

Goodreads Topics and Recommendations

Eitan Angel

July 30, 2019

1 Introduction

2 Exploration

3 Models

4 Applications

Introduction

Problem: Goodreads Topics, Profiles, and Recommendations

Goodreads is a social site for readers. Readers rate, tag, and discuss books.

Topics: ‘Learn’ book topics from reader ratings data.

- Topics are interpretable; user-generated tags assist
- Simple way to create “Because you liked Harry Potter...” product feature.

Profiles: Topics extracted provide a profile for each reader to aid targeted marketing.

Recommendations: Individually-targeted book recommendations for existing users.

- The most popular 10,000 books on Goodreads by ratings with
 - author(s),
 - title(s),
 - language,
 - count of ratings,
 - number of ratings
- Book ratings by over 50,000 users from 1 – 5.
- User-generated tags: A count for each of the top 100 tags for each of the 10,000 books.

Nonnegative Matrix Factorization

For user-ratings matrix V , make a low-rank approximation

$$V \approx WH$$

- Choose number of latent features
- W a matrix of user latent features
- H a matrix of book latent features
- $A = WH$ “fills-in” blank ratings in V
- Recommend top values of rows of A
- H rows weight tags for topic extraction

Ratings Matrix V

Sam	?	4	?	1
Jim	2	5	?	?
Lex	5	?	?	4
Kat	4	1	3	5
Bob	?	?	4	5

authors	title
Suzanne Collins	The Hunger Games (The ...)
J.K. Rowling, Mary GrandPré	Harry Potter and the Sorcerer's...
Stephenie Meyer	Twilight (Twilight, #1)
Harper Lee	To Kill a Mockingbird
F. Scott Fitzgerald	The Great Gatsby
John Green	The Fault in Our Stars
Veronica Roth	Divergent (Divergent, #1)
J.R.R. Tolkien	The Hobbit
Jane Austen	Pride and Prejudice
J.D. Salinger	The Catcher in the Rye

Table: The most popular books on Goodreads.

authors	title
Bill Watterson	The Complete Calvin and Hobbes
J.K. Rowling, ...	Harry Potter Boxed Set, Books 1-5
Brandon Sanderson	Words of Radiance (The Stormlight ...)
Francine Rivers	Mark of the Lion Trilogy
Anonymous ...	ESV Study Bible
Bill Watterson	It's a Magical World: A Calvin and ...
Bill Watterson	There's Treasure Everywhere: A Calvin ...
J.K. Rowling	Harry Potter Boxset (Harry Potter, #1-7)
J.K. Rowling	Harry Potter Collection (Harry Potter, #1-6)
Bill Watterson	The Indispensable Calvin and Hobbes

Table: Calvin & Hobbes and Harry Potter dominate the average ratings.

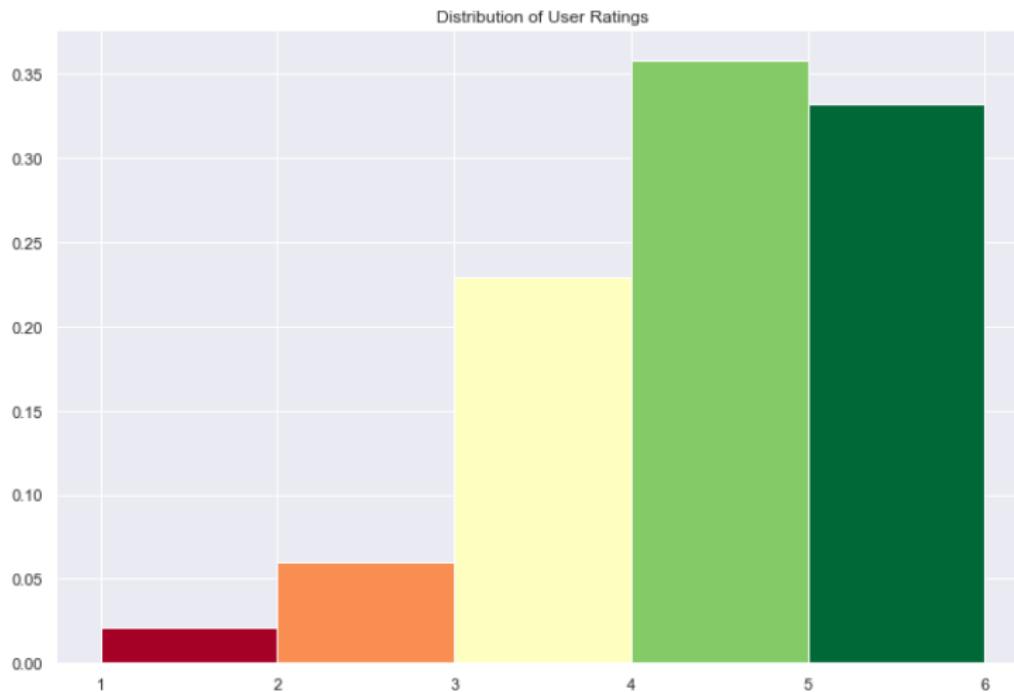


Figure: Ratings of 4 or 5 are by far most common.

