Datasheet for NYTBestsellers

Motivation for Dataset Creation

Why was the dataset created? This dataset was created to support investigation into films that were made as adaptations of novels.
What (other) tasks could the dataset be used for? This Dataset is of all books that have been on the NewYork Times Bestsellers list between 2008-06-08 (the earliest available record) and that time it was collected in November 2021 and may be used for other investigations into bestsellers
Has the dataset been used for any tasks already? This datasheet was used for a school project. Please see https://github.com/MRWilliamsGit/BooksToMoviesML
Who funded the creation of the dataset? This dataset was created for free from free resources.
Any other comments?

Dataset Composition

What are the instances? Each observation is of a book that was on the New York Times Bestsellers List between 2008-06-08 (the earliest available record) and time of collection in November 2021.
Are relationships between instances made explicit in the data? N/A
How many instances of each type are there? There are 34063 observations.
What data does each instance consist of? Each observation includes the book title, description, contributor, author, contributor note, price, target age-group, publisher, isbns, its history on the bestseller list, and links to New York Times reviews.
Is everything included or does the data rely on external resources? This data was collected from the New York Times. Prices will change, the links provided to reviews may expire, and the rank on the bestseller is based on figures provided by publishers and verified by the NYT.
Are there recommended data splits or evaluation measures? N/A
What experiments were initially run on this dataset? This datasheet was used for a school project. Please see https://github.com/MRWilliamsGit/BooksToMoviesML
Any other comments?

Data Collection Process

How was the data collected? This data was collected via the New York Times' own API.
Who was involved in the data collection process? This data was collected by a master's student, for free.
Over what time-frame was the data collected? This data was collected in one batch in November of 2021.
How was the data associated with each instance acquired? Data is presumed to be provided by the publishers of the individual books.
Does the dataset contain all possible instances? This dataset should include all instances of books that have been present on the New York Times Bestsellers list between 2008-06-08 (the earliest available record) and time of collection in November 2021.
If the dataset is a sample, then what is the population? The New York Times has been publishing a bestsellers list since October 12, 1931, so there are decades of instances missing.
Is there information missing from the dataset and why? The features that are most often missing are contributor, price, and target age group. Contributor is often

Datasheet for NYTBestsellers

missing simply because there was no other contributor besides the author. Price is often missing for unknown reasons. Target age is only present for children's books.
Are there any known errors, sources of noise, or redundancies in the data? N/A
Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? No preprocessing was done on this dataset after it was pulled.
Was the "raw" data saved in addition to the preprocessed/cleaned data? N/A
Is the preprocessing software available? N/A
Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? N/A
Any other comments?

Dataset Distribution

How is the dataset distributed? This dataset is hosted on Github at https://github.com/MRWilliamsGit/BooksToMoviesML
When will the dataset be released/first distributed? November 2021
What license (if any) is it distributed under? N/A
Are there any fees or access/export restrictions? No
Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?
This dataset is only hosted on https://github.com/MRWilliamsGit/BooksToMoviesML
Will the dataset be updated? This dataset will most likely not be updated.
If the dataset becomes obsolete how will this be communicated? This dataset will always be representative of a November 2021 pull from The New York Times.
Is there a repository to link to any/all papers/systems that use this dataset? https://github.com/MRWilliamsGit/BooksToMoviesML
If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? Not at this time
Any other comments?

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? N/A
If it relates to other ethically protected subjects, have appropriate obligations been met? N/A
If it relates to people, were there any ethical review applications/reviews/approvals? N/A
If it relates to people, were they told what the dataset would be used for and did they consent? N/A
If it relates to people, could this dataset expose people to harm or legal action? N/A
If it relates to people, does it unfairly advantage or disadvantage a particular social group? N/A
If it relates to people, were they provided with privacy guarantees? N/A
Does the dataset comply with the EU General Data Protection Regulation (GDPR)? N/A
Does the dataset contain information that might be considered sensitive or confidential? N/A

Datasheet for NYTBestsellers

Does the dataset contain information that might be considered inappropriate or offensive? N/A
Any other comments?