

The ROOTS search tool - motivations and implementation

Aleksandra Piktus
piktus@huggingface.co

The ROOTS corpus was developed during the BigScience project with the purpose of training the multilingual, large language model—BLOOM. The **ROOTS search tool**—a search engine giving access to all document in the ROOTS corpus is available at <https://huggingface.co/spaces/bigscience-data/scisearch>. In this document we describe the motivations and technical details of the ROOTS search tool implementation.

1. Motivations

The current NLP landscape is dominated by research showcasing a spectacular and ever increasing boost in the performance of language models (LMs), initiated by the introduction of the transformer architecture [15]. Modern language models such as GPT-3 [4] and BLOOM [3] perform impressively well on generation tasks, often producing text indistinguishable from that written by a person. However, the inner workings of such systems are notoriously obscure, with the lack of visibility into their training corpora emerging as a common culprit. This state of affairs was the main motivator behind the BigScience project¹—a collaborative research endeavor initiated by Hugging Face, aiming to explore the area of large-scale language modelling through training an open-access, massively multilingual language model now known as BLOOM. We subscribe to this ethos and propose the ROOTS search tool—a search engine indexing all document in the ROOTS corpus [8]. We want to give researchers the ability to explore the documents forming ROOTS. We believe this is important due to reasons we outline below.

Data quality. The question of the quality of large scale textual datasets is a valid one in its own right, even before we start thinking about using them to train LMs. Papers often give only cursory descriptions of their data sources or how they were filtered to maintain a certain quality bar and curb undesired phenomena. A no-code tool allowing to browse through the corpus broadens the audience and can help us surface problematic instances and gain a better understanding of the overall text quality and potential corpus-level blindspots.

Model Interpretability. The ability to search through the training corpus of a language model helps explain and interpret its generations, e.g. it might allow us to uncover cases of generations directly memorized from the training data (BLOOM seems like a pretty good *Lorem ipsum* generator²) as well as cases of hallucinations where certain names or concepts don't appear in the training data at all.

¹<https://bigscience.huggingface.co/>

²See the BLOOM demo at https://huggingface.co/spaces/huggingface/bloom_demo

ROOTS language tag	# documents	Data size in GB	# passages	Index size in GB	Analyzer
zh, zhs, zht	88,814,841	168	111,284,681	682	zh
indic	84,982,982	68	100,810,124	95	whitespace
en	77,010,827	449	695,521,432	731	en
es	67,005,817	165	267,542,136	253	es
fr	58,847,091	195	299,938,546	292	fr
vi	34,110,375	41	76,164,552	70	whitespace
pt	31,969,891	74	122,221,863	115	pt
code	26,176,998	166	365,424,222	198	whitespace
ar	15,234,080	71	68,509,441	90	ar
id	12,514,253	19	29,531,873	26	id
ca	6,142,390	17	26,844,600	29	es
eu	5,149,797	2	6,219,039	4	whitespace
nigercongo	1,162,568	1	1,462,238	1	whitespace
total	597,936,751	1,436	2,171,474,747	2,586	

Table 1: Each row represents a single index we’ve built. The Data size column represents the size of compressed data and may not match numbers presented in the ROOTS paper [8].

Knowledge. The task of extracting and using facts contained in large scale, textual knowledge sources is what knowledge-intensive NLP has been attempting to solve with retriever-reader architectures. In a relatively recent trend, researchers have been looking at large scale LMs as *autoregressive* search engines able to surface the *knowledge* they acquire in the training process [12, 10, 2]. Papers in this niche mostly concern themselves with modelling challenges, the question we find more pressing though pertains to the quality of knowledge present in the training corpora - which can be explored using our tool.

Better tooling. Large scale text corpora pose many challenges to people who would like to access and analyze them. First their sheer scale (1.6TB in case of ROOTS) introduces resource requirements that may be hard to acquire by individual researchers, it also constitutes a barrier of entry for people unfamiliar with programmatic ways of accessing large scale data. Next come legal challenges—text, especially that scraped from the internet may be subject to copyright and licences which makes people who collect the data reluctant to re-releasing it for fear of infringing on any laws. As a result, corpora used to train modern LMs are rarely shared with the public. One common pattern is to open-source tools enabling the reproduction of the corpus rather than the corpus itself, e.g. in [16]. BigScience’s meticulous and principled approach to data governance [6] gives us a unique opportunity to work with actual data while following the guidelines of the responsible AI licence [5]. In this context, it is worth mentioning a related, although perhaps less principled effort which resulted in opening a C4 [13] search tool by AllenAI: <https://c4-search.apps.allenai.org/>.

2. Data

The ROOTS corpus [8] is a high-quality, heterogeneous and multilingual text corpus collected as part of the BigScience project to train BLOOM. The full dataset is open to the members of the BigScience Data organization on the Hugging Face hub. To inquire about gaining access to the organisation, please consult this Google Form. The ROOTS corpus consists of 1.6TB of textual data in 46 natural languages and 13 programming languages. The data is organized in

