Book Recommendation & Intelligence Engine (B.R.I.E)

Lakshayy Dua, Sethuraman Annamalai, Supreet Takkar

SFU BIG DATA

Introduction

Motivation

- Being avid readers, we understand that the question that plagues each one of us is – "Which book should I read next?"
- A large portion of the reading community depends on the following sources for recommendation:
  - *Word of mouth* – The suggestions may be biased towards the recommender's choice
  - *Bestseller List* – These lists don't know reader's taste
  - *e-commerce websites* – Recommendations are just based on similar books and the book reviews are for the entire service experience.
- There isn't any dedicated data science product that caters to the needs of everyone involved in the publishing industry with intelligent recommendations and in-depth analytics.
- B.R.I.E has been created as a full-fledged interactive application with rich features to address all these needs.

Product Features

- **A hybrid recommendation engine –** Augments reader's taste with similar books to give high-quality recommendations.
- **Content Dissection –** The content of the book is analyzed to provide specific ratios of different genres involved in the book.
- **Smart Book Viewer –** Provides the reader with interesting stats to help decide if the user might like it.
- **In-depth analytics –** Useful dashboards for readers, authors and publishers to view book statistics from a different perspective.

Pipeline & Tools Used

Data Collection – Multiple web crawlers were implemented along with REST API scripts to collect data of about 30,000 books from multiple sources such as Goodreads, Amazon, Riffle, Wikipedia, etc.

Data Cleaning – A large volume of text data (reviews, comments, description, etc.) was cleaned using different mechanisms such as stop-words and proper-nouns removal, lemmatization, etc.

Data Integration – The data was modelled carefully based on different use-cases and it was spit into two storages - MySQL and MongoDB to scale to larger amounts of data efficiently and easily.

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.

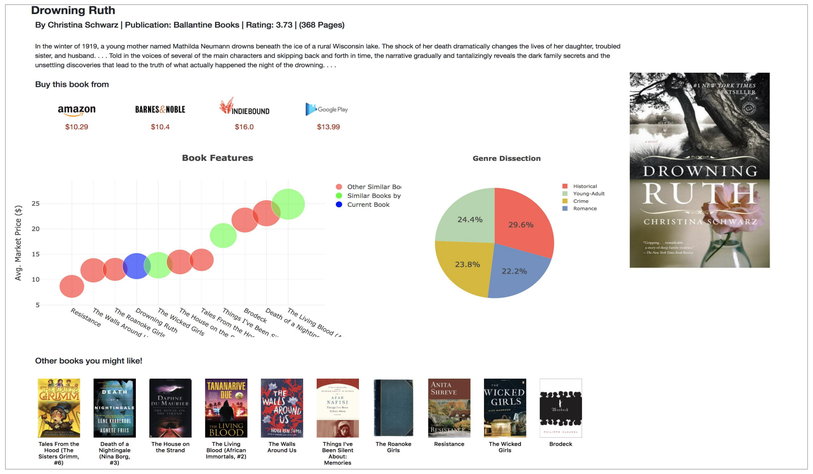
UI/UX & Visualization – The Django web framework is used to host the entire product. The front-end is written in HTML, JavaScript and CSS with bootstrap implementation. The reports are dynamically generated using Plotly.

Methodology

Smart Book Viewer

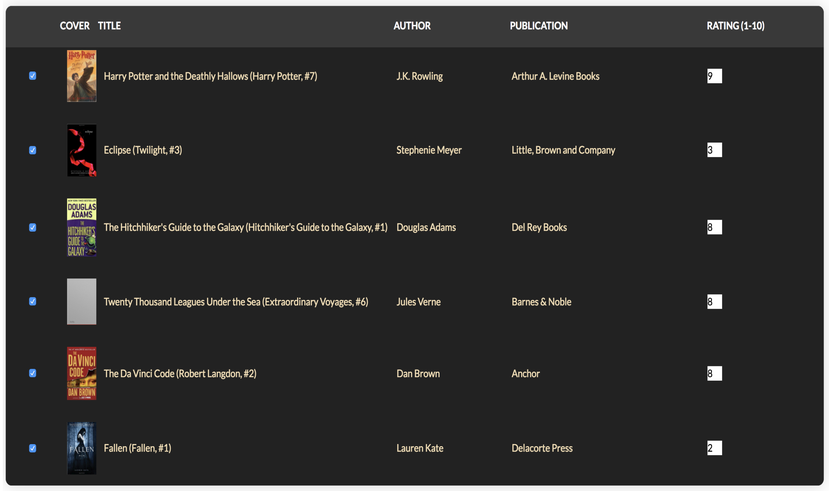
- The reader can search for any book in the catalog and visit the detailed analysis page of each book.
- A detailed description for the book is provided.
- Prices obtained from different e-commerce websites are provided for comparison.

