COMMON CRAWL

Common Crawl and Languages on the Web

Sebastian Nagel sebastian@commoncrawl.org

1st Workshop on Multilingual Data Quality Signals
Conference on Language Modeling (COLM)
Palais des Congrès, Montréal, Canada
10 October 2025

Common Crawl – A Brief Introduction

- About Common Crawl
- Papers Citing Common Crawl
- Papers Citing Common Crawl
- Data Overview
- Data Formats
- Data Formats – Indexes and Metadata
- Data Downloads

A Relevant Sample of the Public Web

Data Collection and Sampling – History and Metrics

Language Identification in a Web crawler

Summary and Outlook

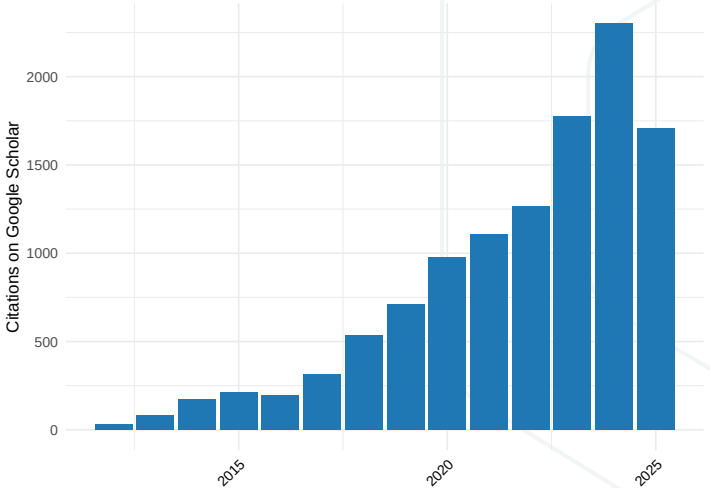
About Common Crawl

- We're a non-profit that makes web data accessible to programmers and data scientists
- Started in 2007 by Gil Elbaz
- Hosted as Open Data set on Amazon Web Services [1, 2]
- Natural language processing, language modeling, web science, information retrieval, semantic web, internet security research, …
- Used for training language models since 2013 (WMT'13 [3], GloVe [4])

Papers Citing Common Crawl

https://github.com/commoncrawl/cc-citations

Papers Citing Common Crawl



Data Overview

- Over 300 billion web pages spanning 17 years (2008 – 2025)
- 2.5 billion pages added each month
- More than 100 crawl archives released to date
- 10.1 PiB of data (Sept 2025)

Data Formats

WARC – web page captures

- Content payload
- HTTP headers
- Connection metadata (datetime, IP address)

ARC – predecessor of WARC

- Used to store data until 2012 (conversion to WARC in preparation)

WAT – HTML metadata and links

WET – plain text extracted from HTML

- No removal of boilerplate content
- No markup preserved
- 10% of the WARC size

Data Formats – Indexes and Metadata

Index – URL, metadata and WARC record location

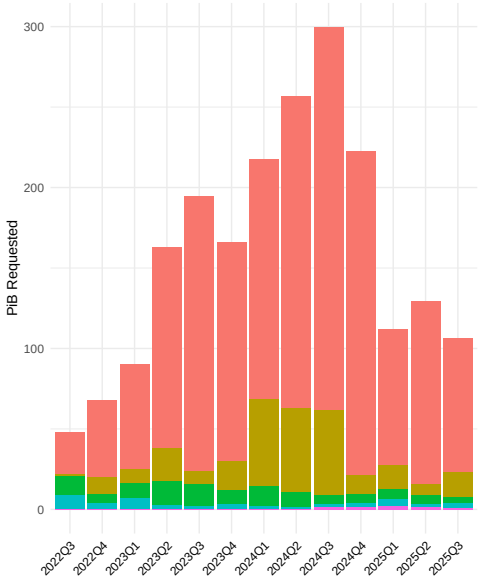
- CDX format – powers our wayback machine (index.commoncrawl.org)
  - Optimized to look up single URLs or domains
- Columnar format (Parquet) [5, 6]
  - SQL queries and aggregations
  - Cheap and scalable, bulk lookup for millions of URLs
  - Using big data tools (Spark, Hive, Presto, Trino, Athena) or DuckDB

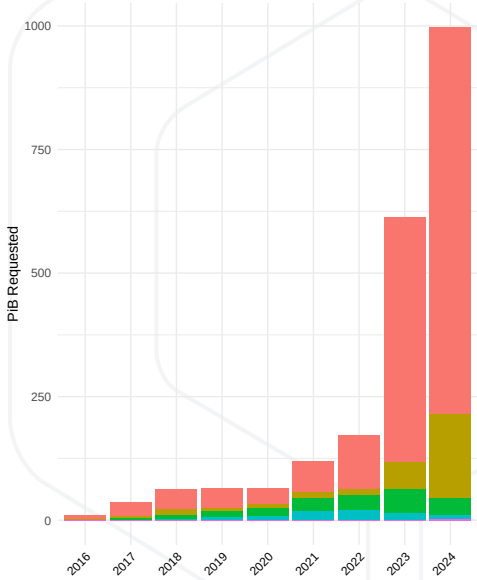
Webgraph and ranks

- Aggregated on host and domain level
- commoncrawl.github.io/cc-webgraph-statistics

Crawl metrics (commoncrawl.github.io/cc-crawl-statistics)

Data Downloads





Sampling Web Pages – Targets and Objectives

- A *representative* sample of the World Wide Web,
  or the crawlable subset (robots.txt) of the public web
- Covering various site categories, topics, languages, geographic regions, …
- A compromise between breadth and in-depth coverage of individual sites
- Balancing freshness – new pages and revisits
- With a focus on text – HTML, PDFs and other textual formats,
  avoid images, video, executables, …

The Need for Sampling

Why sampling and prioritization are necessary? Why not just follow links?

