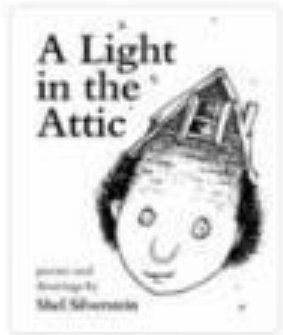


All products

1000 results - showing 1 to 20.

Warning! This is a demo website for web scraping purposes. Prices and ratings he

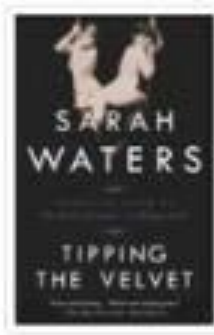


A Light in the ...

£51.77

✓ In stock

Add to basket



Tipping the Velvet

£53.74

✓ In stock

Add to basket

Web Scrape + LLM Demo

Demonstrating a system that scrapes product data from a website, generates AI-powered summaries, and exposes the enriched data via a REST API.

Web Scrape + LLM Demo

- **Scrape 10 items from Books to Scrape website**

Chose Books to Scrape for its static HTML and reliable structure

- **Generate AI summaries using OpenAI GPT-3.5-turbo**

Implemented backoff retry and placeholder fallback on rate limits

- **Serve scraped and summarized data via a REST API with FastAPI**

Dockerized the scraper, summarizer, and API services for easy deployment

- **Demonstrated API endpoints for retrieving product data**

Showcase Swagger UI at <http://localhost:8000/docs>

- **Displayed sample JSON output with title, price, and summary**

Formatted the output to be easily readable and usable

Objective & Scope



Scrape 10 items from a public site

Extract product data like title, price, URL, and rating from the Books to Scrape website



