CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

PROJET N°6
PARCOURS « DATA SCIENTIST »

ETUDIANT : SAHEL TAHERIAN
MENTOR : YANNICK SERGE
EVALUATEUR :

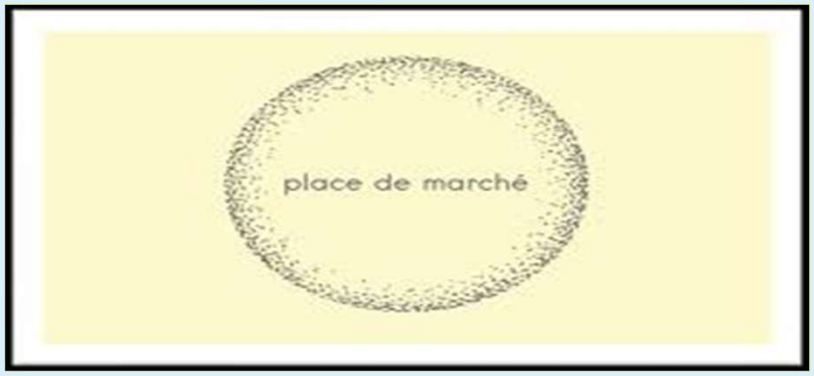
SOUTENANCE DE PROJET
26 NOVEMBRE 2021

sommaire


Présentation de la problématique
Présentation Des Données
Données Textuelles
Données Visuelles
Classification & Clustering
Conclusion

Problématique

Classification automatique de produits



- Marketplace e-commerce proposant des produits à la vente
- Les données des produits : des descriptions textuelles et des images
- Attribution manuelle des catégories : fastidieuse et peu fiable
- Le volume des articles est très petit
- Automatiser cette tâche
- Prétraiter les description des produits et leurs images dans le but de réaliser un clustering

Mission du projet

Etudier la faisabilité d'une automatisation de la classification des produits à partir de leur nom, description, et d'une photo

Contraintes

mettre en œuvre un algorithme de type SIFT / ORB / SURF

PRESENTATION DES DONNEES

- **DECOUVERTE DES DONNEES**
- **DONNÉES MANQUANTES**
- **ANALYSE DES CATEGORIES**

DECOUVERTE DES DONNEES

DATASET

- Uniq_id
- Crawl_timestamp
- Product_url
- Product_name
- Product_category_tree
- Pid Retail_price
- Discounted_price
- Image
- Is_FK_Advantage_product
- Description
- Product_rating
- Overall_rating
- Brand
- Product_specifications

1050 lignes
15 colonnes

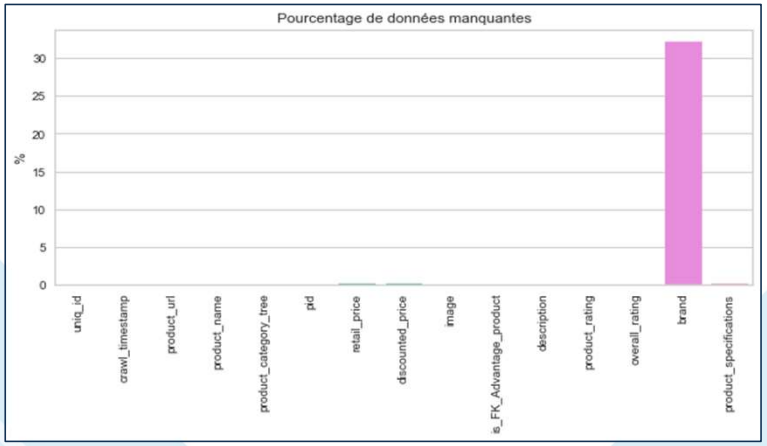
Exemple du premier article de notre jeu de données

'["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Y..."]'

'Specifications of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton bath towel Ideal For Men, Women, Boys, Girls Model ID asvtwl322 Color Red, Yellow, Blue Size Mediam Dimensions Length 30 inch Width 60 inch In the Box Number of Contents in Sales Package 3 Sales Package 3 Bath Towel'

Données manquantes

- ❑ **Suppression des colonnes inutiles**
- ❑ **Imputation des valeurs manquantes de "brand" par " "**



ANALYSE DES CATEGORIES

product_category_tree → **Arborescence complète d'un article**

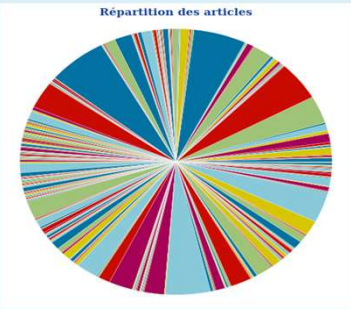
'["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> ..."]'

