الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الشهيد حمه لخضر الوادي
كلية العلوم الدقيقة
قسم الإعلام الآلي

مذكرة نهاية تخرج ضمن متطلبات الحصول على شهادة
ليسانس أكاديمي

خدمة البحث في إعلانات الجرائد «OfferSeek»

من إنجاز الطالب:
عيادي أحمد ياسين

تحت إشراف:
غنابزية أحمد

نوقشت يوم 25\ 05 \ 2025 أمام اللجنة المكونة من الأساتذة
خلايفة عبد الناصر ..... رئيسا
خالد سلطاني ..... مقررا

السنة الجامعية: 2025/2024

الملخص

تشكل المعلومات الإعلانية المنشورة في الصحف مصدرًا مهمًا للباحثين عن فرص العمل، العقارات، المناقصات، وغيرها من الخدمات .ورغم التحول الرقمي، ما تزال العديد من الإعلانات تُنشر بشكل رئيسي عبر الصحف الورقية، مما يصعب الوصول إليها والبحث فيها بشكل فعال.

بناءً على ذلك، يهدف هذا المشروع إلى رقمنة محتوى الإعلانات المنشورة في الصحف اليومية وتوفيرها ضمن منصة رقمية ذكية، تتيح للمستخدمين تصفح الإعلانات، البحث فيها وتلقي تنبيهات دورية حسب اهتماماتهم.

تم تطوير هذا المشروع وفق خطة مرحلية تعتمد على أدوات حديثة وتكنولوجيا مفتوحة المصدر وذلك من أجل تقليل التكاليف وتحقيق استقرار وأداء عالٍ، بدءًا من جمع أرشيف الصحف وتحليلها، مرورًا باستخراج الصفحات الإعلانية والبيانات منها تلقائيًا وذلك باستخدام تقنيات OCRوالذكاء الاصطناعي، ووصولًا إلى توفير واجهات مراجعة بشرية وقاعدة بيانات منظمة وتطبيق موبايل تفاعلي.

من خلال هذا النظام، يمكن ربط مصادر البيانات غير المهيكلة ملفات (PDF)بمنصة بحث ديناميكية سهلة الاستخدام، مما يفتح المجال أمام تطوير خدمات مستقبلية تعتمد على تحليل البيانات وتخصيص المحتوى حسب المستخدم، مع الحفاظ على جودة المعلومات ودقة التصنيف.

Abstract

Advertisement sections in printed newspapers remain a valuable source of information for individuals seeking job opportunities, tenders, real estate offers, and more. Despite the ongoing shift toward digital media, many critical announcements are still published exclusively in traditional print format, making them difficult to access, browse, and track efficiently.

This project aims to digitize newspaper advertisement content and present it through a smart, user-friendly digital platform. The platform allows users to explore newly published ads, search using multiple filters (keywords, location, type, date), save favorites, and receive customized notifications based on their interests.

The development process followed a structured, multi-phase approach using modern, cost-effective, open-source technologies to ensure stability, scalability, and high performance. The stages included building an initial archive of PDF newspapers, automating daily updates, identifying and extracting ad-specific pages, converting them into structured data using OCR and AI tools, validating and cleaning the data via a human review interface, and storing the final records in an indexed database. A mobile application was then developed to provide seamless access to the data.

By bridging the gap between unstructured print media and dynamic digital search, this platform provides a practical and scalable solution that enhances access to classified information, opening the door to future AI-powered personalization and analytics in the digital transformation of public data.

Résumé

Les sections d'annonces publiées dans les journaux papier représentent encore une source précieuse d'informations pour les personnes à la recherche d'opportunités d'emploi, d'appels d'offres, d'offres immobilières, et autres services. Malgré la transition vers le numérique, de nombreuses annonces importantes continuent d'être diffusées exclusivement au format imprimé, ce qui rend leur consultation et leur suivi difficiles.

Ce projet vise à numériser le contenu des annonces publiées dans les journaux et à le proposer via une plateforme numérique intelligente et conviviale. Celle-ci permet aux utilisateurs de parcourir les nouvelles annonces, d'effectuer des recherches multi-critères (mots-clés, wilaya, type, date), d'enregistrer leurs annonces favorites et de recevoir des notifications personnalisées selon leurs préférences.

Le développement du système s'est basé sur une approche en plusieurs étapes, utilisant des technologies modernes, open source et peu coûteuses afin de garantir stabilité, performance et évolutivité. Les phases ont inclus la constitution d'un archivage initial de journaux au format PDF, l'automatisation de la mise à jour quotidienne, l'identification et l'extraction des pages contenant des annonces, la conversion des contenus en données structurées à l'aide d'OCR et d'outils d'intelligence artificielle, la vérification manuelle via une interface de validation, et enfin, l'enregistrement des données finales dans une base de données indexée. Une application mobile a été développée pour offrir un accès fluide aux utilisateurs.

En reliant les sources d'information non structurées à une plateforme numérique dynamique, ce projet propose une solution pratique et extensible, favorisant l'accès à l'information et préparant le terrain à de futures améliorations basées sur l'IA et la personnalisation des contenus.

إهداء

إلى شعب فلسطين الأبيّ،
إلى من سطّروا بصمودهم أروع ملاحم العزّة والكرامة،
إلى المجاهدين في الثغور، الحراس الساهرين على الأرض والعقيدة،
إلى من جعلوا من أرواحهم درعًا لفلسطين، ومن جراحهم طريقًا للنصر...
لكم كل الدعاء، وكل الحروف.

كان الله في عونكم،
ولتحيا فلسطين حرّة أبية، من البحر إلى النهر.

الفهرس

قائمة الأشكال

قائمة الجداول

الفصل الأول: تعريفات ودراسة الموجود

1 مقدمـــــــــــــة

تعد إعلانات المناقصات من أبرز الوسائل التي تعتمدها المؤسسات العامة والخاصة للإعلان عن المشاريع الجديدة. غير أن تتبّع هذه الإعلانات يواجه تحديات كبيرة، خاصةً عندما تُنشر في الجرائد الورقية. وتكمن المشكلة في صعوبة الوصول إليها بشكل رقمي، بالإضافة إلى غياب تنظيم معلوماتي يتيح للمستخدم الوصول السريع والدقيق إلى المحتوى الذي يهمّه.

يستدعي هذا الواقع ضرورة التحول الرقمي في عملية أرشفة وتحليل هذه الإعلانات، بما يتيح للمستخدمين والمهنيين إمكانية تصفح الإعلانات، البحث عنها بناءً على معايير متنوعة مثل نوع المناقصة، تاريخ النشر، والولاية، بالإضافة إلى تلقي إشعارات لحظة حدوث تحديثات أو نشر إعلانات جديدة .من خلال تطبيق هذه التقنيات الحديثة، يصبح الوصول إلى إعلانات المناقصات أكثر كفاءة وفعالية.

الهدف من المشروع هو بناء نظام رقمي مبتكر يقوم بجمع إعلانات المناقصات من الجرائد الورقية بشكل آلي، وتنظيم هذه الإعلانات في قاعدة بيانات قابلة للبحث والاستعلام .بالإضافة إلى ذلك، يهدف المشروع إلى تقديم واجهة تفاعلية تسهل على المستخدمين تصفح الإعلانات، وتوفر إمكانية البحث المتقدم وتنبيهات حول المناقصات التي قد تهمهم.

2 دراسة الموجود

1.2 بيئة العمل الحالية

في الجزائر، تعتمد المؤسسات والإدارات العمومية على الجرائد الورقية كوسيلة رسمية للإعلان عن المناقصات العمومية .يتم نشر هذه الإعلانات بشكل يومي أو أسبوعي حسب نشاط المؤسسة وطبيعة المشاريع، وغالبًا ما تكون هذه الإعلانات موزعة ضمن صفحات مخصصة داخل الجريدة و بدون تصنيف رقمي قابل للبحث، لذلك يعاني المتعاملون الاقتصاديون(مؤسسات، مقاولات، موردون ...)من صعوبات كثيرة نذكر منها:

- تتبع كل الجرائد بصفة يومية.
- فرز الإعلانات ذات الصلة بمجالهم الجغرافي أو التقني.
- احترام آجال التقديم بسبب التأخر في الوصول إلى المعلومة.

2.2 سيرورة العمل التقليدي

العملية التقليدية لمتابعة المناقصات تعتمد على ما يلي:

1. الاشتراك في عدد كبير من الجرائد ورقيًا أو على شكل PDF من مواقع الجرائد.
2. الفحص اليدوي لكل عدد يومي للبحث عن الإعلانات ذات الصلة.
3. النسخ أو الطباعة للاحتفاظ بالإعلانات المهمة.
4. حفظ المعلومات في ملفات أو دفاتر غير قابلة للبحث المهيكل.

هذا النمط بدائي ويستنزف الوقت والموارد، كما أنه لا يتيح تحليلًا أو تصنيفًا آليًا للإعلانات.

3.2 التطبيقات والحلول الحالية

رغم وجود بعض المبادرات الرقمية لمتابعة إعلانات المناقصات، إلا أن معظمها يعاني من محدودية واضحة من حيث التغطية والوظائف. إضافة إلى ذلك، تفتقر هذه الحلول إلى استخدام تقنيات الذكاء الاصطناعي لاستخراج المعلومات بشكل آلي، ولا توفر خدمات تخصيص النتائج أو إرسال تنبيهات ذكية حسب اهتمامات المستخدم .كما أن بعضها لا يتيح أرشفة دقيقة للتواريخ أو لواجهات بحث متقدمة، مما يحدّ من فعاليتها بالنسبة للمهنيين الذين يعتمدون على هذه الإعلانات من أجل اتخاذ قراراتهم التي تكون مبنية على معلومات دقيقة ومحدثة.

التطبيق / المنصة	المزايا	النقائص
مواقع الجرائد الرسمية	توفر الأعداد بصيغة PDF	لا توجد خاصية بحث فعالة أو تصنيف واضح لإعلانات المناقصات
بعض المواقع الخاصة بجمع المناقصات	تقدم قوائم جاهزة للمناقصات	تغطية غير شاملة، مصادر محدودة، غالبًا ما تكون خدماتها مدفوعة
وسائل التواصل (فايسبوك...)	إشعارات سريعة وشبه فورية	غير منظمة، غير موثوقة، تغطية عشوائية وغير قابلة للأرشفة الدقيقة

جدول 1 مقارنة بين الحلول الرقمية الحالية لمتابعة إعلانات المناقصات

4.2 المجتمع المستهدف

1.4.2 المؤسسات الاقتصادية

- مقاولات بناء، توريد، خدمات، دراسات...
- تعتمد على المناقصات كمصدر رئيسي للأعمال.
- تبحث عن حلول فعالة لتصفية وتحليل الإعلانات.

2.4.2 الإدارات

- بعض الإدارات ترغب في تتبع المناقصات المنشورة من طرف الغير لتقييم المنافسة أو ضبط برامجها.

3.4.2 وسطاء الأعمال

- مكاتب استشارات، محاسبون، موثقون...، والذين يعملون على تزويد زبائنهم بمعلومات عن الفرص المتاحة.

5.2 مقارنة بين العمل التقليدي والعمل الرقمي المقترح

العنصر	العمل التقليدي	الحل الرقمي المقترح عبر المنصة
مصدر المعلومات	الجرائد الورقية أو PDF اليومية	أرشيف رقمي منظم حسب الجريدة والتاريخ
طريقة البحث	قراءة يدوية لكل عدد	محرك بحث داخلي وتصنيفات متعددة (نوع، ولاية، تاريخ...)
الزمن اللازم للبحث	طويل – يتطلب تصفح عدد كبير من الصفحات يدوياً	سريع – يعتمد على تقنيات OCR وذكاء اصطناعي لفهم الإعلانات
التحديث	يدوي (قراءة الجرائد يوميًا)	آلي (تحديث تلقائي يومي أو أسبوعي للجرائد الجديدة)
إشعارات بالفرص الجديدة	غير متوفرة	متوفرة – عبر التطبيق حسب اهتمامات المستخدم
تنظيم الإعلانات	غير منظم – يعتمد على الحفظ الفردي	قاعدة بيانات مؤشرة – قابلة للفرز والتحليل
إمكانية التتبع	صعبة – لا توجد طريقة لمتابعة المناقصة بعد نشرها	ممكنة – متابعة وتذكير تلقائي بآجال المشاركة
الموثوقية والدقة	عرضة للأخطاء والنسيان	تدقيق يدوي بعد الاستخراج لضمان جودة المعلومات

جدول 2 مقارنة بين العمل التقليدي والعمل الرقمي المقترح

3 تقنية Web Scraping

1.3 تعريفها

يُعرف مصطلح "Web Scraping" (تجريف الويب) بأنه عملية منهجية وآلية من أجل استخلاص البيانات والمعلومات المنظمة وغير المنظمة من مواقع الويب. فهذه العملية تتجاوز مجرد نسخ ولصق المحتوى يدويًا، حيث تستخدم خوارزميات وبروتوكولات محددة للتفاعل مع بنية مواقع الويب بطريقة تحاكي تصفح المستخدم البشري أو تتجاوزه في بعض الأحيان.

يستخدم "Web Scraping" برامج نصية (Scripts) مكتوبة بلغات برمجة مثل بايثون أو جافا أو غيرها، أو أدوات برمجية متخصصة تقوم بأتمتة عملية الوصول إلى صفحات الويب، وتحليل نموذج كائن المستند DOM (Document Object Model) الخاص بها، وتحديد واستخراج البيانات المطلوبة بناءً على قواعد وأنماط محددة مسبقًا بشكل منظم وفعال.

يمكن تشبيه عملية "Web Scraping" بعملية مسح أو تجريف رقمي لسطح موقع الويب لاستخلاص العناصر والمعلومات القيمة التي يحتوي عليها. تتضمن هذه العملية العديد من الخطوات التقنية المعقدة، بدءًا من إرسال طلبات HTTP (Hypertext Transfer Protocol) إلى خوادم الويب لجلب صفحاتHTML ، مرورًا بتحليل هيكل صفحة الويب باستخدام محللات HTML (HTML Parsers)، وتحديد العناصر المراد استخراجها باستخدام محددات CSS (Cascading Style Sheets Selectors)أو تعبيرات XPath (XML Path Language)، وصولًا إلى تحويل البيانات المستخرجة إلى تنسيق مهيكل وسهل الاستخدام.

تتضمن هذه العملية ايضا تحليل لهيكل صفحة الويب، وتحديد العناصر المراد استخراجها بناءً على معايير محددة (مثل النصوص، الروابط التشعبية(Hyperlinks) ، الصور، الجداول، نماذج البيانات(Data Forms) ، وغيرها من العناصر المرئية وغير المرئية في كود المصدر، ثم حفظ هذه البيانات في تنسيق منظم مثل قواعد البيانات العلائقية (Relational Databases)، أو ملفات CSV (Comma Separated Values)، أو JSON (JavaScript Object Notation)، أو حتى ملفات XML (Extensible Markup Language)لتسهيل تحليلها واستخدامها في تطبيقات أخرى.

2.3 استخداماتها وتطبيقاتها في العصر الرقمي

يشهد "Web Scraping" انتشارًا واسعًا نظرًا للكم الهائل من البيانات المتاحة على الإنترنت. حيث تتعدد استخداماته وتطبيقاته في مختلف المجالات، ومن أبرزها:

- تحليل المنافسين: جمع بيانات حول أسعار المنتجات، واستراتيجيات التسويق، وتقييمات العملاء للمنافسين للحصول على رؤى تنافسية.
- مراقبة الأسعار: تتبع تغيرات أسعار المنتجات عبر مواقع التجارة الإلكترونية المختلفة لاتخاذ قرارات شراء أو بيع.
- تجميع الأخبار والمعلومات: جمع المقالات الإخبارية، والتقارير، والبيانات من مصادر متعددة لتوفير نظرة شاملة حول موضوع معين.
- أبحاث السوق: استخلاص بيانات حول اتجاهات السوق، وآراء المستهلكين، والمنتجات الرائجة لاتخاذ قرارات استراتيجية.
- التنقيب عن البيانات :(Data Mining) جمع كميات كبيرة من البيانات من الويب لاستخدامها في تدريب نماذج الذكاء الاصطناعي وتحليل الأنماط.
- التسويق الرقمي: جمع بيانات حول العملاء المحتملين، وتحليل سلوك المستخدمين، وأتمتة بعض مهام التسويق.
- تحسين محركات البحث :(SEO) تحليل مواقع الويب المنافسة لفهم استراتيجياتهم وتحديد فرص التحسين.

3.3 أهمية فهم شروط الخدمة وسياسات الاستخدام لمواقع الويب في Web Scraping

عند الشروع في أي عملية"Web Scraping" ، من الضروري للغاية تخصيص وقت وجهد كافيين لمراجعة وفهم شروط الخدمة (Terms of Service - ToS) وسياسات الاستخدام (Usage Policies) الخاصة بالمواقع المستهدفة. هذه الوثائق القانونية تحدد القواعد والقيود التي يفرضها مالكو الموقع على كيفية تفاعل المستخدمين والبرامج الآلية مع محتواهم وخدماتهم. تجاهل هذه الشروط قد يؤدي إلى عواقب وخيمة على المستويات القانونية والتقنية والأخلاقية.

1.3.3 الأبعاد القانونية

- تجنب الحظر (Avoiding Blocking) :غالبًا ما تتضمن شروط الخدمة بنودًا تحظر استخدام برامج آلية "Bots"أو "Spiders" للوصول إلى الموقع بطرق قد تؤثر على أدائه أو بنيته التحتية. قد تقوم المواقع بتطبيق آليات تقنية للكشف عن هذه الأنشطة وحظر عناوين IP أو الوكلاء (Proxies) المستخدمة في عمليات "Scraping"المخالفة. فهم الشروط قد يساعد في تصميم عمليات "Scraping" تحترم قيود معدل الطلبات والسلوك المقبول لتجنب الحظر.
- احترام ملف Robots.txt: على الرغم من أنه ليس وثيقة قانونية ملزمة بشكل صارم، إلا أن ملف robots.txtهو معيار صناعي يشير إلى الأقسام التي لا ينبغي على برامج الروبوت الآلية الوصول إليها. غالبًا ما يُنصح باحترام هذا الملف كحد أدنى من السلوك الأخلاقي والتقني، وقد يُذكر في شروط الخدمة كجزء من السياسات المقبولة.

