Bilan des travaux du projet tuteuré du 28/09/2024

Sujets : Utilisation de l'IA et de l'OCR pour la classification automatique de documents

Objectif général : Proposition un pipeline de classification automatique assistée par l'IA des document de la cours des comptes.

Objectifs spécifiques :
- Identification des classes de document de la cours des comptes.
- Étude comparative des modèles d'IA pour la classification des documents structurés, semi-structurés et non structurés.
- Proposition d'une approche de classification automatique des documents pour la cours des comptes.

Résultats obtenus :

- Les classes de document de la cour des comptes identifiées sont : *Rapports d'audit, rapports annuels, observations, jugements, référés, comptes rendus, décisions et rapports thématiques.*
- Après une étude comparative, on a:
    - Le modèle **Pix2Text** : détecter et extraire le texte des images (OCR).
    - Le modèle **Transformers** pour la classification des documents.
- Les modèles **Pix2Text** et **Transformers** implémentés et testés. Il reste à présent à implémenter une interface graphique pour leur usage.

Travaux en cours :

- La rédaction du rapport a également commencé.
- Le stockage du texte extrait dans une base de données.

NB : une rencontre est prévue ce soir à 20h pour faire le bilan de la semaine et préciser les objectifs de la semaine prochaine.

Difficultés :

- Les données de la cour des comptes ne sont pas toujours disponibles. Nous travaillons actuellement avec des données de test.