M1 Data Mining Project Proposal: Analyzing Reader Patterns and Genre Associations in the Goodreads Dataset*

[*] CS 4412: Data Mining

1st David Holland
Student at Kennesaw State University
Marietta GA
dholla36@students.kennesaw.edu

***Abstract*—This proposal outlines a data mining project focused on the 2017 Goodreads dataset. This study aims to discover relationships between book genres, user behaviors, and how readers structure their reviews. We will be applying different techniques such as Association Rule Mining, K-Means Clustering, and Latent Dirichlet Allocation. This proposal seeks to understand how readers associate different genres in an attempt to define different segments of the reading community.**

***Index Terms*—Data Mining, Goodreads, Association Rules, Clustering, Text Mining, Pattern Discovery**

I. Introduction

The digital age has transformed reading from a personal, solo activity into a social activity. Various platforms like Goodreads and Fable provide a massive database of user-generated data through ratings, reviews, and custom shelves. While many similar projects exist to predict user's rating there is more value in understanding how users engage with literature. This project focuses on the UCSD Goodreads dataset to find these patterns.

II. Description

This project will utilize the Goodreads Book Datasets (2017) curated by Julian McAuley at UCSD, Rishabh Misra and Ndapap Nakashole

A. A. Source and Scale

- **Source:** https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html
- **Size:** The full dataset contains metadata for 2.36 million books and 15.7 million detailed review texts. For this project, a subset (e.g., the "Young Adult" or "History" category) will be utilized to maintain computational feasibility while ensuring a sample size of over 100,000 records.
- **Format:** Data is provided in the CSV and json format, requiring significant preprocessing and flattening.

B. Key Features and Schema

The dataset is split into two primary components:

1) **Metadata (*goodreads_books.json*):** Contains `book_id`, `title`, `authors`, `average_rating`, and importantly, `popular_shelves` (user-defined tags).
2) **Reviews (*goodreads_reviews.json*):** Contains `user_id`, `book_id`, `rating`, `review_text`, and `n_votes`.
3) **User Interactions (*goodreads_interactions.csv*):** Contains `user_id`, `book_id`, `is_read`, `rating`, and `is_reviewed`.

III. C. Data Quality and Preprocessing

Known issues include many user-defined shelves that share common tags like "to-read" and varying review lengths. This is most likely caused from Goodread's ability to easily add books to user's "to-read" shelve. Preprocessing will involve tokenization, stop-word removal for text mining and transforming the nested JSON shelves into a transaction-style for association mining.

IV. Discovery Questions

This project moves beyond prediction models to investigate the following questions the team had:

- What are the hidden associations between specfic book genres and user-defined "shelves"? (Do readers of "Romantasy (Romance Fantasy) also show high interset in "Historical Romance" or other fantsay genres?)
- Can we identify distinct Reader Personas based on interaction metrics? By clustering users based on their average rating variance, review frequency, and review length our goal is to discover if it is possible to classify readers as a critical reviewer or casual consumer.
- What type of patterns emerge from highly-voted reviews compared to low-voted ones?

Identify applicable funding agency here. If none, delete this.

V. Planned Techniques

The analysis will follow a multistage data mining pipeline as visualized in Figure 1.

Analysis Pipeline Diagram

JSON Parsing → Cleaning → Feature Extraction → [Mining Modules] → Evaluation

Fig. 1. Planned Data Mining Workflow.

A. Association Rule Mining (FP-Growth)

We will treat each user's shelves as a transaction. Using the FP-Growth Algorithm to discover rules with high lift and confidence to identify non-obvious genre relationships. Our goal is to move beyond the standard "Fantasy" label and discover more sub-genre patterns.

B. Clustering (K-Means)

To answer the second discovery question, we will apply K-Means clustering on standarized numerical features derived the user dataset. We will be using the "Elbow Method" and "Silhoutette Score" to determine the optimal number of reader segments.

C. Text Mining (LDA - Topic Modeling)

Using the Latent Dirichlet Allocation (LDA) algorithm, we will perform topic modeling on the `review_text`. This will allow us to discover the primary topics of discussion within different genres without pre-defining what those topics are.

VI. Preliminary Timeline

The project will proceed according to the following milestones

TABLE I
Project Milestone Timeline

Milestone	Period	Key Activities
M2	Weeks 8	Applying KDD process and Association Rule Mining
M3	Weeks 11	Clustering Analyses, Classification Techniques, detecting Anomalies, and Communicating results
M4	Weeks 14	Text Mining (LDA), Final Report

A. Anticipated Challenges

- Our primary challenge is lack of experience in analyzing data sets.
- Another challene is the volume of data. The Goodreads dataset is comprised of multiple csv and jsons files that is multiple gigabytes compressed. Handling this in local memory will require using pandas chunking or focusing on smaller genre subsets. Text data can also be "noisy" and will require filtering to produce meaningful topics.
- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

B. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(**??**)", not "Eq. (**??**)" or "equation (**??**)", except at the beginning of a sentence: "Equation (**??**) is . . ."

Acknowledgment

The author would like to thank Julian McAuley for providing the Goodreads dataset and Kennesaw State University for the computational resources.

References

[1] Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys'18.

[2] Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL'19