Generate AI summaries

Use OpenAI GPT-3.5-turbo to generate concise summaries for each product



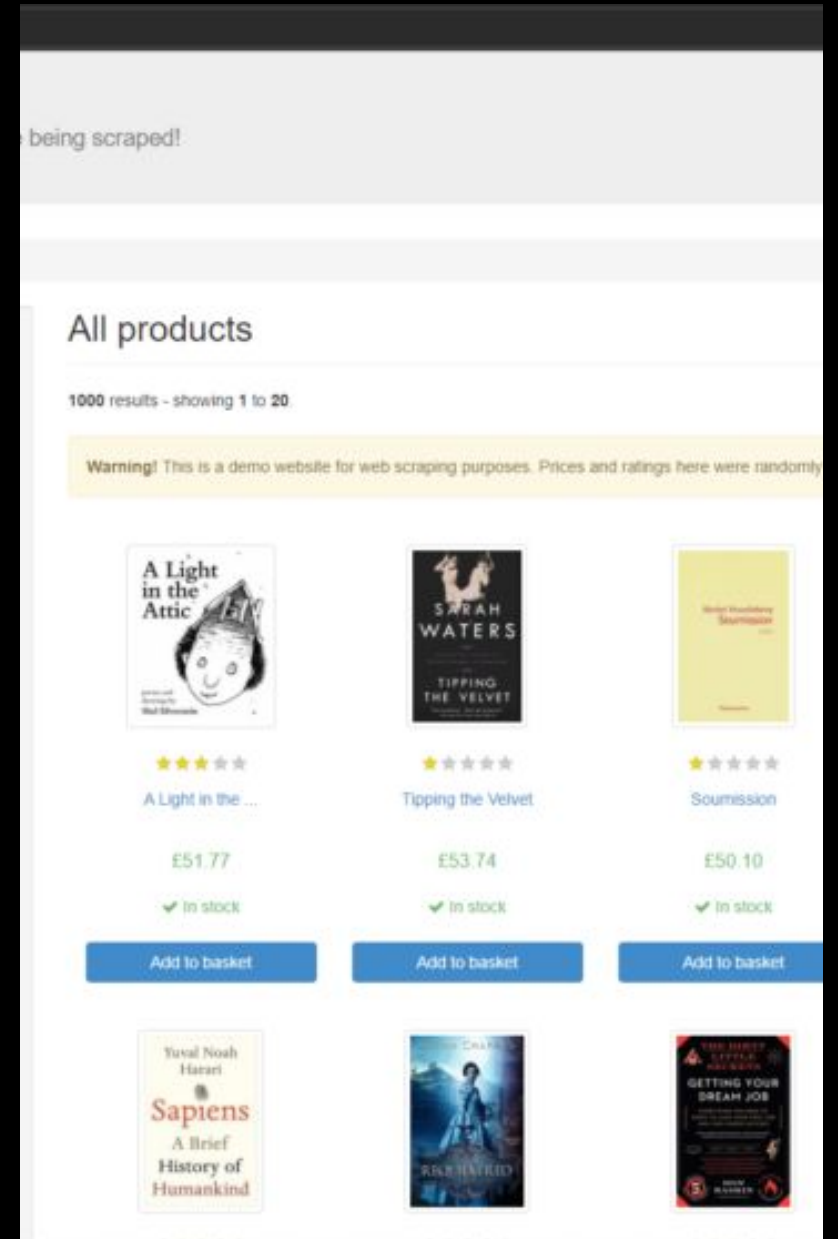
Serve via a REST API

Develop a FastAPI service to expose the scraped and summarized data through a RESTful API

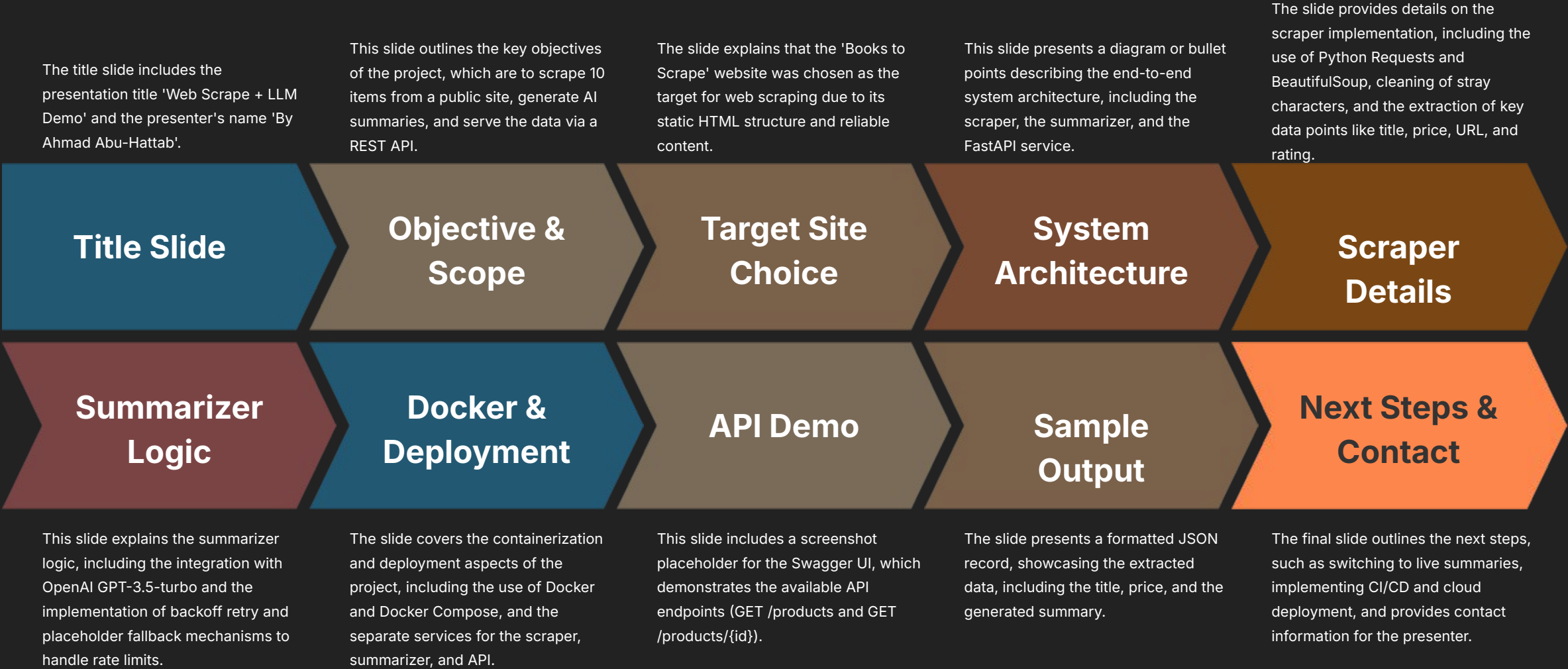
The goal is to demonstrate a complete workflow of web scraping, language model-based enrichment, and API-driven delivery of the data.

