CMHSS Coursework

Joseph Dunne

Question 1

Introduction

To answer this question, I combined multiple sources of data, producing a table of MP questions with their relevant referenced constituency. Constructing this table allowed various statistics and visualisations to be produced to show to what extent different MPs ask questions related to their own constituency.

My method consisted of recognising entities from MP written questions with which I extracted their location and determined what constituency they belonged. I then stored and examined the data by visualising the number of questions from each constituency, and the proportion of which were related to itself. I also investigated the difference in Cabinet Ministers from their questions asked.

Question Data

To get MP question data I used the UK parliament API for written questions with a modified version of the given SPARQL query from this assignment. This modification was to exclude questions from members of the House of Lords. I called the SPARQL query using the SPARQLWrapper library, and the results were stored within a pandas dataframe.

Entity Recognition and Disambiguation

To determine whether questions related to a given constituency, I ran named entity recognition and disambiguation (NER+NED). To do this, I considered several methods and empirically determined their effectiveness for the given data. Due to my own familiarity, and to keep the code for this project consistent, I favoured approaches achievable within Python.

The first option I considered was TAGME, an API which was developed by researchers for efficient processing of short pieces of text. Despite it being utilized in multiple pieces of recent research, a paper comparing end-to-end entity linkers found it to be one of the lowest-performing (1).

Two other systems that are known to be better performing are WAT and REL (2) (3). I tested both of them on MP questions and found their extraction of entities to be suitable. REL, the more modern entity linker, had been shown to outperform WAT, however, because of errors occurring from the API and the duration of tagging, it deterred my selection of it. My chosen NER+NED technique would be applied to 27,000 questions, so processing would take a significant amount of time. For REL, this would likely be an order of magnitude larger. For the same reason, this deterred me from looking into Stanford's NER+NED tool, with the paper outlining WAT stating that Stanford's NER was "orders of magnitude slower" (2).

For both these reasons I chose WAT as the tool for NER+NED. Although, it is not state of the art, its quick computation made it feasible to process all of the questions within seven hours. The findings of this report are therefore limited by its quality and I would direct future research with greater resources to apply a more current approach.

MP Constituency Data

From the question data alone, it did not contain for a given question, the constituency of the MP asking it. To add this, I wrote a SPARQL query to Wikidata in which I selected all MPs from the 58th parliament of the UK and the constituency they represented. This returned some duplicates, due to MPs changing party or being suspended. In the former case, I removed them (as they would still represent the same constituency), and in the latter, I kept them (as either the MP or their replacement could have asked questions).

Linking Entities to Constituencies

To determine what constituency each entity was located in, if at all, there were two methods I considered. The first method was to take each entity's Wikidata information, and use the property P131, “located in administrative territorial entity”. Following this, I would detect if there was a match to any parliamentary constituency. The second method was to take the coordinates of the entity's location and see which constituency it would correspond to on a map.

There are limitations in both approaches. The first is limited because many locations mentioned will not have a P131 property, or that property won't point exactly to a parliamentary constitution. Take for example HM Prison Reading, whose P131 points to Reading. Reading contains two parliamentary constituencies: Reading West and Reading East, hence it would be ambiguous as to which constituency the prison belonged to. Using geographical locations would be able to solve this problem, as HM Prison Reading, would have a single point location and therefore assignable to a single constituency, without ambiguity.

Despite this, using geographical locations has its own drawbacks. Firstly, is that a geographical location can span multiple parliamentary constituencies. By using this method its location would be reduced to only one. This could be problematic for entities such as the “Kingdom of England”, whose coordinate location on Wikidata is located in the parliamentary constituency of Banbury. Consequently, this method would be susceptible to exaggerating particular constituencies when larger regions have their coordinates located in them.

Additionally, certain Wikidata items have multiple locations. The River Tees for example has three locations, where the one which is picked would influence which parlimentary

constituency it is said to belong. In these cases, I would have to decide on an approach by which I assigned them to a constituency or to remove them entirely.

Furthermore, specific departments of government such as the Department of Transport or the Cabinet Office, which have a physical location would have their location extracted and used. When an MP refers to such entities, they will most likely be referring to them in an abstract sense rather than their physical presence. Hence, I would have to take action to exclude such occurrences from biasing the results.

