The Book Club Project

Book Recommender System

Tais Pancier - December 18, 2018

Introduction

A book club is a group of people who meet to discuss a book or books that they have read and express their opinions, likes, dislikes, etc. In today's hectic life, people don't have much time to meet in person to discuss books.

The purpose of this project is to create a virtual book club with a system that will recommend books to read based on ratings and also other readers to discuss books with.

This model can be implemented by a book store wanting to increase their sales or by a library wanting to increase traffic and community engagement.

Dataset

The dataset obtained in Kaggle consists of 3 main files:

- Book metadata: ten thousand XML files
- User ratings for each book: csv file
- Book tags: csv file

https://www.kaggle.com/zygmunt/goodbooks-10k/home

https://github.com/zygmuntz/goodbooks-10k

The book metadata (author, language, year, etc) dataset contains ten thousand books. The ratings dataset contains reviews for all the book by user id.

Ratings go from one to five.

Both book IDs and user IDs are contiguous. For books, they are 1-10000, for users, 1-53424. All users have made at least two ratings.

There are also books marked to read by the users (wish list) and tags.

Data Wrangling

The XML files were read using Python object xml.etree.ElementTree and saved in 2 datasets: books and authors.

The raw books dataset had 10000 rows and 34 columns. The dataset after cleaning has 10000 rows, 10 columns.

The raw ratings dataset had 5976479 rows and 3 columns. The dataset after cleaning has 705914 rows and 3 columns. Some rows were removed because they contained books that were not in the 10K book dataset.

The following steps were performed to clean the dataset:

- identified columns that wouldn't be used in the analysis and could be dropped (some had null values)
- changed type in numeric columns to integer or float depending on the column content
- renamed columns to differentiate between book id in metadata and wish id (book id in "to read" - wish list)
- removed ratings related to books not in the 10K books dataset
- missing value in language code was replaced by "eng" since the country was GB

All clean datasets were exported to csv to be used in EDA and Machine Learning model.

Exploratory Data Analysis

The following steps were performed as part of the initial analysis:

1) Summary statistics for Books - numeric columns:

- ratings_count mean is much higher than text_reviews_count, indicating that people tend to rate the books more often than providing a text review
- average rating is high in the 10K books set

	average_rating	ratings_count	text_reviews_count
count	10000.000000	1.000000e+04	10000.000000
mean	4.002224	5.404310e+04	2385.520300
std	0.254406	1.574966e+05	5069.490366
min	2.470000	2.718000e+03	3.000000
25%	3.850000	1.357700e+04	542.000000
50%	4.020000	2.116850e+04	1138.500000
75%	4.180000	4.108200e+04	2267.000000
max	4.820000	4.784860e+06	141502.000000

2) Summary statistics for Books - categorical columns:

- only country in the dataset is GB
- “Selected Poems” title has the highest mode
- 25 unique language codes, with “eng" being the most common

	book_id	title	country_code	language_code
count	10000	10000	10000	10000
unique	10000	9964	1	25
top	33288638	Selected Poems	GB	eng
freq	1	4	10000	7432

3) Summary statistics for Authors - numeric columns:

	average_rating	ratings_count	text_reviews_count
count	6273.000000	6.273000e+03	6273.000000
mean	3.986536	4.098591e+05	19463.129284
std	0.212449	1.016630e+06	36111.330007
min	2.840000	7.034000e+03	3.000000
25%	3.850000	4.288500e+04	2890.000000
50%	4.000000	1.304810e+05	8290.000000
75%	4.130000	3.812910e+05	21019.000000
max	4.670000	1.804086e+07	434746.000000

4) Summary statistics for Authors - categorical columns:

- 3888 unique authors

	author_id	name
count	6273	6273
unique	3888	3888
top	15872	Dr. Seuss
freq	7	7

5) Summary statistics for Ratings - numeric column:

- Ratings median is high for the 10K books
- Ratings vary from 1 to 5

	rating
count	705914.000000
mean	3.965466
std	1.007613
min	1.000000
25%	3.000000
50%	4.000000
75%	5.000000
max	5.000000

