Spacerini: Plug-and-play Search Engines with Pyserini and Hugging Face

Christopher Akiki[1,2] **Odunayo Ogundepo**[3] **Aleksandra Piktus**[4,5]

Xinyu Zhang[3] **Akintunde Oladipo**[3] **Jimmy Lin**[3] **Martin Potthast**[1,2]

[1] Leipzig University [2] ScaDS.AI [3] University of Waterloo
[4] Hugging Face [5] Sapienza University

Abstract

We present Spacerini, a tool that integrates the Pyserini toolkit for reproducible information retrieval research with Hugging Face to enable the seamless construction and deployment of interactive search engines. Spacerini makes state-of-the-art sparse and dense retrieval models more accessible to non-IR practitioners while minimizing deployment effort. This is useful for NLP researchers who want to better understand and validate their research by performing qualitative analyses of training corpora, for IR researchers who want to demonstrate new retrieval models integrated into the growing Pyserini ecosystem, and for third parties reproducing the work of other researchers. Spacerini is open-source[1] and includes utilities for loading, preprocessing, indexing, and deploying search engines locally and remotely. We demonstrate a portfolio of 13 search engines created with Spacerini for different use cases.[2]

1 Introduction

The commoditization of data has transformed data-driven computer science in general (Hey et al., 2009) and machine learning (ML) and natural language processing (NLP) in particular (Mitchell et al., 2022). The race to train ever-larger language models depends so much on having access to immense amounts of text (Hoffmann et al., 2022) that the datasets have become, as Bender et al. (2021) claims, both technically (Bender et al., 2021) and methodologically (Jo and Gebru, 2020) "too big to document". In practice, this often leads to an approach of training first and asking questions later (Akiki et al., 2022), which again is an example of a convenience experiment (Krohs, 2012), a research approach that depends on the availability of a resource and the ease of a method rather than its suitability to the problem at hand. In this sense, Beaulieu and Leonelli (2021) believe it is important to distinguish between the availability of data and its appropriateness, especially in light of the misconception that web data represents all human experience and is immune to the ever-widening *digital divide* (Leonelli, 2020). They suggest that the divide not only exists, but limits the representativeness of the web, which in turn reinforces the biases of the artifacts that use it (Bender et al., 2021).

Being unable to easily audit large datasets incentivizes researchers to release models trained on data they do not truly understand (Mitchell et al., 2022) leading to model behaviors that are hard to study, predict or trace (Mitchell et al., 2022; Siddiqui et al., 2022; Akyurek et al., 2022). This is especially problematic in light of the potential real-world harms that ensue (Weidinger et al., 2021; Hutchinson et al., 2020; Founta et al., 2018; Fast et al., 2016). Being able to properly understand the limitations of our datasets and qualitatively explore them in an ad-hoc fashion is a necessary first step toward understanding the behavior, harmful biases and failure modes of the artifacts that build upon them. Understanding the training data is therefore a critical step in the process of releasing and auditing large language models (Mökander et al., 2023).

