Book Recommendation and Intelligence Engine [B.R.I.E]

Sethuraman Annnamalai (sannamal@sfu.ca)
Lakshayy Dua (ldua@sfu.ca)
Supreet Kaur Takkar (stakkar@sfu.ca)

I. Motivation and Background

"What book should I read next?" - The question that plagues any avid reader at one point or another. The majority of the reading community feels that there is a dearth in the availability of a data science product dedicated to cater the needs of not only the readers but also other critical players in the publishing industry such as authors and publishers.

Another subtle issue that many readers face is in determining different types of content in a book. Any given book is predominantly classified into a single genre. But any reader can attest to the fact that a given book contains multiple genre content in different proportions. Predicting these proportions would go a long way in determining if a book would be liked by a particular reader.

Existing book recommendation mechanisms can be complicated and biased. Some of the most commonly used sources to get recommendations are:

1) *Word of Mouth* : The recommendations given by a fellow reader is bound to be biased to his/her personal interests.
2) *Bestseller List* : This is just a list, and the books on this list does not reflect the reader's taste.
3) *E-Commerce Recommendations* : The recommendations provided by popular e-commerce websites (eg., Amazon) is influenced by the items purchased by the user. These purchases do not necessarily mean the book is bought for self-use. Moreover, the ratings and the reviews provided by the user is not exclusive for the book but for the entire shopping experience.

II. Problem Statement

We propose BRIE (Book Recommendation and Intelligence Engine) a dedicated data science product for books in order to tackle the above mentioned problems.

We aim to create BRIE as an interactive data analytics product for books with the following features:

1) *A Hybrid Recommendation Engine* : We propose to create a lightweight recommendation system for books that aims to eliminate as much bias as possible in its suggestions. A system that combines user similarity, item similarity and the reader's taste to give reliable recommendations.
2) *Content Dissection* : The content of the book is analyzed to provide specific ratios of different genres involved in the book.
3) *Smart Book Viewer* : An analytical page for each book that allows users to view a book for self-analyzing their affinity towards the content of the book.
4) *In-Depth Analytics* : Useful dashboards for readers, authors and publishers to view book statistics from a different perspective.

Simon Fraser University

III. Data Science Pipeline

BRIE is composed of the following fundamental building blocks of a data science product:

1) *Data Collection* : Numerous web-scrapers and REST API python scripts were deployed to collect data from more than 10 different sources such as Amazon, Goodreads, Wikipedia, Riffle, Readgeek, etc. Some of the most basic information of a book includes - title, author, language, description (obtained from 5 different sources), reviews, comments, etc. Data for about 30,000 books was collected which amounted to approximately 4 GB of raw and uncleaned data.
2) *Data Cleaning* : All the above collected data heavily involved textual content (descriptions, reviews and comments). These large blocks of texts needed to be processed into certain formats for the recommender system. Before this could be achieved, the text had to be cleaned to remove stop words and proper nouns which was followed by basic lemmatization.
3) *Data Integration* : The data was modeled carefully based on different use-cases and it was spit into two storages :
   a) MySQL : Used to store part of the data for uses cases that required frequent joins across multiple tables.
   b) MongoDB : Used to store part of the data that can be used as a data dump.

   With this modeling in place we were able to achieve faster data retrieval and also the ability to handle larger volumes of data if necessary.
4) *Data Science & Intelligence* : Variations of text classification models are used to dissect the genres of each book. A multinomial text model is implemented to come up with accurate book recommendations. The entire recommendation algorithm runs on top of Spark.

5) *UI/UX & Visualization* : The Django web framework is used to host the entire product. The front-end is written in HTML, JavaScript and CSS with bootstrap implementation. The reports are dynamically generated using Plotly so that when additional data is introduced into the storage no extra work needs to be put into generating new dashboards.

IV. Methodology

The working of the most important features that make BRIE unique are explained in detail in this section:

1) ***Smart Book Viewer:*** The user can search for any book in the catalog based on the book's title. A unique web page is rendered for each book on clicking the book. This page primarily contains the most basic information about the book such as title, author, description, book cover, etc. Pricing analytics (prices scraped from different e-commerce websites) is given for the user to select the vendor based on the price if he/she wishes to purchase the book.

