Common Crawl: open web data for everybody

Laurie Burchell

laurie@commoncrawl.org

COMMON CRAWL

Overview

- What is the Common Crawl Foundation?
- What is the data like?
- A representative sample?
- Ongoing research: language diversity
- Next steps

slides

What is the Common Crawl Foundation?

COMMON CRAWL

What is the Common Crawl Foundation?

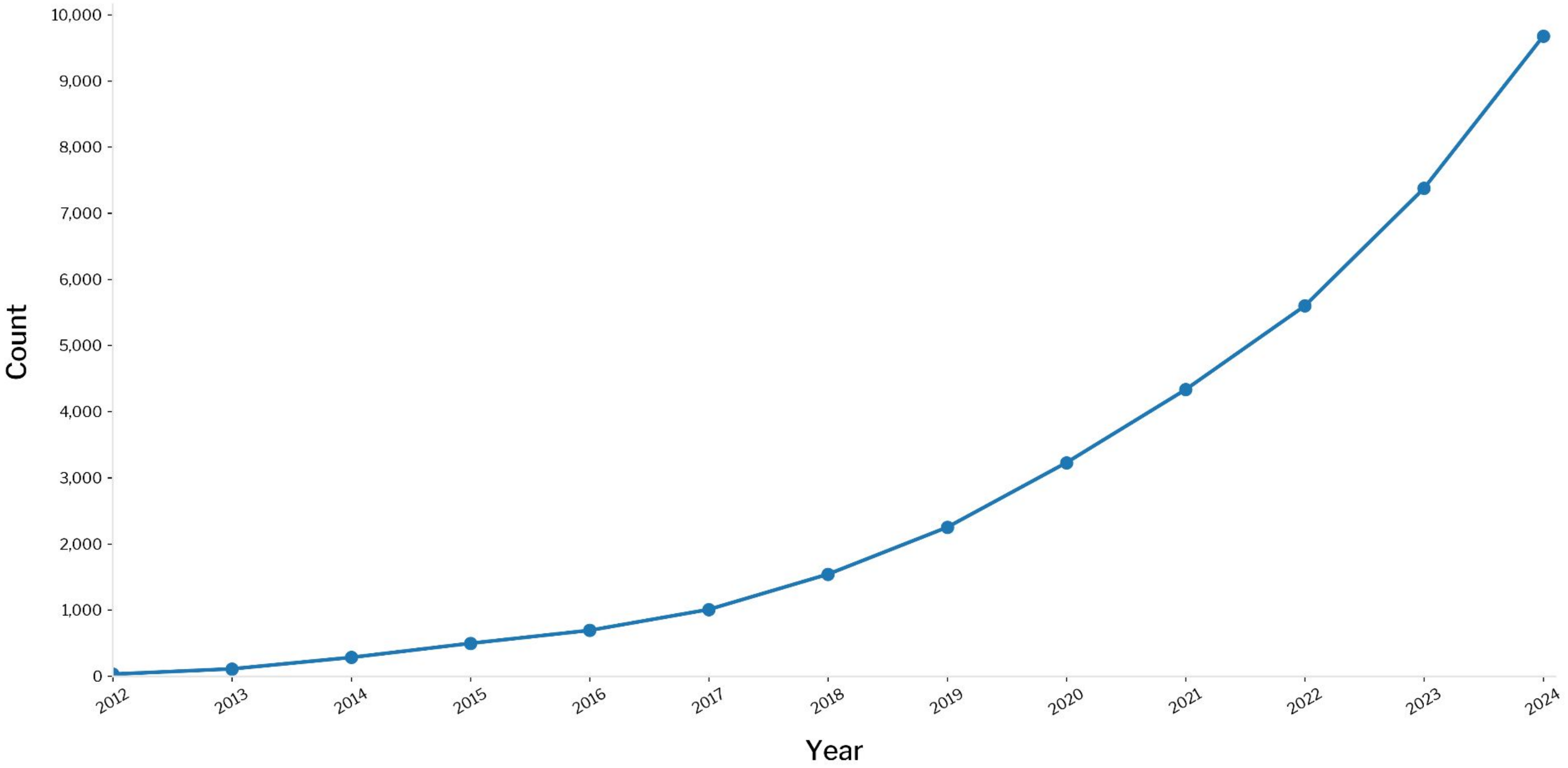
- Non-profit founded in 2007 by Gil Elbaz
- Mission: **make web data accessible** to researchers, developers
- Small (but growing!) team of staff and volunteers
- Part of AWS Open Datasets Program: free to access!

