Kathleen Hablutzel

Professor Kinnaird

CSC 294 Computational Machine Learning

26 March 2021

Building a Book Recommender for Goodreads Users

Goodreads is a popular website where avid readers rate books they have read and find new book recommendations based on their individual tastes in literature. The website harnesses its large community of users to recommend books when they have been rated highly by other users with similar preferences. In this project, we examined a large database of Goodreads book ratings as posted on Kaggle, with the goal of building a recommender system which finds three new book recommendations for users within the dataset. We chose this particular recommender task because individual book preferences are extremely subjective, often span many genres, and can vary greatly depending on many factors such as content, author styles, and changing user preferences over time, so we were curious what patterns our recommender would find beyond typical categories such as genres, authors, and series.

We selected the original dataset for its accessibility on Kaggle and its large size. The dataset was originally posted by Kaggle users Bahram Jannesar and Soroush Ghaderi, who scraped book ratings from Goodreads using the Goodreads Python library and the Goodreads API [1]. The data have been updated many times since June 2020, with the last update in December 2020. The dataset as posted on Kaggle contains two types of data files: user ratings and book information. However, we only utilized the first two user ratings files, which contained book ratings given by the first approximately 2000 users in the dataset.

In the original files, each observation is one rating, with three columns for the book name, a numerical ID of the user giving the rating, and the rating on a five-point scale from “it was amazing” to “did not like it”. We translated the ratings onto a discrete numerical scale from 1 (“did not like it”) to 5 (“it was amazing”), and pivoted the data such that each row is an individual user and each column is a different book, with ratings as values. Since most users had only rated a very small subset of books within the dataset, the data became very sparse, so we slightly reduced this sparsity by selecting only books with 20 or more recommendations and users who had rated 5 or more of those books. Still, our 494 user by 436 book dataframe contained only 18827 total ratings, for a total density of 8.7%, so our data were quite sparse (see Figure 1).

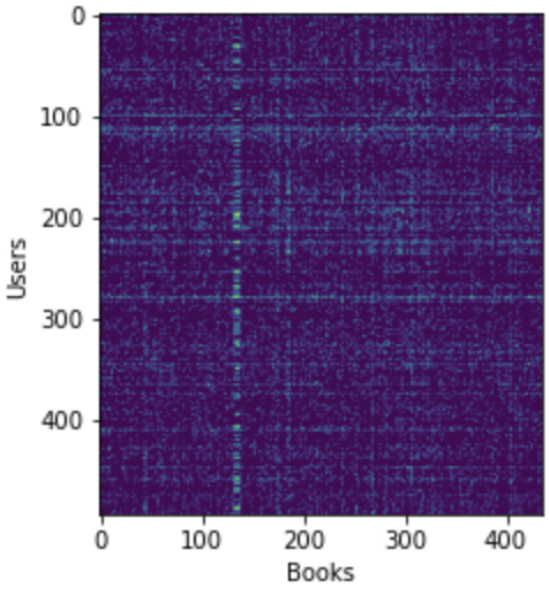


Figure 1. Our dataframe of user ratings shows some popularly rated books and some avid readers, but otherwise is quite sparse.

In order to recommend three new books for each user, we built a recommender system using Singular Value Decomposition (SVD). The recommender reduces the data to five dimensions of variation, a significant reduction compared to the original 436 dimensions from

the 436 unique books in the dataset, then projects the reduced data back into the original 494 by 436 shape. In doing so, the recommender generalizes patterns of variation into five dimensions, then expands upon those patterns to form predictions for users' ratings of all the books in the dataset. We chose SVD over Principal Component Analysis (PCA) for our dimension reduction technique because SVD works most consistently sparse matrices like this dataset and does not require the data to be centered around 0. However, the recommendation procedure would be very similar with PCA. Additionally, we chose to reduce to five dimensions of variation because this number would be a major reduction from the original 436 dimensions, but should still maintain enough of the original variation in the data to produce reasonable predictions.

Once the recommender has formed predictions for users' ratings of all the books in the dataset, our recommender removes ratings for books that a user has already rated, leaving only *new* recommendations for books the user had not rated in the original data. The recommender also removes projected ratings of 0, where the predictions recreated some of the sparsity of the original data, and rounds other ratings to the nearest whole number within the original 1-5 rating scale. From there, the recommender simply selects the three new recommendations with the highest predicted ratings for each user. The output is a dataframe with one row per user, with columns for the user ID, the first recommendation's title, the first recommendation's predicted rating, and the same for the second and third recommendations (see Figure 2a). If a user receives fewer than three new recommendations, removed book ratings will show up with a title of "None" and a rating of "N/A" (see Figure 2b).

To subjectively assess the quality of our system's recommendations, we examined our system's performance for two users in the dataset.

