



CLASSIFIEZ AUTOMATIQUEMENT DES BIEN DE CONSOMATION





SOMMAIRE

- I. Problématique
2. Présentation des données
3. Détails des prétraitements, extractions de caractéristiques et résultats de l'étude de faisabilité
 - I. Traitement du Langage Naturel (NLP)
 2. Traitement des images
4. Classification supervisée des images avec data augmentation
5. Utilisation de l'API : présentation du test



I. Problématique

Contexte : Data Scientist au sein de l'entreprise "Place de Marché", qui prévoit de lancer une marketplace e-commerce. Sur ce site, les vendeurs pourront proposer des articles aux acheteurs en publiant une photo et une description de chaque produit.

Mission : Actuellement, l'attribution des catégories des articles est effectuée manuellement par les vendeurs, ce qui entraîne une certaine imprécision et une fiabilité limitée. La mission sera de réaliser une étude de faisabilité d'un moteur de classification automatique d'articles, en utilisant leurs images et descriptions. Je serai également chargé d'automatiser et d'optimiser ce processus pour garantir une classification plus précise et cohérente des produits.

Objectif : Automatiser l'attribution des catégories des articles pour améliorer la fiabilité et simplifier l'expérience utilisateur. Étudier la faisabilité d'un moteur de classification utilisant les descriptions, les noms des produits et les images fournies par les vendeurs, tout en garantissant un niveau de précision suffisant.

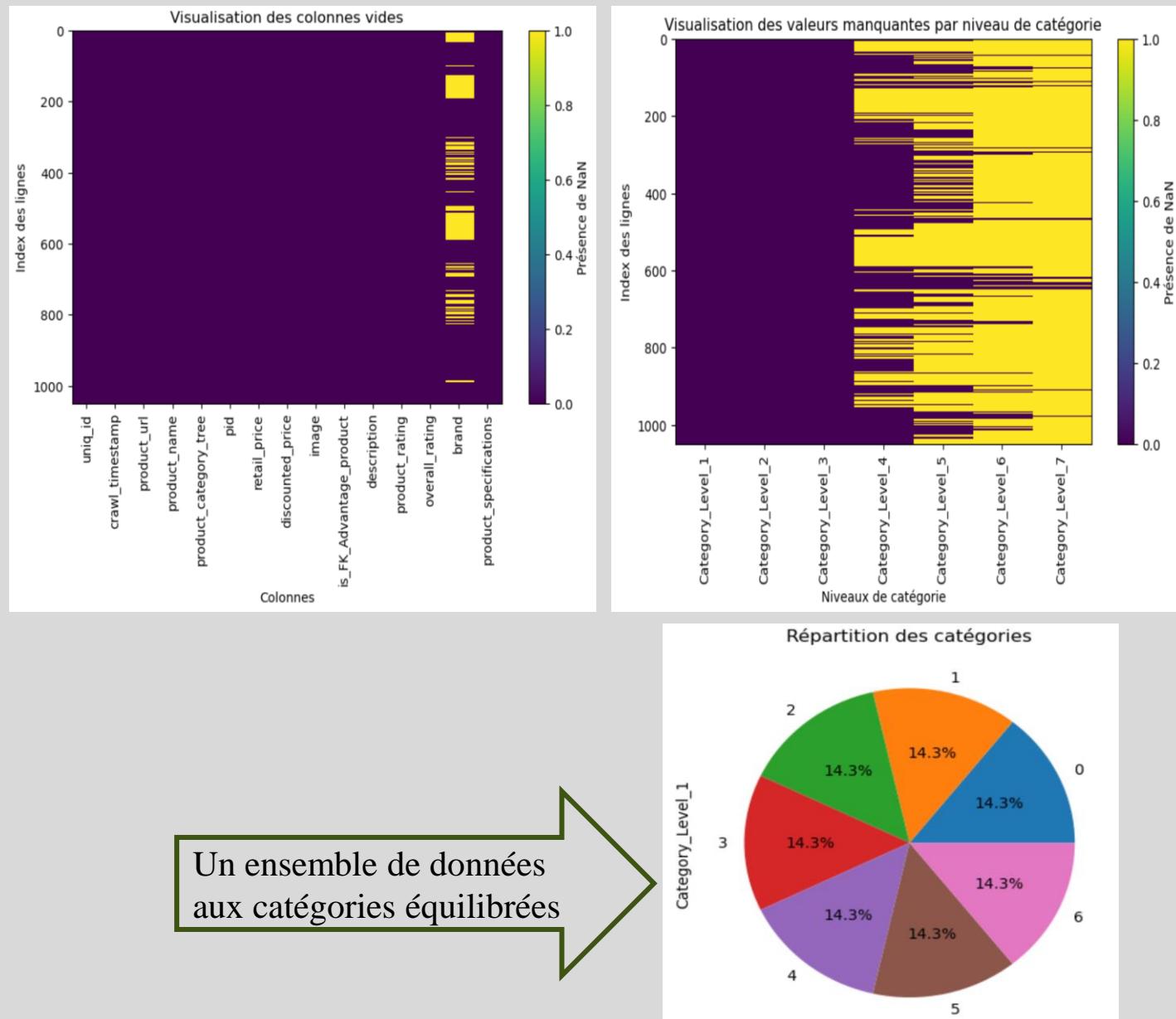


2. Présentation des données

Base de données comprenant 15 variables et 1050 enregistrements, avec 2% de valeurs manquantes.

Variables intéressantes :

- Nom du produit (entre 2 et 27 mots).
- Description du produit (entre 18 et 589 mots).
- Arbre des catégories (7 niveaux, entre 7 et 349 catégories).
- Images.





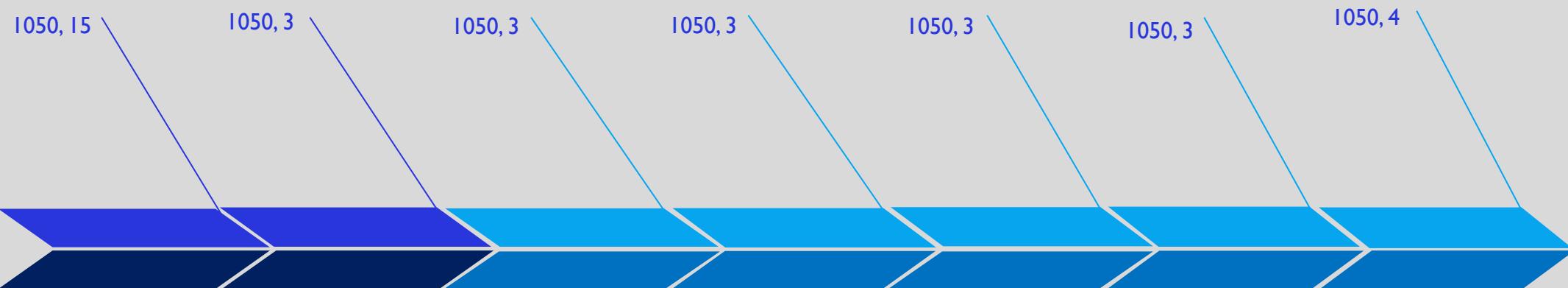
3. Détails des prétraitements, extractions de caractéristiques et résultats de l'étude de faisabilité

Procérons à la classification des données images et textes.

	Pré traitement	Extraction et description des caractéristiques pour la construction d'un vecteur numérique	Réduction de dimension	Clustering	Visualisation	Evaluation
Données textuelles	Extraction des tokens, nettoyage et création d'un vocabulaire	<ul style="list-style-type: none">❖ Bags of Words : Count-vectorizer , TF-IDF❖ Words Embedding : Word2Vec, BERT, USE	TruncatedSVD TSNE ACP	Algorithme de classification Kmeans	TSNE à l'aide de l'ACP	Calcul de l'Indice de Rand Ajusté (ARI)
Données graphiques (image)	Extraction des images et réduction de leur taille	<ul style="list-style-type: none">❖ Bags of visual word : SIFT❖ Embedding : CNN				



Les données textuelles ont été préparées de la manière suivante.



