Web Scraping and Optical Character Recognition(OCR)

Piyawat Chuangkrud

Installation

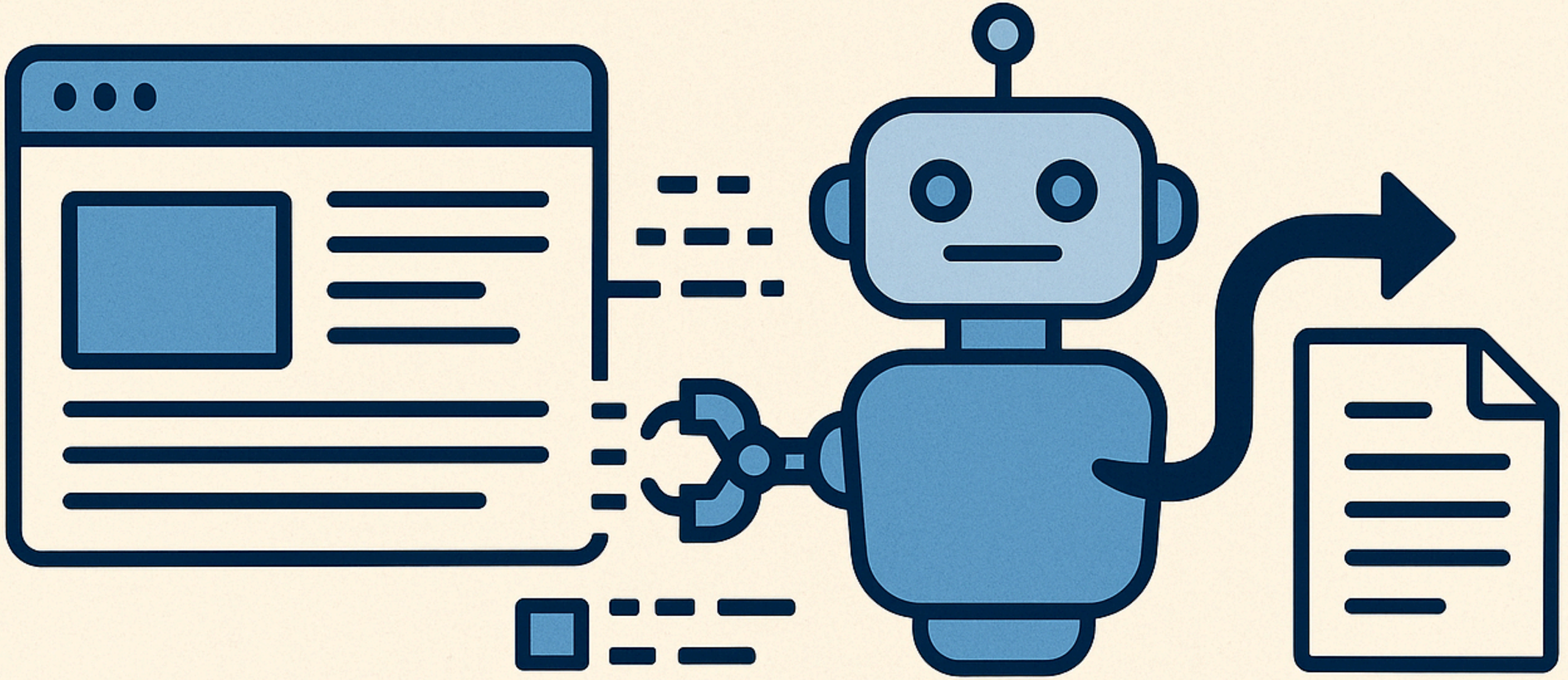
1. **git clone https://github.com/chuangk-p/MOC-workshop.git**
2. **cd MOC-workshop**
3. **pip install -r requirements.txt**
4. **crawl4ai-setup**
5. **playwright install**
6. **crawl4ai-doctor**

7. **if use Ubuntu**
   sudo apt-get install poppler-utils
   if use MacOS
   brew install poppler

WEB SCRAPING

How can we communicate with a website?

WHAT IS WEB SCRAPING?



