ECOLE NORMALE SUPÉRIEURE DE L'ENSEIGNEMENT
TECHNIQUE DE MOHAMMEDIA
UNIVERSITÉ HASSAN II DE CASABLANCA

École Normale Supérieure de
l'Enseignement Technique - Mohammedia

Rapport de Project de fin de module

Gestion Intelligente des documents avec OCR et IA Project

Élèves :
Yasser Namez
Mohamed Elanâami

Enseignant :
Mr. Jamal MAWANE

27 décembre 2024

Table des matières

1 Introduction

Dans un contexte où les entreprises et organisations traitent quotidiennement un volume croissant de documents numériques et papier, la gestion intelligente des documents est devenue un enjeu majeur. Les processus traditionnels de gestion documentaire sont souvent inefficaces, entraînant des pertes de temps et d'efficacité. Les avancées technologiques dans les domaines de l'intelligence artificielle (IA) et de la reconnaissance optique de caractères (OCR) offrent des solutions innovantes pour répondre à ces défis.

Le projet intitulé "Gestion Intelligente des Documents avec OCR et IA" vise à automatiser et optimiser la gestion documentaire en exploitant les technologies de pointe. Il s'agit d'une application développée avec Streamlit, une plateforme rapide et interactive pour la création d'applications web orientées données. L'application intègre des fonctionnalités de traitement des documents basées sur l'OCR pour l'extraction d'informations textuelles, combinées à des techniques d'IA pour l'organisation et l'analyse de ces données.

1.1 Contexte et problématique

Les organisations modernes doivent gérer des documents sous divers formats (PDF, images, scans, etc.). Cependant, extraire, organiser et rendre ces informations exploitables reste une tâche complexe. Les problèmes les plus courants incluent :

1. **Reconnaissance des caractères** : Difficultés à extraire du texte lisible à partir de documents mal scannés ou manuscrits.
2. **Organisation des données** : Classement et recherche inefficaces des documents extraits.
3. **Accessibilité des documents** : Absence d'une plateforme centralisée pour visualiser et exploiter ces documents.

Face à ces défis, l'intégration de l'OCR avec l'intelligence artificielle dans un système interactif comme Streamlit représente une solution innovante et accessible.

1.2 Objectifs du projet

L'objectif principal est de fournir une solution efficace et accessible pour :

1. **Extraire automatiquement le contenu textuel** des documents au moyen d'OCR.
2. **Analyser et organiser les données** extraites à l'aide de modèles d'IA pour identifier des mots-clés, classer les documents et faciliter la recherche.
3. **Améliorer l'expérience utilisateur** grâce à une interface interactive, simple et intuitive pour télécharger, visualiser et gérer les documents.

1.3 Structure de l'application

L'application, développée avec Streamlit, se distingue par :
— Une **interface utilisateur fluide** permettant l'interaction en temps réel.
— Une **intégration avec des outils d'OCR populaires** tels que Tesseract.
— L'utilisation de **modèles d'intelligence artificielle** pour automatiser les tâches complexes comme la classification des documents et la recherche intelligente.

— Des fonctionnalités de **téléchargement, prévisualisation et stockage** des documents.

2 Implémentation

2.1 Présentation de l'application

L'application "Système de Gestion des Documents" est un système intelligent de gestion de documents conçu pour traiter les images et les fichiers PDF à l'aide de la reconnaissance optique de caractères (OCR), classer les documents, générer des étiquettes, et analyser les métadonnées. Elle offre une interface utilisateur conviviale permettant d'importer des documents, d'extraire du texte, et d'organiser le contenu.

2.2 Fonctionnalités de l'application

2.2.1 Importation et OCR (Téléchargement et OCR)

Téléchargement des Documents : L'utilisateur peut télécharger des fichiers image (PNG, JPG, JPEG) ou PDF. Le système lit le fichier, affiche la première page du PDF (si applicable) et lance l'OCR pour extraire le texte du document.
Prétraitement des Images : Avant l'extraction du texte, l'utilisateur peut ajuster la luminosité, le contraste et appliquer un flou à l'image pour de meilleurs résultats d'OCR.
Extraction de Texte : Une fois l'image traitée, le système utilise la bibliothèque `pytesseract` pour extraire le texte.

