AI DATA INTERN - PYTHON JAVA SPRING BOOT

Harshini Akunuri

Submission: Scalable Matching of Copyright Records to Reference Images

Objective:

To develop a scalable method that retrieves the correct copyright registration image(s) based on a provided spreadsheet of copyright data.

Method Overview:

We implemented a Python-based solution that uses Optical Character Recognition (OCR) and fuzzy matching logic to connect spreadsheet entries with reference images.

Steps:

1. OCR Processing: Each image is processed using Tesseract OCR to extract the registration number, title, and claimant (if available).

2. Normalization: Registration numbers are standardized using a normalization function.

3. Matching Logic: For every row in the spreadsheet, we compare its data (registration number, title, claimant) against the OCR-extracted data from each image. Fuzzy scoring is applied to handle variations in formatting and partial matches.

4. Best Match Selection: The image with the highest average score is selected, provided its score exceeds a confidence threshold (≥ 50).

5. Output Generation: The final matches are saved in a file named images.csv , listing the original title, matched image, and confidence score.

Technologies Used:

- Python
- Tesseract OCR (via pytesseract)
- Fuzzy matching for similarity scoring
- Pandas for data handling

Sample Output:

The output CSV includes:

- Spreadsheet Title
- Matched Image Filename
- Matched Title (from OCR)
- Claimant
- Registration Info
- Confidence Score

This method is scalable, efficient, and easily adaptable to larger datasets or additional OCR pipelines.