IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

# BOOK RECOMMENDATION SYSTEM JUST READ IT!

**Mr. D JAYARAM[1], Dr. G. N. R. PRASAD[2], ISHIKA GUPTA[3], KAVYA KONDI[4]**

[1] Asst. Professor, Dept. of Information Technology, CBIT(A), Hyderabad – 500 075, India

[2] Sr. Asst. Professor, Dept. of MCA,CBIT(A), Hyderabad – 500 075, India

[3] B.E (IT) – VI Semester, CBIT(A), Hyderabad – 500 075, India

[4] B.E (IT) – VI Semester, CBIT(A), Hyderabad – 500 075, India

**Abstract:** Books play a very important role in every person's life by introducing them to a world of imagination, providing knowledge of the outside world, improving their reading, writing, and speaking skills as well as boosting memory and intelligence, all of which are quite necessary for different aspects of life. There exist many potential readers, but due to the abundance of information present on the internet, many of these people find it extremely hard to search for books that they might like and might inculcate in them a habit of reading, which is always encouraged. This could result in a huge loss as a lack of readings results in poor language skills, cultural ignorance, and fear of books. Furthermore, many people ask for book recommendations from their friends, neighbors, and families who might not always suggest the right book as they do not have knowledge about the numerous books that are available. If we plan to buy any new book, we normally ask our friends, research about the book, check the book ratings on the internet, find books that have similar content, and then we make our decision. How convenient if all this process was taken care of automatically and recommend the book efficiently? A recommendation system is an answer to this question. Recommendation System (RS) is software that suggests similar items to a purchaser based on their earlier purchases or preferences. The amount of information available on the internet is quite a lot and finding relevant information can become very difficult. Recommendation systems aim to solve such kinds of problems. With the help of recommendation systems, we can find relevant information quickly and easily. Many recommendation systems are also used in commercial websites to sell their products. Consequently, the main aim of our paper is to build a book recommendation web application. The web application can be used by anyone and does not require any login making it much more accessible and easy to use. The user needs to just enter the title of the book that they have read and liked before, and based on the genre and average rating given to the book. The top ten books will be recommended to the user that is the most similar to the book that they have entered.

The technologies used in this paper would be the python programming language for preprocessing the data, exploring the data, and building the machine learning model itself. Many python modules such as pandas and matplotlib and seaborn will be used for handling and visualizing the data. The algorithm that will be used to find books similar to the book entered by the user is the K-Means nearest neighbor algorithm. To provide a proper and appealing interface, this project is going to be a web application that will be developed using Flask. The initial exploration and preprocessing of the data will be done through Google Colab. The dataset used is 'books_1.Best_Books_Ever' from the 'Goodreads' dataset that contains attributes such as booke, title, author, rating, ISBN, genres, characters, awards, num ratings, ratingByStars, setting, bbeScore, bbeVotes etc.

**Keywords:** K-Means nearest neighbor; machine learning; Books; Receommendation;

## 1 INTRODUCTION

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. A recommendation system is a subclass of machine learning whose goal is to generate meaningful recommendations to users for items or products that might interest them. Some real life examples of recommendation systems include: suggestion of books on Amazon or movies on Netflix. A book recommendation system is a type of recommendation system that recommends book to the user based on their interest. There are 2 major techniques used in recommendation systems: Content based filtering and Collaborative filtering. Content based filtering: Content-based filtering, makes recommendations based on user preferences for product features. Collaborative filtering: Collaborative filtering mimics user-to-user recommendations. It predicts users' preferences as a linear, weighted combination of other user preferences. Content-based filtering outperforms user collaborative filtering. Items are more similar and make more sense than users similarities, and thus we have used the content based filtering method approach.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

Book recommendation systems are usually used in E-Commerce websites where books are recommended to the user after they have purchased a certain book. Apart from this, book recommendation systems can also be used in libraries and book shops where the librarian or the book shop owner can give recommendations to its borrowers and customers.