	User	Book 1	Rating 1	Book 2	Rating 2	Book 3	Rating 3
0	0	A Game of Thrones (A Song of Ice and Fire, #1)	5	Hamlet	4	Animal Farm	4
1	1	The Nightingale	1	The Subtle Knife (His Dark Materials, #2)	1	The Golden Compass (His Dark Materials, #1)	1
2	2	Wuthering Heights	4	Of Mice and Men	4	Where the Sidewalk Ends	4
3	3	The Book Thief	1	The Nightingale	1	The Great Gatsby	1
4	4	The Catcher in the Rye	1	The Great Gatsby	1	Memoirs of a Geisha	1
...	...	...	...	...	...	...	...
489	489	To Kill a Mockingbird	2	Pride and Prejudice	2	Memoirs of a Geisha	2
490	490	None	N/A	None	N/A	None	N/A
491	491	The Sun Also Rises	2	The Adventures of Huckleberry Finn	3	Brave New World	2
492	492	The Stranger	1	The Great Gatsby	2	The Color Purple	1
493	493	The Great Gatsby	1	Jane Eyre	1	Angels & Demons (Robert Langdon, #1)	1

Figure 2a: An excerpt from the dataframe of recommendations for all 494 users.

	User	Book 1	Rating 1	Book 2	Rating 2	Book 3	Rating 3
490	490	None	N/A	None	N/A	None	N/A

Figure 2b: Removed books—those which the user had already rated or with predicted ratings of 0—appear with title "None" and rating "N/A" when users receive fewer than three new recommendations.

The first user, user number 0 in our cleaned data, is an avid reader who has rated 123 books on Goodreads (see Figure 3a). The user loves fantasy novels, having given 5-star ratings to most of the *Harry Potter*, *Lord of the Rings,* and *A Song of Ice and Fire* series. They also enjoy classic science fiction reads such as *The Hitchhiker's Guide to the Galaxy* and *Ender's Game*, and dystopian political fiction such as *1984*, *Fahrenheit 451*, and *The Handmaid's Tale*. User 0 appears to be enrolled in an advanced high school English course, having reviewed numerous classic high school English reads such as *Great Expectations, Of Mice and Men, Pride and Prejudice, The Adventures of Huckleberry Finn, The Great Gatsby, The Scarlet Letter,* and

Their Eyes Were Watching God, in addition to giving 4-star reviews to two Shakespeare works *A Midsummer Night's Dream* and *Romeo and Juliet*.

	Book	Rating
0	1984	5.0
1	A Clash of Kings (A Song of Ice and Fire, #2)	5.0
2	A Dance with Dragons (A Song of Ice and Fire, #5)	5.0
3	A Feast for Crows (A Song of Ice and Fire, #4)	4.0
4	A Midsummer Night's Dream	4.0
5	A Short History of Nearly Everything	5.0
6	A Storm of Swords (A Song of Ice and Fire, #3)	5.0
7	A Thousand Splendid Suns	5.0
8	A Time to Kill (Jake Brigance, #1)	2.0
9	A Visit from the Goon Squad	5.0

Figure 3a: An excerpt of User 0's 123 book ratings.

	User	Book 1	Rating 1	Book 2	Rating 2	Book 3	Rating 3
0	0	A Game of Thrones (A Song of Ice and Fire, #1)	5	Hamlet	4	Animal Farm	4

Figure 3b: Recommendations for user 0.

In light of all these reviews, our recommender's suggestions of *A Game of Thrones*, *Hamlet*, and *Animal Farm* are no surprise (see Figure 3b). Since the user has rated many fantasy books and gave mostly 5-star ratings to the second through fifth books of the *A Song of Ice and Fire* series, we were not surprised that our system recommended the first book in the series, *A Game of Thrones*, with a rating of 5. As our peer reviewer Sophia Hager noted, the user likely has already read the first book and simply has not entered a rating on Goodreads [2]. The second recommendation, *Hamlet*, follows this user's theme of classic high school English reads, though the only 4-star recommendation makes sense considering that the user has given *A Midsummer Night's Dream* and *Romeo and Juliet* the same rating. Finally, the third recommendation, *Animal Farm*, seems very much in line with this user's reading history, considering their interest in

dystopian political literature. We might have expected a 5-star rating since they gave a top rating to another of Orwell's works, *1984*, but the user did give only 4 stars to *A Handmaid's Tale*, so a 4-star rating is also reasonable to expect of the user among these classic dystopian works. Overall, all three recommendations were very good fits for this user.

