**Rapport du Projet 2 - Big Data**

# Identités

* Matricule : 16MK292
* Nom, PostNom & Prénom : MUNUNGA KAWIKA Dieudonné
* Promotion : M1 Data Science
* Cours : Big Data

# Intitulé et but du projet

Intitulé Projet : Système de recommandation d'images basé sur les préférences utilisateur

L'objectif de ce projet est de concevoir un système de recommandation d'images personnalisé en fonction des préférences des utilisateurs. Le système met en œuvre plusieurs techniques de collecte, annotation, analyse, visualisation et recommandation de données image.

# Sources des données et licences

Les images utilisées dans ce projet ont été téléchargées depuis des sites libres de droits :

* Pexels : <https://www.pexels.com>
* Pixabay : <https://pixabay.com>
* Unsplash : <https://unsplash.com>

Les images respectent les licences libres fournies par ces plateformes.

# Taille des données

* Nombre total d’images téléchargées : 300
* Format des images : JPG

# Informations stockées pour chaque image

Chaque image est décrite à travers les éléments suivants :

* Nom du fichier
* Taille de l’image (largeur, hauteur)
* Orientation (portrait, paysage, carré)
* Métadonnées EXIF extraites (modèle d’appareil, date, etc.)
* Couleurs dominantes (extraites via KMeans)
* Tags (simulés aléatoirement)

Exemple de fichier généré : metadata\_annotated.json

# Informations concernant les préférences de l’utilisateur

Pour chaque utilisateur, nous avons stocké :

* Liste des images aimées
* Tags préférés (issus des images aimées)
* Orientations préférées
* Couleurs dominantes préférées

Exemple de fichier généré : user\_profiles.json

# Modèles d’exploration de données et d’apprentissage machine utilisés

* Extraction des couleurs dominantes : Algorithme KMeans (Scikit-learn)
* Simulation de profils utilisateurs : Génération aléatoire
* Système de recommandation
* Filtrage basé sur le contenu
* Similarité calculée sur les tags communs

Les modules de visualisation ont été réalisés avec Matplotlib et Seaborn.

# Auto-évaluation du travail

Le système fonctionne correctement avec :

* Collecte et annotation automatisée
* Génération correcte des métadonnées et des profils utilisateurs
* Recommandations pertinentes basées sur les préférences taguées

Limites et pistes d’amélioration :

* Enrichir le système avec des tags réels via des API de reconnaissance d’image (ex : Google Vision, Clarifai)
* Ajouter un filtrage hybride combinant couleurs, orientation et tags
* Augmenter le nombre d’utilisateurs pour un filtrage collaboratif
* Intégrer un système d'évaluation quantitative des recommandations

# Remarques concernant les séances pratiques et les exercices

Les exercices réalisés en TP ont permis de bien comprendre :

* Les techniques de manipulation des fichiers images et des métadonnées
* L’utilisation des librairies de data science (Pandas, Scikit-learn, Seaborn)
* La structuration des modules de traitement en plusieurs étapes

# Conclusion

Ce projet a permis de développer un pipeline complet allant de la collecte des données jusqu’à la recommandation personnalisée. Les outils big data et d’apprentissage machine permettent de construire des systèmes recommandateurs simples mais efficaces, même avec des jeux de données simulés.

# Bibliographie

* Documentation Scikit-learn : <https://scikit-learn.org/>
* Documentation Pillow : <https://pillow.readthedocs.io/>
* Pexels API : <https://www.pexels.com/api/>
* Wikipédia : EXIF <https://en.wikipedia.org/wiki/Exif>

# Fichiers joints

Lien GitHub du projet : (à compléter)

Sources des images : (à compléter)

Notebook de projet : Dieudonné\_Projet2\_Big\_Data.ipynb