**AMAZON ML CHALLENGE 2024 SOLUTION**

**BY- OOPs We Did It!!**

Steps are:-

1. **Installing PaddleOCR** for text extraction from the images using the pip library.
2. The code in **“TextExtractionAndCSVWithExtractedText.ipynb”** handles **downloading images from URLs**, saving them to a specified folder, and creating a placeholder image if the download fails. It supports multiprocessing for faster downloads and uses the `*tqdm*` library to display progress. It can also retrieve images for further processing by converting them into a format suitable for OCR (optical character recognition).
3. The code in **“TextExtractionAndCSVWithExtractedText.ipynb”** processes a dataset of image links to extract text using OCR (Optical Character Recognition). It defines a function to get text from each image, handles errors, and updates a DataFrame with the extracted text. It saves the DataFrame at specified intervals (`**save\_interval**`) to prevent data loss during processing. The final DataFrame, which includes the extracted text, is saved to a CSV file after processing all images.
4. The code in **“TextExtractionAndCSVWithExtractedText.ipynb”** reads two CSV files (**`train\_text\_1.csv` and `train\_text\_2.csv**`) into DataFrames, concatenates them into a single DataFrame (`df4`), and saves the combined DataFrame as `train\_text\_3.csv`. The shape of the combined DataFrame is checked using `df4.shape`, and the file is saved without including the index column.(We were distributing the dataset in variable batches in order to match CPU/GPU(optional) runtime, so joined now.)
5. The code in **“TextExtractionAndCSVWithExtractedText.ipynb”** saves the `df2` DataFrame to two different file paths as CSV files, excluding the index column (`index=False`). It then prints the shape of `df2`, which shows the number of rows and columns in the DataFrame.
6. **Now open an empty folder and open integrated terminal:**

**Installing Ollma**(a LLM Library for handling requests).

a. **Install Homebrew**:

/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"

b. **Install Ollama**:

brew install ollama

c. **Verify**:

ollama --version

1. We are using **gemma2:2b**:-

ollama run gemma2:2b

1. Upload the resultant file with all the parts of the csv now combined.

1. **app.py**: Handles OCR (Optical Character Recognition) on images, extracts text from them, and updates a DataFrame with this extracted text. It processes images in batches, saves progress periodically, and saves the final result in a CSV file.

2. **join.py**: Merges multiple CSV files containing partial results into a single DataFrame and saves the combined data to a new CSV file.

3. **batch.py**: Splits a large dataset into smaller batches, saves each batch as a separate CSV file, and ensures manageable file sizes for further processing.

4.**requirements.txt**:The command pip install -r requirements.txt installs Python packages listed in the requirements.txt file containing (langchain-core

langchain-ollama).

1. **‘’test-text-output-k.csv”** (k depends on the number of batches made) files are executed after doing 1.batch.py(python batch.py) 2.app.py (python app.py)
2. After training the final\_test\_output.csv is created using join.py .
3. Using “eda.ipynb” file we have cleaned data and final\_test\_output\_eda10.csv is generated.
4. Verified the final\_test\_output\_eda.csv file with the sanity.py file and sample\_test\_out.csv
5. It’s all done.

For any queries:

Contact:

ABHISHEK TRIPATHI(<https://www.linkedin.com/in/abhishek-tripathi-5a76a5299/>)

ARYAMAN DEV KUMAR(<https://www.linkedin.com/in/aryaman-kumar-071956318/>)

AYUSH ARYAN SINGH(https://www.linkedin.com/in/ayush-aryan-singh-877989204/)

AYUSH PRIYADARSHI(<https://www.linkedin.com/in/ayush-priyadarshi-9481b9282/>)

Mail to:

ABHISHEK TRIPATHI([221210006@nitdelhi.ac.in](mailto:221210006@gmail.com))

ARYAMAN DEV KUMAR([221210025@nitdelhi.ac.in](mailto:221210025@nitdelhi.ac.in))

AYUSH ARYAN SINGH([221210026@nitdelhi.ac.in](mailto:221210026@nitdelhi.ac.in))

AYUSH PRIYADARSHI(221210027@nitdelhi.ac.in )