Projet 6 : Classifiez automatiquement des biens de consommation

27/09/2024

Soukaina GUAOUA ELJADDI

Parcours Data Scientist
OpenClassrooms

Plan:

- ❏ Problématique et présentation du jeu de données
- ❏ Prétraitements, extractions de features
- ❏ Etude de faisabilité
- ❏ Classification supervisée
- ❏ Test de l’API
- ❏ Conclusion

Problématique

Contexte : Entreprise "**Place de marché**" est une marketplace e-commerce où des vendeurs proposent des articles (**photo** + **description)** avec **attribution manuelle de la catégorie** du produit.

Objectif :

- **Automatiser** la tâche **d'attribution de la catégorie.**
- **Élargir** leur **gamme de produits** à l'épicerie fine (**API**).

Missions :

- Étudier la **faisabilité** d'un **moteur de classification** des articles en différentes catégories.
- Réaliser une **classification supervisée** à partir **des images.**

Présentation du jeu de données




- 1050 articles
-
- 15 colonnes :
  - Identifiant : Id, nom, catégorie, marque, description du produit
  - Prix / Prix réduit
  - Image
  - Évaluation, etc

- 7 catégories : 'Ameublement', 'Soins pour bébé', 'Montres', 'Décoration intérieure et besoins festifs', 'Cuisine et salle à manger', 'Beauté et soins personnels', 'Ordinateurs'.

Présentation du jeu de données

dataset csv

uniq_id	3c4ca34c50a5437a1bcc42b72fc1351f
crawl_timestamp	2015-12-01 12:40:44 +0000
product_url	http://www.flipkart.com/printland-pmr1902-cera...
product_name	Printland PMR1902 Ceramic Mug
product_category_tree	["Kitchen & Dining >> Coffee Mugs >> Printland...
pid	MUGEBFGFGZJZGMG6
retail_price	650.0
discounted_price	299.0
image	3c4ca34c50a5437a1bcc42b72fc1351f.jpg
is_FK_Advantage_product	False
description	Printland PMR1902 Ceramic Mug (350 ml)\r\n ...
product_rating	No rating available
overall_rating	No rating available
brand	NaN
product_specifications	{"product_specification"=>[{"key"=>"Type", "va...

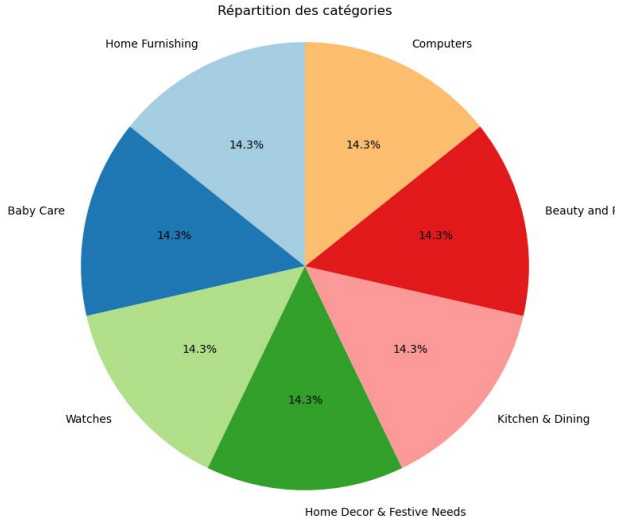
Image

Target: product_category_tree (Kitchen & Dining)

Features textes: description

Présentation du jeu de données

	Category	Count
0	Home Furnishing	150
1	Baby Care	150
2	Watches	150
3	Home Decor & Festive Needs	150
4	Kitchen & Dining	150
5	Beauty and Personal Care	150
6	Computers	150



La **distribution** des catégories est parfaitement **équilibrée**

Idéal pour la formation de **modèles** de **Machine learning**

Partie 1 : Prétraitements, extractions de features

Nettoyage du texte (ponctuation, mots de liaison, mise en minuscules) (NLTK):

