Book Recommender – Final Project

Title Page

The book recommender analysis was conducted for World's Best Book store by Team 2 Data Scientist as an attempt to determine the most appropriate recommender system for customer's use. The current recommendation system was not accurately aligned with the interests and wants of customers. Upon review of books and user rating datasets a case comparison was made with the Iliad to show the results utilizing four recommendation models. It has been determined that the best of the four recommendation systems is the content-based model. It is the recommendation of Team 2 Data Scientist for World's Best Bookstore to utilize a hybrid recommender system of the content-based system in combination with one or more of the models reviewed.

Specification

Problem

There is an existence of many recommendation tools to assist readers with receiving recommendation of the next book they should invest their time and money in reading. "A recommender system has been defined as software tools and techniques providing suggestions for items to be of use to a user (Ricci et al. 2011). The idea of personalized and intelligent agents, search engines and recommender systems has been widely accepted as solutions towards overcoming information retrieval challenges arising from information overload" (Montaner et al. 2003). Some of those tools are based upon algorithms that look at billions of data points in collaboration with a social barometer of subscribed readers rating their favorite reading materials. Another tool utilizes a reader's purchase history to recommend the next project. All those platforms rely on the fundamental question of asking readers what types of books they enjoy reading. The World's Best Bookstore Incorporated (WBB) have received feedback from their readers that the recommendation system they use does not match their preferences. When the recommendations do not match the likelihood of a customer purchasing from their store or returning to their store diminishes. According to "one study (Chen 2008) shows that consumers were more interested in books labeled "customers who bought this book also bought" than books marked "recommended by the bookstore staff" (Alharthi, 2017)

WBB has hired Team 2 Data Scientist Incorporated to find the best tool that is able to personalize book recommendations to customers based upon their

interest, improving user experience. WBB would like to become better aligned with the customer's interest. There are six main categories or types of book recommender tools using algorithms. Those types are collaborative filtering (CF), content-based recommenders (CB), demographic-based recommenders, social RSs, context-aware RSs, and RSs using association rules. Each recommendation tool has advantages and disadvantages based upon the needs of the user.

Hypotheses

If WBB provides book readers with a recommender based upon reader ratings, then readers will have more accurate book recommendations and a better user experience.

Data for analysis

The data used for this analysis includes two main datasets. Those datasets contain the original books csv with the ratings csv from Kaggle.com.

The dataset is a file containing the following fields:

Feature	Description
book_id	Identifier in dataset
title	Title of book
authors	Book authors
average_rating	Mean of book ratings
language_code	Language book was written in
num_pages	Number of pages in book
ratings_count	Count of user ratings
text_reviews_count	Count of user reviews
publication_date	Date book was published
publisher	Book publisher name
publication_month	Month, book was published
publication_day	Day of the week, book was published
publication_year	Year, book was published

Observation and Visualization

Books Dataset Observation

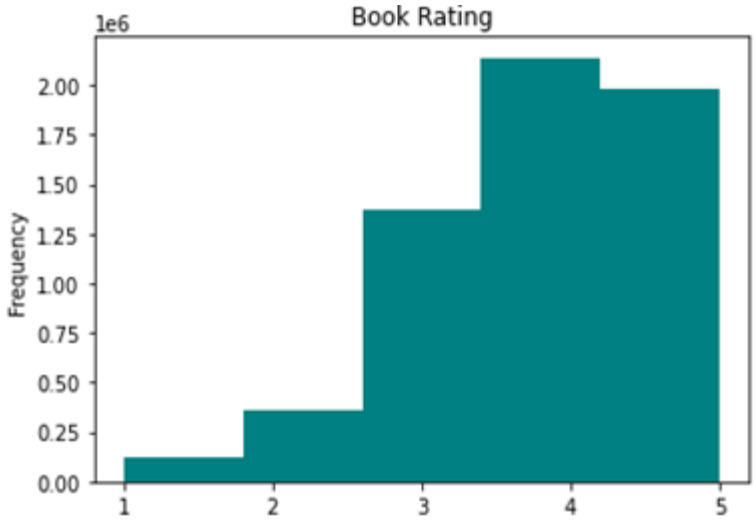
- Total number of books: 10,000
- Total number of ratings: 176, 953, 301
- Book with most ratings: Twilight - 4,597,666 ratings
- Book with highest average rating: The Complete Calvin and Hobbes - 4.82
- 50% of books have less than 750 ratings while 25% have greater than 5000 ratings, we subset the data to only include ratings greater than 750 to avoid skewed data

