**AI Supplier Search Agent - Development Guide**

**Project Overview**

**Business Goal:**

Develop an AI-powered web searcher and scraper to identify and categorize industrial suppliers across multiple verticals, such as automotive, lighting, construction, and electronics.

**Technical Goals:**

1. **Web Scraping** – Extract supplier data from directories, company websites, LinkedIn, and trade portals.
2. **Industry Classification** – Categorize suppliers based on industry and record metadata.
3. **Data Cleaning & Enrichment** – Remove duplicates, standardize formats, and validate data.
4. **Search & Ranking** – Implement a query system to filter suppliers based on location, certifications, and capabilities.
5. **UI/API Integration** – Provide structured supplier data via an intuitive UI or API.

**Phase 1: Web Scraping**

**1. Data Sources**

* **Mfg.com** – Online supplier database
* **Alibaba.com** – B2B marketplace (Asia)
* **IndiaMART & TradeIndia** – B2B marketplaces (India)
* **ThomasNet** – B2B marketplace (North America)
* **Xometry** – On-demand manufacturing marketplace

**2. Web Scraping Techniques**

**a. Scrapy (Preferred for large-scale scraping)**

* Asynchronous requests handling
* Good for well-structured websites

**b. Selenium (For dynamic websites with JavaScript rendering)**

* Required for LinkedIn & JavaScript-heavy pages

**c. BeautifulSoup (For parsing HTML efficiently)**

* Used for simple static pages

**3. Data Schema**

Each supplier's data will be structured as:

{

"Company Name": "ABC Manufacturing Co.",

"Website": "https://abcmanufacturing.com",

"Country": "USA",

"Industries Served": ["Automotive", "Construction"],

"Manufacturing Processes": ["Die-Casting", "Forging"],

"Certifications": ["ISO 9001", "CE"],

"Customers": ["Ford", "GM"],

"Metadata": {

"# Employees": 500,

"Annual Revenue": "$10M-$50M"

}

}

**4. Handling Anti-Bot Measures**

* **Use Rotating Proxies** (e.g., ScraperAPI, BrightData)
* **Randomize User-Agents**
* **Set Request Delays**
* **Use CAPTCHA Solving Services**

**5. Data Cleaning & Storage**

* Remove duplicates & normalize text formats
* Convert categorical data into standardized labels
* Store data in **MongoDB (NoSQL)** or **PostgreSQL (Relational DB)**

**Phase 2: AI-Powered Supplier Search & Ranking**

**1. AI-Based Supplier Ranking**

* Use **TF-IDF & BERT-based NLP models** to assess credibility
* Rank suppliers based on:
  + Completeness of data
  + Number of certifications
  + Presence in multiple directories

**2. Automated Contact Retrieval**

* Use **Regex & NLP** to extract emails and phone numbers
* Cross-check against official company websites

**3. Real-Time Monitoring & Updates**

* Implement **cron jobs** to scrape periodically
* Use **WebSockets** to update data dynamically

**4. Search API & UI**

* **Backend:** FastAPI/Flask to handle search queries
* **Frontend:** React/Next.js for a user-friendly supplier search interface

**Technology Stack**

| **Component** | **Technology Choices** |
| --- | --- |
| **Scraping** | Scrapy, Selenium, BeautifulSoup |
| **Database** | MongoDB, PostgreSQL |
| **AI & NLP** | Spacy, BERT, Scikit-learn |
| **Backend API** | FastAPI, Flask |
| **Frontend** | React, Next.js |

**Next Steps**

1. Choose the first website to scrape (Alibaba/IndiaMART recommended)
2. Implement the data extraction module
3. Set up database schema & storage
4. Begin AI model integration for ranking
5. Build API/UI for querying suppliers

This guide provides a detailed roadmap for building an **AI-powered Supplier Search Agent**. Let me know if you need refinements or additional details! 🚀