TETYS
Topics Evolution That You See

POLITECNICO
MILANO 1863
DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

The CORD-19 Topic Visualizer

Exploring the evolution of research topics during the COVID-19 pandemic

Francesco Invernici

November 10, 2023

SFSCON2023 - NOI TECHPARK, Bolzano, Italy

TETYS - About us

TETYS
Topics Evolution That You See

Stefano Ceri

Full Professor

Databases and Web Technologies

Two ERC Grants

Anna Bernasconi

Assistant Professor

Bioinformatics, Databases, and Data Science

Francesco Invernici

1^{st} year PhD Candidate

Knowledge Management, Databases, and NLP

Funded by:

NGI SEARCH

POLITECNICO MILANO 1863

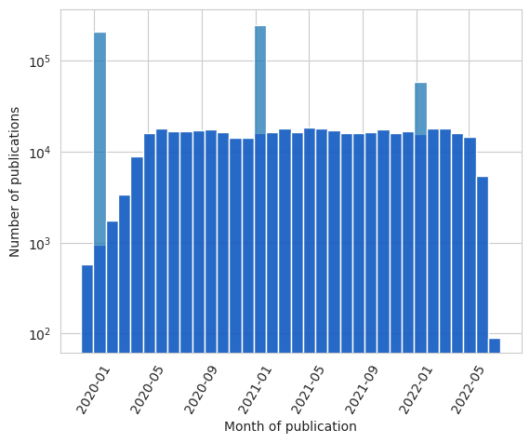
Supported by:

CW2

Information Overload in the COVID-19 pandemic

The effects of COVID-19 pandemic on scientific publications

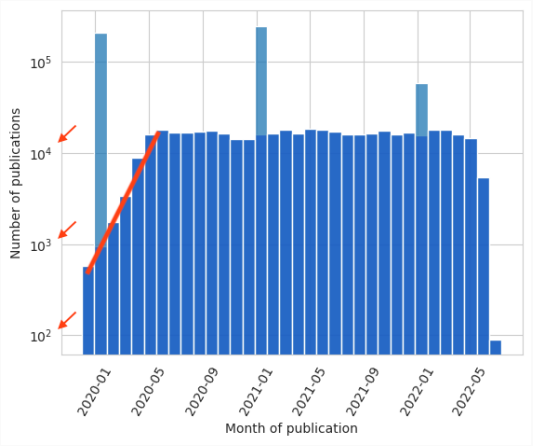
- Outpouring of daily, ever-changing information



Monthly number of publications in CORD-19

The effects of COVID-19 pandemic on scientific publications

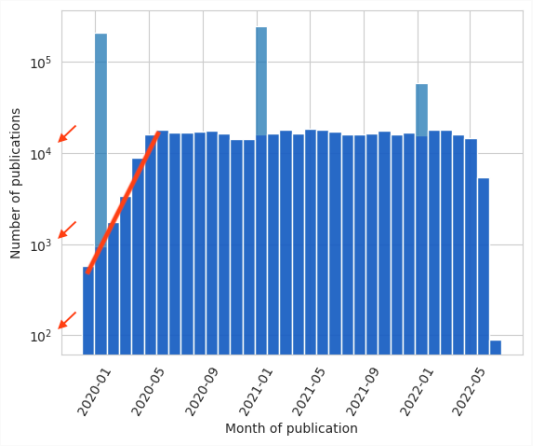
- Outpouring of daily, ever-changing information
- Exponential growth of the number of scientific publications related to COVID-19



Monthly number of publications in CORD-19

The effects of COVID-19 pandemic on scientific publications

- Outpouring of daily, ever-changing information
- Exponential growth of the number of scientific publications related to COVID-19
- Many initiatives to support research: free and non-free datasets, **open-access corpora**, and literature hubs

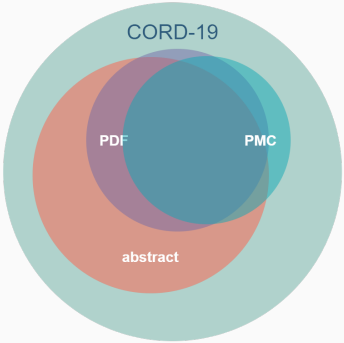


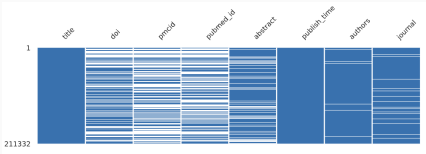
Monthly number of publications in CORD-19

