Mining United Nations General Assembly Debates

Natural Language Processing
Project 1: final presentation

Team 13: Debates-3MB

Mateusz Grzyb
298820

Mateusz Krzyziński
305739

Bartłomiej Sobieski
305830

Mikołaj Spytek
305753

December 13th, 2023

United Nations General Assembly (UN GA)

United Nations (UN):

- international organization established after World War II in 1945 to prevent future wars
- primary goals: maintain world peace, protect human rights, promote nations' cooperation
- at formation - 51 member states; as of 2023 - 193 member states - most sovereign states

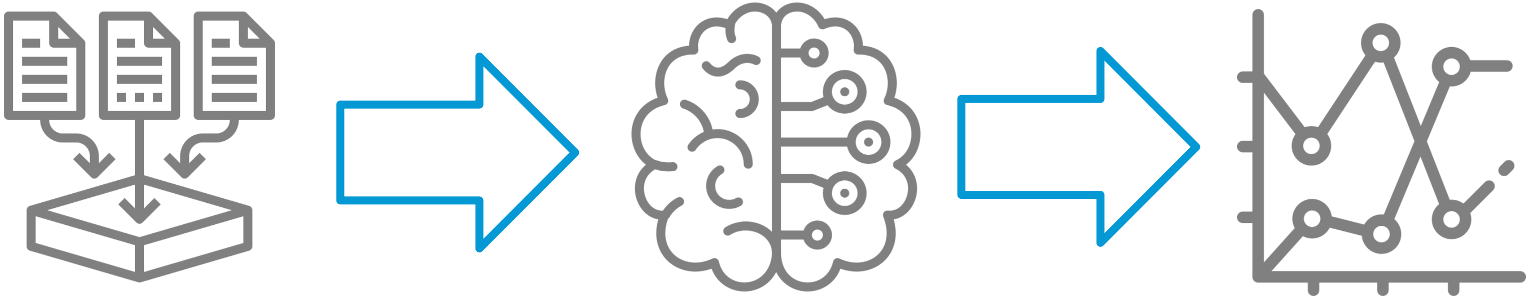
General Assembly (GA):

- central policy-making and representative organ of the UN
- takes place in yearly sessions; gathers all UN members
- general debate - during the opening of each new session
- transcripts of all general debates are publicly available

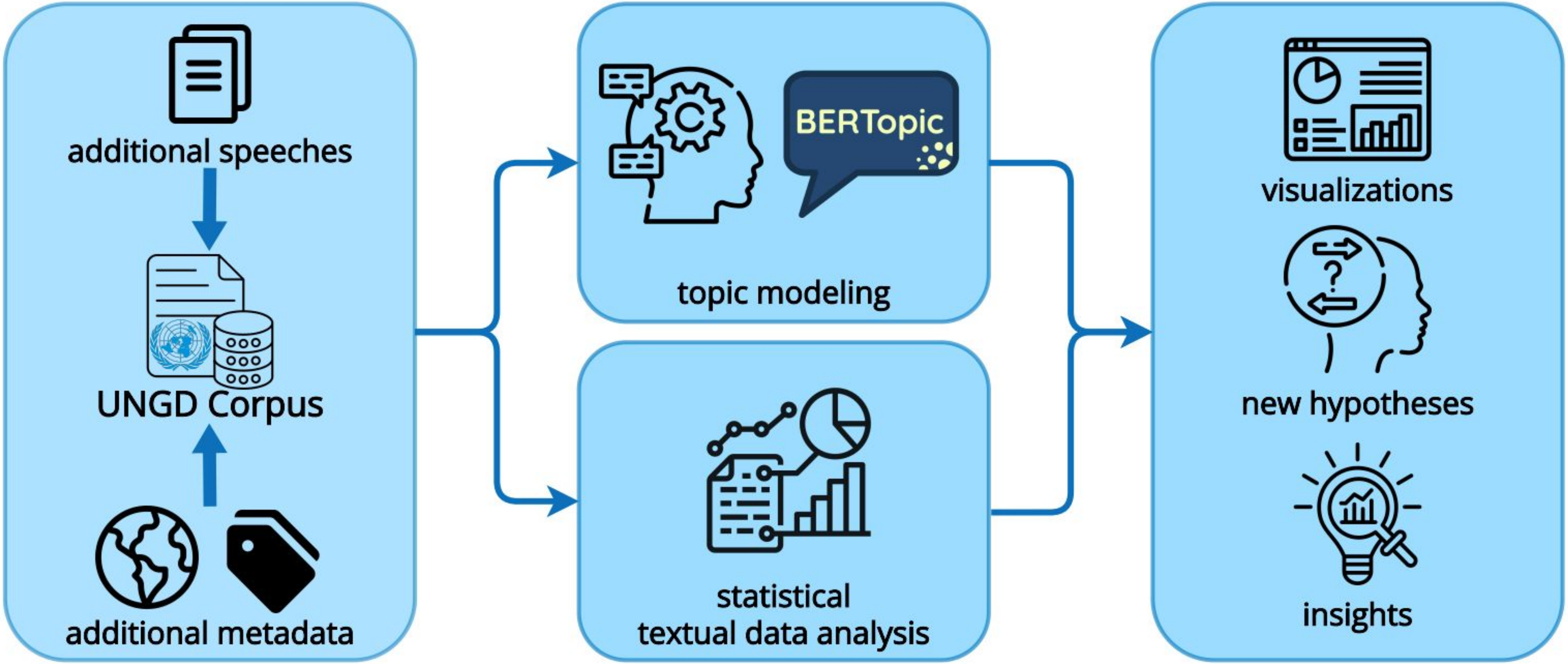
United Nations

Goals of the project

- Preparing a complete UN GA debates corpus (1946-2023) together with metadata
- Enriching this metadata based on additional sources (e.g. Gross Domestic Product)
- Exploring the gathered data using statistical text analysis; visualizing the results
- Applying state-of-the-art topic modelling techniques based on transformer models



Solution diagram



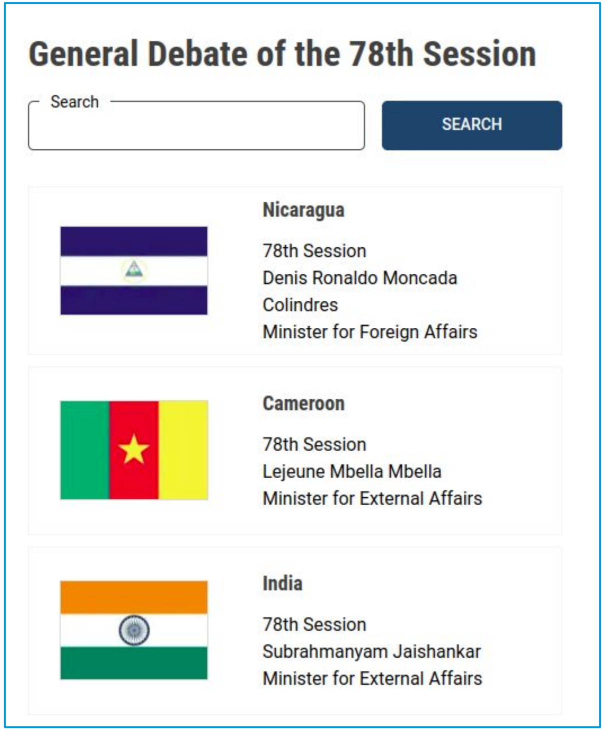
Gathering 2023 statements - web scraping

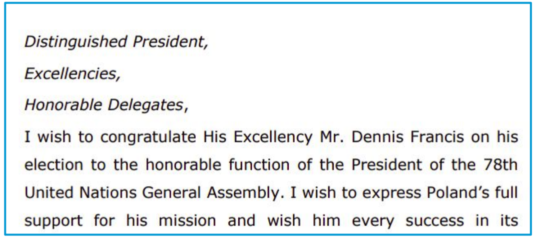
Technology stack:

- ❑ selenium
- ❑ pdfplumber

Source:

[United Nations website]





Distinguished President,

Excellencies,

Honorable Delegates,

I wish to congratulate His Excellency Mr. Dennis Francis on his election to the honorable function of the President of the 78th United Nations General Assembly. I wish to express Poland's full support for his mission and wish him every success in its

Cleaning the dataset

Garbage in, garbage out

original dataset

year	session	ISO code	country	speaking person name	speaking person position
2022	77	BRA	Brazil	Jair Bolsonaro	President
			...		