- An average “monthly” crawl includes 2.5 billion page captures with
  - 500+ billion links
  - 20+ billion unique URLs linked (excluding media links)
- Up to 2.5 billion URLs listed in a single sitemap (sitemap index) [7]

Need to sample given

- Limited resources
- Requirements for crawler politeness: do not overload a single web site
- It’s easy to get lost in the wrong corner of the web!

Sampling By “Budgeting Domains”

Our algorithm, which we call Spam Tracking and Avoidance through Reputation (STAR), dynamically allocates the budget of allowable pages for each domain and all of its subdomains in proportion to the number of in-degree links from other domains.

[...] we found that spam could be “deterred” by budgeting the number of allowed pages per PLD [pay-level domain] based on domain reputation

(“IRLbot: Scaling to 6 Billion Pages and Beyond” [8])

Stratified Domain-level Sampling

- Domain-level ranks define a “budget” per pay-level domain
- Limiting the max. number of sampled URLs/pages and subdomains
- We use “harmonic centrality” [9, 10, 11] ranks derived from hyperlink graphs aggregated on the domain-level
  - Domain: one level below the registry suffix, e.g.:

`commoncrawl.org`
`data.gov.uk`
`mastodon.au`
`nsw.gov.au`
`betterhealth.vic.gov.au`

(`au`, `gov.au` and `vic.gov.au` are ICANN suffixes in the Public Suffix List [12])

Domain-Level Graph-Based Ranking Example

pos	hc	pr	rev. domain
1	71	297	edu.stanford
2	78	285	edu.harvard
3	90	392	edu.mit
4	135	588	edu.berkeley
5	157	757	edu.psu
6	167	515	edu.cornell
7	203	522	edu.cmu
8	213	978	edu.princeton
9	228	998	edu.utexas
10	236	818	edu.columbia
11	239	1011	edu.yale
12	249	1063	edu.wisc
13	268	1050	edu.washington
14	292	1358	edu.brookings*
15	300	1405	edu.usc
16	349	2076	edu.ncsu
17	352	1243	edu.si*
18	391	1824	edu.georgetown
19	397	1248	edu.academia*
20	398	1010	edu.uchicago

rank	QS World [13]
1	MIT
4	Harvard
6	Stanford
10	Caltech
11	U. Pennsylvania
12	Berkeley (UCB)
16	Cornell
21	Chicago
22	Princeton
23	Yale
32	Johns Hopkins
34	Columbia
42	UCLA
43	NYU
44	Michigan-Ann Arbor
50	Northwestern
58	Carnegie Mellon
61	Duke
66	Texas at Austin
69	Illinois

rank	Forbes [14]
1	Princeton
2	Stanford
3	MIT
4	Yale
5	Berkeley
6	Columbia
7	U. Pennsylvania
8	Harvard
9	Rice
10	Cornell
11	Northwestern
12	Johns Hopkins
13	UCLA
14	Chicago
15	Vanderbilt
16	Dartmouth College
17	Williams College
18	Brown
19	Claremont McKenna
20	Duke

Domain-Level Ranking Example (Explanations)

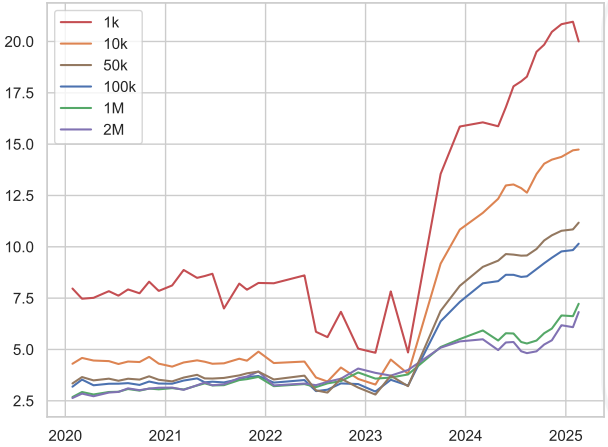
- Top-N `.edu` domains ranked by harmonic centrality (or pagerank) calculated on CCF's domain-level hyperlink graphs [15]
- Reverse domain name notation [16]
- Order by harmonic centrality ("hc") [9, 11]
  - Ranks are shown, not scores
  - PageRank rank [17], for comparison
  - Global ranks – domains below all top-level domains, not only `.edu`
- `.edu` TLD includes not only universities (*)
- Compared with university rankings by QS World [13] and Forbes [14]
- Data from January 2025 (university rankings and CCF web graphs)

