Data Preprocessing and Cleaning Report

This report outlines the data preprocessing and cleaning steps taken on a dataset involving product reviews and information. It involves a comprehensive workflow from data loading, cleaning, preprocessing, to database integration.

Data Preparation

Initially, necessary libraries for data manipulation and natural language processing, such as pandas, numpy, and NLTK, were imported. The data, initially in JSON format, was converted to CSV for easier manipulation. Duplicates were removed to ensure data uniqueness.

Cleaning and Preprocessing

Specific cleaning operations were applied to columns like Average_Review and Brand_Name to standardize the data. For NLP, text data underwent tokenization, punctuation removal, conversion to lowercase, stopwords removal, and lemmatization. These steps are crucial for preparing the data for meaningful analysis or machine learning applications.

Database Integration

The cleaned and processed data was structured into a relational database schema, involving operations such as creating tables and inserting data into a SQLite database. This allows for efficient data retrieval and storage.

Challenges Faced

I faced challenged when integrating with MySQL. Despite efforts, I encountered connectivity issues that prompted me to shift to SQLite.

Conclusion

I took an effective approach to preprocessing and cleaning product review data, highlighting the use of various Python libraries and SQL for data management. The process enhances the data's value for further analysis or machine learning tasks.

Database Schema Diagram