Github Repo Insights Dataset

Abhist Yadav

April 2024

1 Dataset

Github Repo Insights Dataset "The GitHub Repo Insights Dataset offers a comprehensive collection of data pertaining to various GitHub repositories. The motivation behind curating this dataset stems from the profound impact of open-source development on the global software landscape. GitHub has emerged as a central hub for developers to share code, collaborate on projects, and advance technology collectively. By compiling data from GitHub repositories related to different topics, we aim to facilitate deeper insights into the trends, technologies, and community engagement strategies driving innovation across various domains. By providing access to data such as project metadata, contributor activity, stars, and repository URLs, this dataset serves as a valuable resource for researchers, developers, and data enthusiasts."

2 Template

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to facilitate research or analysis related to GitHub activity and repository popularity across various topics. Its purpose likely includes studying trends in user engagement, identifying popular repositories, and potentially analyzing the relationship between stars (indicative of popularity) and other factors like user activity or repository characteristics. The gap it fills could be providing structured data for researchers or analysts interested in understanding GitHub dynamics and trends across different domains.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created as part of a project within our institution for a course-related activity.

Who funded the creation of the dataset?

NA

Any other comments? NA

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances within the dataset represent GitHub repositories of 30 topics. Each instance corresponds to a specific repository hosted on GitHub, containing information such as the username of the repository owner, the repository name, the number of stars received, and the repository URL. The dataset is homogeneous, consisting solely of repositories, and does not include multiple types of instances like users, interactions between users.

How many instances are there in total (of each type, if appropriate)?

Since the dataset comprises GitHub repositories, the total number of instances corresponds to the total number of repositories included in the dataset. Each of the 30 csv files contains 20 top entries based on number of stars.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset represents a sample of instances rather than encompassing all possible GitHub repositories. This sample, consisting of 600 instances across 30 CSV files (each representing a different topic).There is no special consideration made while choosing the topics. Each topic is chosen in alphabetical order.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in the dataset consists of structured data, specifically featuring information related to GitHub repositories. This structured data includes the following features:

Username: The username of the repository owner.

Repo Name: The name of the repository.

Stars: The number of stars received by the repository, indicative of its popularity.

Repo URL: The URL or link to the repository on GitHub.

These features provide essential information about each GitHub repository and can be utilized for various analytical purposes, such as studying user engagement, identifying popular repositories, or exploring trends across different topics on GitHub.

Is there a label or target associated with each instance? If so, please provide a description.

There is no explicit label or target associated with each instance in the dataset. The dataset is primarily focused on describing and analyzing

GitHub repositories rather than predicting a particular outcome or target variable.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The dataset contains some missing values in certain columns across some CSV files. Some repositories may have had missing data for certain columns due to the nature of web scraping. For instance, if a repository owner has chosen to keep certain details private or has restricted access to certain information, the corresponding data fields in the dataset may contain missing values. Web scraping involves extracting data from HTML web pages, and the structure of these pages can vary widely across different repositories. As a result, inconsistencies or missing values may occur due to changes in the layout or organization of the web pages being scraped.During the web scraping process, issues such as parsing errors or incomplete data extraction may result in missing values in the dataset. This could occur due to variations in data formatting or unexpected changes in the website's markup.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

NA

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No specific recommendations

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Noise in the dataset could stem from various sources, including irrelevant or extraneous information scraped from web pages, typographical errors, or inconsistencies in data formatting.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Since the dataset was created via web scraping, it relies on external resources, specifically GitHub repositories and their associated web pages. There are no guarantees that the external resources (GitHub repositories) will remain constant over time. Changes to repository URLs, content, or access permissions could affect the dataset's reliability and consistency. Future users of the dataset should be aware of any restrictions associated with accessing or using the external re-

sources. These restrictions could include licenses such as GPL, MIT, or proprietary licenses, which dictate how the data can be used and redistributed.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset consists of publicly available information from GitHub repositories, such as usernames, repository names, stars, and repository URLs. Since this information is publicly accessible on GitHub, it is unlikely to contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Since the dataset consists of information related to GitHub repositories, such as usernames, repository names, stars, and repository URLs, it is unlikely to contain data that, if viewed directly, might be offensive, insulting, or threatening. However, the content of individual repositories on GitHub could potentially include offensive or sensitive material, depending on the nature of the repository and its content.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people indirectly through the inclusion of usernames associated with GitHub repositories. Each entry in the dataset includes the username of the repository owner, providing information about the individuals or organizations behind the repositories. Therefore, the dataset does relate to people, albeit in an indirect manner through their association with GitHub repositories.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not appear to directly identify sub-populations by characteristics such as age or gender. Instead, it primarily focuses on information related to GitHub repositories, such as usernames, repository names, stars, and repository URLs.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

It's unlikely that the dataset would directly identify individuals, as it primarily consists of information related to GitHub repositories. However, indirect identification of individuals could potentially occur if the usernames in the dataset contain personally identifiable information (PII) or if combined with other external data sources.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic

data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset itself does not appear to contain sensitive data such as racial or ethnic origins, sexual orientations, religious beliefs, political opinions, financial or health data, biometric or genetic data, government identification numbers, or criminal history.