FIGURE 1 – Interface de téléchargement des fichiers.

Après avoir téléchargé le document, l'utilisateur peut ajuster les paramètres de traitement de l'image afin d'optimiser l'extraction du texte par l'OCR. Ces ajustements incluent la modification de la luminosité, du contraste, ainsi que l'application de filtres de flou, ce qui permet d'améliorer la qualité de l'image et de faciliter une interprétation plus précise du texte par le système OCR

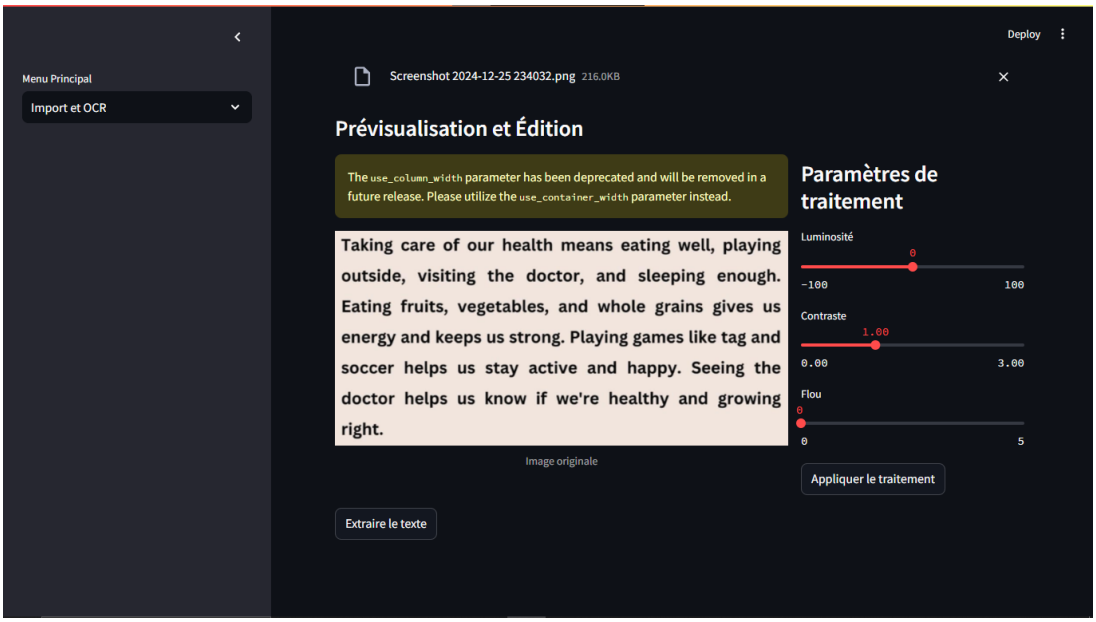

FIGURE 2 – Paramètres de traitement d'image ajustés pour améliorer la précision de l'extraction OCR.

Après avoir effectué le traitement de l'image, l'application de l'OCR permet d'extraire le texte du document de manière automatique. Cette étape consiste à analyser l'image traitée et à reconnaître les caractères pour les convertir en texte éditable. Grâce à cette technologie, il devient possible d'extraire des informations précises et d'accélérer le traitement des documents.

