COMMON CRAWL

Data Quality, Language Coverage and Ethical Considerations in Web Crawling

Sebastian Nagel sebastian@commoncrawl.org
Pedro Ortiz Suarez pedro@commoncrawl.org

HPLT & NLPL Winter School on Pretraining Data Quality and Multilingual LLM Evaluation, February 3–5, 2025

A Brief Introduction

- About Common Crawl
- Data Overview

Data Collection and Sampling of Web Pages

Data Collection and Sampling – History and Metrics

Data Usage and Formats

Common Crawl and NLP/ML

On Linguistic and Cultural Diversity

Ethical Considerations and Crawler Politeness

Questions and References

Appendix – Tools and Examples

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, web science, information retrieval, semantic web, internet security research, …
- Used as training data: today LLMs, 2014 GloVe word embeddings [3]

Data Overview

- Over 275 billion web pages spanning 17 years (2008 – 2025/now)
- 2.5 – 3 billion pages added each month
- More than 100 crawl archives released to date
- 9 PiB of data (end of 2024)

Sampling Web Pages – Targets and Objectives

- We want a *diverse* sample of web pages
- Representative for the WWW,
  or the crawlable subset (robots.txt) of the public web
- A compromise between breadth (geographically, by language, topic, etc.)
  and in-depth coverage of individual sites
- Fair use recommends to collect only a sample of the pages of one site
- Balance between new pages and revisits (page visited in earlier crawls)
- Focus on HTML, allow PDFs and other textual formats,
  avoid multimedia, binaries, executables, etc.

The Need for Sampling

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

- An average “monthly” crawl includes 3 billion page captures with
  - 500+ billion links
  - 25+ billion unique URLs linked
- Up to 2.5 billion URLs listed in a single sitemap (sitemap index) [4]

Need to select a diverse and representative 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!

Stratified Domain-level Sampling

Domain-level ranks

- Define a “budget” [5] per registered domain
  - How many URLs/pages are sampled per domain
  - Domain: one level below the registry suffix, e.g. `w.org`, `data.gov.uk`)
- Are used during URL discovery to sample sitemaps or home pages (top-ranking domains: always, decreasing likelihood for lower ranks)
- Are “projected” to the page-level by inlink count or OPIC [6]
  - Rank the pages within a domain
  - ! We have no absolute “page quality metrics” comparing two pages from two different domains
  - ! A page-level graph would be too large: too costly to build and rank

Domain-Level Graph-Based Ranking Example

- Top-N `.edu` domains ranked by harmonic centrality (or pagerank) calculated on CC's domain-level hyperlink graphs [7]
- Reverse domain name notation [8]
- Order by harmonic centrality ("hc") [9, 10]
  - ranks are shown not scores
  - PageRank rank [11], too
  - global ranks over domains below all top-level domains, not only `.edu`
- Includes not only universities (*)
- Compared with university rankings by QS World [12] and Forbes [13]

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 [12]
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 [13]
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

Current Status and Configuration i

Configuration of per-domain budgets (2024)

- Top 20 domains (e.g., wikipedia.org, microsoft.com)
  - max. 30 million URLs
  - 180k URLs per single host (subdomain)
  - 360k subdomains
- Long tail (below rank 80M or yet unseen)
  - max. 500 URLs, 400 per subdomain, 4 subdomains
- Log distribution between top and tail
  - pos 1000: max. 1.2M, 36k per subdomain, 4k subdomains
  - pos 1M: max. 2.5k, 2.2k per subdomain, 20 subdomains

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

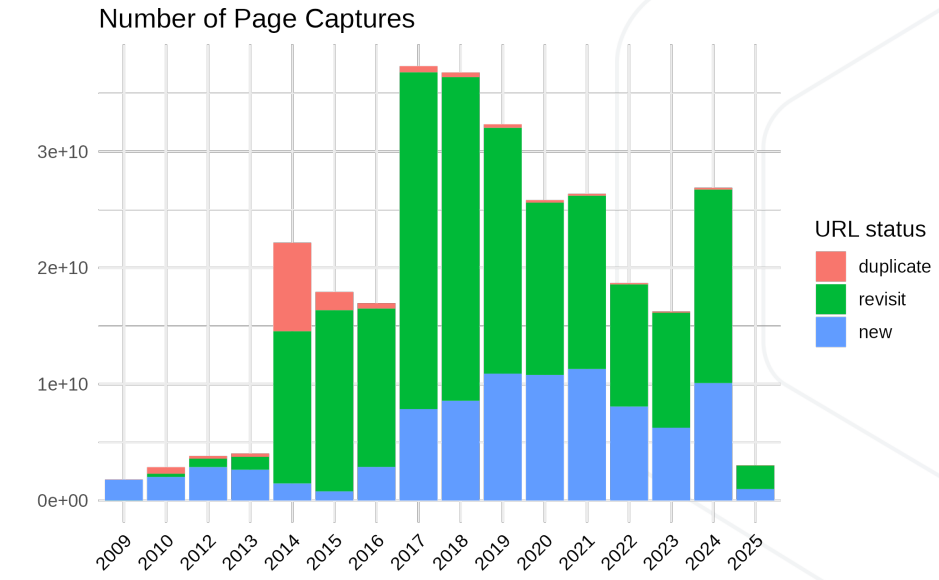
	crawler	seeds / link prioritization	revisit policy
2008–2009	Nutch	list based	independent,
2012	in-house	page rank	yearly crawls
2013 – 2016	Nutch	(seed donations)	
2017 – now	Nutch	harmonic centrality	30% new, 70% revisits

What Is Representative?

Aspects of representativity:

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

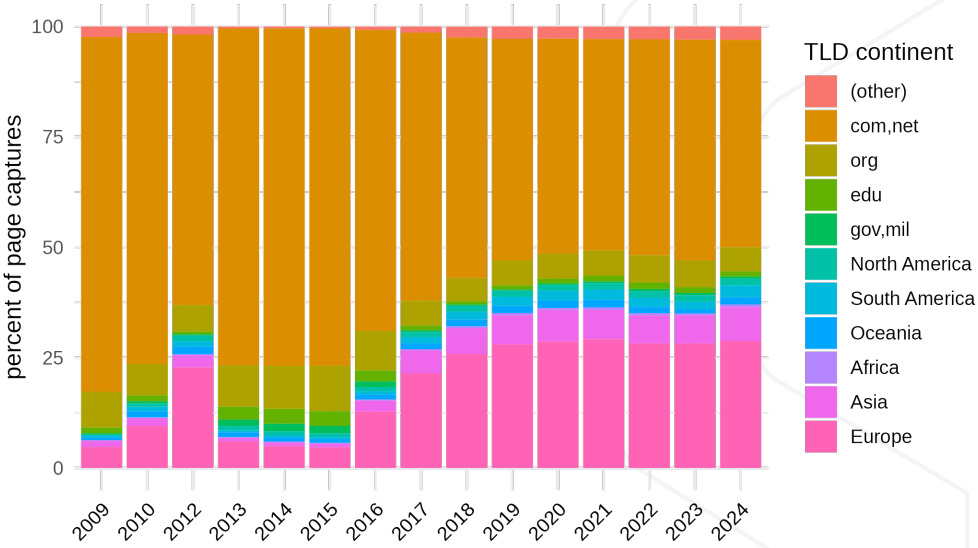
Size and Freshness



Top-Level Domains and Geographical Coverage

% year	com	org	ru	net	de	uk	jp	edu	fr	it	pl	nl	br	cz
(all)	53.10	6.53	4.00	3.82	3.19	2.05	1.60	1.48	1.33	1.31	1.19	1.07	0.92	0.80
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	0.01
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	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	0.76
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	0.10
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	0.06
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	0.05
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	0.52
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	0.85
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	0.93
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	0.99
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	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
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	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	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	1.03

