Using topic modelling to explore case summaries

Author: PhD fellow Asta Sofie Stage Jarlner

This blog post computationally explores the latent topics present in a selection of public case summaries from the Refugee Appeals Board. It showcases how natural language processing can be used for initial exploration of text data as well as some of the considerations that must go into selecting which model to employ. I present a few findings and how one can go about using topics to discover cases, explore topological distribution for a specific group of applicants, and how topics are associated with decisions of the Board.

To live up to the requirements of the UN Refugee Convention and grant refugee status to individuals with a well-founded fear of persecution or real risk of facing torture, every convention state must have a bureaucracy in place to identify those with legitimate claim to asylum. In Denmark, the first instance to evaluate an asylum seeker's claim is the Immigration Service. If the case is rejected, it is automatically appealed to the second instance, The Danish Refugee Appeals Board (See previous blogpost by Høgenhaug and Nielsen), who can either uphold, overturn, or remand the Immigration Service's decision.

On its website (fln.dk/Praksis) the Board publishes summaries of some of the cases they evaluate. In this blogpost, I explore the topological themes in these summaries by employing a natural language processing (NLP) model to a selection of case summaries. This provides an overview of the nature of the cases evaluated by the Board, as well as how outcomes vary across the identified clusters. I employ a topic model, which is an unsupervised machine learning model that can detect semantic structures and discover latent topics within a collection of documents. Computational tools such as topic modelling allow for a fast and consistent analysis of large and complex corpora of text documents though which patterns and nuances non-comprehensible for the human-reader can be found.[1] Especially in the legal domain, researchers can benefit from computational methods, not only due to the vast amounts of text documents that are produced through cases, regulations, and opinions, but also due to their often complex and lengthy nature. Topic models and other explorative approaches to empirical legal research can help identify legal themes, concepts, connections, and patterns and as such help us understand the complex legal landscape and how it has developed over time.

I rely on my colleagues' labour: Researchers at the DATA4ALL project have previously scraped the case summaries from the website, translated, systematised, and quantitatively explored them (See previous blogpost by Katsikouli). While 8000 case summaries have been retrieved from the website, I have focussed on those 731 published case summaries where the final decision fell between 2018 and 2020 and subjected these to a preliminary explorative analysis. Out of these cases, the vast majority concern applications from Somalia (298), Afghanistan (267) and Iraq (66) and only 134 of applicants are labelled female. In total, the Refugee Appeals Board decided to uphold the decision made by the Immigration Service in 567 cases, while 146 were overturned and 17 were remanded to the first instance. In one case concerning a family complex, the Refugee

[1] Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21, no. 3 (2013): 267-297.

Appeals Board decided to reject the adult applicant's application for asylum, whereas the minor applicant conversely was granted protection.

Methodological choices

In this blogpost, I will show how topic modelling can be used to cluster documents based on their topological contents by grouping documents based on their semantic structures. In other words: I subject the case summaries to a topic modelling algorithm and the model finds patterns of words in them. As a result of this clustering exercise, we can see which words make up a topic as well as how present each of the topics are in each of the summaries based on a probability distribution showing how probable it is that a document belongs to a certain topic. Specifically, I use a Hierarchical Stochastic Block Model (hSBM) which draws on community detection to identify clusters within texts and as such represents words and documents as nodes and edges.[2] While there are many kinds of topic model algorithms available, I rely on the hSBM for two main reasons:

First of all, it is better at capturing the hierarchical structures that are present in many real-life datasets compared to e.g., the widely used latent Dirichlet allocation (LDA) model. The LDA assumes that each document contains a mixture of topics, and each topic is a distribution over words. In this model, nodes (documents and words) are independent of each other, and sheer concurrence of words determines the clustering. The hSBM on the other hand is a generative model for networks, which assumes that nodes are assigned to a single community at each level of the hierarchy, and the communities are generated from a hierarchy of nested community structures. This means that a hSBM takes dependencies amongst words and documents into account when detecting clusters and as such calculates a context of each of these words.[3]

Secondly, while many topic models including the LDA require the researcher to select the desired number of topics the model should return, the hSBM model does this itself. It might seem trivial at first, but selecting the appropriate number of topics is by no means inconsequential for the analysis. Setting the number of topics too low will result in distinct topics being fused, while too few clusters will split otherwise coherent topics into multiple topics. This stems from the mathematical assumption that all topics are of uniform size distribution and can compromise the quality of the analysis as fused topics are difficult to infer meaning from and split topics prevent us from seeing

[2] Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann. "A network approach to topic models." *Science advances* 4, no. 7 (2018): eaaq1360.
[3] Gerlach et al., 2018

any thematic connection between clusters.[4] To infer compelling meaning from topics that might be fused or split as well as determining if this might be the case requires the researcher to have extraordinary knowledge of the field. The hSBM, on the other hand, uses a Baysian approach which means that it uses statistical methods to estimate the probability of different numbers of communities. As such, the model itself provides the researcher with a number of topics based on the clusters it can identify in the network it creates from the set of documents it is fed. The number of topics at different levels corresponds to the number of clusters, depending on how aggregated they should be. As such, the model can capture the structure and organisation of the data in a more fine-grained way, by modelling the relationships between topics at different levels of the hierarchy.