Despite the limitations of both approaches, I took the second, as I believed the first approach would result in a significant number of false negatives. In addressing these challenges, I chose to use location data from Wikipedia instead of Wikidata. Wikipedia's coordinate data was more aligned with my goals, with a tendency not to give locations to large geographical regions (such as "Kingdom of England") and few, if any articles with multiple locations. Additionally, I produced a list of entities to exclude, either from being abstract or too large a region to assign to a single constituency.

Geometrical Data

To get geometric data of the UK's constituencies, I used data from Ordnance Survey (4). The reliability of this source is very high, with the UK government stating that, for the UK, it provides "the most accurate and up-to-date geographic data" (5). For Northern Ireland, I used boundary data from the Open Data Portal for Northern Ireland, which is a government website (6). However, the data is from 2008. Despite this, there are the same number and names of constituencies that were used in the 2019 general election.

In determining whether an entity was located in the UK or not I iterated through each entity, checking which constituency its pair of coordinates fell within. Through this, I discovered

many outliers whose coordinates were in bodies of water, such as the constituency of Strangford. To handle them, I let each belong to their closest constituency.

Results

Before attempting to answer the main research question, I examined how many questions were asked per constituency. Figure 1 is a choropleth map with a continuous scale showing the total number of questions asked per constituency. Constituencies which asked no questions from the data are denoted with hatch markings for clarity. There did not appear to be any clear trend by which different regions were more or less likely to ask questions. What was most notable was how few constituencies asked any questions at all, with only 58% doing so. Additionally, it can be observed that a few constituencies asked a vastly greater number of questions than the rest.

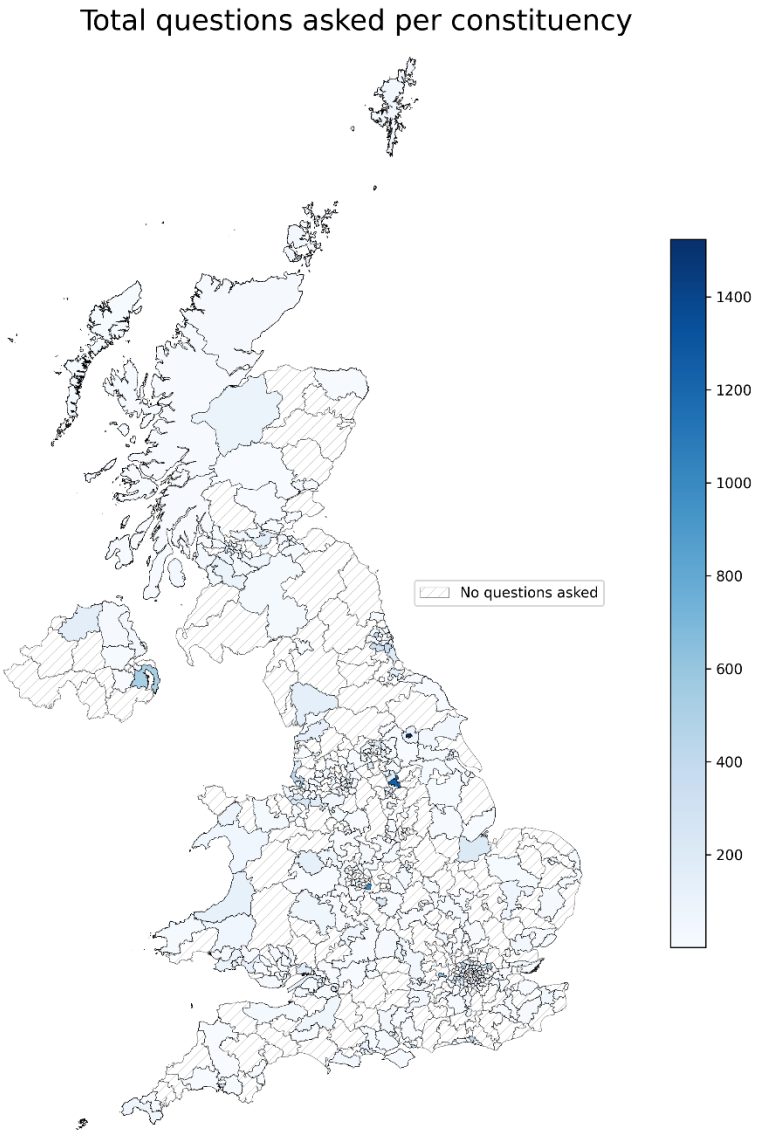

Figure 1

To investigate this inequality further, I plotted a graph of the cumulative proportion of questions asked per constituency, where constituencies are ordered descendingly by the total number of questions asked (Figure 2). From there it was clear the number of questions followed a power law distribution, where 80% of the questions were asked by only 111 constituencies, or 17%. This is roughly consistent with the 80-20 or Pareto principle.

