Compte Rendu du Projet : Test IA Junior

Introduction

L'objectif principal de ce projet était d'extraire automatiquement des informations clés à partir de contrats de bail au format PDF, qu'ils soient en texte brut ou scannés. Ces informations incluent des éléments comme le bailleur, le preneur, l'adresse, la surface, la durée du bail, etc. Pour atteindre cet objectif, j'ai combiné différentes approches, en utilisant des techniques classiques d'extraction de texte et des méthodes plus avancées reposant sur l'intelligence artificielle.

---

2. Différentes Approches Utilisées

Pour l'extraction des textes des fichiers PDF, j'ai expérimenté deux principales approches :

📌 *1ère Approche : Extraction de texte brut (Méthodes basées sur des librairies PDF)*

Objectif : Extraire directement le texte des fichiers PDF contenant du texte sélectionnable.
Outils utilisés : PyMuPDF .
Avantages :

- Extraction rapide et précise (~100% de précision si le texte est bien structuré).
- Facilité de manipulation et de structuration du texte extrait.

📌 *2ème Approche : Extraction via OCR (Reconnaissance Optique de Caractères)*

Objectif : Convertir les PDF scannés en texte exploitable en utilisant l'OCR.
Outils utilisés : pdf2image pour convertir le PDF en image, puis pytesseract (Tesseract OCR) pour extraire le texte.
Avantages :

- Permet d'extraire du texte à partir des documents scannés.
- Compatible avec divers types de PDF, même ceux contenant des images de texte.

Inconvénients :

- L'extraction peut être désordonnée, nécessitant un nettoyage du texte pour le rendre exploitable.

Pour l'extraction des informations dans les fichiers PDF, J'ai expérimenté trois principales approches :

1ère Approche : Extraction des informations via Regex

Objectif : Identifier les informations spécifiques (bailleur, preneur, surface, etc.) dans le texte qui correspondent aux motifs spécifiques

Avantages :

- Efficace pour des documents très structurés, avec une mise en page constante.

Inconvénients :

- Limitée dès que la structure du document change. Les règles doivent être adaptées.

🏷 2ème Approche : : Extraction des informations NLP (Modèle de Reconnaissance d'Entités Nommées - NER)

Objectif ; Identifier automatiquement les entités importantes dans un texte en catégories prédéfinies, telles que le bailleur, le preneur, la surface, etc

Model utilise : Camembert

Avantages :

- **Adaptabilité** à différents formats de texte et contexte, même si la structure du document varie légèrement.
- **Capacité de gestion de texte non structuré**, permettant d'extraire des informations à partir de documents ayant des formats et structures diverses.
-

Inconvénients :

- **Nécessite un dataset annoté** de qualité pour un entraînement efficace
- **Ressources computationnelles élevées**, ce qui peut rendre l'exécution du modèle coûteuse, surtout pour des documents volumineux.
- **Moins performant sur des données très bruitées** ou des textes très mal formatés, comparé à des techniques plus simples comme les expressions régulières.
- **Fine-tuning limité par la taille du dataset** : Si le dataset est trop petit, le fine-tuning peut ne pas suffire pour adapter le modèle efficacement aux spécificités du domaine, menant à un **sur-apprentissage** sur un petit jeu de données, ou à des résultats biaisés et non généralisables.

---

3. Problèmes rencontrés

Lors de l'exécution du projet, plusieurs défis ont été rencontrés :

- **Manque de dataset annoté** : Le manque de données annotées a entravé la performance du modèle NER, qui aurait bénéficié d'un ensemble de données plus complet pour un entraînement efficace affectant la capacité du modèle à généraliser correctement à des documents diversifiés.

- Le temps limité a empêché des essais approfondis et l'optimisation des configurations, particulièrement en ce qui concerne l'entraînement du modèle NER. Le **temps d'entraînement** nécessaire pour un fine-tuning efficace du modèle Camembert a été un facteur contraignant, ce qui n'a pas permis d'explorer pleinement les configurations et d'améliorer la performance du modèle.
- **Coût computationnel élevé** : L'entraînement d'un modèle pré-entraîné comme Camembert pour la reconnaissance d'entités nommées nécessite des ressources computationnelles importantes, en particulier pour le fine-tuning.

---

4. Résultats et Observations

Regex et NER :
Les approches basées sur Regex ont bien fonctionné pour les fichiers bien structurés. Cependant, le manque de données annotées a limité l'efficacité du modèle NER, qui n'a pas pu extraire les informations avec la précision espérée.

---

5. Conclusion et Perspectives

Ce projet a démontré la faisabilité de l'extraction d'informations depuis des contrats de bail en utilisant une combinaison de plusieurs approches techniques. Cependant, le manque de dataset annoté et le temps limité ont restreint la précision du modèle basé sur l'IA. Avec un jeu de données plus important et un fine-tuning plus poussé du modèle Camembert NER, il serait possible d'améliorer considérablement la qualité des extractions. Ce processus pourrait réduire la dépendance aux règles Regex et offrir une solution plus robuste et flexible pour l'extraction d'informations dans des formats variés de contrats de bail.