In summary, hSBM is different from other models that try to identify patterns in data. Where the LDA focuses on finding topics in documents, the hSBM focuses on finding communities in networks. To determine the appropriate number of latent variables, LDA relies on a set number of topics chosen by the analyst, while hSBM uses statistical methods to estimate the best number of communities for each level of the hierarchy.

Before running the topic model on the case summaries however, I first pre-process them to make them ready for computational analysis. Computers do for example not by themselves recognise grammatical variations as the same words and as such the summaries have to be normalised. Further, so-called stopwords, which are commonly used words that may not carry much meaning or relevance for the purposes of text analysis, are removed. While generic stopwords (such as "the," "and," "is," etc.), which are common across domains are straightforward to remove, is it a more iterative process to develop the list of more domain specific stopwords. Commonly used legal words, such as “assessment”, “decision”, “case”, “cf”, are identified by re-running the model and manually elaborating on a list of stopwords, specifically relevant to the asylum jurisprudence domain. Ultimately, removing these words makes the text analysis more accurate and efficient but requires familiarisation with the documents.

[4] Carlsen, Hjalmar Bang, and Snorre Ralund. "Computational grounded theory revisited: From computer-led to computer-assisted text analysis." *Big Data & Society* 9, no. 1 (2022): 20539517221080146.

Results

Figure 1is a visual representation of the clusters identified when the case summaries are subjected to a hSBM topic model. It shows the document nodes i.e. the case summaries (left) and word nodes (right) and how these are clustered in a bipartite network at five levels of aggregation (blue dots). At level one - that is the second level of aggregation - the hSBM produces 15 topics.

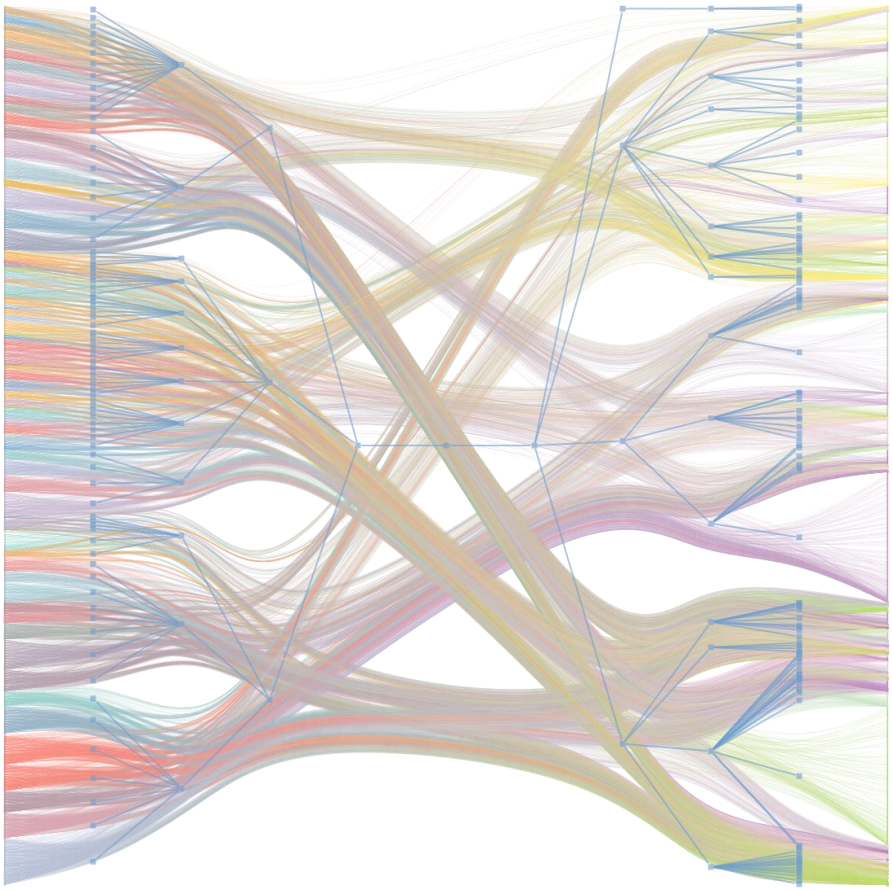


Figure 1

Figure 2 illustrates the words with the highest probability of being used within a topic-cluster.