Problems:

- **typos**, e.g. *Trinidad and Tobado, Swistzerland, Bostwana, United Kindom, ...*
- **inconsistencies**, e.g. *Holy SEE, Vatican, Holy See, Vatican City State, ...*
- **wrong ISO codes**, e.g. *POR* instead of *PRT* for Portugal
- **countries that no longer exist and the ISO codes assigned to them**, e.g. Ukrainian SSR → Ukraine; Yugoslavia (YUG) → multiple smaller countries; German Democratic Republic (DDR) as part of Germany (GER),

... and many more edge cases to consider

Enhancing metadata

year	session	ISO code	country	speaking person name	speaking person position
2022	77	BRA	Brazil	Jair Bolsonaro	President
			...		

cleaned original dataset

GDP	...	HDI
	...	

additional metadata

New covariates:

- population
- total fertility rate
- human development index
- GDP (at constant 2015 US $)
- unemployment rate
- Gini index
- CO_2 emission per capita
- democracy index
- region (6 different regions)
- sub-region (22 different sub-regions)

Metadata sources:

- GapMinder
- World Bank
- Our World in Data
- UN Statistics Division

Text statistics

final corpora

>10k
speech texts

+

10
additional features

for all speeches,
where they are available

statistical text analysis

>60
text statistics

for all speeches

Descriptive statistics:

- counts: #tokens, #unique tokens, #characters, #sentences
- sentence length: mean, median, std
- token length: mean, median, std

Readability measures

Proportion of different POS

Coherence measures

Application

Speech Viewer	Analysis Over Years	Speech Attributes	Speech Comparer

nlp-unga-debates-g2o6gnzttq-lm.a.run.app/

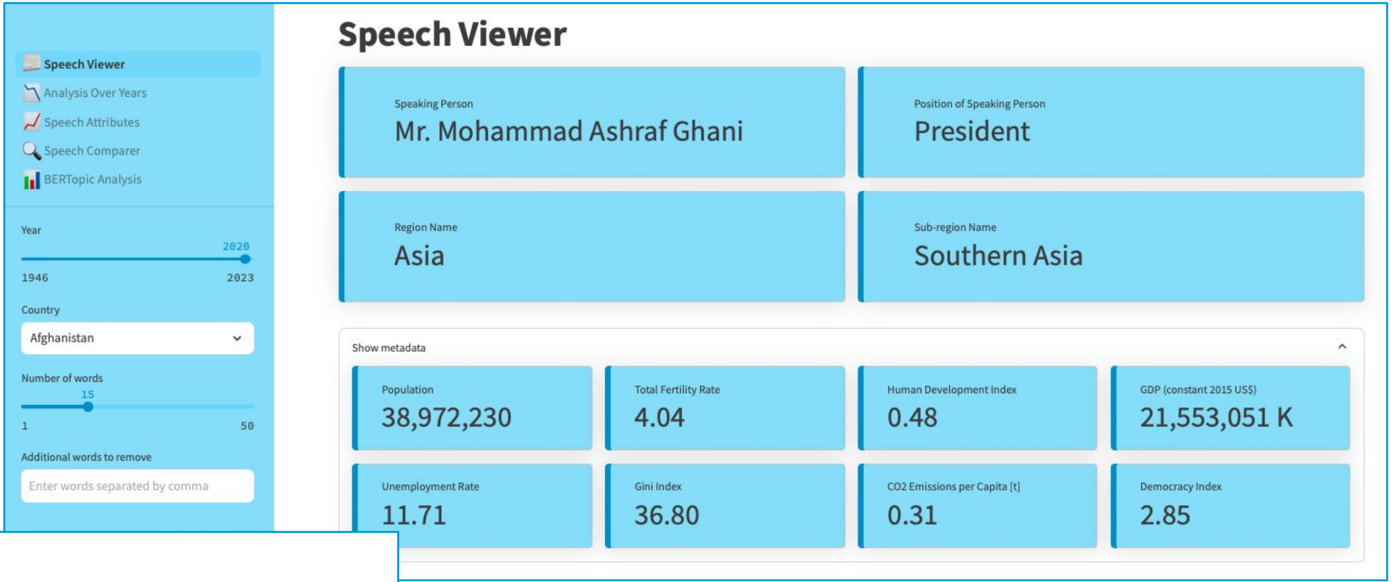
+

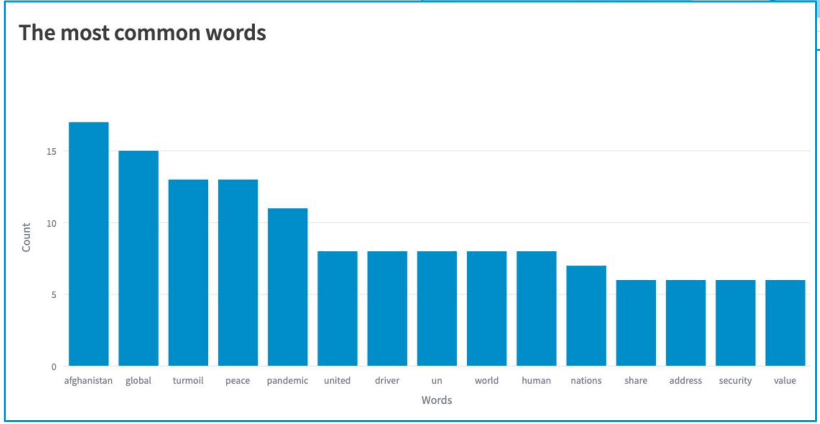
BERTopic Analysis

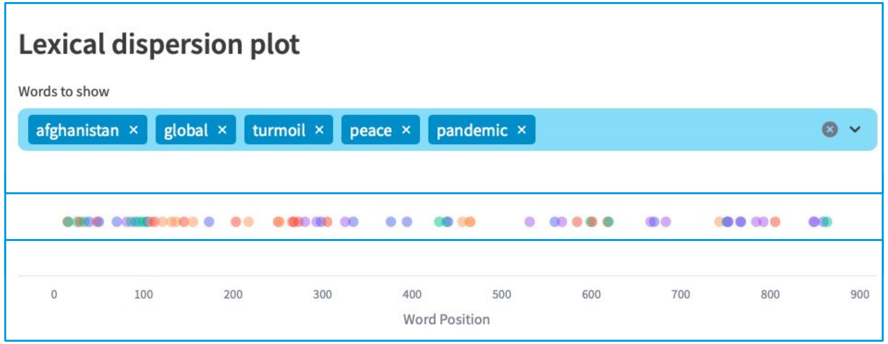
(later)

