Open Data for Open Models

Infrastructure, Standards, and Transparency

AI4IA Conference 2025

Thom Vaughan, Common Crawl Foundation

COMMON CRAWL

What is Common Crawl?

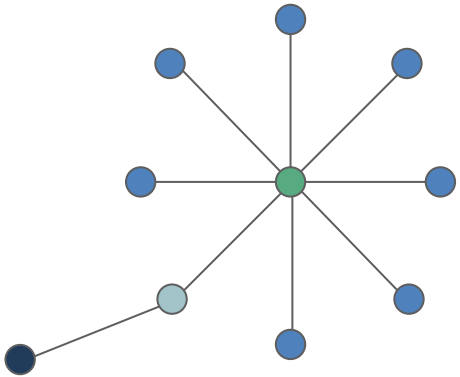
COMMON CRAWL

What is Common Crawl?

- 501(c)(3) nonprofit, started in 2007 by Gil Elbaz
- Free and open corpus containing > 300 billion pages over > 18 years
- Over 10 PB in `WARC` [1] and yet more in `WAT`, `WET`, and `BVGraph` [2]
- Hosted on AWS S3 as an Open Data set, enabled by the AWS Open Data Sponsorship Program
- Cited in > 10,000 research papers [3] [4]
- 3–4 billion new pages added each month
- Broad crawl steered by link-based Harmonic Centrality ranks
- Web Graphs showing the structure and connectivity of the web with host and domain ranks

COMMON CRAWL

[1] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/
[2] https://webgraph.di.unimi.it/docs/it/unimi/dsi/webgraph/BVGraph.html
[3] https://github.com/commoncrawl/cc-citations [4] https://commoncrawl.org/research-papers



$$H(v) = \sum_{u \neq v} \frac{1}{d(v,u)}$$

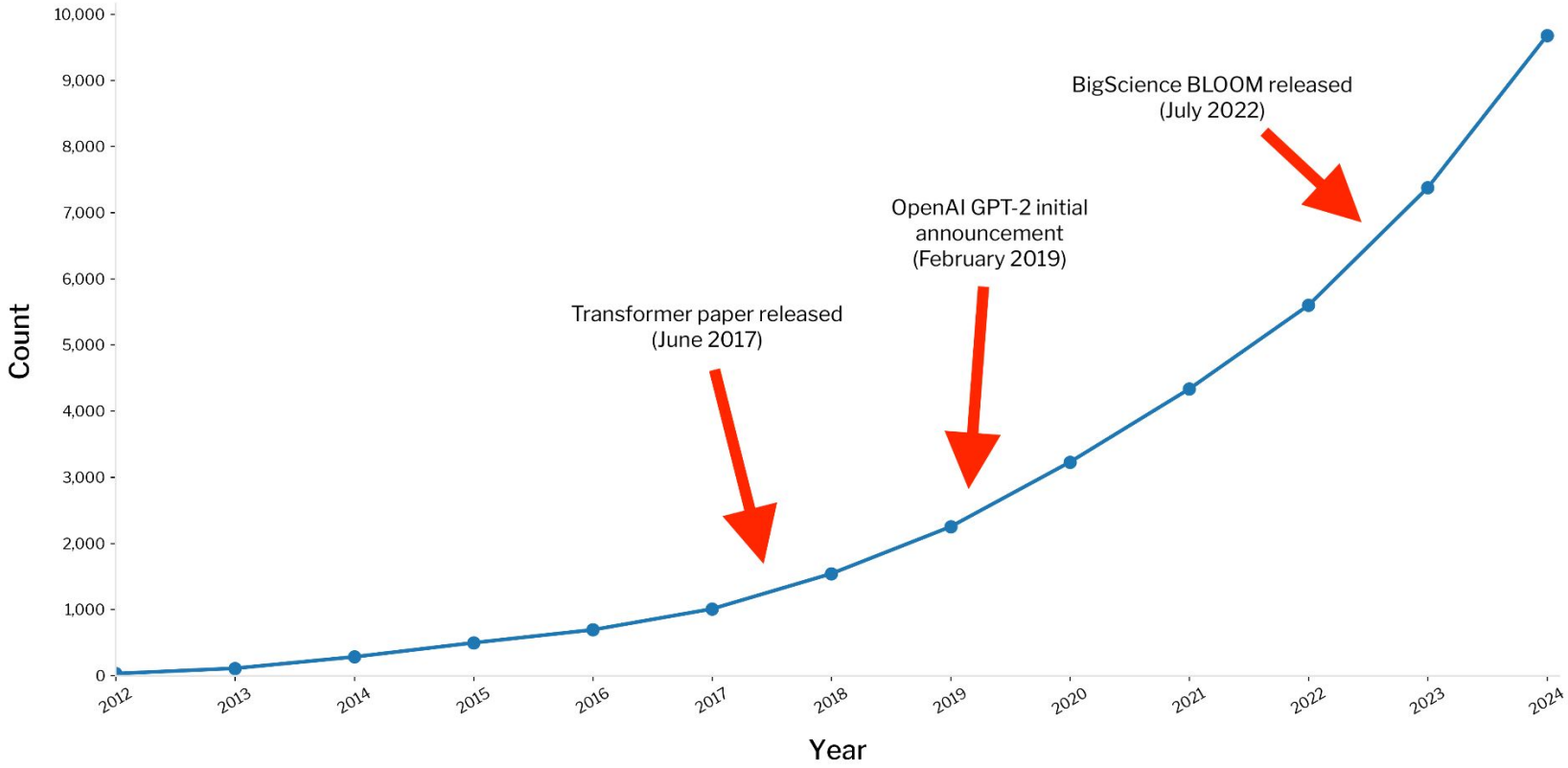
Where $H(v)$ is the **Harmonic Centrality** of vertex v,
and $d(v,u)$ is the shortest path distance between vertices v and u.

