**Classifiez automatiquement des biens de consommation**

**Enoncé :**

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

Sur cette place de marché anglophone, 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 d‘attribution de la catégorie.

Linda, Lead Data Scientist, vous demande donc d'étudier la faisabilité d'un moteur de classification des articles en différentes catégories, à partir du texte (en anglais) et de l’image.

Voici le mail qu’elle vous a envoyé :

*Objet : démarrage de la mission*

*Bonjour,*

*Merci pour ton aide sur ce projet !*

*Ta mission sera de réaliser une étude de faisabilité d'un moteur de classification automatique d’articles, en utilisant leur image et leur description sur le jeu de données d'articles disponible dans la première pièce jointe de ce mail.*

*Pourrais-tu analyser les descriptions textuelles et les images des produits, au travers des étapes suivantes :*

*>Un prétraitement des données texte et image*

*>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*

*>Une analyse du graphique afin de conclure, à l’aide des descriptions ou des images, sur la faisabilité de regrouper automatiquement des produits de même catégorie*

*>Une 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 ainsi la faisabilité de regrouper automatiquement des produits de même catégorie ?*

*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.*

*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).*

*Pour t’aider, en deuxième pièce jointe, tu trouveras un exemple de Notebook mettant en œuvre les approches d’extraction de features d’images sur un autre dataset.*

*Merci encore,*

*Linda*

*PS : J’ai bien vérifié qu’il n’y avait aucune contrainte de propriété intellectuelle sur les données et les images.*

Pour l’approche de type SIFT, n’hésitez pas à consulter le webinaire disponible dans les ressources.

**Etape 1** : Etudiez la faisabilité de classification en utilisant des méthodes NLP basiques

- Réalisez au préalable un prétraitement de vos textes : nettoyage, stemming, lemmatization etc.

- Utilisez des méthodes basiques d’encodage de texte (Bag of Words, TfIdf etc.)

- Utilisez des méthodes de réduction de dimension sur le texte encodé et comparer les résultats avec les catégories réelles

**Etape 2** : Etudiez la faisabilité de classification en utilisant des méthodes NLP plus avancées

- Reprenez la démarche précédente en utilisant cette fois-ci des algorithmes de Word Embeddings (Word2Vec, BERT, Universal Sentence Encoder).

- Comparez vos résultats avec ceux des méthodes NLP plus basiques.

**Etape 3** : Etudiez la faisabilité de classification en utilisant des techniques de traitement d'images basiques

- Analysez les images via différentes méthodes de transformation (niveau de gris, filtrage bruit, contraste etc.).

- Utilisez des algorithmes d'extraction d'e features d’images comme SIFT.

- Utilisez des méthodes de réduction de dimension sur les images encodées et comparer les résultats avec les catégories réelles.

**Etape 4** : Etudiez la faisabilité de classification automatique en utilisant des techniques de traitement d'images plus avancées

- Reprenez la démarche précédente en utilisant cette fois-ci des réseaux de neurones préentraînés.

- Comparez vos résultats avec ceux des méthodes de traitement d’image plus basiques.

Une semaine plus tard, Vous partagez votre travail avec Linda. Elle vous invite désormais à aller plus loin dans l’analyse d’images.

Voici le mail qu’elle vous a envoyé :

*Objet : suite de la mission*

*Bonjour,*

*Merci beaucoup pour ton travail ! Voici la suite de ta mission :*

*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.*

*De plus, nous souhaitons élargir notre gamme de produits à l’épicerie fine.*

*Pour cela, pourrais-tu tester la collecte de produits à base de “champagne” via l’API disponible ici* ***https://rapidapi.com/edamam/api/edamam-food-and-grocery-database*** *?*

*Pourrais-tu ensuite nous proposer un script ou notebook Python permettant une extraction des 10premiers produits dans un fichier “.csv”, contenant pour chaque produit les données suivantes : foodId, label, category, foodContentsLabel, image.*

*Enfin, pourrais-tu formaliser dans un support de présentation de 30 slides maximum au format PDF l’ensemble de ta démarche ainsi que les résultats d’analyse les plus pertinents ?*

*Merci encore, bon courage !*

*Linda*

*PS : En pièce jointe, tu trouveras pour t’aider un exemple de mise en œuvre de classification supervisée sur un autre dataset.*

**Etape 1** : Faites la classification supervisée d'images via CNN Transfer Learning

- Appuyez-vous sur le notebook d’exemple fourni, testez une data augmentation par exemple.

