Tetys

Configurable Topic Modeling Exploration for Big Corpora of Text Documents

Francesco Invernici, Anna Bernasconi, Francesca Curati, Jelena Jakimov, Amirhossein Samavi
Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

URL: gmql.eu/tetys **Code**: github.com/Frinve/TETYS

An open-source platform for democratizing automatic content profiling through topic modeling.

Overview

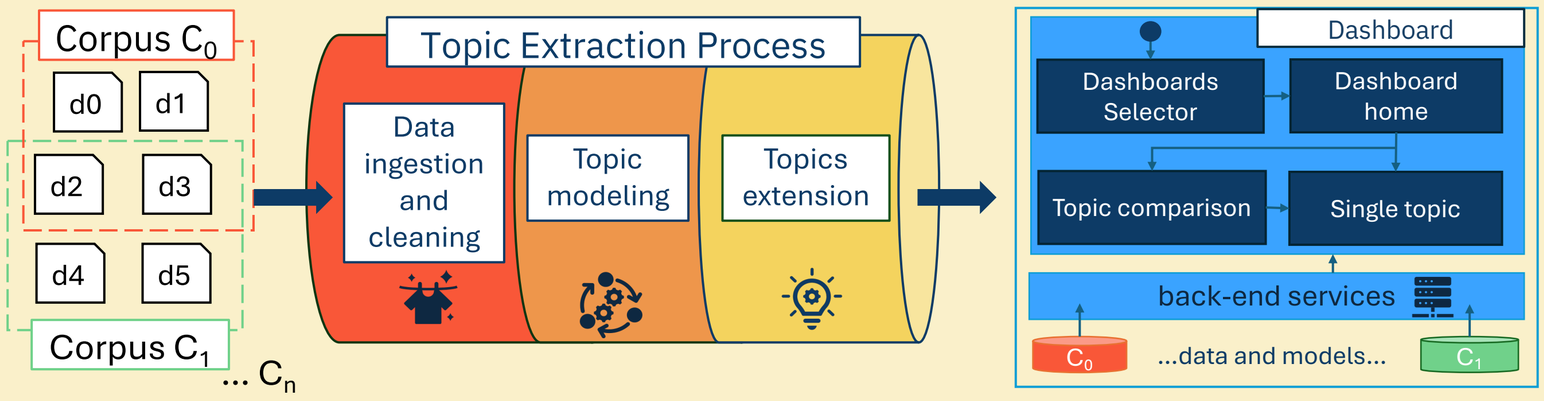
Tetys (Topics' Evolution That You See) is an end-to-end **topic modeling pipeline** that automatically processes **large** text corpora and **generates interactive dashboards** for fast exploration. Tetys uses neural topic modeling[1] enhanced with Large Language Models to identify and visualize key concepts and temporal trends without requiring prior knowledge of content.

Tasks

- ✓ **Make topic modeling projects** from any textual corpus
- ✓ **Discover** topics of interest
- ✓ **Inspect** the characteristics of any topic
- ✓ **Compare** different topic trends

Pipeline

- Data preparation
- Automatic hyper-parameter optimization
- Fitting of the topic model
- Topics extension, enabling:
  - Time series visualization
  - Keyword search functionalities



In this Demonstration:

