Woke Books

a.k.a. Underrepresented Author Book Recommender

Cristine Dewar

cristinedewar@gmail.com
Github:github.com/cristined/woke-books
Linkedin: www.linkedin.com/in/cristinedewar

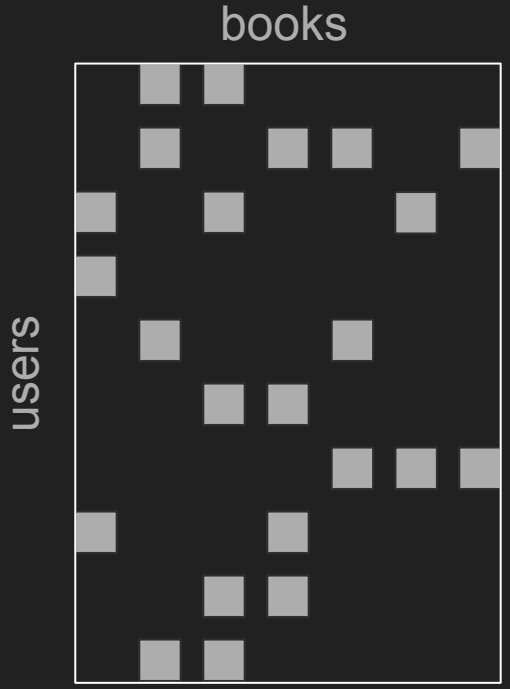
The Mission

Reading books enables you to see the world from another perspective and if we continue to read the same type of authors we box in our world view.

The goal of the project was to build a recommender that would recommend books to a users taste and boost those by races and genders the user has underrepresented in their reading.

Collaborative Filtering

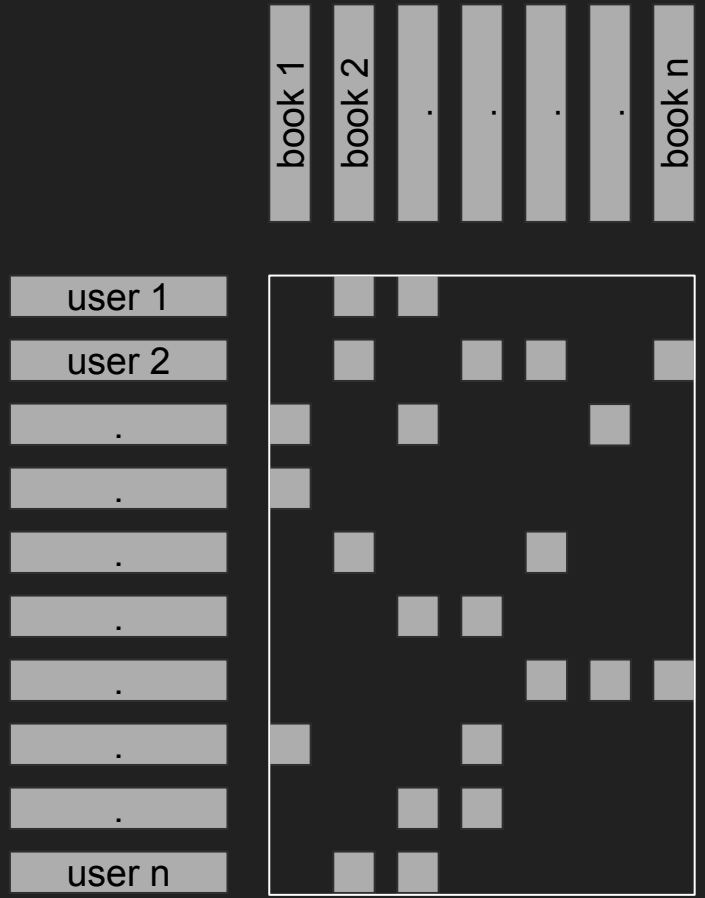
We start with a sparse matrix with user ratings.



