Projet 6

Classifiez automatiquement des biens de consommation

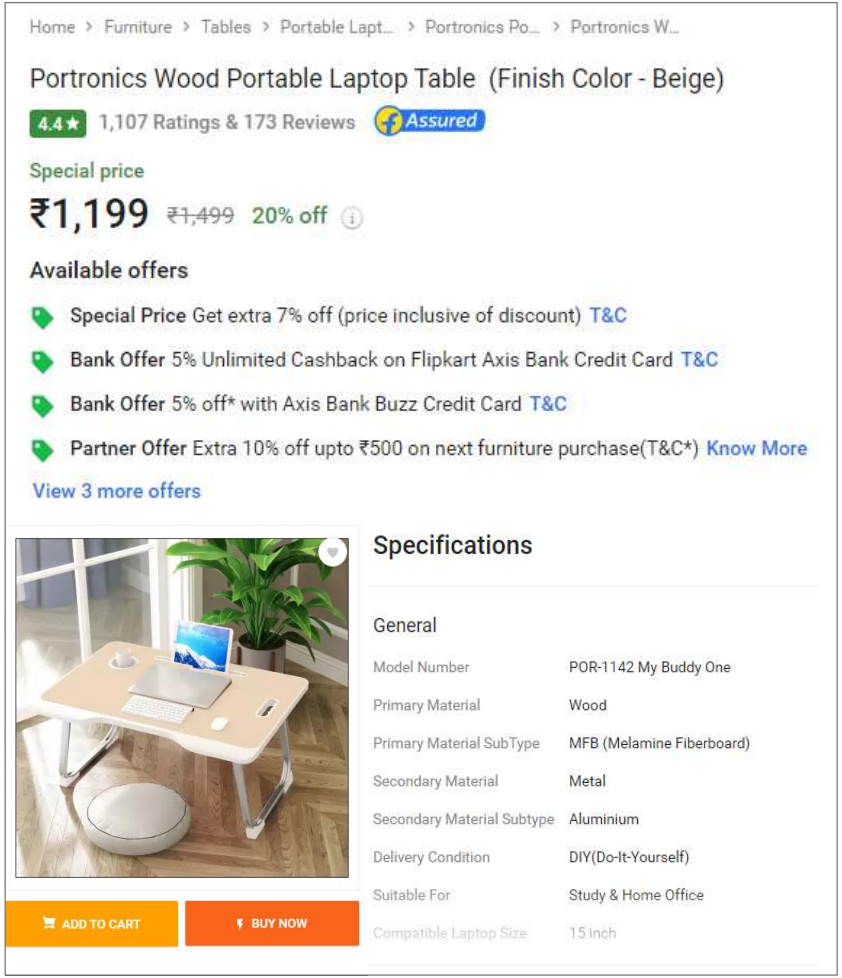
22 Janvier 2022

MBIADOU SALEU G. Raïssa

PLAN

01

CONTEXTE ET JEU DE DONNÉES



Contexte

- « *Place du marché* » : une plateforme d'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 : fastidieuse et peu fiable
- QUESTION : Est-il possible de classifier automatiquement articles en différentes catégories, avec un niveau de précision suffisant ?

