Datasheet for the Project Gutenberg dataset*

Aryaman Suri

March 16, 2024

The Gutenberg Project dataset (Hart 2021) provides a rich source of data for analyzing Villette, a novel authored by Charlotte Brontë. This dataset, curated by the Gutenberg Project, encompasses a wide array of literary texts, including Villette, and serves as a valuable resource for literary analysis and computational studies. Project Gutenberg, was created by Michael S. Hart in 1971. It was one of the earliest initiatives to digitize and archive cultural works to make them freely available to the public. Michael S. Hart is widely regarded as a pioneer of e-books and digital libraries. The project was named after Johannes Gutenberg, the inventor of the movable-type printing press, which revolutionized the production of books in Europe during the 15th century.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *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 Gutenberg Project dataset was created to digitize and archive literary works in the public domain, primarily aiming to make these texts freely accessible to the public for educational and informational purposes. It addressed the gap in accessibility to classic literature and culturally significant texts that were no longer under copyright protection but were still challenging to access.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The Gutenberg Project dataset was created by volunteers associated with Project Gutenberg, a nonprofit organization founded by Michael S. Hart. It was not created by a specific team or research group but rather through collaborative efforts of volunteers worldwide under the umbrella of Project Gutenberg.

*Code and data are available at: https://github.com/aryamansuri/Letters-used-in-Villette.git

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The creation of the Gutenberg Project dataset was primarily funded through donations and volunteer efforts. Since Project Gutenberg operates as a nonprofit organization, it does not rely on external grants for funding. Therefore, there is no associated grant, grantor, or grant number for the creation of the dataset.

4. *Any other comments?*

    - no.

Composition

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

    - The Gutenberg Project dataset comprises instances representing textual documents, predominantly literary works in the public domain, including novels, essays, poetry collections, and plays. While the dataset primarily consists of documents, there may be variations within each text, such as chapters or sections, along with associated metadata like author names, publication dates, genres, and languages.

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

    - The Gutenberg Project dataset contains tens of thousands of instances, with each instance representing a separate textual document. The exact number of instances may vary over time as new texts are added or updated within the dataset.

3. *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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The Gutenberg Project dataset is not a comprehensive collection of all possible instances but rather a curated selection of texts that are in the public domain. The dataset can be considered a sample, but not necessarily a random one, from the larger set of public domain literature. The selection of texts is based on availability, accessibility, and public domain status rather than on specific criteria for representativeness. As such, the dataset may not be fully representative of the entire corpus of public domain literature, but it provides a substantial and diverse collection of texts for analysis and study.

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

    - Each instance within the Gutenberg Project dataset consists of raw textual data, which typically includes the full text of a literary work in the public domain. This raw text may encompass the entire content of a novel, essay, poetry collection, play, or other literary form. The data does not typically include pre-processed features but instead provides the original text as it appears in the source document, preserving the linguistic and stylistic characteristics of the author's writing.

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

    - In the Gutenberg Project dataset, there is typically no specific label or target associated with each instance. The dataset primarily consists of raw textual data, such as novels, essays, poetry collections, and plays, without predefined labels or targets. Each instance represents a standalone text.

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

    - In general, individual instances within the Gutenberg Project dataset typically do not contain missing information. However, occasional errors or omissions may occur during the digitization process, resulting in incomplete sections of text. These instances of missing information are usually unintentional and may arise due to factors such as document degradation, scanning errors, or incomplete transcription.

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

    - In the Gutenberg Project dataset, relationships between individual instances are not typically made explicit. Each instance represents a standalone textual document, and there is usually no inherent linkage or relationship established between different texts within the dataset. Therefore, unlike datasets such as social networks or movie ratings, where explicit relationships exist between users or items, the Gutenberg Project dataset primarily focuses on providing access to individual literary works rather than on capturing relational data between instances.

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

    - The Gutenberg Project dataset does not typically come with recommended data splits for training, development/validation, or testing. Since the dataset primarily consists of textual documents, researchers or practitioners may choose to split the data according to their specific needs or tasks.

