الوصف الشامل لنطاق العمل (Project Scope of Work)

1. الهدف العام للمشروع (Objective)

يستهدف المواقع والمنصات (Cloud-Based Web Scraper & Data Pipeline) بناء نظام سحب بيانات ذكي وسحابي التابعة للجهات الحكومية وشبه الحكومية والشركات المرتبطة بها، لسحب جميع البيانات والمستندات، تصنيفها تلقائياً، وتخزينها في قاعدة بيانات، مع توفير واجهة ربط (API / MCP Server) لربطه بنماذج الذكاء الاصطناعي (Claude مثل) وتسهيل عملية الاستدعاء الآلي.

---

2. نطاق سحب البيانات (Data Extraction Scope)

* المواقع المستهدفة:

مواقع الجهات الحكومية، شبه الحكومية، والشركات التابعة أو المنشأة بواسطة جهات حكومية.

* نوع البيانات المطلوبة للسحب:

* القضايا والنزاعات: تفاصيل القضايا ومحتواها وتصنيفاتها (تجاري، أحوال شخصية، إلخ).

* المستندات واللوائح: القواعد، الأنظمة، اللوائح، والملفات المنشورة بجميع الصيغ (PDF / DOCX).

* حالة الوثائق: تحديد ووسم حالة الوثيقة (سارية / غير سارية / ملغاة) بناءً على ما يعلنه الموقع.

* الأخبار والتحديثات: كل الأخبار والبيانات الصحفية الصادرة عن الجهات.

* التعامل مع الحماية وتسجيل الدخول (Authentication):

دعم السحب من المواقع المغلقة التي تتطلب اسم مستخدم وكلمة مرور (Login Auth Sessions) وتخزين الـ Cookies/Tokens لتجديد الجلسات تلقائياً.

---

3. معالجة وتصنيف البيانات (Data Processing & Structuring)

* التصنيف الآلي (Smart Categorization):

عزل كل فئة بيانات بشكل مستقل لمنع اختلاط البيانات (مثل: الأخبار السياسية منفصلة، الاتفاقيات منفصلة، القضايا بحسب نوعها).

* هيكلة البيانات (Data Structuring):

تحويل النصوص والبيانات غير المنتظمة (Unstructured) إلى صيغ منتظمة (Structured JSON) تشمل:

Title, Content, Date, Category, Source_URL, Document_Status, Metadata

* تنظيف البيانات (Data Cleaning):

إزالة الوسوم البرمجية (HTML tags)، والعناصر المكررة لضمان جودة النص السحابي، الإعلانات.

---

4. البنية السحابية والربط (Cloud Architecture & Integration)

* المحرك السحابي (Cloud Scraper Engine):

رفع السكرايبر على بيئة سحابية (Cloud Server / Serverless Functions) ليعمل بمرونة ودون الحاجة لتشغيله من الجهاز الشخصي.

* واجهة البرمجة والربط (API / MCP Integration):

إنشاء API Endpoint أو MCP Server إرسال رابط (Claude مثل) بحيث يستطيع العميل أو نماذج الذكاء الاصطناعي (URL) لتقوم السحابة بسحب الموقع فوراً وإعادة البيانات أو تخزينها.

* الأمان والامتثال البرمجي (Safe Scraping):

تطبيق تقنيات التأخير الزمني (Rate Limiting) وتدوير الهويات (User Agents) لضمان عدم إجهاد خوادم الجهات المستهدفة أو تسبب أي حظر أو أضرار تقنية لها.

---

5. المرحلة المستقبلية: تفريغ المستندات (OCR Pipeline)

* معالجة الملفات المصورة (Scanned PDFs/Images):

بناء مسار آلي لدمج تقنية التعرف الضوئي على الحروف (OCR) والـ PDFs المصورة لاستخراج النصوص من المستندات غير القابلة للنسخ، لربطها بنفس قاعدة البيانات.

---

6. المخرجات النهائية للمشروع (Deliverables Summary)

* Scraping Engine:

سكرايبر بايثون متطور يدعم التصفح التفاعلي وسحب الملفات وتسجيل الدخول.

* Categorization Pipeline:

نظام فلترة وتصنيف للبيانات المسحوبة بحسب النوع والحالة.

* Cloud API / MCP Connector:

واجهة سحابية تتيح لـ Claude أو للعميل إعطاء أوامر السحب مباشرة عبر الرابط.

* Database Schema:

قاعدة بيانات مهيكلة تحتوي على كل البيانات والأخبار والمستندات المجمعة.

* OCR Module (المرحلة الثانية):

أداة تحويل الملفات الممسوحة ضوئياً إلى نصوص مكتوبة.

(انا هكون وسيط بين المطور والعميل ولي نسبة 50% من الربح النهائي)

***(لايوجد دفع قبل التنفيذ – الدفع خارج منصة

نفذلي وكل التعامل بالخارج – الدفع بعد التسليم من قبل العميل) *** ويجب من المطور عمل تحكم كامل له بداخل الاداه لمنع حالات النصب من قبل العميل

اللي جاهز يتعاون ويكون معي يتواصل واتساب

01280432851