Sous_cat_1

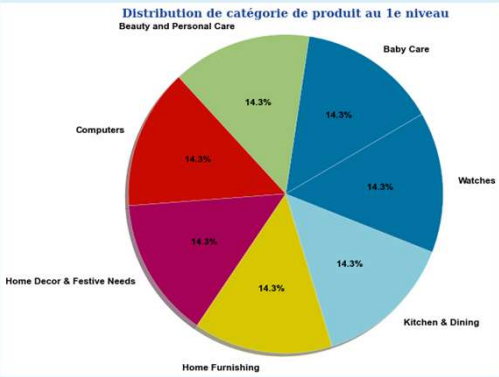
Sous_cat_2

Sous_cat_3

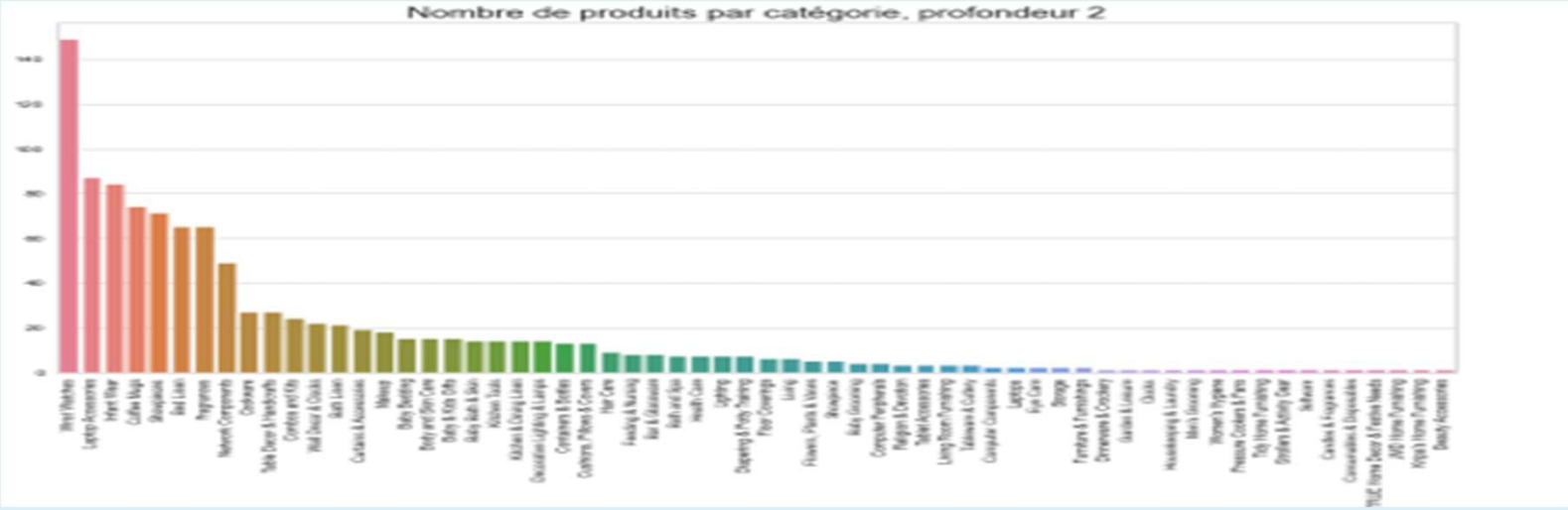
242 Catégories



7 Catégories



62 Catégories



ANALYSE DES CATEGORIES

8/9/2022

DONNEES TEXTUELLES

Pré-traitement du texte

- ❑ **Concaténation des variables textes : "product_name", "description" et "brand"**
- ❑ **Agréger toutes les descriptions ensemble**

Le corpus représente l'ensemble des descriptions de notre jeu de données.

Traitement du texte:
Extraction de variables à partir des données textuelles

- ❑ **Méthode Tf-idf**
- ❑ **Méthode NMF: Non-negative Matrix Factorization**
- ❑ **Méthode LDA: Latent Dirichlet Allocation**

Méthode Tf-idf

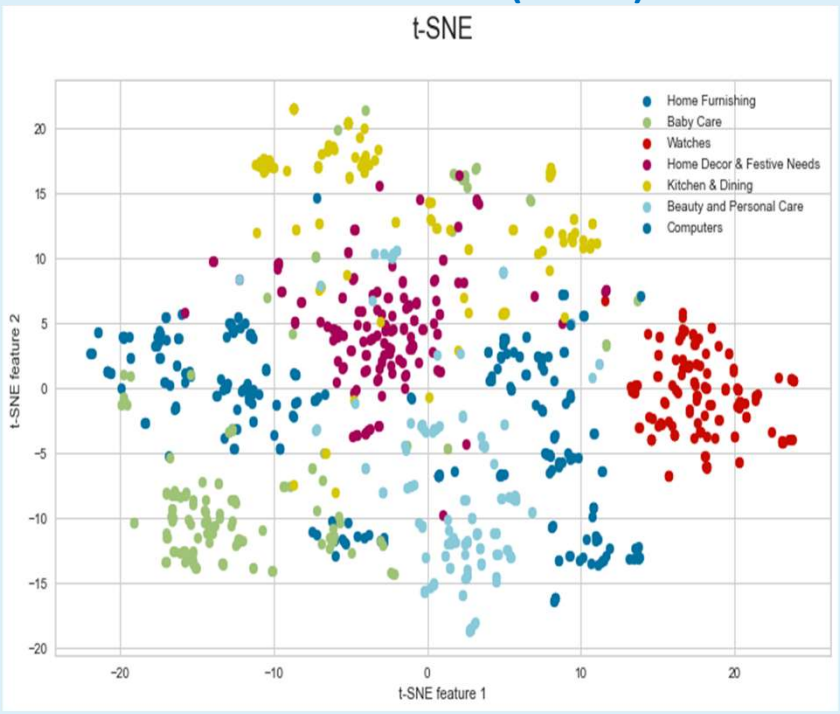
Taille du vocabulaire : 2443
Nombre de « stop word » : 3399

Vectorisation tf-idf

	bow_001	bow_005	bow_01	bow_03	bow_04	bow_05tg	bow_06	bow_085	bow_099	bow_10	...
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.042591	...
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...

5 rows × 2443 columns

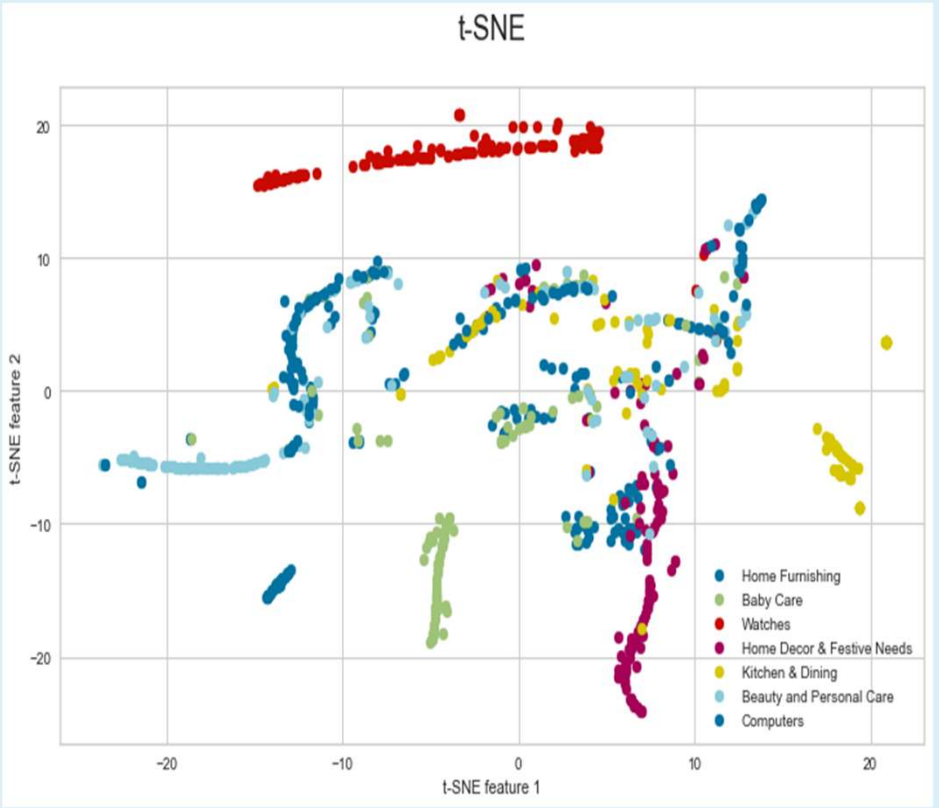
Visualisation via TSNE(TF-IDF)



Non-negative Matrix Factorization (NMF)

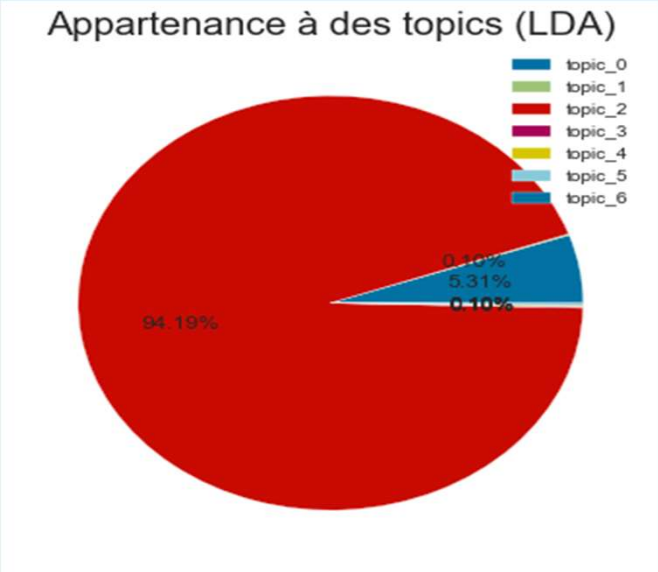
Topic 0: watch analog men women discounts india sonata great maxima boys
Topic 1: set combo com flipkart shipping cash genuine delivery products free
Topic 2: rockmantra mug ceramic porcelain stays thrilling crafting permanent ensuring creation
Topic 3: baby girl boy dress details cotton fabric neck shirt sleeve
Topic 4: cm showpiece best prices 10 handicrafts brass online 30 guarantee
Topic 5: mug coffee ceramic mugs prithish printland tea perfect presented wardrobe
Topic 6: laptop battery cell lapguard hp skin shapes pavilion print mouse

