Data behind LLM

Jie Tang
Department of Computer Science & Technology
Tsinghua University

Table of Contents

Table of Contents

- Data behind large language models
  - Common Crawl
  - WebText and OpenWebText
  - Colossal Clean Crawled Corpus (C4)
  - GPT-3 dataset
  - The Pile
- Documentation for datasets

Data behind large language models

- Large language models are trained on “raw text”. To be highly capable (e.g., have linguistic and world knowledge), this text should span a **broad** range of domains, genres, languages, etc.
- **Web**: a major source.
  - Huge. the Google search index is 100 petabytes (reference).
- **Private datasets** that reside in big companies are even larger than what’s available publicly.
  - WalMart generates 2.5 petabytes of data each hour!

Data behind large language models

- Common Crawl
- WebText and OpenWebText
- Colossal Clean Crawled Corpus (C4)
- GPT-3 dataset
- The Pile

Common Crawl

- **Common Crawl** is a nonprofit organization that crawls the web and provides snapshots that are free to the public.
- Scale: The April 2021 snapshot of Common Crawl has 320 terabytes of data
- Applications: A standard source of data to train many models such as T5, GPT-3, and Gopher.

Common Crawl

Representation harms. Despite the richness of web data, it has been noted in Bender et al, 2021 that:

- Despite the size, large-scale data still has **uneven representation** over the population.
- Internet data overrepresents younger users from developed countries.
- GPT-2's training data is based on Reddit, which according to Pew Internet Research's 2016 survey, 67% of Reddit users in the US are men, 64% between ages 18 and 29.
- 8.8-15% of Wikipedians are female.
- Filtering "bad words" could further marginalize certain populations (e.g., LGBT+).

Takeaway: it is crucial to understand and document the composition of the datasets used to train large language models.

WebText

- WebText: **dataset used in training GPT-2**
- Goal: obtain **diverse but high-quality** dataset.
- Previous work:
  - Datasets were trained on news, Wikipedia, or fiction.
  - Common Crawl contains a lot of junk (gibberish, boilerplate text).
  - Trinh & Le, 2018 selected a tiny subset of Common Crawl based on n-gram overlap with the target task.
- Process for creating WebText:
  - Scraped all outbound links that received at least 3 karma (upvotes).
  - Filtered out Wikipedia to be able to evaluate on Wikipedia-based benchmarks.
  - End result is 40 GB of text.
- WebText was not released by OpenAI

WebText and OpenWebText

- OpenWebText: WebText was replicated (in spirit) by the OpenWebText dataset.
- Process for creating OpenWebText:
  - Extracted all the URLs from the Reddit submissions dataset.
  - Used Facebook's fastText to filter out non-English.
  - Removed near duplicates.
  - End result is 38 GB of text.

Dataset: WebText and OpenWebText

- **Toxicity analysis**.
  - RealToxicityPrompts (Gehman et al. 2020) : a dataset of 100K naturally occurring, sentence-level prompts derived from a large corpus of English web text, paired with toxicity scores from a widelyused toxicity classifier.
  - Analyzed these two datasets and found:
    - 2.1% of OpenWebText has toxicity score >= 50%
    - 4.3% of WebText (from OpenAI) has toxicity score >= 50%
    - News reliability correlates negatively with toxicity (Spearman ρ=−0.35)
    - 3% of OpenWebText comes from banned or quarantined subreddits, e.g., /r/The_Donald and /r/WhiteRights
- Takeaway: We can construct datasets to evaluate toxicity score of large datasets / models.

