Assignment 2

Text as Data

Janine De Vera | 219848

2022-11-17

Introduction

In this assignment, you are asked to use topic modelling to investigate manifestos from the manifesto project maintained by WZB. You can either use the UK manifestos we looked at together in class, or collect your own set of manifestos by choosing the country/countries, year/years and party/parties you are interested in. You should produce a report which includes your code, that addresses the following aspects of creating a topic model, making sure to answer the questions below.

1. Data acquisition, description, and preparation

1.1 Prepare dataframe

For this task, I retrieve a dataset of party manifestos from the **United States and United Kingdom between 1960 and 2022**. Each country and each year will have a unique corpus. In order to create a dataframe where each text is one observation, I follow these steps:

1. Create a list containing all corpora from US and UK for the years of interest.
2. Extract manifestos from each corpus. Some manifestos treat each sentence as one observation while others have the entire text in a single line. For the latter, I separate the full text into sentences and process accordingly (e.g. remove blanks).
3. Construct a dataframe where each observation is a line from the manifesto.
4. Merge the corpus dataframe with the WZB meta dataframe in order to get corresponding information on the year, party, and country.

```
## Connecting to Manifesto Project DB API...
## Connecting to Manifesto Project DB API... corpus version: 2022-1
## Connecting to Manifesto Project DB API...
## Connecting to Manifesto Project DB API... corpus version: 2022-1
## Connecting to Manifesto Project DB API... corpus version: 2022-1
## Connecting to Manifesto Project DB API... corpus version: 2022-1
## Connecting to Manifesto Project DB API... corpus version: 2022-1
## Connecting to Manifesto Project DB API... corpus version: 2022-1
```

```
## # A tibble: 6 x 6
##   corpus_code  edate      partyname    partyabbrev text                  country
##   <chr>        <date>     <chr>        <chr>       <chr>                 <chr>
```

```
## 1 51320_196410 1964-10-15 Labour Party Labour      "\"THE NEW BRITAIN\"~ United~
## 2 51320_196410 1964-10-15 Labour Party Labour      " The British people~ United~
## 3 51320_196410 1964-10-15 Labour Party Labour      " And now, at last, ~ United~
## 4 51320_196410 1964-10-15 Labour Party Labour      " The dying months o~ United~
## 5 51320_196410 1964-10-15 Labour Party Labour      "  A New Britain - m~ United~
## 6 51320_196410 1964-10-15 Labour Party Labour      "  The country needs~ United~
```

Since I am interested in trends in **foreign policy**, I only need a subset of the dataframe above. I filter texts that contain the words *foreign policy* or *foreign*. The research question will be discussed in more detail in the next section.

Below are some information on the dataset that I will use for the rest of the analysis.

Number of texts:

```
## [1] 954
```

Years included:

```
##  [1] 1964 1966 1970 1974 1979 1983 1987 1992 1997 2001 2005 2015 2017 2019 1960
## [16] 1968 1972 1976 1980 1984 1988 1996 2000 2004 2008 2012 2016 2020
```

Countries:

```
## # A tibble: 2 x 1
##   country
##   <chr>
## 1 United Kingdom
## 2 United States
```

Parties by Country:

```
## # A tibble: 16 x 2
##    country        partyname
##    <chr>          <chr>
##  1 United Kingdom Labour Party
##  2 United Kingdom Liberal Party
##  3 United Kingdom Conservative Party
##  4 United Kingdom Social Democratic Party
##  5 United Kingdom Liberal Democrats
##  6 United Kingdom Scottish National Party
##  7 United Kingdom United Kingdom Independence Party
##  8 United Kingdom Green Party of England and Wales
##  9 United Kingdom Social Democratic and Labour Party
## 10 United Kingdom Ulster Unionist Party
## 11 United Kingdom The Party of Wales
## 12 United Kingdom Democratic Unionist Party
## 13 United Kingdom We Ourselves
## 14 United Kingdom Alliance Party of Northern Ireland
## 15 United States  Democratic Party
## 16 United States  Republican Party
```

1.2 Prepare document feature matrix

For the document feature matrix, I treat each line of text as one "document". This is tokenized and pre-processed as follows:

1. Remove punctuation
2. Remove English stop words
3. Lemmatization

It is important to conduct these pre-processing steps since they will affect the analysis later on. I remove punctuation marks and stop words since they do not provide any additional information regarding my research question. I opt for lemmatization so words with similar context are analyzed together. As a final pre-processing step, I rename the rows of the document feature matrix with meaningful IDs, specifically the corpus code from my original dataframe. The document feature matrix has 954 documents and 3,266 features.

```
## Document-feature matrix of: 954 documents, 3,266 features (99.51% sparse) and 0 docvars.
##                 features
## docs             hope blight stalin's brutal intransigence labour's foreign
##   51320_196410.1    1      1        1      1             1        1       1
##   51320_196410.2    0      0        0      0             0        0       1
##   51420_196410.1    0      0        0      0             0        0       1
##   51620_196410.1    0      0        0      0             0        0       1
##   51620_196410.2    0      0        0      0             0        0       1
##   51620_196410.3    0      0        0      0             0        0       1
##                 features
## docs             secretary ernest bevin
##   51320_196410.1         1      1     1
##   51320_196410.2         0      0     0
##   51420_196410.1         0      0     0
##   51620_196410.1         0      0     0
##   51620_196410.2         0      0     0
##   51620_196410.3         0      0     0
## [ reached max_ndoc ... 948 more documents, reached max_nfeat ... 3,256 more features ]
```

As a quick check, I extract the top 10 features of my document feature matrix.

```
##    foreign     policy       will   american      world government   national
##        949        360        252        145         91         79         77
##       must    country     nation
##         72         72         71
```

2. Research question

My main research questions is:

Was there a change in the foreign policy stance of the world's most influential democracies, United States and United Kingdom, during and after the Cold War?

