Recommender System

Mihaela Grigore
Yuliia Nikolaenko
Lara-Anna Wagner

Dataset - BookCrossing

Book rating datasets (3 tables: books, users, raitings) to build a recommendation engine

users.head()

	User-ID int64	Location object	Age
0	1	nyc, new york, usa	nan
1	2	stockton, california, usa	18

ratings.head()

	User-ID int64	ISBN object	Book-Rating
0	276725	034545104X	0
1	276726	0155061224	5

books.head()

	ISBN object	Book-Title object	Book-Author object	Year-Of-Publication object	Publisher object
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada

Merged data sets

- Number of books: 270151
- Number of users: 92106
- Average book rating: 7.63

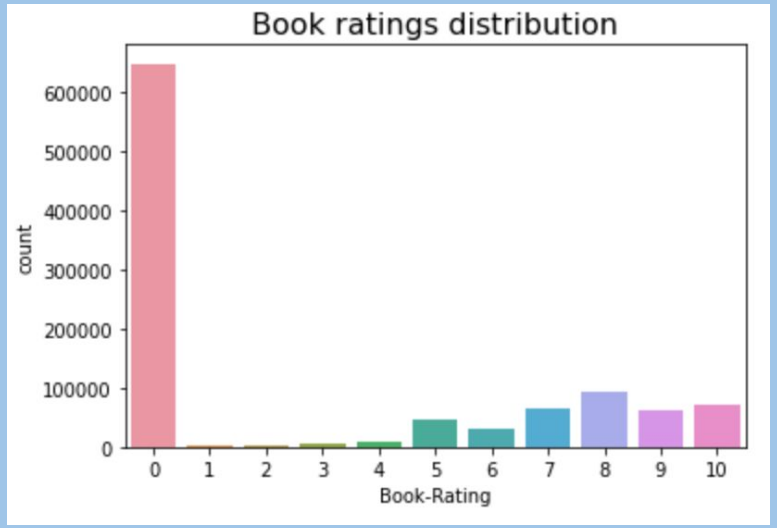
Variables

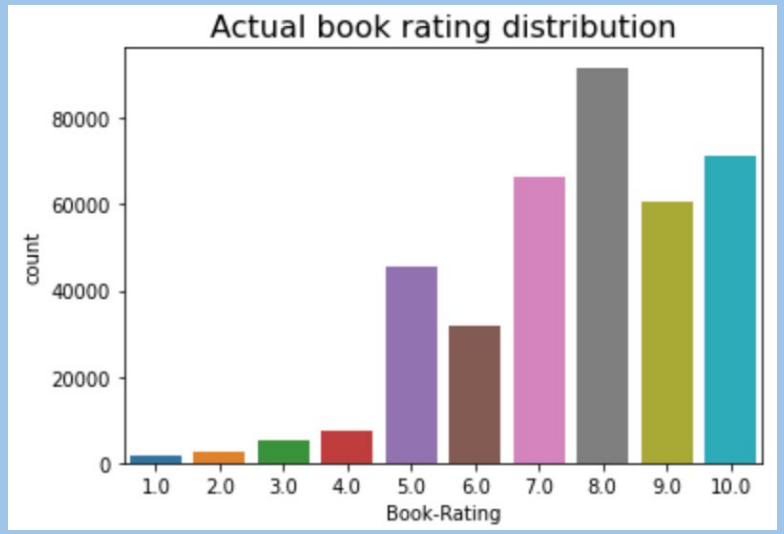
- ISBN
- Book-Title
- Book-Author
- Year-Of-Publication
- Publisher
- Image-URL-S
- Image-URL-M
- Image-URL-L
- User-ID
- Location
- Age
- Book-Rating

Missing or null values

- No mentioning of what 0 value means for rating

- Judging from the distribution => most likely 0 means missing value => drop these entries





Features engineering

1. Dropping Books with no review
2. Dropping images features
3. Imputing missing values (Age 30% of missing values)
4. Extract feature country from the user location
5. Extract sample of size n=8000

Dataset size

Ratings | Users | Books (inner join)	=> 1031136 entries
Drop null reviews	=> 383842 entries (very large for our resources)
Most books have very few reviews & most users gave very few reviews, so we drop those with frequencies < 50	=> 13137 entries (1079 users, 507 books, 13137 non-null reviews)