Visualisation via TSNE(NMF)

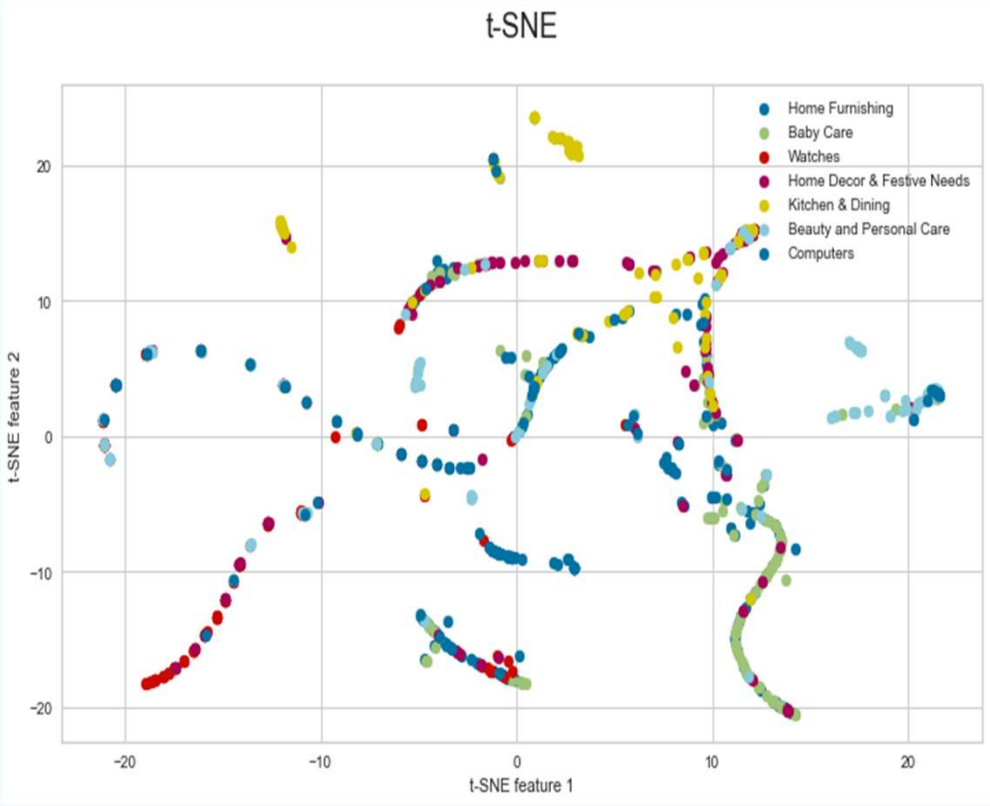


Latent Dirichlet Allocation (LDA)

Topic 0: mug ceramic coffee perfect mugs gift material rockmantra design tea
Topic 1: products delivery free buy cash shipping genuine 30 day guarantee
Topic 2: baby cm cotton pack features specifications general sticker color package
Topic 3: single pizza usb quilts comforters cutter hub steel floral multicolor
Topic 4: laptop adapter battery warranty replacement power charger quality vgn vaio
Topic 5: skin laptop shapes print combo set mouse pad multicolor warranty
Topic 6: cm showpiece material price box towel color features set brass



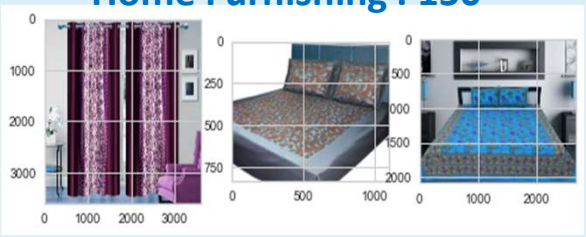
Visualisation via TSNE(LDA)



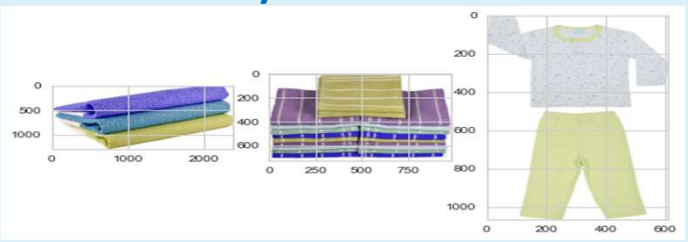
DONNEES VISUELLES

Visualisation: Exemples d'images dans chaque catégorie

Home Furnishing : 150

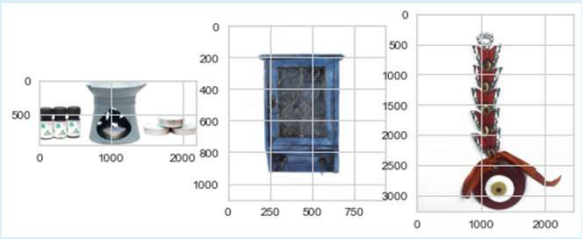


Baby Care : 150

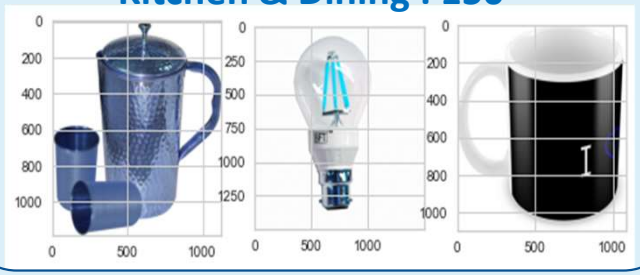


Watches : 150

Home Decor & Festive Needs : 150

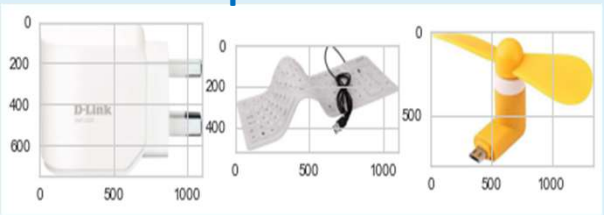


Kitchen & Dining : 150



Beauty and Personal Care : 150

Computers : 150



- Les images sont déjà isolées sur fond blanc (pas besoin de détection d'objet *a priori*)
- Certaines catégories présentent des objets de formes très différentes
- couleurs ou luminosité globale

Pré-traitement des images

- ❑ **Conversion de l'image en niveau de gris**
- ❑ **Egalisation**
- ❑ **Redimensionnement:**
- ❑ **Luminosité et Contraste**

