

Web Scraping Project Documentation

Overview:

The project consists of 4 core elements: Crawler, Scheduler, Apis & Tests.

Crawler: It is a program written in python which crawls a website (<https://books.toscrape.com>) using the popular scraping library beautifulsoup.

Scheduler: It is a python program that runs continuously and runs the crawler after a certain time to find any new or modified data in the website. If it finds any the database (DB) is updated accordingly, and json and csv reports are generated after the scheduler has finished crawling the entire website.

Apis: They are codes written in python using FastAPI to query and retrieve book data stored in the DB.

Tests: They are used to test the apis and the crawler.

- Incode documentation is also provided for better understanding.

Primary Technologies and Libraries used:

Crawler: beautifulsoup

Scheduler: apscheduler

Apis: FastAPI

Tests: pytest

Database: MongoDB (MongoDB Compass)

Detailed Sections of The Project

Crawler:

This module contains an asynchronous web crawler designed to scrape books.toscrape.com (or any similar site) and store book metadata, snapshots, and change logs into a MongoDB database.

It handles:

- Fetching all paginated book links

- Scraping book details
- Detecting content changes via hashing
- Storing snapshots only when content changes
- Parallel crawling with concurrency limits
- Returning a list of all changes detected on a run

Environment Variables

- **BASE_URL** – Base website to crawl
- **CRAWL_CONCURRENCY** – Number of simultaneous requests
- **CRAWL_RETRIES** – Number of retry attempts for failed requests

Defaults:

- BASE_URL = "https://books.toscrape.com"
- CONCURRENCY = 10
- RETRIES = 3

Main Class: Crawler

`__init__(base_url, concurrency)`

Initializes the crawler with:

- Base URL
- Async HTTP client
- Semaphore to limit concurrency

`async fetch(url)`

Fetches HTML from a URL with:

- Concurrency limit
- Automatic retries
- Error logging

Returns: **HTML string**

`async get_all_book_links()`

Scrapes *every page* of the catalog to collect all book detail URLs.

Workflow:

1. Start at page 1
2. Extract <a> links for all books
3. Follow the **Next** button until pages end
4. Deduplicate links while keeping order

Returns: **List of book detail URLs**

`parse_rating(soup)`

Extracts star rating based on CSS class:
"One" → 1, "Two" → 2, ...

Returns: **Integer or None**

`parse_book(html, url)`

Parses a single book page and extracts:

- Title
- Description
- Category
- Prices (incl/excl tax)
- Availability
- Number of reviews
- Image URL
- Rating
- Unique book ID (derived from URL path)
- Timestamp
- Content hash

Returns:

`(book_data_dict, raw_html)`

`async process_book(url)`

Processes a single book.

Behavior:

If book exists in DB

- Compares stored hash vs. scraped hash
- If changed:
 - Create snapshot
 - Update book record
 - Insert change log
 - Return change info
- If unchanged:
 - No snapshot created
 - Returns None

If book is new

- Create snapshot
- Insert book
- Insert change log
- Return change info

Returns: **Change record or None**

`async run(resume=False)`

Main crawling workflow:

1. Get all book links
2. Process each book concurrently
3. Collect change records

Returns: **List of changes detected this run**

Helper Functions

`async crawl_all_books()`

Utility for schedulers:

1. Runs the crawler
2. Fetches all books from DB (without raw HTML)
3. Returns them as a list

`async main()`

Simple script entry point:

```
asyncio.run(c.run())
```

Summary of Data Saved

Books Collection

- Metadata fields (title, category, price, etc.)
- content_hash
- raw_snapshot_id (link to snapshots collection)

Snapshots Collection

- Full HTML
- Timestamp
- Source URL

Change Log Collection

- book_id
- change_type (new / updated)
- timestamps
- old vs new hash (for updates)

Book Model Using Pydantic

```
class Book(BaseModel):  
    id: str = Field(..., description="Unique id")  
    title: str  
    description: Optional[str]  
    category: Optional[str]  
    price_including_tax: Optional[float]  
    price_excluding_tax: Optional[float]  
    availability: Optional[str]  
    num_reviews: Optional[int]  
    image_url: Optional[HttpUrl]  
    rating: Optional[int] # 1-5  
    source_url: HttpUrl  
    crawl_timestamp: datetime  
    content_hash: str  
    raw_snapshot_id: Optional[str]
```

Schedular:

This module manages **automated crawling** of the book website using APScheduler. It triggers the crawler periodically, detects changes, and generates a daily report from those changes.

Logging Setup

A logger named "scheduler" is configured to print timestamped logs to the console to help track job execution and errors.

Function: scheduled_crawl()

An asynchronous task executed by the APScheduler.

Workflow:

1. Logs start of scheduled crawl
2. Creates a Crawler instance
3. Runs the crawler using c.run()
 - o Returns a list of **new or updated books** detected during this run
4. Logs how many changes were detected
5. Calls generate_daily_report(changes)
 - o Generates a report *only* from this run's changes
6. Closes the crawler's HTTP client

Returns:

Nothing, but generates a report and updates the DB.

Function: async_main()

Sets up and starts the APScheduler.

Behavior:

- Creates an AsyncIOScheduler for async tasks

Adds a job:

```
scheduler.add_job(scheduled_crawl, "interval", minutes=15)
```

- → Runs scheduled_crawl() **every 15 minutes**
- Starts the scheduler
- Keeps the program alive indefinitely using await asyncio.Event().wait()

Script Entry Point

If the file is executed directly:

```
asyncio.run(async_main())
```

This launches the scheduler, which will then execute the crawl job every 15 minutes.