Exemple de 'texte': **["**Le chat mangeait tranquillement sous l'arbre, mais il n'a pas vu le chien arriver rapidement.**"]**

Phrase nettoyée : chat mangeait tranquillement sous larbre na vu chien arriver rapidement

Partie 1 : Prétraitements, extractions de features

Tokenisation de la phrase (NLTK) :

Phrase nettoyée : chat mangeait tranquillement sous larbre na vu chien arriver rapidement

Tokens : ['chat', 'mangeait', 'tranquillement', 'sous', 'larbre', 'na', 'vu', 'chien', 'arriver', 'rapidement']

Partie 1 : Prétraitements, extractions de features

Stemming (racines des mots) (NLTK) :

Tokens : ['chat', 'mangeait', 'tranquillement', 'sous', 'larbre', 'na', 'vu', 'chien', 'arriver', 'rapidement']

Stems : ['chat', 'mang', 'tranquill', 'sous', 'larbr', 'na', 'vu', 'chien', 'arriv', 'rapid']

Partie 1 : Prétraitements, extractions de features

Lemmatisation (forme canonique des mots) (NLTK):

Stems : ['chat', 'mang', 'tranquill', 'sous', 'larbr', 'na', 'vu', 'chien', 'arriv', 'rapid']

Lemmas : ['chat', 'mangeait', 'tranquillement', 'sou', 'larbre', 'na', 'vu', 'chien', 'arriver', 'rapidement']

Partie 1 : Prétraitements, extractions de features

Construction de features avec des méthodes NLP basiques :
Bag of Words (Comptage de mots)

Bag-of-Words vocabulaire : ['arriver' 'chat' 'chien' 'larbre' 'mangeait' 'na' 'rapidement' 'sous' 'tranquillement' 'vu']

Bag-of-Words features : [[1 1 1 1 1 1 1 1 1 1]]

Partie 1 : Prétraitements, extractions de features

Construction de features avec des méthodes NLP basiques :
TF-IDF (Fréquence de mots)

TF-IDF vocabulaire : ['arriver' 'chat' 'chien' 'larbre' 'mangeait' 'na' 'rapidement' 'sous' 'tranquillement' 'vu']

TF-IDF features : [[0.31622777 0.31622777 0.31622777 0.31622777 0.31622777 0.31622777 0.31622777 0.31622777 0.31622777 0.31622777]]

Partie 1 : Prétraitements, extractions de features

Construction de features avec des méthodes NLP basiques :
LDA (Latent Dirichlet Allocation)

Phrase nettoyée : chat mangeait tranquillement sous larbre na vu chien arriver rapidement

Topic 0:

mangeait arriver sous na tranquillement rapidement larbre chat vu chien

Topic 1:

chien vu chat larbre rapidement tranquillement na sous arriver mangeait

Partie 1 : Prétraitements, extractions de features

Construction de features avec des méthodes NLP avancées : Word2Vec (Word Embeddings)

- Apprend des représentations vectorielles des mots.
- Utilise les modèles **CBOW** et **Skip-gram** pour capturer les relations entre les mots.
- Capture les similarités sémantiques entre mots (ex: "roi" et "reine").

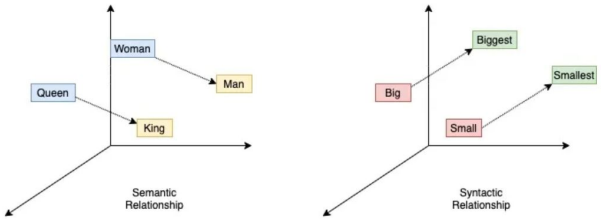


Image by author (Trained Word2Vec Vectors with Semantic and Syntactic relationship).