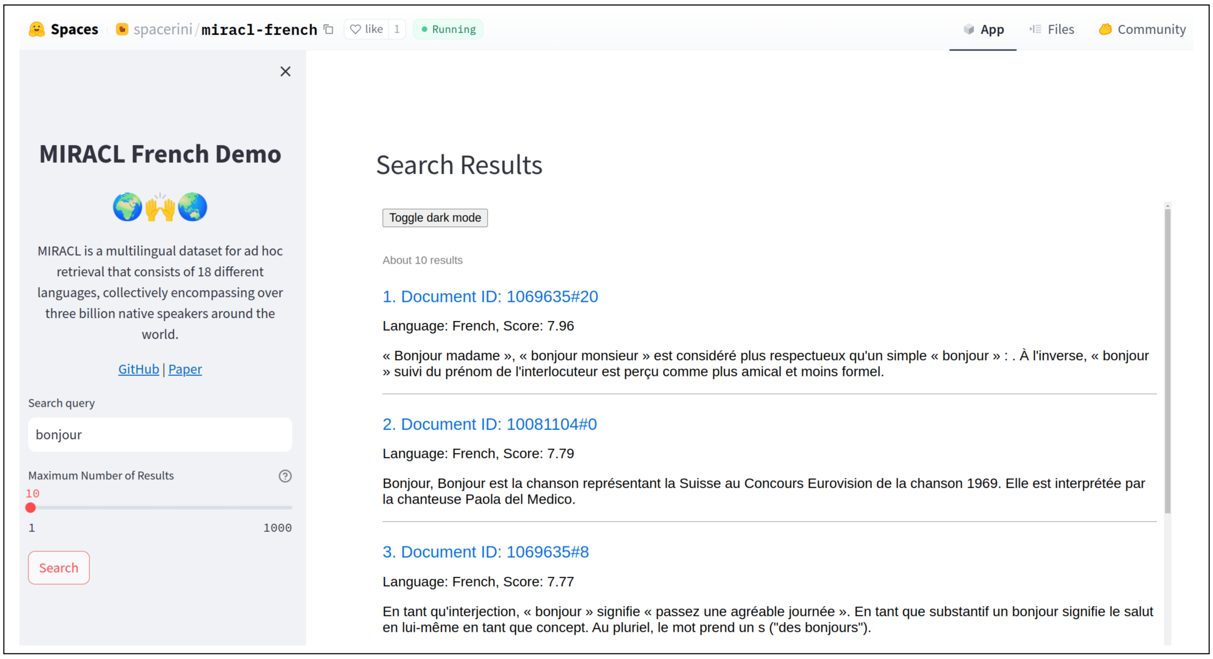
It is from this vantage point that we initially developed Spacerini as an open-source tool for the quick indexing and deployment of shareable search engines, but have also since come to realize its potential in being useful for an even wider audience interested in making their text artifacts searchable. Indeed this includes IR students, Digital Humanists, Shared Tasks organizers, and digital investigative journalists, all of whom seek to quickly deploy ad hoc search engines for their research. We cover these use cases in more detail in Section 4.

Spacerini helps streamline the process of auditing large datasets by allowing users to effortlessly index their text collections and deploy them as interactive search applications that can be easily edited in the browser and shared with all stakeholders. It achieves that by "standing on the shoul-

[1] https://github.com/castorini/hf-spacerini

[2] https://hf.co/spacerini

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 140–148
December 6-10, 2023 ©2023 Association for Computational Linguistics



MIRACL French Demo

MIRACL is a multilingual dataset for ad hoc retrieval that consists of 18 different languages, collectively encompassing over three billion native speakers around the world.

GitHub | Paper

Search query: bonjour

Maximum Number of Results: 10 (1–1000)

Search Results

About 10 results

1. Document ID: 1069635#20

Language: French, Score: 7.96

« Bonjour madame », « bonjour monsieur » est considéré plus respectueux qu'un simple « bonjour » : . À l'inverse, « bonjour » suivi du prénom de l'interlocuteur est perçu comme plus amical et moins formel.

2. Document ID: 10081104#0

Language: French, Score: 7.79

Bonjour, Bonjour est la chanson représentant la Suisse au Concours Eurovision de la chanson 1969. Elle est interprétée par la chanteuse Paola del Medico.

3. Document ID: 1069635#8

Language: French, Score: 7.77

En tant qu'interjection, « bonjour » signifie « passez une agréable journée ». En tant que substantif un bonjour signifie le salut en lui-même en tant que concept. Au pluriel, le mot prend un s ("des bonjours").