COVID-19 Open Research Dataset

CORD-19 datasheet

- Maintained by Allen Institute for AI, in collaboration with public and private entities

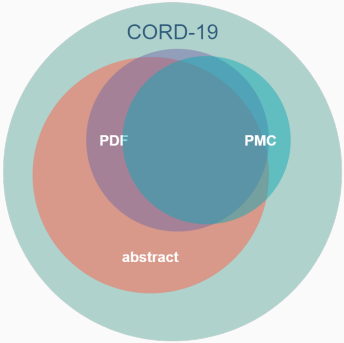


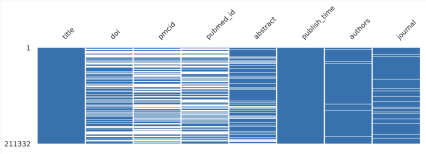


COVID-19 Open Research Dataset

CORD-19 datasheet

- Maintained by Allen Institute for AI, in collaboration with public and private entities
- More than 1 million entries ($\sim$ 1/3 with full text)

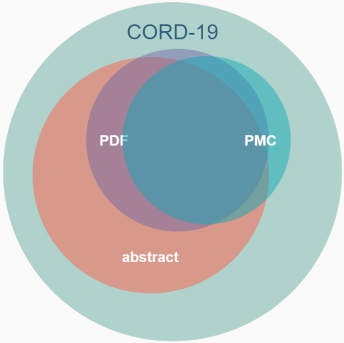


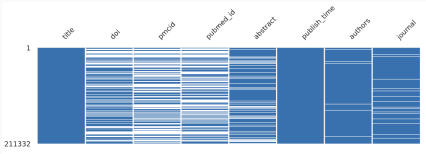


COVID-19 Open Research Dataset

CORD-19 datasheet

- Maintained by Allen Institute for AI, in collaboration with public and private entities
- More than 1 million entries ($\sim$ 1/3 with full text)
- Over 50 thousand journals

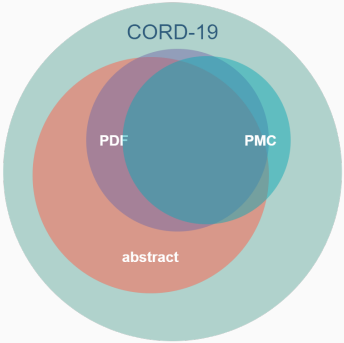


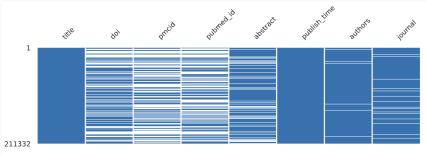


COVID-19 Open Research Dataset

CORD-19 datasheet

- Maintained by Allen Institute for AI, in collaboration with public and private entities
- More than 1 million entries ($\sim$ 1/3 with full text)
- Over 50 thousand journals
- Over 2 million authors

