UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

Interim Report

3204 - Individual Project

Project Name	**:**	Content and Collaborative based Sinhala Book Recommendation System
Supervisor	**:**	Dr. M. G. N. A. S. Fernando

Student Name	**:**	P. N. C. Perera
Index No	**:**	19440677
Registration No	**:**	2019/MCS/067

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Term	Definition
AI	Artificial Intelligence
CTR	Click Through Rate
GUI	Graphical User Interface
KNN	K Nearest Neighbors
IDE	Integrated Development Environment
MS	Microsoft
NLP	Natural Language Processing
OS	Operating System
SDLC	Software Development Life Cycle

1. Introduction

1.1. Chapter Overview

This chapter provides a foreword for the project in terms of background study, problem domain, the main aim, objectives, scope and activities that will be carried out towards the completion of the research. Finally, the chapter concludes with an overview on how the other chapters of the document fit into the project context.

1.2. Background

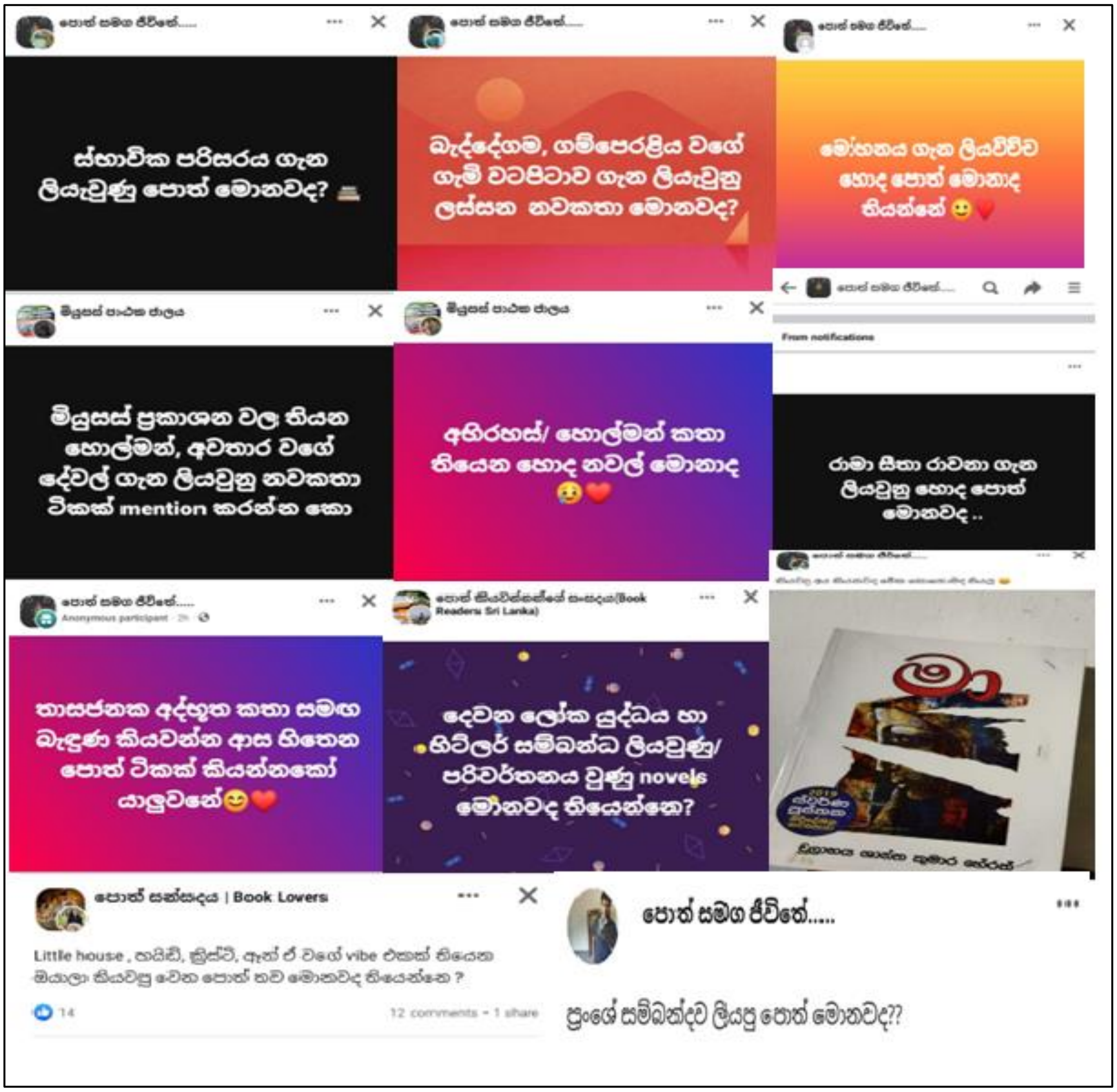
From our childhood, everyone has heard that “Reading makes a man perfect”. People acquire the knowledge by reading a variety of materials. These materials could be a book, an internet article, a newspaper, a magazine, or even a piece of paper, and the gain knowledge by reading these materials is intense. People who read a lot tend to know more about life and are smarter when making decisions and handling difficult situations. (Marappan, 2022) It may not be possible for the reader to “know it all,” but a lot of reading brings man close to perfection. Most of them like to read books as a hobby because it imagines readers' own movie in their mind rather than watching a movie directed by someone.

In today’s world, time has more value and the researchers have no much time to spend on searching for the right articles according to their research domain. (Murali et al., 2019).

Normally, book readers select books by reading some random pages or asking someone to recommended any book. When reading that book, if he finds that the book is not interesting, he will not read any book after that. therefore, it is better to suggest books that he is interested in. With the increase in library collections, it is difficult for readers to quickly find the books they want when choosing books. It is also difficult for readers to find Sinhala books of interest in a short period of time in the face of various bibliographies. Therefore, the user experience of the traditional library borrowing method is poor (Dhanda and Verma, 2016). Due to the Covid-19 pandemic situation and the geographical barriers also it becomes a tremendous challenge for readers (Sarma et al., 2021) to find a relevant book as they do not like to go out and spend time searching books of their preference.

1.3. Motivation

When we navigate through social media specially in Facebook there are so many groups available for almost everything. If you are living in an area, there is a group for that area, if you have an aqua car, there is a group created for aqua car owners. The benefits of such group are you can learn many things and if you have any question you can ask from the group and get it clarified. For Sinhala book readers also, there are so many groups available in the Facebook. You can share what your thought on a specific book or you can see what are latest books released through the groups if you follow those groups. One thing I have noticed is many people asking I have read this particular book; can anyone suggest similar type of books. And some asking what are the books related to Sri Lankan history or related to world war. Some other have asked whether the book is good to read with uploading an image of the book. When considering these three scenarios I thought like it is better to have a system that displays what other users' thoughts about a book and how much of rate could be given to the book and what are other similar books. It gives the motivation to initiate kind of such system for Sinhala book readers.



පොත් සමග ජීවිතේ.....

ස්භාවික පරිසරය ගැන ලියැවුණු පොත් මොනවද? 📚

පොත් සමග ජීවිතේ.....

බැද්දේගම, ගම්පෙරළිය වගේ ගැමි වටපිටාව ගැන ලියැවුනු ලස්සන නවකතා මොනවද?

පොත් සමග ජීවිතේ.....

මෝහනය ගැන ලියවිච්ච හොද පොත් මොනාද තියන්නේ 🙂❤️

මියුසස් පාඨක ජාලය

මියුසස් ප්‍රකාශන වල තියන හොල්මන්, අවතාර වගේ දේවල් ගැන ලියවුනු නවකතා ටිකක් mention කරන්න කො

මියුසස් පාඨක ජාලය

අභිරහස්/ හොල්මන් කතා තියෙන හොද නවල් මොනාද 😢❤️

පොත් සමග ජීවිතේ.....

From notifications

රාමා සීතා රාවනා ගැන ලියවුනු හොද පොත් මොනවද ..

පොත් සමග ජීවිතේ.....

Anonymous participant · 2h

භාසජනක අද්භූත කතා සමඟ බැඳුණ කියවන්න ආස හිතෙන පොත් ටිකක් කියන්නකෝ යාලුවනේ😊❤️

පොත් කියවන්නන්ගේ සංසදය(Book Readers Sri Lanka)

දෙවන ලෝක යුද්ධය හා හිට්ලර් සම්බන්ධ ලියවුණු/ පරිවර්තනය වුණු novels මොනවද තියෙන්නෙ?

පොත් සමග ජීවිතේ.....

පොත් සන්සදය | Book Lovers

Little house , හයිඩි, ක්‍රිස්ටි, ඈන් ඒ වගේ vibe එකක් තියෙන ඔයාලා කියවපු වෙන පොත් තව මොනවද තියෙන්නෙ ?

14

12 comments • 1 share

පොත් සමග ජීවිතේ.....

පුංශේ සම්බන්දව ලියපු පොත් මොනවද??

Figure 1: Questions Users ask from groups

1.4. Problem Domain

Most organizations like Amazon, Ebay have implemented their recommendation system when users buy products online. But almost all the websites are not developed for the buyer's interest; the organizations force add-on sales to buyers by recommending unnecessary and irrelevant products (Sarma et al., 2021) If book recommendation point of view for an instance, if a user has read a book named 'Madol Duwa', he would like to read similar books and there is no Sinhala book recommendation system to address this problem. Additionally, some members of readers groups on Facebook have problems like, is this book good or I have read this book and are there any similar kind of books like the mentioned book?

Many personal book recommendation systems have emerged to conduct effective search based on user rating and interest.

This paper proposed an effective Sinhala book recommendation system for online users that rated a book list using the content and collaboration (hybrid) method. The solution could be used by all Sinhala book readers to find interesting or suitable books without wasting time or money. Authors also could use the system to have an idea of what kind of books readers rate and are interested more and write books accordingly.

1.5. Aim

The main aim of the research is to analyze, design, implement and evaluate an accurate recommendation system related to Sinhala books using content and collaboration algorithms with attractive user-friendly interface which display the searched book details along with already given reviews and suggesting a list of recommended books. The main aim can be further divided to three sub aims as;

- Input Data will be the dataset collected from readers.
- Preprocess by removing null values and unwanted data then apply the content and collaborative algorithm respectively
- Final output data will be displayed in the system.

1.6. Objective

The final outcome of this project is helping Sinhala book readers to find correct and recommend books based on their preferences using content and collaboration algorithm. There is no exact formula for determining how much data would be enough for a recommendation system. Even though capturing many data sets, ends up with manageable data sets after preprocessing and removing null values from the collection list as most online dataset contains parsed data.

The main goal of this research will be achieved by targeting the following objectives.

1. To Collect selected Sinhala book details like title, author, publisher, description, image url and keywords. Online book stores will be used for collecting the details of books.

2. To produce a data set containing user details, selected Sinhala books and rates given by users for those books.

3. To implement web application along with login and registration features.
4. To recommend ten books based on specific field of interest using Content and Collaborative (Hybrid) methodology.
5. To determine the categories preferred by readers so that it motivates authors to write books as per the user's preference.
6. To Increase the number of book readers by recommending books according to their preferences.

1.7. Scope

The scope of the project can be defined as bellow.

1. Sharing a google form containing selected books along with authors and collecting user rates and reviews for those books based on previous readers experience.
2. Applying sentimental analysis for the reviews collected from above and assigning a new rate.
3. Provide facilities for readers to login or register to the system with their email id which is unique.
4. Provide facilities for readers to search a specific book in the repository and it will display the details such as author, publisher and image along with the reviews and rates given by users.
5. Additionally, the system will display a list of recommended Sinhala books based on users' rates and reviews of specific interested field. Content based and Collaborative based (Hybrid) approach will be taken place in order to recommend books.
6. Only Sinhala books are recommended and it is based on users' rates and reviews as well as keywords provided for selected books

We can use library to collect book details, but we will not be able to collect user reviews and rates for selected books. I am using online book store to collect book details but I preferred to get user reviews and rates from users themselves so that they are aware that they have provided the information for the application rather than just coping from the online without their awareness. Until now, around 4000 records have been collected and targeting to collect around 7000 records for the research.

1.8. Resource Requirement

In order to implement and execute the application, following hardware and software requirements should be satisfied.

1.8.1. Hardware requirement

- A Laptop or desktop with core i3 or above processor
- At least 4GB Ram
- At least 30GB

1.8.2. Software requirement

- Python latest version – 3.12.0
- VS code as IDE for implementation and execute the application
- MS Excel and Notepad ++ for viewing and manipulating data
- Stable internet connection for downloading relevant libraries.
- GitHub for storing images and implemented code.

1.9. Chapter Walkthrough

The outline of the chapters are as follows.

1.9.1. Chapter 02: Literature Survey

This chapter will discuss about the review, conducted on the proposed project. It will extensively describe on the stakeholders, the problem, existing solutions, methodologies, and approaches along with their benefits and limitations.

1.9.2. Chapter 03: Methodology

This chapter will discuss about the methodology to be used to implement the solution. The stakeholders, main technology, libraries, prioritized items, how the collected data is analyzed and the how the architecture of the system will be organized will be in detailed discussed. Furthermore, why the selected technology is more suitable than other existing technologies will be clarified.

1.9.3. Chapter 04: Implementation

This chapter covers the implementation stage of the project. Algorithms used and challenges faced and how they are resolved will be discussed in this phrase. Screen shots and code segments for some selected functionalities are also provided to facilitate easier understanding and manipulating over the project implementation.

1.9.4. Chapter 05: Evaluation and Results

The evaluation chapter provides how the results are evaluated based on the feedback collected from Domain experts in this projects Authors. The project will be shown to them and get the feedback for the evaluation. Other than that validation methods will be used to further evaluate the accuracy of the system.

1.9.5. Chapter 06: Conclusion and Future work

The objectives that were able to be successfully achieved will be discussed in conclusion chapter. The challengers and the limitation of implemented system will be highlighted in order for someone to enhance the system.

1.10. Chapter Summary

The chapter began with explanation on background and problem domain of the system. Although many applications have been developed for book recommending systems, most of them are related to English books. Proper applications that satisfy all the requirements with user satisfaction were limited. The main approach is to make an application that help all Sinhala book readers to recommend Sinhala books based on their preference. A goal followed by objectives was defined to make the effort to be success.

2. Literature Review

2.1 Chapter Overview

This main aim of the chapter is, study the existing systems implement for Book recommendation system and find out the limitations. As per the study there are two main models that can be used for the system named as Collaboration filtering and Content based filtering. The chapter concludes by explaining the hybrid model which is the combinations of Content and collaborate filter.

2.2 Problem Domain

In today's world recommendation systems plays significant role for user to find items which they prefer. When you buy any product, it suggests similar items or items which customers buy along with the item you bought. When it comes to book recommendation it help readers to find similar books or books read by other users who has similar preference as you. There are multiple recommendation systems have been implemented for English book. Implement a recommendation system for Sinhala books is kind of a challenge as there are no dataset can be found in many datasets provides. Following are the key research areas to be focused in order to complete the application successfully.

2.2.1 Natural Language Processing (NLP)

Natural Language Processing refers to a branch of Artificial Intelligence (AI) and it gives computers the ability to understand text as well as spoken words which basically human language and act upon commands. NLP has existed for more than 50 years and has roots in the field of linguistics ("What is Natural Language Processing?," n.d.). As you all know 'Siri' in Apple utilize NLP to respond

There are NLP based applications available as follows which understand human text and voice and help computer to make sense of what it to be performed.

2.2.1.1 Sentiment Analysis

This is the process of determining the sentiment or emotional behind a text. As an example, if items have reviews the algorithm can be used to determine how many of reviews given is positive, negative or neutral. It helps to increase the productivity and quality of the item.

2.2.1.2 Machine Translation

This is the process of translating to different language automatically without human intervention. The input is from a different language and the translated to expected language as output. Google Translate is the main widely available technology of NLP which helps users to communicate without language barrier.

2.2.1.3 Named Entry Recognition

The main aim of Named Entry Recognition is to extract phrases in a piece of text into predefined categories such as locations, personal names, organizations and quantities. The input of the model takes as text and the output will be the various named entities along with their start and end positions.

2.2.1.4 Spam Detection

It is not believable that Spam detection could be implemented via NLP technology. But it is identified the best spam detection technologies use text classification capabilities of NLP to scan emails that often indicates spam or phishing. Spam detectors takes email text as input along with other parameters like title, company name, senders name and find they are spam and placed to a specific spam folder.

2.2.1.5 Grammatical Error Correction

Grammatical error correction model encodes grammatical rules to correct the grammar within text. Most popular word processing systems like MS Word and Online grammar checkers like Grammarly use these kind of systems to provide a better writing experience to their users.

2.2.2 Machine Learning Models

Machine learning model is a program which find a pattern from a dataset and make decisions. It helps to train the machine to a model and get the output for a given input and behave like a human but in fastest way. Many models are available in the world which helps human to perform many activities such as NLP, image recognition, NLP recognize the sentence and categorize the while image recognition identify objects like car, dog, computer. The machine learning model perform above NLP and image recognition with train the machine with large amount of dataset. While training the algorithm used to find a pattern or results from the dataset being provided. The output or the pattern usually called as machine learning model.

All machine learning models are break down into two main categories as supervised and unsupervised. Supervise model further categorized as regression and classification.

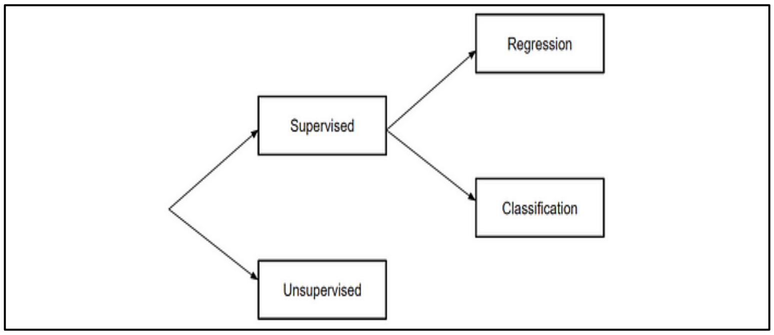


Figure 2: Machine Learning models

2.2.2.1 Supervised

In this model machine is trained with labelled data which means some input data tagged with proper output. After train the model the machine will predict the out for any input data provide. The training data input to the machine work as a supervisor who teaches the machine to predict the output. It can be used to real world applications such as image classification, risk assessment and spam filtering.

2.2.2.1.1 Regression

If there is a relationship between input field and output field, regression algorithm can be used. The algorithm is well supported to predict continuous fields like market trends, whether forecasting. Linear regression, non-linear regression, regression trees, polynomial regression are some of regression algorithms.

2.2.2.1.2 Classification

When the output variable can be categorized which means there are two main classes like, yes-no, male-female, true-false, the classification algorithm can be used. Random forest, Decision tree, Logistic regression and support vector machine are some of classification algorithms.

2.2.2.2 Unsupervised

On the other hand, unsupervised learning is a machine learning techniques models are not able to used supervised simply labelled data. In this model, it needs to find hidden patterns by itself from the data provided. The model needs to be trained with unlabeled data and act without any supervision. The unsupervised model can not be applied to regression or classification problem as we just have input data without output data. The main goal is to find any structure of the dataset, group them based on similarities and apply an algorithm to find similar items. K-means clustering, K-nearest neighbors, Neural Network, Apriori are some of algorithm of unsupervised machine learning.