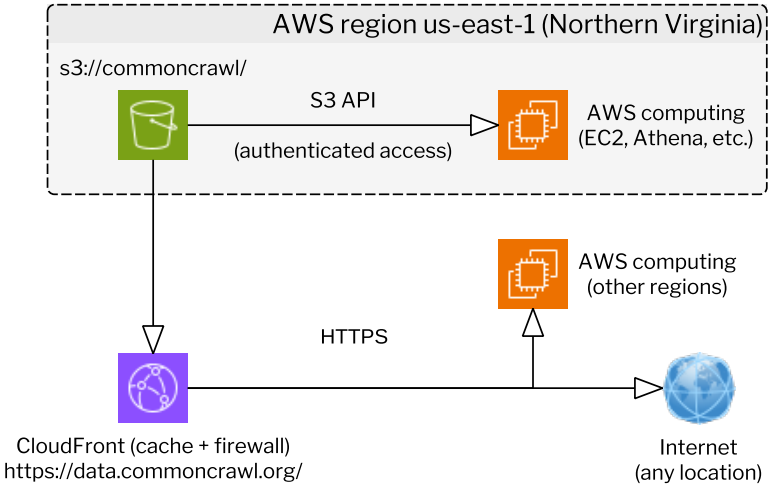
Top-Level Domains and Geographical Coverage



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

Data Location and Access



Data Location and Access

- S3 API to access the data from AWS in the us-east-1 region
  - AWS account and authentication are required
- HTTPS (via CloudFront) from everywhere else (internet and other AWS regions)
  - `https://data.commoncrawl.org/`
  - `https://ds5q9oxwqwsfj.cloudfront.net/`
  - data caching and DDoS protection
- S3 and CloudFront status monitoring

  `https://status.commoncrawl.org/`

cc-downloader

To avoid download errors for our users we created `cc-downloader`, a fast and polite downloader for our data.

It has 2 sub-commands, `download-paths`

```
cc-downloader download-paths CC-MAIN-2024-46 wet path/to/folder
```

And, `download`

```
cc-downloader download path/to/folder/wet.paths.gz path/to/folder
```

This will preserve the tree structure that we use internally by default.

Data Formats

- WARC – web page captures
- ARC – predecessor of WARC
  - used to store data until 2012
  - to be converted to WARC
- WAT – metadata and links
- WET – plain text extracted from HTML
  - no markup or removal of boilerplate content (navigation, header, footer)

Data Formats (ii)

- Index – URL, metadata and WARC record location
  - CDX format – powers our wayback machine (index.commoncrawl.org)
    - good to look up single URLs or domains
  - Columnar format (Parquet) [14, 15]
    - 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)

The WARC format (Web ARChive)

- “Freezes” the internet traffic between a client (web crawler or browser) and web servers at the HTTP protocol level
  - content payload
  - HTTP headers
  - connection metadata (datetime, IP address)
- ISO standard [16, 17]
- WARC I/O modules for many programming languages [18]
- Per-record gzipped: extract single records by offset

Common Crawl WARC Specifics

Specifics of Common Crawl WARC and ARC collections (in difference to other web archivers)

- Only successful fetches (HTTP status 200) recorded
- Separate subsets since 2016 [19]
  - HTTP status other than 200: 404 Not Found, 304 Not Modified, redirects, …
  - robots.txt
- Decoded HTTP content and transfer encoding (decompress, unchunk)
- No page dependencies (images, CSS, etc.)
- Shuffled – every WARC file is a (pseudo-)random sample by its own
- 1 MiB content limit – longer payloads are truncated
  - about 2.5% of all captures are truncated (but 25% of PDF files)

Metadata

- “Replicability” of data collection
  - WARC and HTTP headers
  - robots.txt, 404s and redirects
- Annotations
  - detected MIME
  - identified language (by CDL2) and character set
  - (we need more!)

Common Crawl for NLP/ML through the years

- GloVe [3] and Fasttext [20] were pre-trained using Common Crawl data providing good static embeddings
- But the consensus in 2019 was that web data did not produce good contextualized embeddings [21, 22]
- So use Wikipedia and Books like BERT? [23]
- But maybe web data can work... (RoBERTa/CamemBERT) [24, 25]

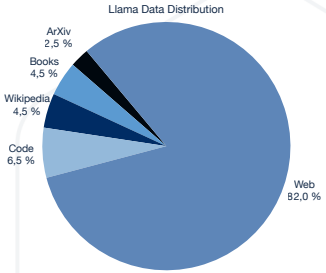
Web Data vs Wikipedia

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		**Average**		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	**89.87**	90.65	**81.88**
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	**89.88**	**91.55**	81.55
Embeddings (with UDPipe Future (tagging, parsing) or LSTM+CRF (NER))													
Wiki	4GB	98.09	92.31	98.74	93.55	96.24	78.91	95.78	89.79	97.21	88.64	91.23	-
CCNet	4GB	**98.22**	**92.93**	99.12	94.65	97.17	**82.61**	**96.74**	89.95	97.81	90.04	**92.30**	-
OSCAR	4GB	98.21	92.77	99.12	**94.92**	97.20	82.47	**96.74**	**90.05**	**97.82**	**90.05**	91.90	-
OSCAR	138GB	98.18	92.77	**99.14**	94.24	**97.26**	82.44	96.52	89.89	97.77	89.84	91.83	-

In the CamemBERT study [24, 25], it was shown that pre-training with samples of Common Crawl filtered data of the same size as Wikipedia yields better scores. Moreover, it was shown that dataset size was key.

Prevalence of web data in LLMs today(ish)

- In practice datasets are not balanced
- Web data is always the cheapest and easiest to get
- Web data is diverse, but definitely not balanced or representative of all the language range.
- Web data always contain unwanted content (fiction, bias, propaganda).
- Programming language code (source code) is becoming ubiquitous in the data mix
- Public domain books and encyclopedic data is also common, but availability varies greatly between languages.



LLaMA Data

Source	Proportion
Web	82%
Code	6.5%
Wikipedia	4.5%
Books	4.5%
ArXiv	2.5%

Multilinguality is hard even with Web data!