The detailed genre dissection of the book is displayed. A bag of words model is initially created. A bag of the most famously used words in each genre is collected by scraping the web. Then this collection is used to determine the score of each word occurring in the book's description by computing the relative document frequency. Then these scores are accumulated for each genre. From these scores the appropriate scores for each genre are determined.

Also, a set of 10 books is displayed which are identified to be most similar with respect to the current book under consideration. This is determined by computing the Jaccard Similarity between different books using title, description, author and publication details of the book. The books having the highest similarity score with respect to the current book are taken and displayed in the web page.

In addition to the features stated above, a special report is provided for every book page to identify the most relevant books amongst the 10 most similar stated above. The bubble chart identifies the most relevant books amongst the 10 using physical features of the books and identifies the most appropriate books that fall within the price and page range with respect to the current book.

2) ***Recommendation Engine:*** The user can search for any book in the catalog. A set of books need to be rated for the engine to start with (the cold start problem persists and the avenues to avoid this are being explored). The user needs to rate the books with a score in the range of 1 to 10. Any score lesser than 5 is considered to be a dislike (negative class) from the user and a like (positive class) otherwise. These selected books are considered to be the training data for the recommendation system. For each of the rated book, its list of 10 most similar books are added to their respective classes to make the training data larger.

The first step in constructing a multinomial naive bayes model for recommendation is to build a probability model for the features (text words) in the training data. The relative frequencies of each word in different documents for each book is calculated for both positive and negative classes. These values are the conditional probabilities with respect to the class labels in computing the conditional probabilities of each book with respect to positive or negative class using bayes theorem.

The following steps are followed to construct multinomial naive bayes model. A Book is made up of a set of documents

$$d_1, d_2, ..., d_n = B$$

Each document is made up of a set of words

$$w_i \in d_i$$

Each book in the training data can either be a like or a dislike by the user. Their posterior probabilities are given as

$$P(like), P(dislike)$$

Bayes theorem is given as

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)}$$

Using this theorem, the posterior probabilities of each of the word can be calculated as follows

$$P(like|w_i) = \frac{P(w_i|like).P(like)}{P(w_i)}$$

By using the naive assumption of the theorem where the occurrence of one word is independent of another word, the conditional probabilities of books with respect to the two binary classes can be computed as follows

$$P(like|B) = \pi_{i=1}^{n} P(w_i|like).P(like)$$

$$P(dislike|B) = \pi_{i=1}^{n} P(w_i|dislike).P(dislike)$$

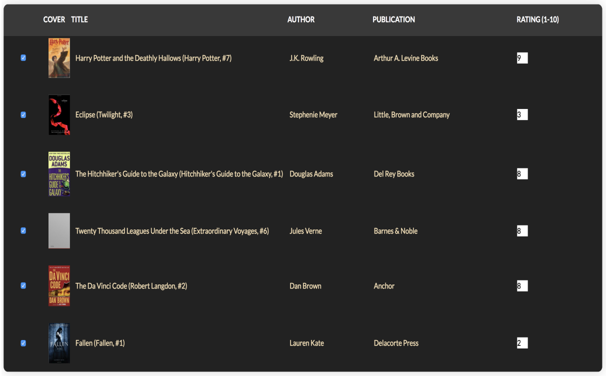
Whichever conditional probability has a higher value, the book is classified into the respective label. In BRIE, after the computation of these conditional probabilities, some additional measures are taken to remove biases by introducing penalties. For example, books belonging to one of the series highly disliked by the user should have its probabilities penalized. After such biases are removed, a score for each book is obtained based on which the recommendations are given to the user.

There are about 30,000 books in BRIE's catalog. In order to make the recommendations as lightweight and scalable when the number of books increases, the multinomial is not built to predict a score for all the books in the catalog. Before the recommendations are given, the user's taste profile is created. Based on the ratings of the selected books and the content dissection of those books, the genre profile of the user is determined. Moreover, from the model created above, the set of words with the highest relative frequencies is presented to the user to showcase the most influential words in the user's taste. Only the books that match the genre profile of the user are taken for score calculations. The books with the highest scores amongst this set are given as recommendations to the user.

V. Data Product

This section explains the working of BRIE from the user's perspective with respect to the steps explained in the previous section. BRIE is developed as a full stack data science product with a fully functional user interface that allows a user to interact with the system to obtain valuable knowledge on the publishing industry and novel book recommendations.

