Classifier automatiquement des biens de consommation

Formation data scientist

Projet 6| Alexis Marceau | Juillet 2022

Sommaire

1. Contexte et problématique

Contexte

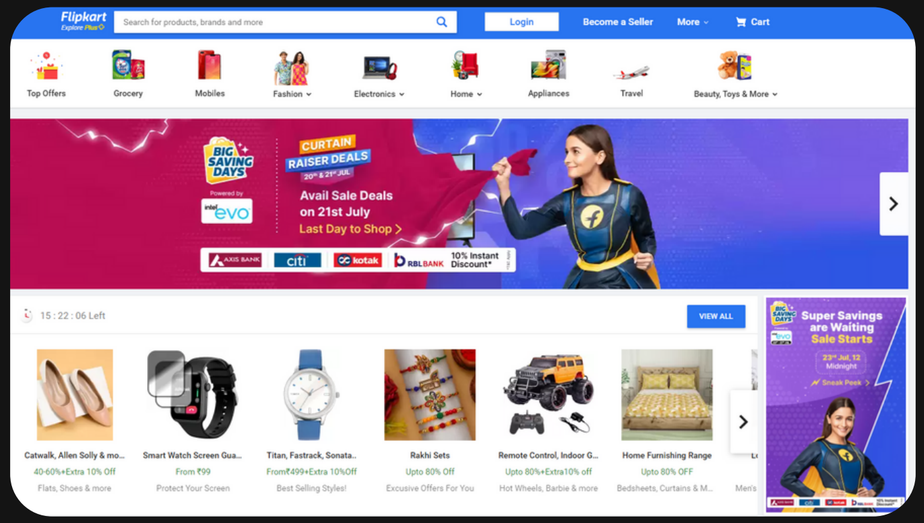
- "Place de marché" souhaite lancer une marketplace e-commerce.
- Des vendeurs proposent des articles à des acheteurs en postant une photo et une description.
- L'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs. C'est une tâche longue et peu fiable.
- Peut-on classifier automatiquement les articles en différentes catégories, avec un niveau de précision suffisant ?

Problématique

La mission est de réaliser une **étude de faisabilité** d'un moteur de classification d'articles basé sur une **image** et une **description textuelle** pour l'automatisation de l'attribution de la catégorie d'un article.

Pas de features algébriques directement exploitable par les modèles de machine learning classique quand on traite du texte ou des images. Il faut en créer.

Données



- Données issues de la base de données flipkart.com

- 1050 produits
- Colonnes : nom, catégorie, image, description, marque, prix...
- Jeux de données très bien rempli

```
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   object
 2   product_url             1050 non-null   object
 3   product_name            1050 non-null   object
 4   product_category_tree   1050 non-null   object
 5   pid                     1050 non-null   object
 6   retail_price            1049 non-null   float64
 7   discounted_price        1049 non-null   float64
 8   image                   1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null   bool
 10  description             1050 non-null   object
 11  product_rating          1050 non-null   object
 12  overall_rating          1050 non-null   object
 13  brand                   712 non-null    object
 14  product_specifications  1049 non-null   object
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```

Sélection des données

Nom du produit:
HMT Sonata Gold Plated Watch For Men Sonata Analog Watch - For Men

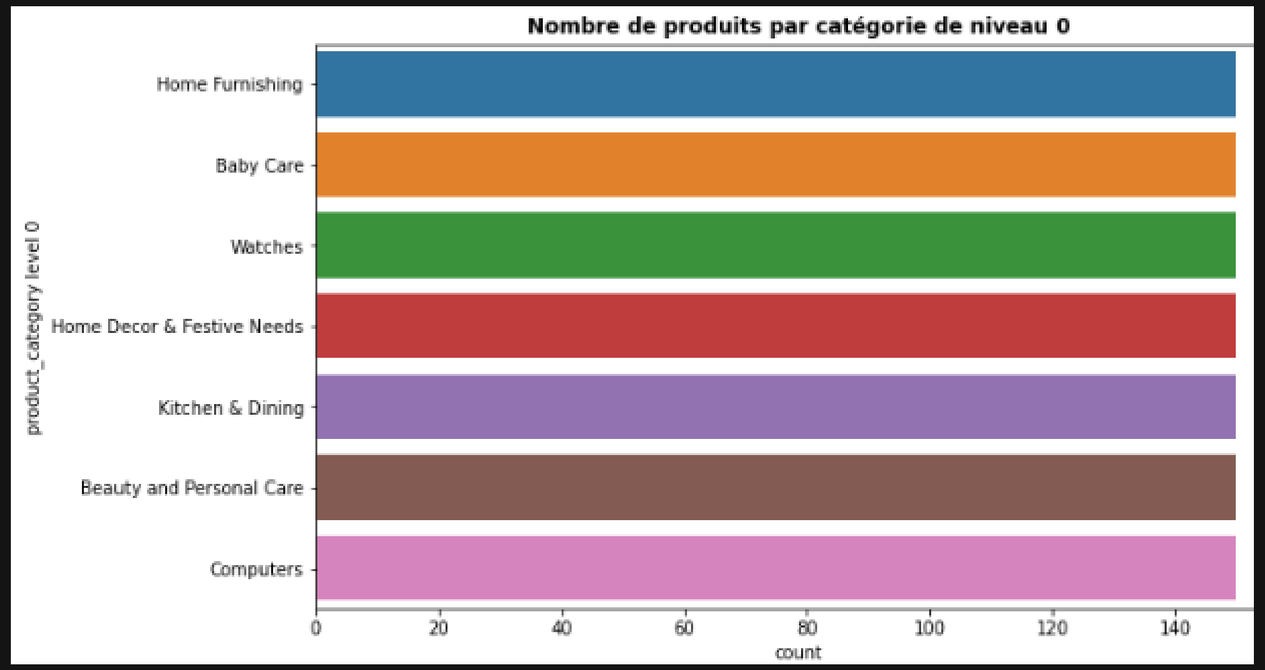
Catégorie: Watches
Description: HMT Sonata Gold Plated Watch For Men Sonata Analog Watch - For Men - Buy HMT Sonata Gold Plated Watch For Men Sonata Analog Watch - For Men Sonata Gold Plated Watch For Men Online at Rs.899 in India Only at Flipkart.com. - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!