Figure 1: Example of one of the many search apps (`hf/spacerini/miracl-french`) deployed as a Hugging Face Space. The Lucene BM25 index is hosted in the same repository as the frontend using `git LFS` and the frontend is based on a `template` which was automatically generated from one of the many Spacerini `cookiecutter` templates. The loading, preprocessing, indexing, and app deployment were made using an end-to-end workflow similar to the one showcased in Section 3.

ders" of battle-tested open-source libraries from the Castorini (Lin et al., 2021), Hugging Face (Abid et al., 2019; Lhoest et al., 2021) and Python ecosystems. It also enhances the interoperability between them to enable quick indexing and free deployment of search interfaces in an easy-to-use package that makes it possible to reduce the overhead typically involved in operationalizing data governance frameworks, and allows stakeholders to focus on data analysis rather than data engineering. This is achieved through the modularity of its design that enables data loading (Section 3.1), preprocessing (Section 3.2), dense and sparse indexing (Sections 3.3), as well as the creation (Section 3.1), and free hosting of graphical search interfaces (Section 3.5) for text datasets.

2 Background and Related Work

Large scale, predominantly web-mined text datasets have been proliferating in NLP recently, giving rise to publications (Laurençon et al., 2022; Gao et al., 2021; Ortiz Suárez et al., 2019; Raffel et al., 2022) which often contain interesting analyses of the specific datasets being presented, however, usually lack any comparison to existing resources beyond basic metrics such as sizes of the datasets or languages they contain.

As discussed in Section 1, in the face of an increased scrutiny of the models trained on datasets in question, the topic of data understanding and governance has been gaining more traction, being accepted as an important part of research. Efforts such as those of Mitchell et al. (2022) contribute frameworks for more standardised and reproducible metrics and measurements of datasets, and we position ourselves as a complementary continuation of their work, focusing on a more curatorial and qualitative assessment that might not readily fit under the umbrella of "measurements". We therefore aim to fill the gap in the evaluation landscape by facilitating qualitative, rather than quantitative analysis of large scale data collections.

Similarly to the authors of Gradio (Abid et al., 2019), a Python package for fast development of Machine Learning demos, we believe that the accessibility of data and model analysis tools is crucial to building both the understanding of and the trust in the underlying resources. The potential of relevance-based interfaces to massive textual corpora, the creation of which can be facilitated by leveraging toolkits such as Pyserini (Lin et al., 2021), has previously been tapped into by the researchers at the Allen Institute of AI who propose a C4 (Raffel et al., 2022) search engine[3]. Similar interfaces have also been found useful in more specialised domains, e.g. in COVID-related datasets (Zhang et al., 2020), news quotes (Vuković et al., 2022), or medical literature (Niezni et al., 2022). However, while these solutions are undeniably useful, they remain very contextual: dataset-specific, and project-specific. We believe Spacerini to be the first generalizable tool which proposes an end-to-end pipeline automating the route from raw text to qualitative analysis.

3 Spacerini

Spacerini is a modular framework that integrates Pyserini with the Hugging Face ecosystem to streamline the process of going from any Hugging Face text dataset (or indeed any text dataset)—either local or hosted on the Hugging Face Hub—to a search interface driven by a Pyserini index that can be deployed for free on the Hugging Face Hub. In what follows, we deconstruct an example script[4] to showcase the different features enabled by Spacerini. When run end-to-end, the script pulls a dataset from Hugging Face, pre-processes it, indexes it, creates a gradio-based search interface and deploys that as a Hugging Face Spaces demo. This is only meant as a feature-complete demo, and we don't expect most people to want to integrate every step into their workflows, but rather to cherry-pick and decide what best to use depending on context.