6) Summary statistics for Ratings - categorical columns:

- 705914 ratings for 812 books and 53400 users
- User id 46521 has the highest number of ratings: 35
- Book id 1 (Harry Potter - of course!) has the highest number of ratings: 22806

	user_id	book_id
count	705914	705914
unique	53400	812
top	46521	1
freq	35	22806

7) Summary statistics for Wish List (books to read) - numeric columns:

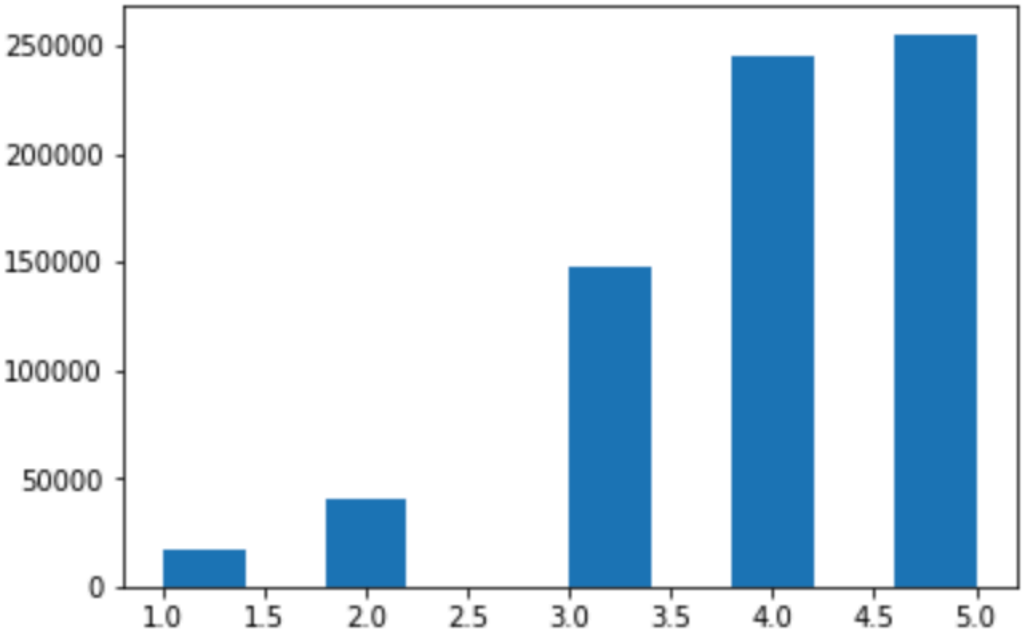
	average_rating	ratings_count	text_reviews_count
count	87493.000000	8.749300e+04	87493.000000
mean	4.028561	2.382774e+05	4522.605980
std	0.306165	5.170892e+05	7601.295942
min	2.800000	5.327000e+03	28.000000
25%	3.820000	1.930100e+04	623.000000
50%	3.970000	4.462600e+04	1603.000000
75%	4.220000	1.658920e+05	4134.000000
max	4.770000	4.607944e+06	58776.000000

8) Summary statistics for Wish List (books to read) - categorical columns:

- user id 12483 has the highest number of books in the wish list: 15
- book id 13 (The Ultimate Hitchhiker's Guide to the Galaxy) shows up 1812 in books to read lists
- author 1077326 (J.K. Rowling) has the highest frequency in wish lists: 7137

	user_id	book_id	authors
count	87493	87493	87493
unique	32806	811	403
top	12483	13	1077326
freq	15	1812	7137

9) Ratings distribution:

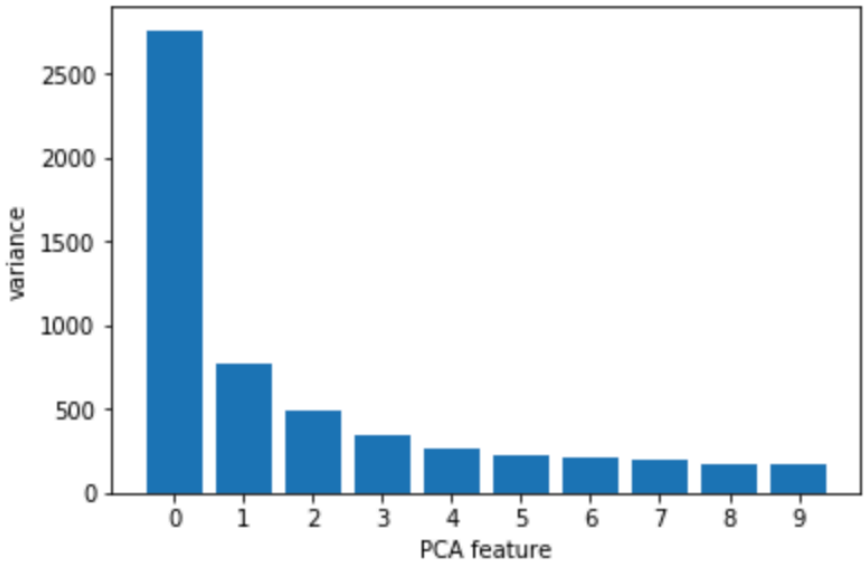


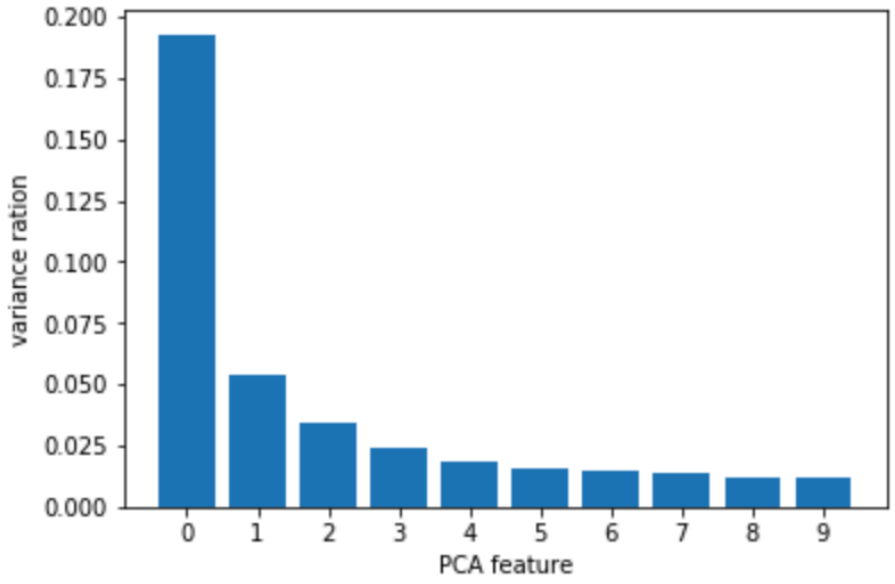
Unsupervised Learning

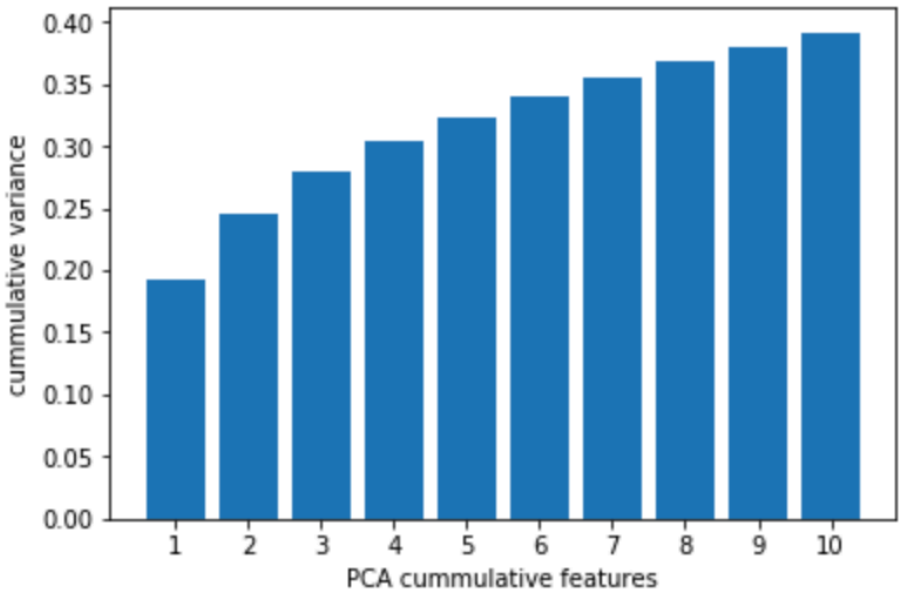
PCA (Principal Component Analysis) was used to investigate the intrinsic dimension.

