**Recommender System Collab Project**

Time Length: 3 Days

Data Set: provided

**Goal of this Collaborative Project:**

-Learn to work in a collaborative environment

-Build a Hybrid recommendation engine to recommend books to different users.

Explore Recommendation Models in details, understand strengths and weaknesses of each model, build a Hybrid model to improve model performance and accuracy.

**About the dataset:**

*Book-Crossing Dataset*

Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

Format:

The Book-Crossing dataset (CSV File) comprises 3 tables.

BX-Users

Contains the users. Note that user IDs (`User-ID`) have been anonymized and map to integers. Demographic data is provided (`Location`, `Age`) if available. Otherwise, these fields contain NULL-values.

BX-Books

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (`Book-Title`, `Book-Author`, `Year-Of-Publication`, `Publisher`), obtained from Amazon Web Services.

Note that in case of several authors, only the first is provided.

URLs linking to cover images are also given, appearing in three different flavours (`Image-URL-S`, `Image-URL-M`, `Image-URL-L`), i.e., small, medium, large. These URLs point to the Amazon web site.

BX-Book-Ratings

Contains the book rating information. Ratings (`Book-Rating`) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

**PROJECT TASKS:**

1) Explore your dataset, perform a simple EDA to pinpoint the main characteristics of your dataset (ex: density of ratings, number of Null values, any duplicates?)

2) MODELLING - Recommendation Engine

A) Content-based Filtering Approach

Use a Content based filtering model to recommend products which are similar to the one that a user has liked in the past.

Calculate the similarity between the profile vector of each user and item vector of each book using 3 different methods:

- Euclidean Distance

- Cosine Similarity

- Pearson’s correlation

Explain in detail the advantages and shortcomings of each similarity measures.

Based on the cosine similarity measures, recommend the top 10 movies to any user present in your dataset.

B) Collaborative Filtering Approach

Use a Collaborative filtering model that uses “Customer Behavior” for recommending books.

- Explore the User-User similarity collaborative filtering approach

- Explore the item-item similarity collaborative filtering approach

Explain in detail the difference between the two approaches.

C) Collaborative Filtering Advanced: Matrix Factorization

In the two previous models (content and basic collaborative filtering approaches), we do not have the ratings for each book given by each user. We must find a way to predict all these missing ratings. For that, we have to find a set of features which can define how a user rates the books. These are called **latent features**.

Explore Matrix Factorization

Useful links:

[https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf](about:blank)

<http://nicolas-hug.com/blog/matrix_facto_1>

<http://nicolas-hug.com/blog/matrix_facto_2>

<http://nicolas-hug.com/blog/matrix_facto_3>

<http://nicolas-hug.com/blog/matrix_facto_4>

Use the Matrix Factorization technique to extract the most important features from the existing features.

Use the Stochastic Gradient Descent to solve your optimization problem.

Explain in detail the advantages and shortcomings of this Matrix factorization approach.

D) Evaluate your Recommendation Engines with the following metrics:

-Recall

-Precision

-RMSE (Root mean squared Error)

-Mean Reciprocal Rank

-MAP at k (Mean average Precision at cutoff k)

-NDCG (Normalized discounted Cumulative Gain)

E) Build a hybrid Recommender Engine that combines Content-based filtering and Collaborative filtering approach

GOAL:

- take advantage from both representation of the content as well as the similarities among users.

- Improve overall recommendation accuracy and power.

3) PRESENTATION

On Friday, starting at 5pm, Present your project, divide it across 5 groups covering the 5 main topics of this project.

1st group: Data Cleaning and EDA

2nd group: Content Based Filtering Approach

3rd group: Collaborative Filtering Approach

4th group: Matrix Factorisation

5th group: Hybrid recommendation Engine

**ADDITIONAL PROJECT INFORMATION/TASK:**

You will have to work as a team to build a powerful Hybrid recommendation Engine.

All your work should be well documented and your code highly readable.

Document your job on Github with all necessary files and Project Description.

Write an article about this project and publish it on Medium.

ADDITIONAL USEFUL LINKS:

<https://www.analyticsvidhya.com/blog/2016/06/quick-guide-build-recommendation-engine-python/>

<https://www.cs.toronto.edu/~rsalakhu/papers/rbmcf.pdf>

<https://arxiv.org/pdf/1511.06939.pdf>

<https://medium.com/@tanaykarmarkar/explainable-restricted-boltzmann-machine-for-collaborative-filtering-6f011035352d>

<https://towardsdatascience.com/paper-summary-matrix-factorization-techniques-for-recommender-systems-82d1a7ace74>