Introduction to Topic Modeling and BERTopic

CLARIAH-AT Summer School:
Machine Learning for Digital Scholarly Editions

8th-12th, September 2025

Selina Galka

Outline

- What is Topic Modeling?
    - Topics and Documents
    - History, Workflow, Topic Modeling Algorithms
    - Use Cases / Examples
- BERTopic
    - How does it work?
    - Components
    - Strengths
- Example: Quick Start with BERTopic

Introduction

- **BERTopic** is a Topic modeling technique
- **Topic modeling**
    - **unsupervised machine learning technique** used in natural language processing to uncover the hidden thematic structure—the main topics—within large collections of texts
    - a **topic** is a **statistically derived cluster of words** that frequently co-occur across documents within a text corpus
    - Quantitative method of text analysis or examining large amounts of text
    - Goal: to discover “hidden” semantic structures

Topic modeling

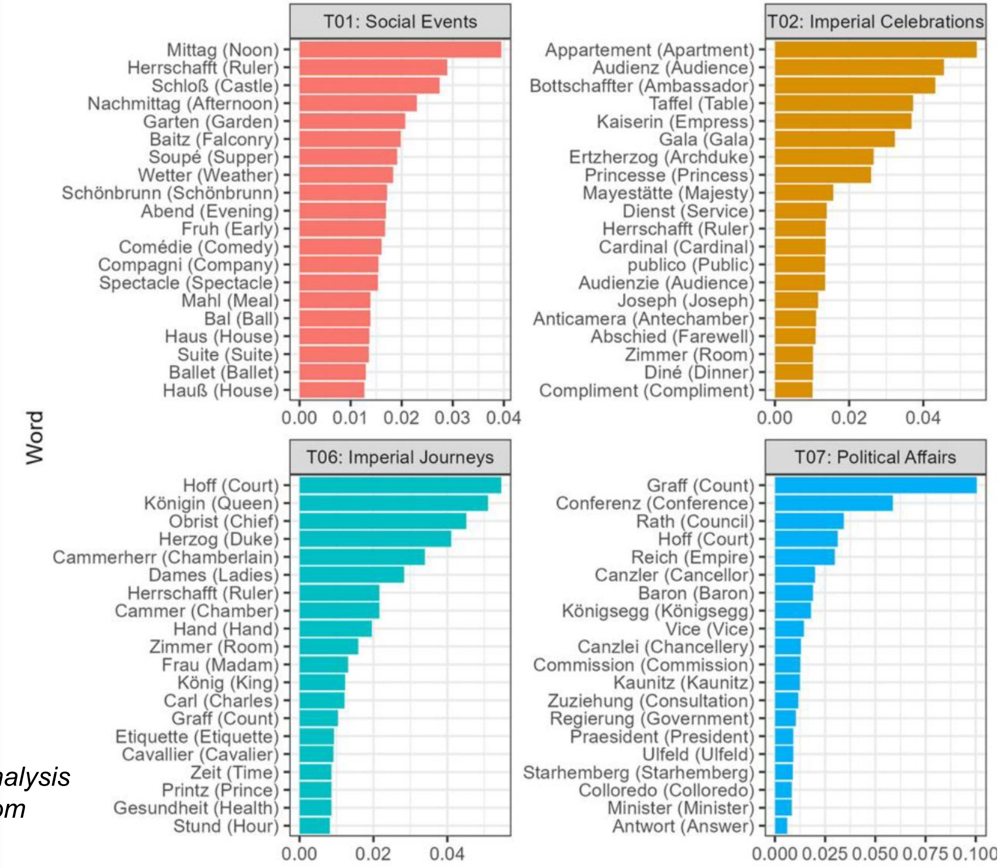
Topic modeling

- Identifying recurring themes, motifs, and discourses automatically
- Without explicit semantic knowledge

"Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time."
David M. Blei (2011)

Topics

- Cluster of Words: Topics are not pre-defined labels, but sets of words
- Each topic consists of a weighted list of words, where the weights reflect the probability that a given word belongs to the topic

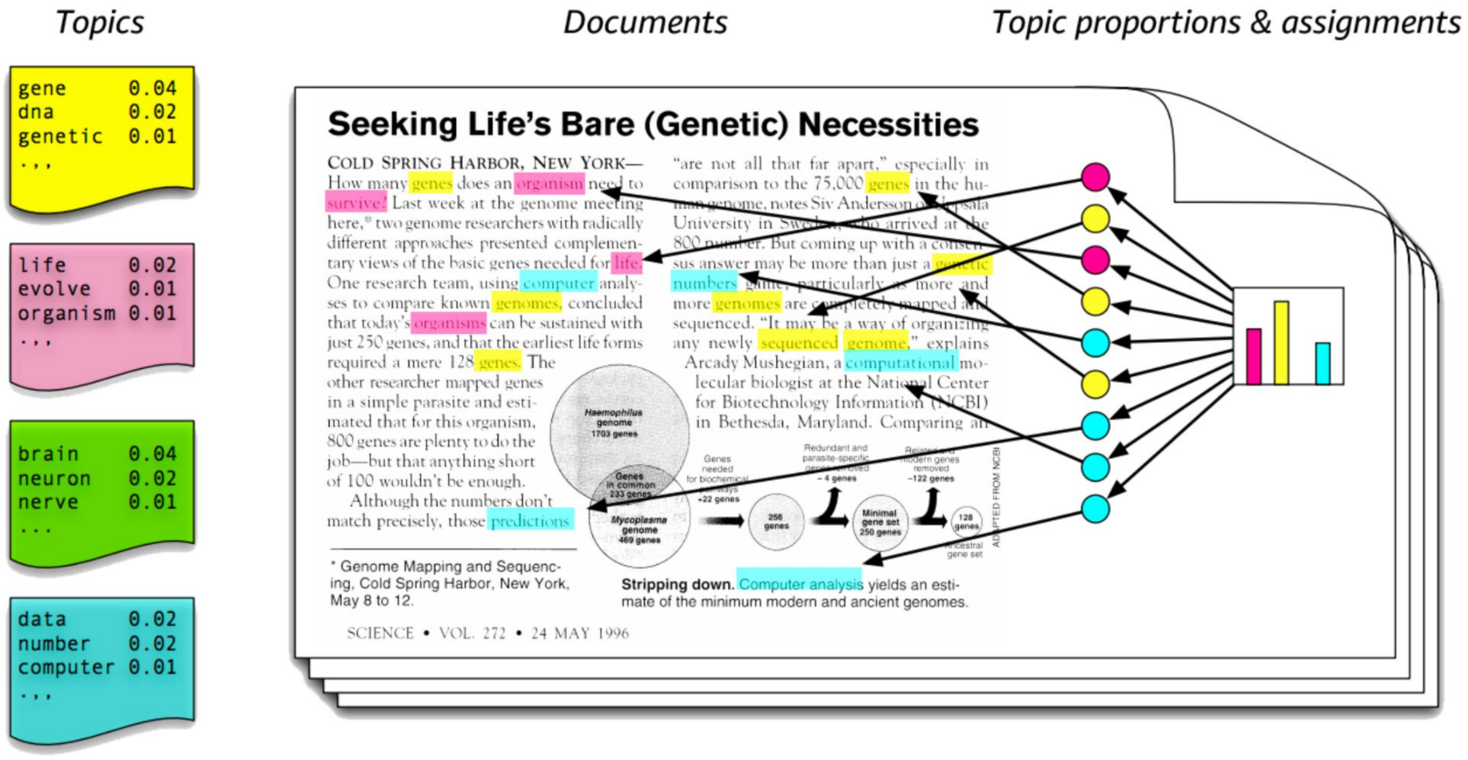


Yamashita/Uchida. 2025. 'Historical Analysis Using Topic modeling: Insights from Khevenhüller's Diary 1742–76'

Documents

A **document** is the fundamental unit of text that the algorithm analyzes to uncover latent (hidden) topics.

