Mining United Nations General Assembly Debates

Natural Language Processing

Project 1: proposal

Team 13: Debates-3MB

Mateusz Grzyb	Mateusz Krzyziński	Bartłomiej Sobieski	Mikołaj Sptyek
298820	305739	305830	305753

November 8th, 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

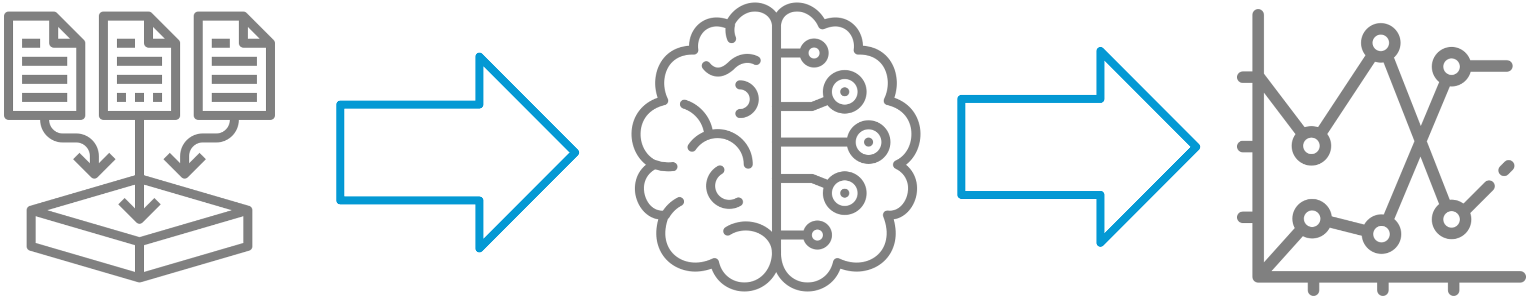
United Nations

Why do we care about UN GA debates?

- **Comparable** - globally comparable textual data on states' foreign policy preferences and priorities over time.
- **Inclusive** - all member states, including smaller and less powerful nations, have the opportunity speak.
- **Explanatory** - the statement present not only positions but also provide explanations and justifications.
- **Unfiltered** - views of governments' opinions that have not been filtered by the media or other sources.