Summary

- scheduled_crawl()
 - Runs crawler → gets changes → generates report
- async_main()
 - Starts APScheduler → schedules the job → keeps program running
- APScheduler interval: **15 minutes**

Alert & Report:

These modules are responsible for creating and sending reports respectively, after the scheduler crawls the entire website at a defined time interval.

Alert

The alert.py file is responsible for sending an email with the generated reports after each scheduler crawl. For testing purposes, an email is sent regardless of whether there are any changes or not.

Report

The report.py file generates csv & json reports every time the scheduler crawls the website. The files contain any change in data of the website or no change message.

APIs:

This module defines the **FastAPI backend** for the Book Scraper system.
It exposes endpoints to retrieve books, book details, and change logs.
It also enforces API key authentication, rate limiting, pagination, and optional filtering.

App Initialization

- Loads environment variables (e.g., API_PORT)
- Creates a FastAPI app with metadata
- Registers global rate limiting via SlowAPI
- Enables CORS for all origins (useful for frontend access)

API Key Authentication

All endpoints include:

```
dependencies=[Depends(get_api_key)]
```

Only clients providing a valid API key can access the API.

Rate Limiting

Each route uses:

```
@limiter.limit("X/hour")
```

Examples:

- /books: **100 requests/hour**
- /books/{book_id}: **5 requests/hour**
- /changes: **100 requests/hour**

This protects the API from abuse.

book_doc_to_resp(doc)

A helper function that extracts only the allowed fields from a MongoDB document before returning it to the client.

It prevents leaking internal fields like raw_snapshot_id or content_hash.

Endpoint: GET /books

Returns a **paginated list of books**, with optional filters and sorting.

Supported Query Parameters:

- category — filter by book category
- min_price, max_price — price range filter
- rating — filter by rating (1–5)
- sort_by — one of:
 - "rating" (desc)
 - "price" (asc)
 - "reviews" (desc)
- page, page_size — pagination controls

Response Body:

```
{  
  "page": 1,  
  "page_size": 20,  
  "total": 1000,  
  "results": [ ...list of books... ]  
}
```

The query is translated into a MongoDB filter, sorted if needed, then paginated manually using skip() and limit().

Endpoint: GET /books/{book_id}

Returns **one book** by its ID.

- Looks up _id in MongoDB
- If not found → returns 404 Book not found
- Otherwise → returns sanitized book document

Endpoint: GET /changes

Returns recent changes (new or updated books) from the change_log collection.

Details:

- Sorted by newest first

- Limited by query parameter limit (default 100)
- Converts `_id` and `book_id` to strings for JSON compatibility

Response:

```
{
  "results": [
    {
      "_id": "log123",
      "book_id": "bk021",
      "change_type": "updated",
      "timestamp": "...",
      "old_hash": "...",
      "new_hash": "..."
    }
  ]
}
```

Running the API

Use the following command:

```
uvicorn api.main:app --host 0.0.0.0 --port <API_PORT> --reload
```

Rate Limiter:

This module sets up **rate limiting** for the FastAPI application using **SlowAPI**.

limiter

```
limiter = Limiter(key_func=get_remote_address)
```

- Creates a global Limiter instance
- `key_func=get_remote_address` → limits requests per client IP address
- Can be used as a decorator on FastAPI routes to enforce rate limits, e.g.:

```
@limiter.limit("100/hour")
```

register_rate_limit(app: FastAPI)

Registers the Limiter with the FastAPI app.

Actions:

1. Attaches the limiter to app.state.limiter
2. Registers a **429 Too Many Requests** exception handler to return a proper response when a client exceeds the limit.

Usage:

```
from fastapi import FastAPI
from api.rate_limit import register_rate_limit

app = FastAPI()
register_rate_limit(app)
```

After this, any route decorated with `@limiter.limit()` is enforced, and exceeding clients get a 429 response.

Summary

- Provides a **global rate limiter** for the API
- Limits are **per client IP**
- Handles **429 errors gracefully**
- Works seamlessly with FastAPI decorators

Tests:

conftest.py

This module provides **fixtures and fake database classes** for testing the FastAPI Books Scraper API without requiring a real MongoDB instance. It enables **unit and integration tests** of API endpoints in isolation.

Key Components

1. Fake MongoDB Classes

- **FakeCursor**
Simulates a Motor cursor with support for `sort()`, `skip()`, `limit()`, and `to_list()`.
- **FakeCollection**
Simulates a MongoDB collection with basic methods:

- `find_one()`, `find()`
- `insert_one()`, `update_one()`
- `count_documents()`
- Supports simple filtering, range queries (e.g., `$gte`, `$lte`), and sorting.
- **FakeDB**
Contains two collections:
 - `books` → stores book documents
 - `change_log` → stores book change records

These classes allow testing database-dependent API logic without a real database.

2. Pytest Fixtures

- **sample_books** → Returns a list of sample book documents for testing.
- **fake_db** → Returns a FakeDB instance preloaded with sample books and change logs.
- **client** → Provides an AsyncClient for testing the FastAPI app with:
 - `get_db` patched to return `fake_db`
 - API key dependency overridden to allow "testapikey"
 - Uses ASGITransport to make requests to the app in-memory

Purpose

This setup allows writing **async API tests** that:

- Query books with filters, sorting, and pagination
- Retrieve individual book details
- Fetch change logs
- Test authentication and rate-limiting logic

All without touching a real database or network.

test_api.py

This module contains **asynchronous pytest tests** for the FastAPI Books Scraper API. It uses the client fixture (with a fake database and overridden API key) to test endpoints **without a real database or network**.

