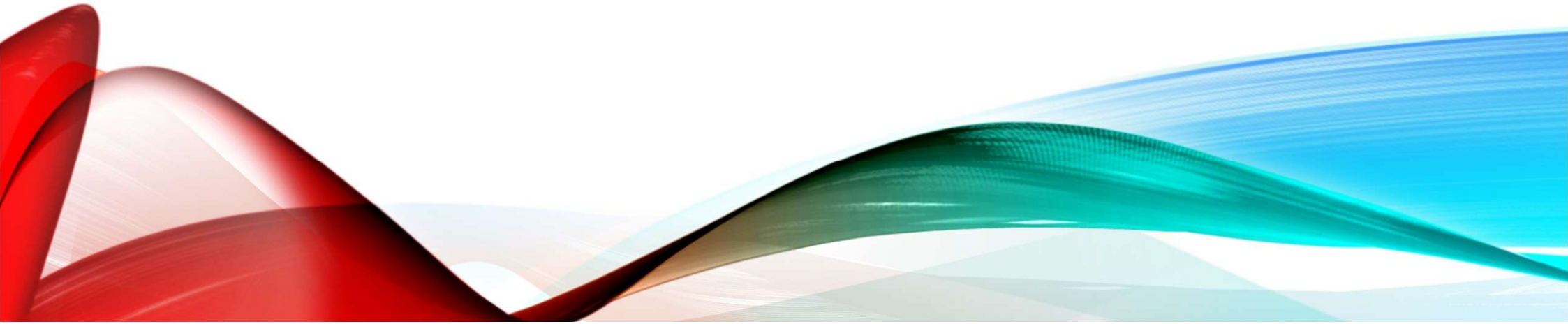
CLASSIFIER AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

SOMMAIRE

- Problématique
- Jeu de données
- Traitement des données textuelles
- Traitement des données visuelles
- Regroupement des deux types de données
- Conclusion

RAPPEL DE LA PROBLÉMATIQUE



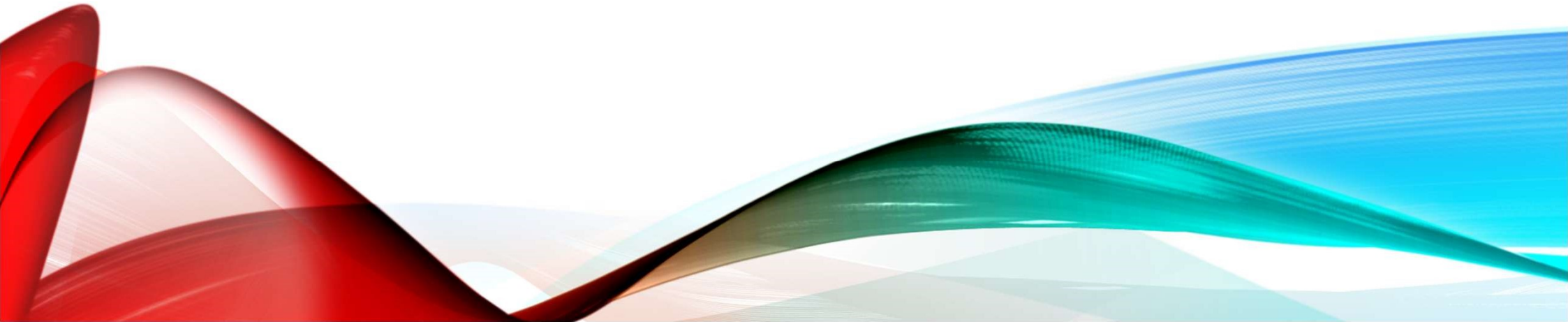
RAPPEL DE LA PROBLÉMATIQUE

Etude de la faisabilité d'un moteur de classification basé sur des photos et descriptions de produits

Détecter les features à utiliser

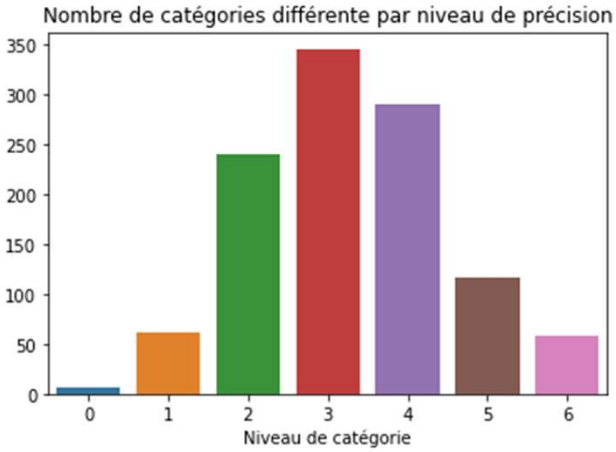
Automatiser l'attribution de la catégorie des articles