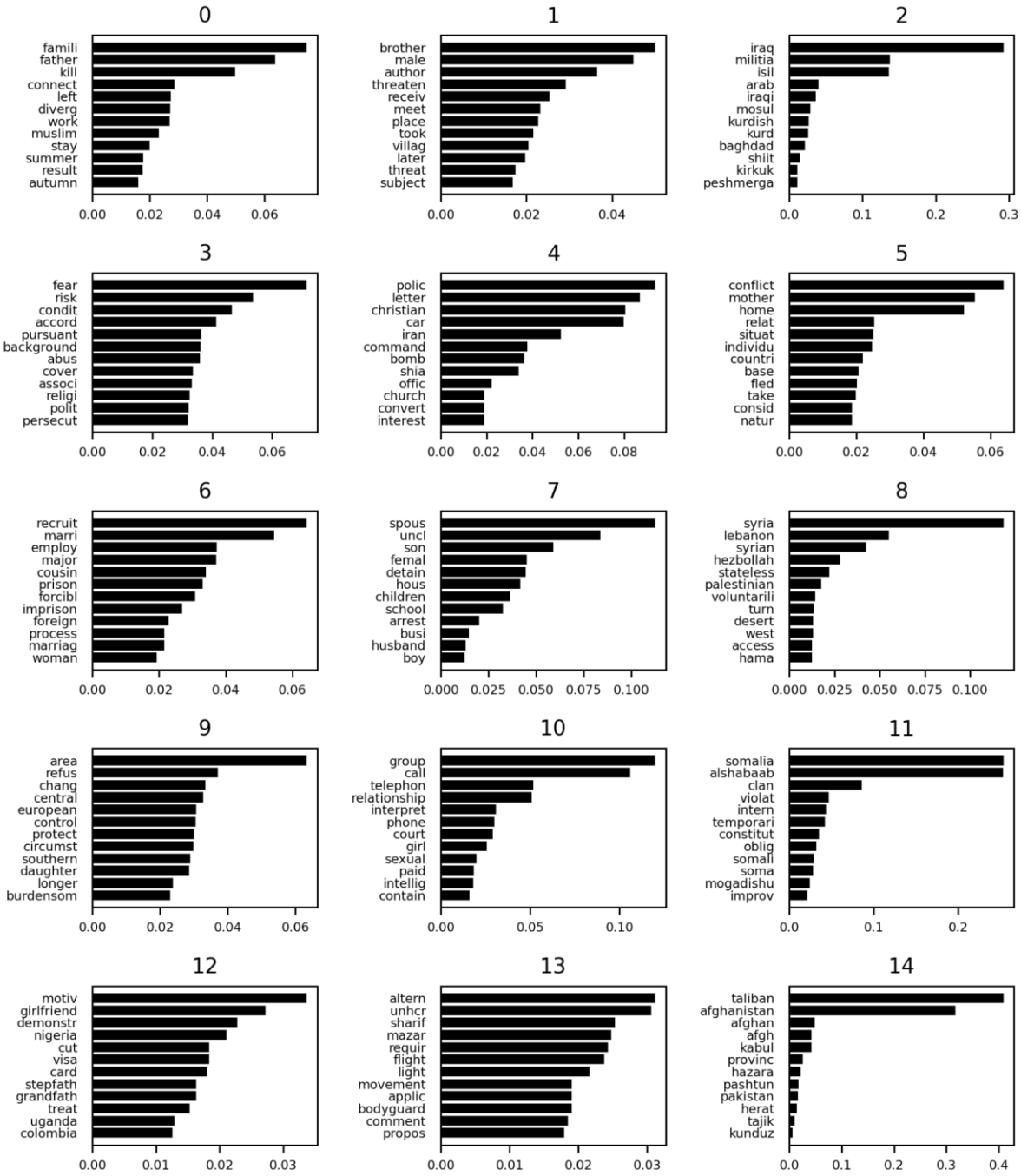


Figure 2

When examining the topics a few findings are of interest for the preliminary analysis. First of all, topic 0 is the largest topic-cluster in the network. Reading the summaries with the highest probability of belonging to this topic provide contextual information of the cases. Case summaries with the highest probability of belonging to topic 0 concern male applicants, largely Sunni Muslim from Afghanistan and Iraq who are not associated with any political or religious associations. As motivation, the fear of being killed by the Taliban or ISIL, supported by the applicants' fathers being so, dominate. Overall, the Refugee Appeals Board finds the applicants to explain divergently,

unclear, and present facts that are unlikely or seem constructed for the occasion and are consequently rejected.