- A **document** is a sequence of words (tokens) that is treated as one item in a corpus (the full collection of documents).
- Depending on the context and application, a "document" could mean:
  - A whole research paper, news article, or book chapter.
  - A paragraph, a sentence, or even a tweet.
  - Any chunk of text chosen as the unit for analysis.



David M. Blei, Probabilisitic Topic Models, in: *Communications of the ACM* 55 (2012).

History

Early Foundations (Pre-2000)

- Latent Semantic Analysis (LSA) (Deerwester et al., 1990)
  - One of the earliest approaches.
  - Used Singular Value Decomposition (SVD) to reduce dimensionality of term-document matrices.
  - Captured word co-occurrence patterns but lacked a probabilistic foundation.

History

Probabilistic Models (2000s)

- Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999)

- Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003)
    - Landmark development: Bayesian probabilistic model.
    - Each document is a mixture of topics; each topic is a distribution over words.
    - Most widely used "classical" topic model.

History

Neural Topic Models (2015–present)

- Neural Variational Inference: Use of variational autoencoders (VAEs) for topic modeling.
    - Example: Neural Variational Document Model (NVDM).
- Neural Topic Models (NTM): Combine deep learning with probabilistic frameworks.
- Word Embedding Integration: Instead of bag-of-words, models use word2vec, GloVe, BERT embeddings.
    - Example: ProdLDA (2017), ETM (Embedded Topic Model, 2018).

History

Transformer & Contextual Models (2020–present)

- BERTopic (2020): Uses BERT embeddings + clustering + class-based TF-IDF for interpretable topics.
- Top2Vec (2020): Jointly learns document and topic embeddings using deep learning.
- Contextualized Topic Models (CTM, 2020): Mix contextual embeddings (BERT) with probabilistic models.

📖 *Hankar, Mustapha, Mohammed Kasri, and Abderrahim Beni-Hssane. 2025. 'A Comprehensive Overview of Topic Modeling: Techniques, Applications and Challenges'. Neurocomputing 628 (May): 129638. https://doi.org/10.1016/j.neucom.2025.129638.*

📖 *Egger, Roman, and Joanne Yu. 2022. 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts'. Frontiers in Sociology 7 (May): 886498. https://doi.org/10.3389/fsoc.2022.886498.*

Data collection

Collection of text data

Preprocessing

e.g. Tokenization, Stopword Removal, Lemmatization (can also happen latter when using embeddings)

Text Representation and Topic modeling

texts are converted into a mathematical form

this varies on the method used: e.g. LDA with bag-of-words concept or BERTopic with embeddings

finding hidden topic structures depending on method used

Interpretation

each topic is a probability distribution of words (or a cluster of embeddings)

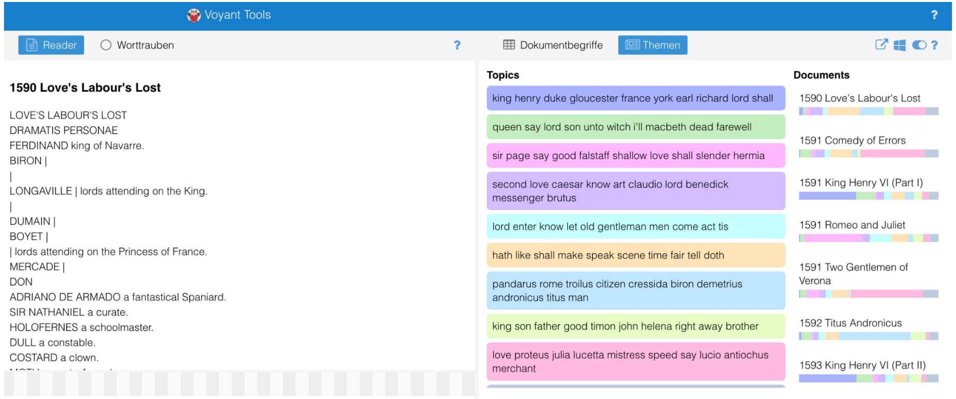
the keywords describe the topic

visualizations (e.g. with pyLDAvis)

Finetuning

e.g. merge or split topics, remove uninformative words (stopwords, jargon), re-label topics for interpretability.

Topic modeling Tools



https://voyant-tools.org/?corpus=shakespeare

- MALLET → Java-based,

  fast implementation of LDA and other models.
- DARIAH Topic Explorer → GUI-based, tailored for humanities research.
- Voyant Tools → Web-based text analysis (includes topic modeling).