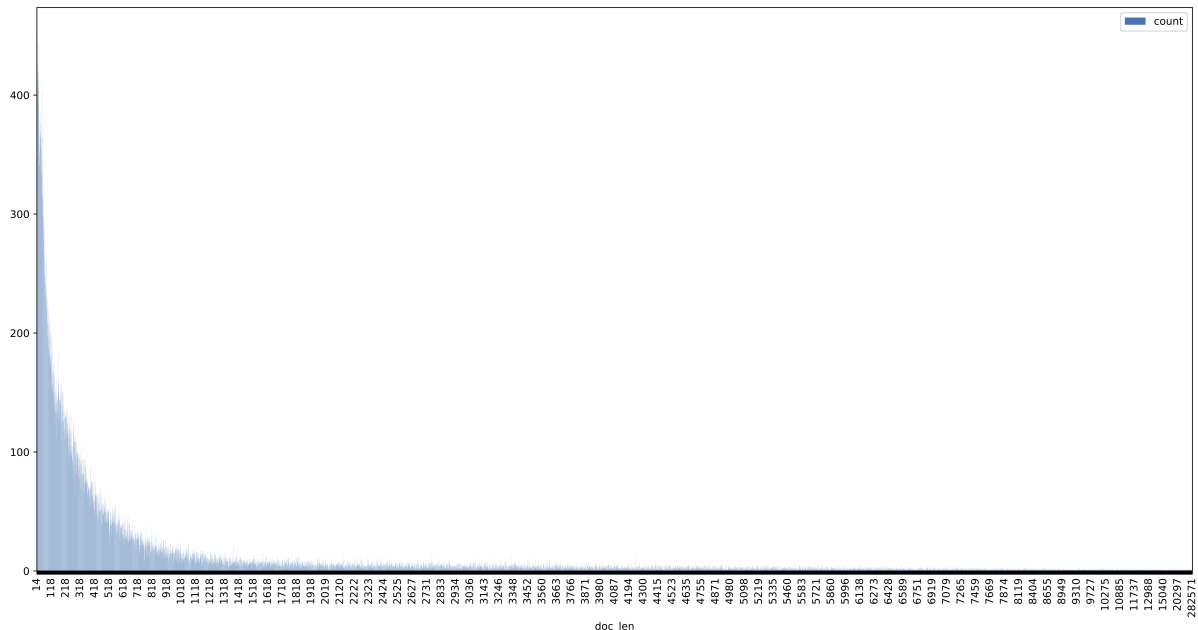


Figure 1: Document lengths distribution in a 100000-element random sample of the ROOTS corpus. The x-axis indicates the document length in characters. The y-axis indicates the number of documents of given length in the sample.

498 datasets, following a common naming pattern—the `bigscience-data` organization name followed by a slash, then the `roots` prefix, followed by the language tag and the data source identifier, delimited by underscores, e.g. `bigscience-data/roots_indic-mr_mkb`. There are two types of language identifiers—those indicating an individual language, e.g. `pt` for Portuguese, `vi` for Vietnamese, and those indicating a language belonging to a language group, e.g. `indic-mr` for Marathi, as part of the Indic language group. Additionally, all programming languages are collected under a common `code` language tag. In our experiments we build 13 indices—one for each individual language, with the exception of Chinese, where we collect 3 independent language tags available in ROOTS (`zh`, `zht`, `zhs`) into a single index; one for `code`, and one per language group (`indic` and `nigercongo`) In Table 1 we present basic information per index. The full list of indexed datasets is available in the appendix.

2.1. Data Governance

The notion of data governance was core to the development of ROOTS—we refer the reader to the FAccT paper summarizing core findings [6] and the preamble of our access form for a more thorough analysis of the guiding principles of the data governance workstream and how they were operationalized during the course of BigScience and after. Here, we briefly share the logic behind sharing the ROOTS search tool publicly. First and foremost, the tool gives no practical way to reconstruct the full corpus. We mitigate the risk of copyright infringements by only displaying 128-word snippets of indexed documents. Wherever possible, we link to source documents by displaying respective URLs (that said, metadata in ROOTS is inconsistent and we only have access to URLs in `pseudocrawl` datasets). Additionally, for each displayed result we provide a link to the ROOTS dataset it came from for further inspection.

2.2. Data preprocessing

Datapoint vs document. The ROOTS corpus was design with the language modeling task in mind. In such a setup, we’re primarily concerned with text understood as stream of words

Document ID: bigscience-data/roots_fr_hal_archives_ouvertes/591107?seg=para_128_8&seg_id=41
 Language: fr
 artystów po roku 1989, zafascynowanego monumentalnymi instalacjami i malarską nową figuracją, która zdominowała scenę artystyczną po przywróceniu ustroju demokratycznego w Polsce; artyści tej generacji wiązali nadzieje na przyciągnięcie do Polski kapitalistycznego rynku sztuki z dziełem jako przedmiotem cennym, często prowokującym, ale i kiczowatym. To surowe stwierdzenie, siłą rzeczy niesprawiedliwe w stosunku do paru artystów dalekich od tego spektakularnego podejścia, nie jest tutaj sądem wartościującym, lecz ikonologiczną refleksją w stylu Erwina Panofsky’ego – nie można było dostrzegać takich wytworów artystycznych jak Petasza w kontekście **polskim** przełomu dwudziestego i dwudziestego pierwszego wieku. Cóż bowiem bardziej odległego od tych aspiracji niż numery Commonpress, drukowane – najczęściej za granicą – na owych kserokopiarkach dawnej generacji, szarawe i bez kontrastów, na wskroś antyestetyczne, spięte zszywkami, których jedynym żywym, czyli kolorowym, akcentem bywała czasami taśma klejąca, wzmacniająca prymitywne życie karteek. Na początku polskiego dwudziestego pierwszego stulecia nie było już śladu zamierzonego „ubóstwa” sztuki, wyznawanego przez artystów po roku 1968: Leszka Przyjemskiego z jego politycznymi

Document ID: bigscience-data/roots_fr_hal_archives_ouvertes/233166?seg=para_128_8&seg_id=239
 Language: fr
 kapitalistycznego rynku sztuki z dziełem jako przedmiotem cennym, często prowokującym, ale i kiczowatym. To surowe stwierdzenie, siłą rzeczy niesprawiedliwe w stosunku do paru artystów dalekich od tego spektakularnego podejścia, nie jest tutaj sądem wartościującym, lecz ikonologiczną refleksją w stylu Erwina Panofsky’ego – nie można było dostrzegać takich wytworów artystycznych jak Petasza w kontekście **polskim** przełomu dwudziestego i dwudziestego pierwszego wieku. Cóż bowiem bardziej odległego od tych aspiracji niż numery Commonpress, drukowane – najczęściej za granicą – na owych kserokopiarkach dawnej generacji, szarawe i bez kontrastów, na wskroś antyestetyczne, spięte zszywkami, których jedynym żywym, czyli kolorowym, akcentem bywała czasami taśma klejąca, wzmacniająca prymitywne życie karteek. Na początku polskiego dwudziestego pierwszego stulecia nie było już śladu zamierzonego „ubóstwa” sztuki, wyznawanego przez artystów po roku 1968: Leszka Przyjemskiego z jego politycznymi

Document ID: bigscience-data/roots_fr_oscar/13723909?seg=para_128_8&seg_id=1
 Language: fr
 sur le configurateur Tesla : et de personnaliser votre Tesla Model S ou Tesla Model X. Pour acheter une Tesla, vous n'aurez pas forcément besoin d'un magasin. Il vous suffit de vous rendre sur le configurateur Tesla : En plus des magasins Tesla en France, vous pourrez rencontrer un représentant commercial sur tout le territoire. Vous souhaitez acheter des protections d'écrans, une console centrale ou un protège clef pour votre Tesla Model S ou Tesla Model X? Nove samochody marki Tesla na **polskim** rynku to oferta obejmujÁ ca 3 modele dostÁ pne we wszystkich salonach dealerskich w cenie od 195 490 PLN (za wersjÁ Tesla Model 3 (409 KM) A1 z silnikiem elektrycznym o mocy 409KM i automatycznÁ skrzyniÁ biegów). Modele zawierajÁ ce oferty promocyjne i wyprzedaÁ%owe zostÁ y oznaczone kolorem czerwonym.