Application

Speech Viewer







Application

Analysis Over Years

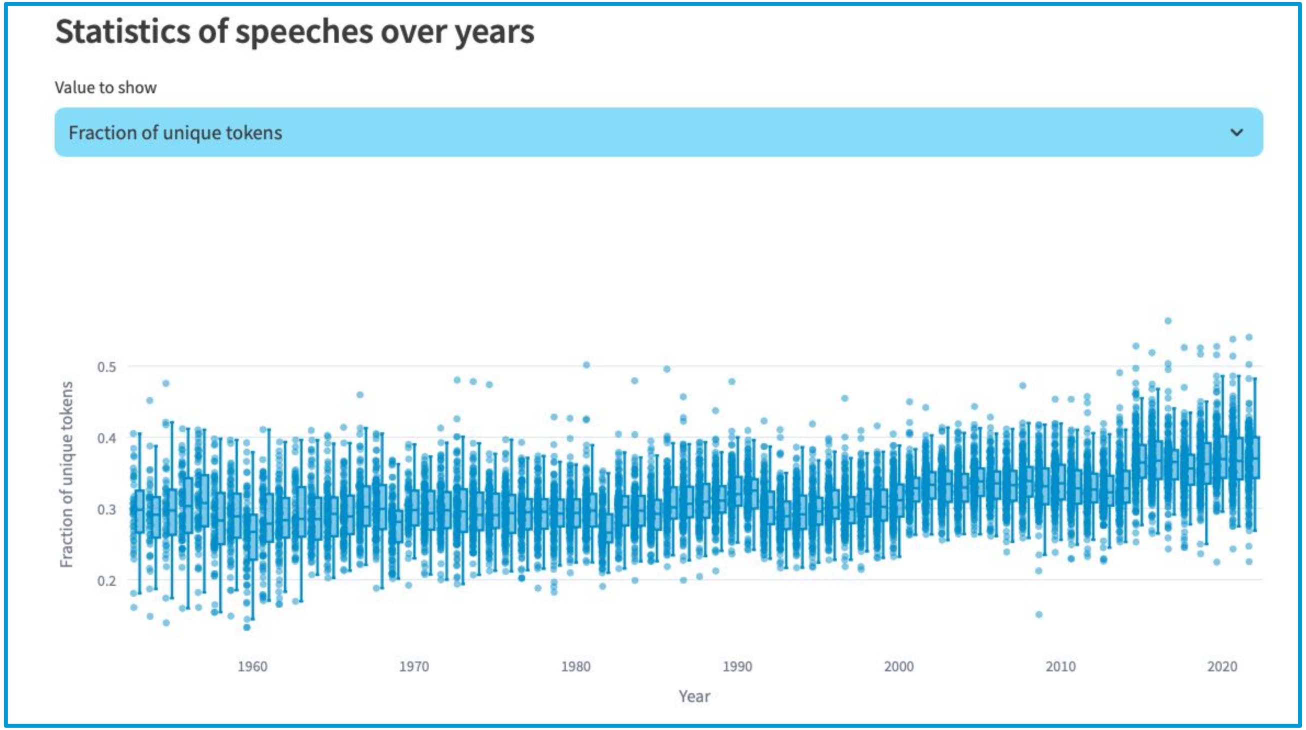
Among the most common words:

- in 1946-1955:

peace, war, Charter, Soviet, USA

- in 2001-2010:

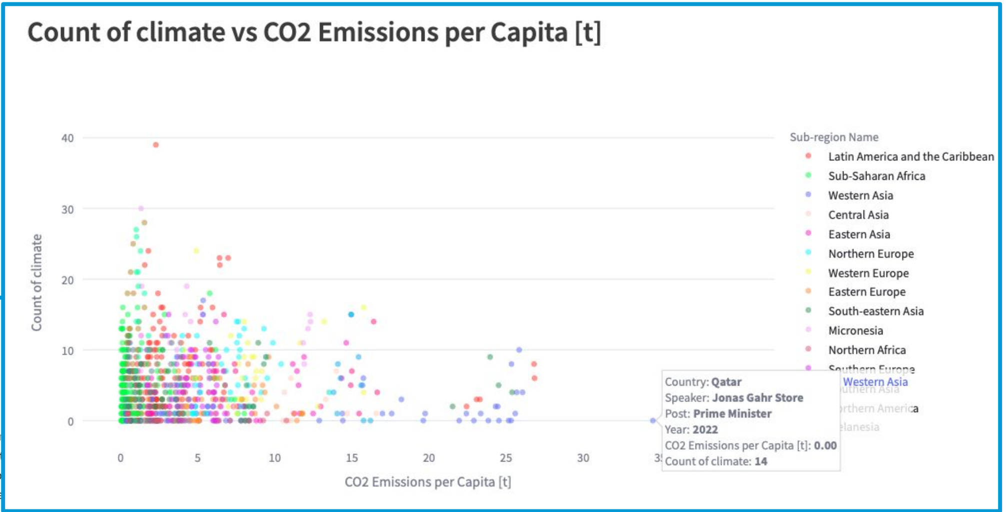
development, security, peace, community, terrorism



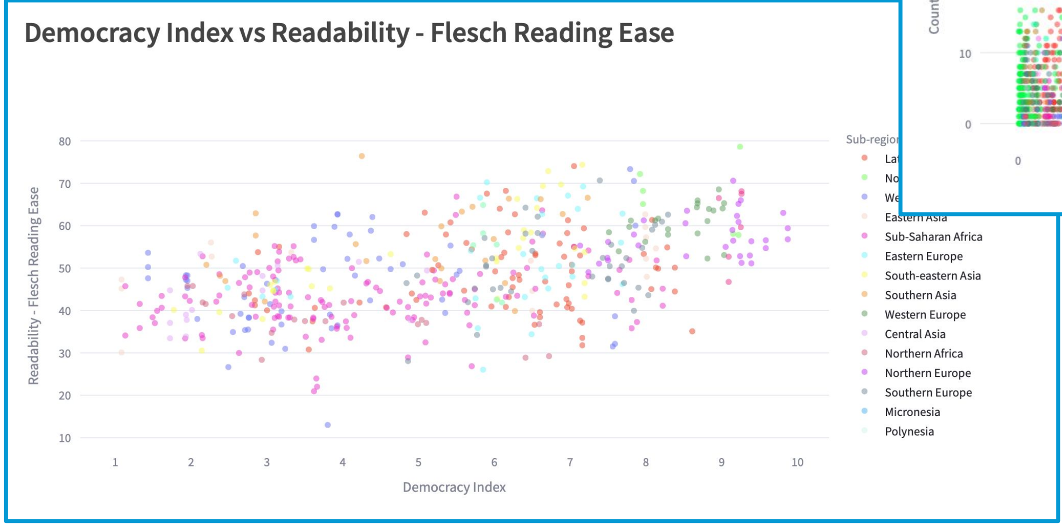
The fraction of unique tokens is increasing → there are more different issues discussed.

Application

Speech Attributes



The countries with high CO2 emissions per capita tend to not speak a lot about climate issues.



The countries with higher democracy index seem to have more readable texts of speeches.

