CLASSIFIER AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

Etude et Présentation: **XXX**

21/01/2024

PROBLÉMATIQUE ACTUELLE DE L'ATTRIBUTION MANUELLE DES CATÉGORIES

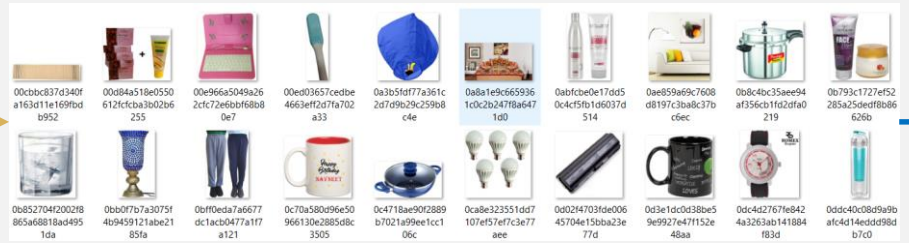
- **Contexte**: Lancement d'une marketplace e-commerce où des vendeurs externes proposeront leurs produits à la vente.
- **Défi**: Attribution manuelle des catégories par les vendeurs - peu fiable et non productive.
- **Objectif global**: Améliorer l'expérience des vendeurs et des clients en automatisant la classification des articles pour faciliter la mise en ligne et la recherche.

PLAN DE LA PRESENTATION

1. PROBLÉMATIQUE ET EXPLORATION DES DONNÉES
2. VARIABLES PERTINENTES
3. ETUDE DE FAISABILITÉ ET SES RESULTATS
4. CLASSIFICATION SUPERIVISÉE DES IMAGES
5. TRANSFER LEARNING: CONCEPT D'UNE TECHNIQUE RECENTE
6. API DE COLLECTE DES DONNÉES
7. BILAN ET PERSPECTIVES

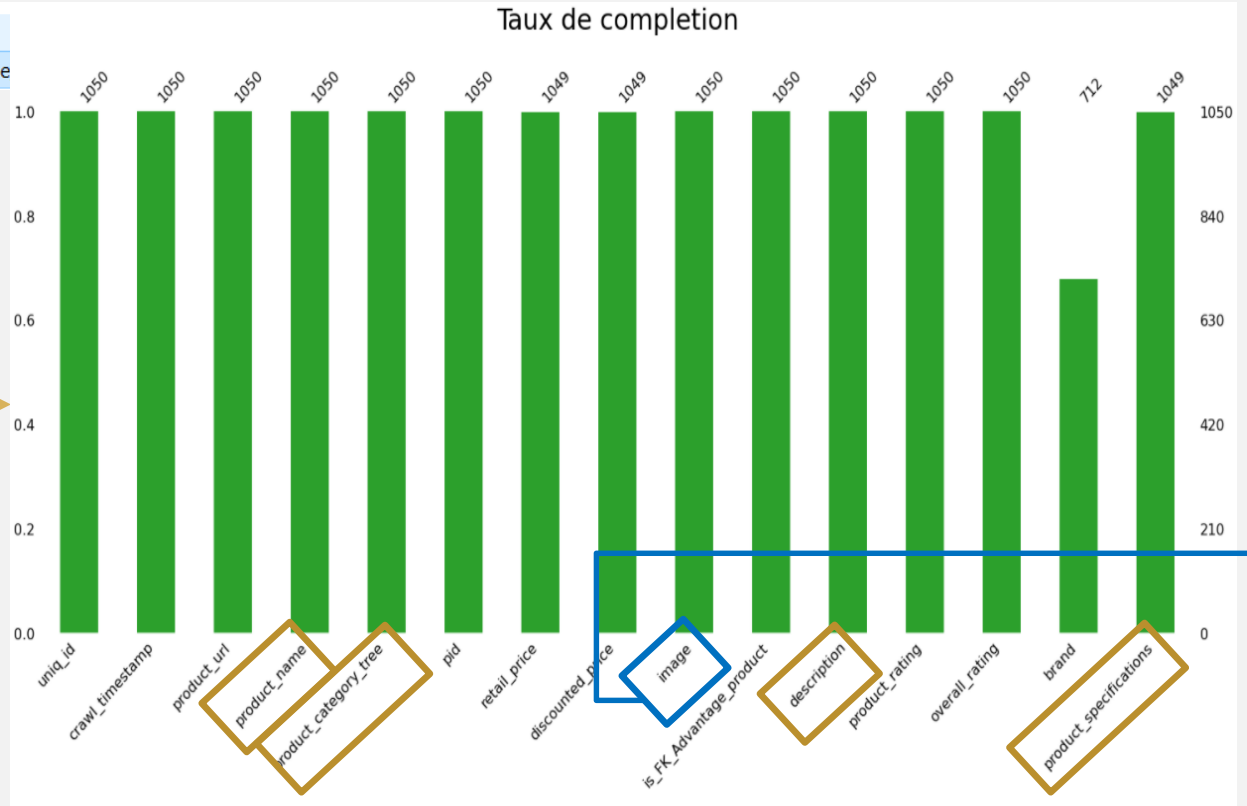
JEU DE DONNEES: EXPLORATION

1050 images



Images

flipkart_com-ecommerce_sample



JEU DE DONNEES: VARIABLES PERTINENTES

Dénomination du produit	Description du produit	Spécifications du produits
• Jack klein BlackLed Digital Watch - For Boys	• Jack klein BlackLed Digital Watch - For Boys - Buy Jack klein BlackLed Digital Watch - For Boys BlackLed Online at Rs.150 in India Only at Flipkart.com. - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!	• {"product_specification"=>[{"key"=>"Chronograph", "value"=>"No"}, {"key"=>"Date Display", "value"=>"No"}, {"key"=>"Altimeter", "value"=>"No {"key"=>"Water Resistant", "value"=>"No"}, {"key"=>"Dial Color", "value"=>"Black"}]}

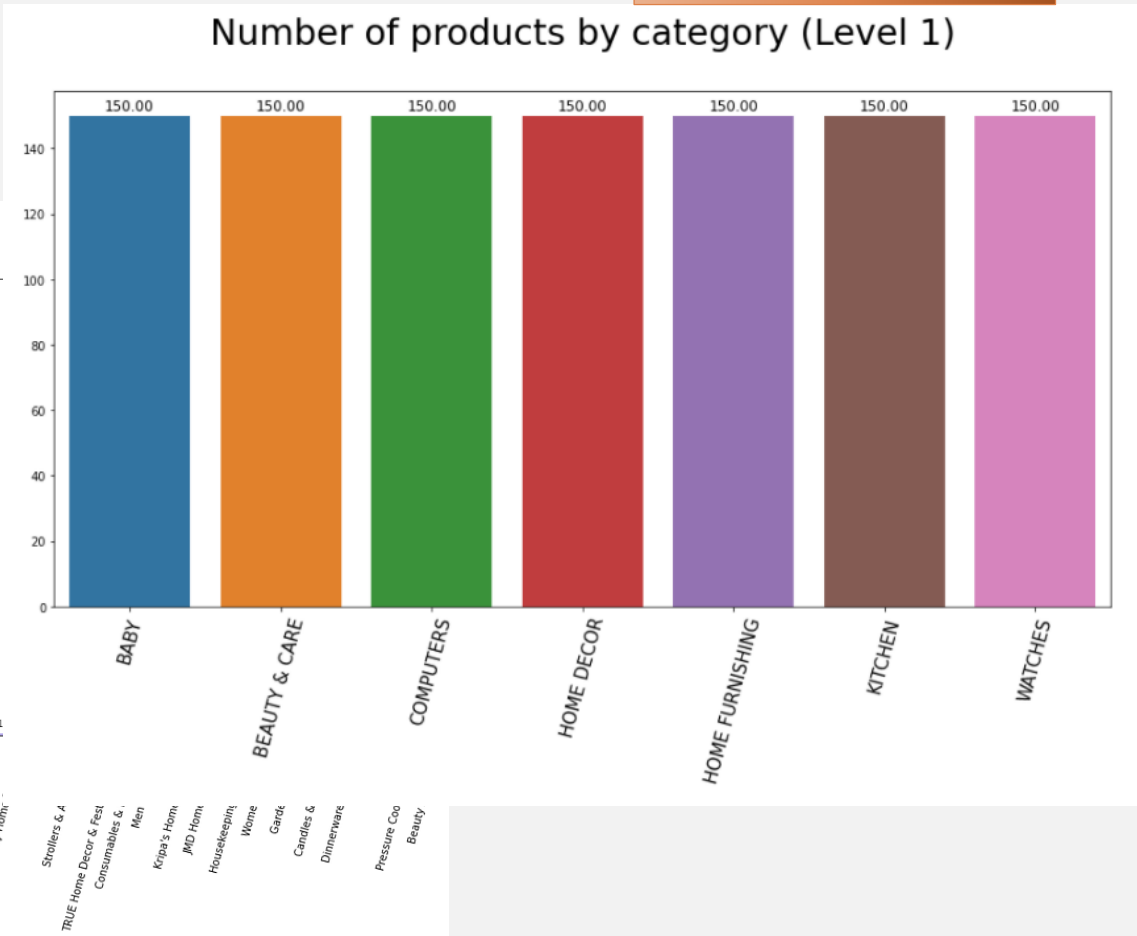
Nom + Description	Nom + Spécifications	Nom + Description + Spécifications