2.3 Existing systems

According to the research (Sarma et al., 2021) they proposed an effective system for recommending books to online users that used the clustering approach to rate books and then found book's similarity to suggest new books. The data set were collected from Good readers book repository of Kaggle for the research. To measure distance and determine similarity between book groups, the suggested system uses the K-means Cosine Distance function and the Cosine Similarity function. This study presented a clustering-based book suggestion framework that utilizes various methodologies including collaborative filtering, hybrid, content-based, knowledge based, and utility-based filtering in order to achieve the highest accuracy. Since the accuracy is a crucial aspect of evaluation, they calculate precision, sensitivity, specificity and F1 score and according to the value they have evaluate the system.

The research (Wadikar et al., 2020) proposes a platform that employs a Convolutional Neural Network (CNN) to recommend books based on two approaches. First approach is, using text processing and the second one is using image classification. In text processing approach, it takes the input from the user as a text and process it. The required data set were taken by performing web scrapped from websites like Amazon and Flipkart and processed separately then converted to csv files. In image classification, a book cover image need to be uploaded and the results are displayed accordingly. The book cover images data set were taken also using web scrapped. They use cosine similarity measure to find the similar books related to the subject or image from the sites. The researches have tried to improvised and modified the traditional recommending system and filtering techniques like content based or collaborative based have not been used in the system. They conducted the experiment to list the similar books from Amazon and Flipkart. But the evaluation and validation process have not been presented in the research.

The study (Ijaz, n.d.) propose how to use machine learning algorithms K-Nearest Neighbor and matrix factorization for the recommendation system. It first gathers the rankings or a preference of books provided by multiple users and then suggests books to different individuals based on various previous tastes and preferences. K-Means Multipathing together with K-Nearest Neighbor is applied on the BX dataset which are collected from the Kaggle official website to achieve the greatest-optimized outcome. To calculate the accuracy of the system predictions it used an ordinary statistical metric named root mean square error (RMSQ). RMSE is a measurement of the variation between the user's real books ratings and the predicted rating for the same books. If the lower the RMSE, the more acceptable the model. An RMSE of zero means the model is absolutely guess the user ratings.

The research (Tian et al., 2019) designs a personalized recommendation system for college library based on hybrid recommendation algorithm which combines both collaborative filtering and content based filtering. According to the algorithm it first classified the readers, then establish user-item scoring matrix, then construct vector space model and finally calculate the similarity among users. The experimental data were collected from Library of Inner Mongolia University of Technology. Since the sparsity is a common problem in Collaborative filtering, the research use clustering to alleviate it. In order to verify the effectiveness of the system it performs the calculation of precision for single algorithm and hybrid algorithm respectively and compared.

The research (Shah, 2019) explain the algorithm collaborative filtering with memory based and model based. A user can either enter rate or sentence which ultimately calculate rate by determining the nature of the sentence using natural language processing. Even the paper discussed various methods that can be used to build a recommendation system, item-based collaboration approach was used for the implementation. It uses "goodbooks10k" dataset in Kaggle for training and test the accuracy and Mean Absolute Error (MEA) is used to determine the quality of the system.

According to (Mercy Milcah Y et al., 2020) they demonstrate a recommendation model that involves Metrix Factorization as a collaborative filtering solution and with further application of artificial intelligence over the previously obtained results from collaborative filtering. The paper presents six types of recommending systems that can be used by user friendly resources or websites or personalized recommending systems. They are collaborative, content based, demographic based, Utility based, Knowledge based and hybrid recommender systems. They also addressed the collaborative problems such as spacity and cold start by combining the system with content based and make it as hybrid.

(Wang et al., 2018) implemented Content based recommend system which gets the information about the scientific article and suggest most appropriate conferences or journals. After desiding the mode of feature acquisition, the content-based filtering approach was used to predict through softmax regression which is more generic approach of logistic regression.

It provides two kind of recommendation results. The first method is 'One class' and it recommends only one journal or conference. The other method is 'Three class' and it recommends three candidate journals or conferences. For the evaluation Chi-square, MI and IG are implemented to make comparisons for feature selection.

Collaborative algorithm is the most desired and widely implemented as well as one of most matured algorithms that are available in the industry. It is mainly based on the assumption that users who liked items in the past will like in the future. And also, users would like similar kind of items as they wanted in the past. The approach builds the model based on rating given by other users for a particular book and users past behavior towards the system. One of the drawbacks of this algorithm is that it needs a tremendous amount of real time user data. Other than that sparsity, cold start and scalability are some of limitation of the approach. But user-item scoring matrix and clustering can be used to alleviate the sparsity problem as it allows re grouping all the books based on the rating and user preference datasets.

Content based algorithm is based on description of the item and the profile of the user's preference. It compares various candidate items with the books previously borrowed or rated by the user and the best matching books will be recommended. The method can be used when a new user login to the system and search for a particular book. The according to the category of the book, a recommended list can be displayed. Some of the draw backs are, it filters the entire set of books from the data set based on the content thus it hinders the performance and it does not help to find out the content quality of the book and it has low accuracy.

Combining any of two types of recommending systems is known as Hybrid recommender system. This is the most demanded method used by many industries as it combines the strength of more than two types of recommending systems while eliminate weaknesses that were there when only one recommended system is used. Since Collaborative based and content-based filtering algorithm having limitations when they used respectively, Hybrid algorithm will be used in proposed system in order to produce efficient and effective book recommendation

Even though several research papers have been published related to book recommendation system, All of them related to English books and no research paper was found related to Sinhala Book Recommendation.

2.4 Chapter Summary

The literature review chapter contains what are the existing system available along with tools and technologies used in these systems. As explained, most of the system have implemented mainly either collaboration based or content based specially for English books. While implementing these systems, it is highlighted the limitations of those systems so that the limitations can be addressed in proposed system.

3. Design and Methodology

3.1 Chapter Overview

Earlier Literature review chapter helped to identified what are similar systems available in the world and what are the limitations in those applications. The main focus in the chapter is how the problem is analyzed and identify the methodologies that can be used to implement the system. Further, architectural diagram will be explained along with why hybrid-based application is focused rather than one particular model will be discussed in details.

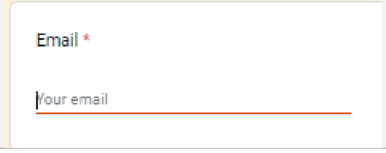
3.2 Software Design Approach

Since the requirement is clearly understand and it is not changing every time, waterfall method can be used as a design methodology. The research is conduced based on the method and the progress of each phrase will be explained

3.3 Data Set

For any machine learning application, a dataset plays a significant role. While going through some initial process I was trying to find a data set from popular dataset providers like Kaggle. All those data providers have lot of datasets related to book recommendation system. But the limit is all of them related to English books. There were no data set for Sinhala books. Therefore, a google form is created to collect the dataset. The google form is shared in all book readers groups in Facebook and able to collect fair amount of data which can be used to build a model.

Following fields are listed in the form to be provided by the book readers.

Input Field	Usage	Google form
Email	Unique id to differentiate the user	

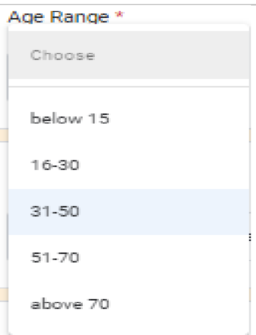
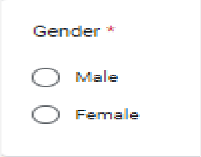
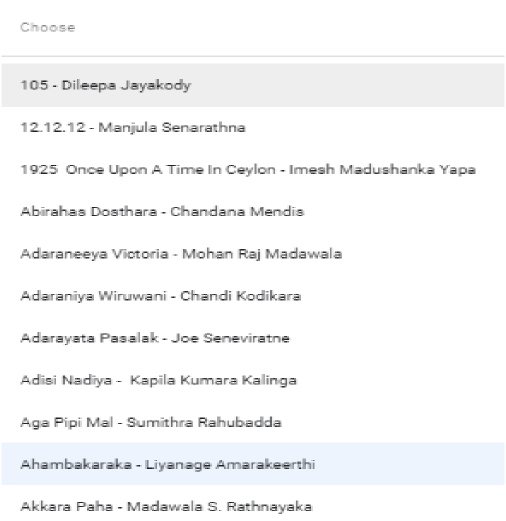
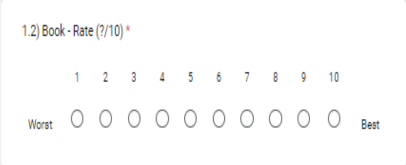
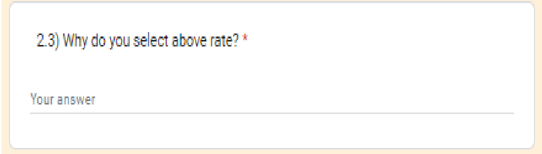
Age Range	Drop down list with age range	 Age Range * Choose below 15 16-30 31-50 51-70 above 70
Gender	Radio button with two fields of male and female	 Gender * ○ Male ○ Female
Book	This will be a dropdown list which contains 304 books	 Choose 105 - Dileepa Jayakody 12.12.12 - Manjula Senarathna 1925 Once Upon A Time In Ceylon - Imesh Madushanka Yapa Abirahas Dosthara - Chandana Mendis Adaraneeya Victoria - Mohan Raj Madawala Adaraniya Wiruwani - Chandi Kodikara Adarayata Pasalak - Joe Seneviratne Adisi Nadiya - Kapila Kumara Kalinga Aga Pipi Mal - Sumithra Rahubadda Ahambakaraka - Liyanage Amarakeerthi Akkara Paha - Madawala S. Rathnayaka
Rate	This is a range field from 1 to 10	 1.2) Book - Rate (?/10) * 1 2 3 4 5 6 7 8 9 10 Worst ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Best
Reason	The input will be considered as a review and taken for sentiment analysis	 2.3) Why do you select above rate? * Your answer

Table 1: Google form with fields details

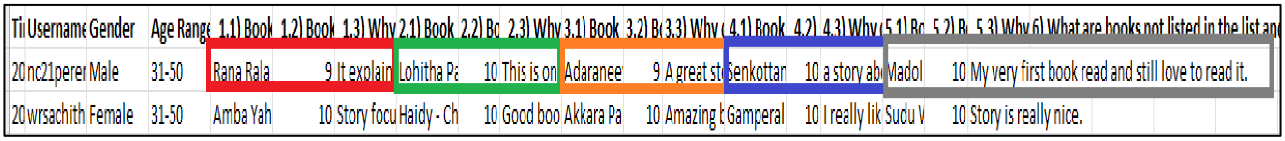
Initial stage of the implementation, an input text filed is given to provide books. But readers entered different data for same value for example, ‘Madol duwa’, ‘Madol Duuwa’. Therefore, a list is created for readers to be selected. In additionaly, if any preferred books are not listed, those books can be included at the end of the form so that the books can be added in to the list by admins in future.

3.3.1 Validate Data Set

So far more than 800 users have entered data for the google sheet provided and shared in Facebook groups. In the form there are fields like rate and review. If the user entered higher rate and give positive review, it means the data properly entered. If the user entered higher value and provide a negative review, these rates of the data should be considered by comparing a review rate which could be calculated for the reviews via sentiment analysis method. Then the rate given by the user and the rate calculated via sentiment analysis can be compared and made sure the dataset is suitable for the system.

3.3.2 Format Data Set

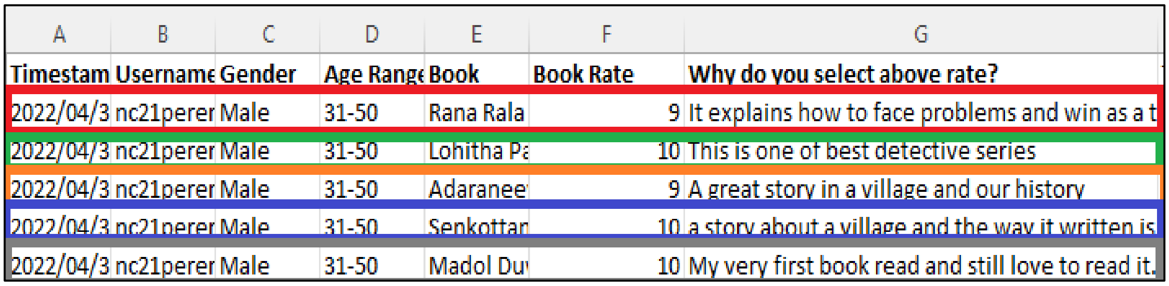
When we analyze the dataset, it contains one user with multiple columns which means the selected books are repeated as bellow.



Ti	Username	Gender	Age Range	1.1) Book	1.2) Book	1.3) Why	2.1) Book	2.2) Bo	2.3) Why	3.1) Book	3.2) B(	3.3) Why (	4.1) Book	4.2)	4.3) Why (	5.1) Bo	5.2) B	5.3) Why	6) What are books not listed in the list and
20	nc21perer	Male	31-50	Rana Rala	9	It explain	Lohitha Pa	10	This is on	Adaranee'	9	A great sto	Senkottan	10	a story abo	Madol	10	My very first book read and still love to read it.	
20	wrsachith	Female	31-50	Amba Yah	10	Story focu	Haidy - Ch	10	Good boo	Akkara Pa	10	Amazing b	Gamperal	10	I really lik	Sudu V	10	Story is really nice.	

Figure 3: Collected data sample

In Order to apply and algorithm data should be formatted as below. The data set should be preprocessed and python libraries could be used for the below format.



A	B	C	D	E	F	G
Timestam	Username	Gender	Age Range	Book	Book Rate	Why do you select above rate?
2022/04/3	nc21perer	Male	31-50	Rana Rala	9	It explains how to face problems and win as a t
2022/04/3	nc21perer	Male	31-50	Lohitha Pa	10	This is one of best detective series
2022/04/3	nc21perer	Male	31-50	Adaranee'	9	A great story in a village and our history
2022/04/3	nc21perer	Male	31-50	Senkottan	10	a story about a village and the way it written is
2022/04/3	nc21perer	Male	31-50	Madol Du	10	My very first book read and still love to read it.

Figure 4: Expected format to apply algorithms

3.4 Preprocessing

Every data set provided by many providers like Kaggle, to be preprocessed and captures the data we required for the algorithm. Since the data combines with stop words, numbers, links which do not have proper meaning, these to be removed. But as the first step, all reviews which were written in Sinhala to be converted to English.

3.4.1 Convert Sinhala review to English

Once the google form is shared user is able to enter data in both English and Sinhala. Most data were entered via English but few ware entered via Sinhala. Since the data review entered via Sinhala is considerable amount, we are not going to ignore but applying google translator by python on Sinhala reviews can be converted to English so that we can utilize those reviews as well for the machine learning algorithm.

3.4.2 Sentiment Analysis for reviews

The reviews provided by the users are valuable to the system as these data will be displayed along with the book details when a user search a particular book. Additionally, reviews can be used to validate the rate. The sentiment analysis process analyzes the reviewed text and give a rate which can be compared with original rate given by the user. For the model to be trained and predict the books, the mean of both rates can be taken as a reasonable rate.

3.5 Architectural Diagram

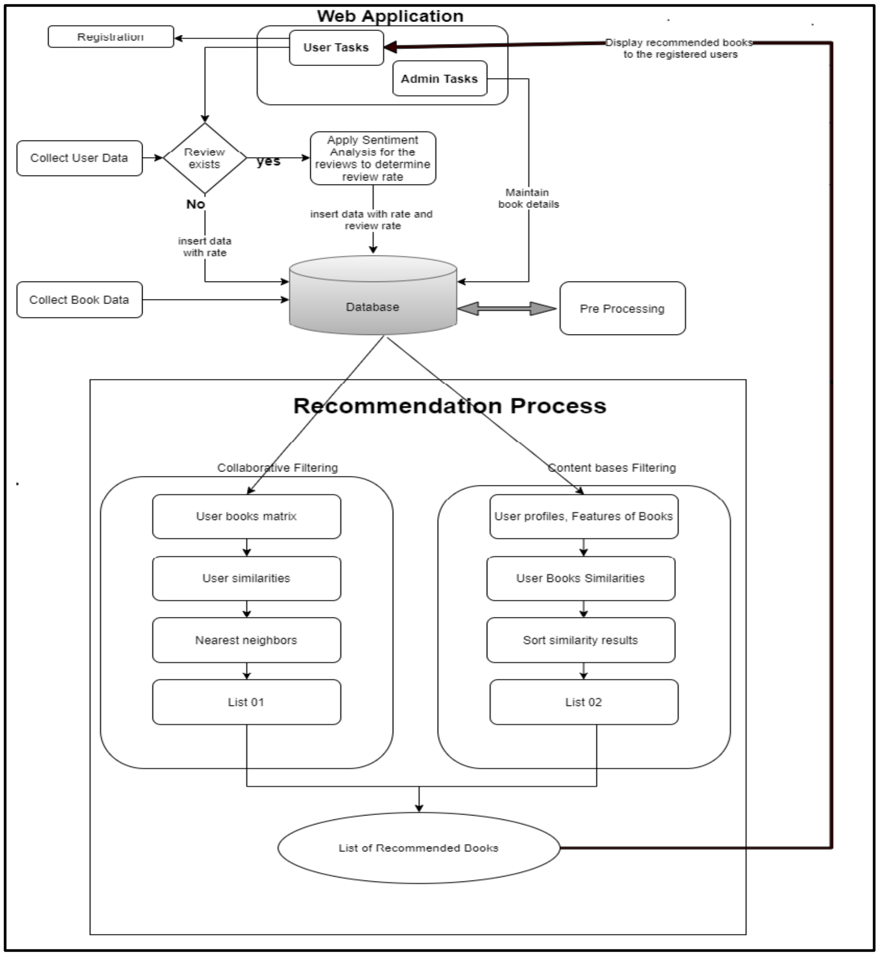


Figure 5: Architectural Diagram

According to the above architectural diagram, initially data is collected and then it needs to be categorized and store in separate 3 tables as Book_Details, User_Details and Rate_Details. All user related information will be stored in User_Details table. Additional book data like ISBN, publisher, year should be added in Book_Details table to display the details of the book when a user search for a particular book. Ratings given for a book by users are stored in Rate_Details table. If a review exists for a book, a separate review rate will be assigned for the review after completing of the sentimental analysis and include it in Rate_Details table as a separate column. For the final calculation, mean value of review rate and normal rate will be taken. If there is no review given, normal rate will be taken for the calculation.

Before starting the recommendation process, collected data need to be preprocessed in order to remove unwanted data like null rate values. And data like giving maximum rate for just one and only book should be eliminated as those kinds of books should not be recommended.

Then the process of applying the algorithms will be taken place. As per the paper (Murali et al., 2019) More than 250 research paper recommender systems were published and the quantity of research papers published every day is increasing rapidly. Thus, it needs an efficient searching and filtering mechanism to choose the quality research papers, so that the effort and time of researchers can be saved.

One of the main two methods used to implement the system is Collaborate filtering. The method is used to recommend a user with best books in their domain according to the queries and preferences based on the similarities found from other users. (Murali et al., 2019). It will find the adjacent neighbors of a customer based on the ratings given by the other users. In user based collaborative algorithm, first we need to build a matrix upon users and books with the rating given. Then cosine similarity, which is one of the techniques of K Nearest Neighbor (KNN) will be computed for each user in the matrix. The KNN is a machine learning algorithm to find clusters of similar users based on common book ratings. The cosine similarity first collects books in which is evaluated by all the users in the nearest neighbors, and then the candidate list which the target user has rated or reviewed is removed. Finally, a list of recommended books will be generated based on the similarity.

