NLP Assignment

Boladeres, Ian

2024-02-13

Loading packages and data

```
## Warning in .recacheSubclasses(def@className, def, env): undefined
subclass
## "ndiMatrix" of class "replValueSp"; definition not updated

## Package version: 3.3.1
## Unicode version: 13.0
## ICU version: 69.1

## Parallel computing: 12 of 12 threads used.

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

## Warning in .recacheSubclasses(def@className, def, env): undefined
subclass
## "ndiMatrix" of class "replValueSp"; definition not updated

##
## Attaching package: 'readtext'

## The following object is masked from 'package:quanteda':
##
##     texts

## Loading required package: proxyC

##
## Attaching package: 'proxyC'

## The following object is masked from 'package:stats':
##
##     dist

##
## Attaching package: 'seededlda'

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

## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Code

1st I construct the corpus. The variable that contains the text is called “text”.

```
UK_2019_corpus <- corpus(UK_2019, text_field = "text")
```

Then, I tokenize the text. The first part of the code deletes punctuation, numbers and symbols that do not increase the explanatory power of our model. Next, I compound names that might appear such as Great Britain, climate change or the European Union. Finally, the last part of the code deletes common words used in English and I decided to remove padding, deleting the empty spaces, so are not counted, at the expense of altering the length of the original text, but I believe that keeping the original length of the text it is not important for our analysis.

```
UK_2019_tokens <- UK_2019_corpus %>%
  tokens(remove_punct = TRUE, remove_numbers = TRUE, remove_symbol = TRUE) %>%
  tokens_compound(pattern = phrase(c('United States', 'United Kingdom','European Union','European Commission','Great Britain','climate change','Hong Kong'))) %>%
  tokens_select(pattern = stopwords("en"), selection = "remove", padding = FALSE)

UK_2019_dfm <-dfm(UK_2019_tokens)
```

I decided to divide the data frame by the most relevant political parties, the Conservatives, Labour, Scottish National Party and Liberal Democrats; leaving aside minority parties such as the Greens. This will be useful to use visual representation of the most important topics for each selected party.

```
UK_2019_dfm_party <- UK_2019_corpus %>%
  corpus_subset(party %in% c("Conservative","Labour","Scottish National Party","Liberal Democrats")) %>%
  tokens(remove_punct = TRUE, remove_numbers = TRUE, remove_symbol =
```

```
TRUE) %>%
  tokens_compound(pattern = phrase(c('United States', 'United
Kingdom','European Union','European Commission','Great Britain','climate
change','Hong Kong'))) %>%
  tokens_select(pattern = stopwords("en"), selection = "remove", padding
= FALSE) %>%
  dfm() %>%
  dfm_group(groups = party)
UK_2019_dfm_party <- dfm_remove(UK_2019_dfm_party,
c("s","sir","gentleman","hon","lady"))

nfeat(UK_2019_dfm)

## [1] 44433
```

44432 tokens seems enough tokens to analyse.

```
topfeatures(UK_2019_dfm, 40)

##         hon  government      people           s         can       right
house
##       49317       37312       33195       31765       27433       26321
25741
##    minister      friend        deal      member         one          us
work
##       25712       21689       19103       17432       16709       15741
15212
##        need        time        many     support          uk        also
make
##       14781       14426       14382       13324       13261       13115
12866
##   secretary     country     members        said        want         now
just
##       12683       12468       12313       12289       12170       12054
11835
##       prime        know         way        made   important       state
say
##       11604       11482       11479       11334       11072       11041
10799
##       point         get       years          eu   gentleman
##       10708       10687       10675       10451       10372