3.1 Loading Data

All our workflows are backed by the Hugging Face `datasets` library (Lhoest et al., 2021), itself based on the extremely efficient Apache Arrow format. `Datasets` is a mature library which provides a standardized interface to any tabular dataset, in particular, to tens of thousands of community datasets hosted on the Hugging Face Hub[5]. The `datasets` library gives fine-grained control over the lifecycle of tabular datasets, which we choose to abstract away through a set of opinionated data loading functions that cover the use cases we deem relevant to information retrieval. We also add new functionality, such as the ability to load any document dataset from the `ir_datasets` library using for example the following one-liner to load MS MARCO (Nguyen et al., 2016) as a Hugging Face `datasets.Dataset` object using a function from the `data` subpackage:

```
from spacerini.data import
    load_ir_dataset

hf_dset = load_ir_dataset("msmarco-
    passage")
```

We include wrappers to load database tables, pandas DataFrames (pandas development team, 2020; Wes McKinney, 2010), and text datasets on disk, as well as the ability to load any dataset either in memory-mapped mode or in streaming mode: the former makes it possible to handle larger-than-memory datasets, and the latter larger-than-disk datasets that can be streamed from a remote location such as the Hugging Face Hub.

3.2 Pre-processing

Spacerini also provices a `preprocess` subpackage which offers a range of customizable preprocessing options for preparing datasets. This module includes a sharding utility that enables the partitioning of large datasets into smaller, more manageable chunks for efficient parallel processing.

```
from spacerini.preprocesss import
    shard_dataset
shard_dataset(
    hf_dataset=hf_dset,
    shard_size="1GB",
    column_to_index="text",
    shards_paths="msmarco-shards",
)
```

3.3 Indexing

Spacerini's `index` subpackage leverages Pyserini to provide very efficient Lucene indexing and allow users to easily and quickly index large datasets, either sharded in the pre-processing step, or any text format accepted by Pyserini, and streaming text datasets, such as those returned by Spacerini's `data` subpackage. This subpackage also exposes several tokenization options using existing language-specific analyzers[6] as well as Hugging Face subword tokenizers (MOI et al., 2022).

```
from spacerini.index import
    index_json_shards

index_json_shards(
    shards_path="msmarco-shards",
    index_path="app/index",
)
```

[3] https://c4-search.apps.allenai.org/

[4] https://github.com/castorini/hf-spacerini/blob/main/examples/scripts/gradio-demo.py

[5] https://hf.co/datasets

[6] https://lucene.apache.org/core/9_5_0/analysis/common/

We also provide wrappers to Pyserini's dense and hybrid retrieval functionality through the `spacerini.index.encode` subpackage.

3.4 Template-based Search Interfaces

Having indexed a collection, one can easily spin up a frontend using the `frontend` subpackage and one of many provided templates[7]. These are built using `cookiecutter`,[8] a Python templating library for software projects. We provide a few batteries-included frontend templates based both on the Gradio[9] and Streamlit [10] demo app frameworks, both of which are natively supported by Hugging Face `Spaces`. Figure 1 showcases a search engine built using one of our Streamlit templates.

```
from spacerini.frontend import
    create_app

cookiecutter_vars = {
    "dset_text_field": "text",
    "metadata_field": "docid",
    "space_title": "MS MARCO Search",
    "local_app": "app"
}
create_app(
    template="gradio-vanilla",
    extra_context_dict=cookiecutter_vars
    ,
    output_dir=".",
)
```

3.5 Deployment to Hugging Face Spaces

The local apps developed in the previous subsection can then be pushed to Hugging Face Spaces and hosted there for free. One can then further customize the running app from the browser, for example to add functionality not provided by the chosen template. The goal of the templates is to provide a useful starting point in the form of a running app that users can further customize with interface features useful for their own workflows.

```
from spacerini.frontend import
    create_space_from_local

create_space_from_local(
    space_slug="msmarco-passage-search,
    organization="spacerini,
    space_sdk="gradio",
    local_dir=LOCAL_APP,
    delete_after_push=True,
)
```

3.6 Sharing Indexes as Hugging Face Datasets

Orthogonal to the workflow presented so far, is the ability to upload Lucene indexes to the Hugging Face Hub using shareable dataset repositories and enabling reproducible retrieval experiments.

```
from spacerini.index import
    push_index_to_hub

push_index_to_hub(
    dataset_slug="lucene-english-
    analyzer-msmarco",
    index_path="index",
)
```