Ce qui nous intéresse :

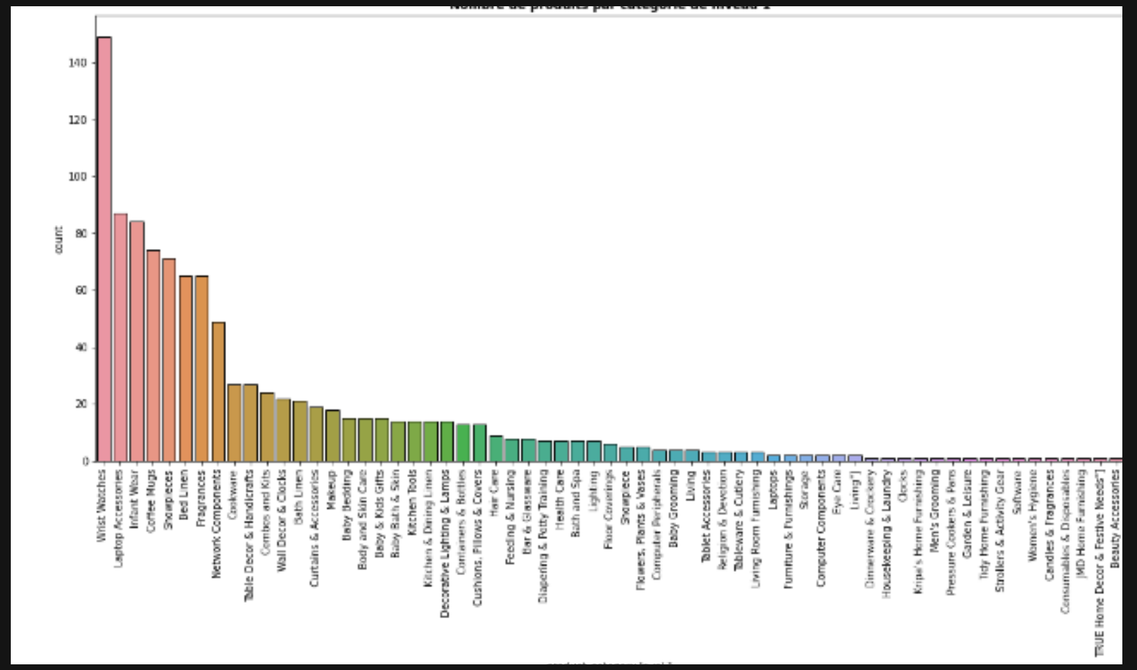
- Données textuelles : nom, description et catégories produit
- Données images : image du produit

Données : catégories

- Organisées sous forme d'arbre à **plusieurs niveaux**
- Extraction niveau 0 et niveau 1 :



niveau 0 : 7 catégories homogènes



niveau 1 : 63 catégories

➡ Focalisation sur les catégories de niveau 0

Démarche générale

01

Preprocessing

02

Extraction des features

03

Réduction de dimension

04

Clustering

04

Evaluation de la correspondance des clusters avec les vraies catégories

2. Analyse textuelle

Nettoyage du texte

Texte avant traitement:

HMT Sonata Gold Plated Watch For Men Sonata Analog Watch - For Men - Buy HMT Sonata Gold Plated Watch For Men Sonata Analog Watch - For Men Sonata Gold Plated Watch For Men Online at Rs.899 in India Only at Flipkart.com. - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!

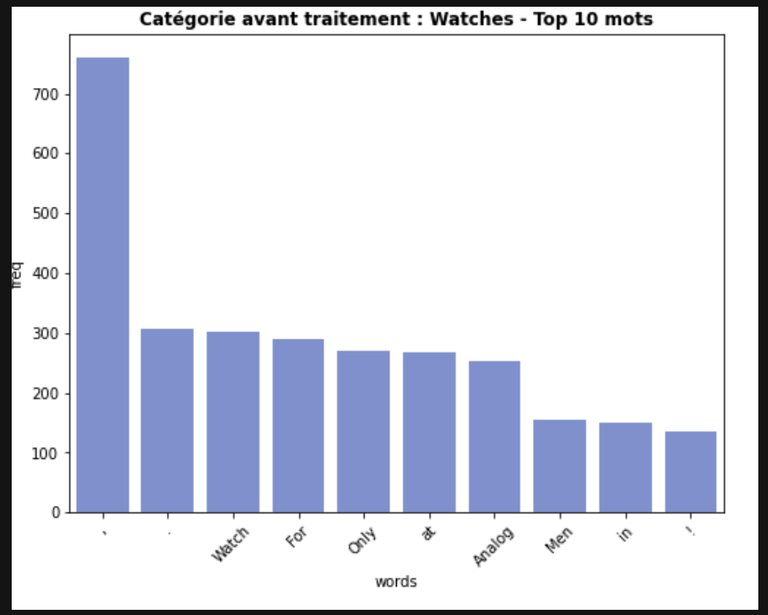
Après traitement:

hmt sonata gold plated watch men sonata analog watch men buy hmt sonata gold plated watch men sonata analog watch men sonata gold plated watch men online rs.899 india great discount genuine product replacement guarantee free shipping cash delivery

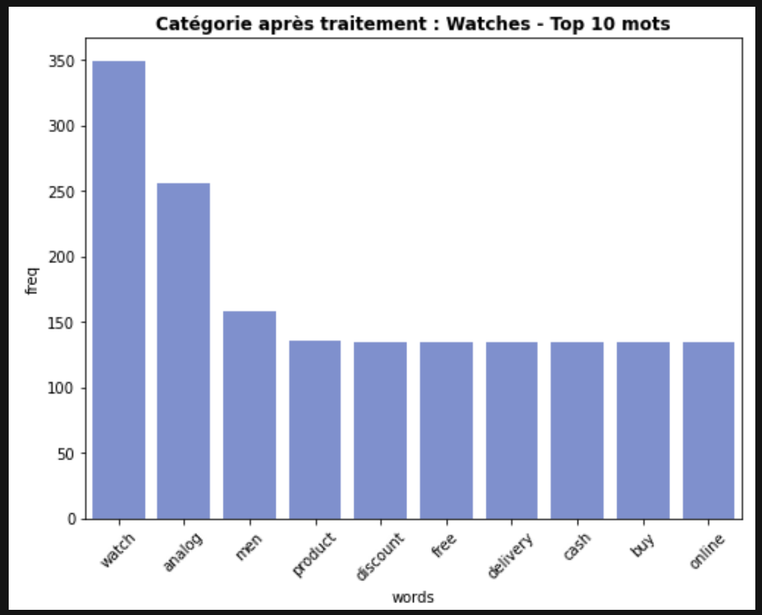
➡ Utilisation de la librairie NLTK

- Suppression de la ponctuation
- Tokenisation : découpage en mots
- Suppression des stopwords : mots très courants mais qui n'apportent pas de valeur informative pour la compréhension du texte (the, for, this...)
- Lemmatisation : réduction des mots à leurs forme canonique (am, are, is... => be)

Nettoyage du texte



Avant traitement



Après traitement

Extraction des features texte

- Techniques de types bag of words :
  - CountVectorizer : compte le nombre de fois qu'un mot apparait dans un document.
  - TF-IDF : (Term Frequency - Inverse Document Frequency) : score qui permet d'accorder plus d'importance à des mots plus rares.
- Techniques de types word embeddings :
  - Word2Vec : Un seul vecteur pour chaque mot. Les différents sens du mot sont combinés en un seul vecteur.
  - BERT et USE: Plusieurs représentations vectorielles pour le même mot, en fonction du contexte dans lequel le mot est utilisé.

Réduction de dimension

Techniques de types bag of words :

- Création de "matrices creuses" : peut biaiser les algorithmes qui considèrent ainsi que les observations à zéro (qui sont présentes en majorité) représente une information à prendre en considération. PCA non adapté, SVD possible.

