Science of Computer Programming 234 (2024) 103079

Contents lists available at ScienceDirect

Science of Computer Programming

journal homepage: www.elsevier.com/locate/scico

Original software publication

HFCommunity: An extraction process and relational database to analyze Hugging Face Hub data

Adem Ait [a,*], Javier Luis Cánovas Izquierdo [a], Jordi Cabot [b]

[a] IN3 – UOC, Barcelona, Spain
[b] Luxembourg Institute of Science and Technology, University of Luxembourg, Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Keywords:
Mining software repositories
Data analysis
Hugging Face

ABSTRACT

Social coding platforms such as GitHub or GitLab have become the *de facto* standard for developing Open-Source Software (OSS) projects. With the emergence of Machine Learning (ML), platforms specifically designed for hosting and developing ML-based projects have appeared, being Hugging Face Hub (HFH) one of the most popular ones. HFH aims at sharing datasets, pre-trained ML models and the applications built with them. With over 400 K repositories, and growing fast, HFH is becoming a promising source of empirical data on all aspects of ML project development. However, apart from the API provided by the platform, there are no easy-to-use solutions to collect the data, nor prepackaged datasets to explore the different facets of HFH. We present HFCommunity, an extraction process for HFH data and a relational database to facilitate an empirical analysis on the growing number of ML projects.

Code metadata

Code metadata description	
Current code version	*v1.1*
Permanent link to code/repository used for this code version	*https://github.com/ScienceofComputerProgramming/SCICO-D-23-00219*
Permanent link to Reproducible Capsule	*https://github.com/SOM-Research/HFCommunity*
Legal Code License	*Creative Commons Attribution-ShareAlike 4.0 International License*
Code versioning system used	*Git*
Software code languages, tools, and services used	*Python, Git and MariaDB*
Compilation requirements, operating environments and dependencies	*Python v3.9.12, MariaDB v10.3.39, Git v2.20.1*
If available, link to developer documentation/manual	*Website: https://som-research.github.io/HFCommunity* *Manual: https://som-research.github.io/HFCommunity/docs/*
Support email for questions	*ait_mimoune@uoc.edu*

1. Motivation and significance

Hugging Face Hub (HFH) is a social coding platform created as a specific solution for hosting and developing ML-based projects. In contrast with other general-purpose platforms such as GitHub or GitLab, HFH aims at sharing datasets, pre-trained ML models

* Corresponding author.
E-mail addresses: aait_mimoune@uoc.edu (A. Ait), jcanovasi@uoc.edu (J.L. Cánovas Izquierdo), jordi.cabot@list.lu (J. Cabot).

https://doi.org/10.1016/j.scico.2024.103079
Received 25 July 2023; Received in revised form 3 January 2024; Accepted 8 January 2024

Available online 10 January 2024

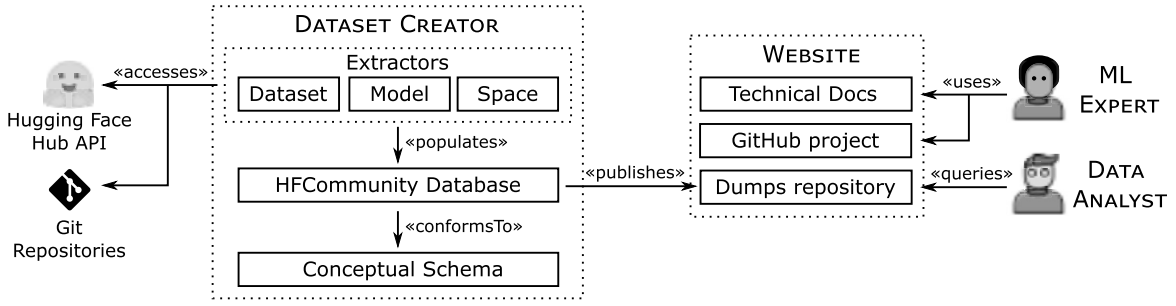

Fig. 1. Architecture of HFCommunity.

and applications built with them (*spaces* in HFH terminology). As of July 2023, HFH hosts more than 400 K repositories, and this number is growing fast. However, current access to HFH is only available programmatically via the official API, which hampers the collection and exploration of the different facets of the platform. Furthermore, the HFH API does not provide information about the evolution of files (e.g., commits and modifications), which is a key aspect for understanding the development of ML projects.

In this software paper, we present the extensions to the architecture, extraction process and database behind HFCommunity [1], a tool that enables researchers to discover HFH data and community insights by collecting and sharing HFH data via a relational database. HFCommunity provides domain-specific concepts such as models, datasets, and spaces, thus facilitating its exploration and querying via SQL-like languages.