Tested Endpoints

1. **GET /books** – List books with optional filters, sorting, and pagination
2. **GET /books/{book_id}** – Retrieve a single book by ID
3. **GET /changes** – Retrieve recent changes (new or updated books)
4. **Authentication and rate limiting**

Tests Overview

Test	Purpose
test_list_books_no_filters	Returns all books when no filters are applied
test_list_books_category_filter	Returns only books in a specific category
test_list_books_price_range	Returns books within a specified price range
test_list_books_sorting_and_pagination	Validates sorting by rating and pagination
test_get_book_by_id_found	Returns correct book when ID exists
test_get_book_by_id_not_found	Returns 404 when book ID does not exist
test_changes_endpoint_default_limit	Returns all changes with default limit
test_changes_endpoint_with_limit	Returns only the specified number of change entries
test_unauthorized_access	Verifies 401 Unauthorized when API key is missing
test_validation_error_for_bad_query_param	Verifies 422 response for invalid query parameters
test_rate_limit_hit	Simulates rate limiting, expecting 429 if limit is exceeded

Key Features

- Uses **async tests** (@pytest.mark.asyncio) with AsyncClient
- Validates **response status codes**, payload structure, and content
- Covers **filtering, sorting, pagination, and authentication**
- Tests **rate limiting behavior** using SlowAPI configuration
- Fully isolated with **fake database and API key override**

Summary

These tests ensure that:

1. API endpoints behave correctly with various query parameters.
2. Authentication and rate limiting are enforced.
3. Edge cases (missing book, invalid input, query limits) are handled gracefully.

The test suite can be run with:
pytest tests/test_api.py

test_crawler.py

This module contains **asynchronous tests** for the Crawler class in the Books Scraper system.

Tested Functionality

get_all_book_links()

- Ensures the crawler correctly **collects all book detail URLs** from paginated catalogue pages.
- Uses **monkeypatching** to simulate HTTP responses, avoiding actual network requests.

Test Overview

Test	Purpose
test_get_all_book_links_monkeypatched	Simulates two catalogue pages with next links and multiple books. Verifies that Crawler.get_all_book_links(): <ul style="list-style-type: none">• Follows pagination• Extracts all book URLs• Preserves correct order

Key Features

- **monkeypatch** replaces Crawler.fetch with fake_fetch returning predefined HTML.
- Tests **HTML parsing, pagination, and URL joining** without external dependencies.
- Verifies that the crawler correctly returns all expected links.

Summary

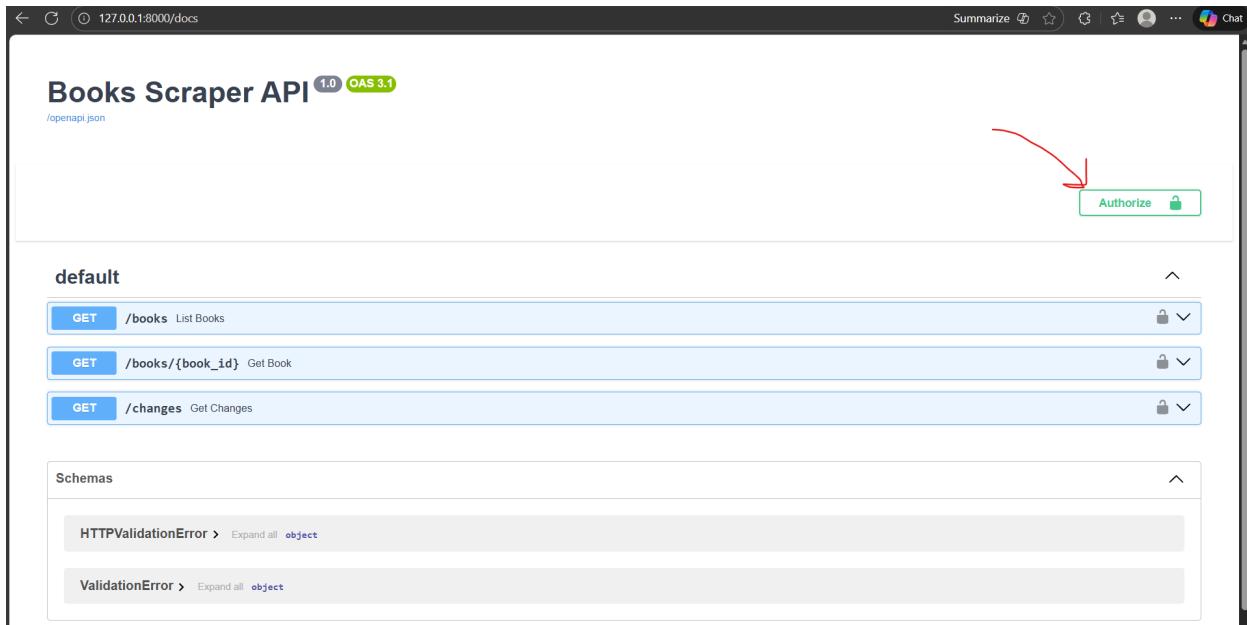
This test ensures the crawler:

1. Handles multi-page catalogues.
2. Extracts book URLs from the HTML correctly.
3. Returns a deduplicated, ordered list of links.

