Dataset Transformation Documentation

Introduction

This document outlines the steps and transformations applied to the initial dataset, leading to the creation of a final cleaned dataset suitable for machine learning model training.

1. Initial Dataset Loading

Action: The dataset was loaded using Pandas' `read_csv` function with low memory optimization to handle a large number of columns and rows.
Initial State: The dataset contained **24,257 rows** and **158 columns**.

2. Duplicate Handling

Action:

- **Exact Duplicates:** Removed rows that were exact duplicates across all columns.
- **Near-Duplicates:** Identified potential duplicate rows where certain key columns (e.g., `Title`, `Authors`, `Year`) matched, but other columns had slight differences.
- **Conflict Resolution:**
  - For numeric columns, the mean of conflicting values was taken.
  - For categorical columns, the most common value (mode) was selected.
  - When both journal and preprint versions were present, the journal version was retained.

Outcome: After removing duplicates, the dataset was reduced to **23,411 rows**.

3. Dropping Columns with More Than 97.5% Null Values

Action: Columns where more than 97.5% of the rows contained null, zero, or missing values were identified and dropped, with exceptions made for critical columns such as `msm_counts` and `wikipedia_counts`.
Outcome: The number of columns was reduced while retaining key columns necessary for the model.

4. Replacing Null Values with Zeros

Action: For numeric columns where a null value logically indicated an absence (e.g., counts or indicators), null values were replaced with zeros.
Outcome: All appropriate numeric columns had nulls converted to zeros.

5. Dropping Rows with 'Query' in the Title

Action: Rows that contained the word `query` in the `Title` column were identified and dropped, assuming these were likely placeholders or irrelevant entries.
Outcome: **9 rows** were dropped, leaving the dataset with **23,402 rows**.

6. Final Dataset

State: The final dataset contains **23,402 rows** and the selected columns, including critical ones like `msm_counts` and `wikipedia_counts`.
Purpose: The dataset is now prepared for training a machine learning model, with irrelevant or redundant data removed, missing values handled, and consistency ensured.

Summary of Key Changes

- **Rows Reduced:** From 24,257 to 23,402.
- **Columns Reduced:** Retained key columns while dropping those with more than 97.5% null values, except for essential columns like `msm_counts` and `wikipedia_counts`.
- **Key Processes:**
  - Duplicate resolution with conflict handling.
  - Dropping columns based on a 97.5% null/zero threshold while retaining important columns.
  - Handling missing values by setting appropriate nulls to zeros.
  - Removal of potentially irrelevant rows.

This process ensures the dataset is well-suited for feeding into machine learning models, focusing on the most relevant and clean data for analysis.