Transformations Images Grayscale

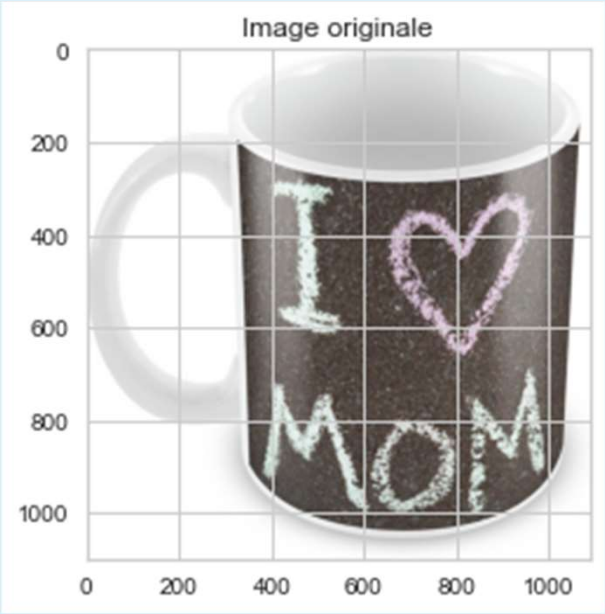
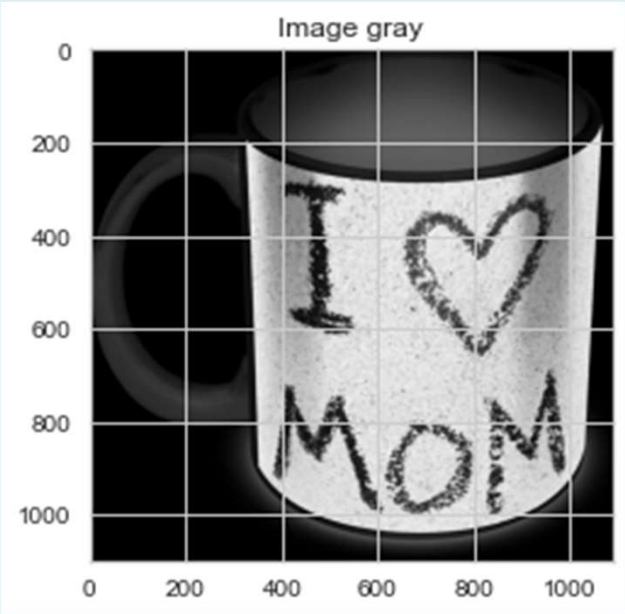