Recommendation tasks

Content Based

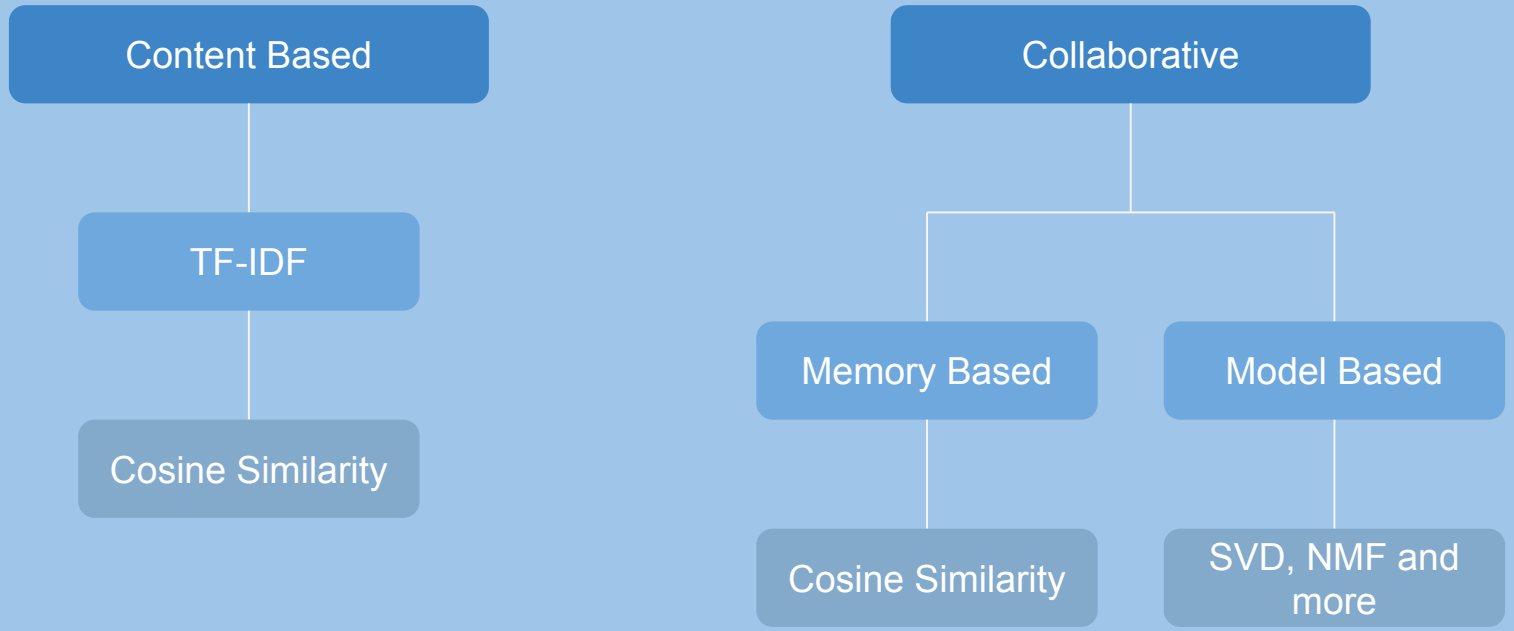
To build a recommender that suggests books based on similar book titles

Collaborative: Memory Based

To build the recommender that suggest the books and other users based on the book grading

Collaborative Filtering: Model based

Recommender System



Test User

150968

Age: 47

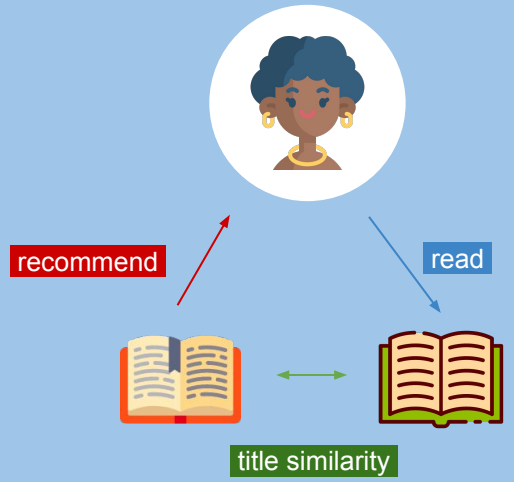
Location: Manchester, New Hampshire, USA

