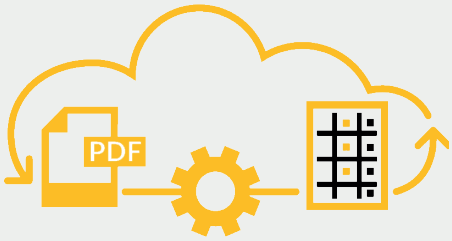
Par Qianwen GUAN,
Consultante Data Scientist, StarClay

Extraction de données tabulaires dans les fichiers PDF

Comment bien choisir ses outils ?

Lorsque nous travaillons avec des Modèles de Langage (LLMs), les documents PDF posent des défis uniques. Conçus avant tout pour un affichage visuel optimal,
ces fichiers, non-structurés, ne sont pas organisés de manière logique pour une exploitation machine.
Cela rend le parsing — ou l'extraction structurée des données — indispensable pour une exploitation efficace.



Dans le cadre d'un projet avec l'un de nos clients, nous avons développé un Proof of Concept (POC) visant à créer une solution innovante de Q&A sur des rapports financiers. Ces rapports incluent des données tabulaires et graphiques, nécessitant une précision d'extraction élevée.
La première étape de ce POC a été de choisir et d'intégrer les meilleures technologies pour le parsing de fichiers PDF, afin de fournir une base fiable et exploitable pour notre solution.

Quels sont les enjeux du parsing de PDF ?

Dans le domaine de l'extraction de données documentaires, le parsing des PDF est une étape technologique critique, notamment lorsqu'il s'agit de traiter des contenus complexes comme des tableaux et des graphiques. Contrairement à une base de données, les PDF n'organisent pas leurs informations de manière directement exploitable pour les machines.

Cela complique leur intégration dans des systèmes d'IA avancés, comme les LLMs, qui requièrent des données claires, précises et bien structurées.

Nous avons réalisé une analyse comparative approfondie pour évaluer les capacités des technologies actuelles pour le parsing multimodal des PDF.
Deux solutions se distinguent : **Unstructured (open-source) et LlamaParse, une solution commerciale avancée.**

Technologies de Parsing : Comparaison entre Unstructured et LlamaParse

Critères	Unstructured	LlamaParse
Type	Open-source avec une version commerciale payante pour les entreprises	Solution commerciale avancée Gratuit jusqu'à 1000 crédits/jour
Modèles utilisés	OCR + Transformers (YOLOX, Detectron 2)	OCR + Modèles avancés : GPT-4o, Claude 3.5 sonnet, Gemini 1.5 pro...
Format pris en charge	PDF, Word, PowerPoint, images...	PDF, Word, PowerPoint, Excel, images...
Format de sortie	Texte, HTML	Texte, JSON, Markdown
Facilité d'intégration	Requiert une configuration technique, adapté aux développeurs	Solution clé en main, intégration simplifiée avec support client

Critères d'évaluation des performances

Pour évaluer la qualité de l'extraction, nous avons retenu trois critères principaux :

1. **Intégrité de la structure du tableau :** Le tableau extrait conserve sa forme, sa taille et son organisation.
2. **Exactitude des en-têtes :** Les titres des colonnes sont correctement identifiés.
3. **Fidélité des contenus des cellules :** Les valeurs des cellules sont extraites sans erreurs significatives.

Quels sont les enjeux du parsing de PDF ?

LlamaParse

1. Anthropic Claude 3.5 Sonnet :

- Plus grande fidélité globale avec les tableaux PDF.
- Structure des tableaux et en-têtes extraites avec précision.
- Erreurs minimes sur les contenus des cellules.
- **Coût :** 20 crédits par page ($60 / 1000 pages).

2. Mode Premium de LlamaParse :

- Performance proche de Claude 3.5 Sonnet.
- **Coût :** 15 crédits par page ($45 / 1000 pages).

3. OpenAI GPT-4o :

- Moins précis que Claude 3.5.
- Erreurs significatives sur la structure des tableaux et les contenus des cellules.
- **Coût** : 10 crédits par page ($30 / 1000 pages).

Unstructured (Open-source)

- Performance variable selon les types de tableaux.
- Moins performant pour les données complexes et les PDF issus de conversion depuis Word (DOCX).
- **Coût** : Gratuit.

Nos recommandations :
une stratégie hybride pour optimiser le parsing PDF

Pour connecter vos données tabulaires aux LLMs, voici une approche optimisée en termes de coûts :

- Utilisez des **solutions open-source gratuites** comme Unstructured pour les contenus simples ou textuels.
- Adoptez des outils avancés comme **LlamaParse** (Claude 3.5) pour les tableaux complexes et critiques.

Le choix de la technologie dépendra toujours de vos besoins spécifiques : complexité des documents, budget, niveau de précision attendu et objectifs finaux.

Par Qianwen GUAN,
Consultante Data Scientist, StarClay

Références

https://docs.cloud.llamaindex.ai/llamaparse/getting_started

https://github.com/Unstructured-IO/unstructured

starclay

www.starclay.com
hello@starclay.com