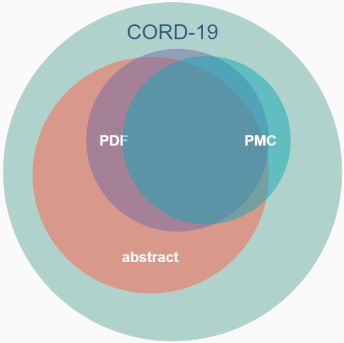


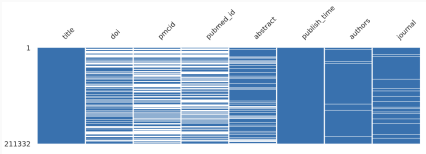


COVID-19 Open Research Dataset

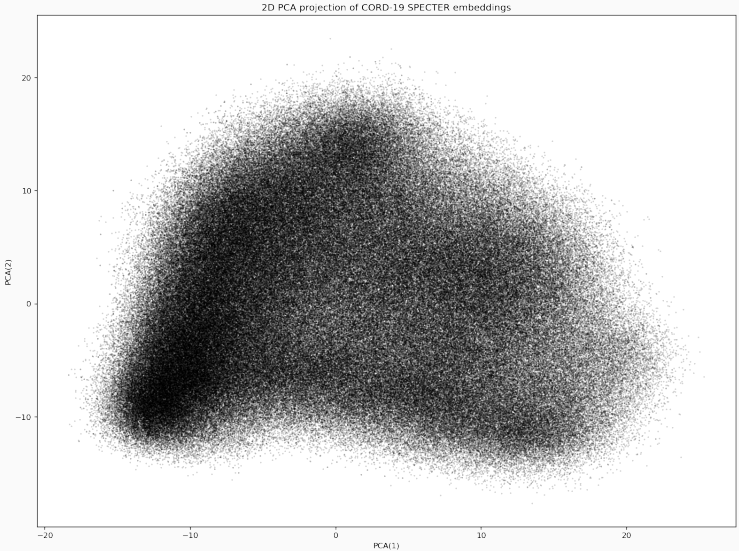
CORD-19 datasheet

- Maintained by Allen Institute for AI, in collaboration with public and private entities
- More than 1 million entries ($\sim$ 1/3 with full text)
- Over 50 thousand journals
- Over 2 million authors
- From December 2019 to June 2022



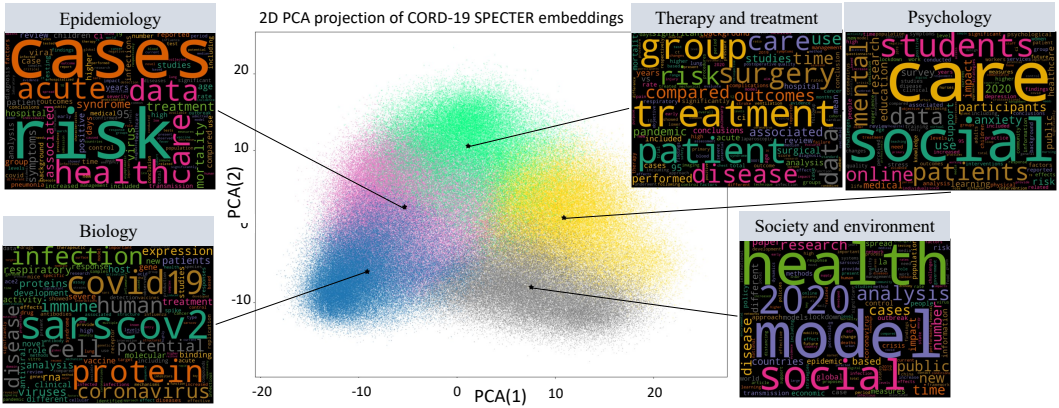


COVID-19 Open Research Dataset



What's inside?

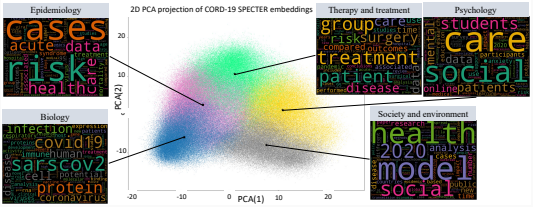
A simple topic model



Results of the preliminary clustering analysis

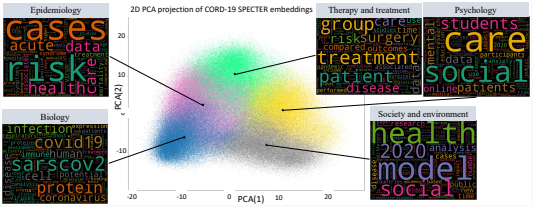
Topic modeling

Topic modeling is a natural language processing technique that automatically identifies and extracts **latent thematic structures** from a collection of documents, providing valuable insights for organizing and **understanding large corpora**.



Topic modeling

By adding the temporal dimension, it is possible to study the **dynamics** of the topics, their change of intensity, and their trends.