UK_2019_dfm <- dfm_remove(UK_2019_dfm, c("s","sir","gentleman","hon"))
```

The “topfeatures” helps to assess if we left out some expressions that should be compounded. I realise that “s” appears 31765, probably a result of deleting symbols and specifically the “Apostrophe”. This shouldn’t be a problem for our analysis, but I decided to delete the “s” to reduce the corpus size. In the next part of the code, I realized that some other common expressions such as “sir”, “gentleman”, “hon” or “lady” appeared a lot in the text and are parliamentary formality. Therefore, I decided to delete them.

The next step is to prepare our frequency analysis. I decided to keep words that appear at least in the 80% and delete words that appear too much.

Table 1

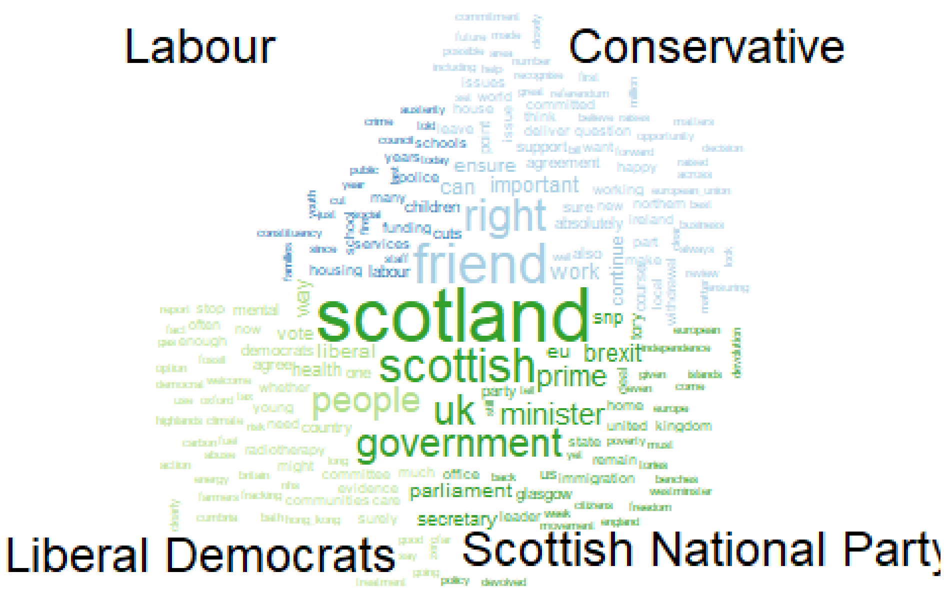
```
UK_2019_dfm_freq<-dfm_trim(UK_2019_dfm, min_termfreq = 0.8,
max_termfreq=0.99, termfreq_type = "quantile")
frequency <- textstat_frequency(UK_2019_dfm_freq, n = 50)
head(frequency, 50)
```

```
##           feature frequency rank docfreq group
## 1             key      1847    1    1586   all
## 2          agreed      1843    2    1584   all
## 3          energy      1842    3    1205   all
## 4         friends      1832    4    1540   all
## 5          forces      1822    5    1065   all
## 6           raise      1821    6    1611   all
## 7        backstop      1817    7    1034   all
## 8          living      1815    8    1408   all
## 9  climate_change      1815    8    1029   all
## 10     importance      1811   10    1586   all
## 11          wrong      1810   11    1537   all
## 12      yesterday      1810   11    1556   all
## 13        discuss      1807   13    1687   all
## 14          short      1802   14    1553   all
## 15          found      1800   15    1526   all
## 16           four      1798   16    1527   all
## 17            met      1791   17    1565   all
## 18         called      1783   18    1545   all
## 19        control      1779   19    1350   all
## 20         needed      1778   20    1552   all
## 21           lost      1776   21    1455   all
## 22          words      1776   21    1506   all
## 23       everyone      1773   23    1523   all
## 24     experience      1770   24    1438   all
## 25           came      1770   24    1561   all
## 26          plans      1762   26    1475   all
## 27           open      1758   27    1501   all
## 28         safety      1755   28    1177   all
## 29         follow      1753   29    1627   all
## 30        matters      1752   30    1526   all
## 31     assessment      1746   31    1477   all
## 32     everything      1738   32    1587   all
## 33        chamber      1728   33    1476   all
## 34   relationship      1725   34    1355   all
## 35            due      1722   35    1534   all
## 36        improve      1718   36    1449   all
## 37        exactly      1707   37    1577   all
## 38        meeting      1706   38    1428   all
## 39         border      1699   39    1166   all
## 40      potential      1697   40    1443   all
```