Figure 3: Language contamination - we issue a query against the French index, however, the results are predominantly in Polish.

3. Findings

As stated above, the main motivation for creating the ROOTS search engine is to allow researchers to gain insights into the datasets used to train BLOOM. Some findings pertaining to PII leakage and language contamination that we’ve observed when testing the tool can be examined in Figure 2 and figure 3. We have also observed evidence of low quality text, e.g. racist slurs, sexually explicit language or hate speech—interestingly, they often come from datasets containing movie subtitles. On one hand, the nature of the data source introduces an extra layer of parenthesis around the controversial content—as readers we understand movie subtitles are less reliable as a knowledge source and may be consciously breaking standards of respectable communication. At the same time, a language model trained on this data isn’t provided with any metadata which would allow it to classify such datapoints as coming from a specific context and therefore make any judgements on the quality of the source.

4. Implementation

We host 13 sparse, BM25 [14] indices, built for all main languages and groups of languages in the ROOTS corpus using Pyserini [9]. We host the index server on a GCP instance. We use native Lucene⁴ analyzers available via Pyserini API to perform text tokenization. Table 1 contains information on data and analyzers used to tokenize the text. We index the data using the following sample command.

```

1 python -m pyserini.index.lucene
2     --collection JsonCollection \
3     --input bigscience-data-pyserini/en/ \
4     --index bigscience-data-index/en/ \
5     --generator DefaultLuceneDocumentGenerator \
6     --threads 64 \
7     --storePositions \
8     --storeDocvectors \
9     --storeRaw \
10    --language en

```

The ROOTS search tool UI allows people to search in a specific language, in all languages (then results are surfaced separately for each language) and to detect the query language on the backend side and then serve results accordingly. The language identification logic uses a FastText-based classifier [7] available on the Hugging Face hub. The user can control how many results they want to see. We apply PII redaction to all datasets on the backend side. We surface links to external URLs whenever available, and links to respective ROOTS dataset on

⁴<https://lucene.apache.org/>

the Hugging Face hub for each result. We enable flagging of suspicious results—in such case, we store the reports in the backend and may use them to inform future data filtering efforts. We use Gradio [1] frontend served via Hugging Face Spaces to host the demo. The front end code can be accessed here, the Spaces demo is available here and the Community is open for discussions here.

5. Next steps

The current tool is heavily influenced by the UX of popular search engines. In the future we intend to review classic corpus analysis tools for ideas of different possible treatments. We would like to add more quantitative information, e.g. term frequency information, number of hits, co-occurrence statistics and others. We’re also review a possibility of adding an exact search retrieval (potentially supported by different backend). We welcome the feedback and suggestions from discussions in the Community tab of the demo. We’re also pursuing a spin-off collaboration with Pyserini to make large scale indexing and hosting of textual data even more seamless.

6. Acknowledgements

All throughout building the tool, the project was supported by the amazing folks at the Big-Science Data Visualization Working Group, especially Christopher Akiki, Anna Rogers, Gerard Dupont and Yacine Jernite. The project wouldn’t be the same without insights from the Pyserini team—Jimmy Lin, Odunayo Ogundepo and Xinyu (Crystina) Zhang. Big thanks to the Gradio team, especially Pete Allen for their support on the frontend side, and the Hugging Face infra team for answering my questions regarding hosting the tool.

References

- [1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.
- [2] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers, 2022.
- [3] BigScience Workshop. Bloom (revision 4ab0472), 2022.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [5] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, page 778–788, New York, NY, USA, 2022. Association for Computing Machinery.

- [6] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 2206–2222, New York, NY, USA, 2022. Association for Computing Machinery.
- [7] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [8] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [9] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.
- [10] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum*, 55(1), jul 2021.
- [11] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache.
- [12] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [14] S. Robertson. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [16] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.

A. Appendix

A list of all datasets available on the Hugging Face hub, indexed and available in the ROOTS search tool.

```
1 bigscience-data/roots_ar_arabench
2 bigscience-data/roots_ar_arabic_billion_words
3 bigscience-data/roots_ar_brad_2
4 bigscience-data/roots_ar_habibi
5 bigscience-data/roots_ar_kalimat
6 bigscience-data/roots_ar_ksucca
7 bigscience-data/roots_ar_labr
8 bigscience-data/roots_ar_multi_un_2
9 bigscience-data/roots_ar_open_subtitles
10 bigscience-data/roots_ar_openiti_proc
11 bigscience-data/roots_ar_opus100
12 bigscience-data/roots_ar_oscar
13 bigscience-data/roots_ar_pseudocrawl-filtered_595_mawdoo3_com
14 bigscience-data/roots_ar_qedcorpus
15 bigscience-data/roots_ar_sanad
16 bigscience-data/roots_ar_tashkeela
17 bigscience-data/roots_ar_ted_talks_iwslt
18 bigscience-data/roots_ar_uncorpus
19 bigscience-data/roots_ar_wikibooks
20 bigscience-data/roots_ar_wikinews
21 bigscience-data/roots_ar_wikipedia
22 bigscience-data/roots_ar_wikiquote
23 bigscience-data/roots_ar_wikisource
24 bigscience-data/roots_ar_wikiversity
25 bigscience-data/roots_ar_wiktory
26 bigscience-data/roots_ca_catalan_general_crawling
27 bigscience-data/roots_ca_catalan_government_crawling
28 bigscience-data/roots_ca_catalan_textual_corpus
29 bigscience-data/roots_ca_enriched_conllu_ancora_for_ml_training
30 bigscience-data/roots_ca_open_subtitles
31 bigscience-data/roots_ca_opus100
32 bigscience-data/roots_ca_oscar
33 bigscience-data/roots_ca_parlament_parla
34 bigscience-data/roots_ca_tecla
35 bigscience-data/roots_ca_ted_talks_iwslt
36 bigscience-data/roots_ca_vilaquad
37 bigscience-data/roots_ca_viquiquad
38 bigscience-data/roots_ca_wikibooks
39 bigscience-data/roots_ca_wikimedia
40 bigscience-data/roots_ca_wikinews
41 bigscience-data/roots_ca_wikipedia
42 bigscience-data/roots_ca_wikiquote
43 bigscience-data/roots_ca_wikisource
44 bigscience-data/roots_ca_wiktory_filtered
45 bigscience-data/roots_ca_xquad_ca
46 bigscience-data/roots_code_github
47 bigscience-data/roots_code_stackexchange
48 bigscience-data/roots_en_book_dash_books
49 bigscience-data/roots_en_multi_un_2
50 bigscience-data/roots_en_no_code_stackexchange
51 bigscience-data/roots_en_odiencorp
52 bigscience-data/roots_en_open_subtitles
53 bigscience-data/roots_en_oscar
54 bigscience-data/roots_en_project_gutenberg
55 bigscience-data/roots_en_pseudocrawl-filtered_159_www_postcrescent_com
56 bigscience-data/roots_en_pseudocrawl-filtered_304_www_semana_com
57 bigscience-data/roots_en_pseudocrawl-filtered_339_www_actasanitaria_com
58 bigscience-data/roots_en_pseudocrawl-filtered_395_www_evwind_es
```

