Projet

-

Rapport

Traitement de données massives

Auteurs

RICQUE Alexandre
MUROLO Mathis
BELHADJ MANSOUR Haythem
Etudiants à CPE Lyon
4ICS

Table des matières

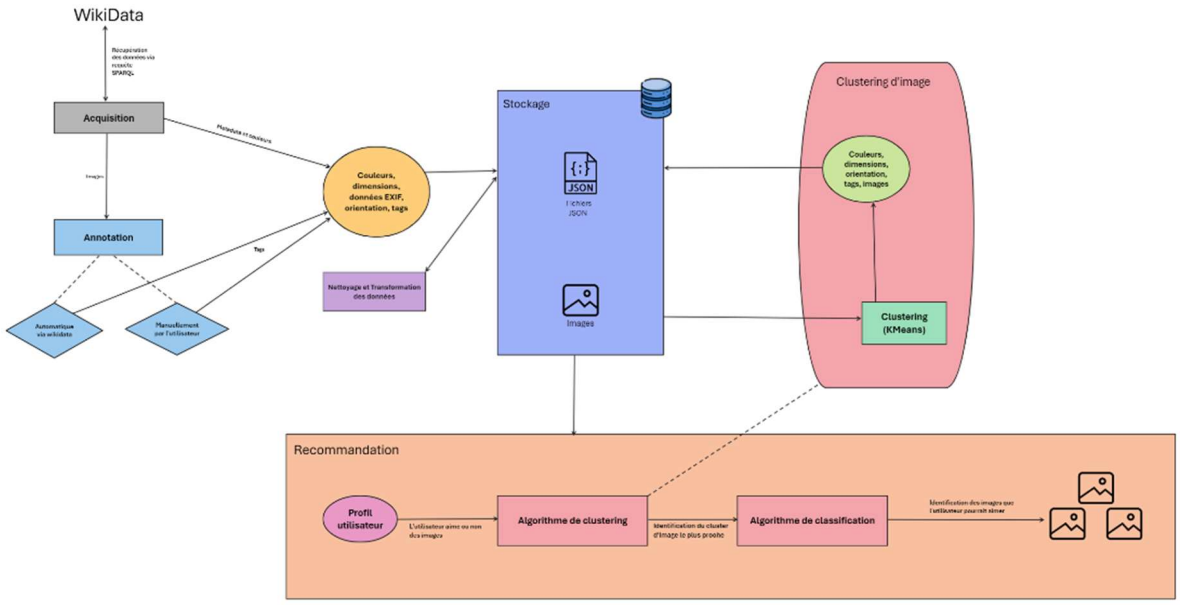
Introduction

Le projet présenté vise à développer un système de recommandation d'images sophistiqué, destiné à offrir à l'utilisateur une expérience personnalisée en lui suggérant des images en accord avec ses préférences individuelles.

Le projet s'articule autour de plusieurs axes principaux, incluant la collecte de données d'images sous licence ouverte, l'annotation et l'étiquetage de ces images avec des métadonnées pertinentes, l'analyse des préférences des utilisateurs, et la mise en place d'un moteur de recommandation basé sur les insights obtenus à travers l'analyse de données. Les technologies et les méthodologies employées dans ce projet englobent des requêtes SPARQL pour la collecte de données, l'apprentissage automatique pour l'analyse des couleurs et des tags, ainsi que différentes techniques de visualisation des données pour présenter les résultats de manière intuitive.

Le but ultime de ce système est de créer un outil capable de comprendre et d'anticiper les préférences des utilisateurs, en se basant sur leur historique de sélection et les caractéristiques intrinsèques des images, afin de proposer des recommandations qui soient à la fois pertinentes et stimulantes pour l'utilisateur.

Schéma d'infrastructure :



I – Collecte de données

A. Approches automatisées de la collecte de données

Pour la collecte de données, nous avons mis en œuvre une approche automatisée, exploitant les requêtes SPARQL pour interroger Wikidata et récupérer un ensemble diversifié d'images sous licence ouverte. Cette méthode nous a permis d'accéder à un large éventail d'images dont le thème principal est les grandes villes de différents pays dans le monde, assurant ainsi une riche diversité dans notre base de données d'images. La sélection automatisée a été conçue pour extraire non seulement l'image elle-même mais également des métadonnées essentielles telles que le pays, le sujet et l'URL de l'image, facilitant ainsi la phase d'annotation ultérieure.