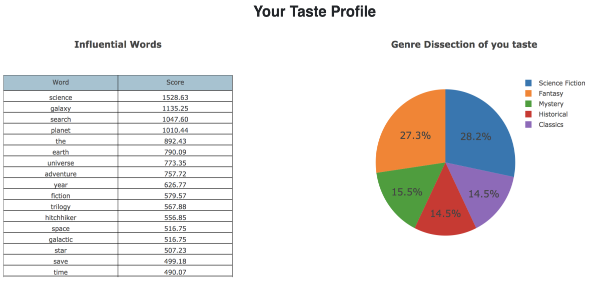
*1) **Recommendation Engine:*** A simple book title feature is provided to enable the user to search any book in the catalog. As shown in Fig. 1. the searched books are rendered, and rating of 1-10 is required for each of the selected books. The selected books are pinned and additional searches can be performed.



	COVER	TITLE	AUTHOR	PUBLICATION	RATING (1-10)
☑		Harry Potter and the Deathly Hallows (Harry Potter, #7)	J.K. Rowling	Arthur A. Levine Books	9
☑		Eclipse (Twilight, #3)	Stephenie Meyer	Little, Brown and Company	3
☑		The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the Galaxy, #1)	Douglas Adams	Del Rey Books	8
☑		Twenty Thousand Leagues Under the Sea (Extraordinary Voyages, #6)	Jules Verne	Barnes & Noble	8
☑		The Da Vinci Code (Robert Langdon, #2)	Dan Brown	Anchor	8
☑		Fallen (Fallen, #1)	Lauren Kate	Delacorte Press	2

Fig. 1. Books selected and the ratings given by a reader

Based on these selected books and ratings, a user profile is generated to make the reader understand his/her taste better as shown in Fig. 2.



Word	Score
science	1528.63
galaxy	1135.25
search	1047.60
planet	1010.44
the	892.43
earth	790.09
universe	773.35
adventure	757.72
year	626.77
fiction	579.57
trilogy	567.88
hitchhiker	556.85
space	516.75
galactic	516.75
star	507.23
save	499.18
time	490.07

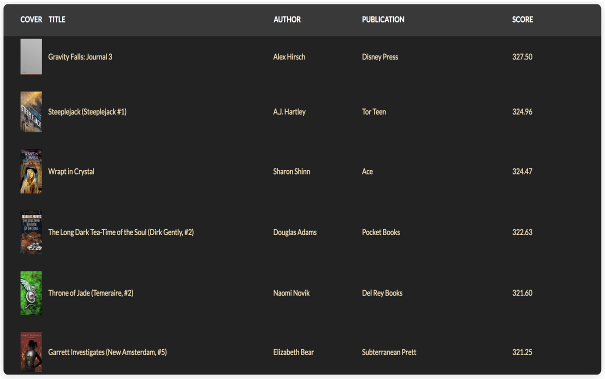
Fig. 2. A sample taste profile for a user

Once the taste profile of the reader is conveyed, using the multinomial text model explained in the previous section the most relevant books are recommended to the user with a score as shown in Fig. 3.

*2) **Smart Book Viewer:*** Each book has its own "smart view" page. When the user searches for a book in the catalog and clicks one of the links of the resulting books, the "smart view" page for the book is rendered.

As explained in the previous section, the components that make up the smart book viewer are:

1) Basic book details
2) Price analytics

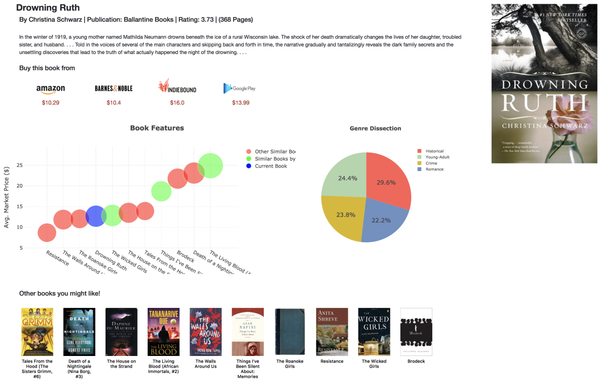


COVER	TITLE	AUTHOR	PUBLICATION	SCORE
	Gravity Falls: Journal 3	Alex Hirsch	Disney Press	327.50
	Steeplejack (Steeplejack #1)	A.J. Hartley	Tor Teen	324.96
	Wrapt in Crystal	Sharon Shinn	Ace	324.47
	The Long Dark Tea-Time of the Soul (Dirk Gently, #2)	Douglas Adams	Pocket Books	322.63
	Throne of Jade (Temeraire, #2)	Naomi Novik	Del Rey Books	321.60
	Garrett Investigates (New Amsterdam, #5)	Elizabeth Bear	Subterranean Prett	321.25