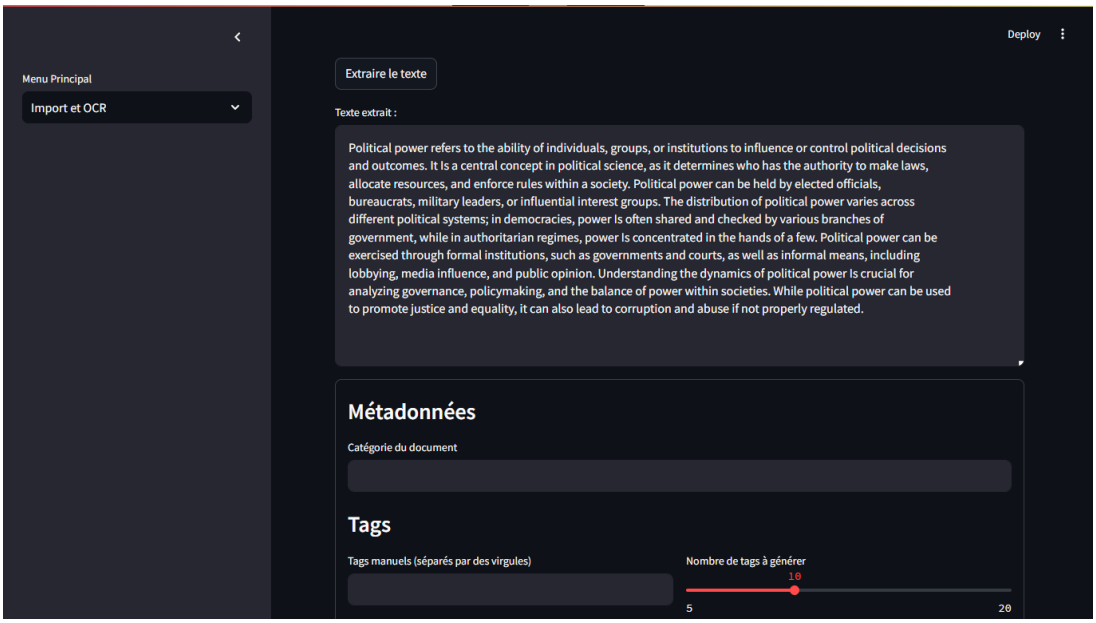

FIGURE 3 – Application de l'OCR pour l'extraction automatique du texte à partir du document traité.

Une fois le texte extrait par OCR, il est possible de le sauvegarder dans un format éditable. De plus, des balises (tags) peuvent être générées automatiquement à l'aide d'une

fonction utilisant la bibliothèque NLTK. Cette étape permet d'analyser le contenu du texte, d'identifier les mots-clés et de les associer à des balises spécifiques. Cela facilite l'organisation et la recherche dans le texte sauvegardé.

