Léo Gonnet 4ETI
Hugo Marta CLBD

Compte rendu

Projet de système de recommandation

Data Mining

Sommaire

Objectifs du projet

Ce projet a pour but final de recommander des images à un utilisateur en fonction de ses préférences. Pour cela il lui sera demandé d'aimer certaines photos parmi un panel proposé. Pour simplifier la compréhension du projet nous avons découpé le problème en plusieurs parties :

- Collecte de données
- Etiquetage et annotation
- Analyse de données
- Visualisation des données
- Système de recommandation

Léo Gonnet 4ETI
Hugo Marta CLBD

Sources des données

Nous avons choisi d'utiliser les images (libres de droit) de Wikimedia. La base de données est assez conséquente pour nos besoins, contrairement à celle de Wikidata qui ne nous avait pas satisfaite de par son petit nombre de résultats par requête.
La documentation de Wikimedia nous apprend à manier son API par url en modifiant différents paramètres.
On note l'importance d'ajouter un header spécifique aux requêtes pour wikimedia, qui permet de passer outre la limite de requête maximum par seconde.

On arrive ensuite à automatiser la récupération des informations des images et certaines de leurs étiquettes judicieusement choisies parmi celles des réponses aux requêtes. Ces étiquettes sont directement insérées dans notre base de données (voir 'Conception d'une base de données').

Finalement, on télécharge l'ensemble des des images que l'on a retenues dans notre base de données, afin de les analyser et compléter nos informations.

Taille des données

Nous utilisons 200 images lors de nos tests illustrant les mots clés 'cat' et 'car'. Il est cependant tout à fait possible de modifier ces paramètres facilement, comme vu dans la partie 'Source de données', pour atteindre des quantités plus importantes d'images. Toutefois, les traitements deviennent de plus en plus longs. On s'est donc limité à une quantité de 200 images ce qui reste raisonnable.
La taille de 200 images est d'environ 8Go.

Conception d'une base de données

Pour stocker simplement et efficacement les informations des différentes images nous utilisons une base de données à l'aide de DataFrame que nous enregistrons dans un fichier CSV. Ainsi il est possible de la modifier et de la visualiser facilement. Cette base de données appelée 'bdd.csv' présente la structure suivante :

Léo Gonnet 4ETI
Hugo Marta CLBD

ID	sha1	annee	largeur	hauteur	url	taille	couleur
0	ace42e2d7c42305f9ebb5786560c234610c47ec6	2010	3264	2176	https://upload.wikimedia.org/wikipedia/commons/e/e6/126p_beige_jaslo.JPG	grande	[208.8411820616843 213.2644949451424 217.7082865289986
1	8225329ebd3e3860f47463a13f92daebbebfa989	2005	1683	1395	https://upload.wikimedia.org/wikipedia/commons/0/0c/1928_Model_A_Ford.jpg	moyenne	[208.856772409013 213.2773792030541 217.7191632811912
2	d751ae9012a18dab35f5b3a846af9dd8ff0836db	2014	1700	741	https://upload.wikimedia.org/wikipedia/commons/e/e8/1946_Dodge_D24C_4-Door_Sedan_Suicide_Doors_259.jpg	moyenne	[208.8411820616843 213.2644949451424 217.7082865289986
3	0fd0cf9c5dd1c4ffeb3c1f6495c9af35f5726207	2015	5198	2847	https://upload.wikimedia.org/wikipedia/commons/e/e4/1998_Lincoln_Mark_VIII_LSC_in_red%2C_front_left.jpg	grande	[208.8347300215877 213.2579697624305 217.701598272147]
4	5a1c017e4f755196b8145d49b06dc2b86e640e18	2008	1648	956	https://upload.wikimedia.org/wikipedia/commons/2/23/2009_Pontiac_G8_GXP.jpg	moyenne	[209.0708079930384 213.4574283232086 217.8963509991399
5	7532d8bad4fb3a7d7db0fe2a9fe81e3eebbc76fc	2021	2781	1373	https://upload.wikimedia.org/wikipedia/commons/e/e5/2014_Perodua_Axia_SE_Front.jpg	moyenne	[209.0708079930384 213.4574283232086 217.8963509991399
6	ffb96ef329de27013239daabd16b5049df4a10b3	2021	2000	914	https://upload.wikimedia.org/wikipedia/commons/4/46/2015_Silverstone_Classic_-_Aston_Martin_DB4_%2819703112644%29.jpg	moyenne	[209.0708079930384 213.4574283232086 217.8963509991399
7	13934023585e5e51417b47877ebb3ccb152511c9	2017	3455	2081	https://upload.wikimedia.org/wikipedia/commons/d/dc/2016_Mercedes-Benz_GLS_350d_%28X_166%29_4MATIC_wagon_%282017-02-08%29_01.jpg	grande	[209.0708079930384 213.4574283232086 217.8963509991399
8	9c3f3340349ad9b3ebd06d83664738a2d65d3256	2018	4430	2620	https://upload.wikimedia.org/wikipedia/commons/d/de/2017_Alfa_Romeo_Stelvio_Milano_Edizione_TD_Automatic_2.1.jpg	grande	[208.856772409013 213.2773792030541 217.7191632811912
9	2437143206c84895c182e52ead91aeded25aec41	2018	4778	2199	https://upload.wikimedia.org/wikipedia/commons/9/9b/2017_Ford_Mustang_EcoBoost_2.3_Rear.jpg	grande	[208.8204663212324 213.2424006908577 217.6822970639120

