User oriented book recommendation system with user interface

Team 2

312553008-陳沛圻, 312551074-陳柏翰

312553027-林煦, 312551106-曹豈銘

Motivation.

In modern book-selling websites, user feedback, which includes ratings and reviews, can be inconsistent.

To improve personalized recommendations, we'll perform sentiment analysis on user reviews of purchased books and calculate new recommendation scores alongside their original ratings.

New users will be categorized into user clusters based on their chosen tags, and association rules will be generated from these clusters.

The system will then filter and suggest books with the highest recommendation scores.

Problem statement.

Input

The dataset includes reader comments, ratings, and essential book details like publication dates and overall ratings.

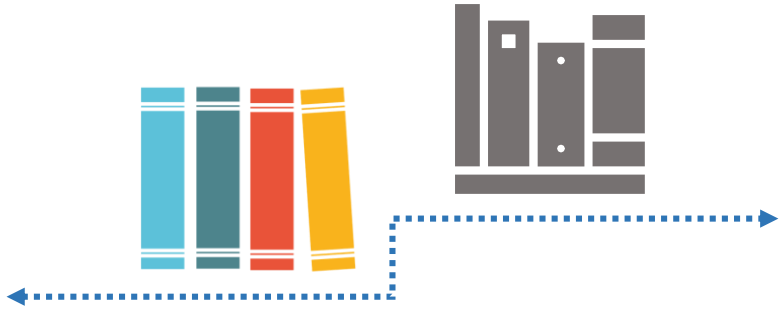
Process

Use readers' purchase data and reviews to cluster and discover item-related association rules for each reader cluster.

Output

Based on readers' historical book purchases and preference categories, recommend books tailored to each reader, thereby increasing platform sales.

Targeted performance.



Evaluation metric

To judge the quality of a model : MRR & MAP

Target

KNN-clustering & Memory-based with
User-based collaborative filtering two models.
Both up to 0.45 accuracy for MRR and 0.40 accuracy for MAP .

Ref : Tian, Yonghong, et al. "College library personalized recommendation system based on hybrid recommendation algorithm." ***Procedia CIRP*** 83 (2019): 490-494.

Evaluation metric.

MRR
Mean Reciprocal Rank

$$MRR = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{p_q}$$

Q : 推薦總數
p_q : 第q位第一個命中位置

MAP
Mean Average Precision

$$MAP = \frac{\sum_{q=1}^{Q} Avep(q)}{Q}$$

$$Avep(q) = \frac{\sum_{k=1}^{n} p(k)\times rel(k)}{\#\ relevant\ \ item}$$

Q: 推薦總數
除了考量第一個命中物品的位置之外，也考量第二個、第三個 … 第N個命中物品的位置，並把每一個命中物品位置的 Top k precision加起來做平均。

Ref : Chen, Mingang, and Pan Liu. "Performance evaluation of recommender systems." International Journal of Performability Engineering 13.8 (2017): 1246.

Data overview.

FROM

kaggle.com

SOURCE

Amazon

NUMBERS

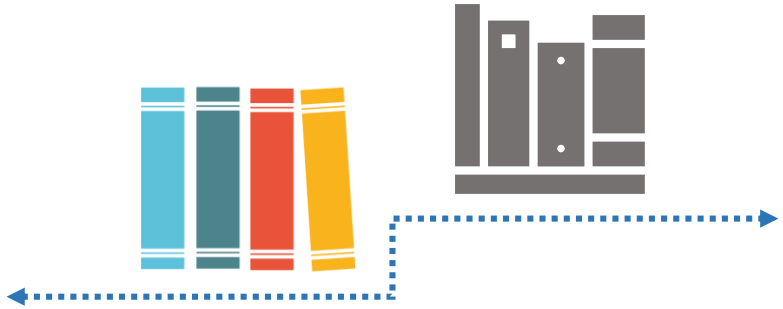
Attributes: 19
Users: 1140
Books: 212404
Ratings: 300000