PRÉSENTATION DU JEU DE DONNÉES

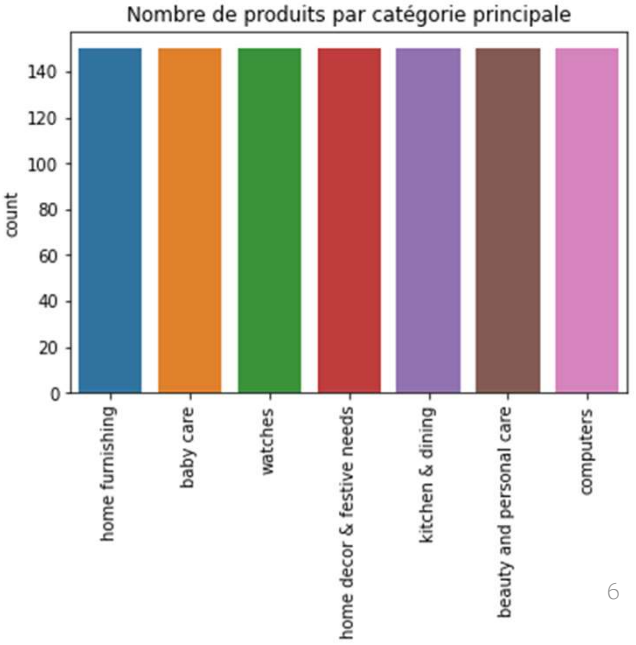


PRÉSENTATION DU JEU DE DONNÉES

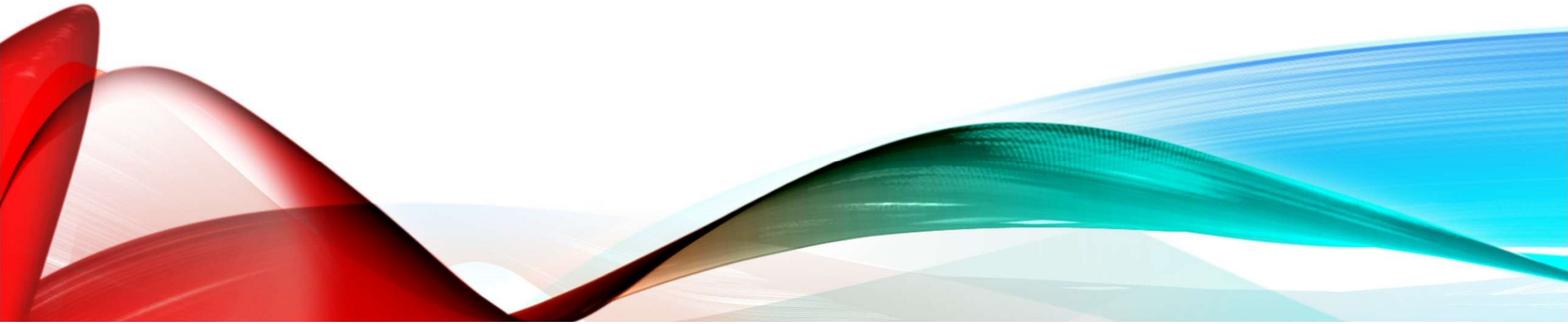
- 1050 produits → Nom, prix, catégories, image, description …
- Arbre des catégories → "home furnishing >> curtains & accessories >> curtains >> elegance polyester multicolor abstract eyelet do..."



7 catégories principales
62 catégories niveau 1
240 catégories niveau 2
345 catégories niveau 3
290 catégories niveau 4
117 catégories niveau 5
58 catégories niveau 6



TRAITEMENT DES DONNÉES TEXTUELLES



TRAITEMENT DES DONNÉES TEXTUELLES

- Chaque objet a une description

'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 curta in enhances the look of the interiors.this curtain is made from 100% high quality polyester fabric.it features an eyelet style stitch with metal ring.it makes the room environment romantic and loving.this curtain is ant- wrinkle and anti shrinkage and have elegant apparance.give your home a bright and modernistic appeal with these designs. the surreal attention is sure to steal hearts. these contemporary ey elet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,specifications of elegance polyester multicolor abstract eyelet door curtain (213 cm in height, pack of 2) general brand elegance designed for door type eyelet model name abstract polyester door curtain set of 2 model id duster25 color multicolor dimensions length 213 cm in the box number of contents in sales package pack of 2 sales package 2 curtains body & design material polyester'

- Transformation en tokens

key, features, of, elegance, polyester, multicolor, abstract, …, sales, package, 2, curtains, body, design, material, polyester

- Utilisation des stopwords : enlever les mots les plus courants

i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself…

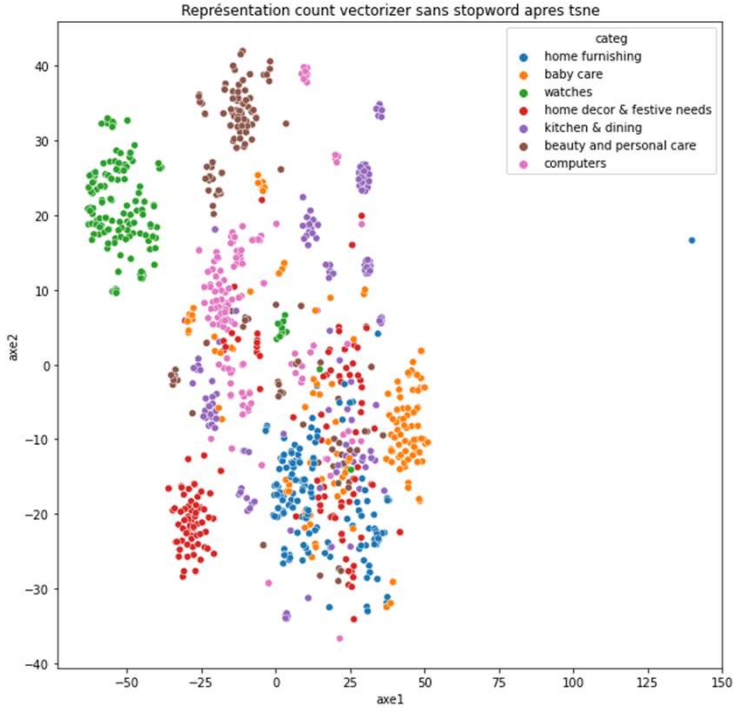
- Utilisation d'un compteur

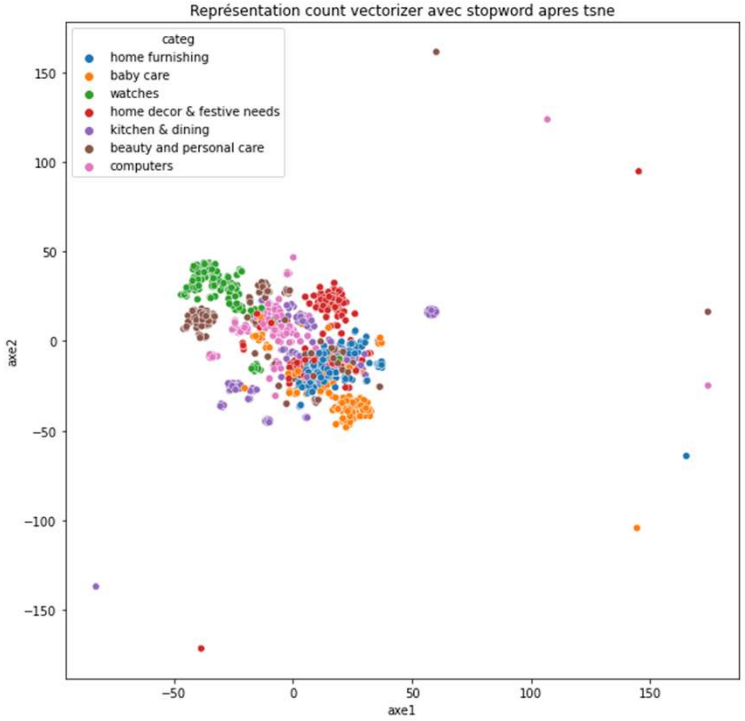
100	2	213	899	abstract	amount	ant	anti	apart	apparance	...	surreal	thing	type	valance	want	welcome	whole	wish	world	wrinkle
1	5	3	1	4	1	1	1	1	1	...	1	1	1	1	1	1	1	1	1	1

TRAITEMENT DES DONNÉES TEXTUELLES

➢ Représentation graphique 2D

Réduction de dimension avec TSNE : garder la proximité locale et perte d'informations avec PCA