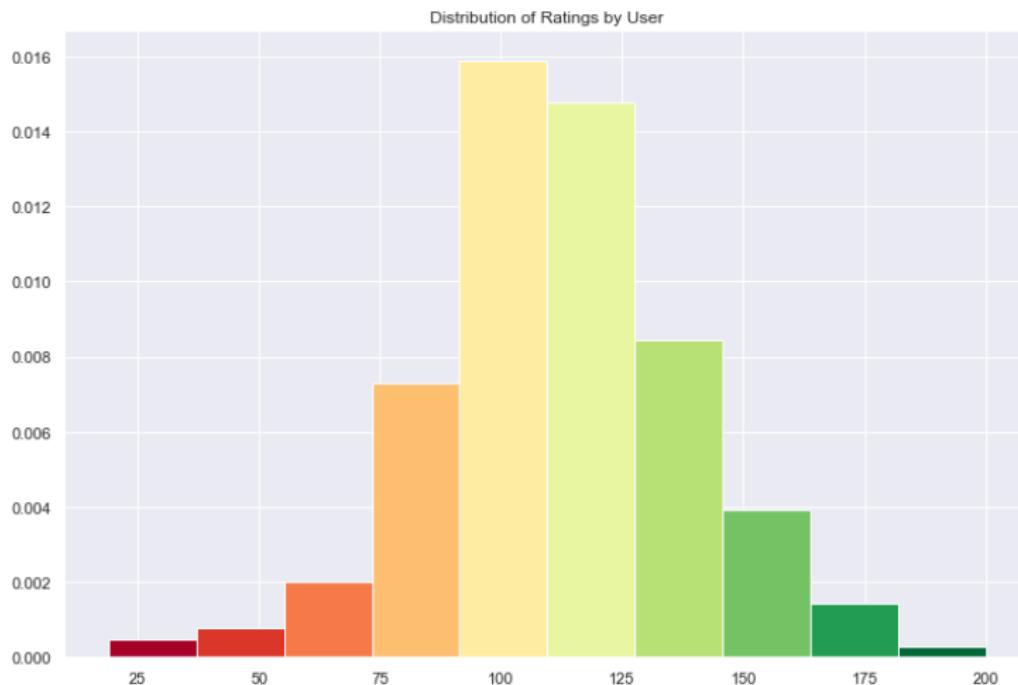


Figure: The range of reviews by user is 19–200.

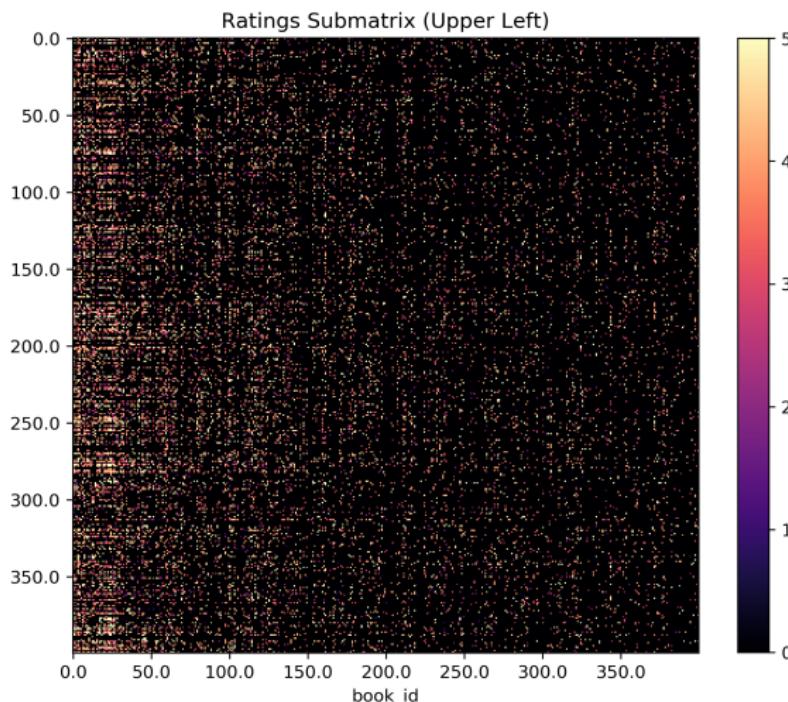


Figure: The first 400 rows and columns of the ratings matrix.