Image originale



Conversion de l'image
en niveau de gris

Amélioration de l'image avec
Egalisation de l'histogramme

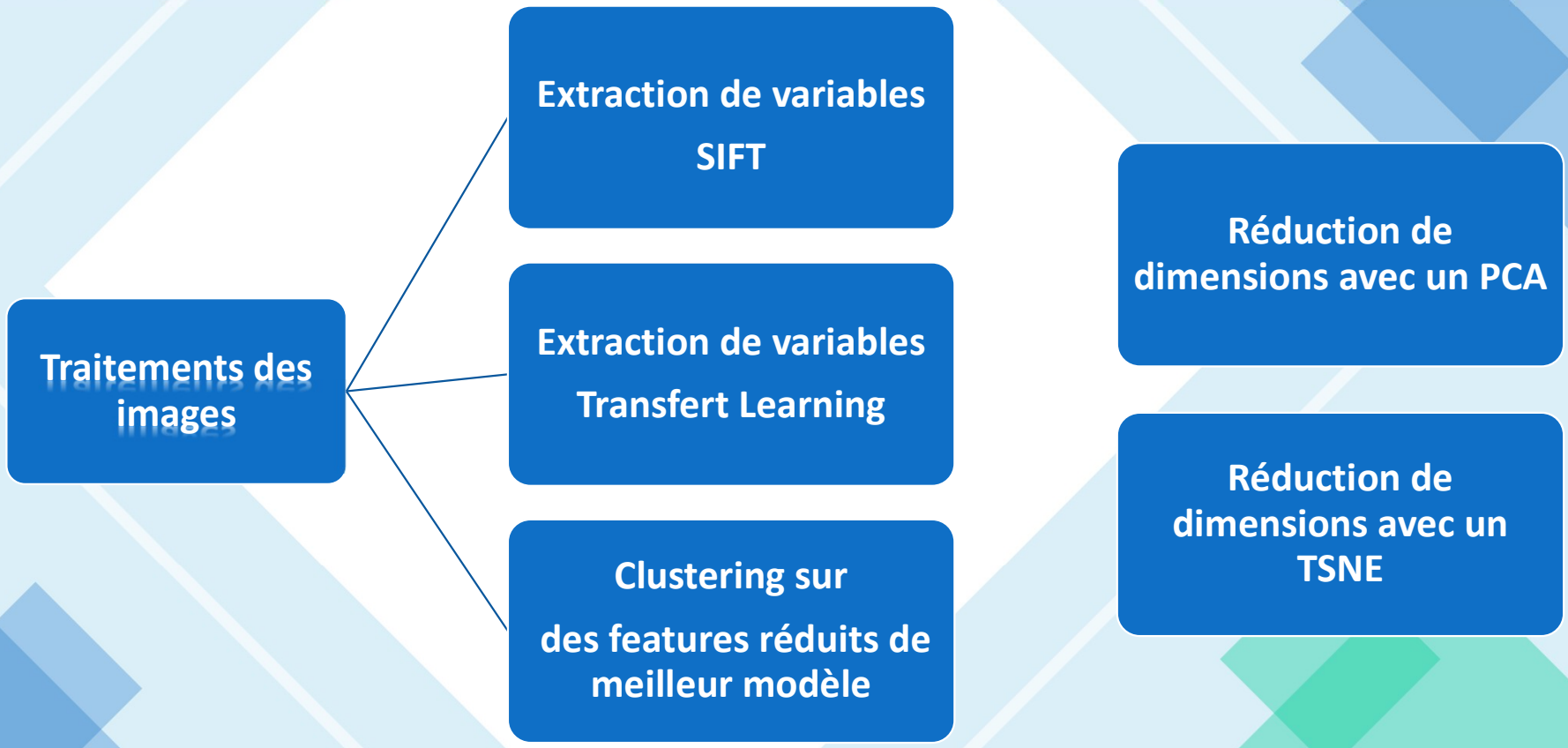
Transformations Images Couleurs

Dimensions initiales de l’image

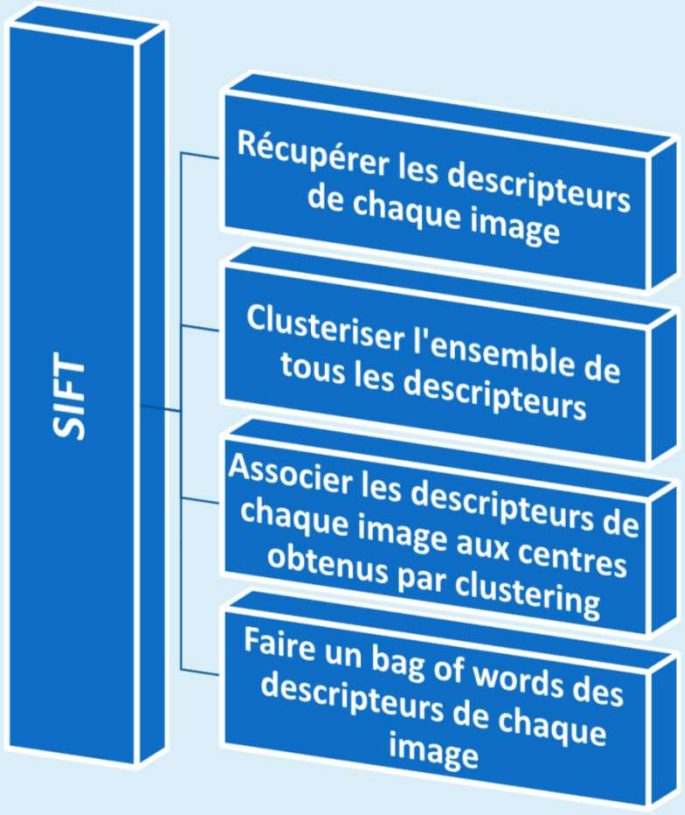
Transformation de la taille
D’images en 224x224