CONTEXTE

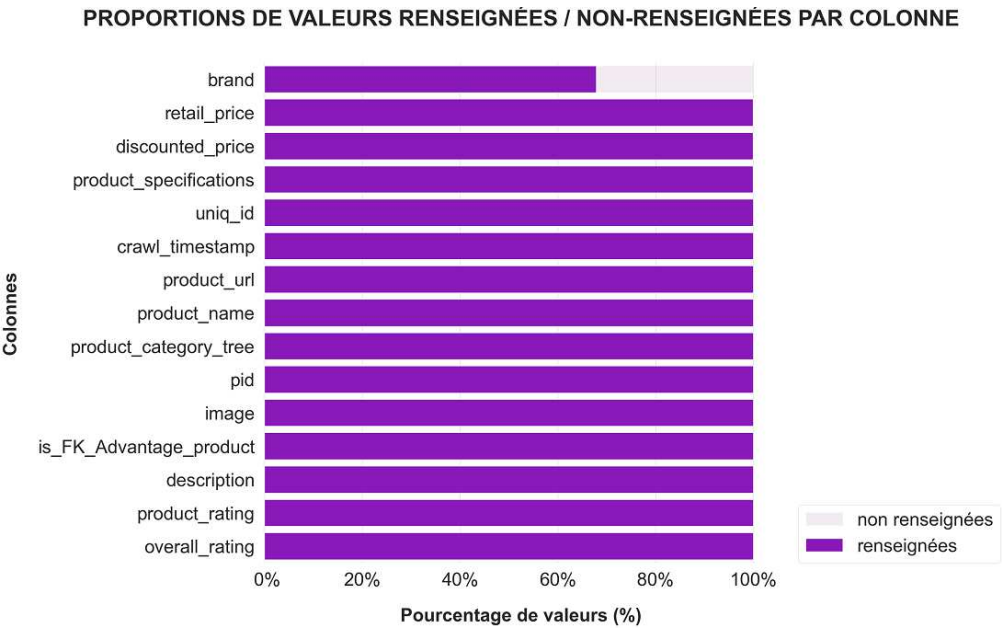
MISSION

- 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

OBJECTIFS

- Améliorer l'expérience des utilisateurs
- Fiabiliser la catégorisation des articles

	Nom du fichier	Nb de lignes	Nb de colonnes	Description
1	sample e-commerce	1050	15	Produits



LES DONNÉES

- Données issues de la base FlipKart
- 1050 produits
- 15 indicateurs couvrant plusieurs types d'informations:
  - Informations produits
  - Informations tarifaires
  - Notes produits
  - Images produits
- Colonnes très bien renseignées

Données visuelles

Données textuelles

product_name

Printland PMR1902 Ceramic Mug

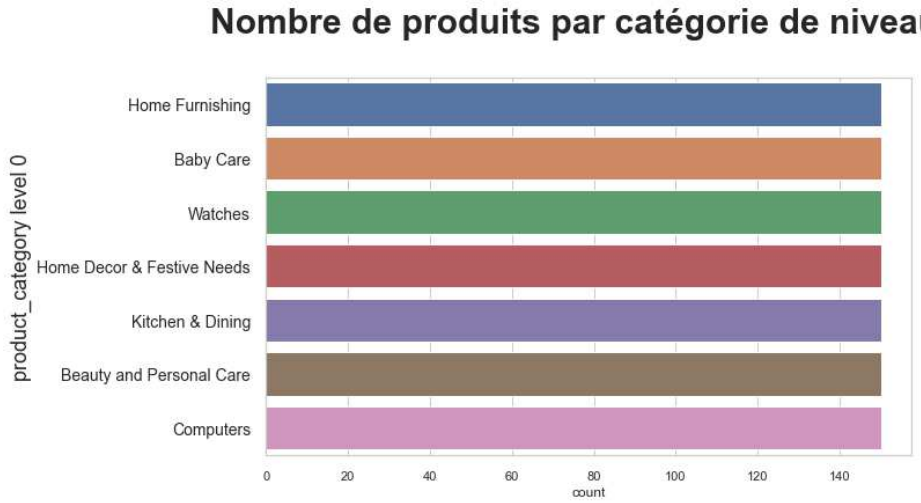
description

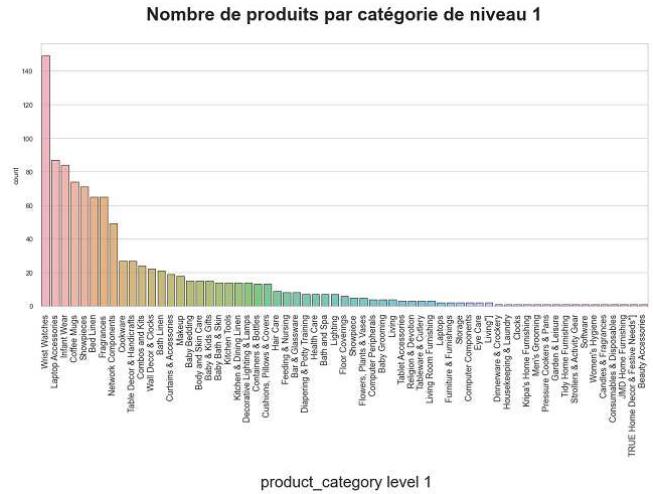
Printland PMR1902 Ceramic Mug (350 ml) Price: Rs. 299 Printland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one. Printland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one.