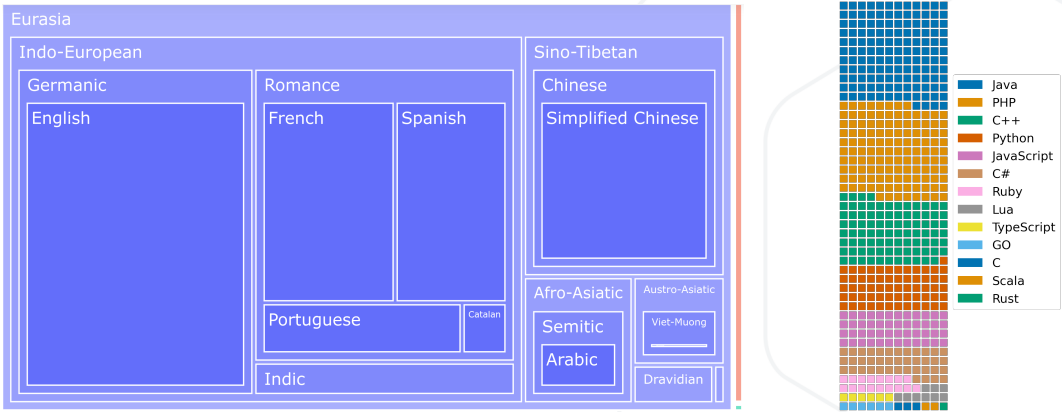
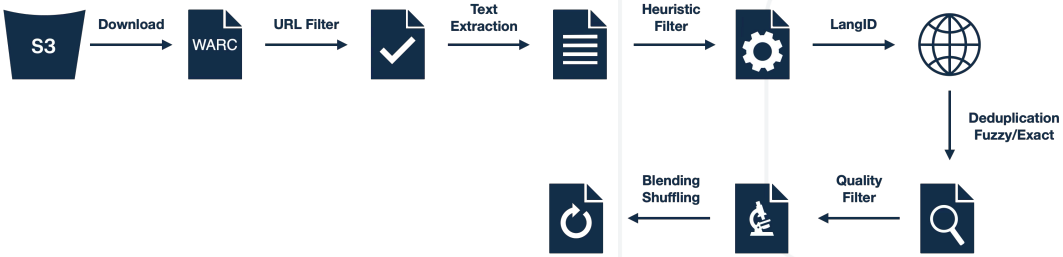
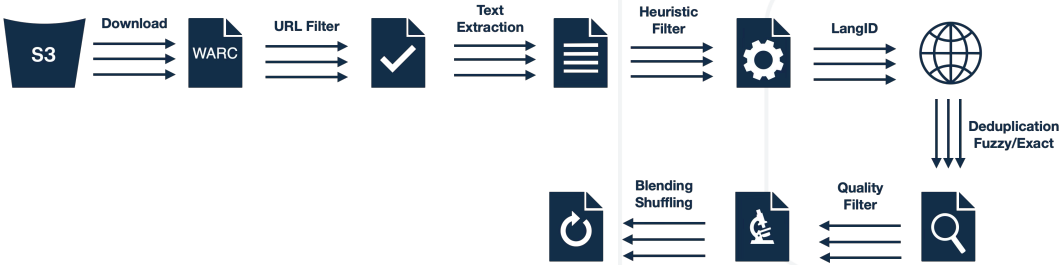


Figure 1: Overview of ROOTS [26] Left: A treemap of natural language representation in number of bytes by language family. Right: A waffle plot of the distribution of programming languages by number of files. One square corresponds approximately to 30,000 files.

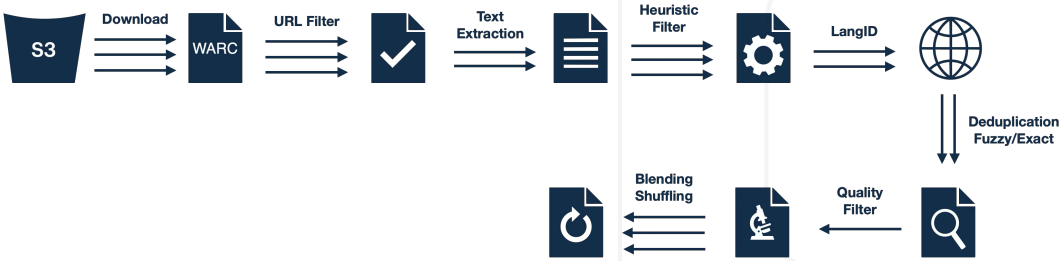
The Anatomy of a Pre-Processing Pipeline



Pipelines are Parallelizable



But Some Components Are Expensive



A Survey of Pre-Processing Pipelines

- OSCAR (First use of LangID treshold for filtering) [27, 28]
- CCNet (Uses KenLM models [29] trained on Wikipedia) [30]
- mC4 (Uses word blocklists) [31]
- Refined Web (Uses Trafilatura for text extraction) [32]
- DataTrove (Uses Trafilatura and aggregates filters) [33]
- Dolma (Works on deduplication) [34]
- NeMo-Curator (Combines many heuristics and filters) [35]
- HPLT (Uses custom pipeline and has bilingual data) [36]
- MADLAD (Uses custom LangID and focuses on Multilinguality) [37]
- GlotCC (Uses Ungoliant [28] plus a new LangID model) [38]
- Community OSCAR (KenLM models trained on Adult Content) [39]
- And many others!

URL Filtering

- Most projects use the UT1 website from Université Toulouse 1 Capitole propose a blacklist managed by Fabrice Prigent: https://dsi.ut-capitole.fr/blacklists/
- We recommend contributing to the project, it is still going!
- But please know that **blocklists are opinionated!**
- Other projects have custom blocklist that were language specific: https://github.com/oscar-project/oscar-blocklists
- We need more language specific lists!

Text Extraction

- **jusText** [40] is a heuristic algorithm based on hand-crafted rules, it uses stopword density
- **Readability** [41] is another heuristic algorithm based on hand-crafted rules. It is implemented in Firefox for providing a “reader view”
- **Resiliparse** [42] The Resiliparse HTML2Text extractor is a heuristic algorithm based on tag rules and regular expressions and it focuses on extraction precision and speed
- **Trafilatura** [43] is a heuristic algorithm employing a cascade of XPath queries for finding the main content, falling back to jusText and Readability (see above) should the extraction fail
- We recommend taking a look at the Webis Group Survey [44] on this topic for a more detailed overview and evaluation

Heuristic Filtering

- Most of the pipelines follow the Gopher Heuristic Filtering [45] which includes:
  - Filtering on document length
  - Filtering on symbol-to-word ratio
  - Removing documents that are mostly lists
  - Requiring that a certain percentage of words contain alphabetic characters
  - Filtering with stop-words list
  - Removing documents with excessive repetition of words
  - Among others
- Other filters might also include things like *cursed regexes* [37]
- We discourage the use of blocklist for words!