```
## 41         powers      1696   41    1088    all
## 42       involved      1694   42    1475    all
## 43     chancellor      1693   43    1329    all
## 44       ensuring      1692   44    1530    all
## 45      following      1688   45    1538    all
## 46      proposals      1686   46    1312    all
## 47       campaign      1685   47    1348    all
## 48         effect      1682   48    1505    all
## 49         almost      1680   49    1492    all
## 50          wants      1675   50    1486    all
```

Figure 1

```
textplot_wordcloud(UK_2019_dfm_party, comparison = TRUE,labelcolor=TRUE,
max_words = 200)
```

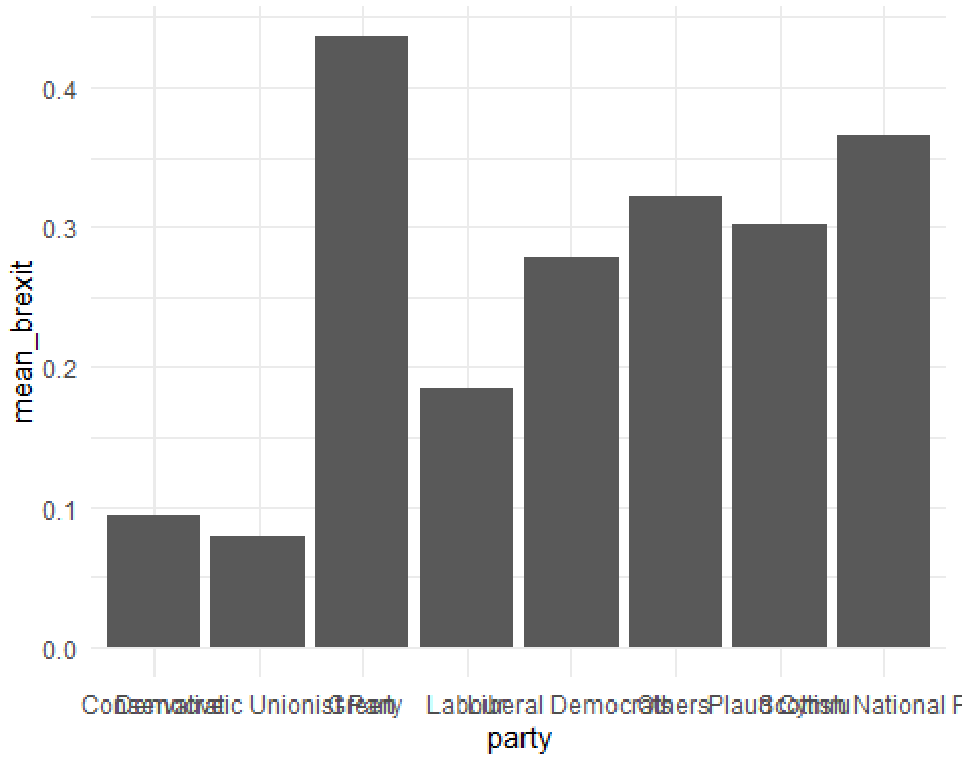


I would like to observe what parties use more times the word Brexit.

Figure 2

