Introduction to Common Crawl

AI Alliance @ IBM One Madison
UN Open Source Week 2025

COMMON CRAWL

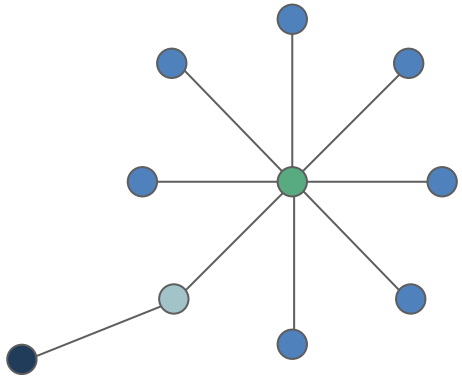
What is Common Crawl?

COMMON CRAWL

What is Common Crawl?

- Free and open corpus containing > 275 billion pages over > 18 years
- 501(c)(3) nonprofit, started in 2007 by Gil Elbaz
- Hosted on AWS S3 as an Open Data set, enabled by the Open Data Sponsorship Program
- Cited in > 10,000 research papers
- 3–4 billion new pages added each month
- Steered by link-based Harmonic Centrality ranks
- Web Graphs showing the structure and connectivity of the web
- We also release the accompanying ranking data for hosts and domains

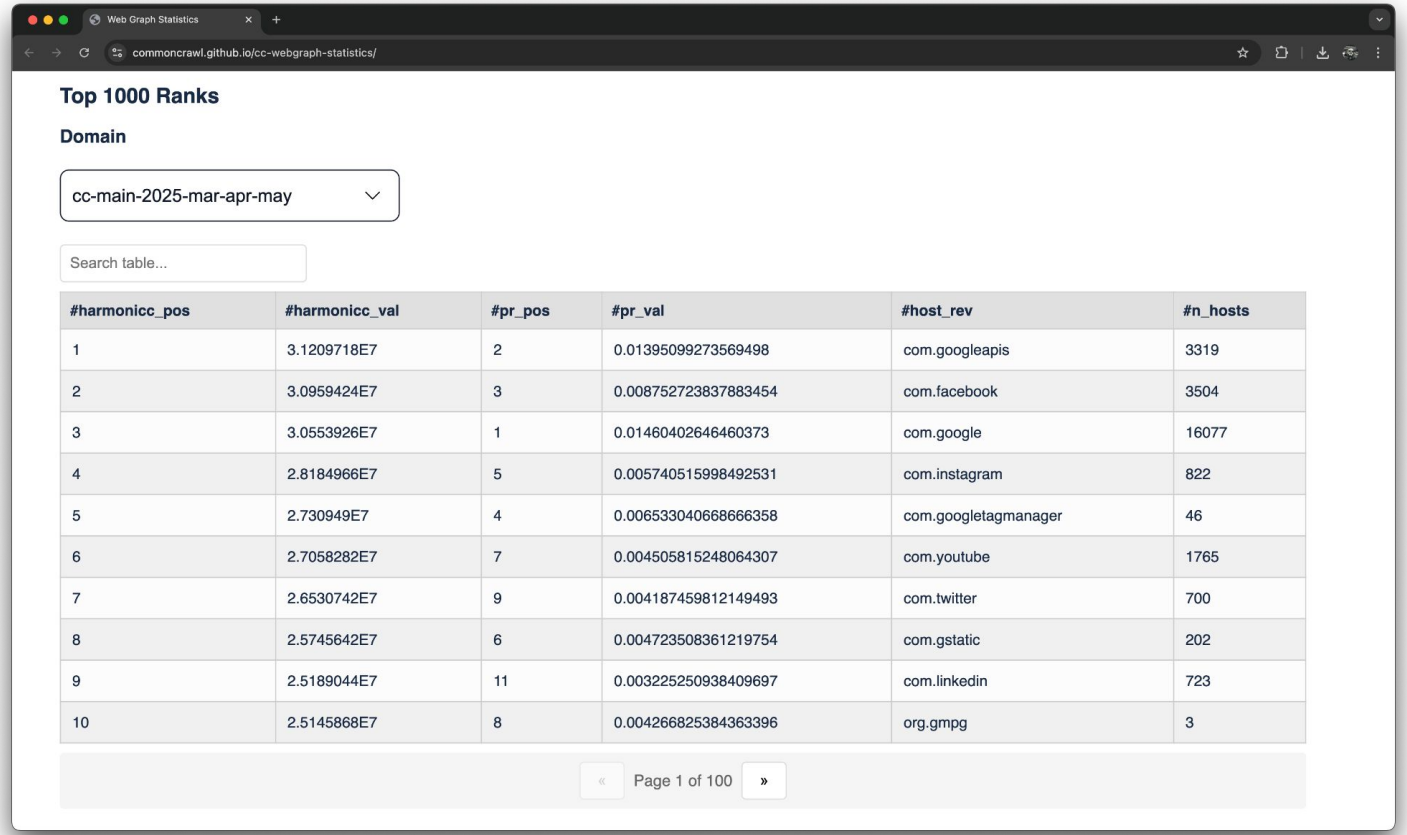
COMMON CRAWL



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

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



Top 1000 Ranks

Domain

cc-main-2025-mar-apr-may

Search table...

#harmonicc_pos	#harmonicc_val	#pr_pos	#pr_val	#host_rev	#n_hosts
1	3.1209718E7	2	0.01395099273569498	com.googleapis	3319
2	3.0959424E7	3	0.008752723837883454	com.facebook	3504
3	3.0553926E7	1	0.01460402646460373	com.google	16077
4	2.8184966E7	5	0.005740515998492531	com.instagram	822
5	2.730949E7	4	0.006533040668666358	com.googletagmanager	46
6	2.7058282E7	7	0.004505815248064307	com.youtube	1765
7	2.6530742E7	9	0.004187459812149493	com.twitter	700
8	2.5745642E7	6	0.004723508361219754	com.gstatic	202
9	2.5189044E7	11	0.003225250938409697	com.linkedin	723
10	2.5145868E7	8	0.004266825384363396	org.gmpg	3

« Page 1 of 100 »