Most works on extraction processes target on general-purpose platforms like GitHub (e.g., GHTorrent [2] and GitHub Archive[1]). There are also API crawler tools such as Prometheus [3] or GHCrawler,[2] which facilitate the retrieval of platform data via its API. Nevertheless, to the best of our knowledge, ours is the first full software pipeline aimed at facilitating the analysis of HFH data, thus complementing these previous approaches.

2. Software description

HFCommunity is a dataset built via a data collection process relying on HFH API and Git. While the former provides information about platform and community activity, the latter covers data about project file evolution. Next, we describe the dataset components and the data collection process.

2.1. Software architecture

Fig. 1 shows the architecture of HFCommunity, which is composed of two main components: the *Dataset Creator* and the *Website*. The former queries HFH data and Git repositories to create a snapshot of HFCommunity (see *HFCommunity Database*), while the latter includes the main documentation and dumps to be downloaded. We foresee that ML Experts may launch the *Dataset Creator* to create their own version of HFCommunity, while Data Analysts may download the dumps to perform their analysis.

2.2. Software functionalities

HFCommunity tool offers the following functionalities:

Extraction Process. This process is performed by the *Dataset Creator*, which includes extractors for HFH data elements (i.e., datasets, models, and spaces) and a database importer to store the extracted data. The extraction process supports incremental updates, that is, only the new data with regard to the last snapshot is actually extracted. The resulting database conforms to the HFCommunity conceptual schema, which includes entities and relationships to query HFH data[3] (e.g., model, dataset or space elements).

Community Analysis. HFCommunity datasets are provided as relational database dumps, which can be queried via SQL-like languages or data analytics tools to enable empirical studies on ML projects. Further insights or analysis can be portrayed by data analysts, being HFCommunity a facilitator of the data extraction process. Section 3 includes illustrative examples.

Tool & Dataset Download. The *Website* provides a landing page with essential information about HFCommunity and links to (1) the technical documentation, which describes how to launch the *Dataset Creator* and its main components; (2) the GitHub repository of the tool with the ready-to-use scripts; and (3) the latest HFCommunity dataset dumps to be downloaded. A new release of a HFCommunity dump is published every month.

[1] https://www.gharchive.org/.

[2] https://github.com/Microsoft/ghcrawler.

[3] More information about the HFCommunity conceptual schema can be found at [1].

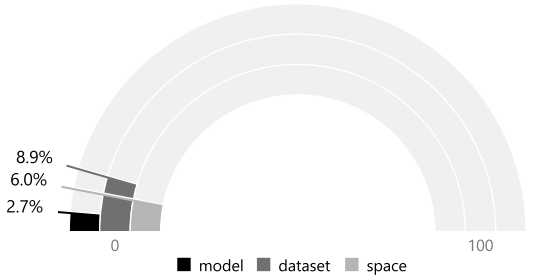

Fig. 2. Percentage of repositories with discussions in HFH.

```
SELECT
  COUNT(DISTINCT d.repo_id) AS num_repos,
  r.type
FROM discussion d
INNER JOIN repository r
ON d.repo_id=r.id
GROUP BY r.type
```

Listing 1: SQL query for metric shown in Fig. 2.

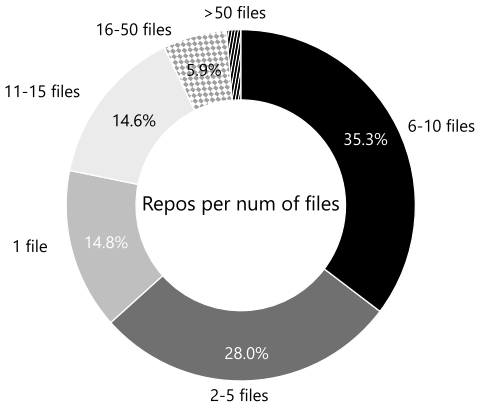

Fig. 3. Number of files in a repository of HFH.

```
SELECT
  files_in_repo AS num_files,
  COUNT(*) AS num_repos
FROM (SELECT repo_id,
             COUNT(*) AS files_in_repo
      FROM file f
      GROUP BY repo_id) s
GROUP BY files_in_repo
```

Listing 2: SQL query for metric shown in Fig. 3.

3. Illustrative examples

Data available in HFCommunity enables the calculation of interesting metrics which may help to understand the dynamics of ML projects in HFH. For instance, discussions enable the collaboration in HFH repositories, thus promoting the communication and interaction between contributors. HFCommunity can be used to measure the usage of discussions according to the repository type. Fig. 2 shows the results of this metric and shows that very few repositories leverage this functionality. This metric can be computed thanks to a simple SQL query, shown in Listing 1.