Swagger UI for API testing:

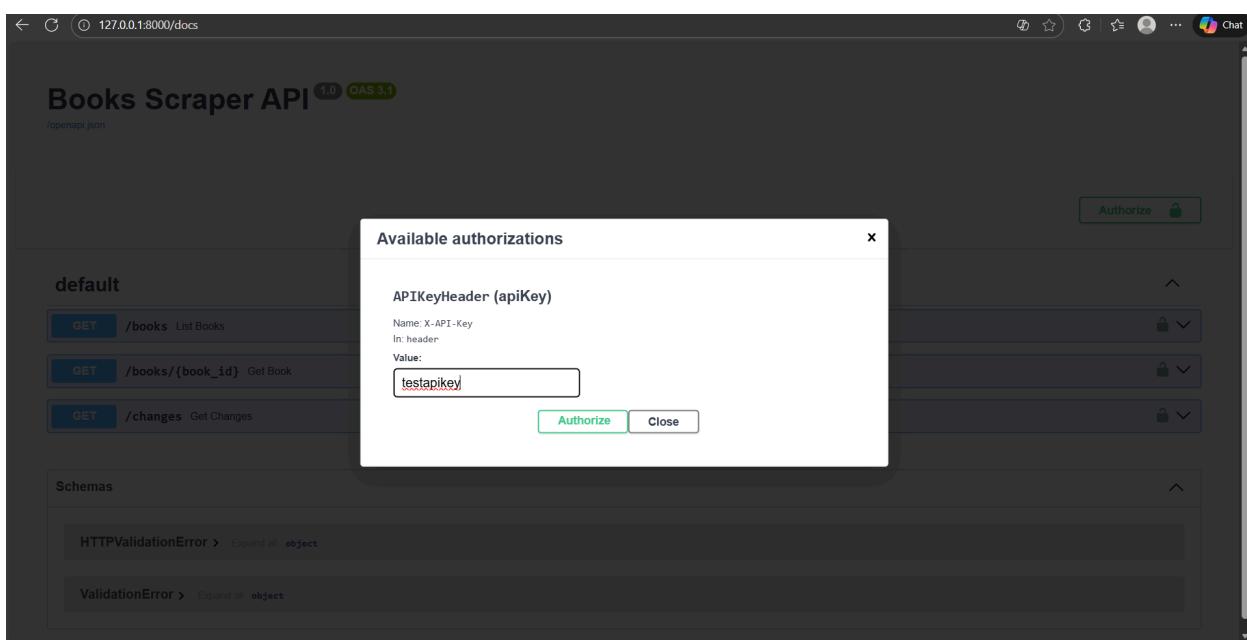
- To use the UI for testing go to the following url (after turning on the server):
<http://127.0.0.1:8000/docs>

- After going to the UI click on the authorize button to enter the authorization header.



The screenshot shows the main interface of the Books Scraper API documentation. At the top right, there is a green "Authorize" button with a lock icon. A red arrow points from the text above to this button. Below the header, there's a section titled "default" containing three "GET" requests: "/books", "/books/{book_id}", and "/changes". Further down, there's a "Schemas" section with two expanded items: "HTTPValidationError" and "ValidationError", each with a "Expand all" link.

- Then input the value “testapikey” as the authentication header value like in the picture below:



The screenshot shows the same API documentation page as before, but with a modal dialog box in the foreground titled "Available authorizations". Inside the dialog, there is one entry: "APIKeyHeader (apiKey)". It shows the "Name" as "X-API-Key", "In" as "header", and "Value" as "testapikey". At the bottom of the dialog are "Authorize" and "Close" buttons. The background of the page is dimmed to indicate it is not active while the dialog is open.

- Now you can test the APIs.

Important Screenshots:

1) Crawler starting to crawl the website.

The screenshot shows the VS Code interface with the terminal tab active. The command `python -m crawler.crawler` is running, and the terminal output shows the crawler listing pages from a website. The output text is as follows:

```

PS D:\Codes\Python\Web Crawler> python -m crawler.crawler
2025-12-01 13:31:26,910 INFO Listing page: https://books.toscrape.com/catalogue/page-1.html
2025-12-01 13:31:28,781 INFO Listing page: https://books.toscrape.com/catalogue/page-2.html
2025-12-01 13:31:29,085 INFO Listing page: https://books.toscrape.com/catalogue/page-3.html
2025-12-01 13:31:29,400 INFO Listing page: https://books.toscrape.com/catalogue/page-4.html
2025-12-01 13:31:29,689 INFO Listing page: https://books.toscrape.com/catalogue/page-5.html
2025-12-01 13:31:30,011 INFO Listing page: https://books.toscrape.com/catalogue/page-6.html
2025-12-01 13:31:30,325 INFO Listing page: https://books.toscrape.com/catalogue/page-7.html
2025-12-01 13:31:30,622 INFO Listing page: https://books.toscrape.com/catalogue/page-8.html
2025-12-01 13:31:30,924 INFO Listing page: https://books.toscrape.com/catalogue/page-9.html
2025-12-01 13:31:31,489 INFO Listing page: https://books.toscrape.com/catalogue/page-10.html
2025-12-01 13:31:31,781 INFO Listing page: https://books.toscrape.com/catalogue/page-11.html
2025-12-01 13:31:32,085 INFO Listing page: https://books.toscrape.com/catalogue/page-12.html
2025-12-01 13:31:32,390 INFO Listing page: https://books.toscrape.com/catalogue/page-13.html

```

2) Crawler inserting books into MongoDB.