COMMON CRAWL

COMMON CRAWL

9.5 PiB

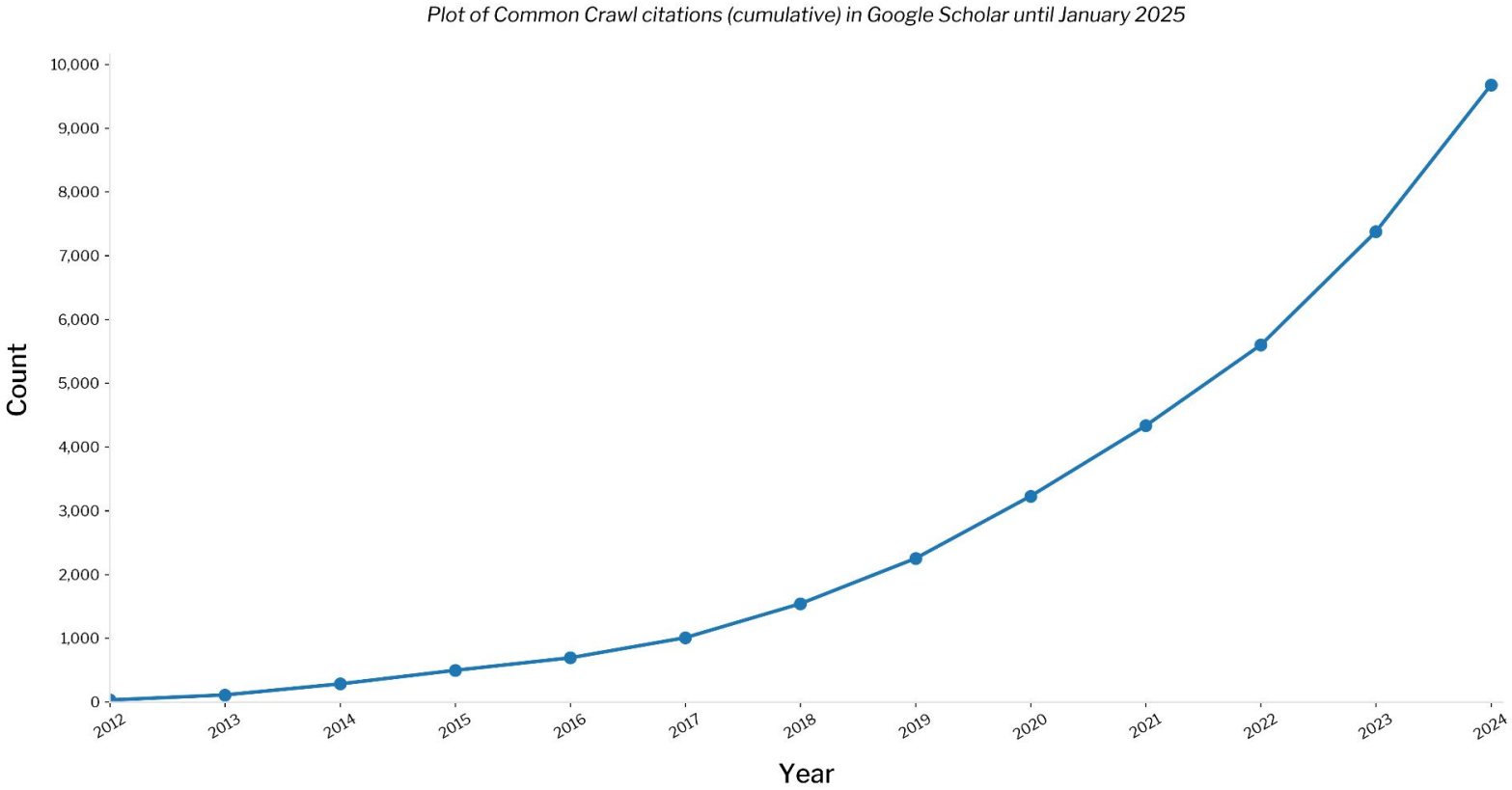
Dataset size as of May 2025.

About 3.45 billion copies of “War and Peace”.
9.5 PiB ≈ 10.696 PB.