- *Content-Dissection* : A bag-of-words model is used to collect the most frequently used words in each genre. Each book's description from various sources is taken into account and a general probability model is created to determine the ratios of different types of content in the book.
- *Similar Books :* A list of the most similar books to the current book is displayed. An Entity Resolution model is utilized to compute the Jaccard Similarity between different books and the pairs of books with the highest similarity scores are chosen.
- *Comparison of Similar Books :* The current book is compared with its most similar books (which were chosen based on content) using other features such as prices and pages to determine the most appropriate book with respect to the current book.



Recommendation Engine

- A set of books are rated (scored out of 10) by the reader. A book with a rating equal or below 5 is considered to be a dislike.

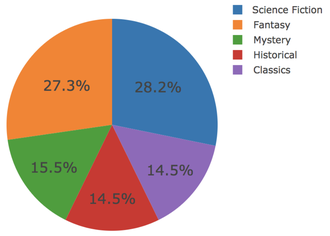


- Based on these ratings, a profile for the user is generated by considering the genres and the "content-dissection" of the books most liked by the reader.
- This profile also contains a set of words that influence the reader's choice the most.

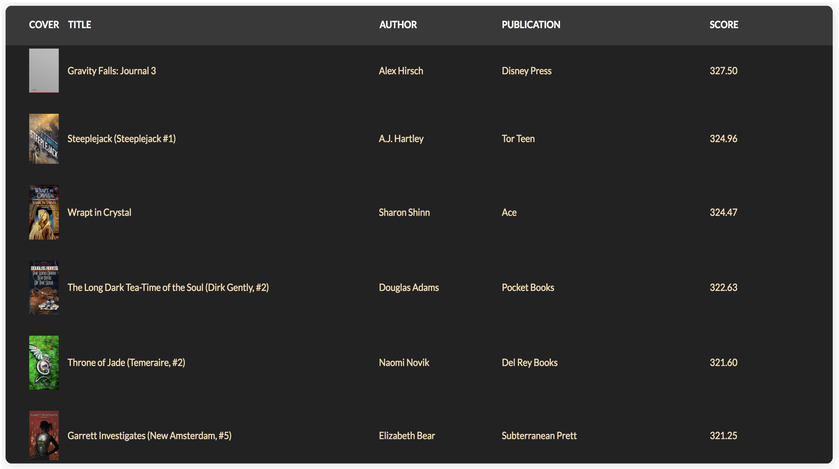
Your Taste Profile

Influential Words

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

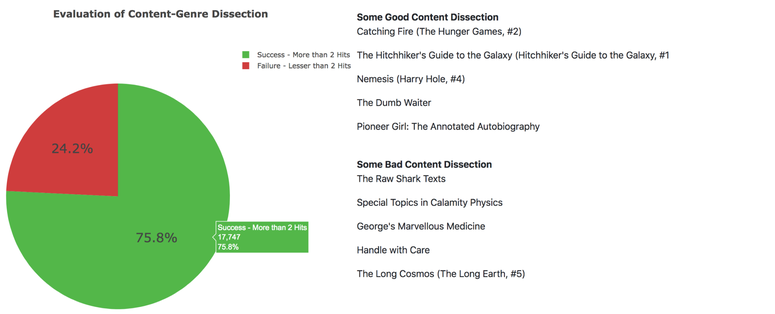


- With this profile, a Multinomial Naïve Bayes model is constructed to give novel recommendations.



Interesting Findings & Results

Evaluation



Nearly 76% of the books were correctly dissected based on genres. Also, it was observed that the most of the misclassified books belonged to topics such as "Physics", "Sports", "Academic", etc. which the system is not yet capable of handling. Currently, the only way to evaluate the recommendation system is getting suggestions from the readers. This is a critical module of work for the future.

Future Work

- B.R.I.E's recommendation was tested on various readers and positive reviews were recorded. Hence, we plan to fine-tune the recommendations and host the application on SFU's cloud.
- Book covers can be used as an additional source of information by applying neural networks for image classification.
- A reader sign-up and account maintenance model can be created to make this a well rounded application.