26544

Age: 37

Location: Woodbridge, Virgina, USA

Content Based

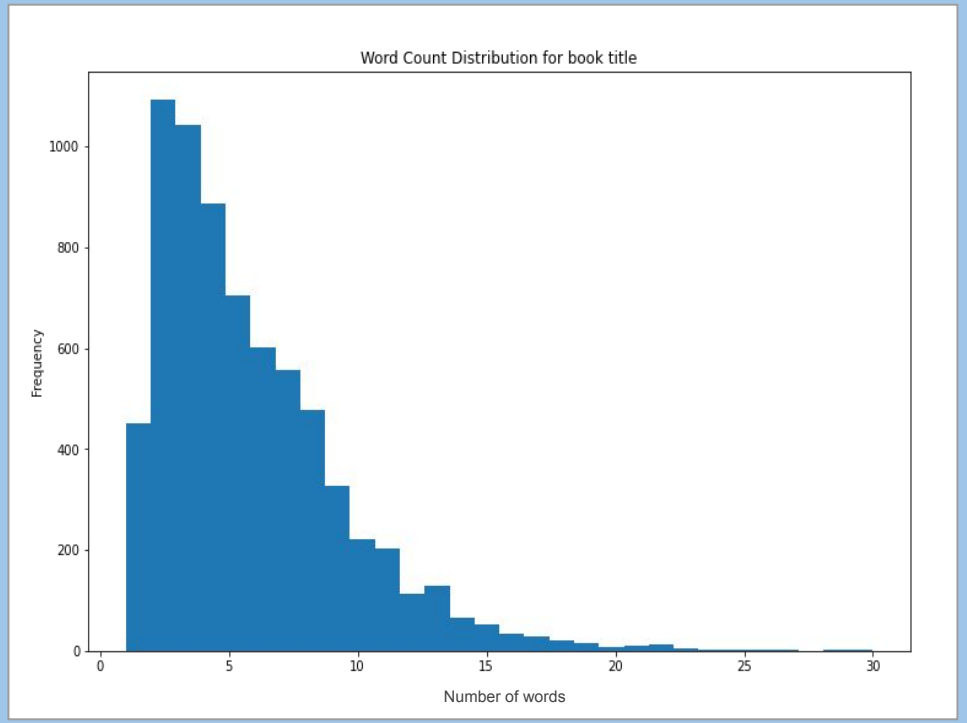


If a user has finished reading a book, which further readings can we recommend to this user based on the **book title similarity**?