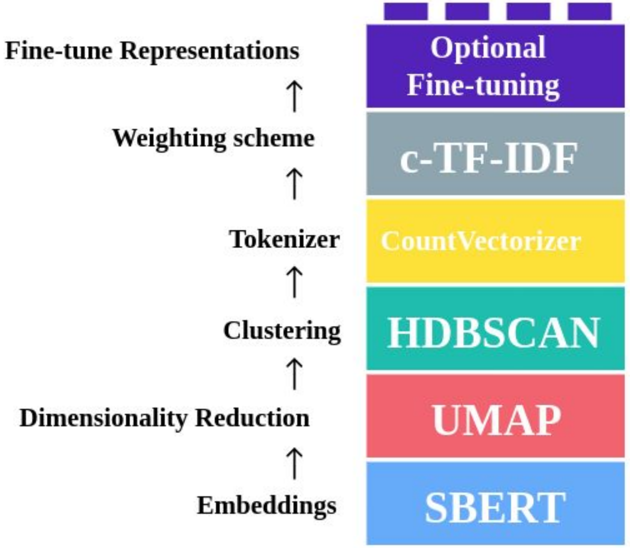
BERTopic - methodology of extracting topics

Used models:

- **LDA -** simple benchmark model
- **BERTopic -** state-of-the-art transformer-based model

  Embeddings based on sentence transformers:
  - **all-Mini-LM-L6-v2**
  - **all-Mini-LM-L12-v2**
  - **all-mpnet-base-v2**

  Embeddings based on BERT:
  - **roberta**
  - **distilbert**
- Compared numerically with topic-modeling metrics

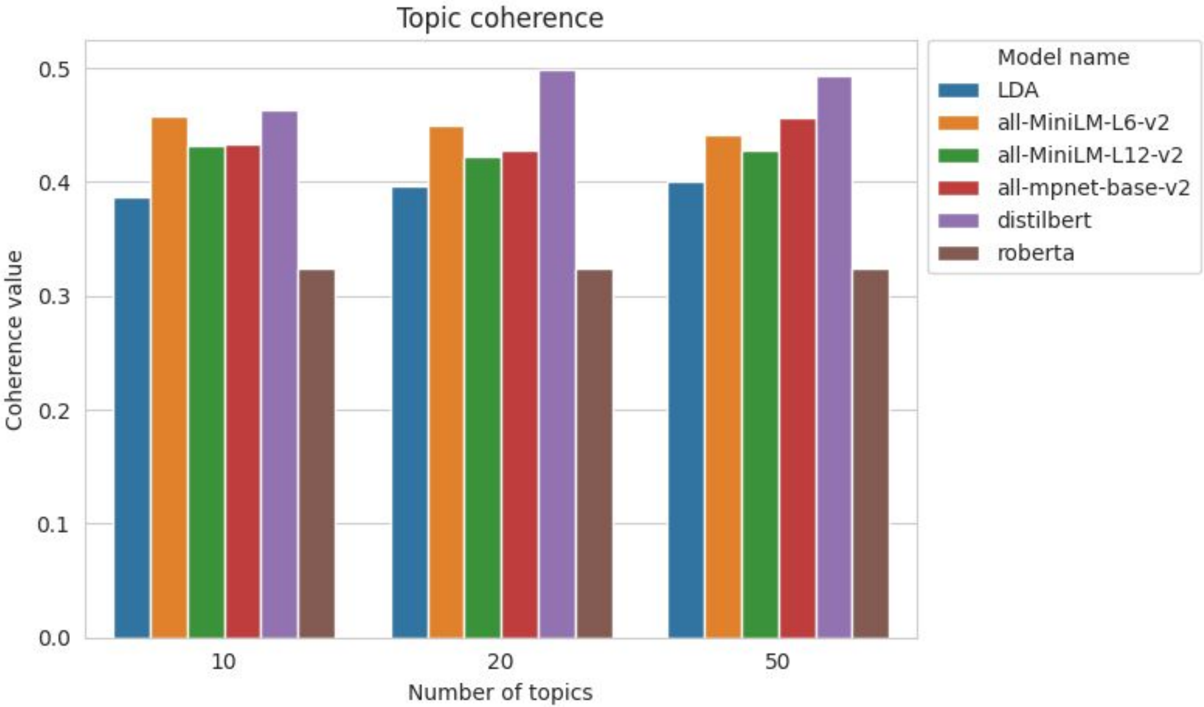


Grootendorst, M. (2022). **BERTopic: Neural topic modeling with a class-based TF-IDF procedure.** arXiv preprint arXiv:2203.05794.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). **Latent dirichlet allocation.** Journal of machine Learning research.

Model evaluation - topic coherence

Topic coherence metrics utilize various statistics drawn from the reference corpus to evaluate how well the extracted topics are 'supported' by it.

In other words, this measure indicates the degree of 'interpretability' of the obtained topics in context of their source.

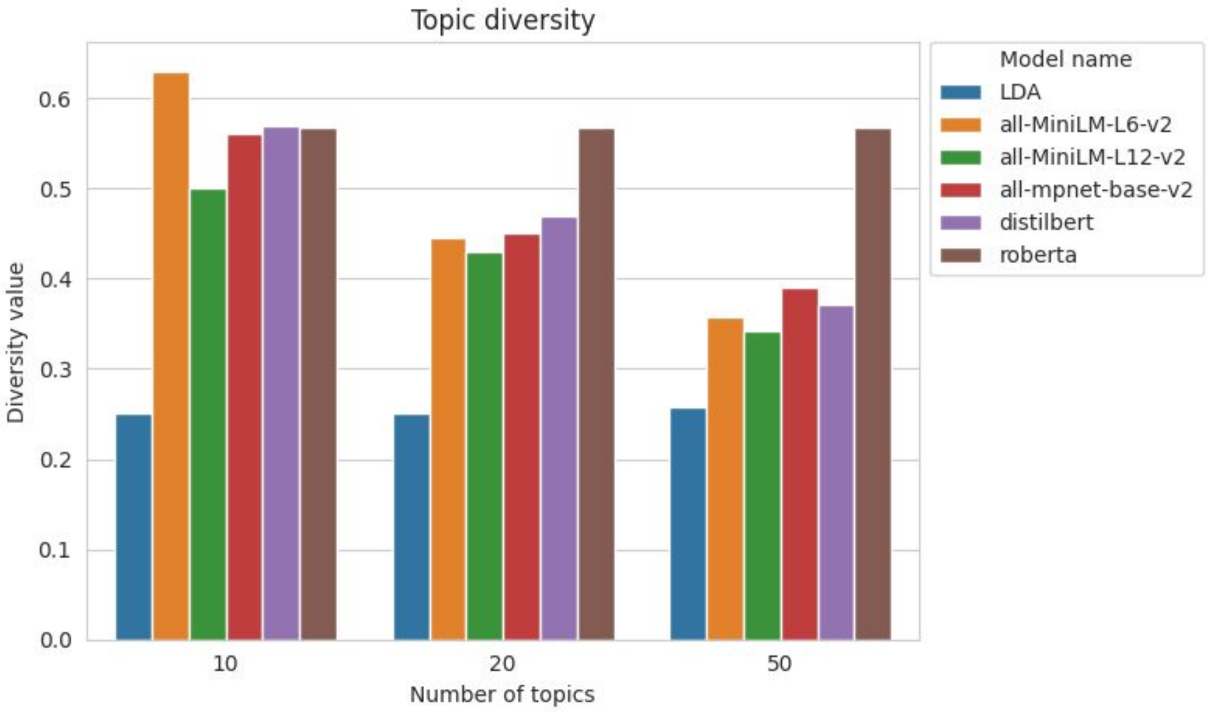


* Results of the BERTopic model with the RoBERTa embeddings are biased due to it extracting only 6 distinct topics

Mimno et al. (2011). **Optimizing Semantic Coherence in Topic Models.** Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 262–272

Model evaluation - topic diversity

In contrast to topic coherence, the topic diversity metric is calculated based solely on the extracted topics. By counting unique words in the top words of each topic and aggregating this information it evaluates how much variability there is among the topics.