Crawler Politeness

- Crawl slowly
  - Further slow down (exponential backoff) if a site responds with errors
- Respect robots.txt rules (Robots Exclusion Protocol – [18]) and
- URI-level metatags (`<meta name=robots value=nofollow>`) [19, 20]

- CCBot identifies itself
  - User-agent string and contact information sent along with requests
  - Crawling from fixed list of IP addresses, publicly announced and verifiable via reverse DNS

Impact of Robots.txt on Crawling



- Share of sites blocking CCBot has grown since 2023
- Higher ranking sites: 10% and more
- Long-tail: disallowed share up from 3% to 7%

Summary Data Collection

- Sample crawls of the public web
- Crawling the public web
- A polite crawler, respecting robots.txt,
  even if we miss non-trivial parts of the web
- Steered by link-based harmonic centrality ranks

A Look Back In Time

Four phases of data collection using different

- Crawler implementations
- Approaches to find and sample (prioritize) seeds and URLs
- Page revisit policies

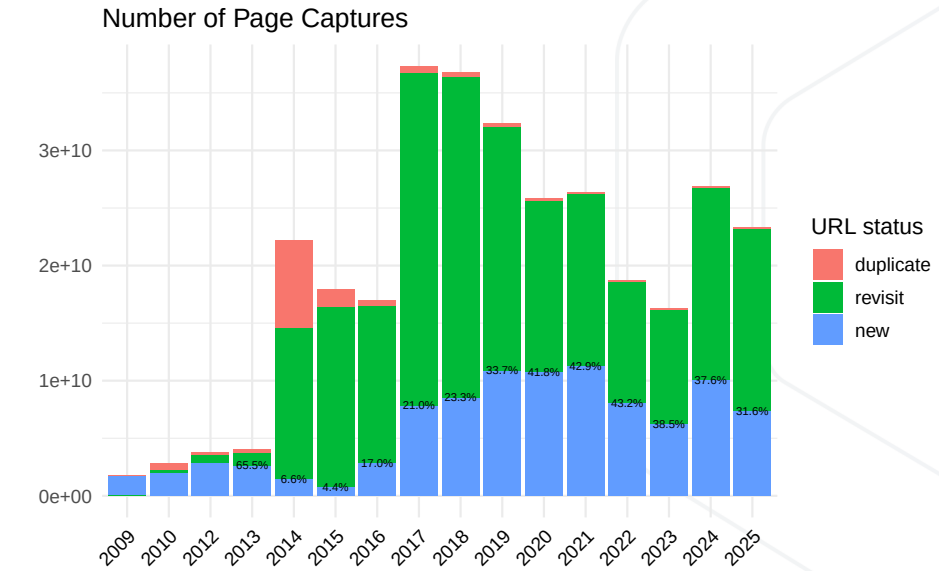
	crawler	seeds / link prioritization	revisit policy
2008–2009	Nutch	list based	
2012	in-house	page rank	
2013 – 2016	Nutch	seed donations, list based	
2017 – now	Nutch	harmonic centrality	30% new, 70% revisits based on page relevance

What Is Representative?

Aspects of representativity:

- Breadth: coverage of unique domains (web sites)
- Depth: per-site coverage
- Freshness (new content)
- Amount of (near-)duplicates (per crawl and over multiple crawls)
- Regional coverage (top-level domains, content languages)
- Content quality
- Applicable for a given data use case?