Any hosted index can then just as easily be downloaded for local use:

```
from spacerini.index import
    load_index_from_hub

index_path = load_index_from_hub("lucene
    -fr-analyzer-")
```

3.7 Search and Pagination

Search features are provided by the `search` subpackage and leverage the memory-mapping feature of Arrow tables to load the entire table of results—no matter how big—only materializing the specific shard that corresponds to the requested result page.

```
from spacerini.search import
    result_indices, result_page

ix = result_indices(
    "Lorem Ipsum",
    num_results=1_000,
    INDEX_PATH,
)

last_results_page = result_page(
    hf_dset,
    ix,
    page=-1,
    results_per_page=20,
).to_pandas()
```

4 Use Cases and Demonstrations

We envision Spacerini to be useful primarily to NLP researchers, students, shared task organizers, data scientists, and data annotators, as well as tech-adjacent and -proficient professionals and laypeople. In what follows, we overview a series of 13 use-cases that we implemented and how they might benefit their respective targeted audience. An overview of all 13 search engines can be found

[7] https://github.com/castorini/hf-spacerini/tree/main/templates

[8] https://github.com/cookiecutter/cookiecutter

[9] https://gradio.app/

[10] https://streamlit.io/

at https://huggingface.co/spacerini; in following inline links to the selected engines, this part of their URL prefix is shortened to '`hf`' for brevity.

4.1 NLP researchers

Spacerini is designed to enable qualitative analysis of large-scale textual corpora without the need for extensive engineering work. It can be used in dataset auditing campaigns, such as those carried out by Kreutzer et al. (2022) or in data annotation efforts. Also relevant here are generative text models whose outputs can be better understood by better understanding the datasets they were trained on. We refer the reader to Section 5 of Piktus et al. (2023) for a detailed exposé of potential use scenarios which include: PII detection, problematic content detection, social representation, benchmark and language contamination detection, as well as plagiarism and memorization detection. An example demo for this context is the index of the XSUM (Narayan et al., 2018) dataset which is indexed and can be explored with the `hf/xsum-search` demo.

4.2 IR researchers

Given its tight integration with Pyserini, Spacerini can also be leveraged by IR researchers to experiment with modifications of their retrieval pipelines in user studies or to deploy demos of their working prototypes. Reproducibility for IR experiments is further enhanced thanks to the index sharing abilities showcased in Section 3.6. As a practical example, Spacerini was used in the context of the BigCode project (Li et al., 2023) to quickly experiment with multiple n-gram tokenization schemes for BM25-based code retrieval; this corresponds to the `hf/spacerini/code-search` demo.

4.3 Linguists

Corpus linguistics relies on qualitative analyses of text corpora to understand language and its many varieties by studying the way it is used (McEnery and Hardie, 2011; Piktus et al., 2023). Some of these empirical analyses can be facilitated by the usage of an inverted index, coupled with the correct querying patterns and frontend elements, both of which are easy to achieve using Spacerini.

4.4 Digital Humanists

Spacerini can also be leveraged by Digital and Computational Humanists, Archivists and Librarians looking to index their collections. Indeed, GLAM (Galleries, libraries, archives, and museums) collections are increasingly being made available as datasets. Furthermore, there is a growing interest in the digital humanities in training and using languages models, as demonstrated by the success of projects such as the *AI for Humanists Project*.[11] In this context, indexing data relevant to these efforts is a difficult task; often project-based and contingent upon precarious funding arrangements. Having a project-agnostic tool like Spacerini could prove valuable to this community and a useful addition to toolkits such as the GLAM Workbench (Sherratt, 2021).

