

Data Scientist

Sous le projet:

*Classifiez automatiquement des biens
de consommation*

Encadré par:

Abdou Karim Kandji

Réalisée par :

Kchaou Mariem

Octobre 2022

Plan



Problématique



Présentation des données



Données textuelles



Données images



Conclusion

- 1
 - 2
 - 3
 - 4
 - 5
- 2

1 Problématique

Plateforme d'e-commerce

- Une plateforme d'e-commerce proposant des produits à la vente
- Les données des produits issus de la base FlipKart incluent des descriptions textuelles et des images
- Attribution manuelle des catégories : fastidieuse et peu fiable
- Catégories déjà renseignées pour un petit volume de produits mais le volume de produit non catégorisés est destiné à s'accroître

⇒ Est-il possible de faire l'étude de faisabilité des produits d'une manière pertinentes ?

2

3

4

5

3

Problématique

Classification automatique de produit

Place de marché

Vendeurs sur
le Marketplace

Articles
(Descriptions
+ Images)

Attribution
catégorie
manuelle



Automatisation
De la tâche

❖ Mission :

Réaliser une étude de faisabilité d'un moteur de classification d'articles basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article pour l'entreprise Place de Marché.

❖ Objectifs :

- Améliorer l'expérience des utilisateurs.
- Fiabiliser la catégorie des articles avec pertinence et précision.

Jeu de données

- 1050 articles
- 15 colonnes par articles:
 - Identifiant : id, nom, catégorie de produit, marque, description
 - Prix / prix soldé
 - Note du produit
 - Image
 - Etc.
- Exemples d'articles:
 - Bracelets de montre
 - Vases
 - Linge de lit
 - Batteries d'ordinateur
 - Etc

Jeu de données

DATASET

uniq_id
crawl_timestamp
product_url
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

Utilisation de 3 features

Exemple d’article de notre jeu de données

```
array(['["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet  
Do..."]',  
      ['["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath  
Towel (3 Bath Towel, Red, Y...")'],  
      ['["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Eurospa Baby Bath Towels >> Eurospa Cotton Terry  
Face Towel Set (20 PIECE FA...")'],  
      ['["Home Furnishing >> Bed Linen >> Bedsheets >> SANTOSH ROYAL FASHION Bedsheets >> SANTOSH ROYAL FASHION  
Cotton Printed King sized ..."]'],
```



0. Home Furnishing

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain, Elegance Polyester Multicolor Abstract Eyelet Door Curtain

1. Baby Care

Specifications of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable

Données textuelles

Analyse des catégories

La feature « product_category_tree » représente l’arborescence complète d’un article.

Reprenons l’exemple de notre premier article dans le jeu de données

On extrait les différents noeuds de l’arborescence

```
array(['["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet  
Do..."]',
```

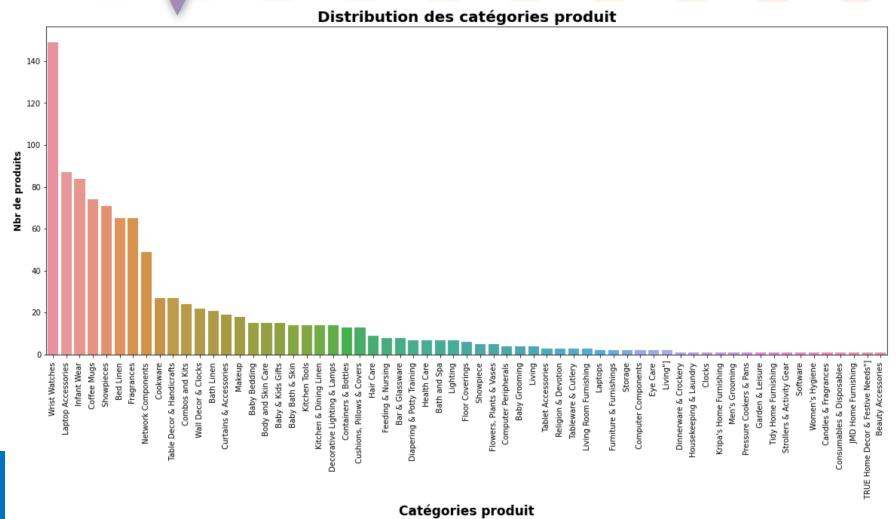
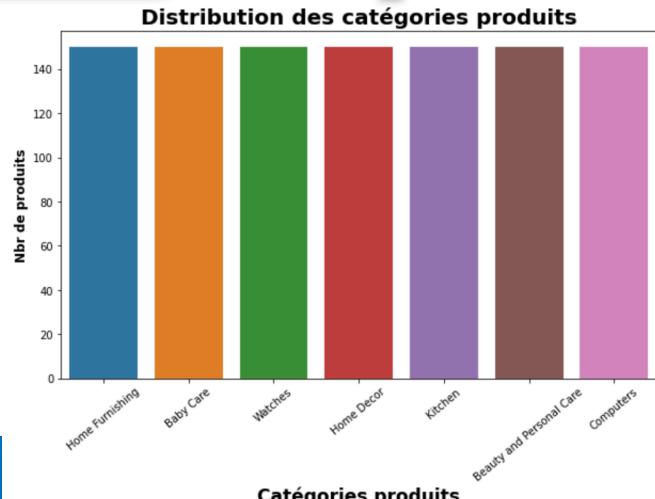
```
array(['["Home Furnishing > Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet  
Do..."]',
```

