Datasheet for Project Gutenberg's 'Stream of Consciousness Literature'*

Quang Mai

April 25, 2024

A datasheet for Quang Mai's Stream of Consciousness Literature (Exploring Word Frequency, Sentiment Value and Mental Health Themes in the Works of Joyce, Woolf, Proust, Mansfield and Eliot from Project Gutenberg), orginal textual data available under Project Gutenberg archive (Johnston and Robinson 2023) ("Project Gutenberg," n.d.). The questions used to form this datasheet were extracted from Gebru et al. (2021). Created using the R statistical programming language (R Core Team 2022).

Table of contents

1 Motivation **2**

2 Composition **2**

3 Collection process **5**

4 Preprocessing/cleaning/labeling **7**

5 Uses **7**

6 Distribution **8**

7 Maintenance **9**

References **11**

*Code and data are available at: https://github.com/ponolite/stream-consciousness-language.git

1 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 dataset or HTML texts were created to be archived as literary texts by Project Gutenberg. We were unable to find a publicly available dataset in a structured format that contained the information on SOC literature and their correlation to mental health themes in the West's late 19th to mid-20th century, thus necessitating a need for these datasets.

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

    - Volunteer contributors and proofreaders of Project Gutenberg initially created the dataset or the HTML text files. Later, the author Quang Mai created his own datasets using the HTML of 9 selected SOC novels available on Project Gutenberg using the package `gutenbergr` (Johnston and Robinson 2023) ("Project Gutenberg," n.d.). These texts were created for the author's own purpose of analyzing their linguistic patterns and contents and their relationship to mental health themes.

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 these datasets were not funded by anyone.

4. *Any other comments?*

    - N/A

2 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.*

    - Each row of the raw dataset represents a the Project Gutenberg identification number for that respective SOC text, a line extracted from the HTML text of that SOC text (for instance, "April is the cruellest month, breading" by T.S. Eliot), the book in which the SOC text originates from and the author's name. As each SOC text spans chapters, each dataset spans multiple rows as there are many lines within that text.

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

- There is only 1 type of data, which is the bibliographic data of each SOC text. To support this data type, there are 5 instances (each combining to represent a text's bibliography, like author's name, book title, each line of the text and book ID).

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 dataset contains all possible instances from the larger set of HTML text data from Project Gutenberg, including all of the abovementioned ones.

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 consists of a number to represent the book ID (9 IDs), a literary line sourced from the 9 HTML texts from Project Gutenberg (167516 lines in total), the names of the SOC texts (9 text names in total, since all authors beside Proust have 2 texts), and the author's name (9 authors).

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

    - Each instance is eponymously labelled, for instance 'book ID' is labelled 'book_id'.

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.*

    - There is no missing information from individual instances since all texts have been in the public domain for a while, thus all data is up-to-date.

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.*

    - Yes, the relationships between individual instances are made explicit through their eponymously titled labels and the book IDs as provided from Project Gutenberg, which helps clarify which data belongs to which individual SOC text.

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 data is split into individual SOC authors. This split ensures a comparative analysis between different demographics of SOC authors later on, for instance, between transnational SOC authors and between gendered SOC authors, enabling a more nuanced analysis of the dataset.

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

    - No, there are no apparent errors or sources of noise. There are redundancies, although these are purposeful and provide added meaning to the data (for example, the repeating book ID that helps ensure certain data instances ony belong to certain books).

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 dataset is strictly self-contained, relying only on metrics provided by the Project Gutenberg. However, it will change depending on changes in legislation and regulation of the archive itself.

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.*

    - The data does not provide any confidential information, since all texts are works of the public domain as disclosed on Project Gutenberg's website ("Project Gutenberg," n.d.).

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.*

    - No, the dataset does not include information that would harm a viewer. However, certain words may trigger certain emotional responses since not all literary SOC texts might align with a viewer's point of view.

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, it does not categorize texts based on sub-populations like age, gender, or region.

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.*

    - It is possible to identify individuals from this dataset because the author names are disclosed, however, this information should be harmless since it's in the public domain.

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.*

    - Based on the diverse literary contents, the dataset might contain sensitive information. However, as stressed, since all information is entered into public domain, this information should be accessible to all.