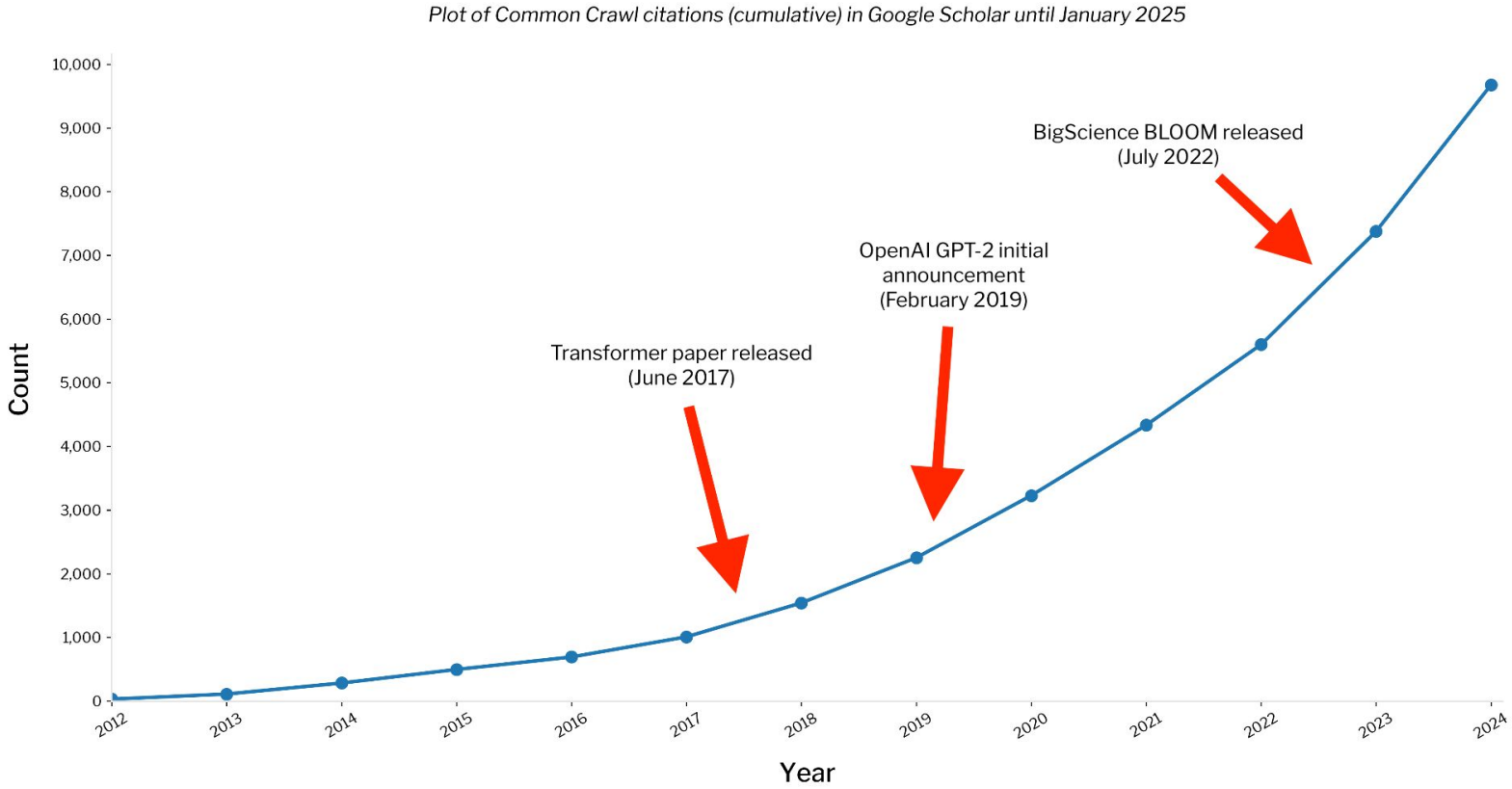
What's it used for?

- Natural language processing
- Web science
- Information retrieval
- Semantic web
- Security research
- Language modelling …

