Moteur de détection automatique de fuites de données

Mise en Situation Professionnelle

Entreprise : Groupe Asten

Étudiante : ANGEA RENZA Keman Ngogang

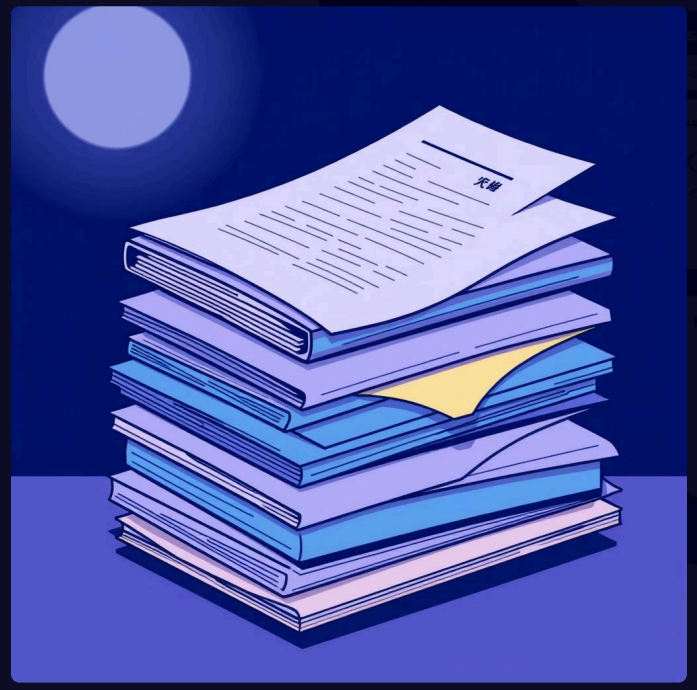
École : ECE Paris

Contexte et Enjeux

Documents Sensibles du Groupe Asten

Le Groupe Asten traite quotidiennement des documents hautement sensibles, incluant :

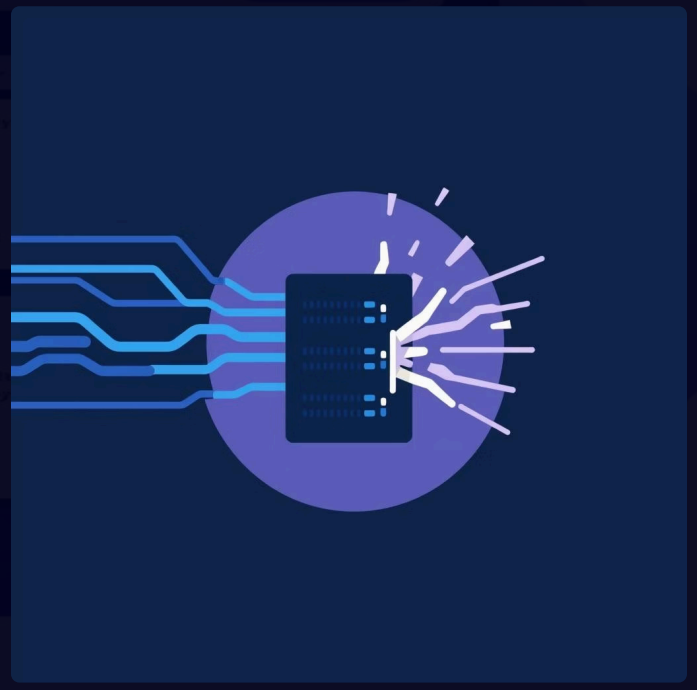
- Rapports internes confidentiels
- Données des Ressources Humaines (RH)
- Informations clients stratégiques



Risques Identifiés

La gestion de ces données présente des risques majeurs :

- **Perte de confidentialité :** Divulgation non autorisée d'informations.
- **Non-conformité réglementaire :** Infractions au RGPD et autres lois.
- **Atteinte à l'image de l'entreprise :** Dommage à la réputation et à la confiance.



Besoin exprimé : Un outil capable de détecter automatiquement les données sensibles avant leur diffusion.

Objectifs du Projet

Ce projet vise à créer une application robuste et intuitive pour la détection des fuites de données.

Analyse de Fichiers

Analyser les fichiers **TXT** et **PDF** pour extraire leur contenu.

Identification Sensible

Identifier les informations sensibles (IBAN, carte bancaire, email, téléphone, etc.).

Calcul de Score

Calculer un score de risque basé sur la nature et la quantité des données détectées.

Décision Automatique

Décider d'une action appropriée : **LOG / WARN / ALERT / BLOCK**.

Interface Utilisateur

Proposer une interface web intuitive et un rapport PDF exportable.

Choix Techniques

Une sélection rigoureuse des technologies a été effectuée pour garantir performance et maintenabilité.

Langage

Python 3.12 : Pour sa polyvalence et son écosystème riche.

