An analysis of self-referential questions and regional

topic differences in the 58th UK Parliament

Damian Elkins

2/2/2024

COMP3517

# 1 - Introduction

The aims of this essay are to give an overview of my approach to the questions posed to me as well as to critically assess both my results and methodology in the context of the questions. The first question tasked me with understanding and finding trends regarding self-referential questions asked in parliament in the first 9 months of 2023. As UK elections happen on a strictly local basis it is important to understand members of parliament’s commitment to the people that put them in office. Many of these constituencies may feel underrepresented in governmental decision-making and if they can understand how their MP is representing them, they can make more informed decisions. The second question was regarding regional trends in topics that appear in the questions asked. A regional understanding of topic distribution allows for intersectional analysis, ensuring that historically underrepresented regions such as the northeast, are having their worries addressed and treated as equal members of society as those in wealthy well represented regions such as London and the southeast. Through applying my methodology I have been able to develop a deeper understanding of both subjects and will further analyze their meanings within their respective sections.

# 2 - Self-Referential Questions

# 2.1 – Methodology

## 2.1.1 – Data Collection

My data collection process had five major phases. The first phase was the broad parliamentary query where I collected all questions and information about them. For this and all other queries I used the SPARQLWrapper module. In phase two I labelled all question texts using lists of places in the UK as well as NER from the SpaCy module. From here I started phase three where I would link all questions to the constituency the PM represents. In phase four I used the Nominatim geocoder from Geopy to gather the coordinates of all places mentioned in the questions with georeferencing. Finally, in phase five I determined which questions contained places within a constituency by using the Shapely module, a GIS of the constituencies, and the coordinates from phase four.

## 2.1.2 – Implementation and Models

In the first phase described above, I gathered the question URI, the question number, the URI of the person asking the question, the text of the question, the date the question was asked, and the MNIS ID of the person asking the question. The MNIS ID is a unique identifier which allows me to link between the parliamentary database and wikidata. This step is quite simple and only involved a single query.

In the second phase I queried wikidata for the names of any cities or constituencies in the UK as well as the names of all countries. This was intended to act as a dumb filter to catch any obvious instances of place mentions. Following this I used the SpaCy module alongside some guidance from a stack overflow source (1) to label the text of the questions. SpaCy used “GPE” tag for geopolitical entities and “LOC” for locations.

In the third phase I linked the questions to the constituency the MP represents. I did this by first querying wikidata for the MNIS ID and name of the constituency represented by the person with that MNIS ID. I then used this as a reference to add the constituency information to the dictionary objects created in the second phase.

In the fourth phase I used a geocoder to gather the coordinates of the place names I found in the previous steps. I chose to use a geocoder instead of a wikidata query because I felt it would be more tolerant and running significantly faster than performing hundreds of queries. I used the geocode Nominatim from the geopy module. Before feeding the place names to Nominatim, I had to process them to gather a list of unique place names from the data, so I am not performing any unnecessary queries. I used these coordinates to cull the list further so that there were only places that fall within the coordinate bounds of the UK. This process was aided by an online tutorial (2).

Finally, in phase five I use a GIS in the geojson format alongside Shapely to determine if a question mentions a place in its own constituency. I chose to use a GIS as it is intuitive to understand as well as being useful for generating future visualizations. I use a GIS I downloaded from the National Institute for Health and Care Research and convert all coordinates into shapely Point objects which I can use in combination with the within() function to determine the constituency the point falls in. I was inspired by a stack overflow answer for this (3). With this information I was able to create a new list where all entries are self-referential questions.

## 2.1.3 – Critical Analysis of Models

The modelling process has two major sections which could cause issues. The first and most notable of these is the NER pipeline. The SpaCy module is very strong but also extremely generalized, I was unable to find an appropriate English language model trained to tag UK geographical locations. I found in my implementation the NER alone was not adequate often missing data. I attempted to adjust for this by including the lists of place names queried from wikidata. I believe my final accuracy in this respect to be at most 60%. This is then compounded by the georeferencing I did to gather place coordinates. This is much harder to estimate my accuracy, so I referenced Milan Gritta’s 2019 PhD thesis on the topic. (4) The strongest performers in the NER geotagging field only reached a maximum F-score of 0.71 on one dataset and an accuracy of 0.66 on another. The best performing models took up to 2 weeks to run and as such I believe my score to be stronger than the worst performance. I believe that more recent advancements will allow me to be closer to those highest performers. In reference to geocoding they use a metric of accuracy at 161km indicating the percent of encodings that were within 161km of their target. Using this as a reference where the best had 76% accuracy, I believe my accuracy to be somewhere around 70%. This is due to the geocoder I used being a lightweight model intended for less strenuous use cases. This places my model’s at representing roughly 42% of the total number of self-references.