59 bigscience-data/roots_en_pseudocrawl-filtered_470_forums_hardwarezone_com_sg
60 bigscience-data/roots_en_pseudocrawl-filtered_483_alvinology_com
61 bigscience-data/roots_en_pseudocrawl-filtered_485_blog_moneysmart_sg
62 bigscience-data/roots_en_pseudocrawl-filtered_487_thesmartlocal_com
63 bigscience-data/roots_en_pseudocrawl-filtered_488_dailyvanity_sg
64 bigscience-data/roots_en_pseudocrawl-filtered_492_www_vivawoman_net
65 bigscience-data/roots_en_pseudocrawl-filtered_497_www_straitstimes_com
66 bigscience-data/roots_en_pseudocrawl-filtered_498_www_channelnewsasia_com
67 bigscience-data/roots_en_pseudocrawl-filtered_499_www_today_com_news
68 bigscience-data/roots_en_pseudocrawl-filtered_500_www_asiaone_com_singapore
69 bigscience-data/roots_en_pseudocrawl-filtered_501_theindependent_sg
70 bigscience-data/roots_en_pseudocrawl-filtered_502_www_ricemedia_co
71 bigscience-data/roots_en_pseudocrawl-filtered_510_timesofindia_indiatimes_com
72 bigscience-data/roots_en_pseudocrawl-filtered_534_www_nairaland_com
73 bigscience-data/roots_en_pseudocrawl-filtered_548_remezcla_com
74 bigscience-data/roots_en_pseudocrawl-filtered_638_globalvoices_org
75 bigscience-data/roots_en_pseudocrawl-filtered_689_www_abc_net_au
76 bigscience-data/roots_en_pseudocrawl-filtered_696_www_oercommons_org
77 bigscience-data/roots_en_qedcorpus
78 bigscience-data/roots_en_royal_society_corpus
79 bigscience-data/roots_en_s2orc_ai2_pdf_parses
80 bigscience-data/roots_en_scielo
81 bigscience-data/roots_en_ted_talks_iwslt
82 bigscience-data/roots_en_the_pile_europarl
83 bigscience-data/roots_en_the_pile_uspto
84 bigscience-data/roots_en_uncorpus
85 bigscience-data/roots_en_wikibooks
86 bigscience-data/roots_en_wikinews
87 bigscience-data/roots_en_wikipedia
88 bigscience-data/roots_en_wikiquote
89 bigscience-data/roots_en_wikiversity
90 bigscience-data/roots_en_wikivoyage
91 bigscience-data/roots_en_wiktictionary
92 bigscience-data/roots_es_multi_un_2
93 bigscience-data/roots_es_open_subtitles
94 bigscience-data/roots_es_oscar
95 bigscience-data/roots_es_project_gutenberg
96 bigscience-data/roots_es_pseudocrawl-filtered_100_www_aporrea_org
97 bigscience-data/roots_es_pseudocrawl-filtered_103_www_elmostrador_cl
98 bigscience-data/roots_es_pseudocrawl-filtered_116_www_latribuna_hn
99 bigscience-data/roots_es_pseudocrawl-filtered_118_www_elheraldo_hn
100 bigscience-data/roots_es_pseudocrawl-filtered_125_www_noticiasde_es
101 bigscience-data/roots_es_pseudocrawl-filtered_130_www_elperiodicomediterraneo_com
102 bigscience-data/roots_es_pseudocrawl-filtered_136_valenciaplaza_com
103 bigscience-data/roots_es_pseudocrawl-filtered_146_www_perfil_com
104 bigscience-data/roots_es_pseudocrawl-filtered_153_financialfood_es
105 bigscience-data/roots_es_pseudocrawl-filtered_157_www_elsoldemexico_com_mx
106 bigscience-data/roots_es_pseudocrawl-filtered_158_www_diariodeleon_es
107 bigscience-data/roots_es_pseudocrawl-filtered_165_www_ticbeat_com
108 bigscience-data/roots_es_pseudocrawl-filtered_167_www_ambientum_com
109 bigscience-data/roots_es_pseudocrawl-filtered_169_www_el_carabobeno_com
110 bigscience-data/roots_es_pseudocrawl-filtered_172_www_rionegro_com_ar
111 bigscience-data/roots_es_pseudocrawl-filtered_181_noticiassin_com
112 bigscience-data/roots_es_pseudocrawl-filtered_182_correodelsur_com
113 bigscience-data/roots_es_pseudocrawl-filtered_189_www_eleconomista_com_mx
114 bigscience-data/roots_es_pseudocrawl-filtered_198_www_eleconomista_es
115 bigscience-data/roots_es_pseudocrawl-filtered_203_www_que_es
116 bigscience-data/roots_es_pseudocrawl-filtered_207_elimpulso_com
117 bigscience-data/roots_es_pseudocrawl-filtered_209_misionesonline_net
118 bigscience-data/roots_es_pseudocrawl-filtered_20_www_clarin_com
119 bigscience-data/roots_es_pseudocrawl-filtered_211_www_elcomercio_com
120 bigscience-data/roots_es_pseudocrawl-filtered_213_www_hola_com

