Machine Learning Engineer Task: Resume Categorization

Objective

Design, implement, and train a machine learning model to automatically categorize resumes based on their domain (e.g., sales, marketing). Following this, develop a script that processes a batch of resumes, categorizes them, and outputs results to both directory structures and a CSV file.

Task Breakdown

1. **Find and Download the Dataset:**
    - Download the dataset from attachment.
2. **Data Exploration and Preprocessing:**
    - Examine the dataset and understand the distribution of different categories.
    - Process the resumes into a format suitable for training (e.g., tokenization, feature extraction).
    - Split the dataset into training, validation, and test sets.
3. **Model Selection and Training:**
    - Select an appropriate machine learning or deep learning model.
    - Implement and train the model using the training set.
    - Evaluate the model's performance using the validation set and refine for accuracy and efficiency.
4. **Script Development:**
    - Create a Python script named `script.py`.
    - The script should accept a directory of resumes, categorize them using the trained model, and move them to their respective category folders (creating folders if necessary).
    - The script should also generate a CSV file named `categorized_resumes.csv` with two columns: filename and category.
5. **Command Line Execution:**

Ensure the script is executable from the command line as follows:

```
python script.py path/to/dir
```

6. **Documentation:**
    - Document the chosen model and the rationale behind its selection.
    - Provide details on preprocessing and feature extraction methods.
    - Include instructions on running the script and expected outputs.
7. **Evaluation Metrics:**
    - Provide metrics like accuracy, precision, recall, and F1-score using the test dataset.
    - Consider adding visualizations or additional insights on the model's performance.

Deliverables

1. Jupyter Notebook or Python scripts detailing the exploration, preprocessing, and model training process.
2. The final trained model file.
3. `script.py`.
4. A sample `categorized_resumes.csv` file.(as a sample output after running your script on a test set)
5. Documentation with instructions.