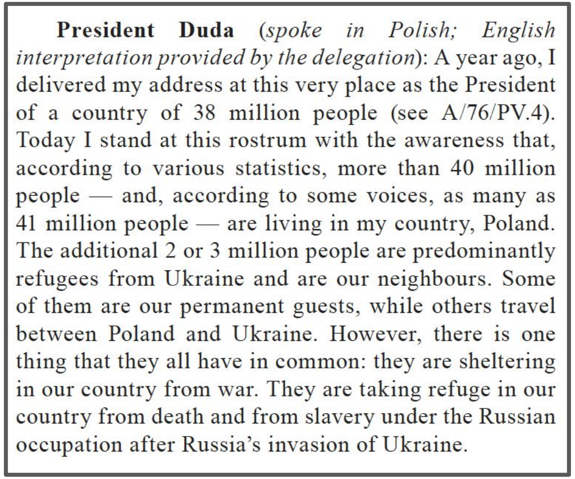
Goals of the project

- 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



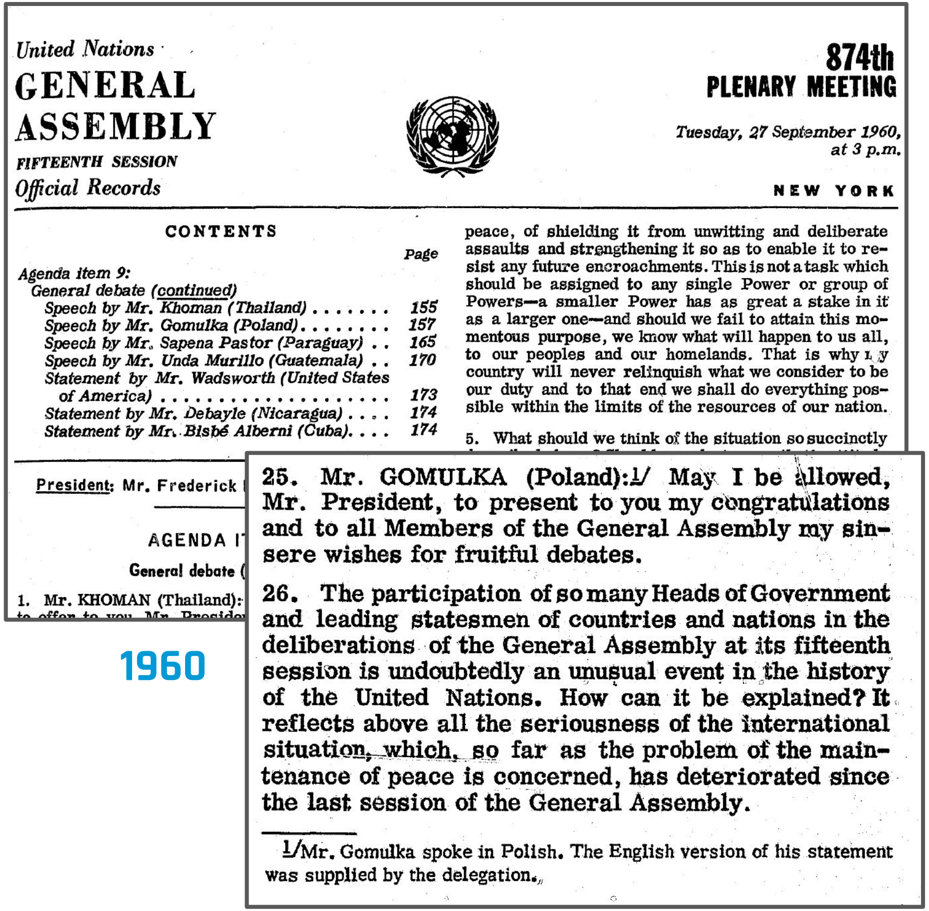
Available data

- stored in United Nations Digital Library
- documents based on plenary meetings, not individual statements
- various formatting, prior to 1992 image copies
- in English (and 5 other official UN languages)



President Duda (*spoke in Polish; English interpretation provided by the delegation*): A year ago, I delivered my address at this very place as the President of a country of 38 million people (see A/76/PV.4). Today I stand at this rostrum with the awareness that, according to various statistics, more than 40 million people — and, according to some voices, as many as 41 million people — are living in my country, Poland. The additional 2 or 3 million people are predominantly refugees from Ukraine and are our neighbours. Some of them are our permanent guests, while others travel between Poland and Ukraine. However, there is one thing that they all have in common: they are sheltering in our country from war. They are taking refuge in our country from death and from slavery under the Russian occupation after Russia's invasion of Ukraine.

2022



United Nations
GENERAL ASSEMBLY
FIFTEENTH SESSION
Official Records

874th PLENARY MEETING

Tuesday, 27 September 1960, at 3 p.m.

NEW YORK

CONTENTS

President: Mr. Frederick

AGENDA I

General debate (

1. Mr. KHOMAN (Thailand):
to offer to you Mr. Preside

peace, of shielding it from unwitting and deliberate assaults and strengthening it so as to enable it to resist any future encroachments. This is not a task which should be assigned to any single Power or group of Powers—a smaller Power has as great a stake in it as a larger one—and should we fail to attain this momentous purpose, we know what will happen to us all, to our peoples and our homelands. That is why ı y country will never relinquish what we consider to be our duty and to that end we shall do everything possible within the limits of the resources of our nation.

5. What should we think of the situation so succinctly

25. Mr. GOMULKA (Poland):1/ May I be allowed, Mr. President, to present to you my congratulations and to all Members of the General Assembly my sinsere wishes for fruitful debates.