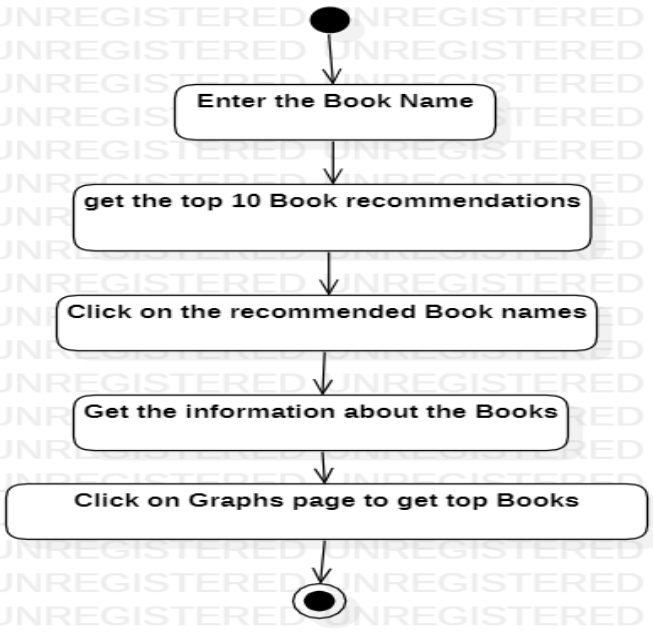
The main driving force of our mini project was to ease the user's task of searching for a good book to read. Books play a significant role in the growth of humans. From memory boost to the treatment of Dyslexia, the importance of books can be seen in every aspect of our day to day life. Books provide us with imagination, knowledge from the outside world, reading, writing and speaking skills. Due to the plethora of information available on the internet, many find it extremely frustrating and difficult to find the right book for them. Many people approach their friends and family for suggestions, who may not always pick the best book for them. This can lead to an aversion or fear of books in potential would-be readers which can hinder their growth as an individual. Thus, 'Just Read It!' was developed to provide an easy, quick and accessible way through which users can get just the perfect book for them to read. It will prevent the user from wasting their time and energy in search for a good book.

To create a book recommendation system, 'Just Read It', that efficiently recommends books to users based on a book that they have previously read and liked before, and provide them insight on the books that they were recommended.

The primary aim and objective of the project is to enable a quick and easy way through which people can get good book suggestions, given the important role that books play in ensuring the holistic development of an individual.

## 2. METHODOLOGY

To build a book recommendation system, we first need to find a dataset. In our project, we found 2 different datasets. Two machine learning models were built by using each dataset and we chose the second dataset as it gave the recommendations more accurately. Data cleaning and preprocessing was done to ensure that quality data would be given to the model to get a quality result. The tasks that were performed for data cleaning and processing were: checking for null or missing values, removing any erroneous values, converting the datatype of columns, and deleting unwanted columns. Once the data was cleaned, it was analysed to understand our data better. Several graphs were plotted by using the Matplotlib and seaborn libraries about the top authors and books in the dataset along with the distribution of the ratings and languages of books. The data was then prepared for the machine learning model by selecting the appropriate features from the dataset, normalizing the values to remove any bias, and converting any categorical data into numerical values. The NearestNeighbors algorithm of sklean.neighbors python module was used to building the machine learning model. The pickle module of python was used for used for serializing and de-serializing the Python object structures so that they could be used in flask and create a front end of the project. Finally, the user can type the title of a book that they have read and liked before to get 10 recommendations that will be there of their interest.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

