Form.io JSON Generation — Data Preprocessing & Training Pipeline

# Objective

Train a model that can generate structured Form.io JSON schemas from input text that combines fields like Name, DisplayName, Description, and more.

# Directory Structure

project/

│

├── sql\_connection/

│ └── db.py # Contains df\_final DataFrame loaded from SQL Server

│

├── logs/

│ └── training\_log.txt # Training logs written here

│

├── model/ # Model outputs will be saved here

│

├── main.py # Main preprocessing and training script

├── README.md # You're reading it

# Setup

Requirements:

- Python 3.8+

- Transformers (pip install transformers datasets)

- PyTorch

- Pandas

# Pipeline Steps

1. Initial Setup:

- Define model name (google/flan-t5-base)

- Create directories for logs and model checkpoints

- Set up basic file logging

2. Load Data:

- Load df\_final from sql\_connection.db

- Validate required columns: Name, DisplayName, Description, Child\_Relationship, Container

- Drop missing or empty values

- Construct 'input\_text' by joining these fields

3. Preprocess Data:

- Rename input\_text to 'description'

- Rename Container to 'formio\_json'

- Convert to HuggingFace Dataset

4. Load Tokenizer & Model:

- Load tokenizer and model from HuggingFace (AutoModelForSeq2SeqLM)

5. Tokenization Function:

- Tokenize both description and formio\_json

- Truncate and pad to max\_length=512

6. Training Arguments:

- learning\_rate=5e-5

- num\_train\_epochs=5

- per\_device\_train\_batch\_size=4

- weight\_decay=0.01

- logging\_steps=10

# Example Input → Output

Input:

Name: User; DisplayName: User Profile; Description: User's personal information; Child\_Relationship: has\_form;

Output (formio\_json):

{

"components": [...],

"submission": {...}

}

# Logs

Logs are written to logs/training\_log.txt and printed to the console during execution.

# Notes

- This script does NOT include validation or inference code.

- Assumes df\_final is available and pre-loaded.

- Can be integrated into a model deployment pipeline.

# Author

Raj – Data Science Developer