The screenshot shows the VS Code interface with the terminal tab active. The command `python -m crawler.crawler` is running, and the terminal output shows the crawler inserting books into a MongoDB database. The output text is as follows:

```

PS D:\Codes\Python\Web Crawler> python -m crawler.crawler
2025-12-01 13:30:14,426 INFO Inserted new book of-mice-and-men_37
2025-12-01 13:30:14,477 INFO Inserted new book off-sides-off-1_38
2025-12-01 13:30:14,453 INFO Inserted new book myriad-prentor-1_36
2025-12-01 13:30:14,454 INFO Inserted new book paradise-lost-paradise-1_45
2025-12-01 13:30:14,512 INFO Inserted new book ms-marvel-vol-1-no-normal-ms-marvel-2014-2015-1_34
2025-12-01 13:30:14,530 INFO Inserted new book meditations_33
2025-12-01 13:30:14,580 INFO Inserted new book lost-among-the-living_31
2025-12-01 13:30:14,601 INFO Inserted new book old-records-never-die-one-mans-quest-for-his-vinyl-and-his-past_39
2025-12-01 13:30:14,747 INFO Inserted new book lord-of-the-flies_30
2025-12-01 13:30:14,748 INFO Inserted new book listen-to-me-fusion-1_29
2025-12-01 13:30:14,883 INFO Inserted new book kitchens-of-the-great-midwest_28
2025-12-01 13:30:14,825 INFO Inserted new book imperfect-harmony_26
2025-12-01 13:30:14,825 INFO Inserted new book icing-aces-hockey-2_25
2025-12-01 13:30:14,885 INFO Inserted new book hawkeye-vol-1-my-life-as-a-weapon-hawkeye-1_24
2025-12-01 13:30:14,897 INFO Inserted new book giant-days-vol-1-giant-days-1_4_22
2025-12-01 13:30:15,057 INFO Inserted new book fruits-basket-vol-1-fruits-basket_1_21
2025-12-01 13:30:15,058 INFO Inserted new book jane-eye_27
2025-12-01 13:30:15,058 INFO Inserted new book frankenstein_20
2025-12-01 13:30:15,059 INFO Inserted new book forever-rockers-the-rocker-12_19
2025-12-01 13:30:15,119 INFO Inserted new book fighting-fate-fighting-6_18
2025-12-01 13:30:15,120 INFO Inserted new book emma_17
2025-12-01 13:30:15,120 INFO Inserted new book eat-pray-love_16
2025-12-01 13:30:15,176 INFO Inserted new book deep-under-walker-security-1_15
2025-12-01 13:30:15,333 INFO Inserted new book charlie-and-the-chocolate-factory-charlie-bucket-1_13
2025-12-01 13:30:15,360 INFO Inserted new book bridget-jones-diary-bridget-jones-1_10
2025-12-01 13:30:15,360 INFO Inserted new book bright-lines_11
2025-12-01 13:30:15,393 INFO Inserted new book bounty-colorado-mountain_7_9
2025-12-01 13:30:15,431 INFO Inserted new book blood-defense-samantha-brinkman_1_8
2025-12-01 13:30:15,432 INFO Inserted new book bleach-vol-1-strawberry-and-the-soul-reapers-bleach-1_7
2025-12-01 13:30:15,433 INFO Inserted new book my-perfect-mistake-over-the-top_1_35
2025-12-01 13:30:15,477 INFO Inserted new book beyond-good-and-evil_6
2025-12-01 13:30:15,477 INFO Inserted new book choosing-our-religion-the-spiritual-lives-of-americas-nones_14
2025-12-01 13:30:15,627 INFO Inserted new book alice-in-wonderland-alices-adventures-in-wonderland-1_5
2025-12-01 13:30:15,628 INFO Inserted new book charitys-cross-charles-towne-bellies-4_12
2025-12-01 13:30:15,651 INFO Inserted new book ajin-demi-human-volume-1-ajin-demi-human-1_4
2025-12-01 13:30:15,654 INFO Inserted new book a-spys-devotion-the-regency-spies-of-london-1_3
2025-12-01 13:30:15,687 INFO Inserted new book 1st-to-die-womens-murder-club_1_2
2025-12-01 13:30:15,697 INFO Inserted new book 1000-places-to-see-before-you-die_1