The second user, user number 2 in our cleaned data, is another avid reader who has rated 130 books on Goodreads (see Figure 3a). This user also seems to be a high school student, but has been a Goodreads user since at least middle school, and has reviewed books at levels ranging from young childhood to high school. The user has reviewed numerous classic young children's books such as *A Light in the Attic*, *Alice's Adventures in Wonderland & Through the Looking-Glass*, *Charlotte's Web*, *Goodnight Moon*, and *The Giving Tree,* and classic books for

	Book	Rating
0	A Farewell to Arms	5.0
1	A Light in the Attic	4.0
2	A Little Princess	5.0
3	A Midsummer Night's Dream	4.0
4	A Prayer for Owen Meany	3.0
5	A Room with a View	5.0
6	A Separate Peace	5.0
7	A Thousand Splendid Suns	5.0
8	A Tree Grows in Brooklyn	5.0
9	A Visit from the Goon Squad	5.0

Figure 4a: An excerpt of User 2's 130 book ratings.

	User	Book 1	Rating 1	Book 2	Rating 2	Book 3	Rating 3
2	2	Wuthering Heights	4	Of Mice and Men	4	Where the Sidewalk Ends	4

Figure 4b: Recommendations for user 2.

slightly older children such as *Anne of Green Gables*, *Bridge to Terabithia*, and *Where the Red Fern Grows*. As the user has grown older, they have reviewed standard middle school reads such as *Flowers for Algernon* and *The Phantom Tollbooth,* and now has started growing into more high school-level classics, such as *Mansfield Park*, *Pride and Prejudice*, *The Catcher in the Rye*, *The Grapes of Wrath*, *and The Secret History*. The user has also developed a fondness for Shakespeare, having reviewed at least five Shakespeare works with 4- or 5-star ratings.

Again, the recommender's suggestions match the user's interests fairly well. The user has been growing into more high school-level classic reads, a category which *Wuthering Heights* and *Of Mice and Men* fit well. Our peer reviewer noticed that the user has already enjoyed similar reads, such as *Pride and Prejudice* and *Mansfield Park* for *Wuthering Heights* and *The Catcher in the Rye* and *The Grapes of Wrath* for *Of Mice and Men* [2]. The user gave 5-star ratings to both *Pride and Prejudice* and *Mansfield Park*, so we could have expected a higher predicted rating for *Wuthering Heights*, but the 4-star expected rating fits well for *Of Mice and Men*, considering that the user gave five stars to *The Catcher in the Rye* but only three stars to *The Grapes of Wrath*. The third recommendation, *Where the Sidewalk Ends*, is a bit different, but it fits the user's affinity for young children's books, and especially other Shel Silverstein works such as *A Light in the Attic* and *The Giving Tree*. Since the user only gave four stars to *A Light in the Attic*, a similar collection of poems, we would reasonably expect the user to give *Where the Sidewalk Ends* a 4-star rating as well. Thus, overall, the recommender gave three solid recommendations to fit this user's interests.

In conclusion, we successfully built a recommender system which recommends new books to Goodreads users within our existing dataset. Our recommender system produces very reasonable recommendations for our two expert users; however, many other recommendations

show only one- to three-star predicted ratings, so the system may be less accurate for some users. Part of this inaccuracy is likely due to the 91.3% sparsity of our dataset; however, we could also attempt to improve our recommender by increasing the number of dimensions in our SVD dimension reduction, thus retaining more of the original variation of the data. Future work could explore the relationship between the number of reduced dimensions and the performance of the recommender, since retaining more variation could provide a more complete picture of the patterns in the data, but retaining too much variation could reproduce the original dataset without any new recommendations. Additionally, it is important to note that any evaluations of the recommender's performance are completely subjective. Thus, while this recommender system will always have room for improvement, we stand by our system as a successful recommender tool for the Goodreads users in our dataset.

Acknowledgements

Many broad approaches were brainstormed with Sophia Hager, who also provided the peer review included in Appendix B. Many thanks to Professor Katherine Kinnaird for her support on this project. Numerous online resources have been crucial to the data wrangling in this project, as cited at the end of `project1_report_figures.ipynb`.

Appendix A

All unit tests passed locally and on Travis.

