PLACE DE MARCHÉ

CLASSIFICATION AUTOMATIQUE DES BIENS DE CONSOMMATION

AOUT 2024 / LOKMAN AALIOUI

FEUILLE DE ROUTE

01 Mise en contexte

02 Exploration des données

03 Etudes de faisabilité texte et images

04 Classification supervisée images

05 Test d’une API

CONTEXTE

- L’entreprise **Place de marché** souhaite lancer une marketplace e-commerce où des vendeurs proposeront des articles à des acheteurs en postant une photo et une description.

- Actuellement, l'attribution de la catégorie d'un article est effectuée manuellement

- Pour faciliter la mise en ligne de nouveaux articles, il faudrait **automatiser** cette tâche d’attribution de la catégorie.

MISSION

- Etudier la faisabilité d’un moteur de classification des articles en différentes catégories, à partir du texte de description d’une part et de l’image d’autre part

- Tester une API de collecte de données

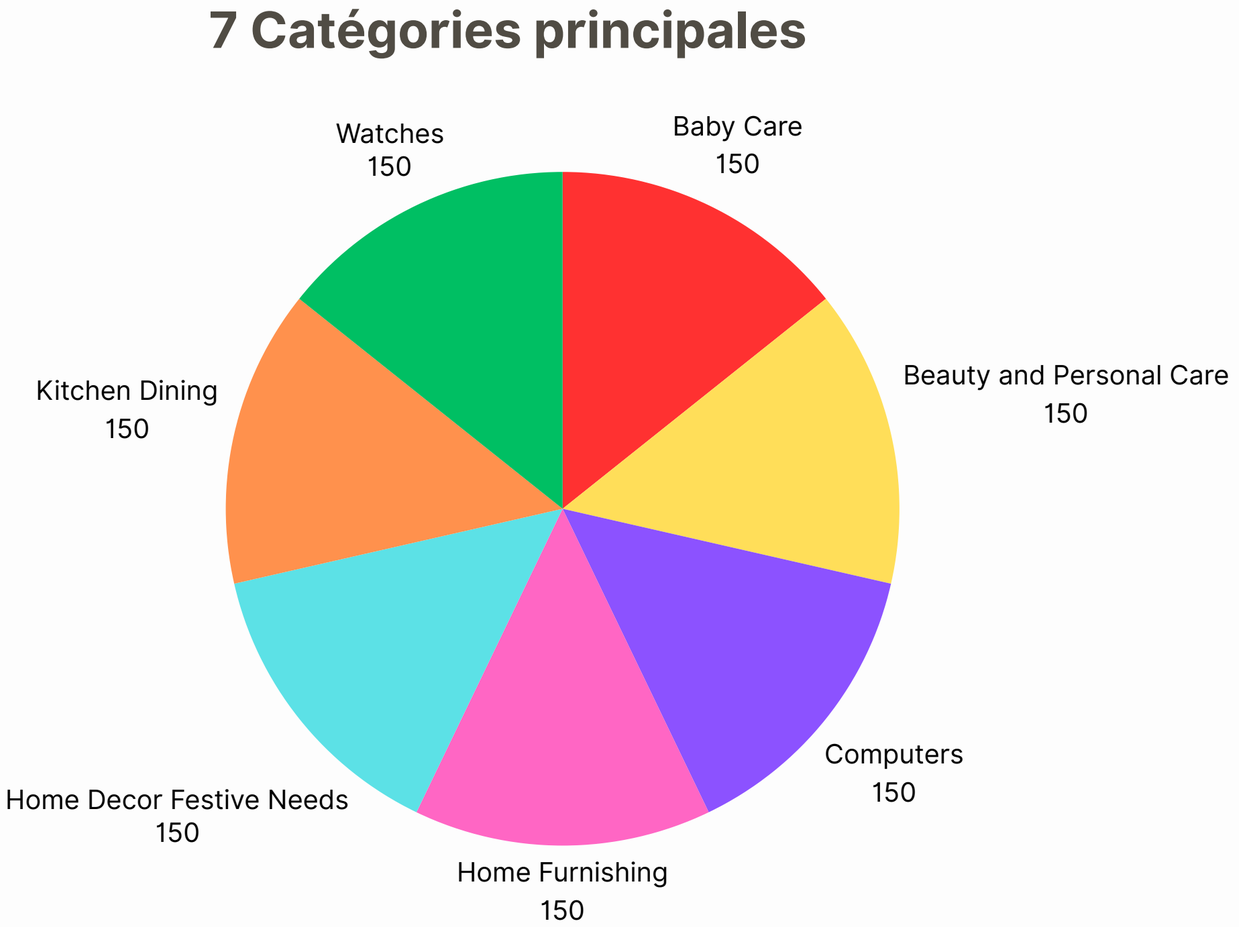
LES DONNÉES

1. Fichier de données :

- 15 variables (Nom, description, catégorie, ...)
- 1050 individus (Chaque ligne correspond à un produit)

2. Dossier d'images :

- 1050 images au format JPG (Chaque image correspond à un produit)



ÉTUDE DE FAISABILITÉ : TEXTE

DÉMARCHE

- Prétraitement des descriptions des produits
- Réduction des dimensions
- Clustering (K-means)
- Calcul de similarité entre les catégories réelles et les clusters