JEU DE DONNEES: Catégories des produits

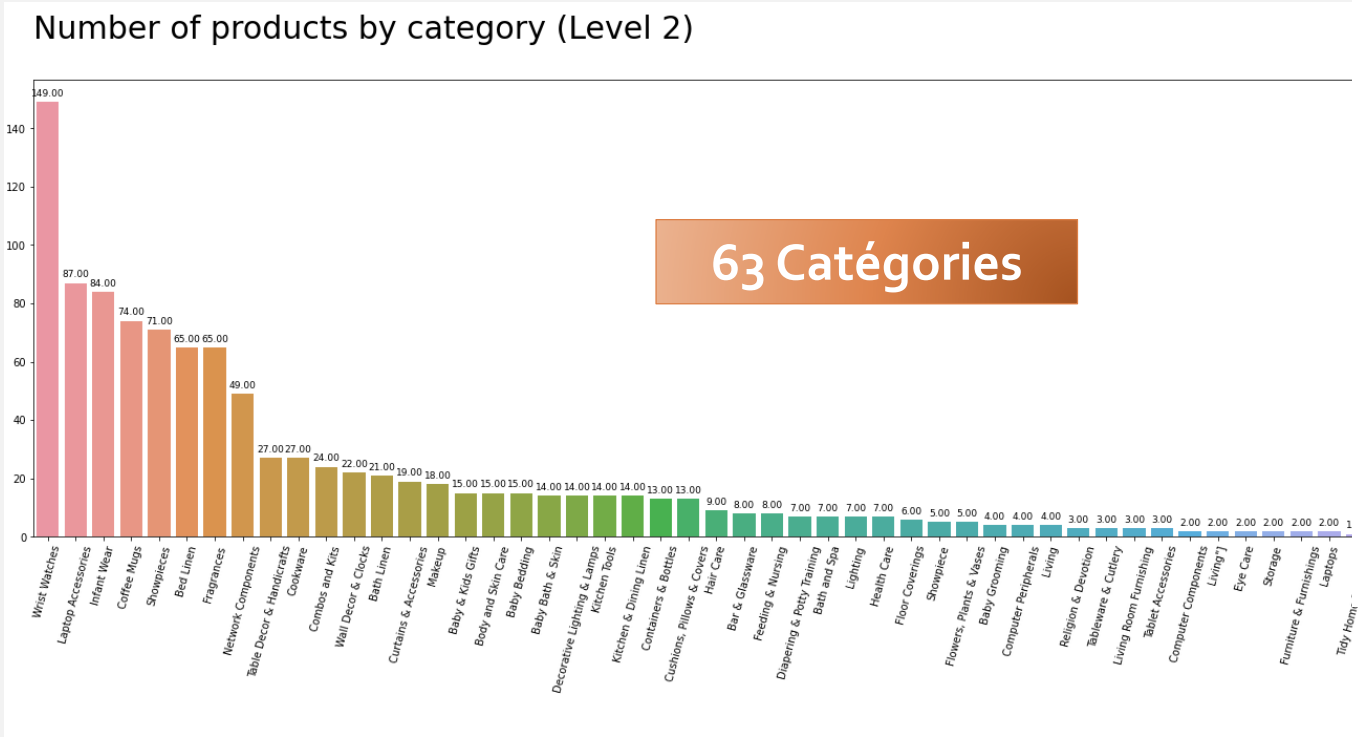
Watches >> Wrist Watches >> Jack klein Wrist Watches

NIVEAU 1 NIVEAU 2 NIVEAU 3

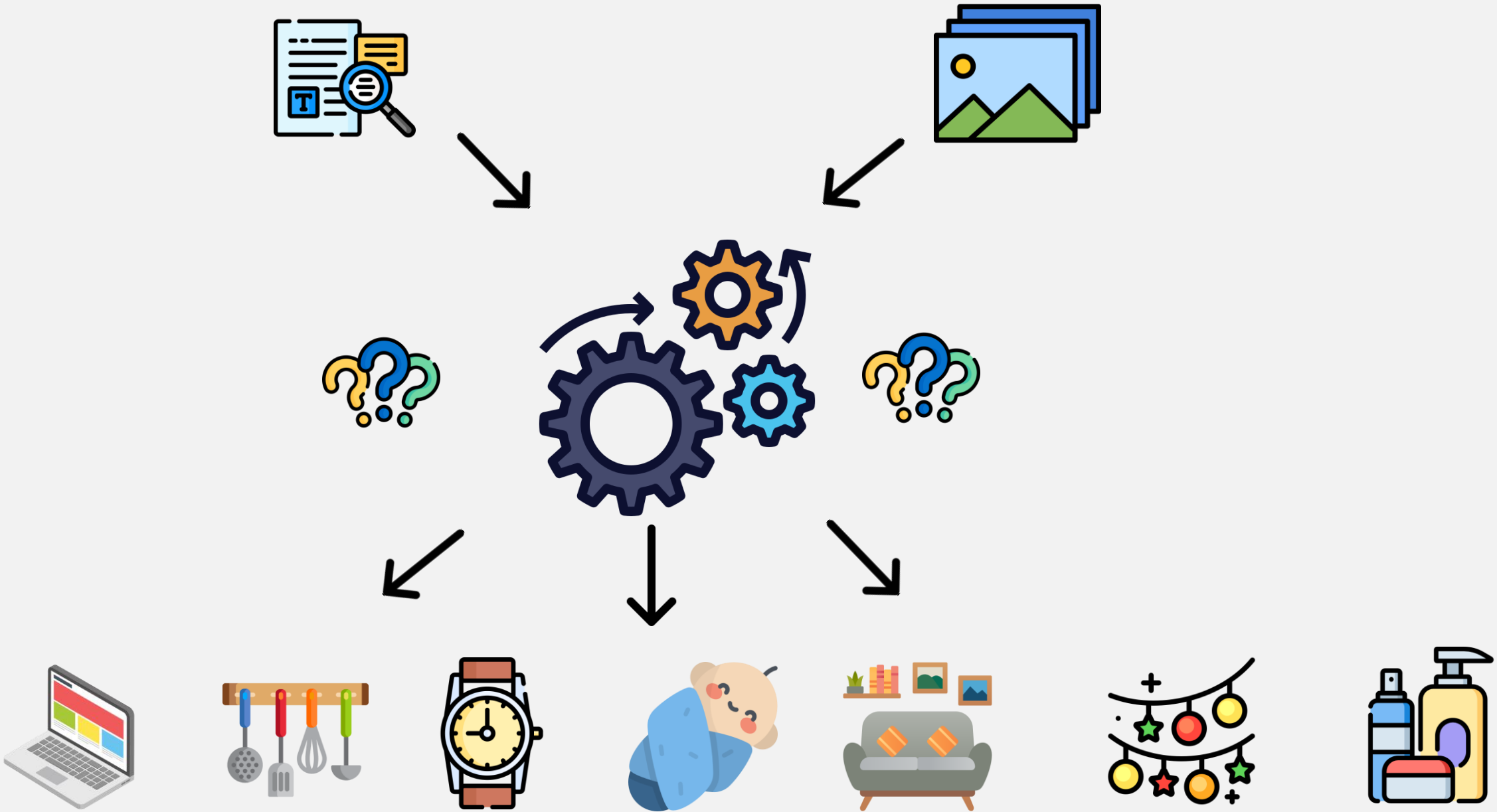
7 Catégories



63 Catégories



ETUDE DE FAISABILITE: PROBLEMATIQUE !