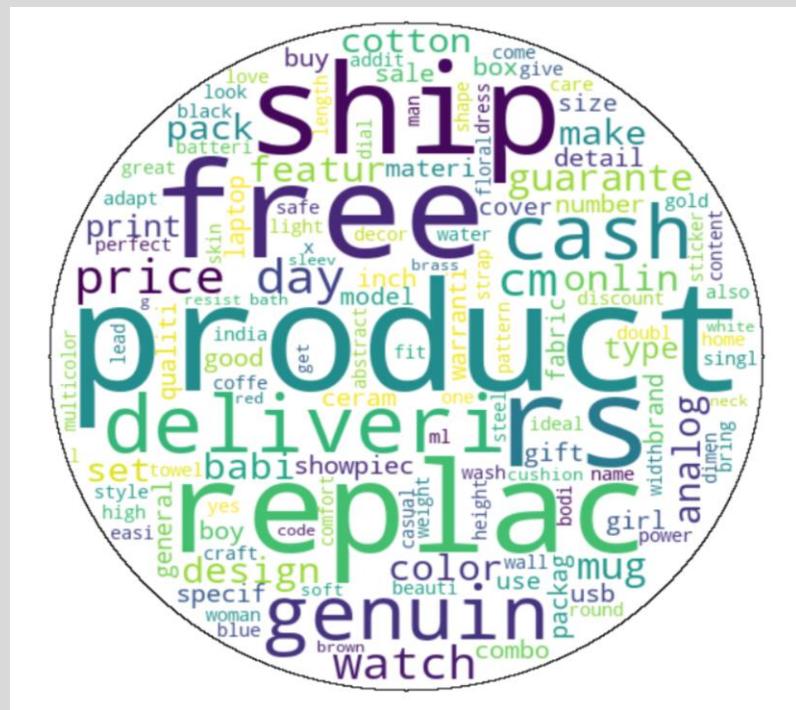
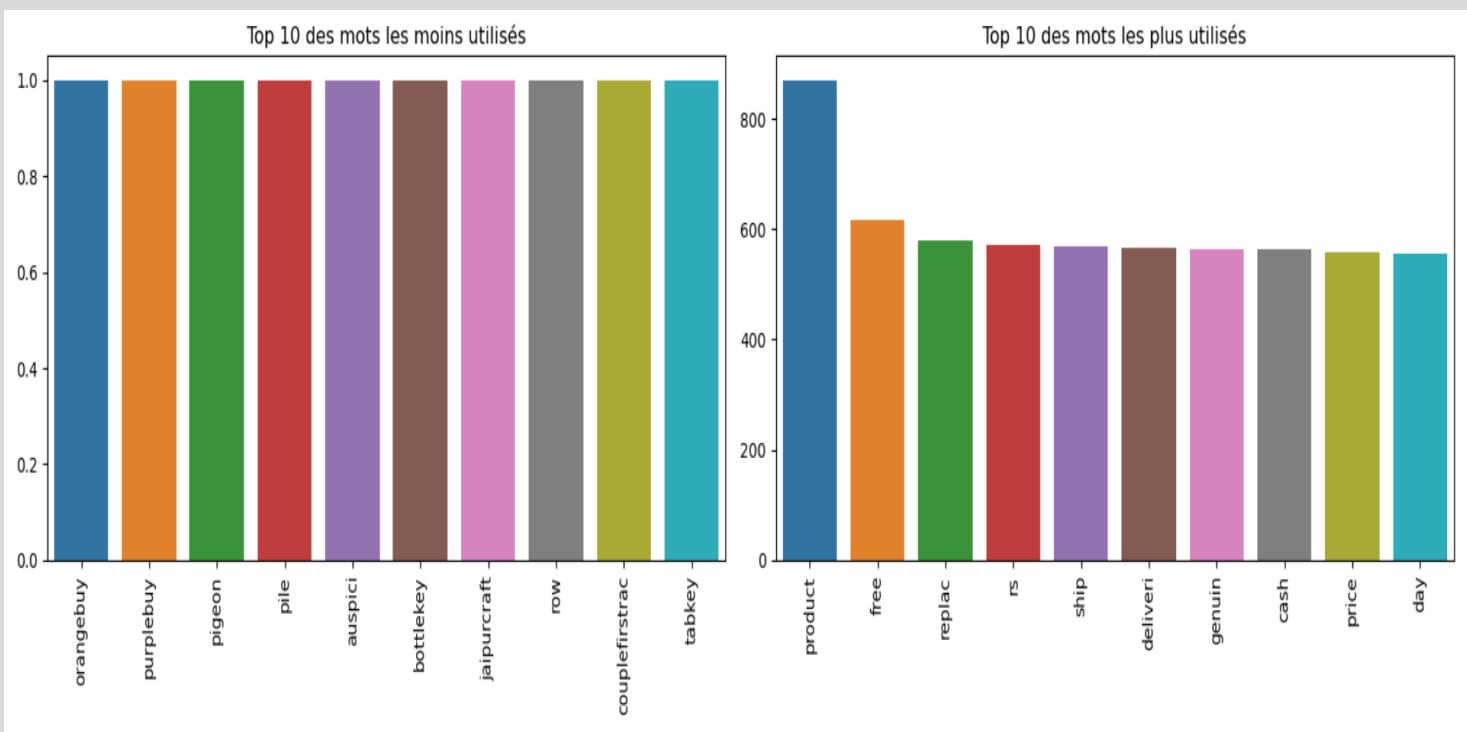
Catégories de produits	Suppression des colonnes non pertinentes	Tokenisation et Lower	Stop word	Stemming	Lemmatizer	Encodage des catégories
Il y a plusieurs catégories et sous-catégories, mais nous ne conservons que la première.	Nous conservons en priorité les descriptions, les images et les catégories.	Suppression des caractères spéciaux et conversion en minuscules	Suppression des stop words et de la ponctuation, et conservation des mots de plus de 2 lettres	Application du stemming pour réduire les mots à leur racine commune	Application de la lemmatisation pour réduire les mots à leur forme canonique ou de base	Création de labels Pour attribuer des codes numériques à chaque catégorie unique



3.1 Traitement du Langage Naturel (NLP)

Extraction et simplification des données textuelles :

Fusion des textes du nom et de la description des produits en une seule variable (compris entre 21 et 593 mots).



Nous pouvons clairement voir les mots les plus utilisés.



□ Approche Bag-of-Words

Définition de CountVectorizer

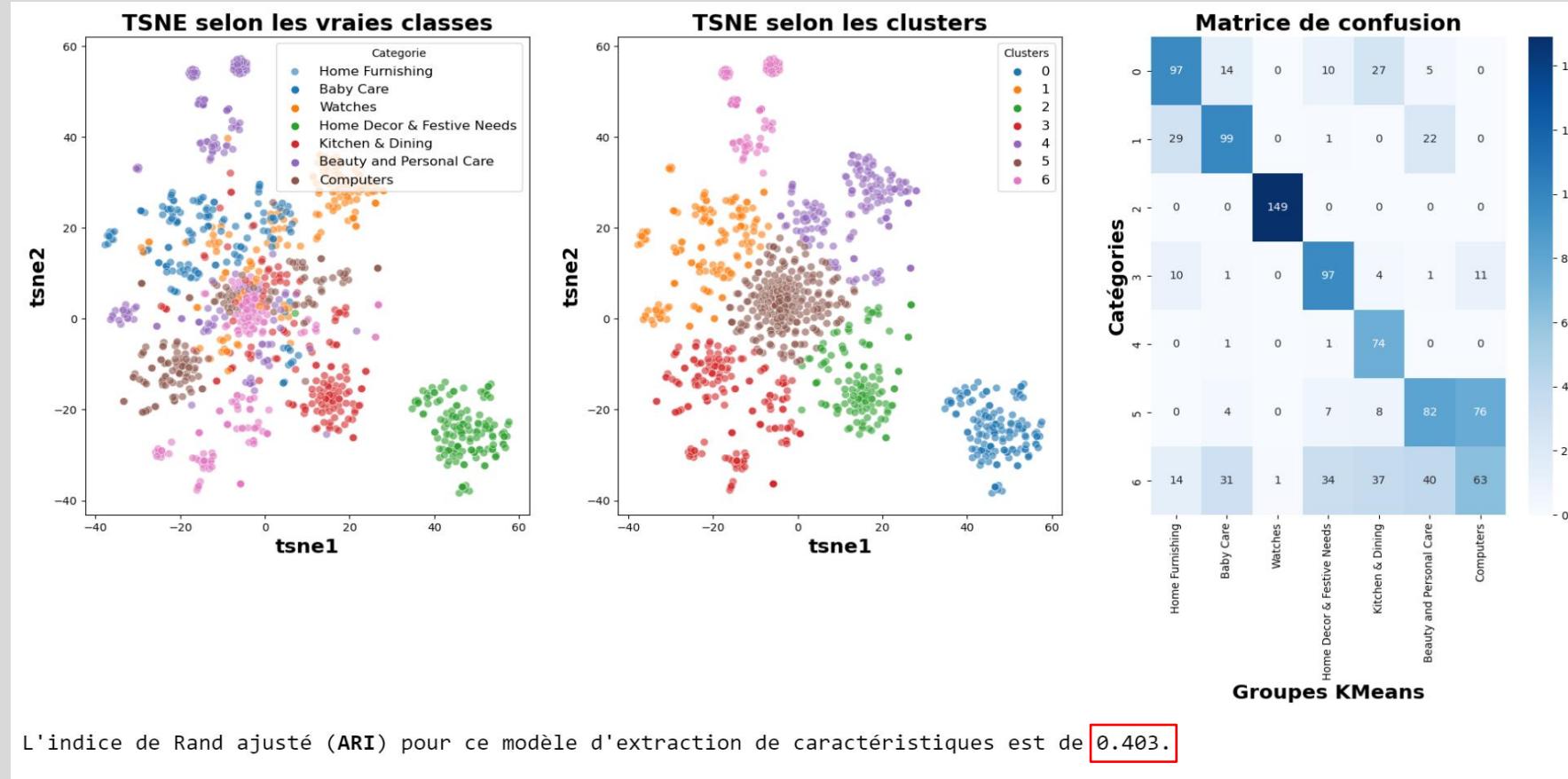
CountVectorizer est une classe de scikit-learn qui convertit une collection de documents textuels en une matrice de comptage de tokens. Chaque document est représenté par un vecteur indiquant la fréquence de chaque mot unique.

[Lien image: <https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c>]

	big	count	create	dataset	differnt	features	hello	james	name	notebook	of	python	this	try	trying	vectorizer	words
0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0
3	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
4	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0



Résultats CountVectorizer



La classification comporte des erreurs et les catégories sont mal attribuées. En effet, la matrice de confusion montre des confusions significatives entre certaines catégories



Définition de TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) est une technique qui évalue l'importance d'un mot dans un document par rapport à un corpus. Contrairement au simple comptage des mots, TF-IDF pondère les mots en tenant compte de leur fréquence dans le document (TF) et de leur rareté dans le corpus (IDF), mettant ainsi en avant les mots significatifs tout en réduisant l'importance des mots courants.