* Results of the BERTopic model with the RoBERTa embeddings are biased due to it extracting only 6 distinct topics

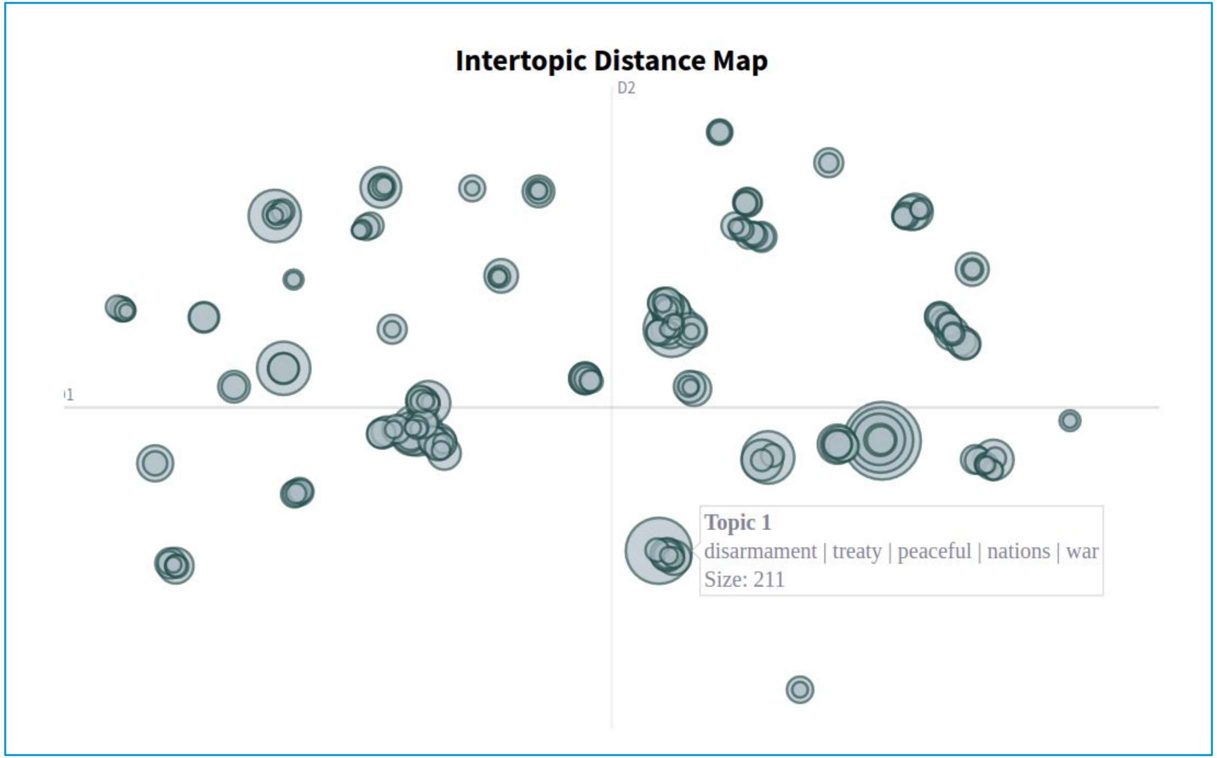
Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). **Topic Modeling in Embedding Spaces.** Transactions of the Association for Computational Linguistics, 8:439–453

High level topic visualization

The best models found **221 topics**

Selected topics can be studied via the intertopic distance map:

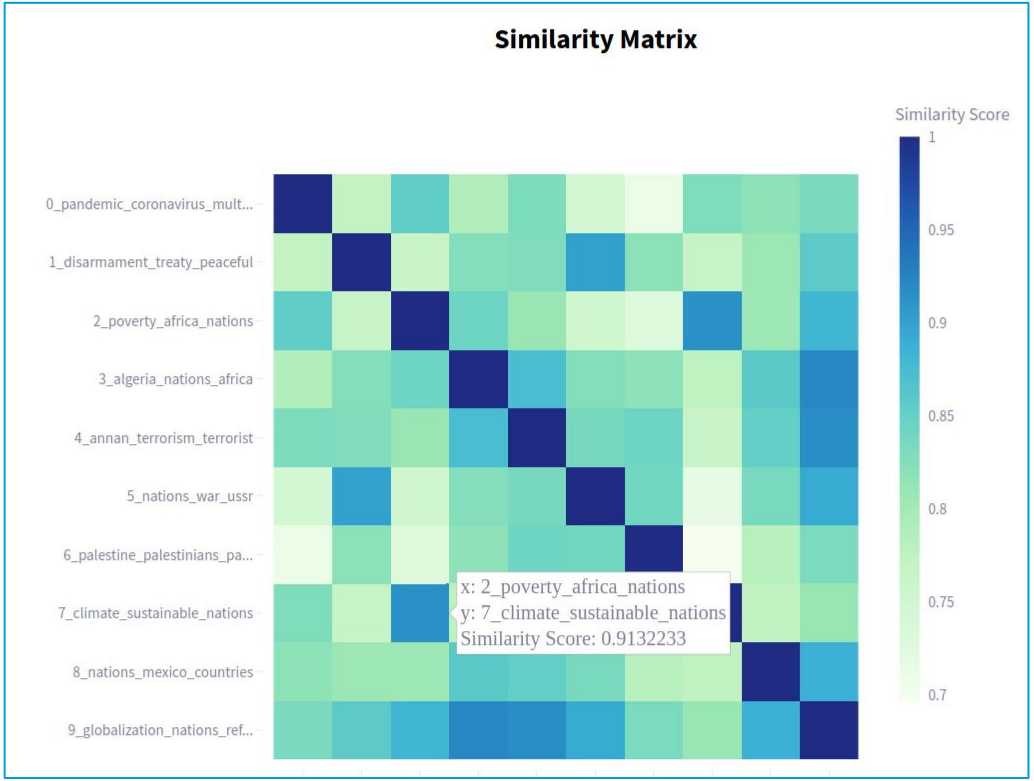
- similar topics are close together
- size of bubble corresponds to number of documents
- the plot is prepared using a dimensionality reduction technique



Topic similarity

The discovered topics can be studied in terms of their semantic similarity:

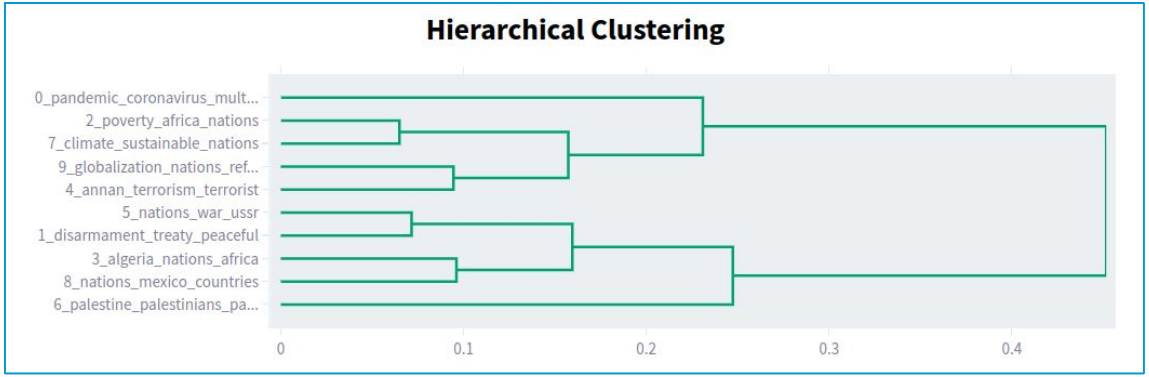
- interactive heatmap highlighting most similar topics



