MOTIVATION.

[question1] 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.
[answer1] BookCorpus was originally created to help train
a neural network that could provide “descriptive explanations for visual content” [39]. Specifically,
BookCorpus trained a sentence embedding model for aligning dialogue sentences frommovie subtitles
with wri_en sentences from a corresponding book. A_er unsupervised training on BookCorpus,
the authors’ encoder model could “map any sentence through the encoder to obtain vector
representations, then score their similarity through an inner product” [39].
[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] BookCorpus was collected by Zhu and Kiros et al. [39] from
the University of Toronto and the Massachuse_s Institute of Technology. Their original paper
includes seven authors, but does not specify who was involved in collecting BookCorpus.
[question3] 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.
[answer3] The original paper by Zhu and Kiros et al. [39]
acknowledges support from the Natural Sciences and Engineering Research Council (NSERC),
the Canadian Institute for Advanced Research (CIFAR), Samsung, Google, and a grant from the
Office of Naval Research (ONR). They do not specify how funding was distributed across these
sources.
It is more difficult to identify funding for the authors who wrote the books in BookCorpus.
Broadly, many authors on Smashwords do make money by selling ebooks to readers (including
on other platforms like Kindle, Audible, Barnes and Noble, and Kobo), although many also write
books as a hobby alongside other occupations. Some books in BookCorpus may have been commissioned
in some way, however, analyzing sources of commission would require further work
[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] 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.
[answer5] BookCorpus consists of text files, each of
which corresponds to a single book from smashwords.com. Zhu and Kiros et al. [39] also provide
two large files in which each row represents a sentence.
[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] In the original dataset described by Zhu and
Kiros et al. [39], BookCorpus contained 11,038 books. However, based on the files we obtained,
there appear to be only 7,185 unique books (excluding romance-all.txtand adventure-all.txt
as explained in 2.2.1).We identified potential duplicates based on file names,which suggested that
2,930 booksmay be duplicated.Using the diffUnix program, we confirmed that BookCorpus contained
duplicate, identical text files for all but five of these books.We manually inspected the five
exceptions:

• 299560.txt(Third Eye Patch), forwhich slightly different versions appeared in the “Thriller” and “Science Fiction” genre folders (only 30 lines differed)
• 529220.txt(On the Rocks), forwhich slightly different versions appeared in the “Literature” and “Science Fiction” genre folders (only the title format differed)
• Hopeless-1.txt, for which identical versions appeared in the “New Adult” and “Young Adult” genre folders, and a truncated version appeared in the “Romance” folder (containing 30% of the full word count)
• u4622.txt, for which identical versions appeared in the “Romance” and “Young Adult” genre folders, and a slightly different version appeared in the “Science Fiction” folder (only 15 added lines)
• u4899.txt, for which a full version appeared in the “Young Adult” folder and a truncated version (containing the first 28 words) appeared in the “Science Fiction” folder
Combined with the diff results, our manual inspection confirmed that each filename represents one unique book, thus BookCorpus contained at most 7,185 unique books.

[question7] 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).

[answer7] Book-Corpus contains free books from smashwords.com which are at least 20,000 words long. Based on metrics from Smashwords [7], 11,038 books (as reported in the original BookCorpus dataset) would have represented approximately 3% of the 336,400 books published on Smashwords as of 2014,while the 7,185 unique books we report would have represented 2%. For reference, as of 2013, the Library of Congress contained 23,592,066 cataloged books [14].We return to the implications of this sample in the discussion (section 5).

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Each book in BookCorpus simply includes the full text from the ebook (o_en including preamble, copyright text, etc.). However, in research that uses BookCorpus, authors have applied a range of different encoding schemes that change the definition of an “instance” (e.g. in GPT-N training, text is encoded using byte-pair encoding).

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] No. The text from each book was originally used for unsupervised training by Zhu and Kiros et al. [39], and the only label-like a_ribute is the genre associated with each book, which is provided by Smashwords.

[question10] 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.

