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 on Hugging Face Spaces. 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 project[1]—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* generator[2]) 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/`

[2] See 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 BigSceince 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

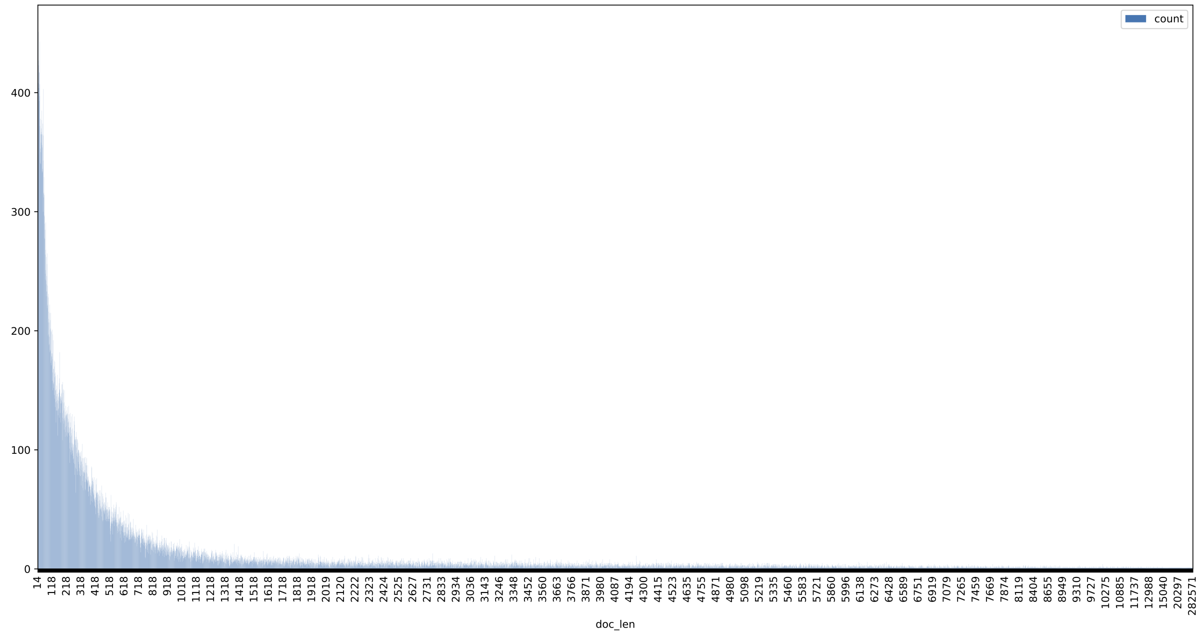


Figure 1: Document lengths distribution in a 100000-element random sample of the ROOTS corpus. The x-axis indicates the document length in characters. The y-axis indicates the number of documents of given length in the sample.

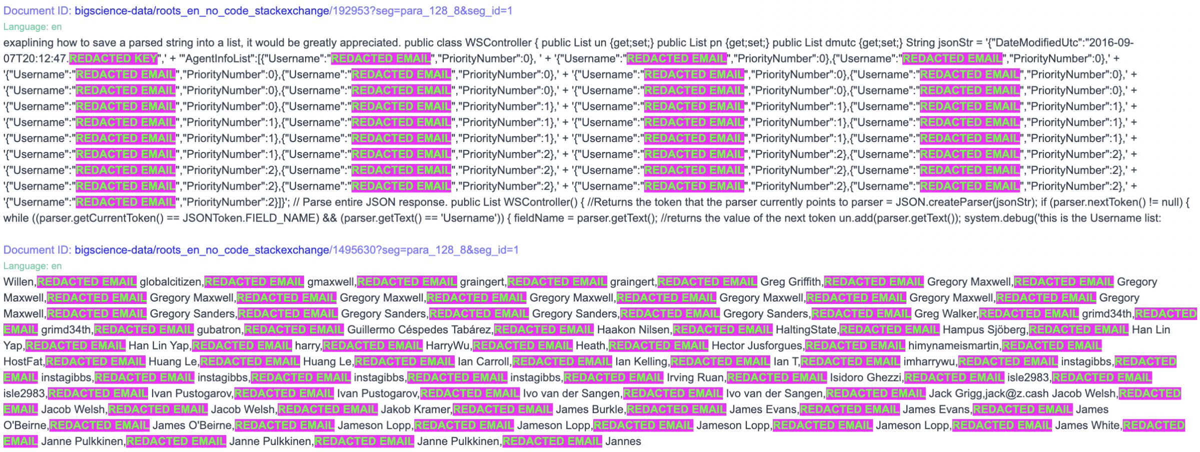
498 datasets, following a common naming pattern—the `bigscience-data` organization name followed by a slash, then the `roots` prefix, followed by the language tag and the data source identifier, delimited by underscores, e.g. `bigscience-data/roots_indic-mr_mkb`. There are two types of language identifiers—those 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_en_no_code_stackexchange/192953?seg=para_128_8&seg_id=1