121 bigscience-data/roots_es_pseudocrawl-filtered_215_www_lainformacion_com
122 bigscience-data/roots_es_pseudocrawl-filtered_219_www_aguasresiduales_info
123 bigscience-data/roots_es_pseudocrawl-filtered_21_www_elperiodicodearagon_com
124 bigscience-data/roots_es_pseudocrawl-filtered_220_www_vanguardia_com_mx
125 bigscience-data/roots_es_pseudocrawl-filtered_223_www_eltambor_es
126 bigscience-data/roots_es_pseudocrawl-filtered_226_www_ole_com_ar
127 bigscience-data/roots_es_pseudocrawl-filtered_229_www_expansion_com
128 bigscience-data/roots_es_pseudocrawl-filtered_231_ojo_pe
129 bigscience-data/roots_es_pseudocrawl-filtered_232_tn_com_ar
130 bigscience-data/roots_es_pseudocrawl-filtered_233_www_dinero_com
131 bigscience-data/roots_es_pseudocrawl-filtered_237_www_cronista_com
132 bigscience-data/roots_es_pseudocrawl-filtered_23_www_elconfidencialdigital_com
133 bigscience-data/roots_es_pseudocrawl-filtered_244_www_df_cl
134 bigscience-data/roots_es_pseudocrawl-filtered_245_www_noticiasdenavarra_com
135 bigscience-data/roots_es_pseudocrawl-filtered_246_www_eldiarionuevodia_com_ar
136 bigscience-data/roots_es_pseudocrawl-filtered_248_www_telesurtv_net
137 bigscience-data/roots_es_pseudocrawl-filtered_249_www_telecinco_es
138 bigscience-data/roots_es_pseudocrawl-filtered_250_www_cooperativa_cl
139 bigscience-data/roots_es_pseudocrawl-filtered_253_www_debate_com_mx
140 bigscience-data/roots_es_pseudocrawl-filtered_254_diario_mx
141 bigscience-data/roots_es_pseudocrawl-filtered_255_elcomercio_pe
142 bigscience-data/roots_es_pseudocrawl-filtered_256_www_laprovincia_es
143 bigscience-data/roots_es_pseudocrawl-filtered_257_www_diaridetarragona_com
144 bigscience-data/roots_es_pseudocrawl-filtered_263_www_lasexta_com
145 bigscience-data/roots_es_pseudocrawl-filtered_267_www_elperiodico_com_es
146 bigscience-data/roots_es_pseudocrawl-filtered_276_radio_uchile_cl
147 bigscience-data/roots_es_pseudocrawl-filtered_277_www_entornointeligente_com
148 bigscience-data/roots_es_pseudocrawl-filtered_280_salamancartvaldia_es
149 bigscience-data/roots_es_pseudocrawl-filtered_286_www_nacion_com
150 bigscience-data/roots_es_pseudocrawl-filtered_287_www_cibercuba_com
151 bigscience-data/roots_es_pseudocrawl-filtered_288_www_marca_com
152 bigscience-data/roots_es_pseudocrawl-filtered_28_www_fayerwayer_com
153 bigscience-data/roots_es_pseudocrawl-filtered_294_www_laopinion_com_co
154 bigscience-data/roots_es_pseudocrawl-filtered_299_www_lne_es
155 bigscience-data/roots_es_pseudocrawl-filtered_30_www_radiocable_com
156 bigscience-data/roots_es_pseudocrawl-filtered_315_lasillavacia_com
157 bigscience-data/roots_es_pseudocrawl-filtered_317_diariocorreo_pe
158 bigscience-data/roots_es_pseudocrawl-filtered_320_www_paginasiete_bo
159 bigscience-data/roots_es_pseudocrawl-filtered_324_gestion_pe
160 bigscience-data/roots_es_pseudocrawl-filtered_325_www_laprensa_hn
161 bigscience-data/roots_es_pseudocrawl-filtered_32_www_elexpresso_com
162 bigscience-data/roots_es_pseudocrawl-filtered_333_www_elmundo_es
163 bigscience-data/roots_es_pseudocrawl-filtered_341_es_cointelegraph_com
164 bigscience-data/roots_es_pseudocrawl-filtered_349_www_eltiempo_com
165 bigscience-data/roots_es_pseudocrawl-filtered_34_www_losandes_com_ar
166 bigscience-data/roots_es_pseudocrawl-filtered_354_www_lagaceta_com_ar
167 bigscience-data/roots_es_pseudocrawl-filtered_359_www_efeverde_com
168 bigscience-data/roots_es_pseudocrawl-filtered_367_elcorreoweb_es
169 bigscience-data/roots_es_pseudocrawl-filtered_373_www_farodevigo_es
170 bigscience-data/roots_es_pseudocrawl-filtered_374_www_talcualdigital_com
171 bigscience-data/roots_es_pseudocrawl-filtered_376_www_elpopular_com_ar
172 bigscience-data/roots_es_pseudocrawl-filtered_381_www_cuartopoder_es
173 bigscience-data/roots_es_pseudocrawl-filtered_386_www_prensalibre_com
174 bigscience-data/roots_es_pseudocrawl-filtered_392_www_muypymes_com
175 bigscience-data/roots_es_pseudocrawl-filtered_396_www_eldiario_es
176 bigscience-data/roots_es_pseudocrawl-filtered_401_www_elperiodicodemexico_com
177 bigscience-data/roots_es_pseudocrawl-filtered_404_www_telam_com_ar
178 bigscience-data/roots_es_pseudocrawl-filtered_405_www_emol_com
179 bigscience-data/roots_es_pseudocrawl-filtered_406_www_americaeconomia_com
180 bigscience-data/roots_es_pseudocrawl-filtered_409_www_proceso_com_mx
181 bigscience-data/roots_es_pseudocrawl-filtered_417_www_radiolaprimerisima_com
182 bigscience-data/roots_es_pseudocrawl-filtered_420_www_retema_es
183 bigscience-data/roots_es_pseudocrawl-filtered_422_www_formulatv_com