26. The participation of so many Heads of Government and leading statesmen of countries and nations in the deliberations of the General Assembly at its fifteenth session is undoubtedly an unusual event in the history of the United Nations. How can it be explained? It reflects above all the seriousness of the international situation, which, so far as the problem of the maintenance of peace is concerned, has deteriorated since the last session of the General Assembly.

1/Mr. Gomulka spoke in Polish. The English version of his statement was supplied by the delegation.

1960

Corpora

UN GD corpus (UNGDC)

first version

- ❑ Baturo et al. (2017)
  - ❑ >7,300 GD statements
  - ❑ years: 1970-2014
  - ❑ dataset available on Kaggle

current version

- ❑ Jankin et al. (2023)
  - ❑ >10,000 GD statements
  - ❑ years: 1946-2022 (*currently*)
  - ❑ dataset available on Harvard Dataverse

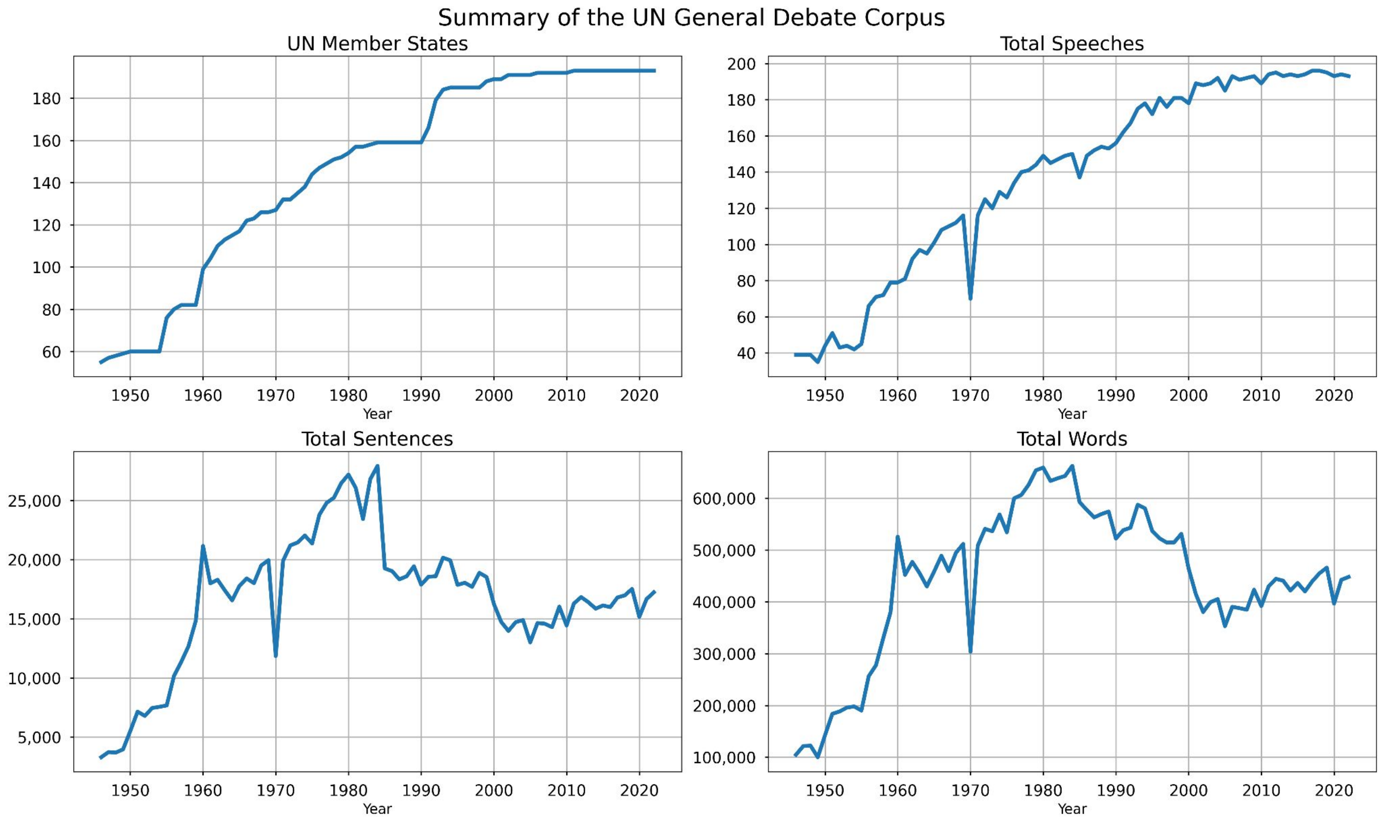
+ basic metadata

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). **Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus.** Research & Politics, 4(2)

Jankin, S., Baturo, A., & Dasandi, N. (2017). **United Nations General Debate Corpus 1946-2022 (Version V11).**

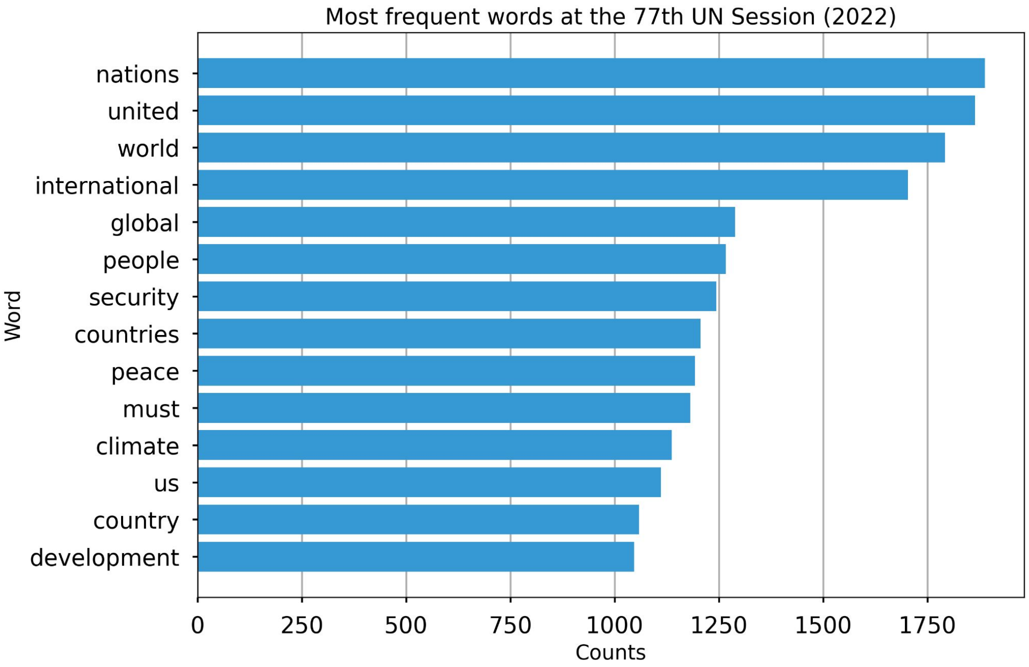
Dasandi, N., Jankin, S., & Baturo, A. (2023). **Words to Unite Nations: The Complete UN General Debate Corpus, 1946-Present.** OSF.