Any other comments? NA

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance in the dataset was acquired through web scraping of GitHub repositories. Web scraping involves programmatically extracting data from HTML web pages, in this case, GitHub pages, to collect information such as usernames, repository names, stars, and repository URLs. Since the data was obtained through web scraping, it can be considered indirectly inferred/derived from other data sources, specifically GitHub's web pages. The information collected was not directly observable but rather extracted from the HTML structure of the web pages.Validation and verification of the scraped data was done through data integrity checks.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data collection for this dataset was conducted using software programs specifically designed for web scraping. These programs utilize techniques to programmatically access and extract information from HTML web pages, in this case, GitHub pages. Overall, the validation of the data collection mechanisms likely involved a combination of software testing, manual verification, and ongoing monitoring to ensure the accuracy, reliability, and integrity of the scraped data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

NA

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

I was directly involved in the data collection process, utilizing web scraping techniques to gather information from GitHub repositories.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the

timeframe in which the data associated with the instances was created.

The creation timeframe of the data associated with the instances would align with the time when the repositories were scraped.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Since the data collection process involved web scraping of publicly available information from GitHub repositories, No formal ethical review processes, such as those conducted by institutional review boards (IRBs), were required.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset indirectly relates to people through the inclusion of usernames associated with GitHub repositories.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

In this case, the data was obtained from a third-party source (GitHub's website) rather than collected directly from the individuals associated with the repositories.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No, the individuals in question were not notified about the data collection.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Since there was no notification or request for consent from the individuals regarding the collection and use of their data, there was no explicit consent obtained from them.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

NA

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

NA

Any other comments? NA

Preprocessing/cleaning/labeling

Was any preprocessing / cleaning / labeling of the data done (e.g., dis-

cretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Addressing missing values in columns like username, repository name, stars, or repository URL by imputation or removal, depending on the extent and impact of missing data.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

NA

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

NA

Any other comments? NA

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No, the dataset hasn't been used for any tasks.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

NA

What (other) tasks could the dataset be used for?

Calculating similarity measures between repositories based on features such as repository names, descriptions, or user interactions to identify similar projects or recommend related repositories to users.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset's composition may not fully represent the diversity of GitHub repositories. Depending on the topics or criteria used for data collection, certain types of repositories or users may be overrepresented while others are underrepresented. Future users should be cautious about generalizing findings based solely on this dataset.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset contains information about GitHub repositories and their associated usernames. However, it's important to avoid using this information for individual profiling or making assumptions about individuals based solely on their association with specific repositories. Attempting to profile individuals based on their GitHub activity could lead to unfair treatment or privacy concerns.

Any other comments? NA

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

As part of a project within an academic institution as part of a course, the dataset will be provided to the course instructor.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

In the form of a github repo link.

When will the dataset be distributed?

27 April 2024

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

NA

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

NA

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

NA

Any other comments? NA

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be supported, hosted, and maintained solely by myself. As the creator of the dataset, I am responsible for ensuring its availability, reliability, and ongoing maintenance. This includes tasks such as data storage, updating, and addressing any issues or inquiries related to the dataset. By maintaining ownership and control of the dataset, I can oversee its use and integrity, ensuring it remains accessible for future reference or analysis.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Via email : y.abhist@iitg.ac.in

Is there an erratum? If so, please provide a link or other access point.

NA

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

NA

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in

question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Since the dataset primarily relates to GitHub repositories and contains information about usernames associated with those repositories, rather than directly identifying individuals, there may not be specific limits on the retention of data associated with the instances. However, if there are any applicable privacy policies or regulations governing the retention of user data, such as those outlined in GitHub's terms of service or relevant data protection laws, it's important to adhere to those requirements.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

NA

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

On the repository of the dataset itself individuals can submit their contributions to the dataset. This platform is a version control system , where users can fork the dataset, make their modifications or additions, and submit pull requests to propose changes.

Any other comments? NA