184 bigscience-data/roots_es_pseudocrawl-filtered_424_www_lavanguardia_com
185 bigscience-data/roots_es_pseudocrawl-filtered_429_cadenaser_com
186 bigscience-data/roots_es_pseudocrawl-filtered_430_www_eldiario_ec
187 bigscience-data/roots_es_pseudocrawl-
filtered_431_www_elperiodicoextremadura_com
188 bigscience-data/roots_es_pseudocrawl-filtered_44_ladiaria_com_uy
189 bigscience-data/roots_es_pseudocrawl-filtered_518_www_elcolombiano_com
190 bigscience-data/roots_es_pseudocrawl-filtered_53_www_expreso_ec
191 bigscience-data/roots_es_pseudocrawl-filtered_56_www_eluniverso_com
192 bigscience-data/roots_es_pseudocrawl-filtered_58_www_levante_emv_com
193 bigscience-data/roots_es_pseudocrawl-filtered_62_www_lapagina_com_sv
194 bigscience-data/roots_es_pseudocrawl-filtered_63_www_lanacion_com_ar
195 bigscience-data/roots_es_pseudocrawl-filtered_641_es_globalvoices_org
196 bigscience-data/roots_es_pseudocrawl-filtered_675_www_espectador_com
197 bigscience-data/roots_es_pseudocrawl-filtered_67_www_elpais_cr
198 bigscience-data/roots_es_pseudocrawl-filtered_71_www_rtve_es
199 bigscience-data/roots_es_pseudocrawl-filtered_78_www_listindiario_com
200 bigscience-data/roots_es_pseudocrawl-filtered_79_www_laopiniondemercia_es
201 bigscience-data/roots_es_pseudocrawl-filtered_86_www_motorpasion_com
202 bigscience-data/roots_es_pseudocrawl-filtered_90_peru_com
203 bigscience-data/roots_es_pseudocrawl-filtered_91_www_diario26_com
204 bigscience-data/roots_es_qedcorpus
205 bigscience-data/roots_es_scielo
206 bigscience-data/roots_es_ted_talks_iwslt
207 bigscience-data/roots_es_the_pile_europarl
208 bigscience-data/roots_es_uncorpus
209 bigscience-data/roots_es_wikibooks
210 bigscience-data/roots_es_wikinews
211 bigscience-data/roots_es_wikipedia
212 bigscience-data/roots_es_wikiquote
213 bigscience-data/roots_es_wikisource
214 bigscience-data/roots_es_wikiversity
215 bigscience-data/roots_es_wikivoyage
216 bigscience-data/roots_es_wiktionary
217 bigscience-data/roots_eu_bsbasque
218 bigscience-data/roots_eu_open_subtitles
219 bigscience-data/roots_eu_opus100
220 bigscience-data/roots_eu_oscar
221 bigscience-data/roots_eu_pseudocrawl-filtered_506_goiena_eus
222 bigscience-data/roots_eu_pseudocrawl-filtered_563_ahotsak_eus
223 bigscience-data/roots_eu_pseudocrawl-filtered_635_www_berria_eus
224 bigscience-data/roots_eu_pseudocrawl-filtered_637_www_argia_eus
225 bigscience-data/roots_eu_ted_talks_iwslt
226 bigscience-data/roots_eu_wikibooks
227 bigscience-data/roots_eu_wikipedia
228 bigscience-data/roots_eu_wikiquote
229 bigscience-data/roots_eu_wikisource
230 bigscience-data/roots_eu_wiktionary
231 bigscience-data/roots_fr_book_dash_books
232 bigscience-data/roots_fr_ester
233 bigscience-data/roots_fr_hal_archives_ouvertes
234 bigscience-data/roots_fr_multi_un_2
235 bigscience-data/roots_fr_open_subtitles
236 bigscience-data/roots_fr_oscar
237 bigscience-data/roots_fr_project_gutenberg
238 bigscience-data/roots_fr_pseudocrawl-filtered_530_www_mediapart_fr
239 bigscience-data/roots_fr_pseudocrawl-filtered_550_www_lemonde_fr
240 bigscience-data/roots_fr_pseudocrawl-filtered_599_fr_globalvoices_org
241 bigscience-data/roots_fr_qedcorpus
242 bigscience-data/roots_fr_ted_talks_iwslt
243 bigscience-data/roots_fr_the_pile_europarl
244 bigscience-data/roots_fr_uncorpus
245 bigscience-data/roots_fr_wikibooks

246 bigscience-data/roots_fr_wikinews
247 bigscience-data/roots_fr_wikipedia
248 bigscience-data/roots_fr_wikiquote
249 bigscience-data/roots_fr_wikisource
250 bigscience-data/roots_fr_wikiversity
251 bigscience-data/roots_fr_wikivoyage
252 bigscience-data/roots_fr_wiktionary
253 bigscience-data/roots_id_indo4b_bppt
254 bigscience-data/roots_id_indo4b_jw300
255 bigscience-data/roots_id_indo4b_kompas
256 bigscience-data/roots_id_indo4b_parallel
257 bigscience-data/roots_id_indo4b_talpc
258 bigscience-data/roots_id_indo4b_tempo
259 bigscience-data/roots_id_indonesian_frog_storytelling_corpus
260 bigscience-data/roots_id_indonesian_news_articles_2017
261 bigscience-data/roots_id_indonesian_news_corpus
262 bigscience-data/roots_id_indonli
263 bigscience-data/roots_id_indosum
264 bigscience-data/roots_id_open_subtitles
265 bigscience-data/roots_id_opus100
266 bigscience-data/roots_id_oscar
267 bigscience-data/roots_id_pseudocrawl-filtered_512_kumparan_com
268 bigscience-data/roots_id_pseudocrawl-filtered_545_www_detik_com
269 bigscience-data/roots_id_pseudocrawl-filtered_549_www_cnnindonesia_com
270 bigscience-data/roots_id_pseudocrawl-filtered_572_tirto_id
271 bigscience-data/roots_id_recibrew
272 bigscience-data/roots_id_ted_talks_iwslt
273 bigscience-data/roots_id_wikibooks
274 bigscience-data/roots_id_wikimedia
275 bigscience-data/roots_id_wikipedia
276 bigscience-data/roots_id_wikiquote
277 bigscience-data/roots_id_wikisource
278 bigscience-data/roots_id_wiktionary
279 bigscience-data/roots_indic-as_opus100
280 bigscience-data/roots_indic-as_samanantar
281 bigscience-data/roots_indic-as_ted_talks_iwslt
282 bigscience-data/roots_indic-as_wikipedia
283 bigscience-data/roots_indic-as_wikisource
284 bigscience-data/roots_indic-as_wiktionary
285 bigscience-data/roots_indic-bn_bangla_lm
286 bigscience-data/roots_indic-bn_bangla_sentiment_classification_datasets
287 bigscience-data/roots_indic-bn_bengali_question_answering
288 bigscience-data/roots_indic-bn_indic_nlp_corpus
289 bigscience-data/roots_indic-bn_mkb
290 bigscience-data/roots_indic-bn_open_subtitles
291 bigscience-data/roots_indic-bn_opus100
292 bigscience-data/roots_indic-bn_oscar
293 bigscience-data/roots_indic-bn_pib
294 bigscience-data/roots_indic-bn_samanantar
295 bigscience-data/roots_indic-bn_ted_talks_iwslt
296 bigscience-data/roots_indic-bn_wikibooks
297 bigscience-data/roots_indic-bn_wikipedia
298 bigscience-data/roots_indic-bn_wikisource
299 bigscience-data/roots_indic-bn_wikivoyage
300 bigscience-data/roots_indic-bn_wiktionary
301 bigscience-data/roots_indic-gu_indic_nlp_corpus
302 bigscience-data/roots_indic-gu_mkb
303 bigscience-data/roots_indic-gu_opus100
304 bigscience-data/roots_indic-gu_pib
305 bigscience-data/roots_indic-gu_samanantar
306 bigscience-data/roots_indic-gu_ted_talks_iwslt
307 bigscience-data/roots_indic-gu_wikipedia
308 bigscience-data/roots_indic-gu_wikiquote