**Etape 2** : Testez une API de collecte de données de change

- Récupérez la clé de l’API et reprenez l’exemple en code Python pour requêter une API.

**Etape 3** : Vérifiez votre travail et préparez-vous à la soutenance

- Structurez votre présentation autour des éléments clés : problématique, jeu de données, prétraitements et extractions de features, résultats de l'étude, classification supervisée et test de l’API.

- Préparez-vous à une discussion constructive, où vous justifierez vos choix méthodologiques face à l'évaluateur. Soyez conscient du temps.

- Anticipez le débriefing pour tirer le maximum de cette expérience d'apprentissage, en préparant des questions à poser à votre évaluateur.

**Livrables :**

1. Un ou des notebooks (ou des fichiers .py) 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

3. Un script Python (notebook ou fichier .py) de test de l’API et le fichier au format “csv” contenant les produits extraits

4. Un support de présentation pour la soutenance, détaillant le travail réalisé (Powerpoint ou équivalent, sauvegardé en pdf, 30 slides maximum) :

> l’étude de faisabilité

> la classification supervisée

> test de l’API

**Référentiel d'évaluation :**

Compétences évaluées :

- Définir la stratégie d’élaboration d’un modèle d'apprentissage profond

- Évaluer la performance des modèles d’apprentissage profond

- Prétraiter des données non structurées de type image

- Prétraiter des données non structurées de type texte

- Réaliser la collecte des données répondant à des critères définis via une API

- Réduire la dimension de données de grande dimension

- Utiliser des techniques appropriées de réduction en deux dimensions

I/ Définir la stratégie de collecte de données en recensant les API disponibles, et réaliser la collecte des données répondant à des critères définis via une API (interface de programmation) en prenant en compte les normes RGPD, afin de les exploiter pour l’élaboration d’un modèle.

CE1 Le candidat a défini sa stratégie de collecte de données et recensé les API disponibles pour le besoin de son projet

CE2 Le candidat a écrit et testé une requête pour obtenir les données via l’API.

CE3 Le candidat a récupéré les seuls champs nécessaires. Dans le cadre de ce projet, il s’agit de : Des champs foodId, label, category, foodContentsLabel, image

CE4 Le candidat applique au moins un filtre sur un des champs nécessaires pour ne collecter que les données ayant les valeurs correspondantes sur ce ou ces champs. Dans le cadre de ce projet, il s’agit de : filtrer sur l’ingrédient (“ingr”) champagne.

CE5 Le candidat a stocké les données collectées via l’API dans un fichier utilisable (ex. : fichier CSV ou pickle).

CE6 Le candidat veille au respect des normes RGPD dans toutes phases de la collecte et du stockage des données. En particulier :

-Il présente les 5 grands principes du RGPD

-Il ne gère que les données nécessaires pour la finalité du projet

II/ Utiliser des techniques appropriées de réduction en deux dimensions de données de grande dimension et les représenter graphiquement afin d'en réaliser l'analyse exploratoire.

CE1 Le candidat a mis en œuvre au moins une technique de réduction de dimension (via LDA, ACP, T-SNE, UMAP ou autre technique).

CE2 Le candidat a réalisé au moins un graphique représentant les données réduites en 2D (par exemple affichage des 2 composantes du T-SNE).

CE3 Le candidat a réalisé et formalisé une analyse du graphique en 2D.

III/ Prétraiter des données non structurées de type texte en prenant en compte les normes liées à la propriété intellectuelle et réaliser un feature engineering adapté aux modèles d'apprentissage afin d’obtenir un jeu de données exploitables.

CE1 Le candidat a nettoyé les champs de texte (suppression de la ponctuation et des mots de liaison, mise en minuscules)

CE2 Le candidat a écrit une fonction permettant de “tokeniser” une phrase.

CE3 Le candidat a écrit une fonction permettant de “stemmer” une phrase.

CE4 Le candidat a écrit une fonction permettant de “lemmatiser” une phrase.

CE5 Le candidat a construit des features ("feature engineering") de type bag-of-words (bag-of-words standard : comptage de mots, et Tf-idf), avec des étapes de nettoyage supplémentaires : seuil de fréquence des mots, normalisation des mots.

CE6 Le candidat a testé une phrase ou un court texte d'exemple, pour illustrer la bonne réalisation des 5 étapes précédentes.