Ratings Dataset Information Observation

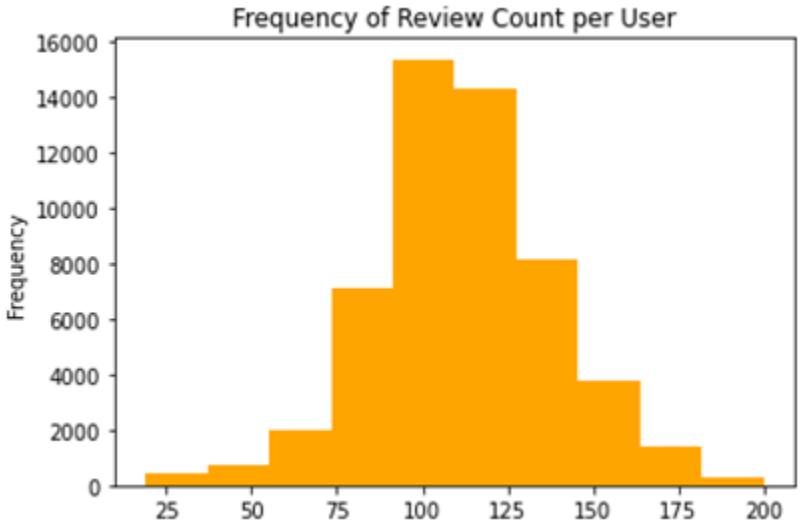
- Total number of ratings: 1, 048, 575
- Percent of total ratings included in this dataset: 0.59%
- Mean: 3.81
- Median: 4.0
- Total number of users in dataset: 13, 123

General Observation

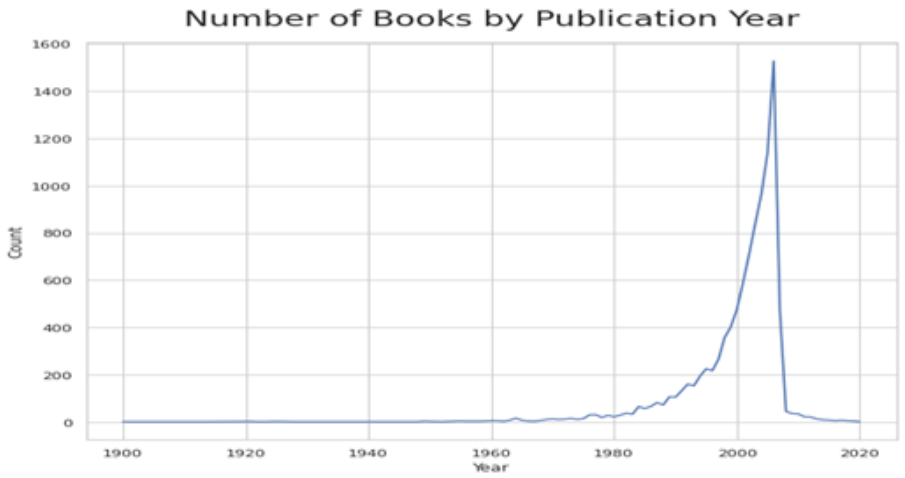
- From 176, 953, 301 ratings in books dataset, only 1, 048, 575 are included here which is about 0.59%
- There are many users who have given 100 or more reviews, but no have given more than 200 reviews, not sure if there is a limit of reviews allowed per user.
- Many readers tend to give high ratings (mean rating = 3.81 & median rating is 4.0). see histogram below. We don't know the reason why but may be its because readers don't want to give negative reviews or read enough reviews before buying and so they end up liking it…and so on.



Also, it's interesting to see no user reviewed book more than 200 times (see histogram below), maybe there is a limit for reviews?