309 bigscience-data/roots_indic-gu_wikisource
310 bigscience-data/roots_indic-gu_wiktionary
311 bigscience-data/roots_indic-hi_iitb_english_hindi_corpus
312 bigscience-data/roots_indic-hi_indic_nlp_corpus
313 bigscience-data/roots_indic-hi_mkb
314 bigscience-data/roots_indic-hi_open_subtitles
315 bigscience-data/roots_indic-hi_opus100
316 bigscience-data/roots_indic-hi_oscar
317 bigscience-data/roots_indic-hi_pib
318 bigscience-data/roots_indic-hi_pseudocrawl-filtered_515_www_aajtak_in
319 bigscience-data/roots_indic-hi_pseudocrawl-filtered_667_www_bhaskar_com
320 bigscience-data/roots_indic-hi_qedcorpus
321 bigscience-data/roots_indic-hi_samanantar
322 bigscience-data/roots_indic-hi_ted_talks_iwslt
323 bigscience-data/roots_indic-hi_wikibooks
324 bigscience-data/roots_indic-hi_wikimedia
325 bigscience-data/roots_indic-hi_wikipedia
326 bigscience-data/roots_indic-hi_wikiquote
327 bigscience-data/roots_indic-hi_wikisource
328 bigscience-data/roots_indic-hi_wikiversity
329 bigscience-data/roots_indic-hi_wikivoyage
330 bigscience-data/roots_indic-hi_wiktionary
331 bigscience-data/roots_indic-kn_indic_nlp_corpus
332 bigscience-data/roots_indic-kn_opus100
333 bigscience-data/roots_indic-kn_samanantar
334 bigscience-data/roots_indic-kn_ted_talks_iwslt
335 bigscience-data/roots_indic-kn_wikipedia
336 bigscience-data/roots_indic-kn_wikiquote
337 bigscience-data/roots_indic-kn_wikisource
338 bigscience-data/roots_indic-kn_wiktionary
339 bigscience-data/roots_indic-ml_indic_nlp_corpus
340 bigscience-data/roots_indic-ml_mkb
341 bigscience-data/roots_indic-ml_open_subtitles
342 bigscience-data/roots_indic-ml_opus100
343 bigscience-data/roots_indic-ml_pib
344 bigscience-data/roots_indic-ml_samanantar
345 bigscience-data/roots_indic-ml_ted_talks_iwslt
346 bigscience-data/roots_indic-ml_wikibooks
347 bigscience-data/roots_indic-ml_wikipedia
348 bigscience-data/roots_indic-ml_wikiquote
349 bigscience-data/roots_indic-ml_wikisource
350 bigscience-data/roots_indic-ml_wiktionary
351 bigscience-data/roots_indic-mr_indic_nlp_corpus
352 bigscience-data/roots_indic-mr_mkb
353 bigscience-data/roots_indic-mr_opus100
354 bigscience-data/roots_indic-mr_pib
355 bigscience-data/roots_indic-mr_samanantar
356 bigscience-data/roots_indic-mr_ted_talks_iwslt
357 bigscience-data/roots_indic-mr_wikibooks
358 bigscience-data/roots_indic-mr_wikipedia
359 bigscience-data/roots_indic-mr_wikiquote
360 bigscience-data/roots_indic-mr_wikisource
361 bigscience-data/roots_indic-mr_wiktionary
362 bigscience-data/roots_indic-
 ne_unsupervised_cross_lingual_representation_learning_at_scale
363 bigscience-data/roots_indic-or_indic_nlp_corpus
364 bigscience-data/roots_indic-or_mkb
365 bigscience-data/roots_indic-or_odiencorp
366 bigscience-data/roots_indic-or_opus100
367 bigscience-data/roots_indic-or_pib
368 bigscience-data/roots_indic-or_samanantar
369 bigscience-data/roots_indic-or_wikipedia
370 bigscience-data/roots_indic-or_wikisource

