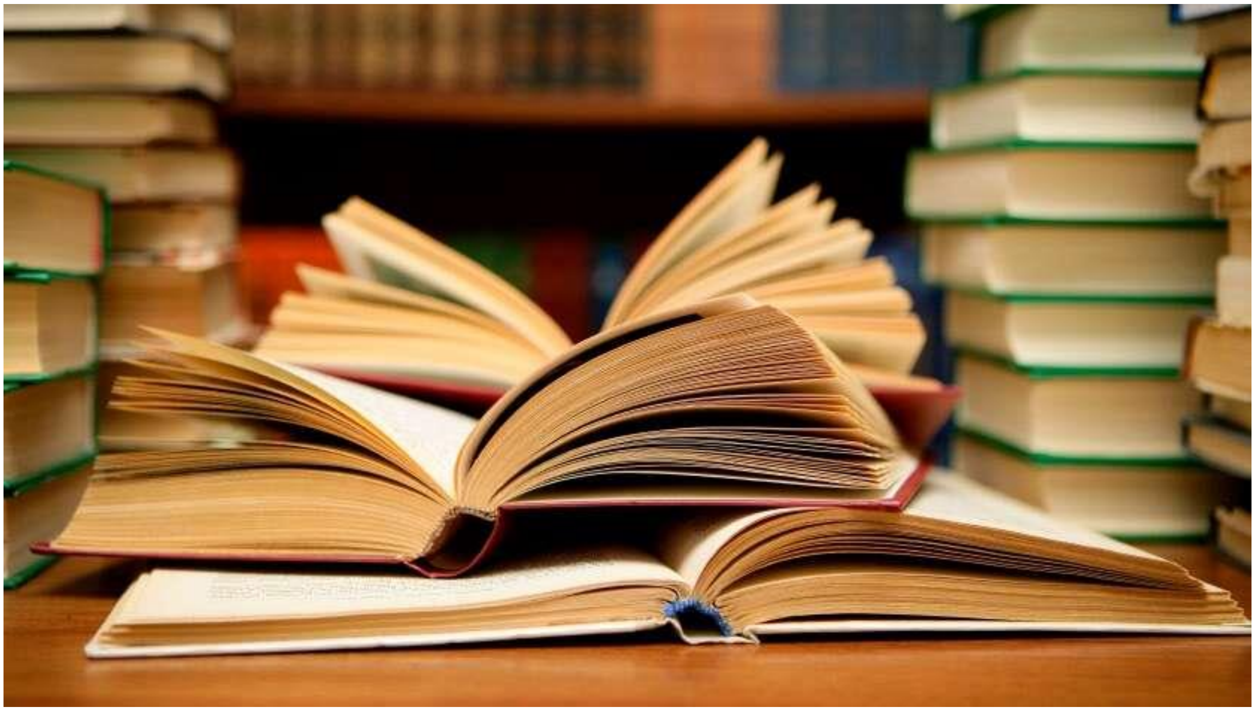
Library Book Recommendation System



Bradley Azegele

Emmanuel Kipkorir

Monicah Iwagit

Belinda Nyamai

Dennis Kimiri

Femi Kamau

1. Business Understanding

1.1 Overview

A library is a collection of information resources, such as books, periodicals, and electronic documents that is organized for use and maintained by a public, institutional, or private body. Library usage is a common trend for an education set-up where people go to gain more insights. Libraries serve an important role in providing access to information and knowledge to members of the community. They can be a valuable resource for students, researchers, and other individuals who are looking for information on a particular topic or subject. In a library system the information about the book a user selects preferably aligns to their interests and reading preferences. Recommender systems are algorithms aimed at suggesting relevant items to users by evaluating and filtering information. They are categorized under the class of personalized information filtering technologies, targeted to help in decision making given large information sets.

