Deal Finder via Web scraping

Built an end-to-end data pipeline that:

- Scrapes data from multiple sources related to Green Steel.
- Summarizes the scraped content using LLM models.
- Stores the summarized articles.
- Exposes an API using FastAPI to serve the data.

Data Sources → Scraping → Summarization → Storage → Serving API

Main Components & Files

1. **Data Ingestion**:

   Directory: ingest/

   Main Files: news_scraper.py, rss_scraper.py, podcast_scraper.py etc.

   Purpose:

- Scrape different sources (Google News, RSS feeds, podcasts, events, tenders).
- Collect raw articles.

   Functions used:

- get_articles_from_google_news()
- get_articles_from_rss()
- get_podcast_data() etc.

2. **Summarization**

   Directory**:** llm/

   Main Files:

- openai_summarizer.py
- local_summarizer.py

   Purpose:

- Send raw articles to LLM (OpenAI or Local Model).
- Generate clean summaries of each article.

   Functions used: summarize_article()

3. **Pipeline Orchestration**

   File: run_pipeline.py

   Purpose**:**

- Load config.
- Call scrapers.
- Call summarizer.
- Aggregate all data.

   Write results to:

- output/articles.json (for persistent storage)
- storage/chroma_store (vector storage using Chroma DB)

   Functions used:

- run_scrapers()
- run_summarizers()
- save_articles_to_file()
- store_articles_in_chroma()

4. **Storage**

4.1 Local File Storage

File: output/articles.json

Purpose: Simple JSON file holding full articles and summaries.

4.2 Vector Storage

Directory: storage/

File: chroma_store.py

Purpose:

- Store embeddings in Chroma DB for semantic search.
- Load articles efficiently.

Functions used: get_articles()

5. **Serving API**

Directory: api/

File: main.py

Purpose:

- Expose your processed data to the outside world via HTTP API.
- Allow users (or your frontend) to query articles.

Main Routes:

Endpoint	Purpose
/	Health check
/articles	Query from Chroma
/json_articles	Load from JSON file

Functions used:

- load_articles() — reads from output/articles.json
- get_articles() — reads from chroma_store