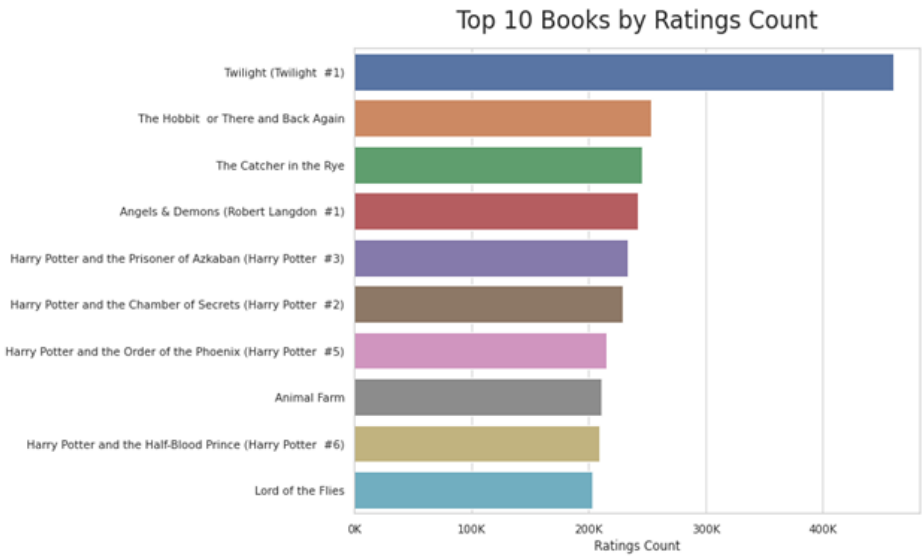


It's not surprising to see most books were published after 20th century. If we start from year 1900 up, see graph below. We had data from the 1500BC, but we removed it to avoid any skewed data



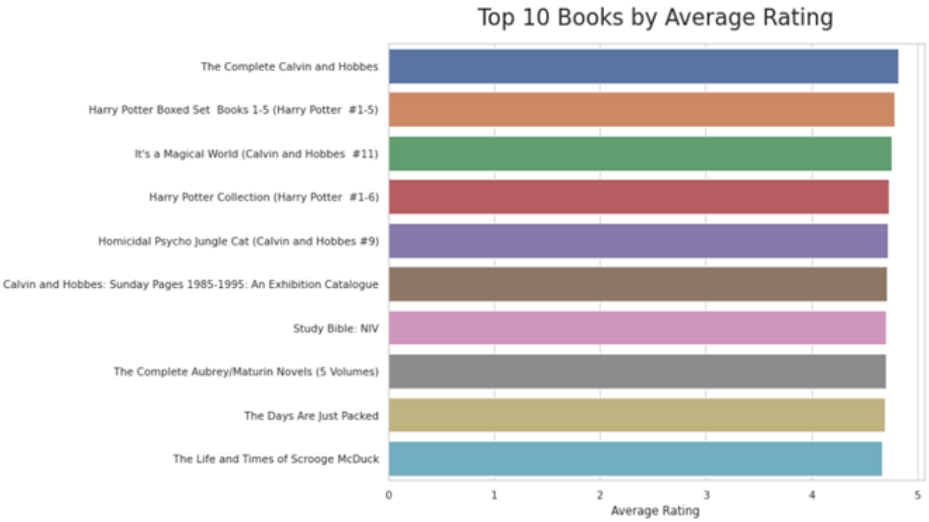
50% of the books have less than 750 ratings, while 25% of books have more than 5000 ratings. We subset the dataset to include books only with 750+ ratings to avoid highly

Top 10 Books by Rating Count



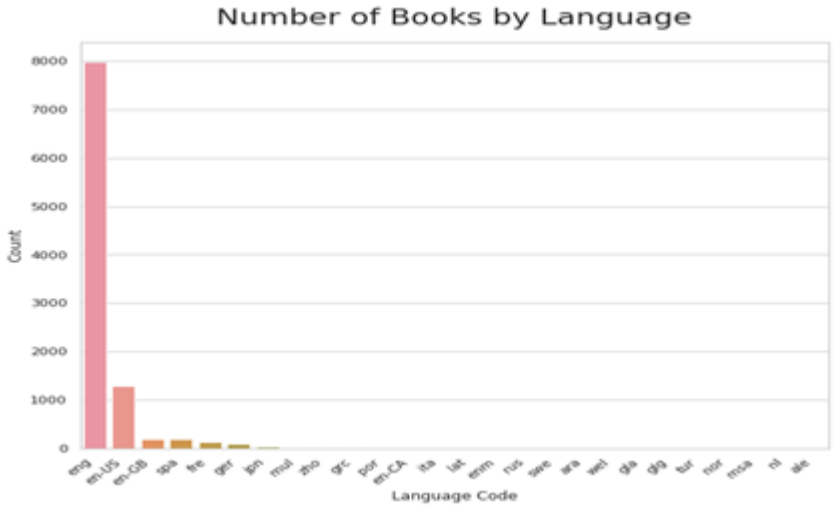
Based on our data and the bar chart above, Twilight has the most ratings with over 4.5 million ratings and the complete Calvin and Hobbes have the most average ratings,

Top 10 Books by Average Rating



Number of books by language

- No wonder over 99% of the books are written in English, we have books written in other languages, but the number is insignificant.