Matrix Factorization

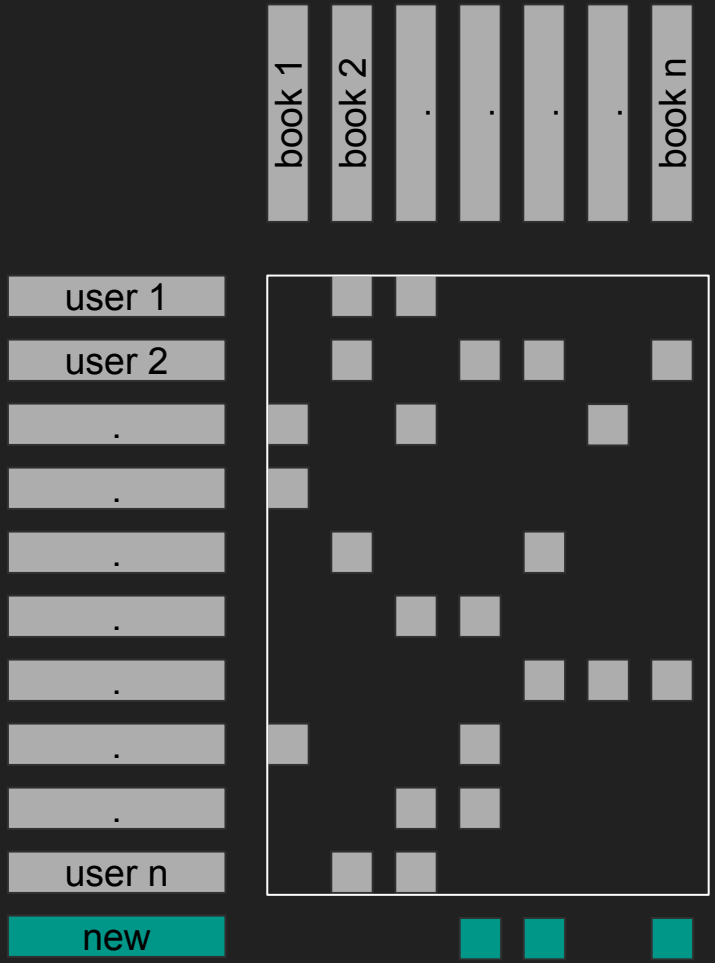
We start with a sparse matrix with user ratings.

Using ALS on Spark we create a user and a book matrix, which can recreate the ratings for the user and predict ratings for unread books for these users.



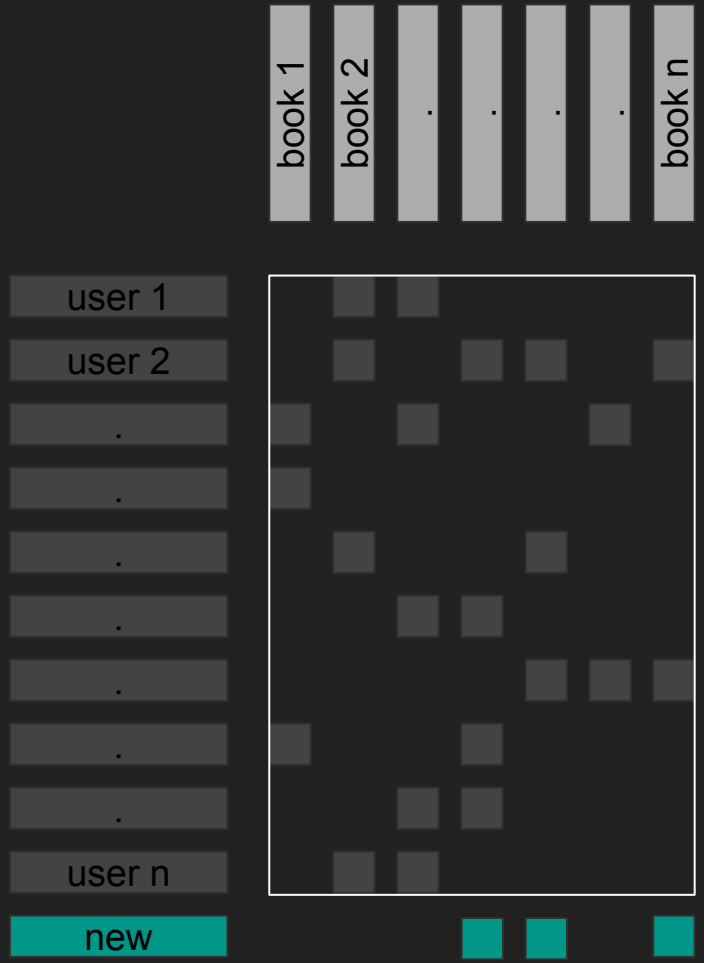
Matrix Factorization

What do we do when a new user signs up?



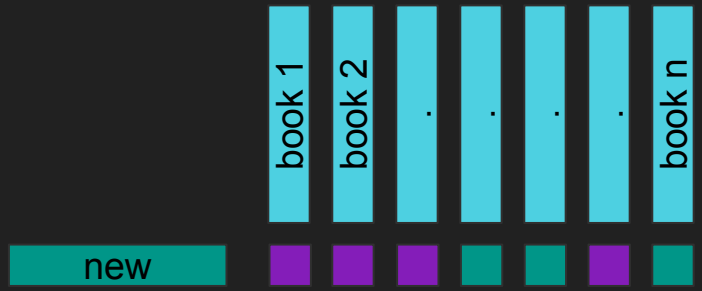
Gradient Descent

The goal was to make the best recommendations based on the sparse matrix of what the user had already rated and the items factor matrix.