The Cold War began in 1947 after World War II and lasted until the fall of the Soviet Union in 1991. It

was a period characterized by geopolitcal tension between the United States and the Soviet Union and their respective allies. Foreign policy priorities of the United States and United Kingdom, WWII allies and two of the most influential democracies, were undoubtedly different now compared to when the world was on the brink of a large scale conflict.

Through topic modeling using party manifestos, I analyze how these priorities changed after the Cold War. Since manifestos are available for several years, it is possible to introduce the time component into the text analysis. Topic models generate phrases and words that it thinks are related. It does so by finding hidden semantic structure in documents and clustering them into similar groups that readers can understand. In this sense, the research question will be answered by extracting prominent topics within the defined time frame.

I use the years 1960-1991 for the Cold War period and 1992-2022 for the post-Cold War period. This gives me roughly 30 years for each time period.

3-4. Topic model development & topic model description

To analyze my research question, I use the topic model Latent Dirichlet Allocation (LDA). As explained above, a topic model uses corpora (a set of documents) to find hidden semantic structures or configurations in texts that give them their intended meaning. Topic models provide a set of words and phrases that it finds to be related.

LDA is a probabilistic model that treats each document as a combination of a fixed number of topics. These topics have a corresponding probability of appearing in a given document. In addition, each word in the corpus also has a probability of appearing in a given topic. The topics are determined by analyzing the co-occurence of these words. LDA produces two matrices. The the **gamma** matrix or **document-topic** matrix provides the probability of each topic appearing in a particular document, while the **beta** matrix or **topic-term** matrix provides the probability of each term belonging to a particular topic.

LDA with four topics

The choice of the number of fixed topics is a hyperparameter - a modeling decision that the researcher has to make. For this LDA, I initially chose **four topics**. Setting this hyperparameter to a small number means that I want my topics to be relatively broad. This makes sense for the analysis because I first want to get an idea of the general themes in foreign policy stance during and after the Cold War. As I examine the words generatad per topic, I may opt to increase the hyperparameter to ask the model to come up with narrower topics.

Other hyperparameters are alpha_W which determines whether the researcher wants documents to be composed of several or few topics and alpha_H which determines whether topics should be composed of several or few words. I leave these hyperparameters at their default values because I already chose the number of topics and I am agnostic towards how many words each topic will contain.

After running the LDA using the document feature matrix from foreign policy-related texts, I converted the results of the gamma matrix into a tidy dataframe and merged it with the original filtered dataframe containing information on foreign policy related text, year, and country. From here, I divided the data into two periods – Cold War and Post-Cold War. Note that I dropped the party variable since it is not a significant angle in my analysis.