## 2.2 – Results and Conclusions

With my model I was able to gather 871 self-referential questions. This represents 3.25% of all total questions asked.

A map of england with red and white colors

Description automatically generatedA map of united kingdom with blue and white colors

Description automatically generatedA map of united kingdom with green lines

Description automatically generated

Across these three maps it is difficult to establish any specific trends. The maps are very spotty which I believe can be explained by the vast number of constituencies that asked very few questions if any at all. 440 of 650 constituencies asked fewer than 20 questions with 294 of those asking none. Those constituencies with the highest ratio tend to ask much fewer questions (quiet constituencies). Of the 50 constituencies with the highest ratio 27 asked a total of fewer than 30 questions with the top 10 containing no constituency that asked more than 27 total questions. It is tempting to say that this indicates that small constituencies are more likely to ask self-referential questions. Of the constituencies that asked fewer than 30 questions (quiet constituencies) and more than 0 questions only 48 asked a self-referential question, representing 24.5% of the total quiet constituencies. This is opposed to the 52% of constituencies that asked more than 30 questions (talkative constituencies) which asked a self-referential question. It is possible that the model I have built is at fault. Quiet constituencies tend to be less well known and as such NER might have a more difficult time recognizing places within them.

The mean ratio of the talkative constituencies was 5.8% while it was 22.1% for the quiet ones. While fewer quiet constituencies ask self-referential ones, those that do are much more likely to ask self-referential questions. When looking at the graphs of these ratios they are both quite strongly skewed but the quiet constituencies are especially skewed. This makes sense as it is much easier to get a higher ratio when you ask fewer questions.

A graph with a line

Description automatically generatedA graph of a number of points

Description automatically generated with medium confidence

A map of united kingdom with green lines

Description automatically generatedA map of england with red and white colors

Description automatically generated It is difficult to draw any broad conclusions about the quiet constituencies as they have such sensitivity to small changes from the model. This is apart from a few standouts in the quieter constituencies, those being Bridgwater and West Somerset who asked the 10th most self-referential questions with only 27 total questions asked and Basildon and Billericay who asked the 14th most self-referential questions with only 18 total questions. Both constituencies are strongly conservative and southern. Indicating that of these quiet constituencies, which are spread across the UK, the southern constituencies are more focused on internal affairs. This is supported by looking at the maps below. The only areas that have a high ratio and a high count in the self-map are in the south. The same is true for the constituencies that are somewhat talkative such as Shrewsbury and Atcham in the west midlands which has yet again a conservative MP. The same is not true however for the more talkative constituencies with an above average self-reference rate such as Newcastle Upon Tyne Central, Battersea, and York Central, all of which are urban and strongly Labour constituencies. It appears that the quieter constituencies that are more self-referential tend to be rural and Conservative while the most talkative that are self-referential tend to be much more often urban areas with Labour leaders.

# 3 – Regional Topic Differences

# 3.1 – Methodology

## 3.1.1 – Data Collection

My data collection process is very similar to my process in 2.1.1 with a few changes. I start by querying parliament for all questions with SPARQLWrapper. I then get the constituency each question asker represents using a wikidata query. From there I gather all the coordinates of the constituencies with another wikidata query. I then use geopandas to gather the region each question is from which will then go on to be processed with genism, nltk, and pyLDAvis to generate and visualize an LDA model.