Recommendation for Users

I have my new user's vector through gradient descent and the predicted ratings for their unread books.



Recommendation for Users

Let’s start with an example user, if we had to create a label for this user it would be “standard nerd”

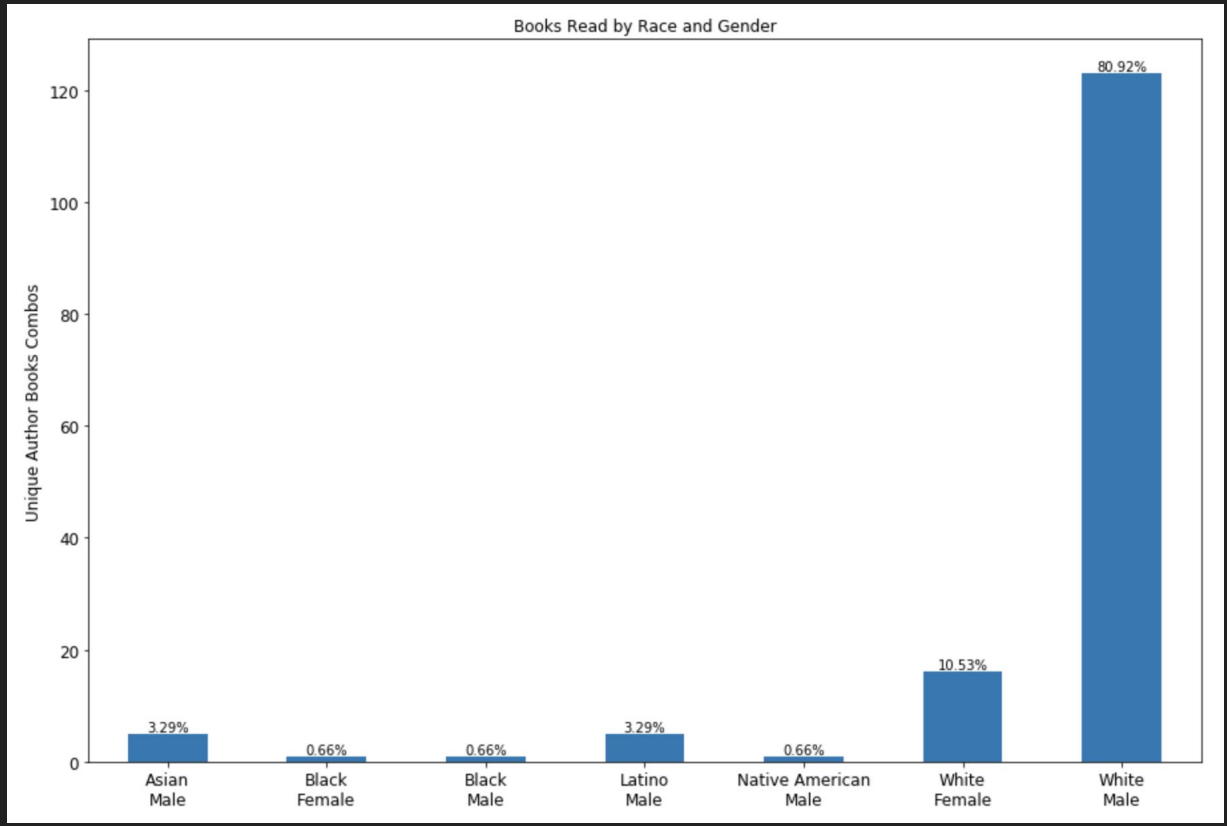
Standard Nerd Recs - Before Boosting

The Fellowship of the Ring (The Lord of the Ri...	J.R.R. Tolkien	WHITE MALE
Harry Potter and the Sorcerer's Stone (Harry P...	J.K. Rowling	WHITE FEMALE
Gravity's Rainbow	Thomas Pynchon	WHITE MALE
Shadows of the Empire (Star Wars)	Steve Perry	WHITE MALE
The Woman in White	Wilkie Collins	WHITE MALE
Vampire Mountain (Cirque Du Freak, #4)	Darren Shan	WHITE MALE
Chapterhouse: Dune (Dune Chronicles #6)	Frank Herbert	WHITE MALE
Fight Club	Chuck Palahniuk	WHITE MALE
World Without End (Kingsbridge, #2)	Ken Follett	WHITE MALE
Salt: A World History	Mark Kurlansky	WHITE MALE

* he would like you to know he has read Lord of the Rings and Dune and not yet rated in Goodreads

Recommendation for Users

The ratings are then boosted by creating a boost between 0 and 1 that is added to the recommendations based on how often each user reads authors of this gender and race compared to a target percentage.