Partie 1 : Prétraitements, extractions de features

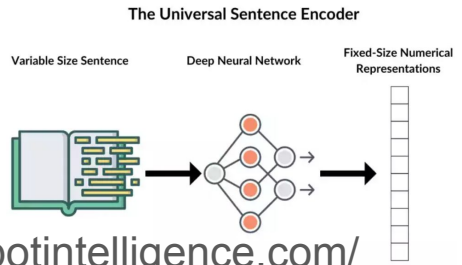
Construction de features avec des méthodes NLP avancées : BERT(Bidirectional Encoder Representations from Transformers)

- Basé sur un modèle **Transformer**, comprend les mots dans leur contexte **bidirectionnel** (avant et après).
- Pré-entraîné sur de grandes quantités de texte via des tâches comme **Masked Language Modeling** et **Next Sentence Prediction**.
- Excellente performance pour les tâches NLP comme la classification de texte.

Partie 1 : Prétraitements, extractions de features

Construction de features avec des méthodes NLP avancées : USE (Universal Sentence Encoder)

- Encode des **phrases entières** en vecteurs, capturant le sens global de la phrase.
- Idéal pour des tâches comme la **similarité de phrases** et la **classification** à un niveau plus large que le mot.



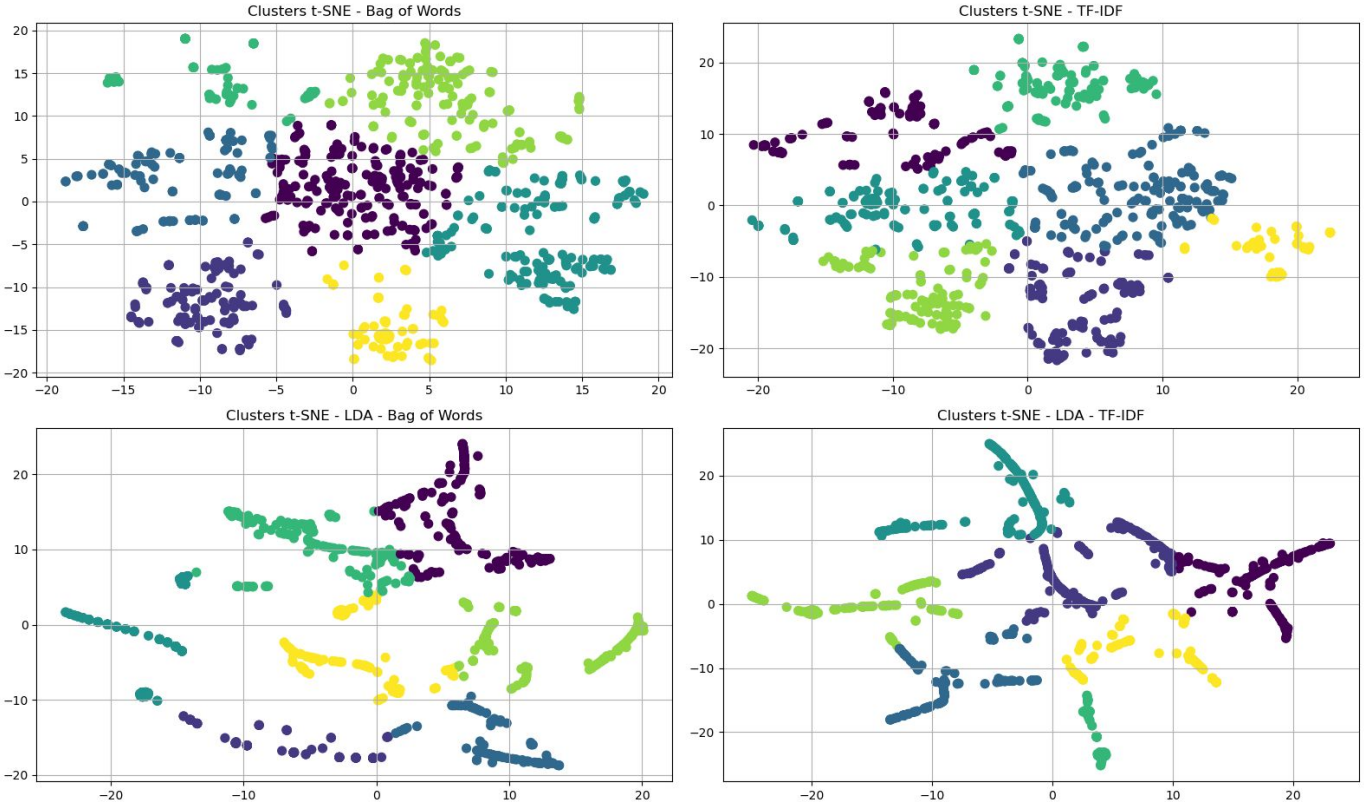
https://spotintelligence.com/

Partie 2 : étude de faisabilité

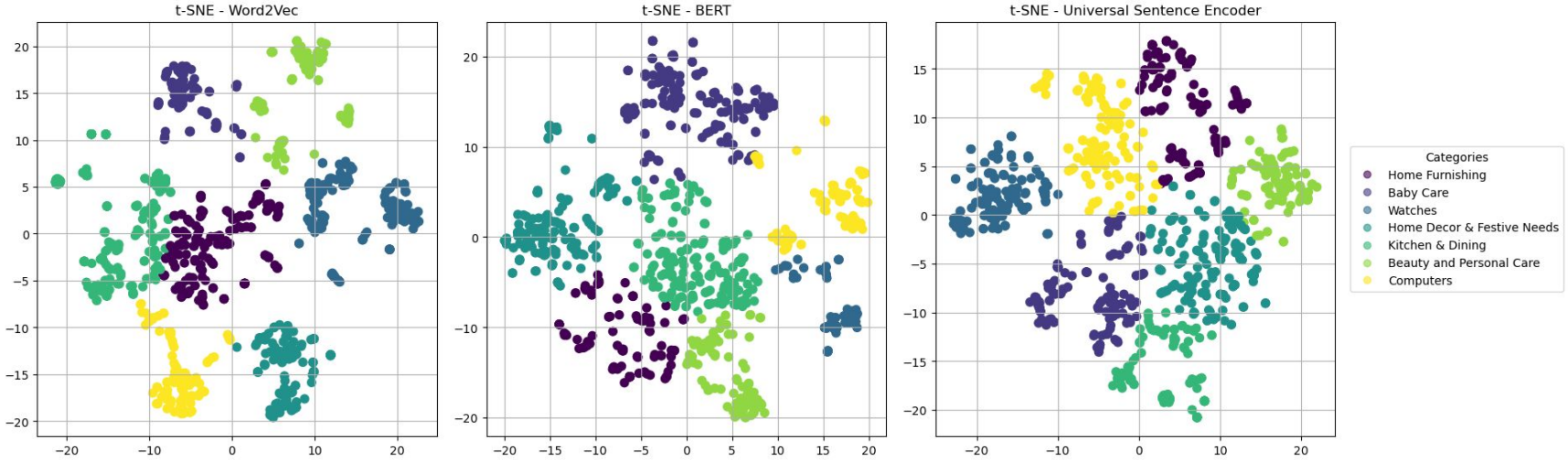
Protocole création espace 2D :

- Réduction de dimensionnalité avec **ACP** et visualisation avec **t-sne**
- Clustering avec **k-means**
- Calcul des métriques de qualité : **Score de silhouette** et **Index de Davies-Bouldin**
- Calcul de l'ARI **(Adjusted Rand Index)** : **Mesure de similarité**, pour évaluer la qualité d'un clustering en comparant un ensemble de **clusters prédits** à des **classes réelles**.

