Fahd Labba

Extracteur de Caractéristiques de Documents Juridiques

21 Février 2023

# Rapport de mini project

## Objective :

Développer un prototype capable de traiter un document juridique numérisé en arabe et d'extraire des caractéristiques clés prédéfinies pertinentes pour un type spécifique d'affaire juridique.

## Approach :

pour développer cette prototype et satisfait les objective donner j’ai suivi ces étapes :

1. **OCR (extracteur de text ):**

   Pour extraire le texte du document, j'ai trouvé que Tesseract était le meilleur modèle raison de sa grande précision et de sa prise en charge de l'extraction de texte arabe.

2. **Gemini (extracteur d’informations):**

   Après l'extraction du texte (étape 1) du document juridique, il est nécessaire d'extraire les caractéristiques clés. C'est pourquoi j'ai utilisé la puissance des LLMs (Large Language Models) dans le domaine du traitement du langage naturel (NLP) pour bien comprendre les mots clés. D'où vient le choix de Gemini (un nouveau LLM de Google qui surpasse GPT-4 dans les benchmarks). De plus, Google fournit une API gratuite pour ce modèle. Donc en utilisant la puissance de ce modèle, j’ai garanti d'extraire toutes les caractéristiques clés avec une très grande précision .

3. **Applications Web :**

   Donc, après l'étape 1 (extraction de texte) et l'étape 2 (extraction de caractéristiques clés), il faut retourner le résultat à l'utilisateur. C'est pourquoi, pour cette étape, j'ai utilisé Streamlit pour développer une application web dont l'utilisateur peut y accéder et télécharger son document juridique. Il recevra ensuite les informations importantes en json format (s'il est possible ) suivantes :

   - Date du document
   - Parties impliquées
   - Termes clés
   - Éléments d'action
   - Sujets du document

   De plus, si l'utilisateur n'a pas le document télécharger sur son PC, il lui suffit de fournir le lien du document.

## Défis :

L'un des défis que je rencontre dans le développement de ce projet est la partie de Prompt Engineering ,en effet les données fournies au modèle sont un peu délicates ( car la langue arabe est l'une des langues les plus difficiles , d'où chaque mot affecte le sens du paragraphe).et il est nécessaire d'extraire toutes les informations importantes et de conserver les idées principales.
Un autre challenge que j'ai rencontré était de choisir le modèle OCR le plus précis pour l'extraction de texte écrit en arabe.

## Suggestions :

L'une des suggestions que j'ai aimé proposer pour améliorer la performance de cette solution est de s'abonner à Google Cloud Vision. En effet, cette plateforme propose plusieurs modèles OCR qui donnent des résultats plus précis que les modèles open source.

L'autre solution consiste à réaliser l'étape d'extraction de texte (étape 1) en utilisant un autre LLM qui donne de bons résultats. Personnellement, j'ai utilisé Gemini-Pro-Vision et j'ai obtenu de bons résultats.

## Technologies Utilisées :

- **Langchain :** pour travailler avec Gemini .
- **Streamlit** : responsable de développement de l'application web .
- **opencv :** pour travailler avec les images .
- **pytesseract :** pour travailler avec notre modèle OCR (tesseract) .