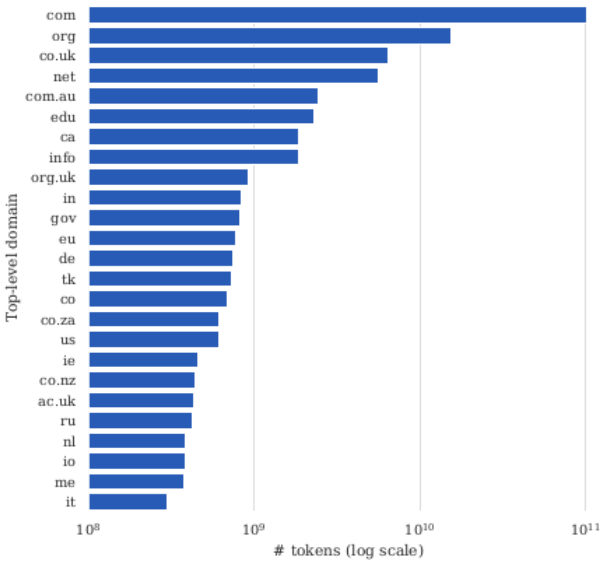
Colossal Clean Crawled Corpus (C4)

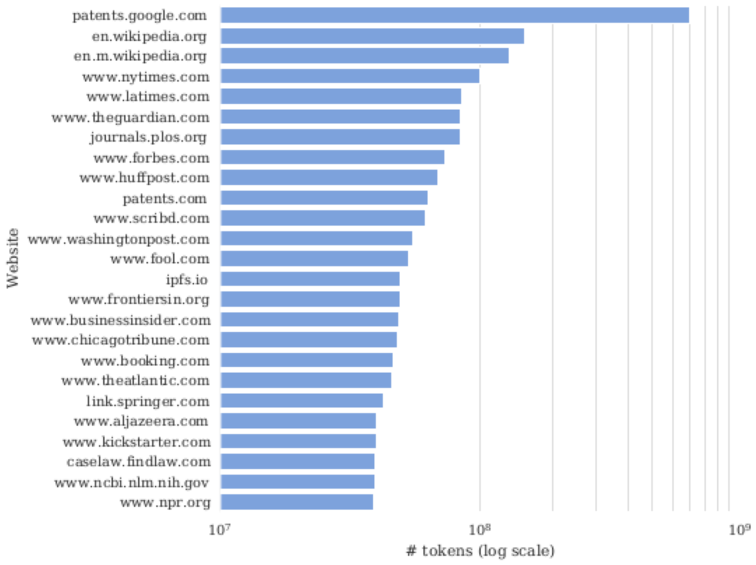
- The Colossal Clean Crawled Corpus (C4) is a larger was created to train the T5 model.
- Processes for constructing C4
  - Started with April 2019 snapshot of Common Crawl (1.4 trillion tokens)
  - Removed “bad words”
  - Removed code (“{“)
  - langdetect to filter out non-English text
  - Resulted in 806 GB of text (156 billion tokens)

Colossal Clean Crawled Corpus

Analysis. Dodge et al. 2021 performed a thorough analysis of the C4 dataset.

- A surprising amount of data from patents.google.com
- 65% pages in the Internet Archive; out of those, 92% pages written in the last decade
- 51.3% pages are hosted in the United States; fewer from India even though lots of English speakers there
- Some text from patents.google.com are automatically created, and thus have systematic errors:
  - Filed in a foreign country's official language (e.g., Japanese) is automatically translated into English
  - Automatically generated from optical character recognition (OCR)





Benchmark data contamination

- When we are evaluating the capabilities of large language models using benchmark data (e.g., question-answer pairs), it makes a difference whether the benchmark data appears in the training data of the language model. If so, then the benchmark performance will be **biased** up.
- Normally, in machine learning, data hygiene (keeping the training data separate from the test) is relatively easy, but **in the case of large language models**, both the training data and benchmark data are derived from the Internet, **it can be difficult to a priori guarantee their separation.**

