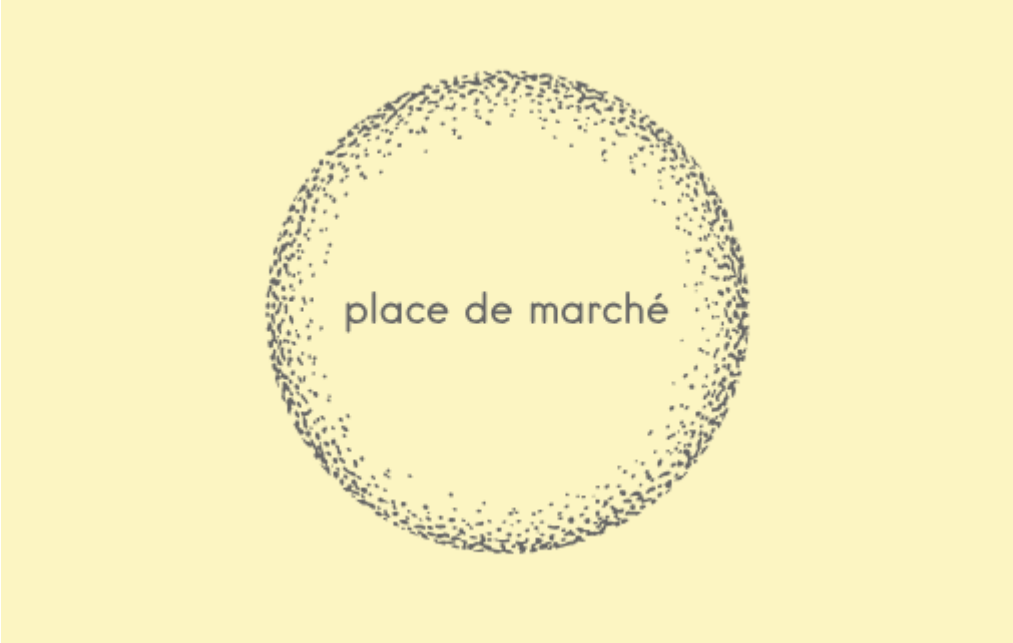
Classifiez automatiquement des biens de consommation

100 heures

Vous êtes Data Scientist au sein de l'entreprise "**Place de marché**", 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**.

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.

Voici le mail qu'elle vous a envoyé.

Bonjour,

Merci pour ton aide sur ce projet !

Ta 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.

Tu dois **analyser le jeu de données** en **réalisant un prétraitement** des descriptions des produits et des images, une **réduction de dimension**, puis un **clustering**. Les résultats de la réduction de dimension et du clustering seront à présenter sous la forme de graphiques en deux dimensions, et confirmés par un calcul de similarité entre les catégories réelles et les clusters. Ces résultats illustreront le fait que les caractéristiques extraites permettent de regrouper des produits de même catégorie.

Pourrais-tu nous démontrer, par cette approche de modélisation, la faisabilité de regrouper automatiquement des produits de même catégorie ?

Voici les contraintes :

> 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 ;
- une approche de type word/sentence embedding classique avec Word2Vec (ou Glove ou FastText) ;
- une approche de type word/sentence embedding avec BERT ;
- une approche de type word/sentence embedding avec USE (Universal Sentence Encoder).

En pièce jointe, tu trouveras **un exemple de mise en œuvre de ces approches sur un autre dataset**. Je t'invite à l'utiliser comme point de départ, **cela va te faire gagner beaucoup de temps** !

> Afin d'extraire les features image, il sera nécessaire de mettre en œuvre :

- un algorithme de type SIFT / ORB / SURF ;
- un algorithme de type CNN Transfer Learning.

Merci encore,

Linda

P.S. : j'insiste sur le fait qu'on n'a pas besoin d'un moteur de classification supervisée à ce stade, mais bien d'une étude de faisabilité !

Pièces jointes :

- premier jeu de données d'articles avec le lien pour télécharger la photo et une description associée
- un notebook d'exemple de mise en œuvre de ces approches

Bon courage !

Livrables

- 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).
- Un support de **présentation** qui présente la démarche et les résultats du clustering.

Pour faciliter votre passage devant le jury, déposez sur la plateforme, dans un dossier zip nommé "***Titre_du_projet_nom_prénom***", votre livrable nommé comme suit : **Nom_Prénom_n° du livrable_nom du livrable_date de démarrage du projet**. Cela donnera :

- *Nom_Prénom_1_notebook_mmaaaa*
- *Nom_Prénom_2_presentation_mmaaaa*

Par exemple, votre premier livrable peut être nommé comme suit *:* *Dupont_Jean_1_notebook_12022*.

Soutenance

La soutenance se déroulera en visioconférence et durera 30 minutes. Elle s'appuiera sur votre deuxième livrable (votre présentation).

- **Présentation** (20 minutes)
  - Rappel de la problématique et présentation du jeu de données (5 minutes).
  - Explication des prétraitements et des résultats du clustering (10 minutes).
  - Conclusion sur la faisabilité du moteur de classification et vos recommandations pour sa création éventuelle (5 minutes).
- **Discussion** (5 minutes)
  L'évaluateur, jouant le rôle de Linda, vous challengera sur vos choix.
- **Débriefing** (5 minutes)
  À la fin de la soutenance, l'évaluateur arrêtera de jouer le rôle de Linda pour vous permettre de débriefer ensemble.

Votre présentation devrait durer 20 minutes (+/- 5 minutes). Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 15 minutes ou au-dessus de 25 minutes peuvent être refusées.

Compétences évaluées

- 💡
  Prétraiter des données image pour obtenir un jeu de données exploitable
- 💡
  Prétraiter des données texte pour obtenir un jeu de données exploitable
- 💡
  Mettre en œuvre des techniques de réduction de dimension

- 
  Représenter graphiquement des données à grandes dimensions