There was a book the has only one review, 25% of books have less than 340 reviews Our model won't be accurate if we include data with low review counts. To make sure we get results that make sense we limited our dataset to only include books with more than 340 reviews.

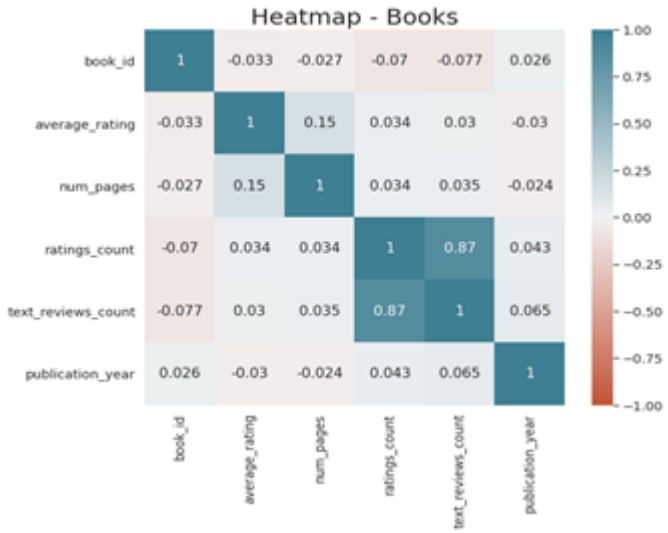
We also looked at authors and how many books they have in this data set and the result is as follows

```
#AuthorsAndBookCounts
books_df.authors.value_counts()
```

```
Stephen King                                        38
Rumiko Takahashi                                    37
P.G. Wodehouse                                      34
Orson Scott Card                                    31
Agatha Christie                                     30
                                                    ..
Anne Easter Smith                                    1
Agatha Christie/Robert Welch Herrick                 1
Nick Flynn/Shirley McPhillips/Philippa Stratton      1
James H. Cone                                        1
Sylvia Plath/Frieda Hughes                           1
Name: authors, Length: 6108, dtype: int64
```

Correlation Matrix

Based upon our questioning of correlations between the books data set attributes, we created a correlation heat map. The heatmap below showed a lack of correlation. Ratings count and text review count has a 0.87 correlation which makes sense, other than that the attributes are not really correlated.



Conclusion from Observation and visualization

- People tend to give high ratings
- Even though people tend to give high ratings, many people don't give a full rating (5Star ratings)

- Most books in this data set are written in English, even though we have books in other languages, their number is insignificant compared to the number of books in English
- Most books are published in the 20th century, we have books starting from 1500BC but again their number is insignificant
- The obvious popular books such as Twilight remains to have the highest rating count
- Surprisingly, many individuals tend to give over 100 reviews in this data set, but no one was giving over 200 which we thought there might be a limit as to how many reviews an individual can give

Analysis and Models

Models and Techniques

I. *Correlation*

A correlation model determines how one variable is linearly related to another. It is a simple and effective way to understand dependency of two variables and therefore is an easy to implement recommendation system. In this model we used Pearson's Correlation Coefficient to suggest book recommendations based on similarities between books. A correlation is represented by numerical values between -1 and 1 with 0 indicating no correlation.

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson's Equation

For all our models, we used *The Iliad* as a test case book to determine accuracy. Clearly, this is subjective. Since there is no numerical test to measure accuracy, we had to rely on common sense and book descriptions to verify if our model's recommendations made sense.

book_id	Correlation	title
1853	1.0	Wild About Books
966	1.0	Angeles & Demons
2680	1.0	Empire 2.0: A Modest Proposal for a United Sta...
3978	1.0	A Winter Haunting (Seasons of Horror #2)
1371	1.0	The Iliad

As illustrated in the table above, the correlation model had successfully identified another version of the Iliad which makes sense. However, the remaining recommendations don't fully align with what we expect a reader of The Iliad to find interesting.

