The Linguistic Features of the Anthropocene

Bringing Big Data and Supercomputing to the Humanities

Aerith Netzer

Northwestern University

April 8, 2025

About me

Aerith Netzer

Digital Publishing and Repository Librarian

Professional Technology and Data Dabbler

Introduction

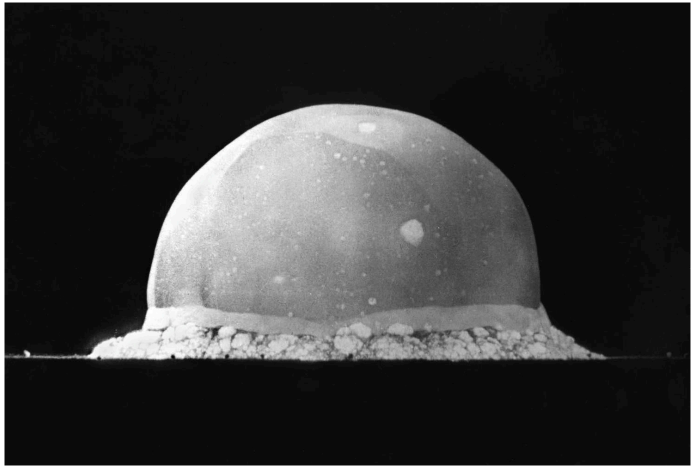
Team Members:

1. James Lee (Professor Northwestern University Libraries/Medill)
2. Han Liu (Professor, McCormick School of Engineering)
3. Lining Mao (PhD Student, McCormick School of Engineering)
4. Kelsey Rydland (Librarian, Northwestern University Libraries)
5. Aerith Netzer (Librarian, Northwestern University Libraries)

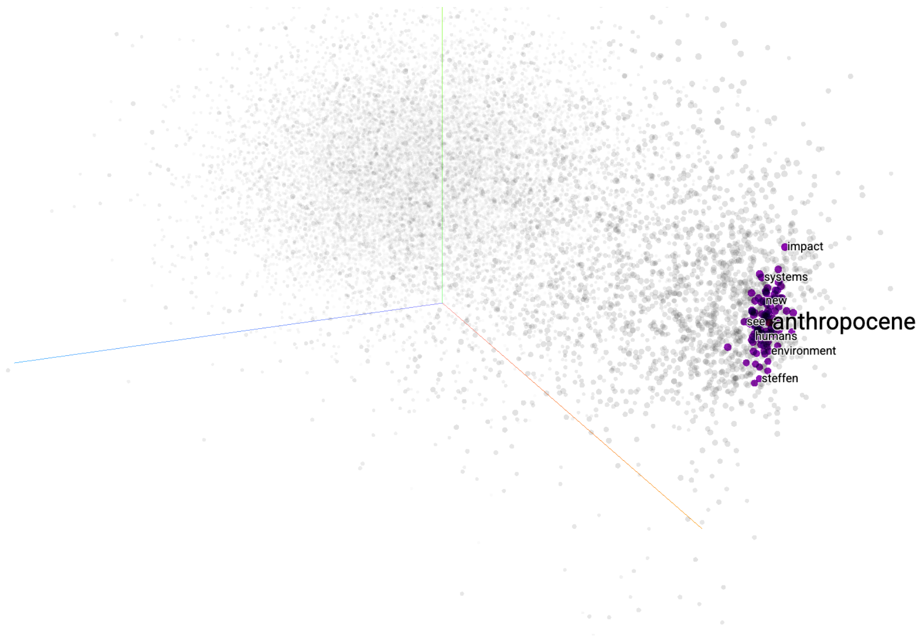
The Problem

What do we wanna know?

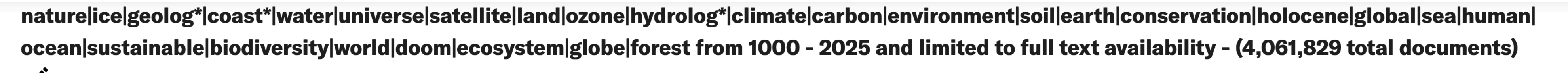
What is the Anthropocene?