categorie_1

categorie_2

63 catégories

7 catégories



1

2

4

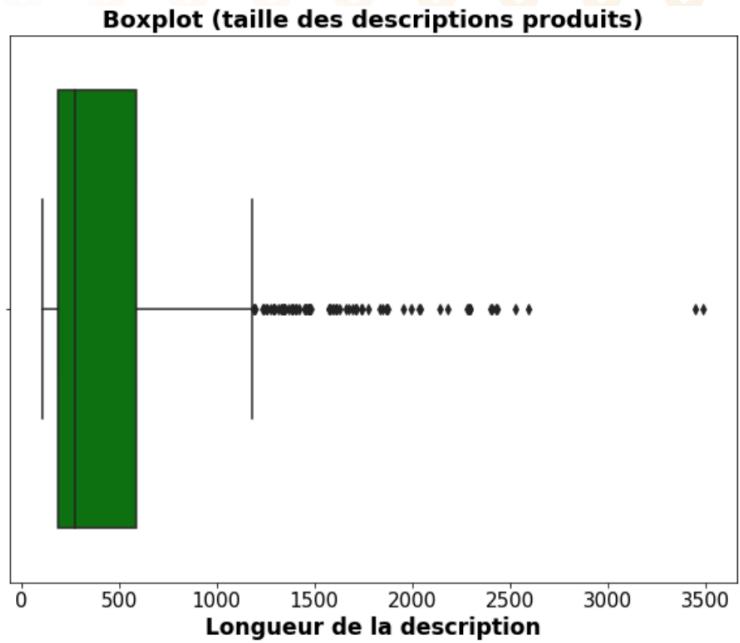
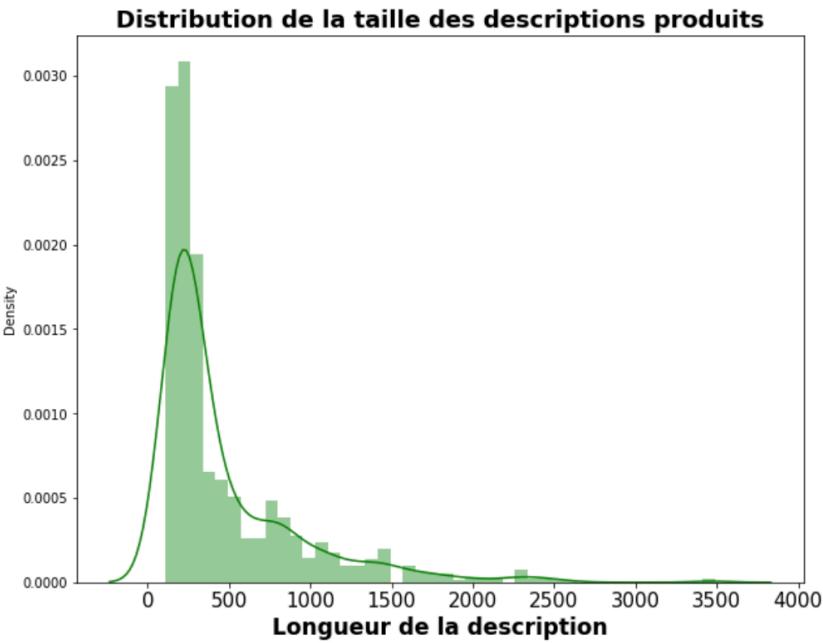
5

7

3

Données textuelles

Les descriptions produits



1

2

4

5

8

Données textuelles

Prétraitement

- Traitements successifs (librairie NLTK)

Minuscule



Tokeniser



Retrait mots de liaison et ponctuation



Lemmatisation/
Racinisation
(Stemming)



Bag of Words/
TF IDF

A => a

I am => 'i', 'am'

An, own, are, in...
, « « :; ...

Am => be
Products => product

vectorisation

- Exemple

'key features of eurospa cotton terry face towel set size: small height: 9 inch gsm: 360,eurospa cotton terry face towel set (20 piece face towel set, assorted) price: rs. 299 eurospa brings to you an exclusively designed, 100% soft cotton towels of export quality. all our products have soft texture that takes care of your skin and gives you that enriched feeling you deserve. eurospa has been exporting its bath towels to lot of renowned brands for last 10 years and is famous for its fine prints, absorbency, softness and durability. note: our product is 100% cotton, so it is susceptible to shrinkage. product color may vary from the picture. size may vary by ±3% wash care: wash in cold water, do not iron, do not bleach, flat dry, wash before first use. size- face towel - 23 cms x 23 cms.,specifications of eurospa cotton terry face towel set (20 piece face towel set, assorted) bath towel features material cotton terry design shuvam general brand eurospa gsm 360 type face towel set model name shuvam20pcftsetassorted ideal for boys, girls, men, women model id shuvam20pcftsetassorted size small color assorted dimensions weight 350 g length 9 inch width 9 inch in the box number of contents in sales package 20 sales package 20 piece face towel set'

'key feature cotton terry face towel set size small height inch cotton terry face towel set piece face towel set assorted price r exclusively designed soft cotton towel export quality product soft texture take care skin give feeling deserve bath towel lot renowned brand last year famous fine print absorbency softness durability note product cotton susceptible shrinkage product color may vary picture size may vary ± wash care wash cold water iron bleach flat dry wash first use size face towel x cotton terry face towel set piece face towel set assorted bath towel feature material cotton terry design general brand type face towel set model name ideal boy girl men woman model id size small color assorted dimension weight ± length inch width inch box number content sale package piece face towel set'