11.3 PB

…and growing by >4 TB each month

Dataset size as of September 2025.

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



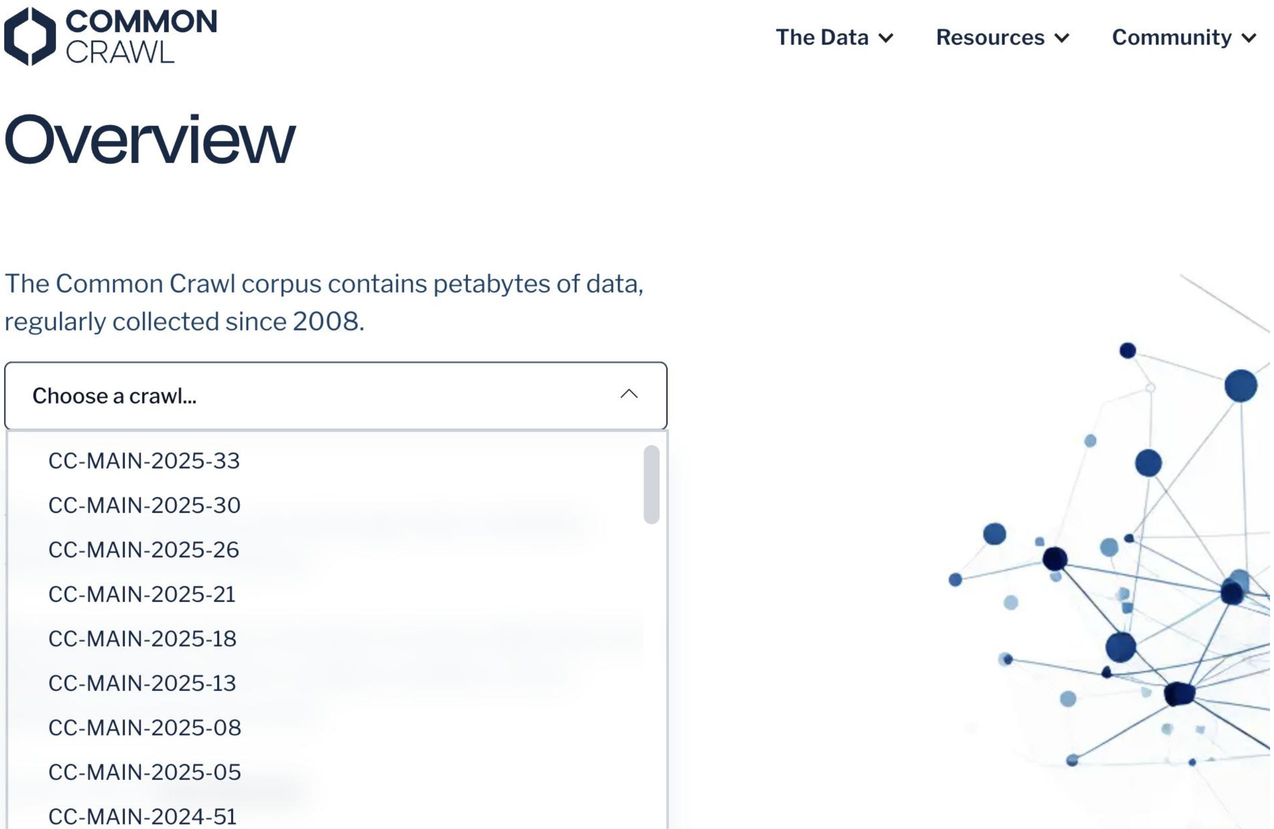
https://commoncrawl.org/research-papers

https://huggingface.co/datasets/commoncrawl/citations

Projects with Common Crawl

- Creating a large-scale, multilingual corpus
- Analysing disappearing links over time
- Detecting misinformation sources
- Censorship of Amazon products
- COVID-19 news mood map

