Israeli-Palestinian conflict made our group wonder whether their hostility were already embedded in their General Debates at the United Nation, and how their policy preferences changed over time. If we can find patterns in their speeches, we may understand their conflicts from a more comprehensive perspective.

Related Study

Jeremy Pressman (2020) 'History in conflict: Israeli–Palestinian speeches at the United Nations, 1998–2016', Mediterranean Politics, 25:4, 476-498, DOI: 10.1080/13629395.2019.1589936

In this paper, Pressman studied the General Debate of both Israel and Palestine from 1998 to 2016, as the result Pressman found that the leaders of both countries covered similar issues. And both countries argued that they are the ones committing for peace, while accusing the other country of invasion. While this paper doesn't use any topic modeling techniques and only include speeches until 2016, it still helped our research by providing background knowledge of the history of Israeli-Palestinian conflict, helping us interpret the results, and providing validation for our findings.

Load required packages

```r
#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()
```

```r
library(tokenizers)
```

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

```r
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.
```

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

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

```r
#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
```

```r
#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 year 1998, because it is the year Palestine first joined the United Nations General Debates.

```
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:

```r
#remove punct, stopwords.. etc
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 include features appeared in at least 5% of documents. Calling method dfm_trim from the quanteda package, and obtain a new document frequency matrix with 2471 features.

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

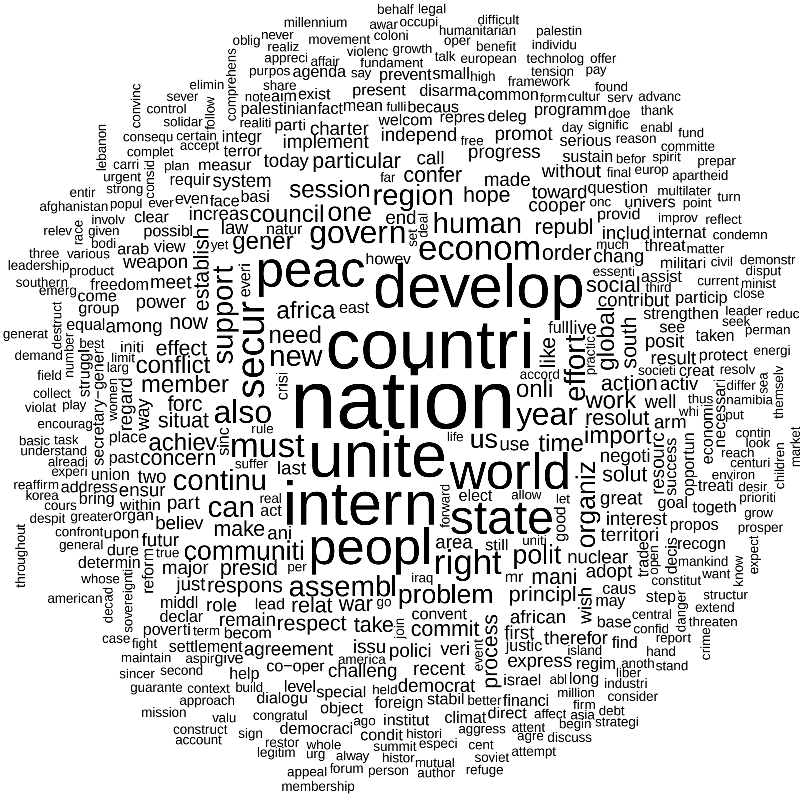
```
## Document-feature matrix of: 8,093 documents, 2,471 features (76.26% sparse) and 3 docvars.
##                    features
## docs               may first convey presid congratul deleg elect session gener
##   ALB_25_1970.txt   5     4      1      3         1     3     1       5     5
##   ARG_25_1970.txt   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.

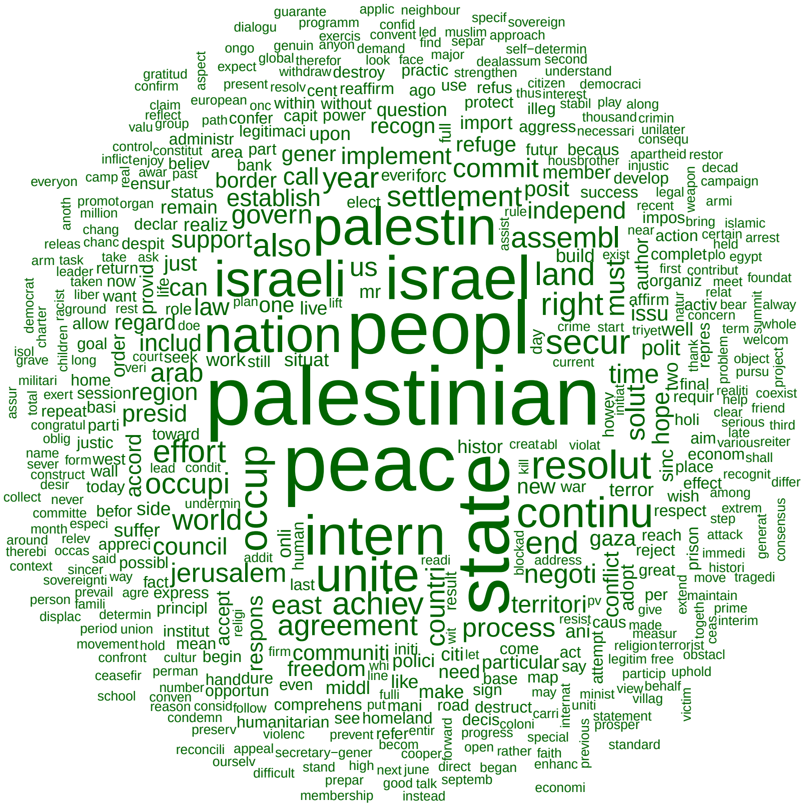
```
#all word based on their word frequency.
textplot_wordcloud(dfm_trimmed, col="black")
```



```r
#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"),]
```