[answer10] Yes.We found 98 empty book files in the folder downloaded from the paper’s website [39]. Also, while the authors collected books longer than 20,000 words, we found that 655 files were shorter than 20,000 words, and 291 were shorter than 10,000words, suggesting that many book fileswere significantly truncated from

their original text.

[question11] 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.

[answer11] No. Grouped into folders by genre, the data implicitly links books in the same genre. We also found that duplicate books are implicitly linked through identical filenames. However, no other relationships are made explicit, such as books by the same author, books in the same series, books set in the same context, books addressing the same event, and/or books using the same characters.

[question12] 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.

[answer12] No. The authors use all books in the dataset for unsupervised training, with no splits or subsamples.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] Yes.While some book files appear to be cleaned of preamble and postscript text, many files still contain this text and various other sources of noise. Of particular concern is that we found many copyright-related sentences, for example:

- "if you're reading this book and did not purchase it, or it was not purchased for your use only, then please return to smashwords.com and purchase your own copy." (n=788)
- "this book remains the copyrighted property of the author, and may not be redistributed to others for commercial or non-commercial purposes..." (n=111)
- "although this is a free book, it remains the copyrighted property of the author, and may not be reproduced, copied and distributed for commercial or non-commercial purposes." (n=109)
- "thank you for respecting the author's work" (n=70)
- "no part of this publication may be copied, reproduced in any format, by any means, electronic or otherwise, without prior consent from the copyright owner and publisher of this book" (n=16)

Here, we note that these sentences represent noise and redundancy, though we return to the issue of copyrights in section 4.6.3. As previously noted, BookCorpus also containsmany duplicate books: of the 7,185 unique books in the dataset, 2,930 occurred more than once. Most of these (N=2,101) books appeared twice, though many were duplicated multiple times, including some books (N=6) with five copies in BookCorpus. See Table 2.

[question14] 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.

[answer14] No. Although Zhu and Kiros et al. [39] maintained a selfcontained version of BookCorpus on their website for some time, there is no longer an "official," publicly-available version. While we were able to obtain the dataset from their website through

a security vulnerability, the public web page about the project now states: “Please visit smashwords.com to collect your own version of BookCorpus” [39]. Thus, researchers who wish to use BookCorpus or a similar dataset must either use a new public version such as BookCorpusOpen [13], or generate a new dataset from Smashwords via “Homemade BookCorpus” [21]. Smashwords is an ebook website that describes itself as “the world’s largest distributor of indie ebooks.”[3] Launched in 2008 with 140 books and 90 authors, by 2014 (the year before BookCorpus was published) the site hosted 336,400 books from 101,300 authors [7]. In 2020, it hosted 556,800 books from 154,100 authors [8].

[question15] 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.

[answer15] Likely no.While we did find personal contact information in the data (see 4.2.15), the books do not appear to contain any other restricted information, especially since authors opt-in to publishing their books.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] Yes. While this topic warrants further research, as preliminary supporting evidence, we found that 537,878 unique sentences (representing 984,028 total occurrences) in BookCorpus contained one or more words in a commonly-used list of “Dirty, Naughty, Obscene, and Otherwise Bad Words” [11]. Inspecting a random sample of these sentences, we found they include some fairly innocuous profanities (e.g. the sentence “oh, shit.” occurred 250 times), some pornographic dialogue, some hateful slurs, and a range of other potentially problematic content. Again, further research is necessary to explore these sentences, especially given that merely using one of these words does not constitute an offensive or insulting sentence. In section 5 we further discuss how some sentences and books may be problematic for various use cases.

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

[SkipA] NO

[question17] 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.

[answer17] Each book is associated with an author however BookCorpus does not identify books by author or any author demographics, and the books_in_sentences folder even aggregates all books into just two files. The books_txt_full folder identifies 16 genres, though we do not consider genres to be subpopulations since they group books rather than authors.

[question18] 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.

[answer18] Likely yes. In reviewing a sample of books, we found that many authors provide personally-identifiable information, o_en in the form of a personal email address for readers interested in contacting them.