Deduplication

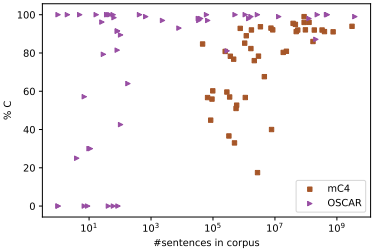
- Exact deduplication:
  - SHA-1 [46]
  - MD5 [47]
  - xxhash64 (an extremely fast non-cryptographic hash algorithm) [48]
- Fuzzy deduplication:
  - MinHash [49]
  - TLSH [50]

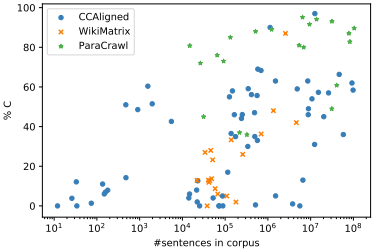
Quality Filtering

- KenLM models trained on Wikipedia [30]
- KenLM models trained on adult content [51]
- PII Indentification
- Detoxification
- Quality classifiers [52, 53]
- Education classifiers [54]

A Human Evaluation of Language Identification

		Parallel			Monolingual	
		CCAligned	ParaCrawl v7.1	WikiMatrix	OSCAR	mC4
	#langs audited / total	65 / 119	21 / 38	20 / 78	51 / 166	48 / 108
	%langs audited	54.62%	55.26%	25.64%	30.72%	44.44%
	#sents audited / total	8037 / 907M	2214 / 521M	1997 / 95M	3517 / 8.4B	5314 / 8.5B
	%sents audited	0.00089%	0.00043%	0.00211%	0.00004%	0.00006%
macro	C	29.25%	76.14%	23.74%	87.21%	72.40%
macro	X	29.46%	19.17%	68.18%	-	-
macro	WL	9.44%	3.43%	6.08%	6.26%	15.98%
macro	NL	31.42%	1.13%	1.60%	6.54%	11.40%
macro	offensive	0.01%	0.00%	0.00%	0.14%	0.06%
macro	porn	5.30%	0.63%	0.00%	0.48%	0.36%
micro	C	53.52%	83.00%	50.58%	98.72%	92.66%
micro	X	32.25%	15.27%	47.10%	-	-
micro	WL	3.60%	1.04%	1.35%	0.52%	2.33%
micro	NL	10.53%	0.69%	0.94%	0.75%	5.01%
micro	offensive	0.00%	0.00%	0.00%	0.18%	0.03%
micro	porn	2.86%	0.33%	0.00%	1.63%	0.08%
	#langs =0% C	7	0	1	7	0
	#langs <50% C	44	4	19	11	9
	#langs >50% NL	13	0	0	7	1
	#langs >50% WL	1	0	0	3	4





Taken from [55]

On the Limitations of Language Identification

#LID classifier	Dataset	Corr.guesses	Ratio	Time
cld2-polyglot	dsl2014	11077/12600	0.879127	0.688s
cld2-python	dsl2014	11072/12600	0.87873	0.595s
cld3-python	dsl2014	10454/12600	0.829683	3.968s
fasttext -m lid.176.bin	dsl2014	11084/12600	0.879683	0.391s
fasttext -m lid.176.bin -l	dsl2014	11023/12600	0.874841	0.358s
fasttext -m lid.176.ftz	dsl2014	10678/12600	0.84746	0.570s
fasttext -m lid.218a.bin	dsl2014	11161/12600	0.885794	2.815s
fasttext -m lid.218a.bin -l	dsl2014	11186/12600	0.887778	2.800s
fasttext -m lid.218a.ftz	dsl2014	11154/12600	0.885238	7.031s
cld2-polyglot	europarl	20775/21000	0.989286	0.695s
cld2-python	europarl	20779/21000	0.989476	0.589s
cld3-python	europarl	20811/21000	0.991	4.300s
fasttext -m lid.176.bin	europarl	20804/21000	0.990667	0.410s
fasttext -m lid.176.bin -l	europarl	20805/21000	0.990714	0.408s
fasttext -m lid.176.ftz	europarl	20681/21000	0.98481	0.537s
fasttext -m lid.218a.bin	europarl	20966/21000	0.998381	3.225s
fasttext -m lid.218a.bin -l	europarl	20963/21000	0.998238	3.175s
fasttext -m lid.218a.ftz	europarl	20969/21000	0.998524	7.578s

#LID classifier	Dataset	Corr.guesses	Ratio	Time
cld2-polyglot	tatoeba	15415/19457	0.79226	0.451s
cld2-python	tatoeba	15417/19457	0.792363	0.312s
cld3-python	tatoeba	12250/19457	0.629593	2.096s
fasttext -m lid.176.bin -l	tatoeba	15108/19457	0.776481	0.149s
fasttext -m lid.176.bin	tatoeba	15317/19457	0.787223	0.169s
fasttext -m lid.176.ftz	tatoeba	13631/19457	0.70057	0.168s
fasttext -m lid.218a.bin -l	tatoeba	15697/19457	0.806753	1.663s
fasttext -m lid.218a.bin	tatoeba	15741/19457	0.809015	1.702s
fasttext -m lid.218a.ftz	tatoeba	15777/19457	0.810865	2.536s
cld2-polyglot	twitter	140240/187461	0.748102	5.972s
cld2-python	twitter	140858/187461	0.751399	3.752s
cld3-python	twitter	111628/187461	0.595473	26.457s
fasttext -m lid.176.bin -l	twitter	144679/187461	0.771782	2.279s
fasttext -m lid.176.bin	twitter	143463/187461	0.765295	2.372s
fasttext -m lid.176.ftz	twitter	139219/187461	0.742656	2.593s
fasttext -m lid.218a.bin -l	twitter	127447/187461	0.679859	19.502s
fasttext -m lid.218a.bin	twitter	121086/187461	0.645926	19.751s
fasttext -m lid.218a.ftz	twitter	120192/187461	0.641157	39.719s

Has the Situation Improved?

- Two new better models have been published since:
  - OpenLID [56]
  - GlotLID [57]
- However, the architecture for the model remains the same (fasttext [20])
- The only project that has proposed a new architecture never disclosed any information about it [58]