Amélioration de la luminosité
et du contraste

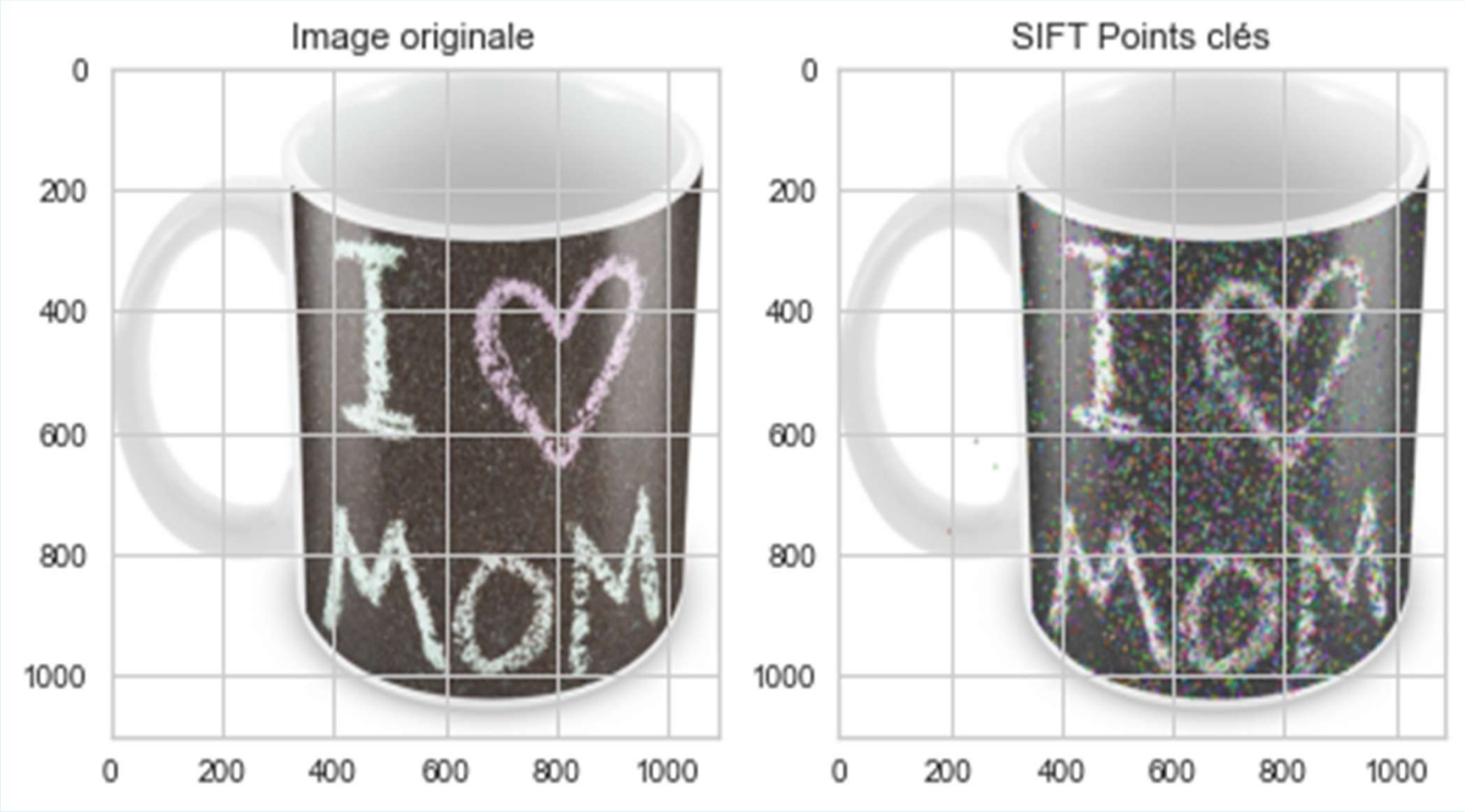
Processus de Traitements des images



Les étapes d'extraction de features par SIFT

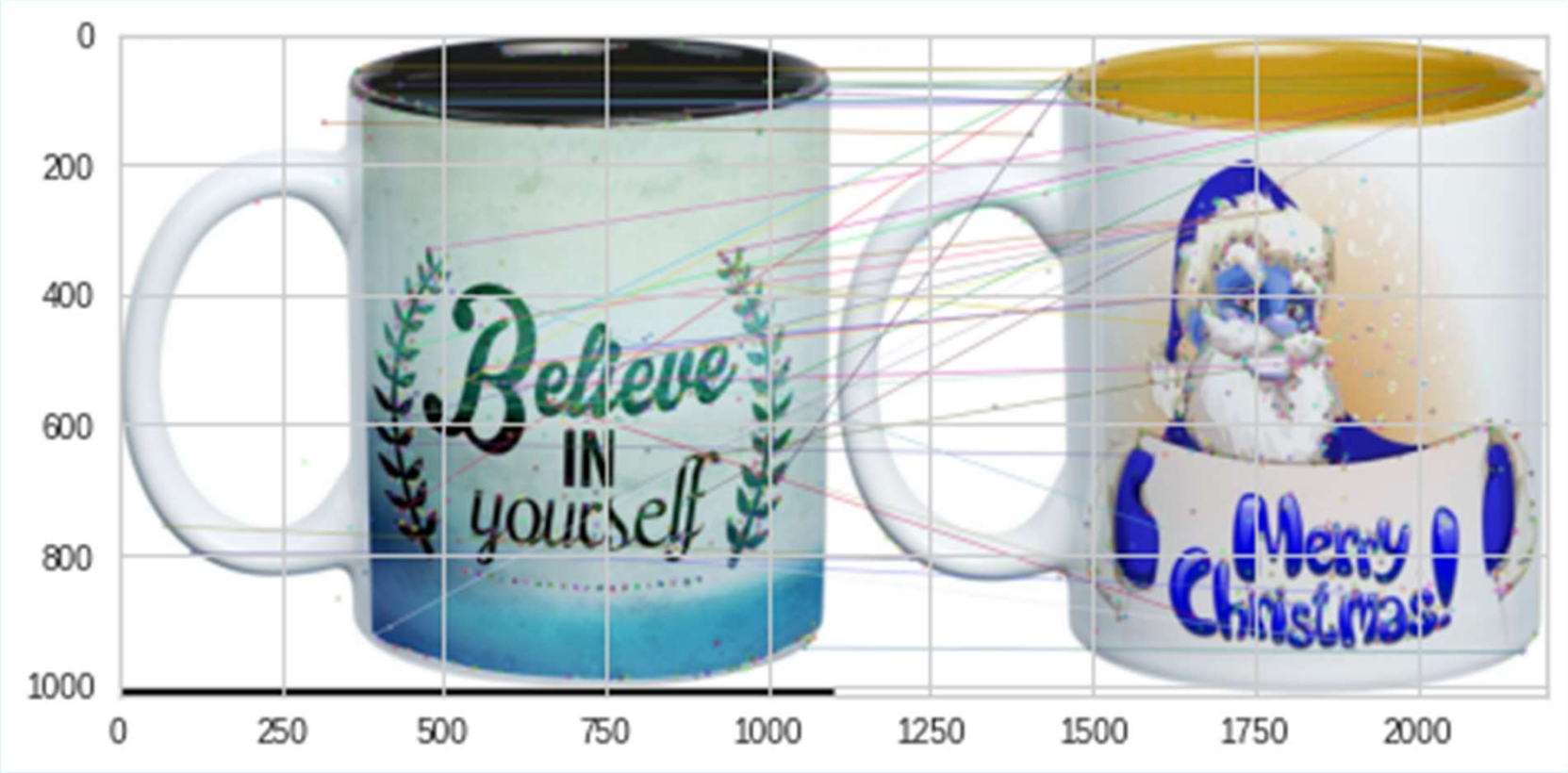


Extraction des caractéristiques par la méthode de SIFT

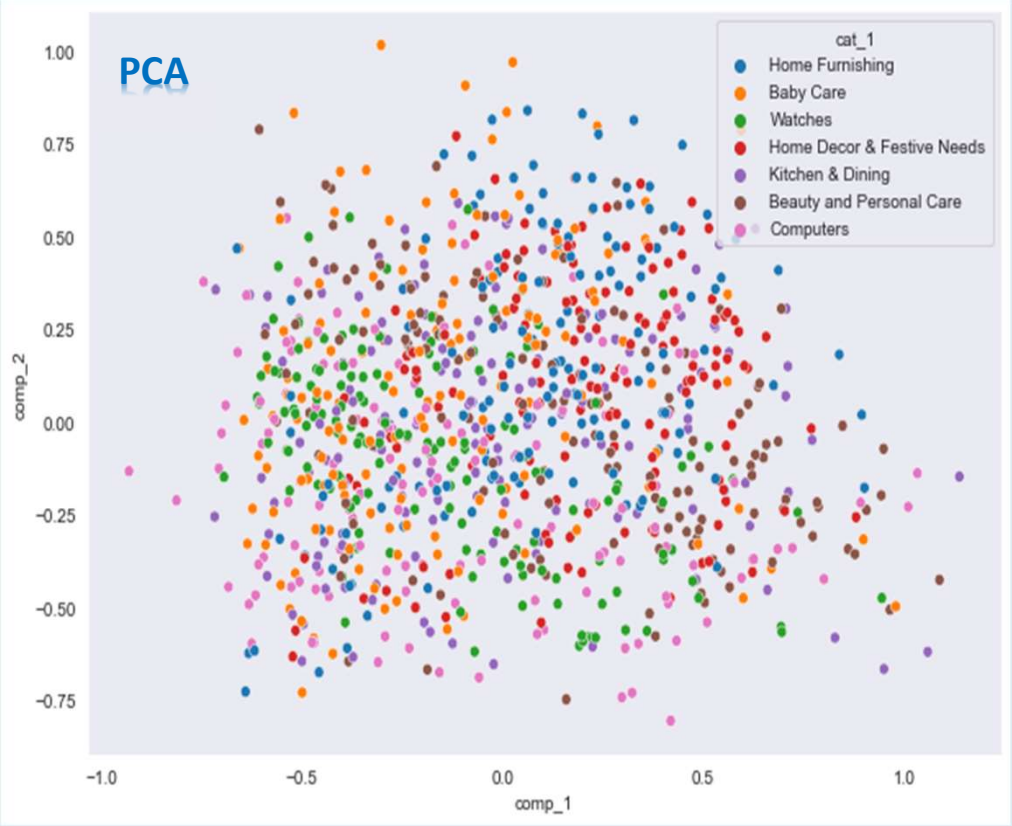


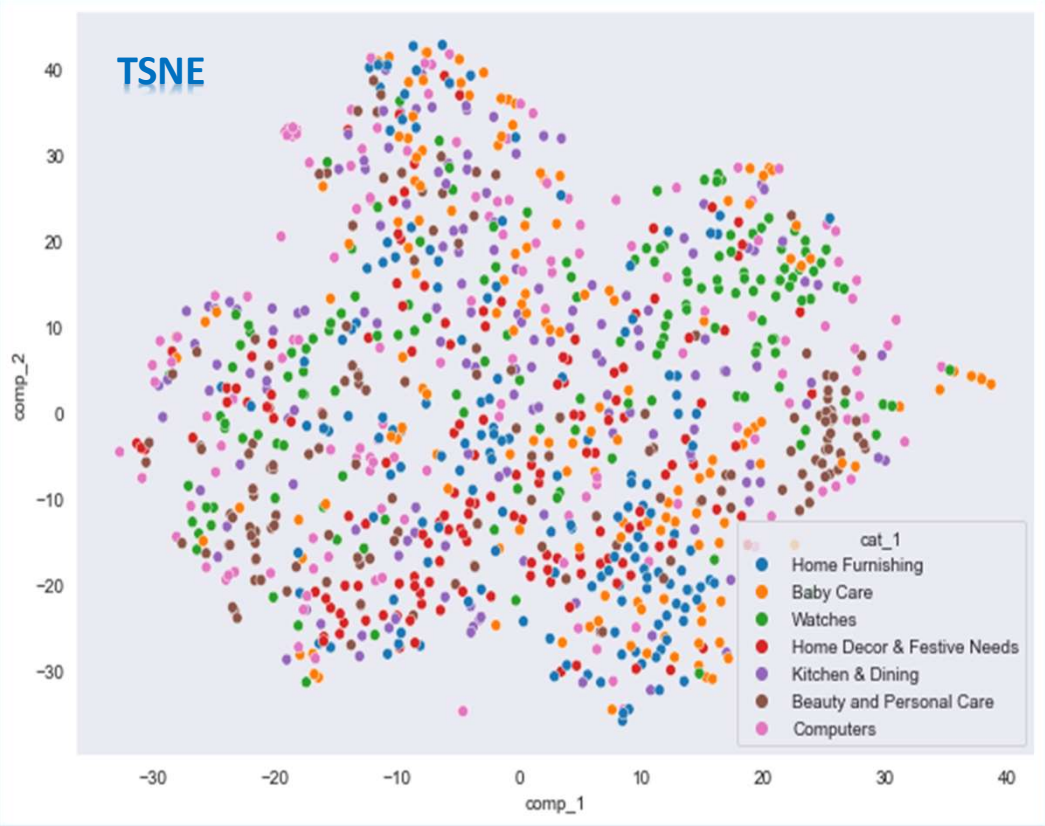

Merry

Correspondance des caractéristiques entre deux objets (Feature matching)