[question19] 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.
[answer19] Yes. The aforementioned
contact information (email addresses) is sensitive personal information.
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] 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.
[answer21] The text for each book was
downloaded from smashwords.com.
[question22] 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?
[answer22] The data was collected
via scraping so_ware. While the original scraping program is not available, replicas (e.g. [21])
operate by first scraping smashwords.com to generate a list of links to free ebooks, downloading
each ebook as an epub file, then converting each epub file into a plain text file.
[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
[answer23] Books were included in the original Book-
Corpus if they were available for free on smashwords.com and longer than 20,000 words, thus representing a non-probabilistic convenience sample. The 20,000 word cutoff likely comes from
the Smashwords interface, which provides a filtering tool to only display books “Over 20K words.”
[question24] 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)?
[answer24] Unknown.
The original paper by Zhu and Kiros et al. [39] does not specify which authors collected and
processed the data, nor how they were compensated.
[question25] 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.

[answer25] Unknown. BookCorpus was originally collected some time before the original paper [39] was presented at the International Conference on Computer Vision (ICCV) in December 2015.

[question26] 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.

[answer26] Likely no. Zhu and Kiros et al. [39] do not mention an Institutional Review Board (IRB) or other ethical review process involved in their original paper.

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

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] Third party. BookCorpus was collected from smashwords.com, not directly from the authors.

[question28] 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.

[answer28] Each book is associated with an author (thus determining that the following three questions should be addressed). Likely no. Discussing BookCorpus in 2016, Richard Lea wrote in The Guardian that "The only problem is that [researchers] didn't ask" [23]. When notified about BookCorpus and its uses, one author from Smashwords said "it didn't even occur to me that a machine could read my book" [23].

[question29] 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.

[answer29] No. While authors on smashwords.com published their books for free, they did not consent to including their work in BookCorpus, and many books contain copyright restrictions intended to prevent redistribution.

As described by Richard Lea in The Guardian [23], many books in BookCorpus include: a copyright declaration that reserves "all rights", specifies that the ebook is "licensed for your personal enjoyment only", and offers the reader thanks for "respecting the hard work of this author"

Considering these copyright declarations, authors did not explicitly consent to include their work in BookCorpus or related datasets. Using the framework of consentful tech [24], a consentful version of BookCorpus would ideally involve author consent that is Freely given, Reversible, Informed, Enthusiastic, and Specific (FRIES).

[question30] 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).

[answer30] Likely no. For example, if an author released a book for free before BookCorpus was collected, then changed the price and/or copyright a_er BookCorpus was collected, the book likely remained in BookCorpus. In fact, preliminary analysis suggests that this is the case for at least 438 books in BookCorpus which are no longer free to download from Smashwords, and would cost $1,182.21 to purchase as of April 2021.

[question31] 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.

[answer31] Likely no. Richard Lea interviewed a handful of authors represented in BookCorpus [23], but we are not aware of any holistic impact analysis.

[question32] Any other comments?

[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] 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.

[answer33] While the original paper by Zhu and Kiros et al. [39] did not use labels for supervised learning, each book is labeled with genres. It appears genres are supplied by authors themselves. Likely yes. The .txt files in BookCorpus seem to have been partially cleaned of some preamble text and postscript text, however, Zhu and Kiros et al. [39] do not mention the specific cleaning steps. Also, many files still contain some preamble and postscript text, including many sentences about licensing and copyrights. For example, the sentence “please do not participate in or encourage piracy of copyrighted materials in violation of the author’s rights” occurs at least 40 times in the BookCorpus books_in_sentences files. Additionally, based on samples we reviewed from the original BookCorpus, the text appears to have been tokenized to some degree (e.g. contractions are split into two words), though we were unable to identify the exact procedure.

[question34] 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.

[answer34] Unknown.

