Rapport Data Mining et Machine Learning

Julien Caposiena — Johan Planchon

Le but de votre projet

Le but de ce projet est de créer un système de recommandation d'images en fonction de préférences utilisateurs, le tout en se formant à toutes les différentes parties constituantes le data mining / machine learning.
Ce projet sera découpé en plusieurs parties, en premier la collecte des données, ensuite l'étiquetage et l'annotation, suivis de l'analyse des données et enfin de la visualisation des données.

Sources des données de vos images et leurs licences

Toutes les images proviennent de Wikidata[1] et sont donc sous licence Libre CC0.[2]

Taille de vos données

Nous avons collecté au total 97 images de chats.

Informations que vous avez décidé de stocker pour chaque image

Pour chaque image, nous avons choisi de garder (les informations en gras sont celles exploitée dans le projet) :

- le mimetype (“Extenstion”)
- la résolution en DPI (“ResolutionUnit”)
- l'offset de l'exif (“ExifOffset”)
- la description de l'image (“ImageDescription")
- la marque de l'appareil ("Make")
- le modèle de l'appareil ("Model")
- le logiciel d'édition photo ("Software")
- la date de prise de la photo ("DateTime")
- la hauteur de l'image ("YResolution")
- le copyright de la photo ("Copyright")

---

[1] https://www.wikidata.org

[2] All structured data from the main, Property, Lexeme, and EntitySchema namespaces is available under the Creative Commons CC0 License; text in the other namespaces is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy.

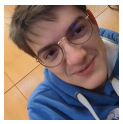
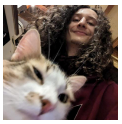
- la largeur de l'image ("XResolution")
- l'artiste qui a pris la photo ("Artist")
- les information sur la position ou a été prise la photo ("GPSInfo")
- l'orientation de la photo ("Orientation")
- la position des composantes de chrominance par rapport à la composante de luminance ("YCbCrPositioning")
- le temps d'exposition de la photo ("ExposureTime")
- le nombre de ligne par bande ("RowsPerStrip")
- le nombre d'octets dans une bande ("StripByteCounts")
- la largeur de l'image ("ImageWidth")
- la longueur de l'image ("ImageLength")
- le mode d'exposition ("ExposureMode")
- la balance des blancs ("WhiteBalance")
- la précision ("Sharpness")
- la distance avec le sujet ("SubjectDistanceRange")
- la largeur de la tuile("TileWidth")
- la longueur de la tuile ("TileLength")
- la compression de l'image ("Compression")
- la configuration planaire ("PlanarConfiguration")
- le contraste ("Contrast")
- la saturation ("Saturation")
- une des couleurs les plus présentes ("Color1"): ajoutée après analyse de l'image
- une des couleurs les plus présentes ("Color2"): ajoutée après analyse de l'image

Informations concernant les préférences de l'utilisateur

On demande à l'utilisateur de tagger les images avec un ou plusieurs mots. Le reste des préférences sera calculé depuis les images. Ici seulement la couleur favorite, cependant cela pourrait être facilement étendu à l'orientation, la taille, le contraste et la saturation favorite.

Modèles d'exploration de données et/ou d'apprentissage machine que vous avez utilisés avec les métriques obtenues.

Nous avons utilisé un modèle KMeans afin d'avoir les couleurs favorites de l'utilisateur. Et nous avons réalisé un barchart pour présenter les couleurs les plus présentes dans les images ainsi que dans les images favorites.

Auto-évaluation

Nous pensons avoir répondu au sujet, cependant le code pourrait être plus complet, analyser plus de paramètres et avoir plus de graphiques. Disposer d'une plus grande quantité de données pourrait être un plus afin de rendre le modèle plus fiable. Cependant nous avons eu beaucoup de mal à trouver de la documentation de qualité en fonction de ce que nous voulions faire, il y a beaucoup d'informations en ligne sur le sujet mais la qualité est très souvent terrible.

Remarques sur les séances pratiques, les exercices et les possibilités d'amélioration

Plutôt que d'axer les TP sur des exemples d'utilisation et de quoi faire, simplifier les exercices pour en placer plus et permettre de comprendre vraiment ce que font les différentes fonctions. Comprendre les nuances entre les différentes fonctions d'activation. Et globalement organiser le cours pour que ce soit plus comme une boîte à outils dans laquelle on peut piocher qui nous présente les outils le plus individuellement possible dans les exercices. De manière à ce que pour le projet il nous suffit d'aller piocher et justifier l'utilisation de chacun des outils plutôt que ici d'en prendre un au hasard et de voir si cela sort un résultat cohérent sans savoir ce qu'il fait ni comment il fonctionne.

Conclusion

Le cours nous à beaucoup appris à chercher sur internet des résultats en rapport avec l'IA, cependant dans ce que nous utilisons lors du cours il reste de trop nombreuses boîtes noires ou l'on a aucune idée de ce qu'elle font ou alors c'est encore trop flou. Je dirais donc que le cours nous aura été utile cependant il n'aura pas forcément répondu à nos attentes de comprendre un peu plus en détail tout ce que nous utilisons en data mining, ce qui nous aura pénalisé sur le projet puisque nous avons perdu beaucoup de temps à se demander quoi utiliser et à chercher des exemples de qualité.