Finding Similar Book Reviews

Algorithms for Massive Datasets (2024/2025)

Antonios Tsipoulakos, Erasmus Student in Computer Science, University of Milan

June 24, 2025

Dataset Description

The project uses the **Amazon Books Reviews** dataset, publicly available on Kaggle [1] under a CC0 license. The version used was downloaded in **June 2025**.

Files considered:

- `Books_rating.csv` : *Main file containing the* `review/text` *field.*

Data Organization and Preprocessing

The dataset was preprocessed as follows:

- Removed entries with missing reviews.
- Tokenized each review into a `set` of lowercase words.
- Used a global variable to limit the sample size for faster execution.

Algorithm and Implementation

We implemented the **Jaccard Similarity** from scratch to detect pairs of similar reviews:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are sets of words.

The algorithm computes similarities between all possible pairs in the selected subset. The top 5 most similar pairs are returned.

Scalability

A global variable controls whether the code uses a small sample or the full dataset. This ensures:

- Quick testing and Colab compatibility
- Scalability to full-size datasets (all logic generalizes)

[1] `https://www.kaggle.com/datasets/rajeevw/amazon-books-reviews`

Experiments and Results

Using the first 100 reviews:

- All pairwise similarities were computed ($\binom{100}{2} = 4950$ pairs).
- Top 5 pairs were printed with their score and review excerpts.

Discussion

The method captures lexical similarity effectively with minimal preprocessing. While limited to word-overlap, Jaccard is a strong baseline. Future improvements may involve TF-IDF or embedding-based similarity.