User Manual

User Manual – Review Scraper Script

Overview

This script scrapes product reviews from **G2** and **Capterra** for a given company and time period.

- **Input:** Company name, start date, end date, and source.
- **Output:** JSON file with structured reviews.

Prerequisites

Before running the script, ensure the following are installed on your system:

- Python 3
- Required libraries (install from `requirements.txt`)

Get the chrome webdriver and place that in the project workspace

Chrome Driver

https://googlechromelabs.github.io/chrome-for-testing/

Get into the Project workspace and use the below command to install all the pre-requisites on cmd.

pip install -r requirements.txt

python -m venv venv

venv\Scripts\Activate.ps1

```
pip install -r requirements.txt
```

Folder Structure

review_scraper/

│

├── scraper.py # Main script

├── requirements.txt # List of dependencies

├── output/ # JSON files saved here

├── config.py # Optional: script settings

└── README.md # Instructions

How to Run the Script

Step 1: Open a Terminal or Command Prompt

Navigate to the project folder:

cd review_scraper

Step 2: Run the Script with Required Inputs

Run the script using the following format:

Standard Syntax

python scraper.py --company "Company Name" --start "YYYY-MM-DD" --end "YYYY-MM-DD" --source SOURCE

Example 1: Scrape reviews from G2

python scraper.py --company "Zoom" --start "2024-01-01" --end "2024-06-30" --source g2

Example 2: Scrape reviews from Capterra

python scraper.py --company "Slack" --start "2024-01-01" --end "2024-06-30" --source capterra

Example 3: Scrape reviews from the third source (bonus)

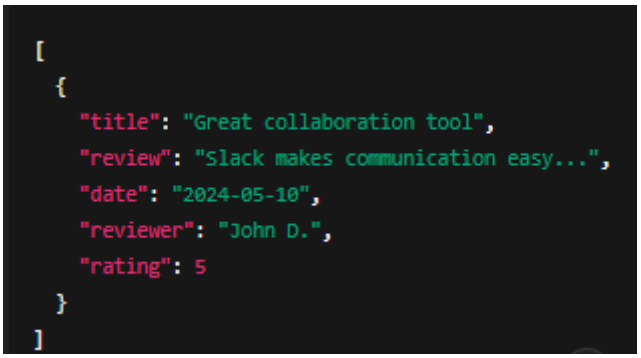
python scraper.py --company "Asana" --start "2024-01-01" --end "2024-06-30" --source thirdsource

Output

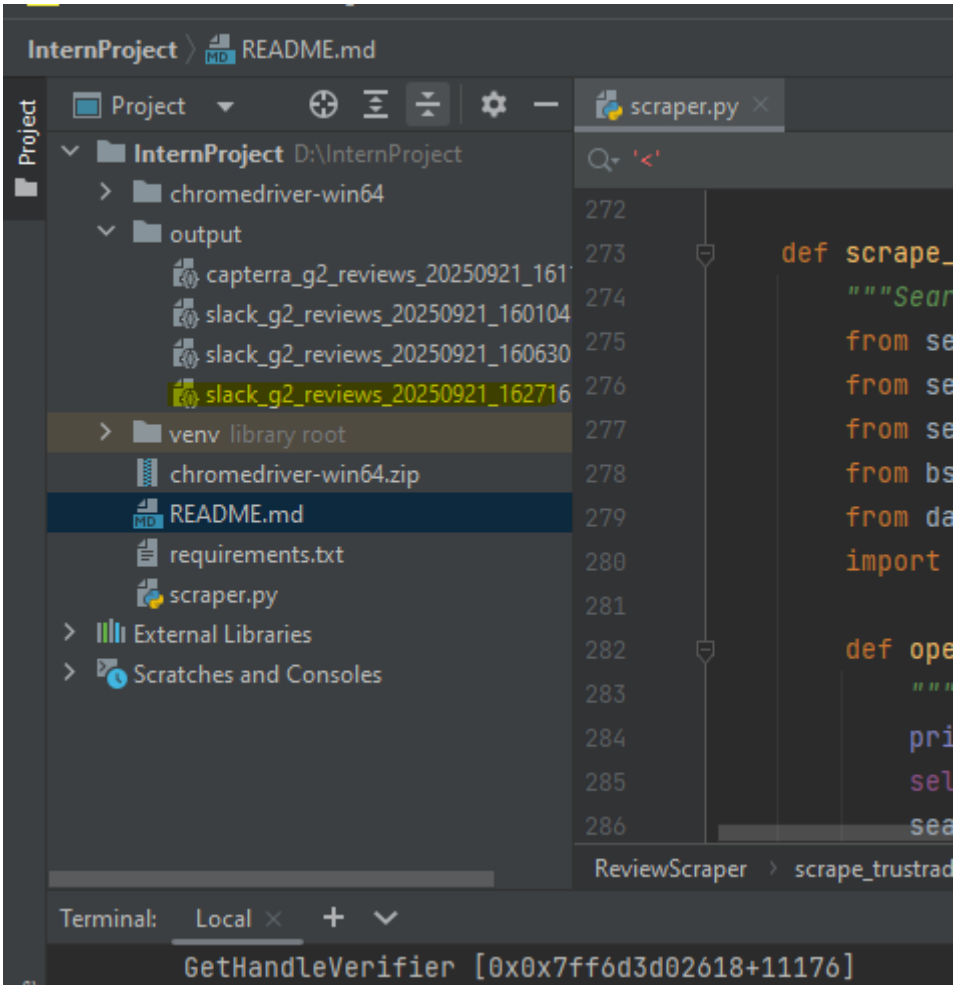
The results are saved in the `output/` folder as a JSON file.