```
brexit <- UK_2019_tokens %>%
  tokens_select(pattern = c("brexit")) %>%
  dfm() %>%
  convert(to = "data.frame") %>%
  select(-c(doc_id)) %>%
  cbind(UK_2019) %>%
  group_by(party) %>%
  summarise(mean_brexit=mean(brexit))
ggplot(brexit) +
```

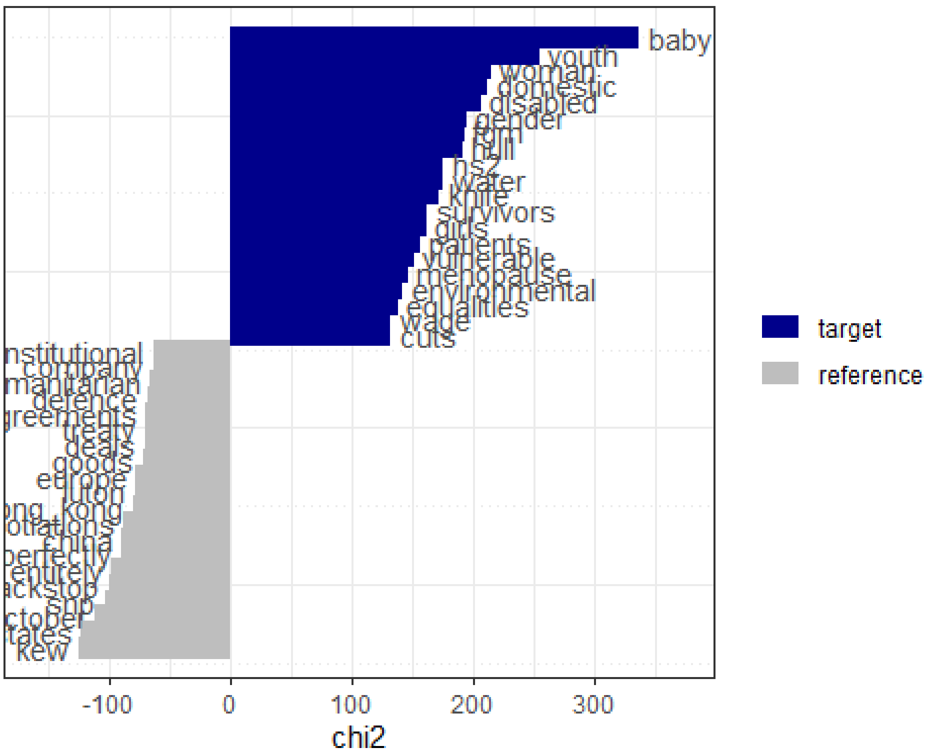
```
  geom_col(aes(party, mean_brexit)) +
  theme_minimal()
```



This graph will show the most recurrent topics by gender.

Figure 3

```
textstat <- textstat_keyness(UK_2019_dfm_freq, docvars(UK_2019_corpus,
"female") == "1")
textplot_keyness(textstat)
```



Building of the topic model. I decided that the model selected 20 topics.

Table 2

```
# topicmodel_UK <- textmodel_lda(UK_2019_dfm_freq, k = 20)
# terms(topicmodel_UK, 20)
```

I had some problems with the printing of the topicmodel_UK

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	online	cuts	supply	treasury	box	project
2	media	budget	tariffs	company	dispatch	city
3	church	councils	farmers	contracts	tomorrow	planning
4	faith	cut	agreements	charge	yesterday	rail
5	freedom	students	steel	hmrc	chancellor	road
6	religious	provision	manufacturing	contract	monday	network
7	stand	per	goods	bank	october	town
8	words	special	products	rates	debates	infrastructure
9	christians	youth	event	paid	discuss	towns
10	white	teachers	potential	costs	raise	line
11	racism	primary	export	buy	deputy	post
12	holocaust	pupils	trading	schemes	shall	rural
13	hate	extra	europe	property	sitting	centre
14	join	quality	assessment	loan	select	bus
15	culture	resources	deals	affordable	cabinet	site
16	jewish	higher	welsh	vat	standing	hs2
17	belief	nursery	exports	insurance	date	projects
18	persecution	maintained	company	rent	raises	greater
19	black	university	industries	value	soon	growth
20	antisemitism	educational	sectors	payments	meeting	residents
	Racism and antisemitism	**Public services: Budget in education and health**	**Trade**	**Fiscal and financial system**	**Parliamentary jargon**	**Intraestructure, transport and development**

Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
immigration	aid	climate_change	backstop	statutory	hospital
courts	hong_kong	energy	border	powers	medical
status	iran	carbon	customs	assembly	patients
commission	conflict	climate	negotiations	devolved	treatment
data	un	emissions	relationship	executive	cancer
judgment	president	zero	declaration	clause	research
electoral	syria	water	alternative	amendments	hospitals
apply	yemen	global	talks	instrument	constituent
individual	nations	target	irish	scrutiny	autism
person	sanctions	environmental	agreed	civil	healthcare
rules	peace	air	proposals	exit	conditions
individuals	humanitarian	net	secure	regulation	condition
application	china	targets	compromise	devolution	patient
supreme	united_states	emergency	negotiate	provisions	drug
inquiry	regime	green	negotiated	draft	clinical
asylum	region	technology	negotiating	matters	nice
migration	partners	industrial	friday	lords	brain
wrong	saudi	clean	certainty	framework	drugs
applications	united	electric	binding	legislative	disease
circumstances	refugees	reduce	deals	required	doctors
Migration	**International Relations**	**Climate change**	**Ireland**	**Parliamentary jargon**	**healthcare**

Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
democracy	officers	universal	got	fire	domestic	john	forces
votes	prison	living	t	safety	experience	football	defence
snp	animal	pension	lost	recommendations	sex	pleasure	armed
tonight	knife	employment	wrong	consultation	girls	sport	commonwealth
democratic	animals	bbc	tory	commission	relationships	chamber	royal
voting	criminal	pensions	truth	inquiry	marriage	david	join
control	welfare	chancellor	bad	buildings	equality	remember	veterans
benches	tackle	paid	failed	published	baby	team	nuclear
front	policing	wage	crisis	grenfell	gender	spoke	opportunities
mps	behaviour	disabled	happened	cladding	employers	congratulate	skills
elections	youth	low	anything	safe	age	welsh	charities
choice	prisons	income	else	body	guidance	chair	organisations
views	probation	rate	worse	residents	men	glasgow	personnel
decide	safe	age	seems	authority	civil	wonderful	challenges
elected	orders	assessment	came	expect	sexual	deputy	supporting
consensus	violent	minimum	reality	reports	everyone	excellent	nation
confidence	force	average	idea	quickly	lgbt	friends	ministry
wants	offenders	petition	failure	paper	woman	came	operation
commons	resources	policies	half	lessons	physical	proud	technology
constitutional	offences	payments	almost	account	experiences	worked	importance
Elections	**Law enforcement**	**Economic policies**	**Non-sense**	**Non-sense**	**Gender and LGBT**	**Sports**	**Defence**

Word embedding. I will build a new token now with padding TRUE to keep the original proportions of the text.