Standard Nerd Recs - After Boosting

Cane River	Lalita Tademy	ASIAN FEMALE
Breath, Eyes, Memory	Edwidge Danticat	LATINO FEMALE
Fullmetal Alchemist, Vol. 1 (Fullmetal Alchemi...	Hiromu Arakawa	ASIAN FEMALE
Love in the Time of Cholera	Gabriel García Márquez	LATINO MALE
Beloved	Toni Morrison	BLACK FEMALE
The Autograph Man	Zadie Smith	BLACK FEMALE
Interpreter of Maladies	Jhumpa Lahiri	ASIAN FEMALE
First They Killed My Father: A Daughter of Cam...	Loung Ung	ASIAN FEMALE
The Atonement Child	Francine Rivers	WHITE FEMALE
The Known World	Edward P. Jones	BLACK MALE

Recommendation for Users

To further personalize the recommendations of the users, I used k-means to cluster the reviews on the aggregated reviews text of the books from amazon enabled me to create categories of books.

The recommendations were shown to the user based on the 5 categories they are most interested in.

Standard Nerd Recs - Top Categories

Collections
['Interpreter of Maladies', 'The Stories of Eva Luna', 'Haroun and the Sea of Stories (Khalifa Brothers, #1)', 'La Dame aux Camélias', 'The Snows of Kilimanjaro and Other Stories', 'Barrel Fever: Stories and Essays', 'Brokeback Mountain']]
==========
Graphic Novels
['Fullmetal Alchemist, Vol. 1 (Fullmetal Alchemist, #1)', 'Bleach, Volume 01', 'V for Vendetta', 'Superman: Birthright', 'Pride and Joy (Runaways, #1)']
==========
Required Reading
['Breath, Eyes, Memory', 'Love in the Time of Cholera', 'Beloved', 'First They Killed My Father: A Daughter of Cambodia Remembers', 'The Atonement Child', 'The Known World', 'Memories of My Melancholy Whores', 'On Beauty', 'Tara Road', 'A Map of the World']
==========
Suspense
["From Potter's Field (Kay Scarpetta, #6)", 'Pardonable Lies (Maisie Dobbs, #3)', 'Carter Beats the Devil', 'Cause of Death (Kay Scarpetta, #7)', 'City of the Beasts (Eagle and Jaguar, #1)', 'Messenger of Truth (Maisie Dobbs, #4)', 'Book of the Dead (Kay Scarpetta, #15)', 'Special Topics in Calamity Physics']
==========
Classics
['The Woman in the Dunes', 'Don Quixote', 'Macbeth', 'A Modest Proposal and Other Satirical Works', 'The Oresteia (Ορέστεια, #1-3)', 'Antigone (The Theban Plays, #3)', 'The Brothers Karamazov', 'The Broken Wings', 'Pale Fire', 'Oedipus Rex (The Theban Plays, #1)']
==========
Top Recommendations not in Your Top Categories
['Cane River', 'The Autograph Man', 'Assassination Vacation', 'Harry Potter and the Order of the Phoenix (Harry Potter, #5)', 'Jesus Freaks: Stories of Those Who Stood for Jesus, the Ultimate Jesus Freaks (Jesus Freaks, #1)', 'Hard Eight (Stephanie Plum, #8)', 'Harry Potter and the Half-Blood Prince (Harry Potter, #6)', 'Deerskin', 'The Secret Garden', 'Everyday Italian: 125 Simple and Delicious Recipes']

We have recommendations!

It is hard to measure the success of recommendations. To do so within the context of this project I used the goodreads accounts and brains of friends to confirm they were getting reasonable recommendations.

Ideally I would like to be able to measure the click through rate on these recommendations so I could find if this would spike their initial interest, whether they marked these books as "to-read", and if they read the book within a certain time period.

Future Work

I would like to continue to improve on the recommender by including user data, time of ratings, and looking deeper into the text/descriptions of the book instead of using reviews alone to create the genres.

Manual classification of the authors was a big limiting factor for scope. Ideally the recommender could expand beyond 10k books, hopefully this would open up the pool of books that would get recommended and boosted up for the user. Also I would like to include LGBTQ and country of origin in the diversity boost score.