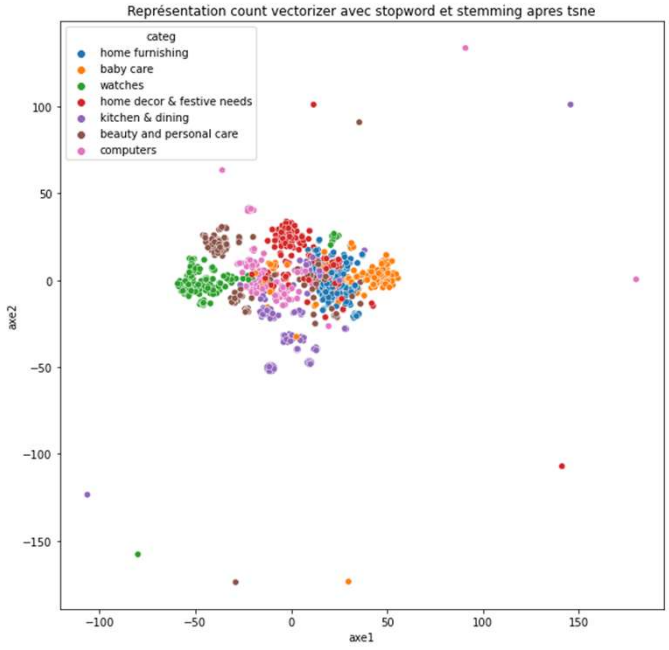


TRAITEMENT DES DONNÉES TEXTUELLES

➢ Nettoyage supplémentaire :

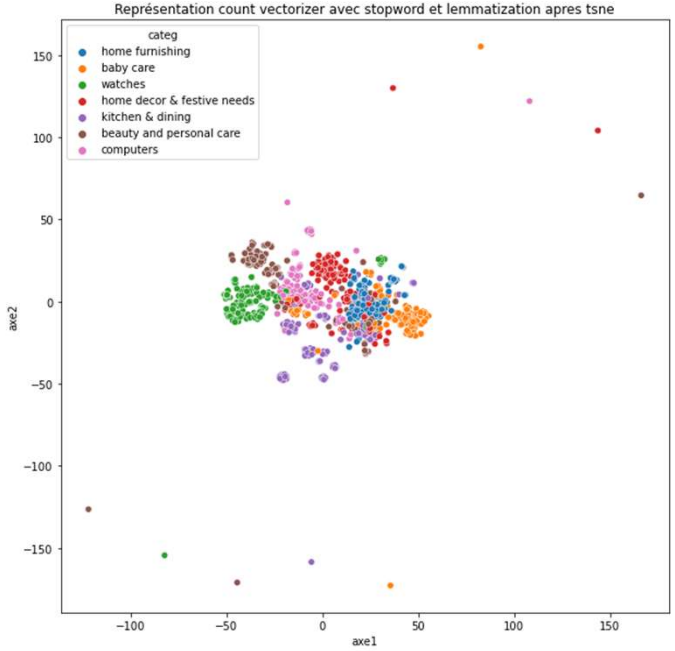
Stemming (suppression de suffixes)

key, featur, eleg, polyest, multicolor, abstract, sale, packag, 2, curtain, bodi, design, materi, polyest



Lemmatization (regroupement par champ lexical)

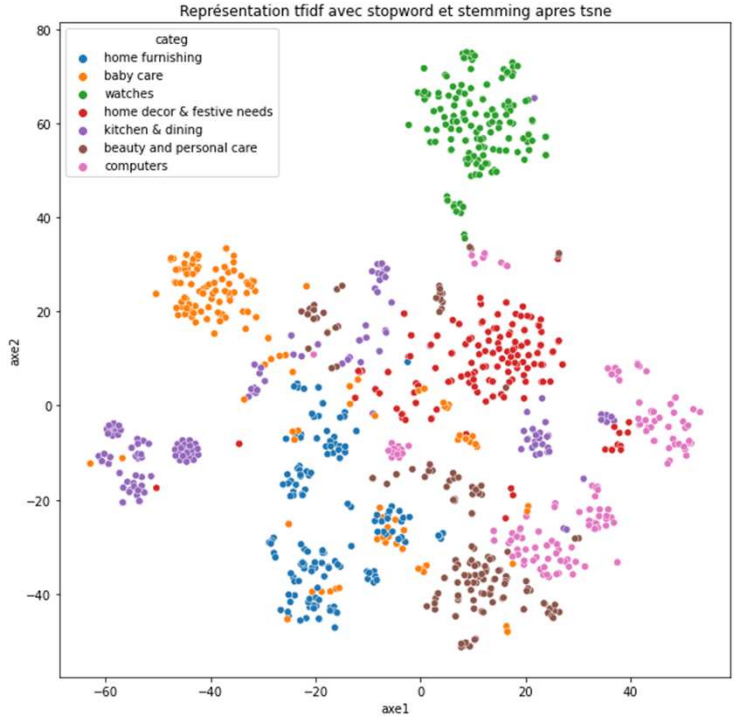
key, feature, elegance, polyester, multicolor, abstract, sale, package, 2, curtain, body, design, material, polyester

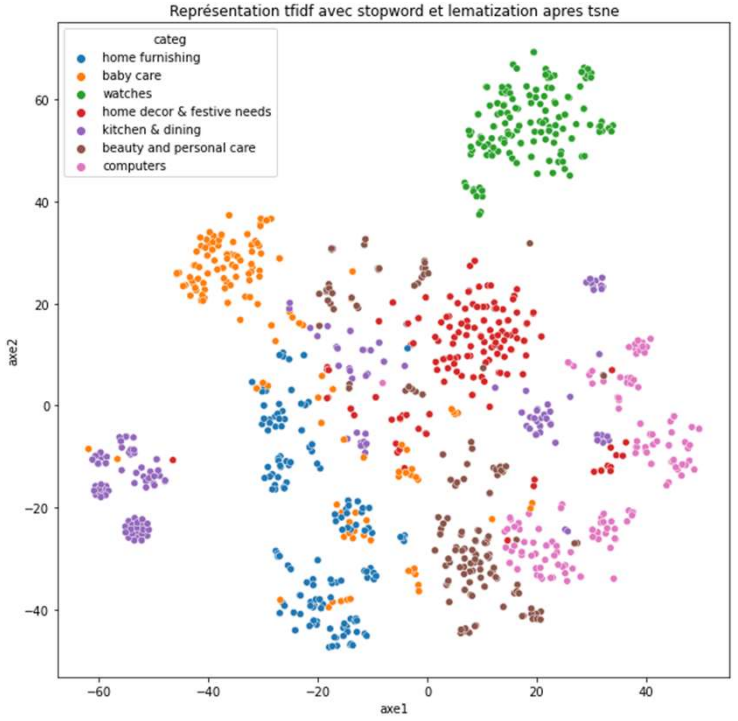


TRAITEMENT DES DONNÉES TEXTUELLES

➢ Nettoyage supplémentaire :

