Audrey Lin
DSCI549
HW1

Goodbooks-10k Dataset

1 Introduction

At 10 years old, I moved from LA to Beijing with my family. In the expat community, everyone was always coming and going, so I couldn't escape the feeling that my time in Beijing was temporary. Anticipating my next, inevitable move, I didn't like to hold onto physical things, because I could only think of the nuisance they'd eventually be to pack and ship. Rather than accumulate possessions, I cataloged experiences through blogging, writing and photographing travels, student life, books, and more. After eight years, I left Beijing--just as I knew I would--and headed back to my home country for undergrad, but I never stopped blogging.

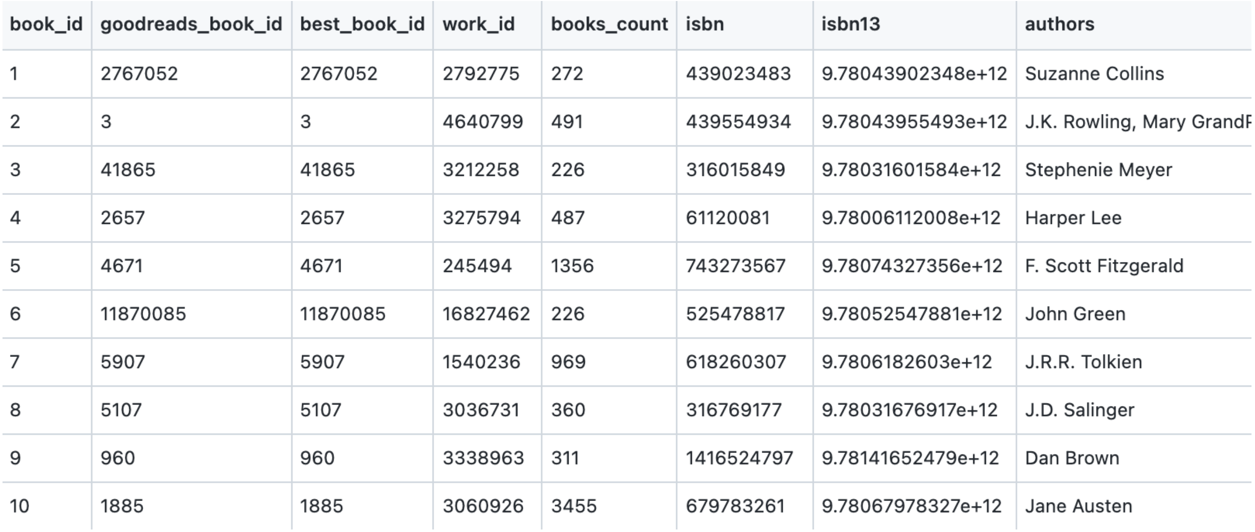
Given my history with blogging (and by extension, social media), I am fascinated with digital culture topics such as online communities, the decentralization of the internet and the widening divide of common truths, the performance inherent in the construction of self on the internet, the distance and intimacy enabled by technology, the relentless pursuit of self-optimization enabled by technology, the underlying biases encoded in technology, and more. Some of my favorite books that touch on these topics include *Trick Mirror* by Jia Tolentino and *Uncanny Valley* by Anna Wiener.

Though I double majored in computer science and linguistics during my undergraduate studies, my liberal arts education gave me space to learn in other disciplines as well--such as user experience design, graphic design, anthropology, and more--through which I could explore the aforementioned topics from a variety of perspectives. When I found out about the MS in Communication Data Science program at USC, I felt that it was the perfect multidisciplinary program for me, tying in my technical academic knowledge and non-technical experience with media.

Despite my experience with programming, I believe there's still a lot for me to learn specific to data science from DSCI549, especially as it is not a class geared toward programming but rather toward high level concepts specific to data science. Programming is just one part of a data science project, but by the end of the semester, I hope to have developed the skills to manage a whole data science project from beginning to end.