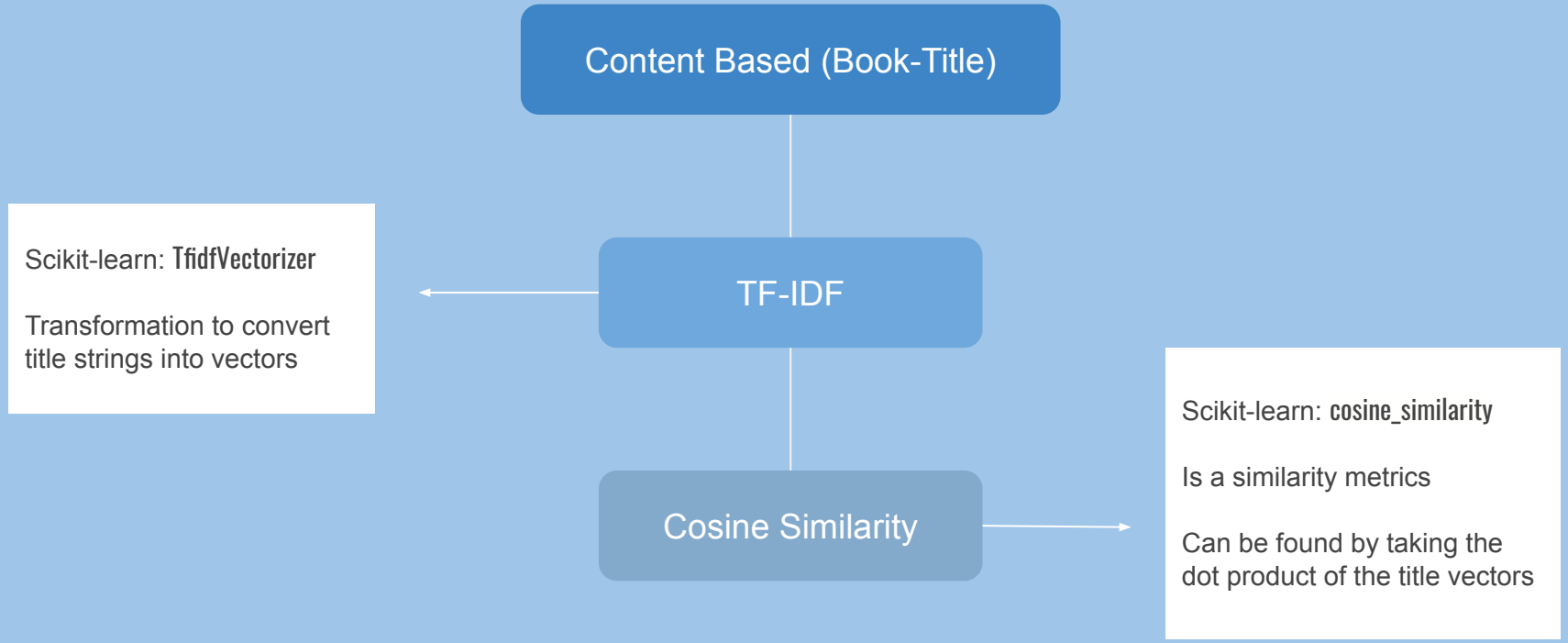
Content Based - *Feature Selection*

Book Features

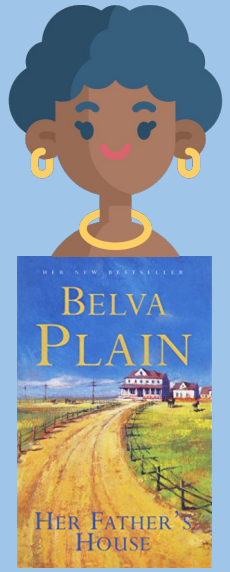
ISBN
Book-Title
Book-Author
Year-Of-Publication
Publisher



Content Based

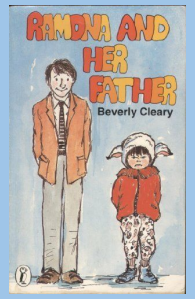


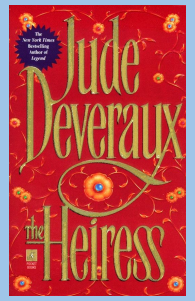
Content Based - *Test Case I*

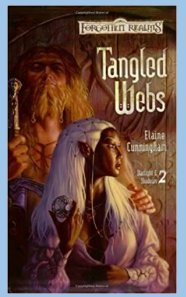


Initial Title:
Her father house

```
----------------------
Recommendation 1:
Ramona and Her Father (Avon Camelot Books (Paperback)) by Beverly Cleary with 0.267 similarity score
----------------------
Recommendation 2:
The Heiress by Jude Deveraux with 0.0 similarity score
----------------------
Recommendation 3:
Tangled Webs (Forgotten Realms: Starlight and Shadows, Book 2) by Elaine Cunningham with 0.0 similarity score
----------------------
```



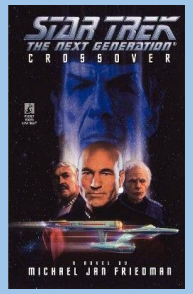


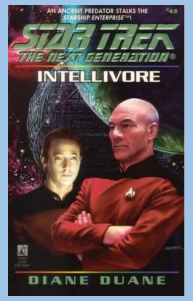


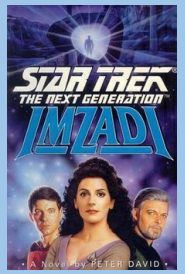
Content Based - *Test Case II*

Initial Title:
Reunion (Star Trek The Next Generation)

