Probing Heterogeneous Pretraining Datasets with Small Curated Datasets

Gregory Yauney Cornell University gyauney@cs.cornell.edu

Emily Reif Google Research ereif@google.com

David Mimno Cornell University mimno@cornell.edu

Why do we need to characterize pretraining datasets?

- Seemingly innocuous dataset curation decisions impact models [1, 2].
- "Quality" filters are ubiquitous but often narrowly select data that is similar to books and Wikipedia articles [2, 3].
- Recent work has found that pretraining data source composition affects downstream performance [2, 4, 5].
- The era of free data is over. We need to be able to make decisions about pretraining dataset composition.

What's in language model pretraining datasets?

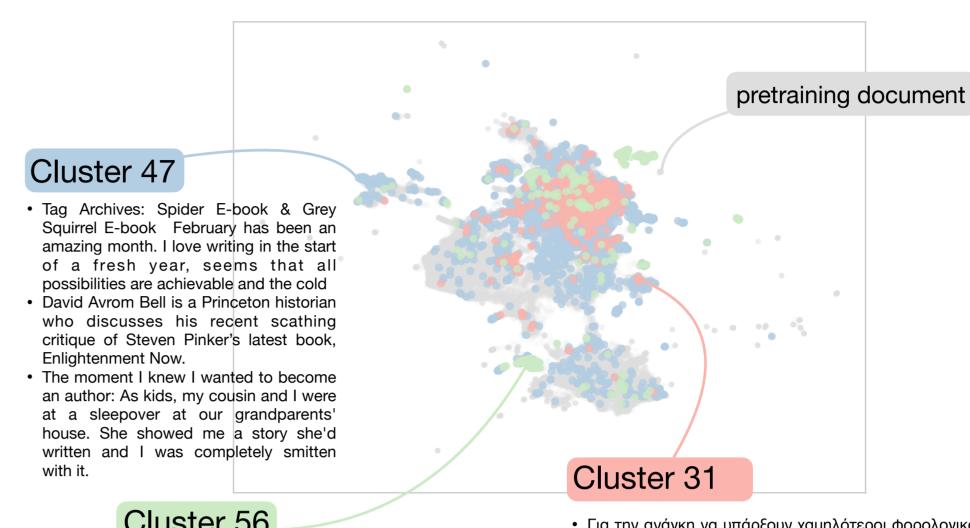
- Scale prevents us from exhaustively describing what is in the datasets [1].
- The Pile and C4 are two of the most-used pretraining datasets. They consist of web-scraped data that is often described only by domain or website of origin.
- We'll use the Pile as a running example. It contains docs from 22 domains.

Related Work

[1] Dodge et al. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. 2021. [2] Longpre et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. 2023.

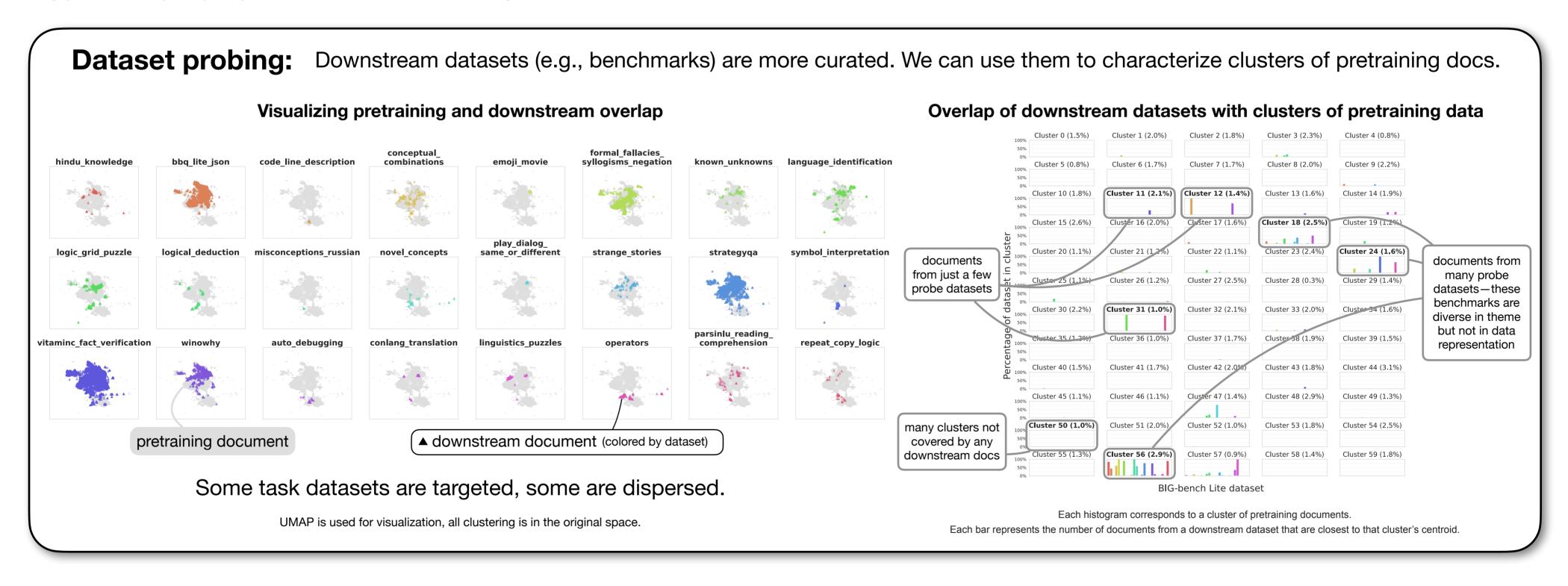
- [3] Chowdhery et al. PaLM: Scaling language modeling with pathways. 2022.
- [4] Xie et al. DoReMi: Optimizing data mixtures speeds up language model pretraining. 2023.
- [5] Nguyen et al. Quality not quantity: On the interaction between dataset design and robustness of CLIP. 2022.