371 bigscience-data/roots_indic-or_wiktionary
372 bigscience-data/roots_indic-pa_indic_nlp_corpus
373 bigscience-data/roots_indic-pa_opus100
374 bigscience-data/roots_indic-pa_pib
375 bigscience-data/roots_indic-pa_samanantar
376 bigscience-data/roots_indic-pa_ted_talks_iwslt
377 bigscience-data/roots_indic-pa_wikibooks
378 bigscience-data/roots_indic-pa_wikipedia
379 bigscience-data/roots_indic-pa_wikisource
380 bigscience-data/roots_indic-pa_wiktionary
381 bigscience-data/roots_indic-ta_indic_nlp_corpus
382 bigscience-data/roots_indic-ta_mkb
383 bigscience-data/roots_indic-ta_open_subtitles
384 bigscience-data/roots_indic-ta_opus100
385 bigscience-data/roots_indic-ta_pib
386 bigscience-data/roots_indic-ta_samanantar
387 bigscience-data/roots_indic-ta_ted_talks_iwslt
388 bigscience-data/roots_indic-ta_wikibooks
389 bigscience-data/roots_indic-ta_wikinews
390 bigscience-data/roots_indic-ta_wikipedia
391 bigscience-data/roots_indic-ta_wikiquote
392 bigscience-data/roots_indic-ta_wikisource
393 bigscience-data/roots_indic-ta_wiktionary_filtered
394 bigscience-data/roots_indic-te_indic_nlp_corpus
395 bigscience-data/roots_indic-te_mkb
396 bigscience-data/roots_indic-te_open_subtitles
397 bigscience-data/roots_indic-te_opus100
398 bigscience-data/roots_indic-te_pib
399 bigscience-data/roots_indic-te_samanantar
400 bigscience-data/roots_indic-te_ted_talks_iwslt
401 bigscience-data/roots_indic-te_wikibooks
402 bigscience-data/roots_indic-te_wikipedia
403 bigscience-data/roots_indic-te_wikiquote
404 bigscience-data/roots_indic-te_wikisource
405 bigscience-data/roots_indic-te_wiktionary
406 bigscience-data/roots_indic-ur_leipzig_wortschatz_urdu-pk_web_2019_sentences
407 bigscience-data/roots_indic-ur_leipzig_wortschatz_urdu_newscrawl_2016_sentences
408 bigscience-data/roots_indic-ur_mkb
409 bigscience-data/roots_indic-ur_open_subtitles
410 bigscience-data/roots_indic-ur_opus100
411 bigscience-data/roots_indic-ur_oscar
412 bigscience-data/roots_indic-ur_pib
413 bigscience-data/roots_indic-ur_ted_talks_iwslt
414 bigscience-data/roots_indic-ur_urdu-monolingual_corpus
415 bigscience-data/roots_indic-ur_wikibooks
416 bigscience-data/roots_indic-ur_wikipedia
417 bigscience-data/roots_indic-ur_wikiquote
418 bigscience-data/roots_indic-ur_wiktionary
419 bigscience-data/roots_nigercongo-ak_aggregated
420 bigscience-data/roots_nigercongo-bm_aggregated
421 bigscience-data/roots_nigercongo-fon_aggregated
422 bigscience-data/roots_nigercongo-ig_aggregated
423 bigscience-data/roots_nigercongo-ki_aggregated
424 bigscience-data/roots_nigercongo-lg_aggregated
425 bigscience-data/roots_nigercongo-ln_aggregated
426 bigscience-data/roots_nigercongo-nso_aggregated
427 bigscience-data/roots_nigercongo-ny_aggregated
428 bigscience-data/roots_nigercongo-rn_aggregated
429 bigscience-data/roots_nigercongo-rw_aggregated
430 bigscience-data/roots_nigercongo-sn_aggregated
431 bigscience-data/roots_nigercongo-st_aggregated
432 bigscience-data/roots_nigercongo-sw_aggregated
433 bigscience-data/roots_nigercongo-tn_aggregated

434 bigscience-data/roots_nigercongo-ts_aggregated
435 bigscience-data/roots_nigercongo-tum_aggregated
436 bigscience-data/roots_nigercongo-tw_aggregated
437 bigscience-data/roots_nigercongo-wo_aggregated
438 bigscience-data/roots_nigercongo-xh_aggregated
439 bigscience-data/roots_nigercongo-yo_aggregated
440 bigscience-data/roots_nigercongo-zu_aggregated
441 bigscience-data/roots_pt_brwac
442 bigscience-data/roots_pt_open_subtitles
443 bigscience-data/roots_pt_opus100
444 bigscience-data/roots_pt_oscar
445 bigscience-data/roots_pt_project_gutenberg
446 bigscience-data/roots_pt_pseudocrawl-filtered_672_pt_globalvoices_org
447 bigscience-data/roots_pt_qedcorpus
448 bigscience-data/roots_pt_scielo
449 bigscience-data/roots_pt_ted_talks_iwslt
450 bigscience-data/roots_pt_the_pile_europarl
451 bigscience-data/roots_pt_wikibooks
452 bigscience-data/roots_pt_wikimedia
453 bigscience-data/roots_pt_wikinews
454 bigscience-data/roots_pt_wikipedia
455 bigscience-data/roots_pt_wikiquote
456 bigscience-data/roots_pt_wikisource
457 bigscience-data/roots_pt_wikiversity
458 bigscience-data/roots_pt_wikivoyage
459 bigscience-data/roots_pt_wiktictionary
460 bigscience-data/roots_vi_binhvq_news_corpus
461 bigscience-data/roots_vi_data_on_covid_19_news_coverage_in_vietnam
462 bigscience-data/roots_vi_open_subtitles
463 bigscience-data/roots_vi_opus100
464 bigscience-data/roots_vi_oscar
465 bigscience-data/roots_vi_ted_talks_iwslt
466 bigscience-data/roots_vi_uit_vsmec
467 bigscience-data/roots_vi_vietai_sat
468 bigscience-data/roots_vi_vietnamese_poetry
469 bigscience-data/roots_vi_vietnamese_students_feedback
470 bigscience-data/roots_vi_vinbigdata_asr_vlsp_2020
471 bigscience-data/roots_vi_vinbigdata_monolingual_vlsp_2020
472 bigscience-data/roots_vi_vinbigdata_mt_vlsp_2020
473 bigscience-data/roots_vi_vntq_corpus_big
474 bigscience-data/roots_vi_wikibooks
475 bigscience-data/roots_vi_wikipedia
476 bigscience-data/roots_vi_wikiquote
477 bigscience-data/roots_vi_wikisource
478 bigscience-data/roots_vi_wikivoyage
479 bigscience-data/roots_vi_wiktictionary
480 bigscience-data/roots_zh-cn_wikipedia
481 bigscience-data/roots_zh-tw_wikipedia
482 bigscience-data/roots_zh-du_reader
483 bigscience-data/roots_zh_multi_un_2
484 bigscience-data/roots_zh_open_subtitles
485 bigscience-data/roots_zh_project_gutenberg
486 bigscience-data/roots_zh_pseudocrawl-filtered_503_www_zaobao_com_sg
487 bigscience-data/roots_zh_pseudocrawl-filtered_674_ai_baidu_com
488 bigscience-data/roots_zh_ted_talks_iwslt
489 bigscience-data/roots_zh_uncorpus
490 bigscience-data/roots_zh_wikibooks
491 bigscience-data/roots_zh_wikinews
492 bigscience-data/roots_zh_wikiquote
493 bigscience-data/roots_zh_wikiversity
494 bigscience-data/roots_zh_wikivoyage
495 bigscience-data/roots_zh_wudaocorpora
496 bigscience-data/roots_zhs_oscar

497 `bigscience-data/roots_zhs_qedcorpus`

498 `bigscience-data/roots_zht_qedcorpus`