4.5 IR Students

Given its low barrier of entry, Spacerini can be a good tool for IR courses, where students could be tasked with developing search engines, by providing an easy-to-deploy frontend interface for their developed retrieval systems which does not even have to be deployed within the same application, as demonstrated by the `hf/chat-noir`, a frontend wrapper for ChatNoir (Bevendorff et al., 2018).

4.6 Shared task organizers

Spacerini can also be leveraged by organizers of shared tasks such as MIRACL (Zhang et al., 2022) and Touché (Bondarenko et al., 2022), who want to help participants explore the datasets without forcing them to download large volumes of data or giving participants full access to the data: it is indeed possible to host the index privately on the Hugging Face Hub and only expose access to it through a search interface. Spacerini can also be used as a platform for participants to deploy working prototypes of their submissions with a unified interface provided by the organizer as a `cookiecutter` template. Example demos for this use case include `hf/miracl-bengali`, `hf/miracl-arabic`, and `hf/miracl-swahili`.

4.7 Tech journalists

Spacerini can help data journalists and digital investigative journalists index, explore, and understand open data, in a similar vein to the functionality provided by the Aleph suite.[12] Providing technical tools to data journalists is a crucial in uncovering matters of public interest, as was evident by role played by the collaborative use of the Neo4j graph

[11] https://www.bertforhumanists.org/
[12] https://docs.alephdata.org/

database in unraveling corrupt networks surrounding tax havens (Díaz-Struck and Cabra, 2018).

4.8 Additional usage patterns

Finally, three features of Hugging Face Spaces make them especially attractive for users: (1) they can leverage private datasets, meaning that one can provide search access to a dataset without sharing the underlying data,(2) they can be seamlessly embedded into HTML, specifically Gradio-based Spaces which can be embedded as *Web Components*[13] so that users can easily integrate a Spacerini-based search feature into their own websites[14], and (3) Gradio-based Spaces expose a FastAPI[15] endpoint that can be queried to access the functionality of the space, making deployed search engines accessible through HTTP calls.

5 Limitations and Future Plans

The main limitation of the off-the-shelf variant of Spacerini is the disk space limit imposed by Hugging Face Spaces, which is currently set to 50 GB for the free tier.[16] While not enough to accommodate entire corpora such as ROOTS or The Pile, such datasets are typically amalgamations of constituent datasets which can each be studied independently. This limit has no bearing on Spacerini search apps deployed locally. Should users still want to get more disk space for their Spaces-hosted indexes, they can either pay for an upgrade to a more appropriate tier or see whether they qualify for a free hardware upgrade through the community grants offered by Hugging Face, in the `Settings` pane of the relevant space.

Planned improvements include automating the creation of dataset cards (or rather "index cards") when pushing an index to the Hugging Face Hub, better documentation, as well as more fine-grained tokenization support.

Please also note that Spacerini is currently in active development and that the stability of its current API and subpackages isn't guaranteed not to involve breaking changes as we converge toward the first stable version release. We also look forward to community contributions both to the codebase and to the frontend templates, as well as in the form of actual use cases of the library that culminate in the deployment of search apps.

6 Conclusion

We presented Spacerini, a modular framework that enables the quick and free deployment and serving of template-based search indexes as interactive applications for ad-hoc exploration of text datasets. The need for such a tool is especially pressing as large language models have come to consume inordinate amounts of text data, reinforcing the need for a qualitative exploration and understanding of datasets to assess them in a way that is impenetrable to quantitative analyses alone.

Spacerini leverages features from both the `Pyserini` toolkit and the Hugging Face ecosystem to facilitate the creation and hosting of user-friendly search systems for text datasets. Users can easily index their collections and deploy them as ad-hoc search interfaces, making the retrieval of relevant data points a quick and efficient process. The user-friendly interface enables non-technical users to effectively search massive datasets, making Spacerini a valuable tool for anyone looking to audit their text collections qualitatively. The framework is open-source and available on GitHub under `gh/castorini/hf-spacerini` and demo search apps can be found under `hf/spacerini`

The key advantage of Spacerini is its ability to simplify the search process, allowing researchers to conduct quick and efficient audits, while abstracting away all the minutiae of indexing data or hosting services. We believe that this provides an opportunity for collaboration and transparency in IR and NLP research. With the creation and sharing of search indexes publicly, practitioners, researchers and the general public can work together to pinpoint problematic content, find duplicates, and identify biases in datasets.