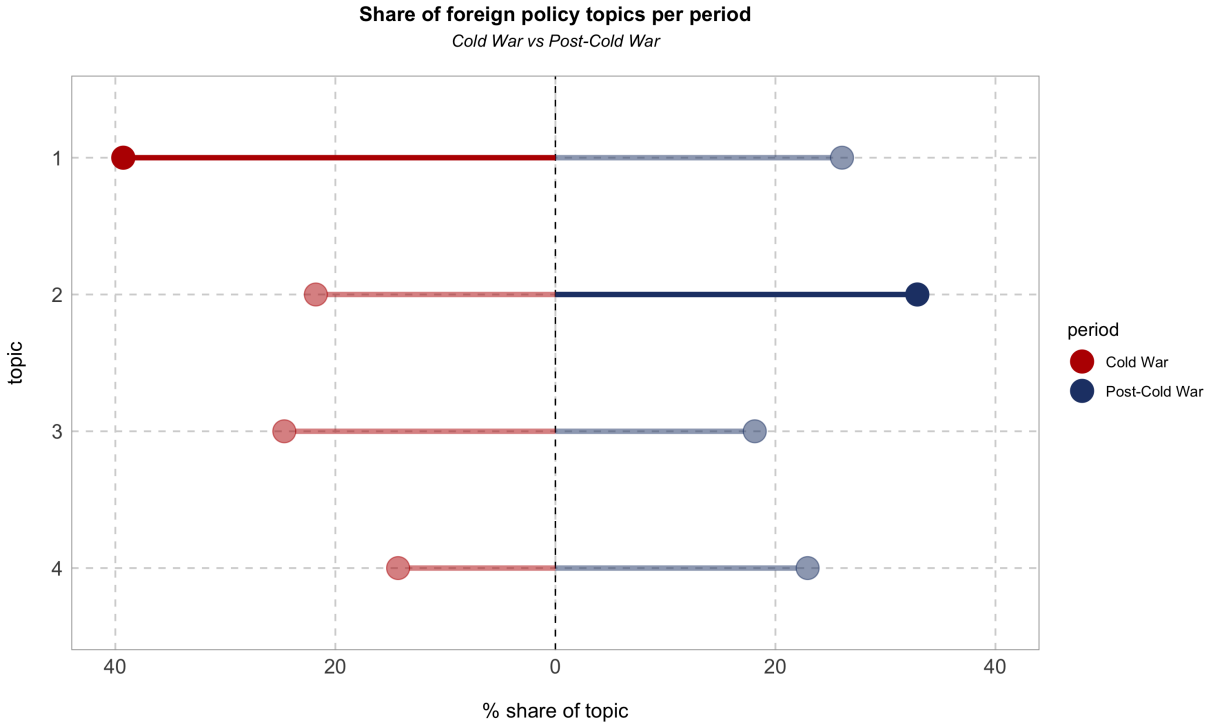
```
## # A tibble: 6 x 7
```

```
##   document_id  topic   gamma edate      country         year period
##   <chr>        <int>   <dbl> <date>     <chr>          <dbl> <chr>
## 1 51320_196410     1 0.00315 1964-10-15 United Kingdom  1964 Cold War
## 2 51320_196410     1 0.00315 1964-10-15 United Kingdom  1964 Cold War
## 3 51320_196410     1 0.00469 1964-10-15 United Kingdom  1964 Cold War
## 4 51320_196410     1 0.00469 1964-10-15 United Kingdom  1964 Cold War
## 5 51420_196410     1 0.00500 1964-10-15 United Kingdom  1964 Cold War
## 6 51620_196410     1 0.986   1964-10-15 United Kingdom  1964 Cold War
```

This shows the number of document-topic pairs per period:

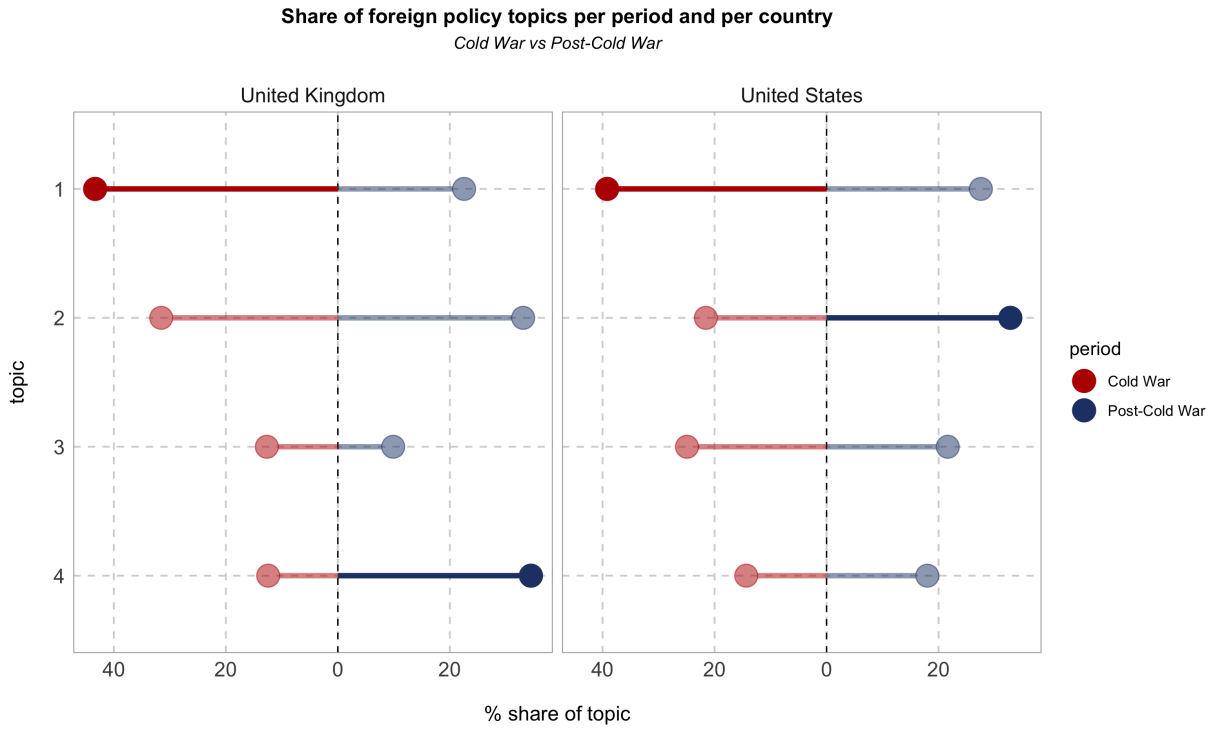
```
## # A tibble: 2 x 2
##   period        counts
##   <chr>          <int>
## 1 Cold War       47888
## 2 Post-Cold War  36800
```

From here, I calculated the share of topics per time period, not distinguishing between countries. Results are shown in the plot below.