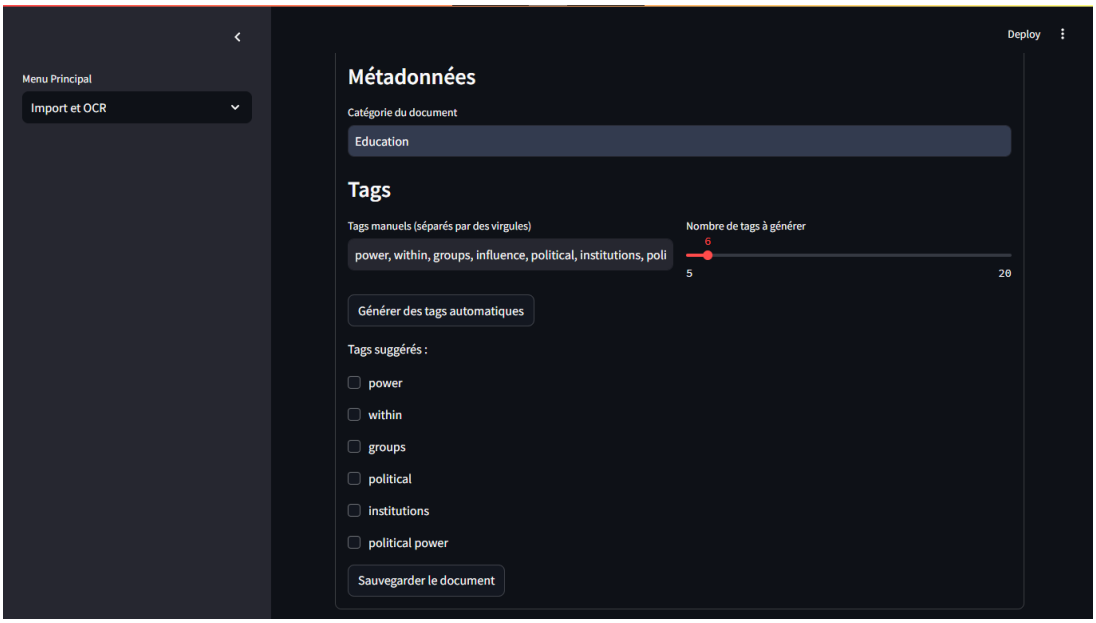

FIGURE 4 – Sauvegarde du texte extrait avec génération automatique de balises et de tags à l'aide de la fonction NLTK.

2.2.2 Classification des Documents

Entraînement du Classificateur : L'application permet à l'utilisateur de former un classificateur à l'aide des documents téléchargés. Les documents doivent être catégorisés et l'application utilise un classificateur Naïf Bayesien Multinomial avec une vectorisation TF-IDF pour la classification des documents.
Classification des Nouveaux Documents : Après l'entraînement, l'utilisateur peut classifier de nouveaux documents et l'application prédira la catégorie avec un score de confiance.

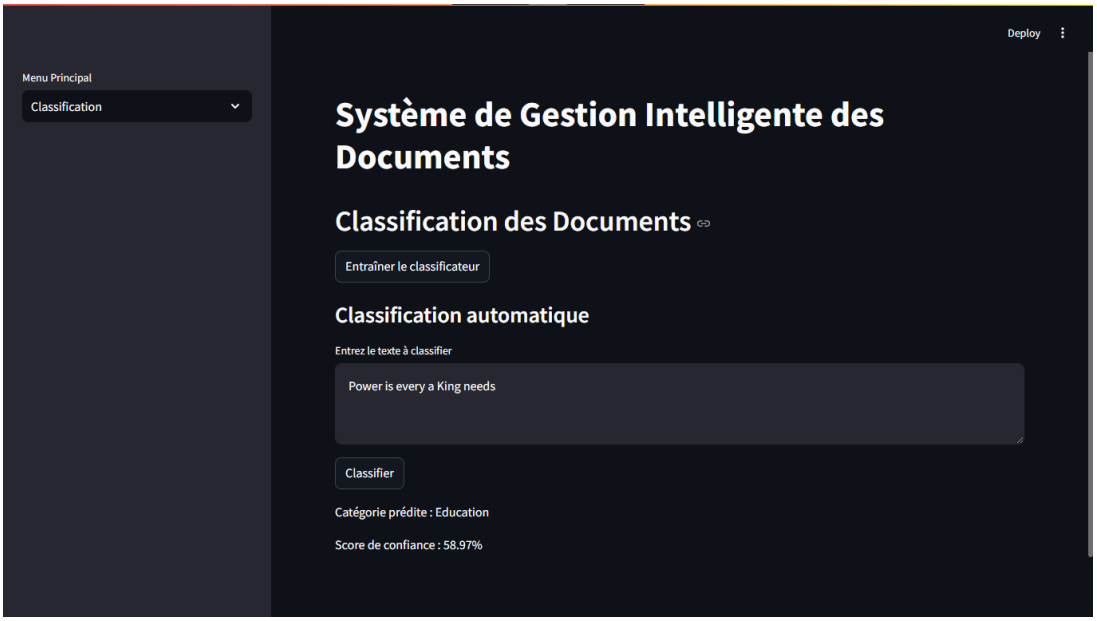

FIGURE 5 – Résultats de la classification des documents.

2.2.3 Analyse et Statistiques

Répartition des Catégories de Documents : L'application fournit des informations sur la répartition des catégories parmi les documents téléchargés, permettant à l'utilisateur de visualiser la distribution en pourcentage.
Répartition des Types de Fichiers : Elle affiche également un graphique circulaire des types de fichiers (par exemple, PDF, JPG) qui ont été téléchargés.