Users have been relying on the search engine to retrieve books given the vast book collection available which involves trying out different keywords, and adjusting them till the required results are obtained. This procedure is favorable to those who know how to represent their requests clearly, especially in a manner quickly interpreted by the engine since they reach their goals soon. However, users’ carefully made decisions and options could not benefit others who have similar requests, although the system has helped to record who has borrowed which books at what time. Hence, search engines are not enough for us to effectively find what we want in a library, and we desire a smarter assistant which could make use of peers’ options.

This project aims towards coming up with a library book recommender system that constitutes an information filtering technique which presents books according to user preferences .By implementing a book recommendation system, the library can provide personalized recommendations to users and help them discover books they are likely to enjoy. This will ensure engagement with various library books and save on the time taken to discover the most suitable book leading to an increase in circulation and overall usage of the library's collection.

1.2 Problem Statement

The tremendous growth and usage of information has led to information overloading where users find it difficult to locate the right information at a specified time. Although there are previous studies conducted on library recommender systems, the datasets used were small compared to the dataset we intend to use hence minimal area coverage. Recommending the right library books is a challenge due to the variety of genres available and the huge collection of books provided. A user finds it difficult to select the most appropriate book that will suit their academic needs, this process consumes a lot of time that the user would have used to sharpen on their desired skills. Additionally, many books in the library are rarely utilized which results in a waste of library resources. Having a personalized recommendation system seeks to predict the preference based on the user's interest, behavior and information. The application of recommender systems in the university library solves the problem of difficulty in choosing books and improves utilization rate of library resources.

1.3 Business Objectives

- Develop a personalized library book recommendation system using different approaches.
- Use Natural Language Preprocessing on book description for implementation in the system.
- Evaluate the accuracy of the recommendation system using Mean Absolute Error to obtain a reasonably low value.

1.4 Metrics of Success

Coming up with a hybrid recommendation system that combines two or more filtering techniques to ensure books are proactively recommended to users.

2. Data Understanding

Data understanding provides a solid foundation for the subsequent steps; data preparation, exploratory data analysis, model deployment and evaluation. The dataset was sourced from link, and was collected by Cai-Nicolas Zinder while in a four week crawl (August-September [2004]) while in a University in Germany. The dataset contains 4 tables named User.csv, Ratings.csv, Books.csv and Books.extra.csv. The description of the datasets is as follows;

- User.csv - it has attributes like user_id, location, age and contains 278858 books.
- Ratings.csv - it has 1149780 books, its attributes are user_id, book_isbn number and rating.
- Books.csv - it contains attributes like author, title, book_isbn number, year of publication, publisher and 271360 books.
- Books.extra.csv - it contains 271,044 books, its attributes are authors, published_date, description, page count, categories, maturity_rating and language.

3. Data Preprocessing

Data preprocessing mainly entails the manipulation of raw data before it can be used to enhance performance. It was conducted through the following processes;