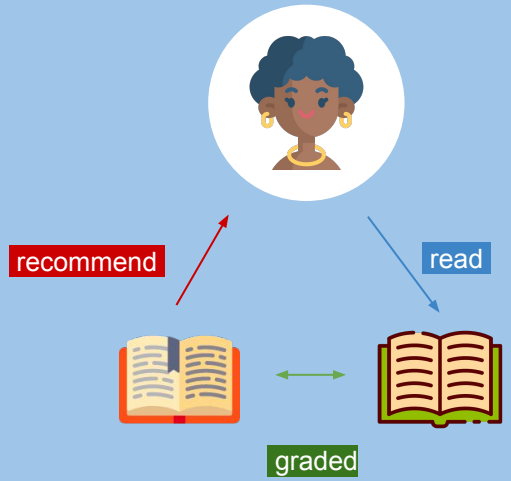
Recommendation 1:
Crossover (Star Trek The Next Generation) by Michael Jan Friedman with 0.713 similarity score
---
Recommendation 2:
Intellivore (Star Trek: The Next Generation) by Diane Duane with 0.713 similarity score
---
Recommendation 3:
Imzadi (Star Trek: The Next Generation) by Peter David with 0.713 similarity score
---







Collaborative: Memory Based



If a user has finished reading a book, which further readings can we recommend to this user based on the **book grading**?

Collaborative: Memory Based

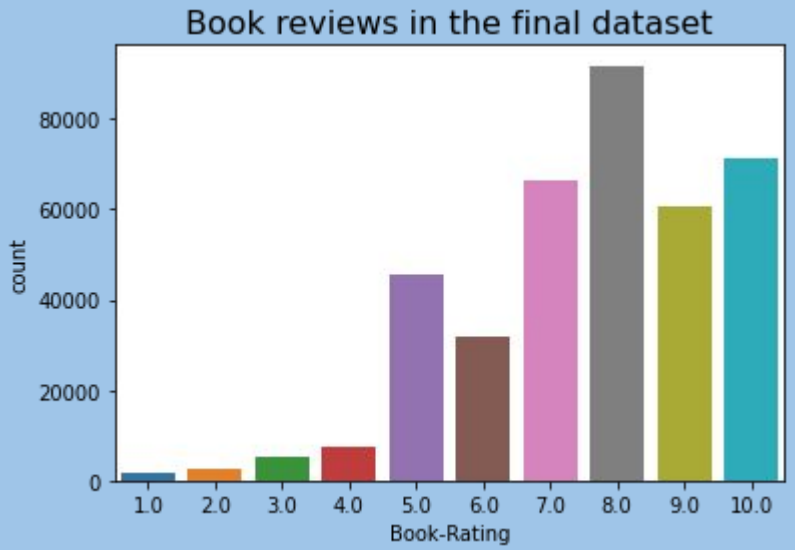
Collaborative

Memory Based

Cosine Similarity

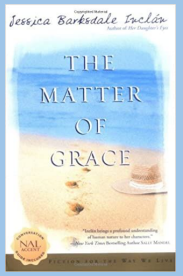
Features

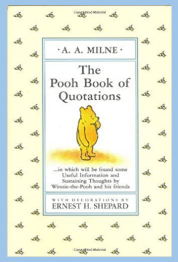
ISBN
Book-Rating
User-ID
Age
Country

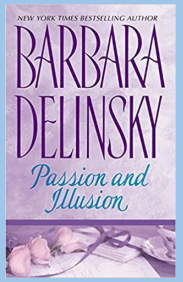


Memory Based - *Test Case I*

Rating based recommender







Memory Based - *Similar users*

183572

3 user
Age: 27
Location: Lafayette, indiana, USA

187003

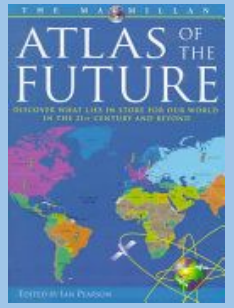
2 user
Age: 51
Location: Manchester, New Hampshire, USA

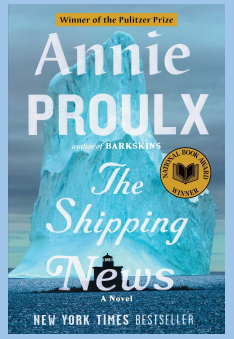
108711

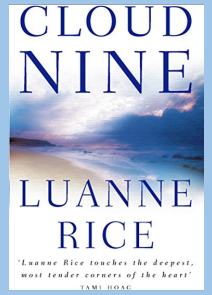
1 user
Age: 39
Location: Burlington, Ontario,Canada

Memory Based - *Test Case II*

Rating based recommender







Memory Based - *Similar users*

168144

3 user
Age: 46
Location: Oceanside, California, USA

54885

2 user
Age: 30
Location: Astoria, New York, USA

123054

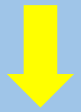
1 user
Age: nan
Location: Houston, Texas, USA

Collaborative Filtering
Model based
...