The Solution


impact
systems
new
see
anthropocene
humans
environment
steffen

Seed Words



nature|ice|geolog*|coast*|water|universe|satellite|land|ozone|hydrolog*|climate|carbon|environment|soil|earth|conservation|holocene|global|sea|human|ocean|sustainable|biodiversity|world|doom|ecosystem|globe|forest from 1000 - 2025 and limited to full text availability - (4,061,829 total documents)

Seed word list allows us to look at articles that concern *themes* of the anthropocene, while not actually mentioning the word “anthropocene”

A Short Detour—Dataset Troubles

There are currently no “authoritative” sources of academic articles that can be used for data and text mining.

We tried:

- OpenAlex — turned off ngrams halfway through our project
- Semantic Scholar — sparse data with dates
- Constellate — great product, will be deprecated in July

The takeaway: We need to support high-quality, open-access datasets, that are the same time beginner-friendly but also can support power users.

Now that we have the data...

We can start building some topic models.

We used BERTopic for it's ease-of-use, especially when integrating RAPIDSAI into the model pipeline.

It also supports time-dependent and class-dependent topic modelling.

About that time I used 70 GPUs on Quest…

Steps for BERTopic:

0. Clean the data — NLTK
1. Create Embeddings—Sentence Transformers
2. Reduce dimension of embeddings—UMAP with RAPIDSAI
3. Cluster the embeddings—HDBSCAN with RAPIDSAI
4. c-TF-IDF over each topic

To run this pipeline over our entire dataset would take a very long time on traditional CPU hardware. Quest free-tier allocation GPU access allowed us to do huge text analysis at no cost.

AWS

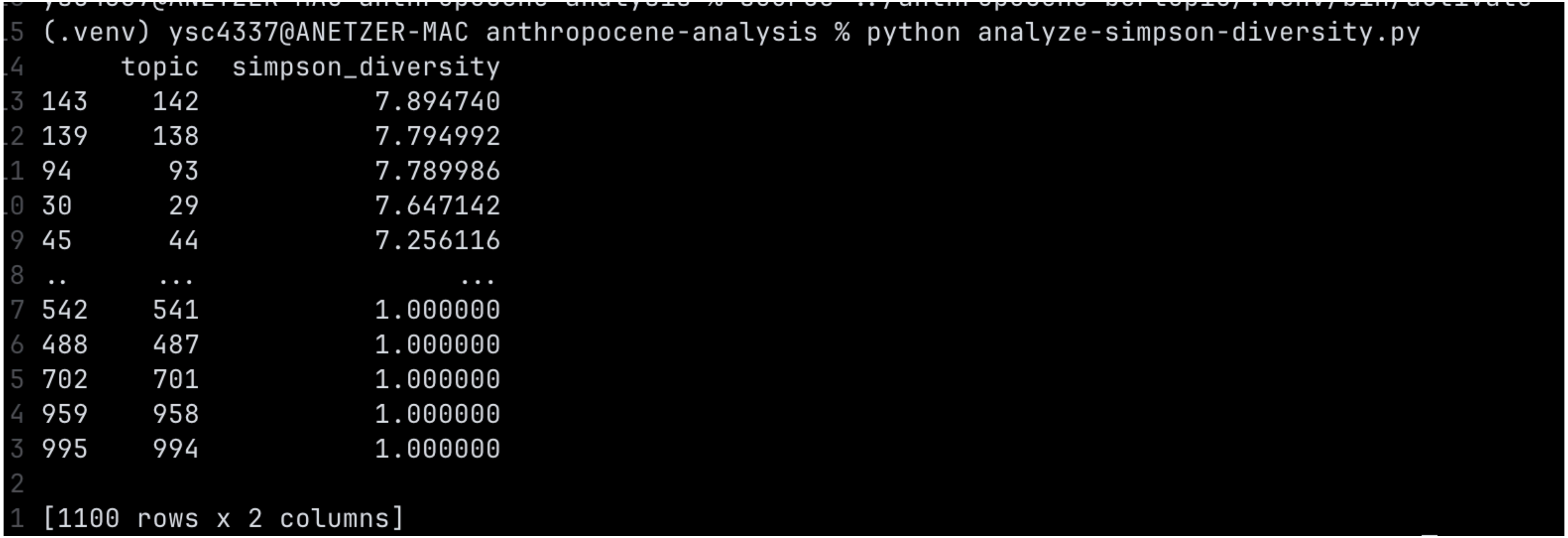
As I would need a lot more compute, and due to the fact that we simply cannot afford to purchase a node on Quest, we moved to AWS.