Beyond this course, my personal ambition is to understand how we can better use technology to make genuine connections with other people, including fostering one-on-one friendships and building online communities that improve quality of life. Part of this also involves sifting through the deluge of information available online, presenting the right information to the right people in a meaningful way, avoiding echo chambers, combating misinformation, and more.

2 Dataset



book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors
1	2767052	2767052	2792775	272	439023483	9.78043902348e+12	Suzanne Collins
2	3	3	4640799	491	439554934	9.78043955493e+12	J.K. Rowling, Mary GrandP
3	41865	41865	3212258	226	316015849	9.78031601584e+12	Stephenie Meyer
4	2657	2657	3275794	487	61120081	9.78006112008e+12	Harper Lee
5	4671	4671	245494	1356	743273567	9.78074327356e+12	F. Scott Fitzgerald
6	11870085	11870085	16827462	226	525478817	9.78052547881e+12	John Green
7	5907	5907	1540236	969	618260307	9.7806182603e+12	J.R.R. Tolkien
8	5107	5107	3036731	360	316769177	9.78031676917e+12	J.D. Salinger
9	960	960	3338963	311	1416524797	9.78141652479e+12	Dan Brown
10	1885	1885	3060926	3455	679783261	9.78067978327e+12	Jane Austen

Figure 1: First 10 rows of data from books.csv of Goodbooks-10k dataset; first eight columns of 23 columns.

The dataset I have selected is Goodbooks-10k, representing 6,000,000 ratings from more than 50,000 users for the 10,000 most popular books on GoodReads, a social website for users to rate and recommend books. (Personally, I have used GoodReads since 2014.) Goodbooks-10k can be manually accessed by downloading the ZIP files on the Goodbooks-10k GitHub repository's release page. For convenience, a sample of the data can be previewed directly on GitHub. Otherwise, the data must be downloaded in order to be viewed, as the original files are too large to be viewed directly on GitHub. The dataset can also be accessed through the Spotlight Python library.

The key files of the dataset include five CSV files of tabular data (see Figure 1), as well as an archive of 10,000 XML files of structured data (one XML file for each book). CSV files and XML files are digital and can be read and analyzed by computer programs, so they are machine processable.

Book metadata contained in the CSV files include book titles, authors, publication year, average ratings, number of overall ratings, number of one star ratings through number of five star ratings, number of witten reviews, language of publication, tags, and more.

This dataset is licensed under the Creative Commons Attribution ShareAlike 4.0 International License (CC BY-SA 4.0). This means that the data can be used commercially and can be adapted, but must be distributed under the same license and given attribution to the original creator of the dataset, Zygmunt Zajac.

3 Project sketch

There are many possible data science projects that could use this dataset:

A book recommendation system could take a book title a reader has enjoyed as input, then recommend a small list of book titles to the reader as an output. This could possibly be done by analyzing user ratings of the input book and linking them to other books that users have also rated highly. It could also be done by analyzing tags of the input book and linking them to other books that have similar tags.

However, it must be noted that GoodReads tags are far from standardized and could benefit from data cleaning, which could involve both standardizing tags as well as sorting tags into further categories, such as genres, target age groups, and author demographics.

Beyond cleaning up the existing data, more data can be collected as well, such as information about book sales. Experts such as librarians, book sellers, professional book reviewers, literature professors, and those working in the publishing industry could help identify other important categories, such as writing style.

With more specific ways to structure the data, different methodologies for book recommendation systems could be implemented and compared, and more ways to ask classification problems could open up as well. Possible classification problems include classifying whether or not a book has been on the New York Times bestseller list, whether a book is classic or contemporary, or classifying a book into a specific genre. For these classification problems, a book title would be the input and the output would be the class.

Statisticians and machine learning experts would be helpful for the technical aspects of both the recommender and classification projects.

4 Works cited

Zajac, Zygmunt. *Goodbooks-10k: a new dataset for book recommendations*, 2017.
https://github.com/zygmuntz/goodbooks-10k