In content-based approach, it is based on the description of the book and a user's preference. The algorithm tries to recommend books which are similar to those that a user rated in the past. Initially it abstracts the features such as title, author, genre of books in the system. Then information such as books user read and the rates given will be considered to create user preference vector. Finally various candidate items are compared with the books previously rated by the user and the best matching books are recommended.

According to the researches most of them use hybrid method which combines both collaborative filtering and content-based filtering. Even though there are multiple strategies to apply the hybrid method, applying content-based and collaborative filtering separately and then combining them together will be adopted as it is more effective in book recommendation system. Therefore, the common book list which are generated from collaborative filtering and content-based filtering will be displayed as the final recommended book list.

3.6 Machine Learning Model

Even though many applications have been implemented using either collaborate based or content based, few applications were implemented using both. How the methodologies can be used in the system will be elaborate below.

3.6.1 Collaborative Filter

The data set contains user, book and rate. As the first step, the data set should be converted to used based mastics. Then cosine similarity can be used to find the recommend books.

3.6.2 Content-based Filter

The data set contains author, review and description. Removing stop words, links and numbers, then combining all together we can create a tag list. Similar books can be selected by applying cosine similarity.

3.6.3 Hybrid Approach

There are multiple ways we can apply hybrid methodology for the dataset. Finding books separately for collaborate and contend based and check common books is one approach and the other approach is, apply collaborate first and then apply content-based for filtered books. Since the first approach will be more feasible and manageable, it will be used as hybrid model.

3.7 Web Application

The web application will be implemented using python flask which is one of library for develop web applications. The basic html with css without any library will not be used as it will be an extra effect to connect with python backend. Mainly register and login page will be there and two mail roles as registered users and admin users will be maintained. Registered users will be able to view the top-rated books, popular books and recommend books where admin uses can maintain book data addition to above features where normal users perform. Top rated books list can be taken from the dataset it self by calculating the highest average rates where popular book list can by taken by calculating the number of rates given for books. For the recommendation list collaborative and contend based filtering will be applied as hybrid approach and the list will be displayed at the bottom of the page after book meta data and reviews with rate provided by users.

3.8 Technology Selection

The main technology along with other related technologies and libraries which will be used for the system is as follows.

Front End	Logic	Data Persistence
Flask web development, one drop at a time HTML CSS B Bootstrap	python scikit learn GOOGLE TRANSLATOR IN PYTHON NATURAL LANGUAGE PROCESSING USING NLTK PYTHON	MySQL MySQL Workbench

Table 2: Technology Stack

3.9 Evaluate

In order to perform the evaluation of the system 2 main authors were contacted. Additionally, more authors will be contacted and more active readers will be contacted to verify and get the feedback from them. Based on the feedback offline evaluation will be conducted. Apart from offline evaluation, quantitate based evaluation will be conducted to increase the accuracy of the system.

As a first phrase, the accuracy of the collaborate filter will be calculated, then content-based accuracy will be calculated. Finally, the accuracy level of combining two approach name hybrids will be calculated. These values can be compared and come to a conclusion which is the best approach.

3.10 Chapter Summary

The chapter explains the main architecture of the system along with the technology which will be used to implement the system. Furthermore, technology stack and the how the evaluation will be conducted explained in the chapter in details. The clarification between the methodologies and why the hybrid method is used is also described in this chapter.

4. Implementation

4.1 Chapter Overview

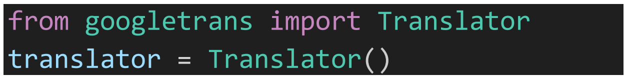
After describing the methodology of the system, the next task is to convert the methodology into a functional prototype. The prototype of the proposed system should address the main objectives that were identified in the first chapter. This chapter will discuss the implementation details individually for identified modules. At the end of the chapter the decisions taken on the low-level implementation would be discussed.

4.2 Preprocessing

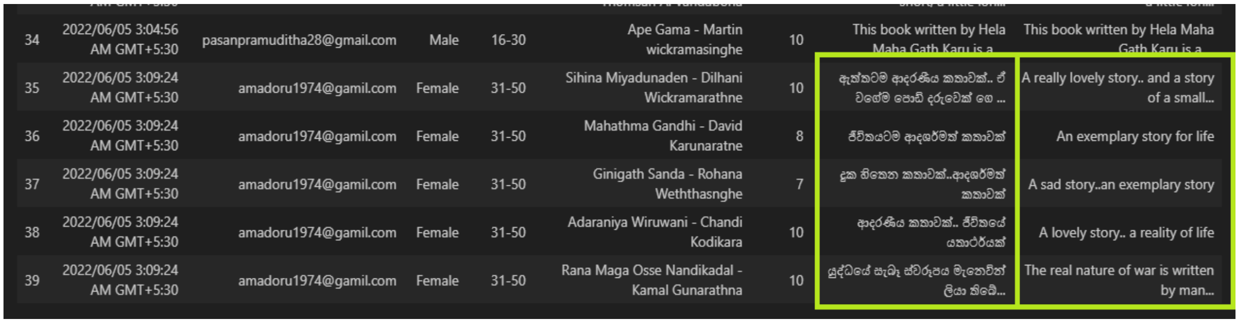
When considering the dataset, it is noted that the dataset contains reviews that were written in Sinhala language. Since we are planning to apply sentiment analysis for the reviews, it is compulsory to have the review data in English language. There for the very first task would be to convert the reviews written in Sinhala to English language.

4.2.1 Language Translation

In order to convert the language, Google Translator library could be used as below. Then all the reviews written in Sinhala will be converted to English reviews which ultimately could be applied sentiment analysis on top of the reviews.



```
from googletrans import Translator
translator = Translator()
```



34	2022/06/05 3:04:56 AM GMT+5:30	pasanpramuditha28@gmail.com	Male	16-30	Ape Gama - Martin wickramasinghe	10	This book written by Hela Maha Gath Karu is a	This book written by Hela Maha Gath Karu is a
35	2022/06/05 3:09:24 AM GMT+5:30	amadoru1974@gamil.com	Female	31-50	Sihina Miyadunaden - Dilhani Wickramarathne	10	ඇත්තටම ආදරණීය කතාවක්.. ඒ වගේම පොඩි දරුවෙක් ගෙ ...	A really lovely story.. and a story of a small...
36	2022/06/05 3:09:24 AM GMT+5:30	amadoru1974@gamil.com	Female	31-50	Mahathma Gandhi - David Karunaratne	8	ජීවිතයටම ආදශර්මත් කතාවක්	An exemplary story for life
37	2022/06/05 3:09:24 AM GMT+5:30	amadoru1974@gamil.com	Female	31-50	Ginigath Sanda - Rohana Weththasnghe	7	දුක හිතෙන කතාවක්..ආදශර්මත් කතාවක්	A sad story..an exemplary story
38	2022/06/05 3:09:24 AM GMT+5:30	amadoru1974@gamil.com	Female	31-50	Adaraniya Wiruwani - Chandi Kodikara	10	ආදරණීය කතාවක්.. ජීවිතයේ යතාර්ථයක්	A lovely story.. a reality of life
39	2022/06/05 3:09:24 AM GMT+5:30	amadoru1974@gamil.com	Female	31-50	Rana Maga Osse Nandikadal - Kamal Gunarathna	10	යුද්ධයේ සැබෑ ස්වරූපය මැනෙවින් ලියා තිබේ...	The real nature of war is written by man...

Figure 6: Review after converting to English

4.2.2 Sentimental analysis

Most online stores like Amazon, AliExpress, Ebay provide a website for users to express their opinions about different items they bought. Since then, it has been established that buying online, 90% of consumers are testing different websites channels to determine the quality of their purchase. To evaluate the text data and then extract the sentiment element from that the field of sentiment analysis is frequently used. From user ratings, suggestions, recommendations and messages, online business websites produce a massive volume of textual data every day.(Wassan et al., 2021)

Sentiment analysis is the process of analyzing a given text and determine if the text means to positive, negative or neutral. It basically helps to understand the human feelings via text. It is one of the Natural Language Processing (NLP) technique used to analyze the text.

As per the research (Tripathy et al., 2015) it says Sentiment analysis is the most prominent branch of natural language processing and it refers to feelings, attitudes, emotions. Most people used to express their sentiments to others through social media, ratings and reviews. Based on the review and the rate other can determine the quality and usability of a product that is sell over internet. The paper presents the comparison of results that is calculated by applying two algorithms Naïve Bayes and Support Vector Machine (SVM). The calculation was based on the dataset taken from a movie dataset.

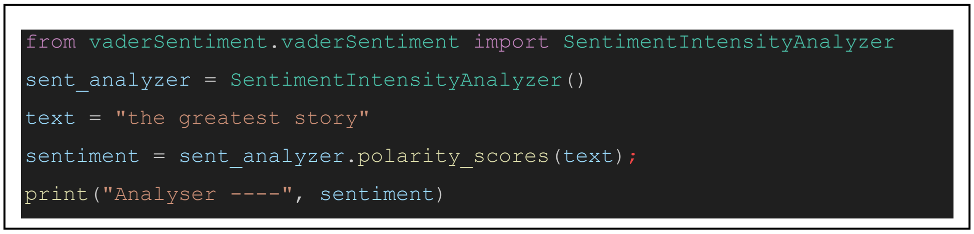
According to the study of the research (Chandrasekaran et al., 2022), it build a sentiment analysis model based on images from social media. For the work they used different transfer learning models, including the VGG-19, ResNet50V2, and DenseNet-121 models, to perform sentiment analysis based on images. As a dataset, Twitter-based images available in the Crowdflower dataset were used which contains URLs of images with their sentiment polarities.

There are several ways to apply sentiment analysis for a text. Following are some of them

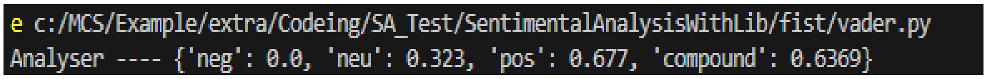
4.2.2.1 Using libraries

4.2.2.1.1 VADER - Valence Aware Dictionary and sEntiment Reasoner

It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is available in the NLTK package and can be directly applied to a text and gives both polarity(positive/negative) and intensity or strength. The feature depends on a dictionary which maps lexical features with sentiment score.



```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
sent_analyzer = SentimentIntensityAnalyzer()
text = "the greatest story"
sentiment = sent_analyzer.polarity_scores(text);
print("Analyser ----", sentiment)
```



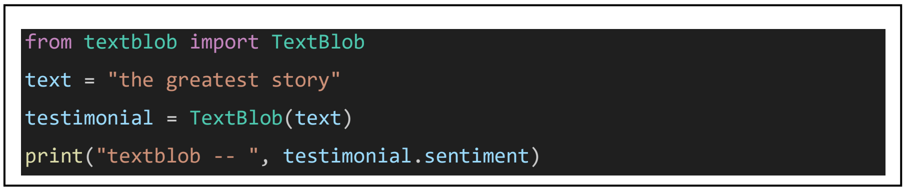
```
e c:/MCS/Example/extra/Codeing/SA_Test/SentimentalAnalysisWithLib/fist/vader.py
Analyser ---- {'neg': 0.0, 'neu': 0.323, 'pos': 0.677, 'compound': 0.6369}
```

Figure 7: VADER result

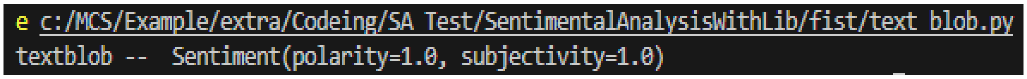
4.2.2.1.2 TextBlob

It is another lexicon-based python library which can be used to process a text and gives two main values polarity and subjectivity. Other than sentiment analysis, the library contains lot of features like noun phrase extraction, tokenization, lemmatization, spelling correction. As per the below example the text contains the word 'greatest' which textblob consider as the sentiment analyser and return positive value 1.0.

Polarity has the value between -1 to 1 where -1 represents the most negative words like 'worst', 'aweful', 'disgusting' while 1 represents most positive words like 'the best', 'excellent'. Subjectivity lies between 0 to 1 where 0 represent factual information while 1 represent more personal opinion. (Barai, 2021)



```
from textblob import TextBlob
text = "the greatest story"
testimonial = TextBlob(text)
print("textblob -- ", testimonial.sentiment)
```



```
e c:/MCS/Example/extra/Codeing/SA Test/SentimentalAnalysisWithLib/fist/text blob.py
textblob --  Sentiment(polarity=1.0, subjectivity=1.0)
```

Figure 8: TextBlob result

4.2.2.1.3 Compare VADER and Textblob

When we check some value in Textblob, it is noted that some text which have more negative values like not and slow, it multiplies -0.5 and -0.3 and gives the polarity of the sentence as a positive value. Another issue of Textblob is, if it finds any negative word in between in a sentence, it gives some polarity other than 0. Due to these issues Textblog could not be considered as one of the best sentiment analyzers.

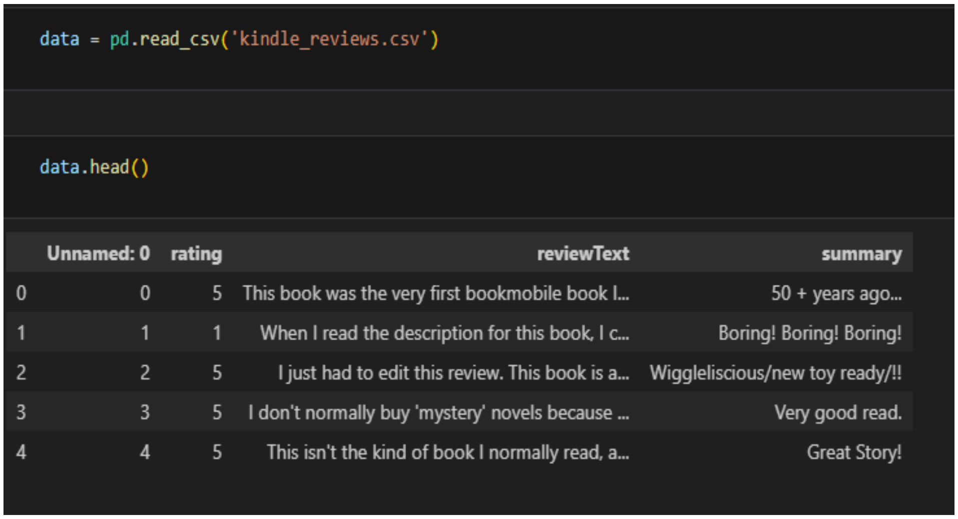
When the same above sentences check with VADER, it gives better result than Textblob. As per the (Barai, 2021) it compares both analyzers and came to a conclusion that Textblob struggled with negative sentences. The discussion further explained that It is not that VADER is better than Textblob in sentiment analysis. But it works better for negative sentences.

As conclusion, there are drawbacks for both of the analyzers. Therefore, it would be more convenient to implement own mechanism for sentiment analyze and predict the value for a given sentence.

4.2.2.2 Using own mechanism

There are some limitations when use any library in our system like not able to customized and not able to understand the logic behind the functionality. There for own mechanism of implementing sentimental analysis would be used. Following are the list of main steps for building the model then implement the pipeline for the model built. (“Machine Learning Project | Classification | Sentiment Analysis | Sinhala - YouTube,” n.d.)

In order to train the model for sentiment analysis, ‘Kindle reviews’ dataset was taken from Kaggle. The dataset looks like as below



```
data = pd.read_csv('kindle_reviews.csv')
```

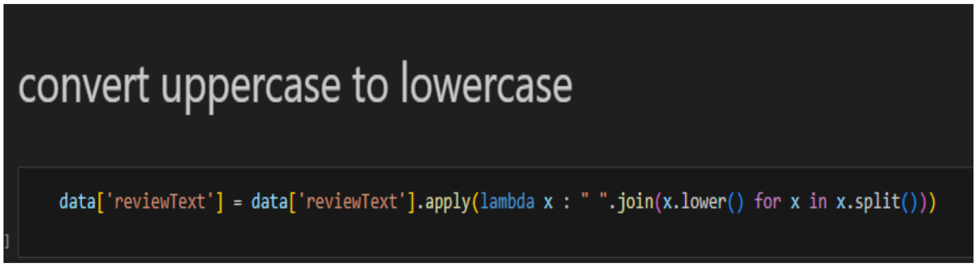
```
data.head()
```

	Unnamed: 0	rating	reviewText	summary
0	0	5	This book was the very first bookmobile book I...	50 + years ago...
1	1	1	When I read the description for this book, I c...	Boring! Boring! Boring!
2	2	5	I just had to edit this review. This book is a...	Wiggleliscious/new toy ready/!!
3	3	5	I don't normally buy 'mystery' novels because ...	Very good read.
4	4	5	This isn't the kind of book I normally read, a...	Great Story!

Figure 9: Kindle Review Data sample

4.2.2.2.1 Convert Uppercase to Lowercase

The review text contains uppercase as well as lowercase. As a first step all the characters to be converted to lowercase so the case sensitiveness can be ignored when comparing values.



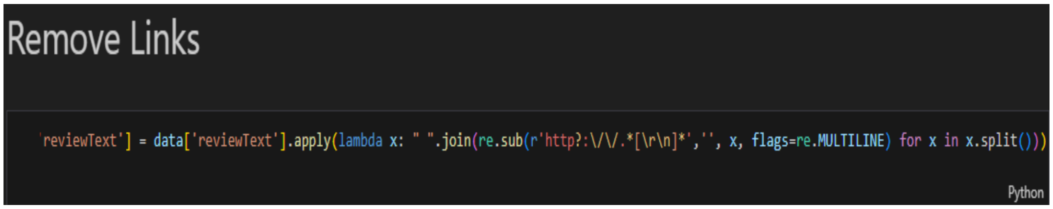
convert uppercase to lowercase

```python
data['reviewText'] = data['reviewText'].apply(lambda x : " ".join(x.lower() for x in x.split()))
```

Figure 10: Code to convert review to upper case

4.2.2.2.2 Remove Links

Since the links do not have any meaning for sentiment analyzer, those links to be removed with below code.



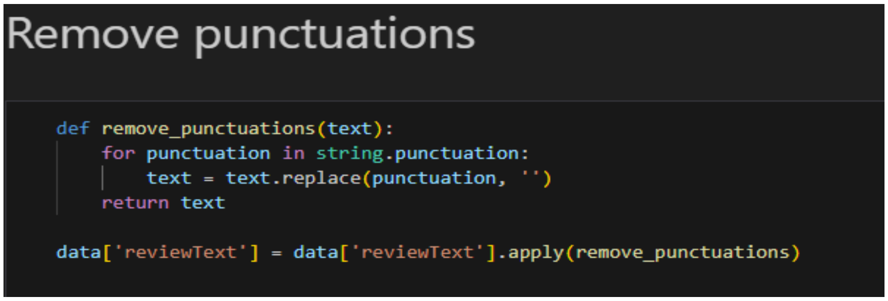
Remove Links

```python
'reviewText'] = data['reviewText'].apply(lambda x: " ".join(re.sub(r'http?:\/\/.*[\r\n]*','', x, flags=re.MULTILINE) for x in x.split()))
```

Figure 11: Code to remove links in reviews

4.2.2.2.3 Remove Punctuations

Since the punctuations also do not have any meaning for sentiment analyzer, all the punctuations to be removed with below code. Punctuation list can be found in string library. A function is defined to remove the punctuations and it is invoked in all the review text.