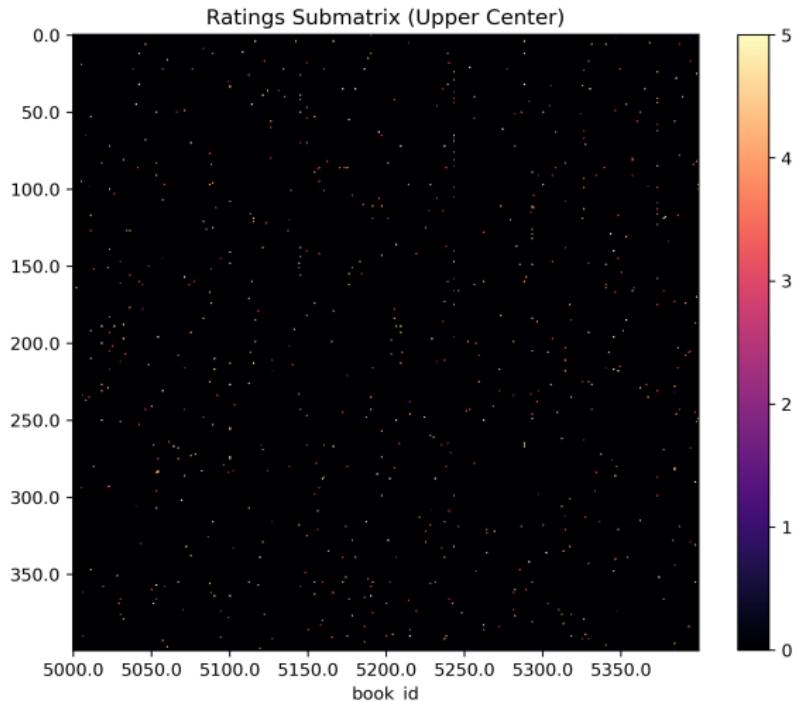


Figure: The upper center of the ratings matrix. `book_id` is ordered by overall ratings. The density of the matrix is 1.12%.

Goodreads Topics and Recommendations

Models

Baseline

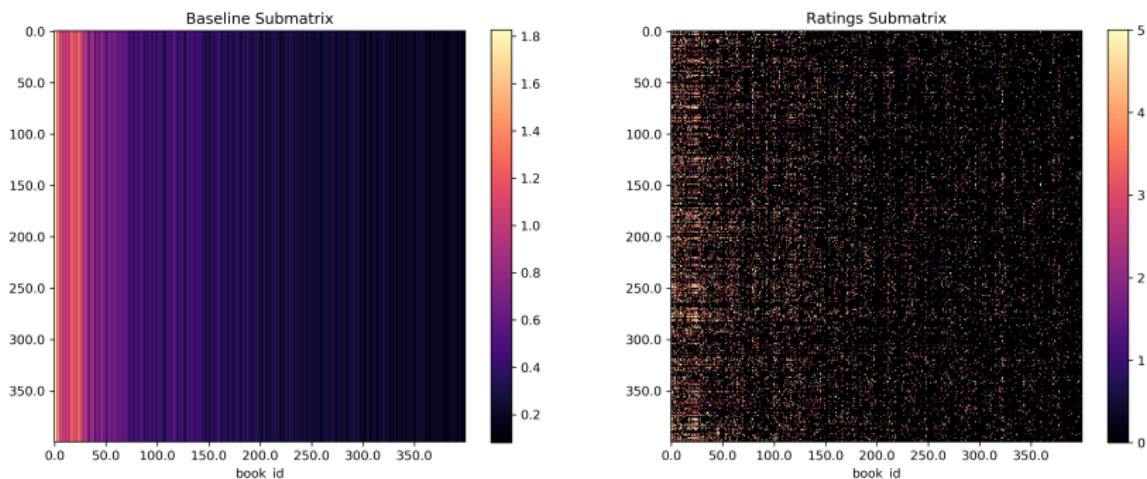


Figure: The mean book rating along each column including no ratings.

Goodreads Topics and Recommendations

Models

$k = 1$

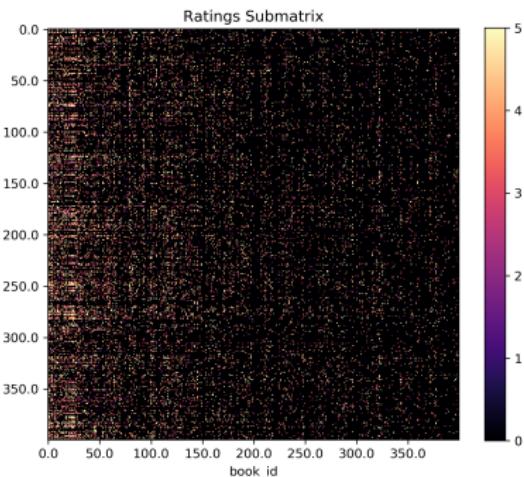
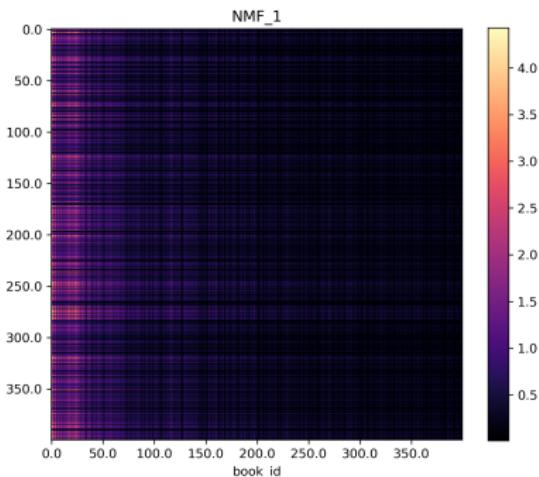


Figure: The matrix reconstruction ($k = 1$).

