Usage Guide - E-commerce Data Extractor

Quick Start

1. Installation

```bash
# Clone the repository
git clone https://github.com/yourusername/ecommerce-data-extractor.git
cd ecommerce-data-extractor

# Install dependencies
pip install -r requirements.txt
```

2. Basic Usage

```bash
# Default scraping (selenium method, max pages from config)
python main.py

# Quick test run
python main.py --test

# Custom configuration
python main.py --method selenium --max-pages 10
```

Command Line Options

Option	Description	Example
`--method`	Scraping method (`selenium` or `requests`)	`--method selenium`
`--max-pages`	Maximum pages to scrape	`--max-pages 50`
`--test`	Run quick connectivity test	`--test`
`--resume`	Resume from backup file	`--resume backup.txt`

Usage Examples

1. Standard Data Collection

```bash
```

```
# Scrape up to 100 pages using selenium
python main.py --method selenium --max-pages 100
```

2. Quick Test Before Full Run

```
bash

# Test site accessibility first
python main.py --test

# If successful, run full scrape
python main.py --max-pages 50
```

3. Resume Interrupted Session

```
bash

# Resume from previous backup
python main.py --resume wine_links_backup.txt --max-pages 200
```

4. Light-weight Scraping

```
bash

# Use requests method for faster, simpler scraping
python main.py --method requests --max-pages 25
```

Output Files

The scraper generates several output files:

- `wine_links.txt` - Extracted product URLs
- `wine_products.csv` - Complete product data
- `wine_links_final_XXX_links.txt` - Final backup with link count

Configuration

Edit `config.py` to customize:

```
python
```

```
# Scraping limits
MAX_PAGES = 500
MAX_RETRIES = 3

# Delays (seconds)
MIN_DELAY = 1
MAX_DELAY = 3

# Output files
LINKS_FILE = 'wine_links.txt'
CSV_OUTPUT_FILE = 'wine_products.csv'
```

Troubleshooting

Common Issues

1. "Driver not found" error:

```
bash

# The script auto-downloads drivers, but if it fails:
pip install --upgrade webdriver-manager
```

2. "No links found" message:

```
bash

# Try test mode first to check connectivity
python main.py --test

# If test passes, try with longer delays in config.py
MIN_DELAY = 3
MAX_DELAY = 6
```

3. Frequent timeouts:

```
bash

# Increase timeout values in config.py
PAGE_LOAD_TIMEOUT = 90
IMPLICIT_WAIT = 15
```

4. Age verification popup:

- The script handles this automatically
- If it fails, try running again (popups are inconsistent)

Performance Tips

For Large-Scale Scraping:

- Use `--max-pages` to limit scope during testing
- Run `--test` first to verify site accessibility
- Monitor logs for error patterns
- Consider running during off-peak hours

For Faster Results:

- Try `requests` method first: `--method requests`
- Reduce `MAX_DELAY` in config for faster operation
- Use smaller page limits for testing: `--max-pages 10`

Data Output Format

CSV Columns:

- `Title` - Product name
- `Price` - Product price with currency
- `Description` - Short product description
- `Origin` - Country of origin
- `Weight` - Product weight
- `Size` - Product size (ML)
- `Category` - Wine category (Red, White, etc.)

Example Output:

```csv
Title,Price,Description,Origin,Weight,Size,Category
"Château Margaux 2015","₨ 45,000.00","Premium Bordeaux red wine","France","750g","750ML","Red Wine"
```

Advanced Usage

Custom Selectors

Modify selectors in `config.py` if the website structure changes:

```python
PRODUCT_TITLE_SELECTORS = [
    "h1.product_title",
    ".product_title",
    "div.summary.entry-summary > h1",
]
```

Extending the Scraper

The modular architecture allows easy extension:

- `base_scraper.py` - Core scraping functionality
- `link_scraper.py` - URL collection logic
- `data_scraper.py` - Product data extraction
- `utility.py` - Helper functions

Rate Limiting & Ethics

This scraper includes built-in rate limiting:

- Random delays between requests (1-3 seconds default)
- Respectful crawling patterns
- Error handling to avoid overwhelming servers

Please use responsibly:

- Don't exceed reasonable request rates
- Respect robots.txt guidelines
- Consider contacting site owners for large-scale data needs

Support

For issues or questions:

1. Check this usage guide
2. Review error logs for specific issues
3. Test with `--test` flag first
4. Try reducing `--max-pages` for testing