## IMPLEMENTATION

The dataset used is 'books.csv' from the 'Goodreads' dataset that contains attributes such as the book id, author of the book, the average rating given to the book, the number of ratings given to the book, the language of the book, and the publisher of the book. As part of data cleaning, The info() function was used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index dtype and column dtypes, non-null values and memory usage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11127 entries, 0 to 11126
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   bookID              11127 non-null  int64
 1   title               11127 non-null  object
 2   authors             11127 non-null  object
 3   average_rating      11127 non-null  object
 4   isbn                11127 non-null  object
 5   isbn13              11127 non-null  object
 6   language_code       11127 non-null  object
 7   num_pages           11127 non-null  object
 8   ratings_count       11127 non-null  int64
 9   text_reviews_count  11127 non-null  int64
 10  publication_date    11127 non-null  object
 11  publisher           11127 non-null  object
 12  Unnamed: 12         4 non-null      object
dtypes: int64(3), object(10)
memory usage: 1.1+ MB
```

**Table : Concise summary of dataframe**

The describe() function was used for giving the statistical information about the numerical columns of the dataset.

| | bookID | ratings_count | text_reviews_count |
|---|---|---|---|
| count | 11127.000000 | 1.112700e+04 | 11127.000000 |
| mean | 21310.938887 | 1.793649e+04 | 541.864474 |
| std | 13093.358023 | 1.124794e+05 | 2576.174610 |
| min | 1.000000 | 0.000000e+00 | 0.000000 |
| 25% | 10287.000000 | 1.040000e+02 | 9.000000 |
| 50% | 20287.000000 | 7.450000e+02 | 47.000000 |
| 75% | 32104.500000 | 4.993500e+03 | 237.500000 |
| max | 45641.000000 | 4.597666e+06 | 94265.000000 |

**Table : Statistical description about dataset**

As seen through the above table, there were 4 records with noisy values in the num_pages column of the dataset. Num_pages represents the number of pages in a book, and these 4 records had the language of the book as a part of it. Since the number of records that had noisy values are negligible when compared to the total number of records in the dataset, they were removed from the dataset. Due to these null values, the datatype of the num_pages column was object, as soon through table, which would not allow us to us to perform numerical computations on it. Thus its datatype was converted from object to float using the astype() function. The same process was done for the same reasons for the average_rating column of the dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

```
8979      eng
4702      eng
5877      eng
3348    en-US
3131      999
7260      998
7532      997
5488      992
3137      992
7371      992
Name: num_pages, dtype: object
```

**Table : Noisy values in dataset**

The data is now clean for us to work with. This can be seen through table

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11123 entries, 4 to 11126
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   index               11123 non-null  int64
 1   bookID              11123 non-null  int64
 2   title               11123 non-null  object
 3   authors             11123 non-null  object
 4   average_rating      11123 non-null  float64
 5   isbn                11123 non-null  object
 6   isbn13              11123 non-null  object
 7   language_code       11123 non-null  object
 8   num_pages           11123 non-null  float64
 9   ratings_count       11123 non-null  int64
 10  text_reviews_count  11123 non-null  int64
 11  publication_date    11123 non-null  object
 12  publisher           11123 non-null  object
dtypes: float64(2), int64(4), object(7)
memory usage: 1.1+ MB
```

**Table : Clean dataset**

Now that the data has been cleaned up, it was analysed to understand the data better. The results of the analysis were visualized using the matplotlib and seaborn modules of python. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Seaborn is a Python data visualization library based on matplotlib. Steps were taken to prepare the data for modeling and the model was created. But since the recommendations were not being given on the basis on genres, another dataset was considered.

The second dataset is named as books_1.Best_Books_Ever.csv. It contains columns such as bookId, title, series, author, rating, description, language, isbn, genres, characters bookFormat, edition, pages, publisher, publishDate, firstPublishDate, awards numRatings, ratingsByStars, likedPercent, setting, coverImg, bbeScore, bbeVotes, and price. Similar tasks were performed for data cleaning as the first dataset.

The first graph that was plotted as a part of data exploration was the top 10 books in terms of the average rating that was given to it. The records of the dataset were first sorted in the descending order of the rating given to the books using the sort_values() function. The first 10 records were retrieved from this sorted list that gave us the top 10 books that had the highest rating given to them. A bar graph was then plotted by taking the rating on the x axis and the title of the book on the y axis. But since there might be a possibility that only a single user had read the book and must have given a high rating to it, this does not give a true picture of the best books present.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