COMMON CRAWL

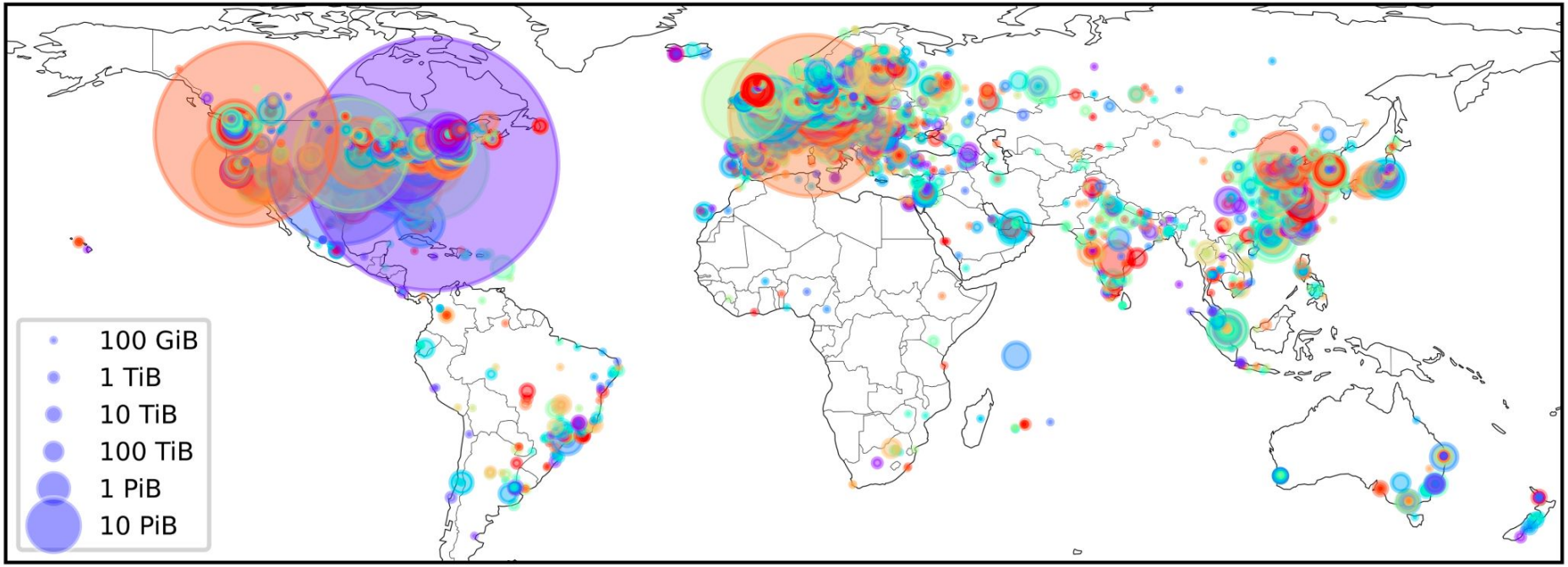


https://commoncrawl.org/research-papers
https://huggingface.co/datasets/commoncrawl/citations