Données textuelles

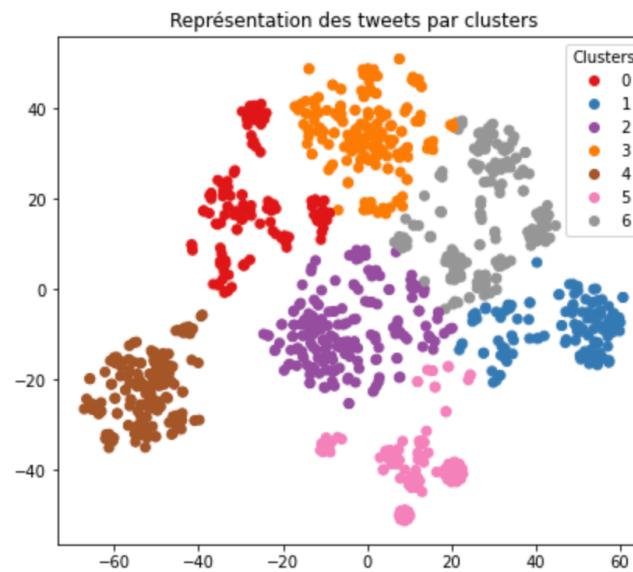
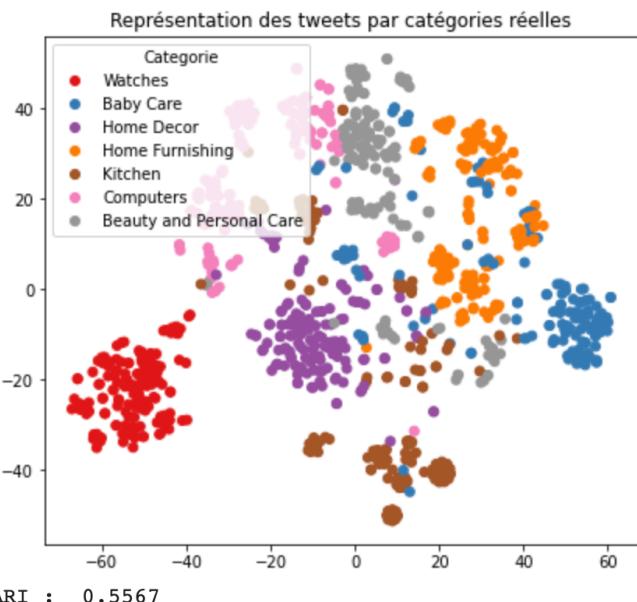
Bag of words -Tf-idf

CountVectorizer :

ARI : 0.3784 time : 10.0

Tf-idf :

ARI : 0.5567 time : 12.0



1

2

4

5 10

Word2Vec

Word2Vec

Réseau de neurones entraîné pour analyser le contexte linguistique des mots.

Utilisation du réseau pré-entraîné

Entrainement sur 100 milliards de mots.

Contient un dictionnaire de 3 millions de mots, vectorisés en dimension 300.

Méthodologie pour vectoriser notre dataset:

Pour chaque document:

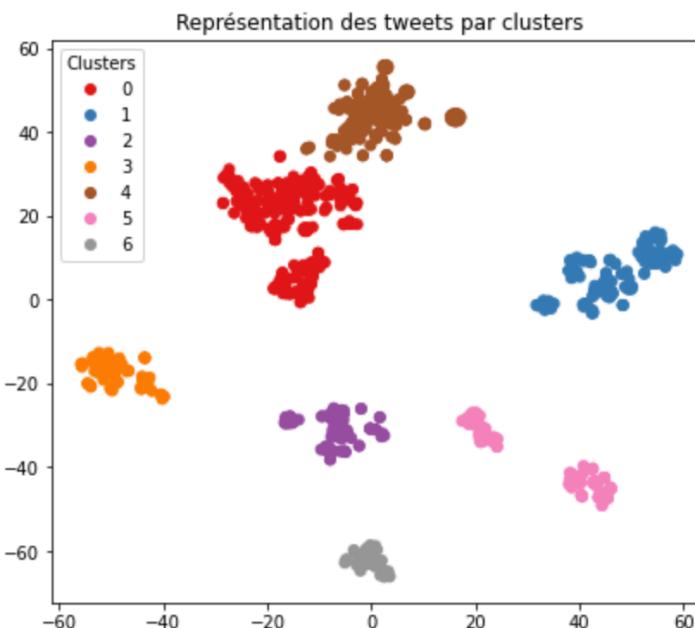
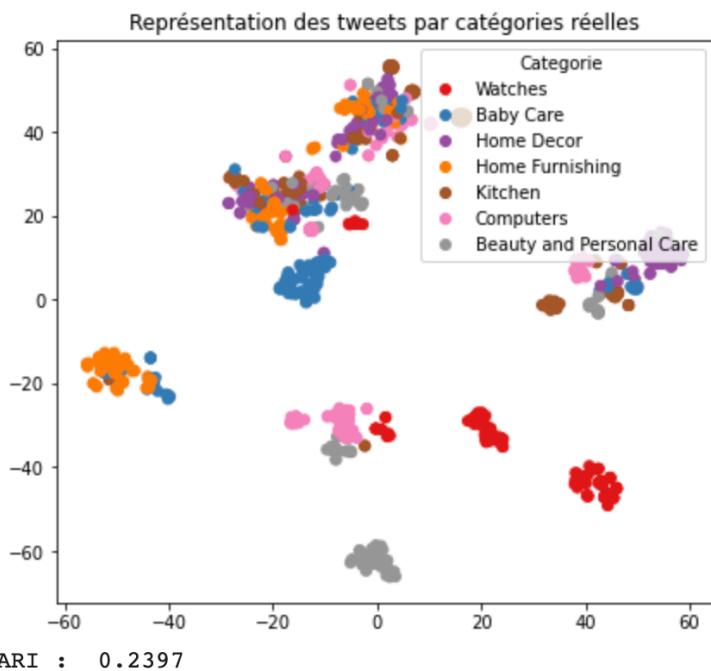
- 1) les tokens (sans stemming) sont convertis en vecteur de dimension 300 (si disponible dans le dictionnaire Word2Vec)
- 2) la moyenne des vecteurs est calculée => obtention d'un vecteur pour le document.

=> Toutes les documents sont décrits par un vecteur de dimension 300.