Tf_idf : calcul des fréquences d'apparition du mot dans la description et la fréquence inverse d'apparition dans toutes les descriptions

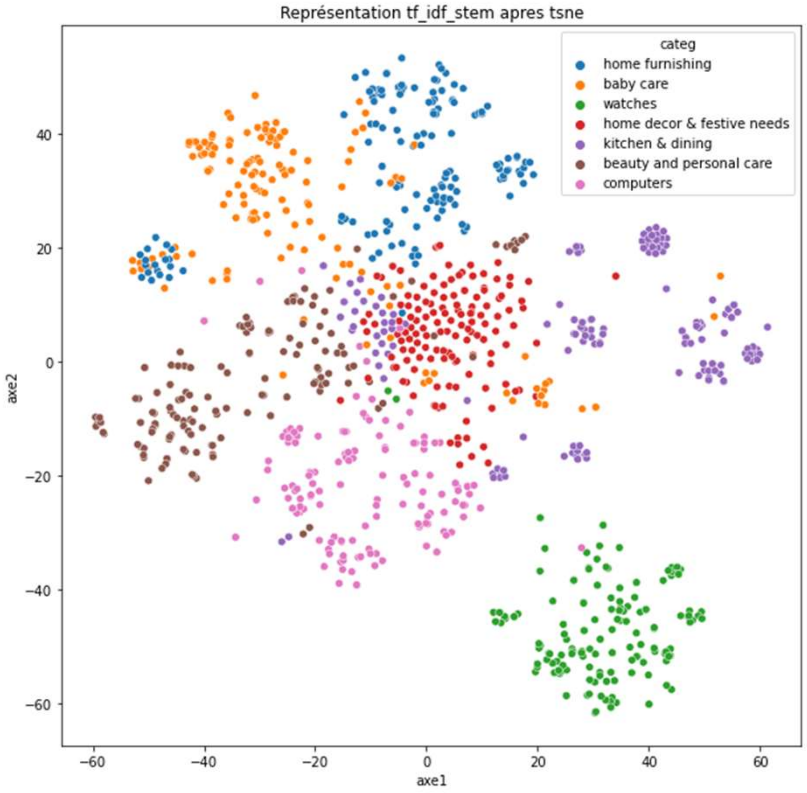




TRAITEMENT DES DONNÉES TEXTUELLES

➢ Optimisation tf-idf :

7 catégories principales => Max_idf à 1/7 ? Termes spécifiques aux catégories

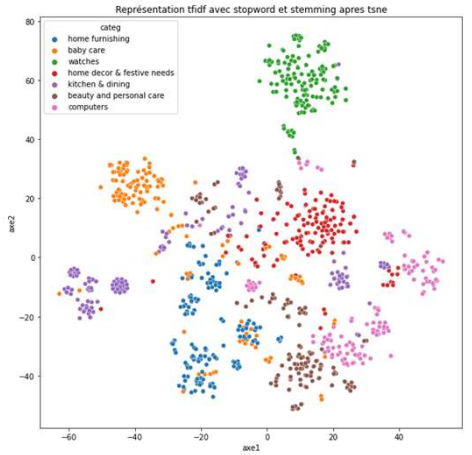


Optimisation par clustering pour différentes valeurs des hyperparamètres du tf_idf

Kmeans à 7 clusters sur résultat du TSNE

➡ max_idf = 0,3
min_idf = 2

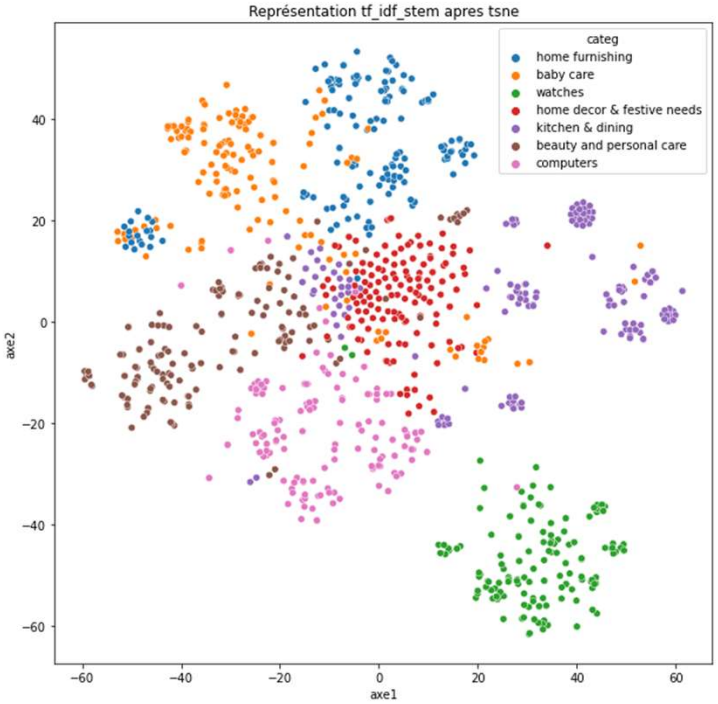
Calcul de l'ARI : maximum de 0,64

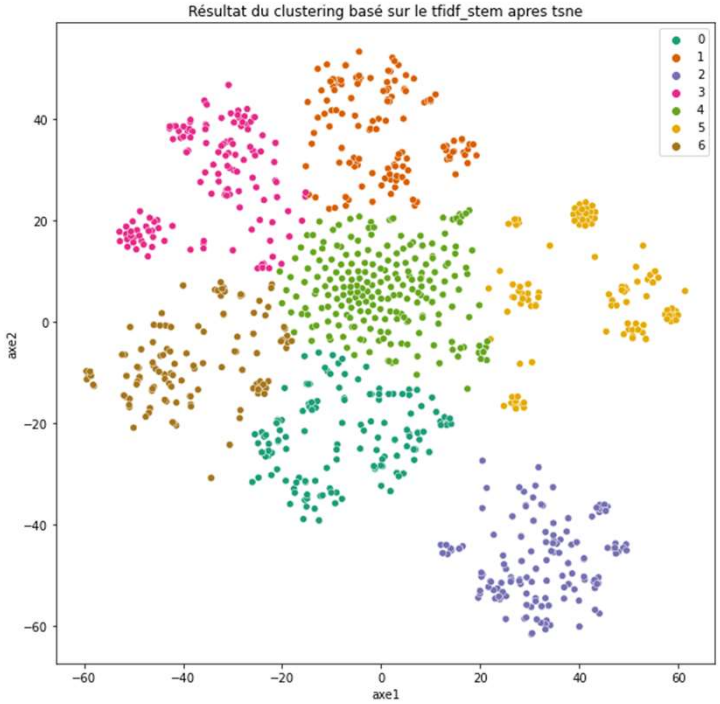


TRAITEMENT DES DONNÉES TEXTUELLES

➢ Optimisation tf-idf :