2.3.3 الأبعاد التقنية

- تحميل زائد على الخوادم (Server Overload) : قد تؤدي عمليات "Scraping" المكثفة وغير المسؤولة إلى إبطاء أداء الموقع أو حتى تعطيله، مما يؤثر سلبًا على تجربة المستخدمين الآخرين. شروط الخدمة غالبًا ما تهدف إلى حماية موارد الموقع من الاستخدام المفرط.
- المنافسة غير العادلة :في بعض الحالات، قد يُنظر إلى "Scraping" بيانات المنافسين بكميات كبيرة واستخدامها بطرق معينة على أنه ممارسة غير أخلاقية أو منافسة غير عادلة، خاصة إذا كان ذلك يتعارض مع شروط الخدمة.

لتجنب هذه الانتهاكات، يُنصح بما يلي:

- قراءة شروط الخدمة وسياسات الاستخدام بعناية :قبل البدء في أي عملية"Scraping" .
- التحقق من وجود أي بنود تحظر "Scraping" أو استخدام برامج آلية.
- فهم القيود المفروضة على استخدام البيانات المستخرجة.
- احترام ملف robots.txt.
- تصميم عمليات "Scraping" تكون مهذبة ومسؤولة.

- التفكير في التواصل مع مالك الموقع لطلب إذن إذا كانت هناك حاجة لاستخراج كميات كبيرة من البيانات أو استخدامها لأغراض تجارية.
- الحصول على استشارة قانونية إذا كانت هناك أي شكوك حول شرعية عملية "Scraping" معينة.

4.3 التحديات التقنية في عمليةWeb Scraping

على الرغم من قوة وفعالية تقنيات"Web Scraping" ، إلا أنها تواجه العديد من التحديات التقنية التي يجب على المطورين والممارسين أخذها في الاعتبار والتغلب عليها لضمان نجاح واستدامة عمليات استخلاص البيانات. من أبرز هذه التحديات:

1.4.3 التغيرات المستمرة في هيكل مواقع الويب

- طبيعة ديناميكية :مواقع الويب ليست ثابتة؛ فهي تخضع لتحديثات وتغييرات مستمرة في تصميمها، وتنسيق بياناتها، وهيكلها (HTML Structure).
- تأثير على محددات العناصر :تعتمد برامج "Scraping" بشكل كبير على محددات العناصر مثل CSS Selectorsو XPathلتحديد واستخراج البيانات المطلوبة. أي تغيير في هيكل HTML يمكن أن يجعل هذه المحددات غير صالحة، مما يؤدي إلى الحصول على بيانات غير صحيحة أو فشل عملية الاستخلاص.
- الحاجة إلى صيانة مستمرة :لمواجهة هذا التحدي، تتطلب برامج "Scraping" مراقبة دورية وتحديثات مستمرة لتكييفها مع التغيرات في هيكل المواقع المستهدفة، ويؤدي هذا إلى استهلاك وقت وجهد كبيرين.

2.4.3 استخدام تقنيات مكافحة "Scraping" (Anti-Scraping Techniques)

- الكشف عن الروبوتات (Bot Detection): تستخدم العديد من مواقع الويب تقنيات متطورة للكشف عن الزيارات الآلية القادمة من برامج "Scraping" وتمييزها عن الزيارات البشرية. تشمل هذه التقنيات تحليل سلوك المستخدم (مثل سرعة التصفح، أنماط النقر)، واستخدام CAPTCHAوreCAPTCHA، وتحليل بصمات المتصفح (Browser Fingerprinting) .
- حظر عناوين IP والوكلاء (IP and Proxy Blocking): عند اكتشاف نشاط "Scraping" مشبوه، قد تقوم المواقع بحظر عناوين IP الخاصة بالخوادم أو الأجهزة التي تقوم بالاستخلاص، أو حتى حظر نطاقات الوكلاء المستخدمة لإخفاء الهوية .
- تحديد معدل الطلبات (Rate Limiting): للحد من تأثير الزيارات الآلية على أداء خوادمها، قد تفرض المواقع قيودًا على عدد الطلبات التي يمكن إرسالها من عنوان IP معين خلال فترة زمنية محددة. تجاوز هذه القيود قد يؤدي إلى حظر مؤقت أو دائم .
- استخدام ملفات تعريف الارتباط (Cookies) وجلسات التصفح (Sessions): تتطلب بعض المواقع التعامل مع ملفات تعريف الارتباط وإدارة جلسات التصفح للوصول إلى البيانات، مما يزيد من تعقيد عملية "Scraping".

- تغيير هياكل الصفحات بشكل ديناميكي :قد تقوم بعض المواقع بتغيير هياكل الصفحات أو إضافة عناصر وهمية بشكل عشوائي بهدف تضليل برامج"Scraping".

3.4.3 المحتوى الديناميكي الذي يتم تحميله بواسطةJavaScript

- التحدي الرئيسي :تعتمد العديد من مواقع الويب الحديثة على لغة JavaScript لإنشاء وتحديث المحتوى بشكل ديناميكي بعد تحميل الصفحة الأولية (Client-Side Rendering) .برامج تحليل HTML التقليدية مثل BeautifulSoupلا يمكنها تنفيذ JavaScript أو الوصول إلى المحتوى الذي يتم إنشاؤه بهذه الطريقة .
- الحلول التقنية :للتغلب على هذا التحدي، يلجأ المطورون إلى استخدام أدوات يمكنها محاكاة متصفح حقيقي وتنفيذJavaScript ، مثل Selenium و Playwright و .Puppeteer هذه الأدوات تسمح للبرنامج بالانتظار حتى يتم تحميل المحتوى الديناميكي قبل استخلاصه .
- تكلفة الأداء :استخدام هذه الأدوات غالبًا ما يكون أكثر استهلاكًا للموارد وأبطأ في الأداء مقارنة بتحليل HTMLالثابت.

وهناك أيضا بعض التحديات الإضافية نذكر منها:

- الحاجة إلى تسجيل الدخول (Authentication) :تتطلب بعض البيانات الوصول إلى حسابات المستخدمين، مما يستلزم تنفيذ آليات تسجيل الدخول في برنامج"Scraping".
- التعامل مع تنسيقات البيانات المختلفة :قد تكون البيانات المطلوبة مضمنة في هياكل HTML معقدة، أو في جداول ديناميكية، أو حتى في ملفات مضمنة (مثل JSON أو XML) تحتاج إلى تحليل إضافي.
- ضمان جودة البيانات :يجب على برامج "Scraping" التعامل مع البيانات غير المتسقة، أو المفقودة، أو بتنسيقات غير متوقعة، وتنفيذ عمليات تنظيف وتوحيد للبيانات المستخرجة.
- التوسع والموثوقية :عند التعامل مع كميات كبيرة من البيانات أو عدد كبير من المواقع، يصبح تصميم نظام "Scraping"قابل للتوسع وموثوق به تحديًا هندسيًا كبيرًا.

4 مكتبات بايثون المستخدمة فيScraping

تُعد لغة البرمجة بايثون (Python) واحدة من أكثر اللغات شيوعًا واستخدامًا في مجال "Web Scraping" نظرًا لما توفره من مكتبات وأدوات قوية ومرنة تسهل عملية استخلاص البيانات من الويب. من أبرز هذه المكتبات:

- مكتبةBeautifulSoup : لتحليل مستندات HTML وXML، وتوفر طريقة سهلة للتنقل في هيكل المستند واستخراج البيانات المطلوبة.
- مكتبة Selenium :أداة لأتمتة تفاعلات المتصفح، مما يسمح بتصفح المواقع التي تعتمد على JavaScript بشكل كبير واستخلاص البيانات بعد تحميلها ديناميكيًا.
- مكتبة Playwright :حديثة مقدمة من شركةMicrosoft ، توفر واجهة برمجة تطبيقات (API) واحدة لأتمتة متصفحات Chromium ، Firefox وWebKit ، وتدعم العديد من الميزات المتقدمة لعمليات "Scraping" المعقدة.

1.4 مكتبة BeautifulSoup

1.1.4 التعريف

تُعد مكتبة BeautifulSoupأداة متقدمة لمعالجة مستندات HTMLو XMLفي لغة بايثون، حيث لا تقتصر وظيفتها على تحليل البنية النصية لهذه المستندات، بل تقوم بتحويلها إلى هيكل بيانات هرمي منظم يسهل التعامل معه برمجياً من خلال كائنات بايثونية .تمثل هذه الكائنات العناصر المختلفة في المستند، مثل الوسوم والخصائص والنصوص .يمكن تشبيه بنية صفحة الويب بشجرة هيكلية معقدة، بينما تؤدي BeautifulSoupدور الوسيط الذي يوجه الباحث ضمن هذه البنية لتحديد واستخلاص البيانات المطلوبة.

توفر المكتبة ايضا إمكانيات مرنة للبحث والاستخلاص اعتماداً على بنية المستند، حيث تتيح تحديد العناصر المستهدفة باستخدام أسماء الوسوم (مثل ، <a>،)، أو من خلال خصائصها (مثل "id="main أو ("class="item، أو حتى وفقاً للعلاقات الهرمية بين العناصر (مثل استخراج جميع الروابط ضمن قسم معين) .

2.1.4 كيفية عمل مكتبةBeautifulSoup

BeautifulSoup لا تعمل بشكل مستقل، بل تحتاج إلى مكتبة أخرى مثل requests لجلب محتوى الصفحة أولًا. ثم تقوم بتحليل (Parsing) المحتوى باستخدام بنيةHTML ، لتمكّنك من استخراج العناصر التي تريدها.

خطوات العمل:

1. إرسال طلب إلى الصفحة باستخدام (requests.get(url.
2. قراءة كود HTML للصفحة الذي تم استرجاعه.
3. تحليل الكود باستخدام BeautifulSoup لإنشاء شجرةDOM.
4. البحث داخل الشجرة باستخدام طرق مثل .()find_all ,() أوCSS Selectors.
5. استخراج البيانات مثل النصوص، الروابط، الصور، الجداول...

ملاحظة : BeautifulSoupتتعامل مع الصفحات الثابتة فقط، والتي لا تحتاج لتشغيل JavaScript لعرض البيانات.

3.1.4 المميزات

- سهولة الاستخدام: تتميز بواجهة برمجة تطبيقات (API) بسيطة وسهلة التعلم، مما يجعلها خيارًا جيدًا للمبتدئين في مجال"Web Scraping" .
- مرونة في التحليل: قدرة BeautifulSoup على التعامل مع HTML/XML سيء التكوين هي ميزة قيمة في عالم الويب الحقيقي، حيث لا تلتزم جميع المواقع بمعايير صارمة. يمكنها محاولة إصلاح الأخطاء الشائعة وإنشاء شجرة تحليلية قابلة للاستخدام قدر الإمكان. يمكنك أيضاً توجيهها لاستخدام محللات مختلفة تكون أكثر تسامحاً مع أنواع معينة من الأخطاء.
- التكامل مع أدوات تحليل أخرى: التكامل مع requestsهو السيناريو الأكثر شيوعًا، حيث تقوم requests بجلب محتوى الويب الخام، وتقوم BeautifulSoup بتحويل هذا المحتوى إلى هيكل قابل للتحليل. يمكن أيضاً

دمجها مع مكتبات أخرى لمعالجة البيانات المستخرجة، مثل pandasلتنظيم البيانات في جداول أو re (التعبيرات النمطية) لإجراء عمليات بحث أكثر دقة داخل النصوص المستخرجة.

- دعم محركات تحليل متعددة: تدعم العديد من محركات تحليل نذكر منها:

1. html. parser:المحلل الافتراضي المدمج في بايثون. سهل الاستخدام ولكنه قد يكون أبطأ وأقل تسامحاً مع الأخطاء مقارنة بالبدائل .
2. lxml :محلل خارجي يعتمد على مكتبةC ، مما يجعله أسرع بكثير ويدعم ميزات إضافية مثل XPath (للاستعلام عن العناصر بطرق أكثر تعقيدا).إذا كنت تتعامل مع كميات كبيرة من البيانات أو تحتاج إلى أداء عالي، فإن lxmlهو الخيار المفضل.
3. html5lib:محلل يتبع معايير HTML5 بدقة، وهو الأكثر تسامحاً مع الأخطاء. قد يكون أبطأ من lxml ولكنه يضمن تحليل المستند بنفس الطريقة التي يعرضها بها المتصفح الحديث.

4.1.4 العيوب

- بطء الأداء: صحيح أن BeautifulSoup قد يكون أبطأ عند التعامل مع المستندات الكبيرة والمعقدة .إذا كان الأداء يمثل مشكلة حقيقية، ففكر في استخدام محلل lxmlالأسرع. أيضاً، حاول أن تكون محدداً قدر الإمكان في عمليات البحث عن العناصر لتقليل الوقت المستغرق في التحليل.
- وظائف محدودة للتفاعل مع صفحات الويب الديناميكية: هذه هي النقطة الأهم BeautifulSoup لا يمكنها تنفيذ JavaScript .عندما تواجه مواقع تعتمد على JavaScript لتحديث المحتوى، فإن البدائل الواضحة هي Seleniumو Playwright. هذه الأدوات تقوم بتشغيل متصفح حقيقي أو محاكي للمتصفح، مما يسمح بتنفيذ JavaScript ورؤية المحتوى الديناميكي بعد تحميله. يمكنك بعد ذلك استخدام BeautifulSoup لتحليل HTML الذي تم إنشاؤه ديناميكياً بواسطة هذه الأدوات. هذا يعني أنك قد تستخدم مزيجاً من الأدوات : Selenium/Playwrightلجلب الصفحة بعد تحميل المحتوى الديناميكي، ثم BeautifulSoup لتحليل هذا المحتوى المستقر.

2.4 مكتبة Selenium

1.2.4 التعريف

Seleniumهي إطار عمل شامل وقوي لأتمتة تفاعلات المستخدم مع تطبيقات الويب عبر مختلف المتصفحات والمنصات، حيث تنشئ جسرًا برمجيًا بين لغة البرمجة والمتصفح الفعلي للسماح بإرسال تعليمات تحاكي إجراءات المستخدم مثل التنقل، والتفاعل مع العناصر (النقر، وملء النماذج، والاختيار)، ومحاكاة حركات الفأرة، والتعامل مع النوافذ والإطارات، وتنفيذJavaScript ، والتقاط لقطات الشاشة، وإدارة ملفات تعريف الارتباط والتخزين المحلي؛ وتكمن أهميتها في Web Scraping في قدرتها على التعامل مع المواقع الديناميكية التي تعتمد على JavaScript لتحديث محتواها بعد التحميل الأولي، وذلك عن طريق تشغيل متصفح حقيقي ينتظر تنفيذ JavaScript وتحميل المحتوى الديناميكي بالكامل، مما يمكن من الوصول إلى هيكل DOM الحديث واستخلاص البيانات بدقة من خلال مكوناتها الأساسية التي تشمل مكتبات العميل بلغات البرمجة المختلفة، وواجهة WebDriver التي تحدد كيفية التواصل

مع المتصفحات، وبرامج تشغيل المتصفحات الخاصة بكل متصفح، وخادم Selenium الاختياري لتوزيع عمليات الأتمتة .

2.2.4 كيفية عمل مكتبةSelenium

Seleniumتقوم بمحاكاة متصفح حقيقي (مثل Google Chrome أوFirefox) ، وتنفذ أوامر كما لو أن إنسانًا يستخدم المتصفح.

خطوات العمل:

1. فتح متصفح حقيقي أو "headless" باستخدام WebDriver(مثل.ChromeDriver)
2. زيارة رابط الموقع باستخدام .(url)get.
3. الانتظار حتى يتم تحميل الصفحة وعناصرها.
4. التفاعل مع الصفحة مثل النقر، ملء نماذج، التمرير.. إلخ.
5. استخراج البيانات من خلال البحث عن عناصر DOM باستخدام ()find_element.أو XPath أو CSS Selectors.

3.2.4 المميزات

- القدرة على التعامل مع المحتوى الديناميكي: يمكن لـ Selenium تنفيذ JavaScript والتفاعل مع العناصر التي يتم تحميلها بعد تحميل الصفحة الأولية، مما يجعلها مناسبة لمواقع الويب الحديثة التي تعتمد على تقنيات مثلAJAX (التواصل مع الخادم وتحميل البيانات في الخلفية دون إعادة تحميل الصفحة كاملة).
- محاكاة تفاعل المستخدم: يمكنها محاكاة إجراءات المستخدم المختلفة مثل النقر على الأزرار، وملء النماذج، والتمرير، مما يفتح الباب لاستخلاص البيانات التي تتطلب تفاعلاً مع الصفحة.
- دعم متصفحات متعددة: تدعم العديد من متصفحات الويب الشائعة مثلChrome ، Firefox، Safari، و . Edge
- إمكانية التكامل مع مكتبات أخرى: يمكن دمجها مع مكتبات تحليل HTML مثل BeautifulSoup لتحليل المحتوى الذي تم تحميله بواسطةSelenium .

4.2.4 العيوب

- استهلاك عالي للموارد: يتطلب تشغيل متصفح حقيقي موارد نظام أكبر (مثل الذاكرة ووحدة المعالجة المركزية) مقارنة بمكتبات تحليل HTML الثابتة.
- بطء الأداء: عملية أتمتة المتصفح أبطأ بشكل عام من مجرد تحليل HTML ثابت باستخدام مكتبات مثل BeautifulSoupأو lxml.
- تعقيد الإعداد والتكوين: قد يكون إعداد وتكوين Selenium أكثر تعقيدًا من مكتبات "Scraping" الأخرى، خاصةً عند الحاجة إلى إدارة برامج تشغيل المتصفحات المختلفة(WebDriver) .