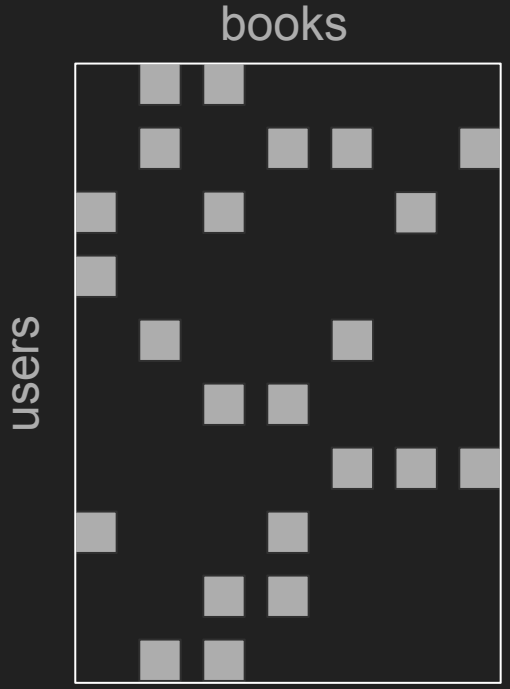
Thank you

cristinedewar@gmail.com
Github:github.com/cristined/woke-books
Linkedin: www.linkedin.com/in/cristinedewar

Appendix

Collaborative Filtering

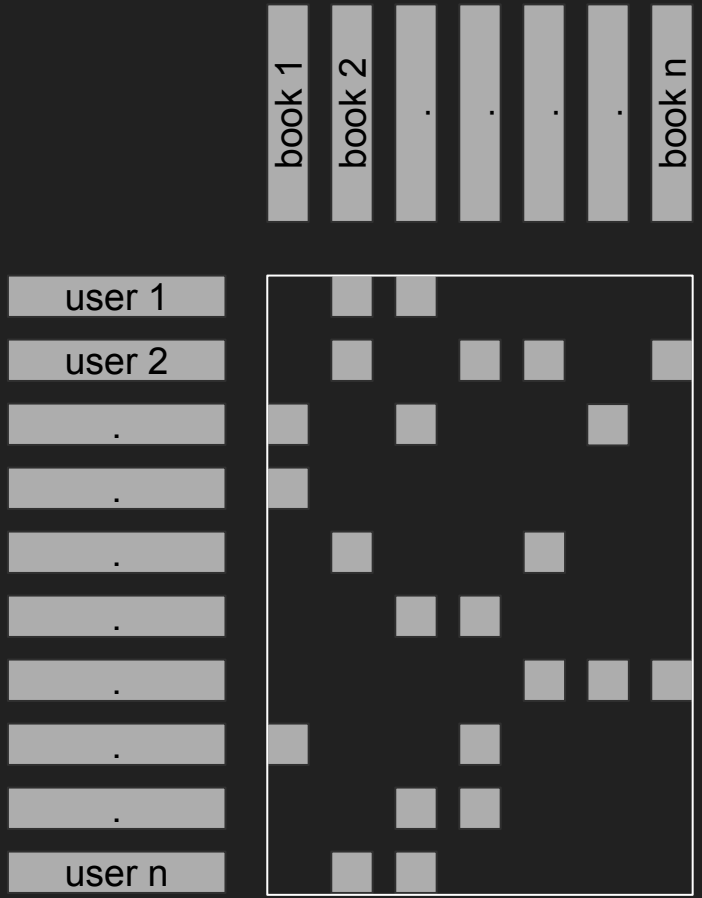
We start with a sparse matrix with user ratings.



Matrix Factorization

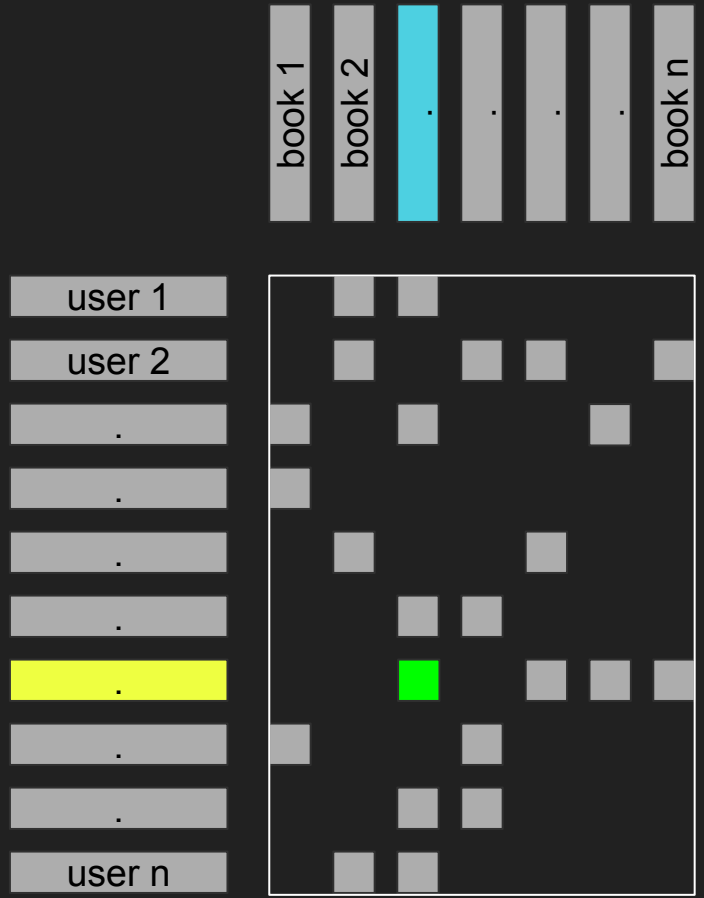
We start with a sparse matrix with user ratings.