Techniques de types word embeddings :

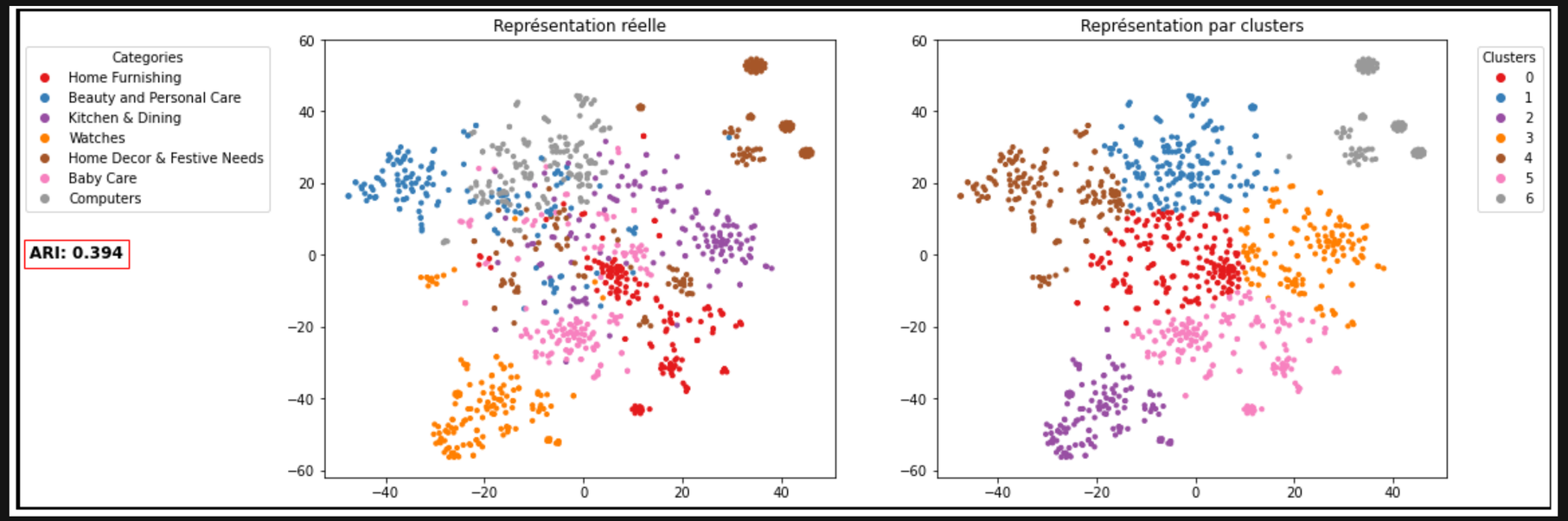
- PCA (95% de la variance des données)

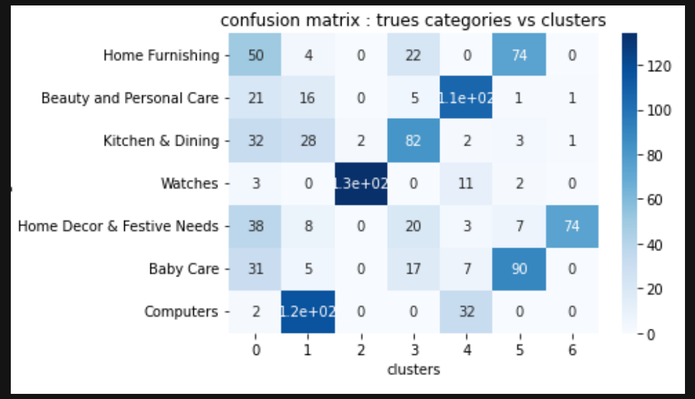
Visualisation :

=> Réduction de dimension en 2 composantes T-SNE pour affichage en 2D

CountVectorizer

TSNE et k-means clustering

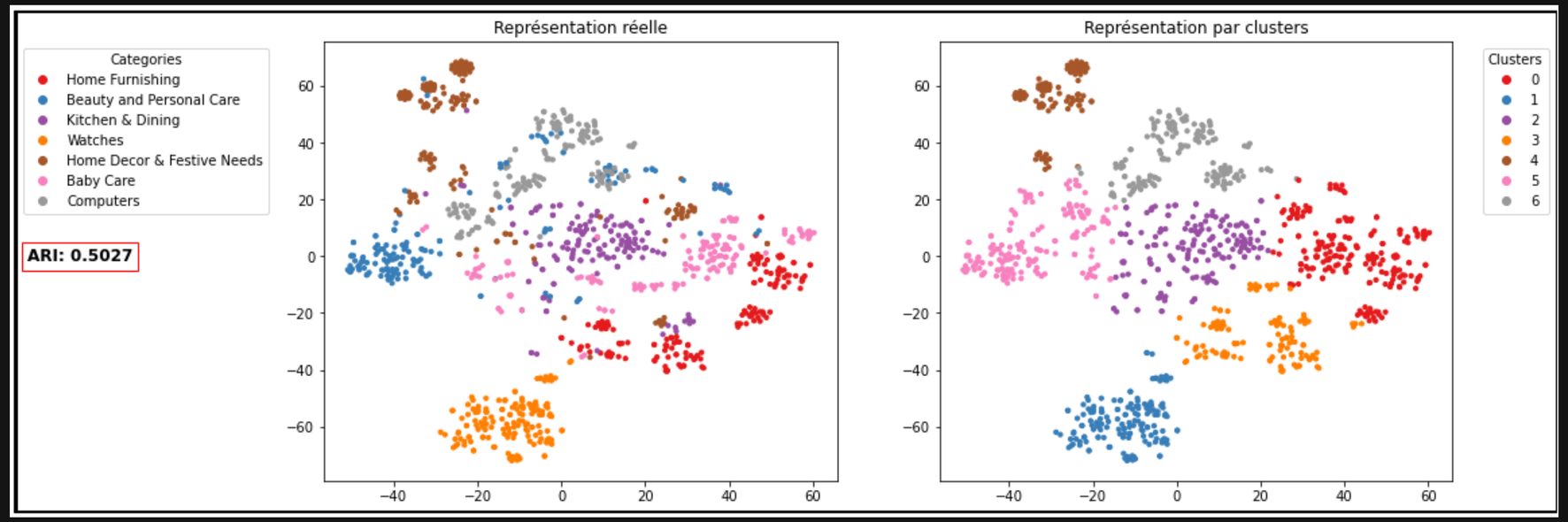


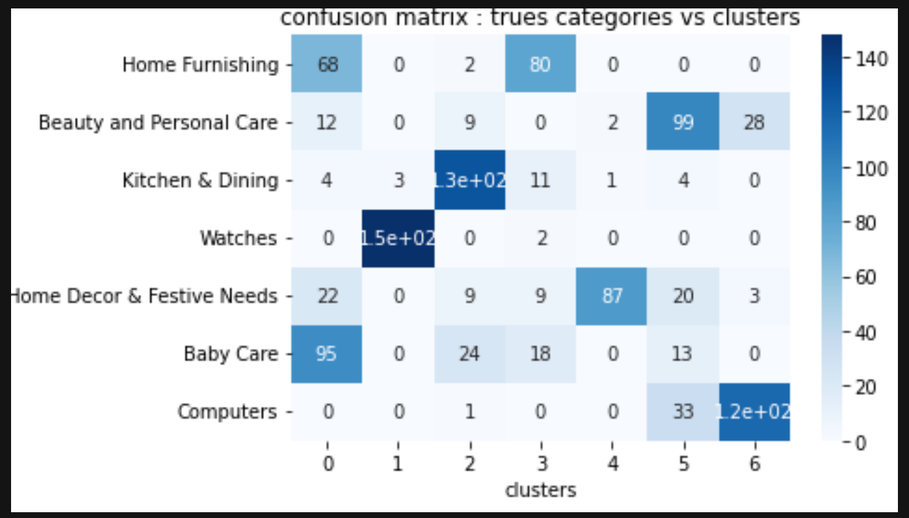


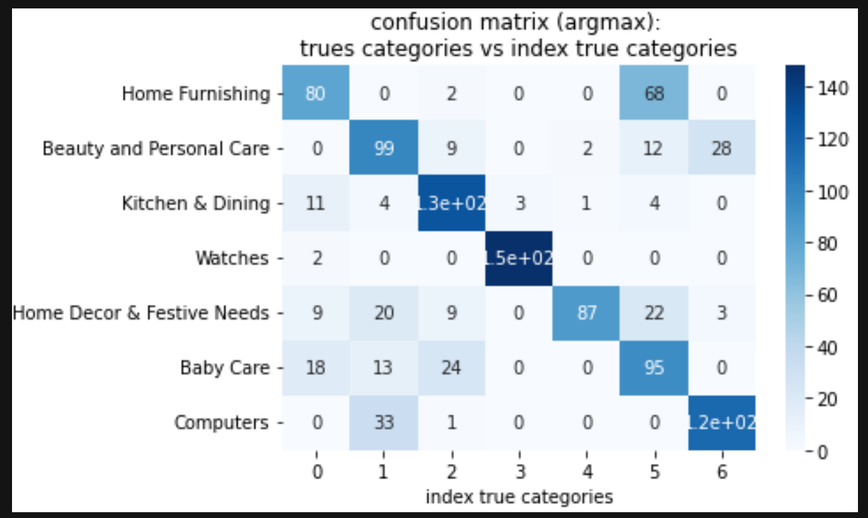
- Attribution des clusters réussi pour :
  - Computers (label 1)
  - Watches (label 2)
  - Beauty and Personal Care (label 4)
  - Home Decor & Festive Needs (label 6)
  - Kitchen & Dining (label 3)
- Confusion pour :
  - Label 5, majoritaire pour Baby care, par déduction label 0 équivalent à Home Furnishing.

