Building Owner Information Scraper Documentation

Overview

A comprehensive web scraping tool built with Streamlit that extracts building owner contact information from Google search results. The application uses advanced AI-powered extraction methods to identify building names, owner details, contact information, and addresses from web pages.

Features

- **Intelligent Web Scraping**: Uses SerpAPI to search Google and extract relevant URLs
- **Dual Extraction Methods**:
  - Regex pattern matching for fast, reliable extraction
  - Google AI (Gemini) for advanced AI-powered information extraction
- **Email Filtering**: Option to process only pages containing email addresses for efficiency
- **Address Recognition**: Specialized Malaysian address format recognition
- **Duplicate Removal**: Fuzzy matching to eliminate duplicate entries
- **Export Options**: Download results as Excel or CSV files
- **Real-time Progress**: Live updates and processing statistics

Prerequisites

API Keys Required

1. **SerpAPI Key** - Get from serpapi.com
2. **Google AI Studio Key** - Get from Google AI Studio

Python Dependencies

streamlit

pandas

requests

beautifulsoup4

python-dotenv

fuzzywuzzy

openpyxl

Installation

1. **Clone or download the script**
2. **Install dependencies**:
3. pip install streamlit pandas requests beautifulsoup4 python-dotenv fuzzywuzzy openpyxl
4. **Create .env file** in the same directory:
5. SERPAPI_KEY=your_serpapi_key_here
   GOOGLE_API_KEY=your_google_ai_key_here
6. **Run the application**:
7. streamlit run main.py

How to Use

1. Configuration

- **Extraction Method**: Choose between Google AI (Gemini Flash) or Regex Only
- **Email Filtering**: Enable to process only pages with email addresses (recommended for efficiency)

2. Search Parameters

- **Keywords**: Enter search terms related to buildings/factories
  - Example: "factory owner email address seremban"

    - Example: "industrial building contact address kuala lumpur"
- **Number of Results**: Specify how many search results to process (1-50)

3. Processing

- Click **"Start Scraping"** to begin
- Monitor real-time progress and statistics
- Review extracted information for each URL

4. Results

- View comprehensive data table with all extracted information
- See processing statistics (URLs processed, emails found, etc.)
- Download results as Excel or CSV files

Core Functions

Search Functions

serpapi_search(query, num_results=10)

- Searches Google using SerpAPI with pagination support
- Returns list of URLs from organic search results
- Includes rate limiting and error handling

Text Extraction

extract_text_from_url(url)

- Fetches webpage content using requests
- Parses HTML with BeautifulSoup
- Removes scripts and styling for clean text extraction

Information Extraction

Regex-Based Extraction

extract_contacts_regex(text)

extract_address_regex(text)

extract_info_from_text(text)

- **Contacts**: Comprehensive email and phone number patterns
- **Addresses**: Malaysian address format recognition including:
    - Street names (Jalan, Lorong, Taman)
    - Postal codes (5-digit format)
    - State names and industrial areas
    - English and Malay address formats

AI-Powered Extraction

call_google_ai_llm(text)

- Uses Google AI Studio (Gemini 1.5 Flash) model
- Structured prompt for consistent information extraction
- Retry logic for rate limiting and error handling
- Safety settings for content filtering

Data Processing

deduplicate_results(data)

- Fuzzy string matching to identify duplicates
- Uses building names and addresses for comparison
- Configurable similarity thresholds (85% for names, 80% for addresses)

Extracted Information Fields

Field	Description	Source
Source URL	Original webpage URL	Direct
Building Name	Factory/building/facility name	Regex + AI
Owner/Manager Name	Contact person or business owner	Regex + AI
Contact Email	Email addresses found on page	Regex + AI
Phone Number	Phone numbers (Malaysian format supported)	Regex + AI
Address	Complete address including postal code	Regex + AI

Address Recognition Patterns

The system recognizes various Malaysian address formats:

Supported Patterns

- **Street Addresses**: Jalan, Lorong, Taman patterns
- **Postal Codes**: 5-digit Malaysian postal codes
- **Industrial Areas**: Kawasan Perindustrian, Industrial Park
- **State Recognition**: All Malaysian states
- **Lot Numbers**: Lot-based addressing
- **Bilingual Support**: English and Malay terms

Example Addresses Recognized

- "Lot 123, Jalan Industri 2, Kawasan Perindustrian Senawang, 70450 Seremban, Negeri Sembilan"
- "No. 45, Taman Perindustrian Shah Alam, 40000 Shah Alam, Selangor"
- "Bangunan ABC, Lorong Teknologi 3, 81300 Johor Bahru, Johor"

Configuration Options

Email Filtering

- **Enabled** (Recommended): Only processes pages containing email addresses
    - Faster processing
    - Reduced API costs
    - More targeted results
- **Disabled**: Processes all pages regardless of email presence

Extraction Methods

- **Google AI (Gemini Flash)**:
    - Advanced AI-powered extraction
    - Higher accuracy for complex formats
    - 15 requests/minute, 1500 requests/day free tier
- **Regex Only**:
    - Fast pattern-based extraction
    - No API costs
    - Good for standard formats

API Limits and Costs

SerpAPI

- Free tier: 100 searches/month
- Paid plans start from $50/month

Google AI Studio (Gemini)

- **Free Tier**:
    - 15 requests per minute
    - 1,500 requests per day
    - 1 million tokens per day
- **Paid Tier**: Pay-per-use after free limits

Error Handling

Common Issues and Solutions

Error	Cause	Solution
API Key Missing	.env file not configured	Set SERPAPI_KEY and GOOGLE_API_KEY
Rate Limit Exceeded	Too many API calls	Wait and retry, or reduce batch size
No Results Found	Keywords too specific	Try broader search terms
Page Loading Failed	Website blocking or timeout	Automatic retry with different headers

Built-in Error Recovery

- **Retry Logic**: Automatic retries for failed requests
- **Timeout Handling**: 10-30 second timeouts for web requests
- **Rate Limiting**: Automatic delays between API calls
- **Graceful Degradation**: Continues processing if individual URLs fail

Output Formats

Excel Export

- Structured spreadsheet with proper column headers
- Formatted for easy data analysis
- Includes all extracted fields and source URLs

CSV Export

- Standard comma-separated format
- Compatible with most data analysis tools
- UTF-8 encoding for international characters

Limitations

Technical Limitations

- Dependent on website structure and content quality
- Subject to API rate limits
- May not work with JavaScript-heavy sites
- Limited to publicly accessible information

Troubleshooting

Common Solutions

1. **Check API Keys**: Ensure .env file is properly configured
2. **Internet Connection**: Verify stable internet connectivity
3. **Search Keywords**: Try different or broader search terms
4. **Clear Browser Cache**: Refresh the Streamlit application
5. **Check API Quotas**: Verify you haven't exceeded daily limits

Debug Mode

Enable detailed logging by checking the expander sections for each processed URL to see:

- Text extraction status
- Email detection results
- AI processing outputs
- Extracted information breakdown