Rapport TDM

VERNIERE Loic

DIB Nassim

Sommaire

I- Introduction

L'objectif de ce projet est de mettre en place un système automatisé d'acquisition, de traitement, d'analyse et de recommandation des images pour un utilisateur en fonction de ces préférences.

Pour ce faire nous utiliserons divers librairies et algorithmes de clustering et de prédiction.

II- Collecte de données

Automatisation de la collecte de données

Il nous a fallu tout d'abord trouver un site proposant des banques de données d'images libre. Nous nous sommes tourne vers le site Kaggle.

Afin de récupérer les images en python, nous avons créé un token l'API que nous appellerons par la suite en python.

Pour automatiser la collecte des données, il nous a fallu importer opendataset dans python puis faire un call API kaggle pointant vers la banque d'image souhaitée

La banque d'image choisi comporte plusieurs catégories d'images (100 images) : avions, voitures, animaux, fleurs… Toutes les images sont classées dans des dossiers nommés en fonction de leurs types.

Les images récupérées ne contiennent pas de métadonnées, il nous faut alors les extraire en utilisant des algorithmes python.

III- Etiquetage et annotations

Grace a notre programme python, toutes les images de notre banque vont être traitées par dossier dans une boucle pour en extraire les informations.

Plusieurs types d'images sont prises en compte tel que les différents formats mais aussi les différentes résolutions.

Les informations que nous voulons extraire de ces images sont donc, une **catégorie**, **une résolution** ainsi qu'une **couleur prédominante**.

Pour ce faire, nous allons utiliser la Library **PIL** afin ouvrir notre image puis nous allons en extraire la taille grâce à une première méthode size ().

Pour ce qui est de la catégorie, nous n'avons qu'a associer le nom du dossier qui correspond à la catégorie.

La couleur prédominante sur chaque image nécessite l'utilisation d'un algorithme de clustering nomme **KMeans.**

Le clustering k-Means est une méthode de quantification vectorielle, issue du traitement du signal, qui est populaire pour l'analyse de clusters dans l'exploration de données. Le clustering k-Means vise à partitionner n observations en k clusters dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche, servant de prototype du cluster.

Il va dont créer des groupes de pixels pour en déduire quelle couleur et la plus présente dans l'image.

Afin de stocker ces données relatives aux images nous avons choisi le format **json** qui nous semble être un bon choix pour les exploiter par la suite.

IV- Analyses et visualisation des données

Création de préférences utilisateurs

Les préférences utilisateurs ont été enregistrées sous la forme de deux listes :

La première contient un échenillons de 10 images, la deuxième contient les préférences de l'utilisateur « Favorite or Not favorite » sur cet échantillon d'image afin de déterminer les images qui pourrait lui plaire.

Suggestions

Une fois les préférences de l'utilisateur définis, nous utilisons **sklearn** afin de prédire quelles images sont susceptible de plaire à notre utilisateur.

Mise en forme

Une fois les données stockées et exploitable, nous les avons créés plusieurs graphs en utilisant matplotlib :

- Graph en battons des couleurs prédominantes
- Graph de résolutions des images les plus fréquentes
- Nombre d'images par catégorie

V- Bibliographie

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://pypi.org/project/webcolors/

https://www.w3schools.com/python/matplotlib_pyplot.asp

VI- Conclusion

Ce projet nous a permis d'appréhender et d'utiliser des librairies crées pour l'exploitation et le traitement des données, grâce à elles nous avons été en mesure de coller aux attentes du projet et de fournir des suggestions correctes selon les gouts d'un utilisateur.

Le traitement des données et le monde de l'image étaient deux choses nouvelles pour nous, c'est pourquoi le projet à demandé un certain investissement personnel afin de gagner en compétences et connaissances sur ces sujets.

Les notions d'analyse de résultats et de visualisation des données nous ont permises d'apporter ce côté visuel et concret à nos algorithmes.

Pour finir nous sommes d''accord sur le fait que ce projet était un très bon moyen de découvrir la manipulation des données.

Cela nous emmène également des questions sur la puissance et la complexité des algorithmes de suggestions qui tournent derrière des outils comme Netflix, Instagram etc… Ici nous avions peu de données à croiser mais si on veut passer à l'échelle mondiale sur des services qui touchent des millions de personnes cela devient très impressionnant.