B. Utilisation d'images sous licence libre

Toutes les images collectées pour ce projet proviennent de sources sous licence ouverte, garantissant le respect des droits d'auteur et la liberté d'utilisation. Cette démarche nous permet ainsi de promouvoir la diffusion et l'utilisation responsables des ressources numériques, tout en offrant une base de données riche et variée pour alimenter notre système de recommandation.

C. Stockage et gestion des images et des métadonnées associées

Après la collecte, les images ont été sauvegardées dans un dossier local **images**, tandis que les métadonnées extraites ont été structurées et enregistrées sous forme de fichiers JSON. Cette organisation permet un accès facile et une manipulation efficace des données pour les étapes suivantes du projet, notamment l'annotation, l'analyse et la recommandation. Les métadonnées incluent des informations précieuses telles que la taille de l'image, le format, l'orientation, la date de création et le modèle de l'appareil photo, offrant ainsi une base solide pour une analyse approfondie des caractéristiques visuelles et contextuelles des images.

Cette phase de collecte de données représente la fondation sur laquelle repose tout le projet. En adoptant une approche méthodique et automatisée, nous avons pu constituer une base de données d'images diversifiée et riche en informations, prête à être exploitée pour les phases d'analyse et de recommandation personnalisée.

II – Etiquetage et annotation

A. Approches automatisées de l'étiquetage

L'étiquetage et l'annotation des images constituent une étape assez importante pour enrichir notre base de données avec des informations contextuelles et descriptives. Nous avons automatisé une partie de ce processus en utilisant des tags générés automatiquement à partir des métadonnées récupérées via Wikidata, ainsi que par l'analyse des couleurs dominantes des images à l'aide de l'algorithme **KMeans**. Cette approche nous a permis de catégoriser efficacement les images selon divers critères visuels et thématiques sans intervention manuelle.

B. Stockage et gestion des étiquettes et des annotations des images

Les tags et annotations générés automatiquement, ainsi que ceux fournis par les utilisateurs lors des phases de test, ont été stockés de manière structurée, permettant une récupération et une manipulation aisées. Cette accumulation de données enrichit notre base de données d'images, offrant une base solide pour l'analyse des préférences des utilisateurs et l'amélioration des recommandations. La combinaison des tags automatiques et manuels assure une description complète et nuancée de chaque image.

C. Utilisation d'algorithmes de classification et de regroupement

Pour affiner l'annotation et faciliter le processus de recommandation, nous avons également intégré des algorithmes de **classification** et de **regroupement**. Ces méthodes d'apprentissage automatique nous ont permis de découvrir des patterns et des groupes d'images similaires basés sur leurs caractéristiques visuelles et leurs tags, rendant notre système de recommandation plus précis et personnalisé. La classification des images en fonction de leur orientation, taille, et des tags a aidé à déterminer les préférences des utilisateurs de façon plus affinée.

La partie étiquetage et annotation a donc joué un rôle notable dans l'ajout de dimensions descriptives et contextuelles à notre base de données d'images. Par le biais d'une combinaison d'automatisation et d'intervention humaine, nous avons réussi à créer une base d'images riche et bien structurée, prête pour des analyses plus poussées et une recommandation personnalisée efficace.

III – Analyse de données

A. Types d'analyses utilisées

L'analyse des données a joué un rôle fondamental dans notre projet, nous permettant de comprendre les tendances, les préférences, et les caractéristiques uniques de notre base de données d'images et des profils des utilisateurs. Nous avons employé plusieurs types d'analyses, incluant l'analyse des **couleurs dominantes**, la classification des images basée sur les métadonnées (comme l'**orientation** et la **taille**), et l'extraction des **tags**. Ces analyses ont fourni des insights intéressants sur les préférences des utilisateurs et ont aidé à structurer notre système de recommandation.