This chart shows the percentage share of each topic based on the sum of gamma probabilities. We see that the single most prominent topic during the Cold War is topic 1 and post-Cold War is topic 2.

I construct a similar chart to see if the same patterns can be observed for UK and US separately.

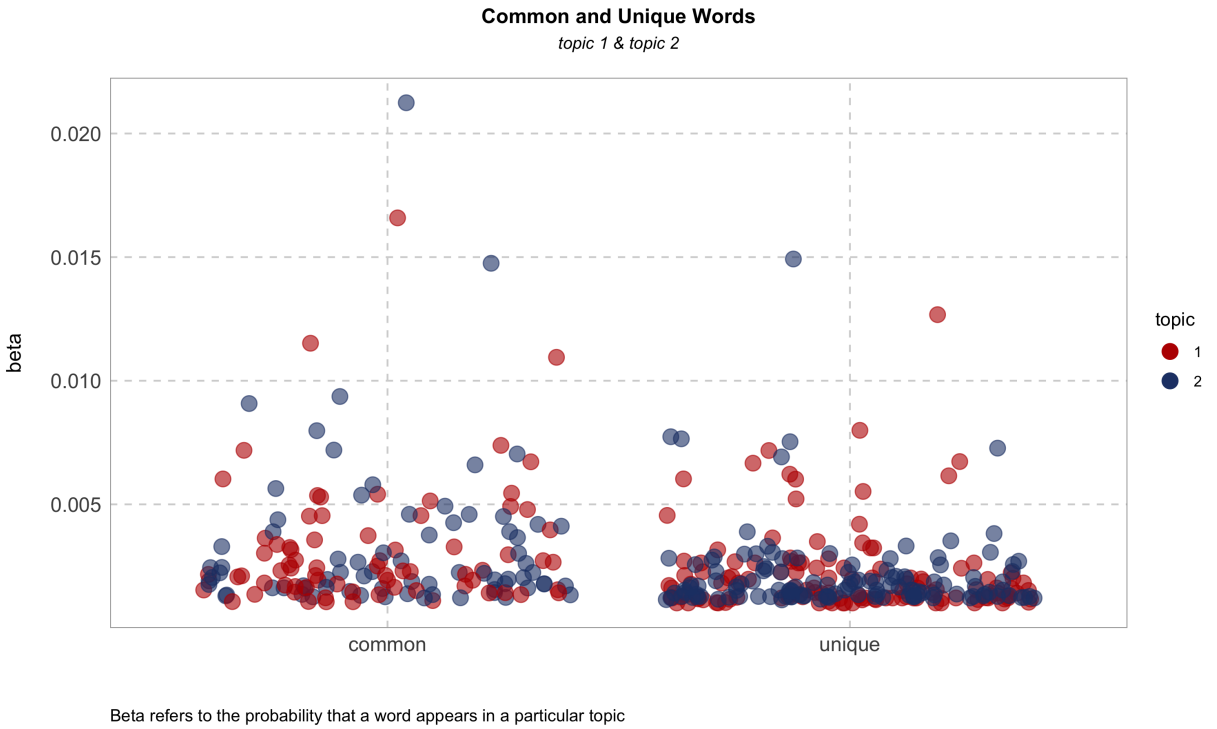


For both countries, topic 1 the most prominent topic during the Cold War. Post-Cold War, topic 4 dominated in the UK, with topic 2 as a very close second, while topic 2 dominated in the US.

Based on this, we assume that we can associate topic 1 with foreign policies during the Cold War and topic 2 with foreign policies after the Cold War. I continue the analysis by looking at the beta or topic-term matrix. This time, I focus on what words make up topics 1 & 2. I extract the top 200 words with the highest probability of occurence per topic.

