Northeastern University, Khoury College of Computer Science

# CS 6220 Data Mining | Project Proposal

Yu Wang
wang.yu25@northeastern.edu

## Objective:

The objective of this project is to develop a book recommendation system that can accurately suggest books to users based on user reading histories (inferred from rating history) and book genre prediction (content-based). The system leverages Book Recommendation Dataset, which contains user demographics, book metadata, and ratings, aiming at recommending the most-likely interested books to the user and providing a personalized reading experience.

## Motivation:

With the rise of web services, recommender systems are becoming more and more important in our daily lives. Right now, book recommendations are often made by looking at simple factors like the genre of books a user reads or what is popular/trending. This recommendation approach does not dig deep into readers' preferences, often leading to generic suggestions that don't quite hit the mark. This project aims at designing a book recommendation system that digs deeper by using information about users' rating histories and book metadata and ratings to make more personalized suggestions. A successful system can help users discover books they really love but might not have found, targeting a better user satisfaction.

## Background:

**Literatures Planning to Review**

- Item-Based Collaborative Filtering Recommendation Algorithms, Sarwar, Karypis, Konstan, and Riedl (2001)
- Matrix Factorization Techniques for Recommender Systems, Koren, Bell, and Volinsky (2009)

- Content-Based Recommendation Systems, Pazzani and Billsus (2007)
- Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, Adomavicius and Tuzhilin (2005)
- Recommender Systems Handbook, Rokach, Shapira, and Kantor (2011)
- Amazon.com recommendations: item-to-item collaborative filtering, Linden, Smith, and York (2003)
- Deep Learning based Recommender System: A Survey and New Perspectives, Shuai Zhang, Lina Yao, Anxin Sun, and Yi Tay (2019)

**Dataset Planning to Use**

The Book Recommendation Dataset is a great fit for this project because it provides detailed data in user demographic, book information, and ratings. This provides a solid foundation for developing a personalized book recommendation system.

“The Book-Crossing dataset comprises 3 files:

- 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.
- 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 flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon web site.
- 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.”

## Proposed Approach:

**Objective to Satisfy**

The project aims to provide personalized book recommendations by analyzing user demographic data, book titles, ratings from the Book Recommendation Dataset, to understand user preferences deeply and suggest books that the user is likely to love.

**Data Pre-processing**

Data pre-processing including handling missing or incorrect values, and data format conversion. Possible steps include but are not limited to:

1. Drop the last several columns containing image URLs which will not be useful for analysis.
2. Some columns may not be loaded correctly, make required corrections to justify the incorrect data loads (e.g. bookAuthor is incorrectly loaded with bookTitle, yearOfPublication incorrectly loaded as 'DK Publishing Inc').
3. Check and handle missing value. Change NaN value to 0.
4. Convert data type to integer (e.g. Age, Year-Of-Publication, Rating).
5. Filter out invalid publication years (e.g. years that are obviously not contemporary).

**Class Imbalance Remediation**

Using a hybrid method that combines oversampling the minority class and undersampling the majority class to optimize the class distribution. Oversampling increases the number of instances in the minority class by generating synthetic samples (e.g. SMOTE). Undersampling reduces the number of instances in the majority class to balance the dataset.

The project may also analyze rating distribution and apply a threshold to focus on relevant data.

**Learning Algorithms**

The project will use content-based filtering to suggest books similar to those a user has liked in the past and is viewing currently.

Consider using the **Word2Vec** model. The Word2Vec model converts each word in a book title into a vector that has a mathematical representation, and looks for the most similar genre based on cosine similarity. The genre most similar to the vector representation of book title will be selected as the genre predictor.

**Result Validation and Evaluation**

Using metrics such as precision and recall to evaluate the performance of the recommendation system.