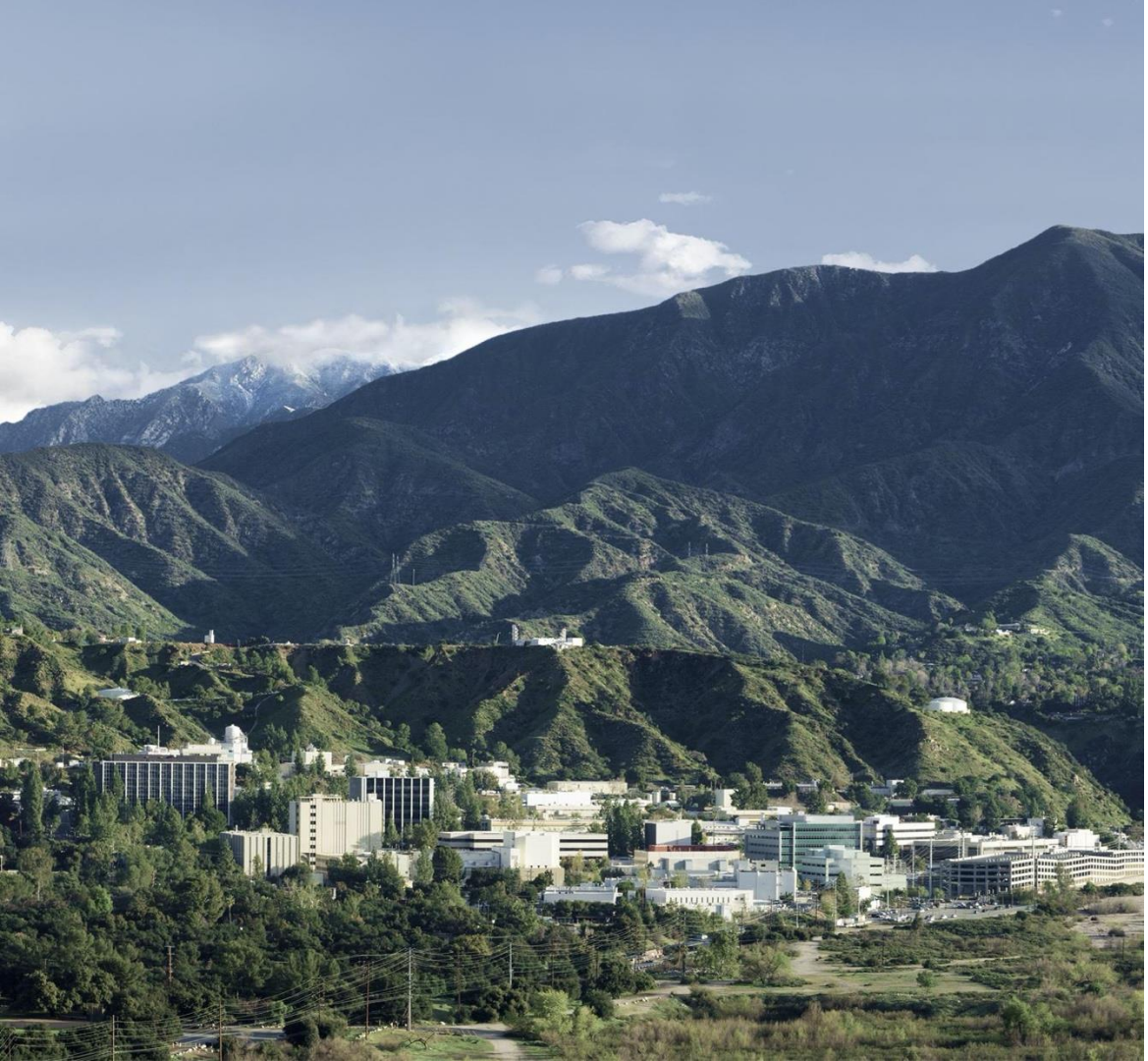


CC-MAIN-2021-31-PDF-UNTRUNCATED

New Corpus of Nearly 8 Million PDFs

May 25, 2023

The research was carried out at the NASA (National Aeronautics and Space Administration) Jet Propulsion Laboratory, California Institute of Technology under a contract with the Defense Advanced Research Projects Agency (DARPA) SafeDocs program.

Jet Propulsion Laboratory
California Institute of Technology

Authors and JPL Team

- Authors (alphabetical): Tim Allison (JPL), Mike Milano (JPL), Ryan Stonebraker (JPL), Peter Wyatt (PDF Association)

- JPL Team: Chris Mattmann (PI), Wayne Burke, Dustin Graf, Anastasia Menshikova, Philip Southam

Debts of Gratitude!

- Sergey Bratus and DARPA's SafeDocs program
- Simson Garfinkel and Digital Corpora (https://digitalcorpora.org/) for publishing this corpus
- Sebastian Nagel and Common Crawl (https://commoncrawl.org/)
- Amazon Open Data Sponsorship Program (https://aws.amazon.com/opendata/open-data-sponsorship-program/)

~8 Million PDFs

- ~8 million Portable Document Format (PDF) files
- Gathered from the web in August/September 2021
- Packaged in 7,933 zip files and freely available:

https://digitalcorpora.org/cc-main-2021-31-pdf-untruncated/

https://pdfa.org/new-large-scale-pdf-corpus-now-publicly-available/

Heaps of Files – Why?

