Projet extraction d'information

Extraction d'informations à partir de documents scannés

Description du projet :

De nombreuses entreprises doivent traiter des centaines de factures papier ou scannées. Extraire manuellement les données importantes (numéro de facture, date, montant, TVA, etc.) est une tâche fastidieuse, sujette à des erreurs humaines.

Ce projet vise à automatiser le processus d'extraction d'informations clés à partir de factures scannées, à l'aide de techniques de Reconnaissance Optique de Caractères (OCR). L'OCR est un processus qui permet de convertir du texte contenu dans une image (ou un PDF) en texte numérique exploitable. Cela concerne aussi bien des fichiers image (JPG, PNG...) que des documents PDF ou numérisés. Le résultat de l'OCR est généralement un bloc de texte brut et non structuré. L'objectif est alors d'extraire automatiquement certaines informations spécifiques à partir de ce texte brut : numéro de facture, date, nom du client, montant total, etc. Pour cela, on utilisera les expressions régulières, un outil puissant permettant de rechercher des motifs (patterns) précis dans un texte, comme des dates, des montants ou des codes. Une fois les données extraites, elles sont enregistrées dans un format structuré, tel qu'un fichier CSV (valeurs séparées par des virgules) ou un fichier JSON, afin de faciliter leur traitement ou archivage ultérieur.

Le projet à réaliser comprend les étapes suivantes :

1. Chargement des documents

- Accepter les fichiers aux formats **image (JPG, PNG, TIFF)**.
- Convertir les pages PDF en images si nécessaire.

2. Prétraitement d'image

- Appliquer des filtres (binarisation, seuils, redressement, nettoyage du bruit).
- Améliorer la lisibilité pour optimiser l'OCR.

3. Extraction du texte avec OCR

- Utiliser un OCR pour l'extraction de caractéristiques visuelles de l'image et obtenir le texte brut présent dans le document.
- Implémentation avec la bibliothèque pytesseract ou easyocr

4. Extraction des données structurées

- Utiliser des **expressions régulières** pour repérer :
    - Numéro de facture
    - Date
    - Raison sociale ou nom du client

    - Montant total
    - TVA, sous-total, autres éléments
- Implémentation avec la bibliothèque re, Spark NLP

5. Structuration des résultats

- Enregistrer les données extraites dans des fichiers JSON ou CSV.
- (Optionnel) Affichage dans un tableau via une interface.

6 Interface utilisateur (bonus)

- Développer une application graphique pour charger un fichier, lancer l'analyse, et afficher les résultats.

➢ **Chaque équipe devra remettre :**

- Le code source commenté (language de progrmmation : python)
- Un rapport d'environs 10 pages détaillants toutes les étapes de réalisation

➢ **Prolongements possibles (pour les groupes avancés)**

- Intégration avec une base de données (SQLite, PostgreSQL)
- Reconnaissance multi-langues

➢ **Date limite de remise du projet :**

La semaine précédant les examens du second semestre.