The LangID Project

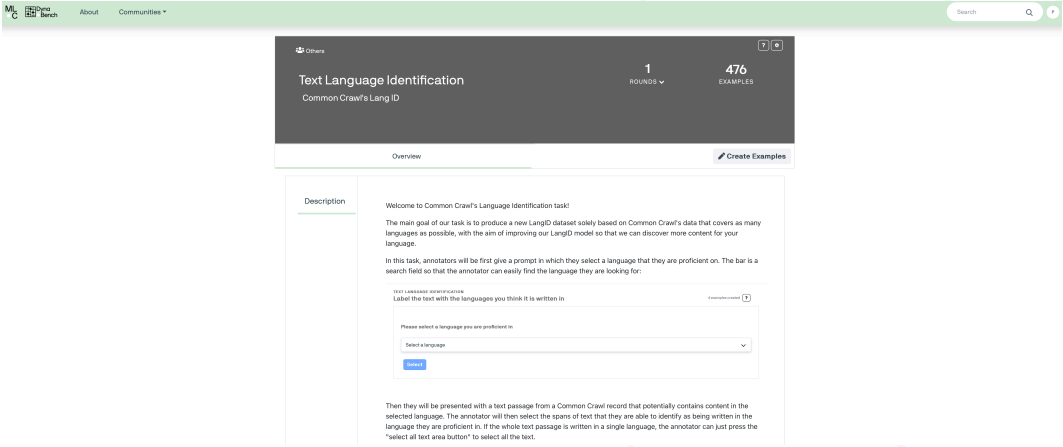


Figure 2: https://dynabench.org/tasks/text-language-identification

The Web-Languages Project

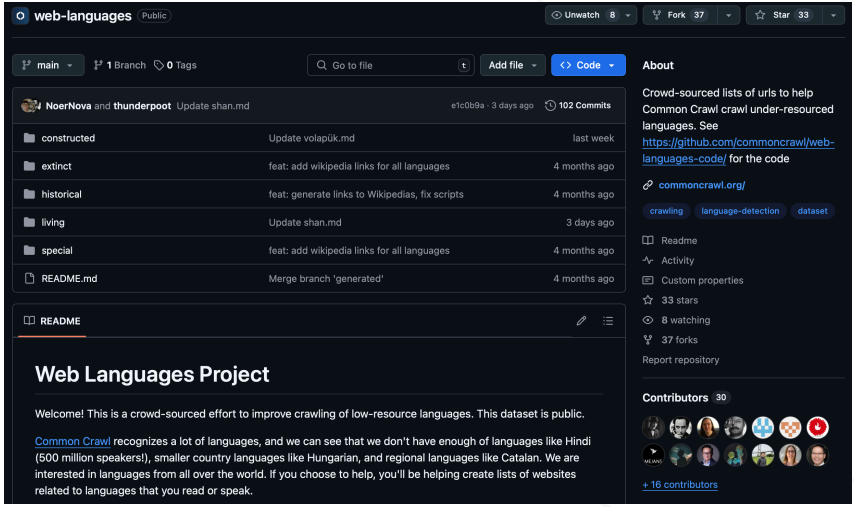


Figure 3: https://github.com/commoncrawl/web-languages

Language Distribution of Recent Crawls

Ethical Considerations

- Environmental impact / resource consumption
- Privacy (GDPR, Data Protection laws)
- Types of data (“Open”, “Public”, “Obtainable”, and “Private”)
- Robots Exclusion Protocol (robots.txt)
- More opt-out vs opt-in (preference signals)
- IAB / IETF recommendations

Crawler Politeness

- Slow crawling
  - (current configuration) min. 3.5 seconds between successive requests to the same host
  - further slow down (exponential backoff) if host responds with errors
- Respect robots.txt rules and
- URI-level metatags (`<meta name=robots value=nofollow>`)

- 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

Robots Exclusion Protocol (REP) - Robots.txt

- a text file `robots.txt` is deployed in the root folder of a web site (eg. `https://example.org/robots.txt`)
- readable for web crawlers (“robots”)
- contains policies whether and how crawlers shall access the site’s content

- a technical solution to coordinate different interests between the owners of content and robots
- a convention based on consensus not a legally binding regulation [59]

Robots.txt – Adoption and Standardization

1994 robots.txt protocol discussed on mailing list [60]

1996 inofficial RFC proposal [61]

- adopted by all major web search engines
- various extensions, conflicting specifications and implementations

2019 RFC draft [62, 63] and reference implementations [64]

2022 RFC 9309 [65]

Robots.txt example (2022)

```
User-agent: Googlebot-News
Disallow: /angebote/

User-agent: *
Disallow: /zeit/
Disallow: /templates/
Disallow: /hp_channels/
Disallow: /send/
Disallow: /suche/
Disallow: /rezepte/suche/
Disallow: */comment-thread?
Disallow: */liveblog-backend*
Disallow: /framebuilder/
Disallow: /campus/framebuilder/

User-agent: Baiduspider
Disallow: /

User-agent: Applebot
Allow: /
Disallow: /cre-1.0/

User-agent: GrapeshotCrawler
crawl-delay: 3

Sitemap: https://www.zeit.de/gsitemaps/index.xml
```

- `https://www.zeit.de/robots.txt`
- visited 2022-08-20
- Googlebot-News and Applebot ev. preferred (more paths allowed)
- Baiduspider penalized
- GrapeshotCrawler [66] to wait 3 seconds between requests
- default rule set excludes templates, duplicated dynamic content or user comments
- improve quality of crawled content and search results!
- announced sitemap provides an up-to-date list of crawlable URLs

Robots.txt example (2024)

```
User-agent: Googlebot-News
Disallow: /angebote/

User-agent: *
Disallow: /zeit/
Disallow: /templates/
Disallow: /hp_channels/
Disallow: /send/
Disallow: /rezepte/suche/
Disallow: */comment-thread?
Disallow: */liveblog-backend*
Disallow: /framebuilder/
Disallow: /campus/framebuilder/
Disallow: /navigation-teasers*

User-agent: Baiduspider
Disallow: /

User-agent: GrapeshotCrawler
crawl-delay: 3

User-agent: GPTBot
Disallow: /

User-agent: Google-Extended
Disallow: /

User-agent: Applebot-Extended
Disallow: /

User-agent: CCBot
Disallow: /

User-agent: Bytespider
Disallow: /

User-agent: anthropic-ai
Disallow: /

User-agent: Claude-Web
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: Timpibot
Disallow: /

User-agent: Meta-ExternalAgent
Disallow: /

User-agent: FacebookBot
Disallow: /

User-agent: Diffbot
Disallow: /

Sitemap: https://www.zeit.de/gsitemaps/index.xml

# Legal notice: zeit.de expressly reserves the right to use its content for comme
# The use of robots or other automated means to access zeit.de or collect or mine
# the express permission of zeit.de is strictly prohibited.
# zeit.de may, in its discretion, permit certain automated access to certain zeit
# If you would like to apply for permission to crawl zeit.de, collect or use data
```

- visited 2024-09-21
- more user-agents penalized
- ‘#‘ starts a comment (until end of line)

Robots.txt – Impact on Web Crawling

- Concerns and research until 2022
  - search engine bias [67, 68]
  - censorship [69]
- ...and after the raise of LLMs and generative AI (since 2023)
  - robots.txt is an acknowledged opt-out protocol [70]
  - widely used which reduces the availability of training data [71, 72]

How widely is the robots.txt for opt-out used and what is the impact?

User-Agents in Top-10k Robots.txt

- 9 years of robots.txt files archived at Common Crawl [73]
  - one crawl analyzed per year (August or September)
- Robots.txt records of 10,000 top-ranking domains
  - ranks from Aug 2022 [74]
  - the domain's "home site" (`domain.com` or `www.domain.com`)
- Missing data points because of
  - site and robots.txt not visited by crawler
  - excluded from crawling on request
  - no robots.txt (HTTP 404) or fetch error
  - domain not registered in years before 2022 or abandoned later

User-Agents in Top-10k Robots.txt (i)

user-agent	2016	2017	2018	2019	2020	2021	2022	2023	2024
(any)	6754	6699	7034	7120	7189	7232	7161	6872	6791
*	6632	6578	6911	7012	7075	7107	7034	6742	6653
googlebot	430	402	461	448	463	467	453	457	431
twitterbot	251	310	363	364	408	447	448	409	392
ahrefsbot	154	164	187	237	287	294	297	286	322
mediapartners-google	369	353	341	335	326	317	297	268	255
adsbot-google	93	97	95	202	218	249	247	246	269
bingbot	173	193	211	226	236	239	241	241	231
mj12bot	127	135	165	180	214	230	224	234	253
semrushbot	36	48	95	145	189	213	220	229	254
baiduspider	190	190	204	214	222	216	213	216	213
yandex	167	170	201	209	200	211	209	210	212
ia_archiver	193	173	191	178	189	187	185	204	196
dotbot	75	86	118	129	152	173	170	202	214
googlebot-news	87	101	125	152	156	165	152	129	126
googlebot-image	128	134	133	147	160	169	148	154	152
slurp	172	177	180	167	168	161	146	131	115
facebookexternalhit	95	95	107	91	119	127	135	117	183
msnbot	171	157	156	147	145	135	113	114	105