16. *Any other comments?*

    - N/A

3 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 uses a mix of information from two sources: individual submission of SOC texts by individual volunteers who went through a rigorous process of transcript submission to Project Gutenberg and voluntary proofreaders that ensure each SOC text underwent professional editing before being archived ("Project Gutenberg," n.d.). The individual submissions are processed using internal submission tool from Project Gutenberg.

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?*

    - Most new Project Gutenberg eBooks are digitized editions of printed books in HTML format, which is validated using the W3C's online validator. Additionally, the archive prefers plain text versions whenever possible due to their universal accessibility, and long-term usability. PDF, Word, and other word processor formats are avoided as master formats due to the challenges involved in converting them to HTML and updating them ("Project Gutenberg," n.d.).

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

    - There was no sampling strategy employed. The data is an inclusive and comprehensive statistical summary of all SOC texts included.

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)?*

    - Data collection or book digitization was conducted by both voluntary contributors and proofreaders from Project Gutenberg, which means no monetary compensation is viable.

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 is collected continuously and there is no set timeframe. However, since Project Gutenberg doesn't accept copyrighted works but only works in the public domain, the majority of SOC texts were published over 95 years ago.

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.*

    - It is unclear whether there was an ethical review of this data, however there is a clearly disclosed submission guideline on Project Gutenberg's website that touched on this aspect ("Project Gutenberg," n.d.)

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 data was obtained via Project Gutenberb, a volunteer archive based in the United States.

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.*

    - No, as the data is anonymized and does not contain any confidential or sensitive information the individuals that form this data were not notified.

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.*

    - All sourced SOC literary texts are in the public domain, thus no consent is required.

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).*

- United States's privacy laws may allow individuals to request the deletion of their data from systems, although it is unclear whether these laws apply to the works in the public domain.

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.*

    - No, no analysis was provided.

12. *Any other comments?*

    - N/A

4 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.*

    - No preprocessing/cleaning/labeling on the data was done by Project Gutenbeg as all archived texts are digitized versions of printed books. They are not new editions.

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 No, it is unclear what software was used to preprocess, clean, and label the data.

4. *Any other comments?*

    - N/A

5 Uses

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

- There is no feasible way of determining how the data was previously or is currently used. Project Gutenberg allows any person to utilize the dataset.

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.*

   - No, there is no public log of any papers or systems that currently use the dataset.

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

   - The data set can be used to conduct topic modeling of the texts involved, especially in regards to SOC literature.

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?*

   - No.

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

   - No. The purpose of Project is to allow the public to use the literary data as they see fit.

6. *Any other comments?*

   - No.

6 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.*

   - The dataset is freely distributed using the `gutenbergr` package accessible using R (Johnston and Robinson 2023). It is also accessible on Project Gutenberg's website found here: https://www.gutenberg.org.

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

   - The dataset is distributed on Project Gutenberg archive.

3. *When will the dataset be distributed?*

    - The dataset has no set schedule for distribution or updates.

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 dataset is provided and licenced under Project Gutenberg. found here: https://www.gutenberg.org.

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.*

    - No, the data is non-profit and in the public domain so no third-parties have a legal claim to the data.

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.*

    - The dataset is governed by the aforementioned Project Gutenberg Licence provided by the archive based in the United States, and may be further restricted by municipal, provincial, or federal legislation.

7. *Any other comments?*

    - N/A

7 Maintenance

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

    - Project Gutenberg solely maintain and update the data, while also supporting and hosting the textual data of SOC literature.

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

    - The archive may be contacted via this mailing address Project Gutenberg Literary Archive Foundation, PO Box 16327, Salt Lake City, UT 84116.

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

    - No.

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)?*

    - Since the literary texts have underwent significant proofreading and editing and they are merely digitized versions of physical copies of books, it's unclear if the dataset or the HTML texts will be updated.

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.*

    - No.

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.*

    - It's unclear if older versions of the dataset continue to be supported, hosted or maintained.

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.*

    - With open data, there are always mechanisms for others to extend, augment, or build upon the dataset. However, there is no formal mechanism to bring these changes to the Project Gutenberg platform or the official dataset. Users may take the dataset and augment it on their own.

8. *Any other comments?*

    - N/A

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.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://docs.ropensci.org/gutenbergr/.

"Project Gutenberg." n.d. https://www.gutenberg.org.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.