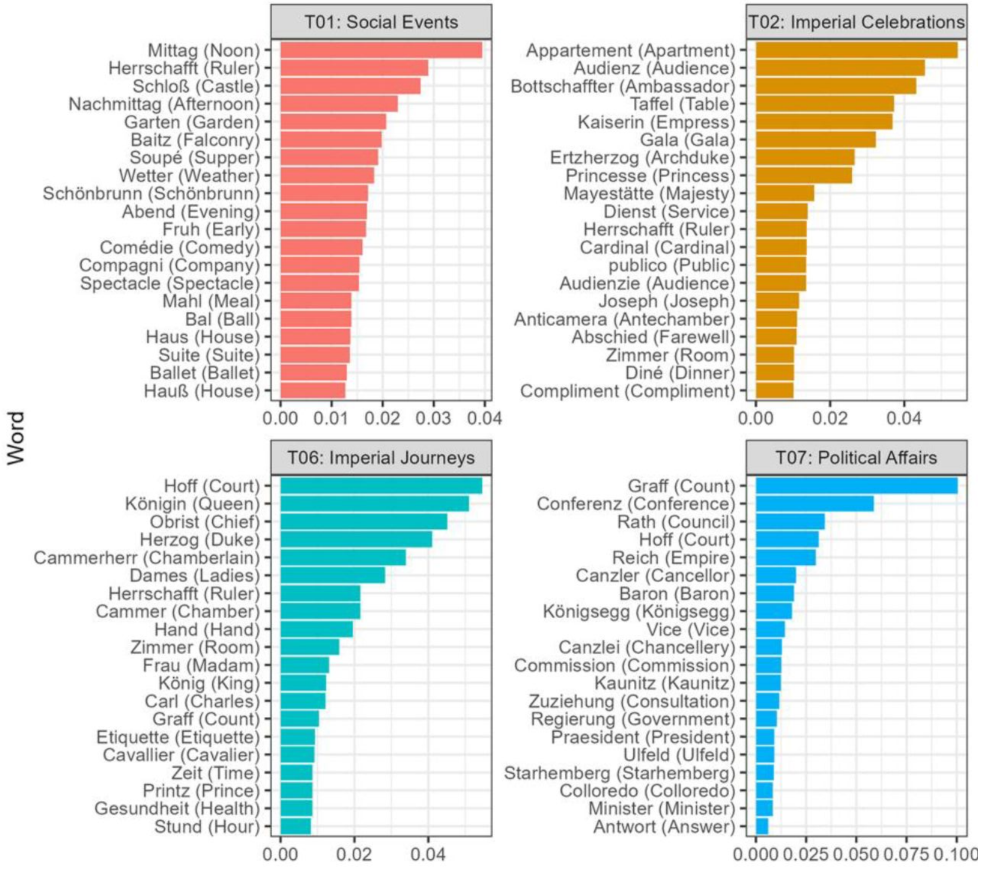
Use Cases of Topic modeling

- Information Retrieval
  - Search by topic/semantic field instead of individual terms
- Recommender systems
  - Suggestions for semantically similar research articles
- Exploration of text collections
  - Research questions from literary and cultural history

Topic modeling Examples

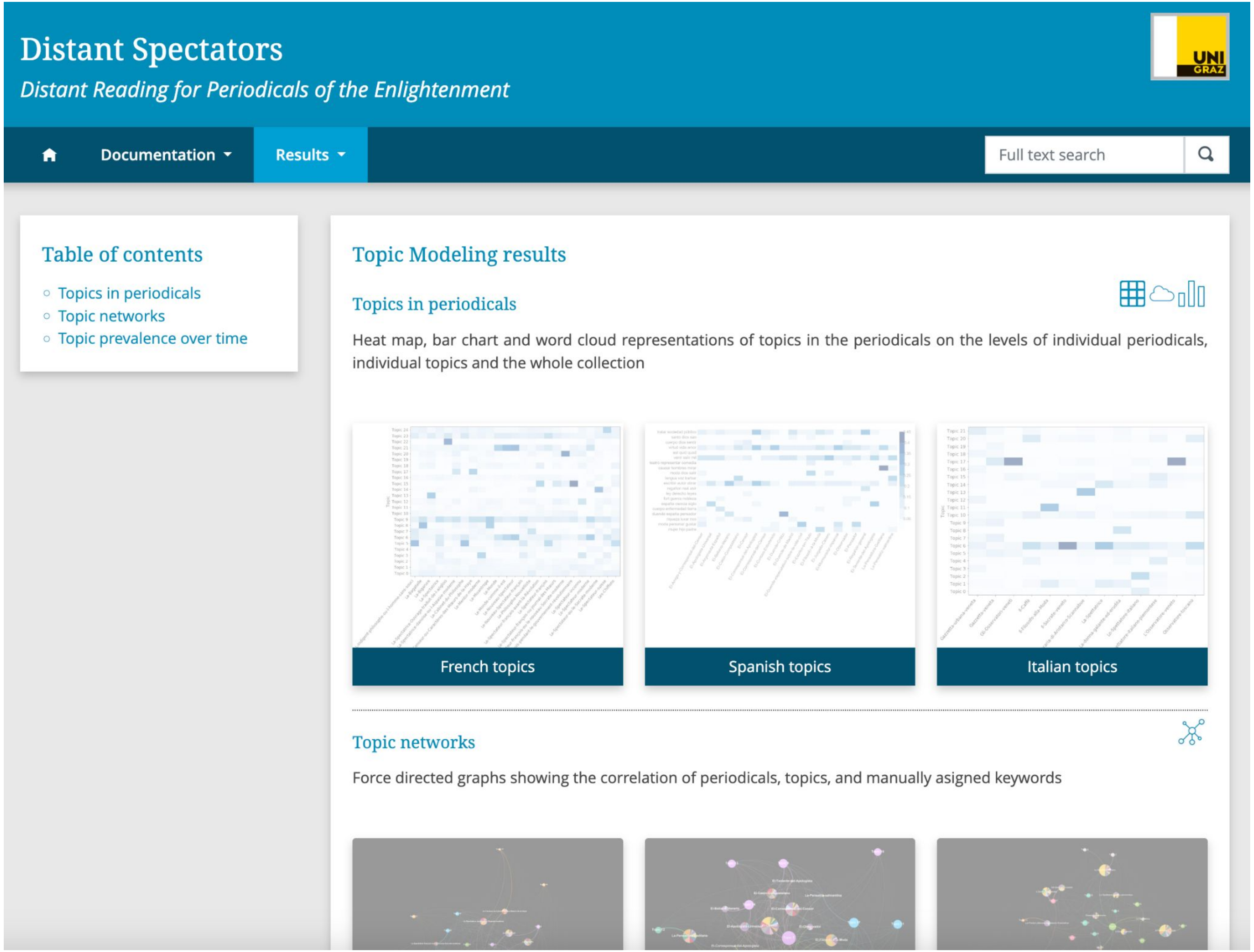
Exploring Topics from Khevenhüller's diary 1742–76

Yamashita, Taisei, and Atsuhiko Uchida. 2025. 'Historical Analysis Using Topic modeling: Insights from Khevenhüller's Diary 1742–76'. *Digital Scholarship in the Humanities*, fqaf036. https://doi.org/10.1093/llc/fqaf036.