ETUDE DE FAISABILITE: BAG OF WORDS (BOW)

Pre-traitement → **Bag of Words: Extraction des features** → **Réduction de dimensionnalité** → **Classement par K-means** → **Evaluation des résultats par rapport aux catégories réelles**

Machine Learning is fascinating!

Nettoyage

machine learning is fascinating

Tokenization

["machine", "learning", "is", "fascinating"]

Suppression des stopwords et Lemmatisation

["machine", "learn", "fascinate“]

Recomposition des tokens

"machine learn fascinate"

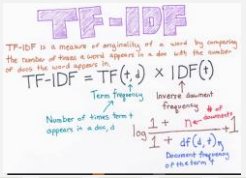
```
1. "Machine learning is fascinating."
2. "Learning algorithm is powerful."
```

Count Vectorizer

Phrase	Machine	Learning	Algorithms	Powerful	Fascinating	is
"Machine learning is fascinating."	1	1	0	0	1	1
"Learning algorithms are powerful."	0	1	1	1	0	1

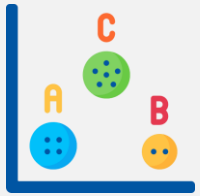
TF-IDF : Term Frequency Inverse Document Frequency

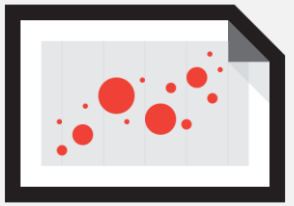
Terme	TF (Phrase 1)	TF (Phrase 2)	IDF (Global)	TF-IDF (Phrase 1)	TF-IDF (Phrase 2)
Machine	1/4 = 0.25	0/4 = 0	log(2/1) = 0.693	0.25 * 0.693 = 0.17325	0 * 0.693 = 0
Learning	1/4 = 0.25	1/4 = 0.25	log(2/2) = 0	0.25 * 0 = 0	0.25 * 0 = 0
Algorithm	0/4 = 0	1/4 = 0.25	log(2/1) = 0.693	0 * 0 = 0	0.25 * 0.693 = 0.17325
Powerful	0/4 = 0	1/4 = 0.25	log(2/1) = 0.693	0 * 0 = 0	0.25 * 0.693 = 0.17325
Fascinating	1/4 = 0.25	0/4 = 0	log(2/1) = 0.693	0.25 * 0.693 = 0.17325	0 * 0.693 = 0
is	1/4 = 0.25	1/4 = 0.25	log(2/2) = 0	0.25 * 0 = 0	0.25 * 0 = 0



- Truncated SVD (99% de variance expliquée).
- TSNE (2 composantes).

- K-means





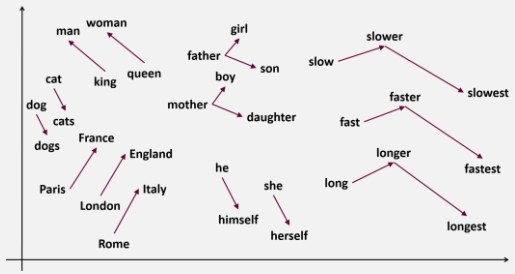
ARI (Adjusted Rand Index) => 1

ETUDE DE FAISABILITE: EMBEDDINGS

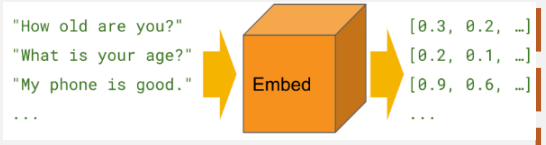
Pre-traitement → **Extraction des features** → **Réduction de dimensionnalité** → **Classement par K-means (Non supervisé: K = 7)** → **Evaluation des résultats par rapport aux catégories réelles**

- **Nettoyage** ✓
- **Tokenization** ✗
- **Suppression des stopwords et Lemmatisation** ✗
- **Recomposition des tokens** ✗

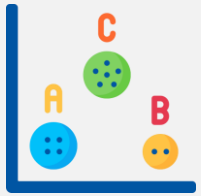
➢ **Word2Vec**



➢ **BERT / USE**



➢ TSNE (2 composantes)



ARI (Adjusted Rand Index) => 1

ETUDE DE FAISABILITE: RESUME

Nom + Description + Spécifications

Etude	Méthode de prétraitement	Dimensions des features	Réduction de dimensionnalité Truncated SVD	Réduction de dimensionnalité TSNE	Classement K-means Et visualisation TSNE
Vectorisation TF-IDF/TF	Avec Lemmatisation et suppression des stopwords	(1050, 4872)	(1050, 325)	(1050, 2)	OK
WORD2VEC	Avec Lemmatisation et suppression des stopwords	(1050, 300)	-	(1050, 2)	OK
BERT	Sans Lemmatisation ou suppression des stopwords	(1050, 512)	-	(1050, 2)	OK
USE	Sans Lemmatisation ou suppression des stopwords	(1050, 512)	-	(1050, 2)	OK

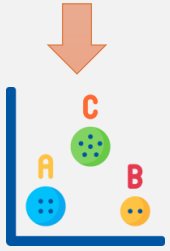
ETUDE DE FAISABILITE: BAG OF VISUAL WORDS (BOVW)

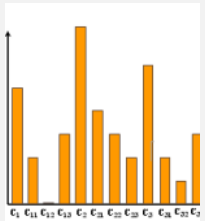
Pre-traitement

- Amélioration du contraste.
- Traitement noir et blanc.

Extraction des features

SIFT (Scale-Invariant Feature Transform)

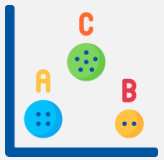




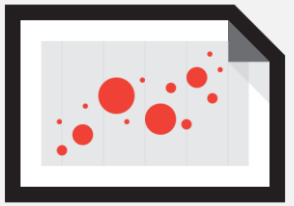
Réduction de dimensionnalité

- PCA (99% de variance expliquée).
- TSNE (2 composantes).

Classement par K-means (Non supervisé: K = 7)



Evaluation des résultats par rapport aux catégories réelles

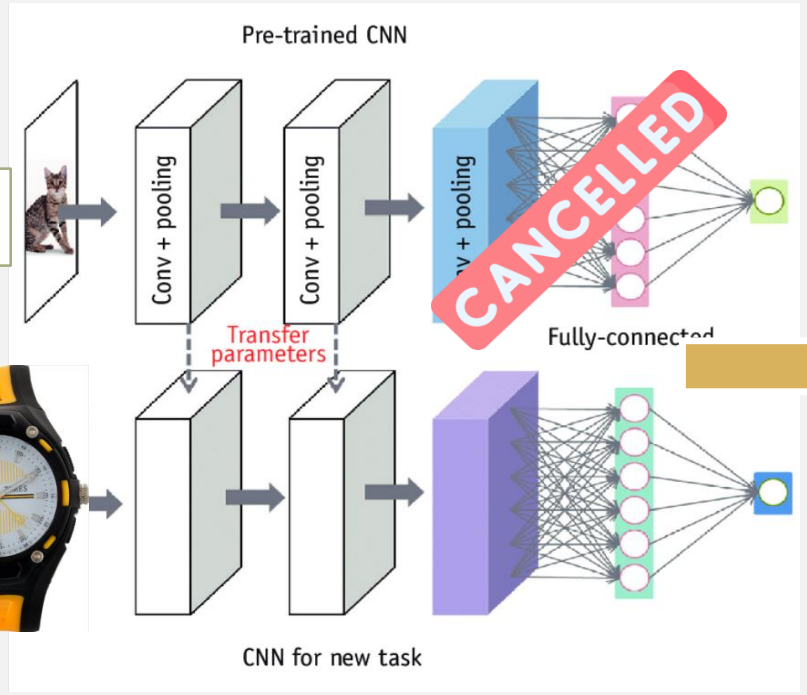


ARI (Adjusted Rand Index) => 1

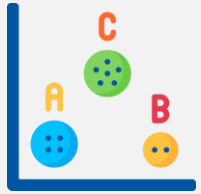
ETUDE DE FAISABILITE: TRANSFER LEARNING (CNN)