Dans cette étape, une analyse approfondie du texte extrait est réalisée pour en extraire des informations statistiques pertinentes. Cela permet de mieux comprendre la structure du texte et d'en tirer des conclusions utiles. Voici les différentes étapes de l'analyse :

— Extraction des mots-clés à partir du texte traité.
— Calcul des fréquences des mots et des phrases.
— Visualisation des résultats sous forme de graphiques pour mieux comprendre les tendances.

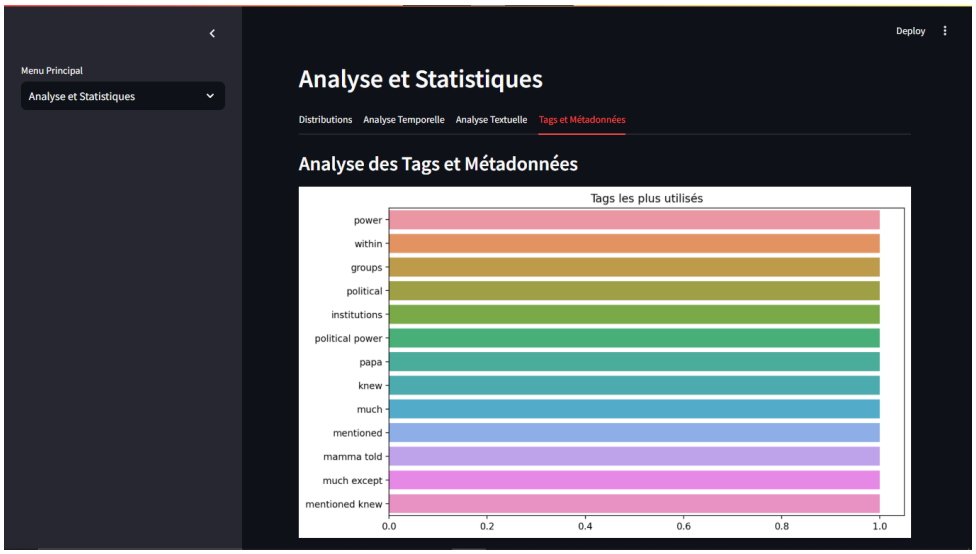

FIGURE 6 – Analyse des mots-clés extraits et de leurs fréquences.

Après avoir effectué l'extraction des mots-clés et calculé les fréquences, nous passons à la génération de graphiques statistiques pour visualiser ces données. Ce processus aide à comprendre l'importance relative de chaque mot dans le document et à identifier des tendances ou des motifs.