Benchmark data contamination in C4

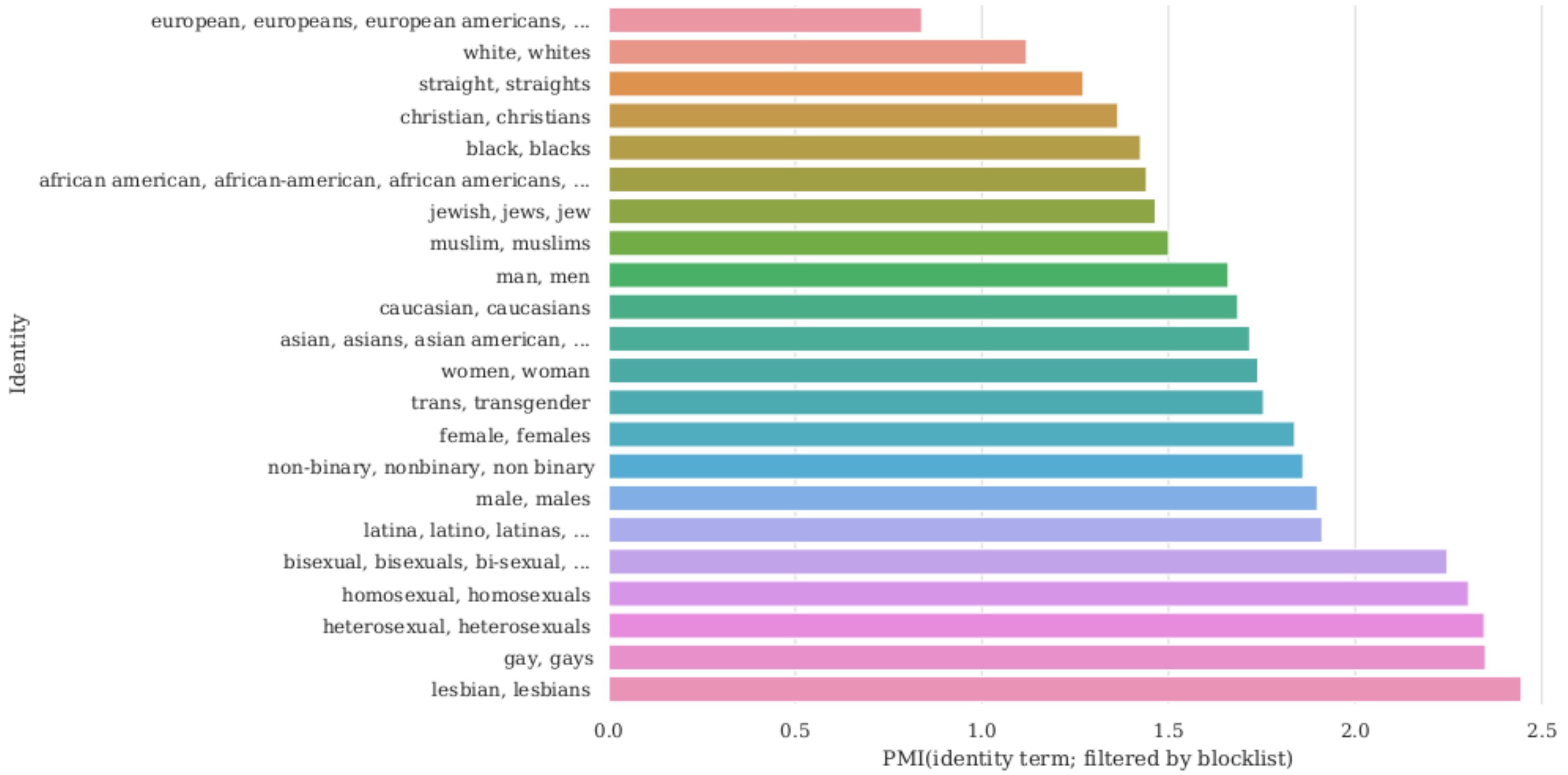
- There are two types of contamination:
  - **Input-and-output contamination**: both the input and output appear in the training data. Varies from 1.87% to 24.88% (XSum is 15.49%).
  - **Input contamination**: the input appears in the training data. Varies from 1.8% to 53.6% (QNLI, which is derived from Wikipedia).
- Note that contamination is not due to hosting datasets (as they are usually stored in a JSON file, not as a webpage).