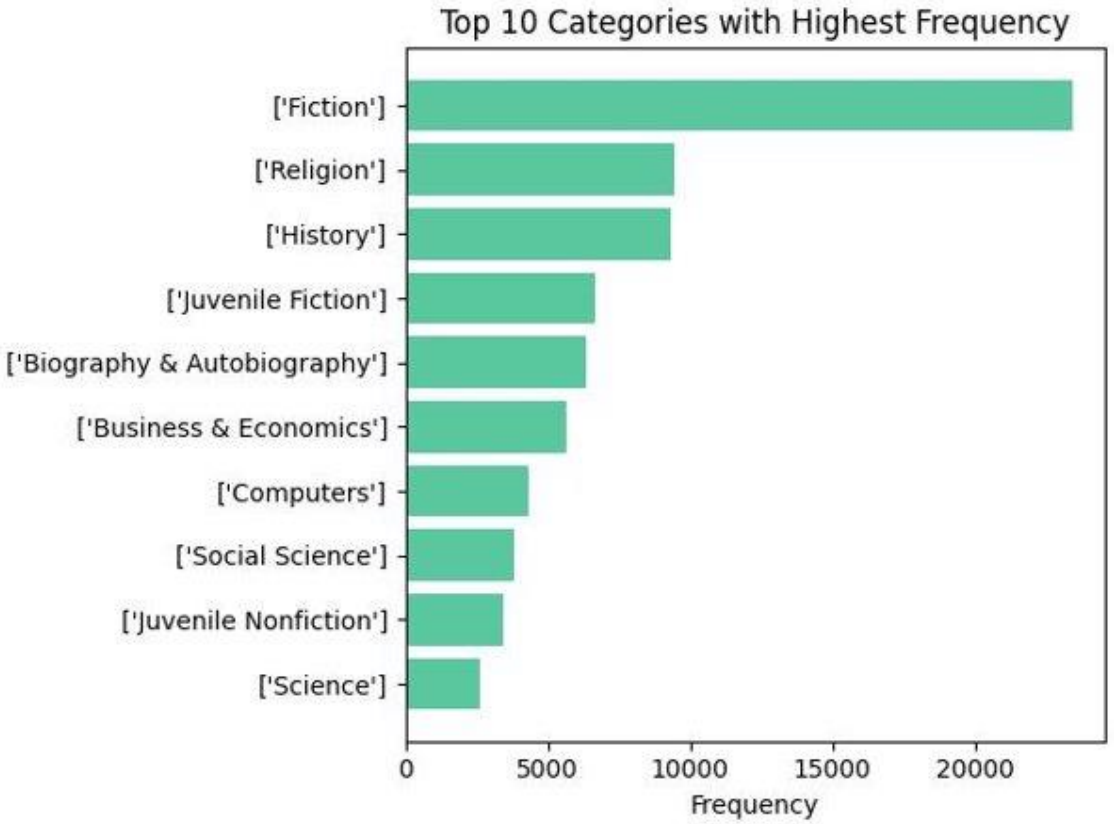
Amazon book review dataset : https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

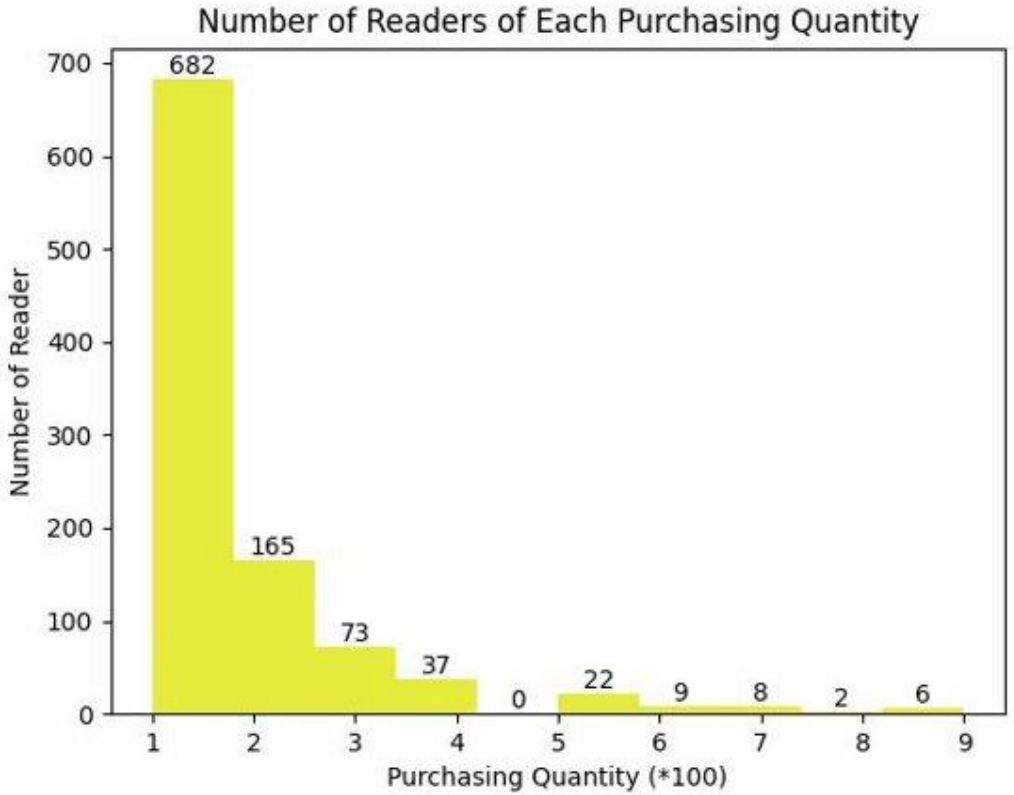
Data description.