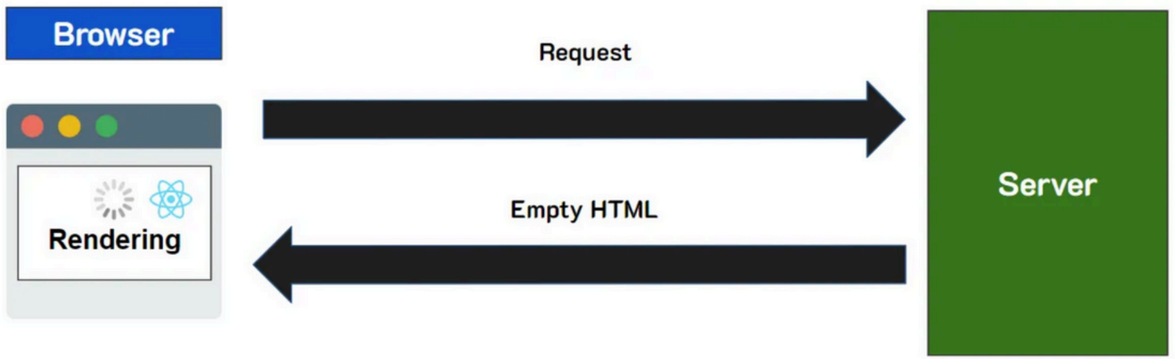
List of Web Scraping Framework

crawl4ai

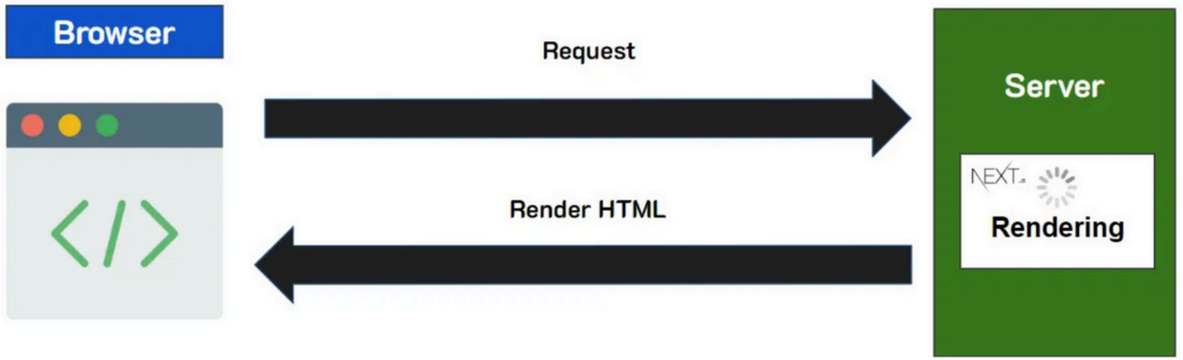
What is Web Scraping

Web scraping is the **process of automatically extracting data from websites**. It involves using software, often referred to as web crawlers or bots, to retrieve and process information from web pages

Client-side Rendering and Server-side Rendering



Client-side Rendering



Server-side Rendering

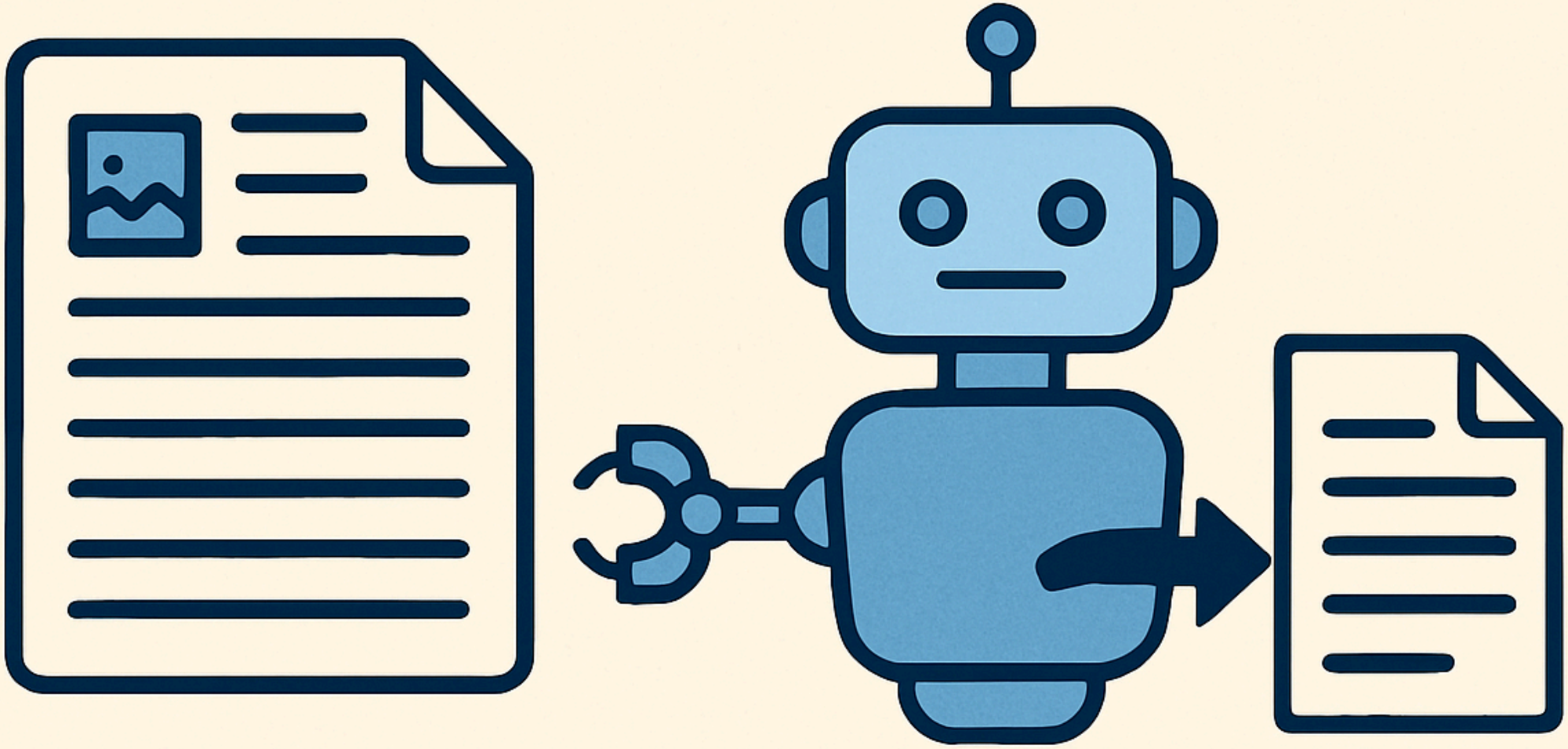
reference: https://kongruksiam.medium.com/รู้จักกับ-csr-ssr-และ-ssg-สำหรับการพัฒนาเว็บแอพพลิเคชั่น-7c99b9974f8e