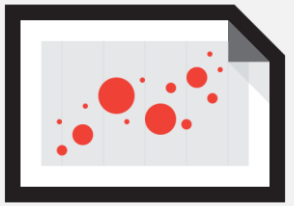
Transfer Learning: VGG16 / RESNET50 → **Réduction de dimensionnalité** → **Classement par K-means (Non supervisé: K = 7)** → **Evaluation des résultats par rapport aux catégories réelles**

1,2 million



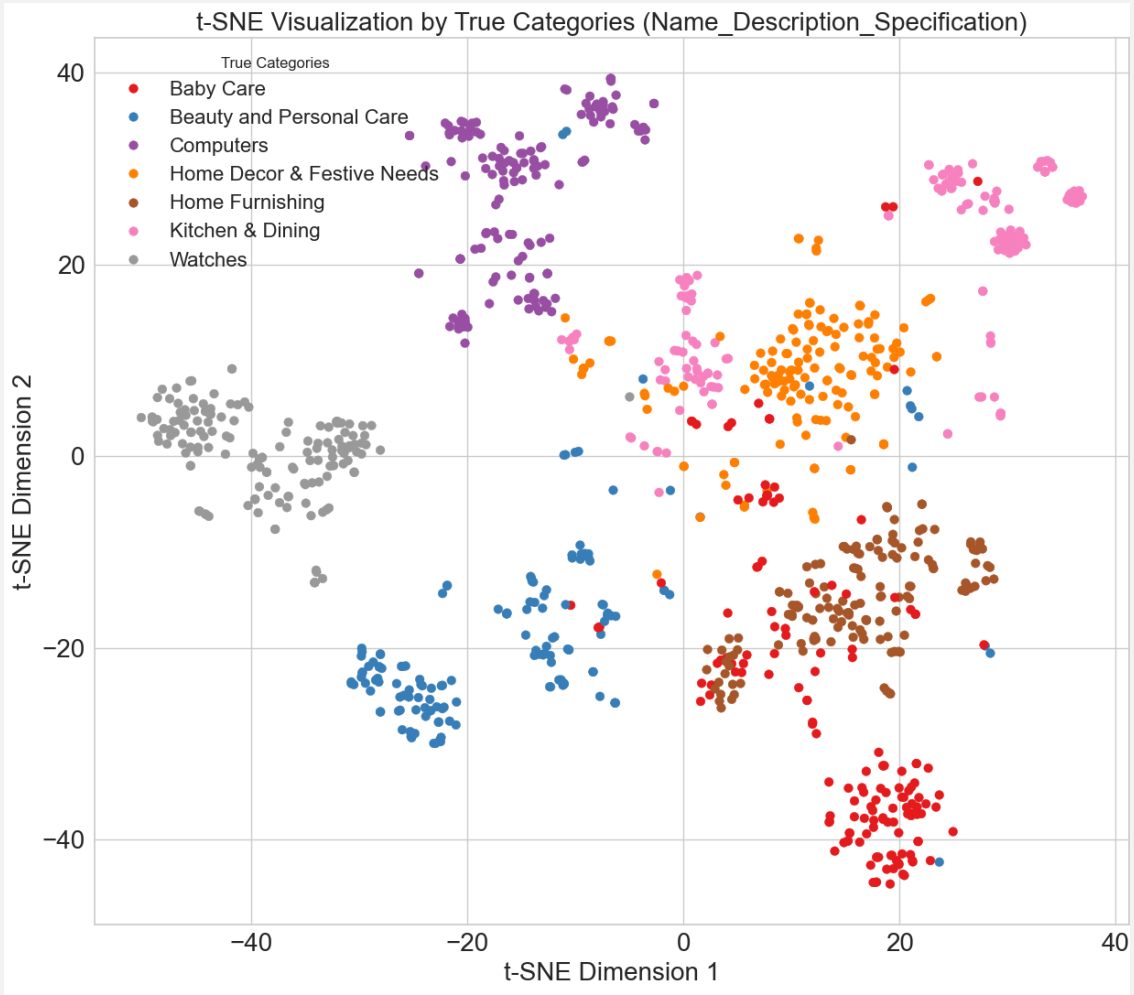
- PCA (99% de variance expliquée).
- TSNE (2 composantes).