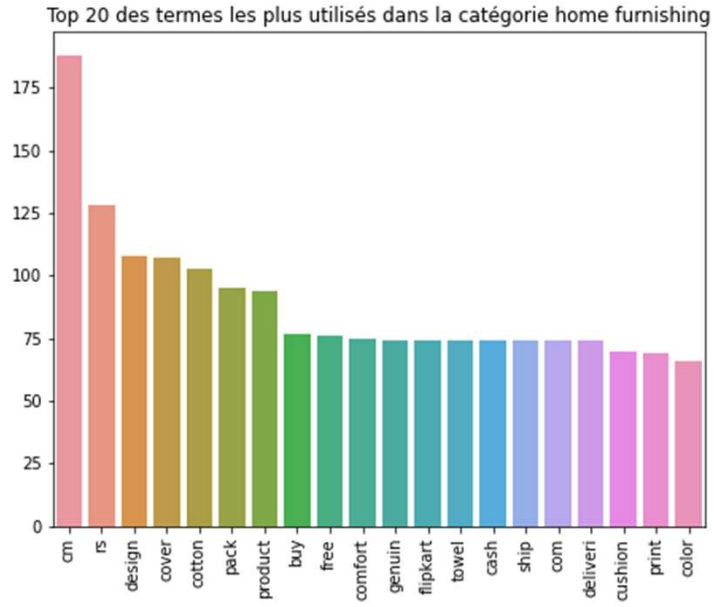
Comparaison avec le clustering du KMeans

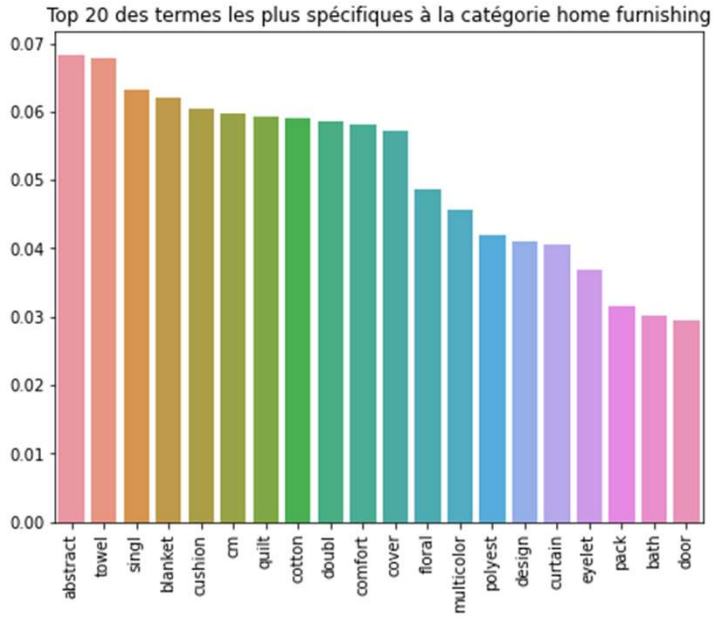




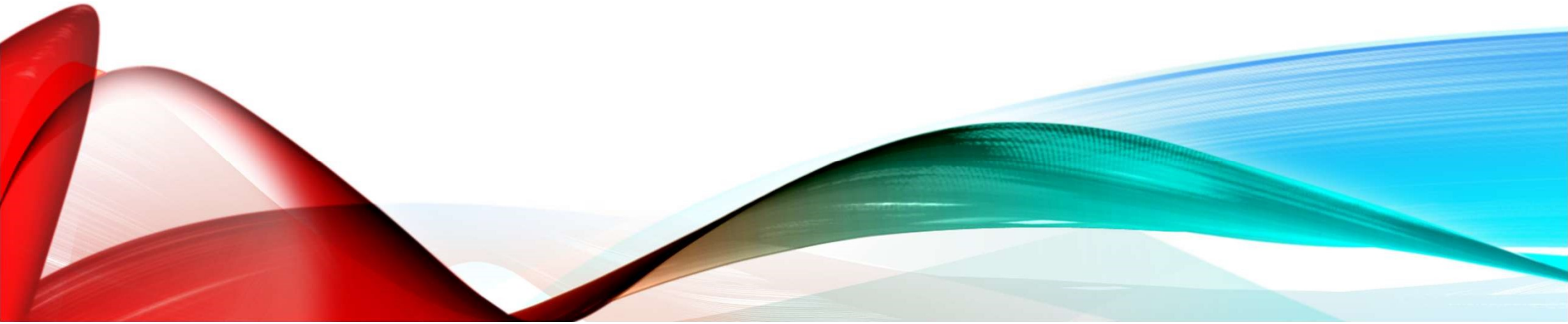
TRAITEMENT DES DONNÉES TEXTUELLES

- Détail des termes les plus présents:





TRAITEMENT DES DONNÉES DE TYPE IMAGE



TRAITEMENT DES DONNÉES DE TYPE IMAGE

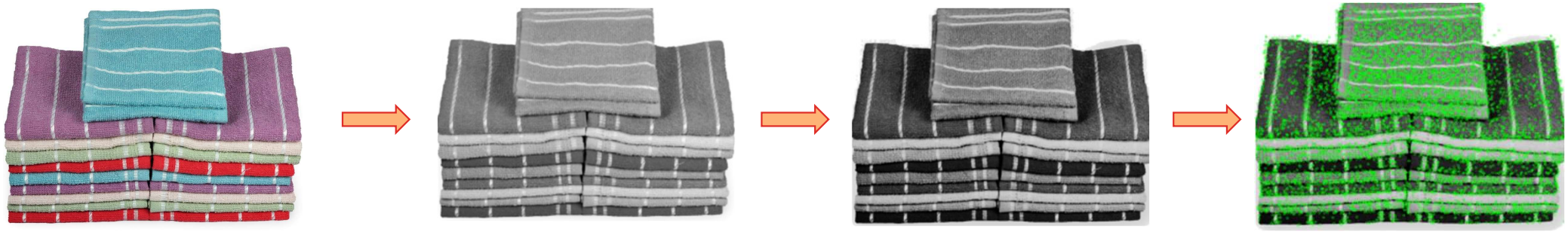
- Traitement par l'algorithme SIFT :