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

Data Requested by GeoIP



COMMON CRAWL

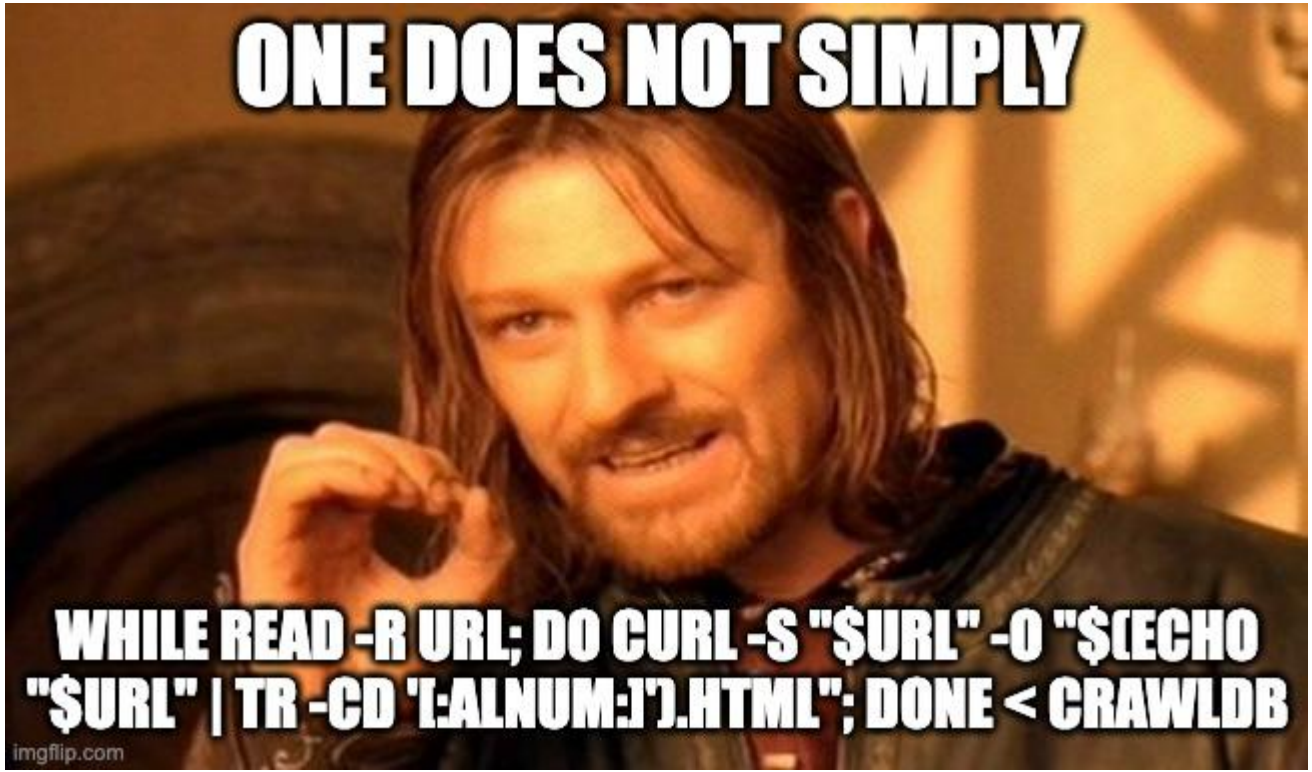
>1 EiB

Total egress in 2024.