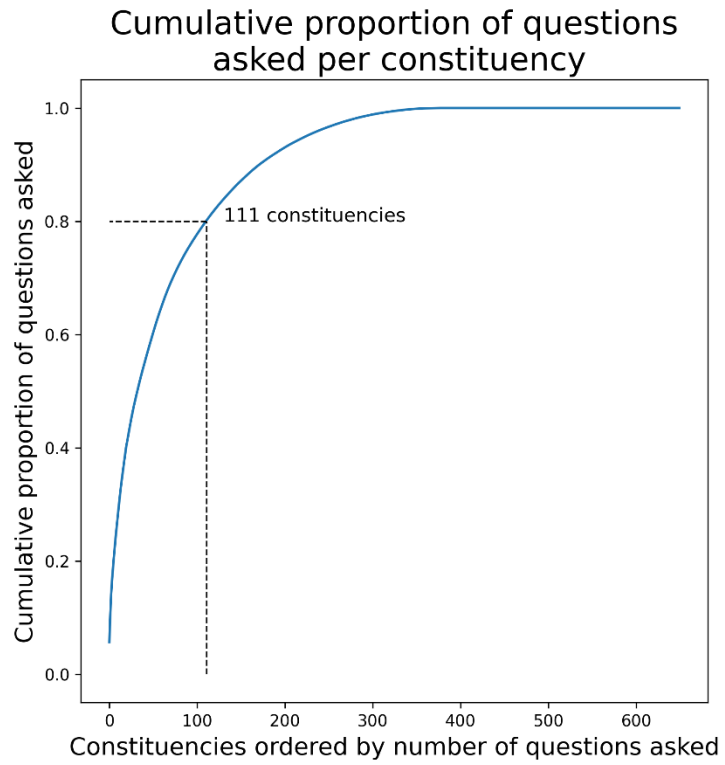

Figure 2

To analyse the extent by which MPs ask questions about their own constituency, I plotted another choropleth graph, with only self-focused questions (Figure 3). This was even sparser than the previous map, with only 21% of MPs asking one or more questions about their own constituency. This lack of data makes it hard to draw any solid conclusions.

For Northern Ireland, it was surprising to see only a relatively modest number of constituency focussed questions. Considering its distance and a strong sense of national identity, I had expected a greater amount. To a lesser extent, I had expected this of Scotland too.