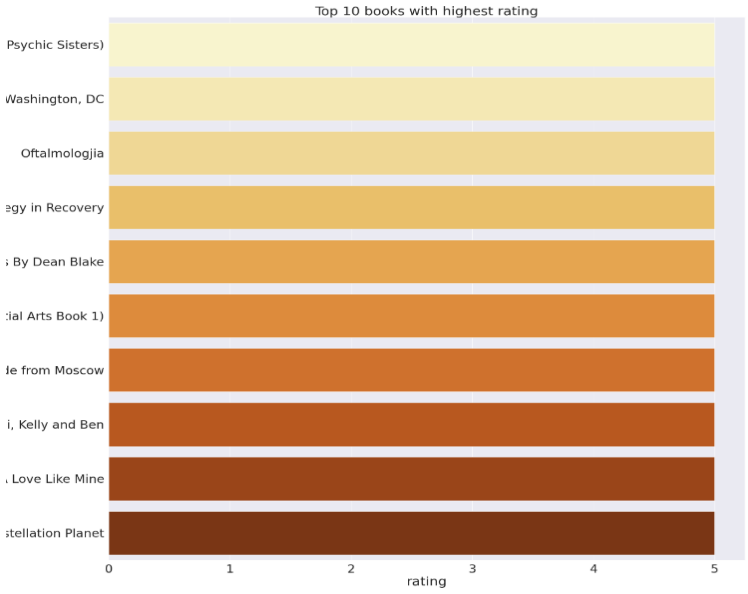



**Fig : Best books in terms of rating**

This plot uses the same methodology as the above graph, but instead of sorting the values in the descending order of their rating, it has been sorted in the descending order of the number of ratings the book received. But even here, there is no way of knowing if all the ratings given to the books are good ratings. There is a possibility that a book received many, but low ratings. Thus this is also not a very good way to know the best books.

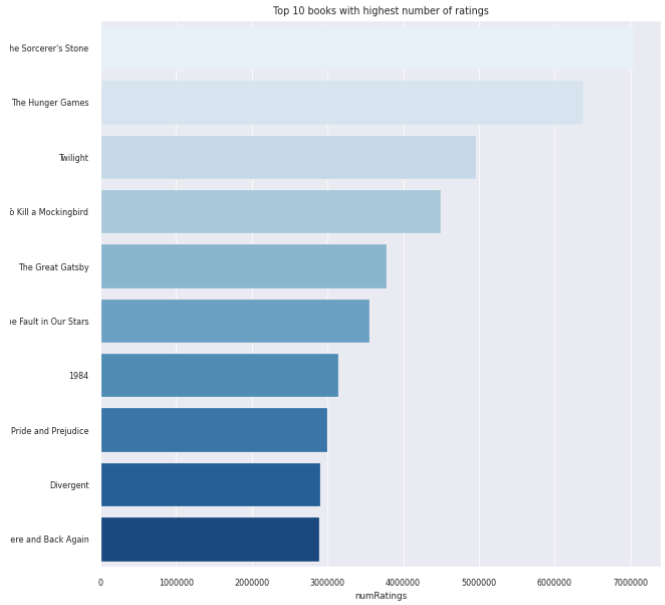



**Fig : Best books in terms of the number of ratings**

By sorting the values in the descending order of the number of ratings first, retrieving the top ten books that have received the most number of ratings, and then sorting it on the basis of the rating given to it, we can get the best books. These books have received many number of high ratings.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

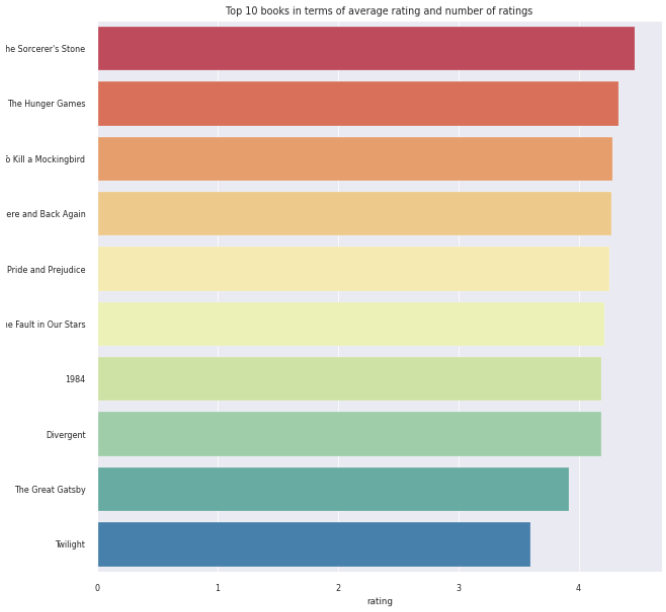



**Fig : Best books**

To know the top ten authors who have written the most number of books, the dataset is grouped on the author name using the groupby() function. The number of books written by each author are then counted using the count() function. Then the top 10 values are retrieved, giving us the authors who have written the most number of books. But writing many books doesn't necessarily indicate if the author is really good or not. An author may have written many low rated books.

