Unsplash Image Scraper

Project URL: https://unsplash.com/

Objective

Develop a Python script to scrape images from Unsplash, categorize and store each image in its own directory, and enhance their metadata using AI via PhotoTag.ai. This system will generate JSON files for each image with detailed metadata, facilitating easy access and readability. The project will explore both serial and multithreaded programming approaches along with AI-enhanced metadata generation.

Project Description

1. Website URL: The main input is the URL to Unsplash, a popular free image repository.
2. Photographer and Category Data: Data about photographers and image categories available on Unsplash.
3. PhotoTag.ai API Key: An API key for accessing PhotoTag.ai to obtain enhanced metadata.

Outputs

1. Directory Structure: Each image will be stored in its own directory, named uniquely, possibly using a combination of the image ID and photographer's name.
2. Stored Images: Each image will be downloaded into its respective directory.
3. Metadata Files: For each image, a JSON file containing both original and AI-enhanced metadata (including titles, descriptions, and tags) will be generated and stored in the same directory as the image.
4. Summary Report: A summary report in CSV format showing the distribution of images and enhancements in metadata.

Phase 1: Serial Implementation

Implement the tasks of web scraping, initial metadata extraction, file storage, and initial JSON file creation using a serial programming approach.

Phase 2: Multithreaded Implementation

Upgrade the script to use multithreading to handle multiple downloads and metadata processing simultaneously.
Compare the performance of this approach with the serial method, focusing on improvements in speed and efficiency.

Phase 3: Integration with PhotoTag.ai

Establish communication with PhotoTag.ai's API to send images and receive back enhanced metadata.
Update the JSON files with the AI-generated metadata for each image, enriching the content with AI insights.
Assess the AI-enhanced metadata for accuracy and relevance.

GitHub Repository

Maintain an active GitHub repository with regular updates, structured commits, and clear documentation of all phases.

README File

Include comprehensive instructions for setup and operation, focusing on API configurations and how each component of the script works.
Explain the structure of the stored data, including how directories are organized and the contents of the JSON files.