3.4 مكتبة Playwright

1.3.4 التعريف

Playwright التي طورتهاMicrosoft ، تمثل نقلة نوعية في مجال أتمتة متصفحات الويب وعمليات Web Scraping، حيث تقدم رؤية موحدة للتحكم في محركات التصفح الرئيسية الثلاثةChromium – ، Firefox، و WebKit –من خلال واجهة برمجة تطبيقات (API) سهلة الاستخدام؛ تهدف هذه المكتبة إلى تبسيط وتسريع عملية أتمتة المهام المختلفة داخل المتصفحات، سواء كانت لأغراض الاختبار أو استخلاص البيانات، وذلك من خلال توفير مجموعة من الميزات المبتكرة التي تعزز الكفاءة بشكل ملحوظ، مما يجعلها أداة قوية للتعامل مع سيناريوهات الويب الحديثة والمعقدة.

2.3.4 كيفية عمل مكتبةPlaywright

PlaywrightتشبهSelenium ، لكنها أكثر حداثة وكفاءة. تعمل أيضًا من خلال محاكاة متصفح حقيقي، لكنها توفر أدوات قوية للتعامل مع الصفحات الحديثة.

خطوات العمل:

1. تشغيل متصفح headless أو مرئي باستخدام محركPlaywright .
2. فتح تبويب جديد (يمكن التعامل مع عدة تبويبات في آنٍ واحد) .
3. زيارة الموقع المطلوب وانتظار تحميل الصفحة تلقائيًا.
4. التفاعل مع العناصر عبر واجهة حديثة تدعم الانتظار الذكي(Waits) .
5. استخراج البيانات من خلال البحث عن العناصر أو قراءة النصوص أو الصفات.

3.3.4 المميزات

- دعم متعدد للمتصفحات: توفر واجهة برمجة تطبيقات موحدة للتحكم في ثلاثة من محركات المتصفحات الرئيسية، مما يبسط عملية تطوير واختبار تطبيقات الويب وعمليات"Scraping" .
- أداء ممتاز: تم تصميم Playwright ليكون سريعًا وفعالًا في أداء عمليات الأتمتة و"Scraping".
- التعامل المتقدم مع المحتوى الديناميكي: توفر ميزات قوية للتعامل مع صفحات الويب الديناميكية، بما في ذلك القدرة على اعتراض طلبات الشبكة وتعديلها.
- دعم مدمج لعمليات الانتظار التلقائي: يمكن لـ Playwright الانتظار تلقائيًا حتى تصبح العناصر المطلوبة مرئية أو قابلة للتفاعل، مما يقلل من مشاكل التوقيت الشائعة في أتمتة المتصفحات.
- تسجيل وتوليد التعليمات البرمجية: توفر أداة لتسجيل تفاعلات المستخدم على المتصفح وتوليد التعليمات البرمجية المقابلة بلغات برمجة مختلفة (بما في ذلكPython) .

4.3.4 العيوب

- حداثة المكتبة: نظرًا لكونها مكتبة أحدث نسبيًا مقارنة بـSelenium ، قد يكون حجم مجتمع المستخدمين والموارد التعليمية المتاحة أقل.
- منحنى تعلم: على الرغم من قوتها، قد يكون لديها منحنى تعلم أكثر حدة قليلاً مقارنة بـ BeautifulSoup للمهام الأساسية.

5 تقنية LLM's وPrompt Engineering

1.5 تعريف(Large Language Model) LLM

يشير مصطلح LLM إلى "Large Language Model" أو النموذج اللغوي. في سياق الذكاء الاصطناعي، النموذج اللغوي هو نموذج حاسوبي تم تدريبه على كميات هائلة من النصوص لفهم وإنتاج اللغة الطبيعية. تعتمد هذه النماذج على تقنيات التعلم العميق، وخاصةً بنية المحولات العصبية(Transformer Architecture) ، التي أظهرت قدرة فائقة على التقاط العلاقات المعقدة في البيانات النصية.

بشكل أساسي، يمكن للنماذج اللغوية التنبؤ بالكلمة التالية في تسلسل نصي بناءً على الكلمات التي سبقتها. من خلال التدريب على مجموعات بيانات ضخمة ومتنوعة من النصوص (مثل الكتب، المقالات، مواقع الويب، وغيرها)، تتعلم هذه النماذج أنماط اللغة، وقواعد النحو، والمعاني الدلالية، وحتى بعض المعارف العامة.

2.5 لمحة تاريخية عن تطورLLMs

- **قبل 2017** :استخدام نماذج بسيطة مثل RNN و LSTMفي المهام اللغوية، مع قدرات محدودة.
- **2017** : ظهور ورقة "Attention is All You Need" وتقديم بنية Transformer من قبلGoogle ، التي شكلت ثورة في معالجة اللغة.
- **2018** : إصدار BERT منGoogle ، نموذج يعتمد على Encoder لفهم السياق بشكل مزدوج الاتجاه.
- **2019:** ظهور GPT-2 منOpenAI ، بنموذج توليدي قوي، وإثارة الجدل حول خطورته.
- **2020–2023**: إصدار GPT-3 ثم GPT-4 منOpenAI ، ونماذج منافسة مثل PaLM من Google و Claudeمن.Anthropic
- **2023–2024 :** تطور النماذج متعددة الوسائط مثل Gemini وGPT-4V، وتوسع استخدامها في البرمجة، الطب، التعليم، وغيرها.

3.5 أنواعLLM

تتنوع النماذج اللغوية الكبيرة (LLM) في بنيتها، وحجمها (عدد المعلمات التي تحتوي عليها)، وطريقة تدريبها، والأغراض التي صممت من أجلها. يمكن تصنيفها بناءً على عدة معايير، من أبرزها:

1.3.5 النماذج متعددة الوسائط (Multimodal LLMs)

قادرة على معالجة أنواع مختلفة من البيانات مثل النصوص، الصور، الصوت والفيديو. أمثلة GPT-4 :(مع تحليل الصور)، و Geminiمن Google .وتعد منصة Google AI Studio[1]مثالًا عمليًا على كيفية استخدام هذه النماذج متعددة الوسائط، حيث تتيح للمطورين التفاعل مع نماذج Gemini بطريقة سهلة وفعالة لتنفيذ مهام تتضمن نصوصًا وصورًا وبيانات مركبة.

2.3.5 النماذج التوليدية (Autoregressive Models)

تعتمد على التنبؤ بالكلمة التالية بناءً على السياق السابق، مثل سلسلة .GPT وهي بارعة في إنتاج نصوص متسلسلة وطبيعية.

3.3.5 النماذج المتعددة اللغات

مدرّبة على بيانات بلغات مختلفة، مما يجعلها قادرة على الفهم والترجمة بين عدة لغات بدقة مقبولة، مثل :mBERT وXLM-R.

4.3.5 النماذج القابلة للتكييف(Fine-tuned Models)

يمكن إعادة تدريبها على مهام محددة مثل التلخيص أو التصنيف. تُستخدم بشكل كبير في التطبيقات العملية المتخصصة.

5.3.5 النماذج مفتوحة المصدر والمغلقة

النماذج المفتوحة مثل LLaMA و Mistral توفر حرية التعديل والاستخدام، بينما المغلقة مثل ChatGPT وClaude تقدم أداءً عاليًا ولكن بقيود ترخيصية وتقنية.

4.5 التحديات التقنية فيLLMs

- **استهلاك الموارد**: تتطلب كميات ضخمة من البيانات والقدرة الحوسبية، مما يجعل تدريبها مكلفًا بيئيًا واقتصاديًا.
- **التحيّز والانحراف**: النماذج قد تتعلم تحيّزات موجودة في البيانات، مما يؤثر على عدالة ودقة المخرجات.
- **السياقات الطويلة**: لا تزال بعض النماذج تجد صعوبة في الاحتفاظ بمعلومات عبر سياقات طويلة.
- **التحديث المعرفي**: لا تُحدث معلوماتها بسهولة، فهي لا تتعلّم من التفاعلات إلا بإعادة التدريب أو التحديث اليدوي.
- **الغموض وعدم التفسير**: من الصعب أحيانًا تفسير لماذا يتخذ النموذج قرارًا معينًا، مما يعيق الاعتماد عليه في مجالات حرجة.

5.5 مستقبل النماذج اللغوية الكبيرة (LLMs)

يُتوقع أن يشهد مستقبل النماذج اللغوية الكبيرة (LLMs) تطورًا ملحوظًا على عدة مستويات. من أبرز الاتجاهات هو التركيز على نماذج أكثر كفاءة، بحيث تصبح أصغر حجمًا وأكثر فعالية، مما يُسهم في تقليل استهلاك الموارد وتسريع الأداء، مع الحفاظ على جودة المخرجات. كما يتجه البحث نحو دمج الذكاء المتعدد الوسائط، حيث تُطوَّر النماذج لتفهم

وتعالج النصوص، الصوت، الصور، الفيديو، وحتى الرموز البرمجية ضمن نموذج موحد. إضافة إلى ذلك، يبرز التخصيص الفردي كأحد الأهداف المستقبلية، بحيث تصبح النماذج قادرة على التفاعل مع المستخدمين وفقًا لتاريخهم وسياقهم الشخصي، مما يعزز من فاعلية التجربة. في المقابل، فإن الانتشار الواسع لهذه النماذج يدفع نحو ضرورة وجود تنظيم قانوني وتشريعي دقيق، لضمان استخدامها بشكل مسؤول والحد من إساءة توظيفها. أخيرًا، من المتوقع أن تُعتمد LLMsبشكل متزايد في المهام الحساسة مثل الطب، القانون، التعليم الرسمي، وصنع القرار، مما يفرض متطلبات أعلى على دقتها وموثوقيتها.

6.5 تعريف Prompt Engineering

Prompt Engineering(هندسة الموجهات أو هندسة المطالبات) هي عملية تصميم وتطوير "الموجهات" أو "المطالبات " (Prompts) الفعالة التي تُقدم للنماذج اللغوية الكبيرة (LLM) لتوجيهها نحو إنتاج الاستجابات المطلوبة بدقة وجودة عالية. "الموجه" هو ببساطة النص المدخل الذي يُقدم للنموذج اللغوي، ويمكن أن يتضمن تعليمات، أسئلة، سياق، أو أمثلة.

تعتبر "Prompt Engineering" مهارة حاسمة في التفاعل معLLM's ، حيث أن جودة الموجه وتصميمه تلعب دورًا كبيرًا في تحديد جودة الاجابة التي سيولدها النموذج. يمكن للموجه المصمم جيدًا أن يستخلص من النموذج إجابات دقيقة، ومبتكرة، ومتوافقة مع السياق المطلوب.

7.5 كيف يساعد Prompt Engineering في تحسين استخدام النماذج اللغوية

تساهم "Prompt Engineering" في تحسين استخدام النماذج اللغوية بعدة طرق، منها:

- **تحسين دقة الاستجابات :**من خلال توفير سياق واضح وتعليمات محددة في الموجه، يمكن توجيه النموذج لتقديم إجابات أكثر دقة وملاءمة للسؤال أو المهمة المطلوبة.
- **توجيه النماذج نحو أنماط إخراج محددة :**يمكن استخدام الموجهات لتحديد النبرة، والأسلوب، والتنسيق المطلوب للاستجابة (مثل طلب كتابة نص بأسلوب رسمي، أو تقديم الإجابة في شكل قائمة نقطية).
- **استخلاص معلومات محددة :**يمكن تصميم الموجهات بطريقة تستهدف استخلاص معلومات معينة من النص أو توليد محتوى يركز على جوانب محددة.
- **تقليل التحيزات :**يمكن أن تساعد الموجهات المصممة بعناية في تقليل احتمالية توليد النماذج لاستجابات متحيزة أو غير مرغوب فيها.
- **تمكين القدرات المتقدمة للنماذج :**يمكن استخدام تقنيات "Prompt Engineering" المتقدمة (مثل استخدام "سلاسل الأفكار" أو "التفكير خطوة بخطوة") لتشجيع النماذج على إجراء عمليات تفكير أكثر تعقيدًا والوصول إلى حلول أفضل للمشكلات.
- **توفير الوقت والجهد :**من خلال الحصول على استجابات عالية الجودة من النماذج في المرة الأولى، يمكن تقليل الحاجة إلى إعادة صياغة الأسئلة أو إجراء تعديلات يدوية على المخرجات.

باختصار، "Prompt Engineering"هي فن وعلم صياغة المدخلات التي توجه النماذج اللغوية الكبيرة لتحقيق أفضل أداء ممكن في مجموعة متنوعة من المهام اللغوية. مع استمرار تطور هذه النماذج، ستزداد أهمية فهم وتطبيق مبادئ "Prompt Engineering" لتحقيق أقصى استفادة من قدراتها.

1.5 تقنية Prompt Engineeringلتوجيه LLMs في مهامScraping

تُعد هندسة المطالبات (Prompt Engineering) تقنية حيوية لتسخير قوة النماذج اللغوية الكبيرة LLMs في سياق استخلاص البيانات من الويب، حيث تهدف إلى تصميم مدخلات نصية دقيقة ومفصلة توجه النموذج اللغوي لأداء مهام استخراج معلومات محددة بكفاءة وفعالية. بدلاً من الاعتماد على طرق تقليدية تعتمد على تحليل HTML/XML أو أتمتة المتصفحات، يمكن للمطالبات المصممة بعناية أن تستفيد من الفهم الدلالي والقدرة على توليد النصوص لدى LLMs لاستخلاص البيانات المطلوبة حتى من الصفحات ذات الهياكل غير المنتظمة أو التي تحتوي على نصوص معقدة. تتضمن هذه العملية صياغة تعليمات واضحة للنموذج تحدد بدقة المعلومات المراد استخلاصها، وتحديد سياق الصفحة أو جزء منها ذي الصلة، وتوفير أمثلة أو تنسيقات للإخراج المطلوب. على سبيل المثال، يمكن توجيه النموذج لاستخراج أسماء المنتجات وأسعارها من وصف منتج نصي، أو لتحديد العناوين الرئيسية والملخصات من مقال إخباري، أو حتى لتحويل جداول بيانات موجودة في نص الصفحة إلى تنسيق منظم مثل JSON أو .CSV تتطلب هندسة المطالبات الفعالة فهمًا لقدرات وقيود النموذج اللغوي المستخدم، بالإضافة إلى معرفة جيدة ببنية ومحتوى صفحات الويب المستهدفة. من خلال التجربة والتكرار، يمكن للمهندسين تحسين المطالبات لزيادة دقة الاستخراج، وتقليل الضوضاء. علاوة على ذلك، يمكن استخدام تقنيات متقدمة في هندسة المطالبات مثل التضمين (few-shot learning) حيث يتم تزويد النموذج بأمثلة قليلة للمهمة المطلوبة داخل المطالبة نفسها، مما يساعده على فهم المهمة بشكل أفضل وتحسين أدائه في استخلاص البيانات.

6 تطبيق موبايل باستخدامKotlin

1.6 تعريف بلغةKotlin

Kotlinهي لغة برمجة حديثة متعددة المنصات، تم تطويرها بواسطة شركة(JetBrains المعروفة بتطوير بيئة التطوير المتكاملة IntelliJ IDEA) تم تصميمها لتكون لغة عملية، موجزة، وآمنة للأنواع، وقابلة للتشغيل التوافقي مع جافا .منذ عام 2019، أعلنت جوجل ان هذه اللغة هي المفضلة لتطوير تطبيقات أندرويد، مما ساهم في انتشارها وشعبيتها بشكل كبير في مجتمع تطوير تطبيقات الموبايل.

2.6 مميزتها

تتميز Kotlin بالعديد من الميزات التي تجعلها خيارًا جذابًا للمطورين، بما في ذلك:

- **التوافقية مع جافا:** يمكنها العمل بسلاسة مع كود جافا الموجود ومكتباتها، مما يسهل عملية الانتقال للمطورين الذين لديهم خبرة في جافا.
- **الإيجاز والوضوح:** تتميز ببنية لغوية موجزة تقلل من كمية التعليمات البرمجية المطلوبة لإنجاز مهمة معينة، مما يجعل الكود أكثر قابلية للقراءة والصيانة.

- **السلامة من الأخطاء الصفرية (Null Safety) :**تعالج مشكلة الأخطاء الناتجة عن القيم الفارغة (NullPointerException)بشكل مباشر من خلال نظام الأنواع الخاص بها، مما يؤدي إلى تطبيقات أكثر استقرارًا وموثوقية.
- **الدعم للبرمجة الوظيفية والشيئية:** تدعم أنماط البرمجة الوظيفية والشيئية، مما يوفر للمطورين مرونة في اختيار الأسلوب الأنسب لحل مشكلاتهم.
- **الدعم القوي من جوجل:** بصفتها اللغة المفضلة لتطوير أندرويد، تحظى بدعم قوي من جوجل من حيث الأدوات والمكتبات والتحديثات.
- **دعم تعدد العمليات المتزامنة (Coroutines) :**توفر آلية سهلة الاستخدام لإدارة العمليات المتزامنة وغير المتزامنة بكفاءة، مما يحسن أداء التطبيقات التي تتطلب تنفيذ مهام في الخلفية.

3.6 كيفية استخدامها في تطوير تطبيقات موبايل

تُعد Kotlinمن أبرز اللغات المستخدمة حاليًا في تطوير تطبيقات نظام التشغيل أندرويد، حيث توفّر بيئة تطوير حديثة مدعومة بمجموعة واسعة من الأدوات والمكتبات التي تُسهّل عملية بناء تطبيقات مبتكرة وعالية الكفاءة .وتتميز Kotlin بتكاملها السلس مع منظومة أندرويد، مما يتيح استخدامها في جميع جوانب تطوير التطبيقات، بما في ذلك:

- **واجهات المستخدم (User Interface - UI) :** يمكن استخدامها لكتابة منطق واجهات المستخدم باستخدام نظام View التقليدي أو باستخدامJetpack Compose ، وهو إطار عمل حديث لبناء واجهات مستخدم تصريحية (Declarative UI) في أندرويد.
- **لتعامل مع البيانات :**تُستخدم Kotlinللتفاعل مع قواعد البيانات المحلية، مثل SQLite، من خلال الاستعانة بمكتبات متقدمة كـ Room Persistence Library، والتي توفر طبقة تجريدية تُسهّل عملية الوصول إلى البيانات وتُعزز من الأمان والكفاءة .كما تُستخدم Kotlinللتكامل مع الخدمات الخلفية (Backend Services)عبر مكتبات مثل Retrofitو Ktor، والتي تتيح إجراء الاتصالات الشبكية (HTTP/REST) بطريقة مرنة وآمنة، مما يسهّل تبادل البيانات مع الخوادم وإدارة الطلبات والاستجابات بشكل فعّال.
- **العمليات في الخلفية:** يمكن استخدام Coroutines في Kotlin لتنفيذ مهام طويلة الأمد في الخلفية دون حظر مؤشر الترابط الرئيسي(Main Thread) ، مما يحافظ على استجابة التطبيق.
- **الخدمات والمكونات الأخرى:** يمكن استخدام Kotlin لتطوير الخدمات(Services) ، ومستقبِلات البث (Broadcast Receivers)، وموفري المحتوى(Content Providers) ، وغيرها من مكونات تطبيقات أندرويد.
- **الاختبار:** تدعم Kotlin كتابة اختبارات الوحدات (Unit Tests) واختبارات التكامل (Integration Tests)لضمان جودة التطبيق واستقراره.