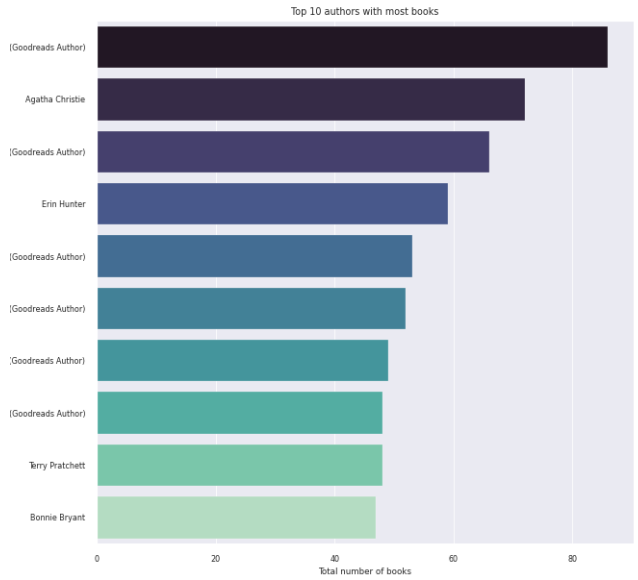



**Fig : Authors with most written books**

To find the authors who have written the highest rated books, the same methodology is used as above, except that instead of counting the number of books an author has written, the mean of all the ratings of the books written by an author is calculated. But even this way is not the best for finding out the best author since the author might have written only a single high rated book.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

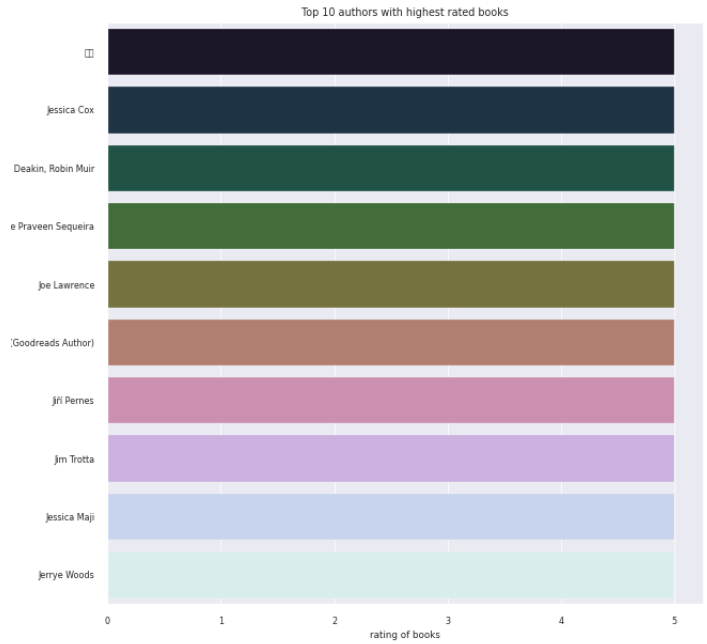



**Fig : Authors with highest rated books**

To find out the top 10 best authors, we first consider only those records which have the authors who have written the most number of books. We group this subset of the dataset on the author names. We then find the mean of the ratings given to the books written by an author. The values are sorted in the descending order of the mean and the top 10 authors are retrieved from this sorted list. This gives us the top 10 authors who have written many high rated books.