3

Données textuelles

Word2Vec



1

2

4

5

12

Bert

Bert

signifie **Bidirectional Encoder Representations from Transformers**, consistant en un *encodeur* pour lire le texte et un *décodeur* pour faire une prédiction. BERT se limite à un encodeur, car son objectif est de créer un modèle de représentation du langage qui sera ensuite utilisable pour des tâches de NLP

Méthodologie pour vectoriser notre dataset:

Pour chaque document:

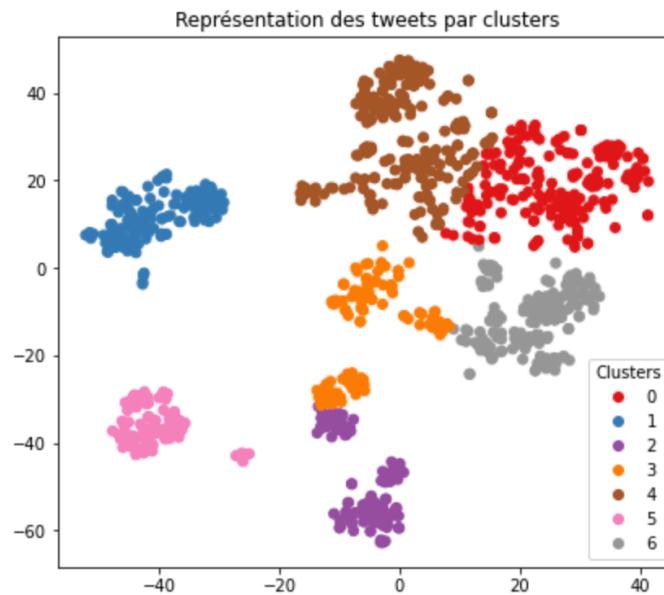
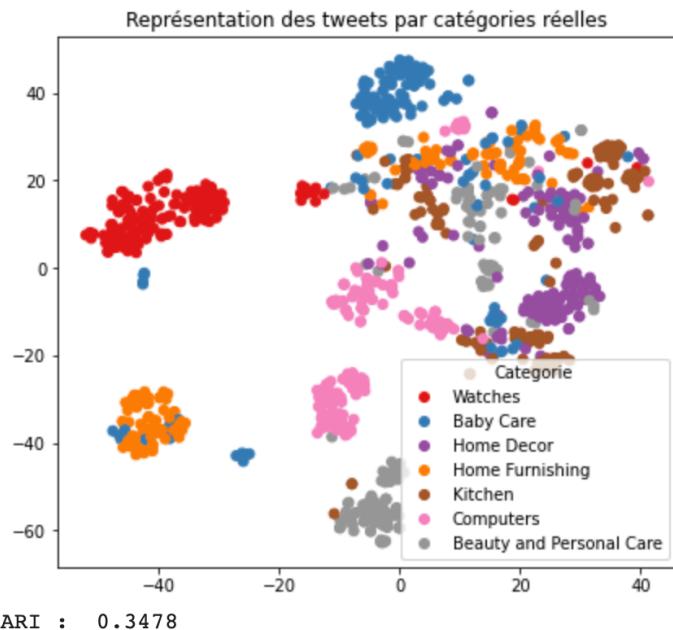
- 1) Tokenisation des mots et ajout de tokens de début et de fin de phrase.
- 2) Marqueur ajouté à chaque phrase pour les distinguer.

=> BERT se base donc sur l'architecture des transformers

3

Données textuelles

Bert



1

2

4

5

14

USE

USE

signifie **Universal Sentence Encoder**. Encode le texte en vecteurs de grande dimension qui peuvent être utilisés pour la classification de texte, la similarité sémantique, le regroupement et d'autres tâches de langage naturel.

Et aussi un Réseaux de neurones pré-entraînés encodant des phrases en vecteurs de dimensions 512 que l'on réduit à 178 dimensions par PCA (95% de variance).

1

2

4

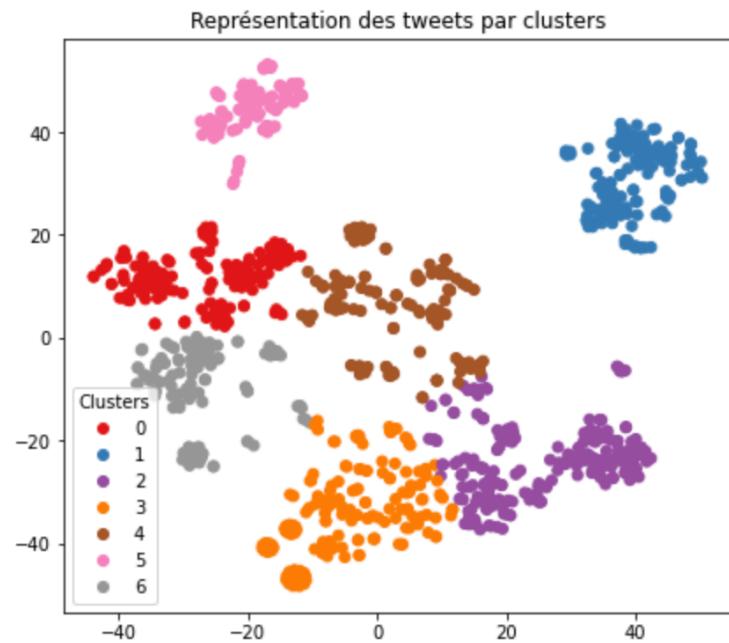
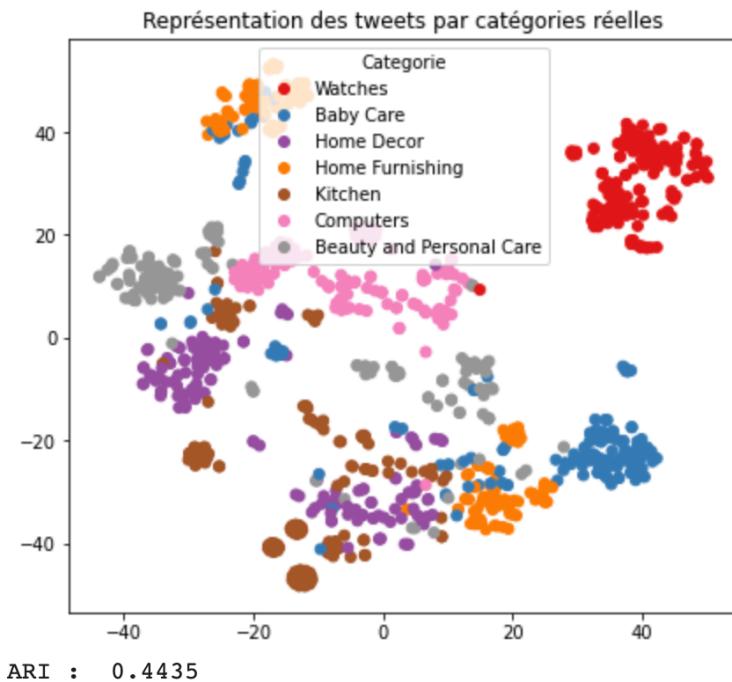
5