Partie 2 : étude de faisabilité



Partie 2 : étude de faisabilité



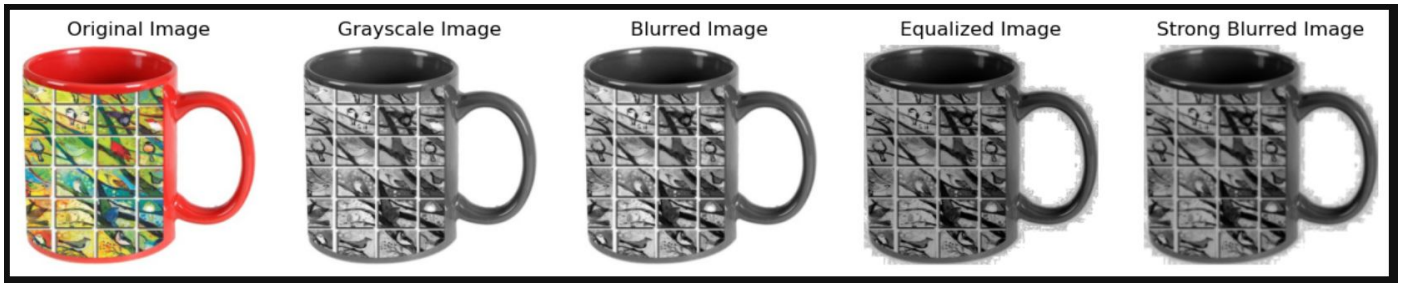
Partie 2 : étude de faisabilité

	Méthodes basiques				Méthodes avancées		
	Bag of Words	TF-IDF	LDA - Bag of Word	LDA - TF-IDF	Word2Vec	BERT	USE
Silhouette Score	0.460	0.444	0.451	0.478	0.467	0.467	0.434
Davies-Bouldin Index	0.714	0.761	0.759	0.707	0.726	0.698	0.758
Adjusted Rand Index (ARI)	0.460	0.471	0.228	0.267	0.365	0.255	0.402

Partie 2 : étude de faisabilité

Prétraitement des Images : OpenCV

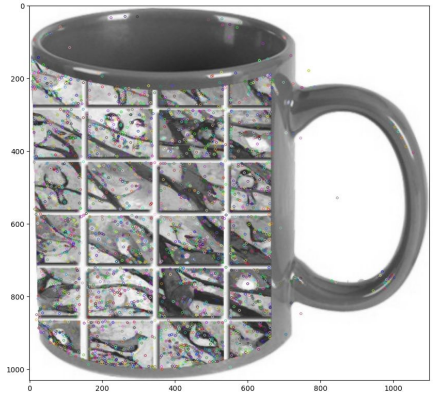
1) **Conversion en niveaux de gris** : simplifier les traitements ultérieurs en réduisant les informations de couleur
2) **Application d'un flou gaussien** : atténuer le bruit
3) **Égalisation d'histogramme** : améliorer le contraste de l'image en redistribuant les niveaux de luminosité
4) **Flou gaussien plus fort** : permettre de voir la différence entre un flou léger et un flou plus prononcé



Partie 2 : étude de faisabilité

Extraction de Features (caractéristiques): SIFT (*Scale-Invariant Feature Transform*)

1) Chargement de l'image en niveaux de gris.
2) Création de l'objet SIFT.
3) Détection des *keypoints* et extraction des descripteurs.
4) Sauvegarde et affichage des *keypoints* sur l'image.
5) Enregistrement des descripteurs pour une utilisation ultérieure.