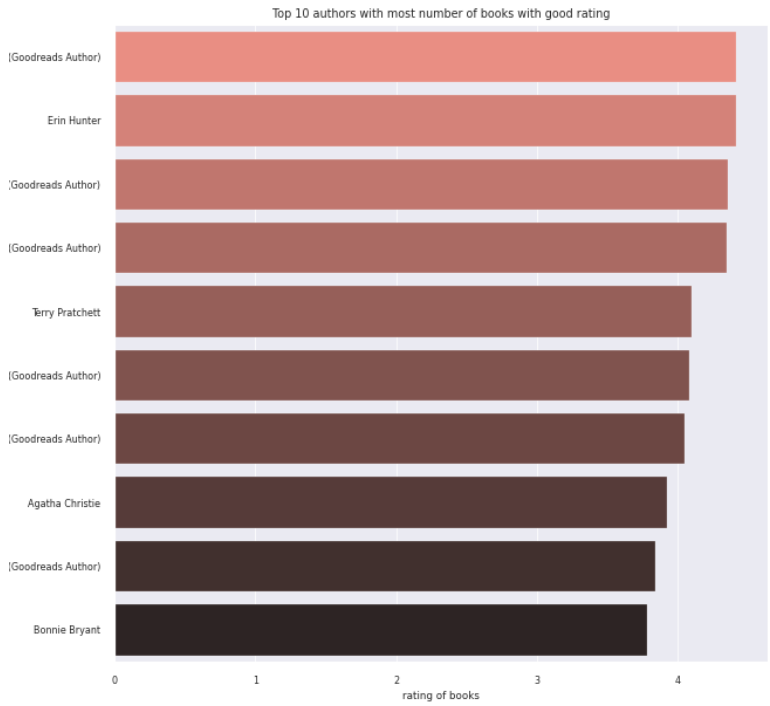



**Fig : Best authors**

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

To know the average distribution of the ratings given to books, a distplot is plotted. A distplot allows you to plot a histogram with a line through it. We can observe that most of the books have been given a rating between 3 and 5 and that there are very few books that have received a very low rating.

## 4 RESULTS

A user defined function bookRecom is written that takes in the title of the book as a parameter. The bookid of this book is taken from the dataset. The set of 10 books that are the most similar to the book passed as a parameter are found out by using idlist. Dist and idlist is returned by model.kneighbors. dist and idlist indicate the distances to the neighbors of each point and their indices. The titles of the recommended books are appended to a list named book_list and this list is returned. As seen through fig (A) and (B), the results from dataset 2 were more accurate when compared to the results of dataset 1. Thus dataset 2 was finally used for the project and its front end.

