Building a Dataset for Drug and Malady Classification

Data Preparation, Transformation, and Integration with AI Models

Ranishree Anegundi

Overview

Objective:

- - Prepare a dataset for drug-malady classification.
- - Transform the dataset for fine-tuning and inference with chat models.

Workflow:

- 1. Dataset Preparation (from Excel).
- 2. Data Transformation to Prompt-Completion format.
- 3. Export to JSONL.
- 4. Integration with AI models.

Dataset Preparation

Source:

- - Medicine descriptions stored in an Excel file.

Steps:

- 1. Load the data using pandas.
- 2. Extract relevant columns: Drug_Name and Reason.
- 3. Assign numerical labels to unique ailments (Reason).

Data Transformation

Objective:

- - Format data for prompt-completion tasks.

Steps:

- 1. Create a 'prompt' column combining Drug_Name and Malady.
- 2. Replace Reason with numerical labels.
- 3. Drop unnecessary columns and rename for clarity.

Exporting to JSONL

Objective:

- - Convert the processed DataFrame to JSONL format for compatibility with AI workflows.

Steps:

- 1. Use to_json with orient='records' and lines=True.
- 2. Save as drug_malady_data.jsonl.

JSONL Transformation for Chat Models

Objective:

- - Format JSONL data for fine-tuning or inference with chat-based models.

Steps:

- 1. Read each JSON object from drug_malady_data.jsonl.
- 2. Transform into messages format expected by chat models.
- 3. Save as drug_malady_data_transformed.jsonl.

Integration with Chat Models

Goal:

- - Utilize the transformed dataset for fine-tuning and real-time drug-malady classification.

Benefits:

- - Structured format for seamless integration.
- - Improved model performance with tailored prompts.

Challenges and Future Steps

Challenges:

- - Handling data inconsistencies.
- - Ensuring model interpretability for medical use cases.

Future Steps:

- 1. Validate the dataset with domain experts.
- 2. Fine-tune the model with additional data.
- 3. Deploy as a real-time API for medical use.

Conclusion

Summary:

- - Successfully transformed a raw dataset into a format suitable for AI-based drug classification.
- - Demonstrated the workflow for preparing, exporting, and transforming data.

Next Steps:

- - Fine-tune the model and test its accuracy.
- - Expand the dataset with more classes and examples.