Our motivation

- Building the **next-generation open-source Web topic explorer**
- Intuitive **dashboard** for topic trends visualization
- Easy-to-drive statistical testing included
- **Self-tuning** topic extraction **pipeline** for custom corpora
- Built on an open-source stack of technologies

Our motivation

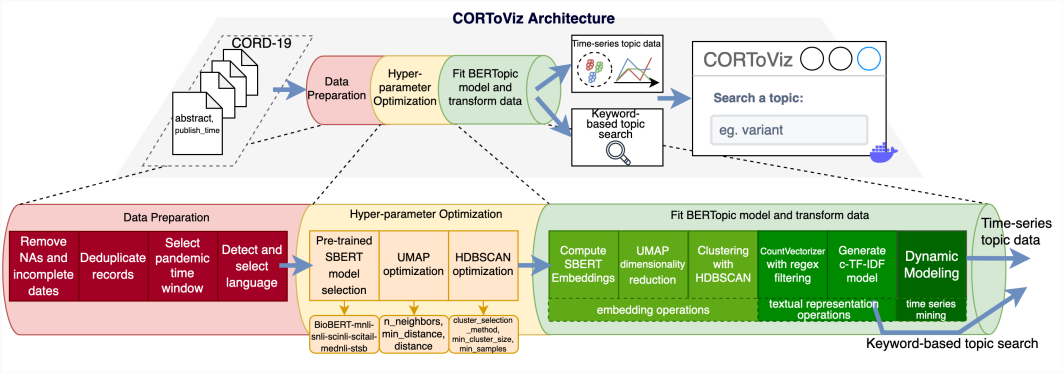
- Building the **next-generation open-source Web topic explorer**
- Intuitive **dashboard** for topic trends visualization
- Easy-to-drive statistical testing included
- **Self-tuning** topic extraction **pipeline** for custom corpora
- Built on an open-source stack of technologies

Re-applicable to many use cases and scenarios **from different domains**:
science of science, journalism, social media analysis, law, ...

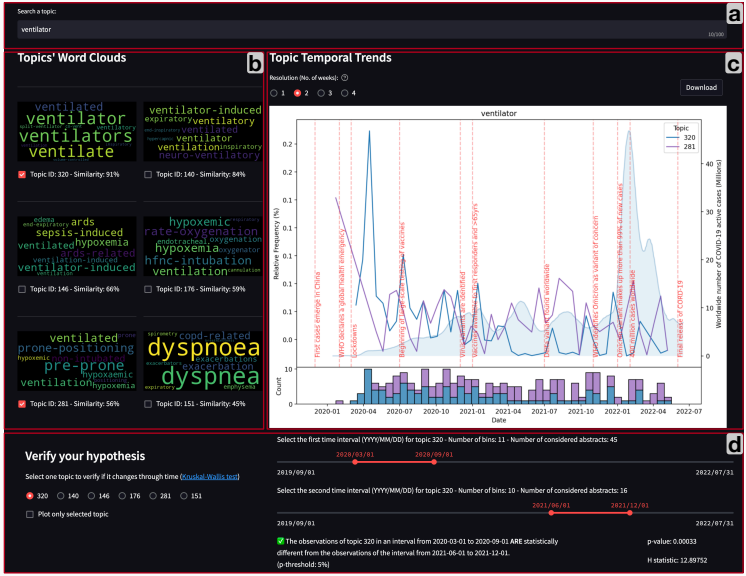
The prototype

The CORToViz Pipeline

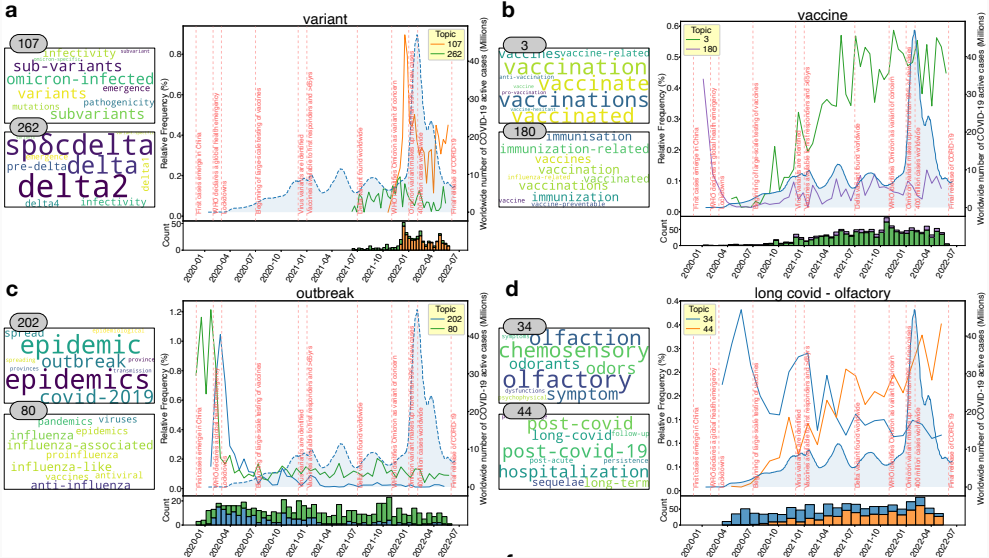
A deep-learning transformer-based pipeline for semantic similarity, textual representation learning, and time-series extraction