Goodreads Topics and Recommendations

Models

$k = 10$

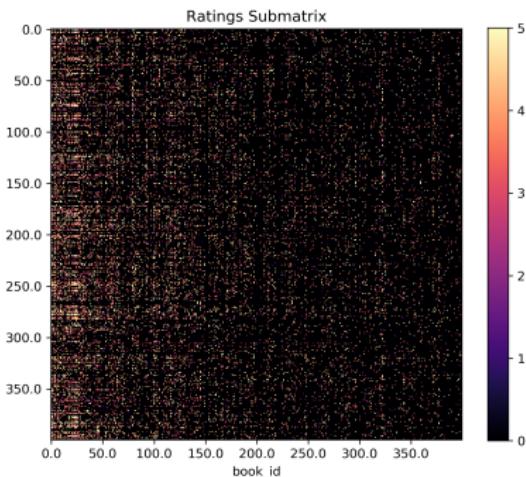
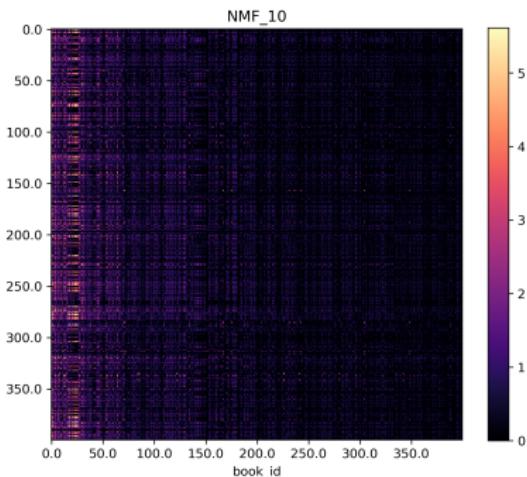


Figure: The matrix reconstruction ($k = 10$).

Goodreads Topics and Recommendations

Models

$k = 25$

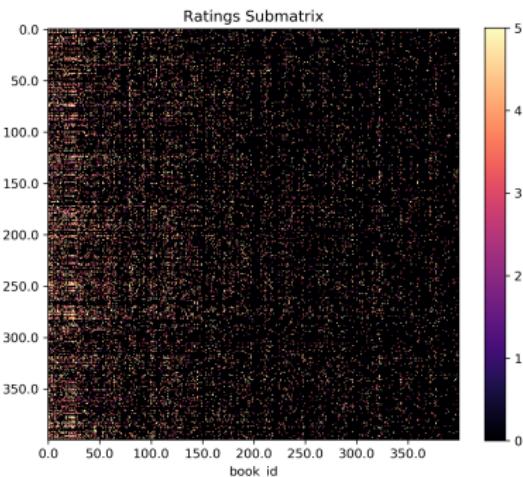
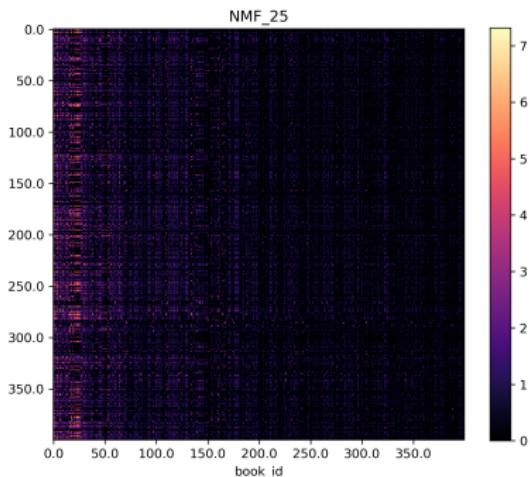


Figure: The matrix reconstruction ($k = 25$).

Goodreads Topics and Recommendations

Models

$k = 10$ Close

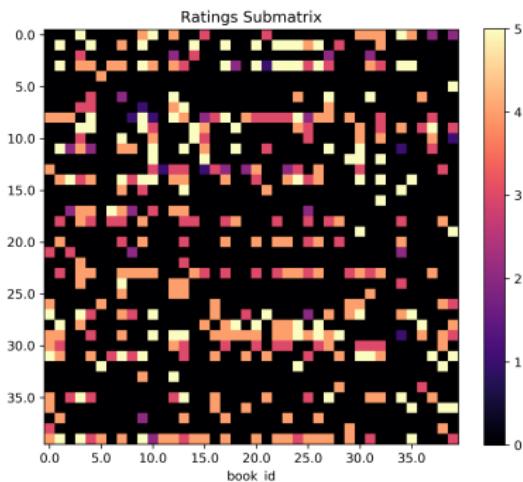
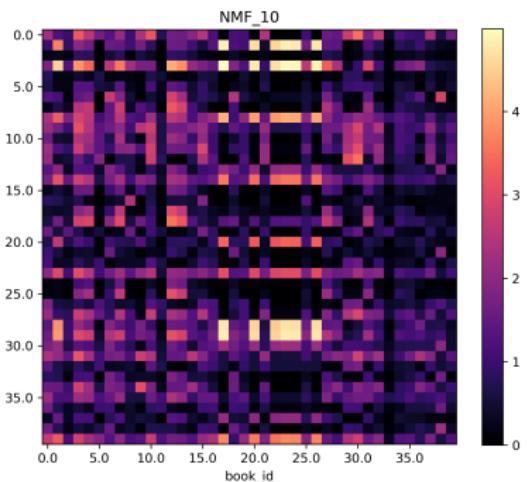


Figure: The matrix reconstruction ($k = 10$).