Pros	***Cons***
Simplest Model	Cold Start
Effective	Cause & Effect

II. *Item-based Collaborative Filtering*

An item-based model looks for similar items a user has previously had a positive interaction with. In the case of our book recommendation system, a positive experience is a previously rated book. Our model uses the K-Nearest Neighbor algorithm from the sklearn python package to determine similarity between books. The kNN model finds clusters of similar books by calculating the distance between vectors of ratings. Cosine correlation and the brute algorithm is used to find the nearest neighbors.

Recommendations for The Iliad – KNN (Item-based)

Title	Distance
Great Jones Street	0.8844422580203157
The Kingdom of God Is Within You	0.8857918369127253
Agile Web Development with Rails: A Pragmatic Guide	0.8860570834204622
The Lord of the Rings (The Lord of the Rings #1-3)	0.8883419447696739
Simply Beautiful Beading: 53 Quick and Easy Projects	0.8933857518508036

Our item-based CF recommender's 5 best recommendations are provided in the above table. Three of the five recommendations seem to be relatively related to our test book, *The Illiad*, however two books are instructional guides and appear to share no contextual or thematic similarities. This is a slight improvement over a pure correlation system but alone does not satisfy a satisfactory personalized recommendation for the customer.

Pros	***Cons***
Dynamic items	Scalability
Accuracy	Sparsity

III. *User-based Collaborative Filtering*

The user-based CF model uses similarity between users' taste in books with the target user. Like the item-based approach, a KNN model is utilized to measure the distance between vectors of books of related users. Unlike the item-based model, this approach measures the distance between vectors of users who have previously given a similar rating for the same books. Because user ratings are highly subjective, a weighted average was used to normalize user ratings to ensure consistency.

Recommendations for The Iliad – KNN (User-based)

Title	Ratings Count
The Lord of the Rings (The Lord of the Rings ...	1618
Agile Web Development with Rails: A Pragmatic ...	1430
God Emperor of Dune (Dune Chronicles #4)	2785

The user-based CF model provided 2 of 3 accurate recommendations. Like the item-based approach, one of the recommendations was an instructional guide that was incorrectly clustered. This inconsistency seems to be a common theme among all models. Although a weighted scale was used to normalize user ratings, there is still a great deal of subjectivity associated with user ratings as users may interpret scales differently and their methodology of what constitutes a good book may slightly differ.

Pros	***Cons***
Easy implementation	Cold-start
Performance	Sparsity

IV. *Content-based model*

Content based models use user provided data or metadata about the user to build a profile about the user and provide recommendations. Similar to collaborative approaches, content-based models measure the distance between vectors to determine the likeness of two items. We decided to keep out model simple to start with by using just the text of each book title as the input for the model. By utilizing natual language processing (NLP), we our model builds vectors for each user/book by extracting high frequency words and removing common stop words.

Title
Harry Potter and the Order of the Phoenix (Har...
Harry Potter and the Chamber of Secrets (Harry...
Harry Potter and the Prisoner of Azkaban (Harr...
Harry Potter Boxed Set Books 1-5 (Harry Potte...)
Unauthorized Harry Potter Book Seven News: "Ha

We can see that all 5 content-based recommendations were for different volumes of Harry Potter. The relationship between our test case and these

recommendations is obvious, all books are fictional fantasy/epic novels. Since we only used keywords extract from titles, we did not expect high accuracy from this model but were pleasantly surprised by the test results. As with all previously tested models, there is plenty of room for improvement.

Pros	***Cons***
No data from other users	Dependent on user tags
Quality over Quantity	Scalability

Recommendation

A review of the analysis we've completed as Team 2 Data Scientist Inc. We have made recommendations that will enhance the current recommendation systems in place at WBB. Based upon our analysis of the four recommendation systems above we have concluded that a content-based model provides the most personalized recommendation to customers out of the four models tested. The inclusion of more features to any of the recommendation systems evaluated such as genres, book summaries, author names, time periods, and story settings would enhance the content model. It is our recommendation to utilize a hybrid model, which would be a combination of 2 or more models to provide optimal recommendations to customers. With an increased review, a simulation should be conducted using the content-based model as the baseline while adding on the combination of other models in a test phase.

References

Alharthi, H., Inkpen, D., & Szpakowicz, S. (2017). A survey of book recommender systems. *Journal of Intelligent Information Systems*, *51*(1), 139–160. https://doi.org/10.1007/s10844-017-0489-9

Tarus, J. K., Niu, Z., & Mustafa, G. (2017). Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. *Artificial Intelligence Review*, *50*(1), 21–48. https://doi.org/10.1007/s10462-017-9539-5