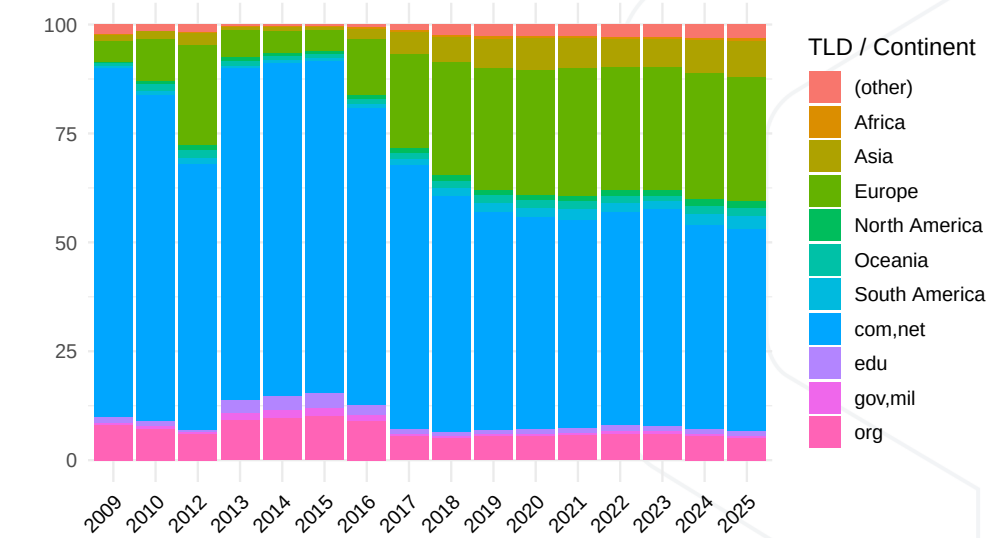
Size and Freshness



Top-Level Domains and Geographical Coverage

year %	com	org	ru	net	de	uk	jp	edu	fr	it	pl	nl	br	au	cz
(all)	52.38	6.42	4.02	3.79	3.26	2.06	1.65	1.45	1.36	1.35	1.23	1.11	0.97	0.82	0.81
2009	71.20	8.10	0.05	9.19	0.05	4.04	0.05	1.28	0.02	0.04	0.19	0.01	0.34	0.60	0.01
2010	68.93	7.14	0.46	6.03	1.51	3.12	0.50	1.31	0.44	0.57	0.48	0.44	0.79	1.02	0.17
2012	55.86	6.02	1.71	5.43	4.75	3.45	1.14	0.61	1.30	1.30	1.79	1.43	0.76	0.98	0.76
2013	73.08	9.40	0.06	3.27	1.12	2.00	0.16	2.91	0.38	0.34	0.02	0.25	0.22	0.63	0.10
2014	73.25	9.75	0.11	3.20	0.81	1.74	0.13	3.36	0.30	0.26	0.16	0.16	0.28	0.51	0.06
2015	73.31	10.25	0.11	3.16	0.76	1.67	0.14	3.34	0.28	0.25	0.14	0.17	0.26	0.51	0.05
2016	64.55	8.96	2.56	3.66	1.97	1.79	0.81	2.48	0.63	0.65	0.49	0.51	0.43	0.60	0.52
2017	56.54	5.64	5.30	4.22	2.83	2.03	2.09	1.02	1.15	1.08	0.99	0.73	0.81	0.70	0.85
2018	50.06	5.28	6.06	4.41	3.49	2.22	2.16	0.75	1.42	1.28	1.37	1.02	0.99	0.78	0.93
2019	46.21	5.68	5.09	3.98	3.88	2.38	1.99	0.89	1.67	1.63	1.51	1.38	1.19	0.92	0.99
2020	44.97	5.68	4.85	3.71	4.03	2.30	1.89	1.08	1.75	1.70	1.52	1.46	1.31	0.98	1.02
2021	44.33	5.71	4.79	3.59	4.15	2.38	1.89	1.28	1.79	1.77	1.59	1.52	1.40	1.05	1.05
2022	45.38	6.16	4.14	3.56	4.21	1.75	1.78	1.49	1.85	1.88	1.66	1.64	1.05	0.80	1.03
2023	46.53	6.06	4.27	3.49	4.30	0.77	1.66	1.31	1.91	2.06	1.64	1.76	0.49	0.41	1.08
2024	43.65	5.51	4.40	3.32	4.03	2.09	2.08	1.14	1.75	1.92	1.75	1.54	1.41	1.05	1.03
2025	43.32	5.13	4.24	3.30	4.15	2.19	2.23	1.06	1.70	1.76	1.66	1.49	1.54	1.16	1.00

Top-Level Domains and Geographical Coverage



Top-Level Domains and Geographical Coverage

% year	(other)	com net	org	edu	gov mil	North America	South America	Oceania	Africa	Asia	Europe
2009	2.37	80.38	8.10	1.28	0.38	0.11	0.49	0.64	0.03	1.48	4.75
2010	1.49	74.96	7.14	1.31	0.51	0.77	0.98	1.41	0.13	1.78	9.53
2012	1.81	61.29	6.02	0.61	0.27	1.31	1.19	1.74	0.23	2.77	22.76
2013	0.40	76.35	9.40	2.91	1.41	1.02	0.42	1.08	0.19	0.85	5.97
2014	0.42	76.45	9.75	3.36	1.69	0.96	0.53	0.87	0.18	0.83	4.96
2015	0.37	76.47	10.25	3.34	1.68	0.84	0.51	0.85	0.17	0.81	4.69
2016	0.74	68.20	8.96	2.48	1.27	0.95	0.81	1.09	0.24	2.48	12.77
2017	1.38	60.76	5.64	1.02	0.42	1.15	1.38	1.33	0.30	5.13	21.49
2018	2.47	54.47	5.28	0.75	0.28	1.40	1.72	1.51	0.40	5.87	25.83
2019	2.80	50.19	5.68	0.89	0.31	1.36	2.04	1.65	0.48	6.62	27.99
2020	2.76	48.68	5.68	1.08	0.35	1.36	2.18	1.66	0.48	7.16	28.61
2021	2.81	47.91	5.71	1.28	0.41	1.45	2.34	1.72	0.50	6.68	29.17
2022	2.88	48.93	6.16	1.49	0.48	1.51	2.08	1.40	0.48	6.37	28.21
2023	2.99	50.02	6.06	1.31	0.47	1.57	1.66	0.94	0.43	6.36	28.21
2024	3.05	46.96	5.51	1.14	0.44	1.59	2.63	1.66	0.60	7.68	28.76
2025	3.20	46.62	5.13	1.06	0.44	1.65	2.79	1.80	0.63	8.31	28.36