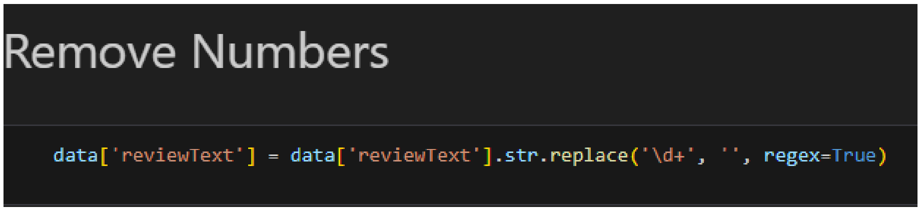
Remove punctuations

```python
def remove_punctuations(text):
    for punctuation in string.punctuation:
        text = text.replace(punctuation, '')
    return text

data['reviewText'] = data['reviewText'].apply(remove_punctuations)
```

Figure 12: Code to remove punctuation in reviews

4.2.2.2.4 Remove Numbers

There were numbers also added in the review test and those were also to be removed as they do not have any meaning for sentiment analyzer process. Removing numbers in a text can be achieved by below code.



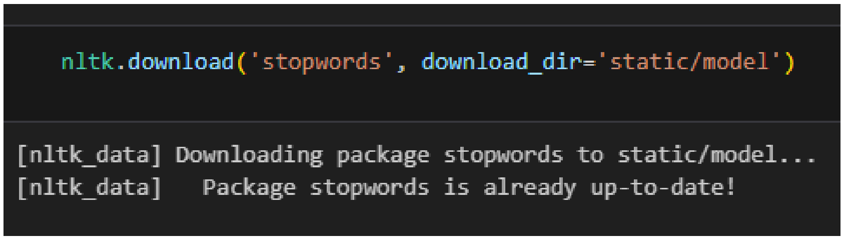
Remove Numbers

```
data['reviewText'] = data['reviewText'].str.replace('\d+', '', regex=True)
```

Figure 13: Code to remove numbers in reviews

4.2.2.2.5 Remove Stop words

There were numbers also added in the review text and those were also to be removed as they do not have any meaning for sentiment analyzer process. The list of stop words can be download from nltk library to a folder specified.

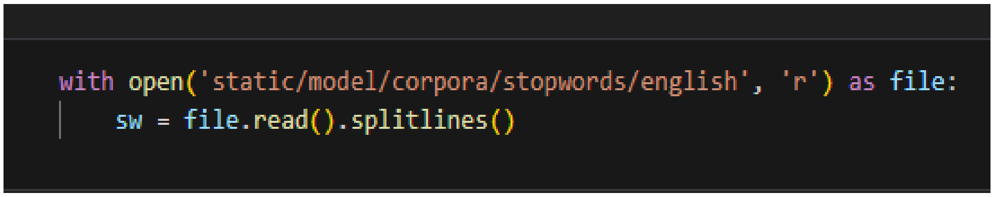


```
nltk.download('stopwords', download_dir='static/model')
```

```
[nltk_data] Downloading package stopwords to static/model...
[nltk_data]   Package stopwords is already up-to-date!
```

Figure 14: download stopwords from nltk library

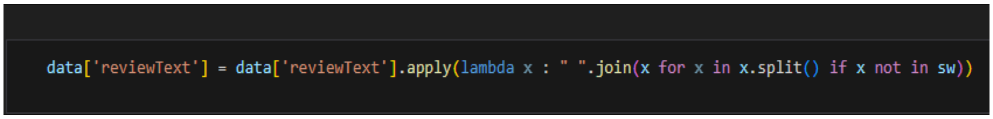
A variable is defined to store the list of stop words as below.



```
with open('static/model/corpora/stopwords/english', 'r') as file:
    sw = file.read().splitlines()
```

Figure 15: Read stopwords and store

Finally, the stop words are removed from the review text with following code

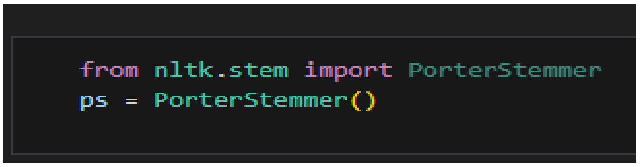


```
data['reviewText'] = data['reviewText'].apply(lambda x : " ".join(x for x in x.split() if x not in sw))
```

Figure 16: Code to remove stopwords in reviews

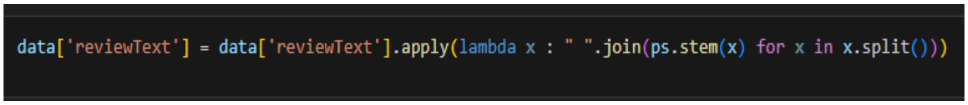
4.2.2.2.6 Apply Stemming

After removing all unnecessary values in the text, the next phrase is converting different verb formats to a common pattern. As an example, write, wrote, written, writing are converted to base form write. This process is called as stemming and the following code snippet will do the conversion.



```
from nltk.stem import PorterStemmer
ps = PorterStemmer()
```

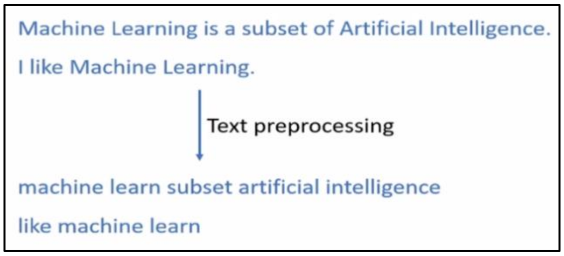
Figure 17: Read the stem and store



```
data['reviewText'] = data['reviewText'].apply(lambda x : " ".join(ps.stem(x) for x in x.split()))
```

Figure 18: Code to apply stemming

After completing the preprocessing part for the reviews, the text came up without uppercases, links, punctuations, numbers and stop words. Finally stemming has been applied to convert all the text to their base form. Following depict shows how the conversion has been done up to now



Machine Learning is a subset of Artificial Intelligence.

I like Machine Learning.

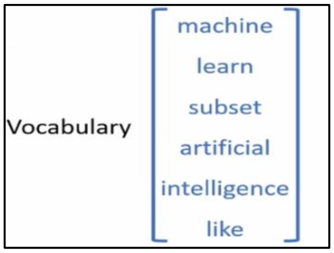
machine learn subset artificial intelligence

like machine learn

Figure 19: The way how the stemming is applied

4.2.2.2.7 Build Vocabulary

In order to build a model, the machine is not able to read and understand the text and they need to converted to numerical values. The building the vocabulary is the process of converting the text to appropriate numerical values. As the first step a unique vocabulary set to be created from the converted text. In the above example, following is the list.



machine
learn
subset
artificial
intelligence
like

Figure 20: Build the vocabulary

4.2.2.2.8 Vectorization

The next step of converting the text to numerical values, is vectorization process. As per the above example, all the sentence could be converted to numerical value which has the length of six (06). The value is same as the length of the vocabulary list. The list contains values which are called as features. In this example there are six features in the vocabulary.

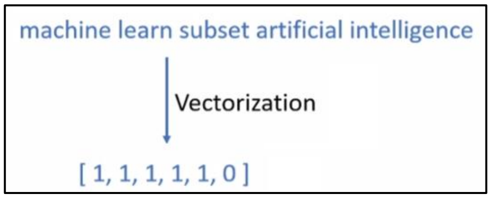


Figure 21: Vectorization 01

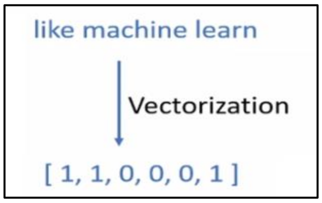
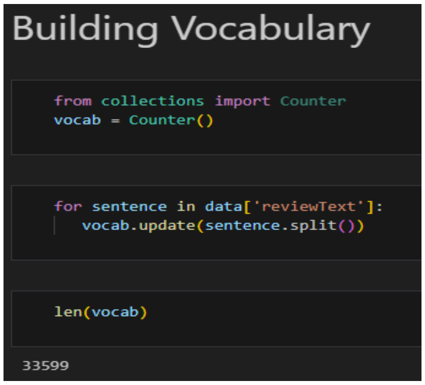


Figure 22: Vectorization 02

After the process, all the reviews will be converted to a numeric value which have the same length. Then the output can be fed to machine learning model.

Following code will check the size of the vocabulary list simply a number of features. The list contains a unique text and the number of times the text is used in the review.



Building Vocabulary

```
from collections import Counter
vocab = Counter()
```

```
for sentence in data['reviewText']:
    vocab.update(sentence.split())
```

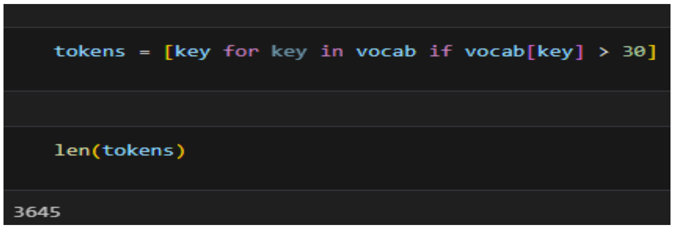
```
len(vocab)
```

```
33599
```

Figure 23: Vocabulary size

As per the above result, there are 33,599 features found in the reviews. In the scene, all the reviews will be represented as numeric value which has the length of 33,599. But the Kindle review data set contains around 12,000 records. If this much of features are used, the model will be over fit. The number of features should be less than the number of records in order for model to be a good one.

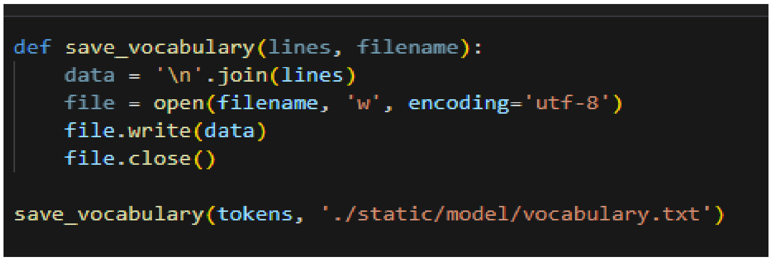
To overcome the issue, the feature selection will be used to reduce the feature count as bellow. Then the feature count is reduced to 3645.



```
tokens = [key for key in vocab if vocab[key] > 30]

len(tokens)

3645
```

Figure 24: Vocabulary refactored size

The final output of vocabulary list will be saved as bellow

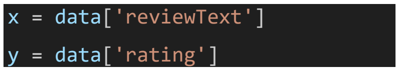


```
def save_vocabulary(lines, filename):
    data = '\n'.join(lines)
    file = open(filename, 'w', encoding='utf-8')
    file.write(data)
    file.close()

save_vocabulary(tokens, './static/model/vocabulary.txt')
```

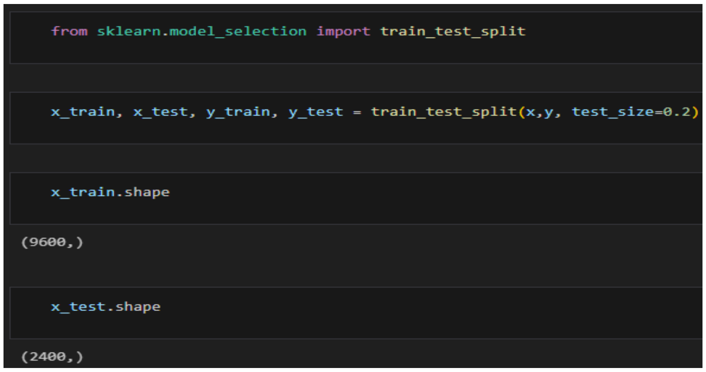
Figure 25: Save vocabulary

When an own method of creating a model for sentiment analysis, accuracy plays a significant role. It gives how the model is accurate as percentage.

Before vectorization, the review dataset to be divided to two main parts as training data and test data. The training data will be used to train the model and the test data will be used to test and get the accuracy of the model.



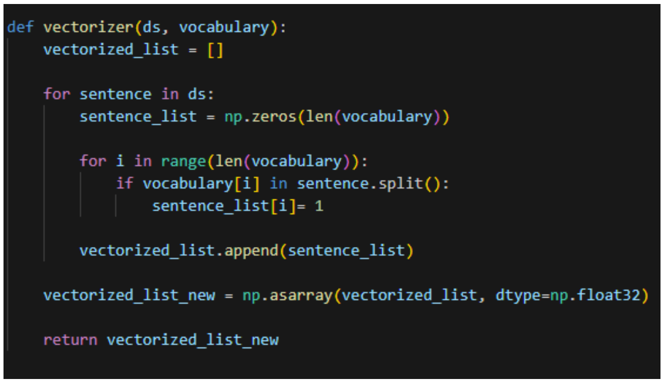
```
x = data['reviewText']
y = data['rating']
```



```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.2)

x_train.shape

(9600,)

x_test.shape

(2400,)
```

Figure 26: Divide train and Test data set

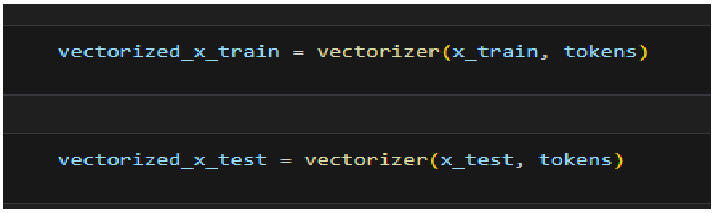
The vectorization can be done via bellow function.



```python
def vectorizer(ds, vocabulary):
    vectorized_list = []

    for sentence in ds:
        sentence_list = np.zeros(len(vocabulary))

        for i in range(len(vocabulary)):
            if vocabulary[i] in sentence.split():
                sentence_list[i]= 1

        vectorized_list.append(sentence_list)

    vectorized_list_new = np.asarray(vectorized_list, dtype=np.float32)

    return vectorized_list_new
```

Figure 27: Vectorization function

Then the function will be invoked for both train and test data as follows. Then all train data reviews and test data reviews will be converted to numeric data set.

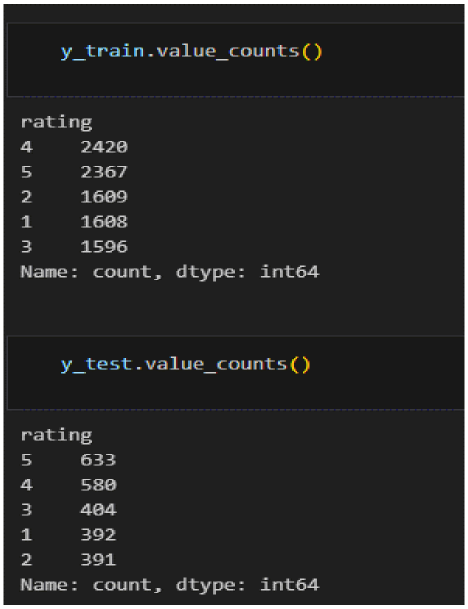


```python
vectorized_x_train = vectorizer(x_train, tokens)
```

```python
vectorized_x_test = vectorizer(x_test, tokens)
```

Figure 28: Apply vectorization function for both train and test data

Note how the values of data set are fairly shared (balanced dataset) for each rate as below



```python
y_train.value_counts()
```

```
rating
4    2420
5    2367
2    1609
1    1608
3    1596
Name: count, dtype: int64
```

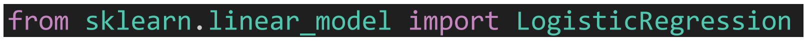
```python
y_test.value_counts()
```

```
rating
5    633
4    580
3    404
1    392
2    391
Name: count, dtype: int64
```

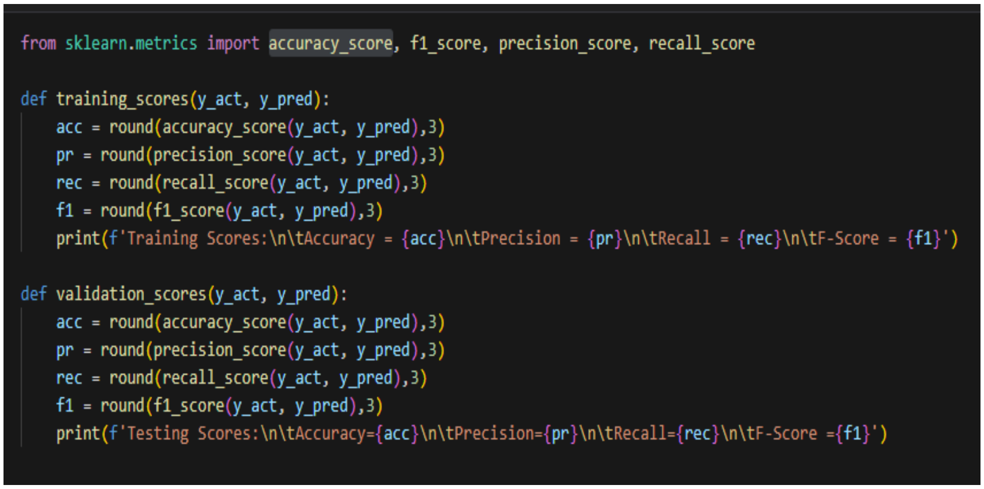
Figure 29: Balanced dataset

4.2.2.2.9 Model training and Evaluation

The next stage of the sentiment analysis process is building a model and evaluation.



```
from sklearn.linear_model import LogisticRegression
```

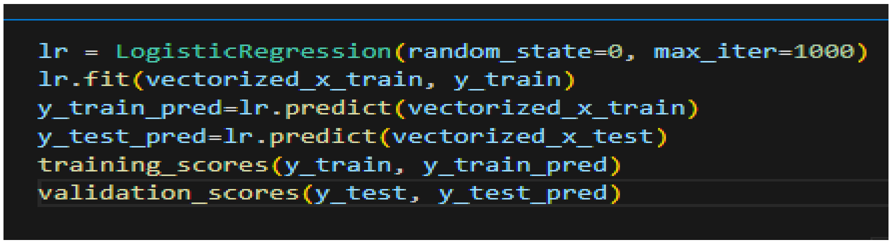


```
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score

def training_scores(y_act, y_pred):
    acc = round(accuracy_score(y_act, y_pred),3)
    pr = round(precision_score(y_act, y_pred),3)
    rec = round(recall_score(y_act, y_pred),3)
    f1 = round(f1_score(y_act, y_pred),3)
    print(f'Training Scores:\n\tAccuracy = {acc}\n\tPrecision = {pr}\n\tRecall = {rec}\n\tF-Score = {f1}')

def validation_scores(y_act, y_pred):
    acc = round(accuracy_score(y_act, y_pred),3)
    pr = round(precision_score(y_act, y_pred),3)
    rec = round(recall_score(y_act, y_pred),3)
    f1 = round(f1_score(y_act, y_pred),3)
    print(f'Testing Scores:\n\tAccuracy={acc}\n\tPrecision={pr}\n\tRecall={rec}\n\tF-Score ={f1}')
```

Figure 30: Functions defined to check the accuracy

4.2.2.2.10 Logistic Regression

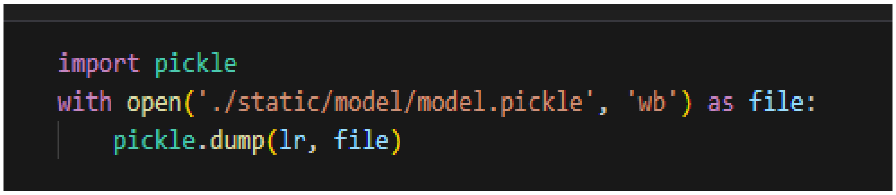
In order to train the model, we use logistic regression as it has the highest accuracy rate among other classification algorithms like decision Tree, Random Forest, Naïve bayes.



```
lr = LogisticRegression(random_state=0, max_iter=1000)
lr.fit(vectorized_x_train, y_train)
y_train_pred=lr.predict(vectorized_x_train)
y_test_pred=lr.predict(vectorized_x_test)
training_scores(y_train, y_train_pred)
validation_scores(y_test, y_test_pred)
```

Figure 31: Apply Logistic regression

After the model is trained properly, it is saved to a location as below

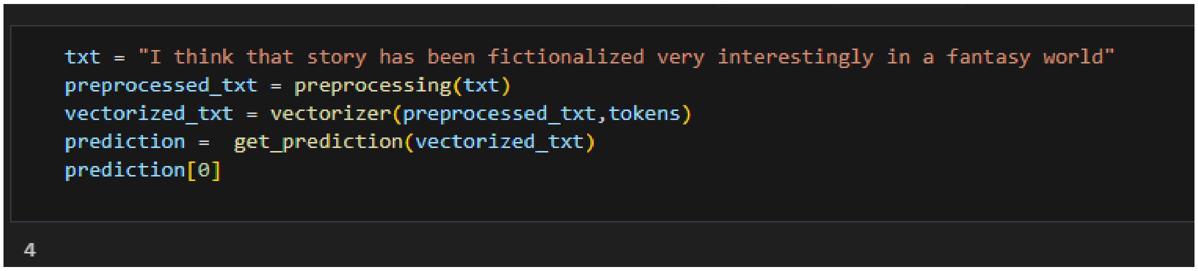


```
import pickle
with open('./static/model/model.pickle', 'wb') as file:
    pickle.dump(lr, file)
```

Figure 32: Save the model

4.2.3 Get Sentiment analysis rate

After the model is build, the sentiment analysis value to be calculate for a given text. There for the given test to be preprocessed, vectorized before get the prediction. Following code will invoke the appropriate function and return the predicted value.



```
txt = "I think that story has been fictionalized very interestingly in a fantasy world"
preprocessed_txt = preprocessing(txt)
vectorized_txt = vectorizer(preprocessed_txt,tokens)
prediction =  get_prediction(vectorized_txt)
prediction[0]
```

```
4
```

Figure 33: Get the review rate

Above result gives positive value 4 out of 5 for the given text. Since the rate calculated for the application is out of 10, it needs to be multiplied by 2 as the review rate.

