T-ESP-700

Big Brain

Lucas FIXARI
Pierre ROCHETTE
William WOZIWODA

Description Fonctionnelle et Technique

1. Introduction

Ce document explique le fonctionnement de l'IA d'analyse documentaire locale, en décrivant chaque partie de l'application, les outils utilisés (en Python) et les critères pour évaluer sa performance.

2. Architecture Générale

L'IA est divisée en plusieurs modules pour organiser, analyser et chercher des informations dans les documents. Les modules sont :

1. **Prétraitement des Données**
2. **Traitement Automatique du Langage Naturel (NLP)**
3. **Domaine de Connaissances (Optionnel)**
4. **Extraction et Intégration Sémantique**
5. **Base de Connaissances**
6. **Interface Utilisateur**

3. Modules Fonctionnels

3.1 Prétraitement des Données

- **Objectif :** Préparer les documents pour les analyser facilement.
- **Technologies :** PyMuPDF pour PDF, python-docx pour DOCX, openpyxl pour Excel, et pytesseract pour lire le texte sur les images.
- **Fonction :** Extraire le texte et les images, transformer en format JSON pour faciliter l'organisation.
- **Critères :** 98 % de précision et un temps de traitement de 5 secondes par document.
- **Plan B :** Utiliser seulement PDF et DOCX si certains formats posent problème.

3.2 Traitement Automatique du Langage Naturel (NLP)

- **Objectif :** Comprendre les documents et les questions posées.
- **Technologies :** spaCy pour l'analyse de texte, Transformers pour des analyses plus avancées.
- **Fonction :** Identifier les mots importants et analyser la grammaire pour comprendre le sens.
- **Critères :** 80 % de précision et réponse en 3 secondes pour des questions simples.
- **Plan B :** Simplifier les questions au début.

3.3 Domaine de Connaissances (Optionnel)

- **Objectif :** Ajouter des informations sur les relations entre concepts pour une compréhension plus précise.
- **Technologies :** rdflib pour gérer les relations, scikit-learn pour améliorer les modèles.
- **Fonction :** Créer une structure de relations pour mieux comprendre le contexte.
- **Critères :** Une bonne organisation des connaissances.

3.4 Extraction et Intégration Sémantique

- **Objectif :** Identifier les relations entre les informations et les organiser dans une base de connaissances.
- **Technologies :** spaCy pour les relations simples, networkx pour créer des liens.
- **Fonction :** Structurer les informations pour donner des réponses contextuelles.
- **Critères :** Temps de réponse de 5 secondes pour traiter les relations.
- **Plan B :** Simplifier si les performances sont limitées.

3.5 Base de Connaissances

- **Objectif :** Stocker et organiser les informations pour une recherche rapide.
- **Technologies :** SQLite pour stocker en local, Whoosh pour créer un moteur de recherche.
- **Fonction :** Organiser les documents et permettre des recherches rapides.
- **Critères :** Temps de réponse de 5 secondes pour une recherche.

- **Plan B :** Simplifier les recherches si besoin.

3.6 Interface Utilisateur

- **Objectif :** Permettre à l'utilisateur de poser des questions et voir les réponses facilement.
- **Technologies :** Flask, FastAPI ou tkinter pour l'interface.
- **Fonction :** Interface pour poser des questions et voir des réponses sous forme de résumés.
- **Critères :** Affichage en moins de 5 secondes.
- **Plan B :** Afficher directement des extraits si les résumés ne fonctionnent pas bien.

4. Synthèse des Technologies Utilisées

Module	Tâche	Technologie
Prétraitement des Données	Extraction de texte	*PyMuPDF, python-docx, openpyxl*
	OCR pour images	*pytesseract*
Traitement NLP	Reconnaissance d'entités	*spaCy*, *Transformers*
Base de Connaissances	Stockage et recherche	*SQLite*, *Whoosh*
Interface Utilisateur	Interface web/GUI	*Flask*, *FastAPI*, *tkinter*