The data



Overview

The Common Crawl corpus contains petabytes of data, regularly collected since 2008.

Choose a crawl...

- CC-MAIN-2025-33
- CC-MAIN-2025-30
- CC-MAIN-2025-26
- CC-MAIN-2025-21
- CC-MAIN-2025-18
- CC-MAIN-2025-13
- CC-MAIN-2025-08
- CC-MAIN-2025-05
- CC-MAIN-2024-51

Access our data via HTTP(S) or AWS:

see commoncrawl.org/get-started

(the most important) Data products

- **WARCs**: web page captures
- CDXJ and columnar **indices**
- **Web Graph**: structure and connectivity

WARCs: web page captures

- Web ARChive format
- Raw crawl data
  - Content payload
  - HTTP headers
  - Connection metadata (datetime, IP address)
- Derived data: WETs (text), WATs (metadata)
- Stats at [cc-crawl-statistics]()

CDXJ and columnar indices

- Different indices for different kinds of access
- CDXJ index: single URLs or domains
- Columnar index (Parquet)
  - SQL queries
  - Big data toolkits

COMMON
CRAWL

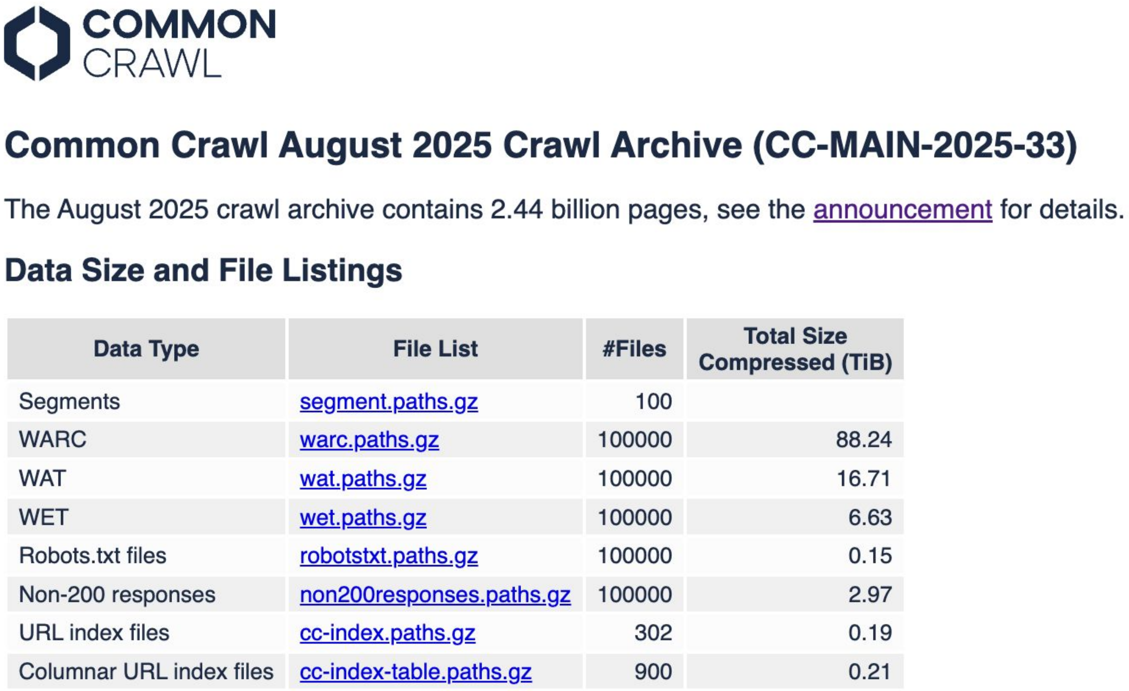
Web Graph

- Structure and connectivity of the web
- Two levels: host and domain
- Tools at cc-webgraph
- Stats at cc-webgraph-statistics

There's more!

