Introduction à l'IA Générative
Projet 2 : Assistant d'Analyse de Documents
2024 - 2025

Issam Falih

Le projet **"Assistant d'Analyse de Documents"** vise à développer un système d'IA capable d'extraire et de traiter du texte à partir de divers formats de documents, notamment des images, des fichiers PDF et des fichiers texte structurés. Le texte extrait sera utilisé pour construire un assistant interactif capable de **résumer des documents, répondre aux questions des utilisateurs et mettre en évidence les informations clés**. Ce projet intègre des techniques de **Reconnaissance Optique de Caractères (OCR), de Traitement du Langage Naturel (NLP) et d'IA Générative**.

1 Objectifs

- Extraire du texte à partir de divers formats de documents (ex. : images, PDF, documents numérisés) en utilisant des outils OCR et de parsing.
- Traiter et structurer le texte extrait pour en faire une base de connaissances interrogeable et consultable.
- Développer un assistant IA capable de :
  - Répondre aux questions des utilisateurs sur la base du contenu extrait.
  - Résumer des documents volumineux.
- Fournir une interface conviviale pour le téléversement de documents et l'interaction avec l'assistant (optionnel).

2 Étapes de mise en œuvre

2.1 Extraction de texte

- Extraire du texte depuis des fichiers PDF en utilisant des bibliothèques comme **PyPDF2** ou **pdfplumber**.
- Effectuer une **OCR** sur les images et les documents numérisés.

- Gérer les cas complexes, comme les **textes manuscrits** ou les **données tabulaires**, grâce à des techniques avancées d'OCR.

2.2 Prétraitement du texte

- Nettoyer et prétraiter le texte extrait (ex. : suppression du bruit, tokenisation).
- Segmenter le texte en **composants logiques** tels que **titres, paragraphes et tableaux**.
- Convertir le texte nettoyé en **embeddings** en utilisant des outils comme **sentence-transformers**.
- Indexer les embeddings avec un **moteur de recherche sémantique** (ex. : FAISS).

2.3 Fonctionnalités de l'assistant

- Fournir des réponses précises aux questions des utilisateurs.
- Générer des **résumés de documents**.
- Extraire dynamiquement **les informations clés**.

2.4 Interface utilisateur

- Développer une **interface web** en utilisant des frameworks comme **Streamlit**.
- Permettre aux utilisateurs de **téléverser des documents, poser des questions et interagir avec l'assistant**.

3 Extensions optionnelles

- **Support multilingue** : permettre l'extraction de texte et la réponse aux questions dans plusieurs langues.
- **Intégration vocale** : permettre aux utilisateurs d'interagir avec l'assistant via des commandes vocales.
- **Mécanisme de feedback** : implémenter un système de retour utilisateur pour affiner les réponses de l'assistant au fil du temps.