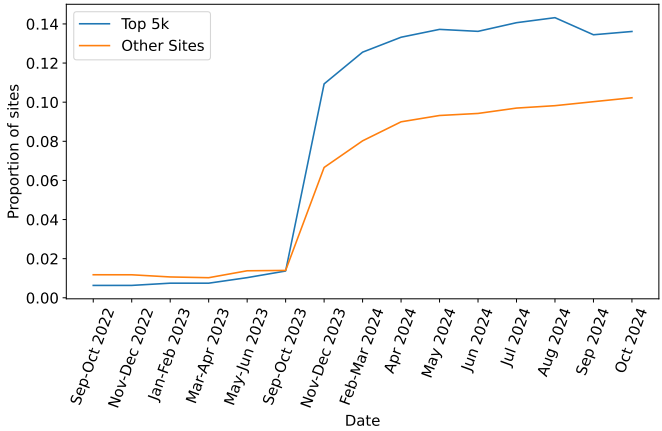
User-Agents in Top-10k Robots.txt (ii)

user-agent	2016	2017	2018	2019	2020	2021	2022	2023	2024
petalbot					13	70	110	121	151
turnitinbot	71	64	65	69	85	83	79	87	154
ccbot	40	34	42	49	62	71	44	300	578
bytespider				2	11	25	24	42	275
omgili	3	3	8	11	17	19	15	32	234
omgilibot	11	13	16	18	17	19	14	33	240
amazonbot					6	14	12	18	203
diffbot	2	2	3	6	6	8	7	8	161
facebookbot							1	5	233
chatgpt-user								167	403
google-extended								52	449
claudebot									305
cohere-ai								6	201
claude-web								2	242
anthropic-ai								23	330
applebot-extended									199
gptbot								614	829
perplexitybot									261

Robots.txt – Impact on Training Data (i)

Liu et al. 2024, Somesite I used to crawl [72]

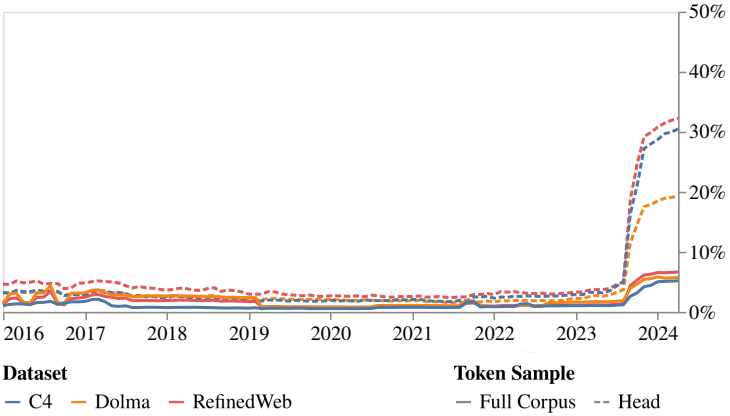
- Proportion of sites that fully disallow any AI-related user agent, broken down by site rank,
- Data: top-5k vs. top-100k Tranco domains, Common Crawl robots.txt captures [73]



Robots.txt – Impact on Training Data (ii)

Percentage of restricted tokens by robots.txt (Longpre et al. 2024, Consent in crisis [71])

- Head: top-2k web sites by token count in C4, Dolma and RefinedWeb
- Full Corpus: 10k randomly sampled sites



AI Training Opt-In and Opt-Out Protocols

Several AI training opt-in/out protocols and initiatives exist:

- "Text & Data Mining Reservation Protocol (TDMRep)", a W3C proposal [75]
  - opt-out mechanism as defined by Digital Single Market EU Directive 2019/790
  - HTTP headers, HTML metadata and `.well-known/tdmrep.json`
- "Vocabulary for Expressing Content Preferences for AI Training", IETF draft [76]
- DECORAIT — A decentralised approach using registries to record opt-in/opt-out choices [77]

AI Training Opt-In and Opt-Out Protocols

Open questions:

- Will there be a single standard, or will there be further fragmentation?
- Can these protocols gain traction or will they meet the same fate as the “Automated Content Access Protocol” (ACAP) [78]
- Lessons from robots.txt and ACAP: simplicity and consensus increase adoption rates

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] 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.

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

References ii

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

[6] Serge Abiteboul, Mihai Preda, and Gregory Cobena. **“Adaptive on-line page importance computation”.** In: (2003). https://dx.doi.org/10.1145/775152.775192.

[7] ***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.

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

References iii

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

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

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

[12] ***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.

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

References iv

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

[15] 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/.

[16] Wikipedia contributors. ***Web ARChive — Wikipedia, The Free Encyclopedia.*** 2021. https://en.wikipedia.org/wiki/Web_ARChive.

[17] ***The WARC Format 1.1.*** https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/.

[18] ***Awesome Web Archiving.***

References v

[19] ***Data Sets Containing Robots.txt Files and Non-200 Responses.*** https://commoncrawl.org/2016/09/robotstxt-and-404-redirect-data-sets/.

[20] Edouard Grave et al. **"Learning Word Vectors for 157 Languages".** In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. https://aclanthology.org/L18-1550/.

[21] Alec Radford et al. **"Language models are unsupervised multitask learners".** In: *OpenAI blog* 1.8 (2019), p. 9. https://aclanthology.org/L18-1550/.

References vi

[22] Trieu H. Trinh and Quoc V. Le. **"A Simple Method for Commonsense Reasoning".** In: *arXiv e-prints*, arXiv:1806.02847 (June 2018), arXiv:1806.02847. doi: 10.48550/arXiv.1806.02847. arXiv: 1806.02847 [cs.AI].

[23] Jacob Devlin et al. **"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".** In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423/.

References vii

[24] Yinhan Liu et al. **“RoBERTa: A Robustly Optimized BERT Pretraining Approach”.** In: *arXiv e-prints*, arXiv:1907.11692 (July 2019), arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692. arXiv: 1907.11692 [cs.CL].

[25] Louis Martin et al. **“CamemBERT: a Tasty French Language Model”.** In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 7203–7219. doi: 10.18653/v1/2020.acl-main.645. https://aclanthology.org/2020.acl-main.645/.

References viii

[26] Hugo Laurençon et al. **“The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset”.** In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 31809–31826. https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

[27] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. **“A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages”.** In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 1703–1714. doi: 10.18653/v1/2020.acl-main.156. https://aclanthology.org/2020.acl-main.156/.

References ix

[28] Julien Abadji et al. **“Towards a Cleaner Document-Oriented Multilingual Crawled Corpus”.** In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 4344–4355. https://aclanthology.org/2022.lrec-1.463/.

[29] Kenneth Heafield. **“KenLM: Faster and Smaller Language Model Queries”.** In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch et al. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 187–197. https://aclanthology.org/W11-2123/.

References x

[30] Guillaume Wenzek et al. **“CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data”.** eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 4003–4012. isbn: 979-10-95546-34-4.
https://aclanthology.org/2020.lrec-1.494/.

[31] Linting Xue et al. **“mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”.** In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 483–498. doi: 10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41/.

[32] Guilherme Penedo et al. **“The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only”.** In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. New Orleans, LA, USA: Curran Associates Inc., 2023.

References xii

[33] Hugging Face. ***DataTrove Library*.** https://github.com/huggingface/datatrove.