TF-IDF

TSNE et k-means clustering



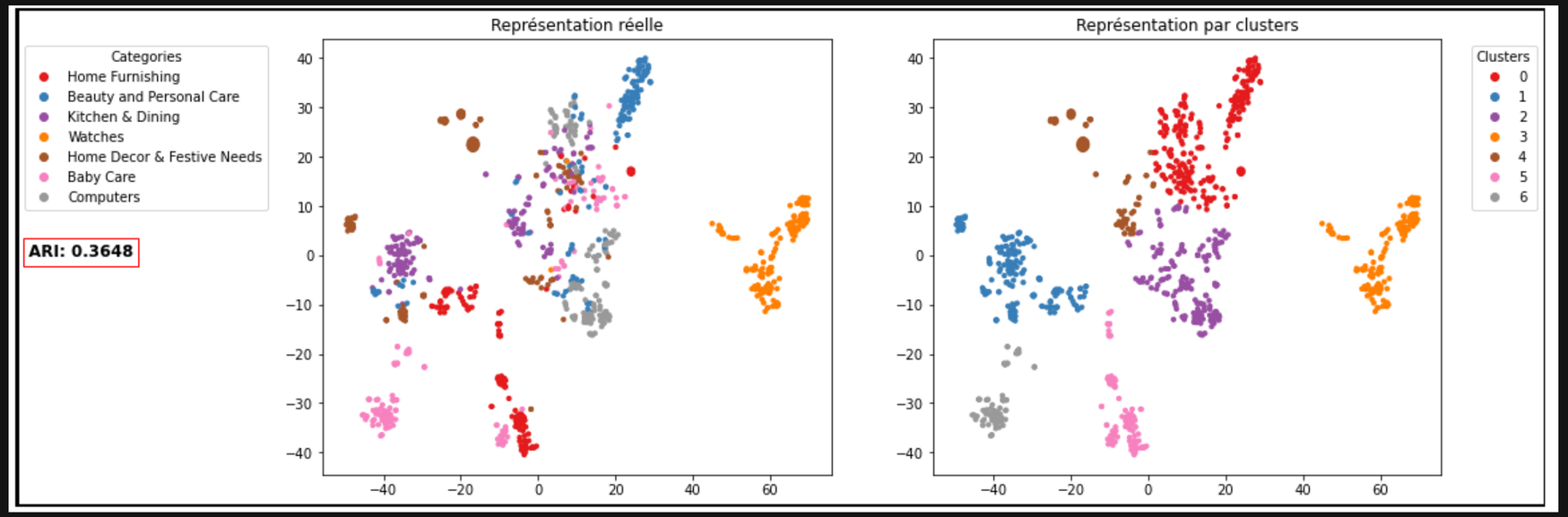


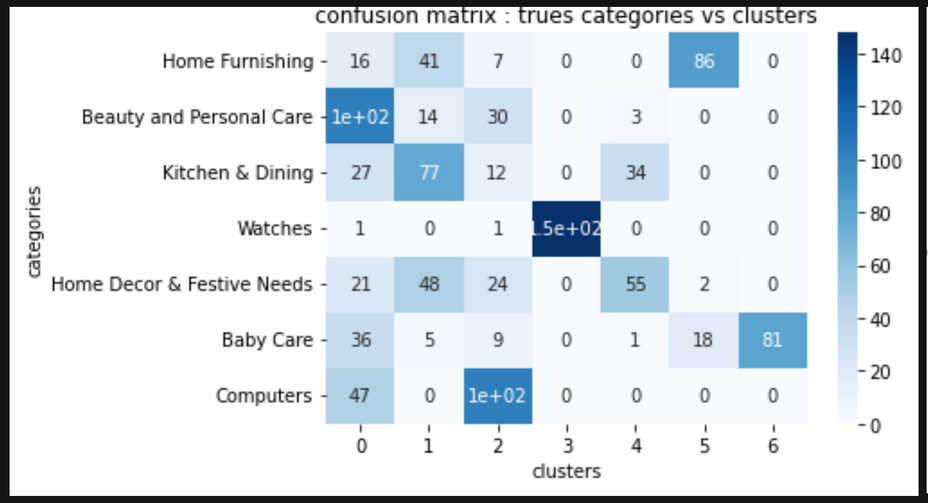


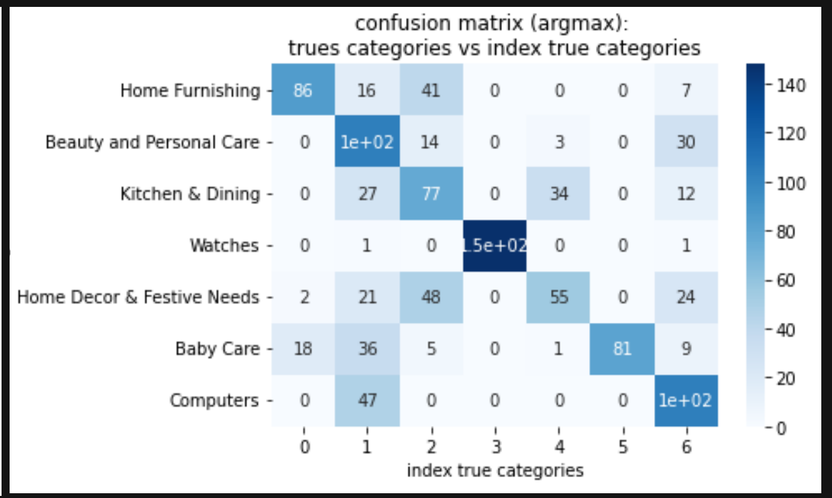
- ARI en hausse par rapport au CountVectorizer
- Attribution des clusters réussi, visible par la diagonale en arrangeant les valeurs de la matrice de confusion

Word2Vec

TSNE et k-means clustering



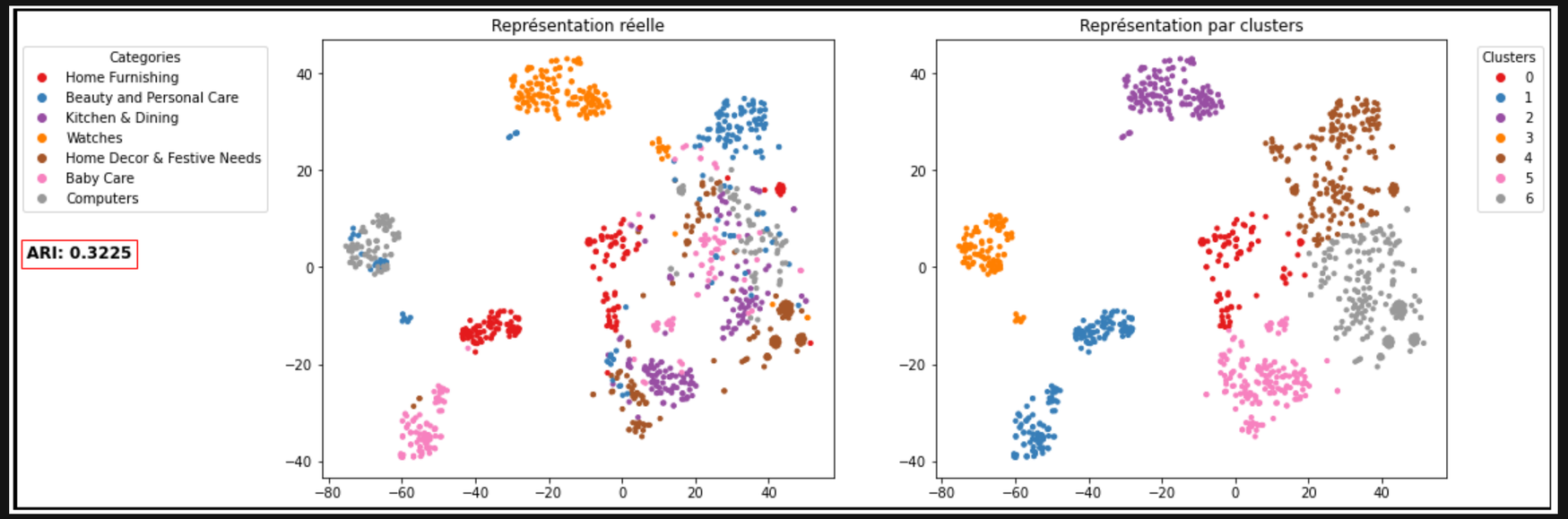


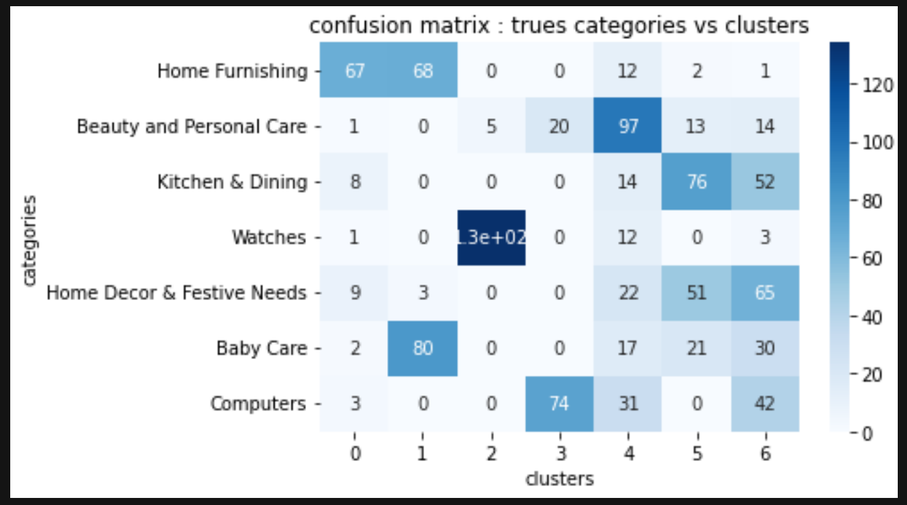


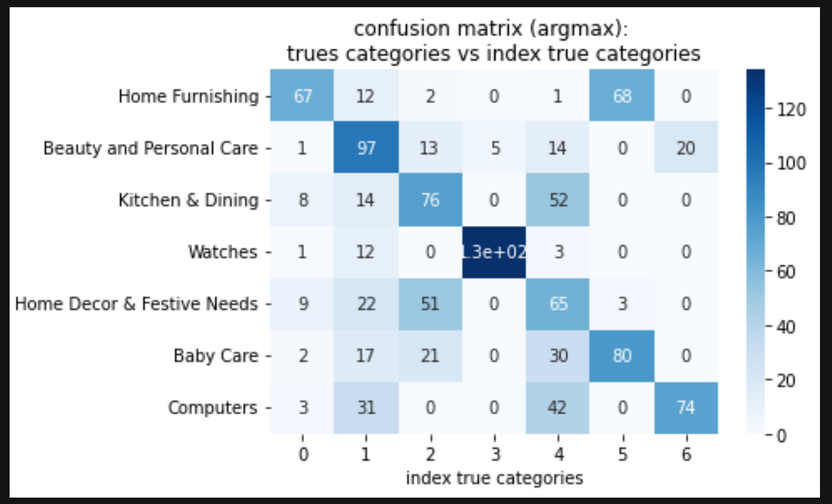
- Attribution des clusters réussi, visible par la diagonale en arrangeant les valeurs de la matrice de confusion

BERT

TSNE et k-means clustering



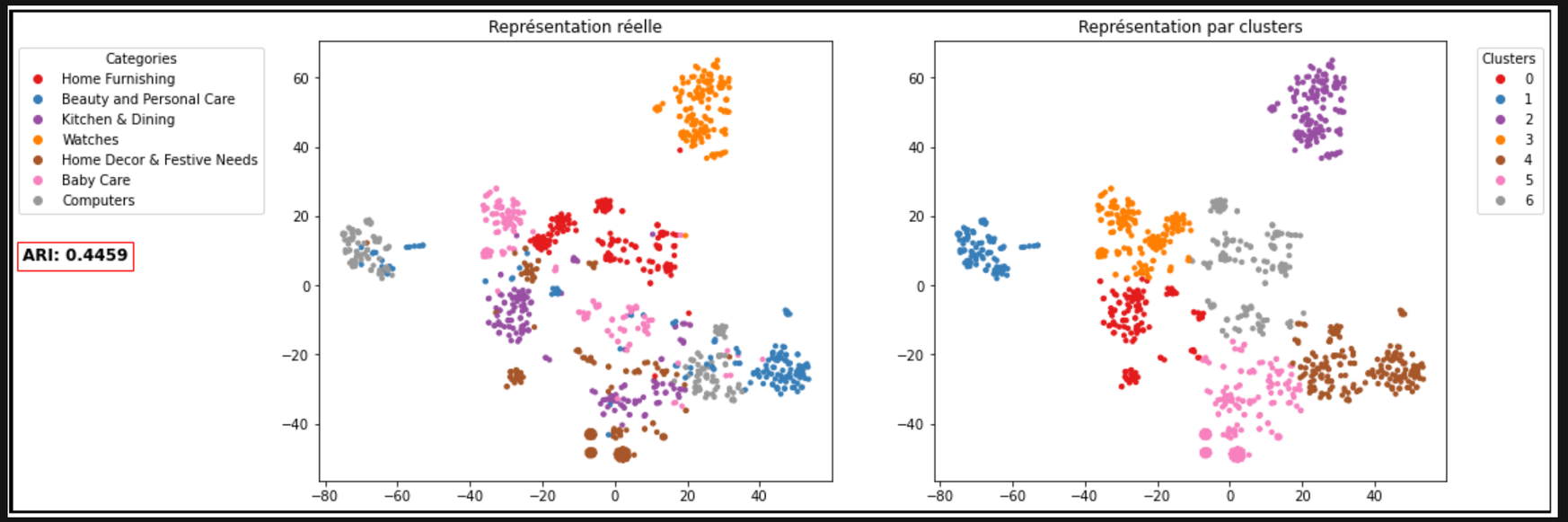


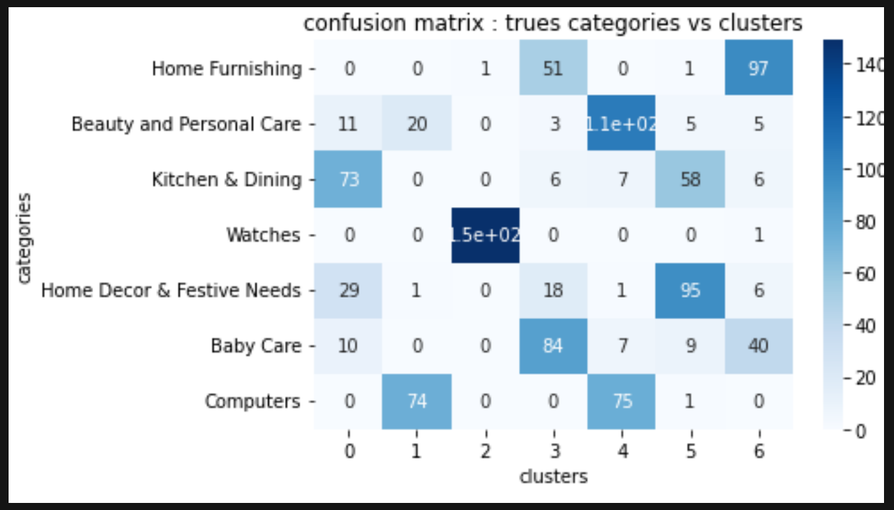