One quintillion bytes, or roughly 372 billion copies of “War and Peace”.

Crawler Perspective

COMMON CRAWL


ONE DOES NOT SIMPLY
WHILE READ -R URL; DO CURL -S "$URL" -O "$(ECHO "$URL" | TR -CD '[:ALNUM:]').HTML"; DONE < CRAWLDB
imgflip.com

Crawler Perspective

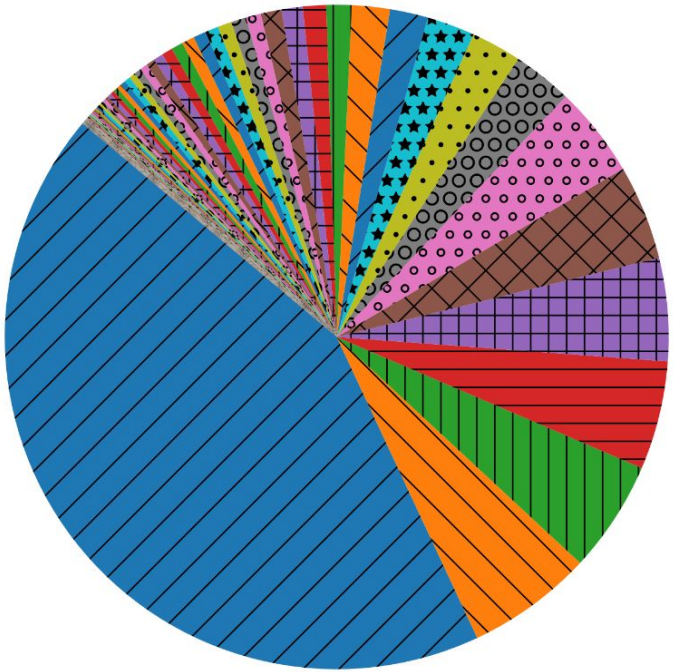
- Politeness is crucial to long-term sustainability, ignoring consent leads to backlash [1]
- Be 👏 transparent 👏 to 👏 build 👏 trust
- Preference signals must be machine-readable [2]
- Purposes differ widely in the “crawl space” [3]
- Opt-Out vs. Opt-In [4]

COMMON CRAWL

[1] https://arxiv.org/abs/2407.14933 [2] https://www.w3.org/TR/tracking-dnt/
[3] https://www.cloudflare.com/en-gb/learning/bots/what-is-a-web-crawler
[4] https://bit.ly/OpenFuture_Rightsholder_OptOuts

