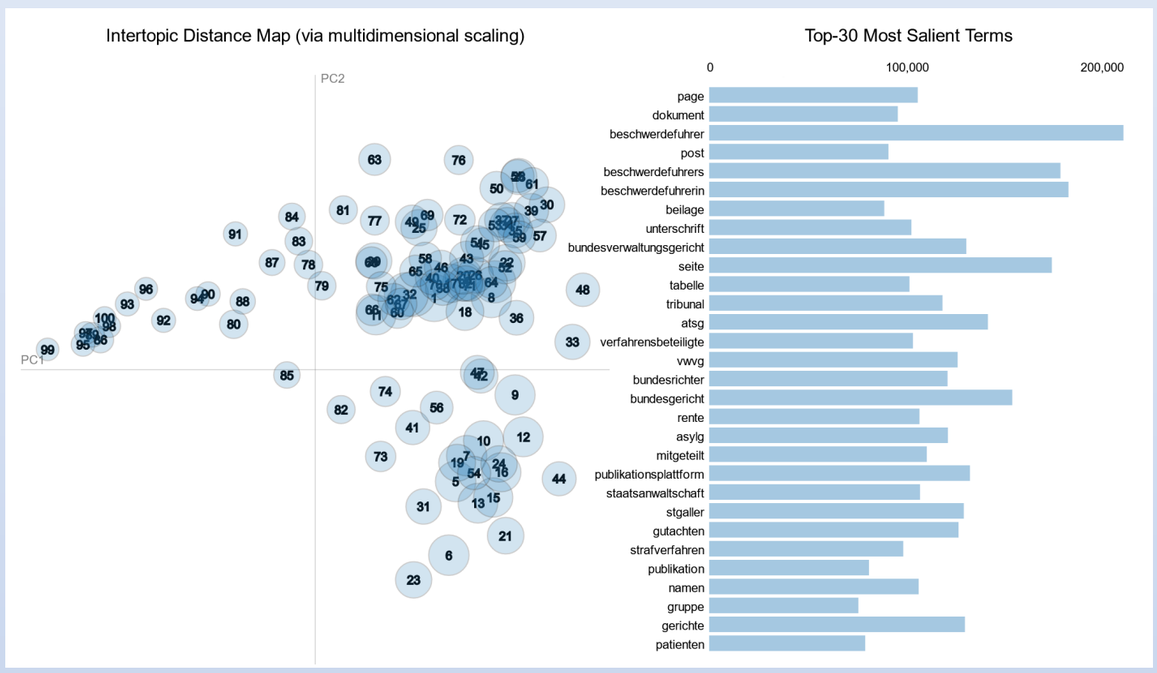
Topic Modeling for Swiss Court Rulings

Goals and Motivation

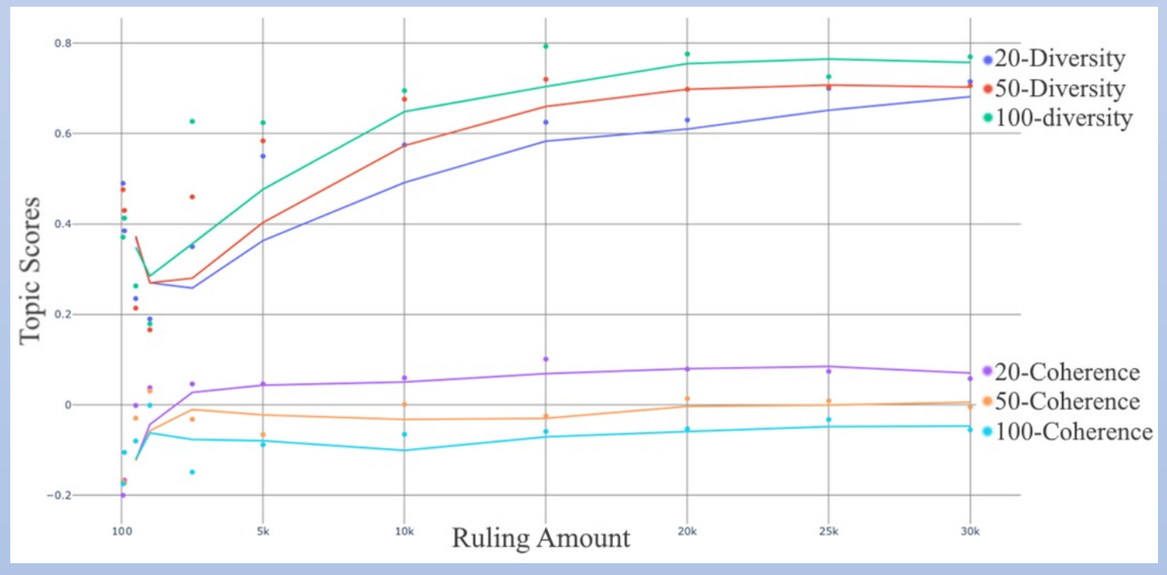
The vast amount of Court rulings makes straight forward analysis for many use cases difficult and time consuming. By performing topic modeling on the dataset, the different rulings can be grouped into meaningful topics and analysis can be performed based on those groupings.

Topic Map



Topic Modeling Scores

Using different metrics such as topic diversity and topic coherence, we were able to deduce, that when increasing the amount of topics, the coherency suffers while the topic diversity prospers. Furthermore increasing the amount of samples always produced better results.



Multilanguage

Multilanguage support proved to be difficult to handle with CTM, LDA and NMF. When using a multilingual training input for the model, the resulting topics often contained similar words, just translated between the languages or the same topics occurred in each language, which reduced the overall amount of topics per language. Lastly, there was the issue of our lacking knowledge of French and Italian which lead to a worse custom stopword removal for these languages. BERTopic offered a much better multilanguage support and is the recommended approach to do any multilingual analysis.

