A close-up of a document

Description automatically generated

Final Project Report

Title: BR recommendation system

Student Name: COULIBALY SOULEY KASSOUM

Student ID: GH1021968

Course: B141 Data Mining

GIT\_HUB\_LINK: https://github.com/juniorskg/recommandation\_system

## *Executive Summary:*

The project aim to make a recommendation system for book which were released 2003 in order to improve customer satisfaction . to do the task we use 2 different dataset. The metric we used to define the similarity between the book is cosine similarity and non \_ numerical values were encoded using TF-IDF Encoding (scikit learn, 2007).

## *Introduction:*

BR is startup online book shop which want to specialize in book recommendations. This project aim in creating a recommendation system for BR customer , we used correlation matrix for simulation measure between each books.

## *Dataset Selection:*

This project contain two data set accessible a this location <https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data?select=Books.csv>

The first data is the books data which contain the ISBN which is the unique number assign to each book , the book title , the book author ,the Year-Of-Publication, the publisher and the image URL S, M and L

The second data data is the rating data which contain the Id of the user which rate the book , the ISBN of the book they rate and there rating (kaggle, 2024)

The first data set contain 271360rows and 8 column

The second data set contain 1149780 rows and 3 columns

## *Data cleaning :*

* I define a function called central\_tendancy\_calculator to handle numerical missing data by calculating the mean , median, skewness, variance and outlier after that the function access between mean and median which is the most suitable central measure to replace the missing data and then it fill in the missing data (Medium, 2023).
* I defined a function called mode\_calculator to handle missing value of categorical data by replacing them by the mode and then I filled in the missing data.
* I removed the useless column in the book dataset ("Image-URL-S","Image-URL-M","Image-URL-L”)
* I checked if the ISBN is unique in the dataset
* I converted the year of publication in to numerical values and I remove the inconsistent year to covert them to null before filling the null values
* Put in the data the average of rating by each book and I put the result instead of book rating
* I removed user id as it is useless for the task
* I merged the two dataset by ISBN
* I only use the data a books which where release after 2003 and I put the ISBN as index of the dataframe
* I encoded non numerical data with Tfidf vectorizer and scaled the numerical data with minmaxscaler and I put them together in one dataframe

## *Cosine similarity:*

* I used the sklearn.metrics.pairwise cosine similarity to compute the cosine similarity and I rename the index to put the “ISBN” as index (scikitlearn, 2022)

## *Recomandation system:*

* I create a a function to give the top n book similar to the book the customer put

Where n is the number of book choose by the customer

# References

kaggle. (2024). *Book Recommendation Dataset.*

Medium. (2023). *Mastering Null Value Handling: A Comprehensive Guide to Replacing Missing Data in Your Dataset.*

scikit learn. (2007). *TfidfVectorizer.*

scikitlearn. (2022). *sklearn.metrics.pairwise.cosine\_similarity.*