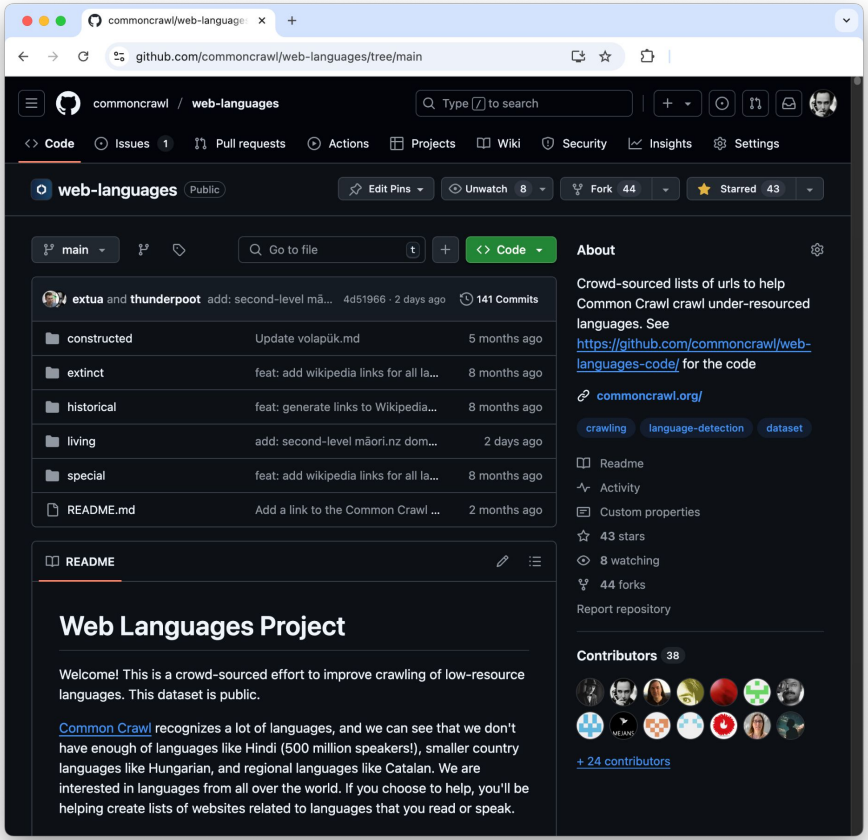
Language Initiatives

COMMON CRAWL



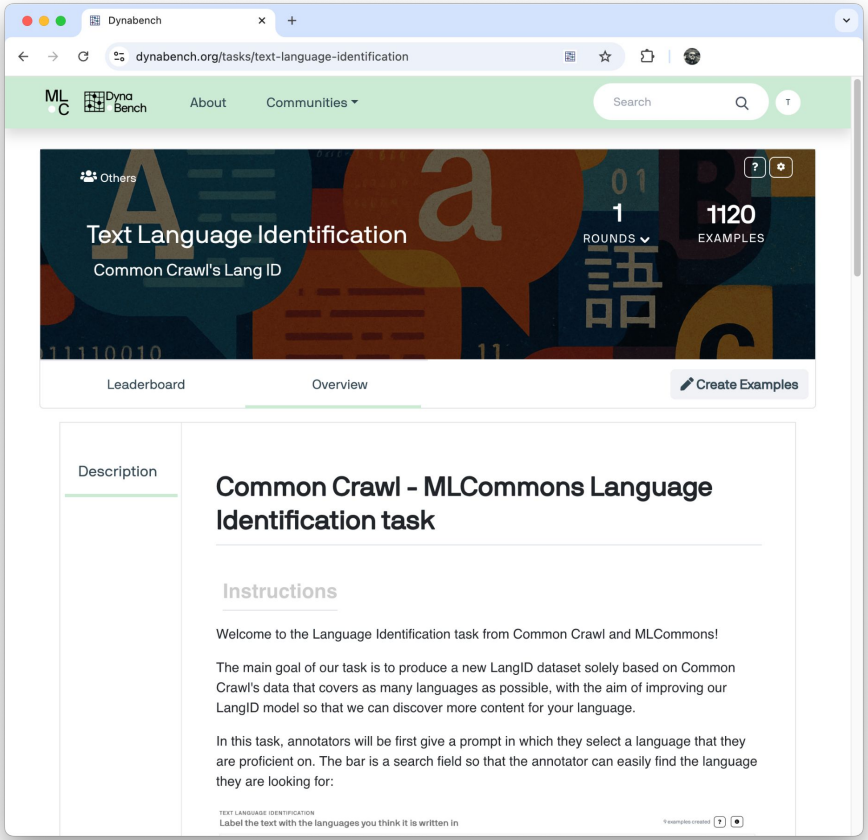
Detected language distribution (averaged) in the last three crawls using CLD2 as the language identifier (**CC-MAIN-2024-46**, **CC-MAIN-2024-51**, and **CC-MAIN-2025-05**)

https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
https://github.com/CLD2Owners/cld2