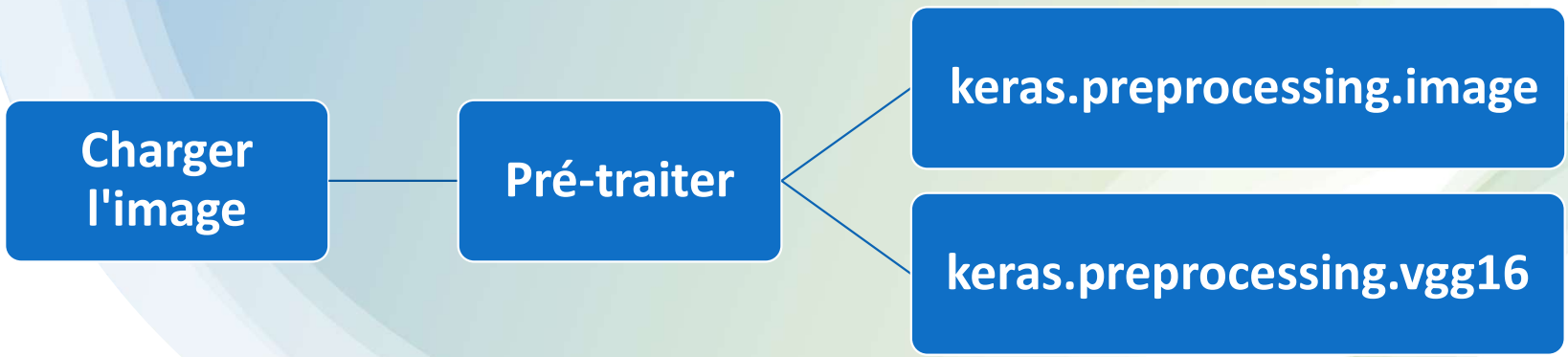
Réduction de dimensions avec un PCA et un TSNE sur SIFT BOW