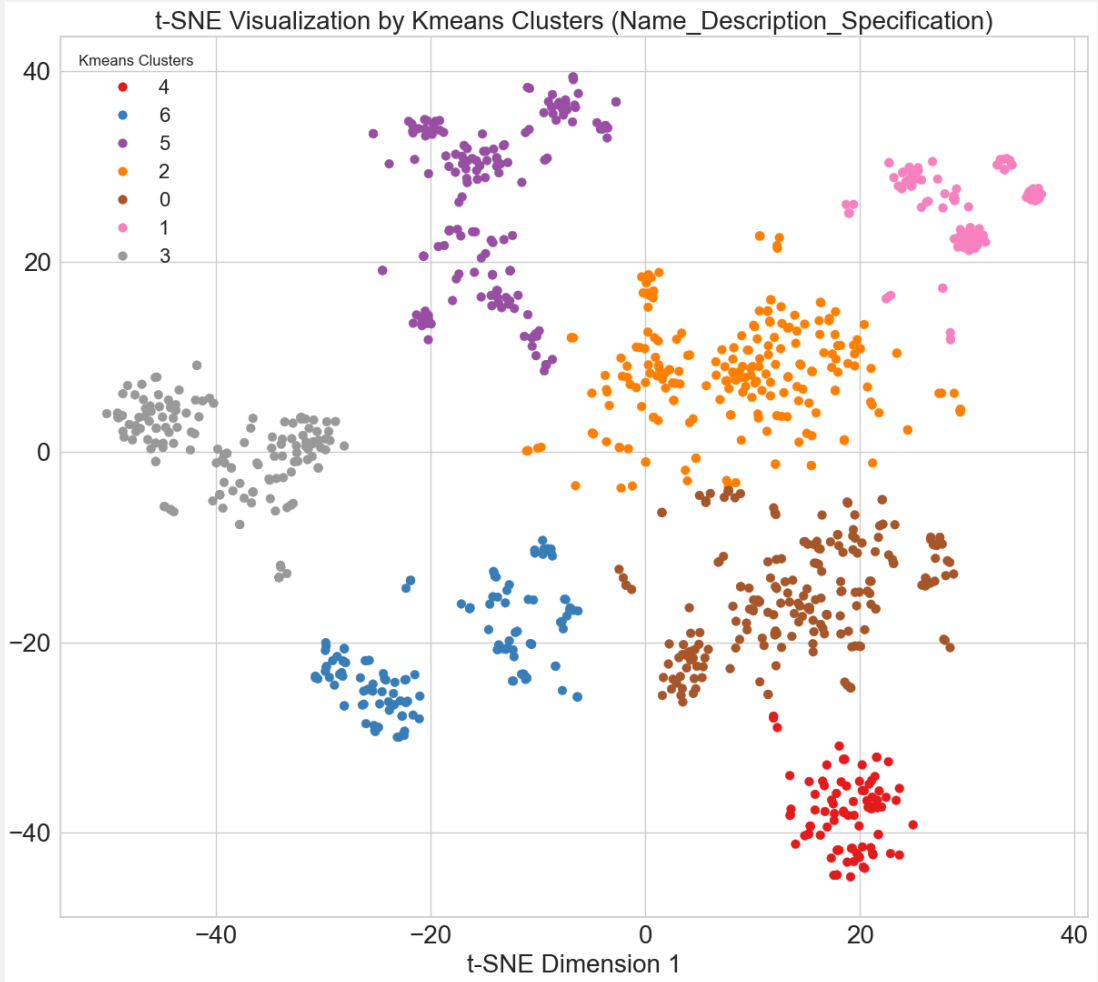




ARI (Adjusted Rand Index) => 1

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS DE TEXTES

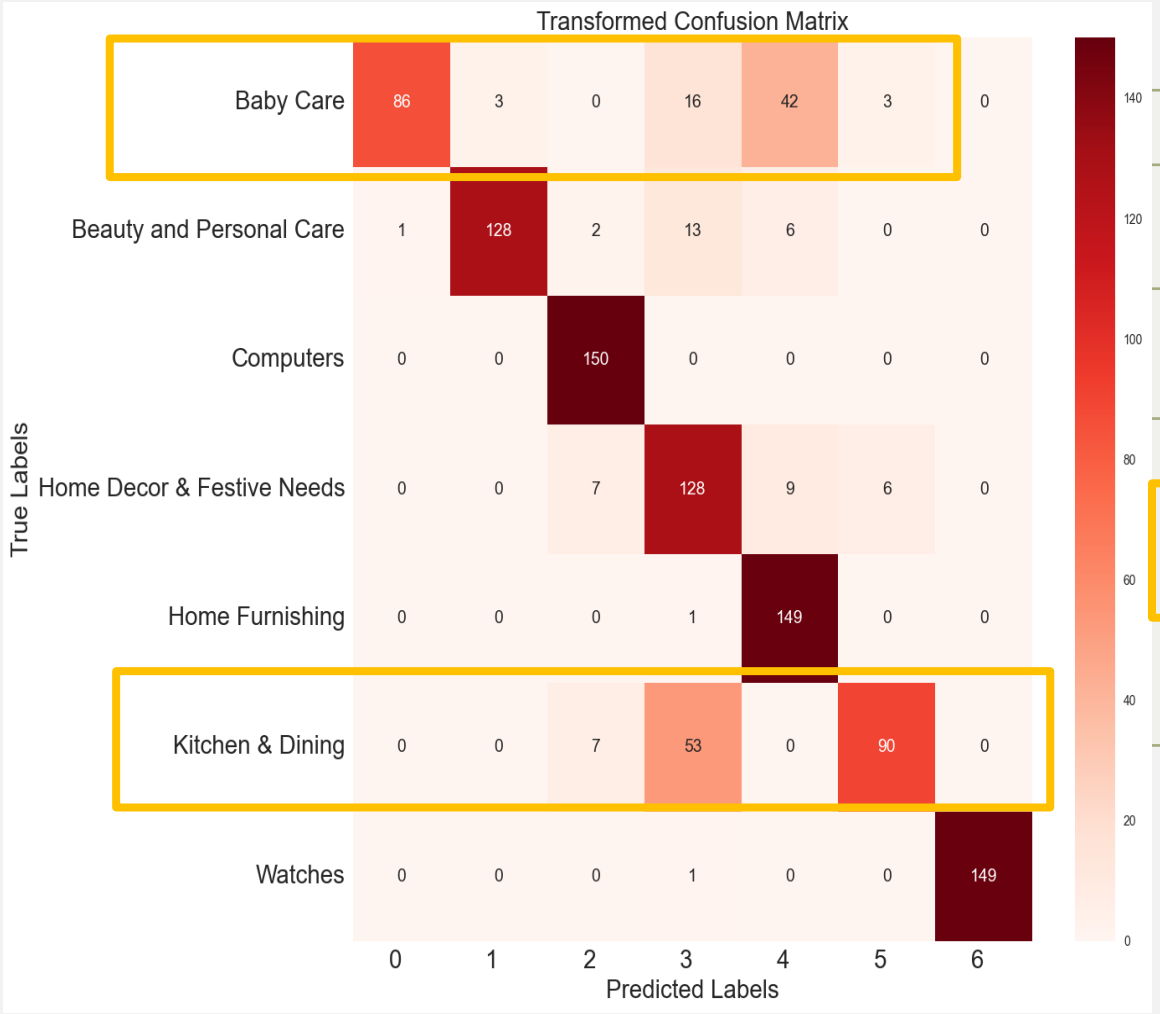




USE

Nom + Description + Spécifications

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS DE TEXTES



Model	Corpus	ARI	Time (s)
BERT	Name_Specification	0.62	914
Count Vectorizer	Name_Specification	0.55	26
TF-IDF	Name_Specification	0.59	27
USE	Name_Description_Specification	**0.69**	30
Word2Vec	Name_Description_Specification	0.64	32

1

USE

1

Nom + Description + Spécifications

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS DE TEXTES

'Key Features of Sparklings Cotton Geometric Double Bedsheet Best Fit For:-Large Size Double Bed (6.5 ft X 6 ft) Thread Count:160 TC Package Content:1 Double Bedsheet With 2 Pillow Cover,Sparklings Cotton Geometric Double Bedsheet (double bedsheet with pillow covers, Multicolor) ...

Baby Care >> Baby Bedding >>
Baby Bedsheets >> Sparklings
Baby Bedsheets >> …

Home furnishing

Classe réelle

Classe prédite

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS DE TEXTES

'Specifications of SOS COMPUTRISED LCD INTRFERENTIAL UNIT-125PROG. MUSCLE SPASM Electrotherapy Device (SOS-121) General Treatment Time 30 mnts or as desired Display Type LCD Number of Programmable Profiles 125 Designed For mussel or nerve contracture Electrodes 8 Warranty Covered...'

Computers

Beauty and Personal Care >> Health Care >> Health Care Devices >> Electrotherapy

Classe prédite

Classe réelle

Nutcase brings you designer water bottle . Nutcase Bottles reflect your style, your attitude. This aluminium water bottle comes with a harness hook , waterproof sticker wrap and a fresh , quirky playful design . Nutcase brings you designer water bottle….