4.3 Collaboration based Filter

Over the past decade, collaborative filtering algorithms have evolved from research algorithms intuitively capturing users’ preferences to algorithms that meet the performance demands of large commercial applications. (Schafer et al., n.d.)

Collaborative approaches make use of the measure of similarity between users. (Roy and Dutta, 2022). The model starts with finding a group or collection of user Y whose preferences, likes, and dislikes are similar to that of user X. Y is called the neighborhood of X. The new items which are liked by most of the users in Y are then recommended to user X. The accuracy of the approach is depending on how efficiency and accuracy the model can find the similarities of the target user. The main drawback of this algorithm is cold start and privacy concern as the user data has to be shared.

Collaborative approach is divided into two main categories named memory based and model based. Memory based again divided to Item based and User based. Following figure depicts all the approached in recommendation system.

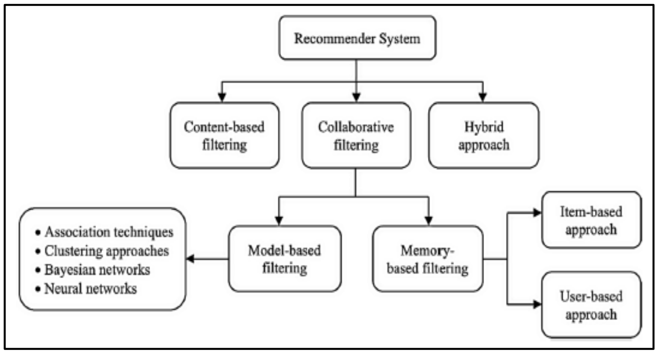


Figure 34: All models related to recommend system

Memory based approach recommend item based on preference of its neighborhood. In this approach to make recommendations for a new user, the user profile must be added to the utility matrix. If the user profile cannot fin, then this approach faces cold start issue. In user based approach, the user rating of a new item is calculated by finding other users from the user neighborhood who has previously rated that same item. If a new item receives positive ratings from the user neighborhood, the new item is recommended to the user. Below figure depicts the user-based filtering approach.(Roy and Dutta, 2022)

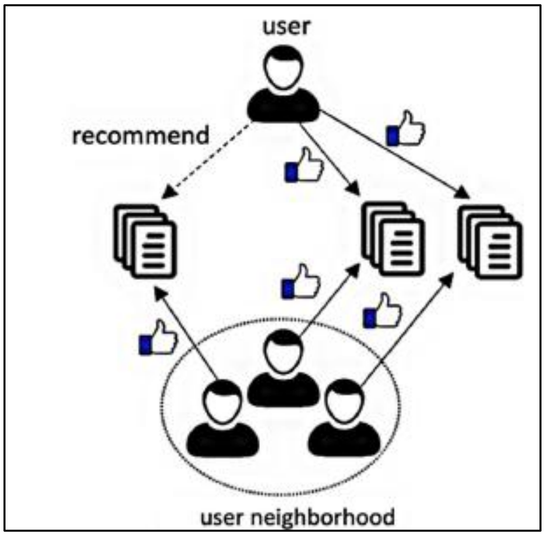


Figure 35: User based Collaborative filter

In the item-based approach, an item-neighborhood is built consisting of all similar items which the user has rated previously. Then that user's rating for a different new item is predicted by calculating the weighted average of all ratings present in a similar item-neighborhood as shown in below figure. (Roy and Dutta, 2022)

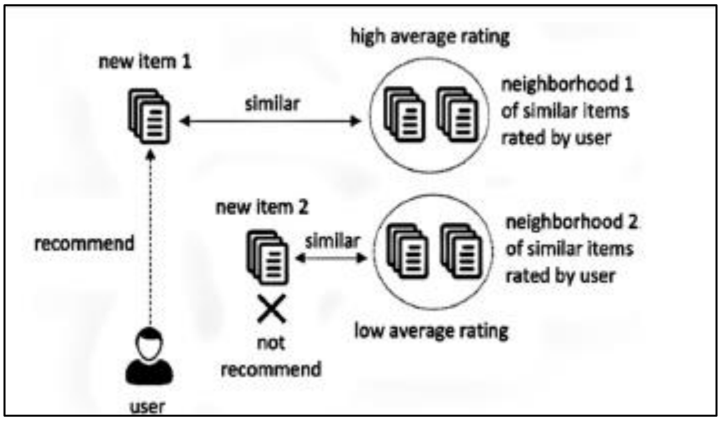
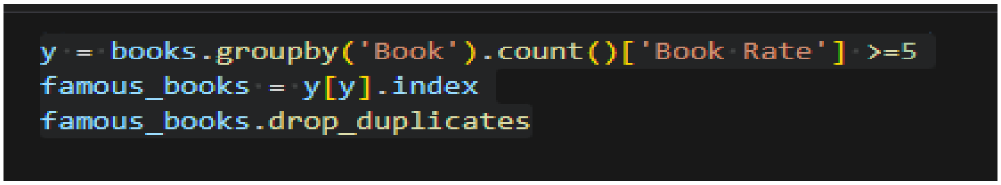


Figure 36: Item based collaborative filter

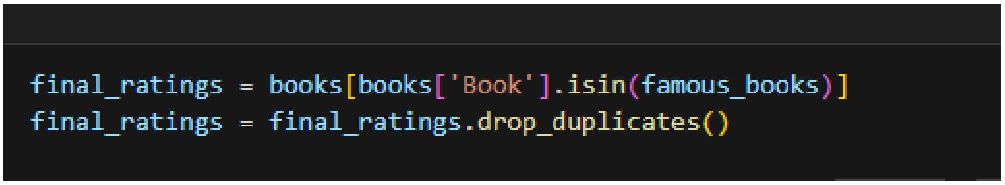
Since Content based filtering cannot discover the quality of an item, collaborative filtering system is used to overcome this problem.

As per the first step, we need to find books that are selected by at least more than 5 users. If no one is selected a book, they should be ignored and should not recommend those books for the users. Below code will remove such books.



```
y = books.groupby('Book').count()['Book Rate'] >=5
famous_books = y[y].index
famous_books.drop_duplicates
```

Figure 37: Find the books that are selected by more than 5 users



```
final_ratings = books[books['Book'].isin(famous_books)]
final_ratings = final_ratings.drop_duplicates()
```

Figure 38:Filter the book list with above selected books

The process of applying the Metrix and how the collaboration filter works is depicted below.

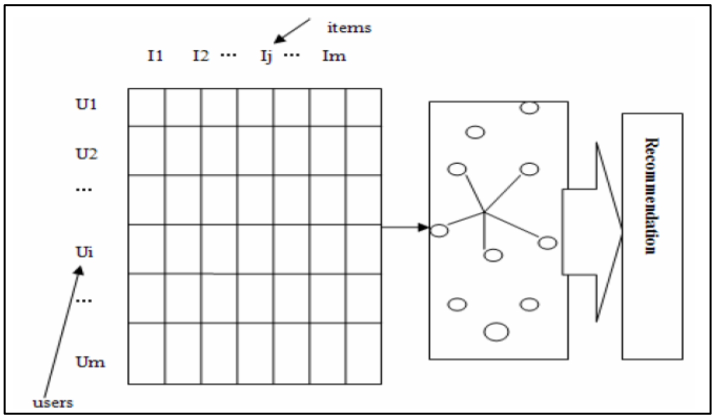
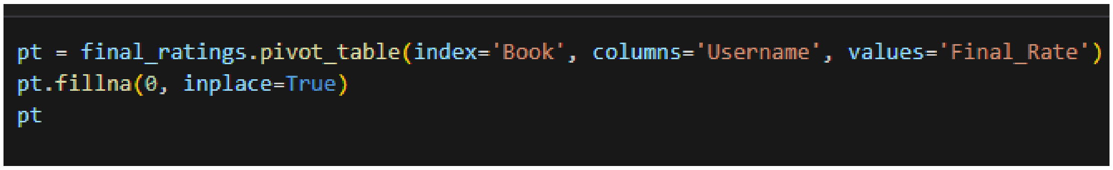


Figure 39:User vs Rate matrix

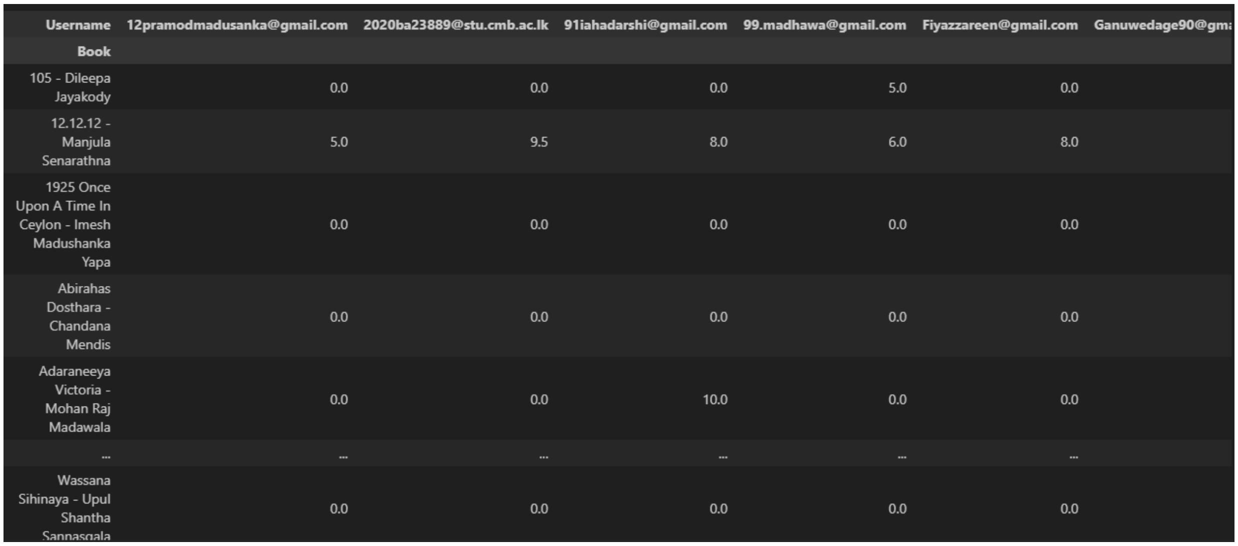
The code snippet to implement the matrix as below



```
pt = final_ratings.pivot_table(index='Book', columns='Username', values='Final_Rate')
pt.fillna(0, inplace=True)
pt
```

Figure 40: Code to implement the matrix

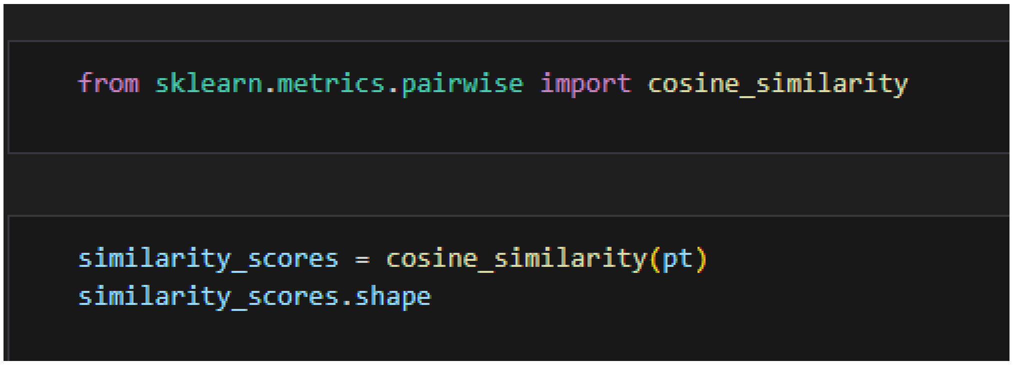
The result of the matrix can be found as below.



Username	12pramodmadusanka@gmail.com	2020ba23889@stu.cmb.ac.lk	91iahadarshi@gmail.com	99.madhawa@gmail.com	Fiyazzareen@gmail.com	Ganuwedage90@gm
Book						
105 - Dileepa Jayakody	0.0	0.0	0.0	5.0	0.0	
12.12.12 - Manjula Senarathna	5.0	9.5	8.0	6.0	8.0	
1925 Once Upon A Time In Ceylon - Imesh Madushanka Yapa	0.0	0.0	0.0	0.0	0.0	
Abirahas Dosthara - Chandana Mendis	0.0	0.0	0.0	0.0	0.0	
Adaraneeya Victoria - Mohan Raj Madawala	0.0	0.0	10.0	0.0	0.0	
...	...	...	...	...	...	
Wassana Sihinaya - Upul Shantha Sannasgala	0.0	0.0	0.0	0.0	0.0	

Figure 41: User vs book rate matrix

Finally, applying cosine similarity for the matrix and finding the similar books as below.

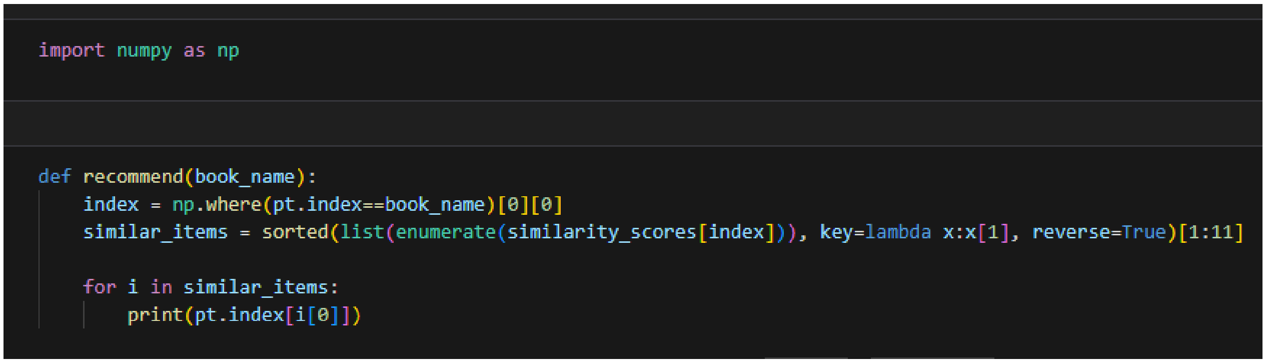


```
from sklearn.metrics.pairwise import cosine_similarity
```

```
similarity_scores = cosine_similarity(pt)
similarity_scores.shape
```

Figure 42:Library for cosine similarity

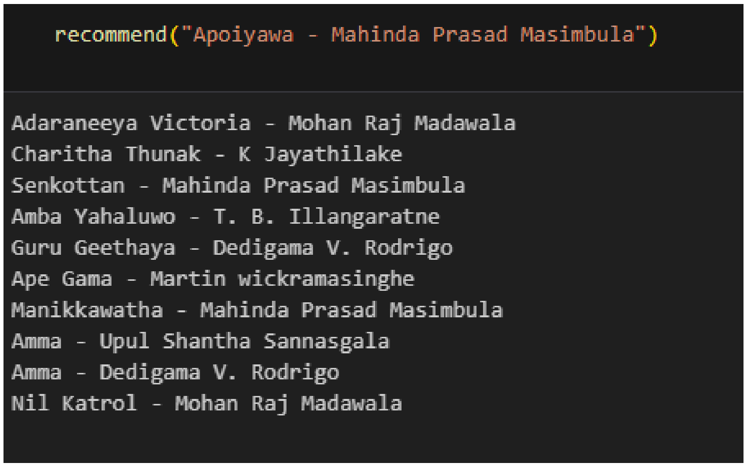
With below function, we can recommend the books.



```
import numpy as np


def recommend(book_name):
    index = np.where(pt.index==book_name)[0][0]
    similar_items = sorted(list(enumerate(similarity_scores[index])), key=lambda x:x[1], reverse=True)[1:11]

    for i in similar_items:
        print(pt.index[i[0]])
```

Figure 43: Function to recommend similar books

Calling the function and getting the book list



```
recommend("Apoiyawa - Mahinda Prasad Masimbula")

Adaraneeya Victoria - Mohan Raj Madawala
Charitha Thunak - K Jayathilake
Senkottan - Mahinda Prasad Masimbula
Amba Yahaluwo - T. B. Illangaratne
Guru Geethaya - Dedigama V. Rodrigo
Ape Gama - Martin wickramasinghe
Manikkawatha - Mahinda Prasad Masimbula
Amma - Upul Shantha Sannasgala
Amma - Dedigama V. Rodrigo
Nil Katrol - Mohan Raj Madawala
```

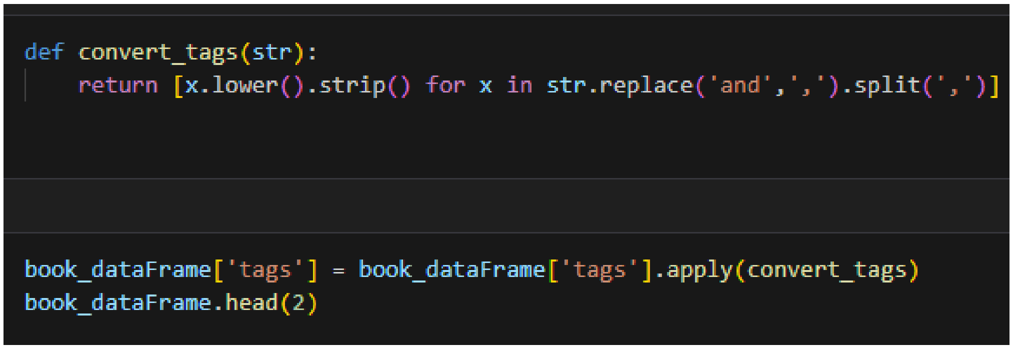
Figure 44: Collaborative filter result

4.4 Content based Filter

The Collaboration filter is totally based on the previous data collected by other users. Content-based filtering uses the assumption that items with similar objective features will be rated similarly. For example, if you liked a web page with the words "tomato sauce," you will like another web page with the words "tomato sauce." (Schafer et al., n.d.)