NCM (Non-negative matrix factorization) was used for the recommendations. The recommender system uses book ratings and the wish list. Based on books in the wish list, it will suggest other books with higher cosine similarity.

1) Intrinsic dimension with PCA:







Based on plots above, intrinsic dimension of 5 features was considered.

2) Recommender System: books

- Objective: suggest books to users based on cosine similarity between the ratings

- Generates matrix (first 5 records) with book ids as rows and user ids as columns

book_id	1	2	3	4	5	6	7	8	9	...	53415	53416	53417	53418	53419	53420	53421	53422	53423	53424
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	...	0.0	0.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0
2	0.0	5.0	0.0	5.0	0.0	0.0	0.0	0.0	4.0	...	0.0	0.0	0.0	0.0	5.0	5.0	5.0	5.0	5.0	5.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	...	0.0	0.0	0.0	0.0	3.0	3.0	0.0	0.0	0.0	4.0
5	0.0	5.0	0.0	4.0	0.0	0.0	3.0	3.0	5.0	...	0.0	0.0	0.0	0.0	3.0	2.0	4.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0

- Algorithm:
  - NMF in sklearn.decomposition
  - Components: 5

- Generates dataframe (first 5 records) with 5 NMF (normalized) features:

book_id	0	1	2	3	4
1	0.000000	0.000000	0.993209	0.000000	0.116345
2	0.493134	0.224524	0.840481	0.000000	0.000000
3	0.018363	0.000000	0.998842	0.044472	0.000000
5	0.000000	0.977175	0.000000	0.108214	0.182810
6	0.020572	0.000000	0.035259	0.000000	0.999166

- Looks in the normalized DataFrame for books in the wish list:
  - User id: 9
  - 3 books in wish list of user 9: 8 (Harry Potter Boxed Set, Books 1-5) , 3476 (Icy Sparks), 112 (Children of Dune (Dune Chronicles #3))

book_id	0	1	2	3	4
8	0.000000	0.988645	0.0	0.147188	0.030282
3476	0.255862	0.953529	0.0	0.109457	0.115479
112	0.000000	0.000000	0.0	0.019758	0.999805

- Drops from normalized dataset books for which user 9 provided ratings

- Computes cosine similarity between books in the wish list and normalized dataset

- Shows 2 recommendations for each book in the wish list, with corresponding book name:
  - wish_id is a book id in the wish list
  - wish_title is a book title in the wish list
  - similar_id is a recommended book id
  - similar_title is a recommended book title

	wish_id	wish_title	similar_id	similar_title	similarity
0	8	Harry Potter Boxed Set, Books 1-5 (Harry Potte...	1420	Hamlet	0.999191
1	8	Harry Potter Boxed Set, Books 1-5 (Harry Potte...	5369	The Amber Room	0.999097
2	3476	Icy Sparks	903	The Egypt Game (Game, #1)	0.998330
3	3476	Icy Sparks	9998	The Woman in the Dunes	0.990695
4	112	Children of Dune (Dune Chronicles #3)	3466	The Wedding (The Notebook, #2)	0.999986
5	112	Children of Dune (Dune Chronicles #3)	2530	Baltasar and Blimunda	0.999980