See our website:
commoncrawl.org

Tools and examples:
github.com/commoncrawl



COMMON CRAWL

Common Crawl August 2025 Crawl Archive (CC-MAIN-2025-33)

The August 2025 crawl archive contains 2.44 billion pages, see the announcement for details.

Data Size and File Listings

Data Type	File List	#Files	Total Size Compressed (TiB)
Segments	segment.paths.gz	100	
WARC	warc.paths.gz	100000	88.24
WAT	wat.paths.gz	100000	16.71
WET	wet.paths.gz	100000	6.63
Robots.txt files	robotstxt.paths.gz	100000	0.15
Non-200 responses	non200responses.paths.gz	100000	2.97
URL index files	cc-index.paths.gz	302	0.19
Columnar URL index files	cc-index-table.paths.gz	900	0.21

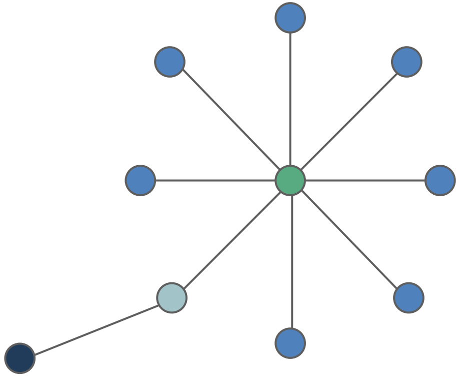
A representative sample of the web?

How does crawling work?

- Start from seeds and spider out
- We crawl politely!
  - Slowly
  - Respecting robots.txt and opt out
  - No log ins
- NB: our crawl is **text-only**

Why sample?

- TL;DR: **the web is big, our resources are finite**
- Per monthly crawl:
  - 2.5 billion page captures
  - 500+ billion links
  - 20+ billion unique URLs linked (excluding media links)
- Also: avoiding overload, spam traps!



