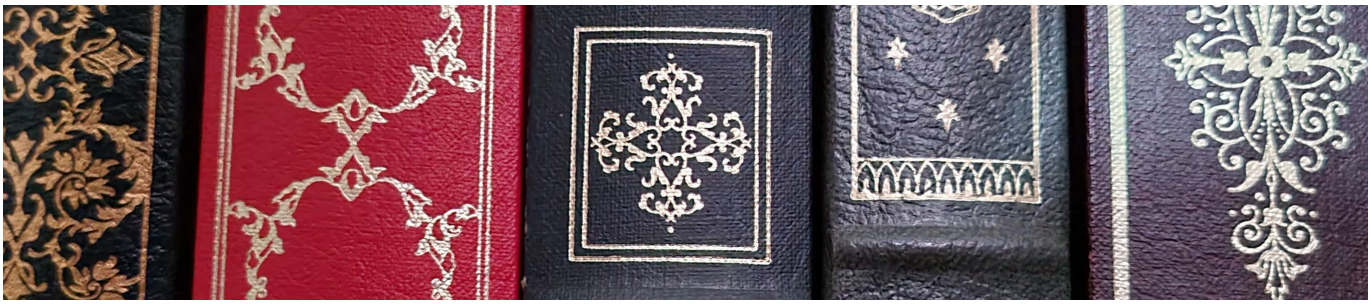

Classic Fiction of the World

Dec 7, 2023 • by Jacob B. Fisher • 5 min read

Insights on the Top 500 Novels

Introduction

I am a university student, ~~and~~, contrary to what many around me seem to believe, university is in fact, for the purpose of studying dusty old books. It may seem ironic for a statistics student to be making that claim, but getting a B.S. does not exclude me from knowing the true purpose of university study, nor does it stop me from achieving that purpose. Being well-read is an excellent goal for anyone and one which typically requires the reading of a large amount of "great" fiction.

The stories that shaped the world into what it is today are worth our time. But how many people's time is it worth? And where in the world are these great stories coming from? Let's take a look at just how bestselling these great works are and if there is any particular region where the best books ~~are~~ from.

Data Collection

Phase 1: Top 500

My list of the top 500 works of fiction comes from the OCLC Library 100 List. Originally, I wanted to include non-fiction, such as the works of Plato and Aristotle as well as religous texts, but since this, ~~the easiest~~ accessible list, only deals with novels, I decided to shift focus to fiction. The books on this list are ranked by number of libraries across the world in which they are found.

The scraping was easy as the entire list of 500 books and their rank, author, and genre are contained on a single page with consistent html to parse. I simply scraped the html, then parsed it with `BeautfiulSoup`. (Note: the books that did not have a genre are labeled as general fiction in my dataset.)

Phase 2: Country

The country data was difficult to find, but this was the primary piece of data I wanted to look at, so I did not give up. I eventually found https://thegreatestbooks.org/ which allows you to search books based on country, but nowhere on the website was that data link to each book. I have to give a big thanks to Shane Sherman, the creator of https://thegreatestbooks.org/ who responded to my inquiry about this and later provided me with this data in a `.json` file. This file provided me with the published year of each book and the country data.

The country data may not be entirely accurate since determining the "nationality" of a book is a bit more difficult than determining the nationality of its author. Some authors change their citizenship througout their lives. For example, Einstein was German-born, but later became a Swiss citizen then an American citizen, but determining if any given work of his is German, Swiss, and/or American may not be as straightforward as looking at his citizenship when he published it. Thus, some of the country data may be disputed, but this is what we are working with.

Phase 3: Bestseller

The data regarding whether or not a work is a bestseller was obtained from the List of best-selling books Wikipedia page. Again, very easy to scrape as all the data are in the first four tables of the page. I simply used the `read_html` function in `pandas`.

Certain books are not listed as bestsellers based on this list that one expects to be. *Don Quioxte*, *The Lord of the Rings* books and others are reported to have very high selling numbers, but due to uncertainty are excluded from the list.

Data Cleaning and Prep

Text data is not fun to clean. Believe or not, the titles of books are not consistent between all people that are interested in books. Some know and include subtitles, others do not. Some will remember the phrase "The Adventures of" in a book title, others will append it unnecessarily. Books written in other languages can have different English transliterations. Did you know that the full, origianl title of the novel commonly refered to as *Robinson Crusoe* by Daniel Defoe is actually *The Life and Strange Surprizing Adventures of Robinson Crusoe, of York, Mariner: Who lived Eight and Twenty Years, all alone in an un-inhabited Island on the Coast of America, near the Mouth of the Great River of Oroonoque; Having been cast on Shore by Shipwreck, wherein all the Men perished*

but himself. With An Account how he was at last as strangely deliver'd by Pyrates. Written by Himself.? You will find different lengths of this title particular across the internet.

All this is to say that cleaning these data so that the dataframes can merge on "Title" took longer than expected, and it required a lot of individual cases to make it work right. Surprisingly and thankfully, no problems occured with any other variables in the data across the three data sources. The main cleaning done was making the titles all lower-case, eliminating the word "the" from the beginning of titles, were the main things that had to be done to ensure as few specific cases needed to be addressed as possible.

Ethical Considerations

Before scraping, I checked the robots.txt files. In addition, since I was scraping from the html, there was no worry of making mulitple requests that might slow the sites down for other users. I see no reason as to why bestselling data and the library rank data should not be readily available in the same dataset. Likewise, since the country a book can be attributed to be from should ideally be known by the average reader, I see no issue with this data being found along with the rest in a single dataset.

Conclusion

There it is. All done. For the code to reproduce the dataset or to find the dataset I created, visit my GitHub repository. Stay tuned for further analysis of this data so that your mind and soul may be expanded.

"Ignorance, the root and stem of every evil." — Plato

JACOB B. FISHER
jbrentfisher@gmail.com

jbfish00
jbfish00

View Source Repo On GitHub