Distant Spectators

Distant Reading for Periodicals of the Enlightenment

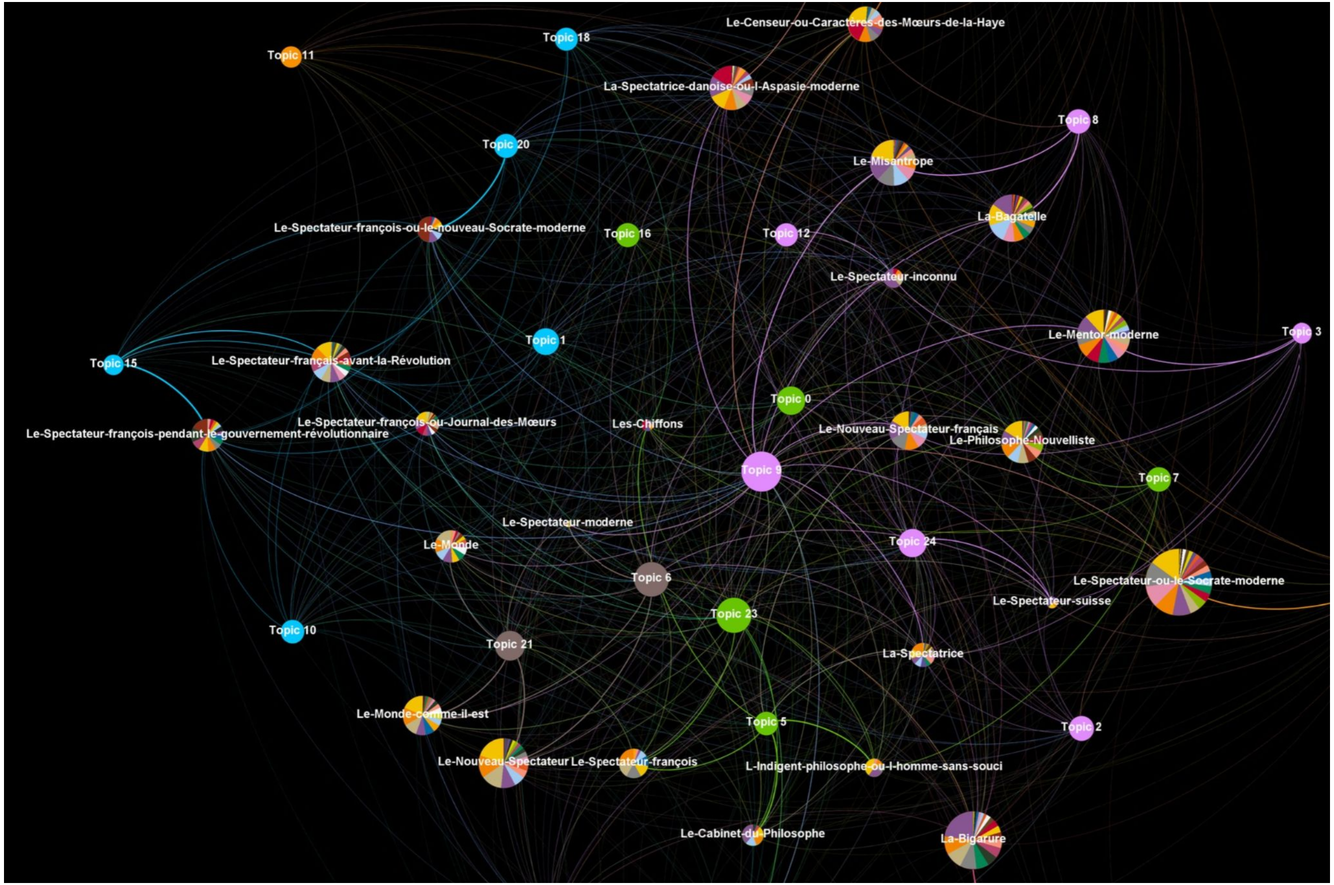


https://gams.uni-graz.at/context:dispecs

Distant Spectators

Distant Reading for Periodicals of the Enlightenment

https://gams.uni-graz.at/context:dispecs



https://gams.uni-graz.at/archive/objects/o:dispecs.result.tm.fr/methods/sdef:TEI/get?mode=topic-network

Table 1 Topics related to dark tourism

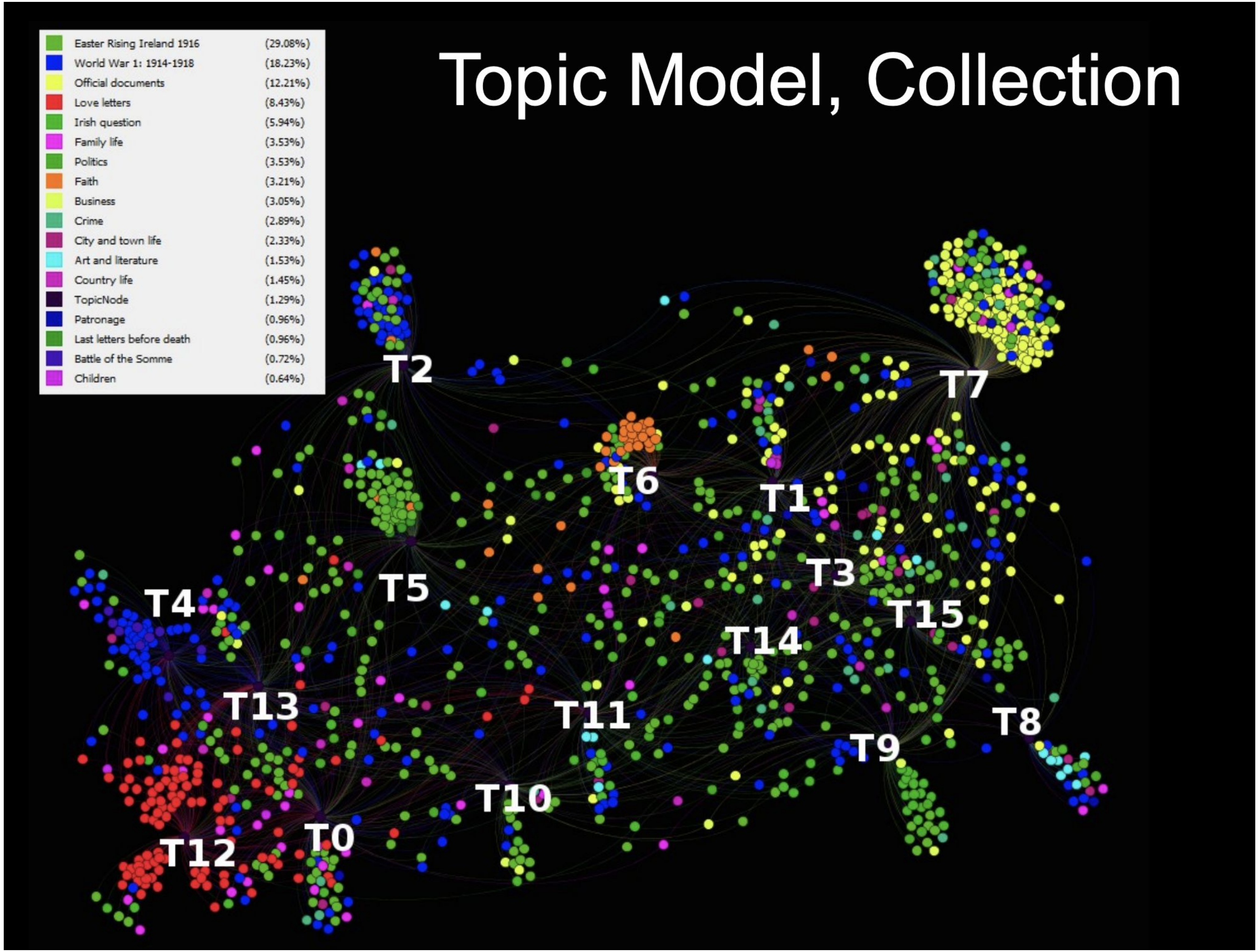
CorEx Topic	*Topic name*	*Keywords*	*NMF Topic*	*Keywords*
1	Chernobyl disaster	chernobyl, pripyat, ukraine, exclusion, chernobylexclusionzone	1	exclusion, chernobyl, pripyat chernobyl, zone ukraine
			2	pripyat, abandon, military, building, wheel
2	Urban exploration	urbex, stalker, urbandecay, urbexworld, utopia		NA
3	Cemetery impressions	cemetery, moscow, taphophile, russia, rookwood	3	cemetery, moscow, russia, vagankovskoe cemetery, novodevichie
4	Dark activities in Philadelphia	history, creepy, survivalofthefittest, philadelphia, gritty		NA
5	Nuclear power of the Soviet Union	nuclear, nuclearaccident, soviet communist, cccp, radioactive		NA
6	Graveyard impressions	graveyard, necropolis, freaks, darkness, fantasy		NA
7	War tourism	war, gunfight, abandoned, seaforts, seacoast		NA
8	Ghost hunting	creepy, wv, paranormalinvestigation, springs sanitarium, sweetspringsresort	4	home, sanitarium, sweet, springs, memorial home
9	Imperial Crypt in Vienna	crypt, vienna, imperial, habsburg, kaisergruft		NA
10	Jeju uprising	dark pictures, emotional, text, april, jejuisland		NA
11	History of Berlin	memory, loving, hearts, berlinwall, textile factory		NA
12	Sedlec Ossuary	kutnahora, bonechurch, sedlecossuary, czech republic		NA

Figure: p. 1240

Egger, Roman, and Joanne Yu. 2021. '**Identifying Hidden Semantic Structures in Instagram Data: A Topic modeling Comparison**'. *Tourism Review* 77 (4): 1234–46. https://doi.org/10.1108/TR-05-2021-0244.