CE7 Le candidat, en complément de la démarche de type “bag-of-words”, a mis en œuvre 3 démarches de word/sentence embedding : Word2Vec (ou Doc2Vec ou Glove ou FastText), BERT, et USE (Universal Sentence Encoder).

CE8 Le candidat s’assure que le texte traité ne relève pas d’une propriété intellectuelle dont l’utilisation ou la modification est interdite.

IV/ Prétraiter des données non structurées de type image en veillant au respect du droit à l’image et réaliser un feature engineering adapté aux modèles d'apprentissage afin d’obtenir un jeu de données exploitables.

CE1 Le candidat a utilisé des librairies spécialisées pour un premier traitement du contraste (ex. : openCV).

CE2 Le candidat a présenté des opérations de retraitement d'images (par exemple passage en gris, filtrage du bruit, égalisation, floutage) sur un ou plusieurs exemples.

CE3 Le candidat a élaboré une fonction permettant d'extraire des features ("feature engineering") de type "bag-of-images" via la génération de descripteurs (algorithmes ORB, ou SIFT, ou SURF).

CE4 Le candidat a élaboré et exécuté une fonction permettant d'extraire des features ("feature engineering") via un algorithme de Transfer Learning basé sur des réseaux de neurones, comme par exemple CNN.

CE5 Le candidat s’assure que les images utilisées ne relèvent pas d’une propriété intellectuelle.

V/ Réduire la dimension de données de grande dimension, afin d'optimiser les temps d’entraînement des modèles.

CE1 Le candidat a justifié la nécessité de la réduction de dimension.

CE2 Le candidat a appliqué une méthode de réduction de dimension adaptée à la roblématique (ex. : ACP).

CE3 Le candidat a justifié le choix des valeurs des paramètres dans la méthode de réduction de dimension retenue (ex. : le nombre de dimensions conservées pour l'ACP).

VI/ Définir la stratégie d’élaboration d’un modèle d'apprentissage profond, concevoir ou ré-utiliser des modèles pré-entraînés (transfer learning) et entraîner des modèles afin de réaliser une analyse prédictive.

CE1 Le candidat a défini sa stratégie d’élaboration d’un modèle pour répondre à un besoin métier (par exemple : choix de conception d’un modèle ou réutilisation de modèles pré-entraînés).

CE2 Le candidat a identifié la ou les cibles.

CE3 Le candidat a réalisé la séparation du jeu de données en jeu d’entraînement, jeu de validation et jeu de test.

CE4 Le candidat s'est assuré qu'il n’y a pas de fuite d’information entre les jeux de données (entraînement, validation et test).

CE5 Le candidat a testé plusieurs modèles d’apprentissage profond (par exemple à l’aide de la librairie Tensorflow / Keras) en partant du plus simple vers les plus complexes.

CE6 Le candidat a mis en oeuvre des modèles à partir de modèles pré-entraînés (technique de Transfer Learning)

VII/ Évaluer la performance des modèles d’apprentissage profond selon différents critères (scores, temps d'entraînement, etc.) afin de choisir le modèle le plus performant pour la problématique métier.

CE1 Le candidat a choisi une métrique adaptée à la problématique métier, et sert à évaluer la performance des modèles

CE2 Le candidat a explicité le choix de la métrique d’évaluation

CE3 Le candidat a évalué la performance d’un modèle de référence et sert de comparaison pour évaluer la performance des modèles plus complexes

CE4 Le candidat a calculé, hormis la métrique choisie, au moins un autre indicateur pour comparer les modèles (par exemple : le temps nécessaire pour l’entraînement du modèle)

CE5 Le candidat a optimisé au moins un des hyperparamètres du modèle choisi (par exemple : le choix de la fonction Loss, le Batch Size, le nombre d'Epochs)

CE6 Le candidat a présenté une synthèse comparative des différents modèles, par exemple sous forme de tableau.

VIII/ Utiliser des techniques d’augmentation des données afin d'améliorer la performance des modèles.

CE1 Le candidat a utilisé plusieurs techniques d’augmentation des données (par ex. pour des images : rotation, changement d’échelle, ajout de bruit…).

CE2 Le candidat a présenté une synthèse comparative des améliorations de performance grâce aux différentes techniques d'augmentation de données utilisées (maîtrise de l’overfitting, meilleur score).