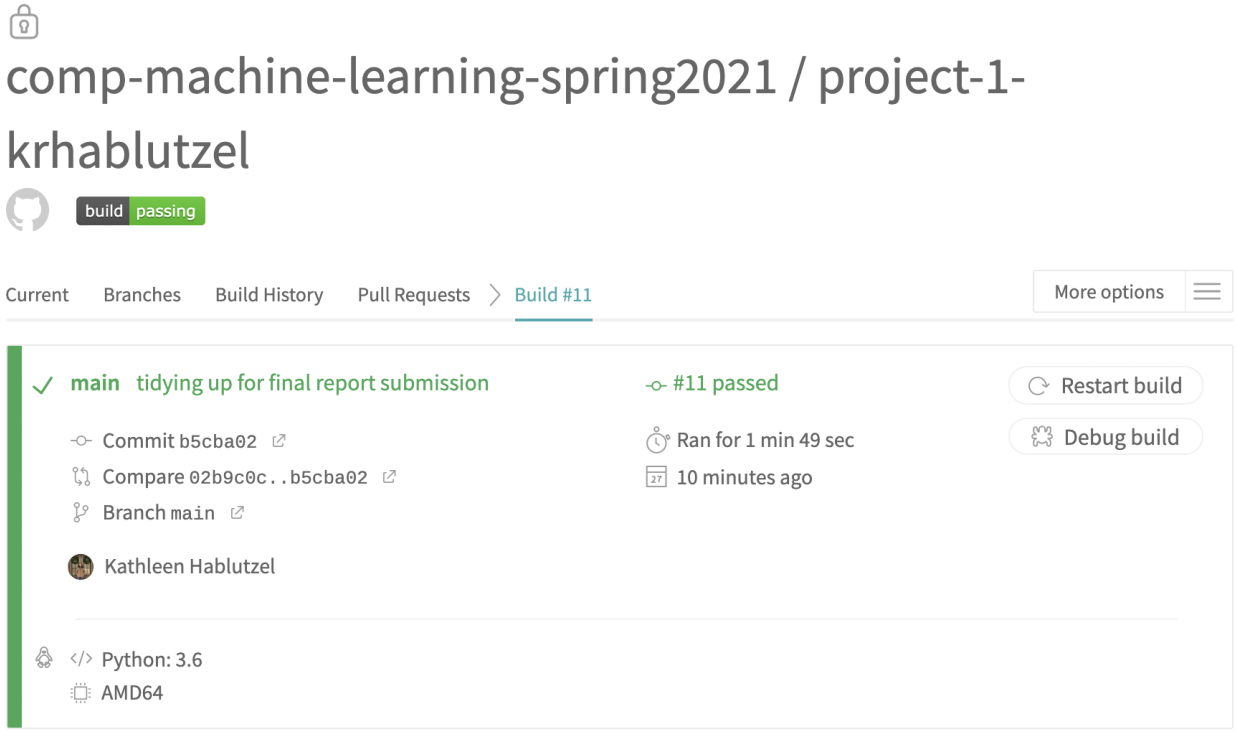
Local Tests

```
(csc294)  krh@Kathleens-MBP  ~/Desktop/CSC294/projects/project-1-krhablutzel  main ±  pytest -v
========================================= test session starts ==========================================
platform darwin -- Python 3.8.5, pytest-6.2.2, py-1.10.0, pluggy-0.13.1 -- /opt/anaconda3/envs/csc294
/bin/python
cachedir: .pytest_cache
rootdir: /Users/krh/Desktop/CSC294/projects/project-1-krhablutzel
collected 4 items

test_project1.py::test_users_type PASSED                                                      [ 25%]
test_project1.py::test_recommendations_type PASSED                                            [ 50%]
test_project1.py::test_recommendations_size PASSED                                            [ 75%]
test_project1.py::test_recommendations_unique PASSED                                          [100%]

========================================== 4 passed in 2.82s ===========================================
(csc294)  krh@Kathleens-MBP  ~/Desktop/CSC294/projects/project-1-krhablutzel  main ±
```

Travis Tests



Appendix B

Sophia Hager provided the following peer review of this recommender's performance:

> "For the first user, I think the recommendations seem pretty appropriate- it looks like they consume a lot of classic literature (like 1984, Fahrenheit 451, Great Expectations, etc), along with a variety of science fiction/fantasy series (A Song of Ice and Fire, Harry Potter, Lord of the Rings,

Hunger Games). A Game of Thrones is a really good recommendation from the data set- they've read the other books in the series (I would actually assume they had read the first book and not entered it in, but based off the given information, this is very fitting). I'm less sure about Hamlet as a recommendation: the only other Shakespeare play they've read is A Midsummer Night's Dream, which they only gave a four. On the other hand, Hamlet is a widely popular play, so I can understand that being a common recommendation. The last recommendation, Animal Farm, seems like it could fit in with the other classics (especially the political fictions of 1984, Handmaid's Tale, and Fahrenheit 451). Overall, I think the books recommended for the first user seem to have a lot in common with their reading history.

"The second user seems like they consume a lot of classics as well, but a lot more books targeted to children, such as The Little Princess and The Secret Garden. Once again, it seems like the books fit within the user's taste: Wuthering Heights is similar to Pride and Prejudice and Mansfield Park, Of Mice and Men is similar to The Catcher in the Rye and The Grapes of Wrath, and Where the Sidewalk Ends fits in both with the poetry book and the children's books." [2]

References

[1] B. Jannesar and S. Ghaderi, *Goodreads Book Datasets With User Rating 10M*, vol. 18, Isfahan, Iran: Kaggle, 2020. Accessed on: Mar. 9, 2021. [Online]. Available: https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m

[2] S. Hager, private communication, Mar. 2021.