

PDF et IA : extraire sans inventer

Mémo d'atelier · OCR, JSON et contrôle des preuves

1. Identifier le PDF

- Texte sélectionnable : essayer l'extraction directe.
- Scan : OCR ; document mixte : --skip-text.
- Une couche OCR mauvaise demande un autre diagnostic ; conserver l'original.

2. Préparer l'OCR

- VM serveur isolée pour les fichiers non fiables.
- Tesseract fra+eng ; vérifier les langues installées.
- Document net, plat, éclairage uniforme ; scan proche de 300 dpi.
- Contrôler visuellement rotation et redressement.

3. Préserver la provenance

- Empreinte du fichier, nom de source, page et versions des outils.
- Un fichier texte par page pour le premier pipeline.
- Pour chaque champ : valeur brute, page, extrait exact.

4. Extraire

- Ollama local au serveur, adresse 127.0.0.1:11434.
- Schéma explicite ; réponse contrôlée avec Pydantic.
- Valeur absente : valeur, page et extrait = null.
- Interdire les champs inattendus et conserver les erreurs.

5. Trois niveaux de contrôle

- Structure : clés et types conformes.
- Provenance : extrait présent dans la page, contenant la valeur.
- Sens : total TTC et date de facture, pas un montant ou une date voisins.
- Lire l'original : OCR, signes, virgules, devise et caractères O/O.

6. Évaluer et revenir en arrière

- Facture fictive témoin ; retirer un champ pour tester null.
- Comparer dix documents à des réponses préparées à la main.
- Valeurs contradictoires entre pages : revue humaine.
- Garder originaux, travail et sorties séparés ; aucun import métier automatique.

Article complet, exercices et sources : blog.gringao.fr/ia-documents-pdf-ocr-json-fiable/