Clustering pretraining datasets is a start



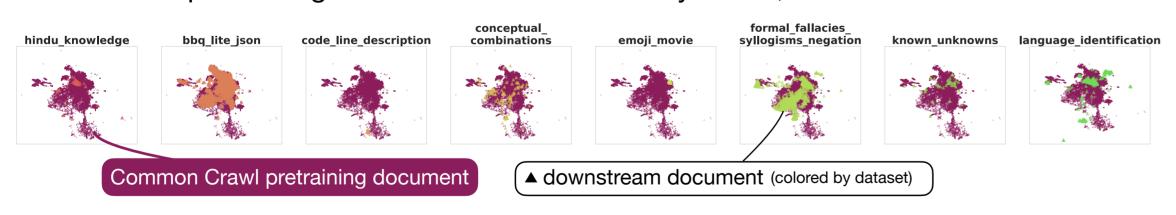
- Cluster 56
- · Some of the designs included in this collection are Pictured. You must have the necessary hardware and software to transfer these designs to your embroidery machine. Shipping for this Item is only \$1.
- · So... I think all it takes to resolve this issue are extra options in the nvidia-settings utility to disable PowerMizer or at least make it run only when on batteries. Is nvidia working on this already
- Re: How do I get my address to show up on the google maps I assume you'd like to learn more about local business listing that can be easily found on Google.com and Google Maps. Am I right?
- Για την ανάγκη να υπάρξουν χαμηλότεροι φορολογικοί συντελεστές, αλλά και βελτίωση της φορολογικής συμμόρφωσης, έκανε λόγο ο Κυριάκος Μητσοτάκης, επισημαίνοντας πως «η Νέα Δημοκρατία είναι το κόμμα το
- Zobacz wideo Spraw a nagrania we wtorek nie zajmowała si e prokuratura, ale, jak ustalilismy w Ministerstwie Spraw Wewn etrznych i Administracji bada j a policja. REKLAMA Funkcjonariusze wyst apili
- Como se había estimado, durante el mes pasado la fábrica de aerogeneradores de media y baja potencia ubicada en la localidad neuquina de Cutral Co ya está dando sus primeros pasos. El emprendimiento

Difficult to characterize clusters. Some are easier to Challenge: qualitatively describe than others.

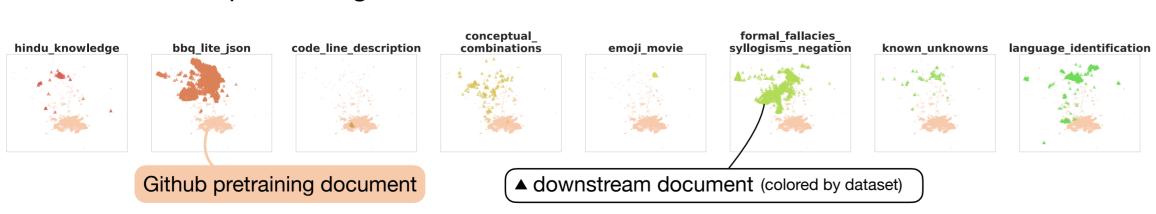


Dataset probing does not simply recover domains: overlap is not always explained by dataset source

Some pretraining metadata domains are very broad, like Common Crawl:



Some pretraining metadata domains are more narrow, like Github:



Cluster 4 Cluster 5 Cluster 9 Cluster 15 clusters contain documents from just a few Cluster 29 domainsdownstream datasets cut across these Cluster 44

Takeaways

- 1. Well-curated small datasets can characterize large web-scraped datasets, complementing current heterogeneity measures.
- 2. Even current broad evaluations are not enough to evaluate models' data.

Future work

- 1. Can we build effective quality filters that use this finer-grained characterization of pretraining data?
- 2. Can we prune pretraining datasets based on data overlap? Or is the seemingly unrelated data necessary for linguistic support?