**28/12/2023 (1.5 hours)**

Watched through the final video and got insights into what is being marked and relevant techniques LDA and NER + NED

**30/12/2023 (2 hours)**

So firstly I have been working on using SPARQL to bring in the Parliamentary database. I did this using sparql wrapper python library. I then removed some additional data, such as type (literal etc), and then paragraph information. I put this in a pandas dataframe. Question 1 is to comment on how often MP’s talk about their own constituency. There is not a lot of information about the sparql endpoint. I will likely have to apply named entity recognition and disambiguation so I know for a given question, what constituency are mentioned. I could then find the MP’s constituency (perhaps pre-compute this) and see how many matches there are. I was thinking also about having a map, colour coded, to show for each constituency, how often its MP’s discuss it.

One limitation I have found so far is that this is just written questions. Perhaps I could use a sparql query to get spoken ones too.

**31/12/2023 (2 hours)**

<https://pypi.org/project/spacy-entity-linker/#description>



A black background with white text

Description automatically generated

<https://tagme.d4science.org/tagme/>

<https://sobigdata.d4science.org/web/tagme/tagme-help>

5fcf52fb-79b2-4718-b7ee-62957df7d7e0-843339462

picked up correct durham.

Id is the curid in wikiedpia like **https://en.wikipedia.org/?curid=32293**

Id is the curid in wikiedpia like **https://en.wikipedia.org/?curid=32293**A screenshot of a computer

Description automatically generated

Id is the curid in wikiedpia like **https://en.wikipedia.org/?curid=32293**

So today I have implemented multiple NED systems in python and managed to get them to extract entities from text, which is stored in a dataframe. I have looked at two systems. The first is <https://pypi.org/project/spacy-entity-linker/#description>. However, through manual testing it seemed to perform poorly. It would identify words such as “department” as a Bollywood film and “State” as software. Instead I looked at the literature and one API that had been used was “TagMe”. I have had a good applying this, and getting an API key, and this has been much more successful. It, also provides a detailed output about its confidence, ID of the Wikipedia article and has documentation <https://sobigdata.d4science.org/web/tagme/tagme-help>. There are also alternatives available on the same website that I could use such as WAT or REL, which are supposed to be more accurate, albeit experimental. Additionally, there is a website showing code with papers <https://paperswithcode.com/paper/pre-training-of-deep-contextualized>, and it has an approach that used a transformer. This would be a bit more of a modern approach <https://github.com/studio-ousia/luke/tree/master/examples/entity_disambiguation>.

Th challenge now, is to be able to link these entities to constituiencies, and also link each person with a constituency too. I could make a new related dataframe with people and there constituency, which I suppose I could mine. It seems more difficult to somehow filter only the place names, and then somehow relate them to a particular constituency.

**01/01/2024 (4 hours)**

When looking at constituencies on the parliament API and people there did not appear to be a clear link between them. Instead I had to use wikidata and have made a related df.

The 58th parliament of the united kingdom, is when these questions are from as it is from 2019 onwards. I therefore use this in my query.

To associate each MP asking a question, with their respective electoral district, I wanted to make a dataframe storing pairs of people and their corresponding district. To do this, I have written a SPARQL query, that extracts each person who has the position of being the 58th parliament of the uk and then their electoral district. When doing this, I got more pairs than the number of elected people for uk parliament (650). Some of these were duplicates (33), but there appears to more (669), still.

The questions from the data extracted contains questions from the house of lords, who are not MP’s and do not represent a constituency. We can alter our sparql query to filter out qnum that start with HL by adding regex with a negative lookahead.

For my related df, which I shall call “**district**”, I then reduce it to only contain mp’s who actually ask questions in the “**question**” df. For a given question, I can now link the **district** of the mp who asked it. The challenge now, is to be able to take the extracted tag’s, filter geographical locations, and then determine whether these locations are in their district.

A screenshot of a computer

Description automatically generated

<https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial>

<https://www.wikidata.org/wiki/Help:Qualifiers>

<https://query.wikidata.org/#SELECT%20%3Fperson%20%3Fdistrict%20%3FpersonLabel%20%3FdistrictLabel%0AWHERE%0A%7B%0A%20%20%3Fperson%20p%3AP39%20%3Fposition.%0A%20%20%3Fposition%20ps%3AP39%20wd%3AQ77685926.%0A%20%20%3Fposition%20pq%3AP768%20%3Fdistrict%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22en%22.%20%7D%0A%7D>

<https://api.parliament.uk/sparql/>

**02/01/2024 (2 hours 45 minutes)**

I have been looking through the data and been analysing why I have more than 650 MP’s (before limiting it to the one’s who asked written questions). MP’s can resign or be suspended , in which case a by election is sometimes run (Mike hill for Hartlepool) (Imran Khan for Wakefield).

Wat supposedely deprecates REL, so should use instead. REL is Wikipedia dump 2021 which could be a limitation if MP’s have stepped down since then.

REL had outdated code and endpoint on the website I used it on, and then when I run it, the API stopped working suddenly and returned Proxy error.

WAT is from 2014 so a bit outdated but is meant to be the replacement to TagMe.

Couldn’t get spacy working with transformer model but both md model and large model were poor (calling Health a music group).

I was thinking about using the location property on wikidata to identify what electoral district a entity belongs, but this could be limited. Take for example, entities that are hierarchically higher than an electoral district, i.e England. These would have a single point still, and thus fall in the boundaries of an electoral district, yet obviously do not constitute one.

Interesting observation for:

To ask the Secretary of State for Health and Social Care, what assessment he has made of the potential impact of the settlement decision of NHS Lancashire and South Cumbria Integrated Care Board not to grant additional funding to hospices on (a) patient care and (b) the financial sustainability of those hospices.

TagMe

A screenshot of a computer

Description automatically generated

Wat

A screen shot of a computer

Description automatically generated

REL

A black background with white text

Description automatically generated