15

3

Données textuelles

USE



1

2

4

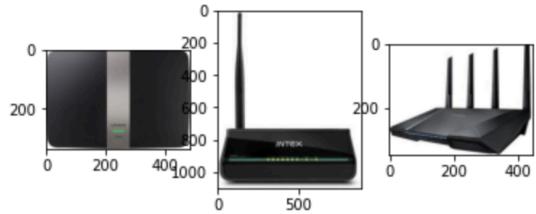
5

16

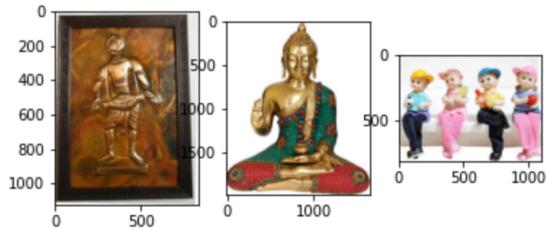
Données images

Les images suivent leur catégories

Computers



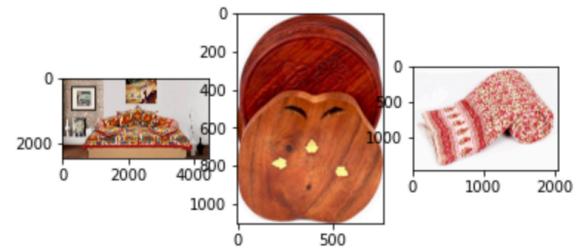
Home Decor



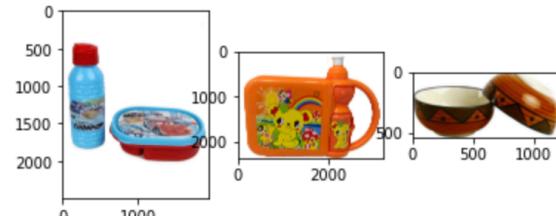
Watches



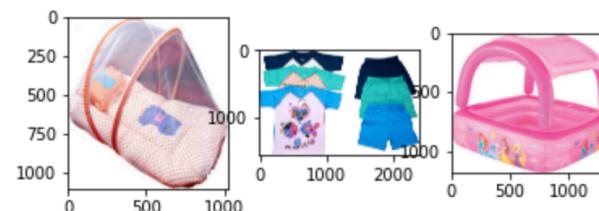
Home Furnishing



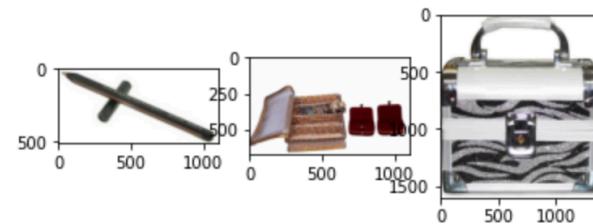
Kitchen



Baby Care



Beauty and Personal Care



SIFT**SIFT**

signifie **scale-invariant feature transform**. est un algorithme utilisé dans le domaine de la vision par ordinateur pour détecter et identifier les éléments similaires entre différentes images numériques (éléments de paysages, objets, personnes, etc.)

Méthodologie pour vectoriser notre dataset:

- 1) Convertir en niveaux de gris.
- 2) Uniformiser l'histogramme des gris.
- 3) Calculer les descripteurs SIFT.

=> obtention de vecteurs à 128 dimensions (plusieurs milliers par image du dataset)

Données images

Détermination et affichage des descripteurs SIFT



Descripteurs : (3098, 128)

```
[[ 4.   8.   2. ...   0.   0.   9. ]
 [ 3.   1.   1. ...   0.   0.   4. ]
 [ 19.   2.   1. ...   0.   4.   56. ]
 ...
 [ 56.  142.  49. ...   0.   0.   0. ]
 [ 5.   0.   0. ...   0.   0.   1. ]
 [ 25.  10.  0. ...   0.   0.   1.]]
```

1

2

3

5

19

Pré-traitement des images via SIFT

0
100
200
300
400
500
600
700
800
900
1000

Nombre de descripteurs : (210796, 128)
temps de traitement SIFT descriptor :

572.11 secondes

SIFT

Nombre de clusters estimés : 459
Création de 459 clusters de descripteurs ...