LES DONNÉES

Ce qui nous intéresse:

- Données textuelles : nom, description, marque et catégorie des produits
- Données visuelles : image des produits (ce sont des images isolées sur fond blanc avec des résolutions et ratio d'aspect variables)

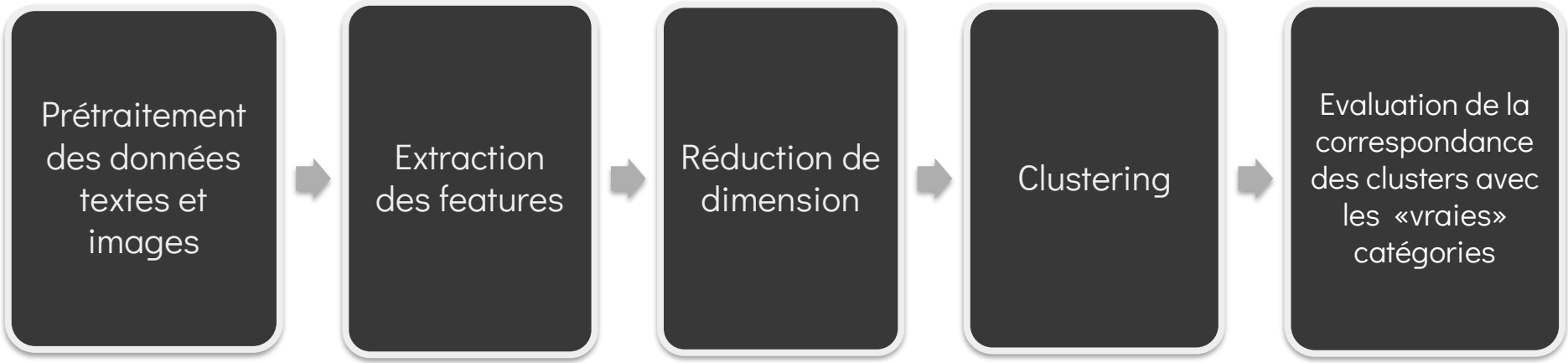




LES CATÉGORIES PRODUIT

- Organisées sous forme d'arbre (sur 6 niveaux)
- 7 catégories de niveau 0 contenant chacune 150 produits
- Nous nous focaliseront uniquement sur les catégories produit de niveau 0

DÉMARCHE


Prétraitement des données textes et images
Extraction des features
Réduction de dimension
Clustering
Evaluation de la correspondance des clusters avec les «vraies» catégories

02

PRÉTRAITEMENT DES DONNÉES ET EXTRACTION DES FEATURES

Environnement de développement

Anaconda

Installation d'Anaconda: plateforme de disbution python la plus populaire

environement virtuel

Mise en place d'un environnement virtuel dédié au projet

Installation des paquets

Installation des paquets nécessaires (numpy, pandas,matplotlib, seaborn, sklearn, scipy, nltk, pillow, opencv-python, etc...) avec la commande **pip install**

Prétraitement des données

Avant Prétraitement

Printland PMR1902 Ceramic Mug Printland PMR1902 Ceramic Mug (350 ml) Price: Rs. 299 Printland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one. Printland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one.

Après Prétraitement

printland pmr1902 ceramic mug printland pmr1902 ceramic mug 350 ml price r 299 printland coffee mug adorable fantastic coffee mug one enjoy morning coffeetea huge mug made ceramic material perfect add kitchen wardrobe look stylish elegant serve teacoffee mug casual get together home also perfect gift presented loved one printland coffee mug adorable fantastic coffee mug one enjoy morning coffeetea huge mug made ceramic material perfect add kitchen wardrobe look stylish elegant serve teacoffee mug casual get together home also perfect gift presented loved one

PRÉTRAITEMENT DES TEXTES

- Utilisation de la librairie NLTK
- 5 étapes :
  - Suppression des ponctuations
  - Tokenisation: découpage en mots
  - Suppression des tokens de taille < 2
  - Suppression des stopwords: les mots très courants qui n'apportent pas de valeur informative pour la compréhension du sens
  - Lemmatisation: réduction des mots à leur lemmes (forme canonique)

FRÉQUENCE DES MOTS/CATÉGORIE

- Wordcloud des 30 mots les plus fréquents par catégorie produit de niveau 0
- Les mots les plus utilisés par catégorie de niveau 0 sont majoritairement distincts

PRÉTRAITEMENT DES IMAGES

- Utilisation des librairies PIL (Pyhton Imaging Library) et OpenCV (Open Compute Vision)
- 5 étapes :

1. Correction de l'exposition (étirement d'histogramme)
2. Correction du contraste (égalisation d'histogramme)
3. Réduction du bruit (filtre)
4. Conversion en niveau de gris
5. Redimensionnement (en 224 * 224)

