100 heures Mis à jour le mardi 30 juin 2020

Modification réalisée le 30 juin 2020 :

Une partie du travail à réaliser sur ce projet nécessitait d'utiliser une API qui n'existe plus, cette partie du projet a donc été supprimée.

Si vous êtes déjà engagé sur le projet et que l'API était encore disponible quand vous réalisiez votre travail, vous pouvez poursuivre sur **l'énoncé de l'ancien projet [https://s3-eu-west-1.amazonaws.com/course.oc-static.com/projects/Data_Scientist_P6/Project+6+Data+Scientist+-+Students+May+2020.pdf]** et présenter l'ensemble de vos livrables pendant la soutenance.
Autrement, vous pouvez démarrer ou poursuivre votre projet en suivant l'énoncé ci-dessous.

Vous êtes Data Scientist au sein de l'entreprise "**Place de marché**", qui souhaite lancer une marketplace e-commerce.

[https://user.oc-static.com/upload/2019/02/24/15510259240381_Projet%20textimage%20logo.png]

Sur la place de marché, des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs et est donc peu fiable. De plus, le volume des articles est pour l'instant très petit.

Pour rendre l'expérience utilisateur des vendeurs (faciliter la mise en ligne de nouveaux articles) et des acheteurs (faciliter la recherche de produits) la plus fluide possible et dans l'optique d'un passage à l'échelle, **il devient nécessaire d'automatiser cette tâche.**

Linda, lead data scientist, vous demande donc d'étudier la faisabilité d'un **moteur de**

classification des articles en différentes catégories, avec un niveau de précision suffisant.

Les données

Linda vous a fourni un premier jeu de données [https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientist/Projet+-+Textimage+DAS+V2/Dataset+projet+pre%CC%81traitement+textes+images.zip] d'articles avec le lien pour télécharger la photo et une description associée.

Votre mission

Votre mission est de **réaliser une première étude de faisabilité d'un moteur de classification** d'articles basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article.

Vous **analyserez le jeu de données** en **réalisant un prétraitement** des images et des descriptions des produits, une **réduction de dimension,** puis un **clustering**. Les résultats du clustering seront présentés sous la forme d'une représentation en deux dimensions à déterminer, qui 'illustrera le fait que les caractéristiques extraites permettent de regrouper des produits de même catégorie.

La représentation graphique vous aidera à convaincre Linda que cette approche de modélisation permettra bien de regrouper des produits de même catégorie.

> Attention, Linda n'a pas besoin d'un moteur de classification à ce stade, mais bien d'une étude de faisabilité !

Contraintes

Linda vous a communiqué la contrainte suivante : afin d'extraire les features, mettre en œuvre *a minima* un algorithme de type SIFT / ORB / SURF.

> Un algorithme de type CNN Transfer Learning peut éventuellement être utilisé en complément, s'il peut apporter un éclairage supplémentaire à la démonstration.

Livrables attendus

- Un **notebook** (ou des fichiers .py) contenant les fonctions permettant le prétraitement des données textes et images ainsi que les résultats du clustering (en y incluant des représentations graphiques au besoin).
- Un support de **présentation** qui présente la démarche et les résultats du clustering.

> Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé

"*P6_nom_prenom*", tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "*P6_01_notebook*", et ainsi de suite.

Modalités de la soutenance

5 min - Rappel de la problématique et présentation du jeu de données (à l'aide de votre support de présentation)

15 min - Explication des prétraitements et des résultats du clustering

5 min - Conclusion sur la faisabilité du moteur de classification et vos recommandations pour sa création éventuelle

5 à 10 minutes de questions-réponses

Compétences évaluées

- Prétraiter des données image pour obtenir un jeu de données exploitable
- Mettre en œuvre des techniques de réduction de dimension
- Représenter graphiquement des données à grandes dimensions
- Prétraiter des données texte pour obtenir un jeu de données exploitable