Product Data Cleaning and Preparation
Etini Akpayang | 2025-02-06

1. Introduction

This project focuses on cleaning and preparing product data for subsequent analysis. The main objectives are to:

- Standardize column names for consistency.
- Handle missing values and duplicate entries.
- Optimize product titles by generating a concise version (≤50 characters) to improve SEO and readability.

2. Methods

Data Loading & Inspection

- **Data Import:**

  The product dataset (3,847 rows, 6 columns) is loaded from a CSV file into an R dataframe (product_df).

- **Initial Data Exploration:**

  Functions like glimpse() and head() are used to inspect the dataset's structure, variable types, and sample entries.

Data Quality Checks

- **Missing Values:**

  The script identifies missing values across columns. For example, missing counts were found in BULLET_POINTS, DESCRIPTION, PRODUCTTYPEID, and ProductLength.

- **Duplicates:**
  A duplicate check on PRODUCTID revealed 217 duplicate entries.

Data Cleaning

- **Column Renaming:**

  All column names are standardized to lowercase with underscores (e.g., PRODUCTID → product_id).

- **Handling Missing Values:**

  Rows with missing critical values in product_type_id or product_length are removed, and missing text fields are replaced with "Unknown."

- **Duplicate Removal:**

  After handling missing values, duplicate rows are removed, reducing the dataset to 3,541 rows.

Title Optimization

- **Shortening Function:**

  A custom R function, shorten_text, is applied to the title column to extract key information and truncate it to 50 characters. This function:

  - Cleans non-text characters.
  - Splits text into sentences and selects the most important one (using textrank_sentences with fallback logic).
  - Removes extra spaces and truncates the result.

- **Application:**
  The function creates a new short_title column, and a title_comparison field is generated to compare original and shortened titles.

Data Export

- The final cleaned dataset, now including the optimized short_title column, is saved as a new CSV file for further analysis and marketing optimization.

3. Results

- **Duplicates:** 217 duplicate entries were removed.
- **Missing Values:** 4,091 missing values were addressed (by either removal or imputation).
- **Title Optimization:** Product titles have been condensed to an average of approximately 41–45 characters.
- **Final Dataset:** Reduced to 3,541 rows with standardized and cleaned data ready for analysis.

4. Conclusion

The cleaning process has effectively standardized column names, removed duplicates, handled missing values, and optimized the product titles. With the dataset now reliable and concise, it is ready for deeper analytical work and strategic marketing decisions.

The technical report work thru can be seen here:
https://drive.google.com/file/d/1z18BUceFYOaFHL8XM-mBjDnTajWKZTp2/view?usp=drive_link