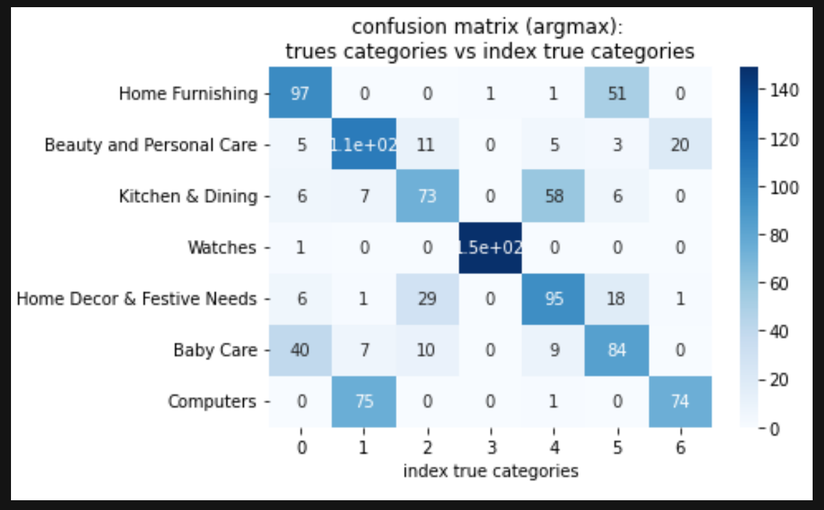
- Classification ambiguë pour Home Furnishing. La visualisation t-Sne semble indiquer deux amas distincts de points pour les carégories Home Furnishing et Babycare contrairement au clustering qui regroupe les points selon un unique cluster

USE

TSNE et k-means clustering





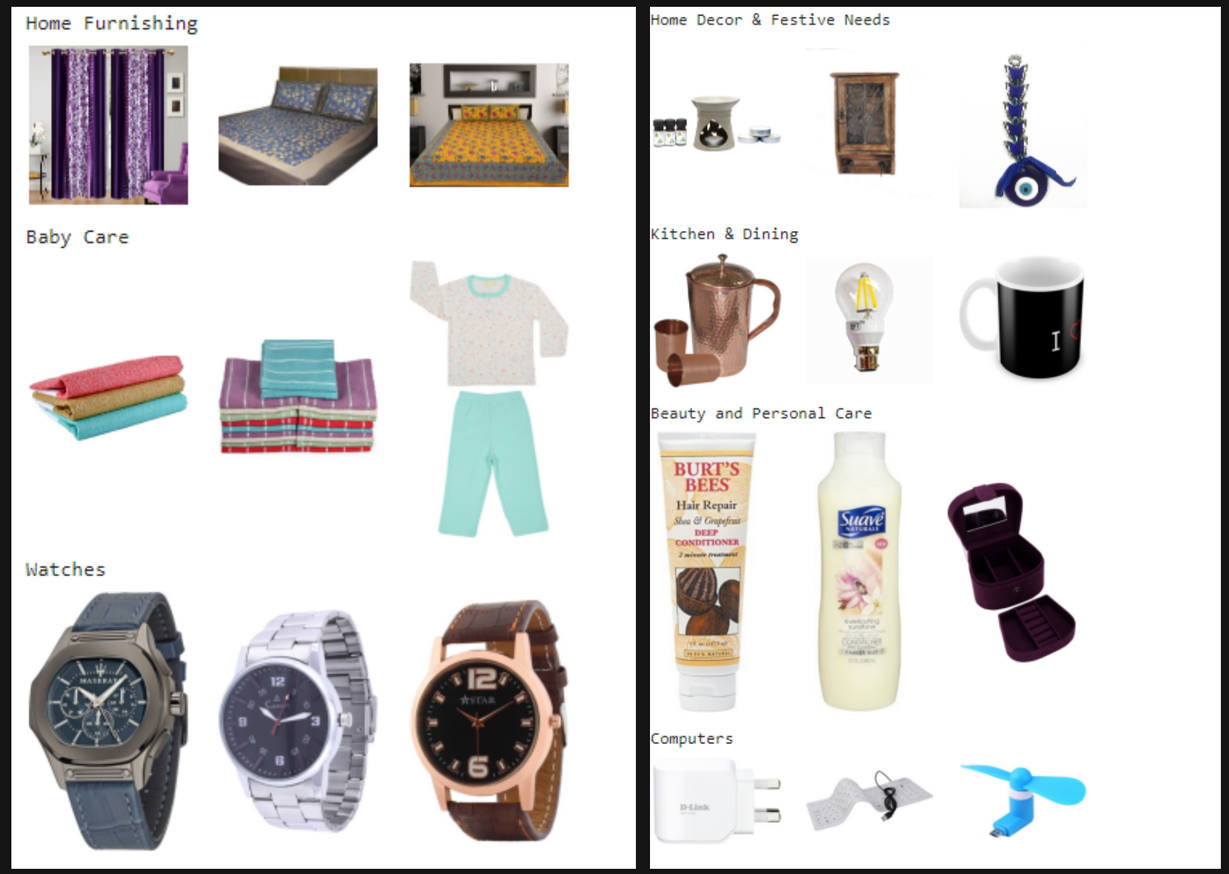


- Attribution des clusters réussi, visible par la diagonale en arrangeant les valeurs de la matrice de confusion. Moins d'erreur de classification cette fois-ci pour la catégorie Home Furnishing

3. Analyse d'images

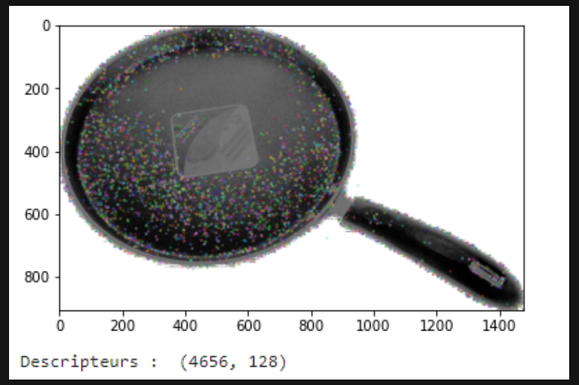
Extraction des features images

- SIFT : (Scale Invariant Feature Transform): permet d'identifier les éléments similaires entres différentes images
- Transfer Learning (VGG-16) : réseau de neurones convolutif pré-entrainé sur ImageNet



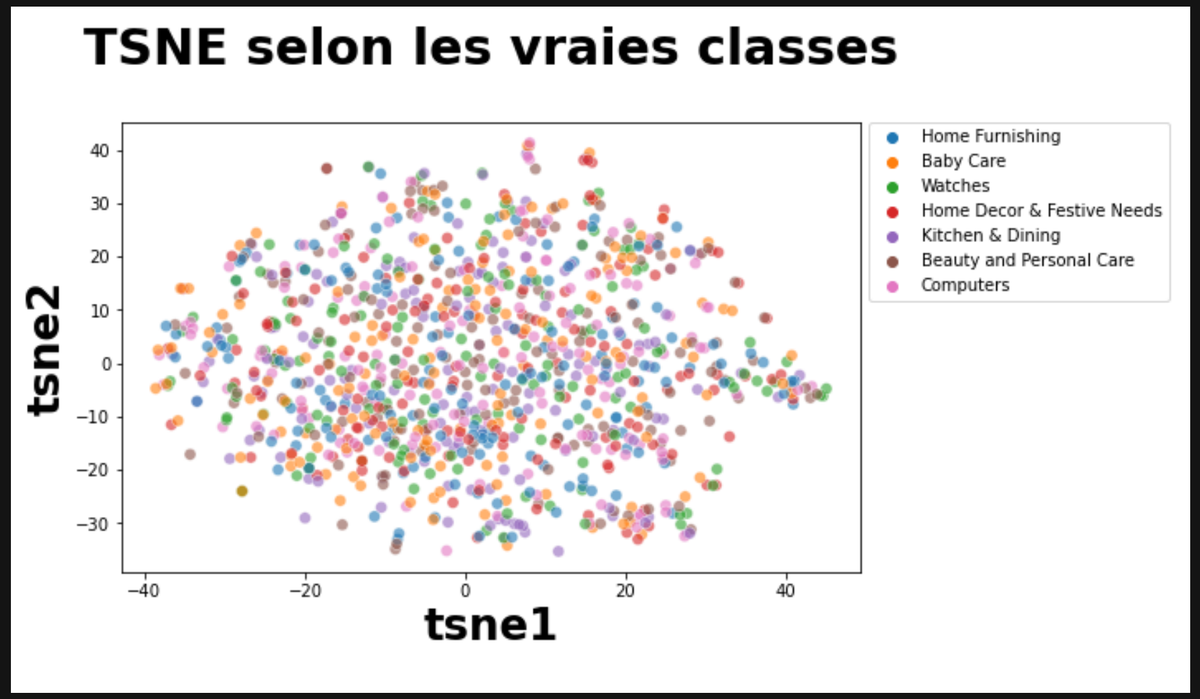
SIFT

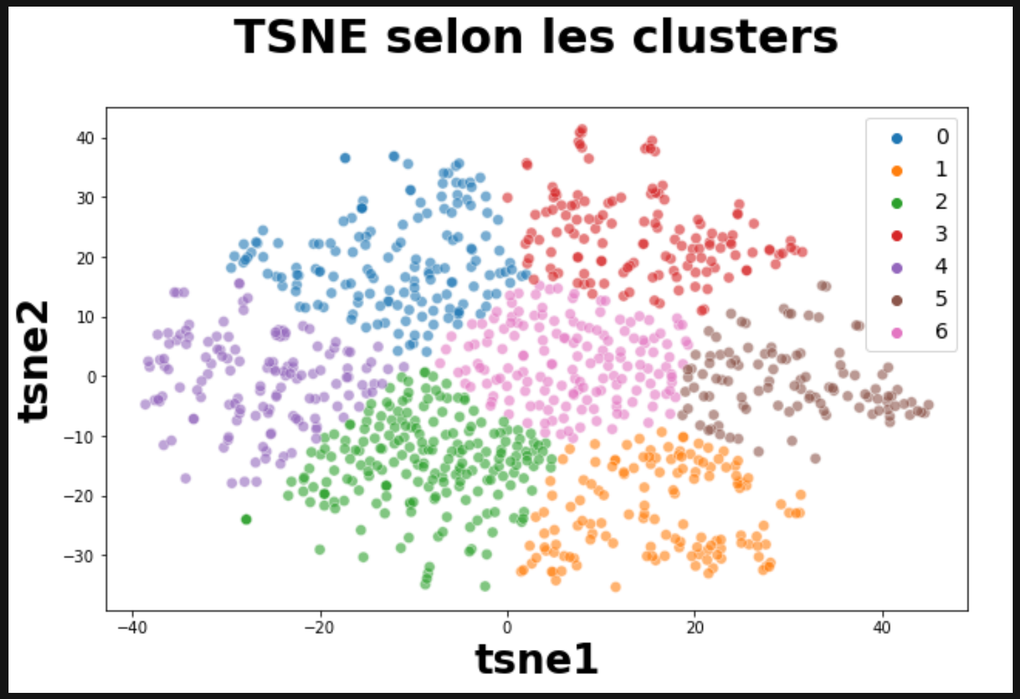
- Prétraitement : passage en niveau de gris et égalisation d'histogramme pour ajuster le contraste
- Récupération des descripteurs



- Clustering de l'ensemble des descripteurs et identification des centres
- Création de l'histogramme de l'image (Bag of Visual Words)
- PCA (95% de la variance expliquée)

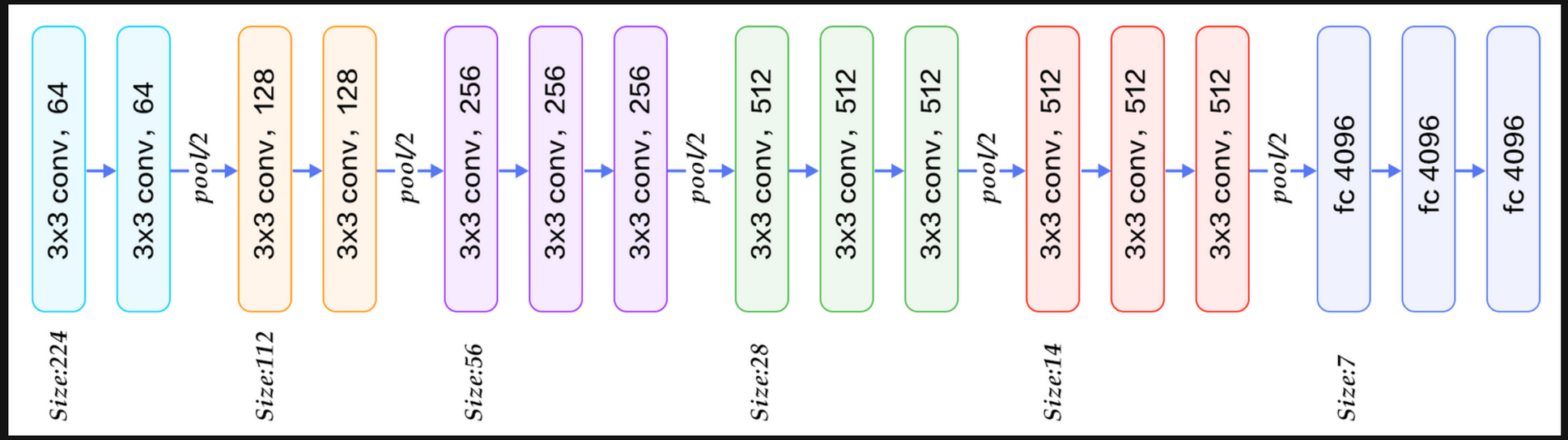
SIFT





- ARI très faible (0.002)
- Mauvais résultats

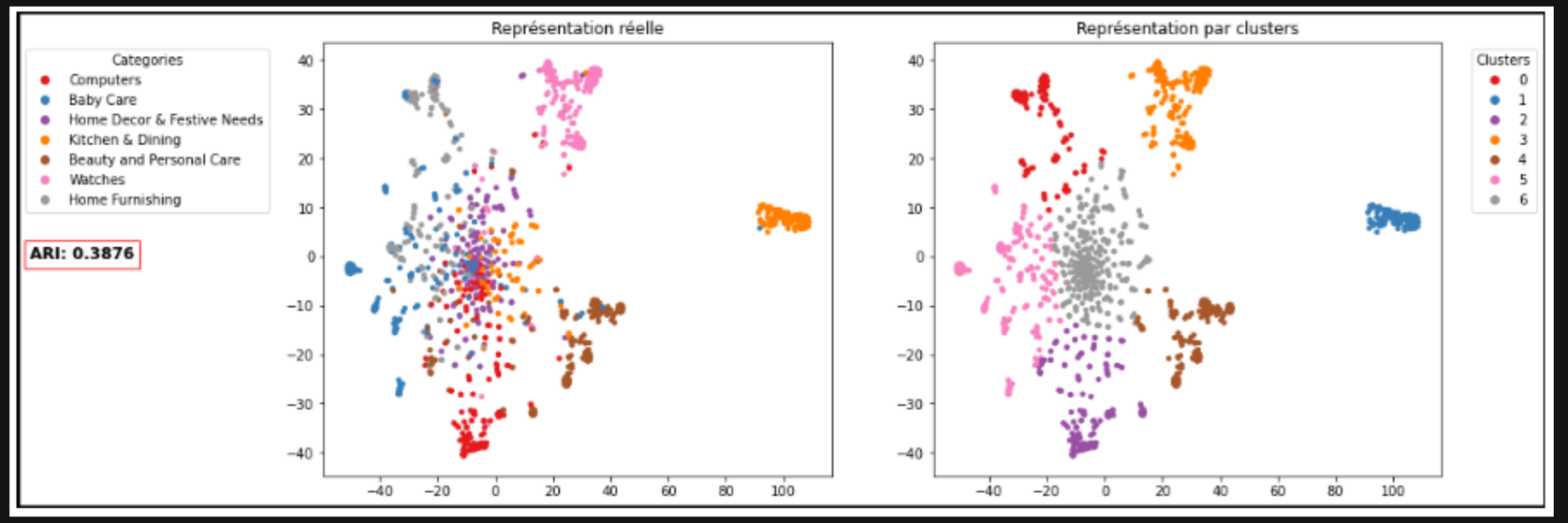
VGG16

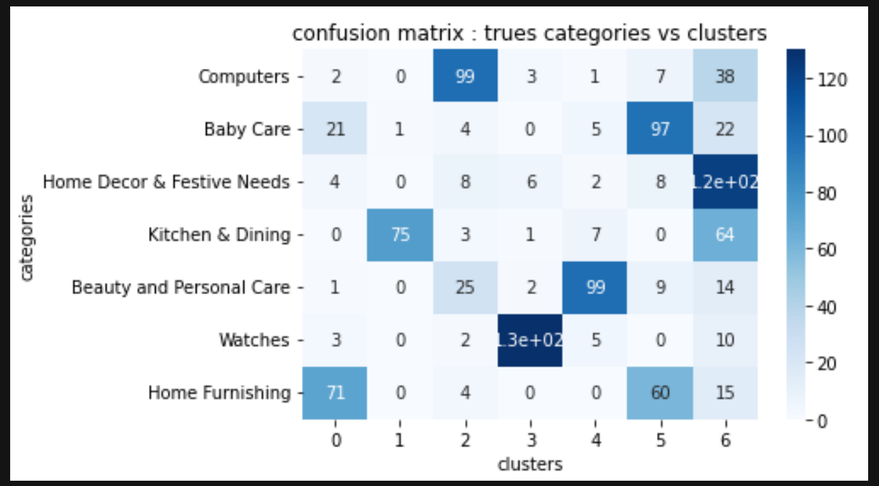


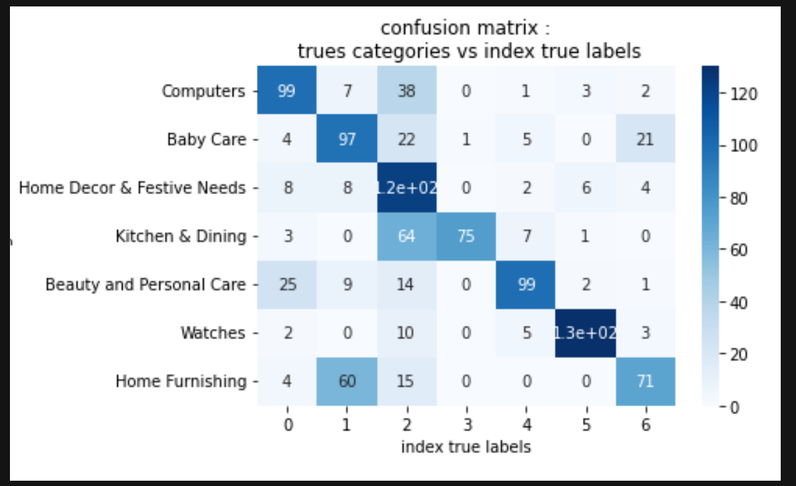
- VGG16 est un réseau de neurone convolutif
- VGG-16 est constitué de plusieurs couches, dont 13 couches de convolution et 3 fully-connected. Il doit donc apprendre les poids de 16 couches.
- Il prend en entrée une image en couleurs de taille 224 × 224 px.
- VGG-16 est pré-entrainé sur ImageNet (base de données de plus de 14 millions d'images labellisées réparties dans plus de 1000 classes)

VGG16

TSNE et k-means clustering







- Confusion pour la catégorie Kitchen & Dining

4. Conclusion et recommandation

Conclusion

- Analyse des données textuelles et visuelles
- Extraction de features en utilisant différentes techniques:
  - textes : BoW, Tf-Idf, Word2Vec, Bert, Use
  - images : SIFT et Transfer learning
- Réduction de dimension
  - PCA et T-SNE (2 dimensions)
- Clustering
  - Kmeans, 7 clusters

Faisabilité

Features	ARI
TF-IDF	0.50
USE	0.45
CountVectorizer	0.39
CNN (vgg16)	0.39
W2Vec	0.35
BERT	0.32
SIFT	0.002

- Le TF-IDF permet une meilleur catégorisation des produits avec un score ARI le plus élevé.
- La classification avec les données images est améliorée en utilisant un algorithme de type CNN

On valide la faisabilité de la mise en œuvre de moteur de classification automatique des produits

Recommandations

- Enrichir la base de données produits
- Combiner un modèle d'image et texte (concaténer les vecteurs de features)