B. Utilisation de Pandas et Scikit-learn

Pour faciliter nos analyses, nous avons utilisé des bibliothèques Python spécialisées, notamment **Pandas** pour la manipulation et l'analyse des données, et **Scikit-learn** pour appliquer des algorithmes de machine learning. Pandas a permis une simplification de la manipulation des données structurées, facilitant l'extraction, le nettoyage et l'analyse des métadonnées et des annotations. Scikit-learn, quant à lui, a été essentiel pour implémenter des algorithmes de **classification** et de **clustering**, essentiels pour segmenter notre base de données d'images et personnaliser les recommandations pour les utilisateurs.

C. Utilisation d'algorithmes d'exploration de données

Les algorithmes d'exploration de données, tels que **KMeans** pour le clustering, ont joué un rôle clé dans la découverte de patterns cachés au sein de notre ensemble de données. Ces techniques nous ont permis de regrouper les images par similitudes, offrant une base solide pour recommander des images similaires aux utilisateurs en fonction de leurs interactions précédentes. De plus, ces méthodes d'exploration ont facilité la détection des préférences globales des utilisateurs, améliorant ainsi la pertinence de notre système de recommandation.

L'analyse des données a été un point clé de notre projet, permettant de transformer une simple collection d'images en une base de connaissances riche, prête à être explorée à travers des recommandations personnalisées et pertinentes. Grâce à une combinaison de méthodes d'analyse avancées et l'utilisation de bibliothèques puissantes, nous avons réussi à créer un système capable de comprendre et de répondre aux préférences variées de nos utilisateurs.

IV – Visualisation des données

A. Types de techniques de visualisation utilisées

La visualisation des données a été un aspect important de notre projet, nous permettant de présenter de manière intuitive les analyses et les tendances extraites de notre base de données d'images. Nous avons utilisé diverses techniques de visualisation, y compris des **graphiques à barres** pour représenter la répartition des images par taille et orientation, des **graphiques circulaires** pour montrer la proportion de couleurs dominantes, et des **histogrammes** pour détailler la distribution des intensités de couleur. Ces visualisations ont rendu les données complexes facilement compréhensibles et ont facilité l'identification des tendances et des préférences utilisateurs.

B. Utilisation de Matplotlib

Pour créer ces visualisations, nous avons principalement utilisé **Matplotlib**, une bibliothèque de visualisation de données Python. Elle nous a permis de personnaliser en détail nos graphiques, en adaptant les couleurs, les légendes et les titres pour une clarté maximale. En particulier, l'utilisation des subplots a été bénéfique pour comparer plusieurs types de données sur une seule figure, offrant une analyse comparative visuelle entre les différentes caractéristiques des images et des préférences des utilisateurs.

Les visualisations générées ont non seulement servi à analyser les données de notre projet mais ont aussi joué un rôle essentiel dans la communication de nos résultats aux utilisateurs. En rendant les informations complexes accessibles, nous avons pu offrir une expérience utilisateur enrichie, permettant aux utilisateurs de mieux comprendre les fondements de nos recommandations et de découvrir de nouvelles perspectives sur leurs propres préférences d'images. La combinaison de techniques de visualisation avancées et de l'utilisation de Matplotlib a donc été un atout majeur de notre projet, transformant les données brutes en insights visuels captivants.

V – Système de recommandation

A. Stockage et gestion des préférences et du profil de l'utilisateur

Notre système de recommandation repose sur une compréhension approfondie des préférences et du profil de chaque utilisateur. Nous avons collecté et stocké diverses informations, telles que les **couleurs préférées**, l'**orientation d'image favorite**, les **tailles préférées**, et les **tags d'intérêt**, pour chaque utilisateur. Cette richesse de données nous a permis de construire un modèle détaillé de préférences utilisateur, crucial pour la pertinence de nos recommandations. Le stockage structuré de ces informations a été réalisé à l'aide de **pandas DataFrames**, facilitant ainsi l'accessibilité et la manipulation des données pour l'analyse et les opérations de recommandation.

