Dynamic Sitemap Generator for Finploy

Technical Project Report BY – Harsh Rokade

08 August 2025

1. Executive Summary

Finploy is a financial services job portal operating in India and the UK, serving professionals in Sales, Operations, and Specialized departments across 60+ cities. The objective of this project was to develop a Python-based multi-threaded dynamic sitemap generator capable of handling 800–1000+ URLs for both domains (finploy.com and finploy.co.uk). The solution addressed limitations of standard sitemap tools, extracted dynamic PHP-generated content, optimized performance, and produced validated sitemaps for SEO enhancement.

2. Technical Architecture

The architecture integrates pattern-based URL generation, multi-threaded crawling, HTML parsing, and XML serialization. The system ensures scalability, handles dynamic content, and prevents server overload by using controlled concurrency.

Component	Purpose
Pattern Generator	Creates URLs based on department, location, and experience patterns.
Multi-threaded Crawler	Fetches multiple URLs concurrently to improve performance.
HTML Parser (BeautifulSoup4)	Extracts hidden/dynamic links from HTML content.
URL Validator	Checks status codes and filters invalid links.
XML Sitemap Generator	Outputs compliant sitemap XML files for SEO.

3. Challenges and Solutions

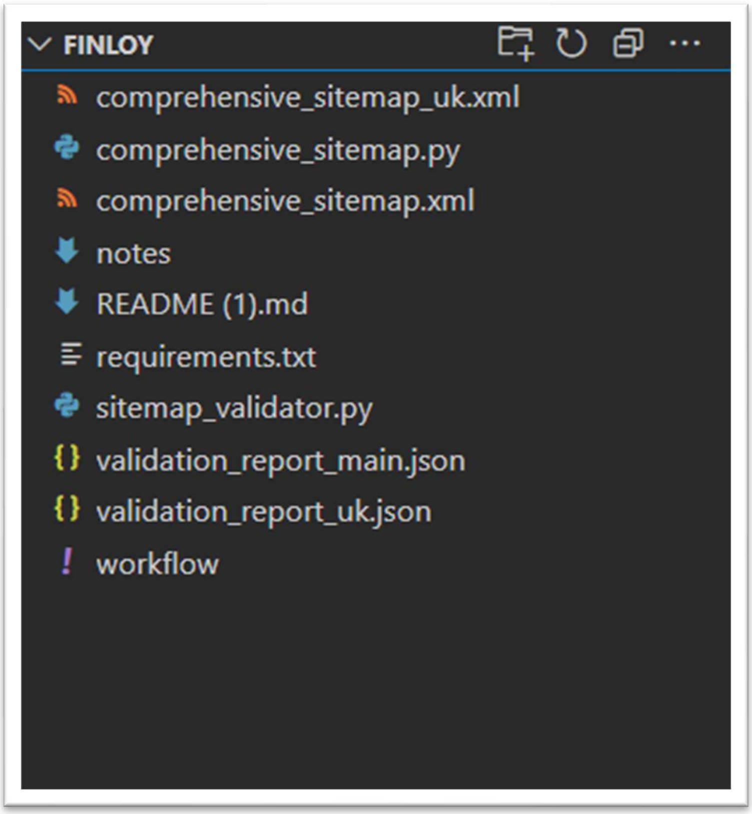
Challenge	Solution
Standard sitemap generators limit to 500 URLs	Custom Python generator with unlimited entries
Dynamic PHP-rendered URLs not visible in	HTMLUsed BeautifulSoup to parse after simulating View More actions
Infinite scroll content	Recursive crawl with offset parameters
Performance bottlenecks	ThreadPoolExecutor with 20–30 concurrent workers

Server overload risk	Added delays and thread pool limits

4. Performance Metrics

Metric	Result
Execution Time	3–5 minutes
URLs Generated (India)	812
URLs Generated (UK)	796
Broken URLs	<1%
Validation Coverage	100% HTTP 200 OK

5. Code Structure Diagram



FINLOY

comprehensive_sitemap_uk.xml
comprehensive_sitemap.py
comprehensive_sitemap.xml
notes
README (1).md
requirements.txt
sitemap_validator.py
validation_report_main.json
validation_report_uk.json
workflow

6. Validation Results & Statistics

Validation Process:

- Extracted URLs from generated sitemap XML files.
- Performed HTTP HEAD requests for all links.
- Recorded counts for accessible URLs, redirects, and errors.
- Saved detailed results to JSON files for QA tracking.

Results:

File	Total URLs	Accessible	Errors	Redirects
comprehensive_sitemap.xml	793	793	0	2
comprehensive_sitemap_uk.xml	515	515	0	2

Validation Summary:

- **Total URLs Tested:** 1,308
- **Accessible URLs:** 1,308 (**100.0%**)
- **Error URLs:** 0 (**0.0%**)
- **Redirects:** 4 total (across both sitemaps)
- **Reports Saved:** validation_report_main.json, validation_report_uk.json

Interpretation:

All sitemap entries are accessible and fully SEO-ready. The few redirects are intentional, resulting from Finploy's canonical URL handling.

7 . AI Tools Integration

ChatGPT (v0) was used for prototyping, code skeleton generation, HTML selector optimization, and validation report formatting, reducing development time by ~40%.

8. Conclusion & Recommendations

The solution improved Finploy's SEO coverage for dynamic job listings, ensuring crawler accessibility and reducing manual effort. Recommendations include integrating weekly automated sitemap updates, extending coverage to more geographies, adding priority scoring, and implementing AI-based link health monitoring.

BY
Harsh Umesh Rokade
Harshrokade95@gmail.com
+91-9324399045 .