Language Coverage

	2018	2019	2020	2021	2022	2023	2024	2025
<other>	6.47	7.31	7.88	7.59	7.51	7.84	7.90	7.91
<unknown>	3.32	2.62	2.35	2.53	2.80	2.84	3.00	2.91
ara	0.76	0.59	0.56	0.60	0.63	0.64	0.66	0.67
ces	1.03	1.04	1.05	1.06	1.03	1.09	1.04	1.02
deu	5.15	5.47	5.57	5.63	5.60	5.79	5.36	5.57
eng	43.96	43.84	43.20	44.98	46.64	45.70	44.40	44.07
fas	0.66	0.59	0.57	0.63	0.63	0.67	0.71	0.72
fra	4.53	4.56	4.53	4.46	4.50	4.64	4.31	4.31
ind	0.75	0.74	0.76	0.83	0.75	0.82	1.01	1.09
ita	2.06	2.31	2.38	2.42	2.48	2.69	2.53	2.31
jpn	5.47	4.81	4.78	4.66	4.68	4.83	5.03	5.09
kor	0.59	0.69	0.76	0.66	0.65	0.67	0.71	0.77
nld	1.50	1.74	1.79	1.80	1.93	2.09	1.85	1.81
pol	1.68	1.70	1.68	1.62	1.62	1.70	1.80	1.75
por	1.99	2.07	2.16	2.15	1.78	1.27	2.14	2.26
rus	9.27	7.41	7.11	7.07	5.90	5.84	6.02	5.97
spa	4.18	4.16	4.25	4.33	4.36	4.58	4.56	4.46
tur	0.90	0.87	0.91	1.00	0.86	0.84	1.18	1.15
vie	0.75	0.74	0.80	0.92	0.92	1.04	1.01	1.03
zho	4.98	6.76	6.91	5.07	4.73	4.41	4.75	5.12

Language Coverage

- 45% of the content is English
- About 5% – Chinese, French, German, Japanese, Russian, Spanish
- Other languages: 2% or below
- Is this balanced?

Language Identification – Crawler Integration

Requirements (summer 2018)

- Java or Java bindings
- AMD64 and ARM CPU architectures, no accelerator (GPU)
- HTML input, arbitrary character encoding (85% UTF-8)
- Fast, both model load time and identification
  - max. 5 ms per HTML page
  - comparison: 9 ms WARC packaging of one page

Evaluation of Language Identifiers

	Languages	Accuracy				CPU Time
Test Set		dsl2014	europarl	twitter	tatoeba	cumul.
Docs		12600	21000	187461	19457	
MiB		3.3	3.4	18.8	0.9	
Languages		10	21	137	143	
cld2 [22]	160	.879	.989	.751	.792	4
cld2 (Java bind) [22, 23, 24]	83	.875	.986	.743	.606	13
cld2 (Java bind) [22, 23, 24]	160	.880	.990	.758	.792	14
cld2 Polyglot (Py bind) [25]	160	.879	.989	.748	.792	14
cld3 [26]	101	.830	.991	.595	.630	29
fasttext [27]	176	.880	.991	.753	.787	4
idNet [28]	463	.682	.997	.721	.680	3:07:07
langid (Python) [29]	97	.790	.992	.695	.604	19:19
langid (C) [30]	97	.790	.992	.695	.604	7
langid (Java) [31]	97	.790	.992	.695	.604	20
optimaize [32]	71	.740	.969	.534	.443	1:25
shuyo-cybozu [33]	52	.761	.993	.620	.420	56

Evaluation of LID – Explanations

- Fair evaluation is difficult
  - Equal number of languages in eval set and supported by LID
  - Mapping language codes
  - Measuring CPU time: no GPU, startup time of interpreter / VM included
- Number of supported (identified) languages
  - 50–80 is not enough
  - more than 200 at cost of speed
  - 150–200 was a good compromise for both accuracy and coverage

About CLD2

- Two variants: 83 or 160 supported languages
- Detects up to 3 languages per document
- Plain text or HTML input, valid UTF-8
- Quadgram Naïve Bayesian classifier with support of
  - Writing system / Unicode script
  - Ignoring web-specific character sequences (*http*, *copyright*)
  - Top-level domain of URL
  - Language codes in HTML and HTTP metadata
  - Character encoding (`BIG5`, `KOI8-R`)
- C++ library, trained on a proprietary corpus; no support for re-training

Integration of CLD2

- Debian / Ubuntu package, compiled library (shared object) for various CPU architectures (`x86_64`, `aarch64`, etc.)
  - 160 languages variant selected per environment variable `LD_PRELOAD=libcld2_full.so`
- Java bindings [24] based on JNA
  - Configure hints (encoding, TLD, etc.)
  - Result pruning: minimum 10 bytes or 2% of the entire text
- +35% CPU time for identification of character encoding and language during WARC packaging

Language Annotations in Common Crawl Data

- WARC metadata record

```
charset-detected: EUC-KR
languages-cld2: {
  "reliable":true,
  "text-bytes":2404,
  "languages":[
    {"code":"ko", "code-iso-639-3":"kor",
     "text-covered":0.97, "score":3735.0, "name":"Korean"},
    {"code":"en", "code-iso-639-3":"eng",
     "text-covered":0.02, "score":1264.0, "name":"ENGLISH"}]}
```

- WAT (same as in WARC metadata record)
- WET

  `WARC-Identified-Content-Language: kor,eng`

- URL indexes: `kor,eng` (up to 3 language codes)
  - Good for exploration, see appendix “Discovery Of Dutch TLDs”

Improving Language Coverage and Balance

Initiatives to improve the language balance

- The Web Languages project
  (`https://github.com/commoncrawl/web-languages`)
  - Crowd-sourced effort to improve crawling of low-resource languages
  - Show the crawler high-quality links to “steer” into the right direction
  - Hopefully, this will influence the harmonic centrality ranks over time

- Better text extracts and language identification

Questions?

