The Linguistic Features of the Anthropocene

Hello everyone, today, my talk concerns investigating the Linguistic Features of the Anthropocene, and while we will discuss preliminary results that we have obtained, a portion of this talk will be about challenges accessing data and compute resources.

Slide

But a little about me, I am Aerith Netzer, the Digital Publishing and Repository Librarian here at Northwestern. I run our journal publishing operations, where we currently have two journals that I am very proud of, *The Bulletin of Applied Transgender Studies* and *Studies in Russian Philosophy, Literature, and Religious Thought*. In addition to these software development activities, I pinch-hit as a data scientist for our team, as I spent the last several years building and implementing recommender and semantic search systems for retail and life sciences companies to accelerate research before joining Northwestern Libraries.

Slide

This project is the result of a very interdisciplinary team. Heading up the project is James Lee, a professor in Medill and an Associate University Librarian for academic innovation. Alongside him is Professor Han Liu and Lining Mao, in McCormick school of engineering. And representing the humanities are Kelsey Rydland and myself, who are both librarians here.

Slide

Let's talk about that time I used 70 GPUs on Quest. If you don't know, Quest is Northwestern's supercomputing center. The research computing team is absolutely incredible, and if you are a humanist who needs access to big compute centers on the cheap, Quest should be the first place you go. Specifically thanks to David Glass and

Emilio Lehoucq for hanging out with me and installing my special little programs that only I use on Quest. So our workflow looks something like this: you clean the data with NLTK, getting out stopwords and non-english words, as it would contribute little meaning to a model. Stopwords are words like “a”, “if”, “the”, that are far too frequent to really contribute any real information to the model. We then create the embeddings of these cleaned texts with the SentenceTransformers library, we used all-MiniLM-L6-v2, a widely used and computationally inexpensive model to get the text vectors. We then use Uniform Manifold Approximation and Projection to bring the high-dimensional space down to a 2-dimensional space for clustering and visualization. Then, we use the very tersely-titled Hierarchical Density-Based Spatial Clustering of Applications with Noise, which essentially means that it can clustering objects hierarchically, allowing us to see which clusters are closest and furthest from one another. Then, we can run class-based Term Frequency-Inverse Document Frequency, or c-TF-IDF, to extract the most representative terms for specific tops or classes of documents. Over millions of documents, this is a very computationally expensive task. So we try to parallelize the workflow over with many submissions to the Quest supercomputing center, allowing us very fast and efficient processing over many nodes. For one of these runs, I sent in about 100 jobs to the Quest supercomputing center. And I guess that was a slow week for quest because it gave me all of the GPUs at once. I took up so many GPUs that apparently some students in the stats department were talking about “what could a librarian” possibly be doing with all of these GPUs, well, now they know, it was this.

Slide

Even so, because we are trying to rapidly iterate on our research, the queue times in Quest were becoming very long for a free-tier user like me. A library does not really have the funds to drop $50,000 on a GPU node, so we needed another option. This ended us up with

AWS, another sponsor of this conference and again I swear I'm not being paid to promote these it's just a really good product. AWS EC2 allowed us near-instant access to massive amounts of compute with very minimal environment setup, but it does cost some money. Not 50,000 dollars, but it's also not free. For a library, we could not justify the price tag of buying a Quest node that would really only be running jobs a quarter of its' lifespan, and that hardware should go to a team here that would hit it 24 hours a day. The happy medium is AWS EC2, allows renting instead of buying space, which, when you only need it for a few hours per day, is more economical.

If you are a humanist working with very big data, the free-tier allocations at Quest are a boon to you. You should take advantage of them, and if you need more compute, but don't want to feel pressured to use your own node 24/7, then use AWS as the intermediate service.

Slide

Defining a good research question is the first step in any good research project. We were given a grant by the Mellon Foundation to use machine learning techniques to investigate how the "Anthropocene" is discussed a long time horizon and between academic disciplines.

