New York Times Hardcover Fiction Bestsellers, 1931-2020

I. Description

The NYT bestseller dataset provides a tabular representation of the fiction bestseller list of *The New York Times* between the years of 1931 and 2020. Previous research using similar data has been limited to partial segments of the list, such as the top 200 longest-running bestsellers since a certain date (Piper and Portelance, 2016) or bestsellers from only particular years (Sorenson, 2007). By contrast, this dataset covers the full list since its inception in 1931, along with each reported work's title, author(s), date of appearance, and rank.

Each row of the dataset is a single "entry" on the list, that is, a single slot for a single week. For each week, there will typically be 10 or 15 works listed. However, since the *Times* has varied the number of bestsellers featured in a given week, there may be 3, 6, 7, 8, or 16. A single "entry" on the list is treated as the basic unit of the dataset so that researchers can easily count the number of weeks that a given book appeared on a list, as well as the first and last weeks that it appeared. Altogether, the dataset features just over 60,000 rows of data available as .tsv (tab-separated values) file titled *nyt_full.tsv*.

I have provided two additional tables derived from this initial table. The first table *nyt_titles.tsv* provides title-level data for every unique title that appears in the full dataset. The second table *hathi_volumes.tsv* provides identifiers for every HathiTrust volume for which a corresponding title could be located in the *nyt_titles* list. The HathiTrust data used for matching was the "volumemeta" dataset described by Ted Underwood *et al*. Fields are as follows:

nyt_full.tsv

year - the year of appearance
week - the weekly issue of the bestseller list
rank - the book's rank on the list for that week
title_id - a unique ID mapping titles to the nyt_titles spreadsheet
title - title of the novel, as reported by the *New York Times*
author - author of the novel, as reported by the *New York Times*

nyt_titles.tsv

id - an arbitrary unique id for the novel
title - the title of the novel, as reported by the *New York Times*
author - the author of the novel, as reported by the *New York Times*
year - the first year that the novel appears on the bestseller list. Note that this year may be different from the publication year.
total_weeks - the total number of weeks the title was on the list
first_week - the first week that the novel appears on the bestseller list
debut_rank - the book's bestseller rank in the week of its first appearance

best_rank - the highest rank achieved by the title while on the list

hathi_volumes.tsv

HTID - unique volume ID from HathiTrust

title_id - the ID for that title in *nyt_titles.tsv*. Note that since HathiTrust is organized around volumes rather than titles, this field contains duplicates, such as in the case of frequently-reprinted works.

II. Relevance

This dataset provides valuable metadata for researchers of 20th century American literature working in fields such as cultural analytics, book and publishing history, and the sociology of literature. In cultural analytics, recent scholarship has used bestseller status as a rough proxy for popularity, enabling researchers to computationally model the textual boundaries between, for instance, popular and prizewinning fiction (Algee-Hewitt and McGurl, 2015; Piper and Portelance, 2016; English, 2016). Previous research of this kind has often relied on the *Publishers Weekly* annual bestseller list. Although *Publishers Weekly* also publishes a weekly list, it is not readily accessible to researchers. In contrast to the *Publishers Weekly* annual list, this dataset reports weekly bestsellers, and therefore captures a much broader subset of the historical literary marketplace.

The larger and more granular *New York Times* dataset presents researchers with a number of potential uses. First of all, existing experiments on bestsellers and prizewinners could be reproduced with this new data. The broader scope of this dataset is likely to dampen the apparent difference between prizewinners and bestsellers, as many prizewinners made it onto the *Times* list without making it onto that of *Publisher's Weekly*. Second, the broader scope of the *Times* list provides a valuable resource for constructing corpora of historical popular literature. Weekly bestsellers have been neglected in humanities corpora relative to yearly bestsellers. Finally, the *Times* list could be used to support ongoing research at the intersection of literary and publishing history. As the most closely-followed public-facing bestseller list, the *Times* list offers insight into the works considered valuable by publishers.

III. Collection and Creation

Data for *New York Times* bestsellers was scraped from Hawes Publications, an online repository that publishes a PDF transcript of the list for every year of the last going back to 1931. Though the Hawes files are high-quality, they are only available as PDF images. Plain text was extracted from the Hawes files programmatically with the open-source Python library *pdfminer*. Though the Hawes files did not come as a structured or tabular dataset, they do report bestseller information in a relatively standardized format. This allowed author, title, date, and rank information to be extracted from the plain text with a mixture of regular expressions and logical operations.