References i

[1] Amazon Web Services. ***Open Data Sponsorship Program.*** https://aws.amazon.com/opendata/open-data-sponsorship-program/.

[2] Amazon Web Services. ***Registry of Open Data on AWS.*** https://registry.opendata.aws/.

[3] Jason R. Smith et al. **"Dirt Cheap Web-Scale Parallel Text from the Common Crawl".** In: (2013), pp. 1374–1383. http://www.aclweb.org/anthology/P13-1135.

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. **"GloVe: Global vectors for word representation".** In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543. https://aclanthology.org/D14-1162.pdf.

References ii

[5] ***Index to WARC Files and URLs in Columnar Format.*** 2018. https://commoncrawl.org/2018/03/index-to-warc-files-and-urls-in-columnar-format/.

[6] Sebastian Nagel. ***Accessing WARC files via SQL.*** Poster at IIPC Web Archiving Conference, 6–7 June 2019, Zagreb, Croatia. 2019. https://digital.library.unt.edu/ark:/67531/metadc1608961/.

[7] ***sitemaps.org.*** https://www.sitemaps.org/protocol.html.

[8] Hsin-Tsang Lee et al. **"IRLbot: Scaling to 6 Billion Pages and Beyond".** In: *ACM Trans. Web* 3.3 (July 2009). ISSN: 1559-1131. https://doi.org/10.1145/1541822.1541823.

[9] Paolo Boldi and Sebastiano Vigna. **"Axioms for Centrality".** In: *CoRR* abs/1308.2140 (2013). http://arxiv.org/abs/1308.2140.

References iii

[10] Paolo Boldi and Sebastiano Vigna. **"In-Core Computation of Geometric Centralities with HyperBall: A Hundred Billion Nodes and Beyond".** In: *2013 IEEE 13th International Conference on Data Mining Workshops* (2013), pp. 621–628. https://vigna.di.unimi.it/papers.php#BoVHB.

[11] Paolo Boldi. ***A modern view of centrality measures.*** 2013. https://www.youtube.com/watch?v=cnGJtGP4gL4.

[12] ***Public Suffix List.*** https://publicsuffix.org/.

[13] ***QS World University Rankings: The top 100 universities in the USA.*** https://www.topuniversities.com/where-to-study/north-america/united-states/ranked-top-100-us-universities.

[14] ***Forbes America's Top Colleges List 2025 - Best US Universities Ranked.*** https://www.forbes.com/top-colleges/.

References iv

[15] ***Host- and Domain-Level Web Graphs October, November, December 2024.*** https://commoncrawl.org/blog/host--and-domain-level-web-graphs-october-november-and-december-2024.

[16] ***Reverse domain name notation.*** https://en.wikipedia.org/wiki/Reverse_domain_name_notation.

[17] ***PageRank.*** https://en.wikipedia.org/wiki/PageRank.

[18] Martijn Koster et al. ***Robots Exclusion Protocol.*** Tech. rep. 9309. Sept. 2022. 12 pp. https://www.rfc-editor.org/info/rfc9309.

[19] Martijn Koster. ***A Standard for Robot Exclusion.*** 1996. https://www.robotstxt.org/meta.html.

References v

[20] Gary Illyes. ***Robots Exclusion Protocol Extension for URI Level Control.*** Internet-Draft draft-illyes-repext-02. Work in Progress. Internet Engineering Task Force, Oct. 2024. 6 pp. https://datatracker.ietf.org/doc/draft-illyes-repext/02/.

[21] ***Statistics of Common Crawl Monthly Archives.*** https://commoncrawl.github.io/cc-crawl-statistics/.

[22] ***CLD2Owners/cld2.*** 2015. https://github.com/CLD2Owners/cld2 (visited on 10/03/2025).

[23] ***cld2 package : Ubuntu.*** https://launchpad.net/ubuntu/+source/cld2.

[24] ***commoncrawl/language-detection-cld2.*** 2018. https://github.com/commoncrawl/language-detection-cld2.

References vi

[25] Rami Alrfou. ***aboSamoor/polyglot.*** 2014. https://github.com/aboSamoor/polyglot.

[26] ***google/cld3.*** 2016. https://github.com/google/cld3.

[27] ***Language identification · fastText.*** 2016. https://fasttext.cc/index.html.

[28] Jonathan Dunn. ***jonathandunn/idNet.*** 2018. https://github.com/jonathandunn/idNet.

[29] saffsd. ***saffsd/langid.py.*** 2011. https://github.com/saffsd/langid.py.

[30] saffsd. ***saffsd/langid.c.*** 2014. https://github.com/saffsd/langid.c (visited on 10/03/2025).

[31] Dawid Weiss. ***carrotsearch/langid-java.*** Nov. 2013. https://github.com/carrotsearch/langid-java.