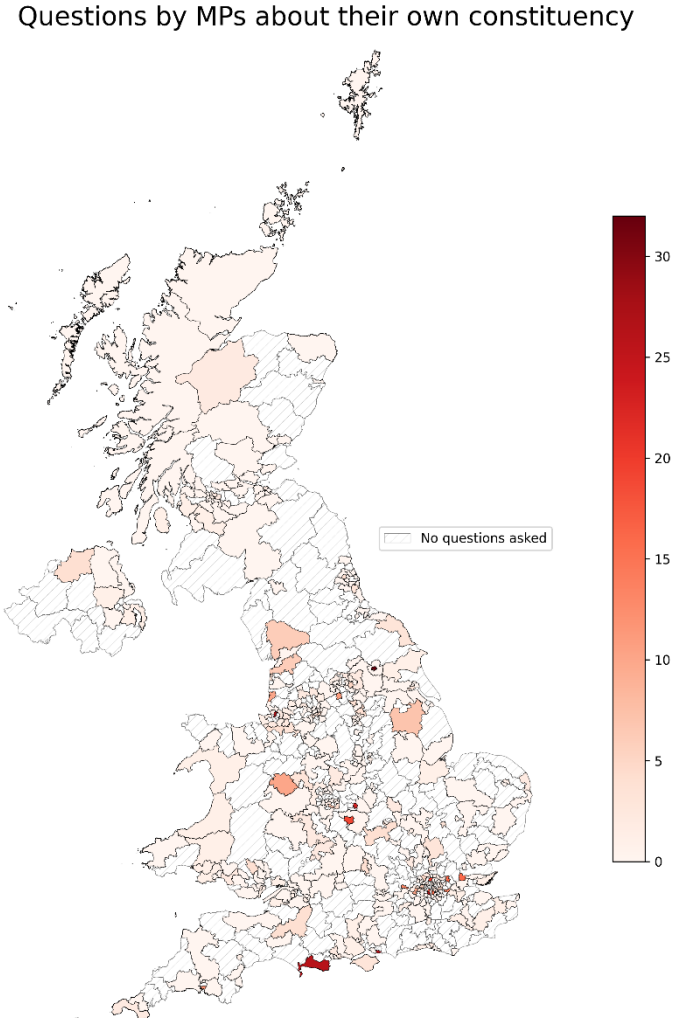

Figure 3

You could argue that the extent by which an MP asks questions about their constituency is more accurately defined as the proportion of related questions they ask rather than the absolute value. A paper that investigated UK constituency focus used this measure so as to quantify it (7). Figure 4 is a choropleth map showing this proportion. One notable change from the previous graph is that of York Central. Previously it was shown to have the greatest number of constituency focussed questions, but as a proportion of its total questions, is quite low.

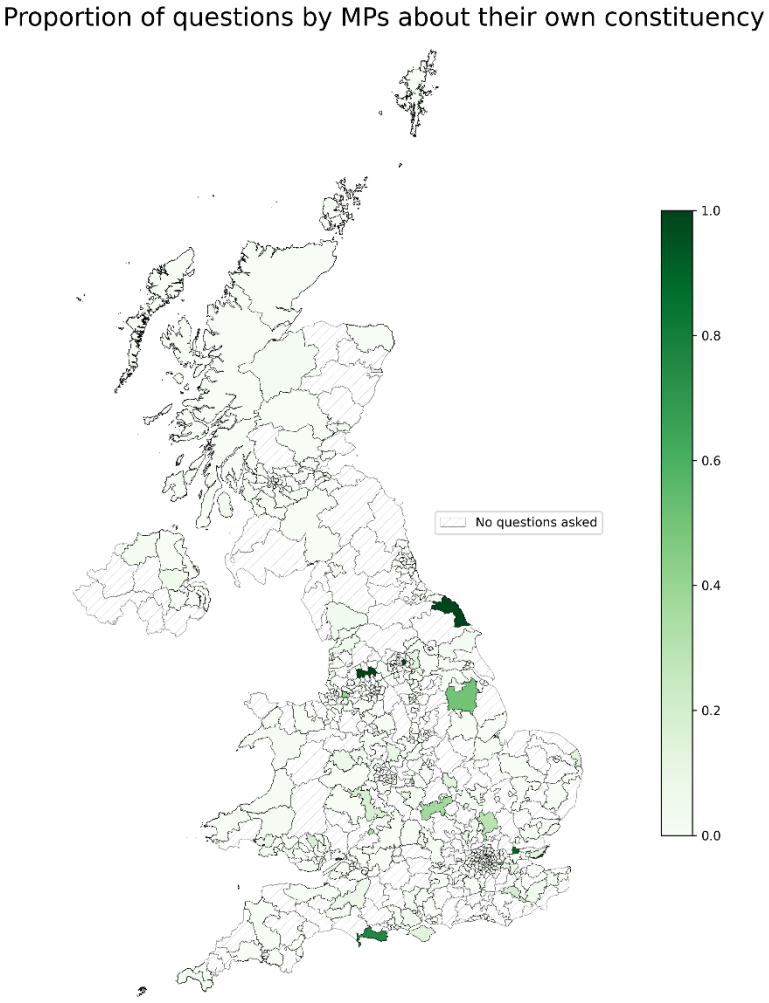

Figure 4

An interesting related question was to what extent cabinet ministers ask questions about their own constituency. Intuitively, I expected this to be less due to their additional responsibilities, and that usually these questions were directed at them.

From Figure 5, it can be observed that the average number of questions asked per MP was significantly lower for MPs in the cabinet. Not only were ministers less likely to submit a written question, but these questions were on average less directed towards their own constituency.

These findings are consistent with the results of the previously mentioned paper (7), where with a p-value less than 0.001, they found that cabinet ministers asked significantly fewer constituency focussed questions when compared to backbenchers.

