Sitemap Crawler Overview

Tools and Language

Language: Python 3 – Chosen for its strong ecosystem of web scraping, HTTP handling, and data processing libraries.

Libraries / Tools Used

- requests – HTTP requests and session management
- BeautifulSoup – HTML/XML parsing
- urllib.parse – URL parsing and joining
- xml.etree.ElementTree – XML sitemap creation
- datetime – Timestamps for sitemap and reports
- time – Crawl delay handling
- logging – Console & file logging
- json – Report generation
- re – Extract URLs from scripts
- collections.deque – Queue management

Key Features

- Multi-source URL discovery (sitemaps, HTML, JavaScript)
- BFS crawling with rate limiting
- Domain filtering & file type exclusion
- Real-time progress tracking

Output Files

- enhanced_sitemap.xml – XML sitemap with priorities
- enhanced_sitemap_report.json – Analytics & stats
- enhanced_sitemap.log – Debug & progress logs

Performance Features

- Session reuse for faster connections
- HEAD requests for quick validation
- Configurable crawl limits & delays
- Error handling with graceful degradation

Specialized for Job Sites

- Pre-defined job keywords & locations
- Careers/jobs page pattern generation
- Priority-based sitemap ordering
- Job-specific validation rules