Roman Bleier:
Topic modeling des Letters of 1916 Briefkorpus

https://gams.uni-graz.at/o:dhd2015.v.005



Edition of Texts from Jerónimo Pizarro

using Mallet for Topic Modelling

automatic generation of **30 topics for a taxonomy**

Fragments (text snippets) are assigned to a topic if the relevance percentage exceeds 11%, based on 1,500 iterations of the model

Arquivo LdoD INICIAR SESSÃO

ACERCA ▾ LEITURA ▾ DOCUMENTOS ▾ EDIÇÕES ▾ PESQUISA ▾ VIRTUAL ▾ PT EN ES

Edição Virtual: Mallet

Editores: Manuel Portela

Sinopse: Edição com taxonomia gerada automaticamente para 30 categorias, sobre o corpus da edição Jerónimo Pizarro, usando o software de geração de tópicos Mallet (integrado no Arquivo LdoD), nomeando cada categoria com as 3 palavras mais relevantes do tópico gerado, associando fragmentos à categoria se a percentagem é superior a 11% e executando 1500 iterações para a geração dos tópicos.

Taxonomia: Mallet

660 Fragmentos:

Search

Número	Título	Categoria	Usa Edições
1	Minha alma é uma orchestra occulta	nevoa leve frio noite dia luz sonho cousas sonhos vida ser alma	->Pizarro
2	Eu não sonho possuir-te	corpo possuir morte mulher sexo mãe sonho cousas sonhos viajar viagens novas vida ser alma	->Pizarro
3	Glorificação das Estereis	amôr paysagem horas mulher sexo mãe sonho cousas sonhos vida ser alma	->Pizarro
4	Nossa Senhora do Silencio	mulher sexo mãe noite dia luz sonho cousas sonhos vida ser alma	->Pizarro

https://ldod.uc.pt/edition/acronym/LdoD-Mallet

Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)

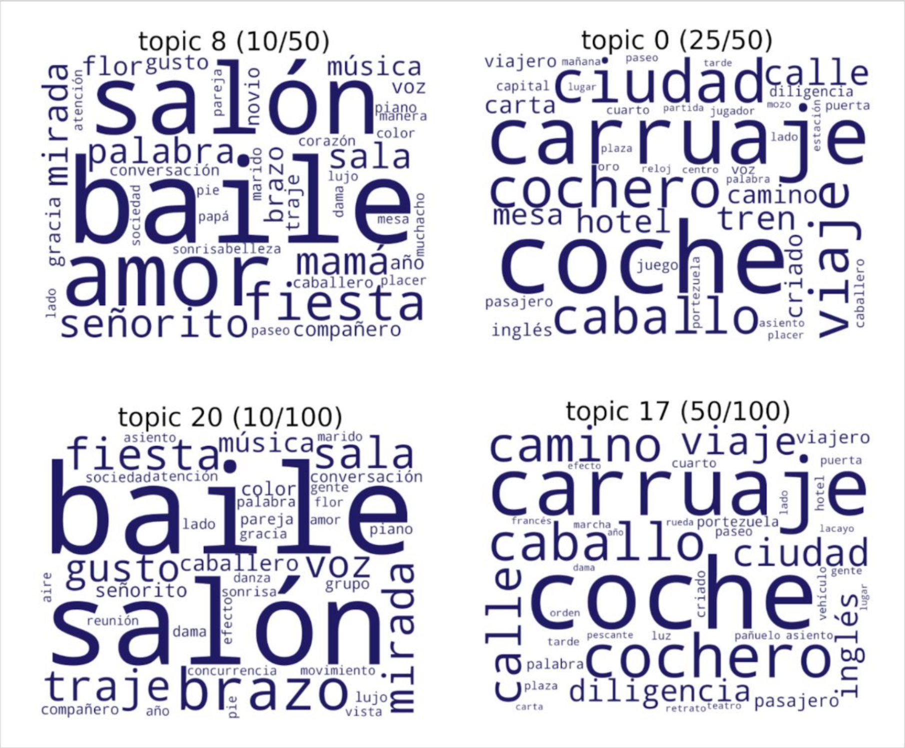


Figure 44: Example topics.

http://side17.i-d-e.de/ch4.html#p648

Henny-Krahmer, Ulrike. 2024. *Genre Analysis and Corpus Design. Nineteenth-Century Spanish-American Novels (1830–1910)*. HTML version, v1.0.1. Schriften des Instituts für Dokumentologie und Editorik 17. Berlin: IDE. DOI: 10.5281/zenodo.10498609

Examples

Navarro-Colorado, Borja. 2018. 'On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry'. *Frontiers in Digital Humanities* 5. https://doi.org/10.3389/fdigh.2018.00015.

Schöch, Christof. 2021. 'Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama'. arXiv:2103.13019. Preprint, arXiv. https://doi.org/10.48550/arXiv.2103.13019.

Völkl, Yvonne, Sanja Sarić, and Martina Scholger. 2022. 'Topic Modeling for the Identification of Gender-Specific Discourse. Virtues and Vices in French and Spanish 18th Century Periodicals'. *Journal of Computational Literary Studies* 1 (1): 1. https://doi.org/10.48694/jcls.108.

Wendel, Luisa, Anna Shadrova, and Alexander Tischbirek. 2022. 'From Modeled Topics to Areas of Law: A Comparative Analysis of Types of Proceedings in the German Federal Constitutional Court'. *German Law Journal* 23 (4): 493–531. https://doi.org/10.1017/glj.2022.39.

Yamashita, Taisei, and Atsuhiko Uchida. 2025. 'Historical Analysis Using Topic modeling: Insights from Khevenhüller's Diary 1742–76'. *Digital Scholarship in the Humanities*, fqaf036. https://doi.org/10.1093/llc/fqaf036.