Dimensions dataset avant réduction PCA : (1050, 459)
Dimensions dataset après réduction PCA : (1050, 381)

Dimension du dataset après Réduction T-SNE

(1050, 3)

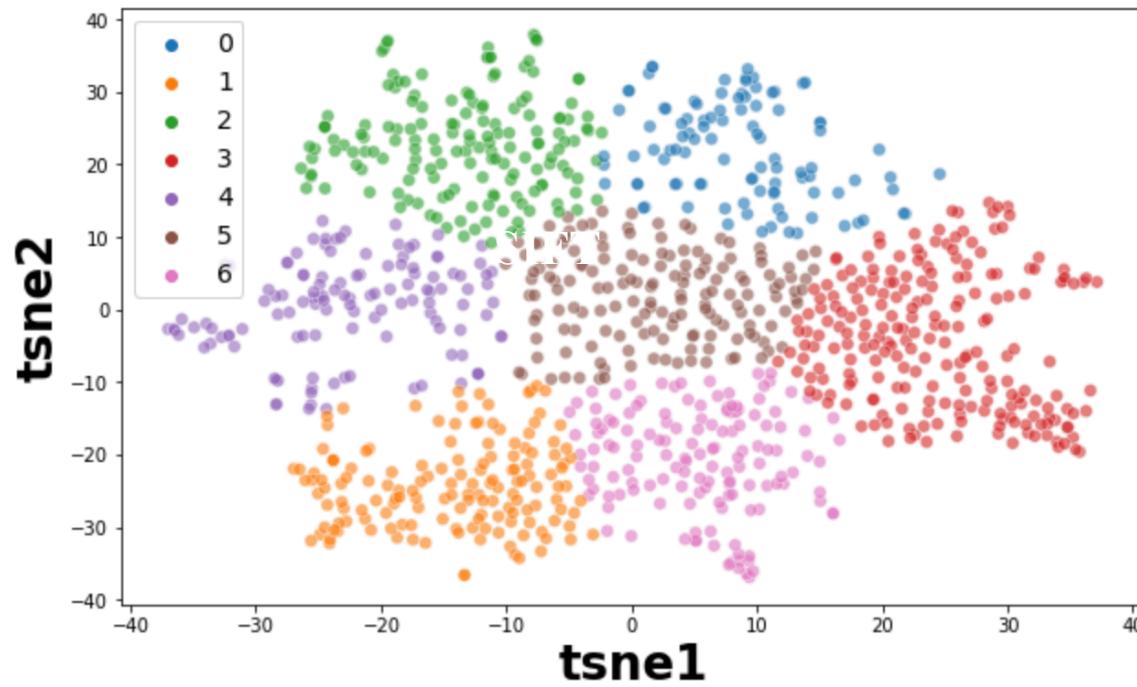
- 1
- 2
- 3
- 5
- 20

4

Données images

SIFT

TSNE selon les clusters



1

2

3

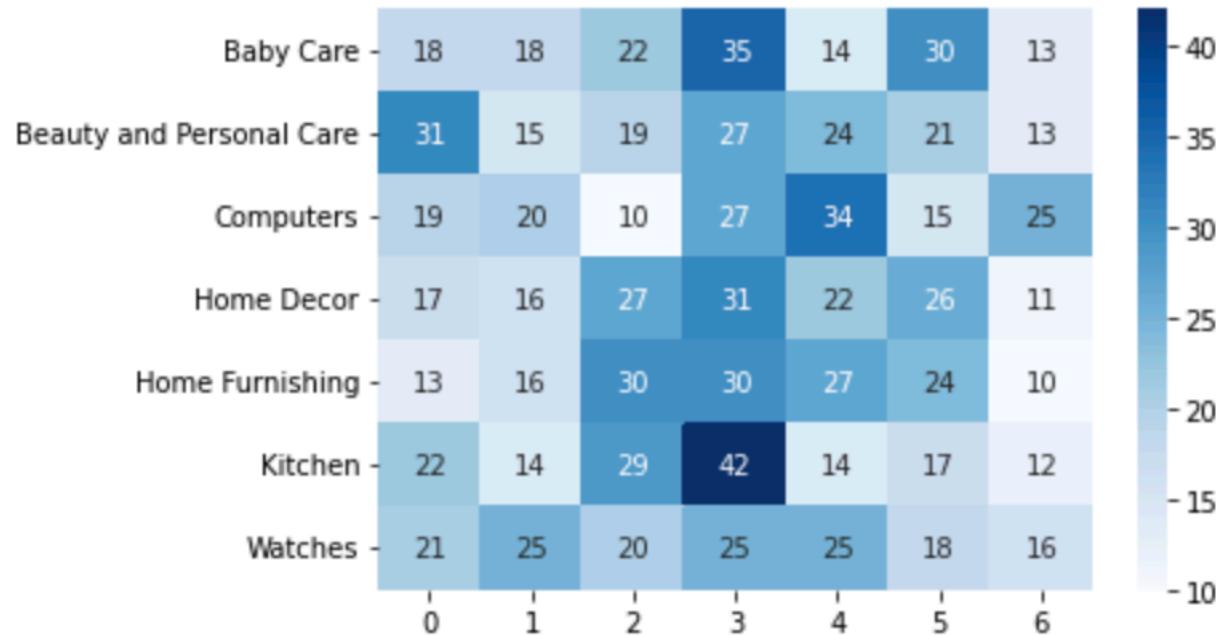
5

21

21

SIFT

Matrice de confusion



1

2

3

5

22

Transfer Learning

ResNet50

ResNet50

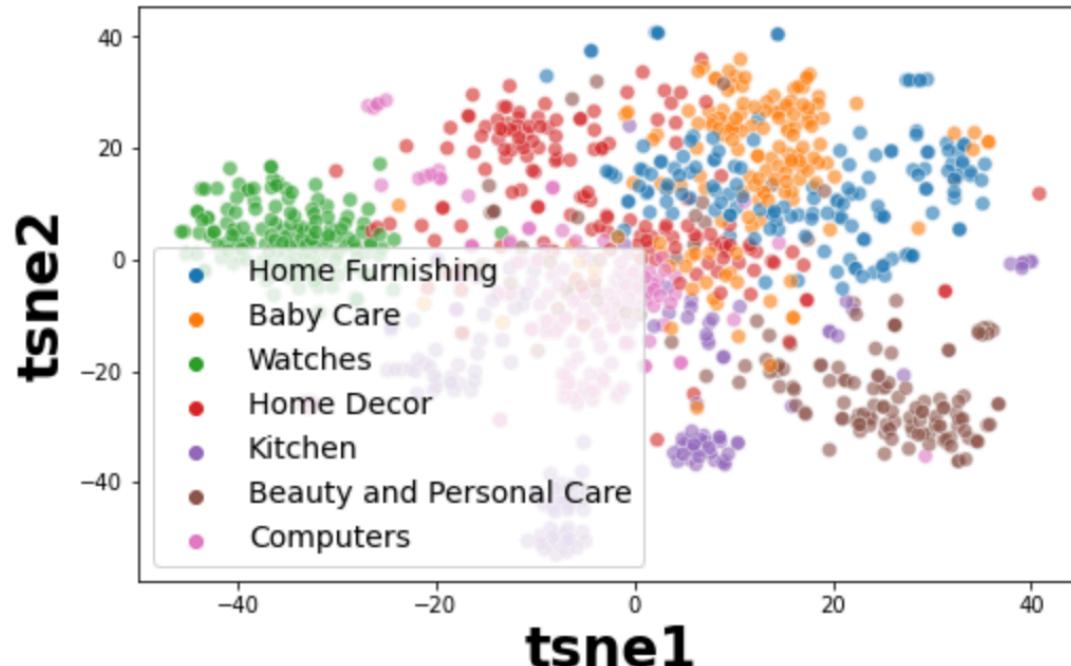
VGG-Net présente deux inconvénients majeurs :

- Il est lent à entraîner.
- Les poids de l'architecture de réseau sont importants (en termes de disque / bande passante), cela peut rendre le déploiement de VGG fastidieux.