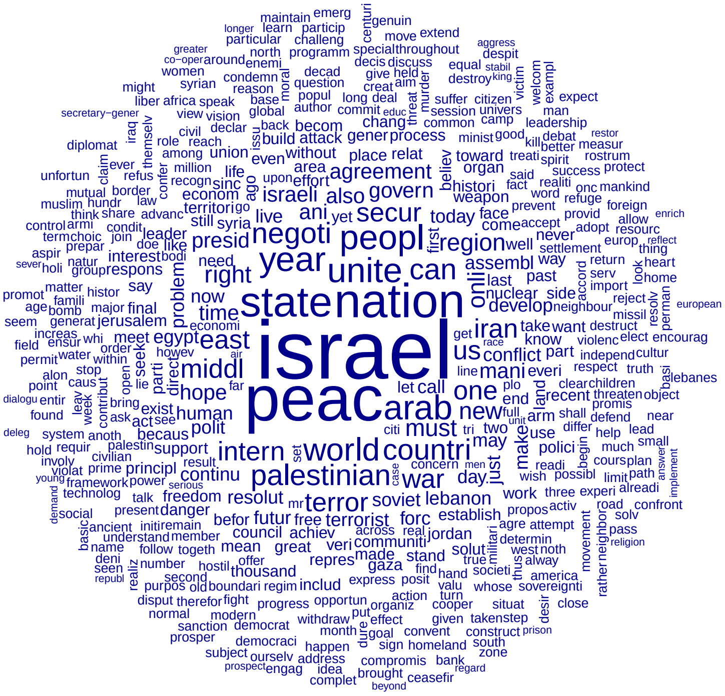
Word Cloud of only Palestine.