BERTopic

BERTopic

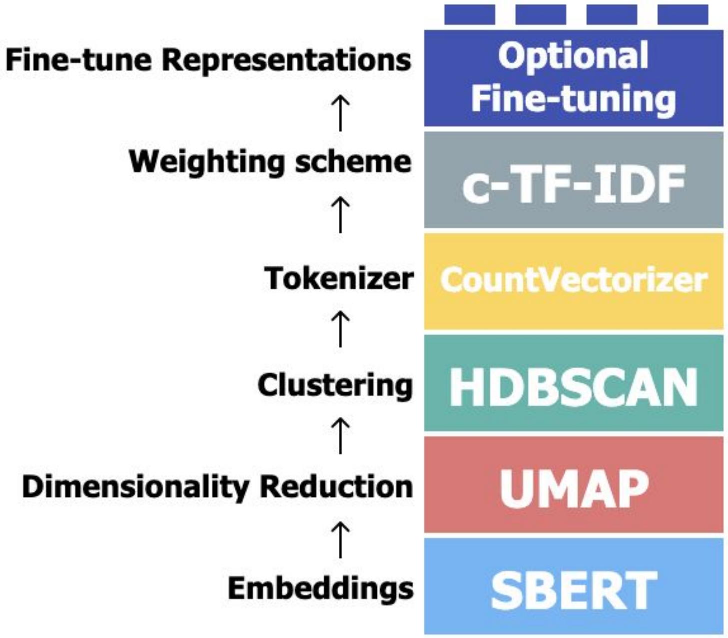
“BERTopic is a modern topic modeling framework that addresses many limitations of traditional approaches.”

- https://bertopic.com/
- Developed by Maarten Grootendorst
  - Documentation: https://maartengr.github.io/BERTopic/index.html
- uses transformer-based embeddings (like BERT) to understand the semantic meaning of documents and clusters them based on their context rather than just word frequency.
- Python library

How BERTopic works

BERTopic can be viewed as a sequence of steps to create its topic representations.

Although these steps are the default, there is some modularity to BERTopic.



https://maartengr.github.io/BERTopic/algorithm/algorithm.html

How BERTopic works

Document Embedding (via BERT or similar)

- Goal: Convert raw text into numerical representations that capture semantic meaning.
- How: Each document is passed through a pre-trained language model (e.g., BERT, RoBERTa, Sentence-BERT).
- Output: High-dimensional dense embeddings (e.g., 768-dimension vectors for BERT).
- Why it matters: These embeddings preserve contextual meaning, allowing for better clustering than simple Bag-of-Words.

https://bertopic.com/#How%20BERTopic%20Works

How BERTopic works

Dimensionality Reduction (UMAP)

- Goal: Reduce the high-dimensional embeddings to a lower-dimensional space for clustering.
- How: BERTopic uses UMAP (Uniform Manifold Approximation and Projection), a nonlinear dimensionality reduction technique.
- Output: 2D or 5D representation of each document that retains semantic structure.
- Why it matters: Lower-dimensional space improves clustering accuracy and speed.

https://bertopic.com/#How%20BERTopic%20Works

How BERTopic works

Clustering (HDBSCAN)

- Goal: Group similar documents into clusters that will become topics.
- How: Uses HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which:
  - Automatically determines the number of clusters
  - Handles noise and outliers
- Output: Cluster labels assigned to documents (e.g., Topic 1, Topic 2, -1 for noise)
- Why it matters: It allows unsupervised, flexible clustering that works well for varied data distributions.

https://bertopic.com/#How%20BERTopic%20Works

How BERTopic works

Topic Representation (c-TF-IDF)

- Goal: Create human-readable topic labels by identifying representative words.
- How: Applies class-based TF-IDF (c-TF-IDF):
  - Treats each cluster as a “class” or “document”
  - Calculates TF-IDF values within clusters instead of globally
- Output: Top keywords per topic
- Why it matters: Ensures that terms strongly associated with a specific cluster are prioritized, making topic descriptions more interpretable.

https://bertopic.com/#How%20BERTopic%20Works

How BERTopic works

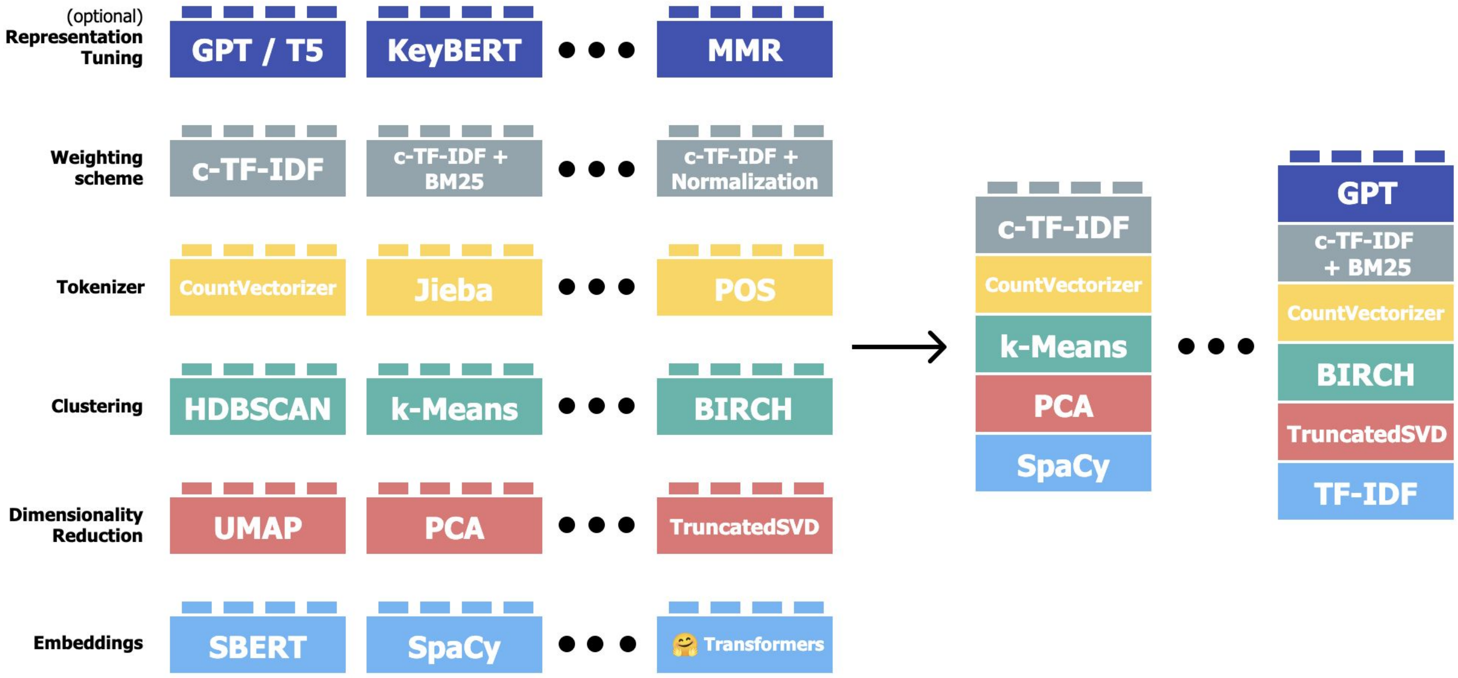
Visualization (Optional)

- Goal: Help users explore topics, their relationships, and importance.
- How: BERTopic offers several visualizations:
  - Intertopic distance map using pyLDAvis
  - Bar charts for top terms per topic
  - Topic hierarchy (tree of merged/split topics)
- Why it matters: Aids in understanding topic relevance and overlaps.