[Lien image: <https://knowledge.dataiku.com/latest/ml-analytics/nlp/concept-natural-language-processing-challenges.html>]

REDUNDANT FEATURES



text

0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

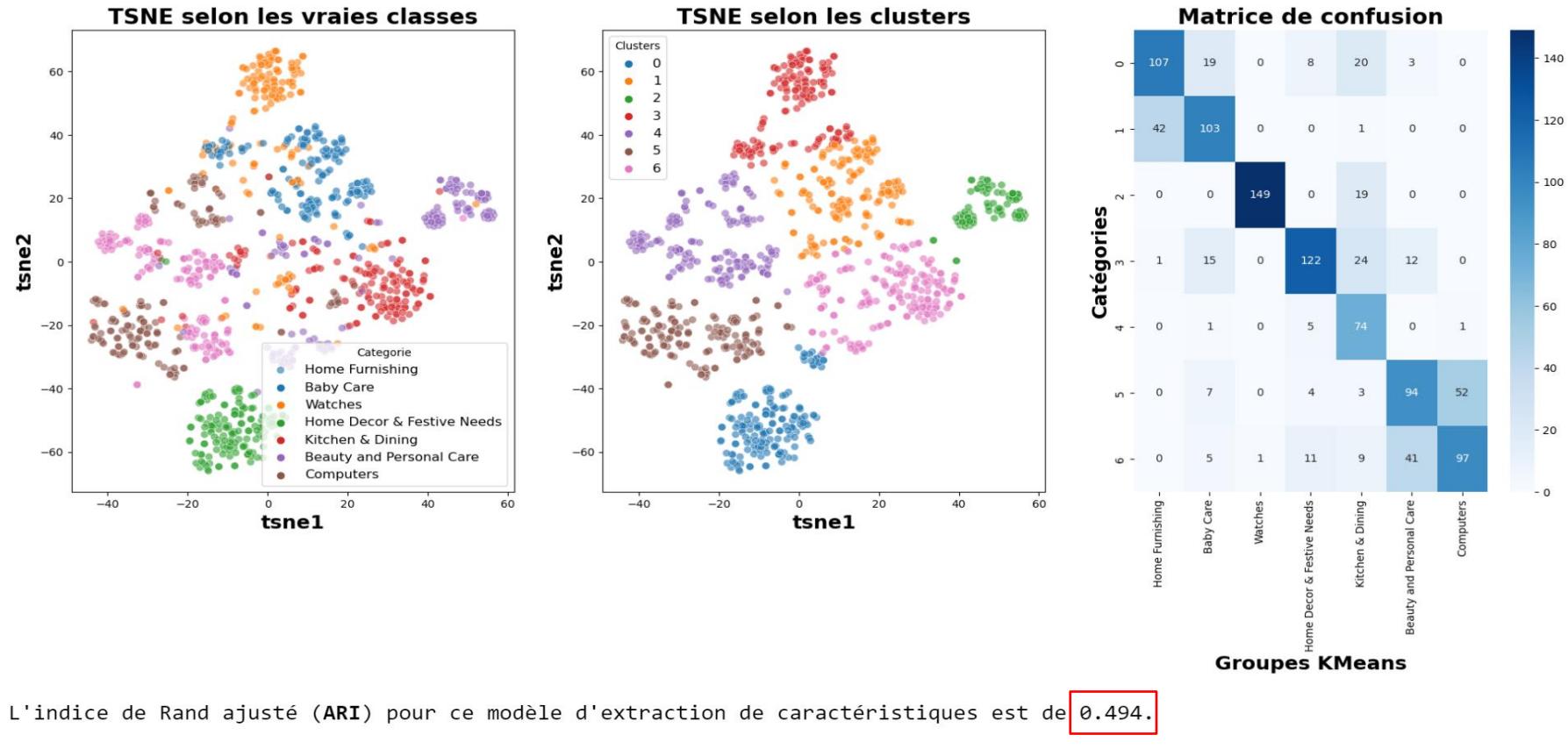
“North” vs. “north”? “north”
“king” vs. “kings”? “king”



	king	was	the	not	But	him	one	north	kings	is	in	he	Eddard	everywhere	different	could	change	but	are	Stark	North	Hodor	without
0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
1	2	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0
2	1	2	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1



Résultats TF-IDF



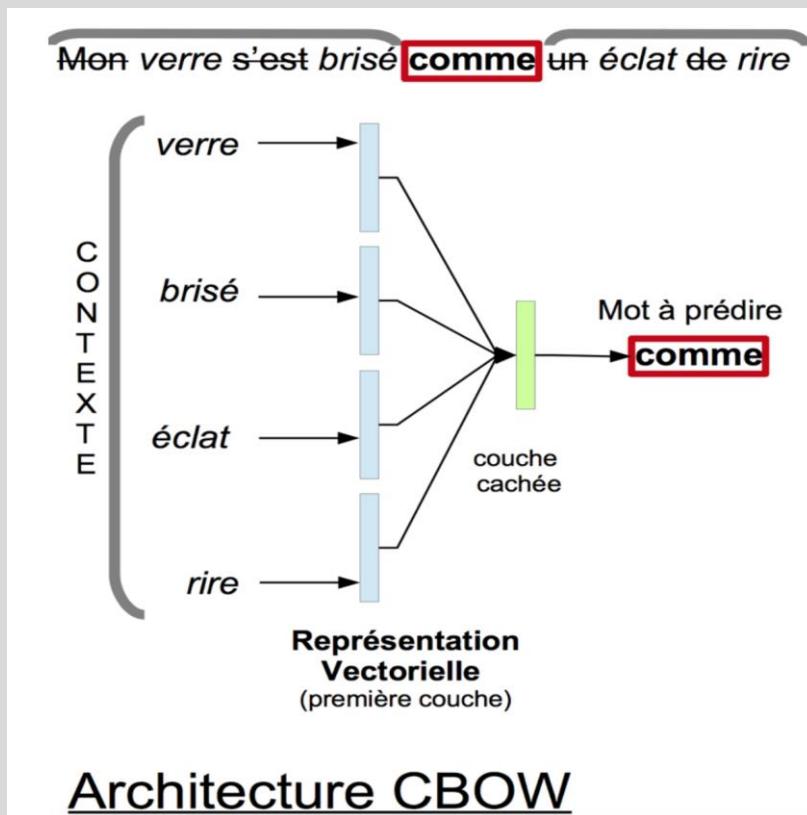
La classification est meilleure qu'avec le CountVectorizer. Les catégories sont relativement bien identifiées par l'algorithme. Cependant, il existe encore des confusions, comme indiqué par la matrice de confusion.



□ Approche Words Embedding

Définition de Word2Vec

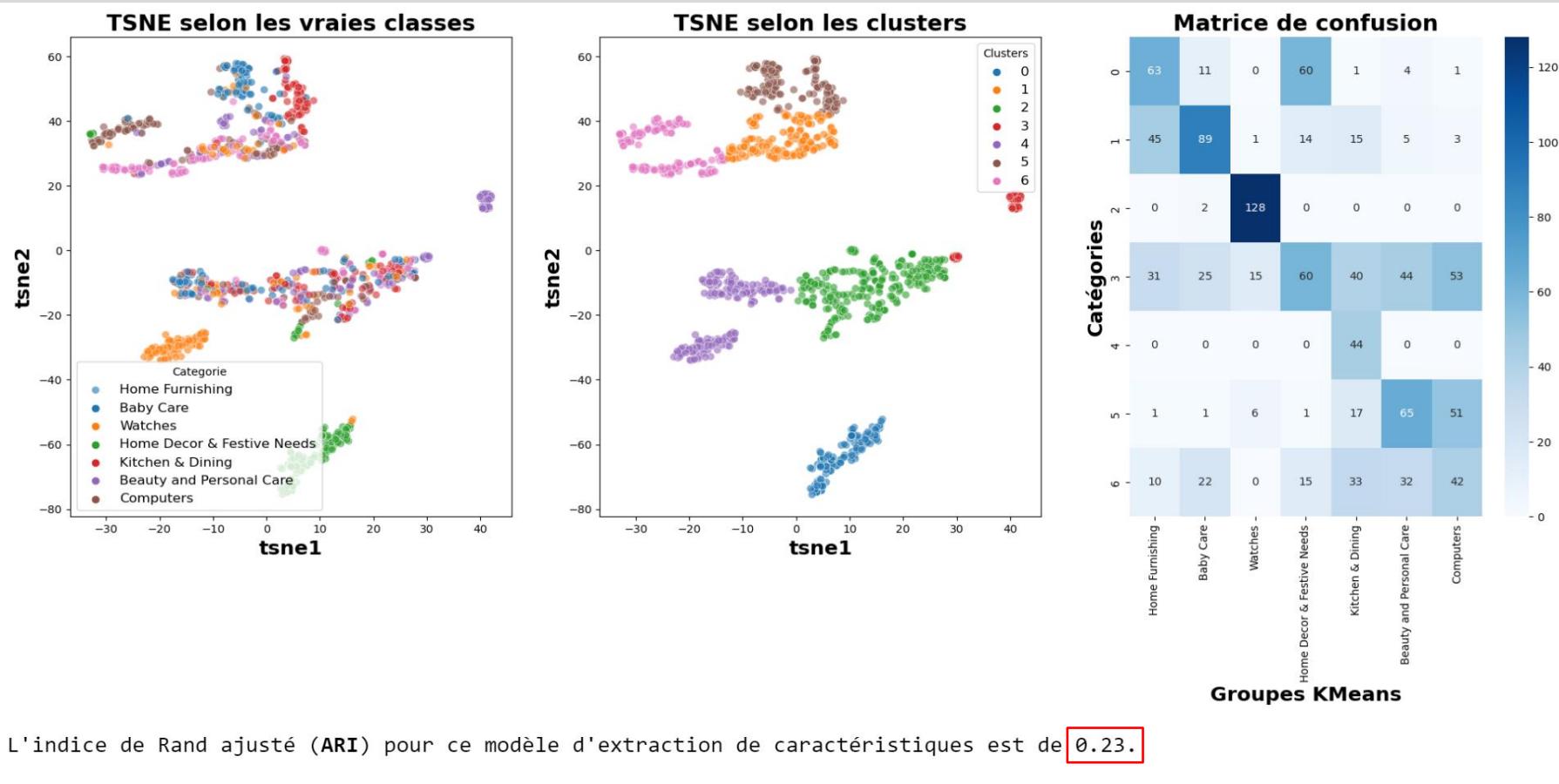
Word2Vec est une méthode avancée qui représente les mots sous forme de vecteurs continus dans un espace vectoriel de dimension réduite. Contrairement à l'approche Bag-of-Words, Word2Vec capture les relations sémantiques entre les mots, de sorte que des mots similaires ont des vecteurs similaires. En utilisant les architectures CBOW et Skip-gram, Word2Vec fait des prédictions de mots basées sur leur contexte. [Lien image: <https://dataanalyticspost.com/Lexique/word2vec/>]



Développé par une équipe de **Google** en **2013** dirigée par **Tomas Mikolov**, il repose sur des **réseaux de neurones à deux couches**.



