

Milestone 1 – Data Collection & Scraping Layer

Project Overview

This milestone focuses on building a complete data scraping and ingestion pipeline using real-world data. The objective was to collect product information from a public website and convert it into a structured dataset for further analysis in upcoming milestones.

Objectives

- 1 Implement web scraping using Python
- 2 Extract product name, price, and ratings
- 3 Clean and normalize raw data
- 4 Generate sentiment labels
- 5 Store structured data in CSV format

Tools & Technologies Used

- 1 Python
- 2 Requests
- 3 BeautifulSoup
- 4 CSV module

Scraping Process

- 1 Sent HTTP request using Requests library.
- 2 Parsed HTML using BeautifulSoup.
- 3 Extracted product details.
- 4 Cleaned and normalized data.
- 5 Generated sentiment labels.
- 6 Saved final dataset as CSV.

Data Schema

Column Name	Type	Description
Product_Name	String	Product title
Price_USD	Float	Product price
Stars	Integer	Rating value
Sentiment	String	Positive / Neutral / Negative
Scraped_Date	Date	Date of scraping

Conclusion

Milestone 1 successfully establishes the foundation for a scalable data pipeline by implementing scraping, preprocessing, and structured storage. This dataset will be used in future milestones.