- Visualisation sur 2 dimensions

ACP

- Réduit les dimensions de 3395 à 509

T-SNE

- Réduit les dimensions de 509 à 2

ARI (Adjusted Rand Index)

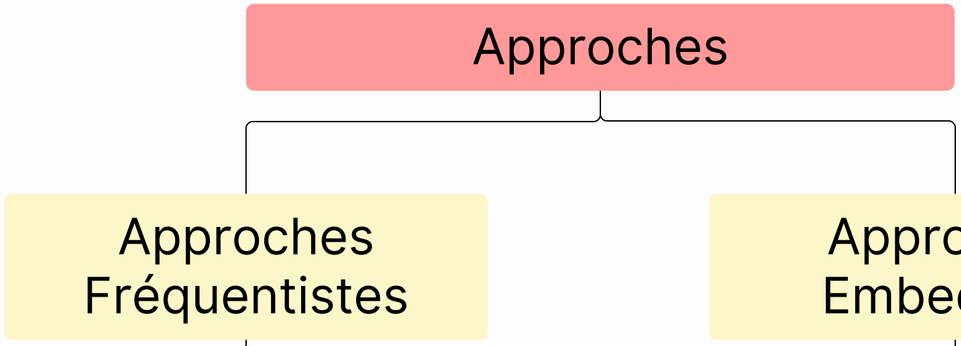
- Evalue la similarité entre deux clusters, compare avec les clusters de réference
- Un score élevé indique une concordance significative

PRÉTRAITEMENTS

Opération	
Fonction	• Convertit le texte en chaîne de caractères • Convertit le texte en minuscule • Supprime les balises HTMLou XML • Supprime les URL • Retire les chiffres • Étiquete les mots avec leur catégorie grammaticale • Filtre les mots pour garder seulement les noms et les verbes de plus de deux lettres

Opération	
RegxpTokenizer	• Divise le texte selon les règles données
nltk_stopwords	• Supprime les mots les plus courant en anglais
WordNetLemmatizer	• Utilise la base de données lexicale WordNet pour lemmatiser les mots, réduit les mots à leur forme canonique ou de base

NATURAL LANGUAGE PROCESSING



Approches Fréquentistes

- **CountVectorizer**

Calcul la fréquence des mots dans le texte

- **TF-IDF**

Calcule la fréquence des mots dans le texte par rapport à la fréquence de ces mots dans le corpus

Approches Embedding

- **Word2Vec**

Transforme les mots en vecteurs, les mots ayant des significations similaires seront proches les uns des autres