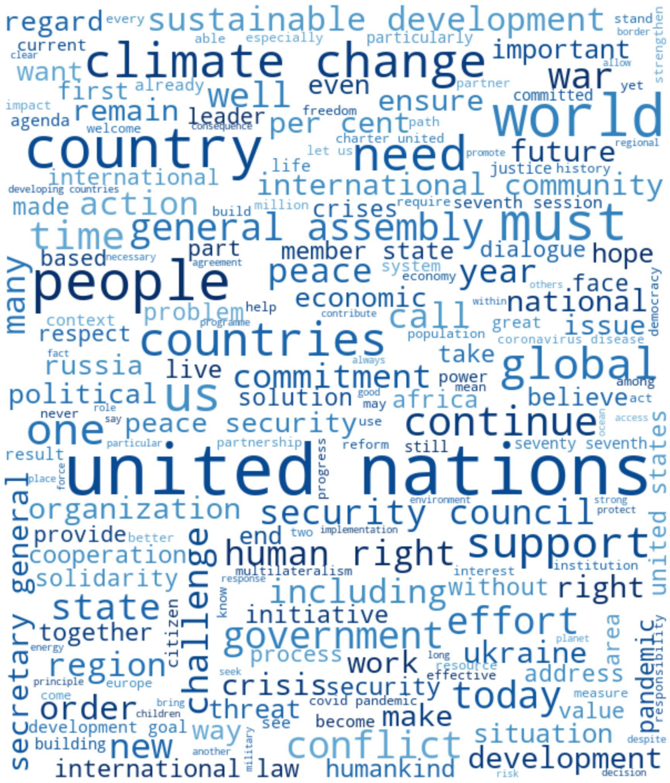
Preliminary Exploratory Data Analysis



Preliminary Exploratory Data Analysis

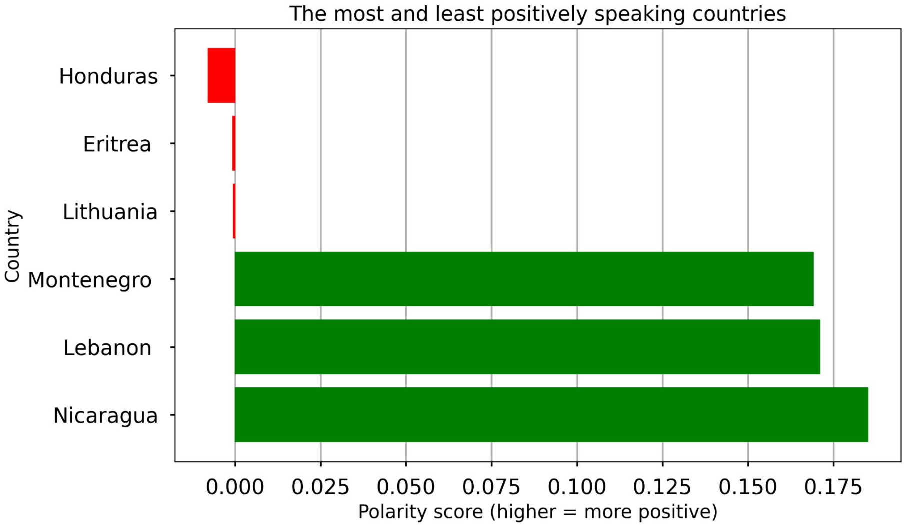
(77th Session, 2022)

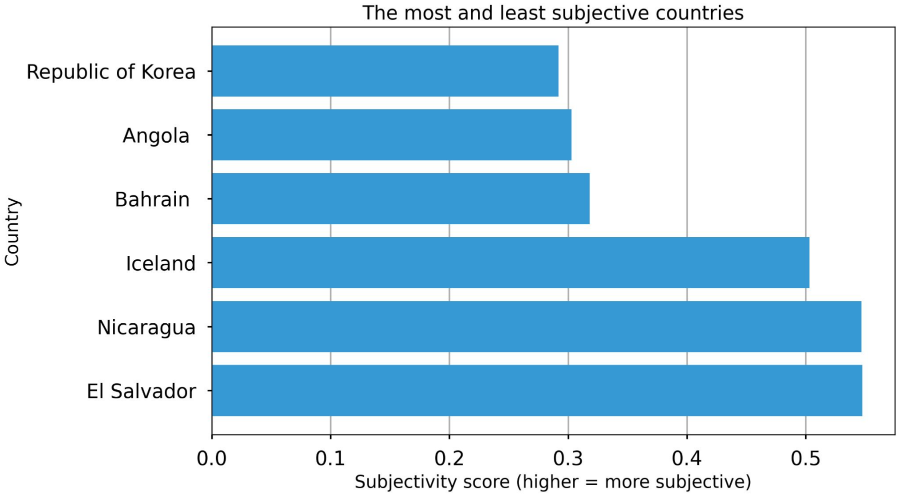




Preliminary Exploratory Data Analysis

(77th Session, 2022)