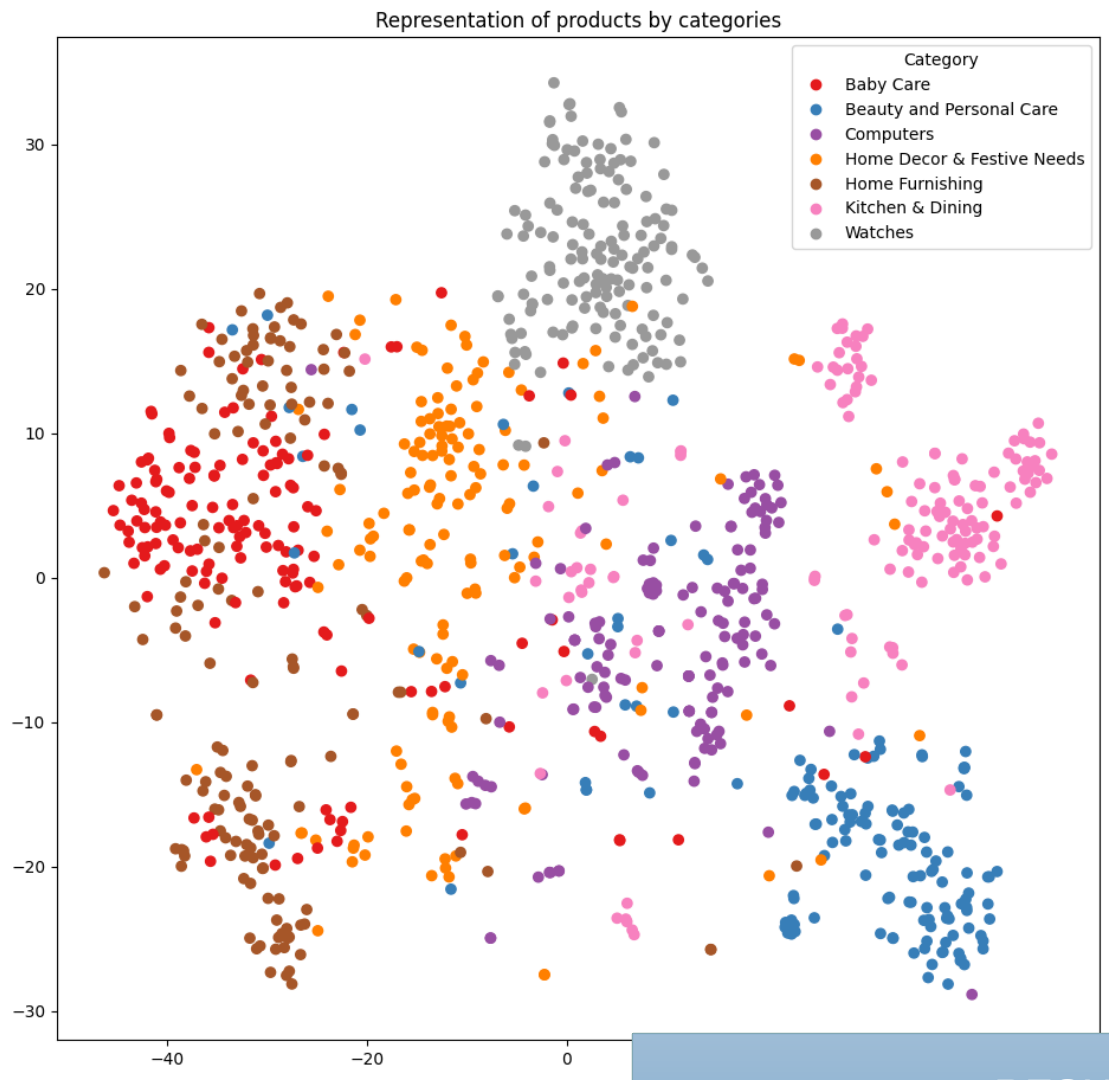
Home decor & Festive needs

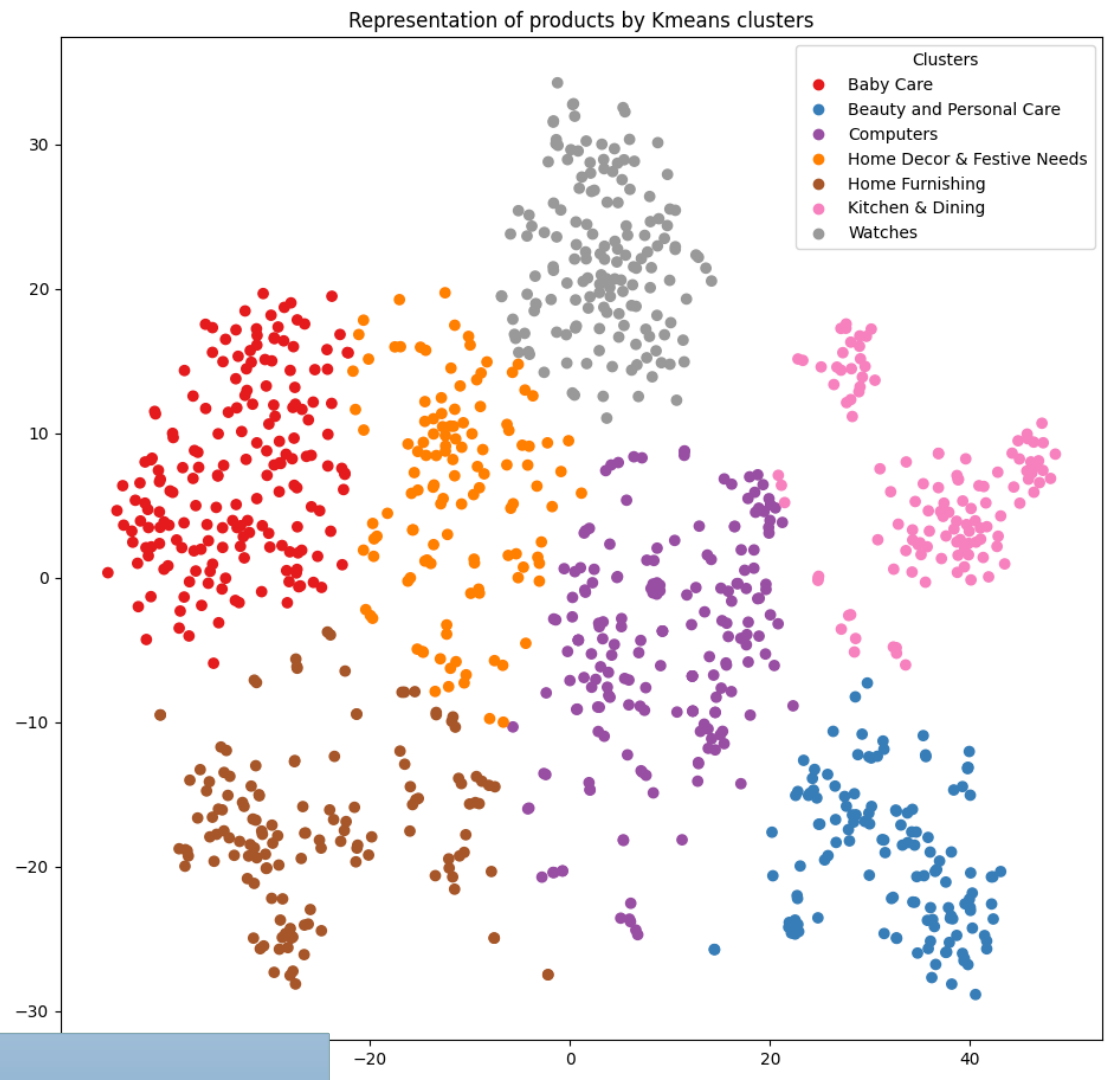
Kitchen & Dining

Classe prédite

Classe réelle

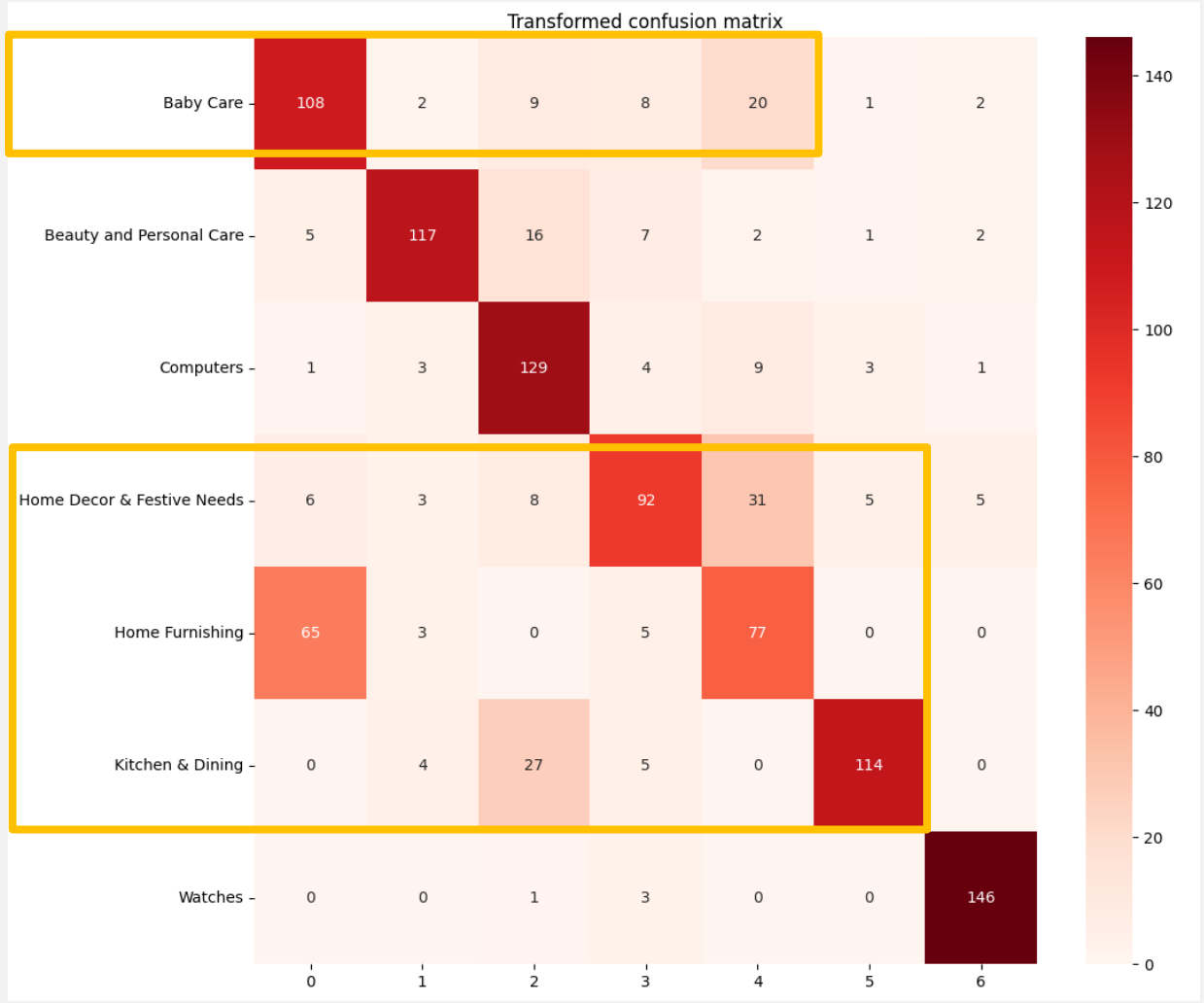
ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS D'IMAGES