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] While the original software is not available, replication attempts provide some software for turning .epub files into .txt files and subsequently cleaning them.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] BookCorpus was originally used to train sentence embeddings for a system meant to provide descriptions of visual content (i.e. to "align" books and movies), but the dataset has since been applied in many different use cases. Namely, BookCorpus has been used to help train more than thirty influential language models [12], including Google's enormously influential BERT model which was shown to be applicable to a wide range of language tasks (e.g. answering questions, language inference, translation, and more).

[question38] 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.

[answer38] On the dataset card for BookCorpus [12], Hugging Face provides a list of more than 30 popular language models that were trained or fine-tuned on the dataset.

[question39] What (other) tasks could the dataset be used for?

[answer39] Given that embedding text and training language models are useful prerequisites for a huge number of language related tasks, the Book-Corpus dataset could in theory be used as part of the pipeline for almost any English language task. However, as discussed below, this work highlights the need for caution when applying this dataset.

[question40] 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?

[answer40] Yes. At the very least, the duplicate books and sampling skews should guide any future uses to curate a subsample of BookCorpus to be_er serve the task at hand.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] We leave this question to be more thoroughly addressed in future work. However, our work strongly suggests that researchers should use BookCorpus with caution for any task, namely due to potential copyright violations, duplicate books, and sampling skews.

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] 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.

[answer43] For some time, Zhu and Kiros et al. [39] distributed

BookCorpus from a web page. The page now states “Please visit smashwords.com to collect your own version of BookCorpus” [39].

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] While there have been various efforts to replicate BookCorpus, one of the more formal efforts is BookCorpusOpen [13], included in the Pile [16] as “BookCorpus2.” Furthermore, GitHub users maintain a “Homemade BookCorpus” repository [21] with various pre-compiled tarballs that contain thousands of pre-collected books.

[question45] When will the dataset be distributed?

[answer45]

[question46] 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.

[answer46] To our knowledge, BookCorpus dataset has never stated any copyright restrictions, however, the same is not true of books within BookCorpus. In reviewing sources of noise in BookCorpus, we found 111 instances of the sentence, “this book remains the copyrighted property of the author, and may not be redistributed to others for commercial or non-commercial purposes.”We also found 109 instances of the sentence “although this is a free book, it remains the copyrighted property of the author, and may not be reproduced, copied and distributed for commercial or non-commercial purposes.” This initial analysis makes clear that the distribution of BookCorpus violated copyright restrictions for many books, though further work from copyright experts will be important for clarifying the nature of these violations. Also, some books in BookCorpus now cost money even though they were free when the original dataset was collected. By matching metadata from Smashwords for 2,680 of the 7,185 unique books in BookCorpus, we found that 406 of these 2,680 books now cost money to download. The total cost to purchase these books as of April 2021 would be $1,182.21, and this represents a lower bound since we only matched metadata for 2,680 of the 7,185 books in BookCorpus.

[question47] 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.

[answer47] Likely no.

[question48] 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.

[answer48] Likely no, notwithstanding the aforementioned copyright restrictions

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] BookCorpus is not formallymaintained or hosted, although a new version called BookCorpusOpen [13] was collected by Shawn Presser and included in the Pile [16]. As BookCorpus is no longer officially maintained, we answer the below questions by focusing on how other researchers have replicated and extended the Book-Corpus data collection approach.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51]

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] No. To our knowledge, Zhu and Kiros et al. [39] have not published any list of corrections or errors in BookCorpus.

[question53] 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)?

[answer53] An updated version of BookCorpus is available as BookCorpusOpen [13]. This updated version was published by Presser, not Zhu and Kiros et al. [39] who created the original BookCorpus.

[question54] 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.

[answer54]

[question55] 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.

[answer55] BookCorpus is no longer available fromthe authors'website,which now tells readers to "visit smashwords.com to collect your own version of BookCorpus" [39].

[question56] 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.

[answer56] Yes, GitHub users maintain a "Homemade BookCorpus" repository [21] that includes so_ware for collecting books from smashwords.com

[question57] Any other comments?

[answer57]