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

    - While the Gutenberg Project dataset strives for accuracy, occasional errors, sources of noise, or redundancies may exist due to factors such as digitization errors, variations in text formatting, or discrepancies in metadata. These errors could include typographical mistakes, formatting inconsistencies, or discrepancies in authorship attribution.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The Gutenberg Project dataset is primarily self-contained and does not heavily rely on external resources. While the dataset may contain links or references to external websites for additional information, there are no guarantees that these resources will remain constant over time. As a result, there are no official archival versions of the complete dataset that include external resources as they existed at the time of creation. Generally, there are no restrictions such as licenses or fees associated with accessing or using the Gutenberg Project dataset. However, users should verify any potential restrictions or licenses associated with accessing external resources linked within the dataset.

11. *Does the dataset contain data that might be considered confidential (for example, 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.*

    - No, the Gutenberg Project dataset typically does not contain data that might be considered confidential. The dataset primarily consists of literary texts that are in the public domain, and therefore, it does not include sensitive or confidential information such as legal privilege, doctor-patient confidentiality, or individuals' non-public communications. The texts within the dataset are generally available for public access and use, and they do not contain personal or confidential data.

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

    - The Gutenberg Project dataset primarily consists of literary works in the public domain, which may include content reflecting historical contexts and societal norms. While the majority of the content is unlikely to be offensive, some texts may contain language or themes that could be considered offensive or distressing to certain individuals or groups due to historical context or cultural differences.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No, the Gutenberg Project dataset typically does not identify specific subpopulations such as age or gender.

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

    - In general, it is highly unlikely to identify individuals directly from the Gutenberg Project dataset. The dataset primarily consists of literary texts in the public domain, which typically do not contain personal information about specific individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.*

    - No, the Gutenberg Project dataset typically does not contain data that might be considered sensitive in the ways described. While some texts may touch on topics such as race, ethnicity, sexual orientation, religion, politics, or geography, they do so within the context of literature and historical narratives rather than providing specific or sensitive personal data about individuals.

16. *Any other comments?*

    - no.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the 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 Gutenberg Project dataset was acquired through the digitization of public domain literary texts. This process involved scanning or transcribing physical copies of books, essays, poetry collections, plays, and other literary works to create digital versions of the texts. The raw textual data available in the dataset represents these digitized versions of the original literary works.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data collection for the Gutenberg Project dataset primarily involves manual human curation and digitization processes. Volunteers transcribe or scan public domain literary texts to create digital versions of the works, which are then made available through the Gutenberg Project website and other platforms.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The Gutenberg Project dataset is not typically a sample from a larger set.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection process for the Gutenberg Project dataset primarily involves volunteers who contribute their time and effort to transcribe or scan public domain literary texts. These volunteers may include students, scholars, enthusiasts, or individuals passionate about preserving and sharing literature. As volunteers, they are typically not compensated monetarily for their contributions but are motivated by a shared interest in digitizing and preserving literary works for public access.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data collection for the Gutenberg Project dataset has occurred over several decades since the project's inception in 1971. The timeframe for data collection extends from the project's earliest efforts to digitize public domain literary texts to ongoing contributions made by volunteers over time.

6. *Were any ethical review processes conducted (for example, 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.*

    - As the Gutenberg Project dataset primarily consists of public domain literary texts, it typically does not involve ethical review processes conducted by institutional review boards (IRBs) or similar bodies. Since the dataset comprises works that are freely available to the public and are not subject to privacy concerns or human subjects research, ethical review processes are generally not applicable.

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

- The Gutenberg Project dataset consists of public domain literary texts that have been digitized by volunteers. The data collection process involves manual transcription or scanning of physical copies of books, essays, poetry collections, plays, and other literary works. Volunteers contribute their time and effort to create digital versions of these texts, which are then made available through the Gutenberg Project website and other platforms.

8. *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.*

    - As the Gutenberg Project primarily involves digitizing public domain literary texts, there typically isn't a process for notifying individuals about data collection since the texts are already publicly available. Since the data consists of texts that are no longer under copyright and are freely accessible to the public, there is no need for individual notification regarding data collection.

9. *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.*

    - There isn't a process for obtaining consent from individuals for the collection and use of their data since all texts are copyright free.

10. *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).*

    - All texts are copyright free.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.*

    - All the texts are copyright free.

