



Datasets: bookcorpus/ **bookcorpus**

like 301

Follow bookcorpus 10

Tasks:



Text Generation



Fill-Mask

Sub-tasks:

language-modeling

masked-language-modeling

Languages:



English

Size:

10M<n<100M

ArXiv:

arxiv:2105.05241

License:



unknown



Dataset card



Files



Community 8

Downloads last month

8,892



Papers with Code



Homepage:

yknzhu.wixsite.com

Size of downloaded dataset files:

1.18 GB



Models trained or fine-tuned on bookcorpus/bookcorpus



google-bert/bert-base-uncased



Fill-Mask • Updated Feb 19, 2024 • 83.8M • • 2.21k



FacebookAI/roberta-large



Fill-Mask • Updated Feb 19, 2024 • 14.9M • 207



distilbert/distilbert-base-uncased



Fill-Mask • Updated May 6, 2024 • 11M • • 656



FacebookAI/roberta-base



Fill-Mask • Updated Feb 19, 2024 • 10.9M • • 477

[Browse 497 models trained on this dataset](#)

💬 pro-grammer/StoryCrafterLLM

🌐 ndhieunguyen/Lang2mol-Diff



🗃 Dataset Viewer

The viewer is disabled because this dataset repo requires arbitrary Python code execution. Please consider removing the [loading_script](#) and relying on [automated data support](#) (you can use [convert to parquet](#) from the datasets library). If this is not possible, please [open a discussion](#) for direct help.

🔗 Dataset Card for BookCorpus

🔗 Dataset Summary

Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story. This work aims to align books to their movie releases in order to provide rich descriptive explanations for visual content that go semantically far beyond the captions available in current datasets.

🔗 Supported Tasks and Leaderboards

[More Information Needed](#)

🔗 Languages

[More Information Needed](#)

🔗 Dataset Structure

🔗 Data Instances

In the original dataset described by [Zhu and Kiros et al.](#), BookCorpus contained 11,038 books. However, based on the files obtained, there appear to be only 7,185

unique books (excluding romance-all.txt and adventure-all.txt as explained in 2.2.1). Potential duplicates were identified based on file names, which suggested that 2,930 books may be duplicated. Using the diff Unix program, it was confirmed that BookCorpus contained duplicate, identical text files for all but five of these books. The five exceptions were manually inspected:

- 299560.txt (Third Eye Patch), for which slightly different versions appeared in the “Thriller” and “Science Fiction” genre folders (only 30 lines differed)
- 529220.txt (On the Rocks), for which 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, the manual inspection confirmed that each filename represents one unique book, thus BookCorpus contained at most 7,185 unique books.

plain_text

- **Size of downloaded dataset files:** 1.18 GB
- **Size of the generated dataset:** 4.85 GB
- **Total amount of disk used:** 6.03 GB

An example of 'train' looks as follows.

```
{  
  "text": "But I traded all my life for some lovin' and some gold"  
}
```

Data Fields

Each book in BookCorpus simply includes the full text from the ebook (often including preamble, copyright text, etc.). However, in research that 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). The data fields are the same among all splits. There is no label or target associated with each instance (book). The text from each book was originally used for unsupervised training by [Zhu and Kiros et al.](#), and the only label-like attribute is the genre associated with each book, which is provided by Smashwords. No relationships between individual instances (books) are made explicit. Grouped into folders by genre, the data implicitly links books in the same genre. It was 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.

plain_text

- text: a string feature.

Data Splits

There are no recommended data splits. The authors use all books in the dataset for unsupervised training, with no splits or subsamples.

name	train
plain_text	74004228

🔗 Dataset Creation

🔗 Curation Rationale

The books in BookCorpus were self-published by authors on smashwords.com, likely with a range of motivations. While we can safely assume that authors publishing free books via smashwords.com had some motivation to share creative works with the world, there is no way to verify they were interested in training AI systems. For example, many authors in BookCorpus explicitly license their books “for [the reader’s] personal enjoyment only,” limiting reproduction and redistribution. 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” [<https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>].

🔗 Source Data

🔗 Initial Data Collection and Normalization

Per [Bandy and Vincent \(2021\)](#), the text for each instance (book) was acquired via download from smashwords.com. The data was collected via scraping software. While the original scraping program is not available, replicas (e.g. <https://github.com/BIGBALLON/cifar-10-cnn>.) 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. 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.” The individuals involved in collecting BookCorpus and their compensation are unknown. The original paper by Zhu and Kiros et al. (<https://yknzhu.wixsite.com/mbweb>) does not specify which authors collected and processed the data, nor how they were compensated. The timeframe over which BookCorpus was collected is unknown as well. BookCorpus was originally collected

some time before the original paper (<https://yknzhu.wixsite.com/mbweb>) was presented at the International Conference on Computer Vision (ICCV) in December 2015. It is unlikely that any ethical review processes were conducted. Zhu and Kiros et al. (<https://yknzhu.wixsite.com/mbweb>) do not mention an Institutional Review Board (IRB) or other ethical review process involved in their original paper. The dataset is related to people because each book is associated with an author (please see the "Personal and Sensitive Information" section for more information on this topic).