The CORD-19 Topics Visualizer



Search sessions



Results

Results - Early COVID-19 pandemic

		2020					2021						2022		
Topic ID	**Description**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	Pandemic outbreak														
174	Italy	4,35	2,07	1,61	1,06	0,88	1,29	0,25	0,45	0,22	0,50	0,21	0,32	0,34	0,28
144	Facemasks and Personal Protective Equipment	1,45	1,03	0,96	1,11	1,17	1,41	0,42	0,71	0,54	0,36	0,59	0,63	0,45	0,33
320	Ventilators	0,73	0,89	0,32	0,43	0,56	0,63	0,45	0,24	0,27	0,28	0,35	0,23	0,33	0,26
279	Management of waste and disinfection	1,27	0,53	0,82	0,66	0,30	0,31	0,58	0,49	0,57	0,52	0,53	0,65	0,34	0,37
136	Self testing	1,12	1,34	1,26	0,66	0,57	0,72	0,91	0,79	0,40	0,56	0,60	0,64	0,70	0,56
131	Postponement and cancellation of surgeries and operations	0,78	1,48	1,36	0,66	1,14	1,12	0,99	0,80	0,48	0,28	0,67	0,50	0,80	0,33
11	Surgeries, laparoscopies and endoscopies	10,25	3,04	3,73	3,32	2,77	1,83	2,48	2,02	2,39	1,70	2,40	2,47	2,29	1,57
	Understanding of the causes of severe disease														
37	Pneumonia and chest scans	7,25	3,55	2,92	2,58	2,62	2,15	1,82	1,29	1,00	0,56	0,60	1,01	0,90	0,56
173	Hyperinflammation, cytokine storm, interleukin and tocilizumab	0	0,95	1,15	0,89	0,88	1,06	0,52	0,68	0,50	0,28	0,67	0,32	0,45	0,44
196	Cardiomyopathy, myocarditis, and cytokines	0,68	1,55	0,96	0,71	0,88	0,90	0,29	0,25	0,40	0,47	0,39	0,55	0,42	0,29
238	Coagulopathies, thrombosis and thromboembolism	0	0,76	0,65	0,73	0,59	0,48	0,28	0,49	0,41	0,24	0,67	0,36	0,47	0,28
34	Symptomps: olfactory and chemosensory disfunctions	1,81	3,08	3,41	1,59	2,22	2,51	1,74	1,82	1,44	1,11	1,77	1,06	1,14	1,00
35	Manifestations of neuroinflammation, neuropathies and encephalitis	0,78	1,90	2,30	2,21	1,76	2,52	2,46	1,47	1,86	1,20	1,51	1,27	1,47	1,54
	Innovative treatments														
23	Hydroxychloroquine and cardiotoxic side effects	4,23	6,42	4,09	3,32	3,66	2,52	1,59	1,25	1,69	0,80	0,81	0,93	0,63	0,56
179	Clinical studies on antihypertensives targeting angiotensin	0,78	1,64	0,96	0,90	0,88	1,04	0,55	0,51	0,46	0,24	0,25	0,25	0,45	0,33
13	Treatments based on heterocyclic compounds	2,13	1,40	1,93	4,15	2,09	2,44	1,82	2,21	2,89	2,41	2,56	2,70	2,35	1,98
22	Therapies based on flavonoids and other phytochemicals	2,90	0,59	1,09	2,21	1,57	1,57	1,65	1,56	2,27	2,78	2,36	3,28	3,16	2,83
8	Efficacy of antimycobacterial and anti-TBC treatments	2,78	1,24	1,15	2,49	1,73	2,26	2,68	2,12	2,75	3,06	3,36	3,49	4,51	4,55
334	Corticosteroids in in-hospital treatment	0	0,30	0,48	0,43	0,44	0,45	0,38	0,31	0,41	0,28	0,37	0,25	0,23	0,29
24	Opioids, medical and non-medical uses in the pandemic	2,21	0,59	1,28	1,42	1,66	1,93	2,27	2,46	1,99	2,01	2,16	2,26	2,26	3,08