- **BERT**

Analyse les mots à gauche et à droite dans la phrase pour en comprendre le sens

- **USE**

Analyse le corpus de texte dans son ensemble pour en comprendre le sens

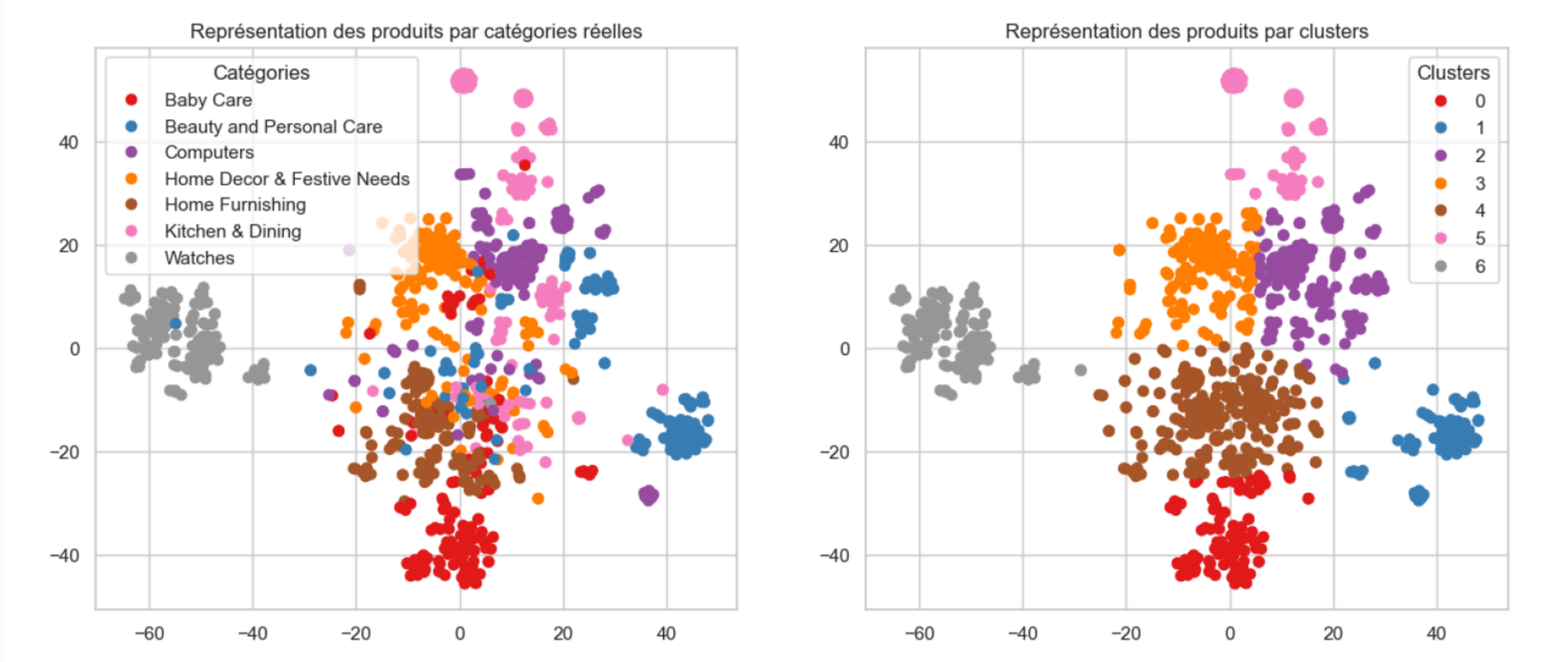
RÉSULTATS

Méthode	Score ARI
CountVectorizer	0.3982
TD-IDF	0.2804
Word2Vec	0.3196
BERT	0.3296
USE	0.3988

- Privilégier **CountVectorizer** si on a des ressources limitées, si nos descriptions sont courtes et simples

- Privilégier **USE** si on des descriptions de produits plus complexes, et si on dispose des ressources nécessaires