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



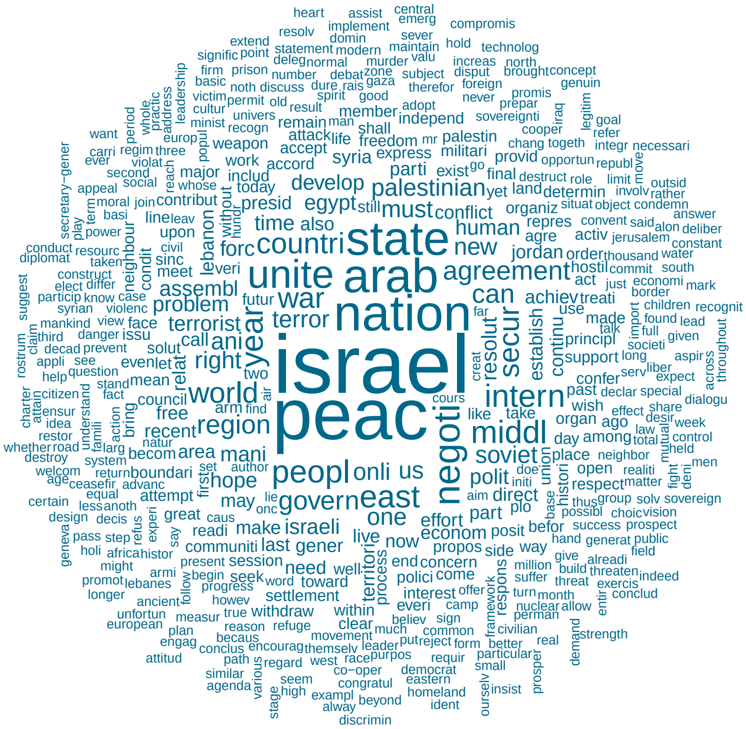
Word Cloud of only Israel

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



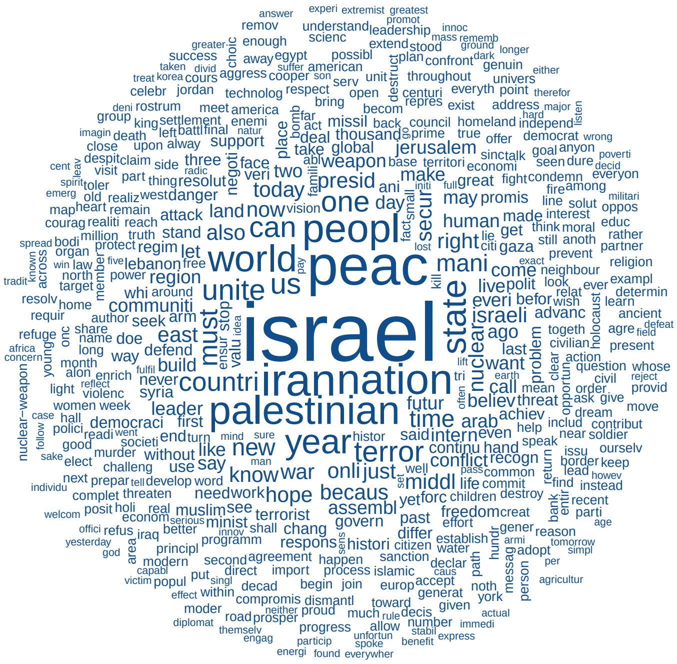
Word Cloud of only Israel before 1998, when Palestine was not in the United nations.

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



Palestine joined the UNDB 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 words "palestinian" and "Iran" appeared more frequently after 1998, while the word "arab" appeared less frequently.

Find distinctive words

```
#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 before 1998 are:

- 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 after 1998 are:

- 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 Palestine's speech in after 1998 are:

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

```
#dfm_trimmed
```

Topic Modelling

LDA:

```
#LDA
######
#Run LDA using quanteda
lda <- textmodel_lda(dfm_trimmed, k = 5)

#Most likely term for each topic
lda.terms <- terms(lda, 5)
lda.terms
```

```
##      topic1   topic2   topic3      topic4        topic5
## [1,] "israel" "israel" "arab"      "palestinian" "peac"
## [2,] "iran"   "nation" "negoti"    "peopl"       "can"
## [3,] "year"   "peac"   "agreement" "peac"        "new"
## [4,] "state"  "unite"  "secur"     "state"       "us"
## [5,] "peopl"  "arab"   "israel"    "intern"      "middl"
```

```
#Topical content matrix
mu <- lda$phi
dim(mu) #5 topics, 2741 words
```

```
## [1]    5 2471
```

```
mu[1:5,1:10]
```

```
##                  may        first       convey       presid    congratul
## topic1 1.969313e-03 3.993841e-03 6.134932e-06 7.736149e-03 6.134932e-06
## topic2 4.354623e-06 7.881868e-04 4.354623e-06 1.180103e-03 4.833632e-04
## topic3 1.712609e-03 1.854734e-03 7.816886e-05 4.626175e-03 7.106260e-06
## topic4 4.454323e-06 4.454323e-06 4.498866e-04 4.057888e-03 1.385294e-03
## topic5 6.181216e-03 2.348952e-03 4.508546e-06 4.508546e-06 8.160468e-04
##                deleg        elect      session         gener      assembl
## topic1 6.134932e-06 6.134932e-06 6.134932e-06 3.742308e-04 1.601217e-03
## topic2 1.785395e-04 4.354623e-06 3.226776e-03 3.139683e-03 5.796003e-03
## topic3 1.073045e-03 7.106260e-06 7.106260e-06 2.991735e-03 7.106260e-06
## topic4 4.454323e-06 1.608011e-03 1.652554e-03 3.968802e-03 7.621347e-03
## topic5 9.062177e-04 2.168611e-03 3.651922e-04 4.508546e-06 1.397649e-04
```

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

```
##      israel        iran        year       state       peopl       world
## 0.046386218 0.021171649 0.015220766 0.012460046 0.012460046 0.011171711
##      nation         one       unite     countri
## 0.010742265 0.009024485 0.008840437 0.008410991
```

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

```
## [1] 70  5
```

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

```
## # A tibble: 5 x 5
##   doc_id      text                                     country    session  year
##   <chr>       <chr>                                    <chr>        <dbl> <dbl>
## 1 ISR_27_197~ "Mr. President, I congratulate you on yo~ ISR_prev_~      27  1972
## 2 ISR_26_197~ "60.\t Mr. President, you come to the le~ ISR_prev_~      26  1971
## 3 ISR_29_197~ "At the outset of my remarks, I wish to ~ ISR_prev_~      29  1974
## 4 ISR_25_197~ "93.\t: Mr. President, your country, Nor~ ISR_prev_~      25  1970
## 5 ISR_28_197~ "52.\t Mr. President, those of us who kn~ ISR_prev_~      28  1973
```

STM

Model Selection LDA vs. STM: we performed LDA as well as STM analysis, and we found that since STM can take in account of medata, which will be useful in further analysis with the help of representative documents, we choose to mainly use STM.

Number of Topics To study the topics of statements of these two countries we decided to use 5 topics. Because we evaluated the results qualitatively, we found that if we use 10 topics there would be overlapped topics. Thus, 5 topics were best to answer our research questions.

STM Topic Model: 2 Topics with Just Israel

Does Israel really changed its debate topic over the years? We can apply the STM model just on Israel to make sure that the arbitrary seperation of ISR at year 1998 is reasonable.

We also need to make sure that the model converges to the optimum.