Résultats Word2Vec



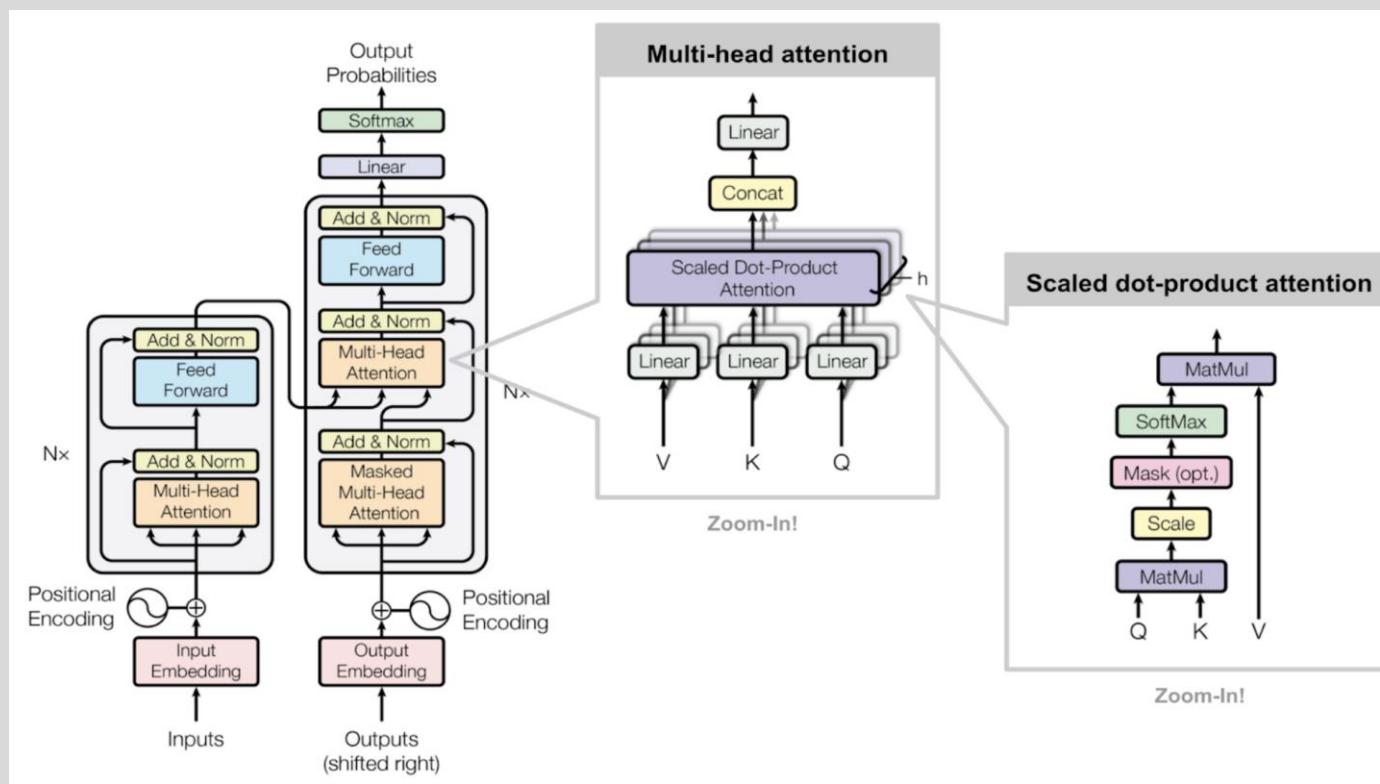
Nous observons une diminution des performances. Les catégories proches sont souvent mal attribuées, comme le montrent les clusters qui se chevauchent dans la visualisation TSNE et les confusions dans la matrice de confusion.



Définition de BERT

BERT (Bidirectional Encoder Representations from Transformers) est une méthode avancée pour obtenir des représentations contextuelles de mots et de phrases. Contrairement à Word2Vec, BERT prend en compte le contexte des mots dans les deux directions (avant et arrière), capturant ainsi des significations plus riches et précises.

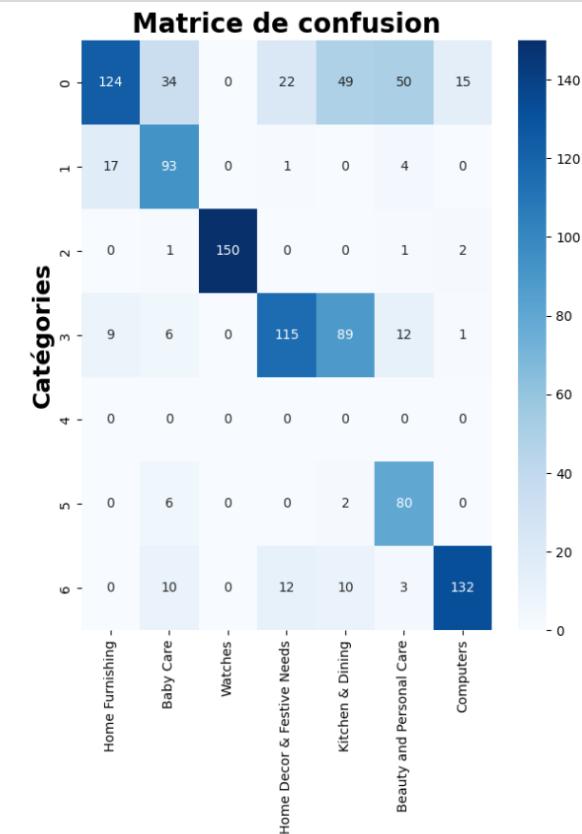
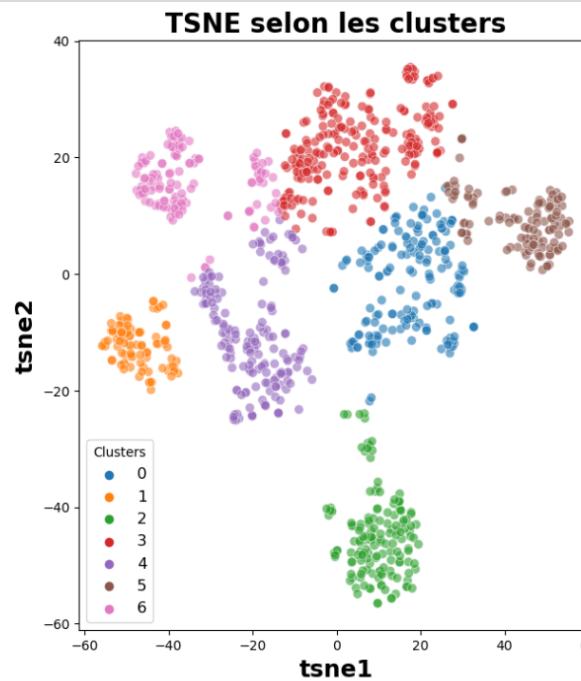
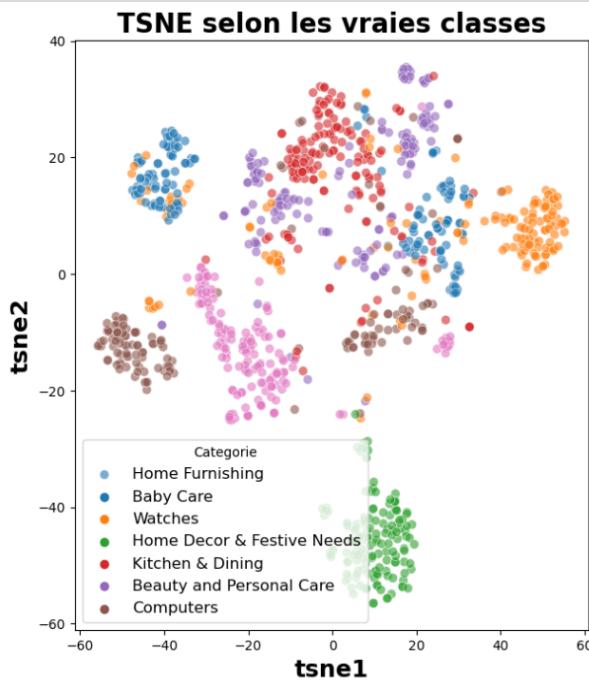
[Lie, image: <https://neptune.ai/blog/bert-and-the-transformer-architecture>]



BERT est un modèle de langage développé par **Google en 2018**. Il utilise l'architecture Transformer, qui repose sur des couches de neurones pour apprendre les relations contextuelles entre les mots ou sous-mots d'un texte. L'architecture comprend plusieurs blocs d'encodage, chacun composé de mécanismes d'attention multi-tête et de couches de feed-forward. Contrairement aux modèles directionnels qui lisent le texte séquentiellement, l'encodeur Transformer de BERT traite la séquence entière de mots simultanément, capturant ainsi le contexte complet des mots à la fois à gauche et à droite. Cela permet à BERT de générer des représentations bidirectionnelles des mots.



Résultats BERT



L'indice de Rand ajusté (ARI) pour ce modèle d'extraction de caractéristiques est de **0.466.**