## 3.1.2 – Implementation and Models

My model starts by querying the questions and assigning them constituencies in the same way as 2.1.2. From this point I query wikidata again and gather a list of coordinates for each constituency that I gathered in the previous step. This involves some data cleaning to simplify the dataframe generated and to store the wikidata response in a usable way. Specifically, wikidata returns coordinates as Point(longitude, latitude) which I turned into lists to be compatible with geopandas. I used a GIS in combination with geopandas to determine which constituencies fall within which region. I used this approach as I felt the specificity of the labels from the P131 property were too varying some being specific boroughs of London and others being vast areas of land. Using a GIS simplified this allowing me to simplify the process and remove the need to merge regions together. I used a dataset from National Institute for Health and Care Research again to gather regional boundaries and performed a spatial join. This is a function in geopandas that allows you to easily merge a GIS and a dataframe on coordinates and GIS geometry. For areas outside of England I assigned them regions by which country they are in once again using a GIS from data.gov.uk. From here I began generating LDA models using the genism and nltk package. I did this following a tutorial from Towards Data Science (5). Each LDA modelling process started with preprocessing text which involved using regular expressions to remove any punctuation or remaining tags such as <p> from the text. From there I generated stop words using the nltk package as well as extending them with words I would see commonly appear in the LDA modelling that were irrelevant to discussion. Next, I created a dictionary as well as a usable corpus. This was then used to build an LDA model of the text which I passed on to the pyLDAvis package which creates easy to use, interactive visualizations. I did this and generated a separate LDA model and visualization for each region allowing me to understand the differences between them easily.

## 3.1.2 – Critical Analysis of Models

This model design is quite straightforward and as such only has a few areas to criticize. First is the decision to have such large regional groupings. Massive regions such as the Northeast would be dominated by a few constituencies that asked many questions as is seen in section 2.2. This means that much of the topics will be much more relevant to people living in those constituencies which tend to be urban environments. There is also the merging of Northern Ireland, Scotland, and Wales as individual regions. This must be done for the sake of having enough training data for the LDA but serves to lessen the diversity of topics across each of the countries. Finally there is the LDA results themselves. When training an LDA it is difficult to determine how many topics to use. In my modelling I may have used to few topics causing some smaller ones to be merged into one, or I may have chosen too many causing larger topics to split into smaller ones appearing less significant. I have tried to get the best parameters for this, but they may still be imperfect. There is also the human error of my own interpretation. With LDA models, it can occasionally be difficult to distinguish between topics. I have done my best to search through multiple models of the same corpus to get the best idea of the topics, however there may still be errors.

## 3.2 – Results and Conclusions

My LDA visualization method creates interactive HTML files, as such I will interpret them into a table below and refer to the visualizations as needed. In viewing my results I found that there were often about 4 topics most important to a region with a few being less important and as such left out of the table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Region | First topic | Second Topic | Third Topic | Fourth topic |
| South west | Healthcare | Defense | Energy | Environment |
| london | Healthcare | Workers’ Rights | International Affairs | Energy |
| south east | Healthcare | International Affairs | Education | Energy |
| East of england | Healthcare | Energy | Environment | International Affairs |
| west midlands | Government Spending | Healthcare | International Affairs | Energy |
| east midlands | Workers’ Rights | Healthcare | Culture | Education |
| North west | Healthcare | Energy | International Affairs | Environment |
| Yorkshire and the Humber | Health | Defense | Energy | Local Economy |
| north east | Education | Healthcare | Defense | Construction |
| scotland | Commonwealth Affairs | Healthcare | Energy | Workers’ Rights |
| wales | Energy | Healthcare | International Affairs | Industry |
| northern ireland | Healthcare | Commonwealth Affairs | Environment | Trade |

There are 2 obvious trends within the data, those being the extreme prevalence of healthcare and energy discussion. The healthcare trend can be linked to the government's current situation regarding the NHS budget across the coming years. The NHS budget increase is planning to be curbed as over recent years the budget has been increasing at a higher rate than the current government is happy with. This has been compounded by the Covid-19 pandemic causing an additional £88 billon in emergency spending. The current government plan aims to cut budget increases from an average of 2.8% to 0.1% across the coming years. (6) Energy is featured as a prominent topic across almost all regions. I believe this has two causes, first of which is the rising cost of living crisis. (7) The second cause was the developments with the Net Zero strategy in 2023. (8)

There are also various regional trends. Most notably is the relationship between the commonwealth and Scotland and Northern Ireland. This is understandable considering the longstanding sentiment of independence shared by the two countries. (9) Tensions were especially high in 2023 for Scotland as the previously announced plans for an independence referendum were halted by the supreme court. The prevalence of the Scottish National Party in the current parliament also likely causes an increase in this independence sentiment, a core belief of the party especially in a post Brexit UK. (10) Amongst the regions it is rare for Healthcare to be pushed out of the largest topic evidencing how prominent this topic is among the Scottish MPs.