Goodreads Topics and Recommendations

Models

$k = 50$ Close

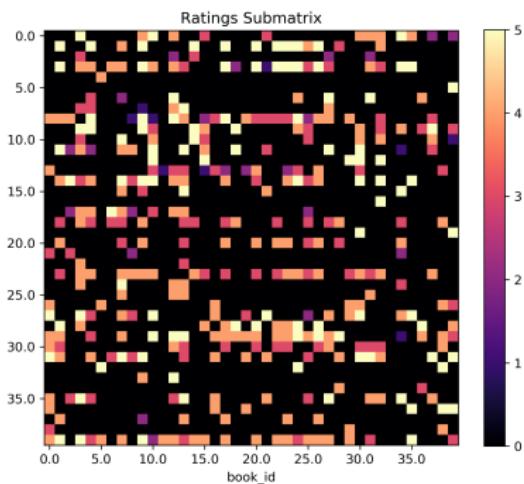
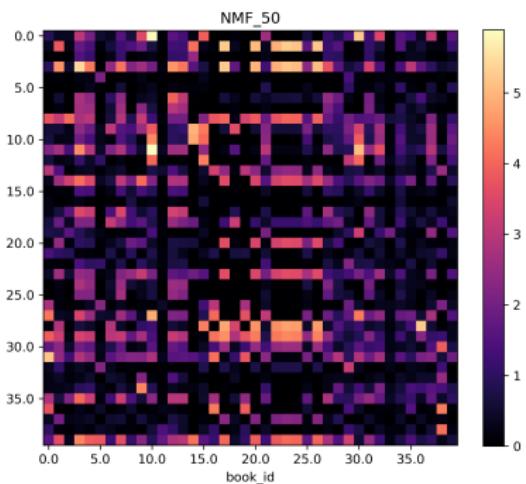


Figure: The matrix reconstruction ($k = 50$).

Goodreads Topics and Recommendations

Models

$k = 250$ Close

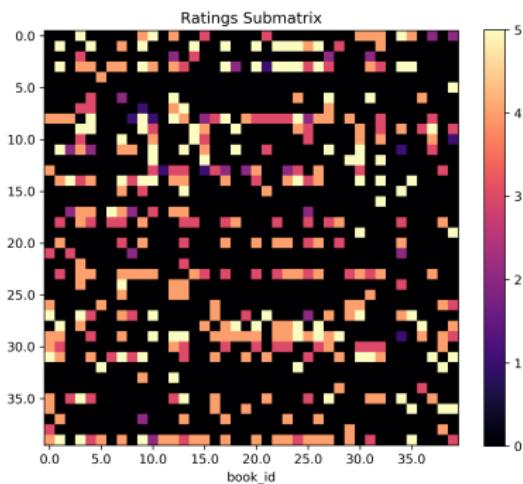
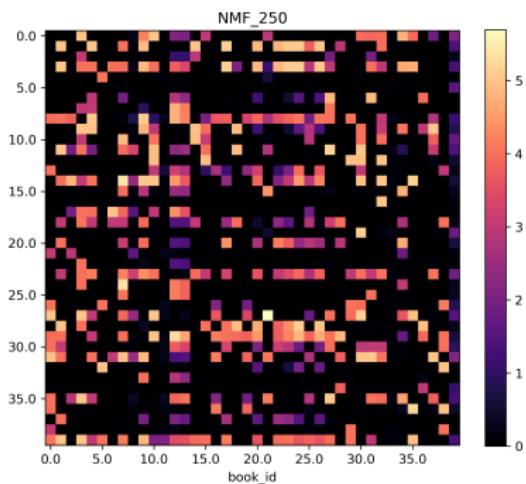


Figure: The matrix reconstruction ($k = 250$).

Goodreads Topics and Recommendations

Models

$k = 10$ Left Center Close

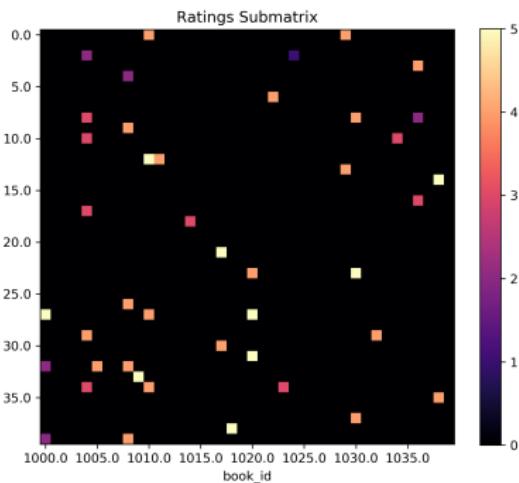
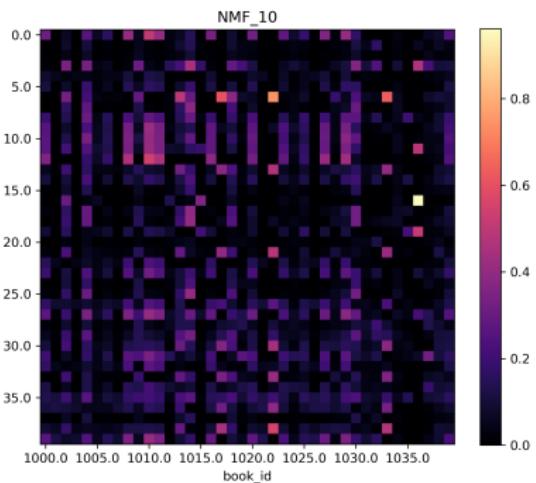