Image originale

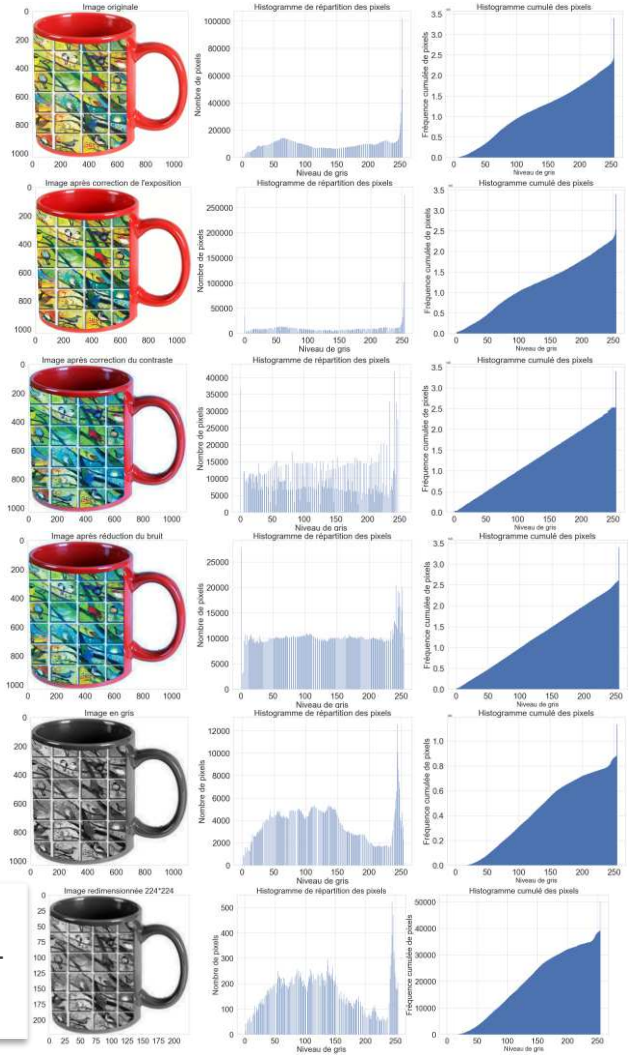
Image après correction de l'exposition

Image après correction du contraste

Image après réduction du bruit (filtre)

Image après conversion en niveau de gris

Image après redimensionnement (224*224)



Extraction des features

EXTRACTION DES FEATURES

Textes :

- Bag Of Words: compte le nombre de fois qu'un mot apparait dans un document
- TF-IDF (Term Frequency - Inverse Document Frequency): les fréquences des mots sont remplacés par des scores TF-IDF
- Word2Vec: technique par plongement de mot. Chaque mot est représenté par un vecteur

Images :

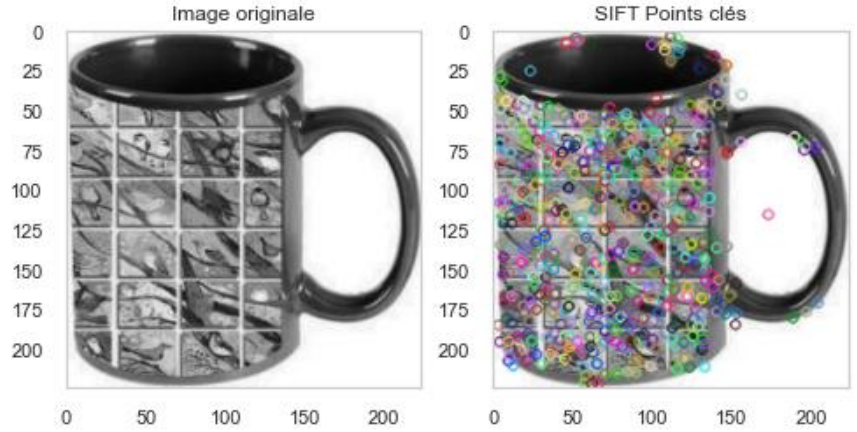
- SIFT (Scale Invariant Feature Transform): permet d' identifier les éléments similaires entre différentes images
- Transfer Learning (VGG-16): réseau de neurones convolutif pré-entrainé sur ImageNet

