Cathelain Yoann

Cereza Ewan

Rapport Projet

Data Mining et Machine Learning

Enseignants : John Samuel

Enseignants : Alexandre Saidi

Introduction

Le but de ce projet, réalisé dans le cadre de notre formation à CPE Lyon Informatique et Réseaux Communicants en majeur Informatique et Développement, est de développer un système de recommandation d'image personnalisé en utilisant des techniques d'apprentissage automatique et d'analyse de données. Ce système vise à recommander des images aux utilisateurs en fonction de différents, de leurs préférences et de leurs interactions passées avec les images. Pour atteindre cet objectif, le projet est divisé en plusieurs tâches, notamment la collecte des données, l'étiquetages et l'annotation, l'analyse de données, la visualisation de données, la construction d'un système de recommandation et enfin les tests.

Dans le cadre de notre formation, ce projet représente une opportunité précieuse d'appliquer les concepts théoriques et les compétences pratiques que nous avons acquises dans le domaine du datamining et machine-learning. En travaillant sur ce projet, nous visons à affirmer nos connaissances et à développer nos compétences en matière de système intelligents.

Le système de recommandation développé dans ce projet permettra aux utilisateurs de découvrir de nouvelles images sélectionner depuis un dataset. Ces images serviront par la suite à construire notre système basé sur des critères d'intérêts, de popularités et de préférences. Grâce à ce système, ce projet vise à améliorer l'expérience des utilisateurs lors de la navigation dans une volumétrie d'image suffisamment conséquente pour rendre ce projet utile et pratique.

Source des données

Le dataset que nous utilisons dans ce projet provient de ce playbook :

- https://www.wikidata.org/wiki/Q144

Ce playbook traîte des races de chien qui sont référencées dans cette collecte de données.

Les images quant à elles, sont toutes libre de droits pour des utilisations pédagogiques mais également professionnel.

Volumétrie de données

Notre projet ce base sur une banque de données contenant 155 éléments. Ces éléments rassemblent des images et des propriétés sur les données en question.

Traitement des données

Afin de rendre ces données plus facilement exploitables lors de la réalisation de ce projet, des scripts ont été mis en place afin d'automatiser des tâches comme la récupération des données depuis le jeu de données présent sur Wikidata et le téléchargement des images dans un dossiers adéquat directement depuis l'url de l'image présent dans chaque élément de notre dataset.

Les limites de cette phase ont été de choisir des données qui pouvaient être exploitées par la suite, nous avons choisi plusieurs banques de données qui avaient attisé notre curiosité mais le manque d'informations concernant les métadonnées nous ont poussé à changer plusieurs fois de dataset. Nous sommes parvenus à trouver un dataset qui nous permettrait facilement de réaliser, mais également d'illustrer les concepts des tâches de ce projet.

Informations des images

Pour chaque image traitée, les informations que nous avons choisi de stocker comprennent :

- Le nom du fichier de l'image pour une référence claire et unique
- La taille de l'image, exprimé sous forme de dimension (longueur, largeur, hauteur), afin de connaitre les dimensions exactes de l'image
- Le format de l'image, indiquant le type de fichier image (JPEG, PNG, etc…), ce qui peut être utile pour la compatibilité et le traitement ultérieur.
- L'orientation de l'image, déterminée en fonction de sa largeur de sa hauteur, classant l'image comment étant en mode paysage, portrait ou carré.
- Les données EXIF de l'image, fournissant des informations supplémentaires sur l'image appelés métadonnées telles que la marque de l'appareil photo, les paramètres d'exposition, et d'autre métadonnées pertinentes.
- Les couleurs dominantes de l'image
- Les tags utilisateurs associés à l'image

Ces informations fournissent une vue d'ensemble de chaque image, allant de ses caractéristiques techniques à des aspects visuel tels que ses couleurs prédominantes, ainsi que ces métadonnées.

Dans le cadre de notre projet toutes ces informations permettent de faciliter l'organisation et l'analyse des images.

Préférences utilisateur

Chaque utilisateur a été simulé pour « liker » quatre images au hasard parmi celles disponibles dans notre jeu de données. Ces images likées servent de points de départ pour recommander d'autres images similaires en termes de coloration. Un algorithme de recommandation calcule la similarité des couleurs entre les images likées et toutes les autres images, puis classe les images en fonction des correspondances.

Les recommandations sont ensuite présentées sous forme de graphique à barres, où les images recommandées sont affichées avec leur scores de similarité correspondants. Ces graphiques offrent une visualisation claire des recommandations pour chaque utilisateur, avec les noms abrégés des images et les scores de similarité associés, permettant aux utilisateurs de comprendre rapidement quelles images leur sont suggérées et à quel point elles sont similaires à leurs préférences antérieures.

Cette approche permet de personnaliser les recommandations d'images en fonction des préférences individuelles de chaque utilisateur, offrant ainsi une expérience plus personnalisée et pertinente. De plus, cette méthode est pérenne car il est facile d'adapter les recommandations en fonction des besoins spécifiques des utilisateurs

Modèles d'exploration de données

Les algorithmes qui ont été exploités dans ce projet sont Kmeans afin de trouver les quatre couleurs prédominantes présentes dans les images. A la suite de ce traitement, on forme des clusters par couleur.

Auto-évaluation du projet

Nous pensons avoir répondu aux attentes premières du projet, cependant nous pourrions pousser l'étude un peu plus loin en procédant notamment à un cluster en nuage de points pour analyser les couleurs de chaque image, et ainsi déterminer la proximité entre elles. De plus, notre gestion de préférences est basée sur des utilisateurs créer par le code qui like de manière aléatoire les images. Une façon de changer ce procédé serait d'utiliser un jeu de données avec des fonctionnalités déjà présentes pour ce sujet ou pousser la réflexion dans ce domaine. Enfin pour la recommandation, nous avons utilisé une façon de faire, mais une étude plus poussée en utilisant les autres moyens pourrait être fait, de même qu'une comparaison à la fin des plusieurs façon de faire afin de voir laquelle est la plus performante.

Remarques sur le module

Rien à redire concernant cette partie.

Conclusion

Ce projet nous a permis de mieux appréhender le domaine du Datamining et Machine-Learning. De plus nous avons eu l'occasion de voir comment mettre en œuvre un système autonome de mise en forme de données grâce à différentes librairies Python.

Nous avons réalisé d'une des différentes manières possible, il pourrait être intéressant de pousser le concept plus loin et de comparer les multiples façon de faire ce projet afin de voir la meilleure solution.