Preprocessing

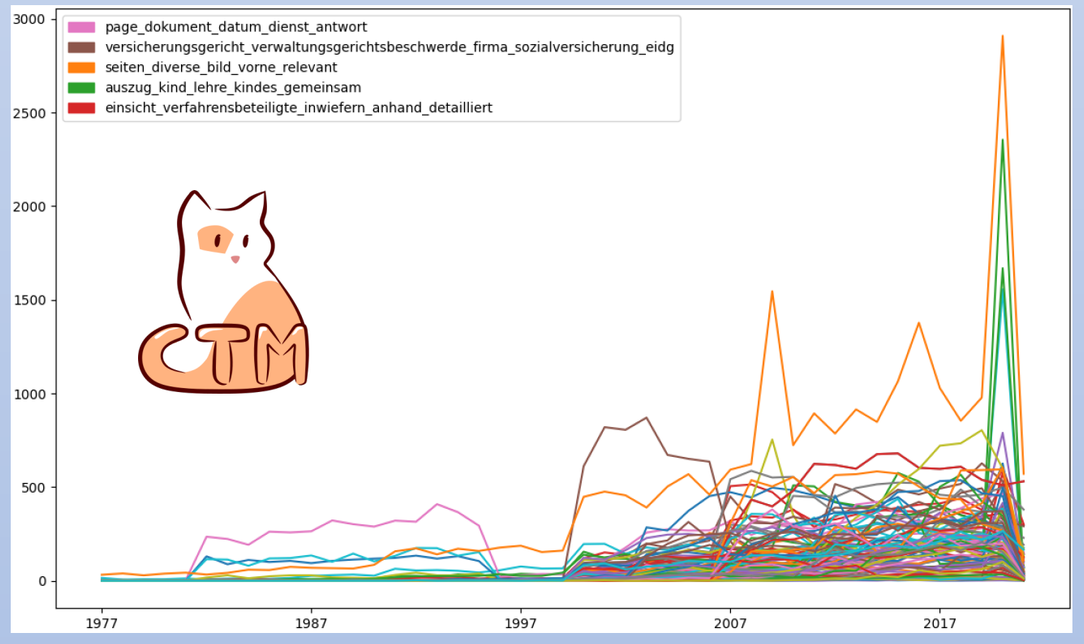
For CTM, LDA and NMF some preprocessing had to be done. This included converting everything to lower case, removing special characters as well as stopword removal. For the stopword removal, we combined pre-existing libraries with custom stopwords specifically selected for the work with court rulings. In the case of CTM, we also applied Lemmatization to the vocabulary of the most frequent words. For BERTopic no preprocessing is recommended.

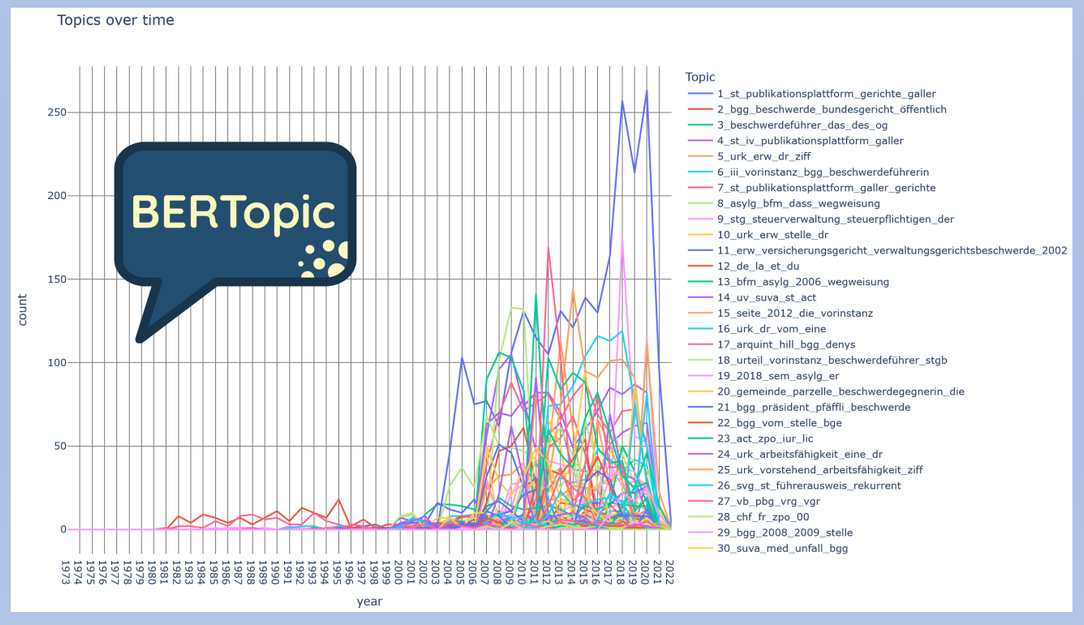
Topic Modeling Approaches

We tried different approaches such as BERTopic, CTM, NMF and LDA. BERTopic proved to be the best fit, as it can be set up quickly and offers good multilanguage support. The topics delivered by CTM were well usable as well, LDA and NMF delivered unsatisfactory results.

Hierarchical Clustering

Topics by year





Natural Language Processing Seminar
Spring Semester 2022
University of Bern
Marius Asadauskas, Renato Rao
Dominik Ummel
Supervised by: Joel Niklaus

CTM Topic Map:

BERT Topic Map:

Bert Hierarchy:

UNIVERSITÄT BERN