Ouverture de l'image

Passage en niveaux de gris

Egalisation de l'histogramme

Détection des points d'intérêts

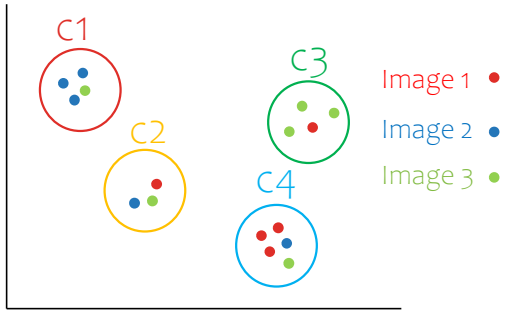


TRAITEMENT DES DONNÉES DE TYPE IMAGE

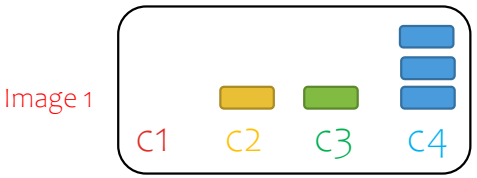
- Traitement par l'algorithme SIFT :

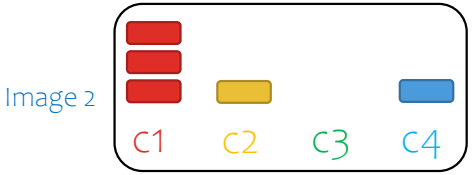
  - Liste des descripteurs des points d'intérêt
    9051349 descripteurs au total

  - Regroupement des descripteurs par clustering
    Utilisation de KMeans avec 3000 clusters
    Chaque cluster regroupe des 'visual words' (descripteurs) proches



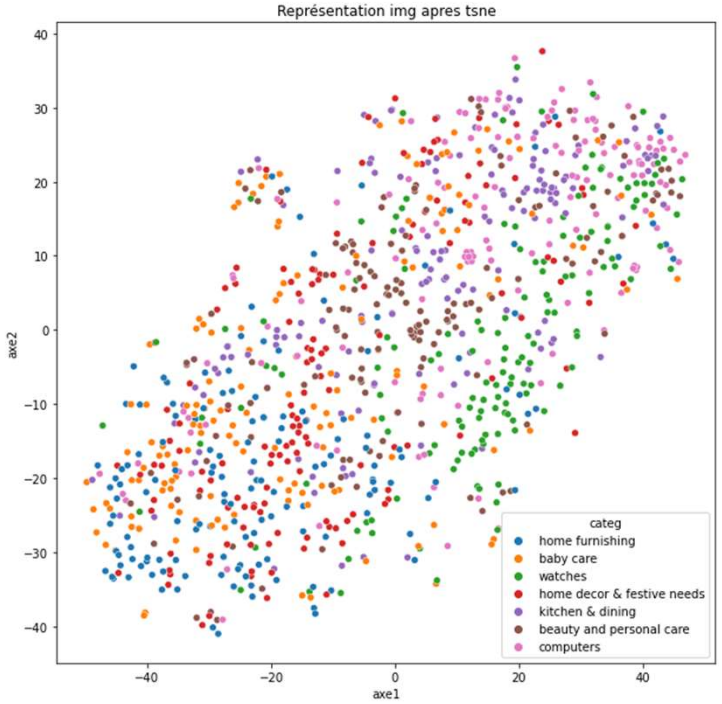
  - Construction d'un histogramme par image basé sur le nombre de fois où apparaissent les clusters dans ses descripteurs

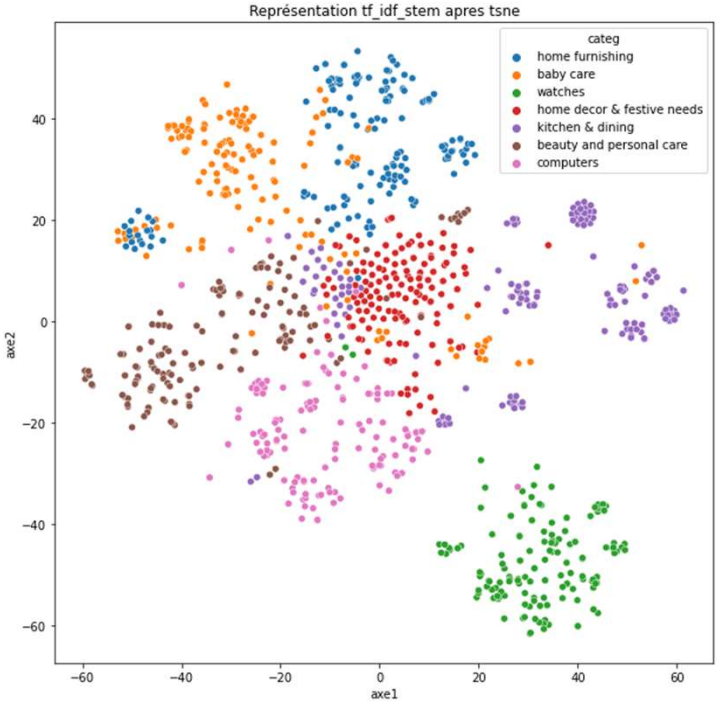




TRAITEMENT DES DONNÉES DE TYPE IMAGE

- Traitement par l'algorithme SIFT :