Figure: The matrix reconstruction ($k = 10$).

Goodreads Topics and Recommendations

Models

$k = 50$ Left Center Close

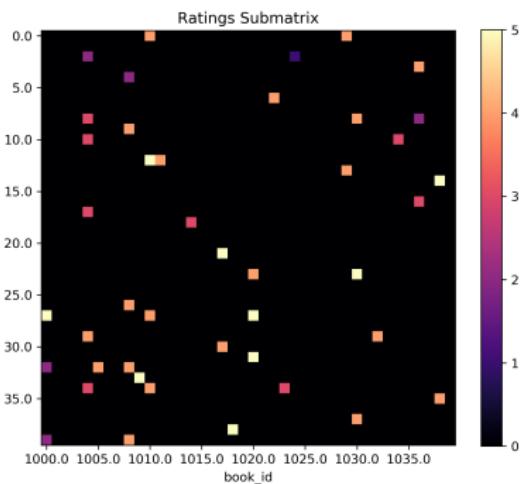
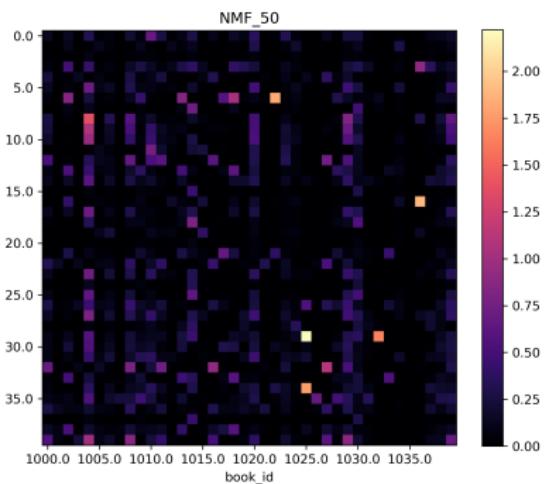


Figure: The matrix reconstruction ($k = 50$).

Goodreads Topics and Recommendations

Models

$k = 250$ Left Center Close

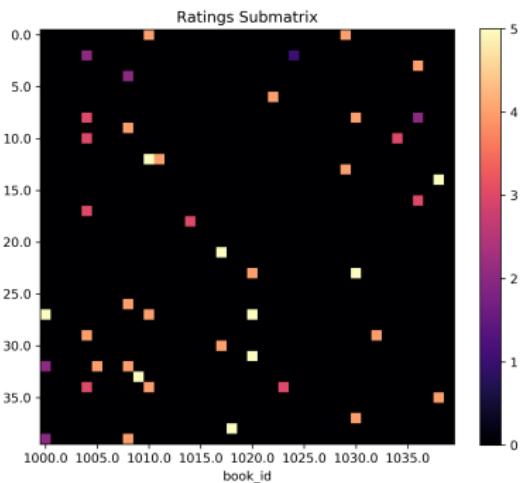
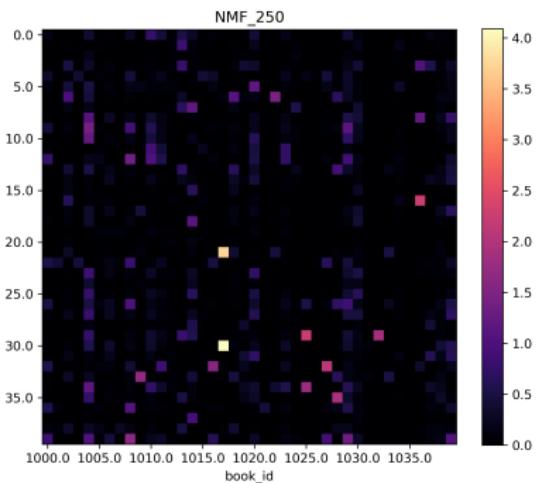


Figure: The matrix reconstruction ($k = 250$).