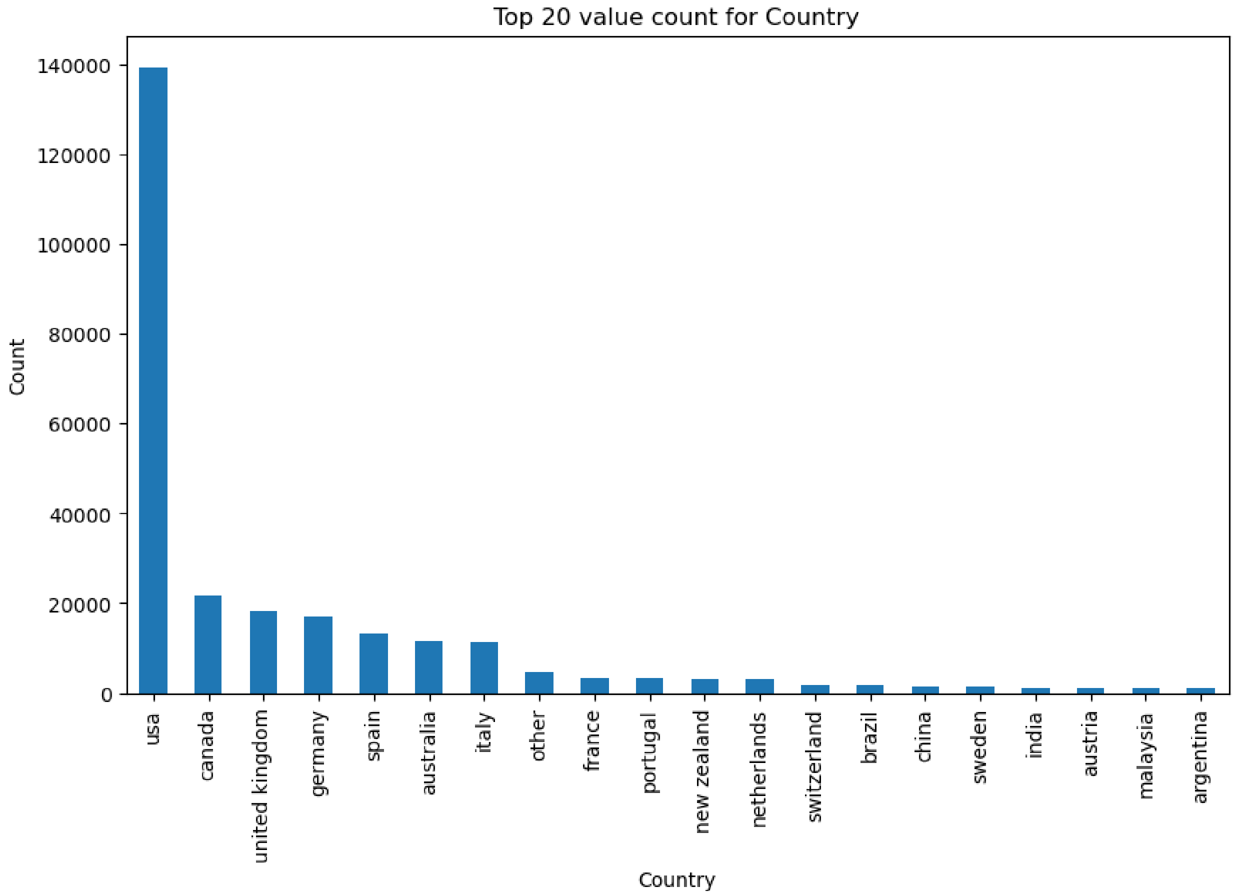
- Selecting data to discover which data sets will be used.
- Cleaning data in order to correct, impute, or remove erroneous values.
- Integrating data to create new datasets for effective analysis.
- Formatting data which involves converting string values that store numbers to numeric values in order to perform mathematical operations.

4. Exploratory Data Analysis

This process is conducted to summarize the main characteristics of the dataset by using visual methods. The process involved exploring the dataset;