HathiTrust volume identifiers were matched based on string comparisons against the "volumemeta" dataset described by Ted Underwood *et al* (2020). I used the *shorttitle* and *author* metadata fields. First, author surnames were extracted heuristically based on spacing and punctuation. Then, title and author fields in both datasets were lowercased and stripped of punctuation. Two works were then considered a match if surnames were exact matches and the *Times* title field overlapped at the *beginning* of the HathiTrust *shorttitle* field. This yielded 4,978 matches. Note that this includes duplicates, as HathiTrust is a volume-level collection. This conservative matching procedure was chosen over a more generous fuzzy matching procedure in order to maximize the accuracy of matches at the expense of recall. Manual inspection suggests that many of the missed matches were in fact absent from the HathiTrust collection, but the exact number of missed potential matches is uncertain.

IV. Ethics

Researchers who use this dataset are encouraged to consider the limitations of drawing historical or cultural conclusions from bestseller data. Bestseller lists are not a transparent window into what the American public was "really reading" at a given historical moment; rather, they reflect editorial decisions about how and what to count. In particular, historical trends on this list are complicated by institutional shifts in book distribution that occurred during the period which it covers. The increased importance of mall stores, chain stores, and retail distributors continually altered the composition of the bookstores surveyed by the *New York Times* (Miller, 2006). As such, the contents of this dataset likely reflect the purchasing habits of only a particular segment of the American population, namely, those that shop at malls and chain bookstores. This population was disproportionately suburban, white, and middle-class for much of the history of the list. The list likely undercounts sales at other outlets, such as independent bookstores and religious stores.

Users of this data should also be aware that hardcover sales at bookstores are especially unrepresentative of the broader book market in the early years of the "paperback revolution" after WWII, when most popular novels were sold in paperback format at non-bookstore outlets like drugstores. These sales are entirely uncounted on bestseller lists, leading to the conspicuous absence of authors like Erle Stanley Gardner and Mickey Spillane, two of the most popular novelists of the early postwar period.

The *Times* only expanded its coverage to include nationwide bestsellers in September of 1945. Before that, entries are based on sales in New York or other metropolitan areas. The exact methods used by the *Times* are not public and the newspaper has come under periodic criticism for its bestseller reporting. For a full discussion of how the bestseller list is constructed, see Miller (2000).

This dataset does not reveal anything that might be considered sensitive. All of the data in this dataset is freely available in publicly-accessible archives, as well as in the pages of the *New York Times* itself.

Works Cited

Algee-Hewitt, Mark and Mark McGurl. "Between Canon and Corpus: Six Perspectives on 20th-Century Novels." *Pamphlets of the Stanford Literary Lab*, Pamphlet 8, 2015.

English, James F. "Now, Not Now: Counting Time in Contemporary Fiction Studies." *Modern Language Quarterly*, Vol. 77, No. 3, 2016, p. 395–418.

Hawes Publications. "Adult New York Times Adult Hardcover Best Seller Listings." *Hawes Publications*. http://www.hawes.com/pastlist.htm

Miller, Laura J. "The Best-Seller List as Marketing Tool and Historical Fiction." *Book History* Vol. 3, 2000, p. 286-304.

Miller, Laura J. *Reluctant Capitalists: Bookselling and the Culture of Consumption*. University of Chicago Press, 2006.

Piper, Andrew and Eva Portelance. "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading." *Cultural Analytics*, 10 May, 2016.

Sinykin, Dan N. "The Conglomerate Era: Publishing, Authorship, and Literary Form, 1965–2007." *Contemporary Literature*, Vol. 58 No. 4, 2017, p. 462-491.

Shinyama, Yusuke, developer. *PDFMiner*. Version 20191010. Nov. 2019. Software. https://pypi.org/project/pdfminer/

Sorensen, Alan T. "Bestseller Lists and Product Variety,' *The Journal of Industrial Economics*, Vol LV, No. 4 (December 2007): 715-738

Underwood, Ted, Patrick Kimutis, and Jessica Witte. "NovelTM Datasets for English-Language Fiction, 1700-2009." *Cultural Analytics*, May 28, 2020. https://doi.org/10.22148/001c.13147