If the user does not have previous data, similar books are not be able to recommended. Even a new book added to the system and it has not been rated, it cannot be recommended. Content based filtering introduced to overcome these problems, even previous data is not found, books can be recommended based on the contents.

For an instance a book contains keywords like 'sherlock Holmes', 'detective' books which have similar keywords will be recommended. The first task of the process is replacing 'and' with a comma (,). It could be done with bellow code snippet.

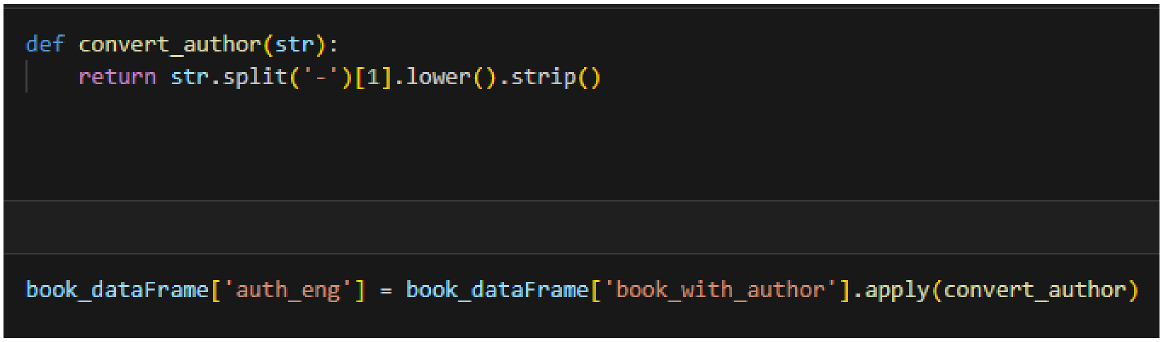


```
def convert_tags(str):
    return [x.lower().strip() for x in str.replace('and',',').split(',')]
```

```
book_dataFrame['tags'] = book_dataFrame['tags'].apply(convert_tags)
book_dataFrame.head(2)
```

Figure 45: replace and with comma

With below code, all authors can be extracted and save in a different field.

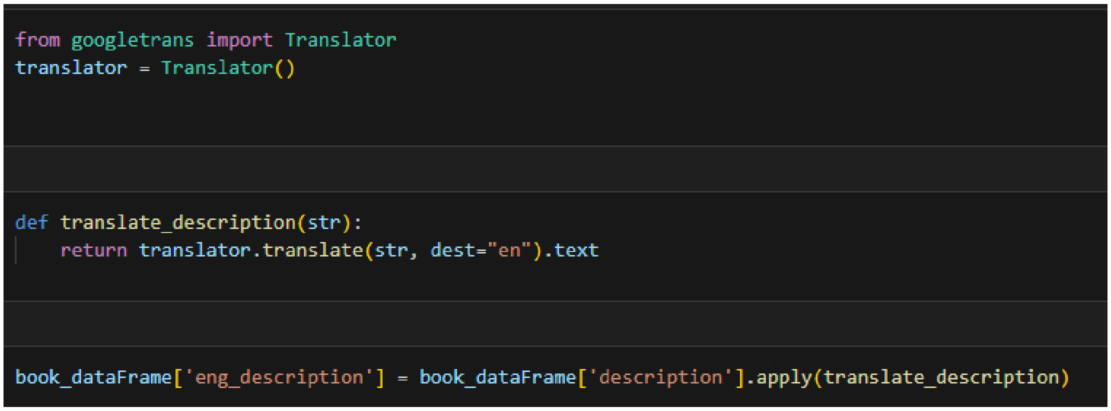


```
def convert_author(str):
    return str.split('-')[1].lower().strip()
```

```
book_dataFrame['auth_eng'] = book_dataFrame['book_with_author'].apply(convert_author)
```

Figure 46: extract authors

Below code will convert Sinhala description to English as the description will be added to the tags



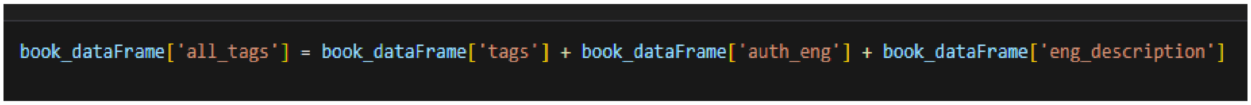
```
from googletrans import Translator
translator = Translator()
```

```
def translate_description(str):
    return translator.translate(str, dest="en").text
```

```
book_dataFrame['eng_description'] = book_dataFrame['description'].apply(translate_description)
```

Figure 47: Convert description to English

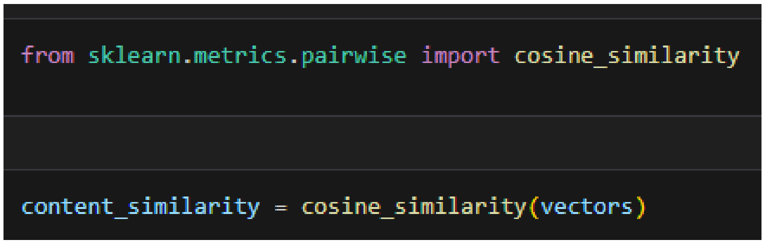
Combine all together as tags as below



```
book_dataFrame['all_tags'] = book_dataFrame['tags'] + book_dataFrame['auth_eng'] + book_dataFrame['eng_description']
```

Figure 48: Combine all together

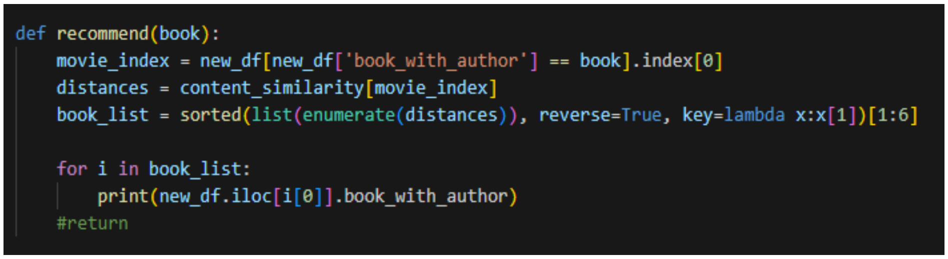
Apply cosine similarity as below



```
from sklearn.metrics.pairwise import cosine_similarity

content_similarity = cosine_similarity(vectors)
```

Figure 49: Apply Cosine similarity

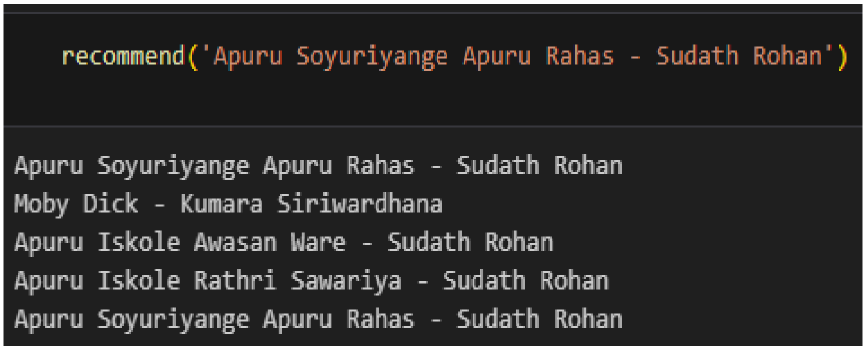
Define a function to recommend content similarity.



```
def recommend(book):
    movie_index = new_df[new_df['book_with_author'] == book].index[0]
    distances = content_similarity[movie_index]
    book_list = sorted(list(enumerate(distances)), reverse=True, key=lambda x:x[1])[1:6]

    for i in book_list:
        print(new_df.iloc[i[0]].book_with_author)
    #return
```

Figure 50: Function to recommend books based on content

Calling the function and getting the book list



```
recommend('Apuru Soyuriyange Apuru Rahas - Sudath Rohan')
```

```
Apuru Soyuriyange Apuru Rahas - Sudath Rohan
Moby Dick - Kumara Siriwardhana
Apuru Iskole Awasan Ware - Sudath Rohan
Apuru Iskole Rathri Sawariya - Sudath Rohan
Apuru Soyuriyange Apuru Rahas - Sudath Rohan
```

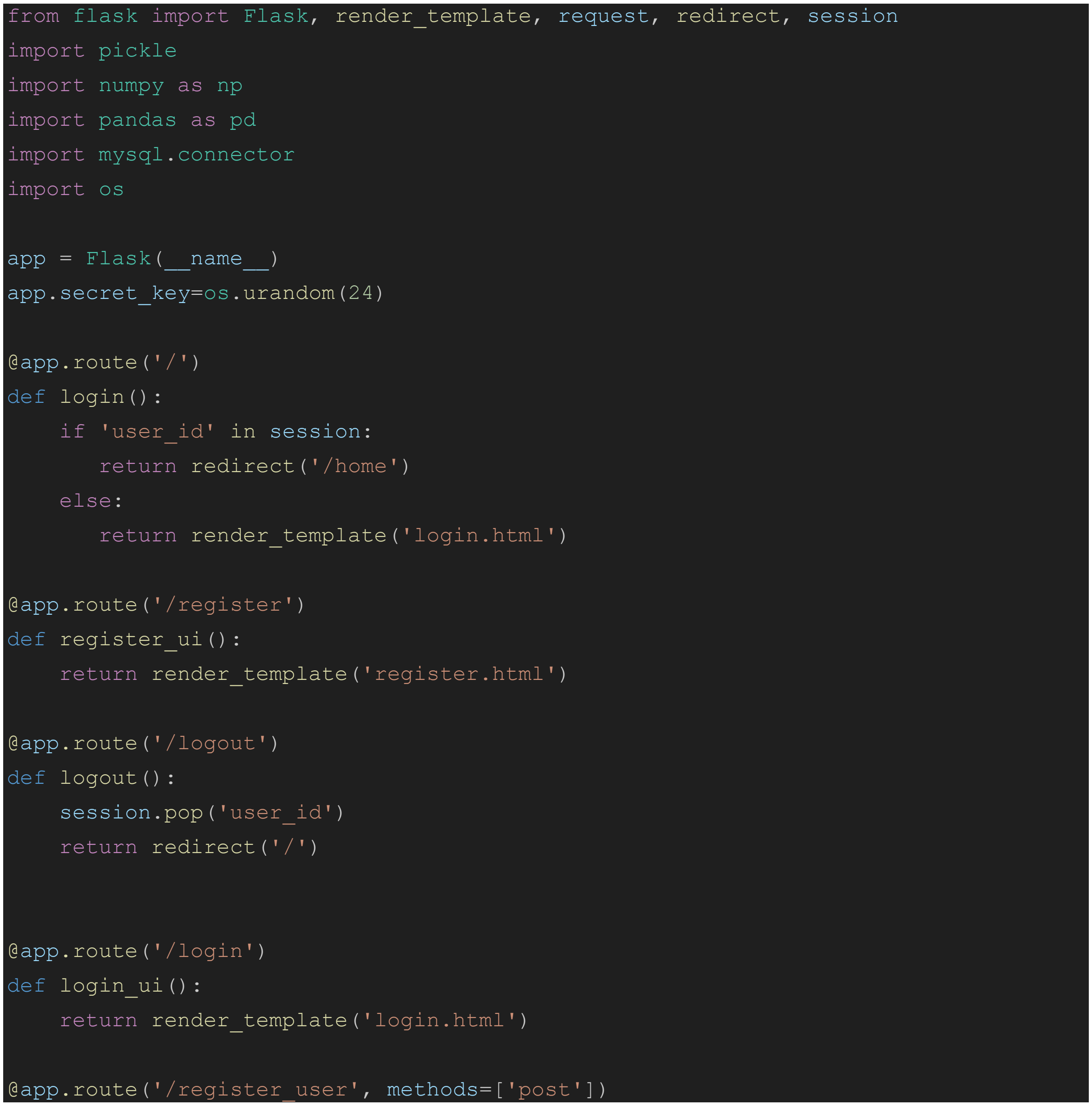
Figure 51: Calling the function and get recommended books

4.5 Web Application

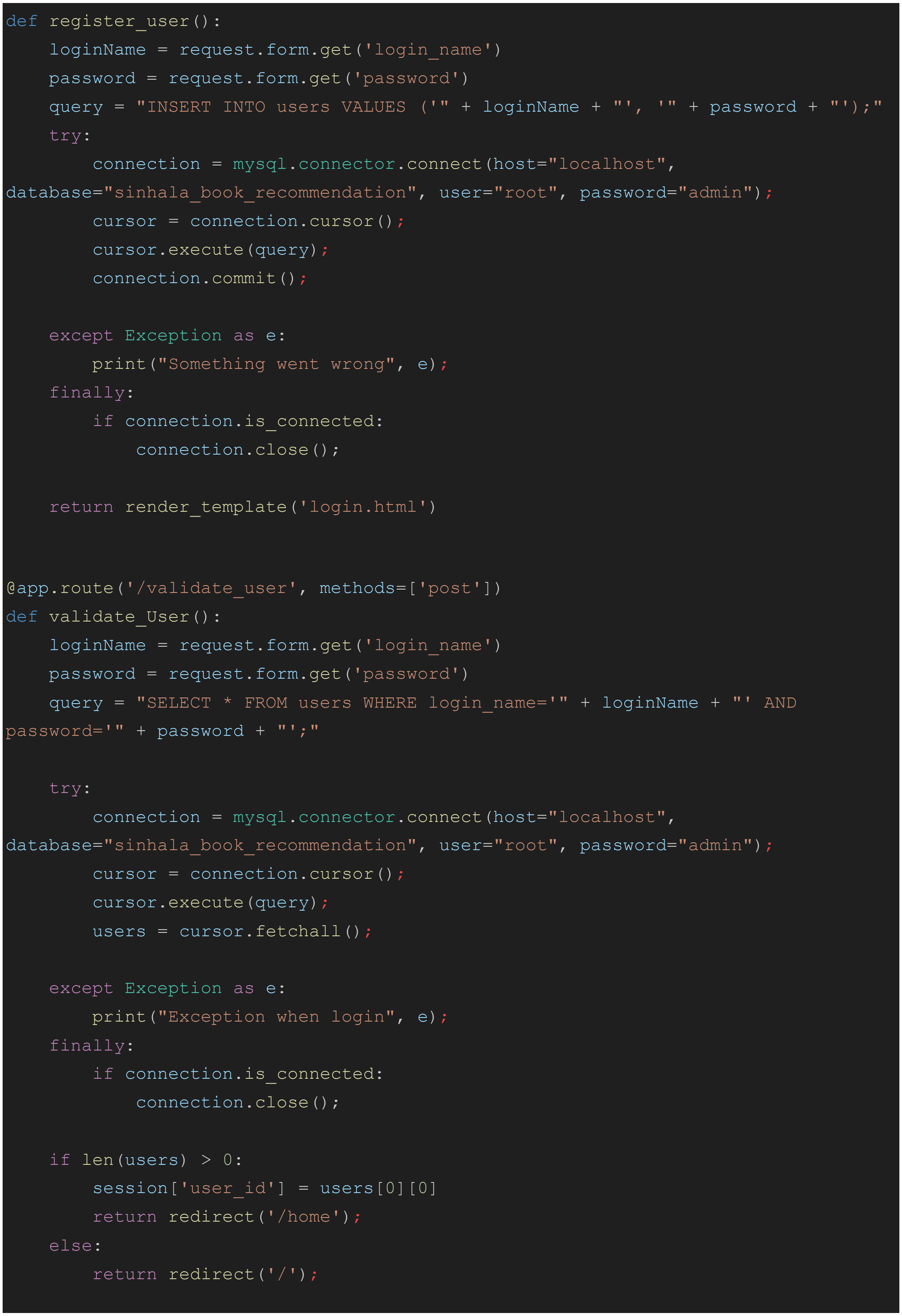
The web application is developed using Flask in python and it contains user interface and the authentication.

4.5.1 User Interface

The graphical user interface (GUI) of the web application is implemented using python flask, HTML, CSS and Bootstrap. The main file which defines the routes of the application is as follows.

