Vous êtes Développeur IA au sein de l’entreprise "The World”, qui souhaite lancer une marketplace e-commerce.

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.

**Elon Musk**, Le Head of AI , vous demande donc d'étudier la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant.

**La 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.

Voici le mail qu’il vous a envoyé.

|  |
| --- |
| Bonjour, |
| Merci pour ton aide sur ce projet !  Ta mission est de réaliser, **dans une première itération, une 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.  Tu dois**analyser les descriptions les images des produits**, au travers des étapes suivantes :   * **Un prétraitement** des données image suivant le cas ; * Une extraction de features ; * Une **réduction en 2 dimensions**, afin de projeter les produits sur un graphique 2D, sous la forme de points dont la couleur correspondra à la catégorie réelle ; * **Analyse du graphique** afin d’en déduire ou pas, à l’aide des descriptions ou des imaVoici les contraintes : * *une approche de type word/sentence embedding avec USE (Universal Sentence Encoder).* * Afin d’extraire les features image, il sera nécessaire de mettre en œuvre : * *3 algorithmes algorithme de type SIFT et ORB et SURF ;* * *3 algorithme de type CNN Transfer Learning + 1 CNN que tu auras implémenté toi-même* * Réalisation d’une **mesure pour confirmer ton analyse visuelle**, en calculant la similarité entre les catégories réelles et les catégories issues d’une segmentation en clusters.   Pourrais-tu nous démontrer, par cette approche, la faisabilité de regrouper automatiquement des produits de même catégorie ?  En pièces jointes, tu trouveras un exemple de mise en œuvre de l’approche de type CNN Transfer Learning d’extraction de features images sur un autre dataset. Je t’invite à l’utiliser comme point de départ, cela va te faire gagner beaucoup de temps !  Merci encore,  Linda  PS : J’ai bien vérifié qu’il n’y avait aucune contrainte de propriété intellectuelle sur les données et les images. |
| Pièces joint :   * premier jeu de données d’articles avec le lien pour télécharger la photo et une description associée |
|  |

Une semaine plus tard, vous partagez votre travail avec Elon, et il vous répond avec enthousiasme avec une autre demande.

|  |
| --- |
| Bonjour, |
| Merci beaucoup pour ton travail !  Bravo d’avoir démontré la faisabilité de regrouper automatiquement des produits de même catégorie !  Maintenant, je te propose de passer à la deuxième itération. Pourrais-tu réaliser une **classification supervisée à partir des images** ? Je souhaiterais que tu mettes en place une **data augmentation** afin d’optimiser le modèle.  Merci encore,  Elon |

**(BONUS)**

Afin d’extraire les features texte, il sera nécessaire de mettre en œuvre :

* *deux approches de type “bag-of-words”, comptage simple de mots et Tf-idf ;*
* *Trois approches de type word/sentence embedding classique avec Word2Vec ( ou Doc2vec ) ou Glove ou FastText) ;*
* *une approche de type word/sentence embedding avec BERT*, la **faisabilité de regrouper automatiquement** des produits de même catégorie ;
* Réalisation d’une **mesure pour confirmer ton analyse visuelle**, en calculant la similarité entre les catégories réelles et les catégories issues d’une segmentation en clusters.

**Contraintes**

* Prétraiter des données image pour obtenir un jeu de données exploitable
* Prétraiter des données texte pour obtenir un jeu de données exploitable
* Représenter graphiquement des données à grandes dimensions
* Mettre en œuvre des techniques de réduction de dimension
* Définir la stratégie d’élaboration d’un modèle d'apprentissage profond
* Évaluer la performance des modèles d’apprentissage profond selon différents critères
* Utiliser des techniques d’augmentation des données

**Livrables attendus**

- Un notebook contenant les fonctions permettant le prétraitement des données images ainsi que les résultats du clustering (en y incluant des représentations graphiques au besoin).

1. Un ou des **notebooks** contenant les fonctions permettant le **prétraitement** et la feature extraction des données textes et images ainsi que les résultats de l’étude de faisabilité (graphiques, mesure de similarité)
2. Un notebook de **classification supervisée des images**