Finally, we emphasize that Spacerini is a first step in the direction of systematic dataset auditing, and more work is still needed to create standardized structures that leverage tools such as ours to properly document the different axes of interest.

Acknowledgments

We are grateful to Lukas Gienapp, Daniel van Strien, Yuvraj Sharma, Omar Sanseviero, Julien Chaumond, Lucain Pouget, Abubakar Abid, and Leandro von Werra for their tireless support and invaluable advice throughout this project.

[13] https://developer.mozilla.org/en-US/docs/Web/Web_Components

[14] e.g.: https://cakiki.github.io/search-engine/

[15] https://github.com/tiangolo/fastapi

[16] https://huggingface.co/docs/hub/spaces-overview#hardware-resources

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilic, and Yacine Jernite. 2022. Bigscience: A case study in the social construction of a multilingual large language model. *CoRR*, abs/2212.04960.

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anne Beaulieu and Sabina Leonelli. 2021. *Data and Society: A Critical Introduction*. Sage.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer.

Emilia Díaz-Struck and Mar Cabra. 2018. *Uncovering International Stories with Data and Collaboration*, pages 55–65. Springer International Publishing, Cham.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *CoRR*, abs/1603.08832.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *CoRR*, abs/1802.00393.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling.

Tony Hey, Stewart Tansley, Kristin Tolle, and Jim Gray. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Ulrich Krohs. 2012. Convenience experimentation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):52–57. Data-Driven Research in the Biological and Biomedical Sciences On Nature and Normativity: Normativity, Teleology, and Mechanism in Biological Explanation.

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. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sabina Leonelli. 2020. Scientific Research and Big Data. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2020 edition. Metaphysics Research Lab, Stanford University.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy V, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Moustafa-Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you! *CoRR*, abs/2305.06161.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. 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.

Tony McEnery and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2022. Measuring data. *CoRR*, abs/2212.05129.

Anthony MOI, Nicolas Patry, Pierric Cistac, Pete, Funtowicz Morgan, Sebastian Pütz, Mishig, Bjarte Johansen, Thomas Wolf, Sylvain Gugger, Clement, Julien Chaumond, Lysandre Debut, François Garillot, Luc Georges, dctelus, JC Louis, MarcusGrass, Taufiquzzaman Peyash, 0xflotus, Alan deLevie, Alexander Mamaev, Arthur, Cameron, Colin Clement, Dagmawi Moges, David Hewitt, Denis Zolotukhin, and Geoffrey Thomas. 2022. huggingface/tokenizers: Rust 0.13.2.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Danna Niezni, Hillel Taub-Tabib, Yuval Harris, Hagit Sason-Bauer, Yakir Amrusi, Dana Azagury, Maytal Avrashami, Shaked Launer-Wachs, Jon Borchardt, M Kusold, Aryeh Tiktinsky, Tom Hope, Yoav Goldberg, and Yosi Shamay. 2022. Extending the boundaries of cancer therapeutic complexity with literature data mining. *bioRxiv*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Tim Sherratt. 2021. Glam workbench.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. 2022. Metadata archaeology: Unearthing data subsets by leveraging training dynamics. *CoRR*, abs/2209.10015.

Vuk Vuković, Akhil Arora, Huan-Cheng Chang, Andreas Spitz, and Robert West. 2022. Quote erat demonstrandum: A web interface for exploring the quotebank corpus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and Jimmy Lin. 2020. Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 31–41, Online. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages.