Classifiez automatiquement des biens de consommation

Présentation du jeu de donnée

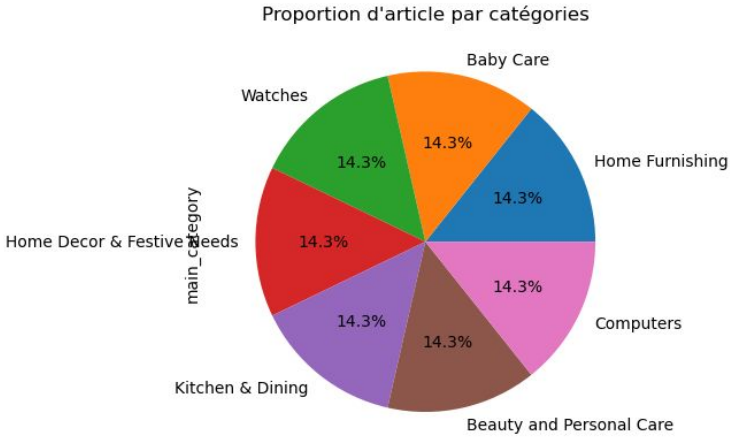
- Table de 1050 produits et de 15 descripteurs
  - On travaille sur 3 descripteurs :
    - product_category_tree
    - image
    - description
- Dossier de 1050 images au format jpg

Objectif du projet

1. Faire une étude de la faisabilité d'une segmentation :
    1.1. A l'aide du traitement naturel du language
        1.1.1. Par Bag-of-Words
            1.1.1.1. Méthode Count Vectorizer
            1.1.1.2. Méthode TF-IDF
        1.1.2. Par word/sentence embedding
            1.1.2.1. Méthode Word2Vec
            1.1.2.2. Méthode BERT
            1.1.2.3. Méthode USE
    1.2. A l'aide du traitement d'image
        1.2.1. Algorithme SIFT
        1.2.2. Algorithme CNN transfer Learning
2. Classification supervisée à partir d'image (par data augmentation)
3. Faire une collecte de produit à partir d'une API

Nettoyage des données

- Récupération des catégories principales :
  - product_categorie_tree :
    - Baby Care
    - Watches
    - Home Furninshing
    - Home Decore & Festive Needs
    - Computers
    - Kitchen & Dining
    - Beauty & Personal Care



Pre-processing des données textes

- Utilisation de la bibliothèque NLTK
- Colonne ‘description’

1. Tokenization
2. Passage en minuscule
3. Suppression des caractères spéciaux (?, !, ., ‘, ...)
4. Suppression des chiffres
5. Suppression des stopwords
6. Lemmatization (improving -> improve)
7. Suppression des mots de longueur 1 (lettre seule)
8. On vérifie que les mots appartiennent au corpus de mots de NLTK

Avant/Après Pre-Processing

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors

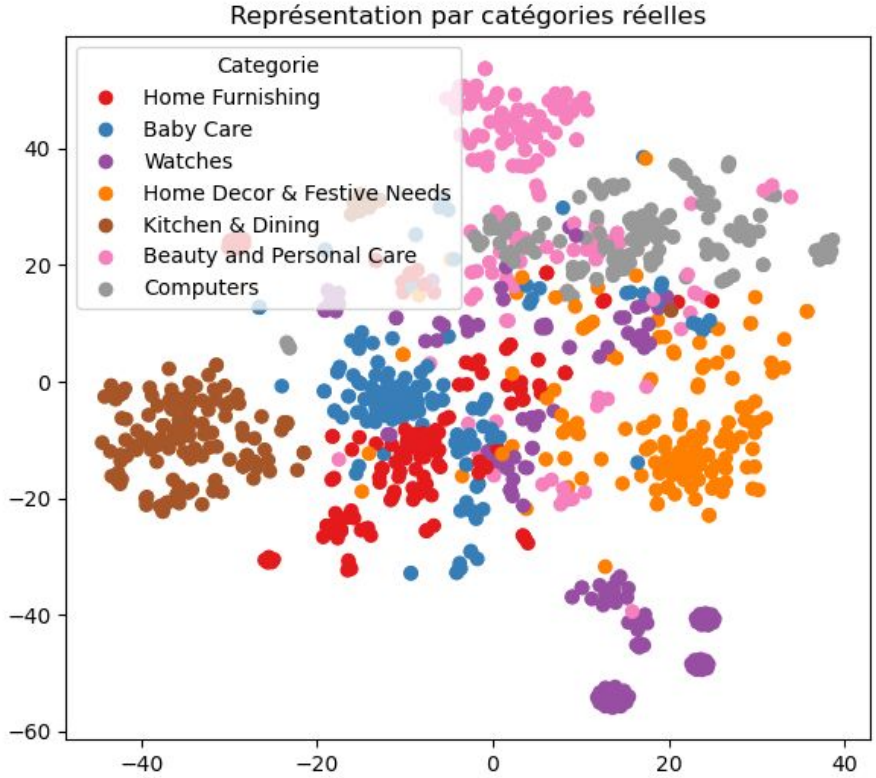
key feature elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract eyelet door curtain height pack price curtain look interior

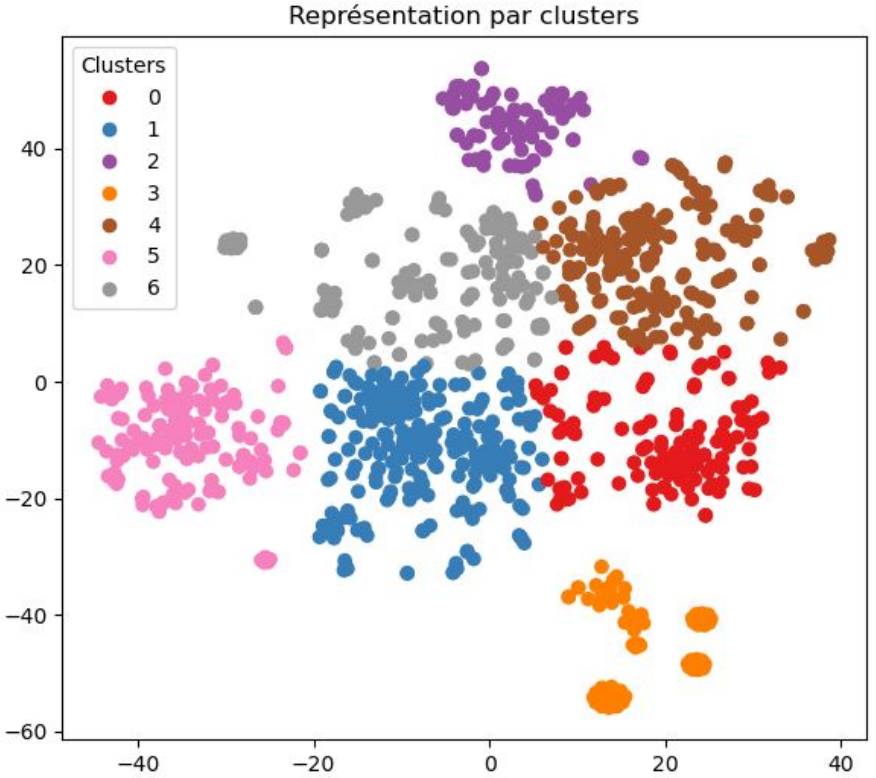
Feature Extraction, Réduction t-sne et Clustering

CountVectorizer

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Résultats





ARI Score : 0.43

Tf-Idf

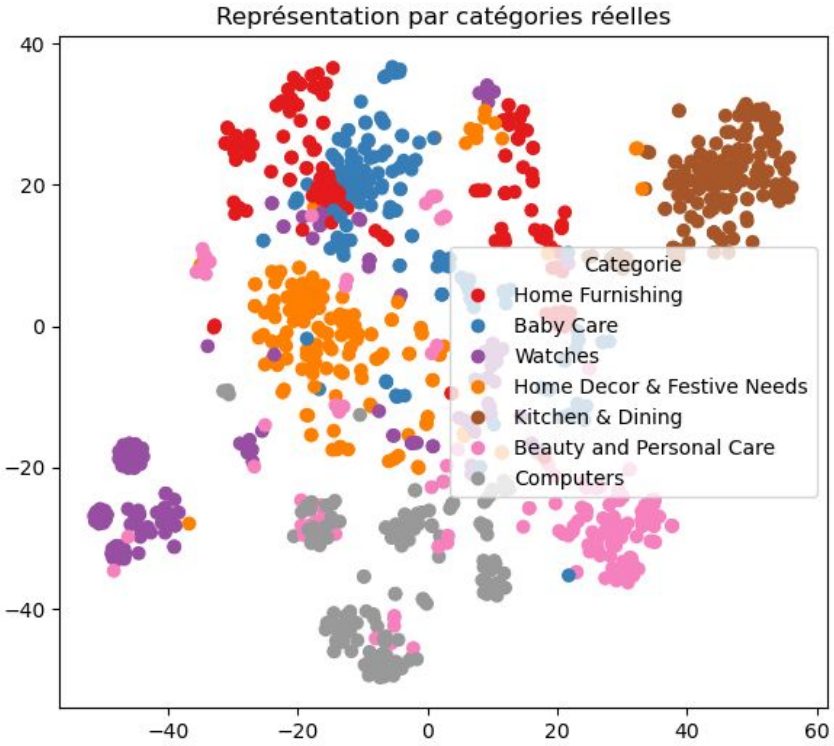
	cups	do	flour	has	in	keyboard	mac	minutes	most	noisy	of	or	prefer	replace	the	windows	you
0	0.61	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.51	0.00	0.00
2	0.00	0.00	0.00	0.00	0.52	0.42	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.42	0.35	0.00	0.00
3	0.00	0.42	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.42	0.42	0.00	0.00	0.42	0.42
4	0.00	0.00	0.00	0.41	0.00	0.33	0.33	0.00	0.41	0.41	0.00	0.00	0.00	0.00	0.54	0.00	0.00

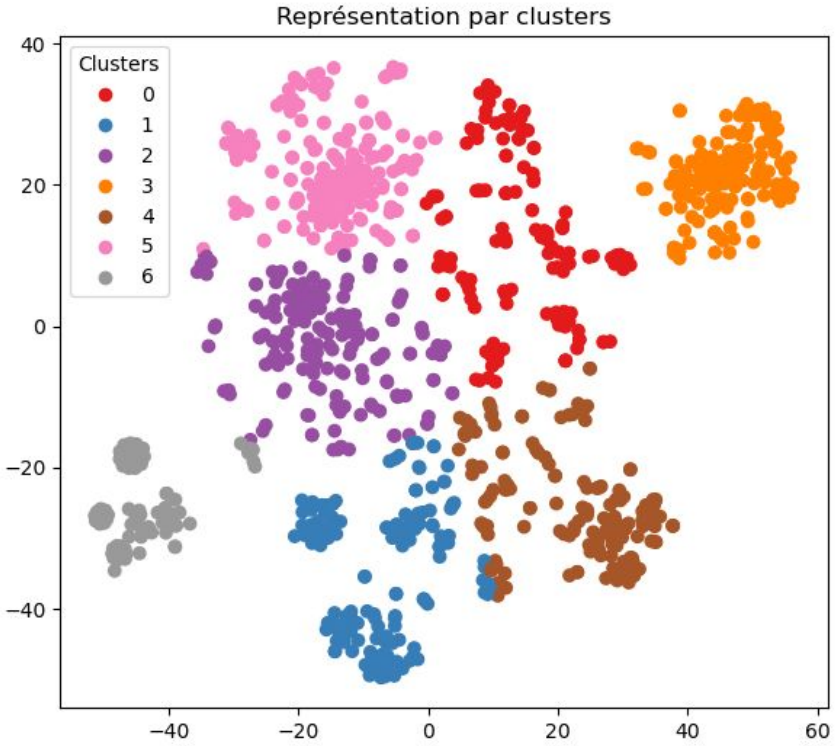
$$TF = \frac{Number\ of\ times\ a\ word\ "X"\ appears\ in\ a\ Document}{Number\ of\ words\ present\ in\ a\ Document}$$

$$IDF = log\left(\frac{Number\ of\ Documents\ present\ in\ a\ Corpus}{Number\ of\ Documents\ where\ word\ "X"\ has\ appeared}\right)$$

$$TF\ IDF\ =\ TF * IDF$$

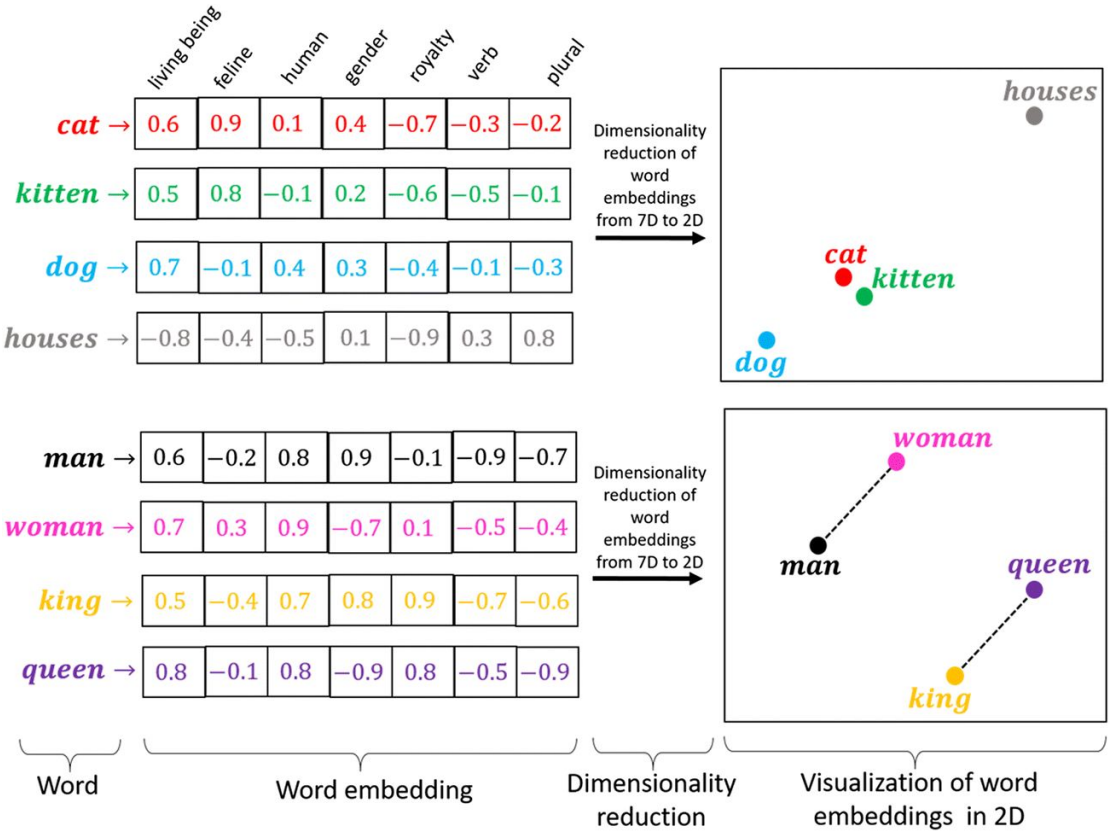
Résultats



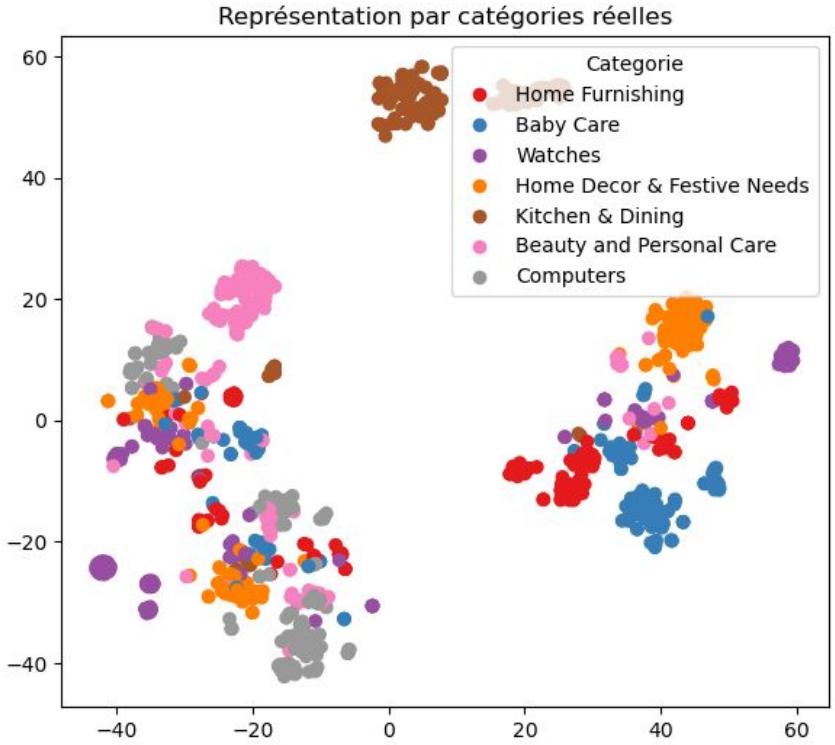


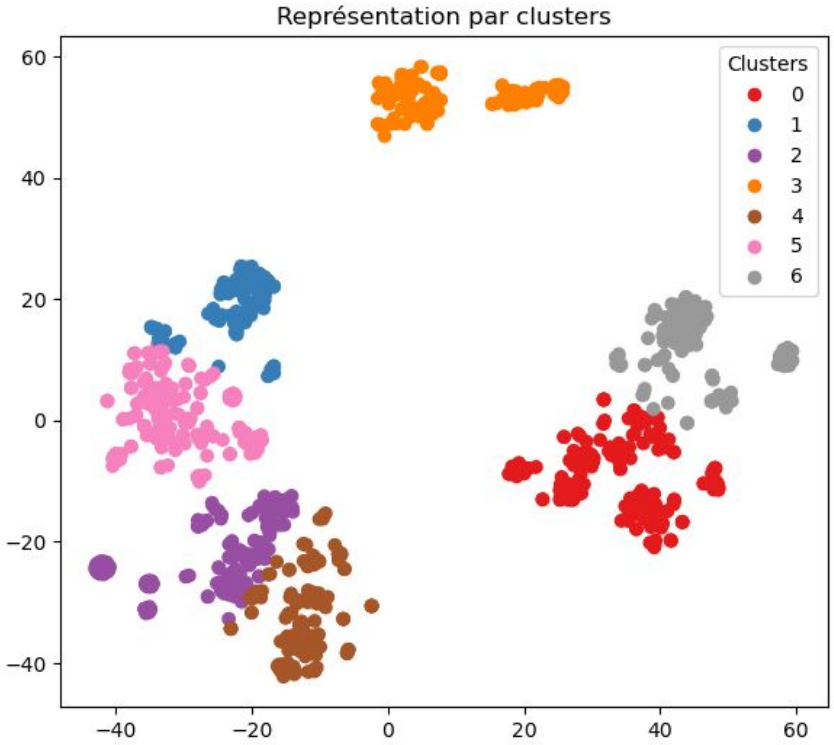
ARI Score : 0.47

Word2Vec


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

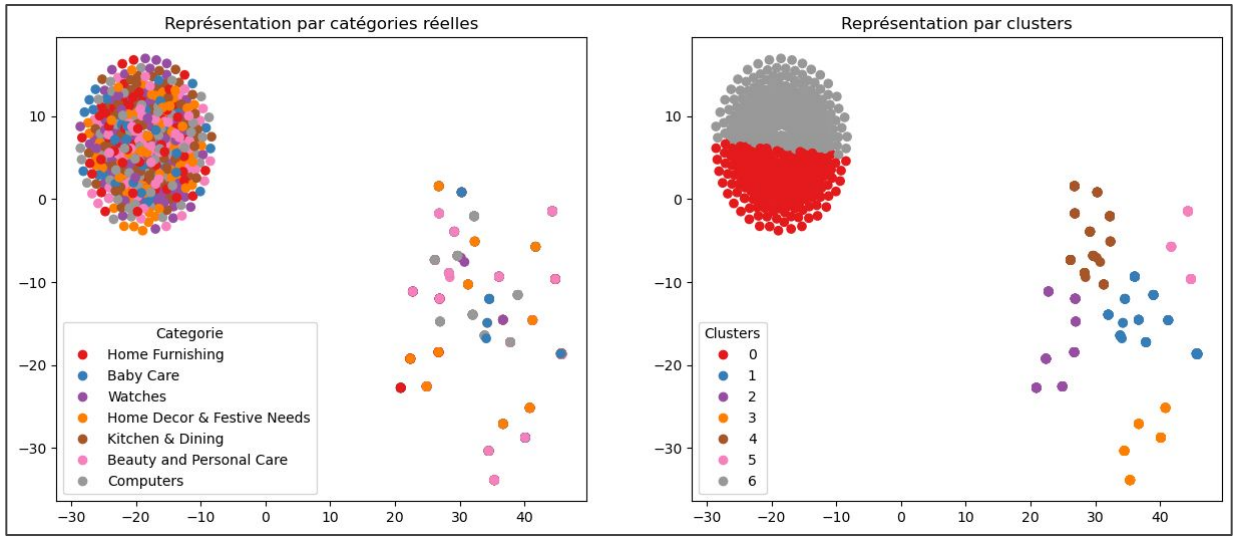
Résultats





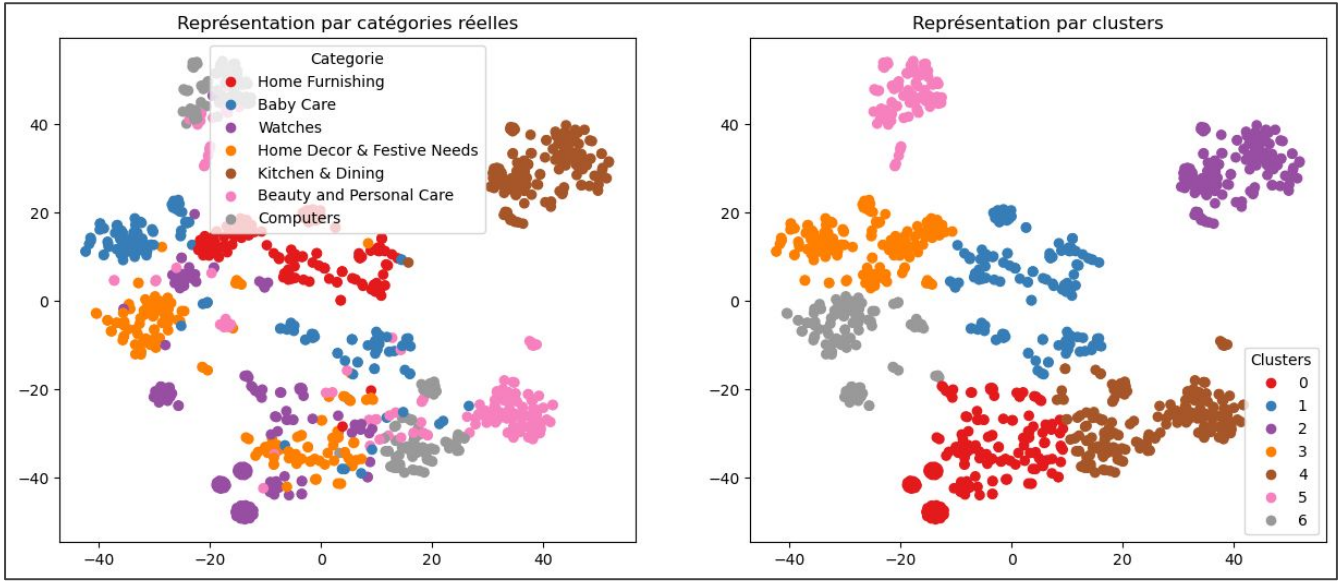
ARI Score : 0.30

BERT



- Modèle développé par Google
- Pas adapté à notre corpus (ARI Score = 0.03)

Universal Sentence Encoder



- Modèle développé par TensorFlow
- Très efficace dans notre cas (ARI Score 0.45)

Feature extraction d’image, réduction t-sne et Clustering

Preprocessing image

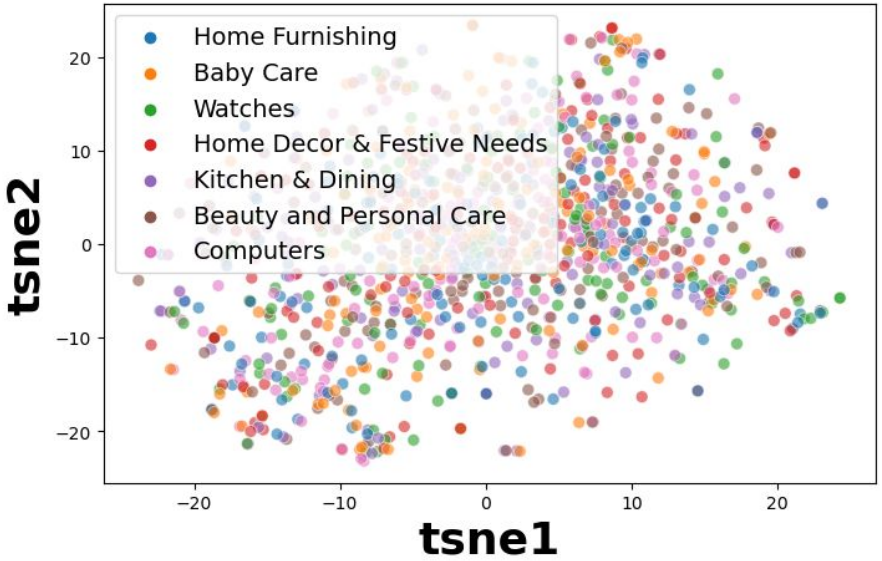
- Conversion en gris :
  - Ignore les couleurs et se concentre sur les reliefs
- Égalisation des couleurs :
  - Mets en relief les bords et élément important de l'image
- Redimensionnement des images en 224x224 :
  - Perte d'information limité
  - Permet au modèle de tourner plus vite

SIFT

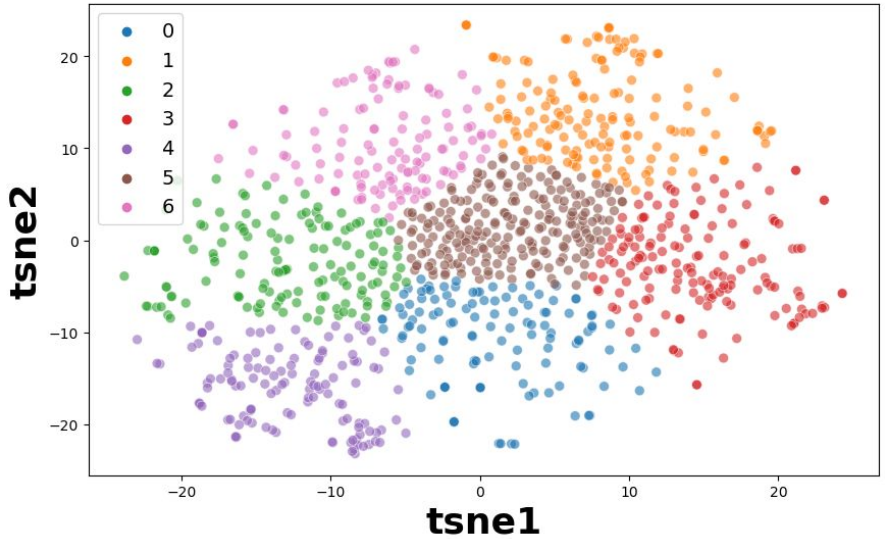
- Création de 314 074 descripteurs de taille 128 chacun
- Création de 560 cluster de descripteur (Clusters servant de “coordonnées” aux descripteurs)
- Création des features images :
  - Pour chaque image on compte le nombre de descripteur dans le premier cluster, puis dans le deuxième , etc... jusqu’au 560ème clusters
  - On a donc à ce stade 1050 images x 560 descripteurs
- Réduction de dimension ACP (passe de 560 descripteurs à 452) :
  - Permet un clustering plus performant et rapide du t-SNE
- Réduction de dimension t-SNE (passe de 452 descripteurs à 2)

Résultats

TSNE selon les vraies classes



TSNE selon les clusters

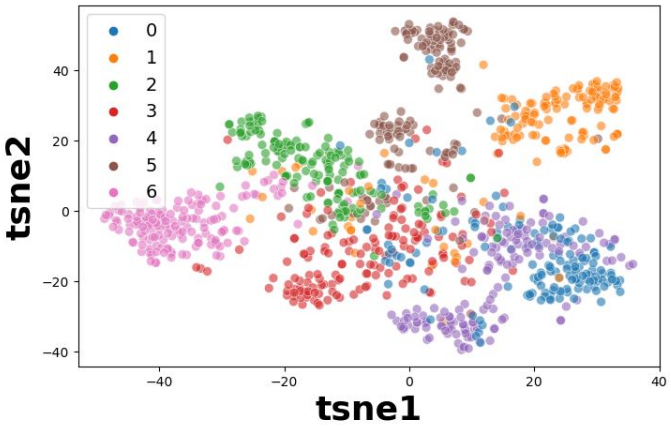


ARI Score : 0 (très mauvais)

Transfer Learning

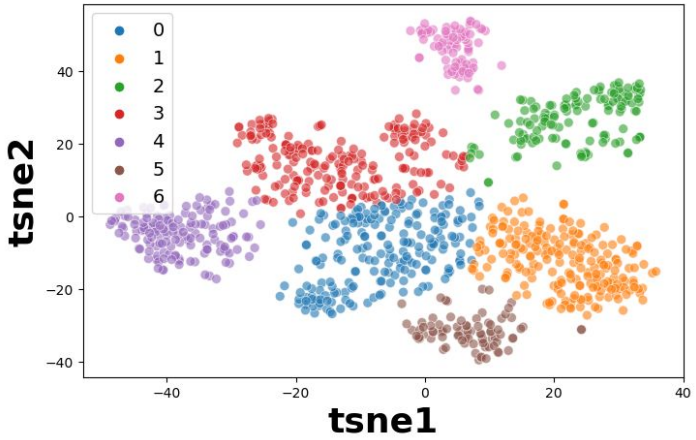
- Modèle de base : VGG16 (sur image 224x224)
- ACP pour réduire la dimension (passant de 4096 à 803):
  - Réduction meilleur et plus rapide pour le t-SNE
- t-SNE pour réduire à 2 dimensions

TSNE selon les vraies classes



ARI Score : 0.48

TSNE selon les clusters



Compréhension des images mal classée

- Image classée dans ‘Décoration Maison’ alors que sa catégorie réelle est Baby Care
- Confusion de certaines catégories en réalité très proche
- Pas assez d’information sur cette image pour que le modèle saisisse la subtilité entre décoration maison et baby care

Catégorie les mieux et moins bien classées

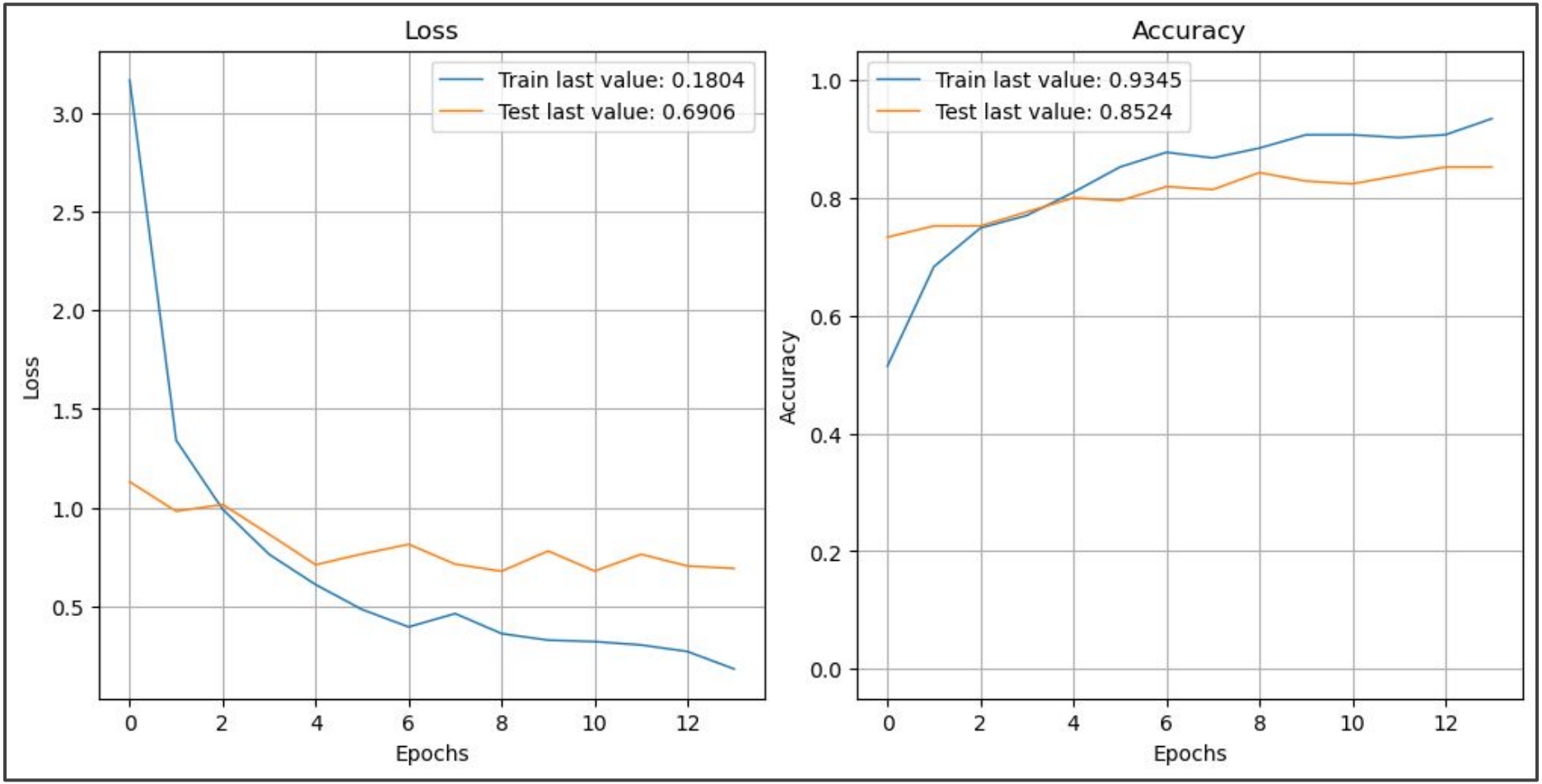
Première conclusion

- 3 méthodes de feature extraction efficace :
  - Tf-Idf
  - CountVectorizer
  - USE
- Faisabilité du clustering par traitement du langage naturel
- Méthodes de features extractions images
  - Modèle SIFT pas concluant
  - Modèle de Transfer Learning très efficace (ARI Score : 0.48)
- Faisabilité du clustering par traitement d'image

