**Prétraiter des données textes pour obtenir un jeu de données exploitable :**

 Vous avez nettoyé les champs de texte (suppression de la ponctuation et des mots de liaison, mise en minuscules)

Vous avez écrit une fonction permettant de “tokeniser” une phrase.

Vous avez écrit une fonction permettant de “stemmer” une phrase.

Vous avez écrit une fonction permettant de “lemmatiser” une phrase.

Vous avez 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.

Vous avez testé une phrase ou un court texte d'exemple, pour illustrer la bonne réalisation des 5 étapes précédentes.

Vous avez, en complément de la démarche de type “bag-of-words”, mis en oeuvre les démarches de word/sentence embedding : Word2Vec (ou Doc2Vec ou Glove ou FastText), BERT, et USE (Universal Sentence Encoder).

 Vous vous êtes assurés que le texte traité ne relève pas d’une propriété intellectuelle dont l’utilisation ou la modification est interdite.

**Prétraiter des données images pour obtenir un jeu de données exploitable.**

Vous avez utilisé des librairies spécialisées pour un premier traitement du contraste (ex. : openCV, pillow ou autre équivalent ).

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

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

 Vous avez 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 ( 3 modèles de transfert Learning + un modèle maison)

Vous vous êtes assurés que les images utilisées ne relèvent pas d’une propriété intellectuelle.

**Mettre en œuvre des techniques de réduction de dimension.**

Vous avez justifié la nécessité de la réduction de dimension.

Vous avez appliqué 3 méthodes de réduction de dimension adaptée à la problématique (ACP,t-SNE, UMAP).

Vous avez justifié le choix des valeurs des paramètres dans la méthode de réduction de dimension retenue

**Représenter graphiquement des données à grandes dimensions.**

Vous avez mis en œuvre les techniques de réduction de dimension ( ACP, T-SNE, UMAP)

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

Vous avez réalisé et formalisé une analyse du graphique en 2D.

**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.**

Vous avez défini votre stratégie de collecte de données et recensé les API disponibles pour le besoin de votre projet

Vous avez écrit et testé une requête pour obtenir les données via l’API.

Vous avez récupéré les seuls champs nécessaires. Dans le cadre de ce projet, comme indiqué dans l’énoncé, il s’agit de :

* Des champs foodId, label, category, foodContentsLabel, image

Vous avez appliqué 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, comme indiqué dans l’énoncé, il s’agit de :

* de filtrer sur l’ingrédient (“ingr”) champagne.

Vous avez stocké les données collectées via l’API dans un fichier utilisable (ex. : fichier CSV ou pickle).

Vous avez veillé au respect des normes RGPD dans toutes phases de la collecte et du stockage des données. En particulier :

* présenter les 5 grands principes du RGPD
* ne gérer que les données nécessaires pour la finalité du projet

**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.**

Vous avez défini votre stratégie d’élaboration d’un modèle pour répondre à un besoin métier (par exemple : choix de conception d’un modèle ré-utilisation de modèles pré-entraînés.

Vous avez identifié la ou les cibles.

Vous avez réalisé la séparation du jeu de données en jeu d’entraînement, jeu de validation et jeu de test.

Vous vous êtes assurés qu'il n’y a pas de fuite d’information entre les deux jeux de données (entraînement, validation et test).

Vous avez 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.

Vous avez mis en oeuvre des modèles à partir de modèles pré-entraînés (technique de Transfer Learning)

**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.**

Vous avez choisi une métrique adaptée à la problématique métier, et sert à évaluer la performance des modèles

Vous avez explicité le choix de la métrique d’évaluation

Vous avez évalué la performance d’un modèle de référence et sert de comparaison pour évaluer la performance des modèles plus complexes

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

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

Vous avez présenté une synthèse comparative des différents modèles, par exemple sous forme de tableau.

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

Vous avez utilisé plusieurs techniques d’augmentation des données (ex. pour des images : rotation, changement d’échelle, ajout de bruit…).

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