5. Critères de Qualité et Tests

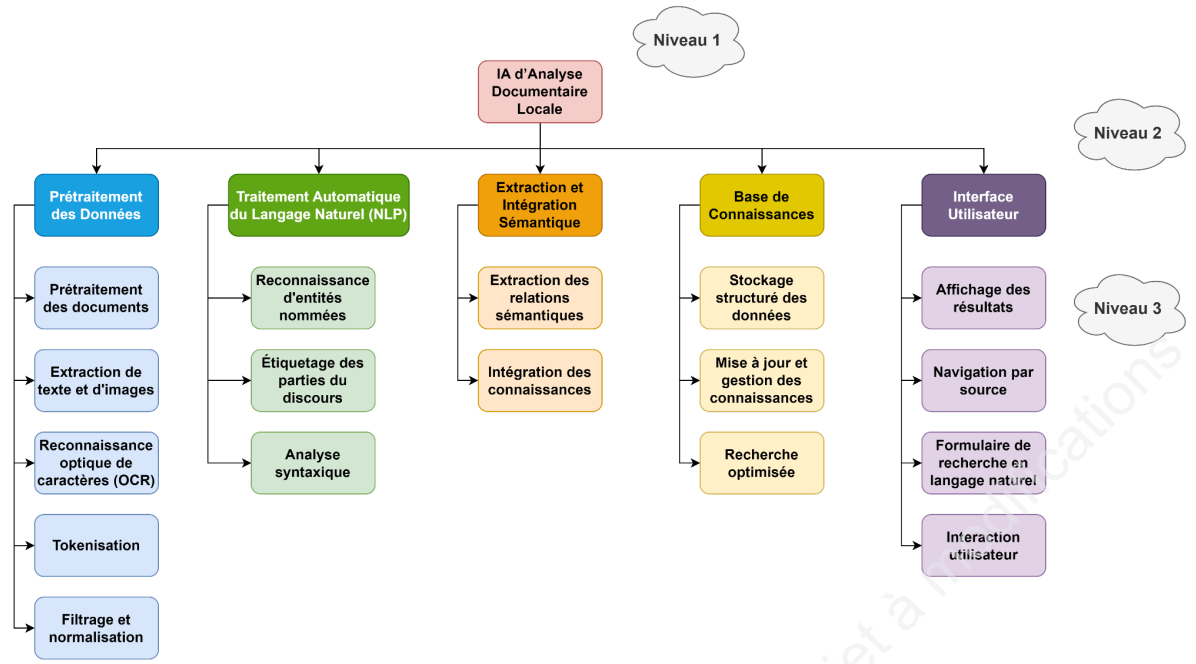
Tests de Qualité

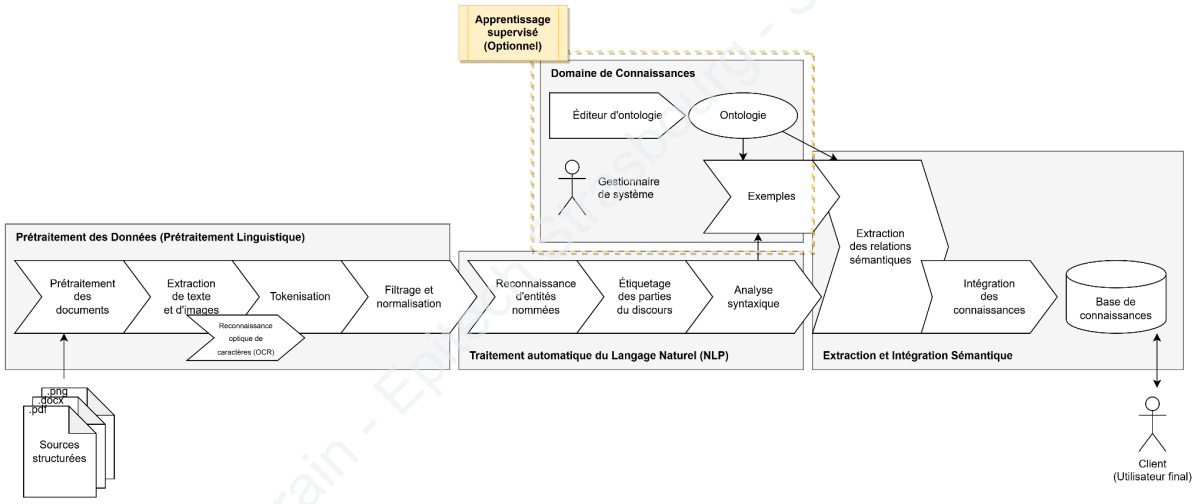
- **Tests de Prétraitement :** Vérifier que les informations sont bien extraites des documents.
- **Tests NLP :** S'assurer que l'IA comprend les questions.
- **Tests de Synthèse :** Vérifier que les résumés sont clairs et précis.
- **Tests d'Indexation :** Contrôler la rapidité de recherche dans les documents.

Indicateurs Clés de Performance (KPIs)

- **Précision d'Extraction :** 98 % pour les documents standards.
- **Temps de Réponse des Recherches :** Moins de 5 secondes.
- **Précision du NLP :** 80 % pour les questions simples.

6. Diagramme de Flux des Modules





Les modules sont connectés en chaîne pour que chaque étape se fasse dans le bon ordre, permettant une recherche efficace et précise.

7. Documentation et Support Technique

Documentation Technique

- **Guide d'installation :** Instructions pour installer les outils nécessaires.
- **Guide d'utilisation :** Explications pour utiliser l'interface.

Support et Maintenance

- **Suivi des Erreurs :** Utilisation de loguru pour surveiller les erreurs et améliorer le système.
- **Mises à jour :** Prévoir des améliorations pour ajouter de nouvelles fonctionnalités ou ajuster la précision.