Personalized Book Recommendations from Amazon Reviews

Background and context to the problem statement:

In today's digital age, online shopping platforms such as Amazon have become go-to destinations for consumers seeking personalized product recommendations. As avid readers turn to these online platforms for book purchases, the sheer volume of available titles can overwhelm readers, hampering their ability to discover relevant books. Our goal is to leverage the wealth of data provided by the *Amazon Books Reviews* dataset to craft a robust recommendation system that simplifies the users' book discovery process. Through our project, we seek to empower users with curated books that resonate with their unique preferences, ultimately enhancing their journey through the world of literature.

Identification and description of the data set:

1. Data identification and source: We will use the "*Amazon Books Reviews*" dataset from Kaggle, a comprehensive collection of two sources. The first source is from the *Amazon Review Dataset*, encompassing feedback from 3 million users on 212,404 distinct books. The second source is built from the *Google Books API*, which provides detailed information about the books mentioned in the first file.

2. Relevance: The dataset contains detailed information about user ratings of various books, which is crucial for developing a recommendation system. It includes user-specific columns such as *user_id*, *profileName, review/helpfulness, and review/time*, alongside multiple attributes related to the books, such as *Id, Title, Price, review/score, review/summary,* and *review/text.*

Proposed ML techniques on applying to solve the problem:

Based on the dataset, we plan to perform data analysis with visualizations, NLP analysis (sentiment analysis, text classification, text clustering), and a recommendation system (Content-Based Filtering, Collaborative Filtering).

1. Data analysis with visualizations
    - Wordcloud for the review texts, bar charts for the top categories, pie charts for the distribution of ratings, and histograms for review lengths.
2. NLP Analysis
    - Sentiment Analysis: Apply sentiment analysis to the review texts to classify them into positive, neutral, or negative sentiments.
    - Text Classification: Classify the reviews into predefined categories based on their sentiment using models like Naive Bayes, Support Vector Machines, or BERT (Bidirectional Encoder Representations from Transformers).
    - Text Clustering: Find patterns or group similar reviews by unsupervised ML methods like K-Means or topic modeling techniques like Latent Dirichlet Allocation (LDA).
3. Recommendation System
    - Content-Based Filtering: From variables such as categories, authors, and titles to recommend similar books, implement TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with cosine similarity.
    - Collaborative Filtering: Implement Singular Value Decomposition (SVD) or memory-based approaches like KNN (k-nearest neighbors) from user-item interaction matrices.