VISUALISATION : COUNTVECTORIZER



MATRICE DE CONFUSION : COUNTVECTORIZER

ÉTUDE DE FAISABILITÉ : IMAGES

PRÉTRAITEMENTS

Opération	
Couleurs	• Charge l’image en niveaux de gris
equalizeHist	• Egalise l’histogramme de l’image
detectAndCompute	• Détecte les points clés et calcule les descripteurs
Résolution	• Standardise les images en 224*224

COMPUTER VISION

Approches

Générateur de descripteurs

Réseaux de Neurones

- **SIFT**

Détecte et décrit les points d'intérêt (keypoints) dans les images, efficace pour détecter des points caractéristiques robustes et distinctifs dans les images

- **CNN Transfer Learning**

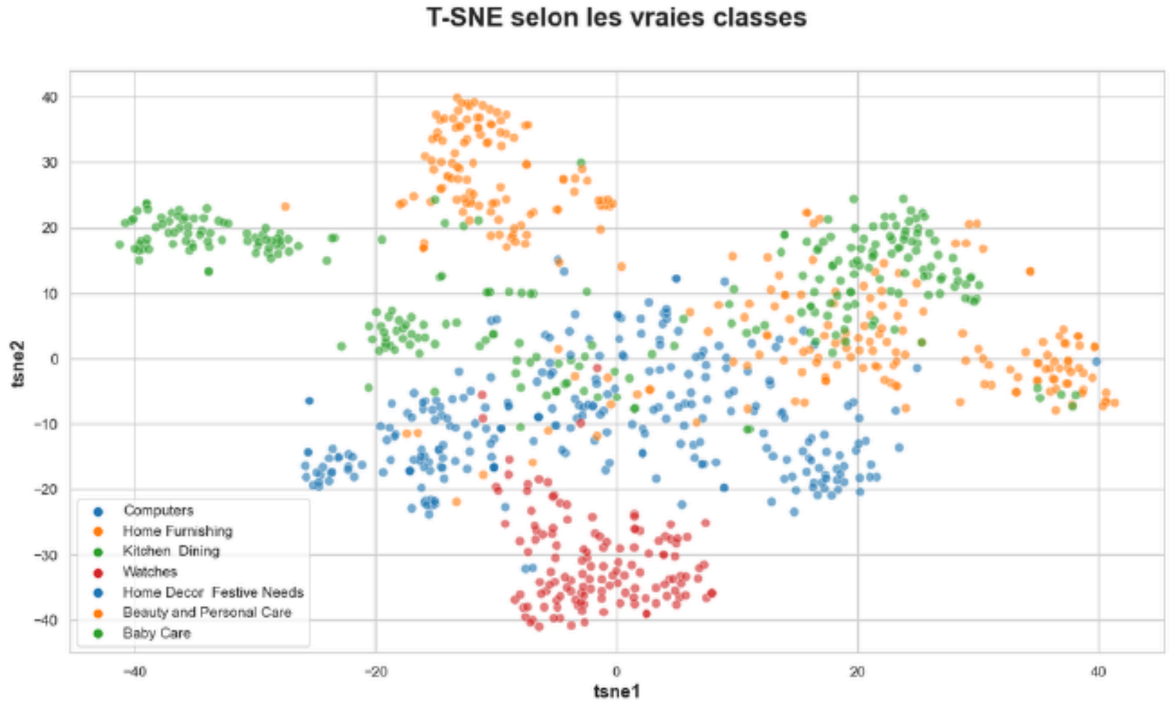
Utilise un modèle pré-entraîné sur un large jeu de données (comme ImageNet) et l'adapte à une nouvelle tâche avec un jeu de données plus petit

RÉSULTATS

Méthode	Score ARI
SIFT	0.0612
CNN Transfer Learning	0.4831

- Privilégier **CNN Transfer Learning** bien qu'il nécessite des ressources de calcul importantes (GPU)