Book Reviews	Book Details
ID	Title
Title	Description
Price	Authors
User_ID	Image
ProfileName	PreviewLink
Review / Helpfulness	Publishers
Review/Score	PublishedDate
Review/Time	InfoLink
Review/Summary	Categories
Review/text	RatingsCount

Data analysis.







Challenge.

Too many type of category

Filter the main categories
Using clustering to create new category type

How to evaluate the similarity of users

Extend features about the preferences of user

Data preprocessing

Cleansing

- Remove unknown user data
- Remove duplicated data

Feature Encoding

- One hot encoding

Feature Extension

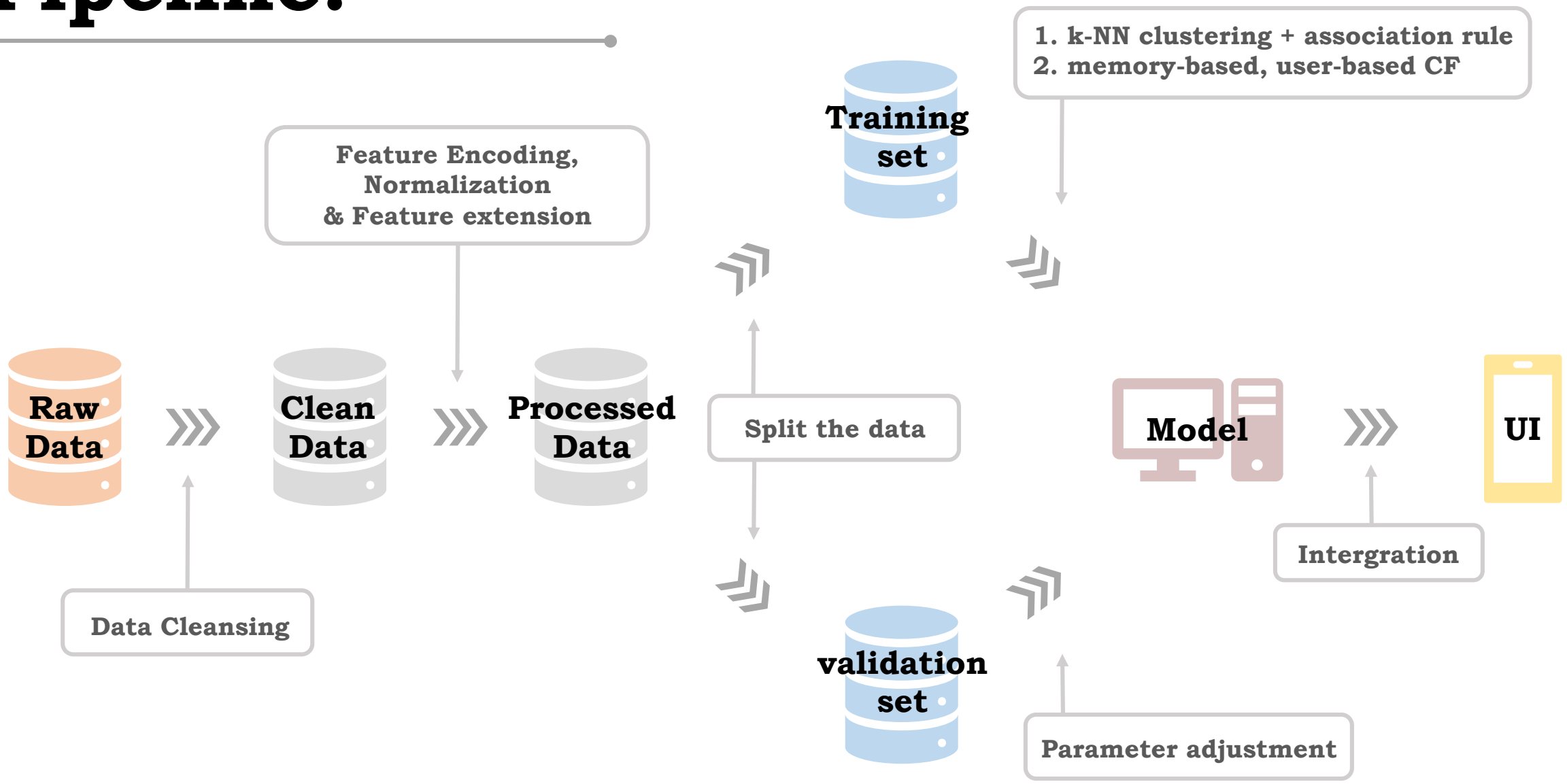
- Fine-tuning BERT for sentiment analysis on review content (revision in next page)
- Calculating new rec. score with review content score & rating