- Le réseau ResNet est plutôt une forme « d'architecture exotique » qui repose sur des modules de micro- architecture.
- Le terme micro-architecture désigne l'ensemble des « blocs de construction » utilisés pour construire le réseau.
- Un ensemble de blocs de construction de micro-architecture est composé de couches standards (Convolution, Pool, etc.)

Transfer Learning :

- Utilisation d'un modèle pré-entraîné sur ImageNet.
- Taille du vecteur par image : 100352.

TSNE selon les vraies classes

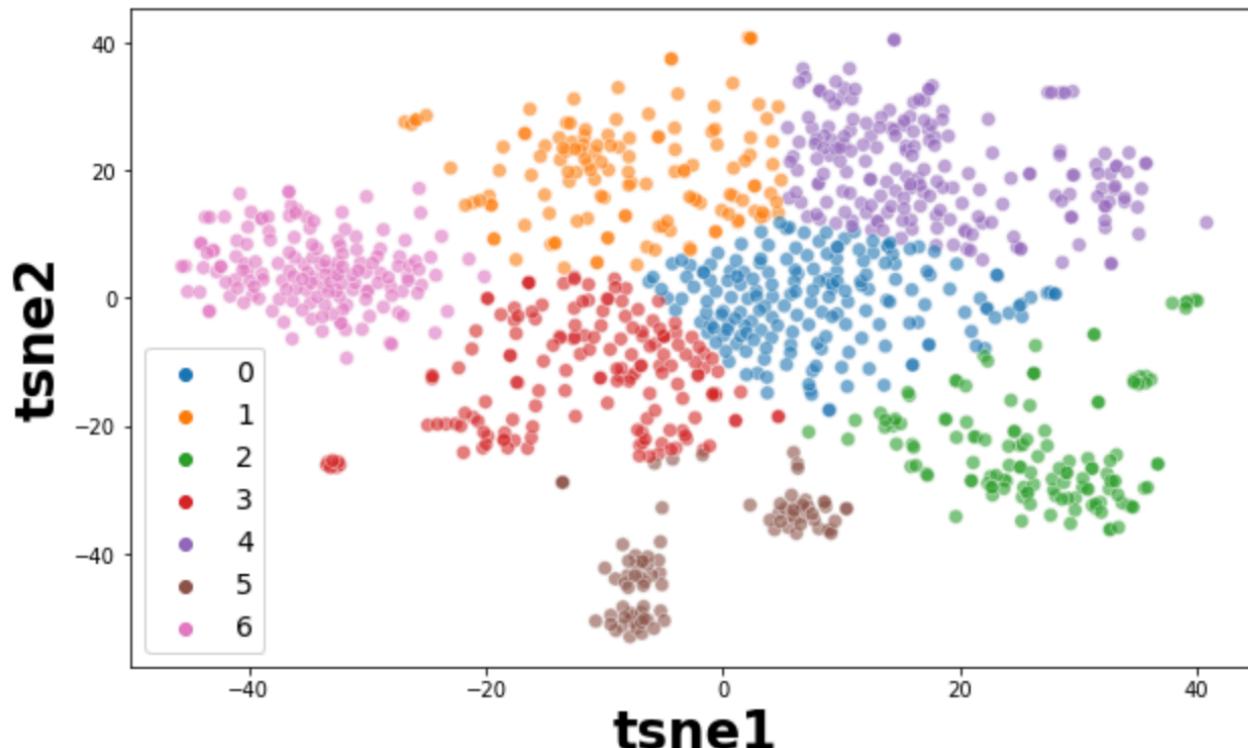


4

Données images

Transfer Learning

ResNet50



1

2

3

5

25

Transfer Learning

VGG16

VGG16

Réseau de neurones convolutif à 16 couches pour la classification d'images.

Réseau pré-entraîné sur 1.3 millions d'images.