Using ALS on Spark we create a user and a book matrix, which can recreate the ratings for the user



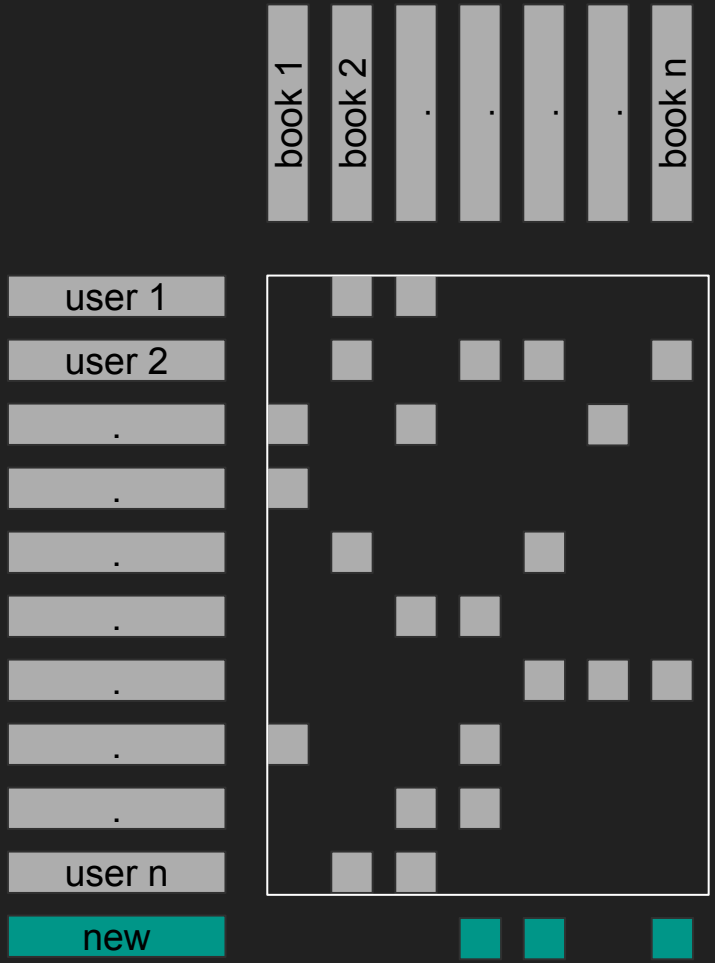
Matrix Factorization

Now we can predict their ratings on unread books. By multiplying the book vector with the user vector.



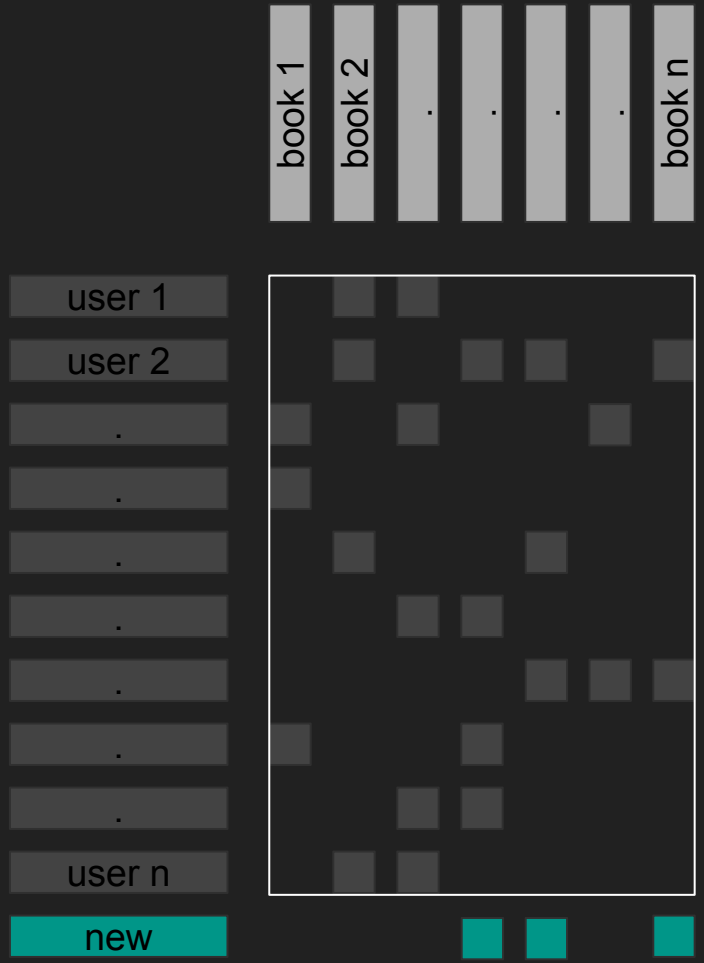
Matrix Factorization

What do we do when a new user signs up?