Language: en

exaplining how to save a parsed string into a list, it would be greatly appreciated. public class WSController { public List un {get;set;} public List pn {get;set;} public List dmutc {get;set;} String jsonStr = '{"DateModifiedUtc":"2016-09-07T20:12:47.REDACTED KEY",' + '"AgentInfoList":[{"Username":"REDACTED EMAIL","PriorityNumber":0}, ' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":0},' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":0},' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":0},' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":0},' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":0},' + '{"Username":"REDACTED EMAIL","PriorityNumber":0},{"Username":"REDACTED EMAIL","PriorityNumber":1},' + '{"Username":"REDACTED EMAIL","PriorityNumber":1},{"Username":"REDACTED EMAIL","PriorityNumber":1},' + '{"Username":"REDACTED EMAIL","PriorityNumber":1},{"Username":"REDACTED EMAIL","PriorityNumber":1},' + '{"Username":"REDACTED EMAIL","PriorityNumber":1},{"Username":"REDACTED EMAIL","PriorityNumber":1},' + '{"Username":"REDACTED EMAIL","PriorityNumber":1},{"Username":"REDACTED EMAIL","PriorityNumber":1},' + '{"Username":"REDACTED EMAIL","PriorityNumber":1},{"Username":"REDACTED EMAIL","PriorityNumber":2},' + '{"Username":"REDACTED EMAIL","PriorityNumber":2},{"Username":"REDACTED EMAIL","PriorityNumber":2},' + '{"Username":"REDACTED EMAIL","PriorityNumber":2},{"Username":"REDACTED EMAIL","PriorityNumber":2},' + '{"Username":"REDACTED EMAIL","PriorityNumber":2},{"Username":"REDACTED EMAIL","PriorityNumber":2},' + '{"Username":"REDACTED EMAIL","PriorityNumber":2},{"Username":"REDACTED EMAIL","PriorityNumber":2},' + '{"Username":"REDACTED EMAIL","PriorityNumber":2}]}'; // Parse entire JSON response. public List WSController() { //Returns the token that the parser currently points to parser = JSON.createParser(jsonStr); if (parser.nextToken() != null) { while ((parser.getCurrentToken() == JSONToken.FIELD_NAME) && (parser.getText() == 'Username')) { fieldName = parser.getText(); //returns the value of the next token un.add(parser.getText()); system.debug('this is the Username list:

Document ID: bigscience-data/roots_en_no_code_stackexchange/1495630?seg=para_128_8&seg_id=1

Language: en

Willen,REDACTED EMAIL globalcitizen,REDACTED EMAIL gmaxwell,REDACTED EMAIL graingert,REDACTED EMAIL graingert,REDACTED EMAIL Greg Griffith,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Maxwell,REDACTED EMAIL Gregory Sanders,REDACTED EMAIL Gregory Sanders,REDACTED EMAIL Gregory Sanders,REDACTED EMAIL Gregory Sanders,REDACTED EMAIL Greg Walker,REDACTED EMAIL grimd34th,REDACTED EMAIL grimd34th,REDACTED EMAIL gubatron,REDACTED EMAIL Guillermo Céspedes Tabárez,REDACTED EMAIL Haakon Nilsen,REDACTED EMAIL HaltingState,REDACTED EMAIL Hampus Sjöberg,REDACTED EMAIL Han Lin Yap,REDACTED EMAIL Han Lin Yap,REDACTED EMAIL harry,REDACTED EMAIL HarryWu,REDACTED EMAIL Heath,REDACTED EMAIL Hector Jusforgues,REDACTED EMAIL himynameismartin,REDACTED EMAIL HostFat,REDACTED EMAIL Huang Le,REDACTED EMAIL Huang Le,REDACTED EMAIL Ian Carroll,REDACTED EMAIL Ian Kelling,REDACTED EMAIL Ian T,REDACTED EMAIL imharrywu,REDACTED EMAIL instagibbs,REDACTED EMAIL instagibbs,REDACTED EMAIL instagibbs,REDACTED EMAIL instagibbs,REDACTED EMAIL instagibbs,REDACTED EMAIL Irving Ruan,REDACTED EMAIL Isidoro Ghezzi,REDACTED EMAIL isle2983,REDACTED EMAIL isle2983,REDACTED EMAIL Ivan Pustogarov,REDACTED EMAIL Ivan Pustogarov,REDACTED EMAIL Ivo van der Sangen,REDACTED EMAIL Ivo van der Sangen,REDACTED EMAIL Jack Grigg,jack@z.cash Jacob Welsh,REDACTED EMAIL Jacob Welsh,REDACTED EMAIL Jacob Welsh,REDACTED EMAIL Jakob Kramer,REDACTED EMAIL James Burkle,REDACTED EMAIL James Evans,REDACTED EMAIL James Evans,REDACTED EMAIL James O'Beirne,REDACTED EMAIL James O'Beirne,REDACTED EMAIL Jameson Lopp,REDACTED EMAIL Jameson Lopp,REDACTED EMAIL Jameson Lopp,REDACTED EMAIL Jameson Lopp,REDACTED EMAIL James White,REDACTED EMAIL Janne Pulkkinen,REDACTED EMAIL Janne Pulkkinen,REDACTED EMAIL Janne Pulkkinen,REDACTED EMAIL Jannes

Figure 2: PII leakage - reuslts for the query `gmail.com`. We indicate the redacted PII with green and pink treatment.

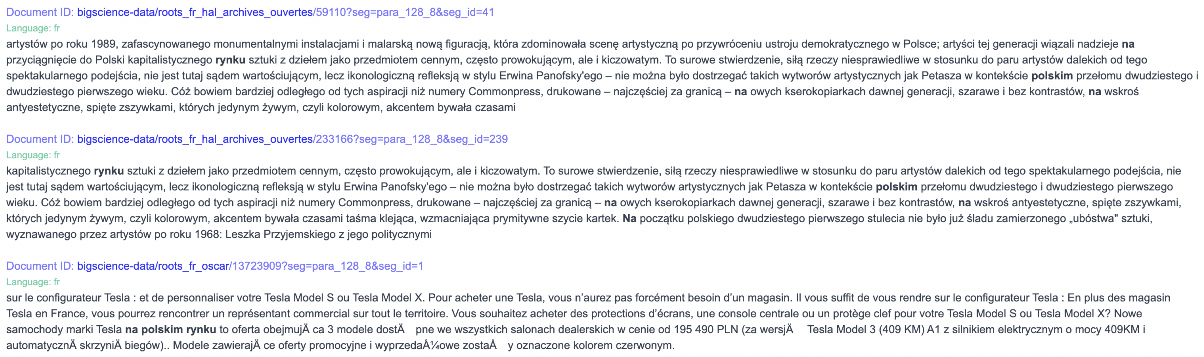
fed to the training process in equally sized batches (potentially with special tokens marking the end of a datapoint, like in the case of BLOOM). From the point of view of retrieval, however, each datapoint is considered a document—a discrete entity, associated with uniform metadata. It turns out the datapoints in ROOTS have vastly varying lengths - see Figure 1 for details. In order to be able to compare and rank documents, we make them more comparable by splitting them into roughly equally sized paragraphs and assign a unique ID to each paragraph.