Target Site Choice

The Books to Scrape website was chosen for this web scraping and language model demo due to its static HTML structure and reliable data. The site offers a wide selection of books across various genres, providing a suitable dataset for showcasing the capabilities of the scraping and summarization components.



System Architecture



Scraper Details



Python Requests + BeautifulSoup

Used the Python Requests library to fetch web pages and BeautifulSoup to parse the HTML content



Cleaned stray characters (Â)

Removed any stray or unexpected characters from the scraped data to ensure clean output



Extracted title, price, URL, Rating

Parsed the HTML to extract the key product details such as title, price, product URL, and rating

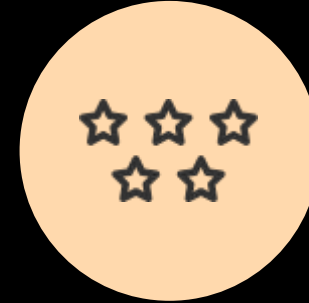
The scraper leveraged Python's powerful web scraping libraries to reliably extract structured data from the target website.

Summarizer Logic



OpenAI GPT-3.5-turbo integration

Used the OpenAI GPT-3.5-turbo language model to generate summaries of the scraped product descriptions



Backoff retry & placeholder fallback on rate limits

Implemented a backoff retry mechanism to handle API rate limits, with a placeholder summary fallback when the model was unavailable

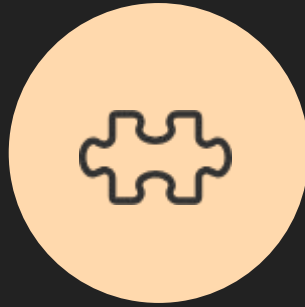
The summarizer leverages the power of large language models to generate concise and informative summaries of the scraped product data, with robust handling of API rate limits to ensure reliable performance.