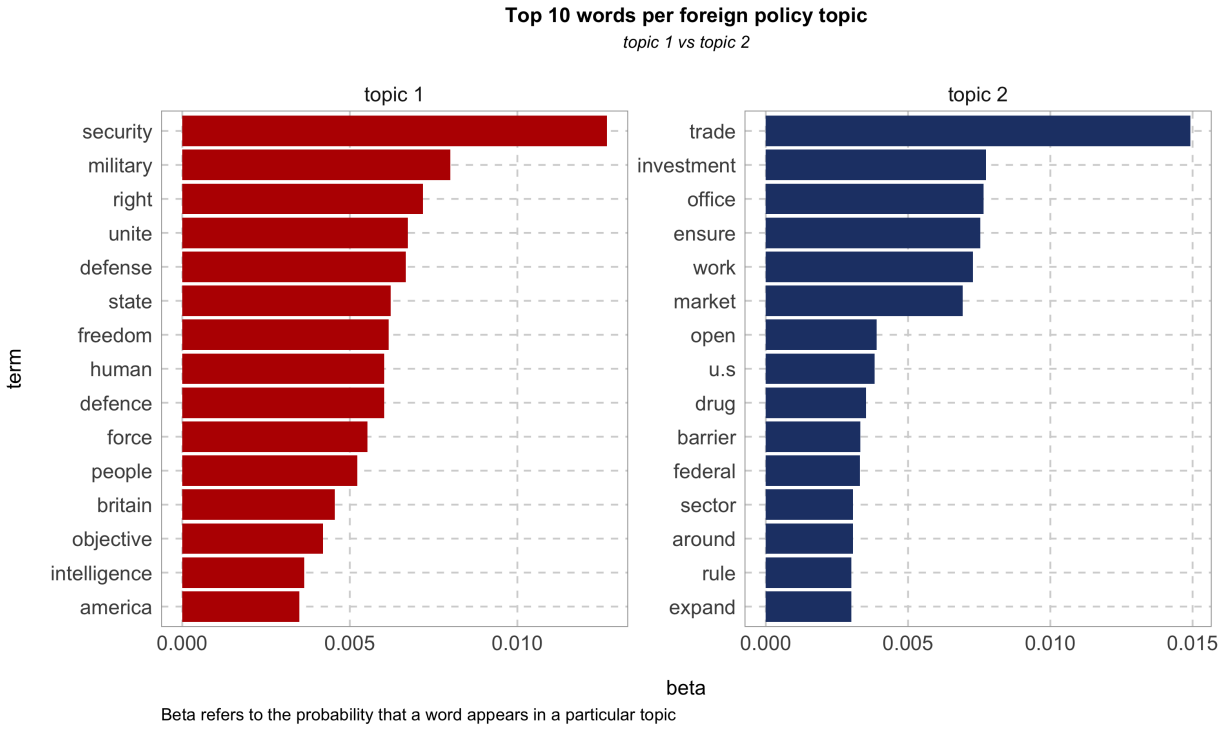
```
## # A tibble: 404 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 foreign    0.0746
##  2     1 policy     0.0420
##  3     1 will       0.0195
##  4     1 american   0.00903
##  5     1 must       0.00897
##  6     1 national   0.00881
##  7     1 assistance 0.00881
##  8     1 world      0.00865
##  9     1 country    0.00805
## 10     1 security   0.00730
## # ... with 394 more rows
```

I then categorize these words by whether they are common or unique to each topic. I do this as a rough check on how similar topics 1 and 2 are. I then visualize the words into a jitter plot which shows whether there are more common or unique words.



We see from the figure that there are less common words as the cluster on the left is more dispersed. Using this information, I assume that topics 1 and 2 are more or less different and can be associated with two different time periods.

As the next step, I take the top 15 words for each topic and visualize them in the plot below.

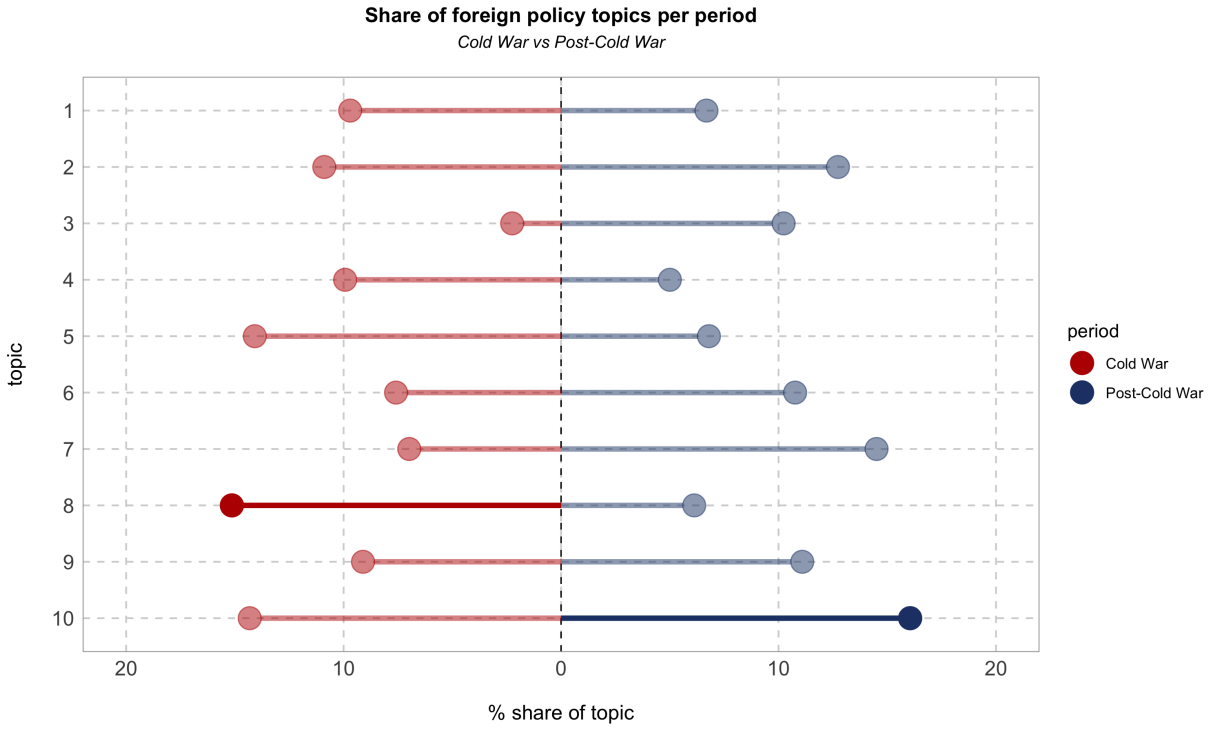


The words under topic 1 appear to be roughly about **national security**, with terms like security, military,