Perhaps the most shocking regional trend is how prevalent education was in the North East. It was one of two regions to include education as a common topic and the only one to have it in the top place. I believe that this is due to a combination of factors. First are the historical inequalities between the north and the south have led to northern schools and students being comparatively disadvantaged. This has been compounded by the Covid-19 pandemic. Recent government spending has also been funneled away from northern schools. Second is the prevalence of Labour MPs in the north east who are more supportive of funding education historically than conservatives have been in recent years. (11)

There was also a prominence of workers’ rights in the east midlands and London. London makes sense as such a strongly Labour voting area would have worker’s rights at the core of its concerns. When inspecting the LDA from the east midlands I noticed how commonly pensions were mentioned. I was unable to discover any trend in the east midlands that would lead to this massive spike in pensions mention.

Much of the discussion was on the topic of international affairs. This mainly regarded asylum seekers in the UK. This was most important to the South East which has the lowest percentage of asylum seekers among migrants to the region with just 2% as you can see below. It is a similar case with the West Midlands who take in 5%. Asylum seekers are a less common topic in the regions that have the highest percent of asylum-seeking migrants such as Wales and the North East which have 15% and 14% respectively and Wales has international affairs as its third most common topic, and it doesn’t make the list in the North East. It seems that most migrant discussion is exclusionary because of this.

(12)A graph of percentages and numbers

Description automatically generated

# Bibliography

1. **(https://stackoverflow.com/users/6082275/gdaras), gdaras.** *StackOverflow.* [Online] October 7, 2018. https://stackoverflow.com/a/52688313.

2. **Panigrahi, Kiran Kumar.** How to get the longitude and latitude of a city using Python? *tutorialspoint.* [Online] August 29, 2023. https://www.tutorialspoint.com/how-to-get-the-longitude-and-latitude-of-a-city-using-python.

3. **(https://stackoverflow.com/users/382489/zebs), Zebs.** *StackOverflow.* [Online] December 27, 2013. https://stackoverflow.com/a/20793808.

4. *Where are you talking about? Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring.* **Gritta, Milan.** s.l. : Apollo - University of Cambridge Repository, 2019.

5. **Kapadia, Sashank.** Topic Modeling in Python: Latent Dirichlet Allocation (LDA). *Towards Data Science.* [Online] 2019. https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

6. **The King's Fund.** The NHS budget and how it has changed. *The King's Fund.* [Online] September 20, 2023. https://www.kingsfund.org.uk/insight-and-analysis/data-and-charts/nhs-budget-nutshell.

7. **National Energy Action.** Timeline of the Energy Crisis. [Online] January 2024. https://www.nea.org.uk/energy-crisis/energy-crisis-timeline/#:~:text=The%20Energy%20Crisis%20has%20been%20driven%20by%20wholesale%20gas%20price,cost-of-living%20crisis..

8. **Burnett, Nuala, Edwards, Tamsin and Watson, Nicole.** The UK’s plans and progress to reach net zero by 2050. *House of Commons Library.* [Online] November 14, 2023. https://commonslibrary.parliament.uk/research-briefings/cbp-9888/.

9. **Hastings, Max.** Northern Ireland and Scotland Are on a Long Road Out of the UK. *The Washington Post.* [Online] February 27, 2023. https://www.washingtonpost.com/business/northern-ireland-and-scotland-are-on-a-long-road-out-of-the-uk/2023/02/24/7e83f914-b465-11ed-94a0-512954d75716\_story.html.

10. **Cochrane, Angus.** Scotland's Yes movement: Inspired or disillusioned? *BBC.* [Online] October 19, 2023. https://www.bbc.co.uk/news/uk-scotland-scotland-politics-66983211.

11. **Learning Cubs.** North-South Education Divide: Impact on Children in the North. [Online] January 23, 2023. https://www.learningcubs.co.uk/resources/north-south-education-divide-impact-on-children-in-the-north#:~:text=The%20North-South%20divide%20in%20education&text=The%20implications%20of%20this%20divide,their%20local%20school%20and%20area..

12. **The Migration Observatory.** Where do migrants live in the UK? [Online] March 24, 2022. https://migrationobservatory.ox.ac.uk/resources/briefings/where-do-migrants-live-in-the-uk/#:~:text=Asylum%20seekers%20awaiting%20a%20decision,of%20England%20and%20South%20West..