COMMON CRAWL

Why is it important?

COMMON CRAWL

Plot of Common Crawl citations (cumulative) in Google Scholar until January 2025



https://commoncrawl.org/research-papers
https://huggingface.co/datasets/commoncrawl/citations
https://github.com/commoncrawl/cc-citations

For comparison, the Hubble Space Telescope is cited in 1,000 papers/year

Common Crawl's Use in AI and Elsewhere

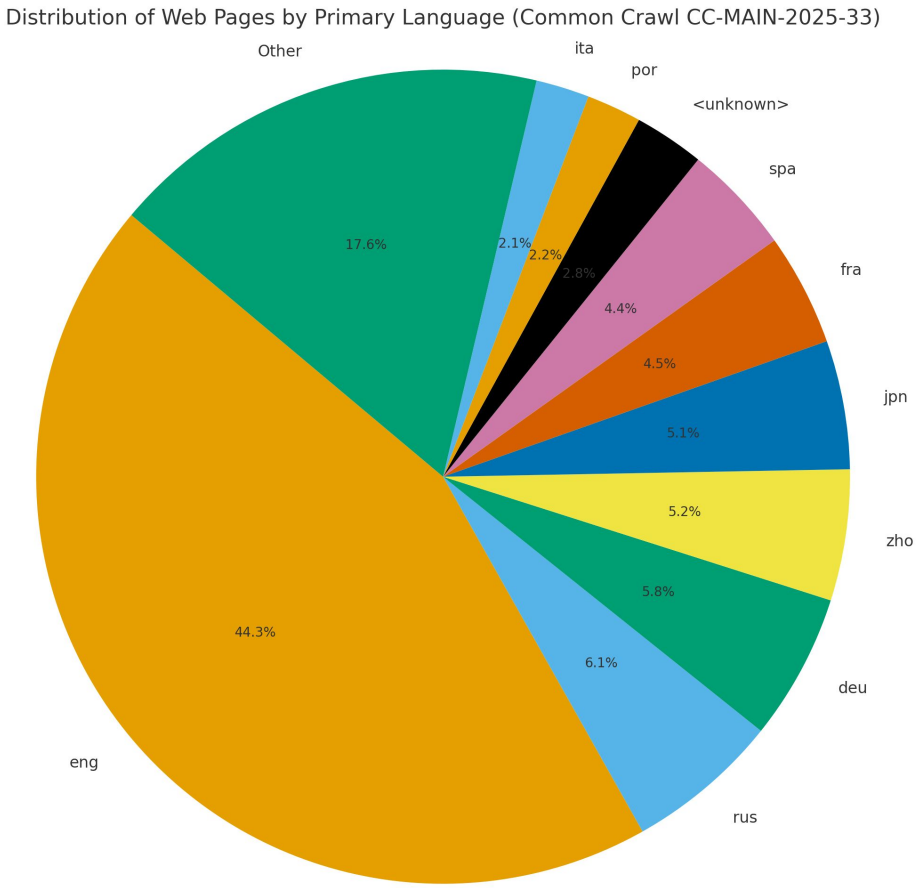
- Common Crawl datasets are instrumental in advancing machine learning and artificial intelligence , particularly in natural language processing and web content analysis. [1]
- Common Crawl is the largest freely available web-crawled dataset and a cornerstone of pre-training data for LLMs, with over 80% of GPT-3's tokens derived from it. [2]
- OSCAR [3], mC4 [4], The Pile [5], and many other foundational datasets are all either *directly derived* from Common Crawl or include large portions of data that originate from it.