While the topics can be explored individually to see the documents making up the clusters it is also possible to explore a certain demographic group and what topics they are most closely associated with and thereafter explore the topic. For example, case summaries concerning applicants from Somalia are not explicitly more associated with one topic but do have a higher probability of belonging to topics 9 and 11 than case summaries concerning complainants from other countries (see Figure 3).

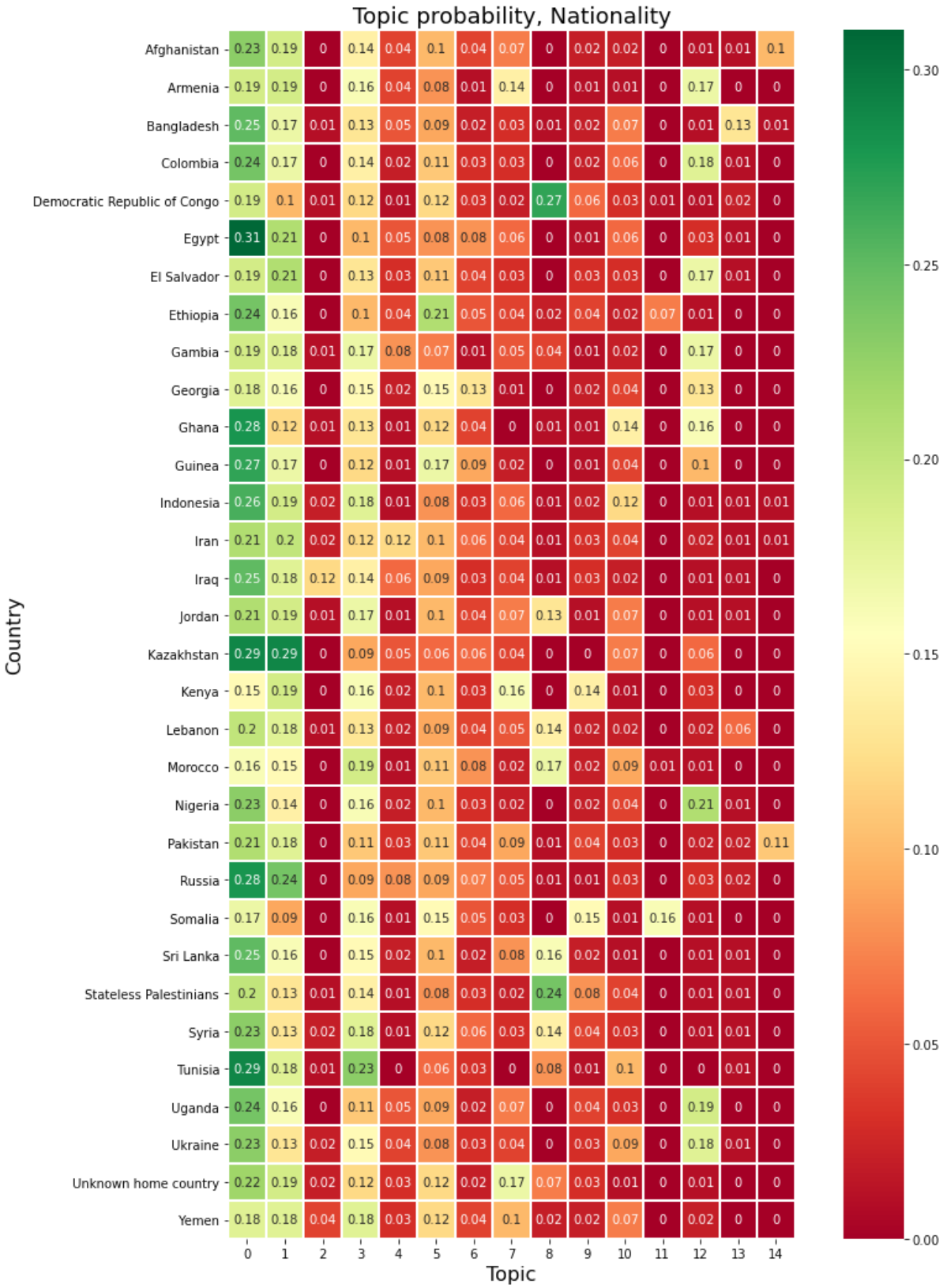


Figure 3

Referring back to Figure 2 topic 11 includes words indicating that Somalia was mentioned in the summary, but also the militant political organisation al-Shabaab. Meanwhile, topic 9 rather relates to the changed circumstances in central and southern Somalia. This is supported by a close reading