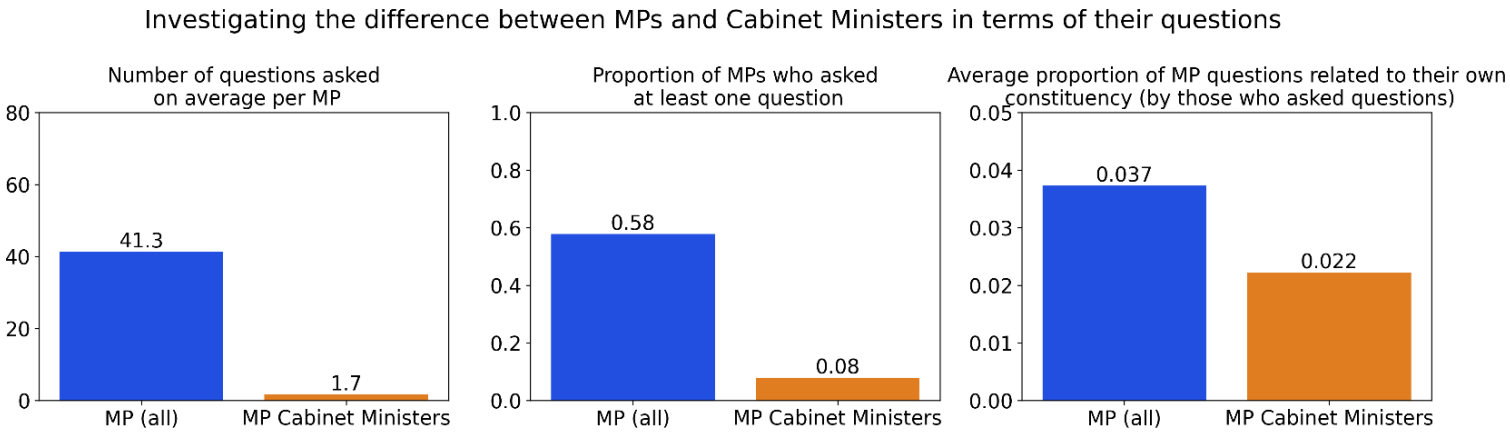

Figure 5

Conclusion

In conclusion, my findings support the idea that most MPs rarely ask questions that are related to their own constituency. During the 9-month period, on average, an MP asked 41 written questions of which 3.7%, or roughly two of them, were constituency focussed. Furthermore, 80% of these constituency focused questions were asked by only 52 of the 650 constituencies.

These findings are quite limited by the initial data, and for further research, it would be sensible to combine data from both spoken and written questions. My main requirement for my choice of NER+NED technique was speed, however a SOTA approach would likely improve the accuracy of the results. Additionally, my approach to removing larger regions was manual, so bias likely exists where they are treated as being in one constituency. These results should therefore be regarded with scepticism and I suggest further analysis is carried out to corroborate them.

Question 2

Geometrical Data

For this question, I needed to use geometric data which spanned larger regions than that of electoral districts. Because of the reliability of OS data, I used it again for England, Wales and Scotland. For Northern Ireland, I used a newer dataset from OSNI of county boundaries (8). The shapefile for this used the Irish National Grid (ING), so I converted it to BNG and merged all three into a single dataframe.

Linking MPs to Regions

To determine what region a particular MP serves, I decided to check which shape the coordinates of their constituency fell within. I picked this approach, as opposed to determining it through a Wikidata property, because it meant I would not have to find a map divided in necessarily the same way.

To get the coordinates of an MP's constituency, I made a SPARQL query to Wikidata. As before, I determined an MP by them holding a position of member of the 58th Parliament of the UK.

Similar to my previous question, a few outliers occurred when linking the MPs to regions. Again, this stemmed from their Wikidata coordinates being in water (i.e. Ross, Skye and Lochaber), and so I simply assigned them to the closest region.

Pre-Processing Question Text

To prepare the data for LDA, I needed a corpus of text. My method to produce this corpus was to concatenate all questions by MPs within a particular region. The corpus would then be a list of each region's concatenated questions.

To pre-process the corpus, textual data was tokenized and stop words removed. Initially, I used stop words from the Python stopwords package. However, this was insufficient in eliminating certain domain-specific artefacts such as list headings. To remove such occurrences and others, I manually added them to the stopwords list. Finally, to ensure greater consistency I applied stemming. Due to my unfamiliarity with natural language processing, I based my approach on code from a tutorial (9).