COMMON CRAWL

[1] https://deepgram.com/ai-glossary/common-crawl-datasets
[2] https://dl.acm.org/doi/10.1145/3630106.3659033
[3] https://huggingface.co/datasets/oscar [4] https://huggingface.co/datasets/mc4 [5] https://pile.eleuther.ai/

Challenges with languages

COMMON CRAWL



https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

COMMON CRAWL

Global Safety

- Multilingual (especially LOTE) content is comparatively sparse [1]
- There are serious risks to deploying LLMs in other languages [2]
- GPT, PaLM2, LLaMA-2-Chat, and Vicuna give unsafe responses for LOTE queries [3]

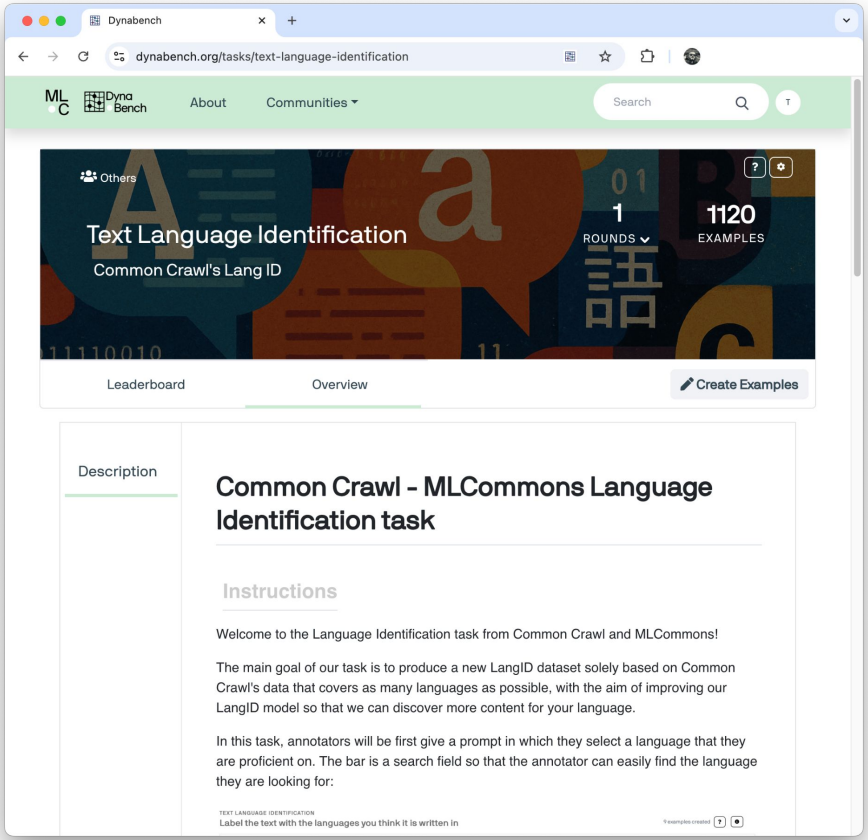
COMMON CRAWL

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
[2] https://arxiv.org/abs/2405.09373 [3] https://arxiv.org/abs/2310.00905