Docker & Deployment



Dockerfile + docker-compose

Containerized the scraper, summarizer, and API services for easy deployment



Services: scraper, summarizer, api

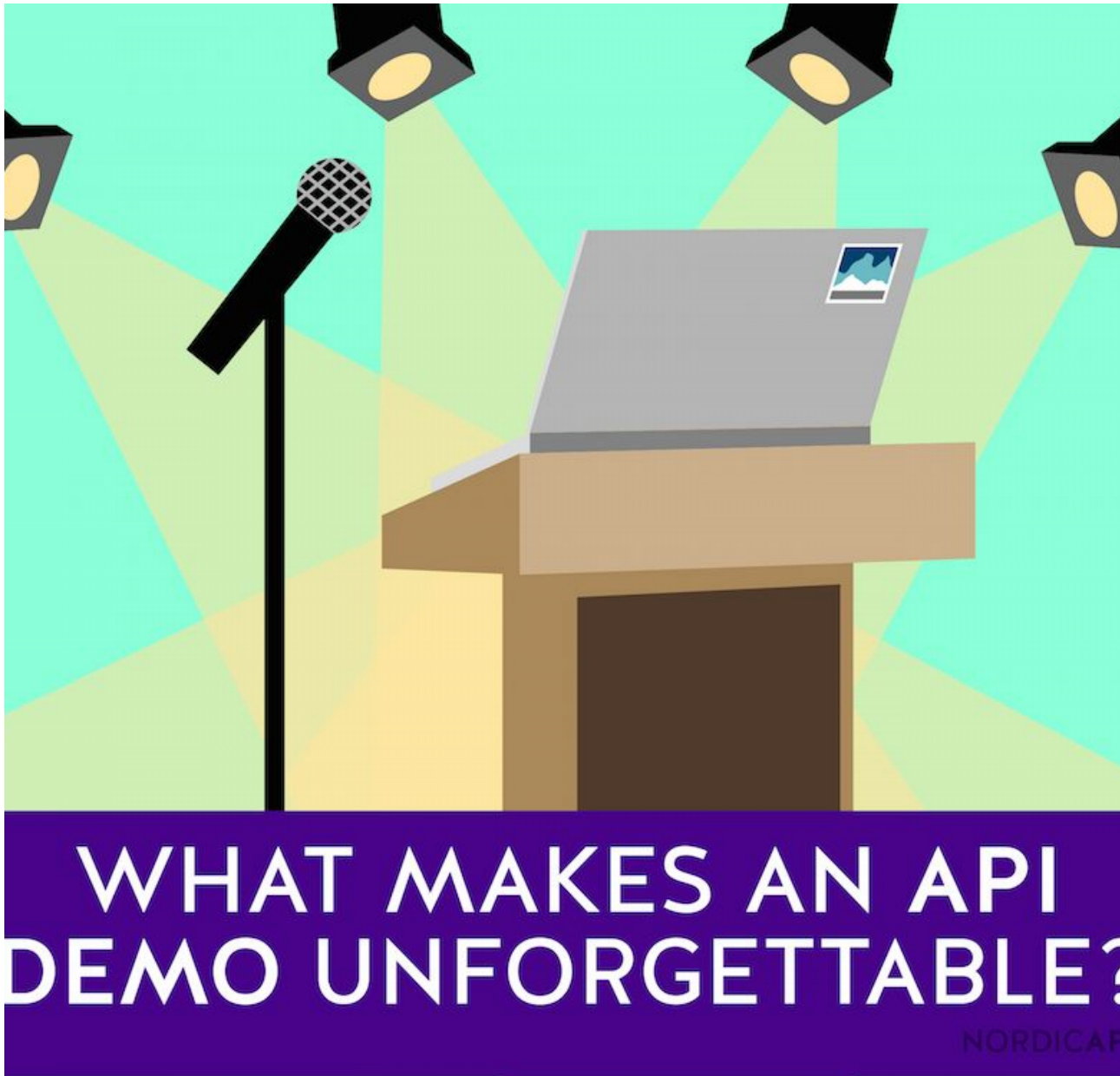
Each component runs in a separate container for modularity and scalability



Versioned configuration

Stored all configuration and environment variables in version control for easy maintenance

The containerized and composable architecture allows for reliable deployment and easy scaling of the web scraping and summarization system.



API Demo

This slide showcases the API demo of the web scraping and language model integration project. It displays a screenshot of the Swagger UI, which provides a user-friendly interface to interact with the API endpoints. The API allows users to retrieve the scraped product data, including the automatically generated summaries.

Sample Output

Book Title

"This is a sample book title from the Books to Scrape website, with its price and a summary generated by the AI model."

Book Price

"\$9.99"

Book Summary

"The AI-generated summary provides a concise overview of the key points and themes of the book, based on the information scraped from the website."

Next Steps & Contact



Switch to live summaries once API quota restored

Utilize the latest API capabilities to generate summaries in real-time once the API quota is increased



CI/CD & cloud deploy

Implement a continuous integration and deployment pipeline to automate the build, test, and deployment process on a cloud platform



Questions? Contact me directly
Provide contact information for users to reach out with any questions or feedback

By implementing these next steps, we can further enhance the functionality and reliability of the web scraping and language model-powered application, making it more accessible and useful for end-users.

Web Scraping in AI T

have purposeful goals.

ML Models

Data Acquisition for Machine Learning Models

Datasets

Diverse and Comprehensive Datasets

Real-Time Data

Real-Time and Up-to-Date Information

Web Scrape + LLM Demo

This presentation demonstrates a web scraping and language model integration project. The goal is to scrape data from the Books to Scrape website, generate AI-powered summaries for each product, and expose the results through a REST API.