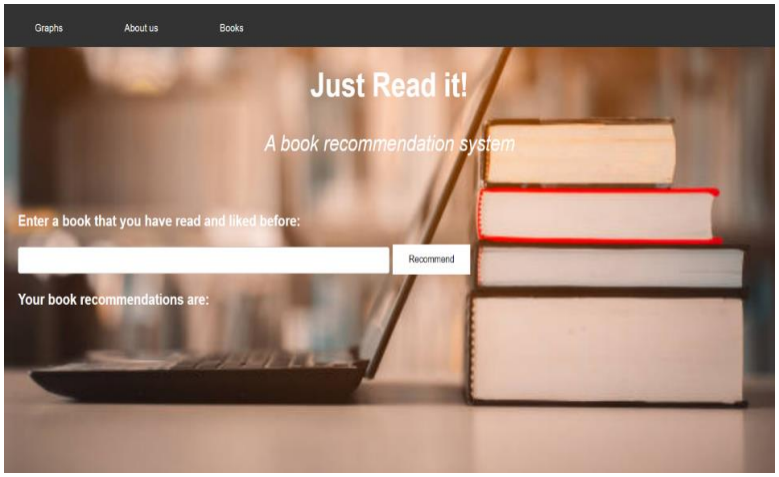



**Fig : Home page**

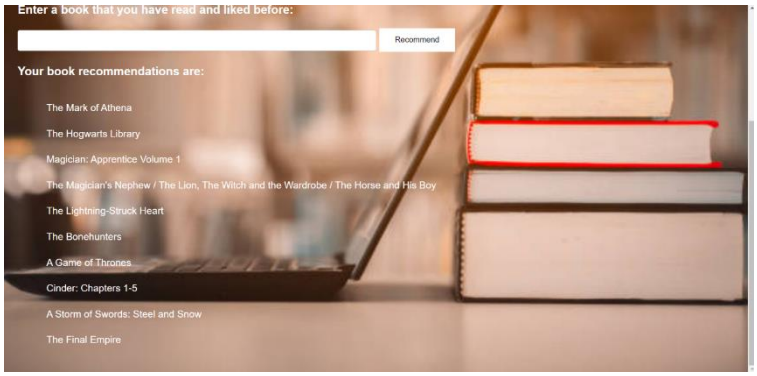



**Fig : Getting book recommendations**

Once the recommendations have been generated, the user can click on any of the recommended books to know more information about the book. As seen through fig 5.6, this information includes the title of the book, its author, the list of genres it belongs to, its description, the rating it received and the awards it won. By gaining access to this information, the user can better understand why this application recommended them a certain book. This provides more insight to the user, which is not provided in other book recommendation system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11602

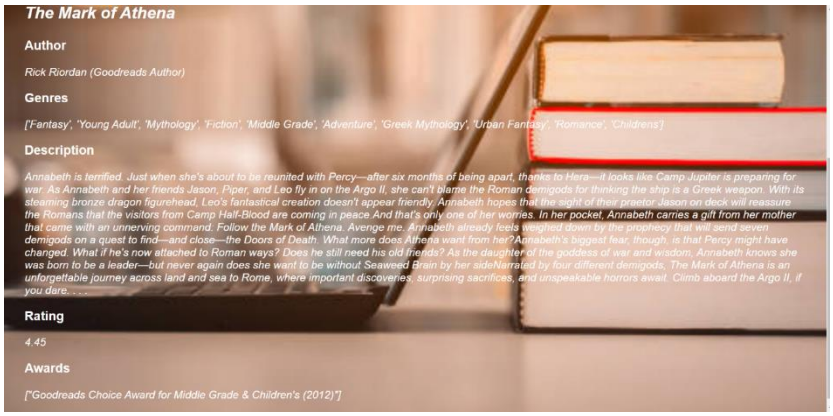



**Fig : Recommended book details**

Apart from this major feature, 'Just Read It!', also allows the user to view the graphs that were plotted as a part of the exploratory analysis of the dataset to know the top 10 best books, and authors. They can also view the distribution of the ratings given to the books, and the most common languages the books are written in.

## 5 CONCLUSION

From the above stated information, it can be concluded that this book recommendation system provides a quick and easy way through which people can quickly get good book suggestions. Knowing how books are important, and the fact that the system can be accessed easily on devices, after it has been deployed, 'Just Read It!' can reach a wider range of audience, which creates a larger impact. Furthermore, since the application does not require a login and does not need the user to build an elaborate profile to get recommendations, the users can get accurate recommendations immediately. The major limitation of this project is that the system can only give results if the user enters a book that is already present in our dataset. We would like to include a search bar that will make it easy for the user to search for a book in our dataset. In addition, we would like 'Just Read It!' to also be a mobile application to increase the accessibility of the system. In machine learning, in the event that we will in general get familiar with a connection between certain highlights and a paired component then we utilize a sigmoid function at the yield layer [8]. In the future, the logistic function can be applied to these types of applications. Furthermore, we would like for people to be able to add new books to our dataset so that our system will be able to recommend the newest of the newest books. After implementing all these features, we would like to fully deploy the project so that it is accessible to everyone through the internet.

## REFERENCES

[1] The Deep Learning Workshop: Learn the Skills You Need to Develop Your Own Next-generation Deep Learning Models with TensorFlow and Keras by Mirza Rahim Baig, Nipun Sadvilkar, and Thomas V. Joseph.
[2]https://www.analyticsvidhya.com/blog/2021/08/a-friendly-guide-to-nlp-bag-of-words-with-python-example/
[3] https://www.kaggle.com/code/ayushmi77al/language-detection-nlp
[4]https://towardsdatascience.com/language-translation-using-python-bd8020772ccc
[5] https://scikit-learn.org/stable/modules/naive_bayes.html
[6] https://www.geeksforgeeks.org/naive-bayes-classifiers/
[7] https://nanonets.com/blog/ocr-with-tesseract/
[8] GNR Prasad, " Implementation of sigmoid function in logistic regression", International Journal of Computer Engineering and Applications, Volume- XIV, Issue - Special Issue, June 2020, ISSN 2321-3469
[9] https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/
[10] https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c