T-ESP-700

Big Brain

Ça manque de contexte et/ou d'information ?

Regarde les autres documents

Lucas FIXARI
Pierre ROCHETTE
William WOZIWODA

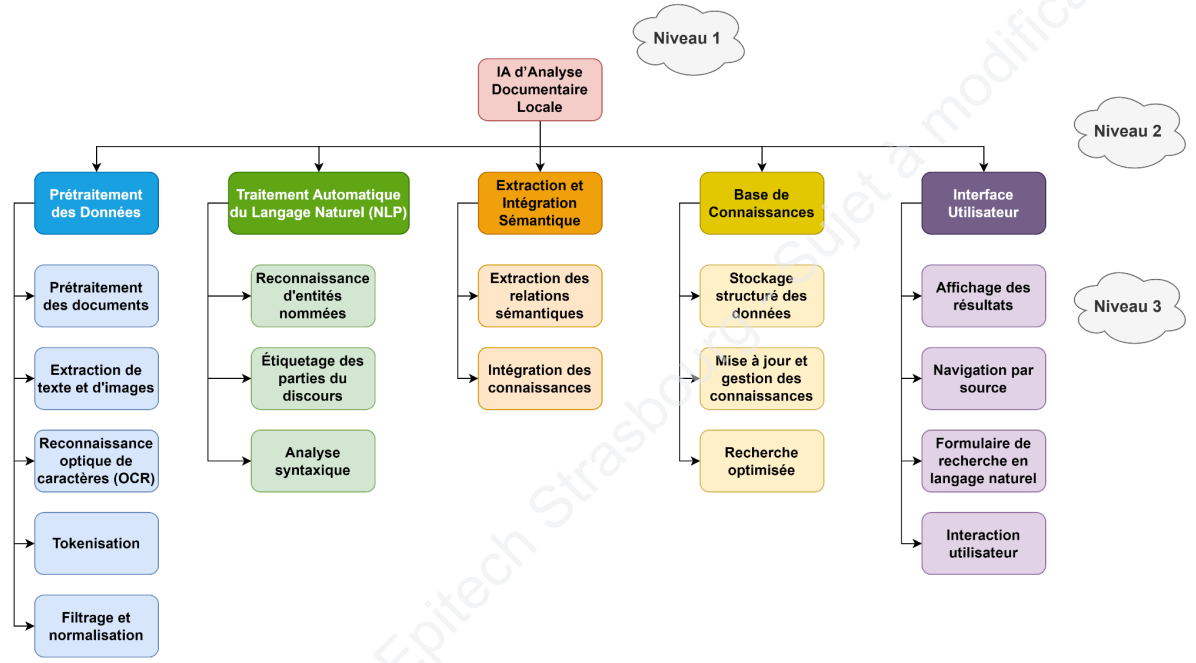
Présentation de l’IA d’Analyse Documentaire

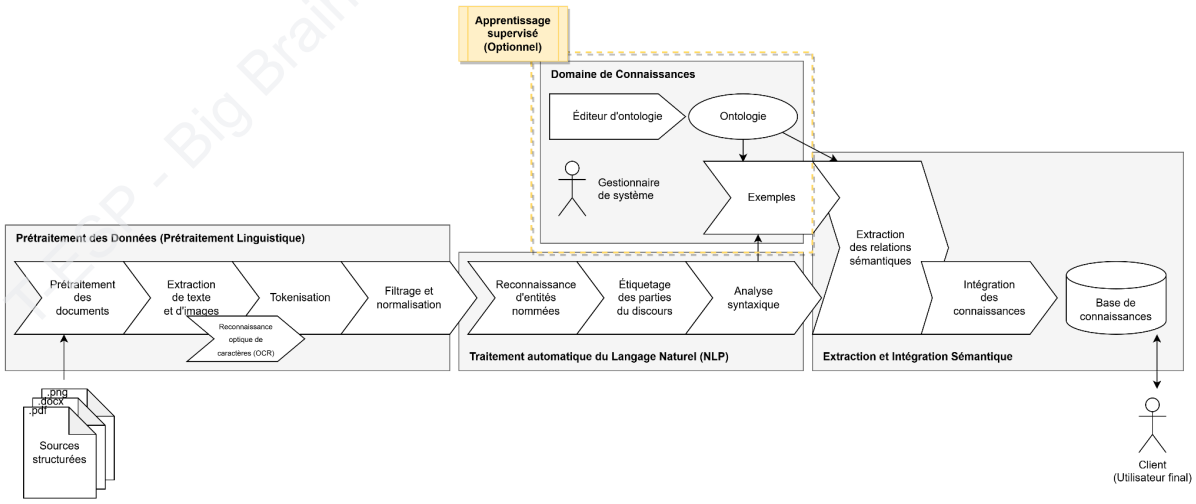
1. Description du Projet

Titre : IA d'Analyse Documentaire Locale pour la Recherche Contextuelle en Langage Naturel