Model-based CF: what it does

Name	Avengers	Star wars	Thor	Spider-man	Iron Man
Alex	4	2	?	5	4
Bob	5	3	4	?	3
Tom	3	?	4	4	3

Model-based CF



Model-based Collaborative Filtering

Matrix factorization algorithms

Single value decomposition (SVD)
Non-negative Matrix Factorization (NMF)
Slope One
Co Clustering

Deep Learning

Neural Networks

Model-based CF: matrix factorization

The idea: preference of user can be determined by a small number of hidden factors.

We can call these factors as **Embeddings.**

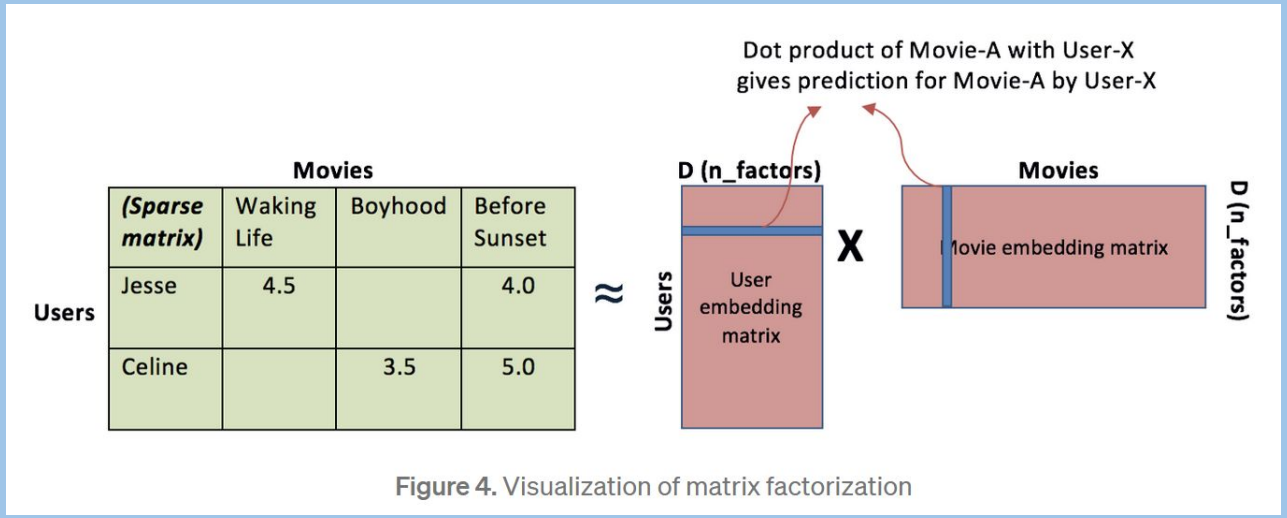


Figure 4. Visualization of matrix factorization

Models comparison

Surprise library -> http://surpriselib.com/

3-fold cross-validation

Reporting RMSE

Algorithm	RMSE (test)	Time (train)	Time (test)
SVD	1.545178	0.650762	0.047195
SVDpp	1.546650	3.752617	0.183756
BaselineOnly	1.559600	0.018354	0.022904
CoClustering	1.656513	0.333316	0.035182
KNNWithMeans	1.682558	0.066995	0.195750
KNNBaseline	1.699507	0.068997	0.240620
KNNWithZScore	1.711343	0.096441	0.222861
KNNBasic	1.803811	0.046665	0.196623
SlopeOne	1.803843	0.037321	0.110750
NormalPredictor	2.338401	0.015554	0.036952
NMF	2.612697	0.798814	0.039230

SVD prediction

```
random_sample[['User-ID', 'ISBN', 'Book-Rating']]
```

	User-ID int64	ISBN object	Book-Rating float64
2215	184339	0061073628	9
2582	214786	0679774386	nan

```
user: 184339     item: 0061073628 r_ui = 9.00   est = 7.91
user: 214786     item: 0679774386 r_ui = nan    est = 7.84
```