```r
isrmeta = metadata[metadata$country != "PSE", ] # Subset, only include Israeli documents
isrtemp = textProcessor(documents = isrmeta$text, metadata = isrmeta) # Preprocessing
```

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

```r
isrout <- prepDocuments(isrtemp$documents, isrtemp$vocab, isrtemp$meta)
```

```
## Removing 2676 of 6581 terms (2676 of 38173 tokens) due to frequency
## Your corpus now has 49 documents, 3905 terms and 35497 tokens.
```

```r
isrmode <- stm(isrout$documents, isrout$vocab, K = 2,
               prevalence = ~s(year), data = isrout$meta) # Run STM
```

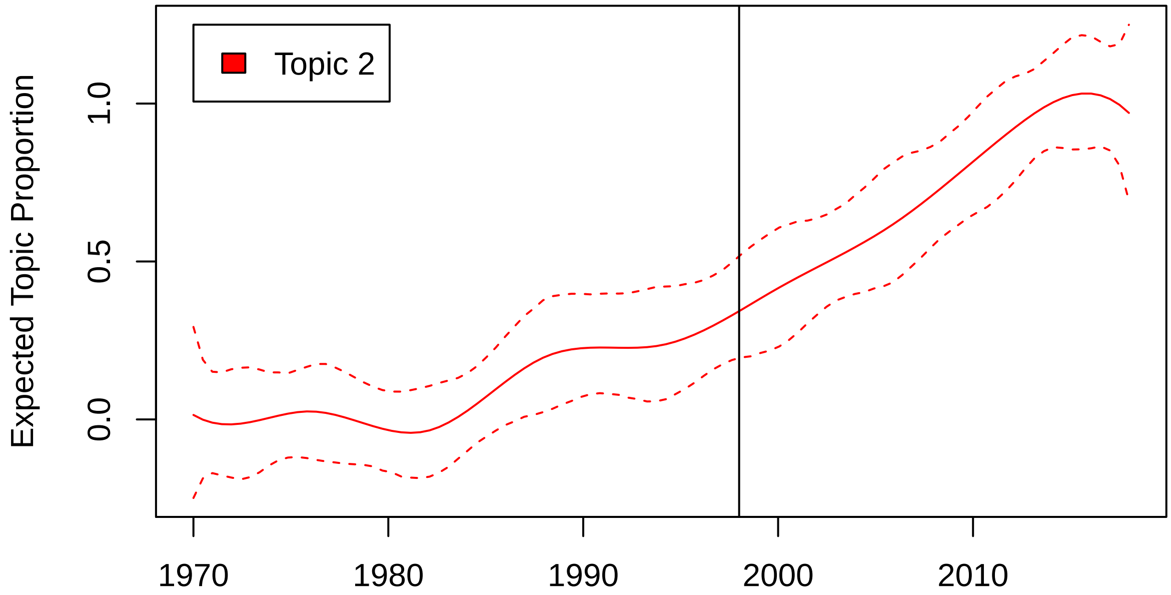
```
## Warning in stm(isrout$documents, isrout$vocab, K = 2, prevalence = ~s(year), :
## K=2 is equivalent to a unidimensional scaling model which you may prefer.
```

```
## 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 = -7.168)
## .................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 2 (approx. per word bound = -7.094, relative change = 1.030e-02)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 3 (approx. per word bound = -7.073, relative change = 3.023e-03)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 4 (approx. per word bound = -7.068, relative change = 6.028e-04)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 5 (approx. per word bound = -7.067, relative change = 1.939e-04)
## Topic 1: israel, peac, nation, state, arab
##  Topic 2: israel, peac, will, palestinian, peopl
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 6 (approx. per word bound = -7.066, relative change = 9.281e-05)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 7 (approx. per word bound = -7.066, relative change = 5.082e-05)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 8 (approx. per word bound = -7.066, relative change = 3.178e-05)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 9 (approx. per word bound = -7.065, relative change = 2.119e-05)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 10 (approx. per word bound = -7.065, relative change = 1.478e-05)
## Topic 1: israel, peac, nation, state, arab
##  Topic 2: israel, peac, will, peopl, palestinian
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 11 (approx. per word bound = -7.065, relative change = 1.067e-05)
## ................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Model Converged
```