Results - Mid and Late COVID-19 pandemic

		2020					2021						2022		
Topic ID	**Description**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	Coronavirus and co-morbidities														
71	Comorbidity and mortality of cancer and malignancies	2,31	3,01	1,57	1,59	1,71	1,01	1,35	0,74	1,24	0,84	0,90	0,36	0,47	0,76
230	Interactions of COVID-19 with HER2 breast cancer	0	0,46	0,32	0,43	0,44	0,54	0,51	0,49	0,45	0,56	0,60	0,58	0,91	0,86
4	Interactions of COVID-19 and diabetes and hyperglycemia	1,82	0,79	1,59	2,32	2,35	2,39	2,98	2,25	3,38	3,41	3,43	4,27	4,74	4,78
220	Vaccine-related and associated myocarditis	0	0	0	0	0	0,26	0,29	0,38	0,49	0,71	0,90	0,76	1,05	1,96
248	Correlation of COVID-19 and Kawasaki disease	0	0,52	0,54	0,76	0,55	0,57	0,68	0,73	0,41	0,47	0,48	0,32	0,25	0,61
52	Interactions of Coccidioidomycosis, fungal infections and COVID-19	0	0,29	0,65	0,40	0,78	0,50	1,12	1,18	2,17	2,51	2,92	2,54	2,80	2,22
	SARS-CoV-2														
150	SARS-CoV-2 receptor bindings and ACE2	2,16	0,36	0,96	0,71	1,24	0,97	0,63	1,17	0,81	0,49	0,81	0,45	0,50	0,56
15	Serology and immunoassays	1,14	2,34	3,46	2,84	3,05	2,63	3,13	3,25	2,42	1,67	1,77	2,06	1,68	1,36
89	Epitopes of antigens for SARS-CoV-2	0	0,64	0,66	0,95	1,03	1,26	1,34	1,45	0,81	1,20	1,12	1,33	0,84	1,05
299	Phenotype, genome and polymorphisms of SARS-CoV-2	0	0,56	0,32	0,33	0,57	0,64	0,55	0,54	0,28	0,72	0,27	0,32	0,47	0,56
124	Variants and substitutions	0	0,31	0,32	0,28	0,29	1,01	1,36	1,13	1,34	1,20	1,12	0,82	0,46	0,66
262	Delta variant	0	0	0	0	0	0	0	0	0,37	0,83	1,33	0,87	0,69	0,44
107	Omicron variant, subvariants and infectivity	0	0	0	0	0	0	0	0	0	0	0,20	4,51	3,15	3,86
	Vaccination														
59	Infection-prevention campaigns and adherence	0,73	1,34	1,15	1,62	2,36	2,58	1,74	1,25	1,35	0,97	1,30	1,13	0,91	0,84
81	Available vaccines, immunization and immunogenicity	1,86	0,52	0,76	0,86	0,59	1,67	0,99	1,10	1,21	1,57	1,35	1,32	1,13	0,98
215	mRNA-based drugs and vaccines	0,73	0	0,42	0,48	0,29	0,67	0,58	0,50	0,55	0,52	0,90	0,77	0,80	0,70
3	Vaccination: for, against and hesitant	0	0,38	0,32	0,81	1,16	2,09	2,98	3,68	4,10	4,72	5,07	4,93	4,51	4,87

TETYS: Next steps

Foreseeable evolution for TETYS:

- Transposition and extension to new contexts and domains
  - Climate change
  - News
  - Social media
  - ...
- Research on new methodologies
  - Large Language Models integration
  - More advanced toolbox for statistical analysis
  - Early detection of potentially interesting topics

Q&A

Thank you for your attention!

Try CORTOViz!

SCAN ME

Any question?

francesco.invernici{at}polimi{dot}it
frinve.com

Preprint and code available at:
Invernici, F., Bernasconi, A., & Ceri, S. (2023).
Exploring the evolution of research topics during the COVID-19 pandemic. ArXiv. /abs/2310.03928