Topic modeling

- One of the **most commonly employed** NLP methods in social science.
- **Used for:**
  - data exploration
  - quantitative analysis
  - comparing documents (semantically)

- **Models:**
  - Latent Dirichlet Allocation
  - Non-negative Matrix Factorization
  - Bayesian Hierarchical Topic Model
  - **BERTopic**

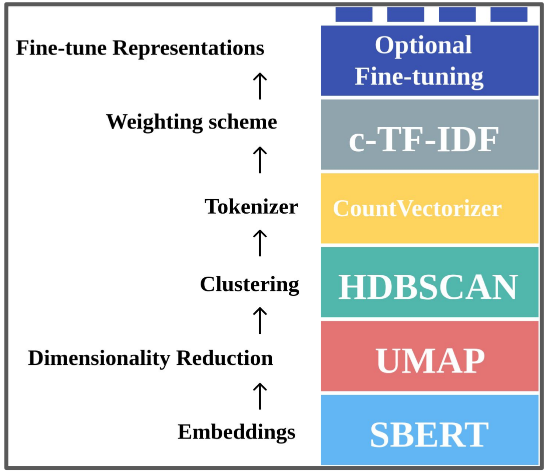
  . . .

- **Packages/frameworks:**
  - Gensim
  - ToModAPI
  - OCTIS
  - ITMT
  - HADES

  . . .

State of the art in topic modeling

- **BERTopic** (Grootendorst, 2022)
- Uses **embeddings from pre-trained transformers** to utilize **semantic meaning** of words in topic selection
- Consists of **4 major steps**:
  - acquisition of document embeddings
  - dimensionality reduction
  - hierarchical clustering
  - creation of topic descriptions