File naming format:

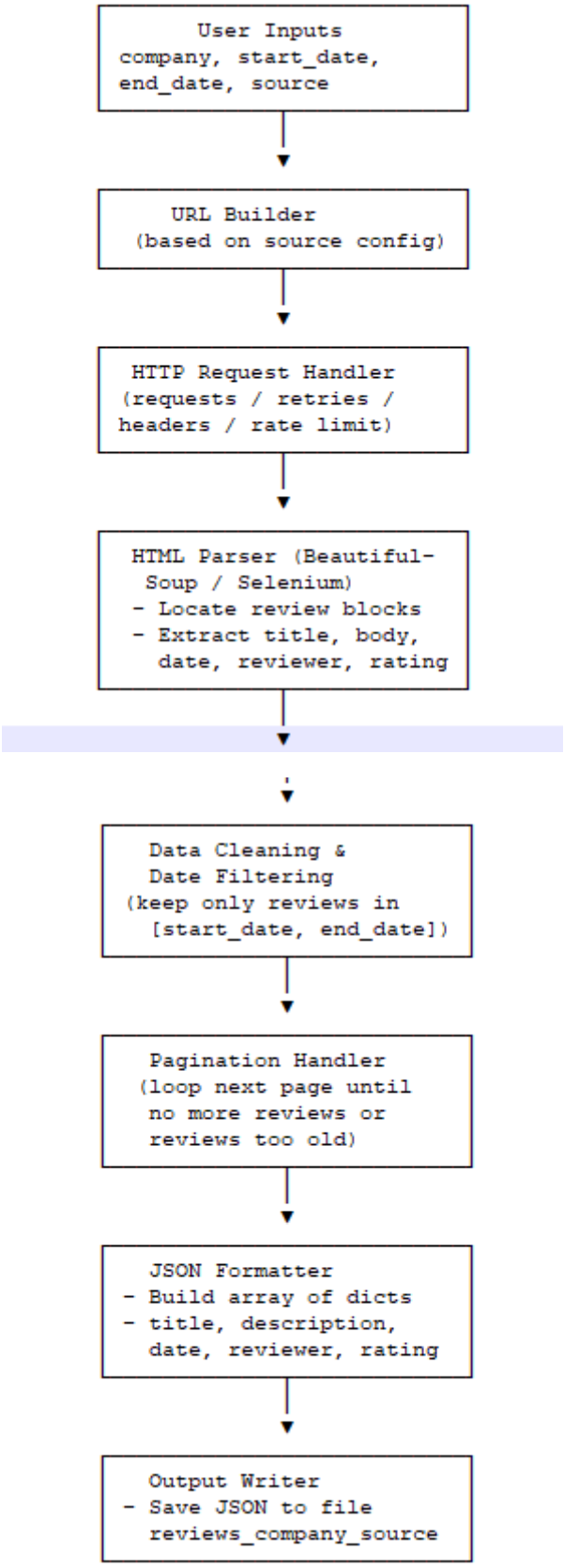
Each JSON file contains an array of reviews with fields like:



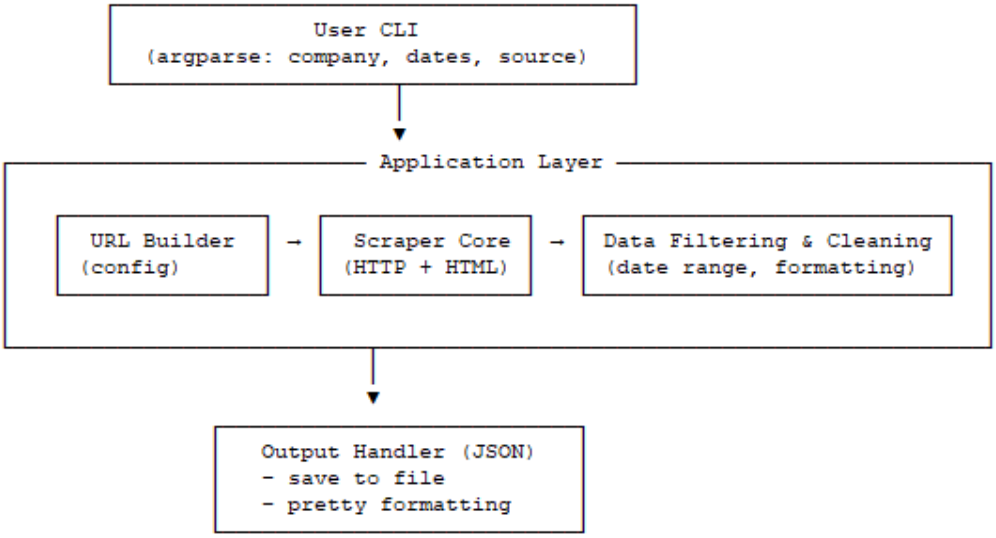
At the end you will get the json file in output directory as below SS.