Slide

But first, what is the Anthropocene? The Anthropocene was coined by Paul Crutzen in 2000 to define the current geological epo ch. As we know from middle school science classes, the Earth's time is measured on geological time, with certain epochs including the Pliocene, Pleistocene, and most recently, the Holocene. When Crutzen first defined the Anthropocene in 2000, it initially was a very specific, geological-minded definition of the current geological era. That is to say, the term "Anthropocene" belonged solidly in the field of Geological and Earth science. Over the next 25 years, this

“anthropocene” term, to borrow a term from Deleuze and Guattari, was deterritorialized. The Anthropocene is now discussed in humanist, physical, social, and medical sciences. In other words, the term “Anthropocene” no longer is solidly a rigid proposed definition of the current geological era, and it seems to me now to be a linguistic framing of the current, for lack of better words, “vibe” of the human noosphere. It seems that Crutzen elagantly put words to a *feeling* that people have latched on to in a variety of academic disciplines.

So, we want to investigate how the Anthropocene is discussed over time. It is my hypothesis that people discussed the feeling that comes with “the Anthropocene” before the term was invented. Of course, humanist, medical, social, and physical research into humanity’s impact on the planet did not begin in 2000, nor will it end whenever the term Anthropocene falls out of favor. Thus, our problem is to see in what ways writers linguistically framed this feeling of the Anthropocene, while not actually using the term Anthropocene.

Slide

Our first step is to identify the meaning of the word Anthropocene using the context of surrounding words. As one would expect, for example, the term “ice” and “environment” have high probabilities of being semantically similar to the “Anthropocene.” Thus, we can intuit that when writers discuss ice, or the environment, they are putting words to this feeling of the Anthropocene, without actually using the word “Anthropocene itself.” Using Word2vec, we created a list of seed words that are most semantically similar to the Anthropocene, with which we wanted to get a dataset of all the works we could find that mention these “seed words.”

Slide

But first, a short detour to talk about the real reason we are here, computation and data for research. There are many many products

out there that sell literary data from academic articles, newspapers, pamphlets, novels, and any other piece of media one could hope to find, but each have distinct problems that are difficult to overcome. At first, we tried Open Alex, which provided article n-grams up until late last year. OurResearch, the company behind it, turned off the n-grams API, completely halting our research and essentially forcing us to start over again. After this, we tried Semantic Scholar for a while, which, while the data behind the text itself is very good, had very sparse timestamp data, which disallowed us from seeing trends over time. We ended up using Constellate, a product from the people behind JSTOR. Constellate acts as more of a "data portal" rather than a place to do very large scale analysis in a unified environment. While Constellate itself will be deprecated in July, I have heard that they will be keeping their Data for Research program, and they are very communicative and helpful in this project. But the point of this is that we need better open access academic articles so that free projects can scrape them and make them available for text mining, without any barriers in licensing or copyright. This is now going to be a shout-out for my Digital Publishing operation, that if your field lacks diamond open access publishing venues, meaning that authors pay no open access fees and readers also pay no fees, get in contact with me and we can talk about setting up a cutting-edge journal for your field that can compete on every front with the big publishers.

Slide

Regardless, now that we have the data, we can start building some actually cool stuff. We chose BERTopic for its' ease of use, modularity, and capacity for time-dependent and class-based modelling. If you are a digital humanist doing big data with text, RAPIDSAI is quite possibly the greatest thing in the world and will change your life. It is a project from NVIDIA, and offers 100% API compatibility with `pandas` and `scikit-learn`, but implemented in CUDA, their GPU parallel processing framework. RAPIDS allowed us to iterate so so so much faster than we could have done otherwise. I

know NVIDIA sponsored this event but I promise they did not pay me to say this but their tooling is so far beyond what anyone else is doing and it is open source, which is very nice.

So now that we have a CUDA-enabled workflow to bring analyzing our time-to-insight from hours to minutes.

Slide

We can move to the results. Here, we took a clue from ecology, quite fitting, I think, concerning the topic, calculate the inverse of the probability that two randomly selected samples will belong to the same species. Here, we have several hundred topics, which represent a "species." JSTOR gives each document a tdmCategory, defining whether the document belongs to arts, the physical science, the medical sciences, *et cetera*. So, within each topic, we want to calculate the probability that two randomly selected documents will belong to the same tdmCategory. This allows us to calculate the diversity, or interdisciplinarity, of a given topic.

Our preliminary results include the following.