RESNET50

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS D'IMAGES



Model	ARI	SILOUHETTE	Time (s)
RESNET50	0.54	0.47	914
VGG16	0.46	0.48	26
SIFT	0.05	0.35	27

RESNET50

ETUDE DE FAISABILITE : RESULTATS DES TRAITEMENTS DE TEXTES

Home Furnishing >> Kitchen & Dining Linen >> Kitchen Linen Sets

Kitchen & dining

Classe réelle

Classe prédite

Home Furnishing

Home Decor & Festive Needs

Classe prédite

Classe réelle

CLASSEMENT SUPERVISE : RESULTATS DES TRAITEMENTS D'IMAGES

- Training: 70%
- Validation: 15%
- Test: 15%

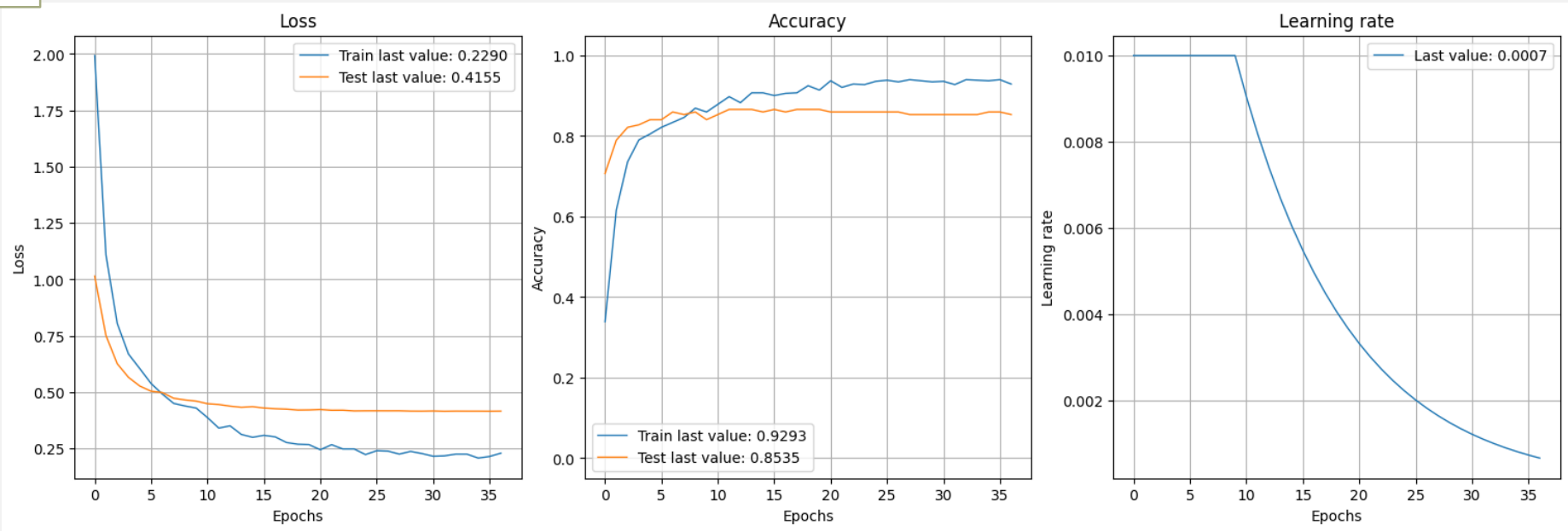
Model	epochs	dropout_rate	optimization	ARI	accuracy validation	accuracy test	loss validation	loss test	training_time (s)	delta_accuracy	delta_loss
ResNet50	17	0.5	adam	0.76	0.89	0.81	0.43	0.91	925.73	-0.08	0.48
ResNet50	37	0.5	sgd	0.68	0.85	0.77	0.42	0.70	2125.00	-0.09	0.28
ResNet50	7	0.5	rmsprop	0.68	0.85	0.78	0.51	0.97	499.98	-0.06	0.46
VGG16	16	0.5	adam	0.61	0.82	0.80	0.67	1.15	2362.39	-0.02	0.48
VGG16	50	0.5	sgd	0.53	0.76	0.71	1.08	1.72	6042.51	-0.06	0.64
VGG16	14	0.5	rmsprop	0.63	0.82	0.78	0.72	1.18	1494.09	-0.04	0.45

Optimisation:
- Adam,
- SGD,
- rmsprop

ARI : 0.68

Accarucy : 0.77

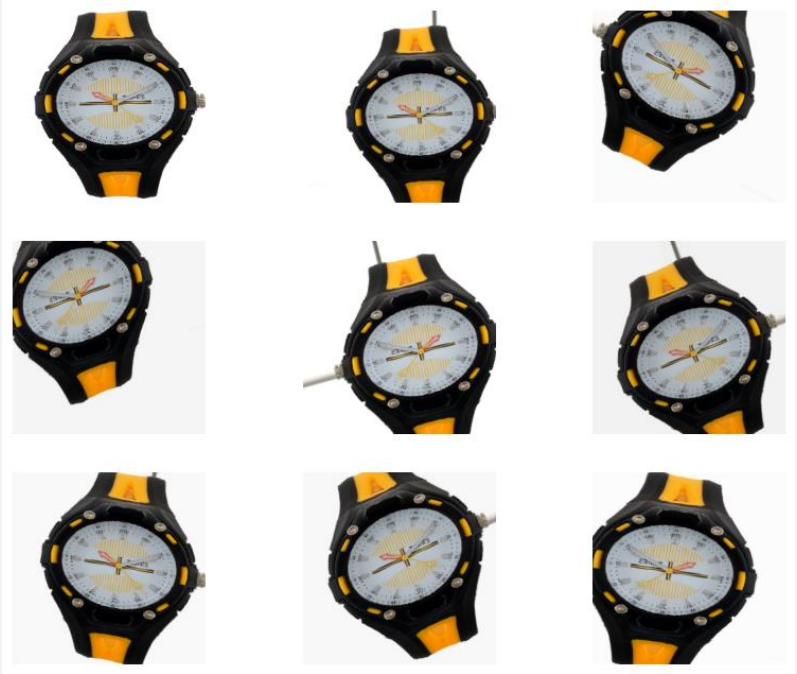
- Perte minimale: 0.42
- Sur-apprentissage minimisé