Models

Scores

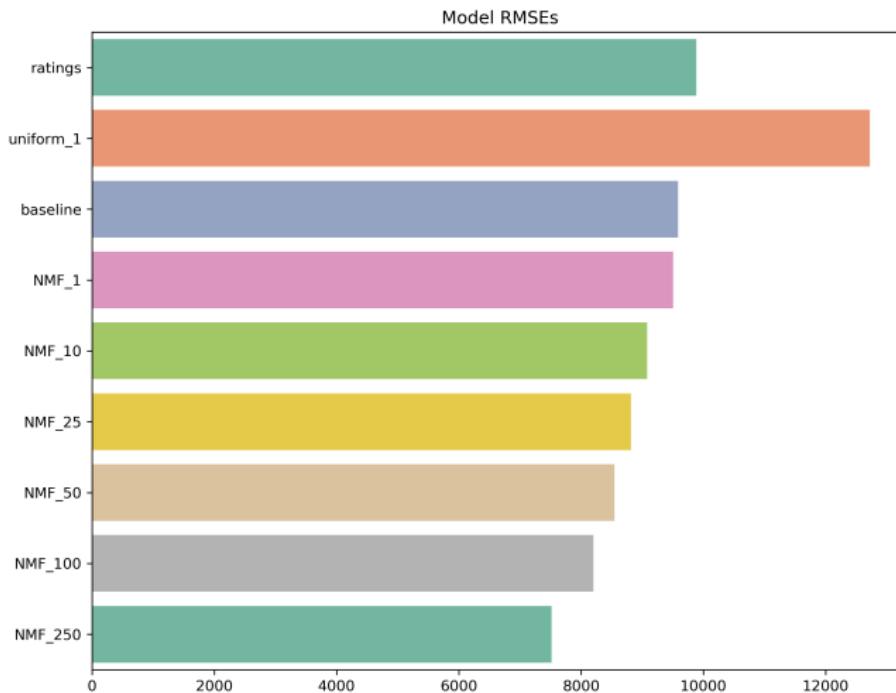


Figure: RMSE scores for each model.

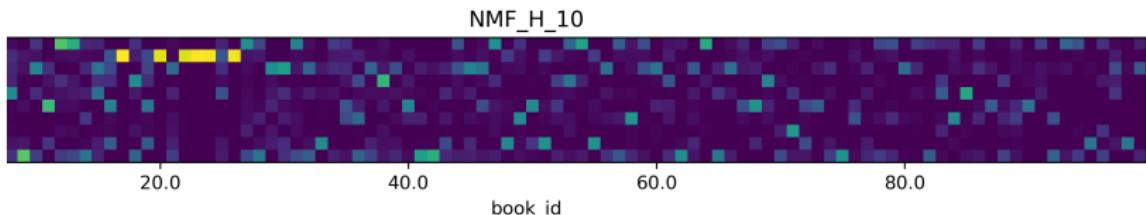


Figure: The book preferred matrix H ($k = 10$).

- Rows h^ℓ of H indicate book correlation with latent topic ℓ .
 - Components of h^ℓ are weights for tags:
 - We have the count $n_{b,t}$ of the top 100 tags for each book
 - The tag frequency r_t^ℓ for tag t in latent factor ℓ is $r_t^\ell = \sum_b h_b^\ell n_{b,t}$.



Figure: Tag cloud for topic 7 ($k = 10$).

authors	title
Stephen King	It
Stephen King, ...	The Stand
Stephen King	The Shining (The Shining #1)
Stephen King	Misery
Stephen King	Carrie
Stephen King	Pet Sematary
Stephen King	'Salem's Lot
Stephen King	Needful Things
Stephen King	The Gunslinger (The Dark Tower, #1)
Stephen King	The Green Mile
Stephen King	The Dead Zone
Stephen King	Firestarter

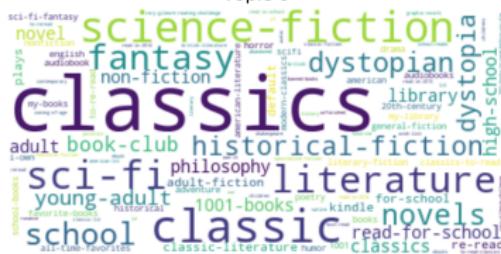
Table: The top-scoring books in topic 7 ($k = 10$) are all Stephen King.

Goodreads Topics and Recommendations

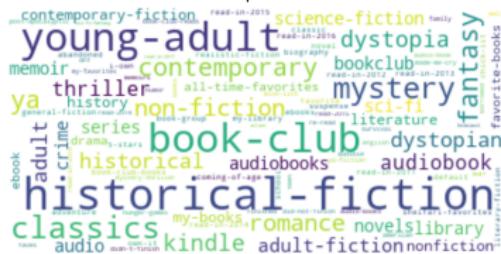
Applications

Topics ($k = 10$)

Topic 0



Topic 2



Topic 2



Topic 3



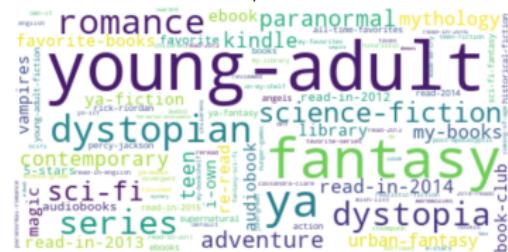
Applications

Topics ($k = 10$)