Harms in C4

- **Representational harms**
  - They look at co-occurrence with ethnicity terms (e.g., *Jewish*) and sentiment-bearing words (e.g., *successful*).
  - *Jewish* has 73.2% positive sentiment, *Arab* has 65.7% positive (7.5% difference).
  - Variation across sites (New York Times had a 4.5% difference, Al Jazeera had 0% difference).
- **Allocational harms**
  - Recall C4 is a filtered version of Common Crawl (only about 10%).
  - Mentions of sexual orientations (e.g., *lesbian*, *gay*) more likely to be filtered out; of those filtered out, non-trivial fraction are non-offensive (e.g., 22% and 36%).
  - Certain dialects are more likely to be filtered (AAE: 42%, Hispanic-aligned English: 32%) than others (White American English: 6.2%)

Allocational harms in C4

- **Pointwise Mutual Information (PMI)** between identity mentions and documents being filtered out by the blocklist. Identities with higher PMI (e.g., lesbian, gay) have higher likelihood of being filtered out.



GPT-3 dataset

1. Selected subset of Common Crawl that's **similar to a reference dataset** (WebText).
   1. Downloaded 41 shards of Common Crawl (2016-2019).
   2. Trained a binary classifier to predict WebText versus Common Crawl.
   3. Sampled (kept) a document with higher probability if classifier deems it more similar to WebText.
2. Performed **fuzzy deduplication** (detect 13-gram overlap, remove window or documents if occurred in <10 training documents), removing data from benchmark datasets.
3. Expanded the diversity of the **data sources** (WebText2, Books1, Books2, Wikipedia).

4. During training, Common Crawl is **downsampled** (Common Crawl is 82% of the dataset, but contributes only 60%).

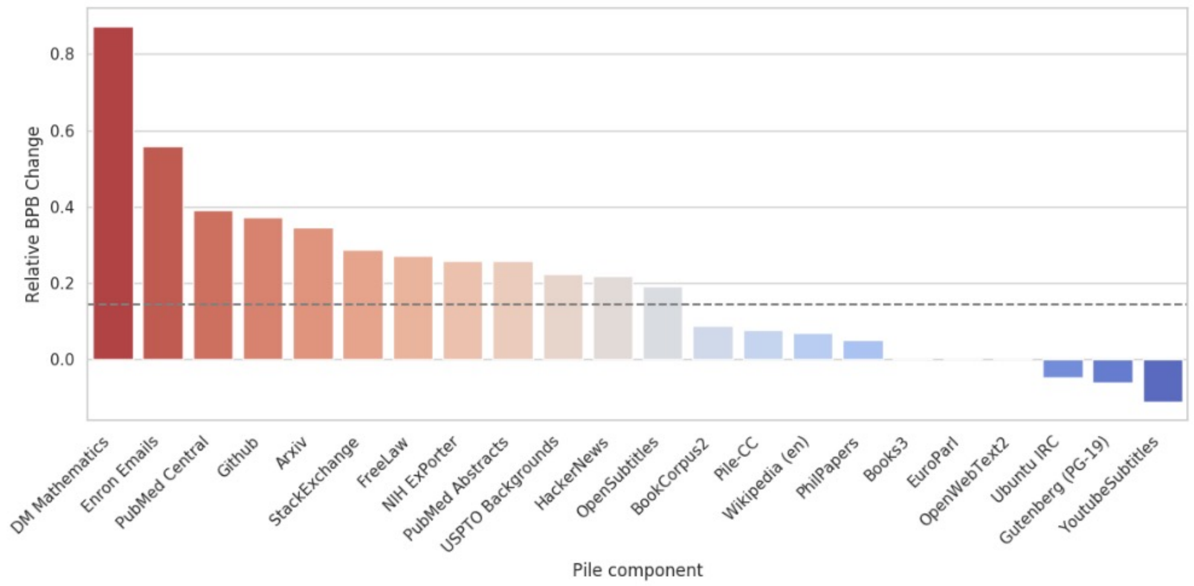
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