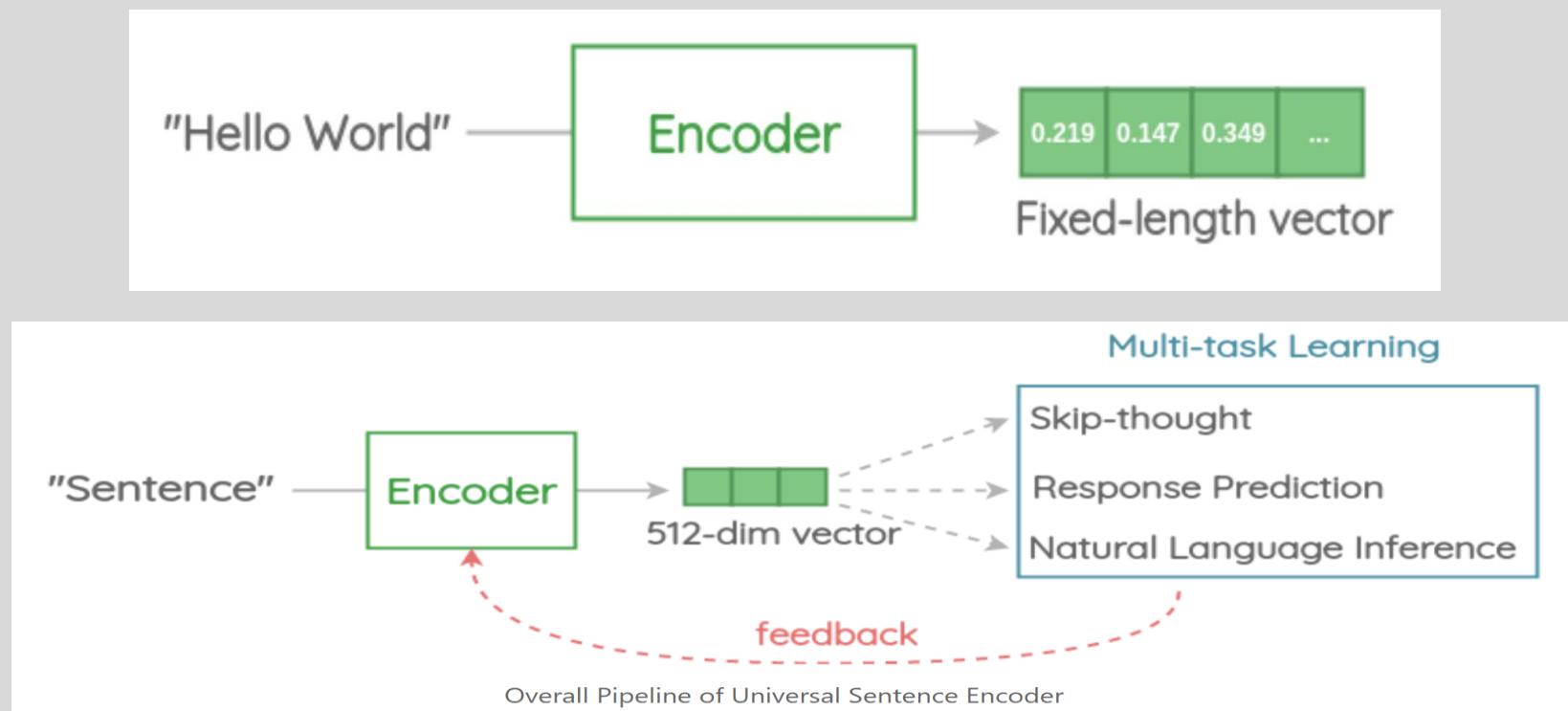
La classification est meilleure qu'avec Word2Vec. Les catégories proches sont relativement bien identifiées par l'algorithme.



Définition de USE

USE (Universal Sentence Encoder), développée par Google, produit des représentations vectorielles de haute qualité pour des phrases entières. Elle est particulièrement utile pour la classification, la similarité et le clustering de texte.

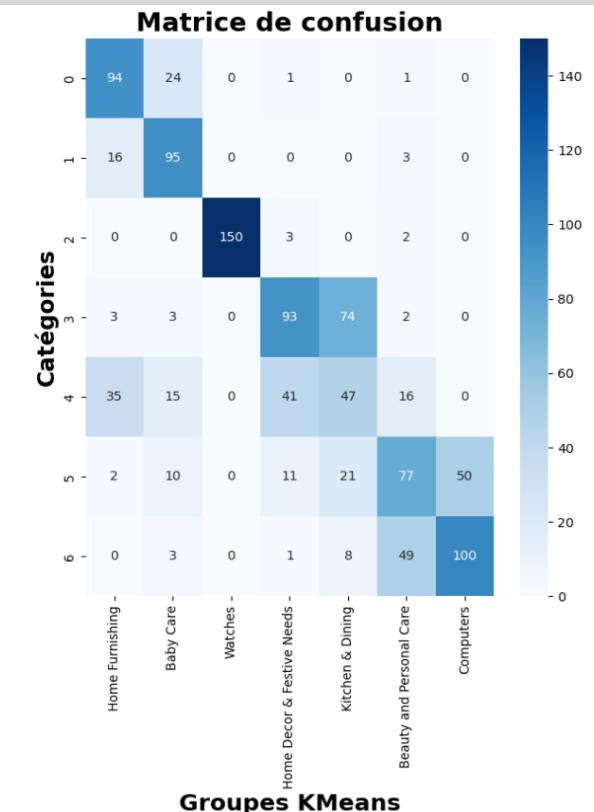
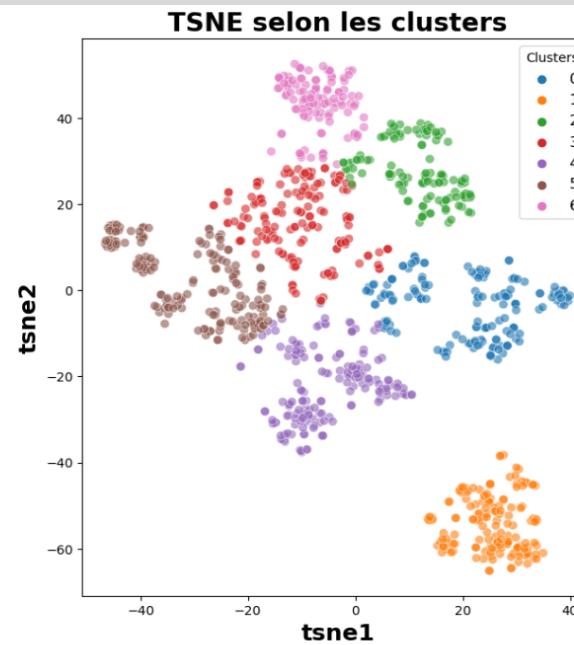
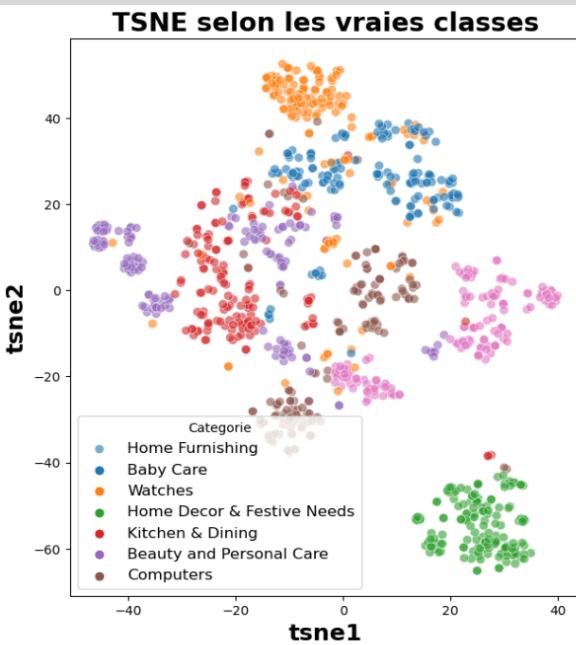
[Lien image: <https://amitness.com/posts/universal-sentence-encoder>]



L'Universal Sentence Encoder (USE) encode chaque phrase en un vecteur de 512 dimensions. En fonctionnant sur plusieurs tâches, il capture les caractéristiques sémantiques essentielles et élimine le bruit, produisant des représentations vectorielles robustes et polyvalentes. Ces vecteurs sont utiles pour diverses applications de traitement du langage naturel telles que la classification de texte, la similarité et l'inférence linguistique.



Résultats USE



L'indice de Rand ajusté (ARI) pour ce modèle d'extraction de caractéristiques est de **0.434.**

Meilleure classification après le TF-IDF et BERT. Le principe d'embedding produit de meilleurs résultats et minimise les matrices vides.

