# AIWR ASSIGNMENT 2

# **Team members** : **[ {Pooshpal Baheti ,PES1UG20CS283} , {Khushi Donthi,PES1UG20CS210} ,{Mihir Santosh Sanjay,PES1UG20CS253}]**

# **Section 1 [dataset] :**

Dataset : Books Dataset  
https://www.kaggle.com/datasets/saurabhbagchi/books-dataset

# **EDA (2 marks)**

EDA is done to :

* + Uncover underlying structure
  + Extract important variables from the dataset
  + Detect outliers and anomalies(if any)
  + Evaluate Missing or Null Data
  + Assess data quality

# **Preprocessing steps of raw data (2 marks)**

* 1. Data Cleaning
  2. Data Transformation - applying it on Book Title
  3. Data Reduction - Drop 3 columns in books\_data
  4. Feature Selection & Importance

# **suitability of Evaluation metrics (2 marks)**

The Evaluation metric used are :

## **Long Tail Plot**

recmetrics.long\_tail\_plot()

The Long Tail plot is used to explore popularity patterns in user-item interaction data. Typically, a small number of items will make up most of the volume of interactions and this is referred to as the "head". The "long tail" typically consists of most products, but make up a small percent of interaction volume.

## **Mar@K and Map@K**

recmetrics.mark()

recmetrics.mark\_plot()

recmetrics.mapk\_plot()

Mean Average Recall at K (Mar@k) measures the recall at the kth recommendations. Mar@k considers the order of recommendations, and penalizes correct recommendations based on the order of the recommendations. Map@k and Mar@k are ideal for evaluating an ordered list of recommendations. .

**Coverage**

recmetrics.prediction\_coverage()

recmetrics.catalog\_coverage()

recmetrics.coverage\_plot()

Coverage is the percent of items that the recommender is able to recommend. It is depicted by this formula.

**Novelty**

recmetrics.novelty()

Novelty measures the capacity of a recommender system to propose novel and unexpected items which a user is unlikely to know about already. It uses the self-information of the recommended item and it calculates the mean self-information per top-N recommended list and averages them over all users.

## **MSE and RMSE**

recmetrics.mse()

recmetrics.rmse()

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used to evaluate the accuracy of predicted values that such as ratings compared to the true value, y. These can also be used to evalaute the reconstruction of a ratings matrix.

# **Results (4 marks)**

MSE: 12.273111617729947

RMSE: 3.503300103863491

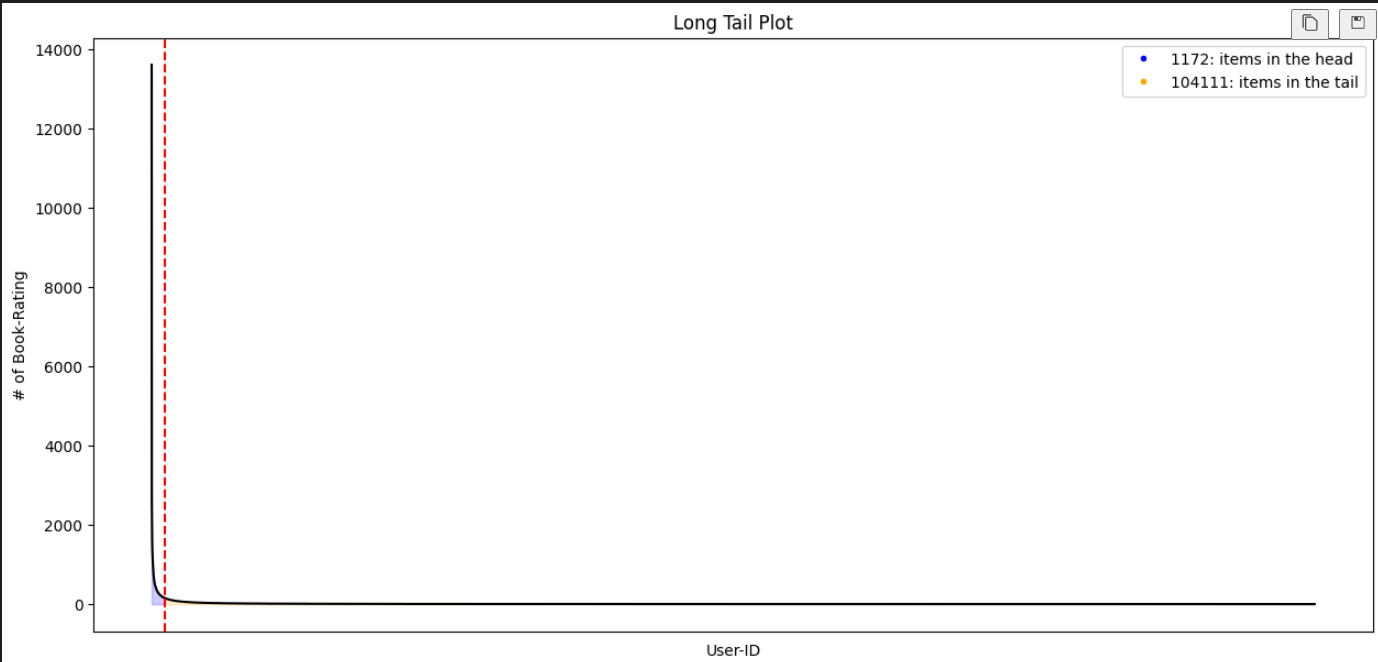
Catalog Coverage Grah in Notebook

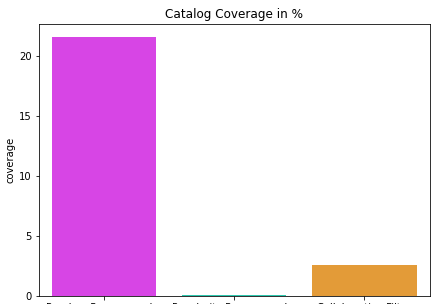
Novelty Measure graph in notebook

# **Handling Domain specific challenges (2 marks)**

1. Book Titles: There was a need to make Titles uniform to understand rating well.
2. Ratings matrix is sparse

# **Demonstration and Report (4 marks) - attach screenshot of your demo**





# 