```

3) Crawler crawled the website and started to insert books into MongoDB.

```

2025-12-01 13:29:28,312 INFO Listing page: https://books.toscrape.com/catalogue/page-22.html
2025-12-01 13:29:28,626 INFO Listing page: https://books.toscrape.com/catalogue/page-23.html
2025-12-01 13:29:29,269 INFO Listing page: https://books.toscrape.com/catalogue/page-24.html
2025-12-01 13:29:29,580 INFO Listing page: https://books.toscrape.com/catalogue/page-25.html
2025-12-01 13:29:30,380 INFO Listing page: https://books.toscrape.com/catalogue/page-26.html
2025-12-01 13:29:30,380 INFO Listing page: https://books.toscrape.com/catalogue/page-27.html
2025-12-01 13:29:30,893 INFO Listing page: https://books.toscrape.com/catalogue/page-28.html
2025-12-01 13:29:31,238 INFO Listing page: https://books.toscrape.com/catalogue/page-29.html
2025-12-01 13:29:31,464 INFO Listing page: https://books.toscrape.com/catalogue/page-30.html
2025-12-01 13:29:31,893 INFO Listing page: https://books.toscrape.com/catalogue/page-31.html
2025-12-01 13:29:32,219 INFO Listing page: https://books.toscrape.com/catalogue/page-32.html
2025-12-01 13:29:32,525 INFO Listing page: https://books.toscrape.com/catalogue/page-33.html
2025-12-01 13:29:32,642 INFO Listing page: https://books.toscrape.com/catalogue/page-34.html
2025-12-01 13:29:33,162 INFO Listing page: https://books.toscrape.com/catalogue/page-35.html
2025-12-01 13:29:33,494 INFO Listing page: https://books.toscrape.com/catalogue/page-36.html
2025-12-01 13:29:33,833 INFO Listing page: https://books.toscrape.com/catalogue/page-37.html
2025-12-01 13:29:34,149 INFO Listing page: https://books.toscrape.com/catalogue/page-38.html
2025-12-01 13:29:34,458 INFO Listing page: https://books.toscrape.com/catalogue/page-39.html
2025-12-01 13:29:34,768 INFO Listing page: https://books.toscrape.com/catalogue/page-40.html
2025-12-01 13:29:35,076 INFO Listing page: https://books.toscrape.com/catalogue/page-41.html
2025-12-01 13:29:35,384 INFO Listing page: https://books.toscrape.com/catalogue/page-42.html
2025-12-01 13:29:35,692 INFO Listing page: https://books.toscrape.com/catalogue/page-43.html
2025-12-01 13:29:36,013 INFO Listing page: https://books.toscrape.com/catalogue/page-44.html
2025-12-01 13:29:36,343 INFO Listing page: https://books.toscrape.com/catalogue/page-45.html
2025-12-01 13:29:36,663 INFO Listing page: https://books.toscrape.com/catalogue/page-46.html
2025-12-01 13:29:37,289 INFO Listing page: https://books.toscrape.com/catalogue/page-47.html
2025-12-01 13:29:37,891 INFO Listing page: https://books.toscrape.com/catalogue/page-48.html
2025-12-01 13:29:38,486 INFO Listing page: https://books.toscrape.com/catalogue/page-49.html
2025-12-01 13:29:39,078 INFO Listing page: https://books.toscrape.com/catalogue/page-50.html
2025-12-01 13:29:39,668 INFO Found 1000 book links
2025-12-01 13:29:39,976 INFO Inserted new book a-light-in-the-attic_1000
2025-12-01 13:29:40,282 INFO Inserted new book starving-hearts-triangular-trade-trilogy_1990
2025-12-01 13:29:40,583 INFO Inserted new book shakespeare-sonnets_989
2025-12-01 13:29:40,845 INFO Inserted new book the-requiem-red_995
2025-12-01 13:29:40,846 INFO Inserted new book the-black-maria_991
2025-12-01 13:29:40,910 INFO Inserted new book the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the-1936-berlin-olympics_992
2025-12-01 13:29:40,912 INFO Inserted new book the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-victoria-woodhull_993
2025-12-01 13:29:40,913 INFO Inserted new book sapiens-a-brief-history-of-humankind_996
2025-12-01 13:29:40,913 INFO Inserted new book summission_998
2025-12-01 13:29:40,915 INFO Inserted new book tipping-the-velvet_999
2025-12-01 13:29:40,916 INFO Inserted new book set-me-free_988