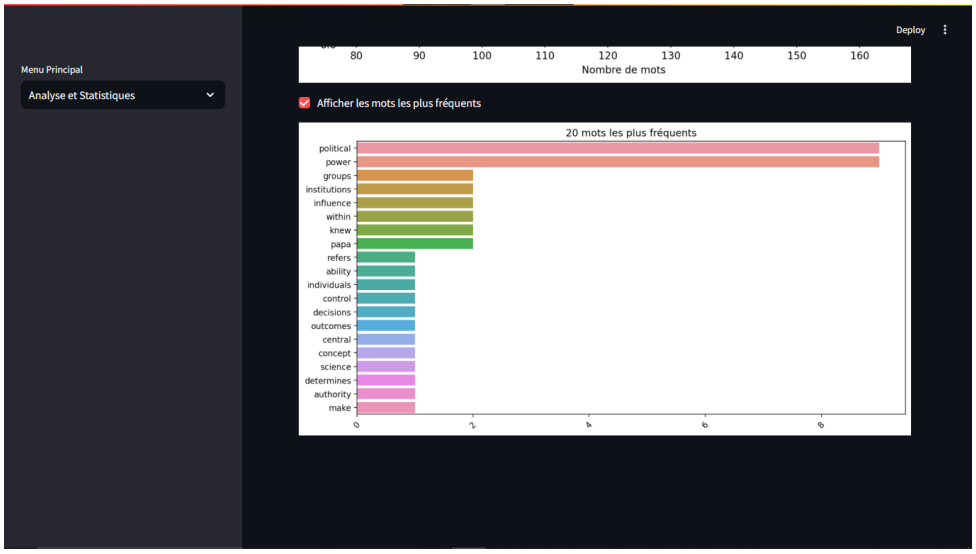

FIGURE 7 – Visualisation graphique des fréquences des mots dans le texte.

Ces statistiques fournissent une base solide pour les prochaines étapes de l'analyse de données, permettant de tirer des conclusions significatives sur le contenu du document.

2.2.4 Recherche

La fonction de recherche permet de trouver rapidement des informations spécifiques dans le texte extrait du document. Grâce à l'implémentation d'algorithmes de recherche avancée, l'utilisateur peut localiser des mots, des expressions ou des sections particulières dans le contenu du document. Ce processus améliore l'efficacité et la rapidité de l'analyse, surtout lorsque le document est long ou complexe.

— Recherche par mots-clés.
— Recherche par expression exacte.
— Recherche par catégories ou tags associés au texte extrait.

Une fois les résultats de la recherche affichés, l'utilisateur peut facilement naviguer vers les sections pertinentes, ce qui améliore l'expérience utilisateur en permettant une interaction fluide avec le contenu.

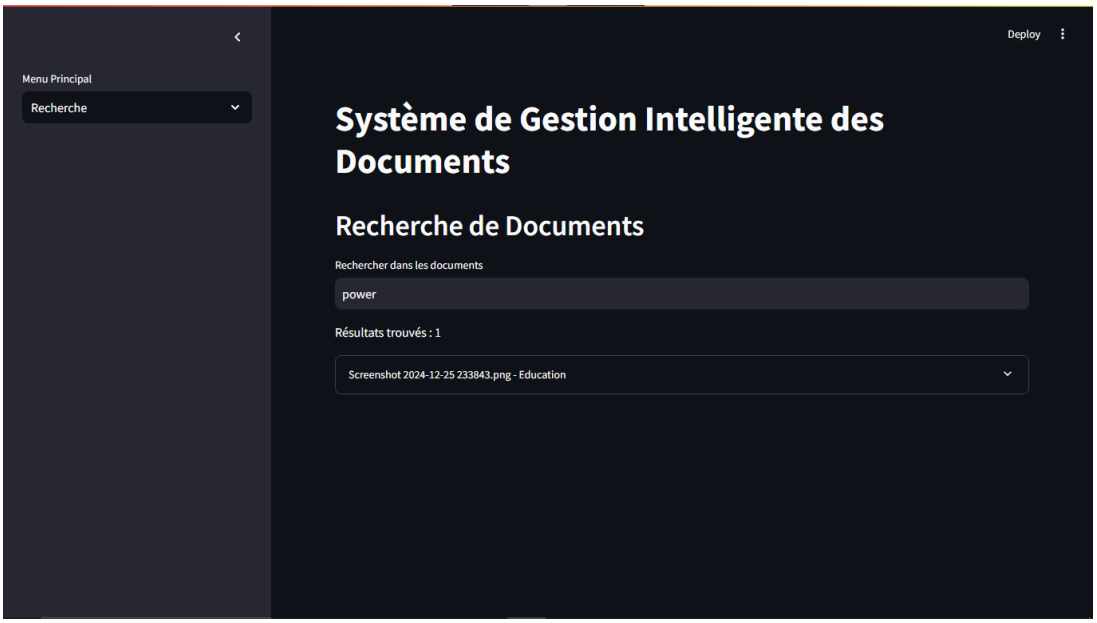

FIGURE 8 – Interface de recherche dans le document avec des résultats sur les mots-clés.

