Anonymisation de documents

Jihad FALFOUL

Travail de Bachelor 2023

Informatique et systèmes de communication - Ingénierie des données

Professeur: Marina NINOSLAV

Expert: Marina NINOSLAV

Description

Le projet d'anonymisation de documents vise à protéger les informations sensibles contenues dans les documents appartenant à des entreprises. Ces informations peuvent inclure des données personnelles, des données financières ou des données commerciales. Une fois anonymisés, les documents peuvent être partagés avec le client afin de permettre une analyse et de produire un rapport RSE. Le besoin commercial du projet est motivé par la nécessité pour les entreprises de pouvoir partager des documents avec leurs clients sans risquer de divulguer des informations sensibles.

Le problème à résoudre est que les documents peuvent être de différents types (PDF, XLS, images, texte, etc.), et que des informations sensibles peuvent être présentes n'importe où dans le document

Déroulement

- Analyse du problème. Lecture des différents papiers concernant la thématique.
- Mise en place de la stratégie afin de répondre aux besoins identifiés durant l'analyse.
- Développement d'un module Python permettant le traitement des fichiers.
- Développement d'un modèle de classification.
- Développement de l'API permettant l'accès au module d'anonymisation.
- Développement d'une interface utilisateur pour rendre l'API accessible.

Résultats

Les trois modules qui constituent cette application sont fonctionnels et faciles à améliorer. De nouvelles fonctionnalités peuvent être ajoutées facilement à tous les niveaux. Ceci permet une grande flexibilité de l'application qui n'est pas publiable pour l'instant, mais qui constitue une base très solide afin de mettre en place des fonctionnalités encore plus avancées.

Document anonymisé

Document non traité

Discussion : Conclusions et perspectives

Cette solution offre une bonne base qui permet une évolution rapide des différentes parties de l'application.
Chaque module est indépendant et a la capacité de tourner seul.
Ils doivent cependant correspondre aux standards de la solution afin qu'ils puissent être utilisés ensemble.