Summary Report: GenAI Dataset Cleaning and Balancing Pipeline (week_3)

Script Title: Automated Data Cleaning and Balancing Pipeline for GenAI Submissions

Language: Python 3

Dependencies: pandas, numpy, json, openpyxl, datetime

Compatible File Types: .csv, .xlsx, .json

Purpose of the Script

This Python script is designed to automate the preprocessing and validation of GenAI submission datasets. It ensures that all entries meet quality, completeness, and format requirements before further use in machine learning or analytics workflows. The script is modular, scalable, and compliant with best practices in data engineering.

What the Code Does

The script performs end-to-end preprocessing by:

- ✓ Validating and standardizing timestamps to ISO format.
- ✓ Assigning unique submission identifiers.
- ✓ Computing missing token counts using submission text word length.
- ✓ Validating origin labels to ensure only 'human' or 'ai' values are present.
- ✓ Balancing the dataset for equal class distribution, crucial for unbiased machine learning.
- ✓ Dropping records missing critical information to maintain data quality.

Each function in the script is purpose-driven and follows a clean, reusable architecture, making the pipeline adaptable for diverse AI datasets.

Core Functions Overview

- load_data() – Reads input data from CSV, XLSX, or JSON formats.
- clean_data() – Cleans, validates, and structures the dataset according to predefined rules.
- update_timestamps() – Converts and corrects timestamp formats and fills missing values.

- assign_unique_ids() – Assigns unique submission IDs to rows with missing or duplicate IDs.
- balance_data() – Balances the dataset between 'human' and 'ai' labels for fairness.
- save_data() – Saves the cleaned dataset in the desired format.

Key Features and Deliverables

- **Data Quality Assurance:** Missing or malformed fields are corrected or imputed.
- **Unique ID Enforcement**: Ensures each submission has a valid and unique submission_id.
- **Timestamp Normalization:** All timestamps are converted to ISO 8601 for consistency.
- **Token Count Auto-Fill:** If missing, token_count is derived from word count.
- **Balanced Dataset Option:** Equalizes 'human' and 'ai' labels for model training readiness.
- **Flexible Output:** Export to .csv, .xlsx, or .json.

Output Example (Console Log)

Cleaned data written to: cleaned_dataset.json
Final shape: (1150, 8)
Now, ALL rows have a unique non-null submission_id, valid timestamp, and required features.

Technical Summary

Created a strong pipeline for data preprocessing to balance, clean, and validate datasets submitted by GenAI. Schema compliance, timestamp normalization, class balance, and export readiness across CSV, Excel, and JSON formats were all guaranteed by the script. Enhanced data reliability for downstream AI model training.