[32] ***optimaize/language-detector.*** 2014.
https://github.com/optimaize/language-detector.

[33] Nakatani Shuyo. ***shuyo/language-detection.*** 2010.
https://github.com/shuyo/language-detection.

[34] ***Common Crawl Index Table (Data).***
https://data.commoncrawl.org/cc-index/table/cc-main/index.html.

[35] ***Interactive SQL - Serverless Query Service - Amazon Athena - AWS.***
https://aws.amazon.com/athena/.

Why English Content Is Favored?

Accept-Language HTTP header: `en-US,en;q=0.5`

- Multi-lingual sites may show English content or redirect to the English (sub)site
- Header is required, randomizing it may cause indeterministic behavior

Crawler is operated from data center located in the US (Northern Virginia)

- Multi-lingual sites may show or redirect to language/region-specific site based on geo-located request IP address
- Content from sites hosted in a geographically close location (given the network topology) are fetched faster and less likely to time out

Solution: crawl from different location

Fetch Time By Top-Level Domain (Nov/Dec 2022)

tld	ms/100kiB	avg. page kiB	ms/page
ca	596.2	148.0	882.1
us	651.2	137.5	895.1
co	664.2	146.3	971.9
dk	667.8	146.1	975.4
com	671.5	152.9	1026.7
ar	720.8	175.6	1265.6
ch	724.9	153.7	1114.1
gov	725.7	122.8	891.4
no	727.2	134.0	974.3
fi	754.0	131.1	988.1
ru	769.7	123.9	953.8
be	785.5	133.6	1049.1
org	793.5	113.5	900.8
edu	810.7	83.3	675.2
gr	818.3	165.0	1350.3
se	819.1	132.0	1080.8
ie	834.6	162.1	1352.9
ua	842.7	124.4	1048.5
at	845.2	132.4	1118.7
fr	849.0	131.0	1111.8
ro	854.7	134.3	1147.5
pl	864.3	123.6	1068.1
sk	870.6	135.4	1179.0
in	877.1	157.6	1381.9
de	891.5	125.0	1114.5
hu	892.6	134.0	1196.3
net	894.2	112.2	1003.0
pt	931.9	137.4	1280.7
cz	964.0	104.3	1005.0
nl	978.7	121.5	1189.3
es	979.3	130.5	1277.8
uk	980.9	91.3	895.7
cl	985.2	141.9	1398.2
eu	1023.4	122.8	1256.9
it	1068.1	130.9	1397.9
info	1107.0	96.5	1068.4
br	1121.1	112.6	1262.1
kr	1256.2	107.9	1355.7
jp	1356.4	91.8	1244.7
id	1538.7	112.9	1737.4
vn	1543.5	126.6	1954.0
ir	1652.8	115.9	1916.3
cn	1838.2	62.9	1156.5

Language Codes in HTML and HTTP

```
<html lang="en">
<doc xml:lang="en">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en-US">

<meta http-equiv="content-language" content="en-GB" />
<meta name="language" content="English">
<meta name="DC.language" content="fr-ca"/>

<span id="1" class="info" lang='en'>
```

HTTP header

```
Content-Language: el-GR
```

Discovery Of Dutch TLDs (i)

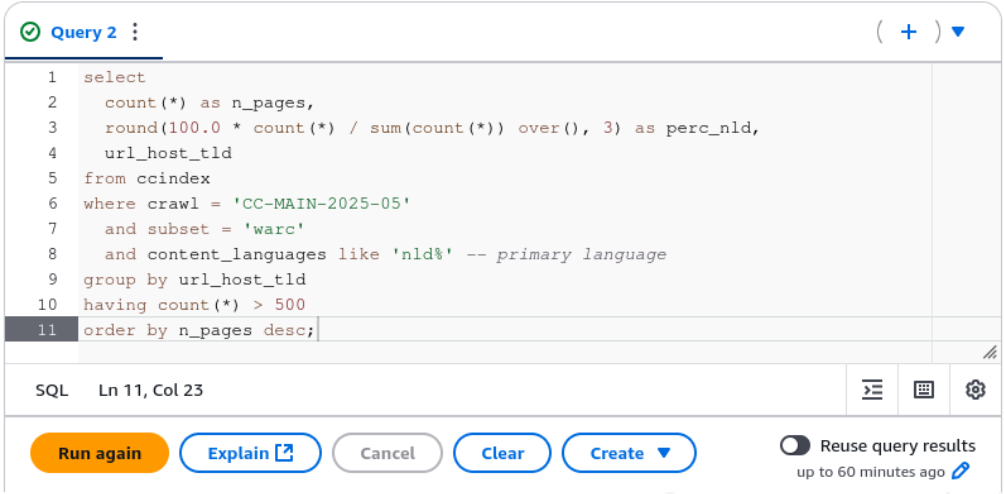
Data in the columnar index

- URL and parts (host name, registered domain, top-level domain, path, query)
- Capture metadata (fetch time, size, WARC record location)
- Content metadata (MIME type, charset, content languages detected by CLD2)