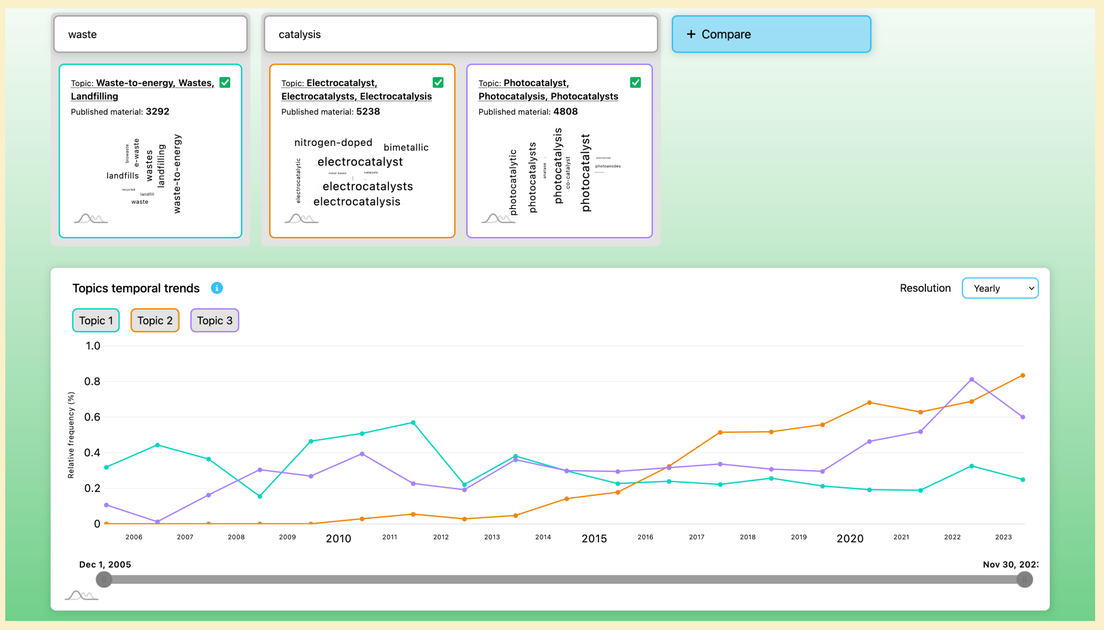
Sustainability Development Goals[2] from 5 areas of interest

- Basic Human Needs and Well-being
- Economic Development and Employment
- Equality and Social Inclusion
- Global Partnership and Peace
- Environment Sustainability

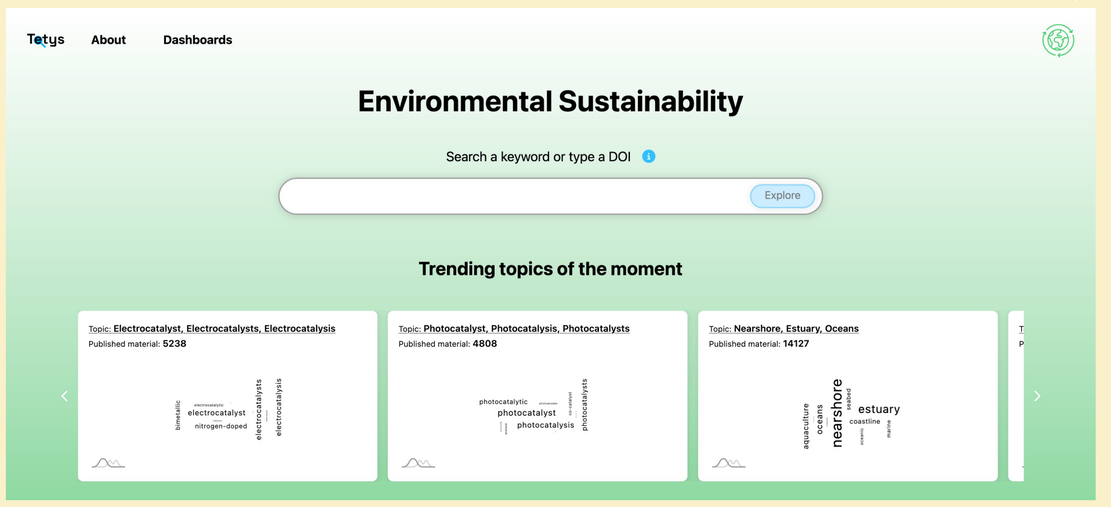
[IN]
>500.000
research papers from 2006 to 2023