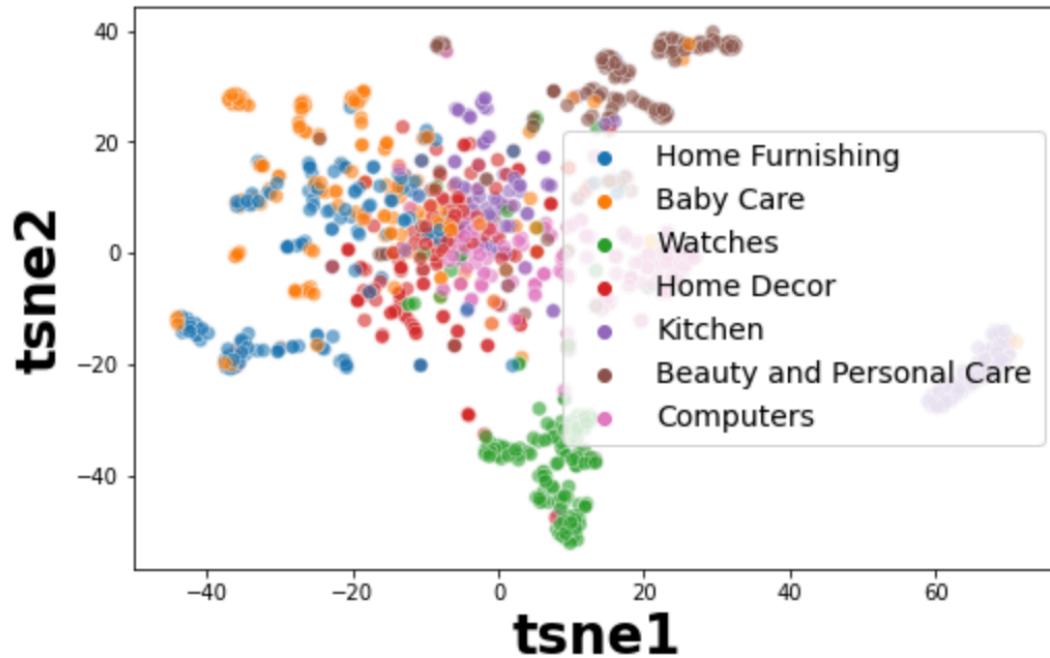
Possibilité de l'utiliser en transfer learning.

Dernière couche : classifieur softmax qui prend un vecteur de dimension 4096 en entrée.

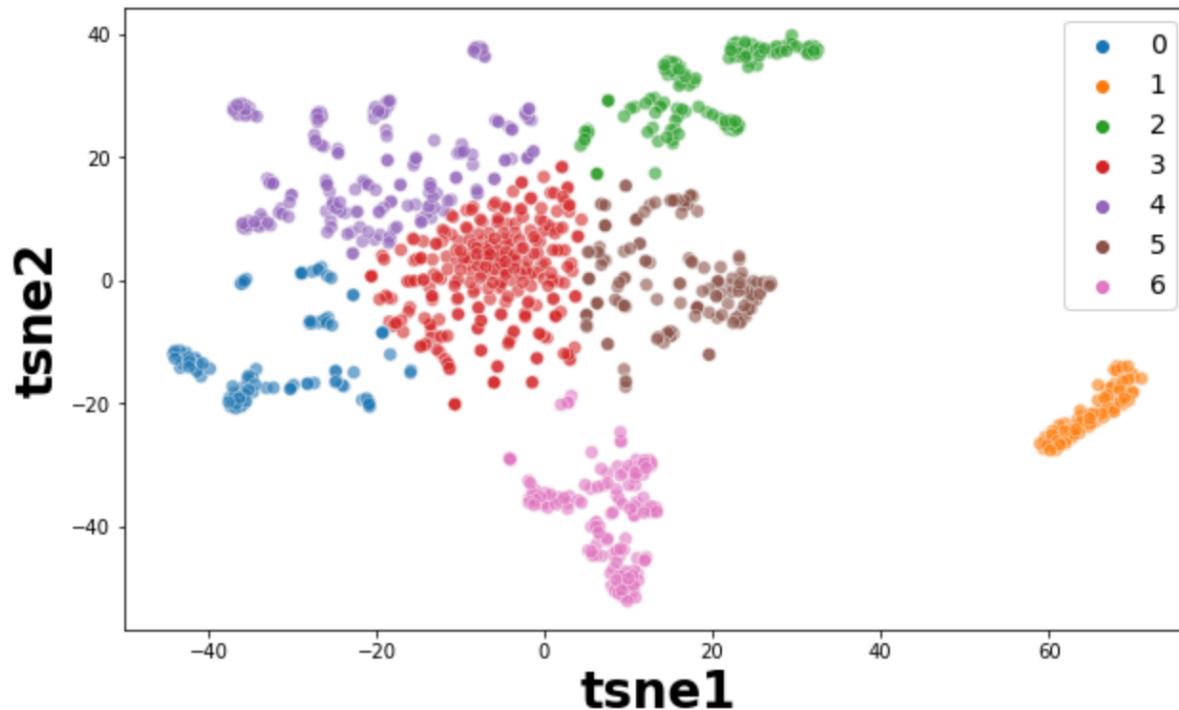
Génération du descripteur (4096 dimensions) d'une image :

- 1) Preprocessing de l'image (en particulier redimensionnement au format 224*224 pixels)
- 2) Calcul du descripteur via le modèle préentraîné.

TSNE selon les vraies classes



TSNE selon les clusters



Conclusion et recommandations

Recommandations

- Proposer une description produit d'une certaine longueur.
- Proposer des photos d'une certaines qualité donnée.
- Enrichir notre base de données.
- Découper certaines catégories en sous catégories.
- Avoir une approche combinée (texte et image).

1

2

3

4

Merci de votre attention