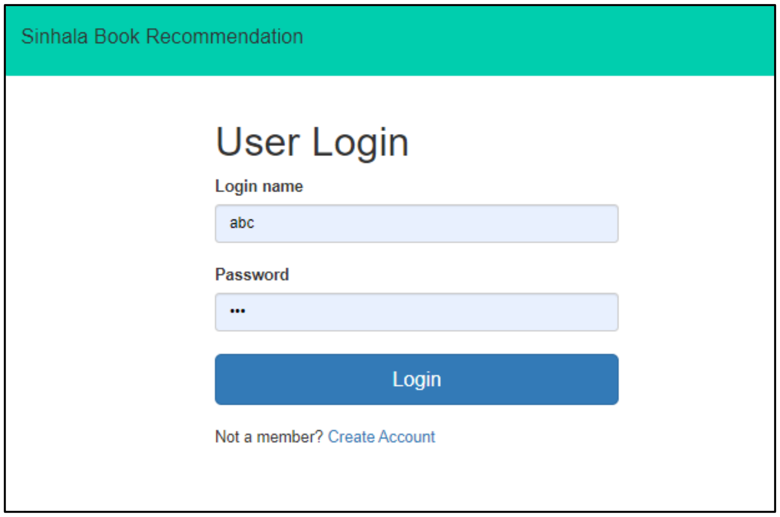


```
from flask import Flask, render_template, request, redirect, session
import pickle
import numpy as np
import pandas as pd
import mysql.connector
import os

app = Flask(__name__)
app.secret_key=os.urandom(24)

@app.route('/')
def login():
    if 'user_id' in session:
       return redirect('/home')
    else:
       return render_template('login.html')

@app.route('/register')
def register_ui():
    return render_template('register.html')

@app.route('/logout')
def logout():
    session.pop('user_id')
    return redirect('/')


@app.route('/login')
def login_ui():
    return render_template('login.html')

@app.route('/register_user', methods=['post'])
```



```
def register_user():
    loginName = request.form.get('login_name')
    password = request.form.get('password')
    query = "INSERT INTO users VALUES ('" + loginName + "', '" + password + "');"
    try:
        connection = mysql.connector.connect(host="localhost",
database="sinhala_book_recommendation", user="root", password="admin");
        cursor = connection.cursor();
        cursor.execute(query);
        connection.commit();

    except Exception as e:
        print("Something went wrong", e);
    finally:
        if connection.is_connected:
            connection.close();

    return render_template('login.html')


@app.route('/validate_user', methods=['post'])
def validate_User():
    loginName = request.form.get('login_name')
    password = request.form.get('password')
    query = "SELECT * FROM users WHERE login_name='" + loginName + "' AND
password='" + password + "';"

    try:
        connection = mysql.connector.connect(host="localhost",
database="sinhala_book_recommendation", user="root", password="admin");
        cursor = connection.cursor();
        cursor.execute(query);
        users = cursor.fetchall();

    except Exception as e:
        print("Exception when login", e);
    finally:
        if connection.is_connected:
            connection.close();

    if len(users) > 0:
        session['user_id'] = users[0][0]
        return redirect('/home');
    else:
        return redirect('/');
```

```
if __name__ == '__main__':
    app.run(debug=True)
```



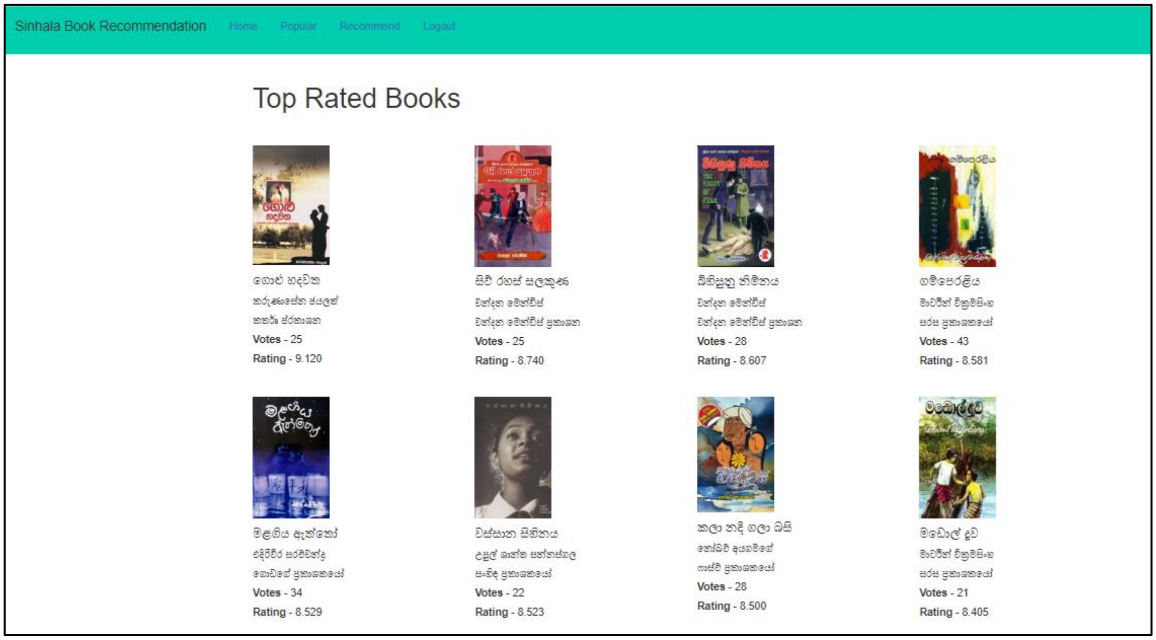
Sinhala Book Recommendation

User Login

Login name: abc

Password: •••

Figure 52: Login Page



Top Rated Books

ගොළු හදවත
කරුණාසේන ජයලත්
කතෘ ප්‍රකාශන
Votes - 25
Rating - 9.120

සිව් රහස් සලකුණ
චන්දන මෙන්ඩිස්
චන්දන මෙන්ඩිස් ප්‍රකාශන
Votes - 25
Rating - 8.740

බිහිසුණු නිම්නය
චන්දන මෙන්ඩිස්
චන්දන මෙන්ඩිස් ප්‍රකාශන
Votes - 28
Rating - 8.607

ගම්පෙරළිය
මාටින් වික්‍රමසිංහ
සරස ප්‍රකාශකයෝ
Votes - 43
Rating - 8.581

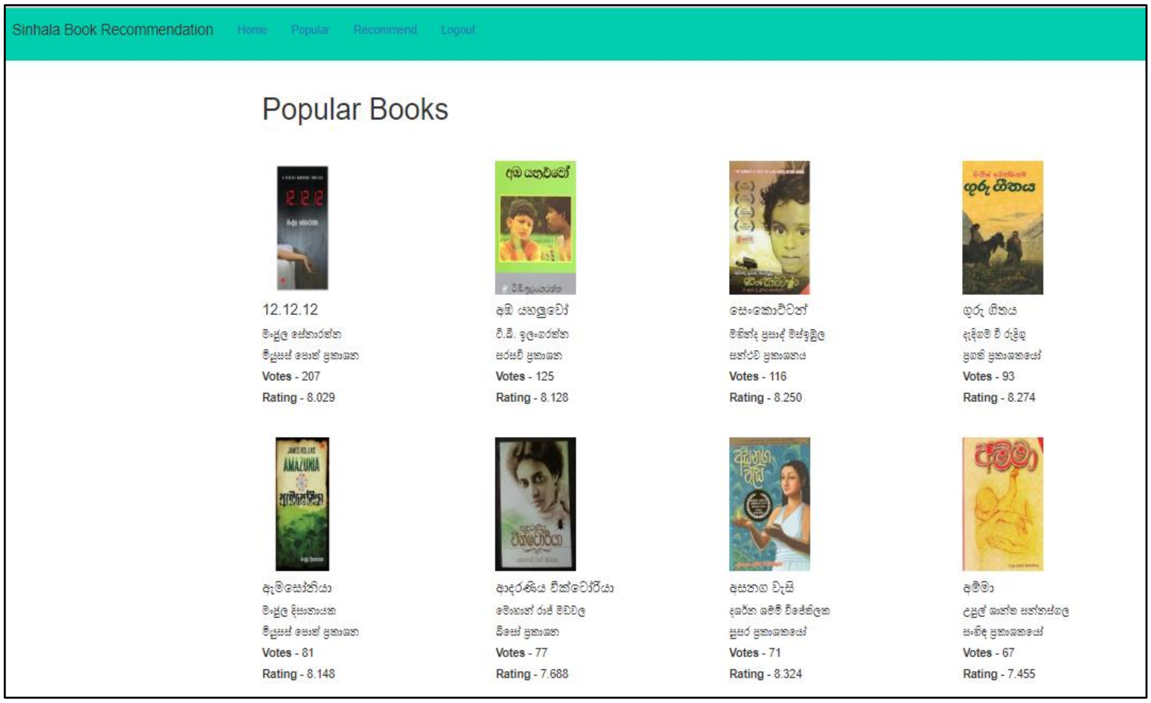
මළගිය ඇත්තෝ
එදිරිවීර සරච්චන්ද්‍ර
ගොඩගේ ප්‍රකාශකයෝ
Votes - 34
Rating - 8.529

වස්සාන සිහිනය
උපුල් ශාන්ත සන්නස්ගල
සංහිඳ ප්‍රකාශකයෝ
Votes - 22
Rating - 8.523

කලා නදී ගලා බසී
හෝබ්ට් අයගමගේ
ෆාස්ට් ප්‍රකාශකයෝ
Votes - 28
Rating - 8.500

මඩොල් දූව
මාටින් වික්‍රමසිංහ
සරස ප්‍රකාශකයෝ
Votes - 21
Rating - 8.405

Figure 53: Top Rated books



Popular Books

12.12.12
මංජුල සේනාරත්න
මියුසස් පොත් ප්‍රකාශන
Votes - 207
Rating - 8.029

අඹ යහළුවෝ
ටී.බී. ඉලංගරත්න
සරසවි ප්‍රකාශන
Votes - 125
Rating - 8.128

සෙංකොට්ටන්
මහින්ද ප්‍රසාද් මස්ඉඹුල
සන්ථව ප්‍රකාශනය
Votes - 116
Rating - 8.250

ගුරු ගීතය
දැදිගම වී රොද්‍රිගු
ප්‍රගති ප්‍රකාශකයෝ
Votes - 93
Rating - 8.274

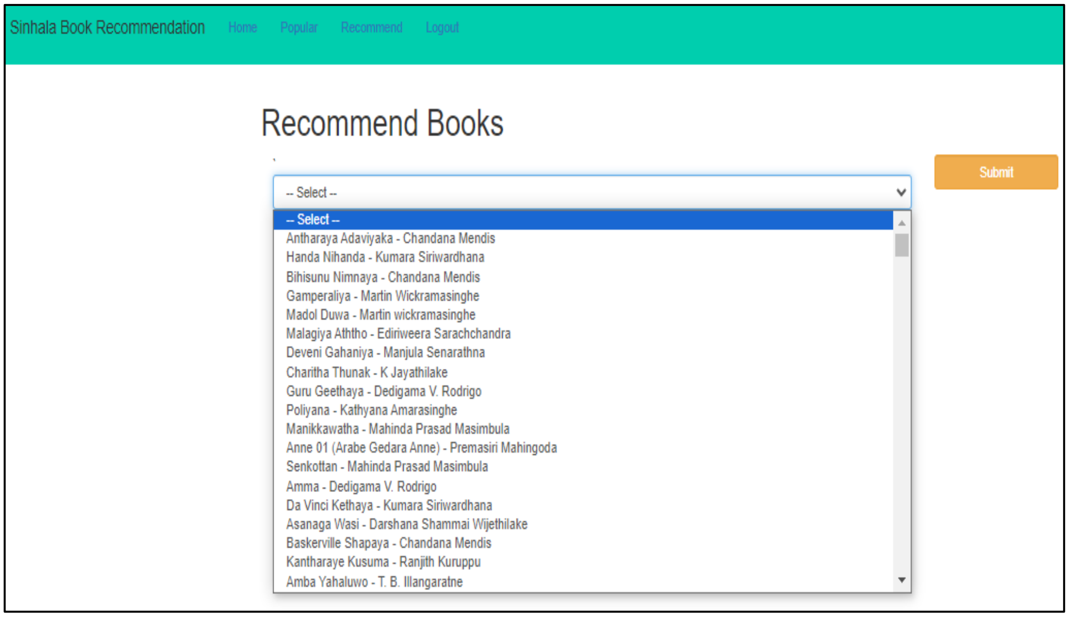
ඇමසෝනියා
මංජුල දිසානායක
මියුසස් පොත් ප්‍රකාශන
Votes - 81
Rating - 8.148

ආදරණීය වික්ටෝරියා
මොහාන් රාජ් මඩවල
ඩිසෝ ප්‍රකාශන
Votes - 77
Rating - 7.688

අසනග වැසි
දර්ශන ශම්මි විජේතිලක
සුසර ප්‍රකාශකයෝ
Votes - 71
Rating - 8.324

අම්මා
උපුල් ශාන්ත සන්නස්ගල
සංහිඳ ප්‍රකාශකයෝ
Votes - 67
Rating - 7.455

Figure 54: Popular Book List

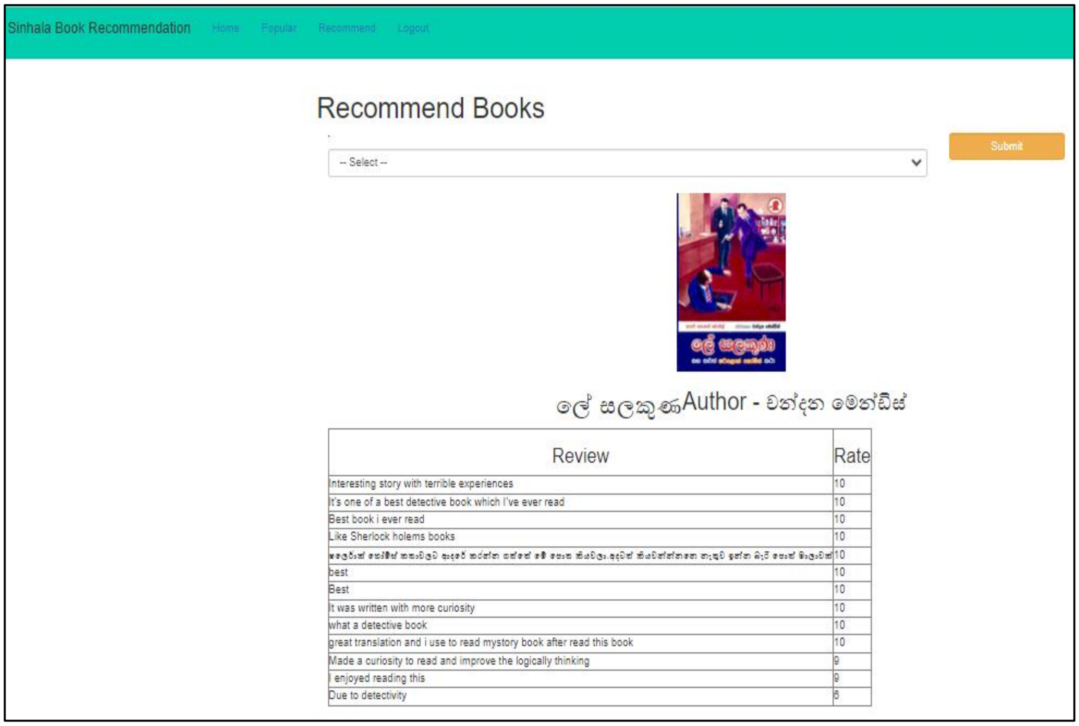


Recommend Books

-- Select --

-- Select --
Antharaya Adaviyaka - Chandana Mendis
Handa Nihanda - Kumara Siriwardhana
Bihisunu Nimnaya - Chandana Mendis
Gamperaliya - Martin Wickramasinghe
Madol Duwa - Martin wickramasinghe
Malagiya Aththo - Ediriweera Sarachchandra
Deveni Gahaniya - Manjula Senarathna
Charitha Thunak - K Jayathilake
Guru Geethaya - Dedigama V. Rodrigo
Poliyana - Kathyana Amarasinghe
Manikkawatha - Mahinda Prasad Masimbula
Anne 01 (Arabe Gedara Anne) - Premasiri Mahingoda
Senkottan - Mahinda Prasad Masimbula
Amma - Dedigama V. Rodrigo
Da Vinci Kethaya - Kumara Siriwardhana
Asanaga Wasi - Darshana Shammai Wijethilake
Baskerville Shapaya - Chandana Mendis
Kantharaye Kusuma - Ranjith Kuruppu
Amba Yahaluwo - T. B. Illangaratne

Figure 55: Recommend book list

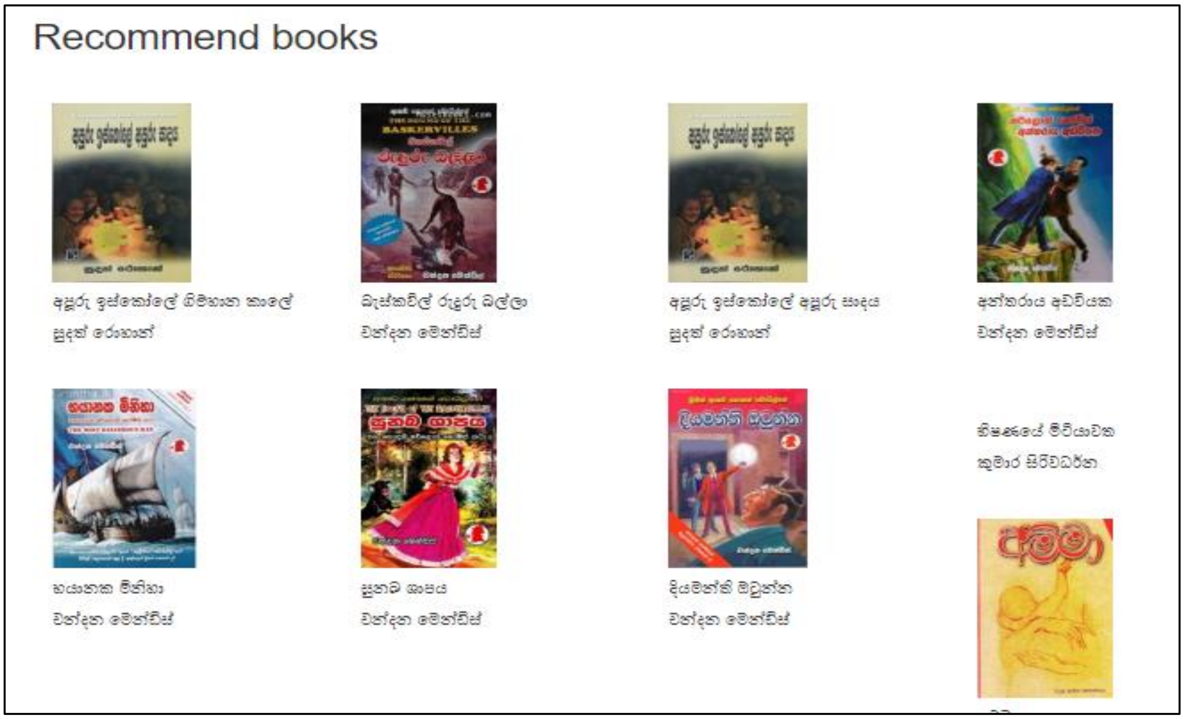


Recommend Books

ලේ සලකුණAuthor - චන්දන මෙන්ඩිස්

Review	Rate
Interesting story with terrible experiences	10
It's one of a best detective book which I've ever read	10
Best book i ever read	10
Like Sherlock holems books	10
ෂෙලොර්ක් හෝම්ස් කතාවලට ආදරේ කරන්න පත්තේ මේ පොත කියවලා.අදටත් කියවන්නනත නැතුව ඉන්න බැරි පොත් මාලාවක්	10
best	10
Best	10
It was written with more curiosity	10
what a detective book	10
great translation and i use to read mystory book after read this book	10
Made a curiosity to read and improve the logically thinking	9
I enjoyed reading this	9
Due to detectivity	6

Figure 57: Selected book with reviews


Recommend books

අද්භූත ඉස්කෝලේ ගිම්හාන කාලේ
සුදත් රොහාන්

බැස්කවිල් රුදුරු බල්ලා
චන්දන මෙන්ඩිස්

අද්භූත ඉස්කෝලේ අද්භූත සංදය
සුදත් රොහාන්

අන්තරාය අඩවියක
චන්දන මෙන්ඩිස්

භයානක මිනිහා
චන්දන මෙන්ඩිස්

සුනඛ ශාපය
චන්දන මෙන්ඩිස්

දියමන්ති ඔටුන්න
චන්දන මෙන්ඩිස්

නිෂණයේ මිටියාවත
කුමාර සිරිවර්ධන

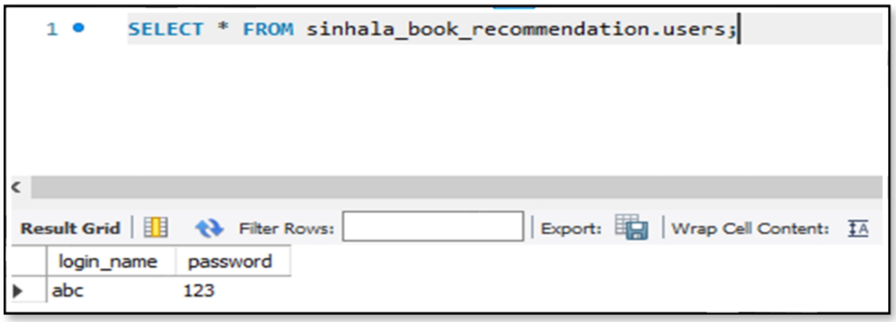
Figure 56: Recommended books

4.5.2 Authenticate

The authentication also integrated with the application so that none of a user can view the data without login to the system and it was done to improve the security of the application. Admin user will be added as the main user who can view all the top rated, most popular and recommended books. User management and book management will be implemented as the project is still under implementation.