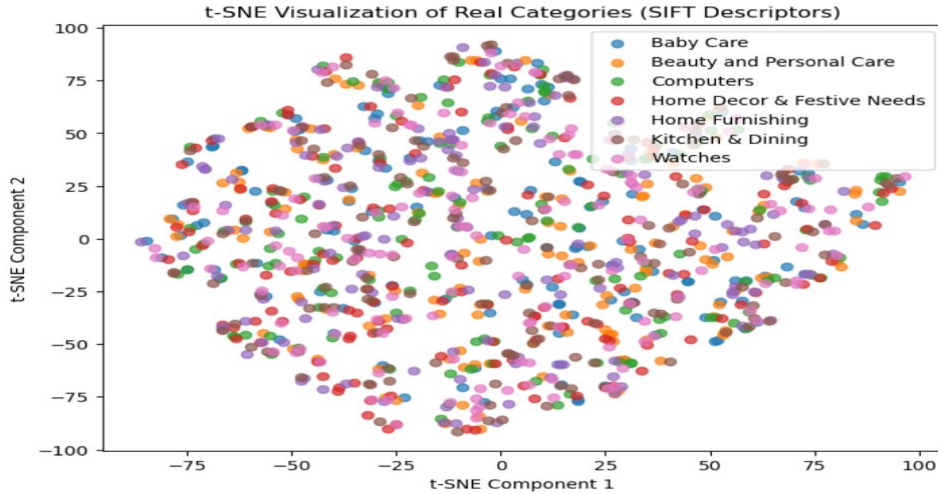


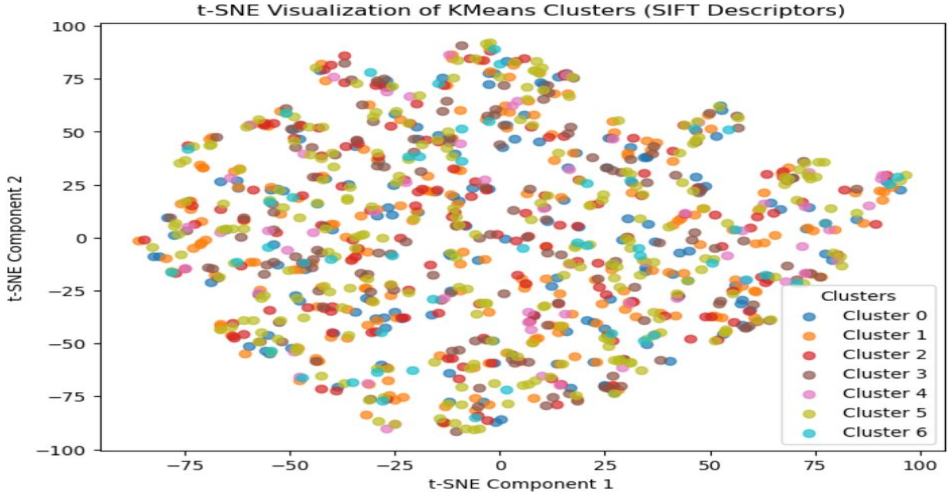
Nombre de features clés (Coffee Mugs) : **3068**

Partie 2 : étude de faisabilité

Visualisation des catégories réelles et des clusters avec SIFT descriptors

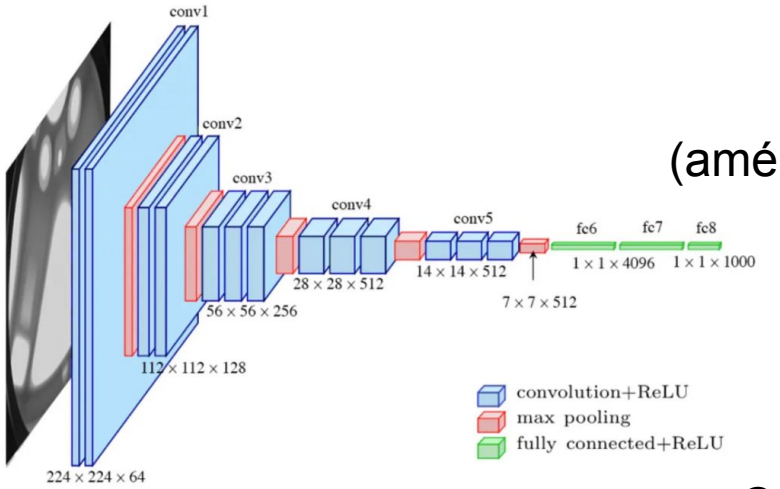
Adjusted Rand Index (ARI) : 0.0019





Partie 2 : étude de faisabilité

Architecture CNN (Convolutional Neural Network) : VGG16



The architecture of VGG16.
Source: Researchgate.net

Couche convolution:
Détecte les patterns d'une image

Couche pooling:
Réduit la taille de l'image
(améliore le temps de calcul + évite le surapprentissage)