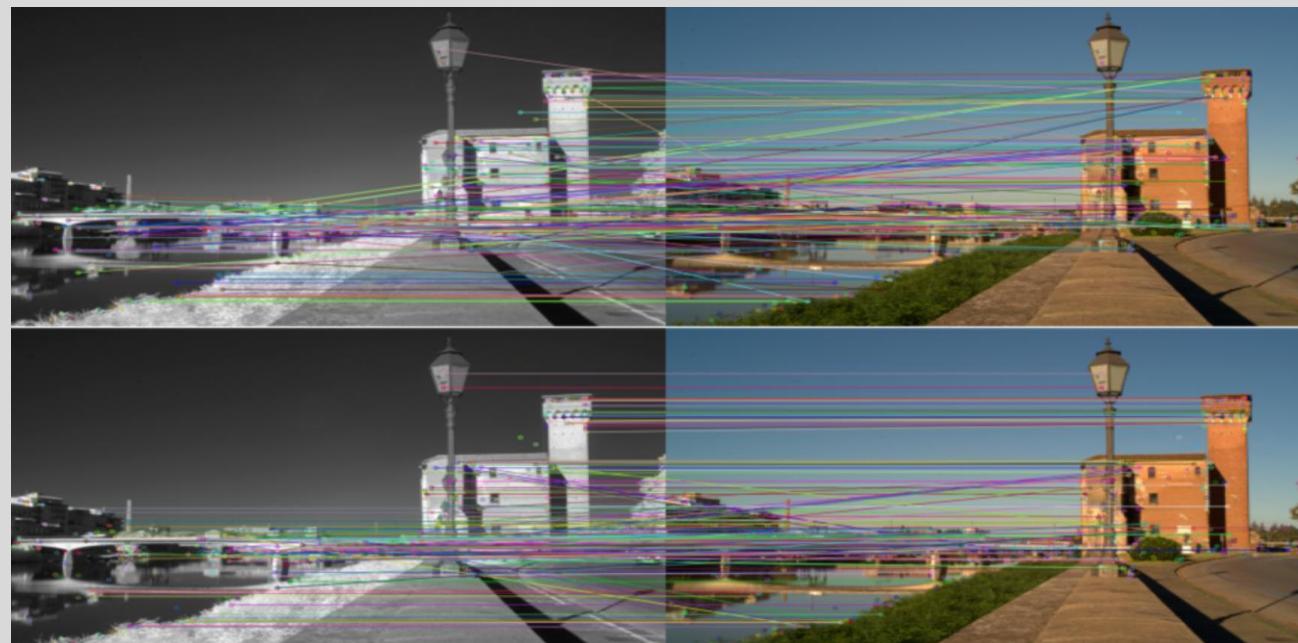


3.2 Traitement des images

Définition de SIFT

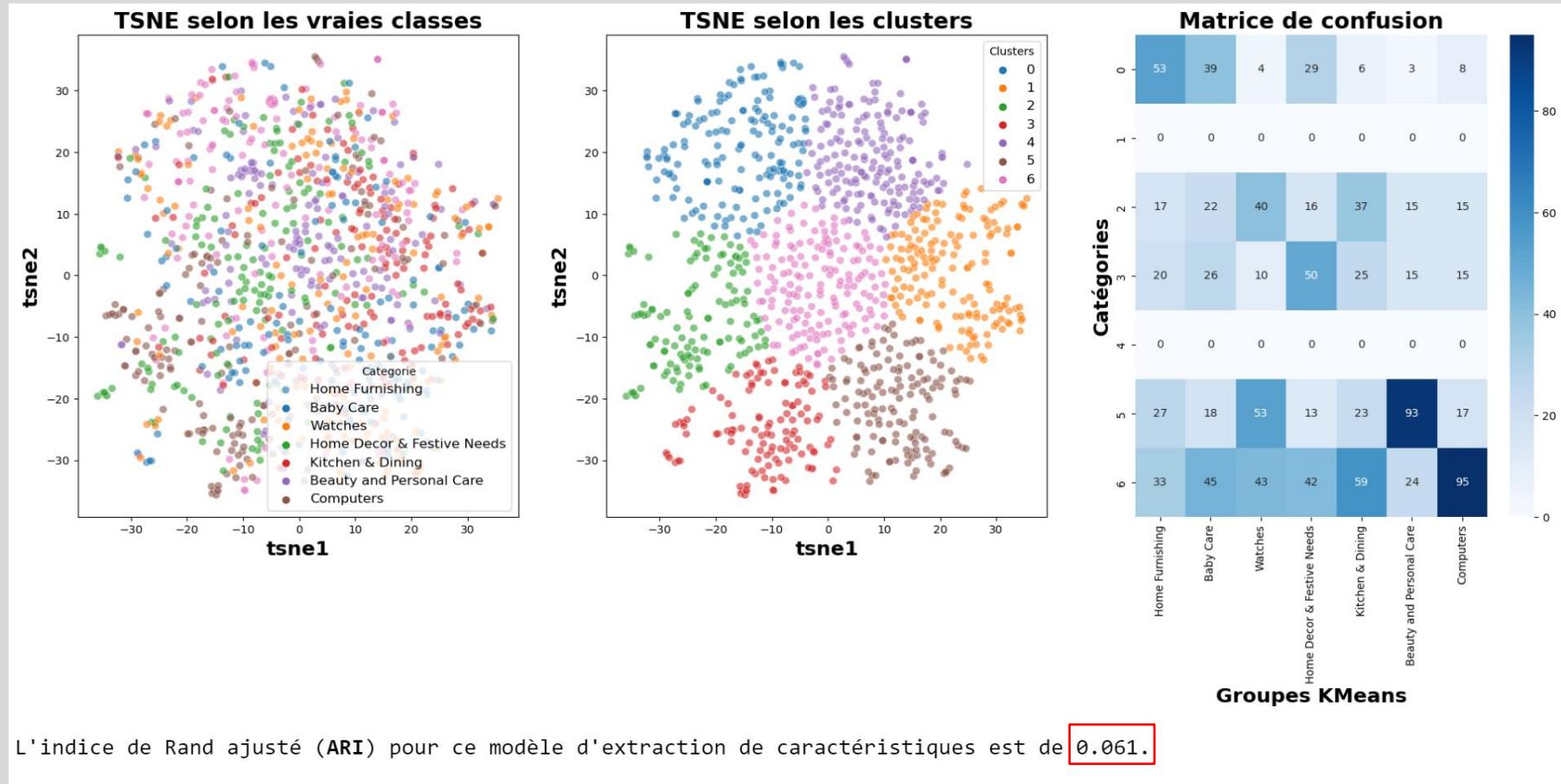
SIFT (Scale-Invariant Feature Transform) est un algorithme développé par David Lowe en 1999 pour détecter et décrire des points d'intérêt dans des images, robuste aux transformations d'échelle et de rotation. Il fonctionne en trois étapes principales : détection des points-clés via des opérations de filtrage à différentes échelles, description des points-clés par normalisation et calcul de vecteurs de caractéristiques, et correspondance des points-clés entre images pour identifier des similarités, utile dans des applications comme la reconnaissance d'objets et la mosaïque d'images.

[Lien image: https://www.researchgate.net/figure/An-example-of-cross-spectral-feature-matching-Top-LSS-and-Bottom-LSSLAT_fig3_309031693]





Résultats SIFT



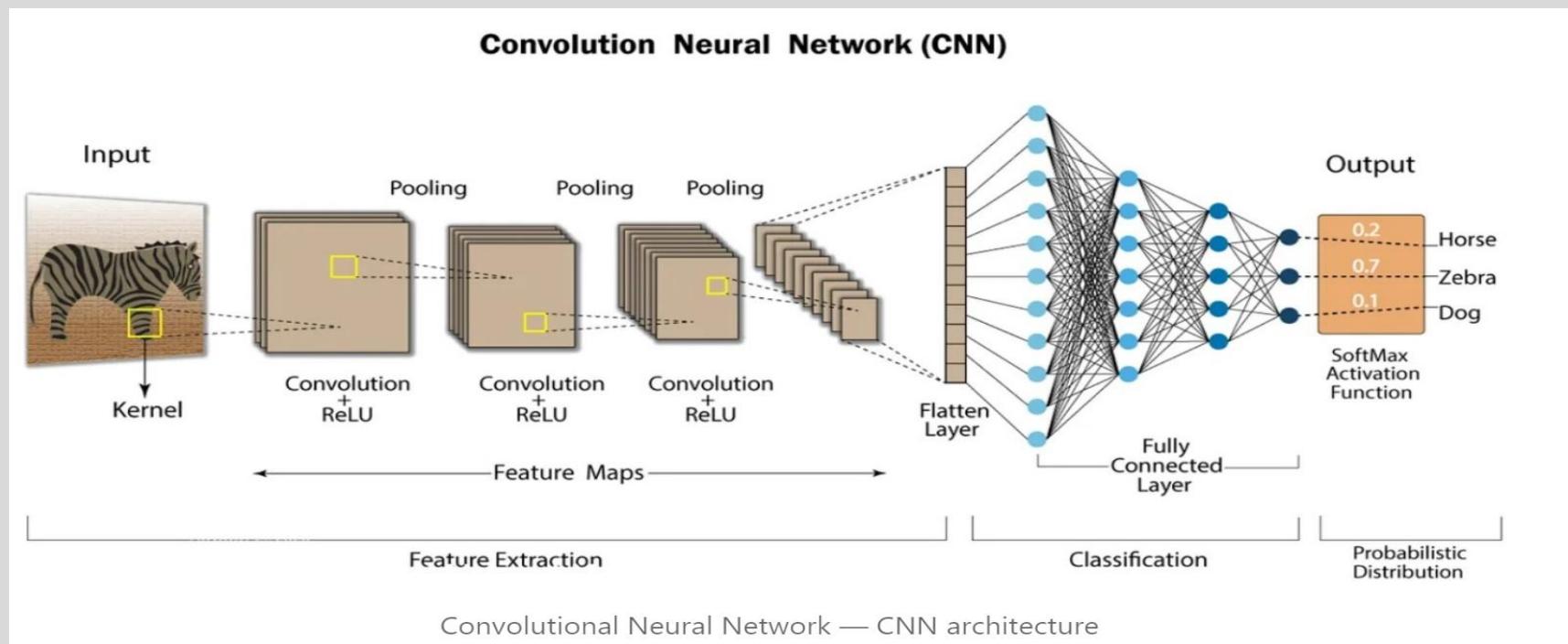
On obtient une très mauvaise classification. Presque aucune image n'est correctement classée, comme le montrent les nombreux points de la matrice de confusion où les catégories sont mal attribuées, et l'indice de Rand ajusté (ARI) très bas de 0.061.



Définition de CNN

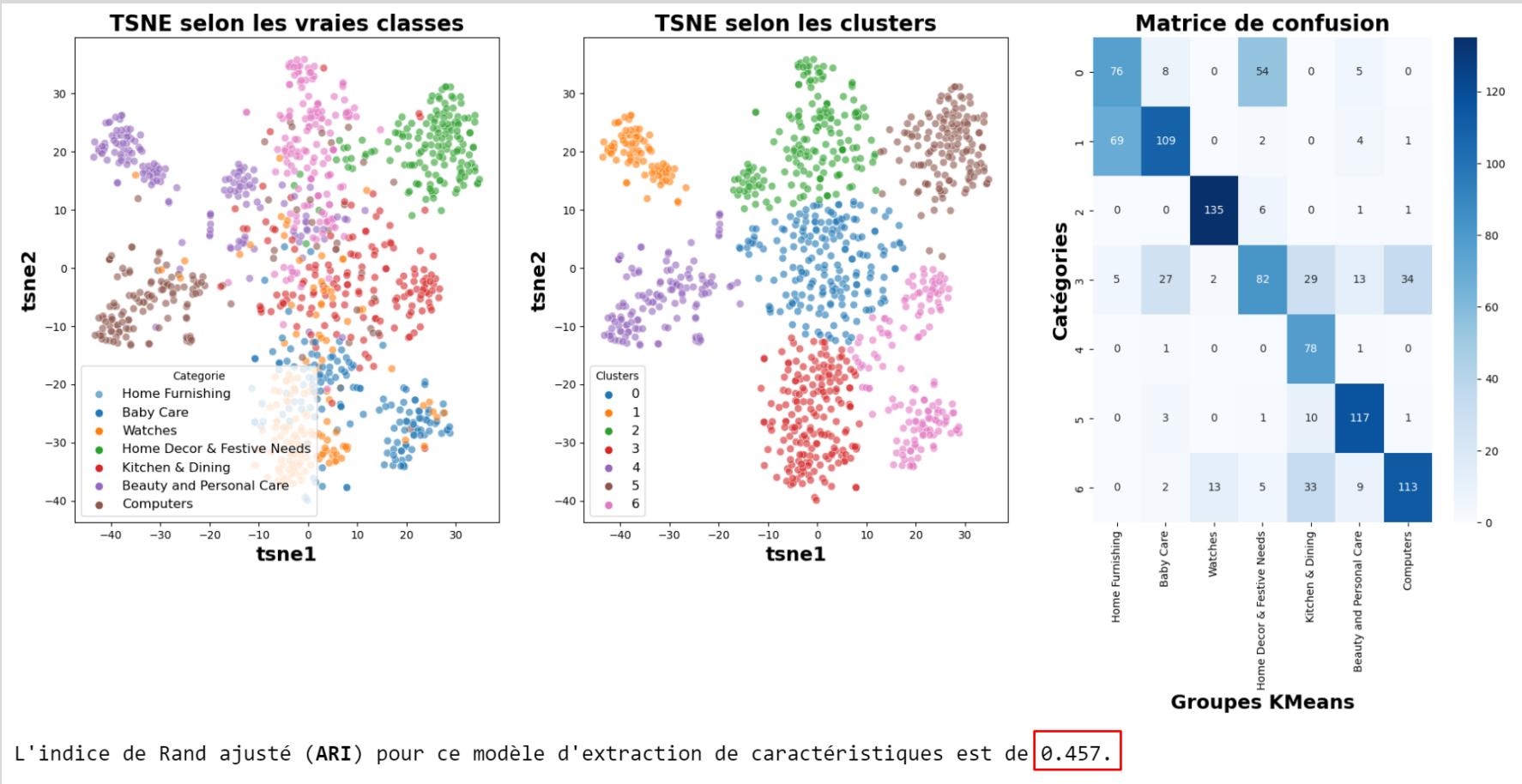
CNN (Convolutional Neural Network) est une classe de réseaux de neurones utilisée principalement pour l'analyse d'images, mais aussi pour d'autres types de données comme l'audio et le texte. Les CNN sont efficaces pour reconnaître des motifs et des structures spatiales. Ils comportent plusieurs couches : convolution pour créer des cartes de caractéristiques, ReLU pour introduire de la non-linéarité, pooling pour réduire la dimensionnalité et conserver les informations essentielles, et une couche complètement connectée pour effectuer des prédictions. La couche Softmax convertit les valeurs en probabilités pour les tâches de classification.