Document Segmentation. We split documents into passages of 128 words each, with 8-word overlap between subsequent paragraphs. Overall, when talking about passages we mean the unit of text used during indexing. When referring to a document or a datapoint we mean a datapoint as it appears originally in one of the ROOTS datasets.

Unique Documents IDs. Each passage comes with an ID which uniquely identifies both the original datapoint and the passage within the datapoint. We adopt the following convention. First, we use the dataset name as defined on the Hugging Face hub, followed by a slash and an ID of the datapoint form which the given dataset came from. We follow this with a question mark introducing two parameters allowing us to uniquely identify the passage within the document. The `seg` parameter describes the segmentation strategy used, and the `seg_id` parameter indicates the index of the given passage under the specified segmentation strategy. E.g. `bigscience-data/roots_en_oscar/54676994?seg=para_128_8&seg_id=5` indicates the fifth passage obtained by splitting the document into 128-word passages with 8-word overlap. The document in question is the 54676994-th document in the `roots_en_oscar` dataset.

PII redaction During preliminary experiments on the ROOTS corpus, OSCAR [11] has been identified as a source of a large amount of documents containing personally identifiable information or PII. A regular-expression-based PII redaction script[3] has been applied to OSCAR prior to BLOOM training. However, the datasets themselves still contain unredacted text. In order to avoid leaking PII in our search engine, we apply an improved variant of the BigScience PII redaction script on the backend side and display results with PII redacted in a visible way - this way one can inspect the data and observe the problem, but personal information are predominantly removed.

[3] `https://github.com/bigscience-workshop/data-preparation/tree/main/preprocessing/training/02_pii`



Document ID: bigscience-data/roots_fr_hal_archives_ouvertes/59110?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

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 szycie kartek. **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 magasin 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? Nowe samochody marki Tesla **na polskim rynku** to oferta obejmujÄ ca 3 modele dostÄ   pne we wszystkich salonach dealerskich w cenie od 195 490 PLN (za wersjÄ   Tesla Model 3 (409 KM) A1 z silnikiem elektrycznym o mocy 409KM i automatycznÄ skrzyniÄ biegów).. Modele zawierajÄ ce oferty promocyjne i wyprzedaÅ¼owe zostaÅ   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 tame, 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 Lucene[4] 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.

```
python -m pyserini.index.lucene
    --collection JsonCollection \
    --input bigscience-data-pyserini/en/ \
    --index bigscience-data-index/en/ \
    --generator DefaultLuceneDocumentGenerator \
    --threads 64 \
    --storePositions \
    --storeDocvectors \
    --storeRaw \
    --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 BigScience 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, Ł ukasz 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.

```
bigscience-data/roots_ar_arabench
bigscience-data/roots_ar_arabic_billion_words
bigscience-data/roots_ar_brad_2
bigscience-data/roots_ar_habibi
bigscience-data/roots_ar_kalimat
bigscience-data/roots_ar_ksucca
bigscience-data/roots_ar_labr
bigscience-data/roots_ar_multi_un_2
bigscience-data/roots_ar_open_subtitles
bigscience-data/roots_ar_openiti_proc
bigscience-data/roots_ar_opus100
bigscience-data/roots_ar_oscar
bigscience-data/roots_ar_pseudocrawl-filtered_595_mawdoo3_com
bigscience-data/roots_ar_qedcorpus
bigscience-data/roots_ar_sanad
bigscience-data/roots_ar_tashkeela
bigscience-data/roots_ar_ted_talks_iwslt
bigscience-data/roots_ar_uncorpus
bigscience-data/roots_ar_wikibooks
bigscience-data/roots_ar_wikinews
bigscience-data/roots_ar_wikipedia
bigscience-data/roots_ar_wikiquote
bigscience-data/roots_ar_wikisource
bigscience-data/roots_ar_wikiversity
bigscience-data/roots_ar_wiktionary
bigscience-data/roots_ca_catalan_general_crawling
bigscience-data/roots_ca_catalan_government_crawling
bigscience-data/roots_ca_catalan_textual_corpus
bigscience-data/roots_ca_enriched_conllu_ancora_for_ml_training
bigscience-data/roots_ca_open_subtitles
bigscience-data/roots_ca_opus100
bigscience-data/roots_ca_oscar
bigscience-data/roots_ca_parlament_parla
bigscience-data/roots_ca_tecla
bigscience-data/roots_ca_ted_talks_iwslt
bigscience-data/roots_ca_vilaquad
bigscience-data/roots_ca_viquiquad
bigscience-data/roots_ca_wikibooks
bigscience-data/roots_ca_wikimedia
bigscience-data/roots_ca_wikinews
bigscience-data/roots_ca_wikipedia
bigscience-data/roots_ca_wikiquote
bigscience-data/roots_ca_wikisource
bigscience-data/roots_ca_wiktionary_filtered
bigscience-data/roots_ca_xquad_ca
bigscience-data/roots_code_github
bigscience-data/roots_code_stackexchange
bigscience-data/roots_en_book_dash_books
bigscience-data/roots_en_multi_un_2
bigscience-data/roots_en_no_code_stackexchange
bigscience-data/roots_en_odiencorp
bigscience-data/roots_en_open_subtitles
bigscience-data/roots_en_oscar
bigscience-data/roots_en_project_gutenberg
bigscience-data/roots_en_pseudocrawl-filtered_159_www_postcrescent_com
bigscience-data/roots_en_pseudocrawl-filtered_304_www_semana_com
bigscience-data/roots_en_pseudocrawl-filtered_339_www_actasanitaria_com
bigscience-data/roots_en_pseudocrawl-filtered_395_www_evwind_es
```