Applying LDA

To apply LDA to the corpus I used the Python library Gensim. For LDA there are two main measures of performance: perplexity and coherence. Perplexity is not correlated with the interpretability of topics and consequently, a model with low perplexity may produce topics that are difficult to interpret. For my use case, being able to interpret the topics was crucial, so as to understand the implications of regional differences. Hence, I opted to use coherence, which measures how semantically similar the most probable words within a topic are.

To optimize coherence, I ran a grid search, finding the best hyperparameter values for alpha, beta and the number of topics. Figure 6 shows the maximum coherence score from the search for each number of topics. From there it can be observed that eight topics produced the highest coherence score, where it begins to fall with any more.

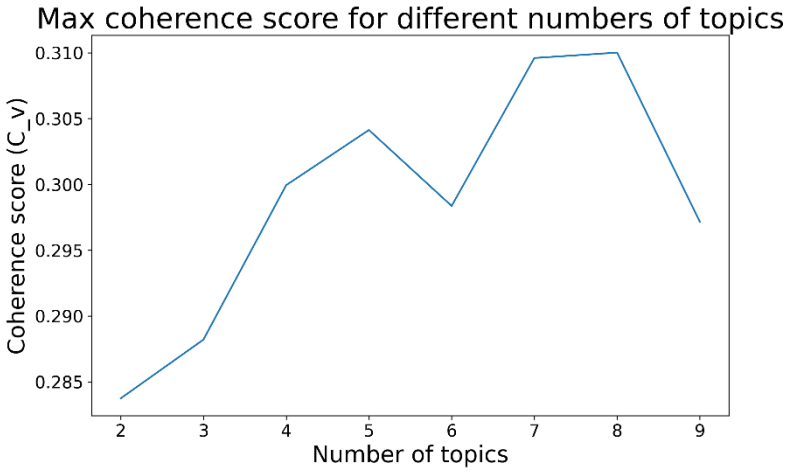

Figure 6

Using the results from the grid search I built an LDA model. Although the words within the topics were semantically similar, there existed many similar topics. “health”, for example, was present as the most popular word for two of them. As a result, this made understanding the difference in what each topic meant, difficult.

To mitigate this effect, I decided to reduce the number of topics. Although I knew this would reduce the coherence between words in a topic, I theorised that with fewer topics altogether, it would reduce the likelihood of similar ones from occurring. Despite reducing the topics to four, I still frequently produced similar occurrences. One observation I made was that the random seed played a large role in whether this occurred or not.

To procure a good seed, I ran the model for a range of seed values, evaluating each based on the semantic similarity between the top word in each topic. In this case, a low similarity was desirable, so as to ensure that the topics were different. From picking the seed with the lowest similarity value, I produced four topics that were meaningful and distinct.

Results

Figure 7 is a word cloud, depicting each topic by the words comprising it, with larger words being more probable. Three of these topics are comparable to the topics of “Education”, “Healthcare” and “Military” which were found by a paper that applied LDA to MP speeches (10). As such, I gave these comparable titles.

Figure 7

To explore the differences in topics between regions, I analysed each topic’s probability with respect to regional questions. Figure 8 to Figure 11, are choropleth maps depicting the probability of a topic for all given regions. When comparing Figure 9 to the rest, it is clear that nearly all of the UK’s MPs asked questions regarding health care. This was interesting, as it seemed to contradict the idea that health care was declining in importance as a topic (10). From investigating why part of Northern Ireland had no probability, it transpired that no MPs within this area asked any questions at all.

Figure 10, shows that questions on Education and Teaching were a somewhat prominent topic within North East England. Examining the data more closely revealed that 307 of the 354 questions from the North East, containing the word “school”, were made by a single MP. Moreover, this MP is the Shadow Secretary of State for Education.

Similarly, for Yorkshire, questions about “defence” were made in majority by the Shadow Secretary of State for Defence. This time it was more unbalanced, with 1,189 of the 1,265 questions being made by this MP. In addition to Yorkshire, Figure 11 shows positive topic probability in South East Wales and South West England.