Fig. 3. List of recommended books to a user

3) Content Dissection
4) Most similar books
5) Most appropriate book report

All these features are rendered in the smart book viewer as shown in Fig. 4.



Drowning Ruth

By Christina Schwarz | Publication: Ballantine Books | Rating: 3.73 | (368 Pages)

In the winter of 1919, a young mother named Mathilda Neumann drowns beneath the ice of a rural Wisconsin lake. The shock of her death dramatically changes the lives of her daughter, troubled sister, and husband. . . . Told in the voices of several of the main characters and skipping back and forth in time, the narrative gradually and tantalizingly reveals the dark family secrets and the unsettling discoveries that lead to the truth of what actually happened the night of the drowning. . . .

Buy this book from

Other books you might like!

Tales From the Hood (The Sisters Grimm, #6)
Death of a Nightingale (Nina Borg, #3)
The House on the Strand
The Living Blood (African Immortals, #2)
The Walls Around Us
Things I've Been Silent About: Memories
The Roanoke Girls
Resistance
The Wicked Girls
Brodeck

Fig. 4. Smart Book Viewer for the book "Drowning Ruth"

VI. Analytics

The goal is not to develop BRIE just as a recommendation engine but rather as a real-time data science product that can be used by anyone in the publishing industry. Hence, apart from the data science involved in the recommendation engine, there are also a set of analytical dashboards on offer which can be used by readers, publishers and authors. Some of the interesting reports are explained in detail in this section.

*1) **Overview Dashboard:*** Fig. 5. showcases some interesting reports that are part of the overview dashboard. The purpose of this dashboard is to provide a bird's eye view of all the books in BRIE's catalog. The figure shows

- The 10 most popular genres based on the number of books in the catalog; It can be seen that "Fantasy", "Historical" and "Mystery" are the genres with the most published books.
- The next plot showcases the trend in the average number of pages in the books published in a few time intervals.