VISUALISATION : CNN TRANSFER LEARNING



IMAGES : CLASSIFICATION SUPERVISÉE

DÉMARCHE

- Prétraitement des images des produits
- Séparation en train et test
- Création des modèles de CNN
- Création du Callback
- Mesure de la précision
- Visualisation

ModelCheckpoint :

- Sauvegarde le modèle à des intervalles réguliers, après chaque époque ou lorsque l'exactitude du modèle s'améliore

Accuracy

- Evalue la précision globale d’un modèle en indiquant la proportion d’éléments correctement classés
- Un score élevé indique une meilleure précision

RÉSEAUX DE NEURONES CONVOLUTIFS (CNN)

Approches

Modèle VGG16

Séquences de couches de convolution et de pooling suivies de couches entièrement connectées, extrayant des caractéristiques locales et réduisant progressivement la dimension

- Classification supervisée simplifiée
- Méthode avec augmentation des données pour l’entrainement du modèle
- ImageDataGenerator avec augmentation de donnée

Modèle VGG19

- Méthode avec augmentation des données pour l’entrainement du modèle

Modèle RESNET50

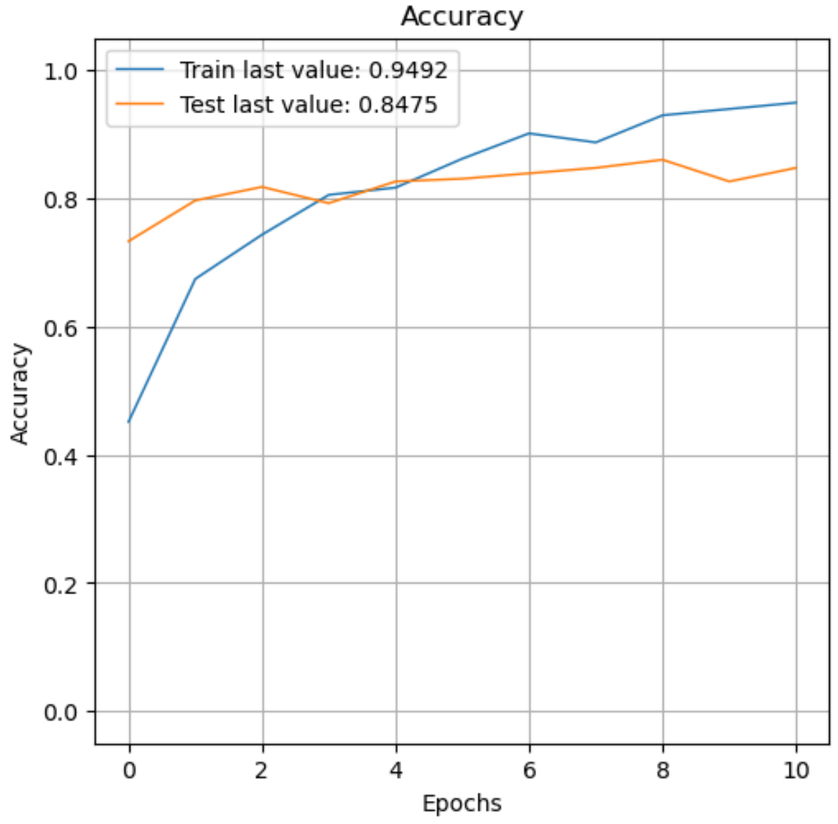
Utilise des blocs résiduels avec connections directes pour faciliter l'appr des réseaux profonds