7 خاتمـــــــــــــة

استهلّينا هذا الفصل بدراسة تحليلية للواقع الميداني، شملت بيئة العمل الحالية وسيرورة نشر إعلانات المناقصات في الجرائد الورقية، مع ما يرافقها من تحديات تتعلّق بصعوبة الوصول، وضعف التنظيم، وتأخر التحديث. كما تم استعراض الحلول والتطبيقات الموجودة، وتحليل نقاط قوتها وضعفها، بما أتاح تحديد احتياجات المجتمع المستهدف وتوجيه مشروعنا نحو تلبية هذه المتطلبات الفعلية.

بعد ذلك، تناولنا المفاهيم والتقنيات المعتمدة في الحلول المقترح، بدءًا من تقنيات Web Scrapingوأدوات Python، مرورًا بدور النماذج اللغوية الكبيرة LLMs و Prompt Engineering في استخراج البيانات من الوثائق غير المنظمة، بما يعزّز دقة وكفاءة النظام.

يمهّد هذا الإطار النظري والتحليلي للانتقال إلى مراحل النمذجة والتنفيذ، على أساس رؤية واضحة ومعطيات واقعية.

الفصل الثاني: نمذجة وتصميم

1 مقدمـــــــــــــة

يُعد توظيف لغة النمذجة الموحدة (UML) في هذا الفصل خطوة محورية توثّق تطوّر المشروع، من خلال إدراج المعطيات الجديدة المستخلصة من الدراسة التحليلية الدقيقة والمحاكاة الواقعية لبيئة النظام .وتُساهم هذه العملية في فهم طبيعة النظام ومتطلباته بشكل شامل ومنهجي، وذلك عبر تقديم مجموعة من المخططات التي تُستخدم في نمذجة وتصميم الأنظمة البرمجية، بما يتيح تمثيلًا بصريًا دقيقًا للجوانب الوظيفية وغير الوظيفية للنظام المستهدف.

2 تعريف لغةUML

هي لغة نمذجة معيارية تُستخدم لتصور وتصميم وتوثيق نظم البرمجيات. كما توفر مجموعة من الرسومات البيانية التي تساعد في توضيح هيكل النظام، سلوك النظام، والتفاعلات بين مكونات النظام. تُستخدم بشكل واسع في هندسة البرمجيات لتسهيل فهم وتطوير نظم البرمجيات المعقدة.

لذلك سنستعين بها في هذا الفصل من خلال التركيز على أربع مخططات رئيسية تُعد أساسية في توصيف الجوانب المختلفة للنظام، وهي:(مخطط الحالة - مخطط النشاط - مخطط التسلسل - مخطط الفئات).[3]

3 تحديد الأطراف الفاعلة في النظام

يوضح الجدول التالي توزيع المهام الرئيسية المرتبطة بمنظومة معالجة الإعلانات المستخرجة من الجرائد، ويبيّن الأدوار المختلفة التي يؤديها كل طرف فاعل ضمن هذه المنظومة. تتنوع المهام بين إجراءات آلية لتنفيذ عمليات جمع البيانات ومعالجتها، وأخرى يدوية تتعلق بمراقبة الجودة وتقديم تجربة استخدام فعّالة للمستخدم النهائي. تم تصنيف الأدوار إلى ثلاث فئات رئيسية: المسؤول(Admin) ، المستخدم(User) ، والعمليات الآلية المجدولة(Cron Job) ، بحيث يساهم كل طرف في مراحل مختلفة من دورة حياة البيانات لضمان دقة المخرجات وكفاءتها.

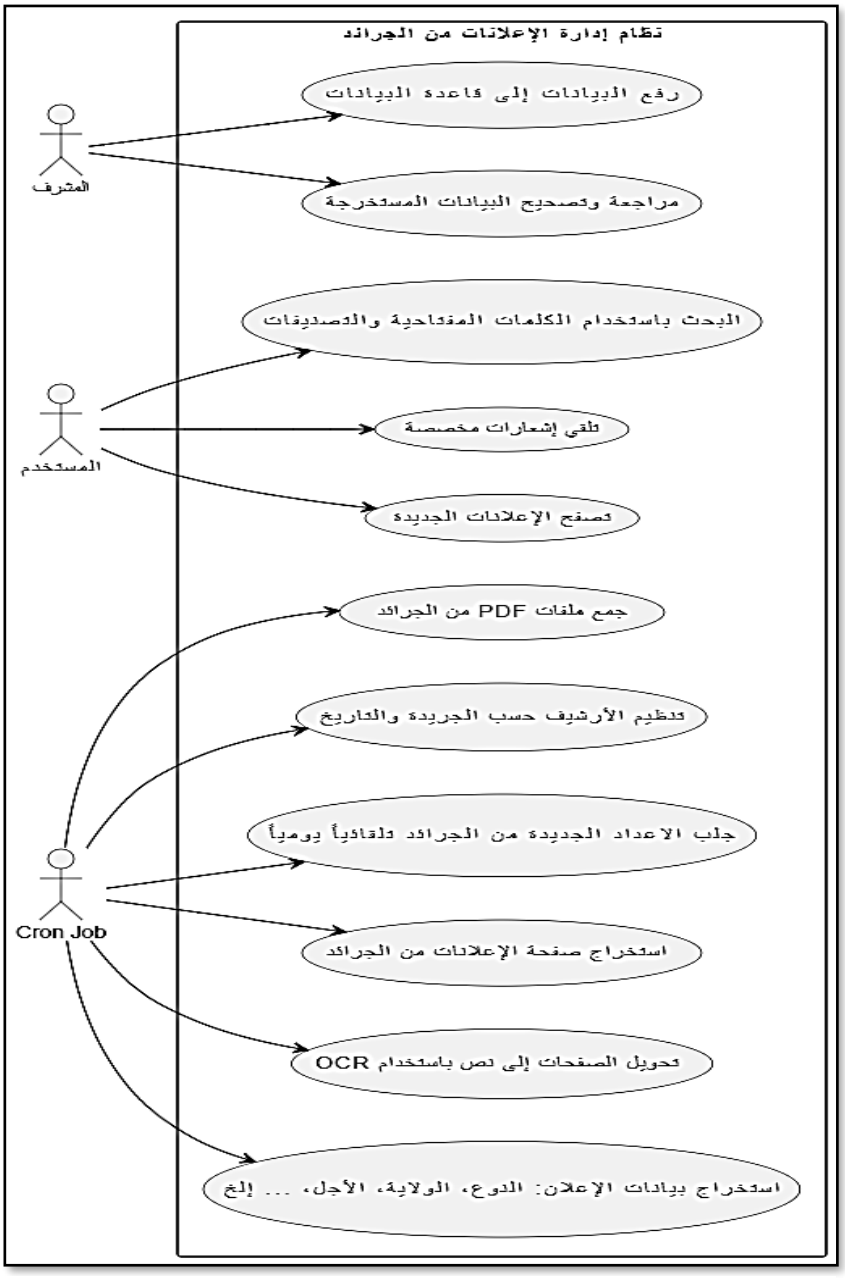
الطرف الفاعل	الدور	المهام الرئيسية
المسؤول (Admin)	مراقبة جودة البيانات والتحكم بالمحتوى النهائي	-مراجعة البيانات المستخرجة - تصحيح الأخطاء أو حذف الإعلانات - تعديل التصنيفات - تأكيد البيانات - رفع البيانات إلى قاعدةSupabase
المستخدم (User)	الاستفادة من النظام عبر التطبيق	-تصفح الإعلانات الجديدة - البحث باستخدام كلمات مفتاحية، ولايات، تواريخ، أنواع - حفظ الإعلانات المفضلة - تلقي إشعارات مخصصة حسب الاهتمامات
Cron Job	تنفيذ المهام الآلية لتجميع ومعالجة البيانات	-جمع ملفات PDF من الجرائد - تنظيم الأرشيف حسب الجريدة والتاريخ - تحديث الجرائد يوميًا - استخراج صفحات الإعلانات - تحويل الصفحات إلى نصوص (OCR) -استخراج بيانات الإعلانات باستخدام الذكاء الاصطناعي

جدول 3 تحديد الاطراف الفاعلة في النظام

4 مخطط حالة الاستخدام

يُركّز في لغة UML على تمثيل سلوك النظام من خلال تصور الحالات المختلفة التي يمكن أن تمر بها الكائنات أو الكيانات ضمن النظام، استجابةً للمحفزات الخارجية أو الأحداث الداخلية .ويُستخدم هذا المخطط لتوضيح كيفية انتقال النظام بين الحالات المختلفة، مما يساهم في فهم أعمق لطبيعة التفاعلات المحتملة مع المستخدمين أو مع العناصر الخارجية الأخرى.

تُبنى مقدمة مخطط الحالة على تحليل دقيق لسلوك النظام، حيث يتم تحديد الحالات الأساسية والتحولات بينها، بناءً على الأحداث أو الشروط التي تؤدي إلى تغيّر في حالة الكائن .وبذلك، يُعد مخطط الحالة أداة فعالة في تصميم الأنظمة التي تتضمن منطقًا معقدًا للسلوك أو تعتمد على تسلسل محدد للاستجابات والانتقالات.



الشكل 1 مخطط حالة الاستخدام

5 مخططات النشاط

يُعتبر مخطط النشاط أحد المخططات الأساسية في لغةUML ، ويُستخدم لتصميم وتوضيح تدفق سير العمليات أو الأنشطة داخل النظام بطريقة مرئية ومنظمة. يركز هذا المخطط على عرض تسلسل الخطوات المنطقية التي يمر بها النظام أو أحد مكوناته لتنفيذ وظيفة معينة، سواء كانت وظيفة بسيطة أو عملية معقدة تتضمن اتخاذ قرارات وتفرعات.

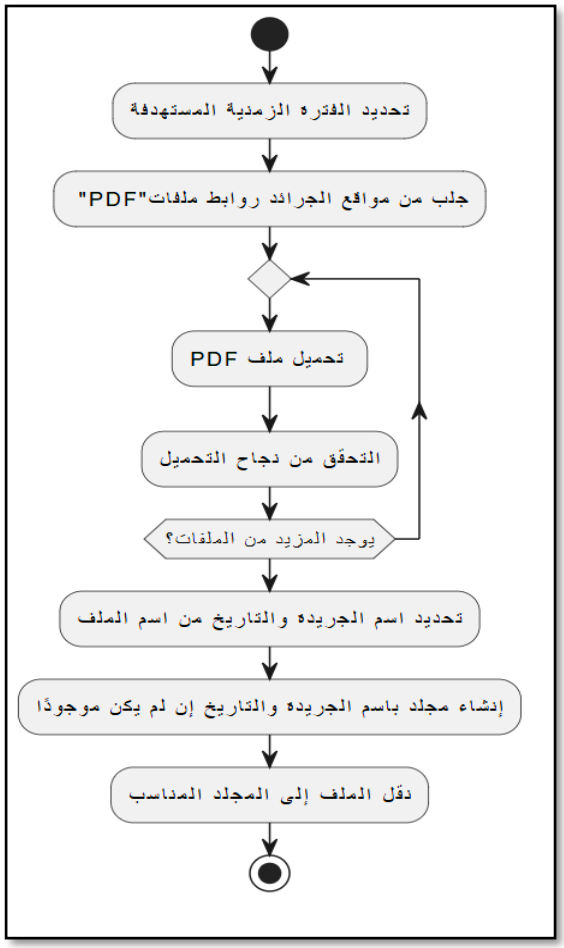
يساعد مخطط النشاط على:

- فهم تسلسل الإجراءات التي تحدث داخل النظام.
- تحديد نقاط التوازي أو التفرع أو الاتخاذ القرارات في سير العمل.
- تحليل تدفق البيانات أو المهام بين المستخدم والنظام أو بين مكونات النظام نفسه.

وغالبًا ما يُستخدم هذا المخطط في مرحلة تحليل النظام لتوثيق العمليات الوظيفية بدقة، كما يُفيد في توصيل الفكرة للمبرمجين والمصممين بوضوح قبل مرحلة التنفيذ.

1.5 مخطط النشاط :جمع وتنظيم ملفات PDF من الجرائد

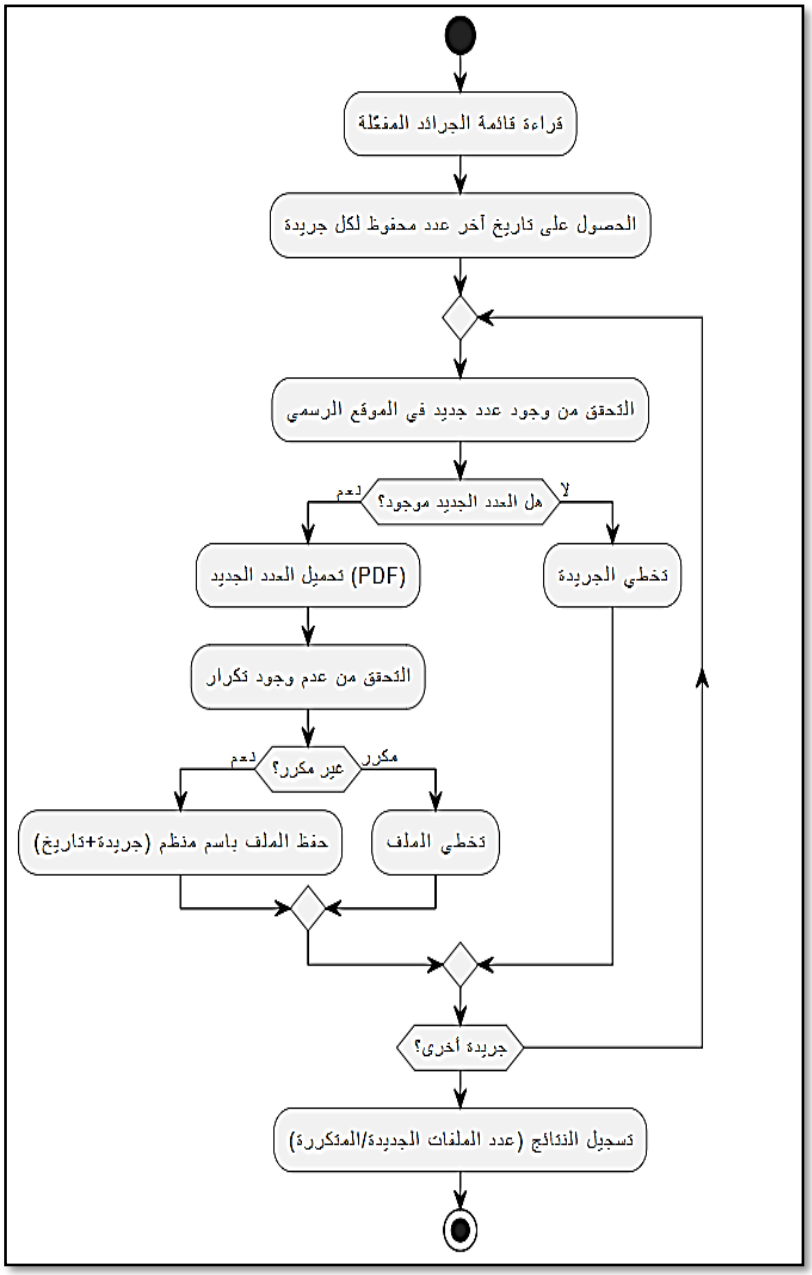
نقوم بتحديد الفترة المستهدفة وبعدها يتم جمع روابط ملفاتPDF ثم يحمّل كل ملف، ليتم نقله إلى مجلد منظم حسب الجريدة والتاريخ.