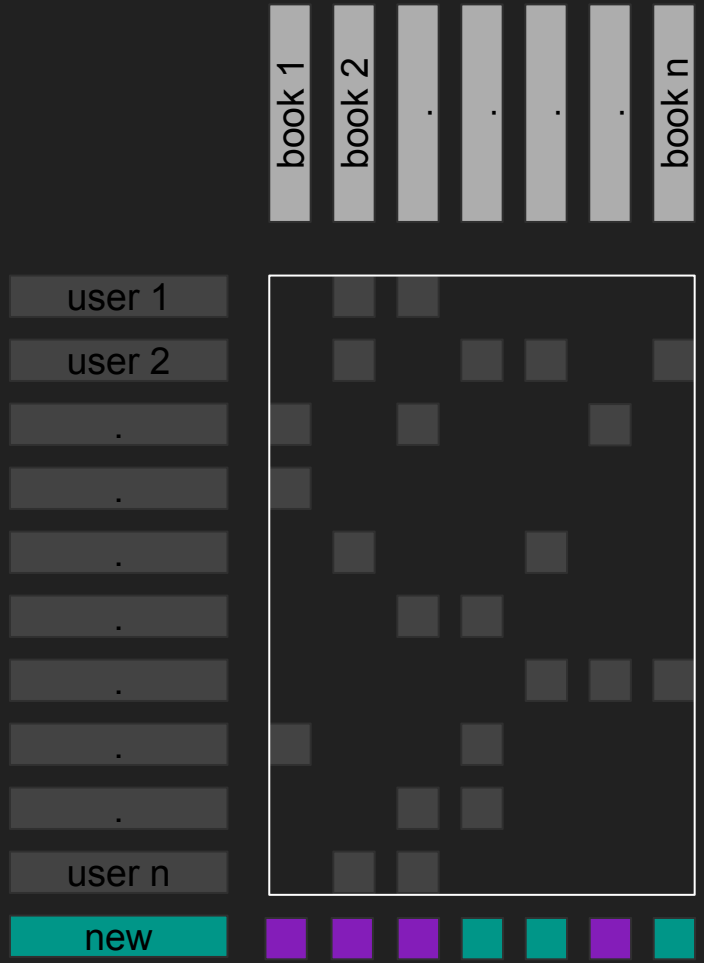
Gradient Descent

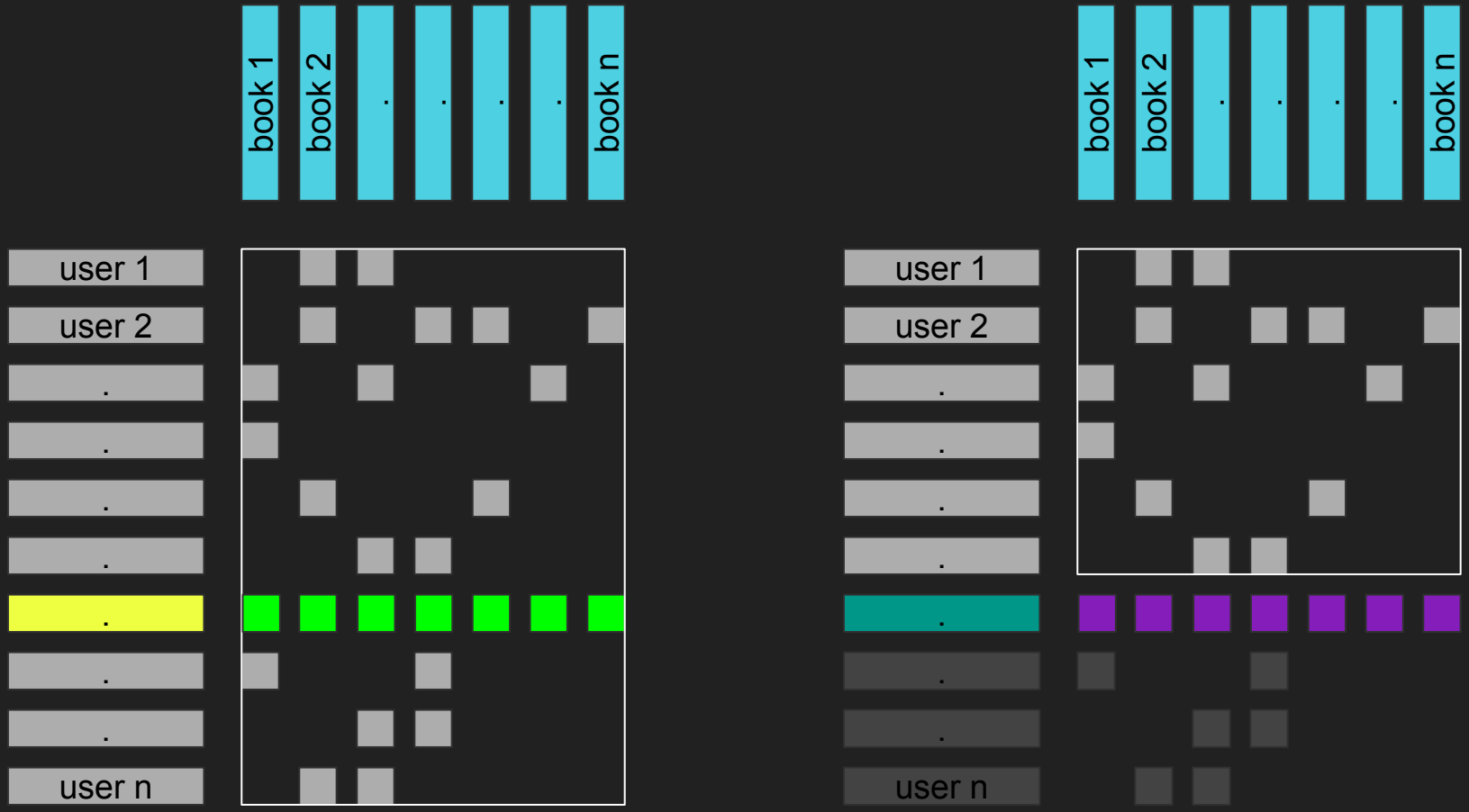
The goal was to make the best recommendations based on the sparse matrix of what the user had already rated and the items factor matrix.



Gradient Descent

It was difficult to find the best metric to determine how close we had gotten to what Spark would have output if we were to have refactored the matrix.