Description : Une IA locale capable d'analyser, indexer et rechercher des informations à partir de divers documents. Conçue pour assurer la confidentialité, elle permet une recherche intuitive en langage naturel et fournit des réponses contextualisées avec les sources documentaires.

2. Structure de Décomposition des Produits (PBS)





Niveau 1 : Solution IA d'Analyse Documentaire Locale

- **Objectif** : Développer une solution d'analyse documentaire sécurisée et locale, offrant une recherche contextuelle rapide et adaptable à tout type d'environnement.

Niveau 2 : Modules Principaux

1. **Prétraitement des Données**
    - **Objectif** : Préparer les documents pour les étapes de traitement NLP et d'indexation en assurant un format de données homogène.
2. **Traitement Automatique du Langage Naturel (NLP)**
    - **Objectif** : Comprendre et analyser le contenu des documents de manière contextuelle, en extrayant les entités et relations sémantiques.
3. **Extraction et Intégration Sémantique**
    - **Objectif** : Extraire les relations sémantiques et intégrer les informations extraites dans une base de connaissances structurée.
4. **Base de Connaissances**
    - **Objectif** : Stocker et organiser les informations pour une recherche rapide et fiable.
5. **Interface Utilisateur**
    - **Objectif** : Fournir une interface interactive pour permettre à l'utilisateur de rechercher et d'afficher les résultats de manière intuitive.

Niveau 3 : Tâches Spécifiques

1. **Prétraitement des Données**
    - **Prétraitement des documents** : Nettoyage de texte, suppression de caractères spéciaux, conversion en texte brut, détection de la langue.
    - **Extraction de texte et d'images** : Utilisation de bibliothèques pour extraire le texte et les images des documents (PDF, DOCX, et XLSX).
    - **Reconnaissance optique de caractères (OCR)** : Utilisation de l'OCR (Tesseract) pour extraire le texte des images (PNG, JPG) scannées.
    - **Tokenisation** : Division du texte en unités analytiques (tokens) pour faciliter l'analyse NLP.
    - **Filtrage et normalisation** : Suppression des stop words et lemmatisation pour normaliser le texte.
2. **Traitement Automatique du Langage Naturel (NLP)**
    - **Reconnaissance d'entités nommées** : Identification des entités clés comme les noms, dates, lieux, etc.
    - **Étiquetage des parties du discours (POS Tagging)** : Analyse grammaticale pour identifier la fonction de chaque mot dans la phrase.
    - **Analyse syntaxique** : Détermination de la structure des phrases pour comprendre les relations entre mots.
3. **Extraction et Intégration Sémantique**
    - **Extraction des relations sémantiques** : Identification des relations entre les entités pour offrir des réponses contextuelles précises.
    - **Intégration des connaissances** : Structuration et stockage des relations extraites dans un format indexé pour faciliter la recherche.

4. **Base de Connaissances**
    - **Stockage structuré des données** : Utilisation de bases de données comme SQLite pour stocker les informations extraites et indexées.
    - **Mise à jour et gestion des connaissances** : Mécanisme de mise à jour des données et gestion des erreurs pour garantir une base de connaissances fiable et à jour.
    - **Recherche optimisée** : Indexation des mots-clés et métadonnées pour des recherches efficaces et rapides.
5. **Interface Utilisateur**
    - **Affichage des résultats** : Interface de visualisation des réponses, incluant les résumés et les extraits pertinents.
    - **Navigation par source** : Possibilité pour l'utilisateur de consulter les documents sources et d'explorer par thème ou type de document.
    - **Formulaire de recherche en langage naturel** : Interface permettant de poser des questions en langage naturel.
    - **Interaction utilisateur** : Interaction en temps réel pour des recherches rapides et une restitution de résultats intuitive.

4. Vidéo de Présentation

Pour une présentation du projet et de ses objectifs, consultez notre vidéo d'interview avec le Community Manager sur YouTube : https://youtu.be/6dElN6iPzHc