[OUT]
Five thematic dashboards

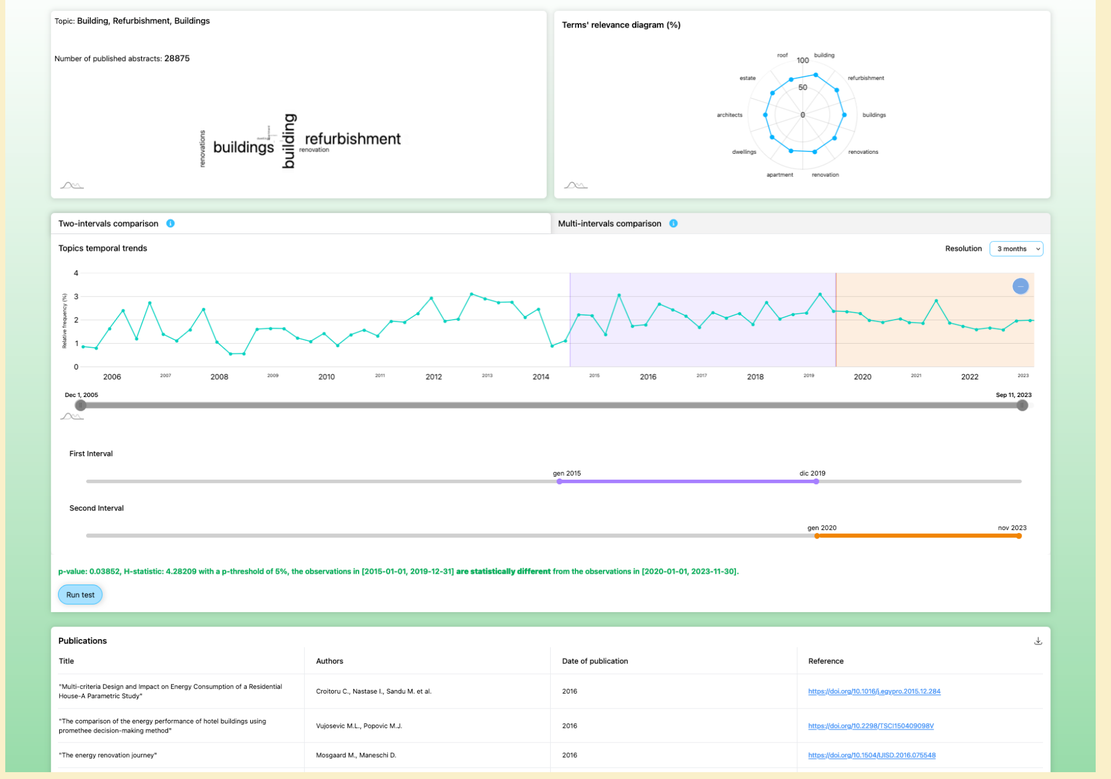
Compare topics across time



Discover research trends



Inspect topics and verify your hypotheses



Evaluation with Users

32.0 (Medium)
NASA-RTLX[3] average workload score

71.3 (Upper-Avg)
System Usability Score[4] by Professional Users

NASA Raw Task Load Index Results
13 students,
8 professionals

Group	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration	WL
Students	46.2	18.5	30.8	32.3	42.3	28.5	33.1
Professionals	43.8	13.8	32.5	21.3	36.3	33.8	30.2
Average	45.3	16.7	31.4	28.1	40.0	30.5	32.0

References

[1] Suzanna Sia et al. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!. In *Proceedings of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 1728– 1736.
[2] United Nations' Sustainable Development Goals (SDG) - https://sdgs.un.org/goals
[3] Mattias Georgsson. 2019. NASA RTLX as a novel assessment for determining cognitive load and user acceptance of expert and user-based evaluation methods exemplified through a mHealth diabetes self-management application evaluation. In *pHealth* 2019. IOS Press, 185–190.
[4] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

Contact: *francesco.invernici@polimi.it*

NGI SEARCH
Funded by the European Union
POLITECNICO MILANO 1863
DATA SCIENCE GROUP
DIPARTIMENTO DI ELETTRONICA INFORMAZIONE E BIOINGEGNERIA
EDBT ICDT BARCELONA 2025