```r
labelTopics(isrmode) # Interprete results
```

```
## Topic 1 Top Words:
##       Highest Prob: israel, peac, nation, state, arab, will, unit
##       FREX: propos, neighbor, co-oper, canal, plo, armistic, geneva
##       Lift: acut, adequ, administ, ambigu, ampli, anarchi, anatoli
##       Score: propos, plo, neighbor, canal, fifty-first, sovereignti, armistic
## Topic 2 Top Words:
##       Highest Prob: israel, peac, will, peopl, palestinian, iran, nation
##       FREX: iran’, iranian, rouhani, uranium, abba, atom, milit
##       Lift: america", anchor, appetit, balfour, benjamin, big, car
##       Score: iran’, ehud, rouhani, isi, trump, ali, abba
```

```
isrmode.ee <- estimateEffect(1:2 ~ s(year), isrmode, meta = isrout$meta) # Estimate Effect
plot(isrmode.ee, "year", method = "continuous", topics = 2) # Plot effect
abline(v = 1998)
```



```
# text(locator(), labels = c("1998")) # Requires interaction
```

There is clearly change in topic after 1998, the year of interest. Now, we will perform our main analysis on these three groups: ISR before 1998, ISR after 1998, and Palestine, which joined the UNGB after 1998.

STM Main Analysis

```
#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.
```

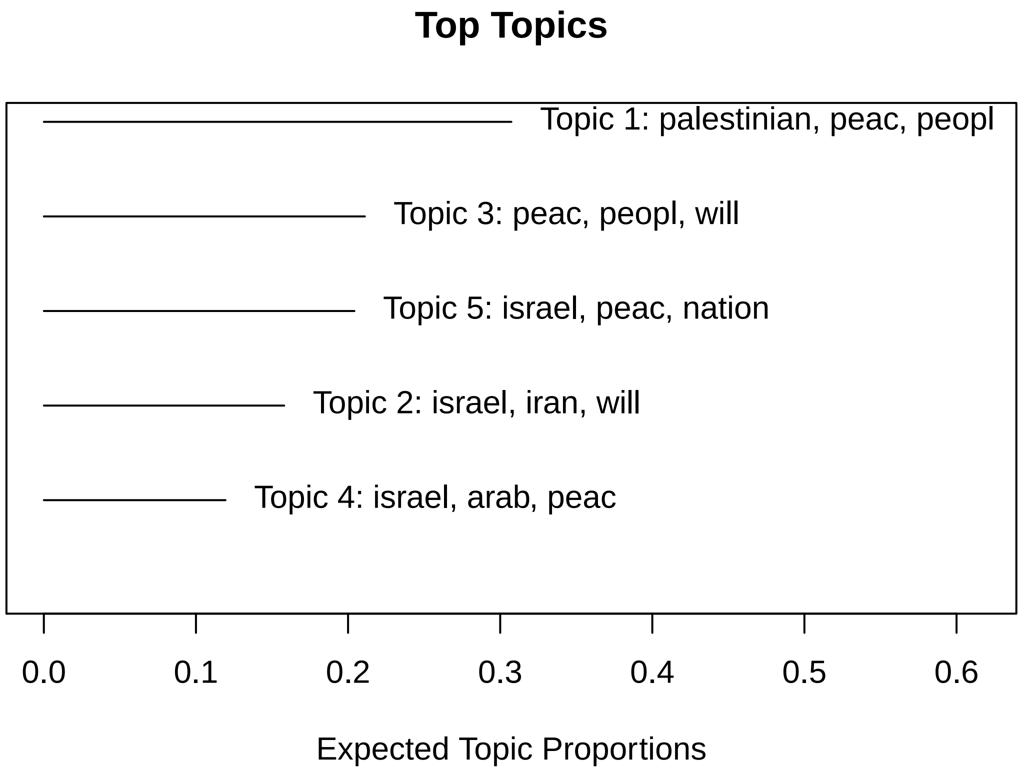
```r
#Let's try to distinguish between topics