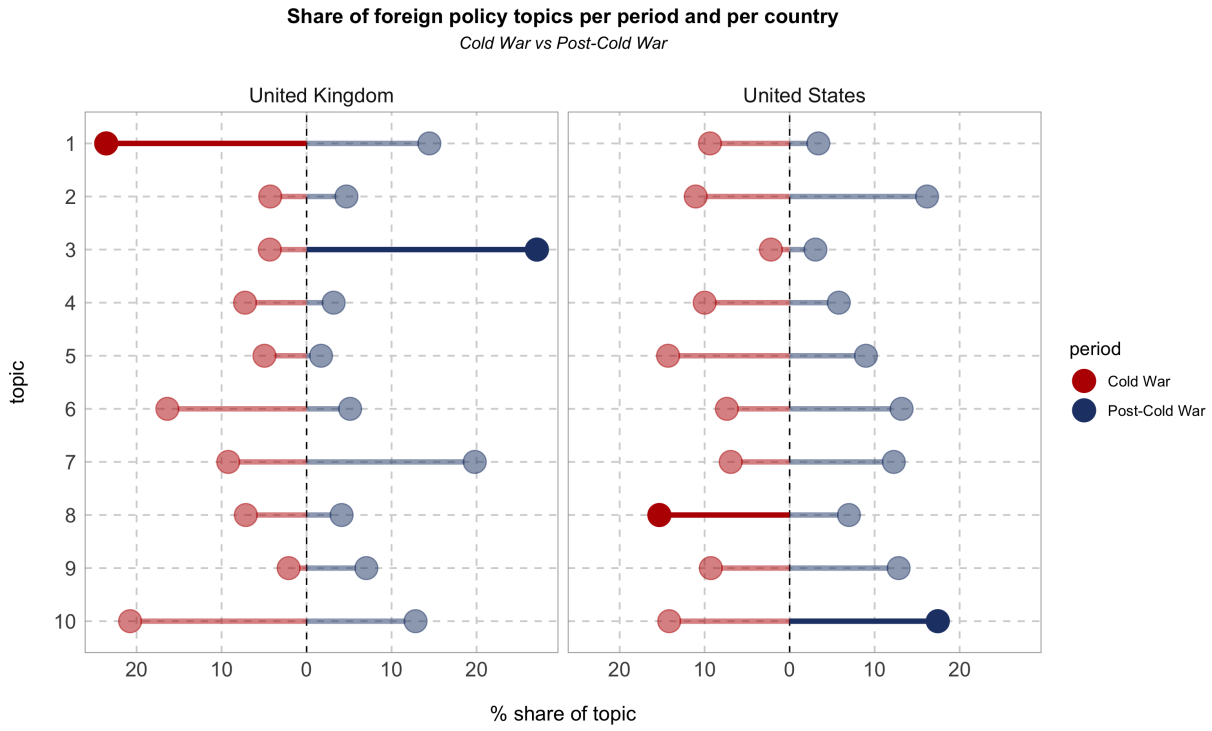
defense/defence, and intelligence. Meanwhile, the words under topic 2 are related to market **openness**. We see terms like trade, investment, market, open, and expand. With these results we can already see that there is indeed a difference between the foreign policies of US and UK during and after the Cold War.

LDA with ten topics

I conduct the same analysis above, this time setting 10 as the fixed number of topics. This is to check whether there will be significant changes in results when we ask the model for topics with a narrower scope.