```
UK_2019_tokens_padding <-tokens(UK_2019_corpus, remove_punct = TRUE,
remove_numbers = TRUE, remove_symbol = TRUE) %>%
  tokens_compound(pattern = phrase(c('United States', 'United
Kingdom','European Union','European Commission','Great Britain','climate
change','Hong Kong'))) %>%
  tokens_select(pattern = stopwords("en"), selection = "remove", padding
= TRUE)
```

I will use the standard window of 6 words.

I will analyse the topic of climate change. For this reason, as keywords I will use "climate_change", "green", "emissions", "climate" and "ecology", a group of words that is closely related to the topic.

```
climate_tokens <- tokens_context(x= UK_2019_tokens_padding, pattern =
c("climate_change","green","emissions","climate"), window = 6L)

## 1021 instances of "climate" found.
## 24 instances of "Climate" found.
## 1403 instances of "climate_change" found.
## 38 instances of "Climate_change" found.
## 374 instances of "Climate_Change" found.
## 988 instances of "emissions" found.
## 7 instances of "Emissions" found.
## 596 instances of "green" found.
## 609 instances of "Green" found.

climate_dfm <-dfm(climate_tokens)
```

Now I build the co-occurrence matrix and the transformation matrix. I will use the pre-trained word embedding set.

```
UK_fcm <- fcm(UK_2019_tokens_padding, context = "window", window = 6,
count = "frequency",tri = FALSE)
transformation_UK <- compute_transform(x = UK_fcm, pre_trained = glove,
weighting = 500)
```

Next, I create the embedding matrix. I am specifically interested in how different parties relate to the different focal terms I selected.