Classification supervisée

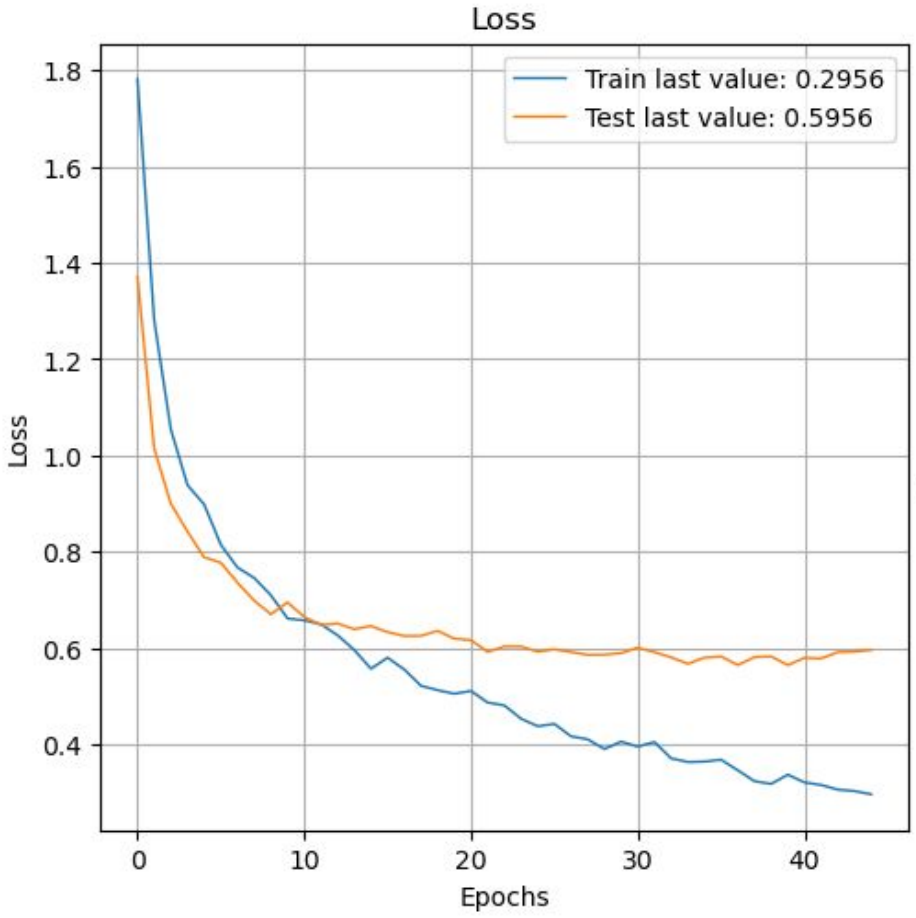
Modèle avec Data Generator

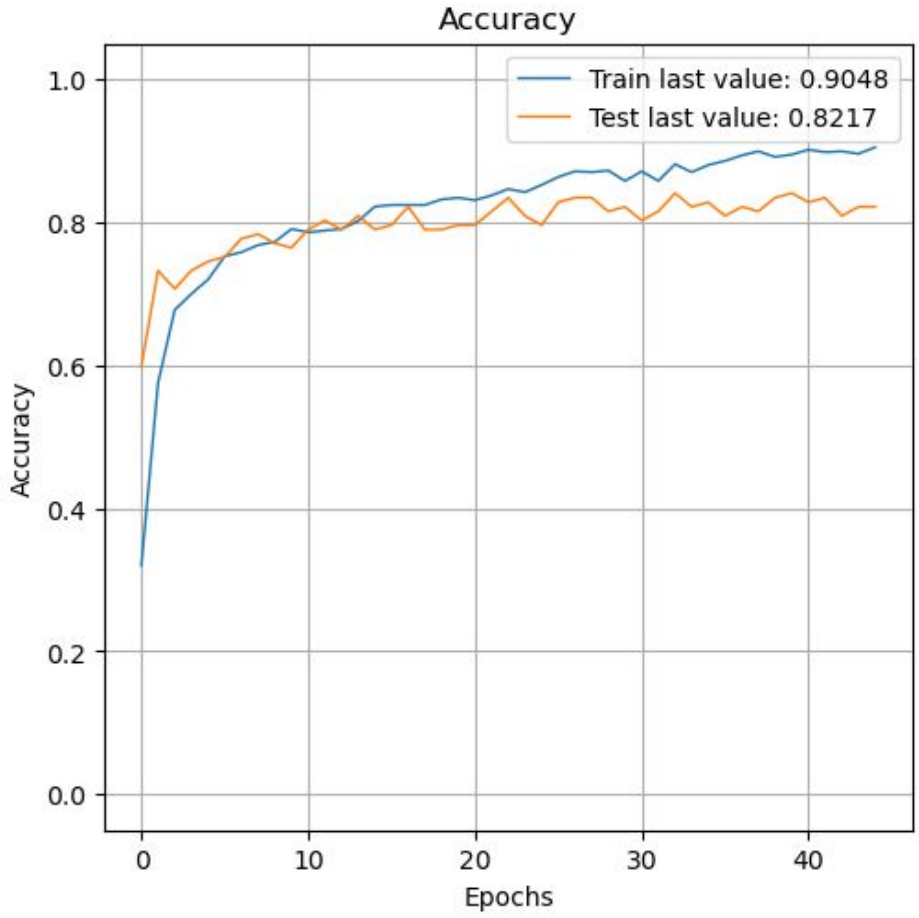
- Modèle VGG16
- Reformatage en 256x256
- Rotation de +- 20 degrés
- Décalage horizontal allant jusqu'à 20% de la taille
- Décalage vertical allant jusqu'à 20% de la taille
- Images retournées horizontalement de manière aléatoire
- Application du One-Hot-Encoding aux labels
- Utilisation de preprocess_input (essentiel pour VGG16):
  - conversion de l'image en numpy
  - 224*224 en entrée
  - mise à l'échelle des valeurs des pixels
- Batch Size = 32
- Epoques = 50
- Taille du Validation set = 0.20


Loss
Train last value: 0.1804
Test last value: 0.6906
Loss
Epochs
Accuracy
Train last value: 0.9345
Test last value: 0.8524
Accuracy
Epochs

Approche avec Data Augmentation intégrée au modèle

- Modèle VGG16
- Retournement horizontal aléatoire avec une proba de 0.5
- Rotation aléatoire de +-10% de l'image initiale
- Zoom aléatoire sur une plage de +-10%
- Normalisation de la taille des pixels (Facilite la convergence du modèle)
- Batch Size = 32
- Epoques = 50
- Taille du Validation set = 0.20


Loss
Train last value: 0.2956
Test last value: 0.5956
Loss
Epochs
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0
10
20
30
40


Accuracy
Train last value: 0.9048
Test last value: 0.8217
Accuracy
Epochs
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40

Collecte API

Information à entrée :

- URL de l'API
- querystring (objet de la requête, ici les produits avec pour ingrédient le champagne)
- headers :
  - X-RapidAPI-Key : clef associée à mon compte
  - X-RapidAPI-Host : Hôte de la requête

Après la requête :

- On récupère un dictionnaire
- Création d'une fonction qui récupère les informations du dictionnaire
- Mise sous forme de Data Frame
- Sélection des 10 premiers produits

Conclusion

- Test de faisabilité pertinent et concluant :
  - NLP :
    - Les méthodes bag-of-words ont le mieux fonctionnés :
      - ARI Score : 0.43 pour CountVectorizer
      - ARI Score : 0.47 pour Tf-Idf
    - USE très efficace également :
      - ARI Score : 0.47
  - Traitement d'image :
    - Transfer Learning très efficace :
      - ARI Score : 0.48
- Classification supervisée :
  - Avec Data Generator :
    - Accuracy de 0.85 sur le Datatest
  - Avec Data augmentation intégrée au modèle :
    - Accuracy de 0.82 sur le Data test

On prendra donc le modèle avec Data Generator car légèrement plus performant.