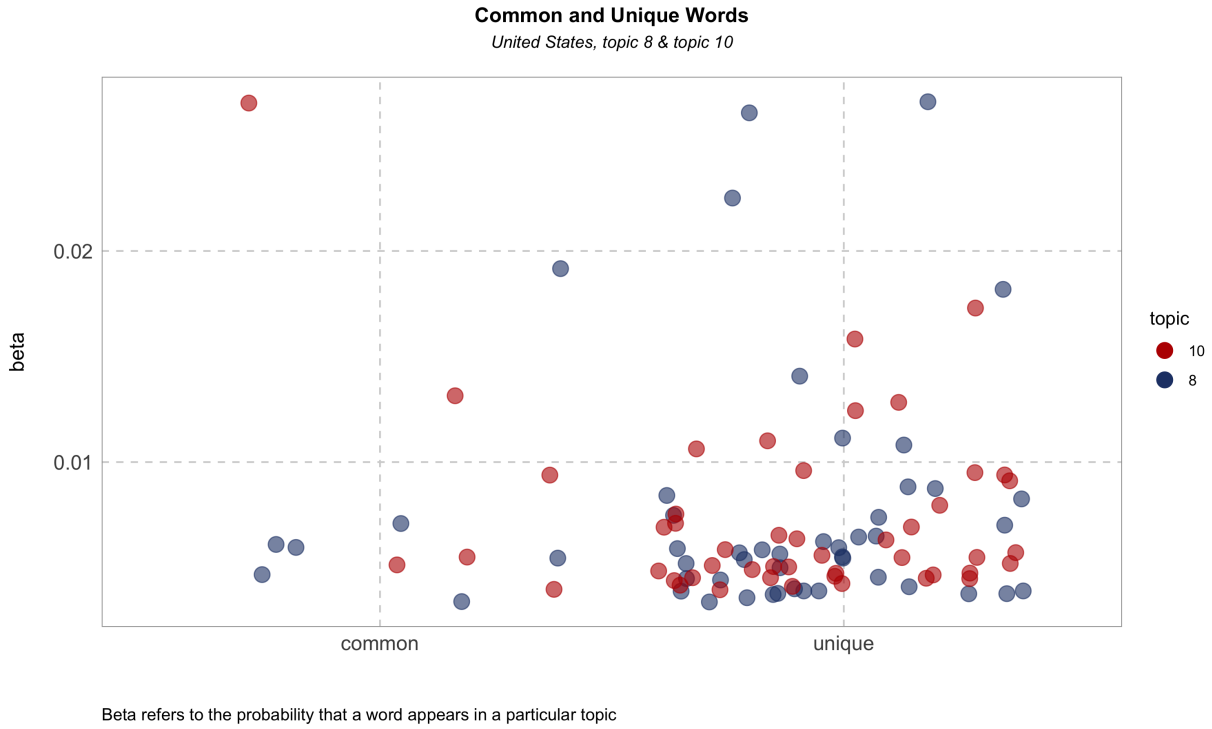


We see that topic 8 is prominent during the Cold War, with topic 5 as a close second. After the Cold War, topic 10 was more prominent.

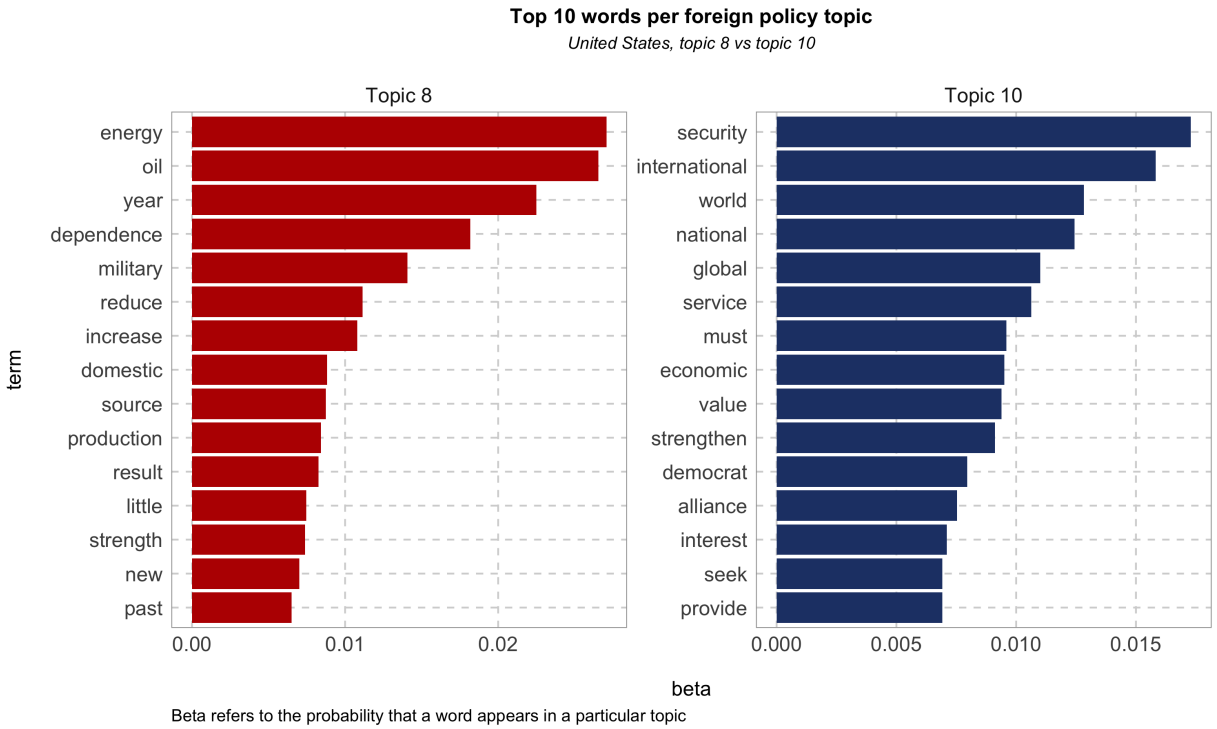


When more topics are used, the differences between UK and US are more pronounced. Topic 1 has the highest probability of appearing in UK documents during the Cold War and topic 3 after the war. In the US, it's topic 8 and 10, respectively.

I look at the difference between topic 8 and 10 as the Cold War for the US. Compared to the the topics generated using in the 4-topic LDA, topics 8 and 10 have less words in common.