DATA AUGMENTATION

- Training: 70%
- Validation: 15%
- Test: 15%

Optimisation:
- Adam,
- SGD,
- rmsprop



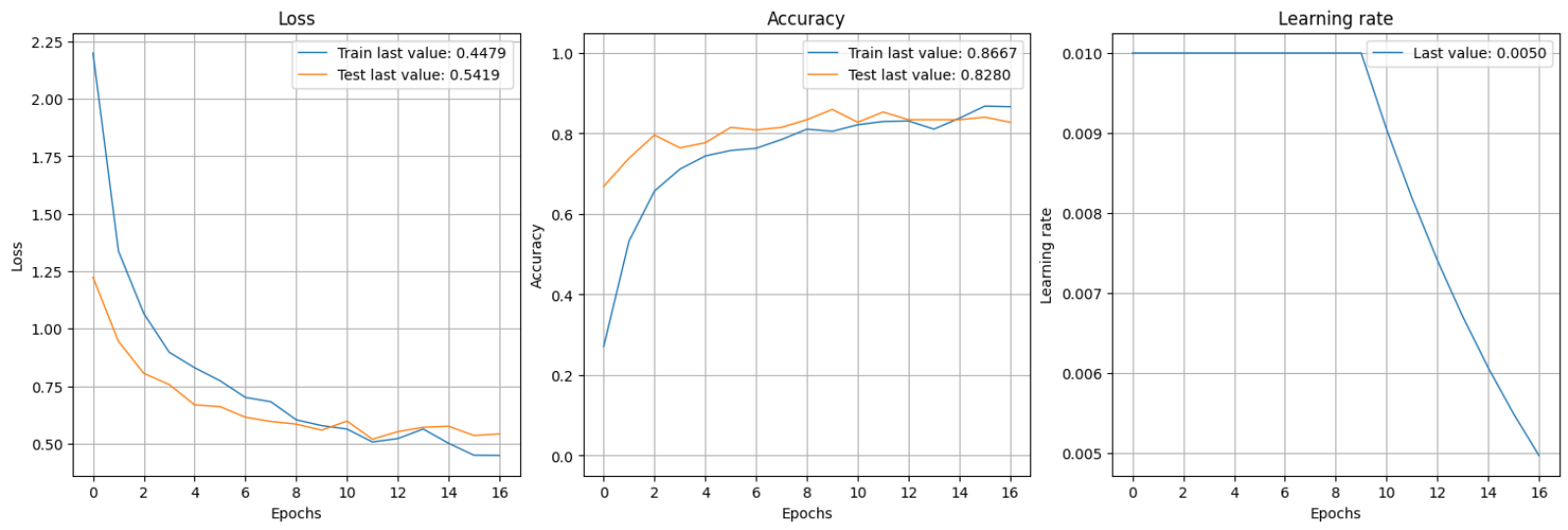
Rotation: 25° maxi,
Ratio de translation en hateur et largeur: 0.2,
Retournement: Oui,
Luminisité:[0.95,1.05],
Zoom =[0.9,1.1],

DATA AUGMENTATION: RESULTATS DES TRAITEMENTS D'IMAGES

Model	epochs	dropout_rate	optimization	accuracy validation	accuracy test	loss validation	loss test	training_time (s)	delta_accuracy	delta_loss
ResNet50	11	0.5	adam	0.87	0.85	0.49	0.58	986.42	-0.02	0.09
ResNet50	17	0.5	sgd	0.84	0.83	0.54	0.55	1456.04	-0.01	0.01
ResNet50	8	0.5	rmsprop	0.83	0.82	0.64	0.61	759.89	-0.01	-0.04
VGG16	10	0.5	adam	0.82	0.82	0.85	0.72	1637.68	0.01	-0.13
VGG16	14	0.5	sgd	0.78	0.78	0.86	0.79	2453.16	-0.00	-0.07
VGG16	12	0.5	rmsprop	0.85	0.84	0.93	0.69	2349.74	-0.01	-0.24

Accarucy : 0.77→ 0.83

Loss : 0.42 → 0.54



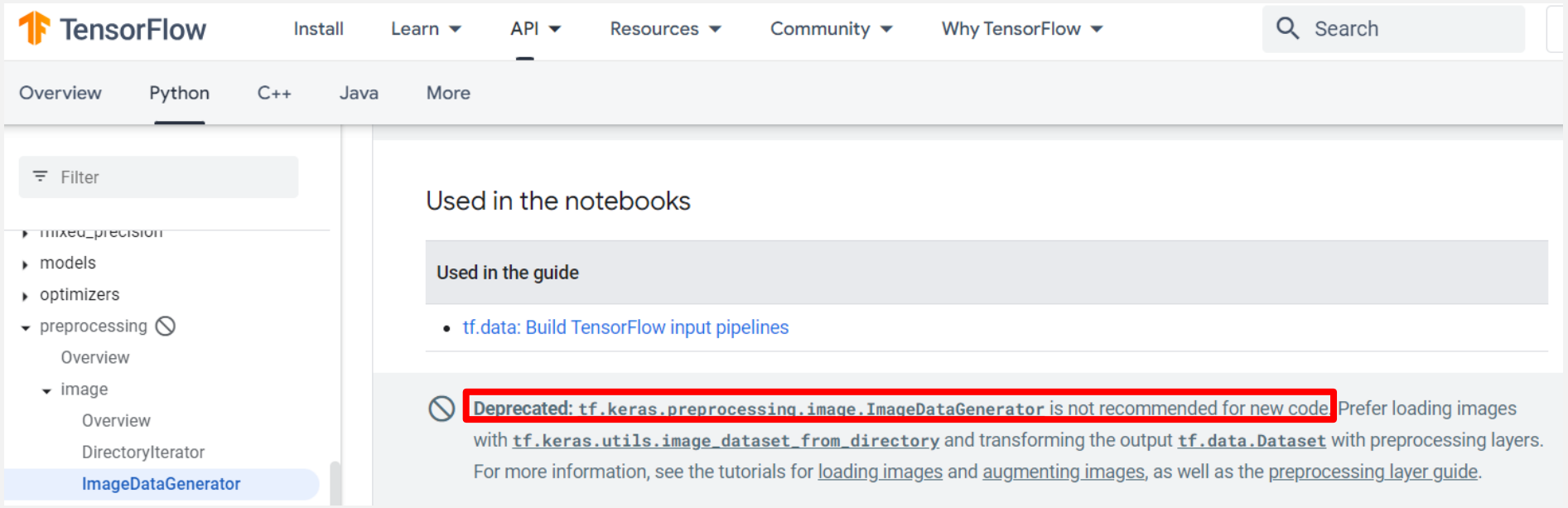
- Eliminer le sur-apprentissage.
- Améliorer l'accuracy.
- Modèle converge plus rapidement.

- Légère dégradation de la valeur de perte.

PREUVE DE CONCEPT D'UNE TECHNIQUE RECENTE

place de marché

tensorflow.keras.preprocessing.image.ImageDataGenerator: en statut «deprecated»
Utilisation de **tensorflow.keras.models.Sequential**



Source: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator

EXTRACTION DES DONNES PAR API

Endpoint de l'API: `https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"`

Requête: HTTP GET	{"ingr": "CHAMPAGNE"}	Limitation: 10 premiers produits

Rappel des grands principes du RGPD
(Règlement général sur la protection des données) :
RGPD est une norme européenne
1 -Ne collectez que les données nécessaires pour atteindre votre objectif
2 -Soyez transparent
3 -Organisez et facilitez l'exercice des droits des personnes
4 -Fixez des durées de conservation
5 -Sécurisez les données et identifiez les risques

source: CNIL

foodId	label	category	foodContentsLabel	image
food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

CONSLUSIONS ET PERSPECTIVES

1. La faisabilité d'un moteur de classification basé sur texte et images a été confirmée.
2. Le moteur de recherche peut être basé sur le modèle Embedding USE pour le traitement de texte (nom du produit, description, spécification) et sur le modèle de transfer learning avec RESNET50 et data augmentation pour les images.
3. Les modèles corrigent des erreurs existantes mais en parallèle peuvent en créer de nouvelles.
4. La classification actuelle se limite aux catégories de niveau 1 ; une étude approfondie est nécessaire pour étendre aux autres niveaux.
5. Afin d'améliorer les performances des modèles il faudra les entrainer sur un data set plus large et envisager les techniques de fine-tuning.

MERCI POUR VOTRE ATTENTION !