Figure 8 is strange in that the graph looks blank, however, South West Wales (SWW) has a probability of 1.4%. When looking back at the wordcloud, it is clear that the topic is concerned with the North West of England (NWE), leaving the question as to why SWW has any probability. I hypothesise that although this topic may have been derived from NWE, the model does not assign it any probability because there are many more tokens in total. NWE has 58,114 tokens, significantly more than SWW’s 5,144. Hence, because of SWW’s single mention of “steel”, the model has seen this as more significant and consequently, prescribed it a probability.

Conclusion

My analysis of the question has shown some regional differences between the questions asked. Health care seemed to be a topic that all regions questioned, whereas Education and Defence were more exclusive. On further inspection of these topics, I found many questions were asked by the Shadow Secretary of the corresponding role. One limitation of my model was the first topic, which did not help explain any meaningful regional differences.

For further research in this area, I would recommend exploring how the coherence measure affects the quality of the produced topics. In my research, I stuck to using C_v, however, there are numerous others which optimising for could yield better results. Additionally, my method to reduce topic-to-topic similarity by procuring a good seed was ad-hoc, and there likely exists a better method to achieve this.

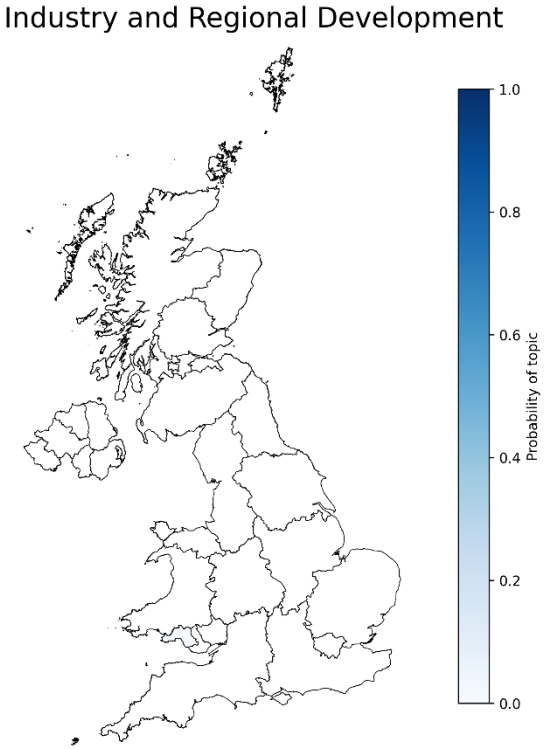

Figure 8

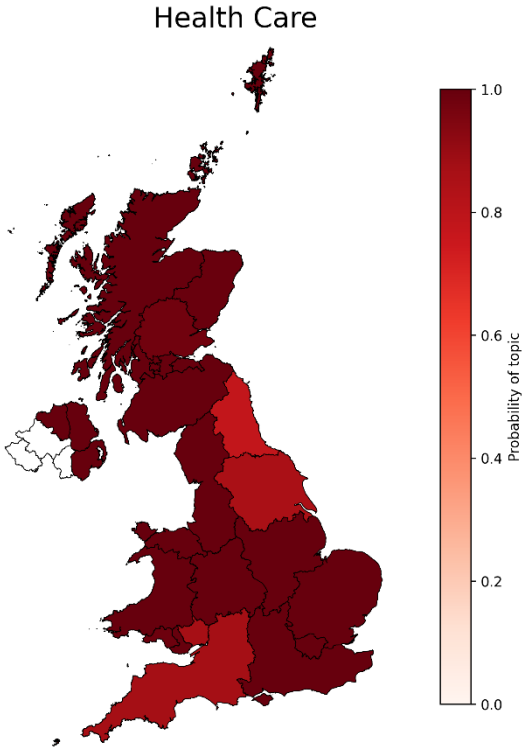

Figure 9

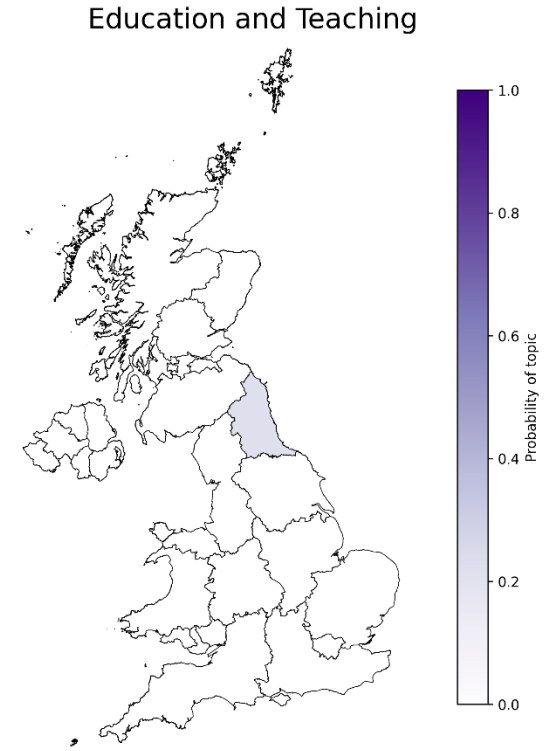

Figure 10

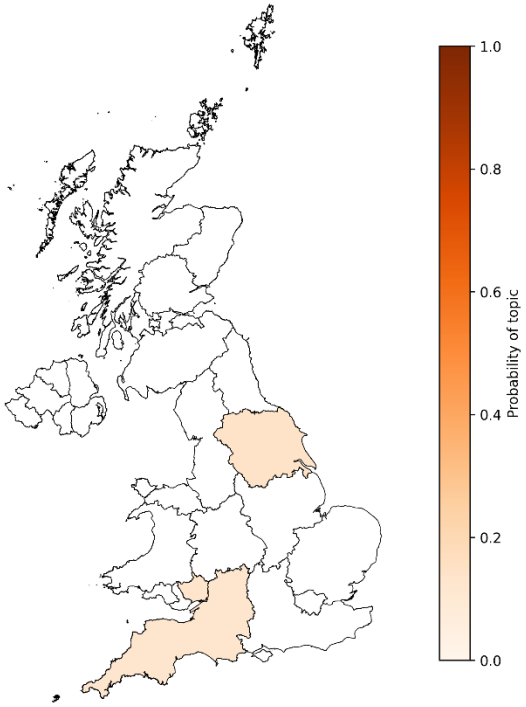

Figure 11

References

1. *A Fair and In-Depth Evaluation of Existing End-to-End Entity Linking Systems.* **Bast, Hannah and Hertel, Matthias and Prange, Natalie.** 2023, arXiv preprint arXiv:2305.14937.