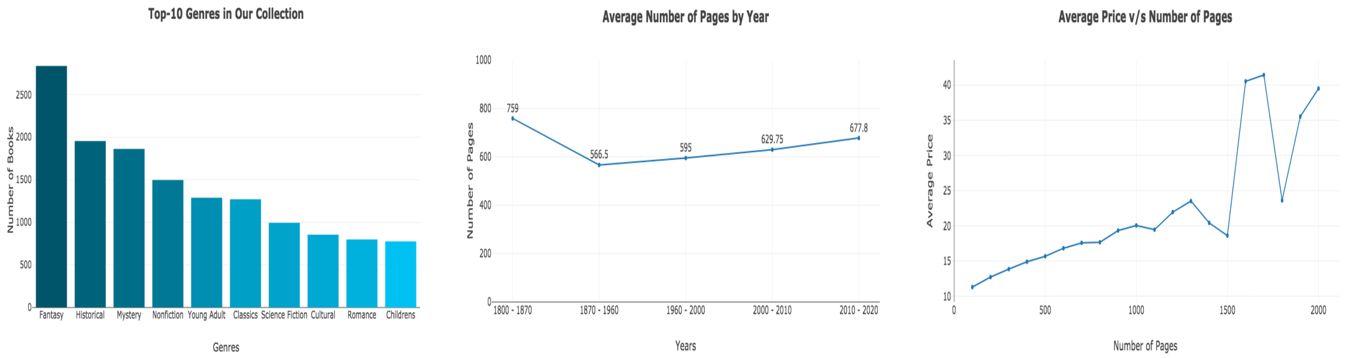


Fig. 5. Some interesting reports in the Overview dashboard

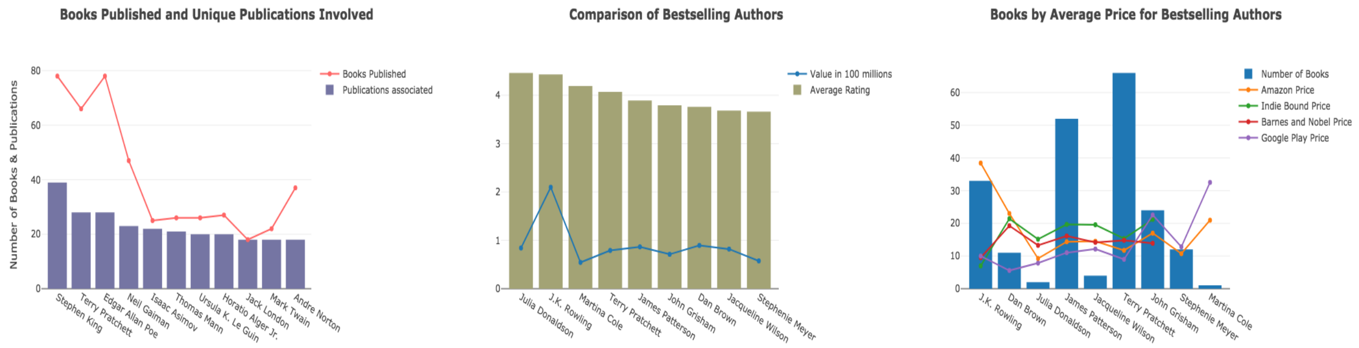


Fig. 6. Some interesting reports in the Authors Dashboard

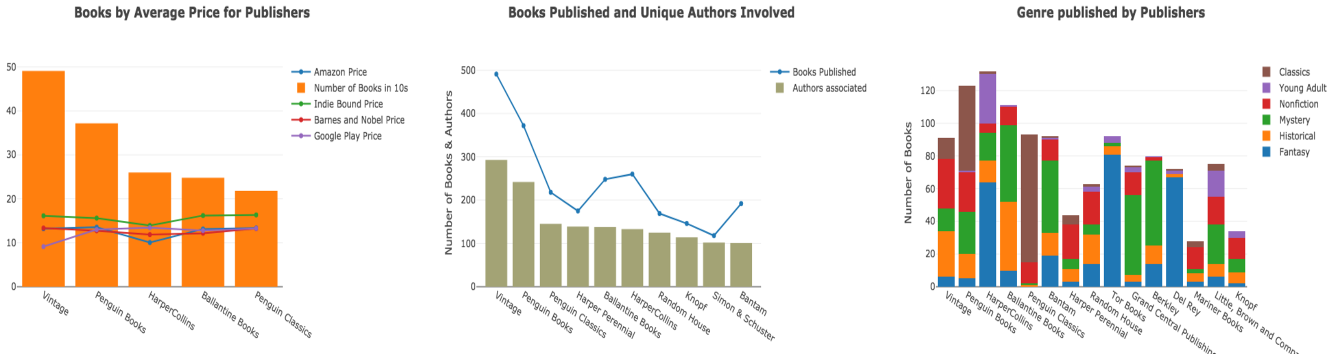


Fig. 7. Some interesting reports in the Publishers Dashboard

A general increasing trend has been observed in the number of pages written per book in the recent years.

- The next trend showcases the book prices with respect to the size of the book in pages. It can be observed in general that both these parameters have a positive correlation as expected.

2) ***Authors Dashboard:*** Fig. 6. showcases some interesting reports that are part of the authors dashboard. The purpose of this dashboard is to discover some interesting information about authors that can be used by readers and publishers. The figure shows

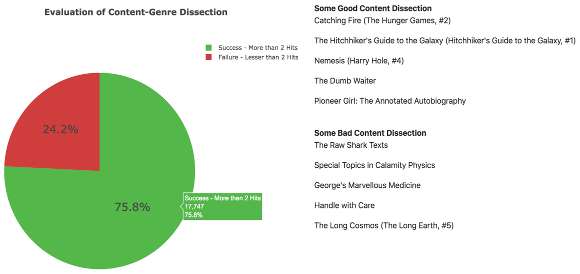
- The first plot shows how many different publications were some of the famous authors have been involved with.
- The next plot depicts a comparison between some bestselling authors with respect to gross revenue generated and the average rating of the books written by them. For example, it can be seen that though J.K. Rowling's net revenue is much higher than Julia Donaldson but Donaldson has a higher average rating for her books. It can be concluded from this report that Julia Donaldson's books are critically acclaimed.
- The next report shows the price comparison of books by popular authors in various e-commerce websites.

