

C4 (Colossal Clean Crawled Corpus)

RANK	DOMAIN	CATEGORY	PERCENT OF ALL TOKENS
1	patents.google.com	Law & Government	0.46%
2	wikipedia.org	News & Media	0.19%
3	scribd.com	News & Media	0.07%
4	nytimes.com	News & Media	0.06%
5	journals.plos.org	Science & Health	0.06%
6	latimes.com	News & Media	0.05%
7	theguardian.com	News & Media	0.05%
8	forbes.com	News & Media	0.05%
9	huffpost.com	News & Media	0.04%
10	patents.com	Law & Government	0.04%
11	washingtonpost.com	News & Media	0.03%
12	coursera.org	Jobs & Education	0.03%
13	fool.com	Business & Industrial	0.03%
14	frontiersin.org	Science & Health	0.03%
15	instructables.com	Technology	0.03%

11,557	mainichi.jp	1.0M	0.0007%
24,508	asahi.com	610k	0.0004%
28,222	mofa.go.jp	550k	0.0003%
28,331	amazon.co.jp	540k	0.0003%
82,779	waseda.jp	230k	0.0001%
156,901	www3.nhk.or.jp	140k	0.00009%
3,405,988	chuo-u.ac.jp	4.8k	0.000003%

“See the websites that make AI bots like ChatGPT sound so smart - Washington Post.”
<https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/> (accessed Aug. 20, 2023).

ライブラリ
(スキップ)