```
UK_dem <- dem(climate_dfm, pre_trained = glove, transform = TRUE,
                transform_matrix = transformation_UK, verbose = TRUE)


UK_embeddings <- dem_group(UK_dem, groups =
                               UK_dem@docvars$party)
dim(UK_embeddings)

## [1]   8 300
```

Now I will find the nearest neighbours.

Table 3

```r
climate_nns <- nns(UK_embeddings, pre_trained = glove, N = 10, candidates
= UK_embeddings@features, as_list = FALSE)
climate_nns <- arrange(climate_nns, target, rank)
print(climate_nns)

## # A tibble: 80 × 4
##    target       feature    rank value
##    <fct>        <chr>     <int> <dbl>
##  1 Conservative next          1 0.582
##  2 Conservative global        2 0.565
##  3 Conservative change        3 0.545
##  4 Conservative way           4 0.532
##  5 Conservative climate       5 0.527
##  6 Conservative future        6 0.526
##  7 Conservative emissions     7 0.518
##  8 Conservative warming       8 0.517
##  9 Conservative make          9 0.514
## 10 Conservative put          10 0.512
## # ℹ 70 more rows

cosine2 = cos_sim(UK_embeddings, pre_trained = glove, features =
c("climate", "economy"), as_list = FALSE)
cosine2