الشكل 2 مخطط نشاط جمع و تنظيم ملفات PDF من الجرائد

2.5 مخطط النشاط :التحديث التلقائي للجرائد

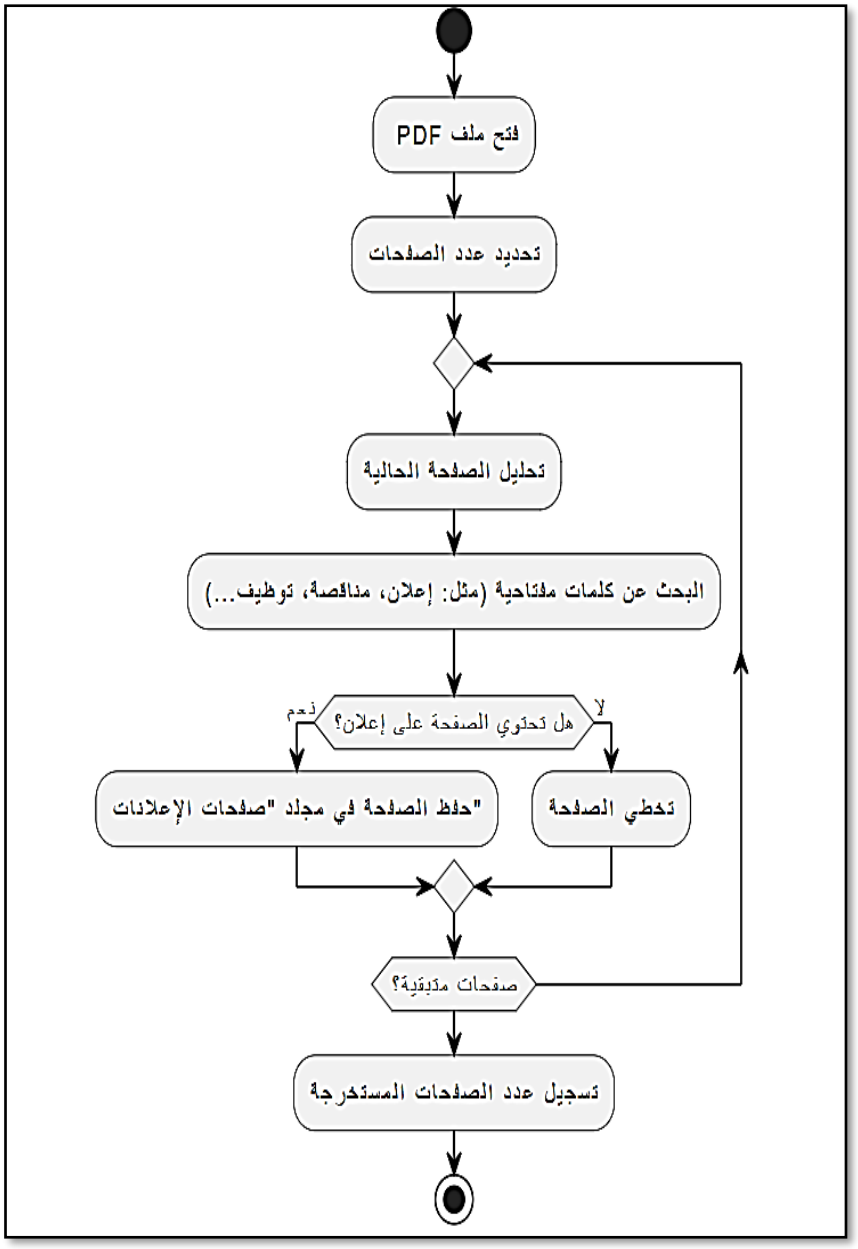
هذا المخطط يغطي عملية التحديث التلقائي للجرائد من خلال تنفيذ مهمة دورية بواسطة Cron Job، حيث يبدأ أولًا بقراءة قائمة الجرائد النشطة في النظام، ثم يستخرج تاريخ آخر عدد محفوظ لكل جريدة. بعد ذلك، يقوم بالتحقق من وجود أعداد جديدة في المواقع الرسمية لتلك الجرائد. في حال توفر عدد جديد، يتم تحميله على شكل ملفPDF ، مع إجراء تحقق من عدم تكراره في الأرشيف الحالي. إذا لم يكن الملف مكررًا، يتم حفظه باسم منظم يحتوي على اسم الجريدة وتاريخ العدد، وإلا يتم تجاهله. تُكرر هذه العملية مع باقي الجرائد في القائمة، وفي النهاية تُسجّل نتائج العملية مثل عدد الملفات الجديدة التي تم حفظها وعدد الملفات المتكررة التي تم تخطيها.



الشكل 3 مخطط لنشاط التحديث التلقائي للجرائد

3.5 مخطط النشاط :استخراج صفحات الإعلانات من ملفاتPDF

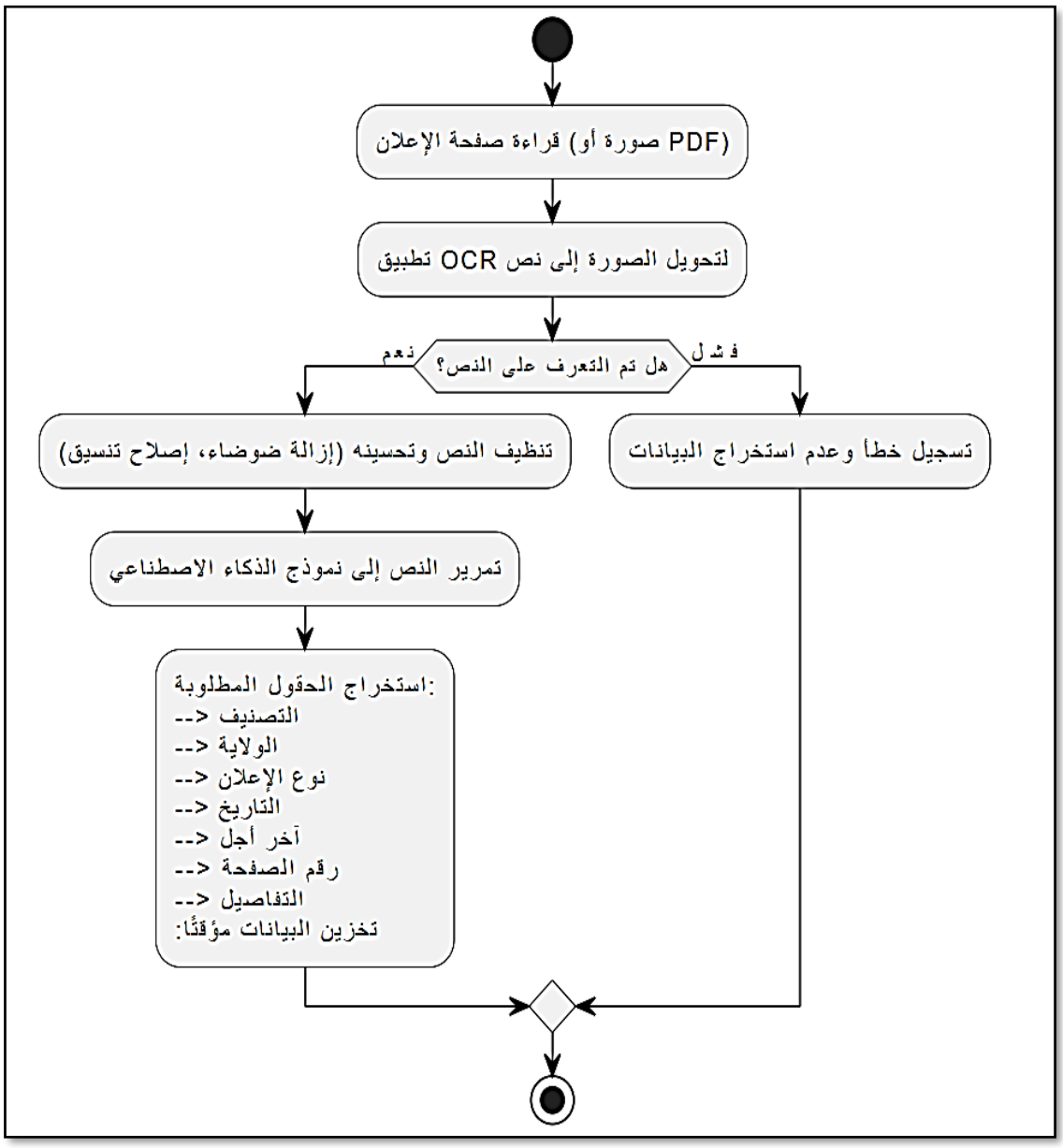
يمثل هذا المخطط خطوات تحديد صفحات الإعلانات داخل كل ملف PDF يتم تحميله من الجرائد. تبدأ العملية بفتح الملف وتحليل كل صفحة على حدة. في كل صفحة، يتم البحث عن كلمات مفتاحية أو أنماط تنسيق تدل على وجود إعلان. إذا تم التعرف على محتوى إعلاني، تُحفظ الصفحة تلقائيًا في مجلد مخصص، أما إذا لم يُكتشف إعلان، فيتم تجاوز الصفحة. تستمر العملية حتى يتم تحليل كل الصفحات، ثم تُسجّل إحصائية بعدد الصفحات المستخرجة.



الشكل 4 مخطط النشاط لاستخراج صفحات الإعلانات من ملفات PDF

4.5 مخطط النشاط :استخراج البيانات من صفحات الإعلانات

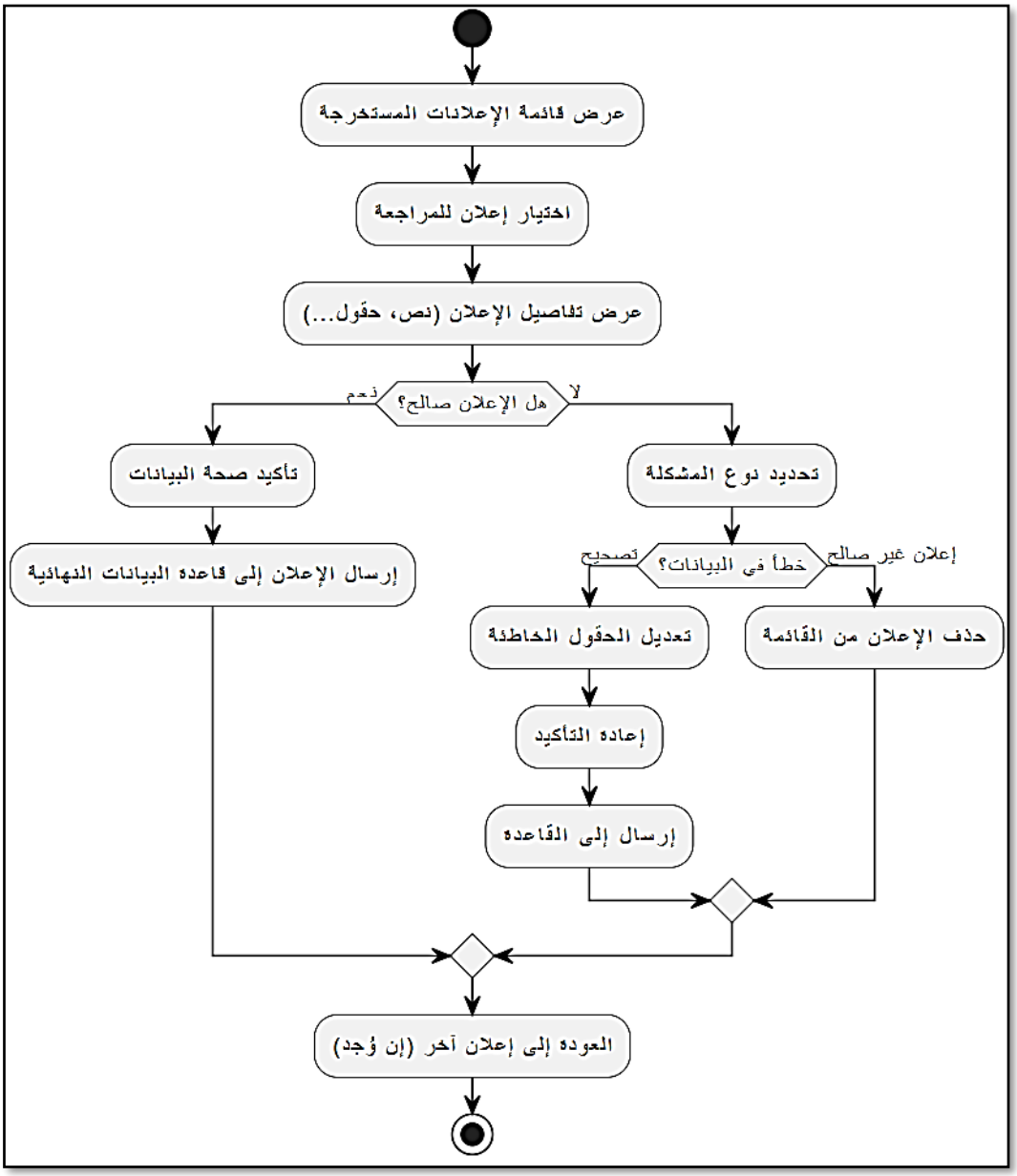
يغطي هذا المخطط عملية استخراج البيانات من صفحات الإعلانات عبر سلسلة من الخطوات التقنية تبدأ بتحويل الصفحة إلى نص باستخدام تقنية OCR. بعد التأكد من الحصول على نص واضح، يُنظّف ويُمرر إلى نموذج ذكاء اصطناعي (Google AI Studio) لاستخلاص الحقول الأساسية مثل نوع الإعلان، الولاية، التصنيف، التاريخ، آخر أجل، ورقم الصفحة. تُخزن هذه البيانات مؤقتًا بانتظار مراجعة المشرف. وإذا فشل التعرف على النص، يتم تسجيل الخطأ لتفادي فقدان البيانات دون علم النظام.



الشكل 5 مخطط النشاط لاستخراج البيانات من صفحات الإعلانات

5.5 مخطط النشاط :مراجعة المسؤول للبيانات

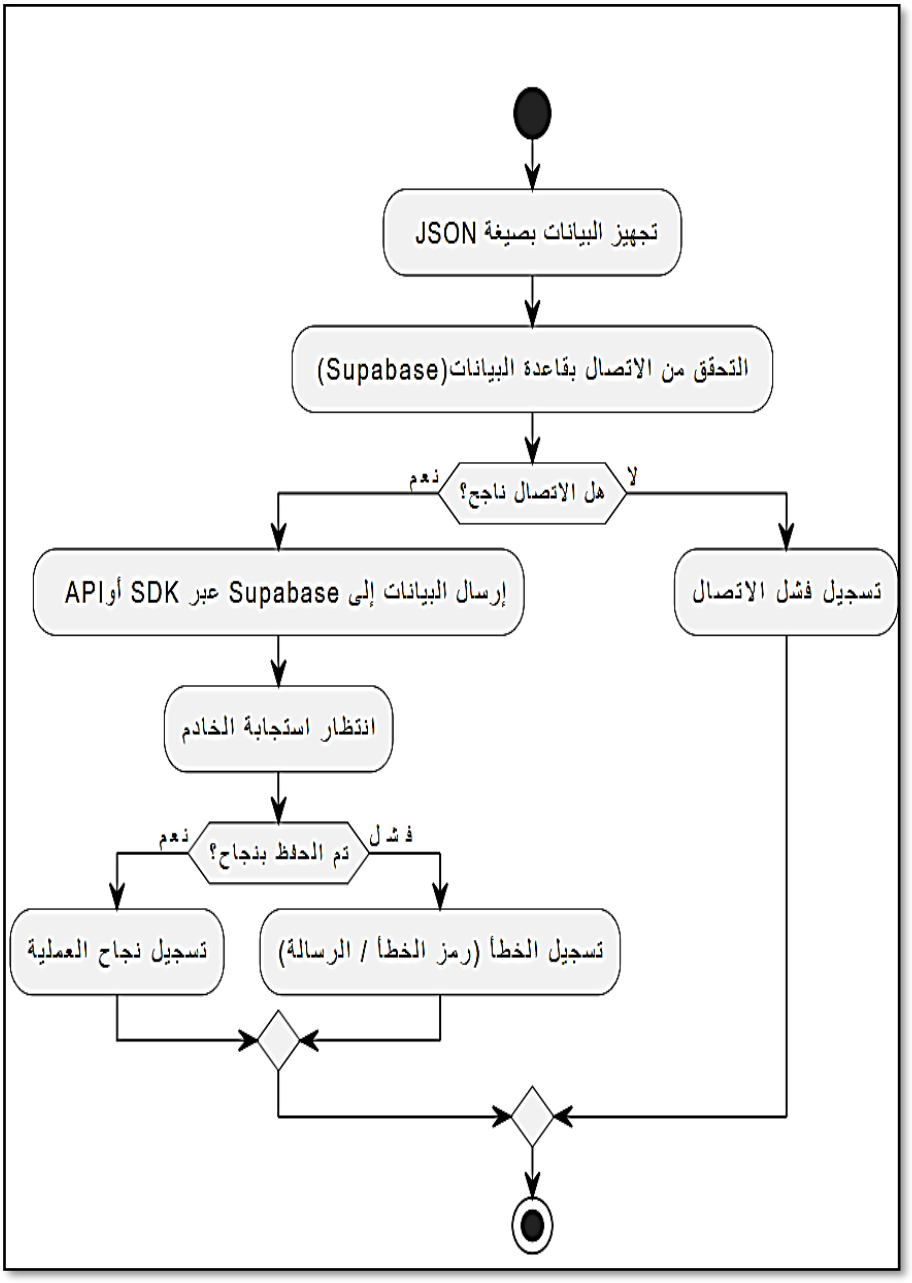
يغطي هذا المخطط عملية المراجعة اليدوية التي يقوم بها المشرف للتأكد من دقة بيانات الإعلانات المستخرجة. تبدأ العملية بعرض قائمة الإعلانات غير المؤكدة، ثم يختار المشرف إعلانًا لمراجعته ويطلع على تفاصيله. إذا كانت البيانات صحيحة، يتم تأكيدها وإرسالها مباشرة إلى قاعدة البيانات النهائية. أما إذا كان الإعلان يحتوي على أخطاء، يمكن للمشرف إما تعديل البيانات الخاطئة وإعادة تأكيدها، أو حذف الإعلان بالكامل إذا كان غير صالح للنشر. تستمر العملية حتى يتم مراجعة جميع الإعلانات في القائمة.