*3) **Publishers Dashboard**:* Fig. 7. showcases some interesting reports that are part of the publishers dashboard. The purpose of this dashboard is to discover some interesting information about publishers that can be used by readers and authors. The figure shows

- The most publications are identified based on the number of books published by the particular house. These publication houses are taken and an average price comparison of the books published by the respective houses is given.
- The next chart showcases how inclined are the popular publication houses in sticking with the same authors to publish new books.
- The stacked bar chart depicts the number of books belonging to various genres produced by popular publication houses.

VII. Evaluation

The content dissection of each book was evaluated for correctness. Each content dissection profile of a book contains 4 genres. Moreover, the actual genre of the book is also scraped during the data collection process. The content dissection is assumed to be correct if the top 2 genres in the dissection profile matched with the actual genres of the book and as a wrong dissection otherwise. Fig. 8. shows the evaluation results of the books in BRIE's catalog.


Evaluation of Content-Genre Dissection
🟩 Success - More than 2 Hits
🟥 Failure - Lesser than 2 Hits
24.2%
75.8%
Success - More than 2 Hits
17,747
75.8%

Some Good Content Dissection
Catching Fire (The Hunger Games, #2)
The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the Galaxy, #1)
Nemesis (Harry Hole, #4)
The Dumb Waiter
Pioneer Girl: The Annotated Autobiography

Some Bad Content Dissection
The Raw Shark Texts
Special Topics in Calamity Physics
George's Marvellous Medicine
Handle with Care
The Long Cosmos (The Long Earth, #5)

Fig. 8. Evaluation Results

As it can be seen, about 74% of the books were accurately dissected into their respective genres. It was observed that most of the books that fell in the 24% of the wrongly classified books belonged to genres such as "sports", "physics", etc. Since BRIE is not equipped to handle these genres these books were misclassified. However the system can easily be extended to handle these genres as well.

As of now the only evaluation model employed to measure the quality of the book recommendation is by asking for feedback from the users using BRIE. It is also worth noting that the recommendation system was periodically updated based on various suggestions given by different beta users.

VIII. Lessons Learnt

There were many critical design decisions that had to be taken during the design and the implementation process of BRIE. This section describes a few of such decisions and the detailed explanation behind these decisions.

1) Multiple web scrapers were implemented to scrape data from numerous sources. The web pages being scraped were dynamic and required browser actions. Selenium was used to automate the actions. This rendered the entire data collection process to be quite slow. Hence, these scrapers were segregated into batches and were run on multiple machines in parallel to reduce the time take in this phase.
2) Multiple avenues of data storage were explored. Google Cloud Platform's BigQuery was tested with the data in hand. Not only the network latency was high but also the pricing plans did not suit BRIE's use cases. Hence a split storage model between local MySQL and MongoDB was chosen.
3) Various charting libraries such as chart.js, chartist.js and d3.js were experimented with. Finally, plotly was decided to be used as the framework for visualization for simple reasons such as being lightweight, ease of usage and the visual appeal.

IX. Future Work

BRIEs recommendation was tested on various readers and many positive reviews were recorded. This gives us further motivation to make the entire system more robust and novel. We propose the following as future modules of work

1) To garner more reviews for making the system better, we propose to host BRIE on SFU's cloud infrastructure.
2) The system works with the data that has been collected during the data collection phase. This introduces some static dependencies. Hence we propose to eliminate these dependencies by creating a dynamic data fetch module for the data that tends to change over a period of time.
3) We propose to solve the cold start problem by implementing clustering mechanisms for books.
4) Book covers can be used as an additional source of information by applying neural networks for image classification.
5) In order to make the entire system a well rounded web application we also propose to create a user account signup and account maintenance setup.

X. Conclusion

BRIE was primarily developed with the motive to bridge the gap in the market for data science applications in the publishing industry and to act as a novel book recommendation system that rarely fails to deliver what the reader expects because we believe in the saying "There is so much more to a book than just reading it".