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

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

Sample by budgeting domains,

guided by **harmonic centrality** rank

COMMON
CRAWL

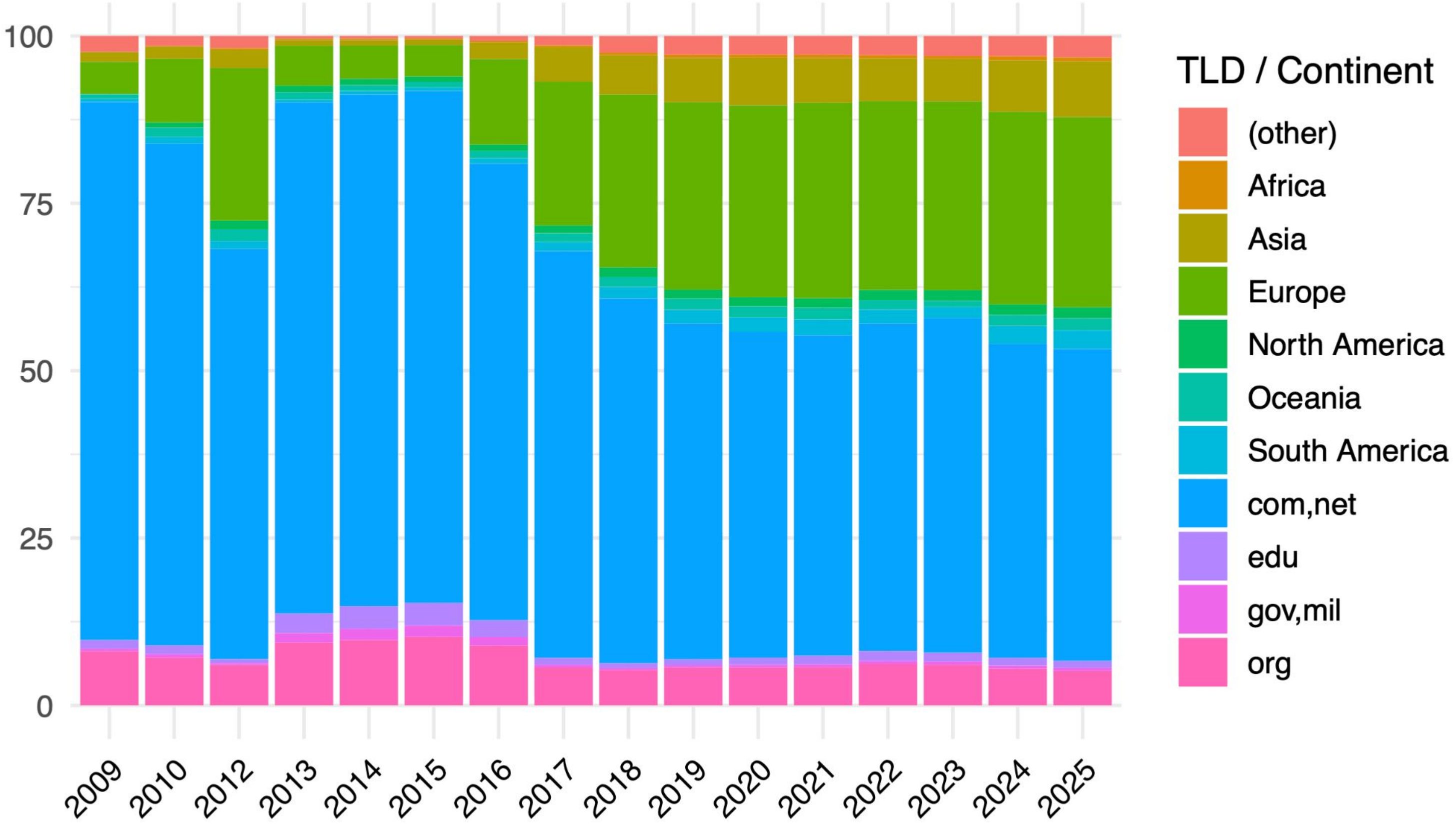
What is ‘representative’?

- Breadth versus depth
- Freshness (new content)
- Amount of (near-)duplicates (per crawl and over multiple crawls)
- Regional coverage (top-level domains, content languages)
- Content ‘quality’

What's the impact?