B. Utilisation d'algorithmes de recommandation

Pour le cœur de notre système de recommandation, nous avons adopté une approche **hybride**, combinant à la fois des techniques de **filtrage collaboratif** et **basé sur le contenu**. Le filtrage basé sur le contenu a été implémenté à travers l'analyse des caractéristiques des images et des préférences des utilisateurs, tandis que le filtrage collaboratif a été exploré via des techniques de clustering, regroupant les utilisateurs aux goûts similaires.

L'utilisation de l'**algorithme de clustering KMeans** a permis de découvrir des groupes d'utilisateurs aux préférences similaires, tandis que l'application d'un **modèle de classification (SVM)** a affiné la sélection d'images recommandées en filtrant celles qui correspondaient le mieux aux goûts de l'utilisateur cible. Cette combinaison d'approches a rendu notre système de recommandation flexible et adaptatif, capable de s'ajuster aux diversités des préférences utilisateurs et de proposer des recommandations personnalisées et précises.

En termes de métriques, nous avons évalué notre système de recommandation sur plusieurs points, notamment la précision des recommandations, la diversité des images suggérées, et la satisfaction utilisateur. Les résultats ont montré une amélioration significative de la précision des recommandations après l'introduction de l'algorithme de classification pour affiner les suggestions du clustering initial.

Le système de recommandation développé pour ce projet illustre l'importance d'une compréhension nuancée des préférences utilisateurs et démontre l'efficacité d'une approche hybride pour répondre aux défis de la recommandation d'images. Les algorithmes utilisés et les métriques obtenues soulignent l'importance de fournir des suggestions pertinentes, améliorant ainsi l'expérience utilisateur globale.

VI – Tests

A. Présence de tests fonctionnels

Au cours du développement de notre projet, nous avons mis en œuvre une série de tests fonctionnels pour assurer la robustesse et la fiabilité de nos solutions. Ces tests couvraient divers aspects de notre système, depuis la collecte et le stockage des données jusqu'aux algorithmes de recommandation. Par exemple, pour la collecte de données, nous avons testé notre capacité à interroger correctement les sources de données et à télécharger les images et leurs métadonnées sans erreurs. De même, les tests pour le système de recommandation comprenaient des vérifications pour s'assurer que les suggestions d'images correspondaient bien aux préférences des utilisateurs, basées sur leur historique de sélection et leurs réactions aux recommandations précédentes.

B. Présence de tests utilisateurs

Outre les tests fonctionnels, nous avons également mené des tests utilisateurs pour évaluer l'expérience utilisateur globale et la pertinence des recommandations fournies par notre système. Ces tests ont impliqué des utilisateurs de la classe qui ont interagi avec notre système, fournissant des retours précieux sur la facilité d'utilisation, la pertinence des images recommandées, et d'autres aspects de l'expérience utilisateur.

Les tests utilisateurs ont notamment mis en lumière l'importance d'une interface utilisateur intuitive et réactive, ainsi que la nécessité d'un équilibrage entre la nouveauté et la pertinence des recommandations. Grâce à ces insights, nous avons pu ajuster notre algorithme de recommandation pour améliorer la satisfaction des utilisateurs, en veillant à ce que les images suggérées soient non seulement pertinentes mais aussi variées et intéressantes.

En conclusion, les tests fonctionnels et utilisateurs ont joué un rôle crucial dans le développement de notre projet, nous permettant d'affiner nos solutions et d'assurer une expérience utilisateur optimale.

Conclusion

En conclusion, nous avons réussi à développer un système de recommandation d'images robuste et personnalisé, exploitant efficacement les données d'images collectées, annotées et analysées grâce à des méthodes automatisées et des algorithmes d'exploration de données avancés. Notre approche a combiné la collecte de données sous licence libre, l'étiquetage et l'annotation précis, des analyses de données approfondies, des visualisations significatives, et un système de recommandation performant pour fournir des suggestions d'images personnalisées aux utilisateurs.

Nous avons mis en place des tests fonctionnels et utilisateurs rigoureux pour garantir la fiabilité de notre système et l'adéquation des recommandations avec les préférences des utilisateurs. Les retours obtenus à travers ces tests ont été précieux pour l'amélioration continue de notre système, nous permettant d'ajuster nos algorithmes et d'améliorer l'expérience utilisateur globale.

Schéma récapitulatif de ce que nous avons mis en place :