- ~~Parser~~ DDL developers
- Security research
- Privacy research
- ML/AI – file structure (parser induction), document structure, categorization
- Digital preservation

Publicly available corpora save researcher/developers the burden, cost and challenges of web crawling and finding the files they want.

8 Million PDFs – Why PDF?

- PDF is ubiquitous
- PDF represents all categories of documents:
  - Books, presentations, CAD, medical, legal, posters, academic papers, …
- PDF represents all languages
- PDF represents all types of content
- PDF (as a file format) is challenging:
  - Is created by many, many different software tools… much variation
  - Is consumed or rendered by many software applications
  - Is leveraged by cyber attackers

8 Million PDFs – Why 8 million?

- Single snapshot in time
- Representative of the public web
  - Equal representation from around the world
- Goldilocks sized corpus
  - Large, yet manageable
  - Larger than all previous PDF-centric corpora
- Complexity requires many PDFs to adequately represent:
  - Classes of document
  - Breadth of content
  - File size, page count
  - PDF file format features and variants
  - …

Common Crawl – An Overview

- Monthly public crawl of a portion of the web – March/April 2023 crawl is ~3 billion files/400TB uncompressed
- Hosted by AWS Open Data Sponsorship Program
- Data stored in Web ARChive (WARC) files
- Files truncated at 1 MB
- Used extensively by big data projects – large language models (LLMs), machine translation, language identification

GPT-3 and C4 (at least)

Commoditization of Large Language Models

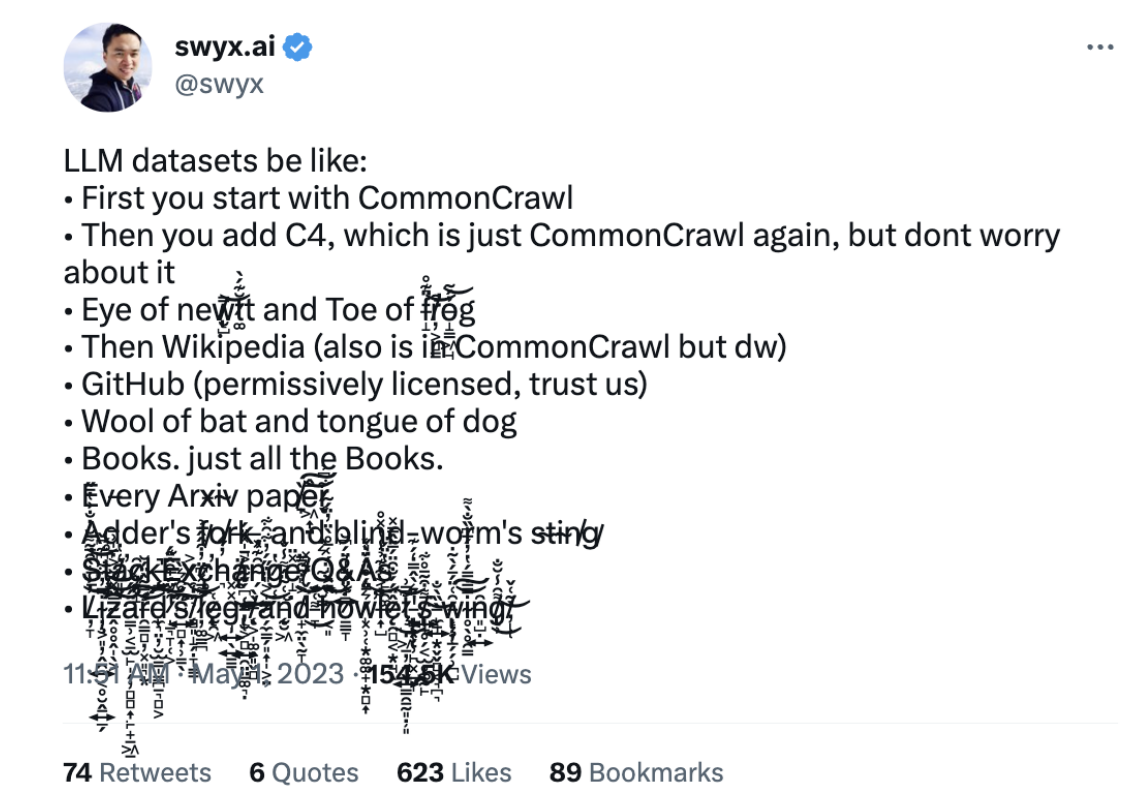
Jul 5, 2022

GPT-3 ushered in a new era of large language models (LLMs) that could generate human-realistic text output. But GPT-3 didn't come from a company with a large and proprietary dataset. Instead, the dataset consisted of:

- 410 billion tokens from the public Common Crawl (60% weight)
- 19 billion tokens from Reddit submissions with a minimum score of 3 (22%)
- 12 billion tokens from "Books1" and 55 billion from "Books2", which are probably books downloaded from the Library Genesis archive (a pirated ebook dataset) (8% each)
- 3 billion tokens from Wikipedia (3%)

https://matt-rickard.com/commoditization-of-large-language-models

CommonCrawl and LLMs



Common Crawl is an important component of the, ahem, secret data sauce of LLMs

https://twitter.com/swyx/status/1653064637611651077

JPL Team’s contributions

- Extracting 6 million files from Common Crawl’s WARC files
- Refetching 2 million files from their original URLs
- Packaging the files into zip files
- Geolocating the source URLs with MaxMind’s free geo-ip database
- Metadata tables

zips

S3 Downloads Browser

corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/zipfiles/ sub-dirs:

- 0000-0999/
- 1000-1999/
- 2000-2999/
- 3000-3999/
- 4000-4999/
- 5000-5999/
- 6000-6999/
- 7000-7999/

corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/zipfiles/0000-0999/ files:

SHOW FILE HASHES

Name	Size	Last Modified
0000.zip	1,266,879,273	2023-03-15 01:24:43Z
0001.zip	1,660,857,091	2023-03-15 01:24:54Z
0002.zip	1,298,887,150	2023-03-15 01:25:08Z
0003.zip	1,234,547,316	2023-03-15 01:25:20Z
0004.zip	1,310,167,182	2023-03-15 01:25:31Z

https://downloads.digitalcorpora.org/corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/zipfiles/

Metadata

corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/metadata/ files:

SHOW FILE HASHES

Name	Size	Last Modified
cc-hosts-20230303.csv.gz	247,695,554	2023-03-07 02:30:47Z
cc-hosts-20230324-1k.csv	104,250	2023-03-25 03:40:02Z
cc-provenance-20230303.csv.gz	1,293,543,097	2023-03-07 02:30:45Z
cc-provenance-20230324-1k.csv	423,859	2023-03-25 03:40:06Z
pdfinfo-20230315.csv.gz	328,982,096	2023-03-25 03:40:14Z
pdfinfo-20230324-1k.csv	187,255	2023-03-25 03:40:18Z

https://downloads.digitalcorpora.org/corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/metadata/

But I don’t want PDFs!

Commoncrawl-fetcher-lite

1) extracts non-truncated files from common crawl
2) creates a list of URLs for truncated files (for later, optional, fetching)

https://github.com/tballison/commoncrawl-fetcher-lite

```
"indices": {
  "paths": [
    "crawl-data/CC-MAIN-2023-06/cc-index.paths.gz",
    "crawl-data/CC-MAIN-2022-49/cc-index.paths.gz"
  ]
},
"recordSelector": {
    "must": {
        "status": [
            {
                "match": "200"
            }
        ]
    },
    "should": {
        "mime_detected": [
            {
                "match": "video/mp4"
            },
            {
                "match": "video/quicktime"
            }
        ]
    }
}
```

Extras

PDF Sizes

Size	Counts
<1kb	7,275
<10kb	109,092
<100kb	1,671,726
<1mb	4,604,023
<10mb	1,692,893
<100mb	210,735
<=1gb	3,686
>1gb	2

Metadata – Two files for each table

- *-1k.csv includes the first 1k rows of the table so that humans can open the file in Excel or similar to get a sense of the data (UTF-8 BOM included)
- *.csv.gz are the full tables for machine reading (no BOM)

Metadata – Provenance

- Source URL
- Common Crawl warc file and offsets
- Whether CC identified the file as truncated or not
- File size and sha256 of fetched/refetched bytes

Metadata – Hosts

- Max Mind's geolite2 database

B	C	D	E	F	G	H
file_name	host	tld	ip_address	country	latitude	longitude
0000000.pdf	augustaarchives.com	com	66.228.60.22	US	33.7485	-84.3871
0000001.pdf	demaniocivico.it	it	185.81.4.146	IT	41.8904	12.5126
0000002.pdf	www.polydepannage.com	com	188.165.112.	FR	48.8582	2.3387
0000003.pdf	community.jisc.ac.uk	uk	52.209.218.9	IE	53.3382	-6.2591
0000004.pdf	www.molinorahue.cl	cl	172.96.161.2	US	34.0584	-118.278
0000005.pdf	www.delock.de	de	217.160.0.15	DE	51.2993	9.491

Metadata – Poppler pdfinfo utility

- About the PDF document (*not content*!):
  - PDF version
  - Creator tool
  - Producer application
  - Creation and Modification dates
  - Number of pages
  - Several other technical PDF features
    - Semantically tagged, web optimized, contains JavaScript, …

Metadata – Apache Tika (coming soon!)

- Automatic language identification
- Out of vocabulary statistic
- Attachments (types and counts)
- Incremental update counts

Recent Related Work

Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, Filip Graliński, “CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data”, ICDAR 2023.

Paper: https://arxiv.org/abs/2304.14953

Scripts: https://github.com/applicaai/CCpdf