COMMON CRAWL

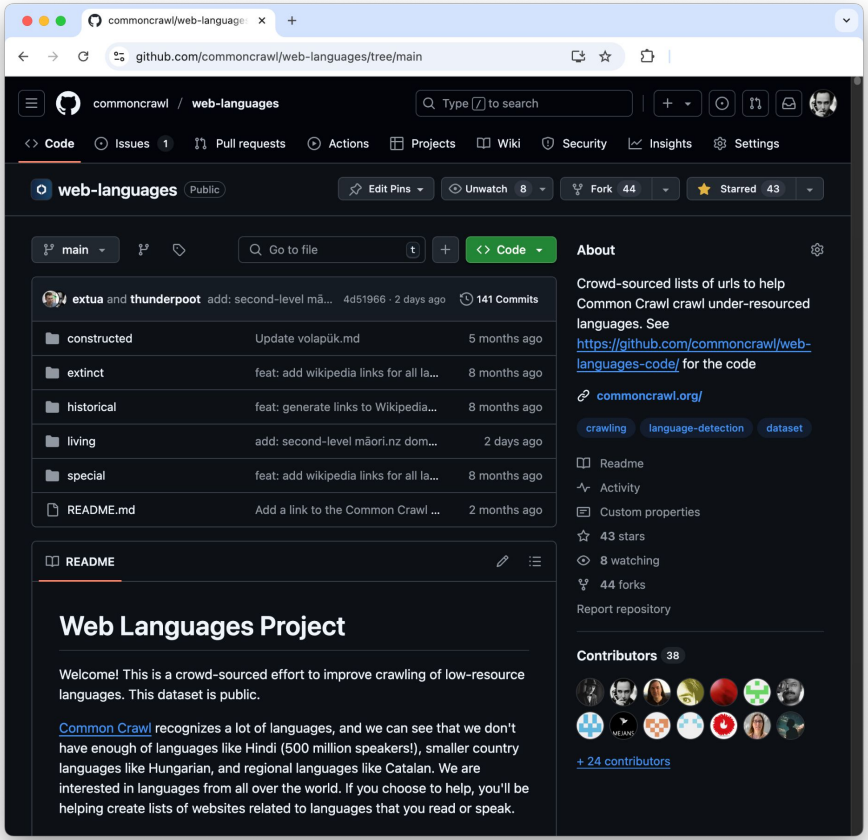
“Hey, buy this car, but the brakes and seat belts only work in English-speaking locales”

Nobody, ever



Language Annotations

COMMON CRAWL



Web Languages

COMMON CRAWL

COLM Workshop

COMMON CRAWL

Open Innovations

COMMON CRAWL

The definition of Open Source

- The Open Source Definition (OSD) from the OSI sets a precedent for clear, permissive licensing. [1]
- Open Data for LLMs needs similar standards to ensure free use, redistribution, and modification of datasets.
- Most datasets used in AI training lack OSD-style guarantees. [2]
- The OSAID builds on OSD principles to include data, models, and training code, explicitly recognising open data as essential to building reproducible and transparent AI systems. [3]
- Without open data, models aren't truly open. Source code is required for open software, so access to open data is necessary for models to be meaningfully open.

COMMON CRAWL

[1] https://opensource.org/osd
[2] https://www.media.mit.edu/publications/a-large-scale-audit-of-dataset-licensing-and-attribution-in-ai/
[3] https://opensource.org/ai/open-source-ai-definition

Collaborators' Work in Signals

- Creative Commons proposes a framework to help content stewards express how they want their works used in AI training: CC Signals [1]
- Groups at the IETF (AIPREF [2] and WEBBOTAUTH [3]) are also working towards formal standards for how websites signal preferences about AI-driven access, and also how automated agents identify and behave accordingly.
- Open Future proposes [4] a move beyond the usual copyright-infringement debates, to think of generative AI as a cultural and social technology that is reshaping how societies access, produce, and value information.

COMMON CRAWL

[1] https://creativecommons.org/ai-and-the-commons/cc-signals/
[2] https://datatracker.ietf.org/wg/aipref/about/ [3] https://datatracker.ietf.org/wg/webbotauth/about/
[4] https://openfuture.eu/publication/beyond-ai-and-copyright/

Conclusions

COMMON CRAWL

Conclusions

- Openness is more than access.
- Web-scale datasets deserve web-scale governance.
- We’re building trust at the infrastructure layer.
- Openness must include everyone.

COMMON CRAWL

Thank you!

You can access these slides with the QR code above.
Please feel free to join us on Discord or in our Google Group
https://discord.gg/njaVFh7avF
https://groups.google.com/g/common-crawl

COMMON CRAWL