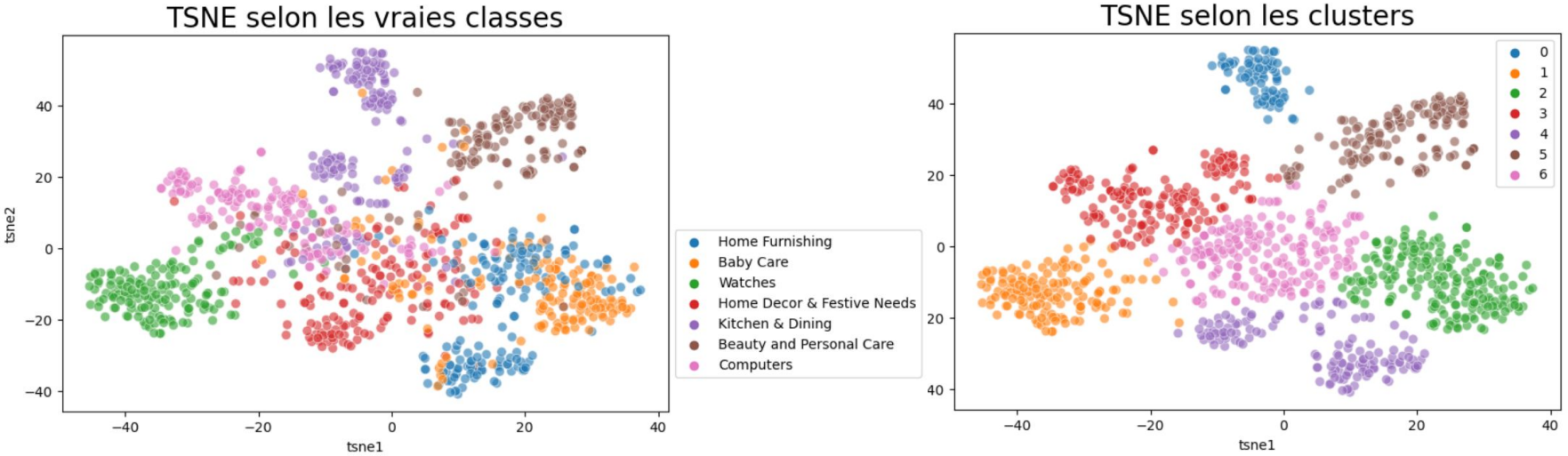
Couche ReLU:
Rend l'architecture non linéaire

Couche fully connected:
Combine tous les patterns appris pour effectuer la classification finale

Partie 2 : étude de faisabilité

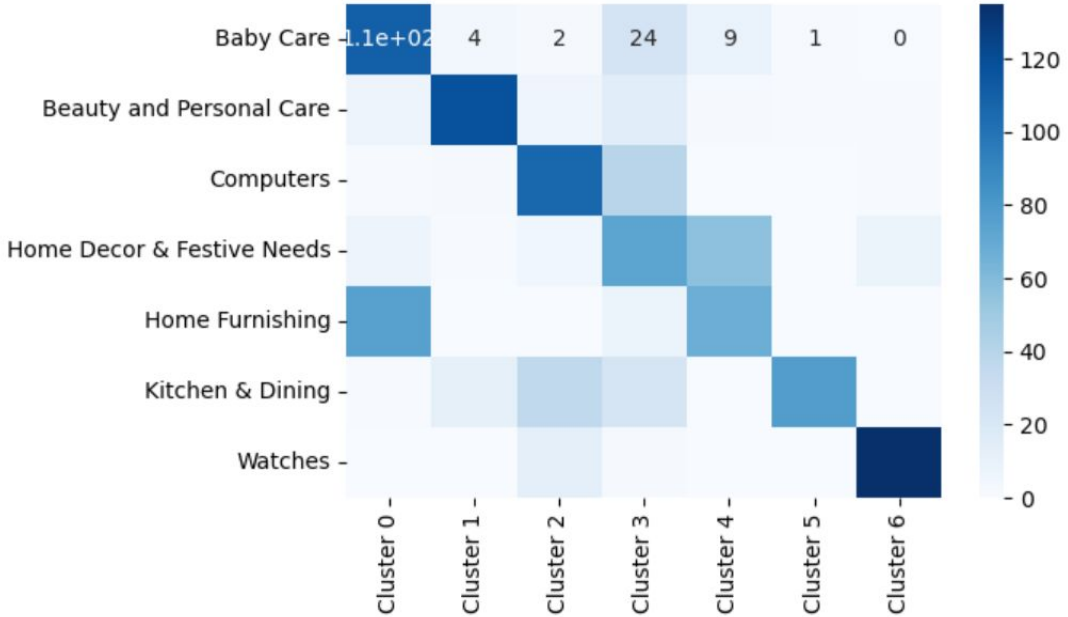
Visualisation des clusters formés par **VGG16**