The Pile

- While a web crawl is a natural place to look for broad data, it’s not the only strategy, and GPT-3 already hinted that it might be productive to look at other sources of higher quality.
- EleutherAI (a nonprofit organization committed to building open language models), pushed this idea even farther. They released The Pile, a dataset for language modeling, where the key idea is to source it from smaller high-quality sources (academic + professional sources).
- **Data composition**.
  - 825 GB English text
  - 22 high-quality datasets

The Pile

- The key idea is to source it from smaller high-quality sources (academic + professional sources).
- contains a lot of information that's not well covered by GPT-3's



They also performed analysis of pejorative content, gender/religion biases. The findings are qualitatively similar to previous work.

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

Takeaway

- The total amount of data out there (web, private data) is massive.
- Training on “all of it” (even Common Crawl) doesn’t work well (not effective use of compute).
- Filtering / curation (OpenWebText, C4, GPT-3 dataset) is needed, but can result in biases.
- Curating non-web high-quality datasets is promising (The Pile).
- Important to carefully document and inspect these datasets.

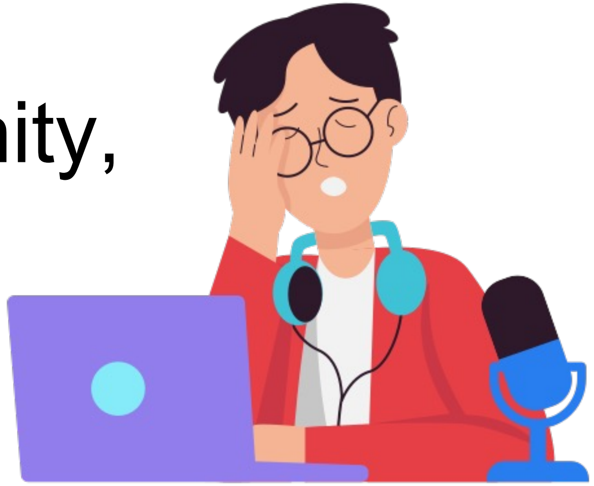
Table of Contents

- Data behind large language models
  - Common Crawl
  - WebText and OpenWebText
  - Colossal Clean Crawled Corpus (C4)
  - GPT-3 dataset
  - The Pile
- Documentation for datasets

Documentation for datasets

- Documentation is important
  - Examples from other fields:
    - **Electronics industry** has a well-established protocol where every component has a datasheet with operating characteristics, test results, recommended and usage.
    - **Nutrition labels**: The FDA mandates that food be labeled with their nutrition content.
- But within the machine learning community,

it has been a fairly ad-hoc process…



Documentation for datasets

- **Datasheets for datasets** (Gebru et al., 2018) is an influential paper that provides community norms around documentation.
- **Data statements** (Bender & Friedman, 2018) is related framework that is more tailored to language datasets.
- The emphasis is on **transparency**.
- Two purposes:
  1. Dataset creators: reflect on decisions, potential harms (e.g., social biases) when creating the dataset.
  2. Dataset consumers: know when the dataset can and can't be used.

Documentation for datasets

A sample of the questions from each category are provided below:

- **Motivation**
  - For what purpose was the dataset created?
- **Composition**
  - What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  - Is any information missing from individual instances?
- **Collection process**
  - How was the data associated with each instance acquired?
  - Who was involved in the data collection process?

Documentation for datasets

- **Preprocessing/cleaning/labeling**
  - Was any preprocessing/cleaning/labeling of the data done?
  - Is the software that was used to preprocess/clean/label the data available?
- **Uses**
  - Has the dataset been used for any tasks already?
  - Are there tasks for which the dataset should not be used?
- **Distribution**
  - How will the dataset will be distributed?
- **Maintenance**
  - Who will be supporting/hosting/maintaining the dataset?
  - Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

As an example, let's look at the datasheet for The Pile.

Thank you！

Jie Tang, KEG, Tsinghua University

http://keg.cs.tsinghua.edu.cn/jietang
https://github.com/THUDM