https://bertopic.com/#How%20BERTopic%20Works

How BERTopic works

You can build your own custom topic modeling workflow!



https://maartengr.github.io/BERTopic/algorithm/algorithm.html#visual-overview

Comparison to LDA

	LDA	BERTopic
Based on	bag-of-words	embeddings
Topic Count	has to be decided before the topic modeling process	generated automatically but can also be decided beforehand
Adjustment of Topic Count	only with another run	subsequent merging, reduction, or renaming possible
Assigning topics to documents	probability distribution per document across all topics	unique top topic per document
Post-processing	limited	very flexible

BERTopic
Topic modeling Techniques

BERTopic: Topic modeling Techniques

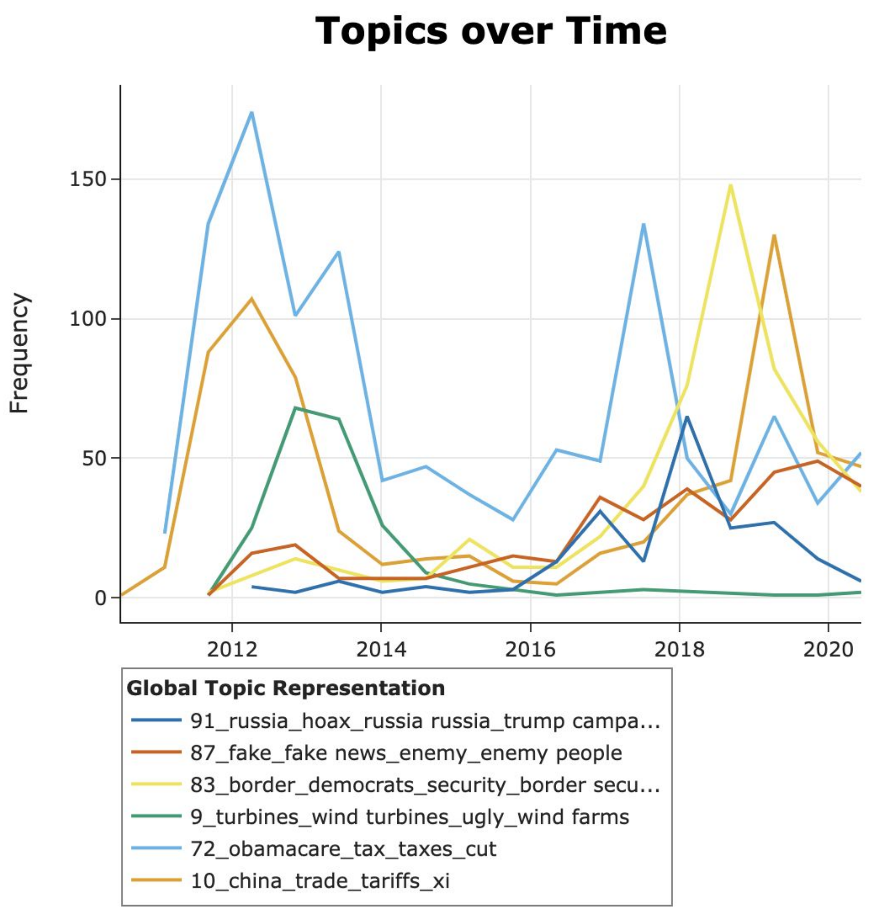
https://maartengr.github.io/BERTopic/index.html

BERTopic supports all kinds of topic modeling techniques:

Guided	Supervised	Semi-supervised
Manual	Multi-topic distributions	Hierarchical
Class-based	Dynamic	Online/Incremental
Multimodal	Multi-aspect	Text Generation/LLM
Zero-shot **(new!)**	Merge Models **(new!)**	Seed Words **(new!)**

Dynamic Topic modeling

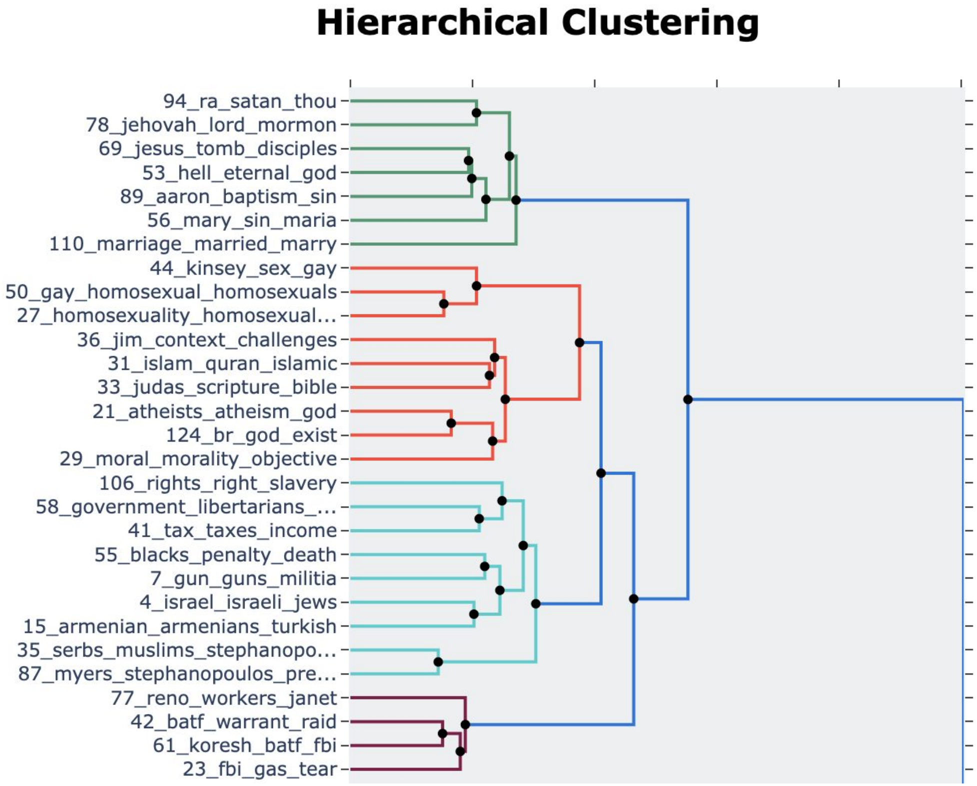
- analyzing the evolution of topics over time
- e.g. in 1995 people may talk differently about environmental awareness than those in 2015
- the topic itself remains the same but the exact representation might differ



https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html

Hierarchical Topic modeling

- understanding, which topics are similar to each other
- insights into sub-topics
- useful, if you want to understand, which topics could be merged



Multimodal Topic modeling

[football, player, players, team, game, sooners, tackle, tackled, ball, stadium]