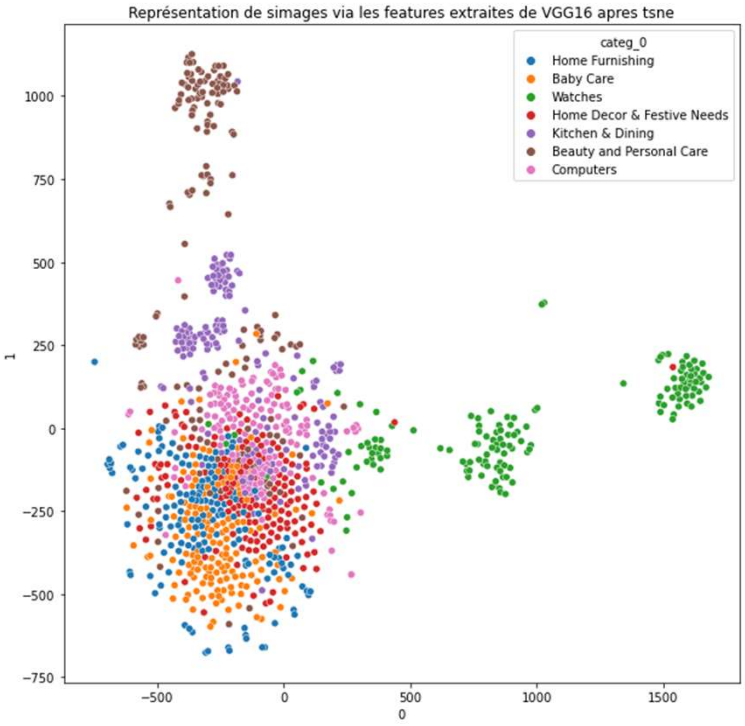




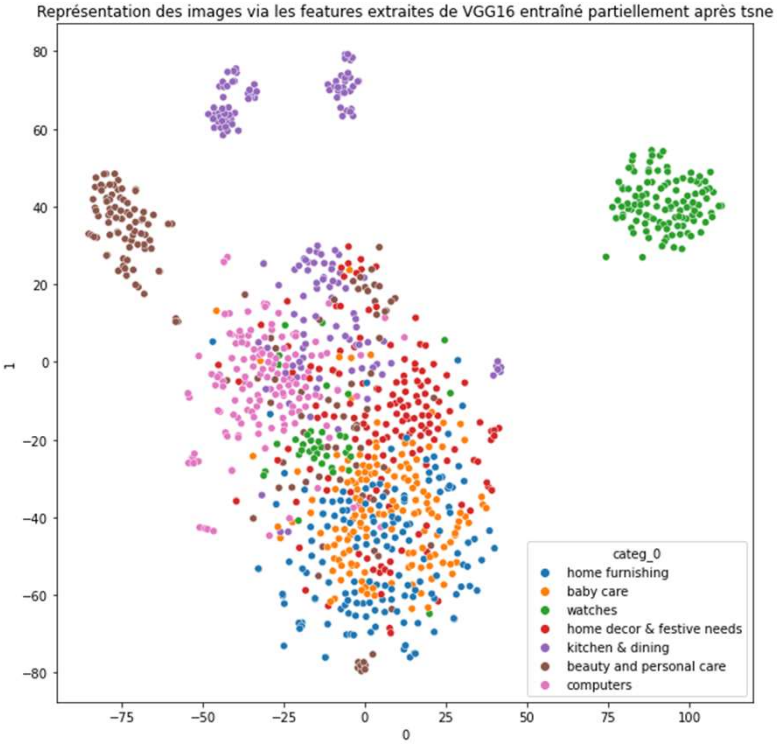
TRAITEMENT DES DONNÉES DE TYPE IMAGE

- Transfert Learning : Utilisation d'un CNN (convolution neural network) type vgg16

Récupération des features directement



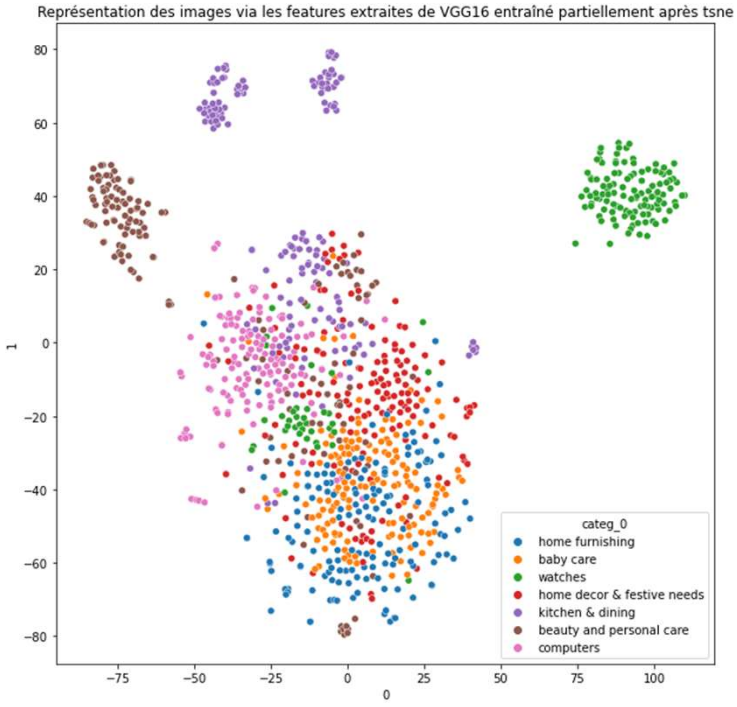
Récupération des features après entraînement partiel

