Datasheet for IMDbCombined

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 specifically of novel adaptations, and could potentially be used for other investigations into novel adaptations. This data should not be used to extrapolate beyond that.
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 a movie released between 1970 and 2021, that was tagged as 'based on novel' on IMDb in November 2021.
Are relationships between instances made explicit in the data? There are ratings for each movie that are relative to each other.
How many instances of each type are there? There are 9778 observations.
What data does each instance consist of? This is data pulled from IMDb's website. It has seven features: the link to the IMDb page, the movie title, the movie release date, the movie rating, genres tagged on IMDb, budget figure estimated by IMDb, domestic gross, worldwide gross, description, keywords, and film runtime.
Is everything included or does the data rely on external resources? This data was collected from a website that is still not the primary source of this information, so this data should not be considered necessarily true.
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 a web-scraping script
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 batches on the same day
How was the data associated with each instance acquired? Some of this data is public knowledge, such as movie title, release date, and runtime. Other data is subjective and labeled by IMDb such as genre. Finally, there is financial data estimated by IMDb that has not been verified.
Does the dataset contain all possible instances? This dataset is all instances of movies tagged as 'based on novel' on IMDb, in November 2021. It should not be considered to contain all actual movies that were made from novels during that time.
If the dataset is a sample, then what is the population? The population of novel adaptations is hard to estimate, but this dataset is heavily weighted with English language, popular films that have been released in movie theaters.

Datasheet for IMDbCombined

Is there information missing from the dataset and why?
The features that are most often missing are financial, because this information is not readily offered by production companies.
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 scraped.
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 scrape of IMDb.
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

Datasheet for IMDbCombined

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