Topic modeling of Images + Text

Other

- Supervised Topic Modeling (using documents with labels, you already have)
- Guided Topic Modeling
- ...

```
seed_topic_list = [["drug", "cancer", "drugs", "doctor"],
                   ["windows", "drive", "dos", "file"],
                   ["space", "launch", "orbit", "lunar"]]
```

https://maartengr.github.io/BERTopic/getting_started/guided/guided.html#example

Example

Example: The memoirs of Countess Schwerin

- Text: Memoirs of the Countess Luise Charlotte von Schwerin (1684-1732)
- Project: https://memoiren.hypotheses.org/
- FWF Project, running from 2022-2025, digital edition
- the french text has been preserved in two manuscripts
    - manuscript A: "L'Histoire De la Vie de madame la comtesse de Scheverin écrite par elle même a ses enfans".
    - manuscript B: "Eigenhändiges Tagebuch der Gräfin Schwerin": "Historie de la vie de Mad. la comtesse de Schwerin, ecrite par elle mème a ces enfans suivant les ordres de son Directeur à Cologne".

```xml
<p xml:id="P.1">
   <anchor corresp="#facs_5" xml:id="normalized_facs_5"/>
   <app>
      <rdg wit="#W">Fürst Khevenhüller <date type="year" when-iso="1896">1896</date>
      </rdg>
   </app>
</p>
<p xml:id="P.2"> « Histoire De la Vie de <persName ref="#P153">madame la comtesse de
      Schwerin</persName>
</p>
<p xml:id="P.3">écrite par elle-même à ses enfants <app>
      <rdg wit="#W">suivant les ordres de son directeur à <placeName ref="#O19"
            >Cologne</placeName>
      </rdg>
   </app> » </p>
<p xml:id="P.4"> Première partie </p>
<p xml:id="P.5">L’histoire de ma vie est si remarquable et si remplie d’événements, tant
   pour leur singularité que par le bruit que les plus considérables ont fait dans le
   monde, que ma curiosité m’a portée en premier lieu d’en rassembler le cours, et <app>
      <rdg wit="#A">deuxièmement</rdg>
      <rdg wit="#W">secondement</rdg>
   </app> pour qu’un jour puissent être instruits par moi-même des suites malheureuses
   (selon le monde) que produit ordinairement une fortune <app>
      <rdg wit="#A">éblouissante</rdg>
   </app> à une jeune personne élevée dans tout ce qui peut plaire à son ambition et à
   ses sens, suites dis-je <anchor corresp="#facs_6" xml:id="normalized_facs_6"
   />d’autant plus dangereuses qu’elles sont imperceptibles à ceux dont l’amour-propre
   ferme les yeux sur leur conduite, et dont ils ne sont détrompés que par une longue
   expérience, d’autant plus rude à être exercée que peu de personnes ont le courage
   d’entreprendre le grand ouvrage de la connaissance de soi-même. Ah ! mon doux et
   divin Jésus, ce n’est que vos compassions et vos abondantes miséricordes qui m’y ont
   entraîné presque malgré moi, que de grâces n’avez-vous pas opposé à ma résistance,
   jusqu’à ce que vous ayez vaincu mon opiniâtreté, et vous m’ayez réduite à une entière
   résignation à votre volonté. Votre saint Nom en soit béni, loué et glorifié. Amen. Je
   puis raisonner des suites fâcheuses où la fortune et l’ambition sont le point de vue,
   et mes tristes réflexions peuvent servir d’exemple à ceux qui comme moi abuseront de
   leurs sentiments et donneront au monde ce qui n’est qu’à Dieu <app>
```

Französischer Lesetext mit Abweichungen

[1] Fürst Khevenhüller 1896

[2] « Histoire De la Vie de madame la comtesse de Schwerin

[3] écrite par elle-même à ses enfants suivant les ordres de son directeur à Cologne »

[4] Première partie

[5] L’histoire de ma vie est si remarquable et si remplie d’événements, tant pour leur singularité que par le bruit que les plus considérables ont fait dans le monde, que ma curiosité m’a portée en premier lieu d’en rassembler le cours, et secondement pour qu’un jour mes chers enfants puissent être instruits par moi-même des suites malheureuses (selon le monde) que produit ordinairement une fortune éblouissante à une jeune personne élevée dans tout ce qui peut plaire à son ambition et à ses sens, suites dis-je d’autant plus dangereuses qu’elles sont imperceptibles à ceux dont l’amour-propre ferme les yeux sur leur conduite, et dont ils ne sont détrompés que par une longue expérience, d’autant plus rude à être exercée que peu de personnes ont le courage d’entreprendre le grand ouvrage de la connaissance de soi-même. Ah ! mon doux et divin Jésus, ce n’est que vos compassions et vos abondantes miséricordes qui m’y ont entraîné presque malgré moi, que de grâces n’avez-vous pas opposé à ma résistance, jusqu’à ce que vous ayez vaincu mon opiniâtreté, et vous m’ayez réduite à une entière résignation à votre volonté. Votre saint Nom en soit béni, loué et glorifié. Amen. Je puis raisonner des suites fâcheuses où la fortune et l’ambition sont le point de vue, et mes tristes réflexions peuvent servir d’exemple à ceux qui comme moi abuseront de leurs sentiments et donneront au monde ce qui n’est qu’à Dieu seul .

The text is encoded with TEI/XML.

Presentation of the text.

BERTopic Quickstart Example

- BERTopic Quickstart Example

```
topics, probs = topic_model.fit_transform(paragraphs)
```

```
2025-09-01 11:07:40,344 - BERTopic - Embedding - Transforming documents to embeddings.
modules.json: 100% 229/229 [00:00<00:00, 20.4kB/s]
config_sentence_transformers.json: 100% 122/122 [00:00<00:00, 10.0kB/s]
README.md: 3.89k/? [00:00<00:00, 334kB/s]
sentence_bert_config.json: 100% 53.0/53.0 [00:00<00:00, 5.00kB/s]
config.json: 100% 645/645 [00:00<00:00, 60.9kB/s]
model.safetensors: 100% 471M/471M [00:06<00:00, 90.0MB/s]
tokenizer_config.json: 100% 480/480 [00:00<00:00, 42.4kB/s]
tokenizer.json: 100% 9.08M/9.08M [00:00<00:00, 64.1MB/s]
special_tokens_map.json: 100% 239/239 [00:00<00:00, 16.9kB/s]
config.json: 100% 190/190 [00:00<00:00, 15.6kB/s]
Batches: 100% 22/22 [01:12<00:00,  3.20s/it]
2025-09-01 11:09:01,312 - BERTopic - Embedding - Completed ✓
2025-09-01 11:09:01,315 - BERTopic - Dimensionality - Fitting the dimensionality reduction algorithm
2025-09-01 11:09:12,874 - BERTopic - Dimensionality - Completed ✓
2025-09-01 11:09:12,875 - BERTopic - Cluster - Start clustering the reduced embeddings
2025-09-01 11:09:12,909 - BERTopic - Cluster - Completed ✓
2025-09-01 11:09:12,922 - BERTopic - Representation - Fine-tuning topics using representation models.
2025-09-01 11:09:13,199 - BERTopic - Representation - Completed ✓
```

Topic modeling and the Memoirs of Countess of Schwerin

- Comparison of the memoirs with the Memoirs of Countess Schwerin with the Memoirs of Madame Guyon using Topic modeling
- maybe use it for indexing and write the topics back to the TEI (work in progress)