```
bigscience-data/roots_en_pseudocrawl-filtered_470_forums_hardwarezone_com_sg
bigscience-data/roots_en_pseudocrawl-filtered_483_alvinology_com
bigscience-data/roots_en_pseudocrawl-filtered_485_blog_moneysmart_sg
bigscience-data/roots_en_pseudocrawl-filtered_487_thesmartlocal_com
bigscience-data/roots_en_pseudocrawl-filtered_488_dailyvanity_sg
bigscience-data/roots_en_pseudocrawl-filtered_492_www_vivawoman_net
bigscience-data/roots_en_pseudocrawl-filtered_497_www_straitstimes_com
bigscience-data/roots_en_pseudocrawl-filtered_498_www_channelnewsasia_com
bigscience-data/roots_en_pseudocrawl-filtered_499_www_today_com_news
bigscience-data/roots_en_pseudocrawl-filtered_500_www_asiaone_com_singapore
bigscience-data/roots_en_pseudocrawl-filtered_501_theindependent_sg
bigscience-data/roots_en_pseudocrawl-filtered_502_www_ricemedia_co
bigscience-data/roots_en_pseudocrawl-filtered_510_timesofindia_indiatimes_com
bigscience-data/roots_en_pseudocrawl-filtered_534_www_nairaland_com
bigscience-data/roots_en_pseudocrawl-filtered_548_remezcla_com
bigscience-data/roots_en_pseudocrawl-filtered_638_globalvoices_org
bigscience-data/roots_en_pseudocrawl-filtered_689_www_abc_net_au
bigscience-data/roots_en_pseudocrawl-filtered_696_www_oercommons_org
bigscience-data/roots_en_qedcorpus
bigscience-data/roots_en_royal_society_corpus
bigscience-data/roots_en_s2orc_ai2_pdf_parses
bigscience-data/roots_en_scielo
bigscience-data/roots_en_ted_talks_iwslt
bigscience-data/roots_en_the_pile_europarl
bigscience-data/roots_en_the_pile_uspto
bigscience-data/roots_en_uncorpus
bigscience-data/roots_en_wikibooks
bigscience-data/roots_en_wikinews
bigscience-data/roots_en_wikipedia
bigscience-data/roots_en_wikiquote
bigscience-data/roots_en_wikiversity
bigscience-data/roots_en_wikivoyage
bigscience-data/roots_en_wiktionary
bigscience-data/roots_es_multi_un_2
bigscience-data/roots_es_open_subtitles
bigscience-data/roots_es_oscar
bigscience-data/roots_es_project_gutenberg
bigscience-data/roots_es_pseudocrawl-filtered_100_www_aporrea_org
bigscience-data/roots_es_pseudocrawl-filtered_103_www_elmostrador_cl
bigscience-data/roots_es_pseudocrawl-filtered_116_www_latribuna_hn
bigscience-data/roots_es_pseudocrawl-filtered_118_www_elheraldo_hn
bigscience-data/roots_es_pseudocrawl-filtered_125_www_noticiasde_es
bigscience-data/roots_es_pseudocrawl-
    filtered_130_www_elperiodicomediterraneo_com
bigscience-data/roots_es_pseudocrawl-filtered_136_valenciaplaza_com
bigscience-data/roots_es_pseudocrawl-filtered_146_www_perfil_com
bigscience-data/roots_es_pseudocrawl-filtered_153_financialfood_es
bigscience-data/roots_es_pseudocrawl-filtered_157_www_elsoldemexico_com_mx
bigscience-data/roots_es_pseudocrawl-filtered_158_www_diariodeleon_es
bigscience-data/roots_es_pseudocrawl-filtered_165_www_ticbeat_com
bigscience-data/roots_es_pseudocrawl-filtered_167_www_ambientum_com
bigscience-data/roots_es_pseudocrawl-filtered_169_www_el_carabobeno_com
bigscience-data/roots_es_pseudocrawl-filtered_172_www_rionegro_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_181_noticiassin_com
bigscience-data/roots_es_pseudocrawl-filtered_182_correodelsur_com
bigscience-data/roots_es_pseudocrawl-filtered_189_www_eleconomista_com_mx
bigscience-data/roots_es_pseudocrawl-filtered_198_www_eleconomista_es
bigscience-data/roots_es_pseudocrawl-filtered_203_www_que_es
bigscience-data/roots_es_pseudocrawl-filtered_207_elimpulso_com
bigscience-data/roots_es_pseudocrawl-filtered_209_misionesonline_net
bigscience-data/roots_es_pseudocrawl-filtered_20_www_clarin_com
bigscience-data/roots_es_pseudocrawl-filtered_211_www_elcomercio_com
bigscience-data/roots_es_pseudocrawl-filtered_213_www_hola_com
```