```

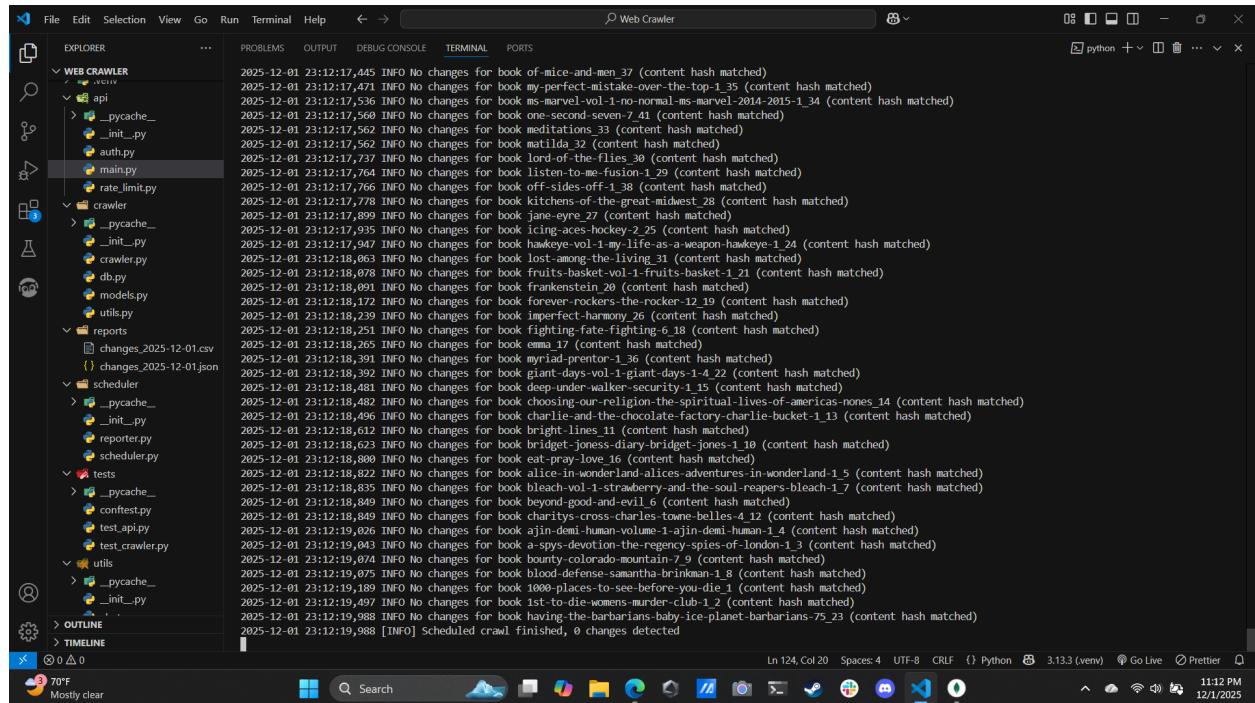
4) Scheduler starting to crawl.

```

(.venv) PS D:\Codes\Python\Web Crawler> python -m scheduler.scheduler
2025-12-01 16:13:12,710 [INFO] Scheduler started (daily at 02:00 UTC)
2025-12-01 16:18:12,746 [INFO] Starting scheduled crawl
2025-12-01 16:18:13,472 INFO Listing page: https://books.toscrape.com/catalogue/page-1.html
2025-12-01 16:18:21,264 INFO Listing page: https://books.toscrape.com/catalogue/page-2.html
2025-12-01 16:18:25,297 INFO Listing page: https://books.toscrape.com/catalogue/page-3.html
2025-12-01 16:18:26,669 INFO Listing page: https://books.toscrape.com/catalogue/page-4.html
2025-12-01 16:18:28,137 INFO Listing page: https://books.toscrape.com/catalogue/page-5.html
2025-12-01 16:18:29,553 INFO Listing page: https://books.toscrape.com/catalogue/page-6.html
2025-12-01 16:18:32,334 INFO Listing page: https://books.toscrape.com/catalogue/page-7.html
2025-12-01 16:18:35,437 INFO Listing page: https://books.toscrape.com/catalogue/page-8.html
2025-12-01 16:18:38,161 INFO Listing page: https://books.toscrape.com/catalogue/page-9.html
2025-12-01 16:18:41,902 INFO Listing page: https://books.toscrape.com/catalogue/page-10.html
2025-12-01 16:18:46,017 INFO Listing page: https://books.toscrape.com/catalogue/page-11.html
2025-12-01 16:18:50,318 INFO Listing page: https://books.toscrape.com/catalogue/page-12.html