Another example is the number of files in HFH repositories, which allows us to visualize the typical number of files found in a repository, and also detect empty repositories. Fig. 3 shows the results of this metric, and the Listing 2 shows the SQL query, revealing that almost a half of HFH repositories have less than 5 files. Furthermore, 14.8% of all repositories have only 1 file, which may reveal toy or test projects, or that the actual development may be performed on other platforms (e.g., GitHub).

4. Impact

Even though HFCommunity has been recently released, we believe it will have a significant impact on the ML and empirical software communities.

Enabling empirical studies for ML projects. We believe that HFCommunity opens the door to new empirical studies focused on ML and AI projects, to complement existing literature (e.g., [4]) and to enable the replication of studies in other platforms (e.g., [5,6]), thus allowing understanding the dynamics of this kind of projects and communities. Furthermore, it could have facilitated the data extraction of recently published studies (e.g., [7,8]).

Performing longitudinal studies in HFH. As HFCommunity is released on a monthly basis, it enables the study of the evolution of the HFH platform, thus allowing its comparison with the evolution of other platforms. For instance, the study of the Diffusion of Innovation [9] may help to study whether the platform is gaining or losing momentum.

Reference conceptual schema for code-hosting platforms. The conceptual schema used in the HFCommunity [1] may be used as a reference schema for other code-hosting platforms, such as GitHub and GitLab, thus helping to create a common representation for their main entities and relationships, and serve as a starting point to offer multi-platform extractors.

5. Conclusions

In this paper, we have presented HFCommunity, a relational database collecting the information about the HFH repositories and community discussions, together with the process to populate this database by extracting information from HFH artifacts and their corresponding Git repositories.

As future work, we are interested in applying NLP-based techniques to extract additional information from repository descriptions, which may be useful to include fine-grained annotation data for explainability analysis [10]. We also plan to leverage on HFCommunity as a data source to perform empirical studies already published in GitHub to compare HFH to other code hosting platforms, along with the evaluation of the effectiveness of our approach. Finally, we will consider other sources to enrich HFCommunity such as PapersWithCode,[4] which hosts ML articles and their metadata, to detect mirror repositories in other platforms.

CRediT authorship contribution statement

Adem Ait: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Javier Luis Cánovas Izquierdo:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Jordi Cabot:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is part of the project TED2021-130331B-I00 funded by MCIN/AEI/10.13039/501100011033; and BESSER, funded by the Luxembourg National Research Fund PEARL program, grant agreement 16544475.

References

[1] A. Ait, J. Cánovas Izquierdo, J. Cabot, HFCommunity: a tool to analyze the Hugging Face Hub Community, in: Int. Conf. on Software Analysis, Evolution and Reengineering, 2023, pp. 728–732.
[2] G. Gousios, The GHTorrent dataset and tool suite, in: Working Conf. on Mining Software Repositories, 2013, pp. 233–236.
[3] A. Jobst, D. Atzberger, T. Cech, W. Scheibel, M. Trapp, J. Döllner, Efficient github crawling using the graphql API, in: Computational Science and Its Applications, vol. 13381, 2022, pp. 662–677.
[4] D. Gonzalez, T. Zimmermann, N. Nagappan, The state of the Ml-universe: 10 years of artificial intelligence & machine learning software development on GitHub, in: Int. Conf. on Mining Software Repositories, 2020, pp. 431–442.
[5] A. Ait, J.L. Cánovas Izquierdo, J. Cabot, An empirical study on the survival rate of GitHub projects, in: Int. Conf. on Mining Software Repositories, 2022, pp. 365–375.
[6] V.N. Subramanian, An empirical study of the first contributions of developers to open source projects on GitHub, in: Int. Conf. on Software Engineering, 2020, pp. 116–118.
[7] W. Jiang, N. Synovic, M. Hyatt, T.R. Schorlemmer, R. Sethi, Y. Lu, G.K. Thiruvathukal, J.C. Davis, An empirical study of pre-trained model reuse in the hugging face deep learning model registry, CoRR, arXiv:2303.02552, 2023.

[4] https://paperswithcode.com/.

[8] J. Castaño, S. Martínez-Fernández, X. Franch, J. Bogner, Exploring the carbon footprint of hugging face's ML models: a repository mining study, CoRR, arXiv: 2305.11164, 2023.
[9] E.M. Rogers, Diffusion of Innovations, 5 ed., Free Press, 2003.
[10] J. Giner-Miguelez, A. Gómez, J. Cabot, DescribeML: a tool for describing machine learning datasets, in: Int. Conf. on Model Driven Engineering Languages and Systems: Companion Proceedings, 2022, pp. 22–26.