```
bigscience-data/roots_es_pseudocrawl-filtered_215_www_lainformacion_com
bigscience-data/roots_es_pseudocrawl-filtered_219_www_aguasresiduales_info
bigscience-data/roots_es_pseudocrawl-filtered_21_www_elperiodicodearagon_com
bigscience-data/roots_es_pseudocrawl-filtered_220_www_vanguardia_com_mx
bigscience-data/roots_es_pseudocrawl-filtered_223_www_eltambor_es
bigscience-data/roots_es_pseudocrawl-filtered_226_www_ole_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_229_www_expansion_com
bigscience-data/roots_es_pseudocrawl-filtered_231_ojo_pe
bigscience-data/roots_es_pseudocrawl-filtered_232_tn_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_233_www_dinero_com
bigscience-data/roots_es_pseudocrawl-filtered_237_www_cronista_com
bigscience-data/roots_es_pseudocrawl-filtered_23_www_elconfidencialdigital_com
bigscience-data/roots_es_pseudocrawl-filtered_244_www_df_cl
bigscience-data/roots_es_pseudocrawl-filtered_245_www_noticiasdenavarra_com
bigscience-data/roots_es_pseudocrawl-filtered_246_www_eldiarionuevodia_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_248_www_telesurtv_net
bigscience-data/roots_es_pseudocrawl-filtered_249_www_telecinco_es
bigscience-data/roots_es_pseudocrawl-filtered_250_www_cooperativa_cl
bigscience-data/roots_es_pseudocrawl-filtered_253_www_debate_com_mx
bigscience-data/roots_es_pseudocrawl-filtered_254_diario_mx
bigscience-data/roots_es_pseudocrawl-filtered_255_elcomercio_pe
bigscience-data/roots_es_pseudocrawl-filtered_256_www_laprovincia_es
bigscience-data/roots_es_pseudocrawl-filtered_257_www_diaridetarragona_com
bigscience-data/roots_es_pseudocrawl-filtered_263_www_lasexta_com
bigscience-data/roots_es_pseudocrawl-filtered_267_www_elperiodico_com_es
bigscience-data/roots_es_pseudocrawl-filtered_276_radio_uchile_cl
bigscience-data/roots_es_pseudocrawl-filtered_277_www_entornointeligente_com
bigscience-data/roots_es_pseudocrawl-filtered_280_salamancartvaldia_es
bigscience-data/roots_es_pseudocrawl-filtered_286_www_nacion_com
bigscience-data/roots_es_pseudocrawl-filtered_287_www_cibercuba_com
bigscience-data/roots_es_pseudocrawl-filtered_288_www_marca_com
bigscience-data/roots_es_pseudocrawl-filtered_28_www_fayerwayer_com
bigscience-data/roots_es_pseudocrawl-filtered_294_www_laopinion_com_co
bigscience-data/roots_es_pseudocrawl-filtered_299_www_lne_es
bigscience-data/roots_es_pseudocrawl-filtered_30_www_radiocable_com
bigscience-data/roots_es_pseudocrawl-filtered_315_lasillavacia_com
bigscience-data/roots_es_pseudocrawl-filtered_317_diariocorreo_pe
bigscience-data/roots_es_pseudocrawl-filtered_320_www_paginasiete_bo
bigscience-data/roots_es_pseudocrawl-filtered_324_gestion_pe
bigscience-data/roots_es_pseudocrawl-filtered_325_www_laprensa_hn
bigscience-data/roots_es_pseudocrawl-filtered_32_www_elexpresso_com
bigscience-data/roots_es_pseudocrawl-filtered_333_www_elmundo_es
bigscience-data/roots_es_pseudocrawl-filtered_341_es_cointelegraph_com
bigscience-data/roots_es_pseudocrawl-filtered_349_www_eltiempo_com
bigscience-data/roots_es_pseudocrawl-filtered_34_www_losandes_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_354_www_lagaceta_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_359_www_efeverde_com
bigscience-data/roots_es_pseudocrawl-filtered_367_elcorreoweb_es
bigscience-data/roots_es_pseudocrawl-filtered_373_www_farodevigo_es
bigscience-data/roots_es_pseudocrawl-filtered_374_www_talcualdigital_com
bigscience-data/roots_es_pseudocrawl-filtered_376_www_elpopular_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_381_www_cuartopoder_es
bigscience-data/roots_es_pseudocrawl-filtered_386_www_prensalibre_com
bigscience-data/roots_es_pseudocrawl-filtered_392_www_muypymes_com
bigscience-data/roots_es_pseudocrawl-filtered_396_www_eldiario_es
bigscience-data/roots_es_pseudocrawl-filtered_401_www_elperiodicodemexico_com
bigscience-data/roots_es_pseudocrawl-filtered_404_www_telam_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_405_www_emol_com
bigscience-data/roots_es_pseudocrawl-filtered_406_www_americaeconomia_com
bigscience-data/roots_es_pseudocrawl-filtered_409_www_proceso_com_mx
bigscience-data/roots_es_pseudocrawl-filtered_417_www_radiolaprimerisima_com
bigscience-data/roots_es_pseudocrawl-filtered_420_www_retema_es
bigscience-data/roots_es_pseudocrawl-filtered_422_www_formulatv_com
```