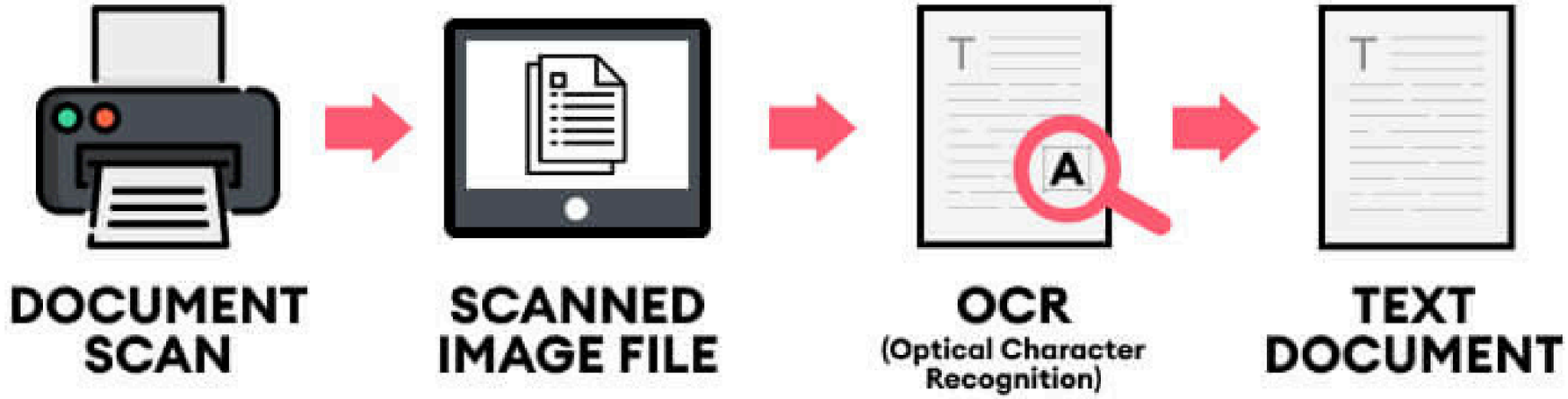
Compare Web Scraping Framework

Feature/Library	Pros	Cons	Client-Side Rendering	Server-Side Rendering	Speed	Default Markdown Output	LLM Integration
BeautifulSoup	Lightweight, fast for static HTML, easy to use, minimal setup	Static only, needs extra libraries (e.g., Requests), no JS rendering	No	Yes	Fastest for static content	No	No
Selenium	Handles dynamic content, supports JS, simulates user actions, bypasses anti-bot	Slow, resource-heavy, complex setup, scripts may break	Yes	Yes	Slowest (browser overhead)	No	No
Trafilatura	Fast, optimized for articles, clean text, Markdown output, multi-language	Less versatile, limited JS rendering	Limited (some JS via integrations)	Yes	Fast (lightweight)	Yes	Partial
Crawl4AI	LLM-focused, fast (6x claimed), async, Markdown output, open-source	Newer, less mature, documentation in progress	Yes	Yes	Very fast (async)	Yes	Yes

Optical Character Recognition (OCR)

WHAT IS OCR?





OCR, or Optical Character Recognition, is a technology that converts images of text into machine-readable text.