BERT Model

- 未fine tuning過的BERT模型並不知道針對書籍評論的情感指數
- 使用cleansing後的redundant data針對BERT模型做fine tuning
- BERT
  - Input: review
  - Label: rating score(1-5)，代表5個等級的情感指數
- 為了避免評論和評分的不一致所造成的影響[Ref]，我們會結合rating score和fine tuning後的BERT模型算出的情感指數，作為推薦指數
  - 推薦指數 = 0.7*情感指數+0.3*rating score
- 為了驗證以上想法是否正確，我們會比較結合兩種feature前後的推薦效果(比較MRR & MAP)

Ref: Aralikatte, Rahul, et al. "Fault in your stars: an analysis of android app reviews." *Proceedings of the acm india joint international conference on data science and management of data*. 2018.

Pipeline.


1. k-NN clustering + association rule
2. memory-based, user-based CF
Training set
Feature Encoding, Normalization & Feature extension
Raw Data
Clean Data
Processed Data
Split the data
Model
UI
Intergration
Data Cleansing
validation set
Parameter adjustment

User interface.

User Input

The interface will fetch the preference of user

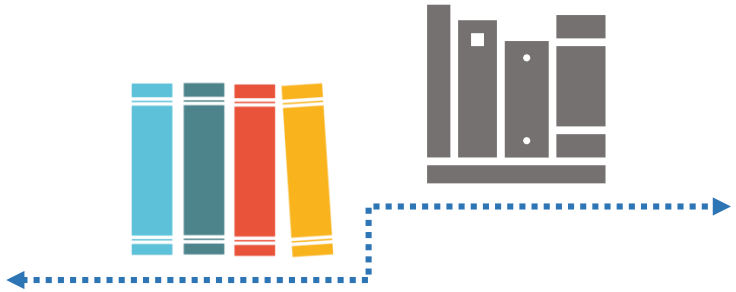
Model

Backend performs rec. calculations

Output

The system will show top 10 books with highest rec. scores

Environment.



Platform

Windows10

Language

python

Development tool

VS code

Schedule.

Action plan	Oct				Nov				Dec			
	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Dataset Determination and Topic Discussion	■	■										
Data Cleansing & Preprocessing		■	■	■								
Feature Engineering				■	■	■						
Clustering Model Evaluation						■	■	■	■	■		
User-Based CF Evaluation						■	■	■	■	■		
Project Report										■	■	■

Thank you.