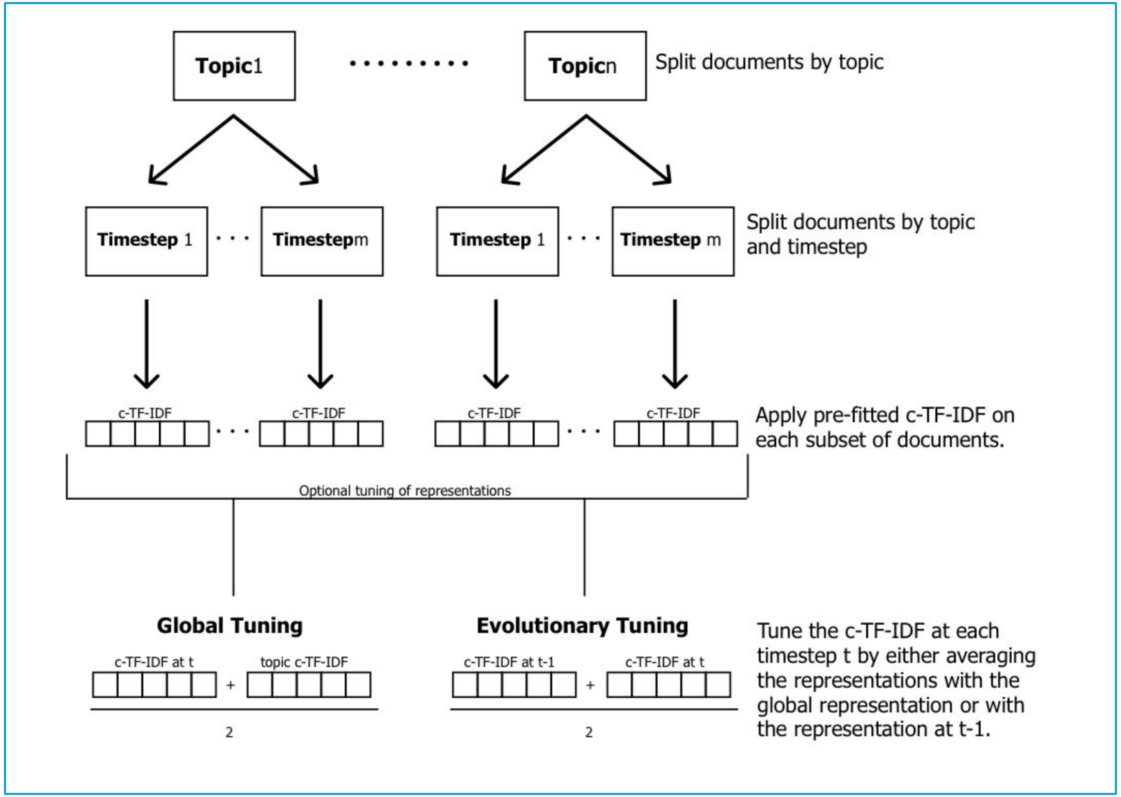
Hierarchical topic modeling

Due to the hierarchical nature of the clustering we can prepare a dendrogram of the discovered topics.

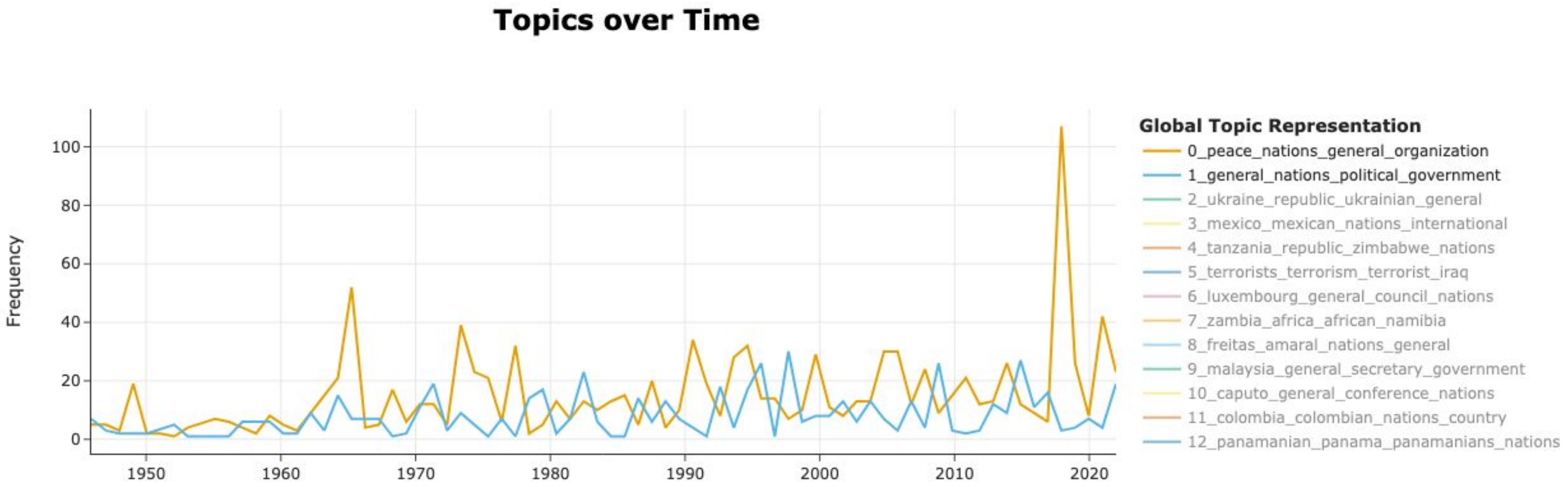
It shows how close together (in the latent topic space, not semantically) to each other are pairs of topics.