الشكل 6 مخطط النشاط لمراجعة المسؤول للبيانات

6.5 مخطط النشاط :رفع البيانات إلى قاعدة البيانات

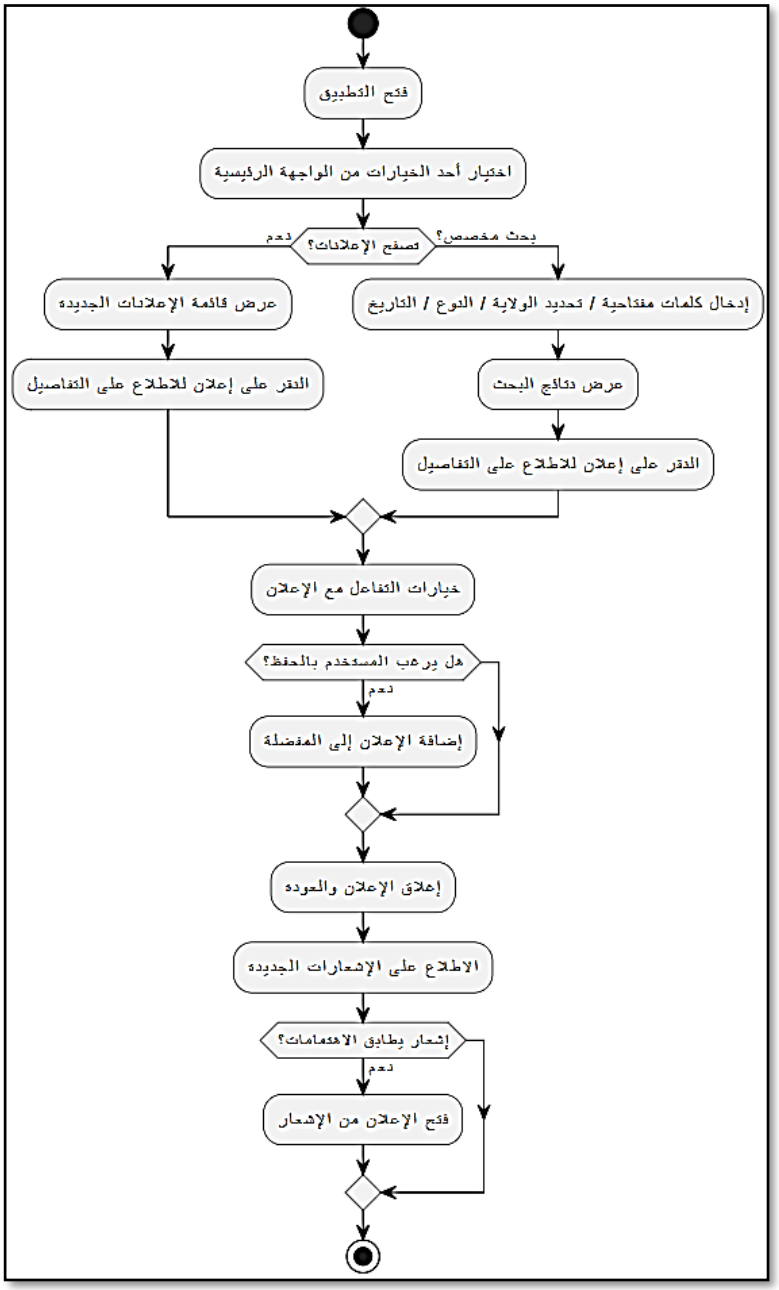
يصف هذا المخطط خطوات رفع البيانات المؤكدة إلى قاعدة البيانات النهائية باستخدام Supabase. تبدأ العملية بتجهيز البيانات في صيغة JSON مرتبة حسب الحقول المطلوبة. ثم يتم التحقق من وجود اتصال نشط مع قاعدة Supabase. إذا كان الاتصال ناجحًا، تُرسل البيانات باستخدام API أو SDK ويتم انتظار استجابة الخادم. إذا كانت الاستجابة إيجابية، تُسجّل العملية كناجحة، أما في حالة الخطأ يتم تسجيل تفاصيل الخطأ (مثل رمز الخطأ أو الرسالة). أما إذا فشل الاتصال من البداية، فيُسجّل الفشل دون محاولة الإرسال.



الشكل 7 مخطط النشاط لرفع البيانات إلى قاعدة البيانات

7.5 مخطط النشاط :تفاعل المستخدم في التطبيق

يمثل هذا المخطط رحلة المستخدم داخل تطبيق الهاتف، حيث يبدأ بتشغيل التطبيق والوصول إلى الواجهة الرئيسية، ثم يختار بين تصفح الإعلانات الجديدة أو تنفيذ بحث مخصص باستخدام معايير مثل الكلمات المفتاحية، الولاية، النوع أو التاريخ. بعد عرض النتائج، يمكنه استعراض تفاصيل أي إعلان، واتخاذ إجراءات مثل حفظه في قائمة المفضلة. كما يطّلع المستخدم على الإشعارات التي تصله بناءً على اهتماماته، ويمكنه فتح الإعلانات مباشرة من خلالها. تهدف هذه الأنشطة إلى تسهيل الوصول السريع والدقيق إلى الإعلانات ذات الصلة.



الشكل 8 مخطط النشاط لتفاعل المستخدم في التطبيق

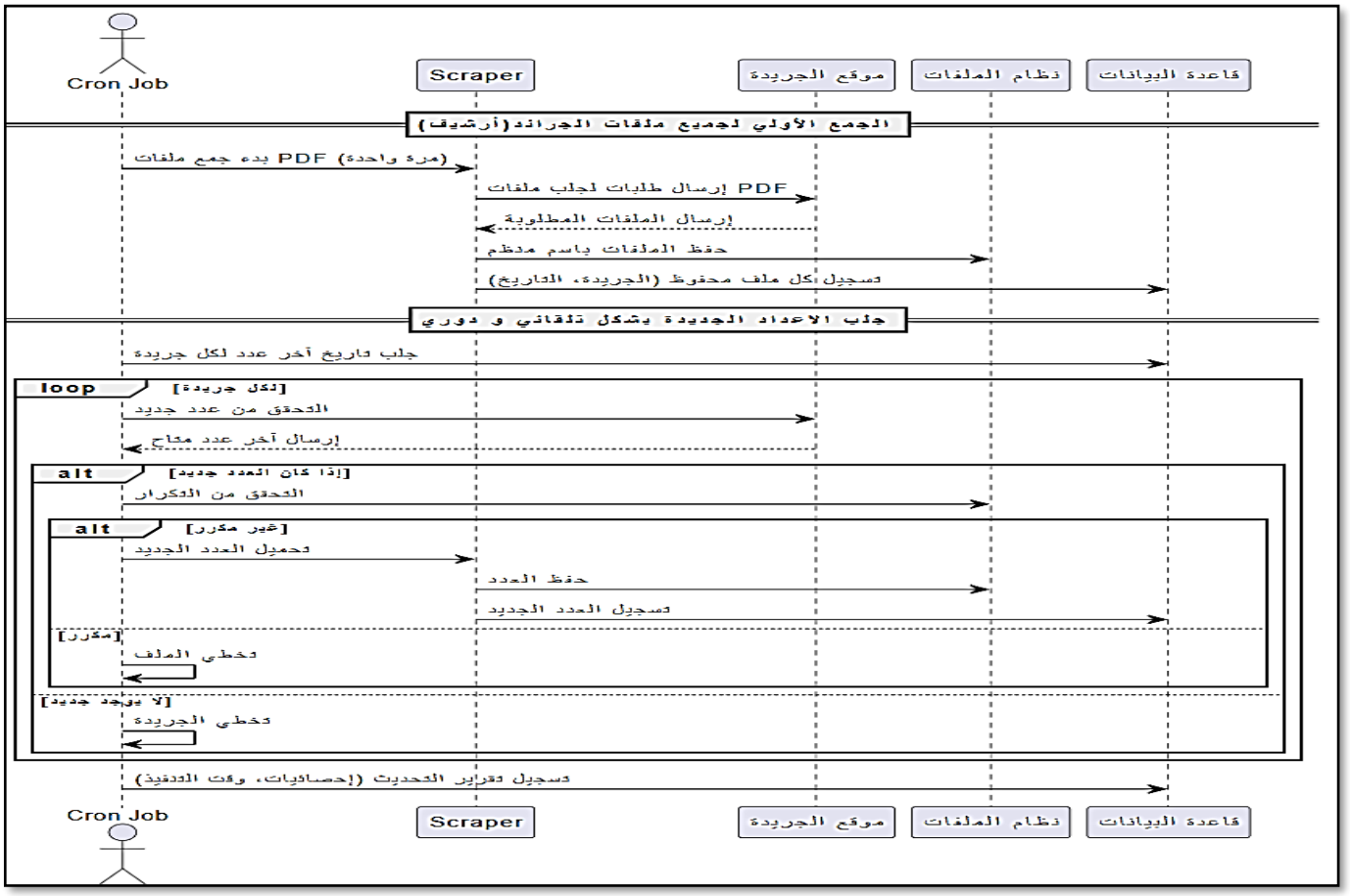
6 مخطط التسلسل

مخطط التسلسل هو نوع من مخططات UML يُستخدم لعرض التفاعل بين الكائنات (Objects) أو (Actors) في النظام بترتيب زمني .يركز على ترتيب الرسائل المتبادلة بين الكيانات المختلفة لتحقيق سيناريو معيّن من الاستخدام، ويُستخدم لتوثيق سلوك النظام بطريقة واضحة.

1.6 مخطط التسلسل لجمع الملفات والتحديث التلقائي

يمثل هذا المخطط تسلسل العمليات في مرحلة جمع ملفات PDF من الجرائد لأول مرة، بالإضافة إلى التحديث التلقائي اليومي أو الأسبوعي لاحقًا. يبدأ التنفيذ بواسطة مهمة Cron Jobالتي تطلق سكربت الجمع الأولي، حيث يتم التواصل مع مواقع الجرائد للحصول على الملفات ضمن نطاق زمني محدد. تُحفظ هذه الملفات في نظام الملفات بشكل منظم حسب الجريدة والتاريخ، ويتم تسجيلها في قاعدة البيانات.

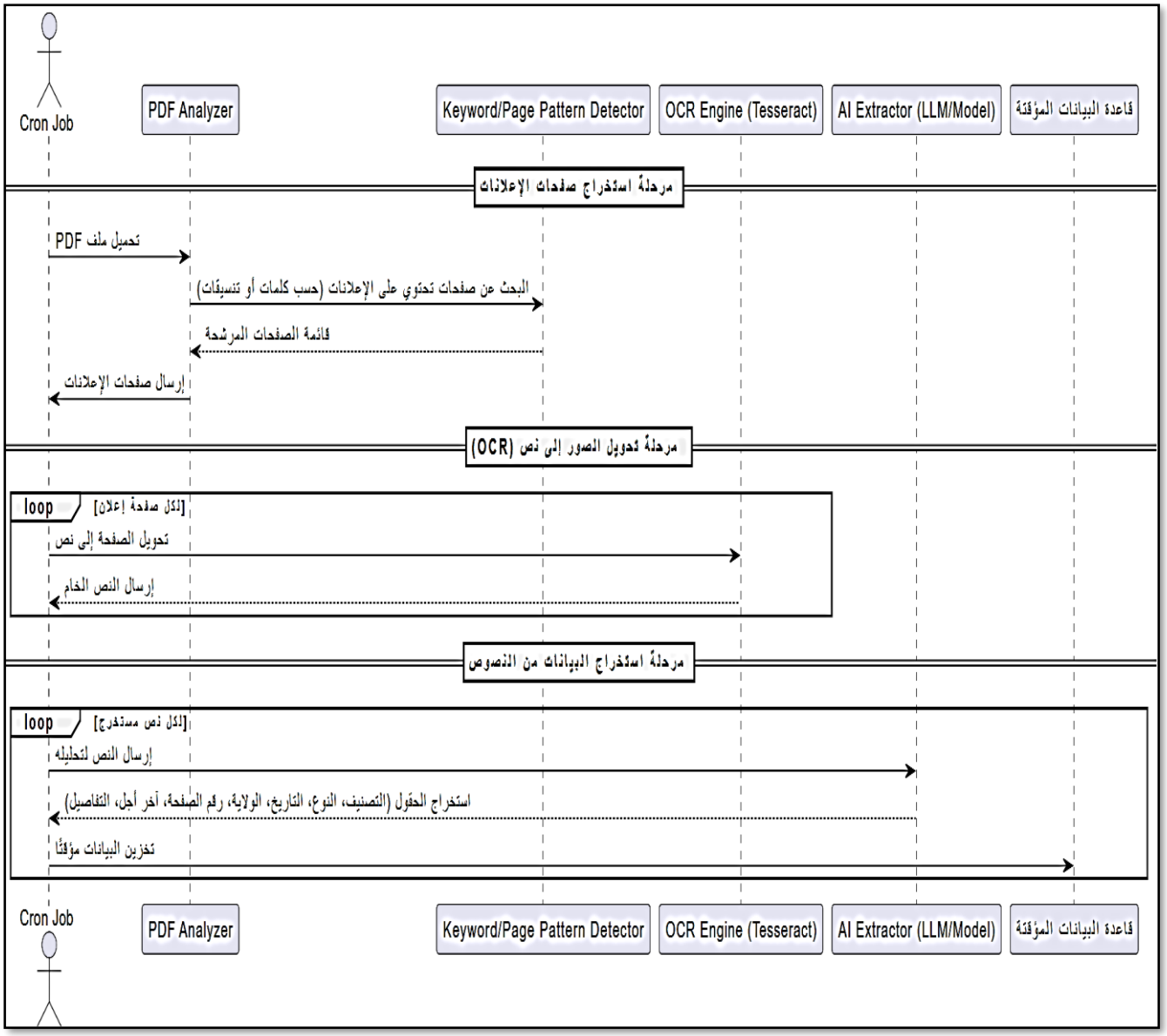
بعد ذلك، يدخل النظام في مرحلة التحديث التلقائي، حيث يتحقق Cron Jobدوريًا من وجود أعداد جديدة لكل جريدة. يتم مقارنة الأعداد الجديدة بالتواريخ المسجلة في قاعدة البيانات، وفي حال توفر عدد جديد غير مكرر، يتم تحميله وحفظه وتحديث القاعدة بالمعلومات الجديدة. إذا لم يكن هناك عدد جديد، أو إذا تبيّن أن العدد موجود مسبقًا، يتم تجاهل العملية للجريدة المعنية. في نهاية كل دورة تحديث، يُسجّل تقرير يلخص النتائج (مثل عدد الأعداد المضافة أو المتكررة ووقت التنفيذ).



الشكل 9 مخطط التسلسل لجمع الملفات والتحديث التلقائي

2.6 مخطط التسلسل لاستخراج الصفحات , OCR و استخراج البيانات

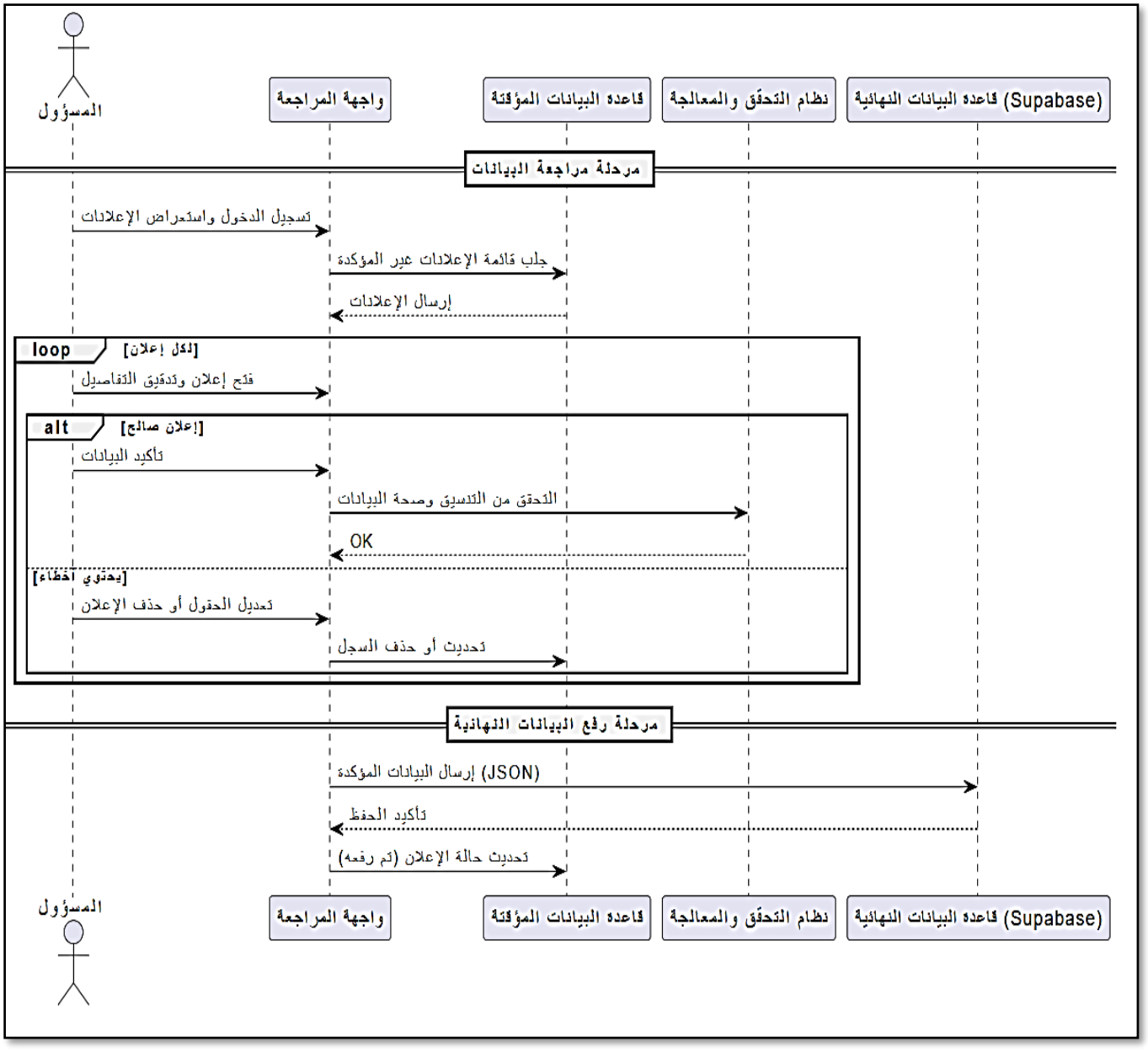
يغطي هذا المخطط ثلاث مراحل مترابطة في معالجة إعلانات الجرائد. تبدأ العملية عندما يقوم Cron Jobبتحميل ملفات PDF وتحليلها عبر وحدة استخراج الصفحات التي تبحث عن الإعلانات اعتمادًا على كلمات مفتاحية أو أنماط تنسيقية. بعد تحديد الصفحات المحتملة، يتم تمرير كل صفحة إلى محرك OCR لتحويلها إلى نص رقمي. بعد الحصول على النصوص، يتم إرسال كل نص إلى نموذج ذكاء اصطناعي مثل (Google AI Studio) لتحليل المحتوى واستخراج الحقول المطلوبة مثل نوع الإعلان، الولاية، التاريخ، التصنيف، رقم الصفحة، آخر أجل، وغيرها من التفاصيل. في النهاية، يتم تخزين البيانات المستخرجة في قاعدة بيانات مؤقتة بانتظار مراجعة المشرف.