[34] Luca Soldaini et al. **"Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research".** In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15725–15788. doi: 10.18653/v1/2024.acl-long.840. https://aclanthology.org/2024.acl-long.840/.

[35] NVIDIA. ***NeMo Curator*.** https://github.com/NVIDIA/NeMo-Curator.

[36] Ona de Gibert et al. **“A New Massive Multilingual Dataset for High-Performance Language Technologies”.** In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 1116–1128. https://aclanthology.org/2024.lrec-main.100/.

[37] Sneha Kudugunta et al. **“MADLAD-400: A Multilingual And Document-Level Large Audited Dataset”.** In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 67284–67296. https://proceedings.neurips.cc/paper_files/paper/2023/file/d49042a5d49818711c401d34172f9900-Paper-Datasets_and_Benchmarks.pdf.

[38] Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. **“GlotCC: An Open Broad-Coverage CommonCrawl Corpus and Pipeline for Minority Languages”.** In: *arXiv e-prints*, arXiv:2410.23825 (Oct. 2024), arXiv:2410.23825. doi: 10.48550/arXiv.2410.23825. arXiv: 2410.23825 [cs.CL].

[39] Manuel Brack et al. **“Community OSCAR: A Community Effort for Multilingual Web Data”.** In: *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*. Ed. by Jonne Sälevä and Abraham Owodunni. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 232–235. doi: 10.18653/v1/2024.mrl-1.19. https://aclanthology.org/2024.mrl-1.19/.

[40] Jan Pomikálek. **“Removing Boilerplate and Duplicate Content from Web Corpora [online]”.** SUPERVISOR: prof. PhDr. Karel Pala, CSc. Doctoral theses, Dissertations. Masaryk University, Faculty of InformaticsBrno, 2011 [cit. 2025-02-03]. https://theses.cz/id/nqo9nn/.

[41] Mozilla. ***Readability*.** https://github.com/buriy/python-readability.

[42] ChatNoir. ***Resiliparse*.** https://resiliparse.chatnoir.eu/en/stable/api/extract/html2text.html.

[43] Adrien Barbaresi. **“Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction”.** In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Ed. by Heng Ji, Jong C. Park, and Rui Xia. Online: Association for Computational Linguistics, Aug. 2021, pp. 122–131. doi: 10.18653/v1/2021.acl-demo.15.
https://aclanthology.org/2021.acl-demo.15/.

[44] Janek Bevendorff et al. **“An Empirical Comparison of Web Content Extraction Algorithms”.** In: *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. Ed. by Hsin-Hsi Chen et al. ACM, July 2023, pp. 2594–2603. isbn: 9781450394086. doi: 10.1145/3539618.3591920.

[45] Jack W. Rae et al. **“Scaling Language Models: Methods, Analysis & Insights from Training Gopher”.** In: *arXiv e-prints*, arXiv:2112.11446 (Dec. 2021), arXiv:2112.11446. doi: 10.48550/arXiv.2112.11446. arXiv: 2112.11446 [cs.CL].

[46] Wikipedia contributors. ***SHA-1 — Wikipedia, The Free Encyclopedia*.** 2025. https://en.wikipedia.org/wiki/SHA-1.

[47] Wikipedia contributors. ***MD5 — Wikipedia, The Free Encyclopedia.*** 2025. https://en.wikipedia.org/wiki/MD5.

[48] xxHash authors. ***xxHash.*** 2025. https://xxhash.com.

[49] A.Z. Broder. **"On the resemblance and containment of documents".** In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. 1997, pp. 21–29. doi: 10.1109/SEQUEN.1997.666900.

[50] Jonathan Oliver, Chun Cheng, and Yanggui Chen. **"TLSH – A Locality Sensitive Hash".** In: *2013 Fourth Cybercrime and Trustworthy Computing Workshop*. 2013, pp. 7–13. doi: 10.1109/CTC.2013.9.

[51] Tim Jansen et al. **“Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data”.** In: *arXiv e-prints*, arXiv:2212.10440 (Dec. 2022), arXiv:2212.10440. doi: 10.48550/arXiv.2212.10440. arXiv: 2212.10440 [cs.CL].

[52] Jeffrey Li et al. **“DataComp-LM: In search of the next generation of training sets for language models”.** In: *arXiv e-prints*, arXiv:2406.11794 (June 2024), arXiv:2406.11794. doi: 10.48550/arXiv.2406.11794. arXiv: 2406.11794 [cs.LG].

References xx

[53] Dan Su et al. **“Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset”.** In: *arXiv e-prints*, arXiv:2412.02595 (Dec. 2024), arXiv:2412.02595. doi: 10.48550/arXiv.2412.02595. arXiv: 2412.02595 [cs.CL].

[54] Guilherme Penedo et al. **“The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale”.** In: *arXiv e-prints*, arXiv:2406.17557 (June 2024), arXiv:2406.17557. doi: 10.48550/arXiv.2406.17557. arXiv: 2406.17557 [cs.CL].

[55] Julia Kreutzer et al. **“Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets”.** In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by Brian Roark and Ani Nenkova, pp. 50–72. doi: 10.1162/tacl_a_00447. https://aclanthology.org/2022.tacl-1.4/.

[56] Laurie Burchell et al. **“An Open Dataset and Model for Language Identification”.** In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 865–879. doi: 10.18653/v1/2023.acl-short.75. https://aclanthology.org/2023.acl-short.75/.

[57] Amir Hossein Kargaran et al. **“GlotLID: Language Identification for Low-Resource Languages”.** In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6155–6218. doi: 10.18653/v1/2023.findings-emnlp.410. https://aclanthology.org/2023.findings-emnlp.410/.

[58] Isaac Caswell et al. **“Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus”.** In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6588–6608. doi: 10.18653/v1/2020.coling-main.579. https://aclanthology.org/2020.coling-main.579/.

[59] MHM Schellekens. **“Are internet robots adequately regulated?”** In: *Computer Law & Security Review* 29.6 (2013), pp. 666–675. doi: https://doi.org/10.1016/j.clsr.2013.09.003. https://www.sciencedirect.com/science/article/pii/S0267364913001659.

[60] Martijn Koster. ***A Standard for Robot Exclusion.*** 1995. https://www.robotstxt.org/.

[61] Martijn Koster. ***A method for web robots control.*** 1996. https://www.robotstxt.org/norobots-rfc.txt.

[62] Martijn Koster et al. ***Robots Exclusion Protocol.*** Internet-Draft draft-koster-rep-00. Work in Progress. Internet Engineering Task Force, July 2019. 10 pp. https://datatracker.ietf.org/doc/draft-koster-rep/00/.

[63] Henner Zeller, Lizzi Sassman, and Gary Illyes. ***Formalizing the robots exclusion protocol specification.*** 2019. https://developers.google.com/search/blog/2019/07/rep-id.

References xxv

[64] ***Google Robots.txt Parser and Matcher Library.***
https://github.com/google/robotstxt.

[65] Martijn Koster et al. ***Robots Exclusion Protocol.*** RFC 9309. Sept. 2022. doi: 10.17487/RFC9309. https://www.rfc-editor.org/info/rfc9309.

[66] ***Oracle Data Cloud Crawler.***
https://www.oracle.com/corporate/acquisitions/grapeshot/crawler.html.