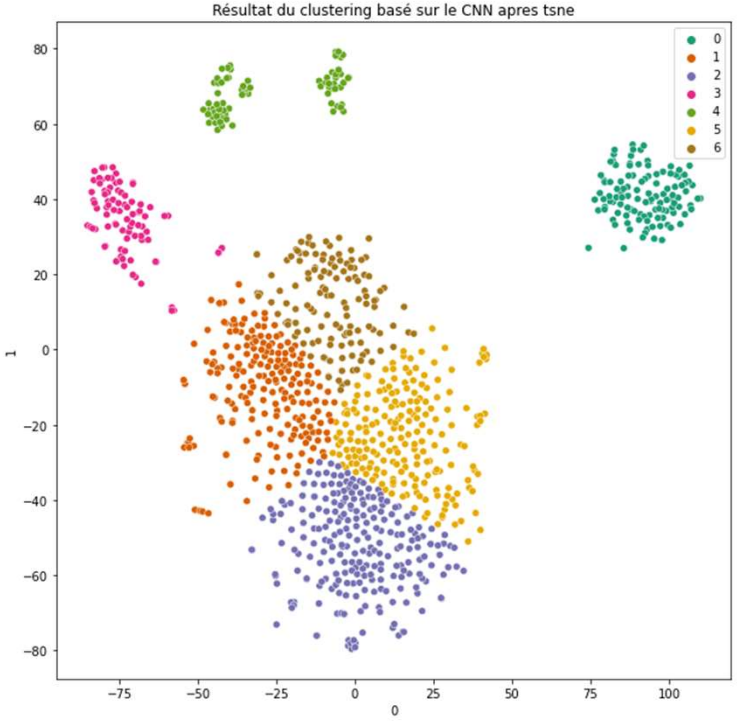


TRAITEMENT DES DONNÉES DE TYPE IMAGE

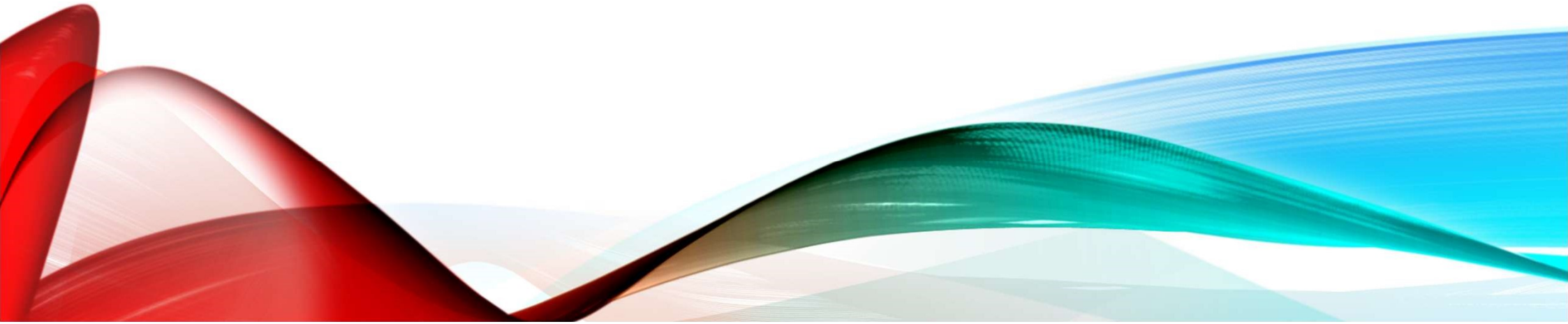
- Transfert Learning : Utilisation d'un CNN (convolution neural network) type vgg16

Comparaison avec le clustering ⟹ ARI = 0,34





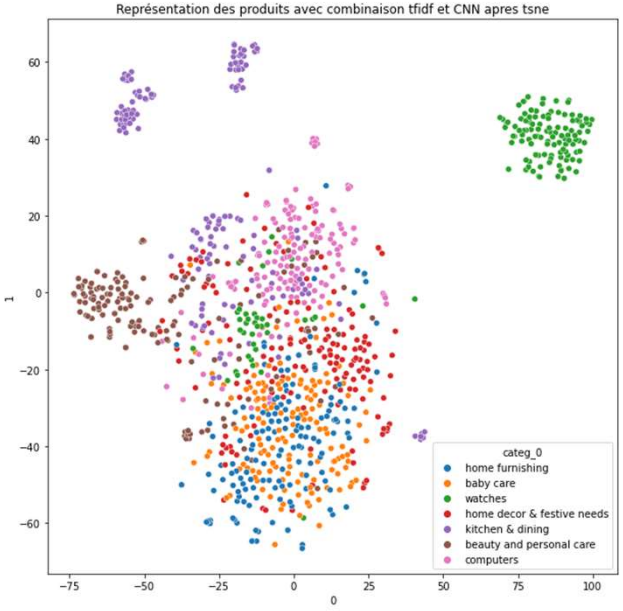
REGROUPEMENT DES DEUX TYPES DE DONNÉES

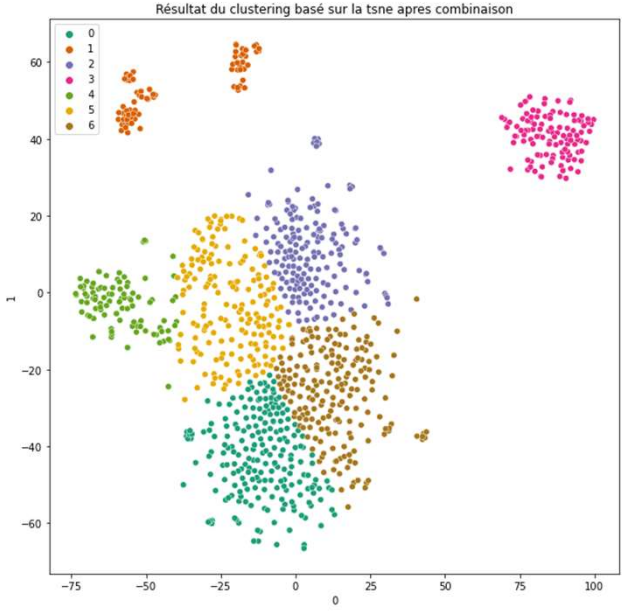


REGROUPEMENT DES DEUX TYPES DE DONNÉES

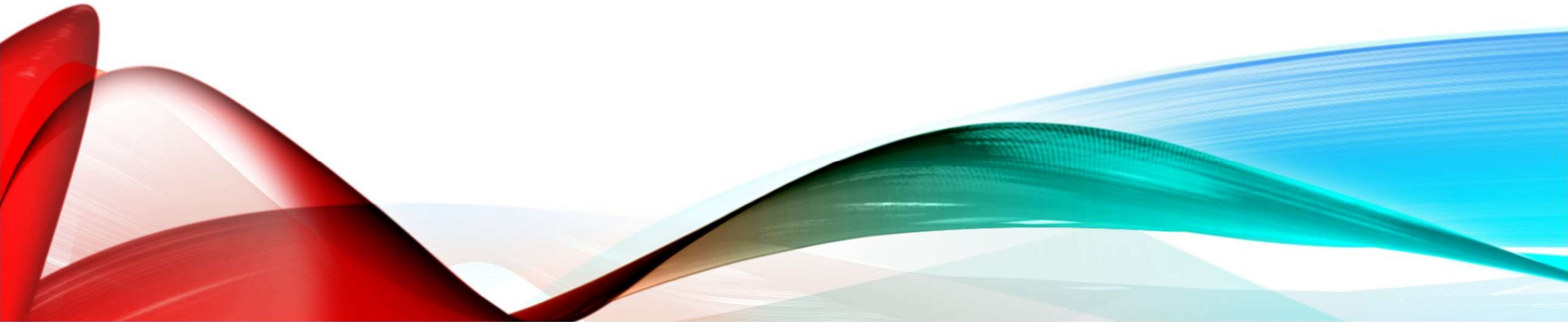
- Utilisation tf_idf et transfert learning :

Proche du resultat du CNN (seulement 10% de features sont dus au texte)
ARI = 0,34





CONCLUSION



CONCLUSION

- Meilleur résultat avec l'utilisation seule de la description (ARI de 0,64 vs 0,34)
  - L'apport des images apporte du 'bruit' et modifie sensiblement les features
  - Algorithme SIFT avec les plus mauvais résultats
  - L'utilisation d'un CNN permet d'améliorer le traitement des images
    - Utilisation de VGG16 ⟹ Tester d'autres modèles
- Modèle de classification (KNN) sans optimisation :
  - Tf_idf seul : accuracy de 0,84
  - Modèle combiné : accuracy de 0,64
- Recommandations :
  - Se concentrer sur les descriptions
  - Utilisation de modèles plus avancés (ex : RNN avec BERT en transfert learning)
  - Si utilisation de photos : se concentrer sur le transfert learning et gérer le souci des dimensions (réduction de dimension des features images, vecteurs plus dense…) / tester d'autres CNN