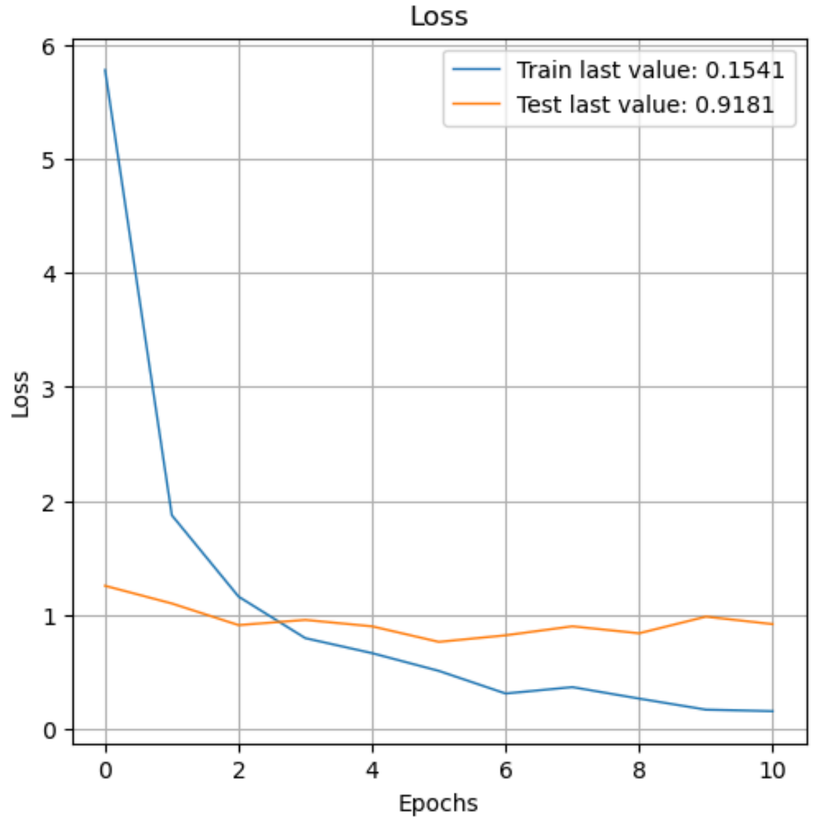
- Méthode avec augmentation des données pour l’entrainement du modèle

RÉSULTATS

Méthode	Validation Accuracy	Test Accuracy	Temps entrainement	Temps validation
VGG16 Classification supervisée simplifiée	0.8608	0.5810	369s	15s
VGG16 ImageDataGenerator avec augmentation	0.8347	0.5429	47s	16s
VGG16 approche avec augmentation intégrée	0.8644	0.5714	61s	25s
VGG19 approche avec augmentation intégrée	0.8305	0.5333	63s	22s
RESNET50 approche avec augmentation intégrée	0.8729	0.6000	27s	9s

RESNET50 APPROCHE AVEC AUGMENTATION DES DONNÉES





- **Performance :**

Le modèle généralise bien aux données non vues, mais il y a un écart notable par rapport à la précision d'entraînement

- **Surapprentissage :**

Une perte de test plus élevée par rapport à la perte d'entraînement indique un certain degré d’overfitting

TEST D'UNE API

DÉMARCHE

- Paramètrage d’un script python sur RapidAPI
- Importation JSON des données
- Transformation en Dataframe
- Requeter sur la catégorie Champagne
- Filtrer en assurant les règles RGPD
- Extraire au format CSV

API Edamam

- Interface de programmation d'application qui permet aux développeurs d'accéder à des données nutritionnelles, des recettes et des informations alimentaires

RGPD

- Utiliser exclusivement les données strictement nécessaires à notre utilisation

AFFICHAGE DES 10 PREMIERS PRODUITS LABELISÉS “CHAMPAGNE”

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
7	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
8	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN
9	food_a79xmnya6togreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour; Salt; Pepper; Boneless, Skinless Chicke...	NaN

CONCLUSIONS

- La classification automatique des produits est possible, la faisabilité a été démontrée pour le texte et l'image avec une approche simple et peu gourmande en ressources (CountVectorizer)
- Les résultats de la classification supervisée sont probants sur un modèle efficace (RESNET50)

- Il est possible pour augmenter la précision, d'utiliser des approches multi-modales en conjuguant données des textes et des images

MERCI POUR VOTRE ATTENTION

AOUT 2024 / LOKMAN AALIOUI