2. *From TagME to WAT: a new entity annotator.* **Piccinno, Francesco and Ferragina, Paolo.** 2014. Proceedings of the first international workshop on Entity recognition \& disambiguation. pp. 55--62.

3. *Autoregressive entity retrieval.* **De Cao, Nicola and Izacard, Gautier and Riedel, Sebastian and Petroni, Fabio.** 2020, arXiv preprint arXiv:2010.00904.

4. **Boundary Line.** ***Ordnance Survey.*** **[Online] [Cited: 15th January 2024.] https://www.ordnancesurvey.co.uk/products/boundary-line#overview.**

5. ordnance survey. ***gov.uk.*** **[Online] [Cited: 15th January 2024.] https://www.gov.uk/government/organisations/ordnance-survey.**

6. admin.opendatani.gov.uk. ***OSNI Open Data - Largescale Boundaries - Parliamentary Constituencies (2008).*** **[Online] [Cited: 15th Janurary 2024.] https://admin.opendatani.gov.uk/dataset/osni-open-data-largescale-boundaries-parliamentary-constituencies-2008.**

7. ***Does constituency focus improve attitudes to MPs? A test for the UK.*** **McKay, Lawrence. 2020, The Journal of Legislative Studies, pp. 1--26.**

8. admin.opendatani.gov.uk. ***OSNI Open Data - Largescale Boundaries - County Boundaries.*** **[Online] [Cited: 18th January 2024.] https://admin.opendatani.gov.uk/dataset/osni-open-data-largescale-boundaries-county-boundaries/resource/137d0589-46fc-4d12-85c2-3941eaf2f165.**

9. Latent Dirichlet Allocation (LDA) with Python. [Online] 18th January 2024. http://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html.

10. ***Complex politics: A quantitative semantic and topological analysis of uk house of commons debates.*** **Gurciullo, Stefano and Smallegan, Michael and Pereda, Mar{\'\i}a and Battiston, Federico and Patania, Alice and Poledna, Sebastian and Hedblom, Daniel and Oztan, Bahattin Tolga and Herzog, Alexander and John, Peter and others. s.l. : arXiv preprint arXiv:1510.03797, 2015.**