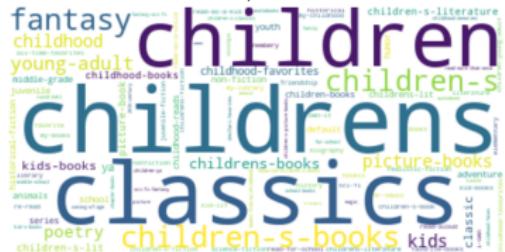
Topic 4



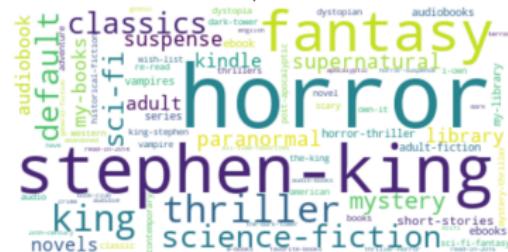
Topic 5



Topic 6



Topic 7



Applications

Topics ($k = 10$)

Topic 8



Topic 5

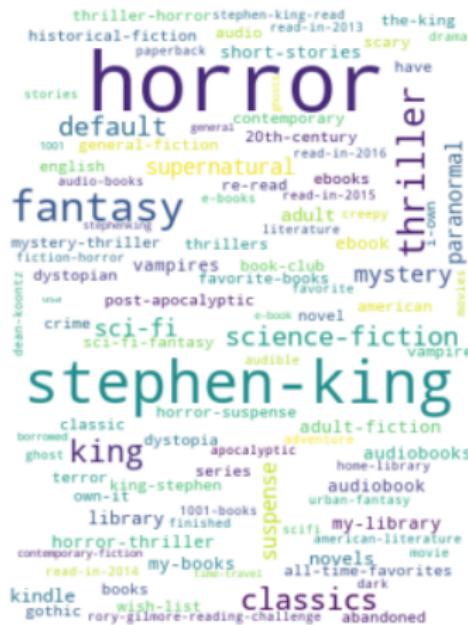


Applications

Topics ($k = 25$)

The topics are finer for greater latent factors k . When $k = 25$ there are two distinct Stephen King topics:

Topic 4



Topic 7



4	7
title	title
The Shining...	The Drawing of the Three (The Dark...)
It	The Waste Lands (The Dark Tower, #3)
Misery	Wizard and Glass (The Dark Tower, #4)
The Stand	Wolves of the Calla (The Dark Tower, #5)
Carrie	The Dark Tower (The Dark Tower, #7)
Pet Sematary	The Gunslinger (The Dark Tower, #1)
'Salem's Lot	Song of Susannah (The Dark Tower, #6)
Needful Things	The Wind Through the Keyhole (The...)
Cujo	The Eyes of the Dragon
Firestarter	The Talisman (The Talisman, #1)
The Dead Zone	Hearts in Atlantis
Christine	Different Seasons

Table: The top books in the $k = 25$ Stephen King topics.

- Rows w_u of W are reader preferences for latent factors
- Interpret latent factors as book topics
- Profile is reader 'preference' for each topic
- Targeted marketing applications

topic_name	9
Harry Potter	0.29
Modern Classics	0.22
Fiction	0.22
Twilight & Fifty Shades	0.12
Austen & Bronts	0.06
Thrillers	0.02
Stephen King	0.00
Children's	0.00
Young Adult	0.00
Fantasy & Sci-Fi	0.00

Table: w_9 describes user 9's preferences for book topics.

- Use a complex model ($k = 250$).
- To recommend for user u , obtain the u^{th} row of the “matrix completion” A via

$$a_u = w_u \cdot H$$

- Choose the top (unrated) values of a_u

authors	title	model	rating
Truman Capote	In Cold Blood	5.21	5.00
Jane Austen	Pride and Prejudice	5.08	5.00
David Sedaris	Me Talk Pretty One Day	5.05	5.00
F. Scott Fitzgerald	The Great Gatsby	5.05	5.00
Mark Haddon	The Curious Incident of the Dog in the Night-Time	5.02	5.00
George Orwell, . . .	1984	4.79	5.00
Sylvia Plath	The Bell Jar	4.54	4.00
Stephenie Meyer	Eclipse (Twilight, #3)	4.41	4.00
Jeffrey Eugenides	Middlesex	4.32	4.00
Stephenie Meyer	Breaking Dawn (Twilight, #4)	4.27	5.00

Table: User 9's top model scores.

authors	title	model	to-read
David Sedaris	When You Are Engulfed in Flames	2.72	NaN
Gabriel Garc...	Love in the Time of Cholera	2.30	True
Jane Austen, ...	Persuasion	2.21	True
Dave Eggers	A Heartbreaking Work of Staggering Genius	2.02	NaN
Michael Chabon	The Amazing Adventures of Kavalier & Clay	2.01	NaN
Jonathan Safran Foer	Extremely Loud and Incredibly Close	1.94	NaN
Sue Monk Kidd	The Secret Life of Bees	1.60	NaN
Kazuo Ishiguro	Never Let Me Go	1.55	NaN
Jeffrey Eugenides	The Virgin Suicides	1.49	NaN

Table: User 9's top recommendations.