```
bigscience-data/roots_es_pseudocrawl-filtered_424_www_lavanguardia_com
bigscience-data/roots_es_pseudocrawl-filtered_429_cadenaser_com
bigscience-data/roots_es_pseudocrawl-filtered_430_www_eldiario_ec
bigscience-data/roots_es_pseudocrawl-
    filtered_431_www_elperiodicoextremadura_com
bigscience-data/roots_es_pseudocrawl-filtered_44_ladiaria_com_uy
bigscience-data/roots_es_pseudocrawl-filtered_518_www_elcolombiano_com
bigscience-data/roots_es_pseudocrawl-filtered_53_www_expreso_ec
bigscience-data/roots_es_pseudocrawl-filtered_56_www_eluniverso_com
bigscience-data/roots_es_pseudocrawl-filtered_58_www_levante_emv_com
bigscience-data/roots_es_pseudocrawl-filtered_62_www_lapagina_com_sv
bigscience-data/roots_es_pseudocrawl-filtered_63_www_lanacion_com_ar
bigscience-data/roots_es_pseudocrawl-filtered_641_es_globalvoices_org
bigscience-data/roots_es_pseudocrawl-filtered_675_www_elespectador_com
bigscience-data/roots_es_pseudocrawl-filtered_67_www_elpais_cr
bigscience-data/roots_es_pseudocrawl-filtered_71_www_rtve_es
bigscience-data/roots_es_pseudocrawl-filtered_78_www_listindiario_com
bigscience-data/roots_es_pseudocrawl-filtered_79_www_laopiniondemurcia_es
bigscience-data/roots_es_pseudocrawl-filtered_86_www_motorpasion_com
bigscience-data/roots_es_pseudocrawl-filtered_90_peru_com
bigscience-data/roots_es_pseudocrawl-filtered_91_www_diario26_com
bigscience-data/roots_es_qedcorpus
bigscience-data/roots_es_scielo
bigscience-data/roots_es_ted_talks_iwslt
bigscience-data/roots_es_the_pile_europarl
bigscience-data/roots_es_uncorpus
bigscience-data/roots_es_wikibooks
bigscience-data/roots_es_wikinews
bigscience-data/roots_es_wikipedia
bigscience-data/roots_es_wikiquote
bigscience-data/roots_es_wikisource
bigscience-data/roots_es_wikiversity
bigscience-data/roots_es_wikivoyage
bigscience-data/roots_es_wiktionary
bigscience-data/roots_eu_bsbasque
bigscience-data/roots_eu_open_subtitles
bigscience-data/roots_eu_opus100
bigscience-data/roots_eu_oscar
bigscience-data/roots_eu_pseudocrawl-filtered_506_goiena_eus
bigscience-data/roots_eu_pseudocrawl-filtered_563_ahotsak_eus
bigscience-data/roots_eu_pseudocrawl-filtered_635_www_berria_eus
bigscience-data/roots_eu_pseudocrawl-filtered_637_www_argia_eus
bigscience-data/roots_eu_ted_talks_iwslt
bigscience-data/roots_eu_wikibooks
bigscience-data/roots_eu_wikipedia
bigscience-data/roots_eu_wikiquote
bigscience-data/roots_eu_wikisource
bigscience-data/roots_eu_wiktionary
bigscience-data/roots_fr_book_dash_books
bigscience-data/roots_fr_ester
bigscience-data/roots_fr_hal_archives_ouvertes
bigscience-data/roots_fr_multi_un_2
bigscience-data/roots_fr_open_subtitles
bigscience-data/roots_fr_oscar
bigscience-data/roots_fr_project_gutenberg
bigscience-data/roots_fr_pseudocrawl-filtered_530_www_mediapart_fr
bigscience-data/roots_fr_pseudocrawl-filtered_550_www_lemonde_fr
bigscience-data/roots_fr_pseudocrawl-filtered_599_fr_globalvoices_org
bigscience-data/roots_fr_qedcorpus
bigscience-data/roots_fr_ted_talks_iwslt
bigscience-data/roots_fr_the_pile_europarl
bigscience-data/roots_fr_uncorpus
bigscience-data/roots_fr_wikibooks
```

```
bigscience-data/roots_fr_wikinews
bigscience-data/roots_fr_wikipedia
bigscience-data/roots_fr_wikiquote
bigscience-data/roots_fr_wikisource
bigscience-data/roots_fr_wikiversity
bigscience-data/roots_fr_wikivoyage
bigscience-data/roots_fr_wiktionary
bigscience-data/roots_id_indo4b_bppt
bigscience-data/roots_id_indo4b_jw300
bigscience-data/roots_id_indo4b_kompas
bigscience-data/roots_id_indo4b_parallel
bigscience-data/roots_id_indo4b_talpco
bigscience-data/roots_id_indo4b_tempo
bigscience-data/roots_id_indonesian_frog_storytelling_corpus
bigscience-data/roots_id_indonesian_news_articles_2017
bigscience-data/roots_id_indonesian_news_corpus
bigscience-data/roots_id_indonli
bigscience-data/roots_id_indosum
bigscience-data/roots_id_open_subtitles
bigscience-data/roots_id_opus100
bigscience-data/roots_id_oscar
bigscience-data/roots_id_pseudocrawl-filtered_512_kumparan_com
bigscience-data/roots_id_pseudocrawl-filtered_545_www_detik_com
bigscience-data/roots_id_pseudocrawl-filtered_549_www_cnnindonesia_com
bigscience-data/roots_id_pseudocrawl-filtered_572_tirto_id
bigscience-data/roots_id_recibrew
bigscience-data/roots_id_ted_talks_iwslt
bigscience-data/roots_id_wikibooks
bigscience-data/roots_id_wikimedia
bigscience-data/roots_id_wikipedia
bigscience-data/roots_id_wikiquote
bigscience-data/roots_id_wikisource
bigscience-data/roots_id_wiktionary
bigscience-data/roots_indic-as_opus100
bigscience-data/roots_indic-as_samanantar
bigscience-data/roots_indic-as_ted_talks_iwslt
bigscience-data/roots_indic-as_wikipedia
bigscience-data/roots_indic-as_wikisource
bigscience-data/roots_indic-as_wiktionary
bigscience-data/roots_indic-bn_bangla_lm
bigscience-data/roots_indic-bn_bangla_sentiment_classification_datasets
bigscience-data/roots_indic-bn_bengali_question_answering
bigscience-data/roots_indic-bn_indic_nlp_corpus
bigscience-data/roots_indic-bn_mkb
bigscience-data/roots_indic-bn_open_subtitles
bigscience-data/roots_indic-bn_opus100
bigscience-data/roots_indic-bn_oscar
bigscience-data/roots_indic-bn_pib
bigscience-data/roots_indic-bn_samanantar
bigscience-data/roots_indic-bn_ted_talks_iwslt
bigscience-data/roots_indic-bn_wikibooks
bigscience-data/roots_indic-bn_wikipedia
bigscience-data/roots_indic-bn_wikisource
bigscience-data/roots_indic-bn_wikivoyage
bigscience-data/roots_indic-bn_wiktionary
bigscience-data/roots_indic-gu_indic_nlp_corpus
bigscience-data/roots_indic-gu_mkb
bigscience-data/roots_indic-gu_opus100
bigscience-data/roots_indic-gu_pib
bigscience-data/roots_indic-gu_samanantar
bigscience-data/roots_indic-gu_ted_talks_iwslt
bigscience-data/roots_indic-gu_wikipedia
bigscience-data/roots_indic-gu_wikiquote
```