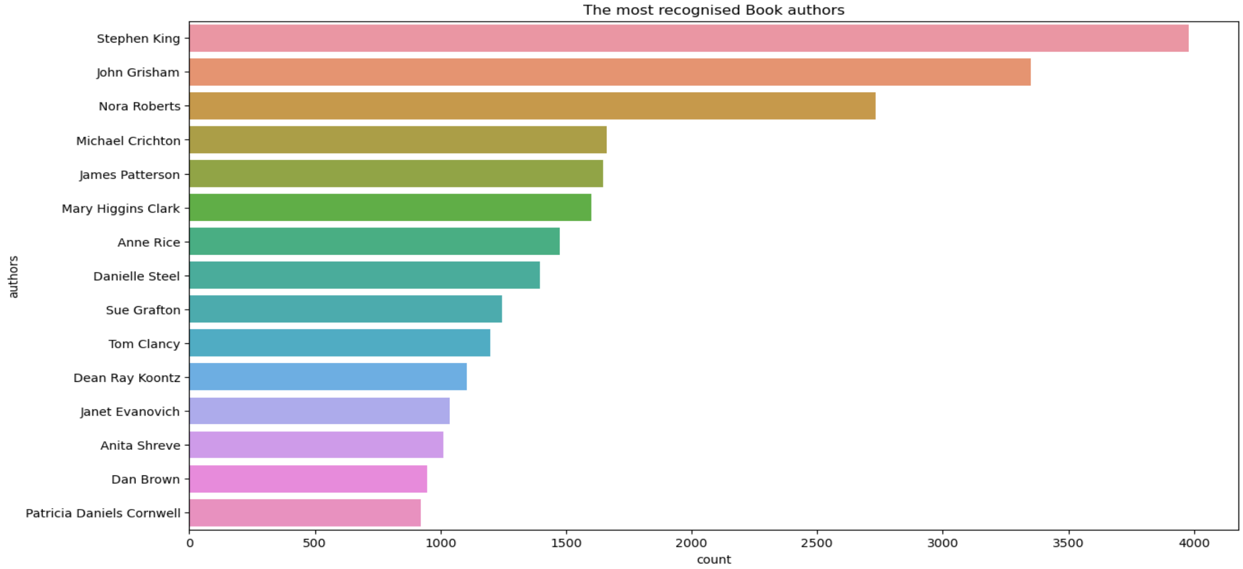
- To discover the country with a huge number of library users.
- To find out the most recognized authors for writing books.
- To check for the most preferred publishers for publishing books.
- To discover the most preferred book category.

Discovering the number of library users per country



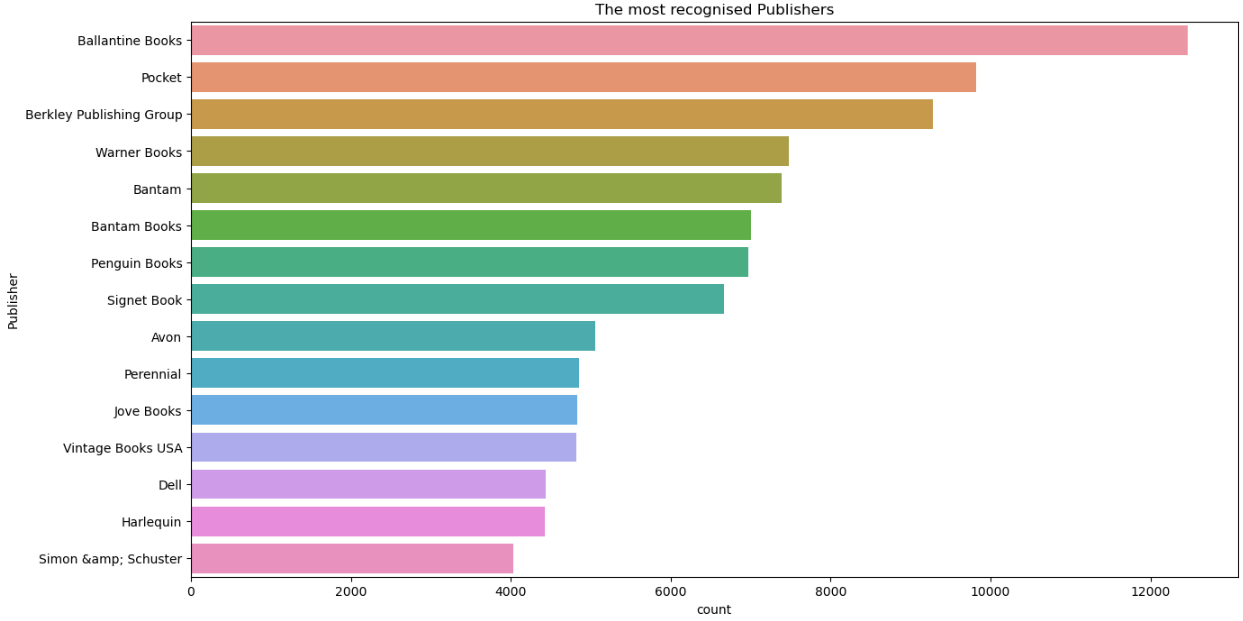
The country with a huge number of library users is the USA while the country with the least number of users is Argentina.