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

```
## 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.
## Completing Iteration 10 (approx. per word bound = -6.921, relative change = 5.984e-05)
## 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, peac, state
##  Topic 5: israel, peac, nation, will, negoti
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 11 (approx. per word bound = -6.920, relative change = 5.624e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 12 (approx. per word bound = -6.920, relative change = 4.874e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 13 (approx. per word bound = -6.920, relative change = 4.790e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 14 (approx. per word bound = -6.919, relative change = 6.601e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 15 (approx. per word bound = -6.918, relative change = 1.023e-04)
## Topic 1: palestinian, peac, state, peopl, intern
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, state, peac, nation
##  Topic 5: israel, peac, nation, will, state
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 16 (approx. per word bound = -6.917, relative change = 1.591e-04)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 17 (approx. per word bound = -6.916, relative change = 2.115e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 18 (approx. per word bound = -6.914, relative change = 2.059e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 19 (approx. per word bound = -6.913, relative change = 1.444e-04)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 20 (approx. per word bound = -6.913, relative change = 8.798e-05)
## Topic 1: palestinian, peac, peopl, state, intern
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, peac, state, nation
##  Topic 5: israel, peac, nation, will, state
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 21 (approx. per word bound = -6.912, relative change = 5.426e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 22 (approx. per word bound = -6.912, relative change = 3.948e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 23 (approx. per word bound = -6.912, relative change = 3.988e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 24 (approx. per word bound = -6.912, relative change = 4.345e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 25 (approx. per word bound = -6.911, relative change = 4.013e-05)
## Topic 1: palestinian, peac, peopl, state, intern
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, peac, state, nation
##  Topic 5: israel, peac, nation, will, state
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 26 (approx. per word bound = -6.911, relative change = 3.897e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 27 (approx. per word bound = -6.911, relative change = 3.952e-05)
## .....................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 28 (approx. per word bound = -6.910, relative change = 3.670e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 29 (approx. per word bound = -6.910, relative change = 3.017e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 30 (approx. per word bound = -6.910, relative change = 2.252e-05)
## Topic 1: palestinian, peac, peopl, state, intern
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, peac, state, nation
##  Topic 5: israel, peac, nation, will, state
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 31 (approx. per word bound = -6.910, relative change = 2.111e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 32 (approx. per word bound = -6.910, relative change = 2.481e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 33 (approx. per word bound = -6.910, relative change = 2.266e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 34 (approx. per word bound = -6.910, relative change = 1.803e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 35 (approx. per word bound = -6.909, relative change = 1.293e-05)
## Topic 1: palestinian, peac, peopl, state, intern
##  Topic 2: israel, iran, will, peac, year
##  Topic 3: peac, peopl, will, nation, new
##  Topic 4: israel, arab, peac, state, nation
##  Topic 5: israel, peac, nation, will, unit
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 36 (approx. per word bound = -6.909, relative change = 1.016e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 37 (approx. per word bound = -6.909, relative change = 1.027e-05)
## ......................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Model Converged
```

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

```
## Topic 1 Top Words:
##       Highest Prob: palestinian, peac, peopl, state, intern, will, israel
##       FREX: occup, occupi, palestin, settler, implement, two-stat, strip
##       Lift: accompli, agreed-upon, albright, anger, ascens, barack, befel
##       Score: occup, sieg, two-stat, settler, al-nakba, occupi, palestine’
## Topic 2 Top Words:
##       Highest Prob: israel, iran, will, peac, year, peopl, nation
##       FREX: iran’, iran, rouhani, uranium, hama, nuclear, iranian
##       Lift: big, demon, deton, drone, finish, goldin, haley
##       Score: iran’, rouhani, uranium, milit, isi, iran, nuclear-arm
## Topic 3 Top Words:
##       Highest Prob: peac, peopl, will, nation, new, can, palestinian
##       FREX: longer, economi, promis, choic, democraci, valu, scienc
##       Lift: wheel, -embrac, barcelona, bastion, ben-gurion, beneath, blown
##       Score: wheel, forty-eighth, fresh, laden, saddam, wealth, ecolog
## Topic 4 Top Words:
##       Highest Prob: israel, arab, peac, state, nation, unit, agreement
##       FREX: boundari, egyptian, ceasefir, canal, jar, propos, egypt
##       Lift: adjud, adjust, airbus, amin, amiti, ampli, antiisrael
##       Score: ehud, neighbor, canal, suez, propos, jar, detent
## Topic 5 Top Words:
##       Highest Prob: israel, peac, nation, will, unit, state, countri
##       FREX: treati, plo, david, soviet, lebanon, camp, co-oper
##       Lift: carter, anatoli, annul, anti-soviet, autonomi, begin", broaden
##       Score: co-oper, neighbor, arab-israel, plo, unifil, moratorium, judaea
```