Bandy and Vincent also assert that while the original paper by Zhu and Kiros et al. (<https://yknzhu.wixsite.com/mbweb>) did not use labels for supervised learning, each book is labeled with genres. It appears genres are supplied by authors themselves. It is likely that some cleaning was done on the BookCorpus dataset. The .txt files in BookCorpus seem to have been partially cleaned of some preamble text and postscript text, however, Zhu and Kiros et al. (<https://yknzhu.wixsite.com/mbweb>) 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 the exact procedure used is unclear. It is unknown if some of the "raw" data was saved in addition to the clean data. While the original software used to clean the BookCorpus dataset is not available, replication attempts provide some software for turning .epub files into .txt files and subsequently cleaning them.

Who are the source language producers?

Per Bandy and Vincent (2021), the data in BookCorpus was produced by self-published authors on smashwords.com and aggregated using scraping software by Zhu and Kiros et al.

Annotations

Annotation process

More Information Needed

Who are the annotators?

More Information Needed

Personal and Sensitive Information

Per Bandy and Vincent (2021), it is unlikely that authors were notified about data collection from their works. Discussing BookCorpus in 2016, Richard Lea wrote in The Guardian that “The only problem is that [researchers] didn’t ask” (<https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>). 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” (<https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>). Authors did not consent to the collection and use of their books. 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 (<https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>), 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 (<https://www.consentfultech.io>), a consent-ful version of BookCorpus would ideally involve author consent that is Freely given, Reversible, Informed, Enthusiastic, and Specific (FRIES). It is unlikely that authors were provided with a mechanism to revoke their consent in the future or for certain uses. For example, if an author released a book for free before BookCorpus was

collected, then changed the price and/or copyright after 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.

Considerations for Using the Data

The composition of BookCorpus or the way it was collected and preprocessed/cleaned/labeled might impact future uses. At the very least, the duplicate books and sampling skews should guide any future uses to curate a subsample of BookCorpus to better serve the task at hand. An analysis of the potential impact of BookCorpus and its use on data subjects has not been conducted. Richard Lea interviewed a handful of authors represented in BookCorpus ([Richard Lea](#)).

Social Impact of Dataset

The dataset contains data that might be considered sensitive. The aforementioned contact information (email addresses) is sensitive personal information.

Discussion of Biases

BookCorpus contains free books from smashwords.com which are at least 20,000 words long. Based on metrics from [Smashwords](#), 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 ([Audrey Fischer](#)).

There are some errors, sources of noise, or redundancies in BookCorpus. 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)

Note that these sentences represent noise and redundancy. As previously noted, BookCorpus also contains many 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.

🔗 Other Known Limitations

There are no export controls or other regulatory restrictions that apply to the dataset or to individual instances. Some information is missing from individual instances (books). 98 empty book files were found in the folder downloaded from [Zhu and Kiros et al.](#) Also, while the authors collected books longer than 20,000 words, 655 files were shorter than 20,000 words, and 291 were shorter than 10,000 words, suggesting that many book files were significantly truncated from their original text.

There were no ethical review processes conducted. [Zhu and Kiros et al.](#) do not mention an Institutional Review Board (IRB) or other ethical review process involved in their original paper. Bandy and Vincent strongly suggest that

researchers should use BookCorpus with caution for any task, namely due to potential copyright violations, duplicate books, and sampling skews.

🔗 Additional Information

🔗 Dataset Curators

More Information Needed

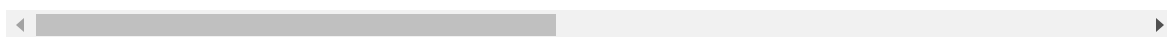
🔗 Licensing Information

The books have been crawled from <https://www.smashwords.com>, see their [terms of service](#) for more information.

A data sheet for this dataset has also been created and published in [Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus](#).


🔗 Citation Information

```
@InProceedings{Zhu_2015_ICCV,  
  title = {Aligning Books and Movies: Towards Story-Like Visual Ex  
  author = {Zhu, Yukun and Kiros, Ryan and Zemel, Rich and Salakh  
  booktitle = {The IEEE International Conference on Computer Visio  
  month = {December},  
  year = {2015}  
}
```



🔗 Contributions

Thanks to [@lewtun](#), [@richarddwang](#), [@lhoestq](#), [@thomwolf](#) for adding this dataset.

 System theme

Company

TOS

Privacy

About

Jobs

Website

Models

Datasets

Spaces

Pricing

Docs