12. *Any other comments?*

    - none.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization 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 remaining questions in this section.*

- The Gutenberg Project dataset typically undergoes minimal preprocessing, primarily involving basic steps such as tokenization and removal of special characters or formatting inconsistencies to improve data quality. These preprocessing steps are aimed at ensuring the readability and usability of the raw textual data rather than performing complex transformations or feature extraction. Labeling of the data is not typically applicable, as the dataset primarily consists of literary texts and does not involve categorizing or labeling instances.

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

   - N/A

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - N/A

4. *Any other comments?*

   - No.

Uses

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

   - Yes, the Gutenberg Project dataset has been extensively used for various natural language processing (NLP) tasks, including text classification, sentiment analysis, language modeling, and authorship attribution. Researchers have leveraged the dataset to develop and evaluate NLP algorithms, train machine learning models, and conduct empirical studies in computational linguistics and digital humanities.

2. *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.*

   - N/A

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

   - The Gutenberg Project dataset can be utilized for tasks such as literary analysis to study authorship styles and thematic elements, genre classification to categorize texts into different literary genres, and language learning by providing access to classic literature in multiple languages.

4. *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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues)*

or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

- The Gutenberg Project dataset's composition of public domain literary texts spans diverse genres, time periods, and cultural contexts, necessitating caution regarding potential biases or stereotypes in the content. Users should assess the representativeness and quality of the texts, considering historical context and potential errors in digitization or transcription.

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

   - The Gutenberg Project dataset may not be well-suited for tasks such as sentiment analysis, speech recognition, or specific domain analysis. Its focus on literary texts with nuanced language and subjective interpretations may pose challenges for tasks requiring straightforward classification or domain-specific knowledge.

6. *Any other comments?*

   - No.

Distribution

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

   - Yes, the Gutenberg Project dataset is typically distributed to third parties outside of the entity responsible for its creation, which includes researchers, educators, developers, and the general public. The dataset is made freely available through the Gutenberg Project website and other platforms for anyone to access, download, and use for research, educational, or personal purposes.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The Gutenberg Project dataset is typically distributed through the Gutenberg Project website and other online platforms in the form of downloadable files, which may be provided in various formats such as plain text, HTML, ePub, or PDF. Users can access and download individual texts or entire collections from the Gutenberg Project's extensive library of public domain literary works.

3. *When will the dataset be distributed?*

   - The Gutenberg Project dataset is continuously available for distribution through the Gutenberg Project website and other online platforms.

4. *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.*

    - The Gutenberg Project dataset is typically distributed under public domain or open access licenses, allowing users to freely access, use, and distribute the data without restrictions. Since the dataset primarily consists of public domain literary texts, there are generally no copyright or intellectual property restrictions associated with its distribution.

5. *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.*

    - As the Gutenberg Project dataset primarily consists of public domain literary texts, third parties typically do not impose intellectual property-based restrictions on the data associated with the instances.

6. *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.*

    - No export control or regulatory restrictions apply.

7. *Any other comments?*

    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The Gutenberg Project dataset is supported, hosted, and maintained by the Gutenberg Project organization, which is a volunteer-driven initiative dedicated to digitizing and preserving public domain literary texts. The project relies on contributions from volunteers worldwide who transcribe, proofread, and digitize texts, as well as maintain the project's website and infrastructure.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The best way to contact the team is by email. contact.gutenberg@yahoo.ca

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

    - There is no erratum

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

    - The Gutenberg Project dataset is regularly updated and maintained by volunteers who contribute new texts, correct errors, and improve the dataset's quality over time. Updates to the dataset, such as corrections to labeling errors or additions of new instances, may occur on an ongoing basis as volunteers transcribe, proofread, and digitize new texts.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.*

    - N/A

6. *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 dataset consumers.*

    - As the Gutenberg Project dataset is continuously updated and maintained with new contributions and improvements, older versions of the dataset may not be explicitly supported or maintained in their entirety. Users can generally access historical versions through alternative sources or archival platforms.

7. *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 dataset consumers? If so, please provide a description.*

    - Yes, the Gutenberg Project dataset welcomes contributions from volunteers who want to extend, augment, or improve the dataset with new texts or corrections. Volunteers can contribute to the project by transcribing, proofreading, or digitizing public domain literary texts and submitting them through the Gutenberg Project website or other designated platforms.

8. *Any other comments?*

    - No.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Hart, Michael S. 2021. "Gutenberg Project, 2021." Project Gutenberg.https://www.gutenberg.org/ .