4.6 Database

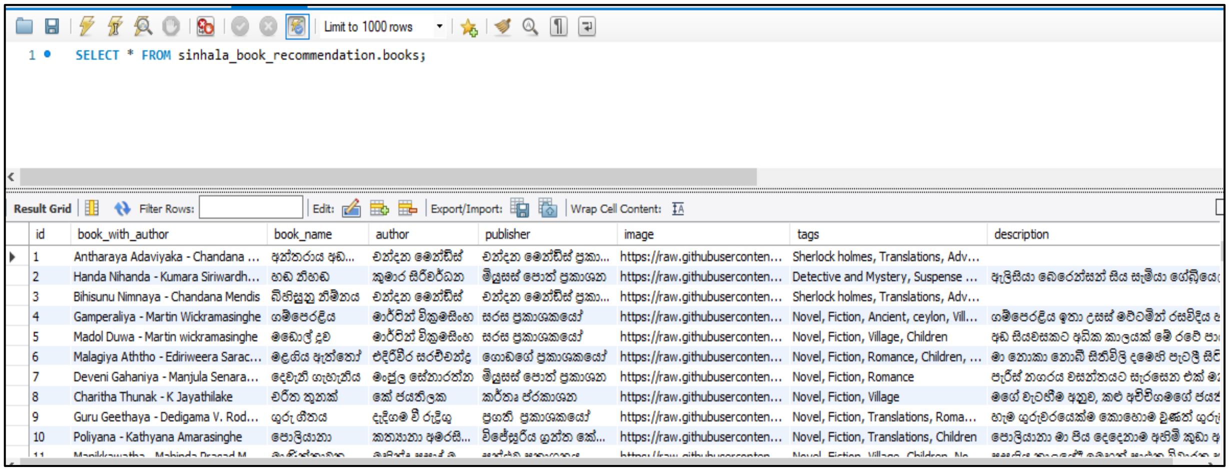
In order to store persistence data like book details, user login details my sql database is used. MySql workbench is used to manage data in mysql databse. Even there are multiple database like oracle, postgres available and can be used for the same purpose, Mysql was used as it is open source and can easily be managed. Mysql connector in python was used to connect the application with the database.



```
SELECT * FROM sinhala_book_recommendation.users;
```

login_name	password
abc	123

Figure 58: User stored data



```
SELECT * FROM sinhala_book_recommendation.books;
```

id	book_with_author	book_name	author	publisher	image	tags	description
1	Antharaya Adaviyaka - Chandana ...	අන්තරාය අඩ...	චන්දන මෙන්ඩිස්	චන්දන මෙන්ඩිස් ප්‍රකා...	https://raw.githubuserconten...	Sherlock holmes, Translations, Adv...	
2	Handa Nihanda - Kumara Siriwardh...	හඬ නිහඬ	කුමාර සිරිවර්ධන	මියුසස් පොත් ප්‍රකාශන	https://raw.githubuserconten...	Detective and Mystery, Suspense ...	ඇලිසියා බෙරෙන්සන් සිය සැමියා ගේබ්‍රියෙ
3	Bihisunu Nimnaya - Chandana Mendis	බිහිසුනු නිම්නය	චන්දන මෙන්ඩිස්	චන්දන මෙන්ඩිස් ප්‍රකා...	https://raw.githubuserconten...	Sherlock holmes, Translations, Adv...	
4	Gamperaliya - Martin Wickramasinghe	ගම්පෙරළිය	මාර්ටින් වික්‍රමසිංහ	සරස ප්‍රකාශකයෝ	https://raw.githubuserconten...	Novel, Fiction, Ancient, ceylon, Vill...	ගම්පෙරළිය ඉතා උසස් මට්ටමින් රසවිඳිය
5	Madol Duwa - Martin wickramasinghe	මඩොල් දූව	මාර්ටින් වික්‍රමසිංහ	සරස ප්‍රකාශකයෝ	https://raw.githubuserconten...	Novel, Fiction, Village, Children	අඩ සියවසකට අධික කාලයක් මේ රටේ පා
6	Malagiya Aththo - Ediriweera Sarac...	මළගිය ඇත්තෝ	එදිරිවීර සරච්චන්ද්‍ර	ගොඩගේ ප්‍රකාශකයෝ	https://raw.githubuserconten...	Novel, Fiction, Romance, Children, ...	මා නොකා නොබී සිතිවිලි දමෙහි පැටලී සිට
7	Deveni Gahaniya - Manjula Senara...	දෙවැනි ගැහැනිය	මංජුල සේනාරත්න	මියුසස් පොත් ප්‍රකාශන	https://raw.githubuserconten...	Novel, Fiction, Romance	පැරිස් නගරය වසන්තයට සැරසෙන එක් මා
8	Charitha Thunak - K Jayathilake	චරිත තුනක්	කේ ජයතිලක	කර්තෘ ප්‍රකාශන	https://raw.githubuserconten...	Novel, Fiction, Village	මගේ වැටහීම අනුව, කළු අච්චිගමගේ ජය
9	Guru Geethaya - Dedigama V. Rod...	ගුරු ගීතය	දැදිගම වී රුද්‍රිගු	ප්‍රගති ප්‍රකාශකයෝ	https://raw.githubuserconten...	Novel, Fiction, Translations, Roma...	හැම ගුරුවරයෙක්ම කොහොම වුණත් ගුරු
10	Poliyana - Kathyana Amarasinghe	පොලියානා	කත්‍යානා අමරසි...	විජේසූරිය ග්‍රන්ථ කේ...	https://raw.githubuserconten...	Novel, Fiction, Translations, Children	පොලියානා මා පිය දෙදෙනාම අහිමි කුඩා අ
11	Manikkawatha - Mahinda Prasad M	[illegible]	[illegible]	[illegible]	https://raw.githubuserconten	Novel, Fiction, Village, Children, No	[illegible]

Figure 59: Saved Book Data

```
import mysql.connector
```

4.7 Chapter Summary

The chapter explained how the implementation was done mainly using python programming language. Initially the data needed to be preprocessed to remove unwanted data. Then it describes what the methods available for sentiment analysis and how the own model was build and trained. The implementation of UI and authentication was described with screen shots. Finally, how the database is connected to the application was explained.

5. Evaluation and Results

5.1 Chapter Overview

Implementing any application without proper testing or evaluation is considered as incomplete system. In industry also when we implement the application, after the QA test we hand over Client for User Acceptance Test (UAT) and get the feedback of the client. The feedback is kind of evaluation of what we have implemented.

5.2 Evaluation

Evaluating a book recommendation system is essential to ensure that it provides meaningful and useful suggestions to users. There are several key metrics and methods available to evaluate the accuracy and the performance of a book recommendation system.

5.2.1 Accuracy Metrics

In terms of accuracy, we can use metrics such as precision, recall, and F1 score.

5.2.1.1 Precision

It calculates the proportion of recommended books which are actually relevant to the user’s preference. High precision means that the system provides relevant and accurate recommendations. It can be calculated by following formula.

Precision = (No of relevant recommendation / Total no of recommendation)

5.2.1.2 Recall

It calculates the proportion of a user's preferred books which were correctly recommended by the system. A high recall means that the system captures a significant portion of the user's preferences. It can be calculated by following formula.

Recall = (No of relevant recommendations / Total no of user's preferred books)

5.2.1.3 F1 Score

The F1 score is the harmonic mean of precision and recall. It provides a balance between these two metrics, considering both false positives and false negatives. A higher F1 score indicates a well-balanced system that both accurately recommends relevant books and captures a significant portion of the user's preferences. It can be calculated by following formula.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- If the precision is high but the recall is low, it means the system provides very accurate recommendations but may miss many relevant books. This will be suitable for users who prioritize quality over quantity.

- If the recall is high but the precision is low, it means the system captures many relevant books but also recommends a lot of irrelevant books. This will be suitable for users who want a broad range of recommendations.

- A high F1 score indicates a well-rounded system that balances precision and recall, offering both accuracy and coverage.

It's important to set an appropriate threshold for relevance when calculating these metrics. What is considered as relevant may different from one recommendation system to another and depend on user preferences.

The accuracy metrics should be considered along with other evaluation metrics, such as user engagement, diversity, user feedback, and online or offline evaluation, to provide a more reasonable assessment of the recommendation system's performance.

5.2.2 User Engagement Metrics

It is the measurement getting by user actions like Click Through Rate (CTR), Conversion Rate and Bounce Rate.

5.2.2.1 Click Through Rate (CTR)

It calculates the percentage of users who clicked on a recommended book.

5.2.2.2 Conversion Rate

It calculates the percentage of users who buy and read the recommended book

5.2.2.3 Bounce Rate

It calculates how many users ignore or not accept the recommended books without interacting the any results.

5.2.3 Diversity Metrics

The main categories are novelty and serendipity. Novelty measures how many diverse and non-redundant book were recommended. It ensures that the system should not give same type of books repeatedly. On the other hand, serendipity measures how often the system recommends books that are not expected but appreciated by users.

5.2.4 User Feedback

In order to collect feedback from users, surveys, reviews or direct interaction can be used to understand their preferences or how satisfy they are about the system. This metrics will be used for the implemented solution to calculate the accuracy and the performance.

5.2.5 Online Evaluation

When the application is deployed to the server and getting the feedback from users is considered online evaluation. Deploying the application in Heroku kind of cloud service requires to be registered and the service is not free, the online evaluation will not be performed.

5.2.6 Benchmarking

It is the comparison the application with other popular recommend system and verify how the application fits to the industry standards.

As per the paper (Raval and Khedkar, 2019) used item based collaboration filtering recommendation system and the Root Mean Square Error (RMSE) and Mean Absolute Error (MSE) methods were used for evaluation respectively. In the final results, their proposed method outperforms all the state-of-art methods.

Even though there are many evaluation metrics are available, the research (Kurmashov et al., 2015) implemented a book recommendation system which gives fast result based on collaborative filtering and use online survey for the evaluation because they realized that there is no database suitable for their task to evaluate the results. Therefore, they have selected independent readers and ask to provide a score from 1 to 10 based on the parameters like quality, convince and ease of use of the recommending system implemented. The higher score indicates the relevance of the recommendation.

There are two options to validate the system named as offline validation and online validation. For offline validation, I will be having user data and performing a standard machine learning training-test split in order to learn and train the model for the evaluation. Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) or any other evaluation function could be used.

The main point here is, this kind of evaluation cannot be done without interacting large amount of data. For online validation, a recommender model will be created based on information taken from other domains which is also called as cross-domain recommender system and test the system with live data. Since the implementing system do not have access to some live system, I will be focusing on finding a data set that is more than enough for the offline validation. In order to perform offline validation for the application, we can make use of the concept of precision-recall. Recall describes, what ratio of items that a user like will be actually recommended. And the precision describes out of all recommended items, how many items user actually will like. The main idea of any recommending system is recommending only items user likes. This is the optimal recommender and My target is to get as close as possible.

In order to validate the model further, expert authors will be contacted. He would check the recommended book list is matches with searched book or the list contain books which is preferred by the user.

According to the proposed solution, expecting accuracy level would be more than 80%. We can increase the accuracy by collecting and allocating more data for training. At the end of the project a user can find a best recommend books according to his preference and the rates given by other users. Once the recommended book is read the user may realize the accuracy of the application and no need to waste time on finding the books in everywhere. Once the model is developed, we can use it to make recommendation for that we need to save the desired model and restore it when we need to do recommendation through it

5.3 Chapter Summary

Evaluate the implemented system was discussed in this chapter with test results. In order to evaluate the system, some experts will be contacted and get their feedback. The accuracy of the evaluation will also be discussed in the phrase.

5.4 Project Plan and Timeline

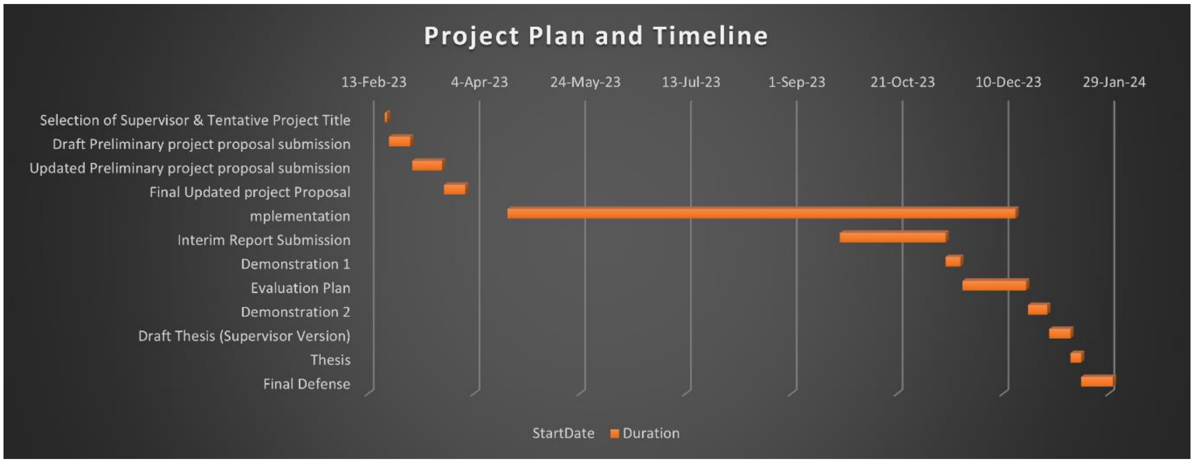


Figure 60: Project Plan and TimeLine

6. Conclusion and Future work

6.1 Chapter Overview

The conclusion chapter presents the final part of the application and discuss the aim and the objectives have been achieved successfully. Furthermore, the limitations and future work will be discussed in order for someone to add the missing feature and enhance the application.

6.2 Conclusion

The idea to implement an application for Sinhala Book recommendation came to the mind by surfing Facebook book related groups. Many people ask so many questions like is this book, I like books related to history and romance and please suggest me some books. The process of implementing such system began by finding any data set. Kaggle and many dataset providers have not provided any data set related to Sinhala books, thus a google form was created to collect the data which was a challenge. Even the form was share among groups, not able to collect data set as expected. Therefore, each member of groups were contacted and shared the form to be filled.

In parallel of collecting data some research papers were read and understand what are the system available and limitation of them. In literature review chapter all the details were discussed. Then the implementation started by learning Python programming language. In order to apply the hybrid model, First Contant based filtering was applied to the dataset and then Collaboration filter was applied on top of that. Based on the feedback provide by the authors or experts the system will be evaluated.

6.3 Limitations

Even most of the features were able to completed as per the proposal, some limitations were found to be implemented as an enhancement.

1. The age and gender fields were not considered for recommendation.
2. The latest books were not considered.
3. If user have not read any books, he is not able to use the system without select at least a book from the list.

6.4 Future Enhancement

The application was implemented as per the proposed system and there are some features that could be added as enhancements for the features of the system.

1. Since there is a drop down to select books, most books were selected from the top of

the list.

2. The drop-down book list was created in English. Most people entered data, suggested to display them in Sinhala as the project is related to Sinhala books.
3. The user should be able to select any books based on some categories like author and genre like history, romance, detective.
4. The dataset was limited like only around 4500 were able to be collected.

6.5 Chapter Summary

The main goals and objectives of the application were defined at the introduction chapter and the conclusion chapter discussed whether all of them have been successfully achieved. The limitation and future enhancements were discussed.

7. References

1. Barai, M.K., 2021. Sentiment Analysis with TextBlob and Vader. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/ (accessed 10.30.23).
2. Chandrasekaran, G., Antoanela, N., Andrei, G., Monica, C., Hemanth, J., 2022. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. Applied Sciences 12, 1030. https://doi.org/10.3390/app12031030
3. Dhanda, M., Verma, V., 2016. Recommender System for Academic Literature with Incremental Dataset. Procedia Computer Science 89, 483–491. https://doi.org/10.1016/j.procs.2016.06.109
4. Ijaz, F., n.d. Book Recommendation System using Machine learning.
5. Kurmashov, N., Latuta, K., Nussipbekov, A., 2015. Online book recommendation system, in: 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO). Presented at the 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO), IEEE, Almaty, Kazakhstan, pp. 1–4. https://doi.org/10.1109/ICECCO.2015.7416895
6. Machine Learning Project | Classification | Sentiment Analysis | Sinhala - YouTube [WWW Document], n.d. URL https://www.youtube.com/playlist?list=PL495mke12zYDPRGhXd6JGY5EUoksIVwYU (accessed 10.29.23).
7. Marappan, R., 2022. Create a Book Recommendation System using Collaborative Filtering. IJMEBAC 1, 44–46. https://doi.org/10.31586/ijmebac.2022.341
8. Mean Absolute Error [WWW Document], n.d. . C3 AI. URL https://c3.ai/glossary/data-science/mean-absolute-error/ (accessed 11.5.23).
9. Mercy Milcah Y, Moorthi K, Jansons Institute of Technology, 2020. AI based Book Recommender System with Hybrid Approach. IJERT V9, IJERTV9IS020416. https://doi.org/10.17577/IJERTV9IS020416
10. Murali, M.V., Vishnu, T.G., Victor, N., 2019. A Collaborative Filtering based Recommender System for Suggesting New Trends in Any Domain of Research, in: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). Presented at the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE, Coimbatore, India, pp. 550–553. https://doi.org/10.1109/ICACCS.2019.8728409
11. Raval, N., Khedkar, V., 2019. A Review Paper On Collaborative Filtering Based Moive Recommedation System 8.
12. Roy, D., Dutta, M., 2022. A systematic review and research perspective on recommender systems. J Big Data 9, 59. https://doi.org/10.1186/s40537-022-00592-5
13. Sarma, D., Mittra, T., Shahadat, M., 2021. Personalized Book Recommendation System using Machine Learning Algorithm. IJACSA 12. https://doi.org/10.14569/IJACSA.2021.0120126
14. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S., n.d. 9 Collaborative Filtering Recommender Systems.
15. Shah, K., 2019. Book Recommendation System using Item based Collaborative Filtering 06.
16. Tian, Y., Zheng, B., Wang, Y., Zhang, Y., Wu, Q., 2019. College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm. Procedia CIRP 83, 490–494. https://doi.org/10.1016/j.procir.2019.04.126
17. Tripathy, A., Agrawal, A., Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science 57, 821–829.

https://doi.org/10.1016/j.procs.2015.07.523

18. Wadikar, D., Kumari, N., Bhat, R., Shirodkar, V., 2020. Book Recommendation Platform using Deep Learning 07.
19. Wang, D., Liang, Y., Xu, D., Feng, X., Guan, R., 2018. A content-based recommender system for computer science publications. Knowledge-Based Systems 157, 1–9. https://doi.org/10.1016/j.knosys.2018.05.001
20. Wassan, S., Chen, X., Shen, T., Waqar, M., Jhanjhi, N., 2021. Amazon Product Sentiment Analysis using Machine Learning Techniques.
21. What is Mean Absolute Error [WWW Document], n.d. . Deepchecks. URL https://deepchecks.com/glossary/mean-absolute-error/ (accessed 11.5.23).
22. What is Natural Language Processing? | IBM [WWW Document], n.d. URL https://www.ibm.com/topics/natural-language-processing (accessed 11.12.23a).
23. What is Natural Language Processing? An Introduction to NLP [WWW Document], n.d. . Enterprise AI. URL https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP (accessed 11.12.23b).