Framework Web

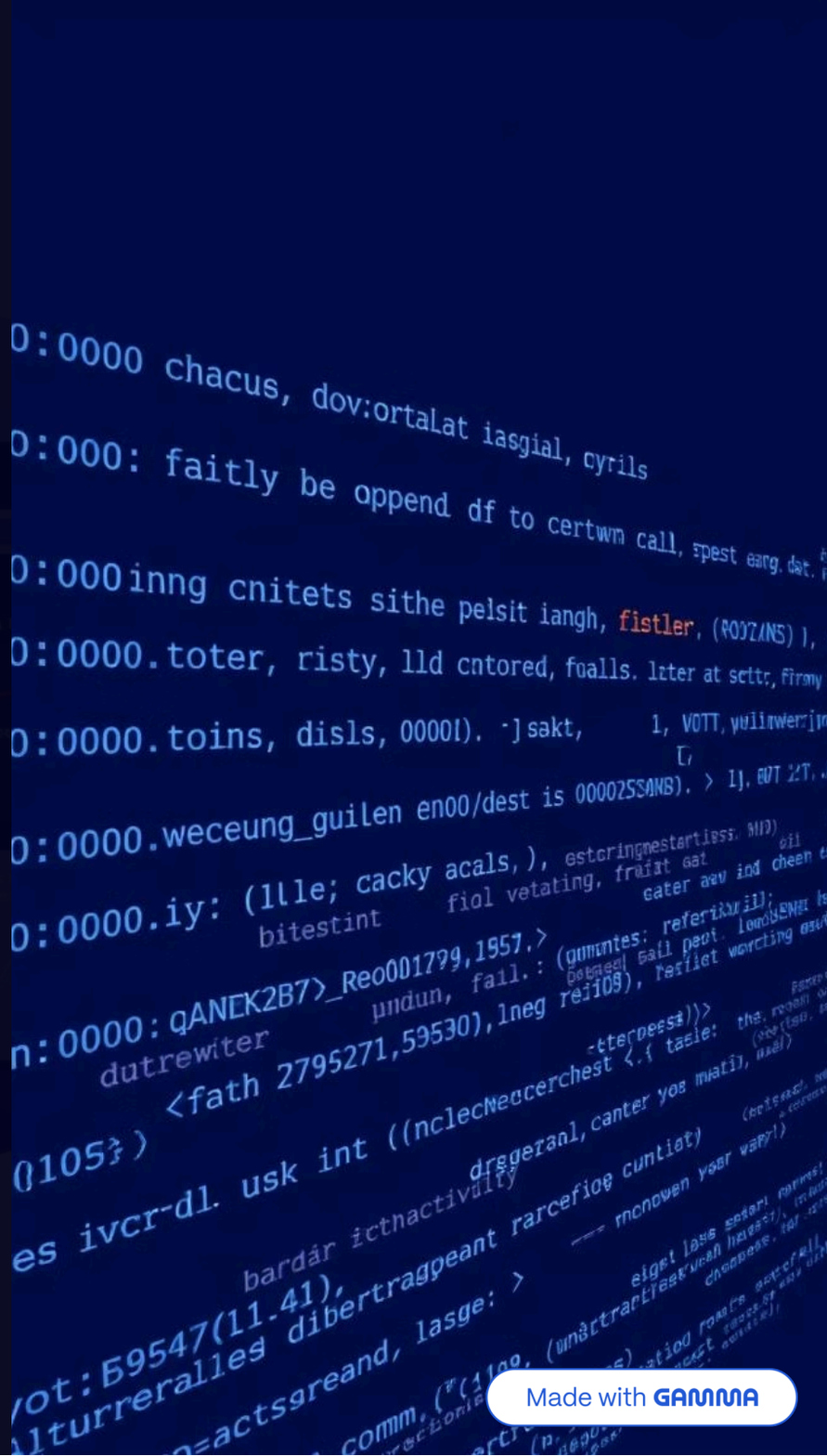
FastAPI : Rapide, moderne, avec auto-documentation via Swagger.

Librairies Clés

- **pdfplumber, pdfminer.six :** Extraction texte PDF
- **regex :** Détection par motifs
- **reportlab :** Génération de PDF
- **pytest :** Tests unitaires

Outils de Développement

- **GitHub Codespaces :** Environnement de développement cloud
- **Git :** Gestion de projet et versionnement



Architecture : Schéma du Flux

Le processus de détection des fuites de données suit un flux logique et structuré.

Upload Fichier

L'utilisateur soumet un fichier (TXT ou PDF).

Extraction Texte

Le contenu textuel est extrait du fichier.

Détection

Les motifs de données sensibles sont identifiés.

Scoring

Un score de risque est calculé.

Résultats

Affichage via l'interface utilisateur et génération de rapport PDF.

Architecture : Organisation des Modules

L'application est organisée en modules distincts pour une meilleure modularité et maintenance.

detection/

- Règles de détection
- Moteur de détection
- Calcul du score
- Fonctionnalités de caviardage

utils/

- Extraction de texte (PDF/TXT)
- Fonctions utilitaires diverses

templates/

- Pages HTML (accueil + résultats)
- Structure des vues web

static/

- Fichiers CSS pour le style
- Scripts JavaScript pour l'interactivité

tests/

- Tests automatiques pour chaque module
- Assurance qualité du code

Fonctionnalités Réalisées

Le projet a abouti à la mise en œuvre de plusieurs fonctionnalités clés, garantissant une solution complète.