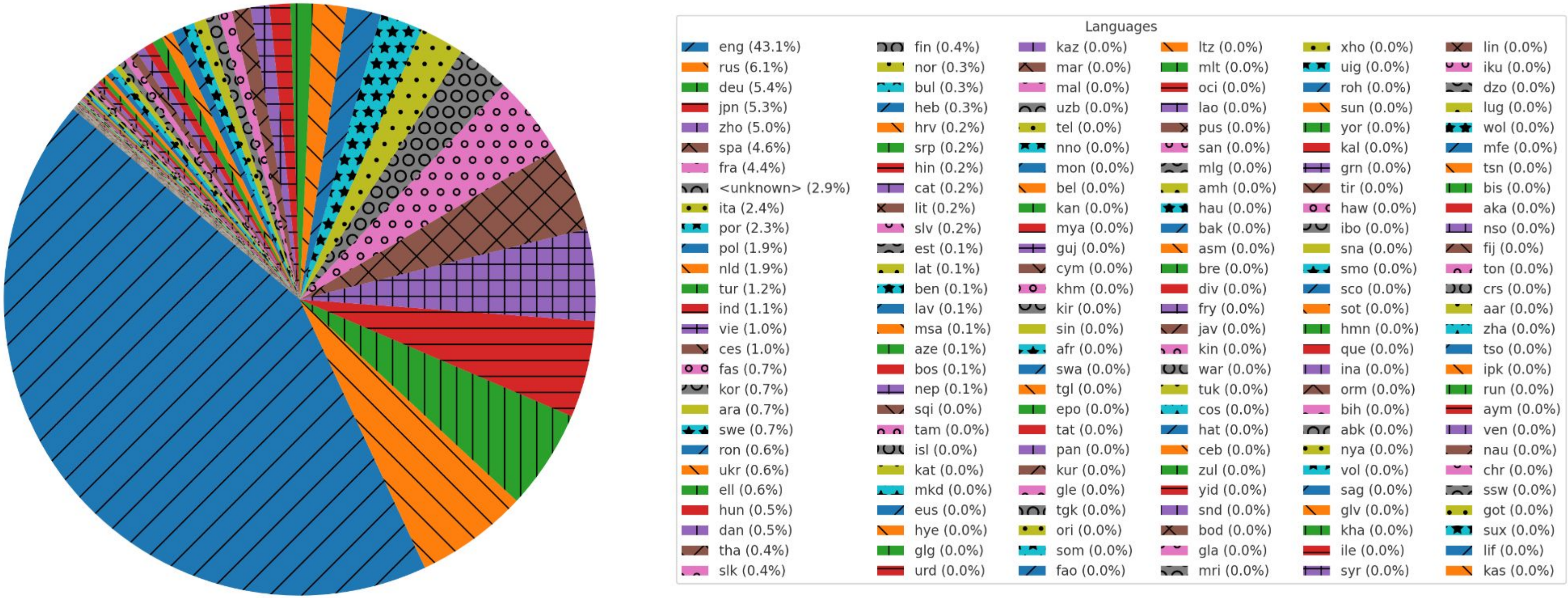
- ‘Representative’ depends on use case!
- We aim for compromise:
  - Breadth versus depth
  - Site categories, topics, languages, geographic regions…
- Focus on USA and Western world (for now…)

Top-Level Domains and Geographical Coverage



Ongoing research: language diversity

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**)

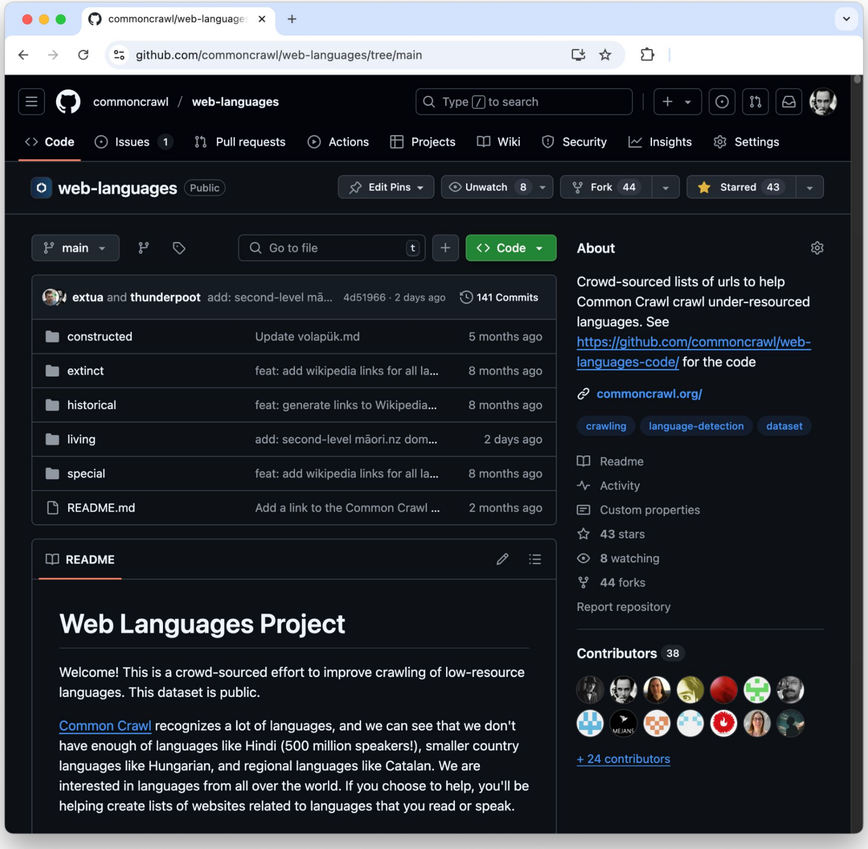
https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
https://github.com/CLD2Owners/cld2

