1. Feature Extraction and Cleaning

- The dataset (books_cleaned.csv) contained book titles with inconsistencies such as special characters (&).
- Cleaning steps performed:
    - Removed special characters and HTML encodings.
- Feature extraction focused solely on book titles, assuming that titles reflect the genre, theme, or similarity to other books.

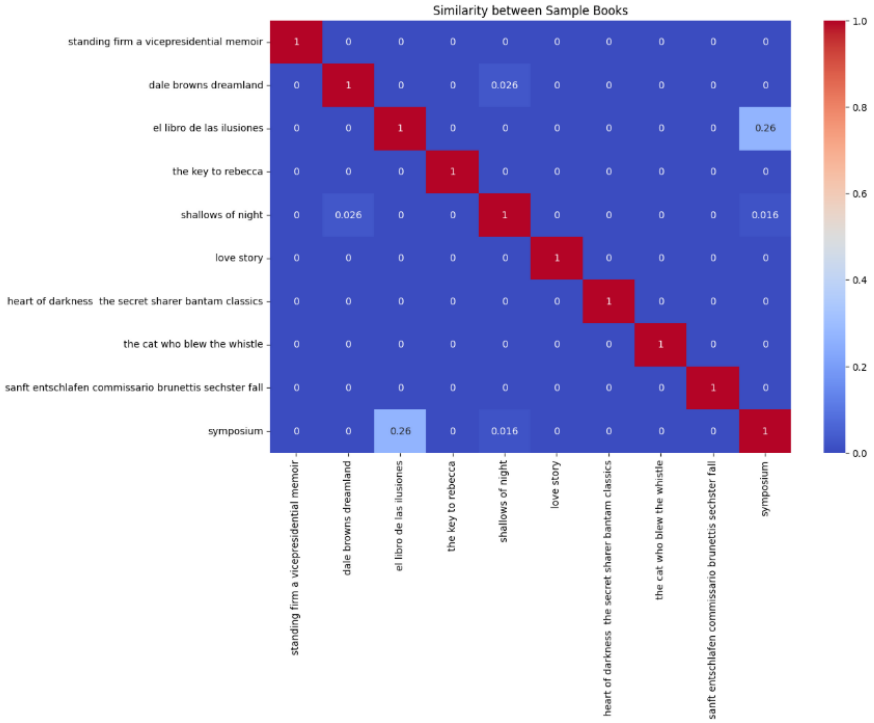
2. Vectorization Technique Used

- Technique used: TF-IDF (Term Frequency-Inverse Document Frequency)
- Reason for choosing TF-IDF:
    - TF-IDF emphasizes important and distinctive words in book titles.
    - It reduces the influence of common words like "the", "of", "a", which would otherwise dominate.
    - This technique helped highlight unique aspects of each title, allowing more intelligent matching.

TF-IDF was selected because it makes the recommendation system more sensitive to the specific nature of the book titles rather than just frequency of words.

3. Similarity Matrix Screenshot

Below is the similarity matrix generated using TF-IDF vectorization and cosine similarity:



Observations:

- Most off-diagonal similarities are low, indicating that most book titles are quite distinct.
- A few titles show moderate similarity, suggesting slight overlap in genres or themes.
- The diagonal elements are all 1, representing perfect similarity of a book with itself.

This matrix provides the foundation for generating recommendations.

4. Output Examples

Top recommendations for **"a matter of honor"**:

Rank	Recommended Book
1	kane abel
2	to cut a long story short
3	middlesex a novel
4	the honor of the queen honor harrington series book 2
5	the professor and the madman
6	mere christianity
7	riding the bus with my sister a true life journey
8	az of behaving badly
9	the gospel of judas a novel
10	citizens a chronicle of the french revolution

5. Insights and Observations

- Series connection: Titles containing similar keywords such as "honor" led to recommendations related to similar series.
- Genre alignment: Books covering historical events, philosophical themes, and biographies were closely linked to the query book.
- Meaningful matching: Even without full book descriptions, TF-IDF captured deeper textual similarities based on important words.

This shows that careful title-based analysis can lead to relevant recommendations without needing full metadata or user reviews.

6. Comparative Notes: Content-Based Filtering vs Collaborative Filtering

Aspect	Content-Based Filtering (used here)	Collaborative Filtering
Based on	Features of the item (text in book titles)	User behavior and interactions (ratings, reviews)
New Item Problem	No issue; new books can be recommended immediately	New books cannot be recommended until they receive ratings
Personalization	Limited; similar items for all users	High personalization based on user taste
Cold Start	No cold start for books	Cold start for new users and new books
Data Needed	Only item metadata (titles)	Requires large user-item interaction data
Example	Recommend books similar to "a matter of honor" based on text similarity	Recommend books that other users who liked "a matter of honor" also liked

Content-based filtering is a good starting point when user interaction data is not available, while collaborative filtering is useful when a rich dataset of user preferences exists.