Enhanced Scraper Framework Documentation

Overview

The enhanced scraper framework provides a robust foundation for scraping job fair information from various sources in Hong Kong. It addresses the key issues identified in the original codebase and adds new features.

Key Features

- **Multiple Scraper Types**: Support for static HTML, dynamic content (Selenium), and API-based scrapers
- **Configurable Update Frequencies**: Hourly, daily, weekly, monthly, or custom schedules
- **Anti-Scraping Measures**: User-agent rotation, proxy support, CAPTCHA handling, and rate limiting detection
- **Robust Error Handling**: Retry mechanisms, exception handling, and comprehensive logging
- **Data Validation**: Pydantic models for data validation and normalization

Changes Made

1. **Fixed Import Issues**: Corrected relative imports in JobsDB and HKTDC scrapers
2. **Enhanced Base Scraper**: Added support for different scraper types, update frequencies, and anti-scraping measures
3. **Added Data Validation**: Created Pydantic models for job fair events
4. **Improved Error Handling**: Added retry mechanisms and better logging
5. **Added Anti-Scraping Utilities**: Created utilities for rotating user agents, proxies, and handling CAPTCHAs

Usage

To create a new scraper, extend the `BaseScraper` class and implement the `scrape()` method:

```
from hk_job_fair_aggregator.scrapers.base import BaseScraper
from hk_job_fair_aggregator.scrapers.scraper_types import ScraperType, UpdateFrequency

class MyScraper(BaseScraper):
    def __init__(self):
        super().__init__(
            name="My Scraper",
            base_url="https://example.com",
            source_id="my_scraper",
            source_type="JOB_PORTAL",
            source_priority="PRIMARY",
```

```
            scraper_type=ScraperType.STATIC,
            update_frequency=UpdateFrequency.DAILY,
            language="EN"
        )

    def scrape(self):
        # Implement scraping logic here
        pass
```

Backward Compatibility

The enhanced framework maintains backward compatibility with existing scrapers. The original functionality is preserved while adding new features.