Increasing language coverage

Two strategies:

1. More **diverse seeds**
2. Better **language detection**

Web Languages



constructed	Update volapük.md	5 months ago
extinct	feat: add wikipedia links for all la...	8 months ago
historical	feat: generate links to Wikipedia...	8 months ago
living	add: second-level māori.nz dom...	2 days ago
special	feat: add wikipedia links for all la...	8 months ago
README.md	Add a link to the Common Crawl ...	2 months ago

README

Web Languages Project

Welcome! This is a crowd-sourced effort to improve crawling of low-resource languages. This dataset is public.

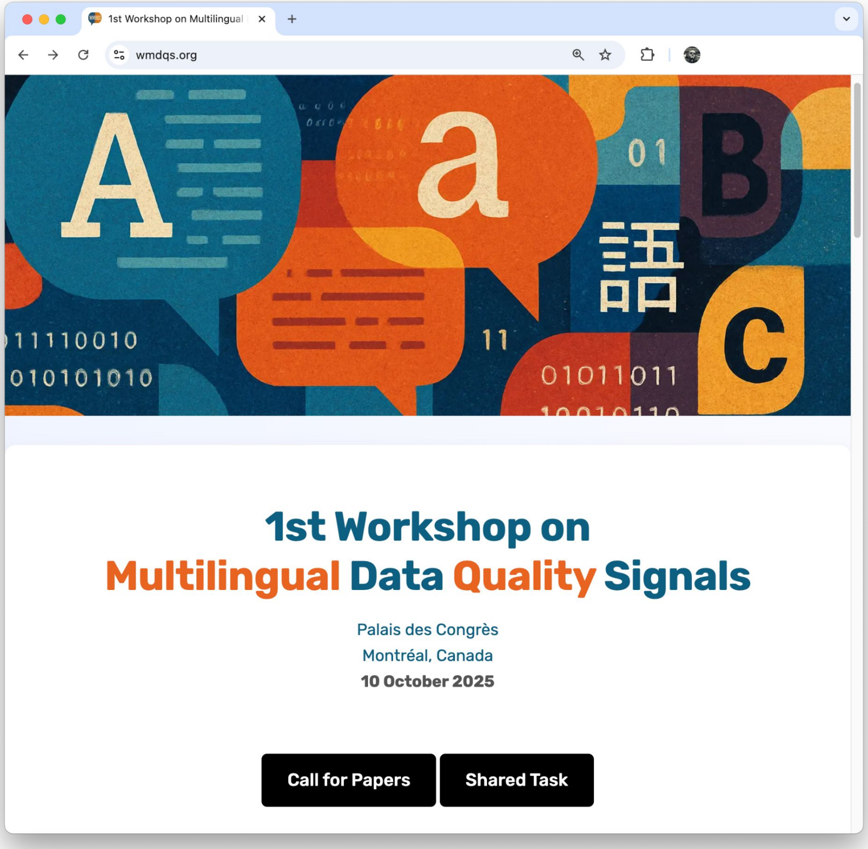
Common Crawl recognizes a lot of languages, and we can see that we don't have enough of languages like Hindi (500 million speakers!), smaller country languages like Hungarian, and regional languages like Catalan. We are interested in languages from all over the world. If you choose to help, you'll be helping create lists of websites related to languages that you read or speak.