Discovering the most popular authors



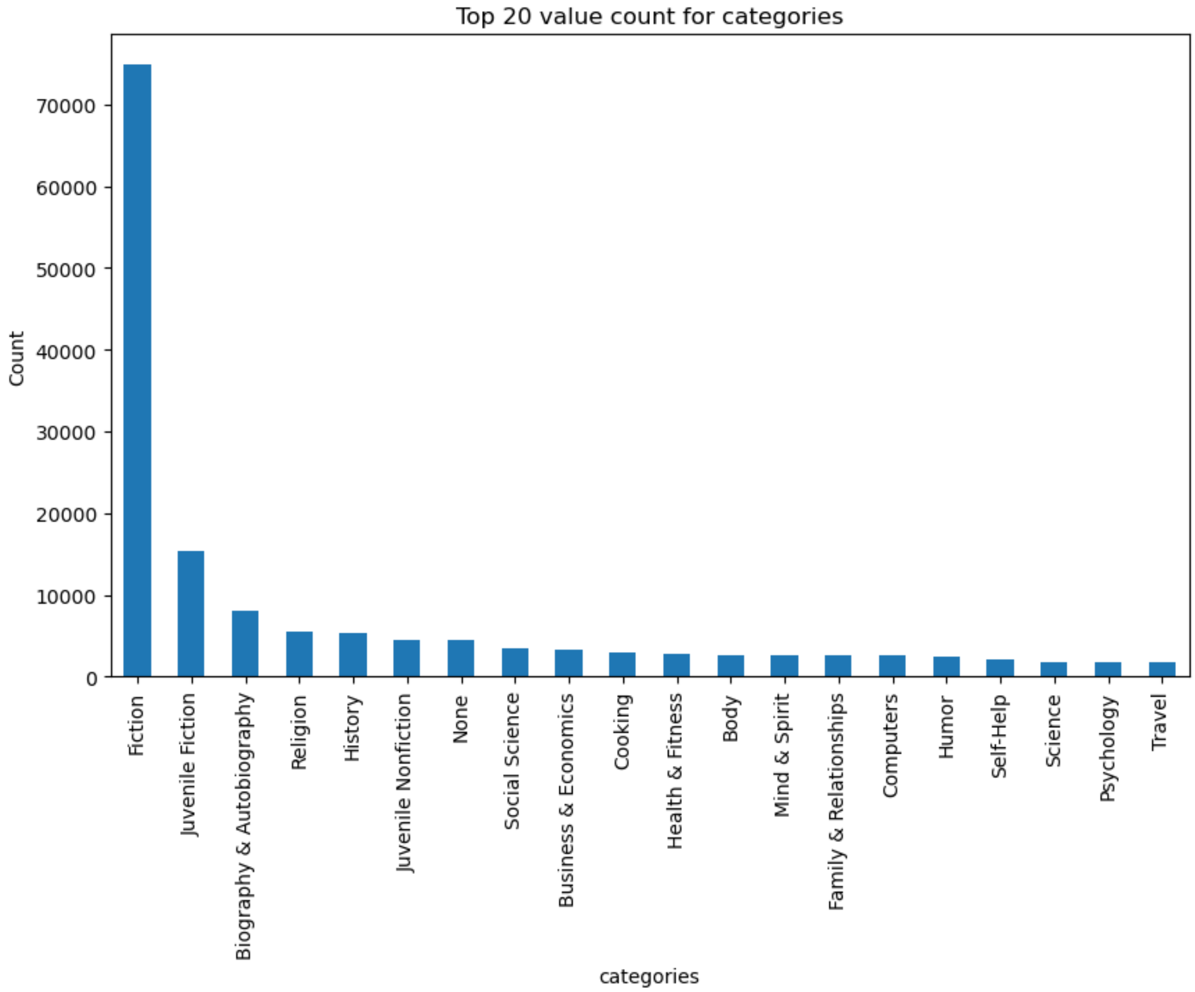
The most popular author is Stephen King while the least preferred author is Patricia Daniels Cornwell which could be attributed to the content they mainly major their writing on.