[67] Y. Sun et al. **"Determining bias to search engines from robots.txt".** In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007*. 2007, pp. 149–155. doi: 10.1109/WI.2007.98.

References xxvi

[68] Santanu Kolay et al. **“A larger scale study of robots.txt”.** In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 1171–1172. https://dl.acm.org/doi/abs/10.1145/1367497.1367711.

[69] Greg Elmer. **“The spam book: On viruses, porn and other anomalies from the dark side of digital culture”.** In: ed. by Jussi Parikka and Tony D. Sampson. Creskill, New Jersey: Hampton Press, 2009. Chap. Robots.txt: The politics of search engine exclusion, pp. 217–227.

[70] Peter Henderson et al. ***Foundation Models and Fair Use.*** 2023. arXiv: 2303.15715 [cs.CY]. https://arxiv.org/abs/2303.15715.

[71] Shayne Longpre et al. ***Consent in Crisis: The Rapid Decline of the AI Data Commons.*** 2024. arXiv: 2407.14933 [cs.CL]. https://arxiv.org/abs/2407.14933.

[72] Enze Liu et al. ***Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers.*** 2024. arXiv: 2411.15091 [cs.HC]. https://arxiv.org/abs/2411.15091.

[73] ***Data Sets Containing Robots.txt Files and Non-200 Responses – Common Crawl.*** https://commoncrawl.org/2016/09/robotstxt-and-404-redirect-data-sets/.

[74] ***Host- and Domain-Level Web Graphs May, June/July and August 2022.*** https://commoncrawl.org/2022/09/host-and-domain-level-web-graphs-may-jun-aug-2022/.

[75] W3C. ***TDM Reservation Protocol (TDMRep).*** 2024. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/.

[76] Thom Vaughan. ***Vocabulary for Expressing Content Preferences for AI Training.*** Internet-Draft draft-vaughan-aipref-vocab-00. Work in Progress. Internet Engineering Task Force, Jan. 2025. 13 pp. https://datatracker.ietf.org/doc/draft-vaughan-aipref-vocab/00/.

[77] Kar Balan et al. **“DECORAIT - DECentralized Opt-in/out Registry for AI Training”.** In: *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production*. CVMP ’23. London, United Kingdom: Association for Computing Machinery, 2023. isbn: 9798400704260. doi: 10.1145/3626495.3626506. https://doi.org/10.1145/3626495.3626506.

[78] Wikipedia contributors. ***Automated Content Access Protocol — Wikipedia, The Free Encyclopedia.*** 2024. https://en.wikipedia.org/w/index.php?title=Automated_Content_Access_Protocol.

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

References xxx

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

Detecting Dutch TLDs With the Columnar Index

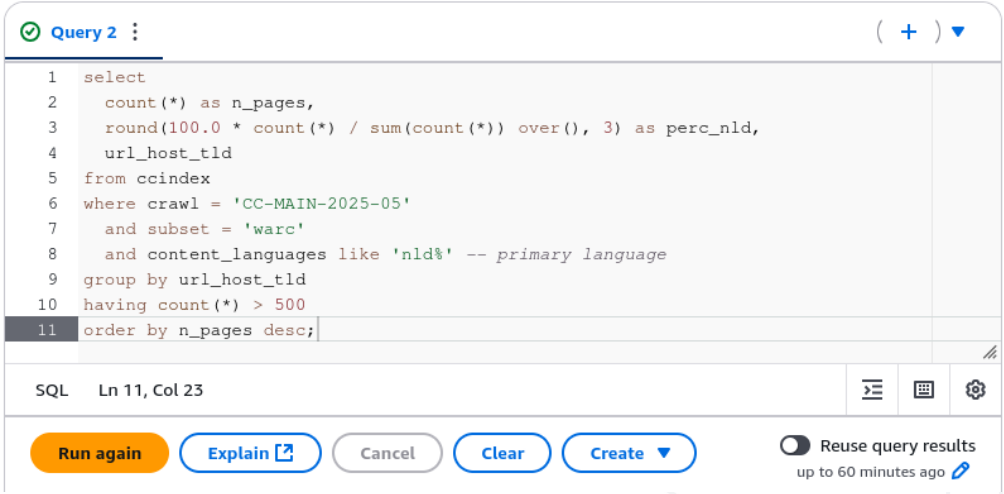
Data in the columnar index

- 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)

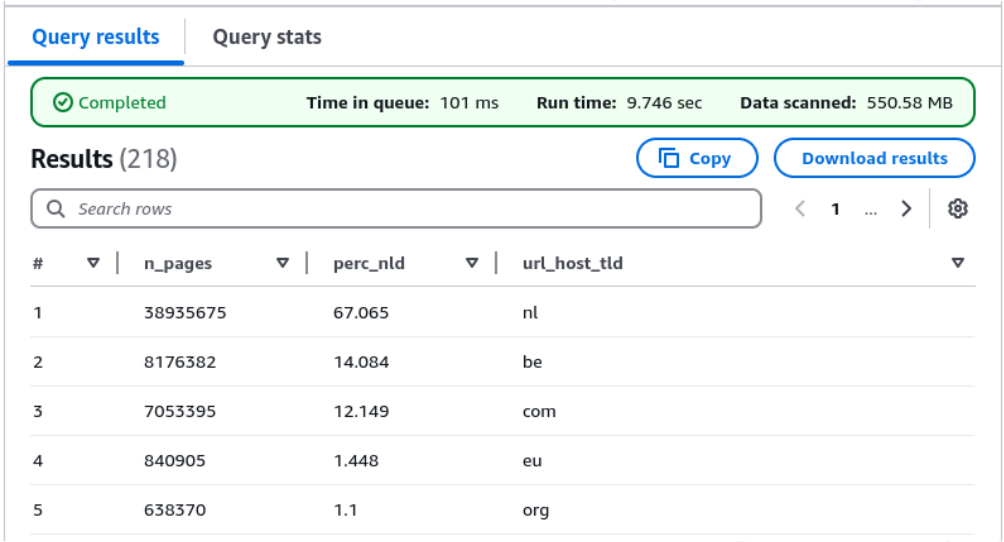
Instructions for setup and example queries in [14, 79]

The following example queries were run with Amazon Athena [80] engine v3, running on Trino (trino.io).

Detecting Dutch TLDs With the Columnar Index



Detecting Dutch TLDs With the Columnar Index



#	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

Detecting Dutch TLDs With the Columnar Index

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

Detecting Dutch TLDs With the Columnar Index

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

1. extract the primary language and other columns we need
2. count the total number of pages per TLD
3. select Dutch content, calculate the percentage and sort the result

Detecting Dutch TLDs With the Columnar Index

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

Detecting Dutch TLDs With the Columnar Index

```
16
17  select
18    n_pages,
19    round(100.0*n_pages/total_tld, 3) as perc_tld,
20    -- calculate the percentage of Dutch pages per TLD
21    round(100.0*n_pages/sum(n_pages) over (), 3) as perc_nld,
22    tld
23  from tmp2
24  where primary_language = 'nld'
25    and n_pages > 500
26  group by tld, n_pages, total_tld
27  order by perc_tld desc;
28
```

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

Detecting Dutch TLDs With the Columnar Index

#	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

Detecting Dutch TLDs With the Columnar Index

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

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