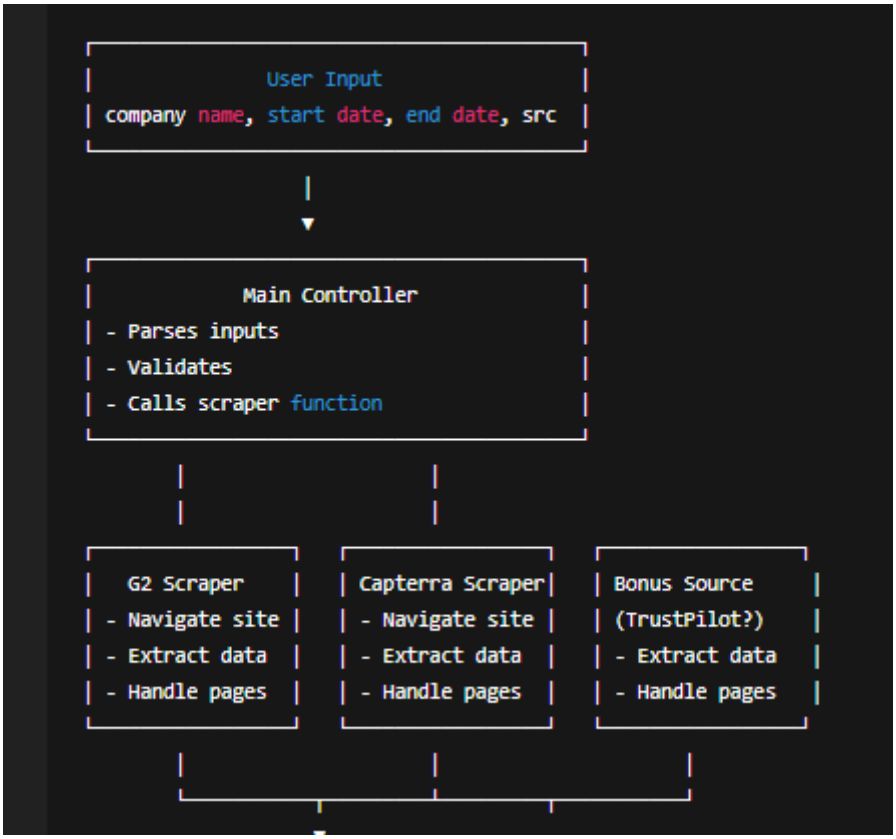


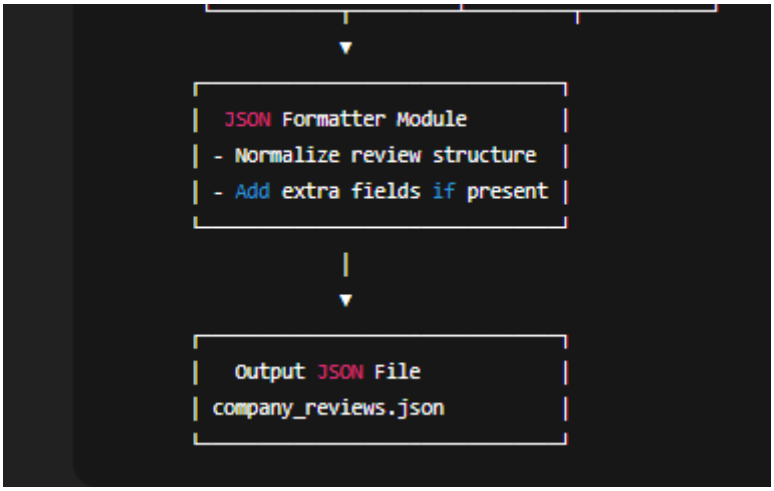
Flow Graph Plan



Architecture


User CLI
(argparse: company, dates, source)
Application Layer
URL Builder
(config)
Scraper Core
(HTTP + HTML)
Data Filtering & Cleaning
(date range, formatting)
Output Handler (JSON)
- save to file
- pretty formatting


User Input
company name, start date, end date, src
Main Controller
- Parses inputs
- Validates
- Calls scraper function
G2 Scraper
- Navigate site
- Extract data
- Handle pages
Capterra Scraper
- Navigate site
- Extract data
- Handle pages
Bonus Source
(TrustPilot?)
- Extract data
- Handle pages


JSON Formatter Module
- Normalize review structure
- Add extra fields if present
Output JSON File
company_reviews.json

********** END **********