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

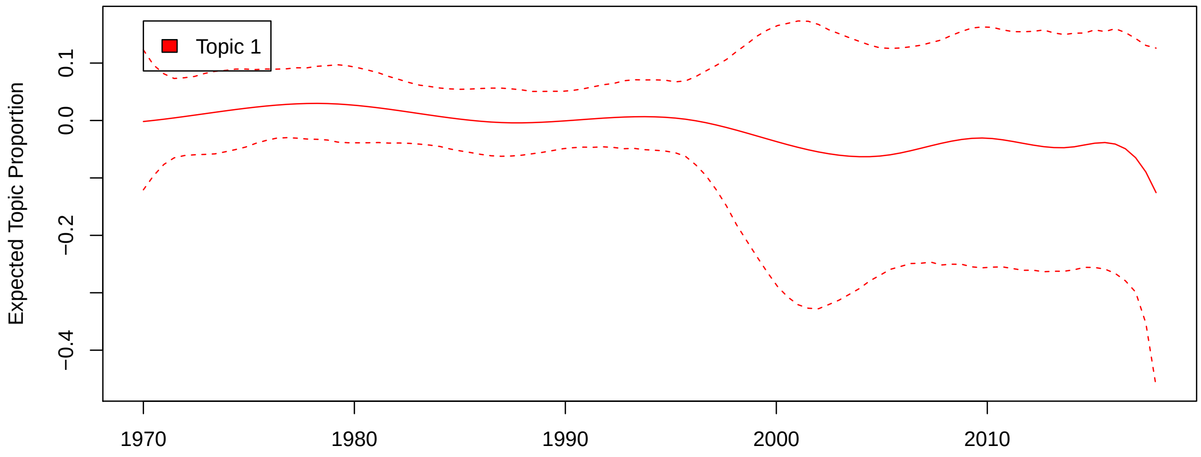


```
topic_words =
  c("palestinian, peac, peopl, state, intern, will, israel",
      "israel, iran, will, peac, year, peopl, nation",
      "peac, peopl, will, nation, new, can, palestinian",
      "israel, arab, peac, state, nation, unit, agreement",
      "israel, peac, nation, will, unit, state, countri"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

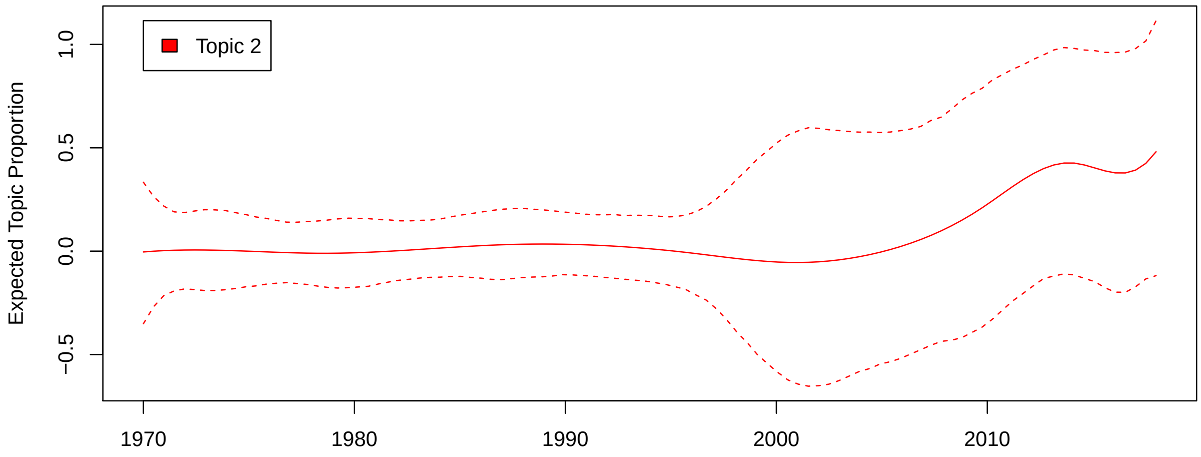
Plot each topic vs. countries, and effect of the topic over the years.

