AI-Powered Web Scraping & Summarization Suite

Project Overview

This project is an AI-powered web scraping and summarization suite that extracts, processes, and summarizes data from Google Search and YouTube using **SerpAPI**. The solution is designed to automate lead generation, research, and content analysis while improving business decision-making efficiency.

Approach(Quantity Driven)

We developed four core tools:

1. **Google Search Scraper**: Extracts search results, including titles, URLs, and snippets, for research and market intelligence.
2. **YouTube Video Scraper**: Fetches video details such as title, channel name, views, and descriptions for content analysis.
3. **AI-Powered Summarization**: Uses **SerpAPI** to summarize extracted data, providing concise insights.
4. **Email CSV Exporter:** Allows users to directly send the result CSV files to their email via the GUI. It uses SMTP protocol and prompts for file selection and recipient address, making data sharing easier and quicker.

****Note: When attempting to send emails via Gmail from the GUI, the app fails with a 535, 5.7.8 Username and Password not accepted error. This occurs because Gmail has restricted the use of direct username-password login for third-party apps.**

Model Selection

- **Web Scraping**: We used **SerpAPI**, a reliable API for extracting search results, ensuring accuracy and consistency.
- **Summarization**: The summarization tool was designed to use **SerpAPI's AI-based processing**, but alternative models like **GPT-based APIs** can be integrated for better NLP-based insights.

Data Preprocessing

- The extracted data is cleaned using **Pandas**, ensuring structured CSV outputs.
- Irrelevant fields are removed, and data formatting (e.g., text cleaning, removing special characters) is applied.
- The summarization tool processes only relevant content to ensure meaningful insights.

Performance Evaluation

- **Data Accuracy**: Verified that extracted data matches real-world search results.
- **Summarization Quality**: Ensured that the generated summaries capture essential insights.
- **Execution Speed**: Optimized API calls to reduce processing time.
- **Error Handling**: Implemented checks to handle API failures (e.g., **404 errors in summarization**).

Business Relevance

This tool streamlines data collection and insight generation, reducing manual research time. It is valuable for **lead generation, competitive analysis, content research, and trend identification**—key aspects of business intelligence and sales strategy.

Conclusion

This project successfully automates the extraction and summarization of business-relevant data, making it a valuable asset for decision-makers. Future enhancements could include integrating **AI-powered text summarization** beyond SerpAPI and developing a **web-based GUI** for better user interaction.