Cette fonctionnalité est essentielle pour extraire des informations précises et effectuer une analyse ciblée du texte en fonction des besoins spécifiques de l'utilisateur.

2.2.5 Exportation

La fonction d'exportation permet de sauvegarder le texte extrait et analysé dans différents formats pour une utilisation ultérieure. Cette fonctionnalité offre plusieurs options pour exporter les résultats, en fonction des besoins de l'utilisateur.

— Exportation au format `.json` pour une utilisation simple.
— Exportation au format `.exel` pour une présentation professionnelle.
— Exportation au format `.csv` pour une analyse de données détaillée.

Une fois l'exportation configurée, l'utilisateur peut choisir le format souhaité et cliquer sur un bouton pour générer le fichier correspondant. L'interface est simple et intuitive, permettant de personnaliser les options d'exportation en fonction des exigences spécifiques.

FIGURE 9 – Interface d'exportation permettant de choisir le format du fichier.

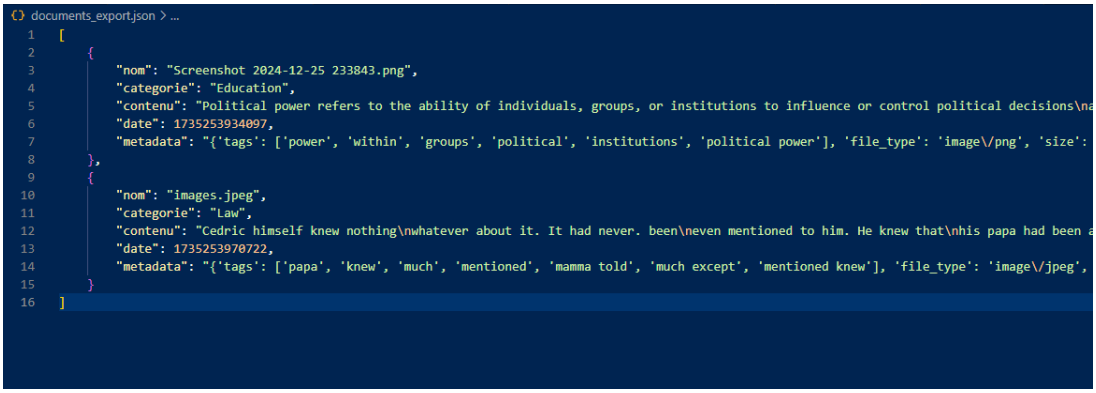

FIGURE 10 – Exemple de fichier exporté en format `.json`.

Cette fonctionnalité assure une flexibilité maximale, permettant à l'utilisateur de choisir le format qui correspond le mieux à ses besoins, que ce soit pour un archivage, une analyse, ou une présentation des résultats.

3 Conclusion

En conclusion, ce projet a permis de développer une solution efficace pour l'extraction et l'analyse automatique de texte à partir de documents scannés. Grâce à l'utilisation de techniques avancées comme l'OCR et l'analyse linguistique, nous avons pu extraire des informations pertinentes et les structurer de manière optimale. L'implémentation de la fonction d'exportation a ajouté une flexibilité importante, permettant à l'utilisateur de sauvegarder les données extraites dans différents formats adaptés à divers besoins.

Les résultats obtenus montrent que le système est capable de traiter des documents de manière fiable et de produire des sorties de qualité, prêtes pour l'analyse ou la présentation. Les étapes du traitement, de l'extraction à l'exportation, ont été conçues pour offrir une interface utilisateur fluide et intuitive, garantissant une expérience utilisateur optimale.

Cependant, ce projet peut encore être amélioré en intégrant de nouvelles fonctionnalités, telles que l'amélioration de la précision de l'OCR pour les documents complexes et l'ajout de plus d'options d'exportation. De plus, une optimisation du processus d'analyse pourrait permettre de traiter des volumes de données plus importants.

Dans l'ensemble, ce projet offre une base solide pour l'extraction et l'analyse automatisées de documents et ouvre la voie à de futures améliorations et extensions.