About

Crowd-sourced lists of urls to help Common Crawl crawl under-resourced languages. See https://github.com/commoncrawl/web-languages-code/ for the code

commoncrawl.org/

crawling language-detection dataset



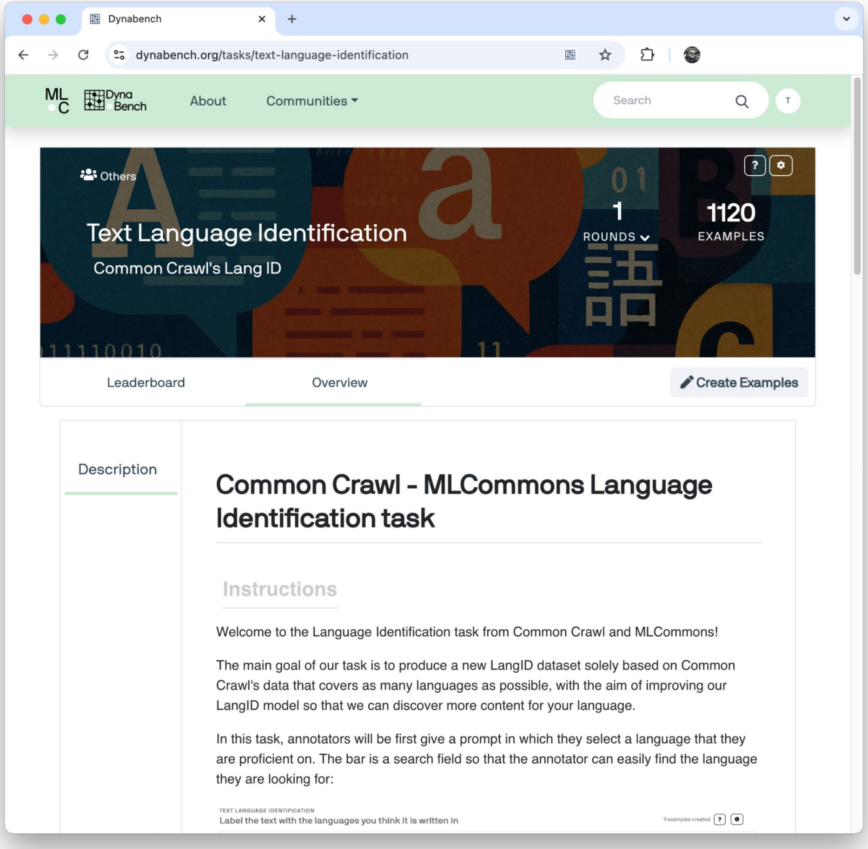
1st Workshop on Multilingual Data Quality Signals

Palais des Congrès
Montréal, Canada
10 October 2025

Research workshop: WMDQS

COMMON CRAWL

Language Identification Shared Task



Text Language Identification

Common Crawl's Lang ID

Common Crawl - MLCommons Language Identification task

Instructions

Welcome to the Language Identification task from Common Crawl and MLCommons!

The main goal of our task is to produce a new LangID dataset solely based on Common Crawl's data that covers as many languages as possible, with the aim of improving our LangID model so that we can discover more content for your language.

In this task, annotators will be first give a prompt in which they select a language that they are proficient on. The bar is a search field so that the annotator can easily find the language they are looking for:

TEXT LANGUAGE IDENTIFICATION

Label the text with the languages you think it is written in

9 examples created

Next steps with Common Crawl

COMMON CRAWL

Getting started

- Visit our website: Get Started guide
- Whirlwind Tour in Python
- Join our community: Discord, contact us!

COMMON
CRAWL

COMMON
CRAWL

Thank you!

Questions?

Laurie Burchell
laurie@commoncrawl.org