[Lien image: <https://www.linkedin.com/pulse/what-convolutional-neural-network-cnn-deep-learning-nafiz-shahriar/>]





Résultats CNN



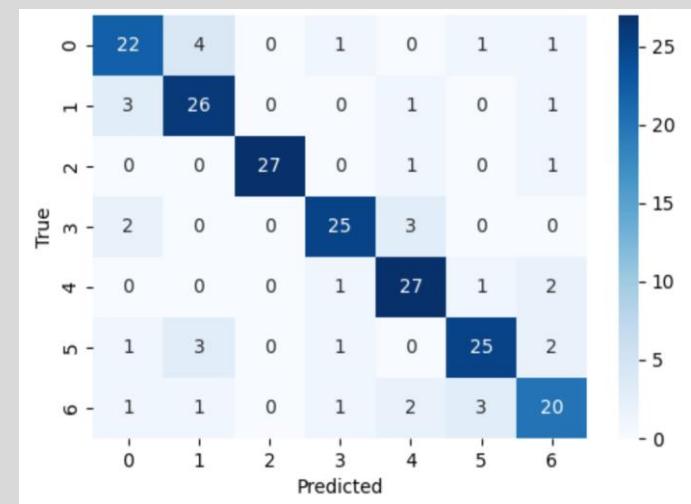
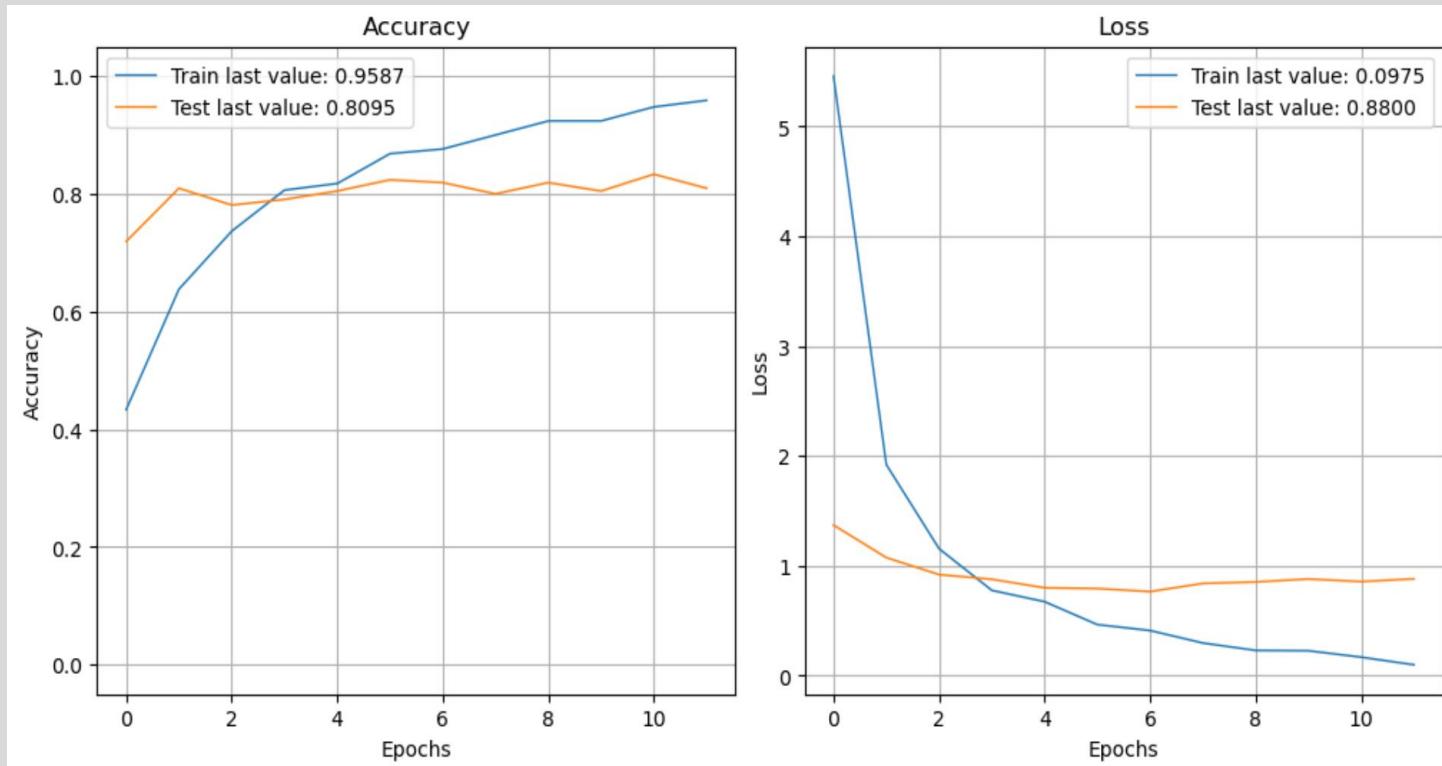
La classification est meilleure pour les données graphiques (photos). Les catégories sont assez bien identifiées par l'algorithme. Le score peut être amélioré en adaptant davantage le modèle à nos données.



4. Classification supervisée des images avec data augmentation.

3 approches sont présentées :

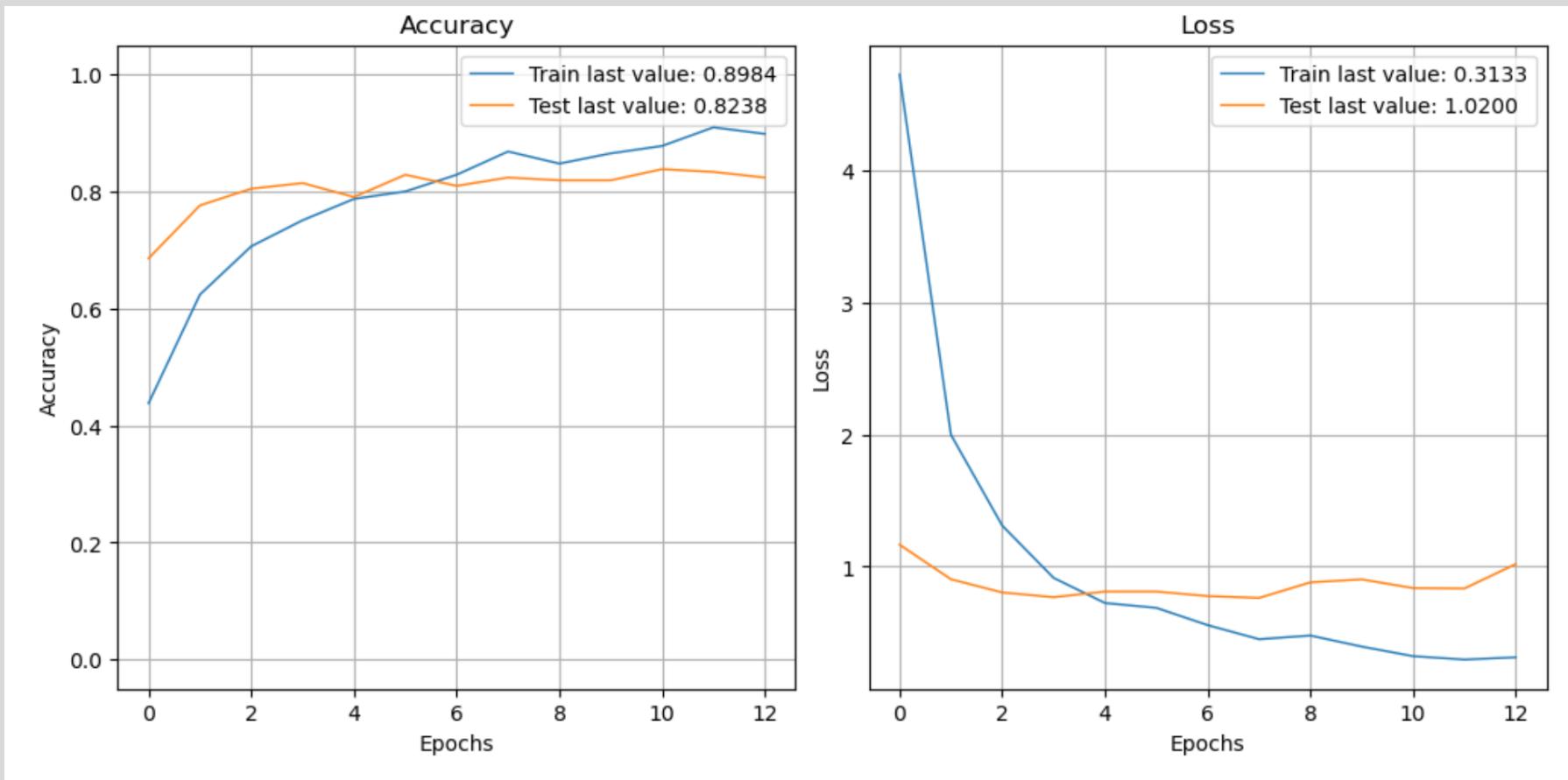
Une approche simple par préparation initiale de l'ensemble des images avant classification supervisée.



