Dataset Facts

Dataset BookCorpus
Instances Per Dataset 7,185 unique books, 11,038 total

Motivation

Original Authors	Zhu and Kiros et al. (2015) [39]
Original Use Case	Sentence embedding
Funding	Google, Samsung, NSERC, CIFAR, ONR

Composition

Sample or Complete	Sample, ≈2% of smashwords.com in 2014
Missing Data	98 empty files, ≤655 truncated files
Sensitive Information	Author email addresses

Collection

Sampling Strategy	Free books with ≥20,000 words
Ethical Review	None stated
Author Consent	None

Cleaning and Labeling

Cleaning Done	None stated, some implicit
Labeling Done	None stated, genres by smashwords.com

Uses and Distribution

Notable Uses	Language models (e.g. GPT [29], BERT [9])
Other Uses	List available on HuggingFace [12]
Original Distribution	Author website (now defunct) [39]
Replicate Distribution	BookCorpusOpen [13]

Maintenance and Evolution

Corrections or Erratum	None
Methods to Extend	“Homemade BookCorpus” [21]
Replicate Maintainers	Shawn Presser [12]

Genres	% of BookCorpus*
Romance 2,881 books	26.1%
Fantasy 1,502 books	13.6%
Vampires 600 books	5.4%

Horror 4.1%	• Teen 3.9%
Adventure 3.5%	• Literature 3.0%
Historical Fiction 1.6%	

Not a significant source of nonfiction.

* Percentages based on directories in books_txt_full. Some books cross-listed.