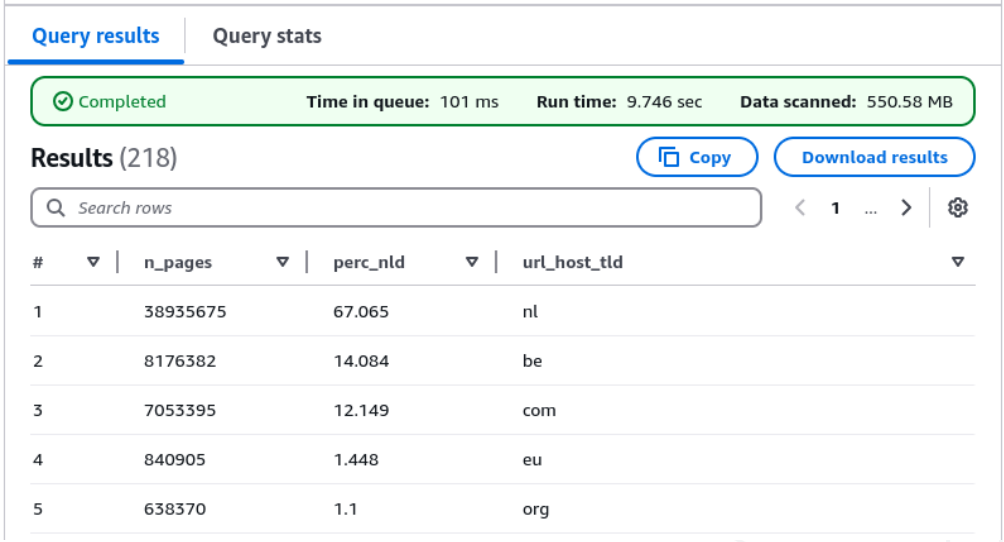
Instructions for setup and example queries in [5, 34]

The following example queries were run with Amazon Athena [35] engine v3, running on Trino (trino.io).

Discovery Of Dutch TLDs (ii)



Discovery Of Dutch TLDs (iii)



Query results | Query stats

Completed — Time in queue: 101 ms — Run time: 9.746 sec — Data scanned: 550.58 MB

Results (218) — Copy — Download results

Search rows — 1 ...

#	n_pages	perc_nld	url_host_tld
1	38935675	67.065	nl
2	8176382	14.084	be
3	7053395	12.149	com
4	840905	1.448	eu
5	638370	1.1	org

Discovery Of Dutch TLDs (iv)

5	638370	1.1	org
6	525475	0.905	net
7	313201	0.539	nu
8	191459	0.33	info
9	182334	0.314	shop
10	165178	0.285	de
11	62564	0.108	fr
12	60202	0.104	tv
13	51763	0.089	online
14	32868	0.057	vlaanderen
15	28607	0.049	co
16	26895	0.046	store
17	26525	0.046	amsterdam

Discovery Of Dutch TLDs (v)

Ok. We found top-level domains with Dutch content.

But there's a lot of noise. Can we do better?

Let's try to sort by the percentage of Dutch pages within a TLD!

We do it in steps

- Extract the primary language and other columns we need
- Count the total number of pages per TLD
- Select Dutch content, calculate the percentage and sort the result

Discovery Of Dutch TLDs (vi)

Query 1

```
with tmp1 as (select
  url_host_tld AS tld,
  regexp_extract(content_languages, '^([a-z]{3})')
     as primary_language
from ccindex
where crawl = 'CC-MAIN-2025-05'
  and subset = 'warc'),

tmp2 as (select
  count(*) as n_pages,
  tld,
  primary_language,
  sum(count(*)) over (partition by tld) as total_tld
from tmp1
group by tld, primary_language)

select
```

SQL Ln 20, Col 35

Discovery Of Dutch TLDs (vii)

```
group by tld, primary_language;

select
  n_pages,
  round(100.0*n_pages/total_tld, 3) as perc_tld,
  -- calculate the percentage of Dutch pages per TLD
  round(100.0*n_pages/sum(n_pages) over (), 3) as perc_nld,
  tld
from tmp2
where primary_language = 'nld'
  and n_pages > 500
group by tld, n_pages, total_tld
order by perc_tld desc;
```

SQL Ln 20, Col 35

Run | Explain | Cancel | Clear | Create

Reuse query results
up to 60 minutes ago

Query results | Query stats

Completed **Time in queue:** 85 ms **Run time:** 14.444 sec **Data scanned:** 543.44 MB

Discovery Of Dutch TLDs (viii)

#	n_pages	perc_tld	perc_nld	tld
1	32868	86.902	0.057	vlaanderen
2	38935675	82.727	67.065	nl
3	1595	82.047	0.003	bauhaus
4	23400	78.979	0.04	frl
5	12572	69.145	0.022	gent
6	26525	57.635	0.046	amsterdam
7	624	53.47	0.001	lease
8	8176382	50.591	14.084	be
9	5799	43.27	0.01	cw
10	21868	32.446	0.038	sr
11	19422	27.775	0.033	brussels
12	313201	21.725	0.539	nu
13	1185	19.856	0.002	auto
14	1237	14.264	0.002	aw

Discovery Of Dutch TLDs (ix)

Done. A list of top-level domains to restrict a crawl to Dutch content.

Curaçao (`.cw`) and Aruba (`.aw`) are part of the Kingdom of the Netherlands. Suriname (`.sr`) was a Dutch colony.