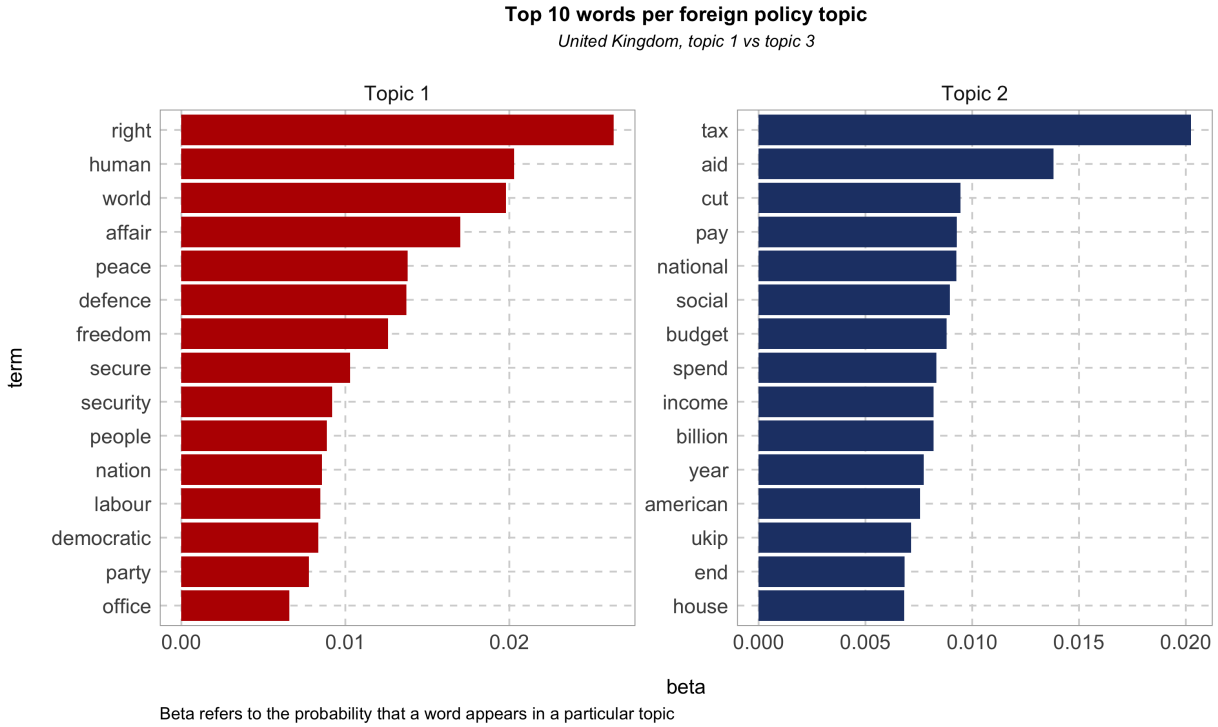


Next, I check the common words for topics 8 and 10.



We see new words that were not in the 4-topic LDA. This shows another foreign policy consideration of the US during the Cold War: **reducing dependence on natural resources**. We see this from the words energy, oil, and source. Post-Cold War, the main consideration appears to be **global security and strengthening international relationships**. We find words such as security, world, economic, and alliance.

The results for UK show different foreign policy priorities. During the Cold War, it appears they focused more on **human rights and world peace**. After the war, focus shifted to **tax cuts and national spending**.



One noticeable difference between the 4-topic LDA and 10-topic LDA is that we are able to form more concrete ideas using the latter. When topics were, we found similar words. After asking the model to generate narrower topics, we actually found coherent phrases.

5. Answering your research question

Going back to the research question:
Was there a change in the foreign policy stance of the world's most influential democracies, United States and United Kingdom, during and after the Cold War?

I find that there was indeed a difference in the foreign policy priorities of the two countries during and after the Cold War. Between 1960-1991, national security, human rights and world peace, and dependence on natural resources were the main concerns. Beginning 1991, focus shifted to openness, global security, international relations, and economic activity. Further, I find that there is a difference between the foreign policy priorities of US and UK, even though they were close allies during the Cold War.

One limitation of this analysis (and using LDA) is the fact that the model's ability to generate meaningful insights is heavily dependent on a hyperparameter that is set by the researcher. Depending on the corpus, the hyperparameter may be highly sensitive - setting it small will generate completely different topics compared to setting it high. However, topic modeling is still a very useful tool in approaching questions that require text analysis. It is a more scientific approach compared to perusing several documents manually.

As to my approach to answering the research question, one limitation is that I was only able to make comparisons across topics and not across words. This is because only the document-topic matrix can be matched to the original dataframe. Time-based comparisons per word was conducted based on the assumption that a certain topic belongs exclusively to a certain period. This is not that big an issue in this case since there were prominent topics for the chosen years, but in cases where the distinction is not as defined, this approach will be less accurate.

As an additional angle, statistical tests could be conducted on group averages (e.g. the probability of topic 1 appearing in period 1 vs period 2) to see if there is empirical evidence supporting the grouping.