Web Languages

COMMON CRAWL



Language Annotations

COMMON CRAWL

Workshop

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

Ethical & Trustworthy AI Data Pipelines

COMMON CRAWL

Ethical & Trustworthy AI Data Pipelines

- Crawler opt-outs are rising, so researchers and model builders must rethink how they gather data, and how that affects model safety and quality
- Sanitised or ambiguous opt-outs dilute user intent. AIPREF [1] aims to preserve real author preferences
- We need open, accountable systems that show *how* and *which* web content is used in model training
- We need a common vocabulary for expressing AI-related content preferences
- We need protocol-level attachments (e.g., metadata, HTTP headers, etc.) to signal those preferences
- We need a standard for resolving conflicting signals across sources
- We need all this in order to rebuild ***trust***.

COMMON CRAWL

[1] https://datatracker.ietf.org/wg/aipref/about/

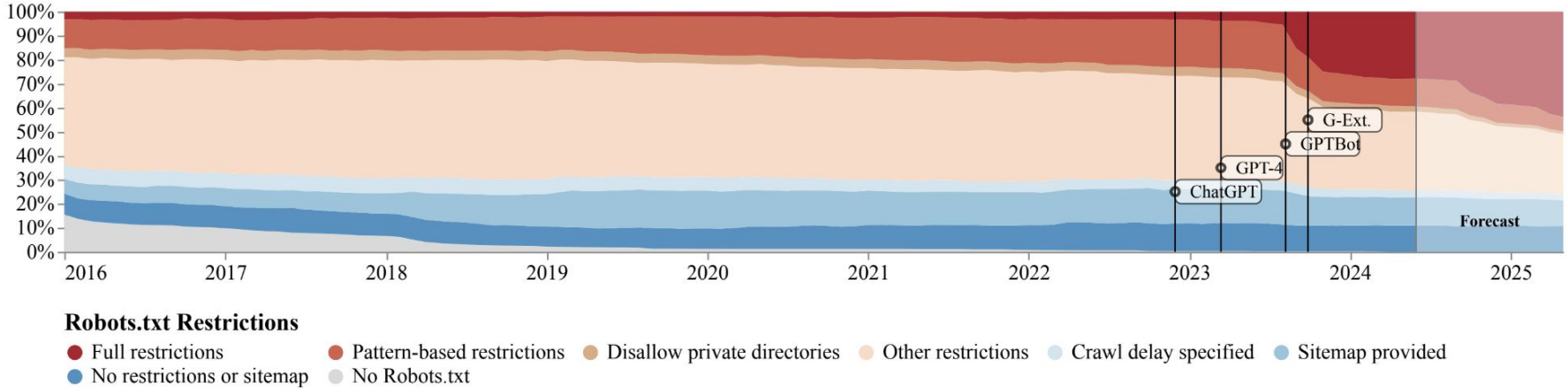
COMMON CRAWL

“If respected or enforced, these restrictions are rapidly biasing the diversity, freshness, and scaling laws for general-purpose AI systems. [...] The foreclosure of much of the open web will impact not only commercial AI, but also non-commercial AI and academic research.”

Consent in Crisis: The Rapid Decline of the AI Data Commons

Shayne Longpre et al., 2024

https://arxiv.org/abs/2407.14933



Consent in Crisis: The Rapid Decline of
the AI Data Commons

Shayne Longpre et al., 2024

https://arxiv.org/abs/2407.14933

COMMON CRAWL

“The only thing you absolutely have to know is the location of the library.”

Albert Einstein

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


I AM AWAKE

Thom Vaughan (Moderator)
Principal Technologist, Common Crawl

Lilith Bat-Leah
DMLR Working Group Co-chair, MLCommons

Dean Wampler
IBM Head of Technology, AI Alliance

Jose Plehn-Dujowich
CEO, BrightQuery

Greg Lindahl
CTO, Common Crawl

Dave Buckley
Senior Policy Manager, OpenMined

Roberto di Cosmo
Director, Software Heritage

- Ethics of large-scale data collection
- Preserving authenticity in user preference signals
- Governance and transparency in AI training data usage
- Collaborative standards across the AI ecosystem
- Building trust in public data pipelines