الشكل 10 مخطط تسلسل لاستخراج الصفحات , OCR و استخراج البيانات

3.6 مخطط التسلسل لمراجعة المسؤول و رفع البيانات

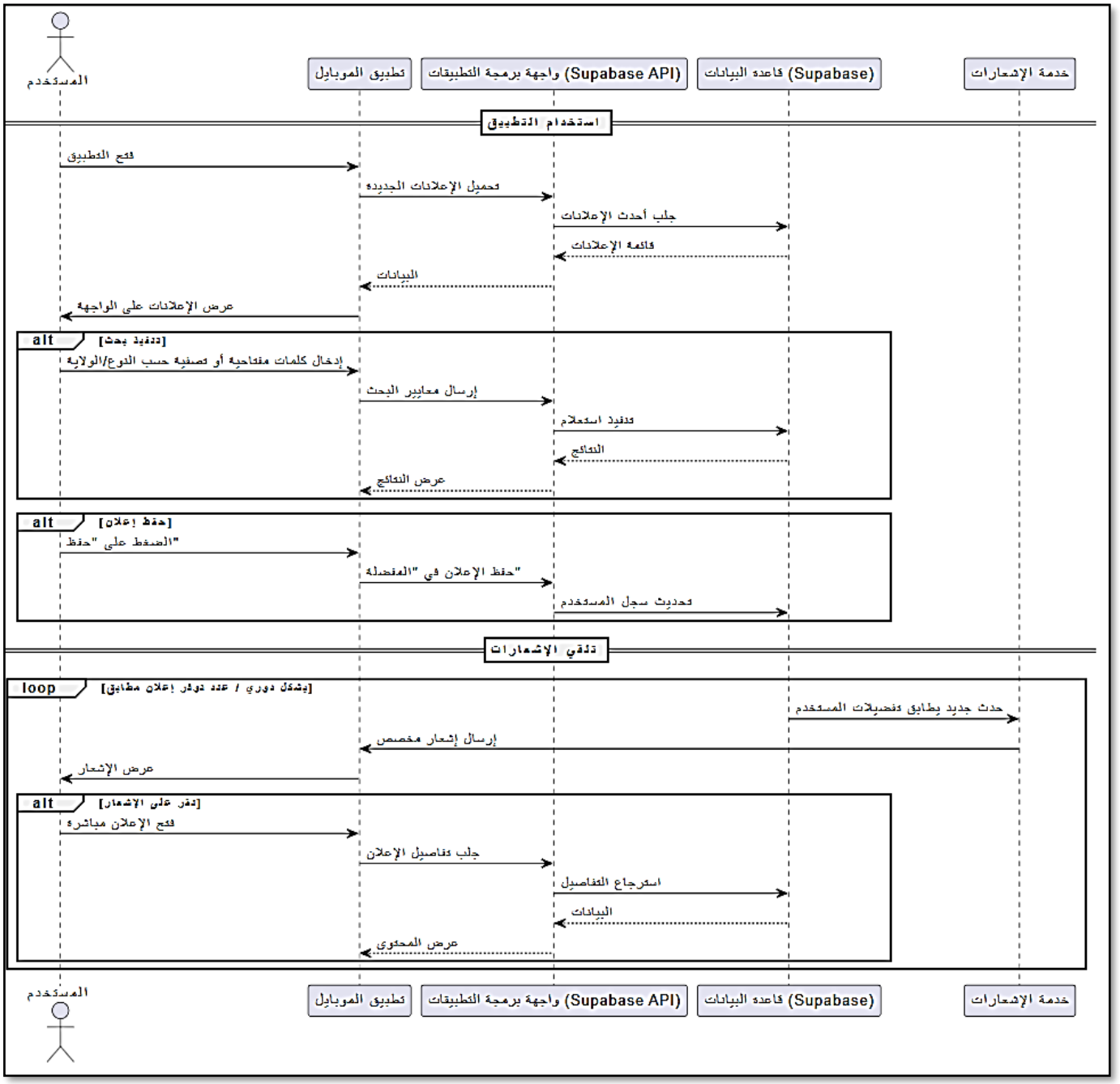
يغطي هذا المخطط مرحلتين متكاملتين تبدأ بمراجعة المشرف وتنتهي برفع البيانات المؤكدة إلى قاعدةSupabase . يبدأ المشرف باستخدام واجهة مراجعة بسيطة تعرض قائمة الإعلانات غير المؤكدة المخزنة مؤقتًا. يقوم بفتح كل إعلان على حدة، ويقرر إما تأكيده (إذا كان صالحًا) أو تعديله أو حذفه (إذا وُجدت أخطاء). عند التأكيد، يتم التحقق من سلامة البيانات من خلال وحدة التحقق، ثم تُرسل البيانات النهائية بصيغة JSON إلىSupabase. بعد حفظها بنجاح، يتم تحديث حالة الإعلان في قاعدة البيانات المؤقتة على أنه تم رفعه.



الشكل 11 مخطط التسلسل لمراجعة المسؤول ورفع البيانات

4.6 مخطط التسلسل لاستخدام التطبيق وتلقي الإشعارات

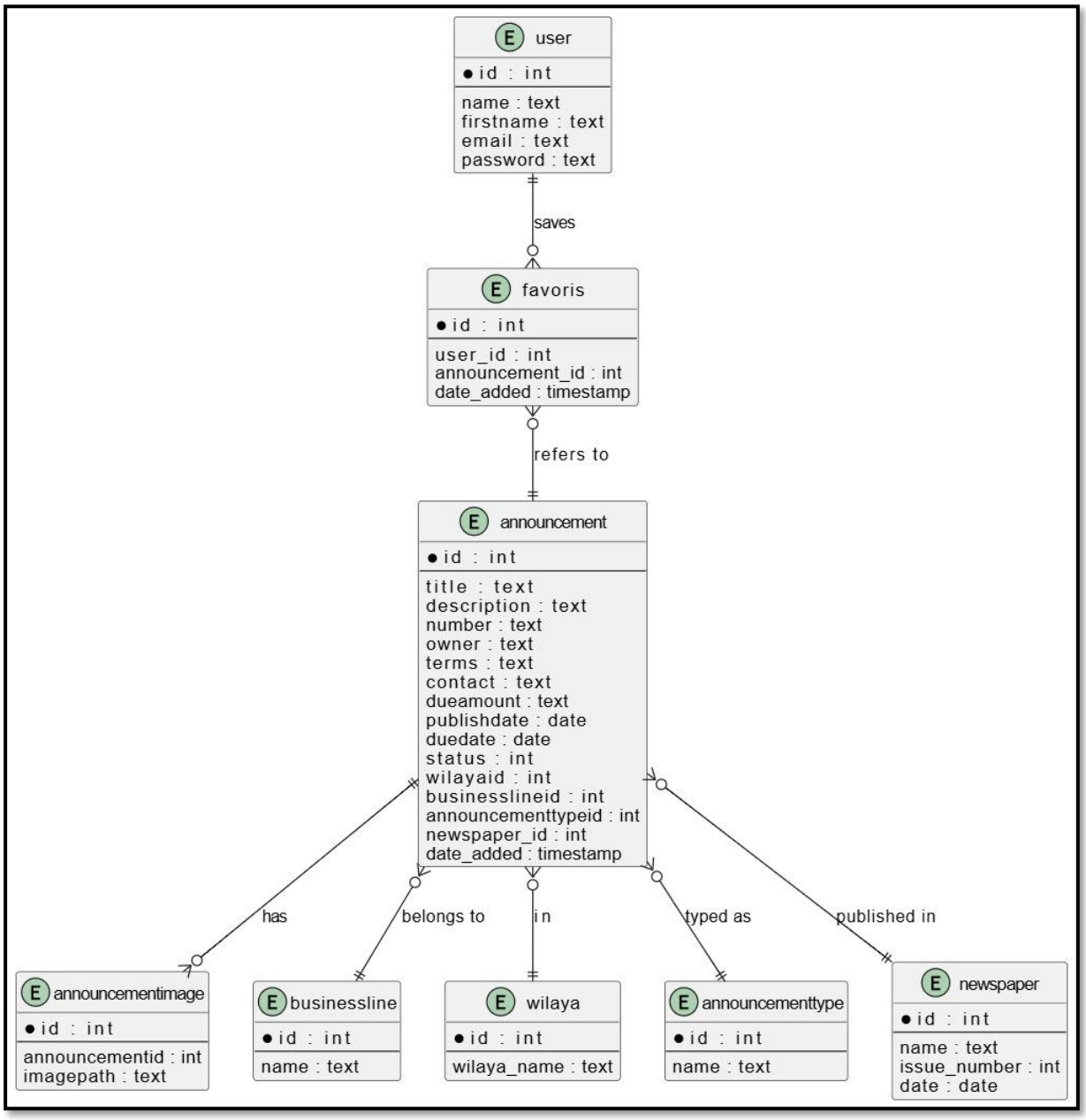
يمثل هذا المخطط تسلسل التفاعلات التي يقوم بها المستخدم عند استخدام التطبيق، مثل فتحه لاستعراض أحدث الإعلانات، تنفيذ عمليات بحث مخصصة، أو حفظ الإعلانات ضمن المفضلة. يتفاعل التطبيق مع Supabase API لجلب أو إرسال البيانات، بينما يتم عرض النتائج في الواجهة. بالتوازي، تعمل خدمة الإشعارات على مراقبة الإعلانات الجديدة ومقارنتها مع اهتمامات المستخدم. عند مطابقة أي إعلان، يتم إرسال إشعار إلى التطبيق، وإذا قام المستخدم بالنقر عليه، يتم فتح الإعلان مباشرة داخل التطبيق، مما يعزز التفاعل السريع مع المحتوى الجديد.



الشكل 12 مخطط التسلسل لاستخدام التطبيق و تلقي الإشعارات

7 مخطط الفئات

يتم في هذه المخطط تحديد الكائنات المختلفة داخل النظام والعلاقات بينها، بما في ذلك الخصائص والعمليات التي تمتلكها تلك الكائنات وكيفية تفاعلها مع بعضها البعض .باستخدام هذه المخططات، يتم تسهيل فهم متطلبات النظام وتصميمه بشكل صحيح وفعال وتحديد الجوانب الرئيسية التي يجب التركيز عليها خلال المراحل اللاحقة من عملية التطوير.



الشكل 13 مخطط الفئات

8 خـــاتـــــــمة

في هذا الفصل، تمّ التركيز على نمذجة المشروع كمرحلة أساسية تمهّد لتطوير التطبيق، وذلك من خلال الإجابة على التساؤلات المرتبطة بالتصور المفاهيمي والبنيوي للنظام .وقد تم اعتماد لغة النمذجة الموحدة (UML) كأداة رئيسية، عبر استخدام مجموعة من المخططات شملت مخططات حالات الاستعمال، تلتها مخططات التسلسل، وانتهاءً بـ مخططات الفئات، مما أتاح بناء تصور شامل ومتكامل حول مكونات النظام وسلوكياته.

أما في الفصل الموالي، فسيتم الانتقال من مرحلة التصور إلى مرحلة التنفيذ، حيث يُركّز على وصف التطبيق وإنجازه تقنيًا استنادًا إلى النماذج المصممة في هذا الفصل.

.

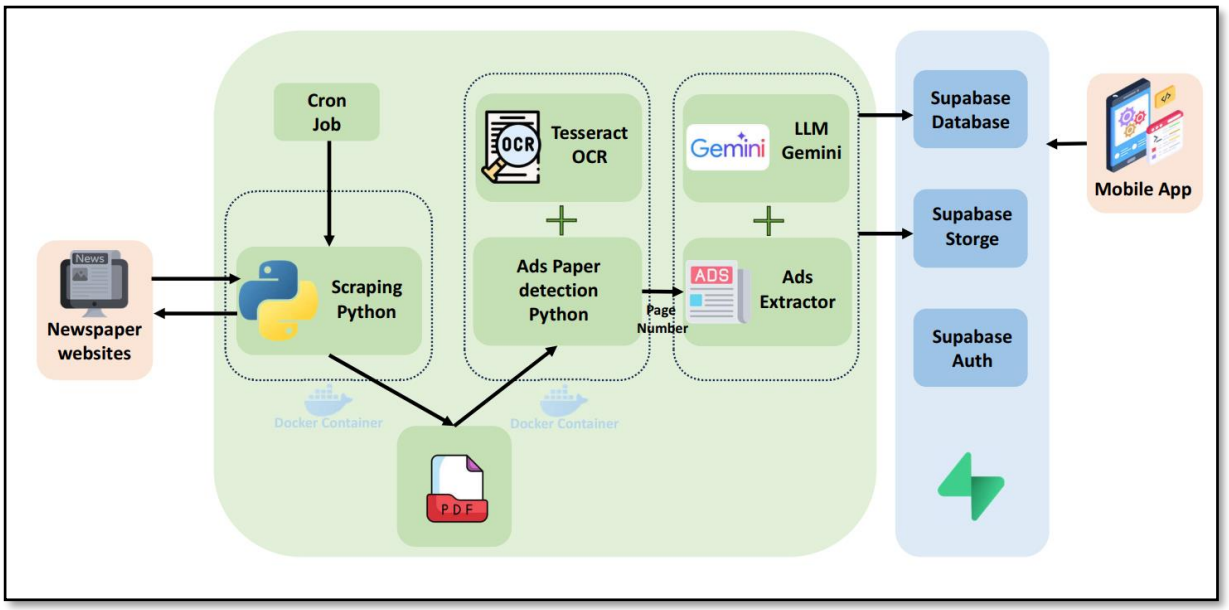
الفصل الثالث: الإنجاز

1 مقدمـــــــــــــة

بعد استكمال مراحل التصميم والنمذجة باستخدام لغة النمذجة الموحدة UML، خصص هذا الفصل لاستعراض الجوانب التقنية المتعلقة بتطوير التطبيق، بدءا من تقديم نبذة موجزة عن لغات البرمجة والأدوات المستخدمة (البيئة التطويرية المعتمدة والتكنولوجيا وقاعدة البيانات)

بعد ذلك، سيتم التركيز على شرح أهم واجهات المستخدم الرسومية للبرنامج من خلال أمثلة توضيحية.

2 مكونات بيئة العمل



الشكل 14 مخطط هندسة النظام

3 اللغات المستخدمة وأطر العمل:

- **Kotlin:** لغة برمجة حديثة، طوّرتها شركة JetBrains وتعمل على آلة جافا الافتراضية (JVM). تُستخدم بشكل شائع في تطوير تطبيقات الأندرويد وتجمع بين البساطة، الأمان، ودعم البرمجة الكائنية والوظيفية.[4]
- **Python :** لغة برمجة تُستخدم على نطاق واسع في تطبيقات الشبكة وتطوير البرامج وعلم البيانات والتعلم الآلي ،يستخدمها المطورون لأنها تتسم بالكفاءة وسهولة التعلم ويمكن تشغيلها على عديد من المنصات المختلفة.[5]

4 التكنولوجيا المستخدمة

- **GitHub**:هي منصة قائمة على السحابة ،حيث يمكنك تخزين الأكواد البرمجية ومشاركتها والعمل مع الآخرين لكتابتها بشكل تعاوني.[6]

- **Continuous Integration (CI)التكامل المستمر**: يعد من الممارسات الأساسية في منهجيات تطوير البرمجيات الحديثة، حيث يقوم المطورون بدمج التعديلات التي يُجرونها على الشفرة البرمجية بشكل منتظم في مستودع مركزي مشترك. وتُخضع كل عملية دمج لسلسلة من إجراءات البناء التلقائي (Automated Build)والاختبارات التلقائية(Automated Tests) ، بهدف الكشف المبكر عن الأخطاء وضمان استقرار النظام. تسهم هذه الممارسة في توفير تغذية راجعة فورية، مما يسمح بمعالجة الأخطاء بسرعة وتقليل التحديات المرتبطة بعمليات الدمج المعقدة. وبهذا، يُساعد التكامل المستمر في رفع جودة البرمجيات، وتحسين كفاءة التطوير، وتقليص الزمن اللازم لاكتشاف وتصحيح المشكلات البرمجية. [7]

- **Continuous Deployment (CD)النشر المستمر**: يشير إلى ممارسة متقدمة في مجال تطوير البرمجيات، يتم من خلالها نشر كل تغيير يتم إدخاله على الشيفرة البرمجية تلقائيًا إلى بيئة الإنتاج، شريطة أن يجتاز هذا التغيير جميع مراحل خط الإنتاج، بما في ذلك الاختبارات التلقائية، دون الحاجة إلى تدخل بشري. لا يُسمح بتمرير التغييرات التي تفشل في الاختبارات إلى بيئة الإنتاج، مما يضمن الحفاظ على استقرار النظام وجودته. وتُساهم هذه المنهجية في توفير دورة تغذية سريعة ومستمرة، كما تُقلل من الضغط المرتبط بإصدارات البرمجيات التقليدية، وتُتيح تنفيذ تحسينات وتحديثات بشكل منتظم وفعال، دون الحاجة للانتظار لفترات زمنية طويلة بين الإصدارات. [8]

- **Docker** :هي منصة تكنولوجيا الحاويات التي تُستخدم لتطوير، نشر وتشغيل التطبيقات بسرعة.[9]

5 شرح النظام (الوصلة بين الأدوات والنظام)

- تم استخدام GitHubكسيرفر مركزي لرفع الكود المصدري وتنظيم إصدارات المشروع عبر الفروع، مع تتبع كل مرحلة من مراحل التطوير بشكل منظم وتعاوني، كما تم استخدامه لربط مستودعات المشروع مع أدوات التكامل التلقائي.

- عبر GitHub Actions، تم إعداد سلسلة مهام تلقائية (Workflow) تبدأ باختبار السكربتات المكتوبة بلغة Pythonوتحزيمها، وتنتهي ببناء صورة Docker وتشغيل اختبارات عليها، ويتم تنفيذ هذه المهام تلقائيًا عند كل تحديث للكود في الفرع الأساسي. استُخدمت ميزة Secretsلحماية بيانات الدخول إلى قواعد البيانات وخدمات الطرف الثالث.