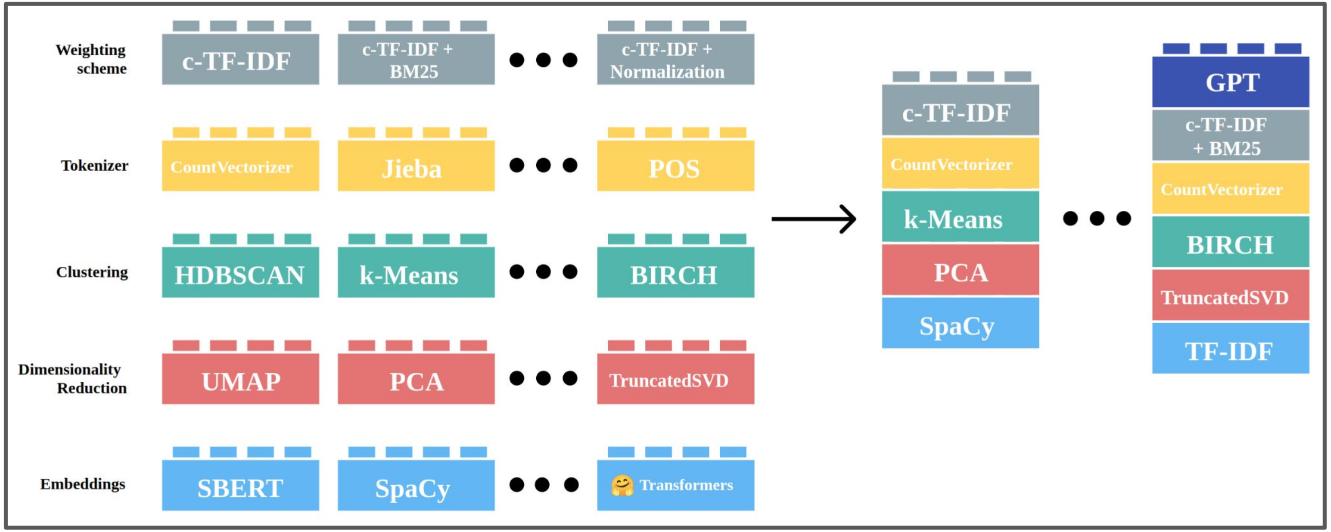


Grootendorst, M. (2022). **BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure.** arXiv Preprint arXiv:2203. 05794.

State of the art in topic modeling

- Even though BERTopic was proposed as a specific model, it can be thought of as a general **framework with swappable parts**.

- Risk minimization – parts with problems with implementation can be changed.



Grootendorst, M. (2022). **BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure.** arXiv Preprint arXiv:2203. 05794.

Dynamic Topic Modeling

- Basic topic modeling techniques find **general themes** present in the entire corpus
- Sometimes the ways these topics are spoken about changes **over time**.
- Dynamic topic modeling allows us to compare these **changes** and more easily track their **evolution** as well as their **relative frequency** in the documents.

- The BERTopic implementation offers the **dynamic topic modeling** framework.

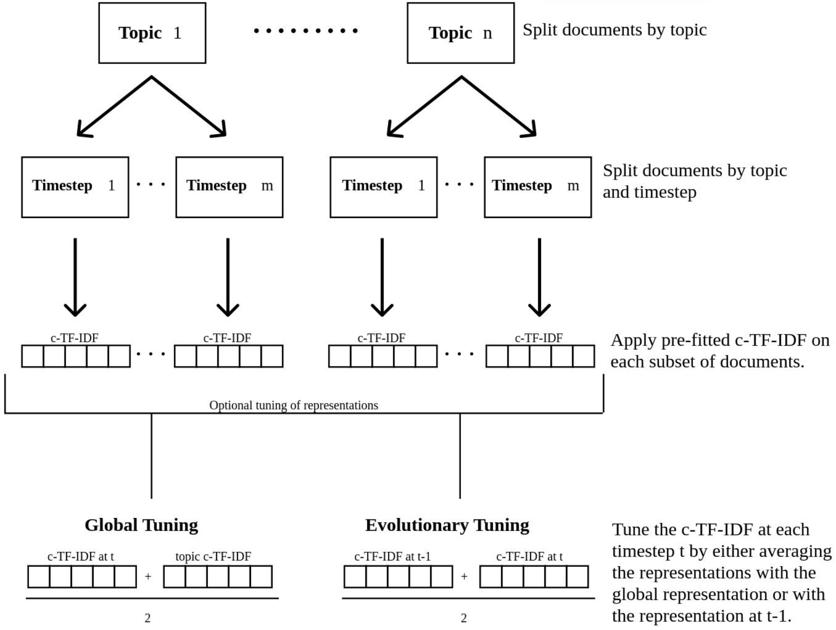

Topic 1
Topic n
Split documents by topic
Timestep 1
Timestep m
Timestep 1
Timestep m
Split documents by topic and timestep
c-TF-IDF
c-TF-IDF
c-TF-IDF
c-TF-IDF
Apply pre-fitted c-TF-IDF on each subset of documents.
Optional tuning of representations
Global Tuning
c-TF-IDF at t
topic c-TF-IDF
2
Evolutionary Tuning
c-TF-IDF at t-1
c-TF-IDF at t
2
Tune the c-TF-IDF at each timestep t by either averaging the representations with the global representation or with the representation at t-1.

Diagram showcasing how to use dynamic topic modeling with BERTopic

Thank you!

Questions?

Team 13: Debates-3MB

Mateusz Grzyb	Mateusz Krzyziński	Bartłomiej Sobieski	Mikołaj Sptyek
298820	305739	305830	305753

November 8th, 2023