```
bigscience-data/roots_indic-gu_wikisource
bigscience-data/roots_indic-gu_wiktionary
bigscience-data/roots_indic-hi_iitb_english_hindi_corpus
bigscience-data/roots_indic-hi_indic_nlp_corpus
bigscience-data/roots_indic-hi_mkb
bigscience-data/roots_indic-hi_open_subtitles
bigscience-data/roots_indic-hi_opus100
bigscience-data/roots_indic-hi_oscar
bigscience-data/roots_indic-hi_pib
bigscience-data/roots_indic-hi_pseudocrawl-filtered_515_www_aajtak_in
bigscience-data/roots_indic-hi_pseudocrawl-filtered_667_www_bhaskar_com
bigscience-data/roots_indic-hi_qedcorpus
bigscience-data/roots_indic-hi_samanantar
bigscience-data/roots_indic-hi_ted_talks_iwslt
bigscience-data/roots_indic-hi_wikibooks
bigscience-data/roots_indic-hi_wikimedia
bigscience-data/roots_indic-hi_wikipedia
bigscience-data/roots_indic-hi_wikiquote
bigscience-data/roots_indic-hi_wikisource
bigscience-data/roots_indic-hi_wikiversity
bigscience-data/roots_indic-hi_wikivoyage
bigscience-data/roots_indic-hi_wiktionary
bigscience-data/roots_indic-kn_indic_nlp_corpus
bigscience-data/roots_indic-kn_opus100
bigscience-data/roots_indic-kn_samanantar
bigscience-data/roots_indic-kn_ted_talks_iwslt
bigscience-data/roots_indic-kn_wikipedia
bigscience-data/roots_indic-kn_wikiquote
bigscience-data/roots_indic-kn_wikisource
bigscience-data/roots_indic-kn_wiktionary
bigscience-data/roots_indic-ml_indic_nlp_corpus
bigscience-data/roots_indic-ml_mkb
bigscience-data/roots_indic-ml_open_subtitles
bigscience-data/roots_indic-ml_opus100
bigscience-data/roots_indic-ml_pib
bigscience-data/roots_indic-ml_samanantar
bigscience-data/roots_indic-ml_ted_talks_iwslt
bigscience-data/roots_indic-ml_wikibooks
bigscience-data/roots_indic-ml_wikipedia
bigscience-data/roots_indic-ml_wikiquote
bigscience-data/roots_indic-ml_wikisource
bigscience-data/roots_indic-ml_wiktionary
bigscience-data/roots_indic-mr_indic_nlp_corpus
bigscience-data/roots_indic-mr_mkb
bigscience-data/roots_indic-mr_opus100
bigscience-data/roots_indic-mr_pib
bigscience-data/roots_indic-mr_samanantar
bigscience-data/roots_indic-mr_ted_talks_iwslt
bigscience-data/roots_indic-mr_wikibooks
bigscience-data/roots_indic-mr_wikipedia
bigscience-data/roots_indic-mr_wikiquote
bigscience-data/roots_indic-mr_wikisource
bigscience-data/roots_indic-mr_wiktionary
bigscience-data/roots_indic-ne_unsupervised_cross_lingual_representation_learning_at_scale
bigscience-data/roots_indic-or_indic_nlp_corpus
bigscience-data/roots_indic-or_mkb
bigscience-data/roots_indic-or_odiencorp
bigscience-data/roots_indic-or_opus100
bigscience-data/roots_indic-or_pib
bigscience-data/roots_indic-or_samanantar
bigscience-data/roots_indic-or_wikipedia
bigscience-data/roots_indic-or_wikisource
```