- تم استخدام Dockerلحاوية السكربتات ومعالجة البيانات عبر مراحل المشروع المختلفة. حيث تسهل الحاويات التحكم في البيئة البرمجية وضمان التوافق والاستقرار عند التشغيل محليًا أو على الخوادم.

- تعتمد قاعدة البيانات المركزية على PostgreSQL، مستضافة عبر Supabase، والتي توفر أيضًا واجهات API جاهزة REST وGraphQL، بالإضافة إلى خدمات التوثيق والتخزين.

- تم تطوير واجهة مراجعة البيانات باستخدام React.jsعبر محرر Visual Studio Code، حيث تتيح للمشرفين مراجعة الإعلانات المستخرجة وتصحيح الأخطاء وتأكيد الصلاحية قبل رفعها إلى قاعدة البيانات النهائية.

- استُخدم Google Tesseract OCRلتحويل صفحات الإعلانات من PDF إلى نصوص قابلة للمعالجة، وتم تعزيز مرحلة الاستخراج عبر نموذج ذكي من Google AI Studioلتحليل النص واستخراج الحقول المطلوبة مثل: التصنيف، الولاية، النوع، التاريخ، وغيرها.

- تم تصميم واجهات المستخدم عبر Figmaلتخطيط تجربة الاستخدام لتطبيق الموبايل وتطبيق الويب، مما ساعد في ضمان سهولة التفاعل وسلاسة التنقل بين المراحل.

- طُوّر تطبيق الأندرويد باستخدام Kotlinعبر بيئة Android Studio، ويتيح للمستخدم تصفح الإعلانات والبحث ضمنها وحفظ المفضلات وتلقي إشعارات مخصصة حسب الاهتمامات. التطبيق متصل مباشرة بـ Supabaseلاسترجاع البيانات ورفع التفاعلات عبر REST API أوSupabase SDK[10] .

6 أدوات التطوير المستخدمة

- **PostgreSQLعبر (Supabase) :** قاعدة البيانات المستخدمة لتخزين الإعلانات بعد معالجتها وتنظيمها Supabase .وفرت واجهات API جاهزة، بالإضافة إلى خدمات التوثيق والتخزين. [11]

- **Visual Studio Code** :بيئة تطوير خفيفة استخدمت لتحرير الكود البرمجي بلغة Python وJavaScript، وتسهيل التكامل مع أدوات التحكم في الإصدارات والتوثيق.[12]

- **Figma:**أداة تصميم واجهات المستخدم، استخدمت لتصميم مخططات تطبيق الويب وتطبيق الهاتف، مما ساهم في تحسين تجربة الاستخدام وتوضيح تصور المشروع بصريا. [2]

- **Android Studio :**بيئة التطوير الرسمية لتطبيقAndroid ، تم استخدامها لتطوير تطبيق المستخدم بلغة Kotlin، مع ربطه بقاعدة البيانات عبر REST API أوSupabase SDK .[13]

7 أهم جداول قاعدة البيانات

1.7 الإعلانات(announcement)

يحتوي على معلومات تفصيلية حول الإعلانات المختلفة ويشمل: رقم الإعلان، العنوان، الوصف، رقم المرجع، المالك، الشروط، معلومات الاتصال، المبلغ المستحق، تاريخ النشر، تاريخ الاستحقاق، الحالة، الولاية، مجال العمل، نوع الإعلان، رقم عدد الجريدة، اسم الجريدة، تاريخ الجريدة، وتاريخ الإضافة.

id	title	description	number	owner	terms	contact	dueamount
3	مناقصة وطنية لتوريد تجهيزات طبية لمستشفى بولاية تيزي وزو	تعلن مديرية الصحة والسكان لولاية تيزي وزو عن فتح مناقصة وطنية مفتوحة لتوريد مجموعة من التجهيزات والمعدات الطبية لفائدة المستشفى المركزي بالولاية.	رقم: 2025/07	مديرية الصحة والسكان لولاية تيزي وزو	يمكن سحب دفتر الشروط من مصلحة الصفقات العمومية بالمديرية. آخر أجل للإيداع هو 2025-06-18 على الساعة 14:00.	026 21 45 78	0
4	إعلان عن طلب عروض لإنجاز مشروع تهيئة حضرية ببلدية سيدي بلعباس	تعلن بلدية سيدي بلعباس عن فتح طلب عروض مفتوح للشركات المؤهلة لإنجاز مشروع تهيئة حضرية يشمل تعبيد الطرق الرئيسية، تركيب أعمدة الإنارة ذات النجاعة الطاقوية، وتهيئة بعض المساحات الخضراء.	رقم: TB/01/2025	بلدية سيدي بلعباس	يجب تقديم العروض في أجل أقصاه 2025-06-20 على الساعة 10:00 صباحًا بمقر البلدية.	tel: 048 55 12 34	0

publishdate	duedate	status	wilayaid	businesslineid	announcementtypeid	newspaper_issue_nu...	newspaper_name	newspaper_date	date_added
18/05/2025	18/06/2025	نشط	15	3	3	العدد 1254	الشروق اليومي	19/05/2025	2025-05-12 12:30:00+02
20/05/2025	20/06/2025	نشط	22	4	3	العدد 987	النهار الجديد	21/05/2025	2025-05-12 12:35:00+02

جدول 4 الإعلانات (announcement)

2.7 صور الإعلانات(announcementimage)

يتضمن معلومات حول الصور المتعلقة بكل إعلان ويشمل: رقم الصورة، رقم الإعلان المرتبط، ومسار الصورة.

id	announcementid	imagepath
3	3	/images/tender_medical_equipment.jpg
4	4	/images/tender_urban_project.jpg

جدول 5 صور الإعلانات (announcementimage)

3.7 مجال العمل(businessline)

يحدد مجال العمل لكل إعلان ويشمل رقم مجال العمل، والاسم.

id	name
3	الصحة
4	الأشغال العمومية

جدول 6 مجال العمل (businessline)

4.7 الولاية(wilaya)

يحدد مجال العمل لكل إعلان ويشمل رقم مجال العمل، والاسم.

id	wilaya_name
15	تيزي وزو
22	سيدي بلعباس

جدول 7 الولاية (wilaya)

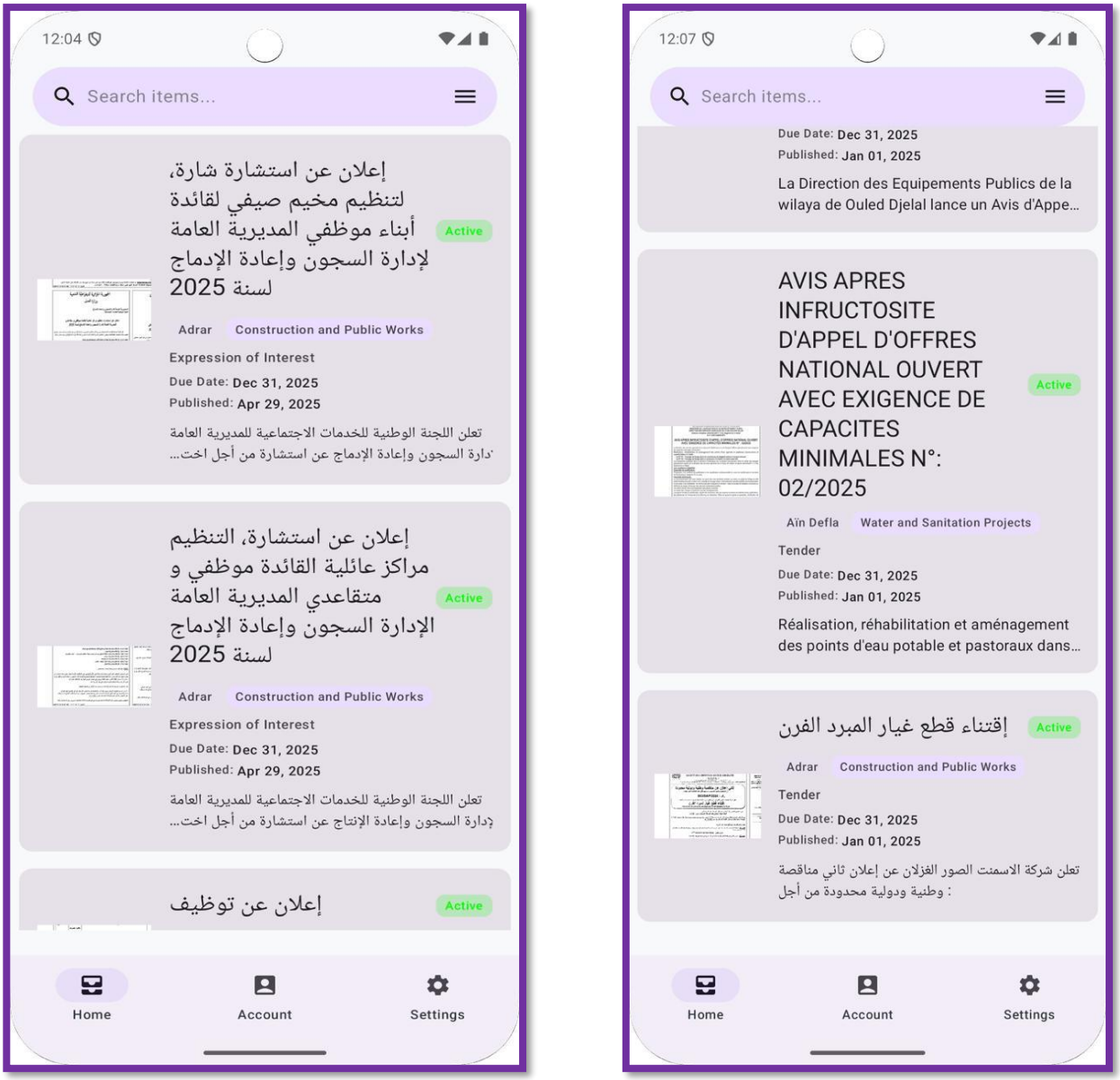
5.7 نوع الإعلان(announcementtype)

يحدد نوع كل إعلان ويشمل: رقم نوع الإعلان، والاسم.

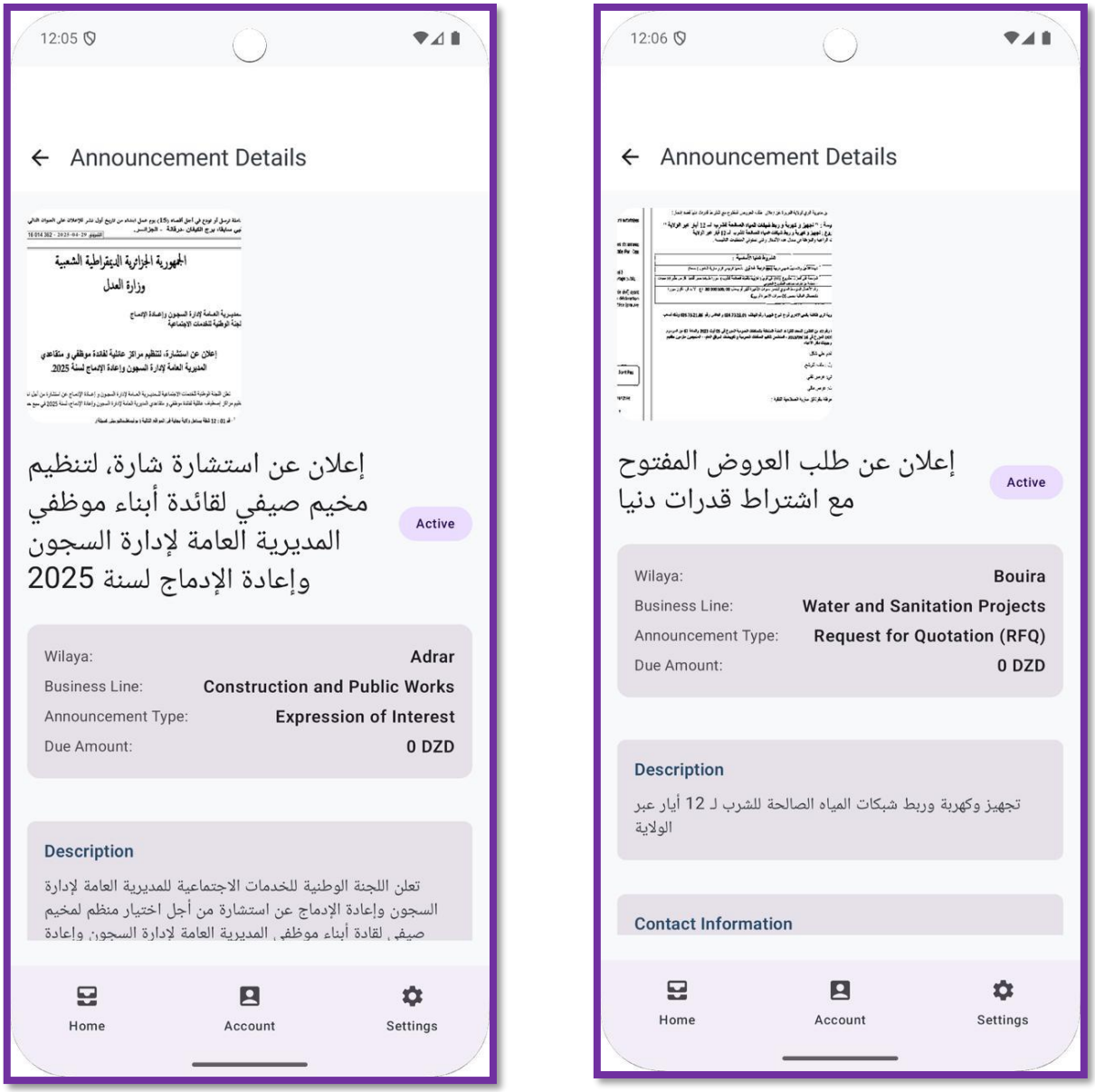
id	name
3	مناقصات

جدول 8 نوع الإعلان (announcementtype)

8 الواجهات الرئيسية لتطبيق "OfferSeek"



الشكل 15 الواجهة الرئيسة



الشكل 16 إظهار تفاصيل الاعلان

الشكل 17 صورة الاعلان

9 خاتمـــــــــــة

بعد استكمال تطوير الواجهات المطلوبة، بات لدينا تطبيق متكاملة يعمل بتناغم ضمن نظام شامل وواحد، محققةً بذلك أهداف دراستنا ومترجمة جهودنا الى واقع ملموس.

الخاتمة العامة

في ختام هذه المذكرة، يمكن القول إن المشروع قد نجح في تحقيق الأهداف المسطرة من خلال توفير حل رقمي متكامل لجمع الإعلانات المنشورة في الجرائد الورقية، وتنظيمها، وتحليلها، ثم عرضها بشكل ذكي ومتاح للمستخدمين عبر تطبيق موبايل سهل الاستخدام. وقد سمح لنا هذا المشروع بدمج عدة تقنيات متقدمة في مجالات مختلفة، بدءًا من **Web Scraping**و**OCR**، مرورًا باستعمال النماذج اللغوية الكبيرة **LLMs** وتقنيات **Prompt Engineering**، وانتهاءً بتطوير واجهة تفاعلية وتطبيق موبايل بلغةKotlin ، مما يبرز التكامل الحقيقي بين الذكاء الاصطناعي وتطوير البرمجيات.

لقد مرت مراحل الإنجاز بعدة خطوات مدروسة، بدايةً بتكوين الأرشيف وتحليل بنية الجرائد، ثم المرور بمرحلة التحديث التلقائي، وتحديد صفحات الإعلانات، وصولًا إلى استخراج البيانات وتصنيفها آليًا باستخدام أدوات الذكاء الاصطناعي. ولم يكن هذا كافيًا دون مرحلة مراجعة الجودة، التي تضمن دقة وصحة المعلومات المقدّمة للمستخدم النهائي. وفي النهاية، تم رفع البيانات إلى قاعدة بيانات منظمة وفعالة، ليتم استغلالها عبر تطبيق موبايل يتيح للمستخدم البحث والتفاعل بسهولة.

إن أهمية هذا المشروع لا تكمن فقط في الجوانب التقنية التي تطرقنا إليها، بل تتعداها إلى المساهمة الفعلية في رقمنة محتوى تقليدي لا يزال بعيدًا عن متناول محركات البحث والمنصات الرقمية. فباستعمال هذا النظام، يمكن تسهيل الوصول إلى محتوى الإعلانات الرسمية والمهمة، وتوفير الوقت والجهد على الباحثين والمهتمين.

ختامًا، فتح لنا هذا المشروع آفاقًا واسعة لفهم التحديات الحقيقية المرتبطة بالتكامل بين البيانات غير المهيكلة والتقنيات الحديثة، ووفر لنا تجربة عملية ثرية تجمع بين التحليل، التصميم، والبرمجة، مما يؤهلنا لمواجهة تحديات السوق الرقمية بثقة وكفاءة.

قائمة المراجع

[1] ❖ G. A. Studio :https://makersuite.google.com.

[2] ❖ F. T. C. I. D. Tool :https://www.figma.com.

[3] ❖ UML ، "،" ،: About the Unified Modeling Language Specification Version 2.5.1 (omg.org).[تاريخ الوصول] .

[4] ❖ K. P. Language: https://kotlinlang.org.

[5] ❖ P. P. Language :https://www.python.org.

[6] ❖ A. G. a. Git: https://docs.github.com/en/get-started/using-git/about-git.

[7] ❖ W. i. C. I. |. Atlassian :https://www.atlassian.com/continuous-delivery/continuous-integration.

[8] ❖ W. i. C. D. |. Atlassian: https://www.atlassian.com/continuous-delivery/continuous-deployment.

[9] ❖ D. E. A. D. f. Developers :https://www.docker.com.

[10] ❖ S. J. Library: https://supabase.com/docs/guides/with-js.

[11] ❖ S. –. T. O. S. F. Alternative :https://supabase.com.

[12] ❖ V. S. C. –. C. E. Redefined :https://code.visualstudio.com.

[13] ❖ A. S. –. A. Developers :https://developer.android.com/studio.