```
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")
```

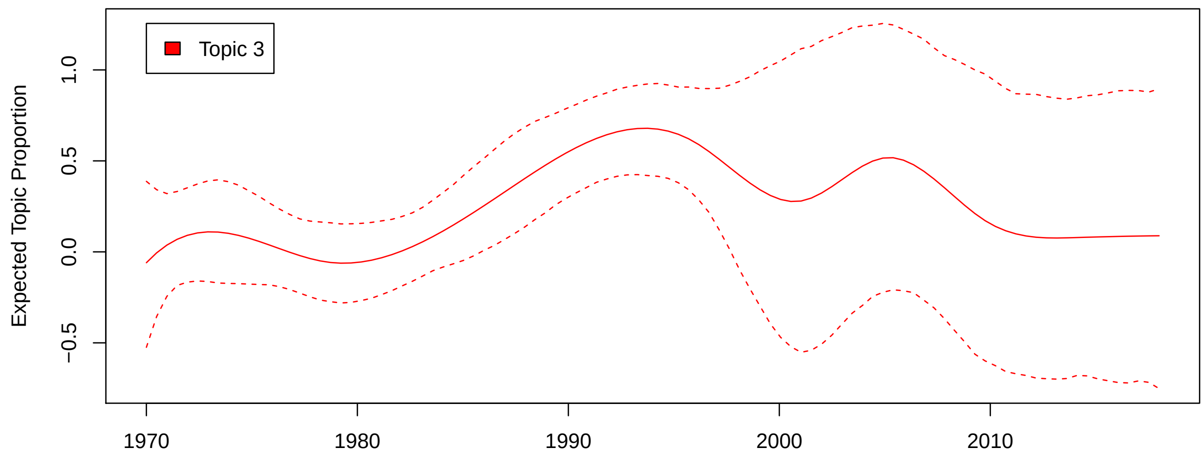
```
for (i in 1:num_topic){
  #plot(model.stm.ee, "country")
  plot(model.stm.ee, "year", method="continuous", topics=i, main = paste("Topic ", i, ": ", topic_words
}
```



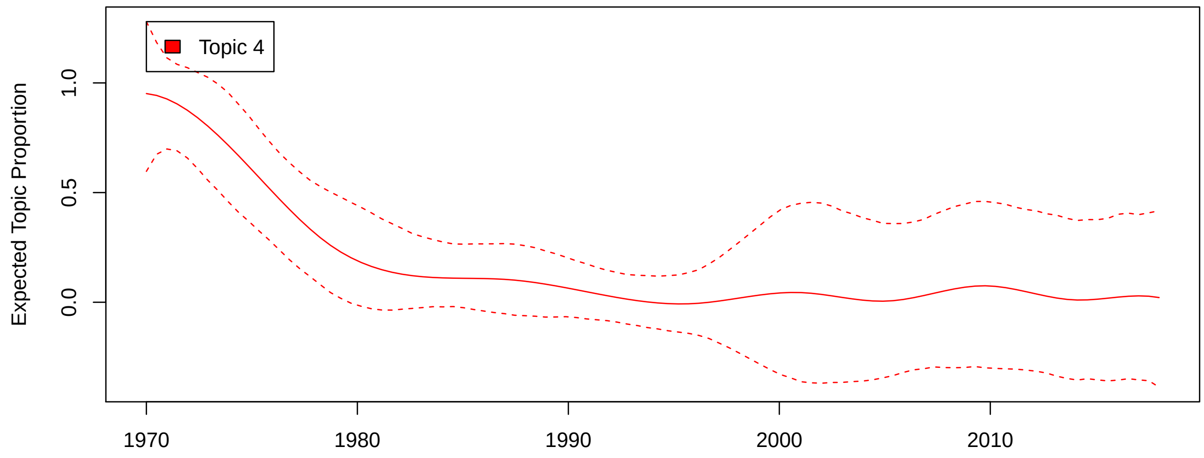
Topic 1 : palestinian, peac, peopl, state, intern, will, israel



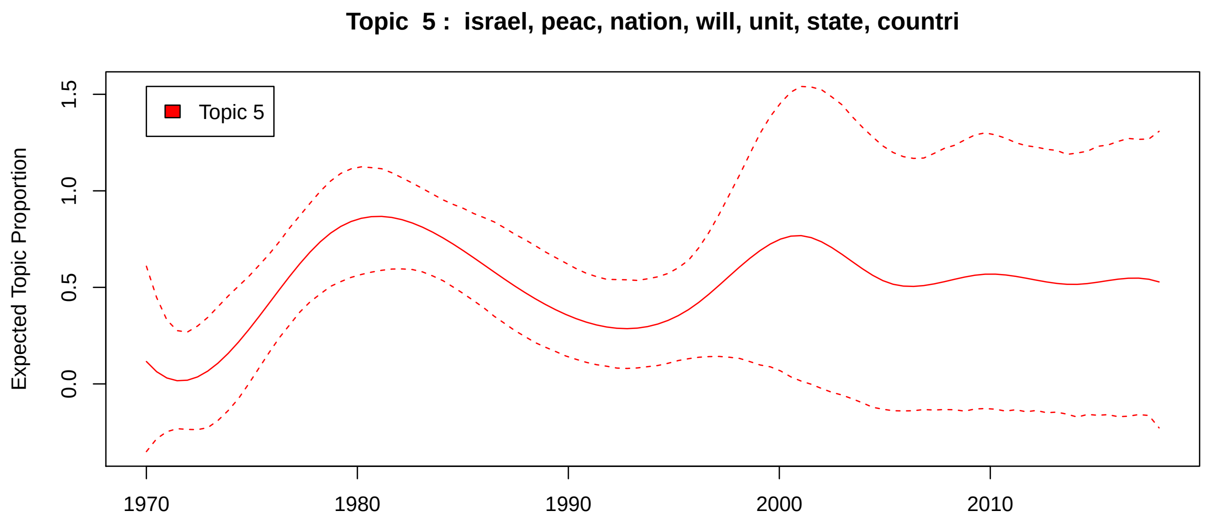
Topic 2 : israel, iran, will, peac, year, peopl, nation



Topic 3 : peac, peopl, will, nation, new, can, palestinian



Topic 4 : israel, arab, peac, state, nation, unit, agreement



Get representative document

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

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

```
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, 5,
                dimnames=list(c(), c("Topic", "Word", "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), "Word"] = topic_words[i]

  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                                                   Word Year Country
## 1     1 palestinian, peac, peopl, state, intern, will, israel 2010     PSE
## 2     1 palestinian, peac, peopl, state, intern, will, israel 2011     PSE
##
representative##nd1 \ntruly reflective of the current international situation. \nThis is especially important in the li
## 2 At the \noutset, I extend my congratulations to you, Sir, on your \nassumption of the presidency o
```

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