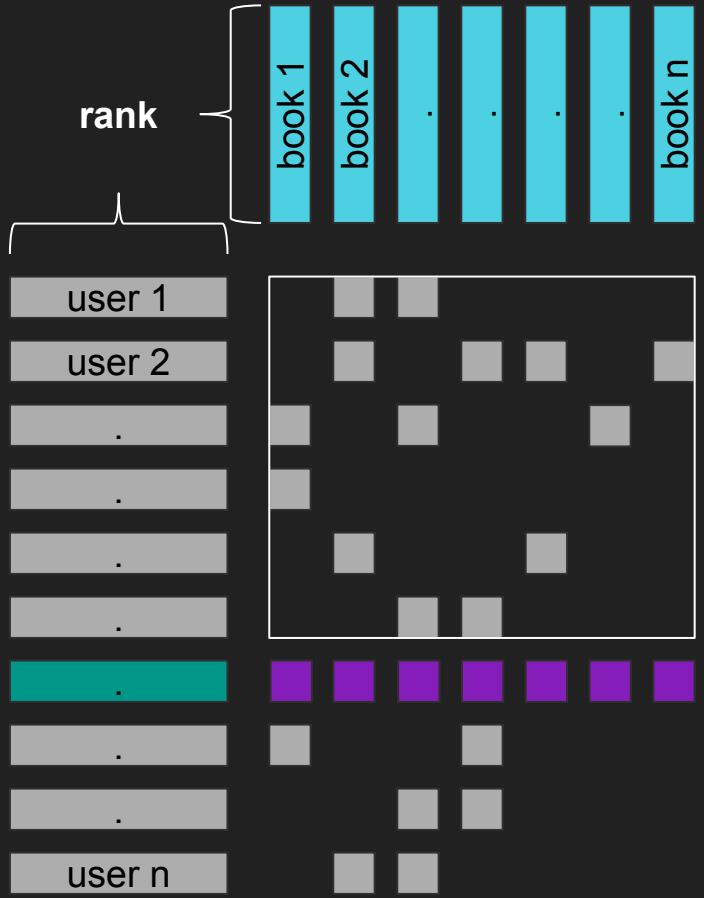




Compare the dot products of the spark user vector vs the gradient descent user vector

Gradient Descent

I looked across different ranks to determine not only what were the best parameters for the gradient descent but what rank of matrix factorization created the best user vector.



Amazon vs Goodreads Ratings