```

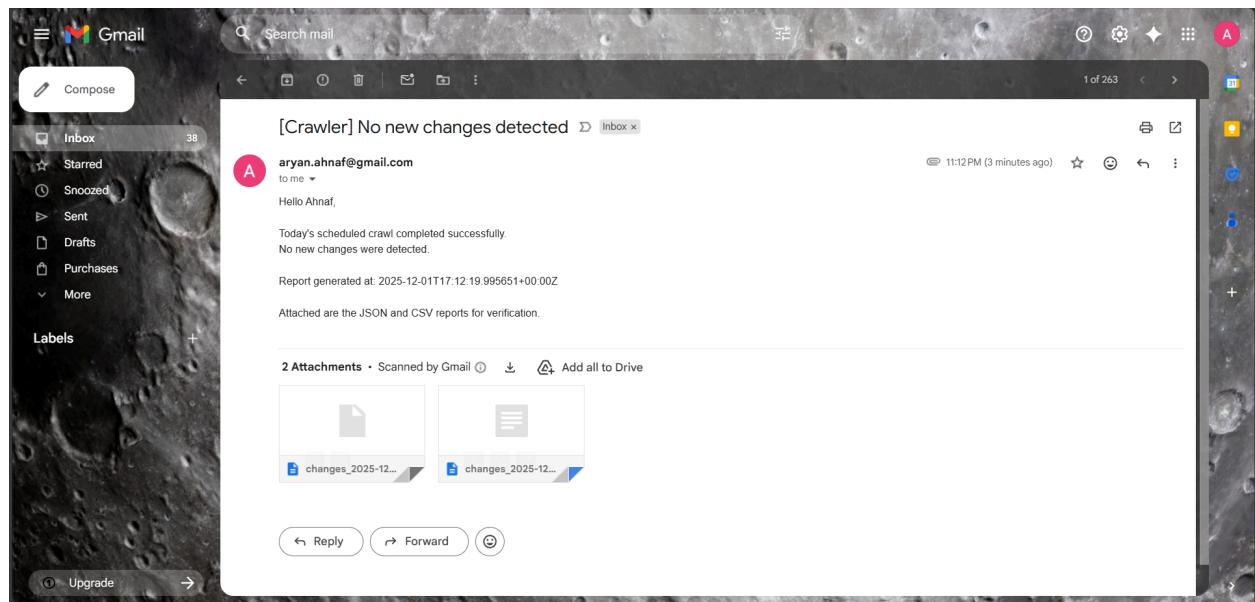
5) The scheduler finished crawling and comparing the hashes of each book to detect any changes and then generated reports and sent an email alert.



```

2025-12-01 23:12:17,445 INFO No changes for book of-mice-and-men_37 (content hash matched)
2025-12-01 23:12:17,471 INFO No changes for book my-perfect-mistake-over-the-top_1_35 (content hash matched)
2025-12-01 23:12:17,536 INFO No changes for book ms-marvel-vol-1-no-normal-ms-marvel-2014-2015-1_34 (content hash matched)
2025-12-01 23:12:17,560 INFO No changes for book one-second-seven-7_41 (content hash matched)
2025-12-01 23:12:17,562 INFO No changes for book meditations_33 (content hash matched)
2025-12-01 23:12:17,562 INFO No changes for book matilda_32 (content hash matched)
2025-12-01 23:12:17,573 INFO No changes for book lord-of-the-flies_30 (content hash matched)
2025-12-01 23:12:17,764 INFO No changes for book listen-to-me-fusion_1_29 (content hash matched)
2025-12-01 23:12:17,766 INFO No changes for book off-sides-off-1_38 (content hash matched)
2025-12-01 23:12:17,778 INFO No changes for book kitchens-of-the-great-midwest_28 (content hash matched)
2025-12-01 23:12:17,935 INFO No changes for book jane-eye_27 (content hash matched)
2025-12-01 23:12:17,935 INFO No changes for book icing-aces-hockey-2_25 (content hash matched)
2025-12-01 23:12:18,063 INFO No changes for book hawkeye-vol-1-my-life-as-a-weapon-hawkeye-1_24 (content hash matched)
2025-12-01 23:12:18,078 INFO No changes for book lost-among-the-living_31 (content hash matched)
2025-12-01 23:12:18,091 INFO No changes for book fruits-basket-vol-1-fruits-basket_1_21 (content hash matched)
2025-12-01 23:12:18,091 INFO No changes for book frankenstein_20 (content hash matched)
2025-12-01 23:12:18,172 INFO No changes for book forever-rockers-the-rocker-12_19 (content hash matched)
2025-12-01 23:12:18,739 INFO No changes for book imperfect-harmony_26 (content hash matched)
2025-12-01 23:12:18,765 INFO No changes for book fighting-fate-fighting-6_18 (content hash matched)
2025-12-01 23:12:18,765 INFO No changes for book emma_17 (content hash matched)
2025-12-01 23:12:18,791 INFO No changes for book myriad-prentor-1_36 (content hash matched)
2025-12-01 23:12:18,392 INFO No changes for book giant-days-vol-1-giant-days-1-4_22 (content hash matched)
2025-12-01 23:12:18,481 INFO No changes for book deep-under-walker-security_1_15 (content hash matched)
2025-12-01 23:12:18,482 INFO No changes for book choosing-our-religion-the-spiritual-lives-of-americas-nones_14 (content hash matched)
2025-12-01 23:12:18,496 INFO No changes for book charlie-and-the-chocolate-factory-charlie-bucket-1_13 (content hash matched)
2025-12-01 23:12:18,612 INFO No changes for book bright-lines_11 (content hash matched)
2025-12-01 23:12:18,623 INFO No changes for book bridget-joness-diary-bridget-jones-1_10 (content hash matched)
2025-12-01 23:12:18,800 INFO No changes for book eat-pray-love_16 (content hash matched)
2025-12-01 23:12:18,835 INFO No changes for book alice-in-wonderland-alices-adventures-in-wonderland-1_5 (content hash matched)
2025-12-01 23:12:18,849 INFO No changes for book bleach-vol-1-strawberry-and-the-soul-reapers-bleach-1_7 (content hash matched)
2025-12-01 23:12:18,849 INFO No changes for book beyond-good-and-evil_6 (content hash matched)
2025-12-01 23:12:19,026 INFO No changes for book ajin-demi-human-volume-1-ajin-demi-human-1_4 (content hash matched)
2025-12-01 23:12:19,043 INFO No changes for book a-spys-devotion-the-regency-spies-of-london-1_3 (content hash matched)
2025-12-01 23:12:19,074 INFO No changes for book bounty-colorado-mountain-7_9 (content hash matched)
2025-12-01 23:12:19,075 INFO No changes for book blood-defense-samantha-brinkman-1_8 (content hash matched)
2025-12-01 23:12:19,189 INFO No changes for book 1000-places-to-see-before-you-die_1 (content hash matched)
2025-12-01 23:12:19,497 INFO No changes for book 1st-to-die-womens-murder-club-1_2 (content hash matched)
2025-12-01 23:12:19,988 INFO No changes for book having-the-barbarians-baby-ice-planet-barbarians-75_23 (content hash matched)
2025-12-01 23:12:19,988 [INFO] Scheduled crawl finished, 0 changes detected

```



6) All 12 tests passed.

The screenshot shows a VS Code interface with the following details:

- File Explorer:** Shows the project structure under "WEB CRAWLER".
- Terminal:** The terminal tab is active, displaying the output of a pytest run.

```
(.venv) PS D:\Codes\Python\Web_Crawler> pytest -v
platform win32 -- Python 3.13.3, pytest-9.0.1, pluggy-1.6.0
rootdir: D:\Codes\Python\Web_Crawler
configfile: pytest.ini
plugins: anyio-4.11.0, asyncio-1.3.0
asyncio mode=Mode.AUTO, debug=False, asyncio_default_fixture_loop_scope=None, asyncio_default_test_loop_scope=function
collected 12 items

tests\test_api.py .....
tests\test_crawler.py .

===== 12 passed in 0.54s =====
```
- Status Bar:** Shows file statistics (Ln 473, Col 33), code analysis (Air: Very Poor), and system information (126 AM, 12/2/2025).