Checking for the most preferred publisher



The most preferred publisher is Ballantine books while the least preferred publisher is Harlequin and Simon and Schuster. This discrepancy could mainly have been caused by the type of books they publish and how regularly they publish them.

Checking for the most preferred book category



The most preferred book category is books based on Fiction while the least preferred is books based on psychology and travel.

6. Modeling

The recommendation systems that were created include;

- Popularity based recommendation system
- Collaborative filtering user-item filtering system
- Collaborative filtering item-item filtering system
- Content based recommendation system
- Hybrid recommendation system

Popularity based recommendation systems work with the trend by using items that are in trend currently. For instance, if any book is usually bought by every new user then there are chances that it may suggest that book to a new user who just signed up to the system.

Book weighted average formula:

Weighted Rating (WR) = $[vR/ (v + m)] + [m C/ (v + m)]$

Where;

v is the number of votes for the books;

m is the minimum votes required to be listed in the chart;

R is the average rating of the book;

C is the mean vote across the whole report.

However, the popularity based recommender provides a general chart of recommended books which are not sensitive to the interests and tastes of a particular user.

Collaborative Filtering is a technique used to make recommendations to Book Readers. It is based on the idea that users can be used to predict the extent to which a particular product or service they have used/experienced will be liked by a new user. There are two categories of collaborative filtering algorithms; memory based and model based.

Model based approach involves building machine learning algorithms to predict user's ratings. They involve dimensionality reduction methods that reduce high dimensional matrices containing an abundant number of missing values with a much smaller matrix. This section mainly compares Singular Value Decomposition (SVD) and Non-Negative matrix factorization (NMF) using the surprise library.

Memory based algorithms apply statistical techniques to the entire dataset to calculate predictions. They can be divided into two main sections: user-item filtering and item-item filtering. The goal of this system is to compare user-item and item-item approaches, try different configurations of parameters, optimize model performance and explore obtained results.

User-item filtering mainly involves finding users that have rated similar content and using their preferences to recommend new items.

Item-based filtering mainly involves finding similar books based on other users' preferences.

Content Based recommendation system that finds the similarity in a book based on the book title, author, publisher and book category and recommends it to a user.

A Hybrid recommendation system brings together ideas from content and collaborative filtering systems. It mainly builds an engine that gives book suggestions to a particular user based on the estimated ratings that it had internally calculated for that user.

6. Model Evaluation

Mean Absolute Error was used to evaluate the accuracy of the recommendation system by measuring the average of the absolute deviance between the actual and estimated ratings given by users. Content-Based had a value of 0.041, Item-based and hybrid system both had a value of 0.041. The hybrid system was however preferred since it is more adaptive to the various preferences a library user may have.

7. Conclusion

The book recommendation system increases the visibility and availability of books in a library by helping new users discover the most efficient books from past users' experience. The most recognized author is Stephen King followed by John Grisham while Harlequin has the most number of books published. The most active library users are in the age bracket between 20 and 30 years. The books category that most users prefer are novels with content mainly based on fiction.

8. Challenges encountered

Handling of sparse information from the datasets was a major challenge since user interactions were not present for majority of the books. Understanding the metric for evaluation was a challenge as well. Since the data consisted of text data, data cleaning was quite hectic for columns like description. Decision making on missing value imputations and outlier treatment was also a time consuming tiresome task.

9. Recommendations

Incorporating data security and privacy measures to protect the personal details of users by encrypting personal data and ensuring strict access controls. More feedback channels should also be implemented to improve accuracy and performance of the recommendation system. For instance, a ratings system or a comments section where people give reviews concerning the various books they have read or borrowed from the library.