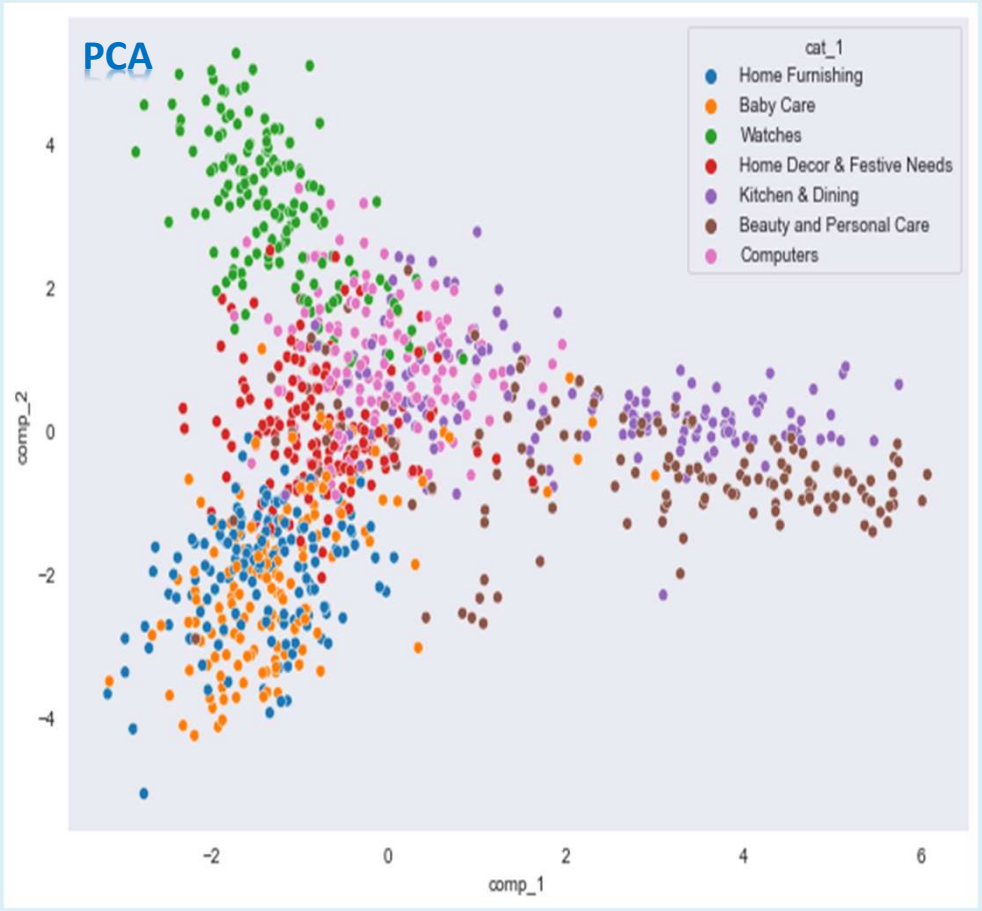


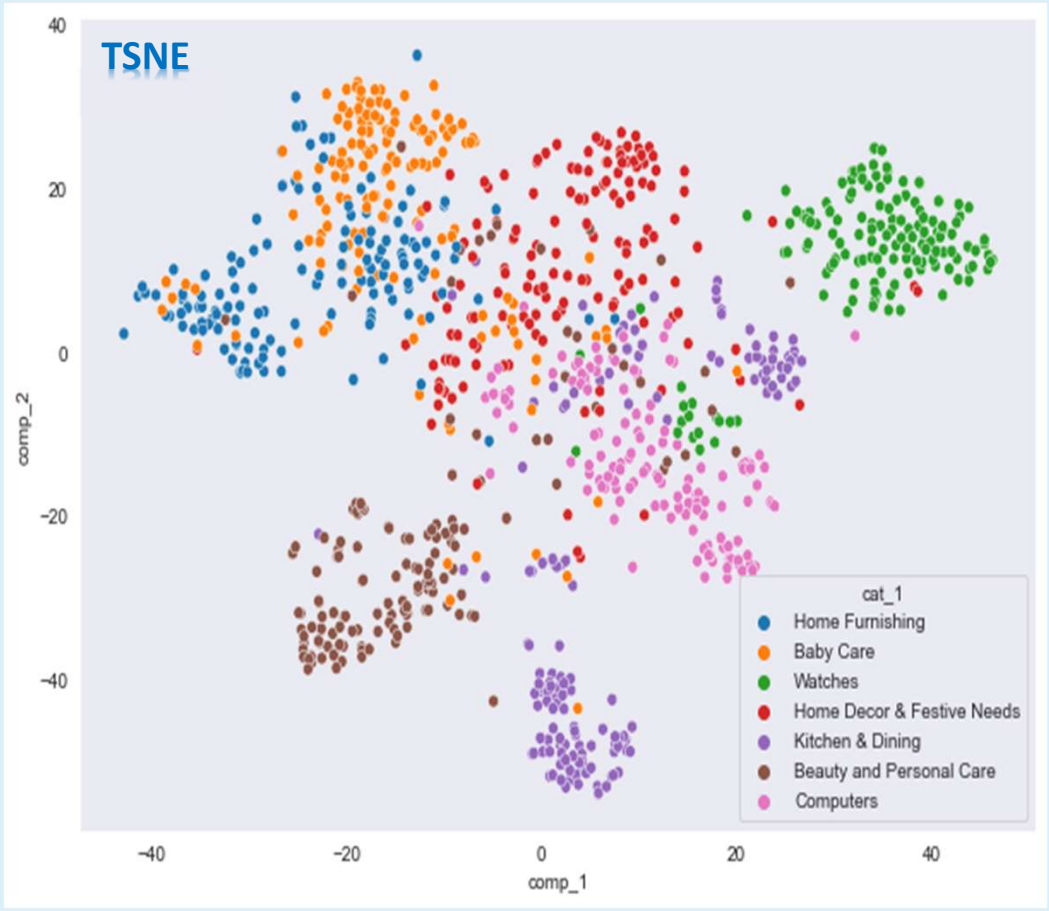
Transfer Learning

Modèle VGG-16 fourni par Keras et pré-entraîné sur ImageNet

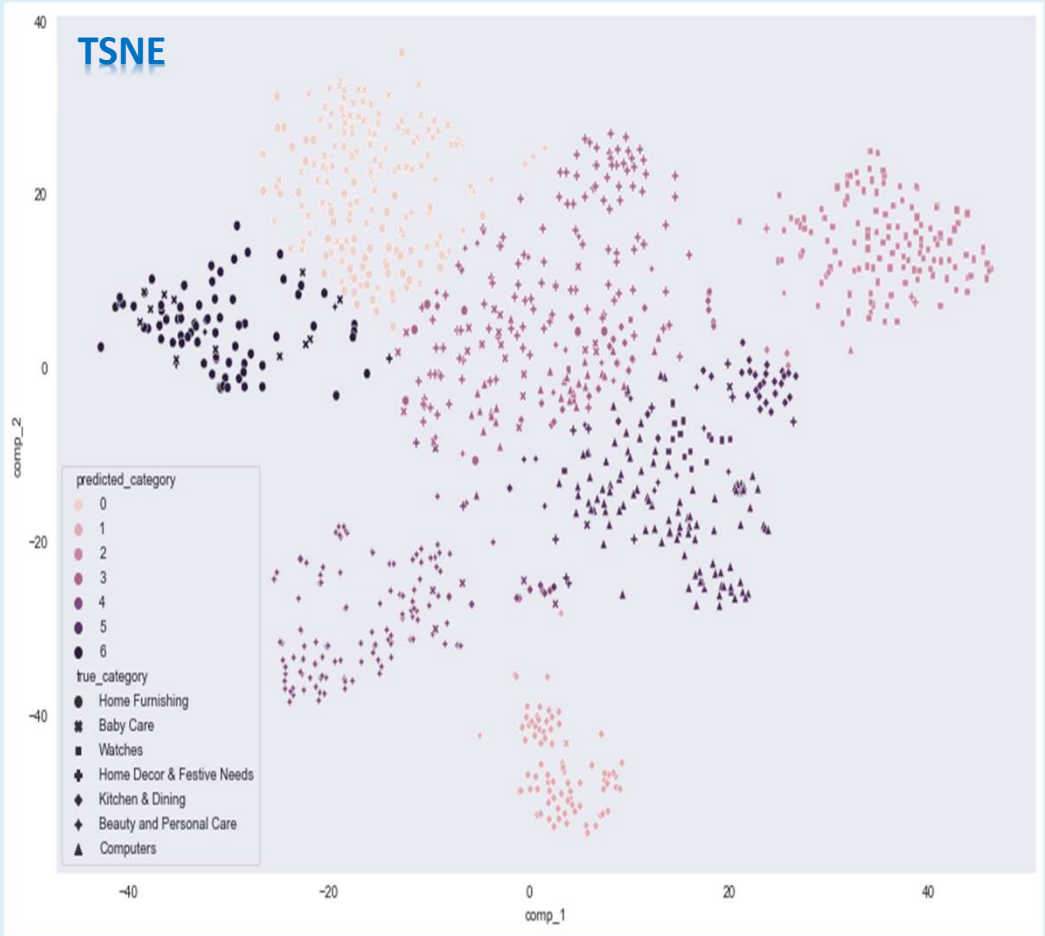


Réduction de dimensions avec un PCA et un TSNE sur Transfer Learning Features



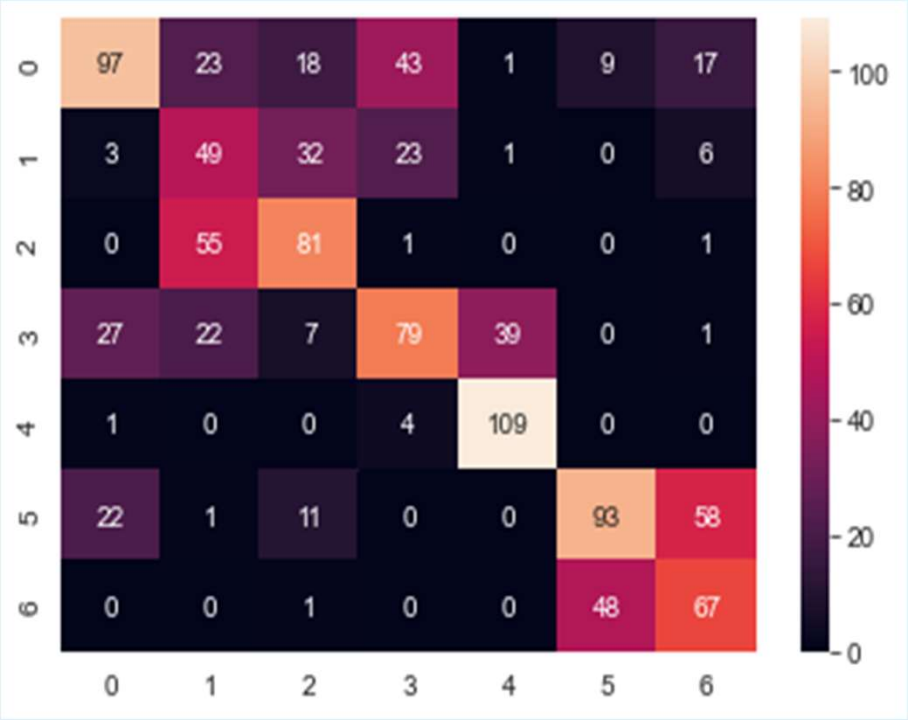


Clustering sur des Features réduits de VGG



Matrice de Confusion

ACP



Score ARI: 0.31

TSNE

Score ARI: 0.44

Conclusion et Perspectives

- Pour les données textuelles, la méthode Tf-idf, nous permets de mieux classifier les catégories de produits;
- Pour les données visuelles, le Transfer Learning avec une réduction de dimansion par un TSNE nous offre la meilleur classification;
- La faisabilité de la classification par cette méthode;
- La faisabilité de la classification automatique peut être amenée à changer si :
  - Le nombre d'image dans chaque catégorie se multiplie;
  - La qualité des descriptions ou des images changeait

Merci de votre attention