##                        target feature      value
## 1                Conservative economy 0.44661377
## 2   Democratic Unionist Party economy 0.27035944
## 3                       Green economy 0.45535820
## 4                      Labour economy 0.42532574
## 5           Liberal Democrats economy 0.43556872
## 6                      Others economy 0.19300598
## 7                 Plaud Cymru economy 0.07338716
## 8     Scottish National Party economy 0.26972045
## 9                Conservative climate 0.52738938
## 10  Democratic Unionist Party climate 0.33507439
## 11                      Green climate 0.54557493
## 12                     Labour climate 0.65728003
## 13          Liberal Democrats climate 0.62302588
## 14                     Others climate 0.19306152
## 15                Plaud Cymru climate 0.17352237
## 16    Scottish National Party climate 0.47637731
```

Questions

1. What were the main topics under discussion in the British House of Commons in 2019?

Table 1 already shows some preliminary results on the main topics discussed in the House of Commons in 2019. In the top 10 most featured words, we find non-sensical words such as key, agreed or important. Those interesting to us are “forces” that might refer to armed forces or military forces, reflecting the growing international stability and the rising relevance of military forces. In the ninth position, we already find “climate change”. If we had data for previous years we might be able to visualize how the importance of this topic has been rising for the last decades.

The word cloud (Figure 1) gives us some clues about which were the relevant topics of discussion by the main political parties. For the Conservative Party is difficult to identify a clear topic, but some words connected to negotiations appear repeatedly, such as “agreement”, “committed”, “ensure” or “support”. Meanwhile, in the case of the Labour Party is clearer, as the words are oriented towards public services in general (public, funding, services), referring to some specific services such as education (schools, children), housing or law enforcement, as well as possible reference to the state of these services with words such as “cuts” or “austerity” that might reflect the worsening of British public services. In the case of the Liberal Democrats we find something similar, but with specific reference to the health system (NHS, health, radiotherapy, treatment) and a special interest in climate-related issues (climate, fossil). Finally, the Scottish National Party makes a lot of references to Scotland, reflecting their heavy regional implementation, and are from the parties selected, those that made more references to Brexit and the EU, perhaps because Brexit was a reality accepted by the rest of the parties but not in the case of the SNP. It is more understandable if we observe a map of the results of the Referendum, in which most of Scotland voted to “Remain” in great contrast to the rest of the UK. Part of the current strategy of the SNP is, in light of the results of the Brexit Referendum, to repeat the Scottish Referendum of independence, expecting that the desire for secession from the UK has increased, and rejoining the European Union. Figure 2 confirms my idea that one of the parties that used the most the word Brexit was the SNP, just behind the Green. Furthermore, another regional party, the Plaud Cymru also used recurrently the word Brexit.

Figure 3 divides the dataset by the gender of the member of the British House of Commons. We can observe a divergence in the topics that are more prominent among female (target) parliamentary members and their male counterparts (reference). Female members made more mentions of baby, youth, woman, disabled, and vulnerable; reflecting the gender roles of British society in which women are responsible for caring for others, to care for those vulnerable. Meanwhile, in the case of males, words related to international politics (negotiations, treaties, deals, Hong Kong, China, ship, United States) are more prominent. This graph is a great representation of the gender biases in politics and the difference between “soft topics” or “soft politics” related to social services; and “hard topics” or “hard politics” related to economy, international relations or defence.

Finally, the analysis of the topic model reveals the following:

Topics 16 and 17 are non-sensical and topics 5 and 11 are composed of parliamentary jargon. Topic 1 is about racism and antisemitism, a relevant topic in the UK, that has gained relevance thanks to the internet and the proliferation of xenophobic attitudes. Furthermore, the recurrent mentions of antisemitism might be due to the scandal with Labour candidate Jeremy Corbyn and whether he was or not an antisemite.

Topic 2 is linked to public services, specifically to education. From the wordcloud we know that one of the parties more vocal about this topic was the Labour Party, denouncing the cuts on the budget and the worsening state of the educational system in the UK. Similarly, topic 12 is about healthcare, being one of the most prominent parties speaking about the healthcare system the Liberal Democrats. To end the public services block, topic 14 is about law enforcement and policing.

All of these public services must be funded by a budget and topic 4 is about this, fiscal discussion and financial jargon. Topic 15 is also about economic policy in general, in specific to pensions, wages and employment. Topic 3 is about trade, making references to agriculture, the manufacturing sector and deals. This is relevant due to the Brexit and the negotiations with the European Union. Connected with this topic is the topic about Ireland (Topic 10), a candent topic in the negotiations with the EU, due to the problems that could arise if the UK decided to build a “strong border” with the Republic of Ireland and how the accession of UK to the EU helped to dissipate the tensions with the Irish population and the IRA. Some features of this topic make references to creating an agreement to solve the problem and reference to the history of anglo-british relations (the Friday, referring to the Good Friday Agreement).

Topic 8 is centred around international relations, although there is no clear focus, there are different features that connect to different subtopics such as relations with China, the UN, humanitarian policy or the Middle East. Topic 20 also refers to international relations specifically to the defence sector, nuclear power, the importance of technology to the military and the Commonwealth. Topic 7 is about immigration, in particular, it seems to deal with asylum seekers.

Topic 6 is about infrastructure, development and transport, the creation of development projects, road and railway expansions. Topic 13 seems to be about elections and democracy, appearing jargon about normative elements of democracy, such as consensus, confidence or control (all features that we can link to positive elements of democracy). That the SNP is in this topic might reflect the strategy of the SNP to repeat the Scottish Referendum of Independence. Topic 19 seems to be about sports and football, but it is not clear.

Finally, topic 18 is about gender and LGBT, with references to sexual and domestic violence, and equality. Topic 9 is about climate change and the green transition for green economic development, specifically to reduce emissions and find alternatives to the energy system.

2. Select one keyword (or a group of keywords) of one of the topics that you have identified in the previous question and examine the extent to which its usage varied across the political parties represented in Parliament

I focus on the topic of climate change. Table 2 gives the results of the embedding by parties. The two parties with features with closer affinity to my selected group of words are first the Liberal Democrats Party and the Greens. A closer look at how both parties treat the same topic, the Liberal Democrats emphasize the topic related to energy, being the closest related word to my word selection, with a value of 0.65. Also, resources are a relevant word in their discourses related to climate change, emphasizing the need and belief of the Liberal Democrats to transition from our current energy model to a greener one. Also the word global might refer to the necessity of cooperation between nations. In the case of the Greens, global is their first word, implying that the necessity for global cooperation is more important in the Green circles than in the Liberal Democrats. Moreover, the word crisis appears in the top 4, a word that does not appear in any other party. This may be an interesting feature of the Green narrative towards climate change, emphasizing the crisis it supposes.

The following two parties for which climate change is a relevant topic are the Labour Party and the Conservatives. Both speak in similar terms about climate change, although in the case of the Conservative Party more words about actions like “change” or “make”, which is a normal feature taking into account that it was the Party in office during 2019. A party in office speaks about the things can or not do, meanwhile, the parties in the opposition are more prone to speak about normative issues or how they would act. Interestingly, this is a feature shared by the SNP, that the most well-connected words are verbs, but in the end, they are also the ruling party in Scotland.

The Democratic Unionist and the Plaud Cymru are the two parties (besides others) that are less connected to the climate change issue, although not that much unconnected. The latter uses words connected to infrastructure projects and development, whereas in the case of the former appear a lot of verbs.