La première colonne ‘ID’ contient l’identifiant unique de l’image. La seconde ‘sha1’ contient le hash de l’image : on l’utilise afin d’identifier une image; en effet le procédé de hachage garantit la quasi unicité de cette clé ce qui nous suffit largement à notre échelle. Ensuite on retrouve l’année pendant laquelle la photo a été prise, ainsi que sa largeur et sa hauteur. On retrouve alors l’url à partir duquel on télécharge l’image. Dans ‘couleur’ on retrouve un triplet de valeurs RGB allant de 0 à 255 représentant la couleur dominante de l’objet principal présent sur l’image. Enfin après avoir classé l’image dans une catégorie de taille : petite, moyenne ou grande; on ajoute cette même catégorie dans la colonne ‘taille’.

Gestion du profil de l’utilisateur

Nous avons créé une classe Utilisateur qui comporte plusieurs fonctions ainsi que certaines données sur les préférences de l’utilisateur. On retrouve notamment son nom, son identifiant et les photos qu’il a aimé. Les fonctions proposées sont les suivantes : ‘like’ qui lui permet d’aimer une photo, ‘recommander’ qui relève les caractéristiques des images aimées et qui propose alors une image qui se rapproche le plus de ses préférences. (à noter que cette image peut volontairement différer avec des paramètres identiques)

Nous avions également prévu de programmer les fonctions charger et enregistrer qui permettent de stocker les données d’un utilisateur dans un fichier JSON mais par manque de temps nous nous sommes concentrés sur d’autres fonctionnalités.

Léo Gonnet 4ETI
Hugo Marta CLBD

Modèle de recommandation d’images

Notre fonction de recommandation est basée sur un calcul de distances. En effet, celle-ci va comparer chaque image aimée par l’utilisateur avec toutes les images téléchargées. Cette distance va varier selon : la couleur dominante de l’objet le plus imposant de l’image, l’année de capture de la photo et la taille de l’image (petite, moyenne et grande).

Notre implémentation en orienté objet permet facilement de faire évoluer notre code. Notre prochain objectif est de mettre en place une interface utilisateur pour que l’utilisateur puisse se connecter à son profil et aimer de nouvelles photos ou bien se faire recommander des images.

Auto-évaluation des résultats obtenus

Par souci d’optimisation du temps d’exécution du programme nous n’affichons pas les images aimées par l’utilisateur mais seulement leurs identifiants. En revanche, nous affichons l’image finale. Nous avons alors vérifié l’efficacité de notre programme en vérifiant manuellement la proximité entre les images aimées et l’image proposée.

Dans l’ensemble le résultat est correct, cependant il s’agit d’un programme basique n’utilisant pas d’IA qui permettrait par exemple de reconnaître l’objet principal et de proposer des images comportant le même objet.

Remarques sur les séances de cours/TP

Nous n’avons pas de remarque particulière par rapport aux séances de cours et de TP. Nous avons largement préféré la partie Projet avec plus de liberté et d’autonomie, toutefois, le temps qui nous était donné était un peu court pour réaliser toutes les fonctionnalités demandées dans le cahier des charges.

Conclusion

En conclusion, ce projet de système de recommandation d'images a été réalisé avec succès en utilisant des techniques de collecte, d'annotation et d'analyse de données. Nous avons utilisé une base de données pour stocker les informations des différentes images et avons développé un modèle de recommandation basé sur un calcul de distances. Nous avons également créé une classe Utilisateur pour gérer les préférences de l'utilisateur et proposer des recommandations personnalisées. Les résultats obtenus ont été auto-évalués et sont satisfaisants.
Ce projet nous a montré l'importance de la collecte et de l'analyse de données pour développer des systèmes de recommandation précis et utiles pour les utilisateurs.