```
bigscience-data/roots_indic-or_wiktionary
bigscience-data/roots_indic-pa_indic_nlp_corpus
bigscience-data/roots_indic-pa_opus100
bigscience-data/roots_indic-pa_pib
bigscience-data/roots_indic-pa_samanantar
bigscience-data/roots_indic-pa_ted_talks_iwslt
bigscience-data/roots_indic-pa_wikibooks
bigscience-data/roots_indic-pa_wikipedia
bigscience-data/roots_indic-pa_wikisource
bigscience-data/roots_indic-pa_wiktionary
bigscience-data/roots_indic-ta_indic_nlp_corpus
bigscience-data/roots_indic-ta_mkb
bigscience-data/roots_indic-ta_open_subtitles
bigscience-data/roots_indic-ta_opus100
bigscience-data/roots_indic-ta_pib
bigscience-data/roots_indic-ta_samanantar
bigscience-data/roots_indic-ta_ted_talks_iwslt
bigscience-data/roots_indic-ta_wikibooks
bigscience-data/roots_indic-ta_wikinews
bigscience-data/roots_indic-ta_wikipedia
bigscience-data/roots_indic-ta_wikiquote
bigscience-data/roots_indic-ta_wikisource
bigscience-data/roots_indic-ta_wiktionary_filtered
bigscience-data/roots_indic-te_indic_nlp_corpus
bigscience-data/roots_indic-te_mkb
bigscience-data/roots_indic-te_open_subtitles
bigscience-data/roots_indic-te_opus100
bigscience-data/roots_indic-te_pib
bigscience-data/roots_indic-te_samanantar
bigscience-data/roots_indic-te_ted_talks_iwslt
bigscience-data/roots_indic-te_wikibooks
bigscience-data/roots_indic-te_wikipedia
bigscience-data/roots_indic-te_wikiquote
bigscience-data/roots_indic-te_wikisource
bigscience-data/roots_indic-te_wiktionary
bigscience-data/roots_indic-ur_leipzig_wortschatz_urdu-pk_web_2019_sentences
bigscience-data/roots_indic-ur_leipzig_wortschatz_urdu_newscrawl_2016_sentences
bigscience-data/roots_indic-ur_mkb
bigscience-data/roots_indic-ur_open_subtitles
bigscience-data/roots_indic-ur_opus100
bigscience-data/roots_indic-ur_oscar
bigscience-data/roots_indic-ur_pib
bigscience-data/roots_indic-ur_ted_talks_iwslt
bigscience-data/roots_indic-ur_urdu-monolingual-corpus
bigscience-data/roots_indic-ur_wikibooks
bigscience-data/roots_indic-ur_wikipedia
bigscience-data/roots_indic-ur_wikiquote
bigscience-data/roots_indic-ur_wiktionary
bigscience-data/roots_nigercongo-ak_aggregated
bigscience-data/roots_nigercongo-bm_aggregated
bigscience-data/roots_nigercongo-fon_aggregated
bigscience-data/roots_nigercongo-ig_aggregated
bigscience-data/roots_nigercongo-ki_aggregated
bigscience-data/roots_nigercongo-lg_aggregated
bigscience-data/roots_nigercongo-ln_aggregated
bigscience-data/roots_nigercongo-nso_aggregated
bigscience-data/roots_nigercongo-ny_aggregated
bigscience-data/roots_nigercongo-rn_aggregated
bigscience-data/roots_nigercongo-rw_aggregated
bigscience-data/roots_nigercongo-sn_aggregated
bigscience-data/roots_nigercongo-st_aggregated
bigscience-data/roots_nigercongo-sw_aggregated
bigscience-data/roots_nigercongo-tn_aggregated
```

```
bigscience-data/roots_nigercongo-ts_aggregated
bigscience-data/roots_nigercongo-tum_aggregated
bigscience-data/roots_nigercongo-tw_aggregated
bigscience-data/roots_nigercongo-wo_aggregated
bigscience-data/roots_nigercongo-xh_aggregated
bigscience-data/roots_nigercongo-yo_aggregated
bigscience-data/roots_nigercongo-zu_aggregated
bigscience-data/roots_pt_brwac
bigscience-data/roots_pt_open_subtitles
bigscience-data/roots_pt_opus100
bigscience-data/roots_pt_oscar
bigscience-data/roots_pt_project_gutenberg
bigscience-data/roots_pt_pseudocrawl-filtered_672_pt_globalvoices_org
bigscience-data/roots_pt_qedcorpus
bigscience-data/roots_pt_scielo
bigscience-data/roots_pt_ted_talks_iwslt
bigscience-data/roots_pt_the_pile_europarl
bigscience-data/roots_pt_wikibooks
bigscience-data/roots_pt_wikimedia
bigscience-data/roots_pt_wikinews
bigscience-data/roots_pt_wikipedia
bigscience-data/roots_pt_wikiquote
bigscience-data/roots_pt_wikisource
bigscience-data/roots_pt_wikiversity
bigscience-data/roots_pt_wikivoyage
bigscience-data/roots_pt_wiktionary
bigscience-data/roots_vi_binhvq_news_corpus
bigscience-data/roots_vi_data_on_covid_19_news_coverage_in_vietnam
bigscience-data/roots_vi_open_subtitles
bigscience-data/roots_vi_opus100
bigscience-data/roots_vi_oscar
bigscience-data/roots_vi_ted_talks_iwslt
bigscience-data/roots_vi_uit_vsmec
bigscience-data/roots_vi_vietai_sat
bigscience-data/roots_vi_vietnamese_poetry
bigscience-data/roots_vi_vietnamese_students_feedback
bigscience-data/roots_vi_vinbigdata_asr_vlsp_2020
bigscience-data/roots_vi_vinbigdata_monolingual_vlsp_2020
bigscience-data/roots_vi_vinbigdata_mt_vlsp_2020
bigscience-data/roots_vi_vntq_corpus_big
bigscience-data/roots_vi_wikibooks
bigscience-data/roots_vi_wikipedia
bigscience-data/roots_vi_wikiquote
bigscience-data/roots_vi_wikisource
bigscience-data/roots_vi_wikivoyage
bigscience-data/roots_vi_wiktionary
bigscience-data/roots_zh-cn_wikipedia
bigscience-data/roots_zh-tw_wikipedia
bigscience-data/roots_zh_du_reader
bigscience-data/roots_zh_multi_un_2
bigscience-data/roots_zh_open_subtitles
bigscience-data/roots_zh_project_gutenberg
bigscience-data/roots_zh_pseudocrawl-filtered_503_www_zaobao_com_sg
bigscience-data/roots_zh_pseudocrawl-filtered_674_ai_baidu_com
bigscience-data/roots_zh_ted_talks_iwslt
bigscience-data/roots_zh_uncorpus
bigscience-data/roots_zh_wikibooks
bigscience-data/roots_zh_wikinews
bigscience-data/roots_zh_wikiquote
bigscience-data/roots_zh_wikiversity
bigscience-data/roots_zh_wikivoyage
bigscience-data/roots_zh_wudaocorpora
bigscience-data/roots_zhs_oscar
```

```
bigscience-data/roots_zhs_qedcorpus
bigscience-data/roots_zht_qedcorpus
```