La matrice de confusion montre une bonne précision globale, avec une accuracy test de 0.81. Cependant, il reste quelques erreurs de classification spécifiques à corriger pour améliorer la performance du modèle.



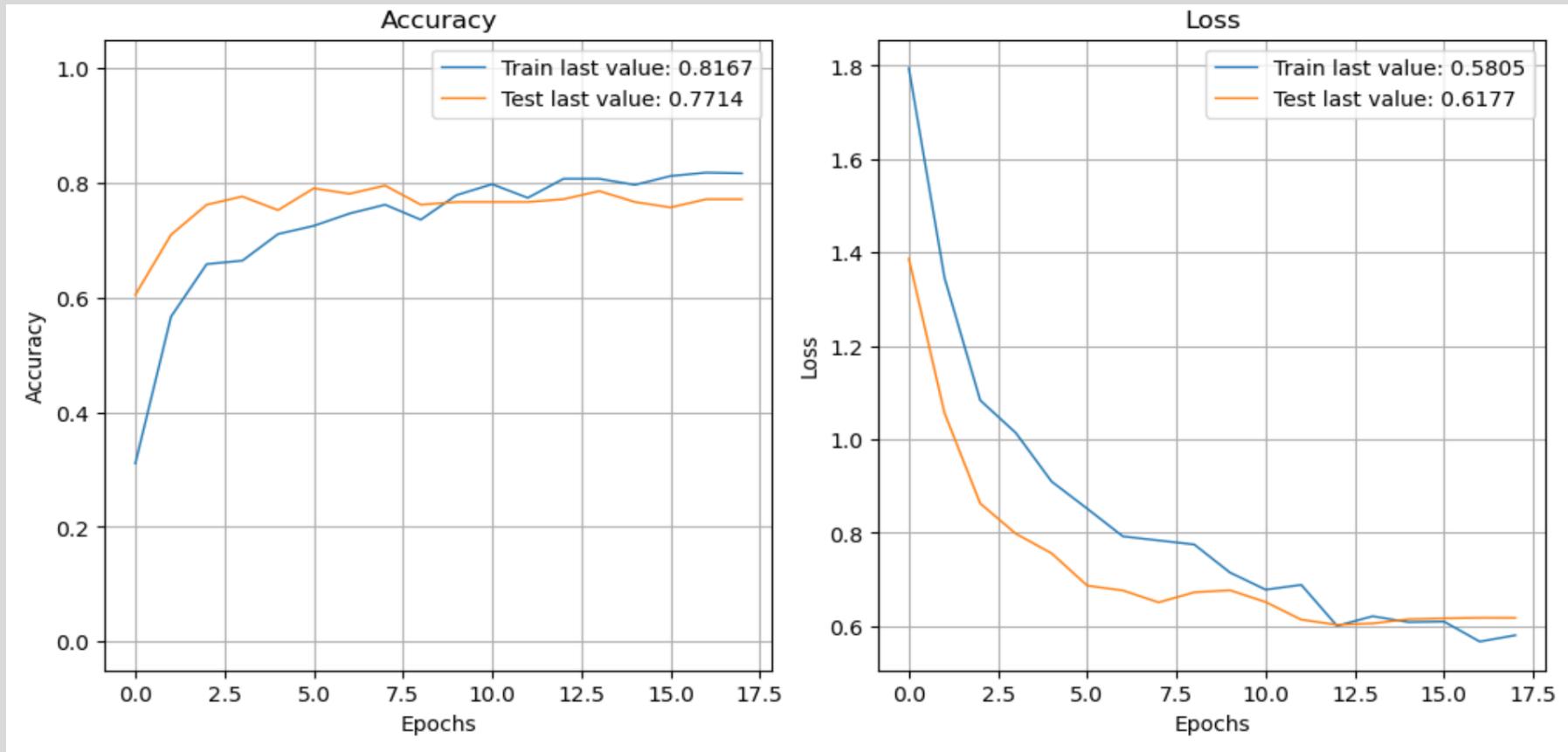
Une approche par data generator, permettant facilement la data augmentation. Les images sont directement récupérées à la volée dans le répertoire des images.



Le modèle fonctionne bien dans l'ensemble, avec une bonne performance de classification. L'accuracy de test de 0.82 est bien meilleure que celle de la première approche. Cependant, il y a un léger surapprentissage qui pourrait être réduit avec des techniques supplémentaires de régularisation et d'augmentation des données.



Une approche par DataSet, avec data augmentation intégrée au modèle via une couche dédiée (tf.keras.layers.experimental.preprocessing).



Le modèle fonctionne bien dans l'ensemble, avec une bonne performance de classification. Cependant, l'accuracy de test de 0.77 est inférieure à celle des deux premières approches. Ainsi, la meilleure approche est la deuxième, celle utilisant ImageDataGenerator avec data augmentation.



5. Utilisation de l'API : présentation du test

API : <https://developer.edamam.com/food-database-api>

- Si on prend juste les 10 premiers produits à base de "champagne" collectés via l'API, le tableau suivant présente les résultats.

	foodId	label	category	foodContentsLabel	image
0	food_ax1n26waalpd9cbc64bjob7pw6hg	Champagne Jelly	Generic meals	Champagne; gelatine; caster sugar; blueberries	NaN
1	food_b4va8u0bb6pf74akh2rtcb3llna9	Champagne Punch	Generic meals	champagne; simple syrup; orange juice; blueber...	NaN
2	food_a4j8wm8ayffl13b45t3c3bk9w4ek	Champagne Sangria	Generic meals	mint leaves; Champagne; orange juice; lemon; ...	NaN
3	food_bba727vaimolf0b8stgoibx7ujei	Champagne Cake	Generic meals	flour; baking powder; salt; butter; sugar; egg...	NaN
4	food_a6mj2obbqy38soat01vrxaqnvet	Champagne Cupcakes	Generic meals	butter; sugar; eggs; champagne; plain yogurt; ...	NaN
5	food_anrtk55a3aac9uactv3wlanz1m02	Champagne Cocktail	Generic meals	sugar; bitters; Champagne; lemon rind; orange ...	NaN
6	food_aoxaf73b3o0lgebpj6wjga6kqhco	Strawberry Champagne	Generic meals	frozen strawberries; champagne; sugar; gourd; ...	NaN
7	food_bncple4a2uagu1b4hov92budz2vs	Champagne Grape	Generic foods	NaN	https://www.edamam.com/food-img/ca5/ca55ac74de...
8	food_bxdqsxkax2vgmpbv8e4ygb6zfnkn	Champagne Vinegar	Generic foods	NaN	https://www.edamam.com/food-img/5f6/5f69b84c39...
9	food_a1huw7sahgjfqva60rgdqbpvcf2s	Champagne Margaritas	Generic meals	Lime Juice; tequila; Triple Sec; sparkling win...	NaN

- Résultats obtenus

	foodId	label	category	foodContentsLabel	image
0	food_bu12urpbuo9v6b4jpvk2a1fh4hh	Champagne Simply Dressed Vinaigrette, Champagne	Packaged foods	FILTERED WATER; CANOLA OIL; CHAMPAGNE AND WHIT...	https://www.edamam.com/food-img/736/736a3e72a6...
1	food_b48c55sagj89z4afsne5dar76h4x	Champagne Vinegar	Packaged foods	CHAMPAGNE VINEGAR DILUTED WITH WATER TO 7% ACI...	https://www.edamam.com/food-img/ad8/ad8c8d6ba8...
2	food_bb0nrgbsr5g4ac2enxb1deyh8	Champagne Mustard	Packaged foods	DIJON MUSTARD (VINEGAR; WATER; MUSTARD SEED; S...	https://www.edamam.com/food-img/775/775b39c0b0...
3	food_arm1rlyb5v81v6arurmjqb6xiy9t	Light Champagne Dressing, Light Champagne	Packaged foods	WATER; SOYBEAN OIL; WHITE WINE (PRESERVED WITH...	https://www.edamam.com/food-img/ee0/ee0475f45b...
4	food_bzb3i0lbxz24nnbv10dsladd965e	Inglenook Champagne	Packaged foods	Dealcoholized Champagne; Water; Grape Juice Co...	https://www.edamam.com/food-img/cb3/cb33b00db...
5	food_byl67wcbbfw82ua6j1n7oa6ago4a	Cola Champagne	Packaged foods	CARBONATED WATER; HIGH FRUCTOSE CORN SYRUP; AR...	https://www.edamam.com/food-img/f82/f82d164536...
6	food_b3ia2z2gav39j6abdapavjka0qz2e	Cola Champagne	Packaged foods	CARBONATED WATER; HIGH FRUCTOSE CORN SYRUP; AR...	https://www.edamam.com/food-img/b4/b4b747a25b...
7	food_bsck64rasshe3aa46w4dbq3ip81	Champagne Reserve Vinegar	Packaged foods	CHAMPAGNE VINEGAR	https://www.edamam.com/food-img/d2/d2dcfa4b43...
8	food_b2y7rqan02gndal9tapb4ldv5j	Girard's Champagne Vinaigrette	Packaged foods	Canola and Soybean Oil; White Wine; Contains S...	https://www.edamam.com/food-img/189/18997efa76...
9	food_b2exrzmbb6rq7za4wsvoeaga4q09	Champagne Vinaigrette Dressing	Packaged foods	CANOLA OIL; HONEY; CHAMPAGNE VINEGAR; WATER; D...	https://www.edamam.com/food-img/c08/c085cc31e2...

Nombre total de produits extraits après nettoyage : 10