AWS allowed us near-instant access to massive amounts of compute, allowing us rapid prototyping of ideas.

The takeaway: Humanities labs with scarce funding should take advantage of Quest, and if they need more compute, use AWS.

Results

Calculating the Discipline-Diversity of a Topic



```
ysc4337@ANETZER-MAC anthropocene-analysis % source ../anthropocene-bertopic/.venv/bin/activate
(.venv) ysc4337@ANETZER-MAC anthropocene-analysis % python analyze-simpson-diversity.py
     topic  simpson_diversity
143    142           7.894740
139    138           7.794992
94      93           7.789986
30      29           7.647142
45      44           7.256116
..     ...                ...
542    541           1.000000
488    487           1.000000
702    701           1.000000
959    958           1.000000
995    994           1.000000

[1100 rows x 2 columns]
```

[(‘association’, np.float64(0.007555375419563376)), (‘profession’, np.float64(0.006520569830456999)), (‘toast’, np.float64(0.005526529570120919)), (‘medical’, np.float64(0.005260710303465564)), (‘meeting’, np.float64(0.00505217653478894)), (‘council’, np.float64(0.0038823767105308936)), (‘president’, np.float64(0.003877622025285873)), (‘thanks’, np.float64(0.003505203347242533)), (‘medicine’, np.float64(0.0033858136674474563)), (‘resolution’, np.float64(0.0033416475263302408))]

[(‘canal’, np.float64(0.03677396809378495)), (‘panama’, np.float64(0.03031833597591905)), (‘isthmus’, np.float64(0.012744569049003538)), (‘isthmian’, np.float64(0.012026460550519142)), (‘tonnage’, np.float64(0.011712563668802538)), (‘pacific’, np.float64(0.008868659034418953)), (‘traffic’, np.float64(0.008366589607552702)), (‘treaty’, np.float64(0.007901303044322398)), (‘route’, np.float64(0.006845348256367414)), (‘waterway’, np.float64(0.00650534662751136))]

[(‘hygiene’, np.float64(0.035404014517456596)), (‘temperance’, np.float64(0.021750063084633557)), (‘teaching’, np.float64(0.013291206119155578)), (‘elementary’, np.float64(0.009932335060235685)), (‘education’, np.float64(0.009872305934518627)), (‘instruction’, np.float64(0.009442139657912706)), (‘school’, np.float64(0.008437368185719257)), (‘health’, np.float64(0.007976727868443733)), (‘training’, np.float64(0.007538580290303486)), (‘taught’, np.float64(0.007200627352757128))]

Interpretation of the Data

Top 3 Topic c-TF-IDF scores:

1. association, profession, toast, medical, meeting, council, president, thanks, medicine, resolution
2. canal, panama, isthmus, isthmian, tonnage, pacific, traffic, waterway, treaty, route
3. hygiene, temperance, teaching, elementary, education, instruction, school, health, training, taught

The Core Takeaway

Themes of professional associations, sea-trade, and health/hygiene education bridge the gap between the humanities and the sciences