Upload de Fichiers

Prise en charge des fichiers **TXT** et **PDF**.

Détection Sensible

Identification des données : **IBAN, PAN, EMAIL, PHONE, NIR, SIRET**.

Score et Décision

Calcul d'un score de risque et décision automatique (LOG/WARN/ALERT/BLOCK).

Interface Claire

Affichage des résultats avec résumé, tableau des détections, aperçu caviardé et thermomètre de risque dynamique.

Export de Rapports

Export des analyses au format **JSON** et **PDF**.

Journalisation

Enregistrement des logs dans analysis.log.

Tests Unitaires

Tous les tests unitaires (**8/8**) ont été réussis, assurant la fiabilité.

Exemple d'Utilisation

L'utilisateur télécharge un fichier (PDF ou texte) via l'interface.

L'outil analyse le contenu :

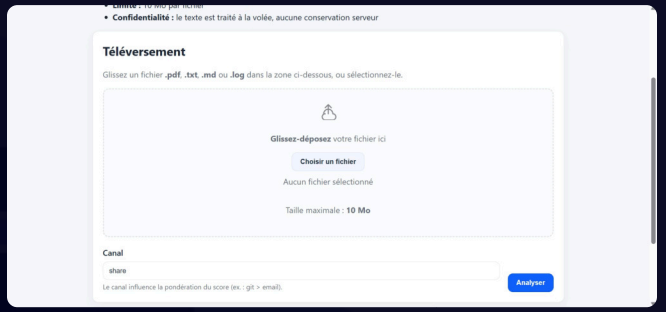
- Exemple : un IBAN, une carte bancaire et un email sont détectés.
- Un score de **21** est calculé.
- La décision automatique est : **BLOCK**.

L'interface affiche clairement :

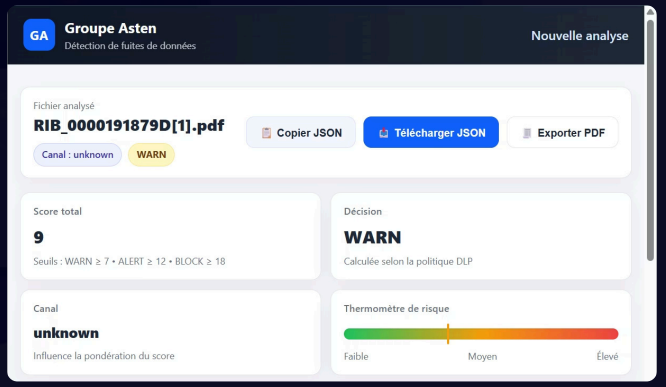
- Le tableau des détections.
- Le score et la décision prise.
- Un thermomètre de risque visuel.

L'utilisateur peut ensuite :

- Copier ou télécharger le rapport **JSON**.
- Exporter un rapport **PDF** détaillé.



(Capture d'écran de l'interface utilisateur pour le téléchargement des fichiers)



(Capture d'écran de l'interface utilisateur pour le résultat d'analyse)

Défis et Solutions

Difficultés Rencontrées ⚠️

- **Conflit de dépendances :** Entre pdfplumber et pdfminer.six.
- **Erreurs d'import et indentation :** Problèmes liés à l'environnement Python.
- **Thermomètre dynamique :** Ajustements complexes en CSS et JavaScript.
- **Rapport PDF :** Travail minutieux sur la mise en page et le caviardage.

Solutions Apportées ✅

- **Versions fixées :** Résolution des conflits de dépendances par la fixation des versions.
- **Nettoyage et PYTHONPATH :** Correction des erreurs d'import et d'indentation.
- **Optimisation UI/UX :** Ajustement précis du thermomètre dynamique.
- **ReportLab :** Utilisation de ReportLab pour un export PDF formaté et lisible.
- **Design UI :** Interface sobre et responsive, aux couleurs du Groupe Asten.
- **Journalisation :** Mise en place d'une politique de logs dans analysis.log.
- **Caviardage :** Caviardage automatique dans l'aperçu pour protéger les données.

Réflexion Critique et Conclusion

Points Forts 👍

- Projet fonctionnel de bout en bout (upload → détection → export PDF).
- Interface simple, compréhensible par un non-technique.
- Code modulaire et testé.

Limites Actuelles 🚧

- Pas de déploiement Docker ni HTTPS.
- Pas de dashboard de logs temps réel.
- Règles limitées aux données françaises (IBAN_FR, NIR…).

Pistes d'Amélioration 🚀

- Ajout d'une limite de taille d'upload (10 Mo max).
- Intégration avec un SIEM (ELK / Splunk).
- Dashboard d'administration (Grafana/Loki).
- Extension des règles (autres formats IBAN, passeport, etc.).

✔️ **Le projet a permis de développer un moteur DLP fonctionnel qui répond au besoin initial de Groupe Asten.**

✔️ **L'outil détecte les fuites de données, alerte et génère des rapports exploitables.**

✔️ La solution est prête pour une première utilisation interne, avec des pistes claires d'amélioration pour une mise en production.