2) Recommender System: users

- Objective: suggest other readers to discuss about books based on cosine similarity between the ratings

- Generates matrix (first 5 records) with user ids as rows and book ids as columns

user_id	1	2	3	5	6	8	10	11	13	...	9854	9864	9865	9912	9913	9914	9915	9943	9957	9998
1	0.0	0.0	0.0	0.0	0.0	0.0	4.0	5.0	4.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	5.0	0.0	5.0	0.0	4.0	5.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	5.0	0.0	4.0	0.0	4.0	5.0	4.0	4.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- Algorithm:
  - NMF in sklearn.decomposition
  - Components: 5

- Generates dataframe (first 5 records) with 5 NMF (normalized) features:

user_id	0	1	2	3	4
1	0.000000	0.354922	0.059074	0.933028	0.0
2	0.781226	0.594015	0.021383	0.190721	0.0
3	0.000000	1.000000	0.000000	0.000000	0.0
4	0.603214	0.738703	0.107796	0.280767	0.0
5	0.000000	0.000000	0.000000	0.000000	1.0

- Looks in the normalized DataFrame by user and calculate cosine similarity:
  - User id: 9
  - User_ids below are the recommended readers with largest similarity:

```
user_id
9        1.000000
6445     0.999219
52779    0.998622
4200     0.997966
9992     0.997526
50147    0.997012
14205    0.996186
29404    0.995741
5335     0.995724
1979     0.995468
```

- Shows other users together with the highest and lowest book ratings (example below for similar user id = 6445):

    - Highest ratings:

	user_id	book_id	rating	title	country_code	language_code	average_rating	ratings_count	text_reviews_count	authors	author_name	work_id
304673	6445	24	5	In a Sunburned Country	GB	eng	4.05	58749	3667	7	Bill Bryson	2611786
45446	6445	36	5	The Lord of the Rings: Weapons and Warfare	GB	eng	4.53	18788	42	5448409	Chris Smith	4414
304685	6445	119	5	The Lord of the Rings: The Art of The Fellowsh...	GB	en-US	4.59	24343	64	60	Gary Russell	4479
304684	6445	13	5	The Ultimate Hitchhiker's Guide to the Galaxy	GB	eng	4.37	222943	3477	4	Douglas Adams	135328
304682	6445	93	5	Heidi	GB	eng	3.97	144537	1827	49	Johanna Spyri	1738595

    - Lowest ratings:

	user_id	book_id	rating	title	country_code	language_code	average_rating	ratings_count	text_reviews_count	authors	author_name	work_id
69477	6445	1381	2	The Odyssey	GB	eng	3.73	670663	5468	903	Homer	3356006
36616	6445	26	3	The Lost Continent: Travels in Small-Town America	GB	en-US	3.83	40286	1901	7	Bill Bryson	1888943
211795	6445	34	3	The Fellowship of the Ring (The Lord of the Ri...	GB	eng	4.34	1768461	10768	656983	J.R.R. Tolkien	3204327
304668	6445	30	3	J.R.R. Tolkien 4-Book Boxed Set: The Hobbit an...	GB	eng	4.59	90944	1425	656983	J.R.R. Tolkien	89369
77275	6445	291	4	The Broken Wings	GB	eng	3.93	5613	461	6466154	Kahlil Gibran	1676536

Interpretation of data above: User 9 will get the message that based on their ratings, user 9 would enjoy chatting with user 6445. User 6445 highest and lowest rated books are shown to give a starting point for the conversation.

Proposed Next Steps

- Improve the model so it addresses diversity, serendipity, novelty, robustness and scalability in the book recommendations

- Collect feedback on the recommendations to evaluate the mode

- Collect additional data related to the ratings, for example time stamp. With the time stamp it's possible to implicitly determine if a user rated a book after it was recommended to them, which is an indication of a good performance.

- Implement ranking evaluation methods such as Spearman, Kendall or Utility