of the summaries with the highest probability of belonging to these two topics indicating that they are largely concerning re-evaluations of temporary residence permit pursuant to section 7 (1) of the Aliens Act. 2, cf. the practice after the Sufi and Elmi v United Kingdom case. This decision from the ECHR led to an increase in residence permits granted to Somali applicants, as deportation to southern and central Somalia was found to constitute a violation of Denmark's international obligations, including Article 3 of the European Convention on Human Rights. The introduction of section 19 of the Aliens Act of March 2019, however, led to a need for re-evaluation of many of those cases involving people who fled their country due to general levels of violence. According to section 19 subsection 1 of the Aliens Act, a residence permit pursuant to section 7 must be revoked when the condition that granted a person asylum is no longer present - unless it would be in conflict with Denmark's international obligations. Alas, a changed assessment of the situation in southern and central Somalia caused a large number of cases to be re-examined. As such, the two topics are also more probable of being present in case summaries from 2019 than other years (see Figure 3).

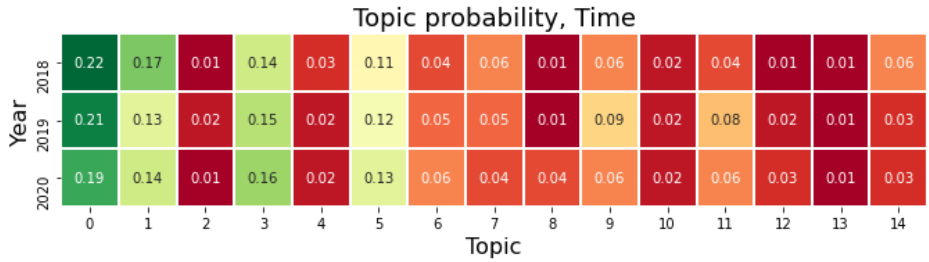
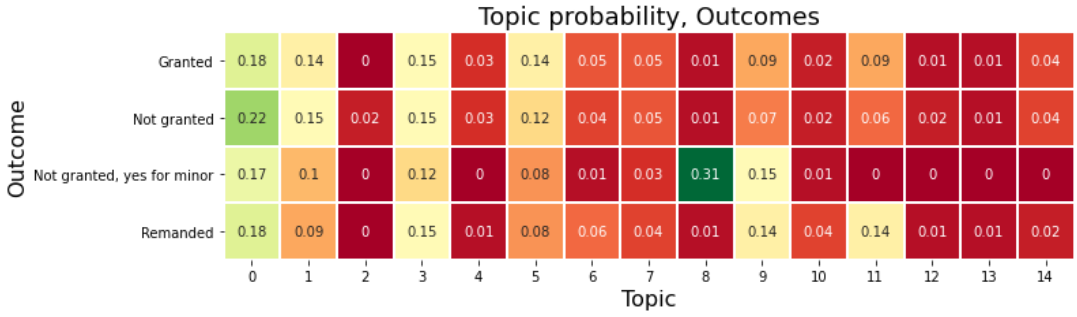


Figure 4

Looking at the topic probability distribution in case summaries and the subsequent outcomes (Figure 4), it can generally be observed that a few topics (5, 9, and 11) are associated with a slightly higher rate of first instance decisions being overturned than upheld (topic 6 is associated with an increase in probability of being granted asylum that is only 0.01 and is too small to infer meaning from). The cases with the highest probability of belonging to topic 5 concerns complainants from Somalia, Syria and Afghanistan. Of the 20 case summaries with the highest probability of belonging to topic 5, 13 of those applicants were ultimately granted asylum in Denmark.



Reflections

While the dataset used for this preliminary analysis is relatively small and thus compromises the quality of the conclusions that can be drawn - especially concerning national differences - it showcases how computational topological clustering can contribute with useful insight into the patterns existing in Danish refugee status determination practices. Clustering documents across many topics and in a hierarchy is not a feasible task for a human researcher. While a dataset with more and longer cases would only exacerbate the infeasibility of the task for a human researcher, a hSBM topic model would be able to return an even more fine-grained analysis of the cases that are taken up by the Refugee Appeals Board and how these are evaluated. This demonstrates the potential computational methods and NLP holds for empirical research in the legal domain.