Lazada Web Product Scraper

08.06.2025

Author: Ya Min Thu

Overview and Objective

This project implements a dynamic product scraper for **Lazada Singapore**, using **Python** and **Playwright** to handle JavaScript-heavy pages. It supports **keyword search**, **pagination**, **price filtering**, and optional **category selection**. Extracted fields mainly include **product title**, **price**, **number of reviews, product url and seller location**, and are saved in **CSV format** for easy analysis. Brand filtering and sorting were deferred due to the complexity of consistently extracting brand data across diverse categories.

The scraper supports Coupang's strategic focus on Southeast Asia's fast-growing e-commerce market by enabling real-time data collection from Lazada. This data supports **market research**, **pricing intelligence**, and **competitive monitoring**, providing valuable insights into consumer behavior and pricing strategies.

Technical Approach and Architecture

- Developed in Python using Playwright, automating a headless browser to simulate user interactions and navigate paginated listings.

- Key product data fields collected:
  `product_name`, `price`, `discount_percentage` (heuristic-based), `review_count`, `discount_tag_line`, `product_url`, `location`, `quantity_sold`, `category`, `scraped_at`

- Implements anti-bot countermeasures, including rotating user agents and randomized request delays to avoid detection and IP blocking.

- Modular codebase:
  - `scraper`: browser automation and data extraction
  - `config`: runtime parameters
  - `analyzer`: post-scrape data summarization
  - `utils`: logging and output directory management

- Data is output as CSV files, enabling integration with internal tools and analytics pipelines.

Challenges and Future Opportunities

- **Brand extraction** is complex due to inconsistent naming conventions across categories, so brand-based filtering and sorting were excluded to maintain data integrity.
- **Discount formats** vary dynamically during campaigns (e.g., "Voucher saves $x," "Subsidized $x off"), requiring current heuristic parsing. Future improvements will integrate **NLP-based models** to extract and interpret discounts more robustly.
- Planned upgrades include applying **transformer-based NLP models** for:
    - Automated **brand recognition**
    - **Attribute normalization**
    - Enhanced **product insights** for advanced competitive intelligence and price tracking at scale

Data Extraction and Testing

- Extracted data fields per product:
  `product_name`, `price`, `discount_percentage`, `review_count`, `discount_tag_line`, `product_url`, `location`, `quantity_sold`, `category`, `scraped_at`
- Data is stored as **CSV files** in the `output/` folder for validation which can be easily ingested to a data pipeline.
- Discounts are currently parsed using heuristics but require future enhancements through **pattern recognition** or **NLP** to adapt to Lazada's dynamic formats during sales campaigns.
- Rigorous testing was conducted across categories such as "smartphones," "headphones," and "shoes" to ensure broad coverage and robustness.
- Integrated **pagination handling**, **rate-limiting**, and **randomized delays** to avoid IP blocking.
- Output quality was validated through manual CSV inspection, log audits, and browser session monitoring.

Scalability and Customization

Modular, configurable scraper supports keyword and category-driven data collection tailored to specific business needs (e.g., Global Pricing team).

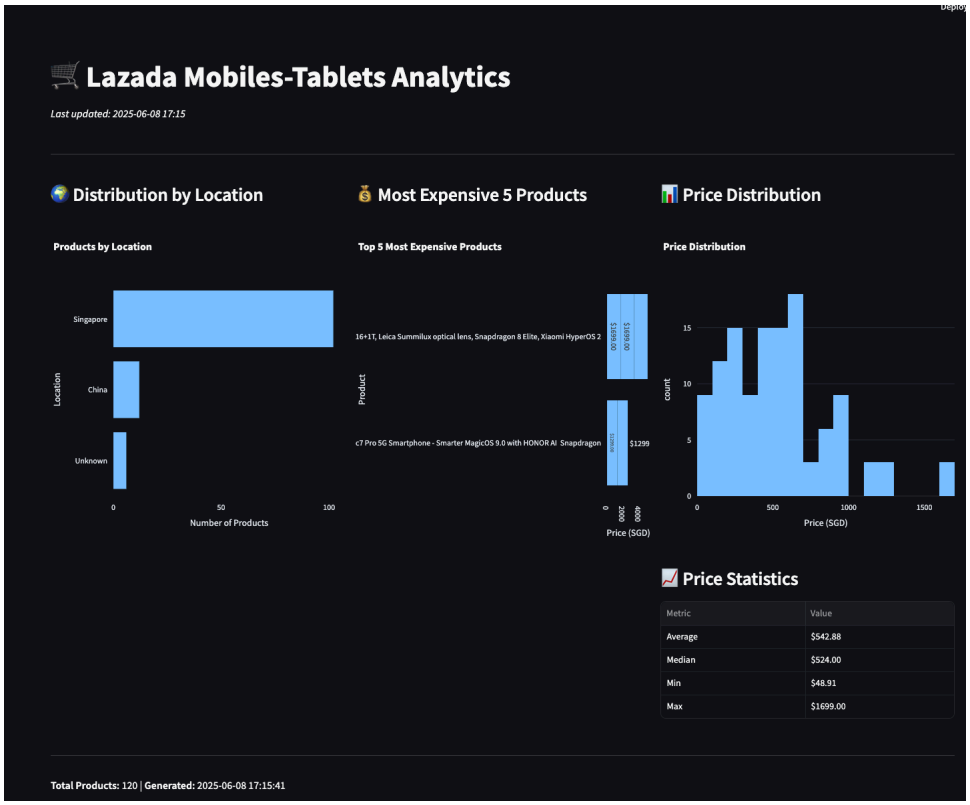
Currently scrapes only search result pages; planned extension to a **multi-step workflow**:

1. Scrape listings metadata
2. Scrape product pages for detailed reviews and seller metrics

Enables richer analysis such as **sentiment analysis**, **seller trust scoring**, and granular **competitive intelligence**.

Data Summary and Analytics

- The built-in **analyzer** module (triggered with `--analyze`) generates summary reports on:
    - Total products scraped
    - Price statistics (min, max, mean, median)
    - Price distribution across brackets
    - Top products by price
- Sample reports are saved in the `output/` directory.



- An **interactive Streamlit dashboard** provides visual exploration of price distributions, product locations, and other summary statistics.