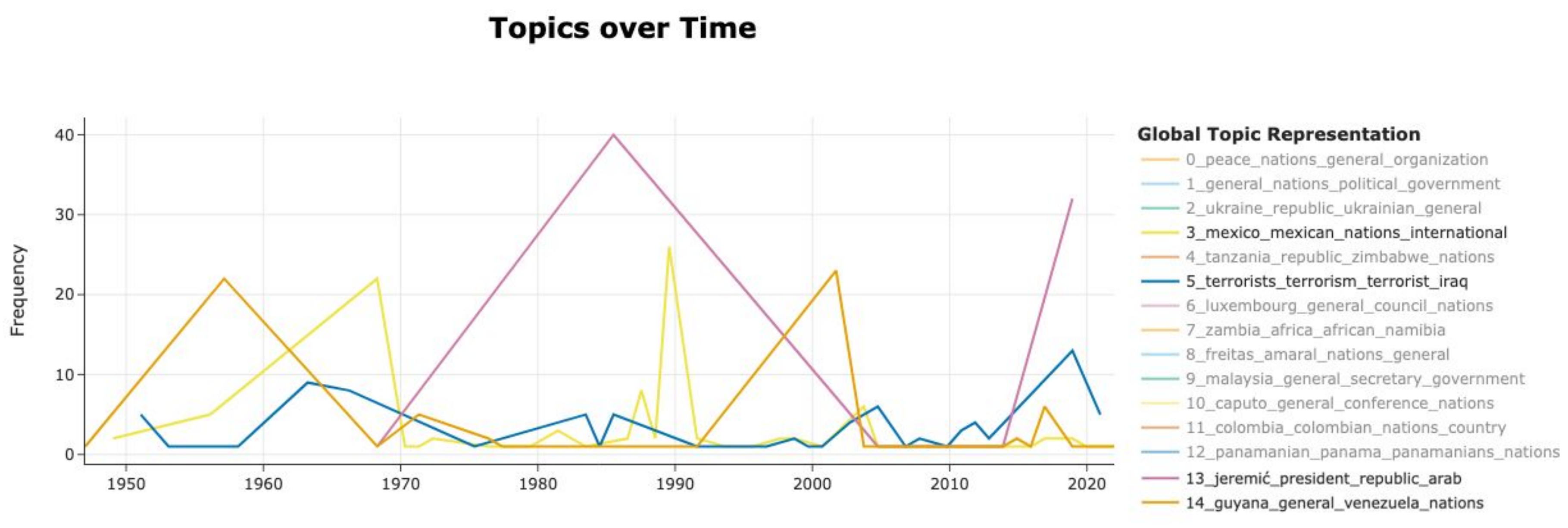
Dynamic topic modeling



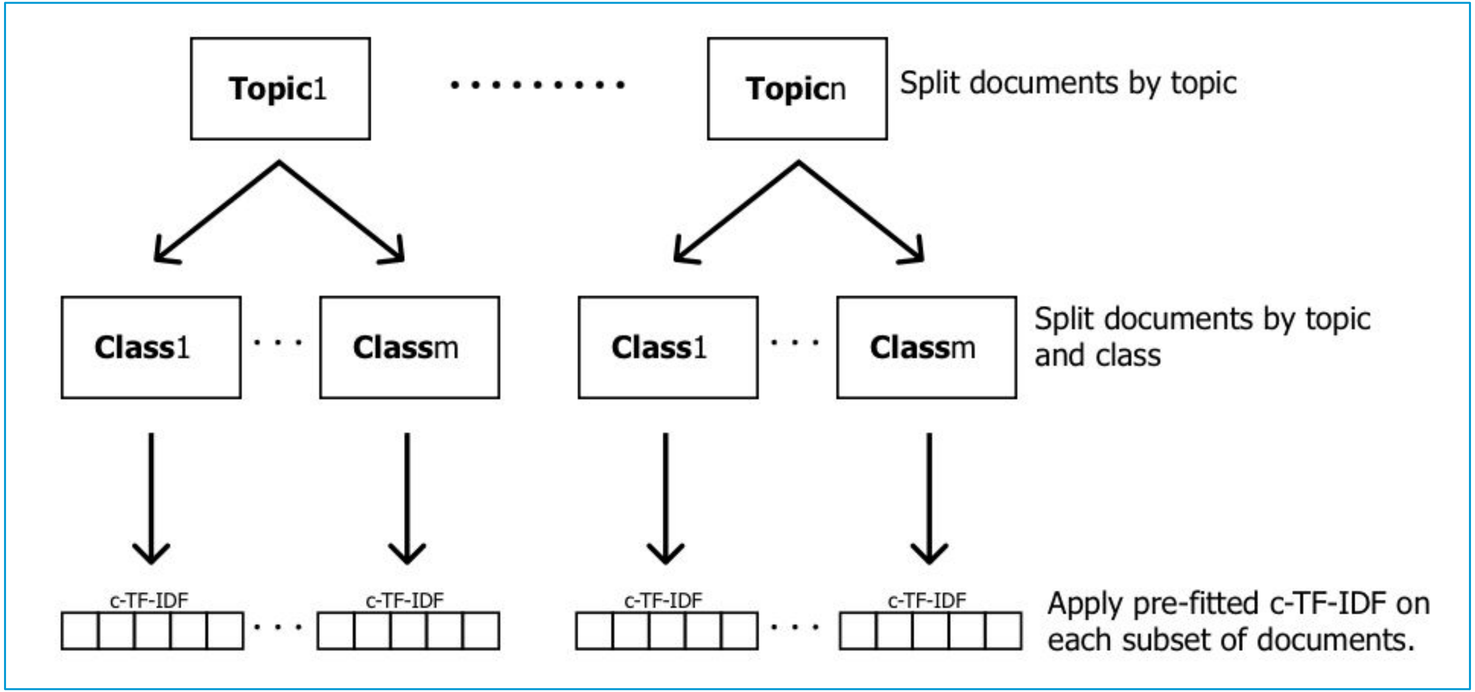
Dynamic topic modeling



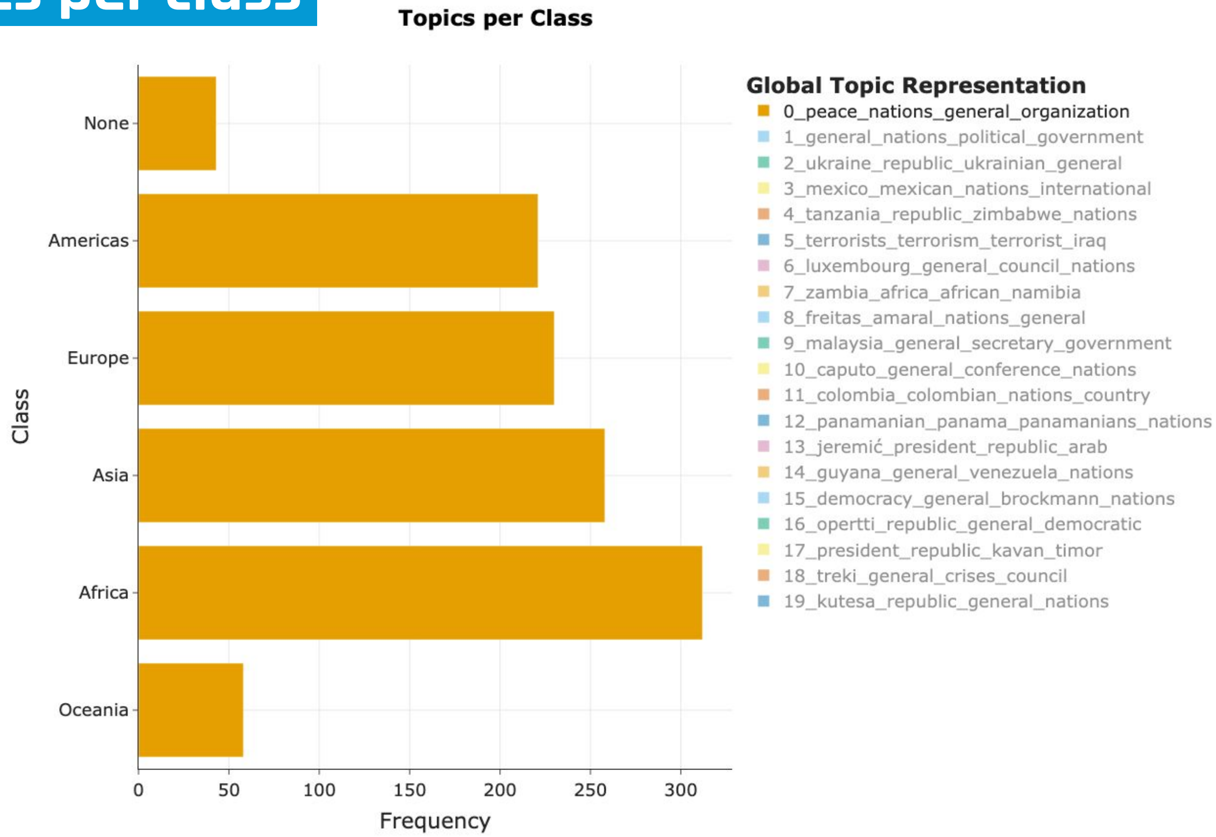
Dynamic topic modeling



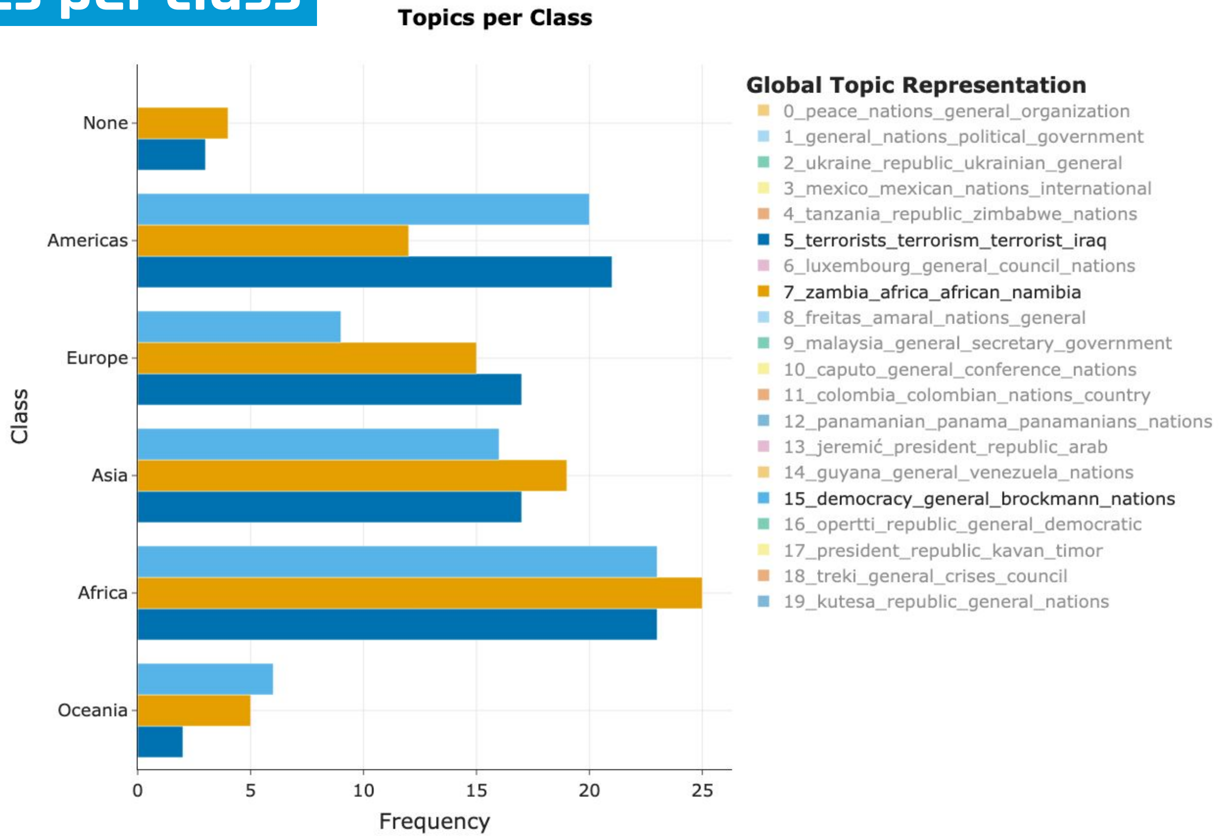
Topics per class



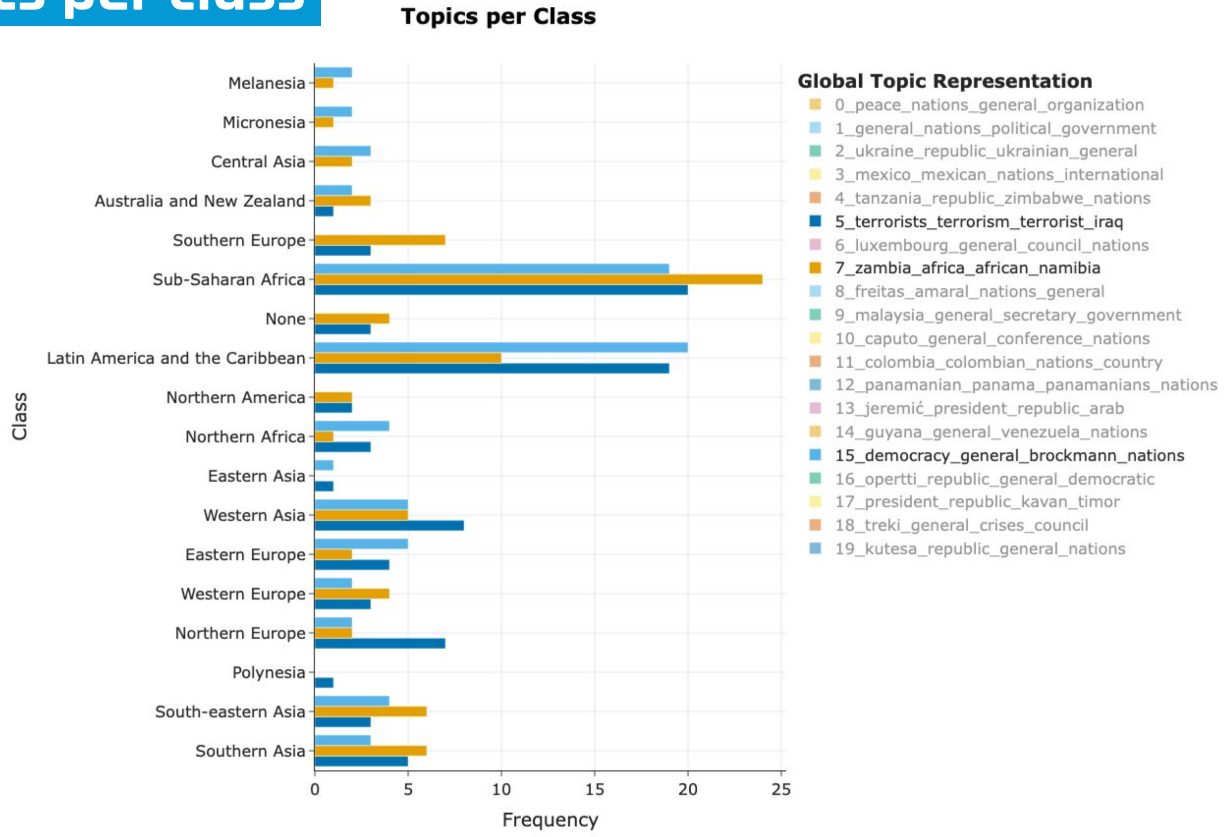
Topics per class



Topics per class



Topics per class



Summary

- **Dataset** The UN GA general debates corpus has been carefully sanitized, supplemented with the latest speeches and enriched with over 10 additional geopolitical features.
- **Application** An interactive and containerized application has been developed to enable viewing both the raw transcripts and over 60 related text statistics together with their visualizations.
- **Modeling** The BERTopic topic modeling pipeline has been applied together with 5 different embedding methods and compared with the LDA baseline using topic coherence and diversity metrics, achieving superior results.
- **Analyses** The best BERTopic pipeline variant has been exploited to prepare a variety of topic modelling analyses, such as topic similarity, hierarchical dependencies and temporal changes.

Limitations and future work

- Lack of broad political science knowledge – analyses at a basic level
  → **preparing instructions on how to use the application for political scientists (end users)**

- Very large corpora and often long, complicated speeches
  → **preparing summarization of speech texts**
  (experiments with both extractive and abstractive summaries)

- Limiting the text features under consideration to purely statistical ones
  → **preparing additional semantic features**
  (sentiment analysis of polarity and subjectivity of all speeches)

Thank you!

Questions?

Team 13: Debates-3MB

Mateusz Grzyb
298820

Mateusz Krzyziński
305739

Bartłomiej Sobieski
305830

Mikołaj Spytek
305753

December 13th, 2023