ARI: **0.439**



Partie 2 : étude de faisabilité

La classe la plus prédite est **Watches** suivi de **Beauty and Personal Care** et **Computers**



Partie 3 : Classification supervisée

Train (60%): permet d'entraîner le modèle.

Validation (20%): permet d'ajuster les hyperparamètres du modèle et d'éviter le surapprentissage (**overfitting**).

Test (20%): sert à évaluer la performance finale du modèle de manière totalement indépendante.

Stratifiés: permet de s'assurer que la répartition des classes (catégories) dans les ensembles créés est **proportionnellement la même** que dans l'ensemble de données d'origine.

Partie 3 : Classification supervisée

Comparaison des résultats : **Nb Epochs = 10, batch_size = 32**

	test_accuracy	test_loss	temps (s)
Simple	**0.2952**	**3.1414**	**4**
Data generator avec data augmentation	**0.1429**	**1.9484**	**141**
DataSet, sans data augmentation	**0.4305**	**1.7090**	**5**
DataSet, avec data augmentation	**0.1198**	**1.9579**	**139**

Partie 4 : Test de l’API

Objectif : Collecte de données ⇒ Nouvelle catégorie (à base de Champagne)

Première étape : Envoyer une requête pour récupérer des produits liés à "champagne" via l'API Spoonacular.

API:https://api.spoonacular.com/food/products/search?query=champagne&apiKey={api_key}

Deuxième étape : Sauvegarder les résultats de cette requête dans un fichier CSV.

Troisième étape : Afficher ces résultats sous forme de tableau.

Partie 4 : Test de l’API

	foodId	label	category	\
0	10461678	Terrine de canard au champagne et miel	N/A	
1	2049910	Champagne Waris-Larmandier Brut Racines de Trois	N/A	
2	5885230	Champagne Leguillette Romelot Cepages d'Autref...	N/A	
3	9734138	Pate de higado de pato con champagne	N/A	
4	5884966	Champagne Leclerc Briant Blanc de Meuniers Bru...	N/A	
5	6405920	Lionne Royale Brut Champagne 750ml	N/A	
6	10117102	Champagne & strawberries marshmallow covered i...	N/A	
7	461061	Champagne Collet Brut Vintage Collection Privee	N/A	
8	11429974	Galantine de dinde à la fine champagne	N/A	
9	10304956	Tartinade de framboises et peches au champagne	N/A	

	foodContentsLabel	image
0	N/A	https://img.spoonacular.com/products/10461678-...
1	N/A	https://img.spoonacular.com/products/2049910-3...
2	N/A	https://img.spoonacular.com/products/5885230-3...
3	N/A	https://img.spoonacular.com/products/9734138-3...
4	N/A	https://img.spoonacular.com/products/5884966-3...
5	N/A	https://img.spoonacular.com/products/6405920-3...
6	N/A	https://img.spoonacular.com/products/10117102-...
7	N/A	https://img.spoonacular.com/products/461061-31...
8	N/A	https://img.spoonacular.com/products/11429974-...
9	N/A	https://img.spoonacular.com/products/10304956-...

Conclusion

- La faisabilité d'un **moteur de classification automatique** est validé.
- La classification **supervisée à partir des images** est bien.
- L'approche la plus performante pour la classification supervisée est **DataSet sans data augmentation**.