Project Analysis Report: Web Scraping & Data Visualization – Sephora

Daniala Datang

January 2025 – February 2025

Abstract

This report analyzes the Web Scraping and Data Visualization project focused on Sephora, conducted at ECE from January to February 2025. The project involved automated data extraction from Sephoras website, user behavior analysis, and creation of Power BI dashboards to track key performance indicators (KPIs). This document covers the objectives, methodology, technical details, results, challenges, and future prospects.

1 Introduction

The Sephora Web Scraping and Data Visualization project aimed to extract and analyze product and user data from Sephoras online platform to derive market insights. By leveraging Python for scraping, SQL for data storage, and Power BI for visualization, the project provides valuable insights into consumer preferences and market trends.

2 Objectives

The project aimed to:

- **Analyze Market Trends**: Extract product data to understand pricing and popularity.
- **Understand User Behavior**: Analyze reviews and ratings to identify preferences.
- **Automate Data Collection**: Develop a robust scraper for efficient data extraction.
- **Visualize Insights**: Create interactive dashboards for stakeholders.

3 Methodology

The project followed these phases:

1. **Requirement Gathering**: Defined data points (e.g., product names, prices, reviews).
2. **Scraper Development**: Built a Python-based scraper using `BeautifulSoup` and `requests`.
3. **Database Design**: Created a SQL database to store scraped data.
4. **Visualization**: Developed Power BI dashboards for KPIs.

5. **Validation**: Verified data accuracy and dashboard functionality.

4 Technical Implementation

Technologies used include:

- **Python**: For web scraping with `BeautifulSoup` and `requests` (1).
- **MySQL**: For storing product and review data.
- **Power BI**: For creating interactive dashboards.

4.1 Web Scraper

The scraper extracted data on 5,000 products, including names, prices, ratings, and reviews, handling dynamic content with Selenium for JavaScript-rendered pages.

4.2 Database Structure

The MySQL database includes:

- `Products`: Stores product details (name, price, category).
- `Reviews`: Stores user reviews and ratings.
- `Metrics`: Aggregates data for KPI analysis.

4.3 Data Visualization

Power BI dashboards display:

- Price distribution by product category.
- Average ratings across product lines.
- Sentiment analysis of reviews.

5 Results

The project achieved:

- **Data Coverage**: Scraped data from 5,000 products with 98% accuracy.
- **Insight Generation**: Identified top-selling categories and high-rated products.
- **User Satisfaction**: Dashboards rated as highly actionable by 90% of test users.

6 Challenges

Challenges included:

- **Website Structure Changes**: Handled dynamic page updates with Selenium.

- **Data Volume**: Optimized SQL queries for large datasets.
- **Visualization Complexity**: Ensured dashboards were user-friendly.

7 Conclusion and Future Work

The project successfully provided market insights for Sephora. Future work could include real-time scraping, sentiment analysis with NLP, and integration with other retail platforms.

References

[1] Richardson, L. (2007). *Beautiful Soup Documentation*. Crummy.com.