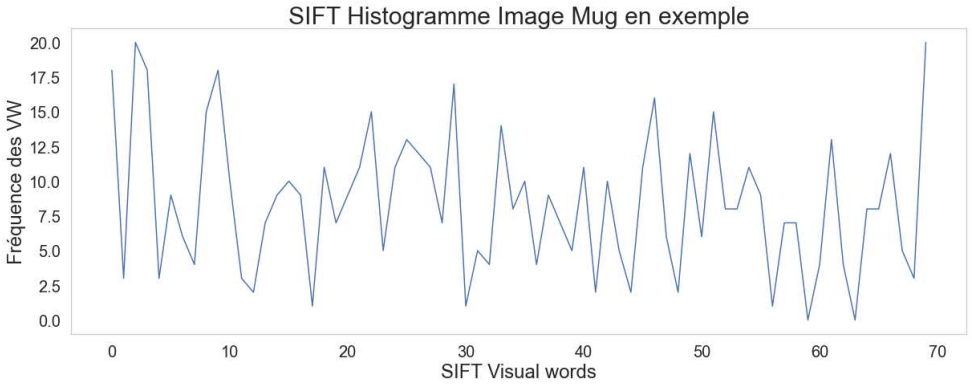
SIFT (3 ÉTAPES)

Récupération des descripteurs de l'image

Clustering de l'ensemble des descripteurs et identification des centres (utilisés comme vocabulaire du dictionnaire visuel)

Construction de l'histogramme de l'image (Bag of Visual Words)





VGG-16 (TRANSFER LEARNING)

Architecture de VGG-16

- VGG-16 est une version du réseau de neurones convolutif VGG-Net.
- 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
- Utilisation du modèle VGG16 pré-entraîné sur ImageNet

03

Réduction de dimension

Techniques utilisées

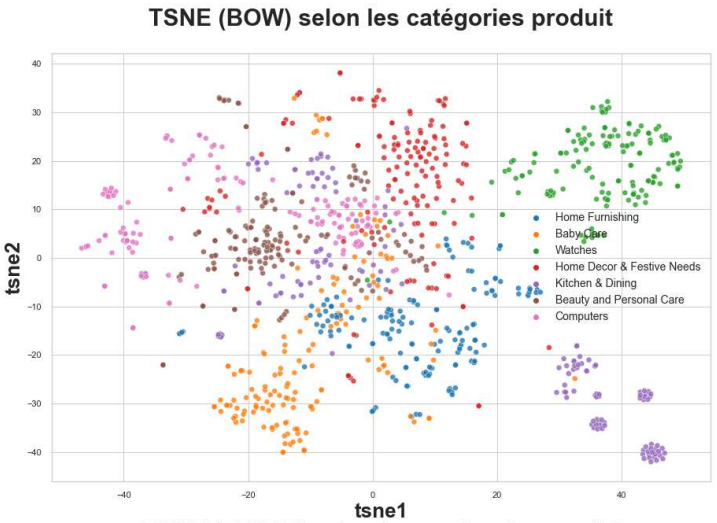
ACP (PCA en anglais)

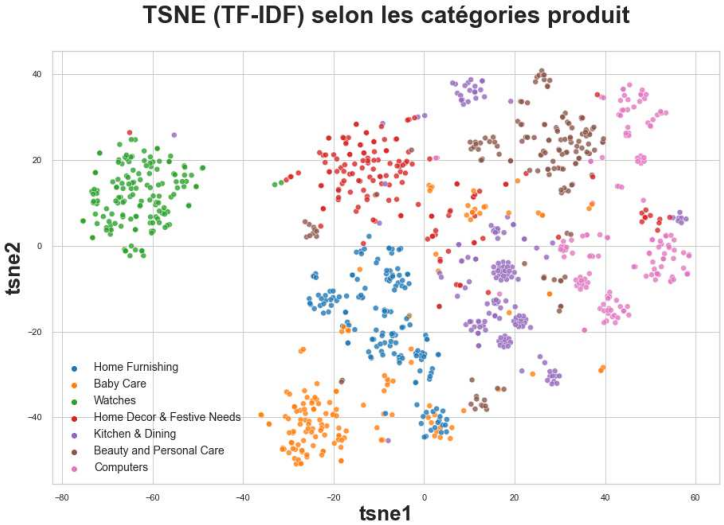
L'**A**nalyse en **C**omposantes **P**rincipales est une méthode largement utilisée en réduction de dimension qui cherche à représenter les données dans un hyperplan proche de sorte à **conserver au maximum la variance** du nuage de données. En d'autres termes, il s'agit de représenter les données dans un sous-espace de plus petite dimension maximisant l'inertie totale du nuage projeté dans cet espace.

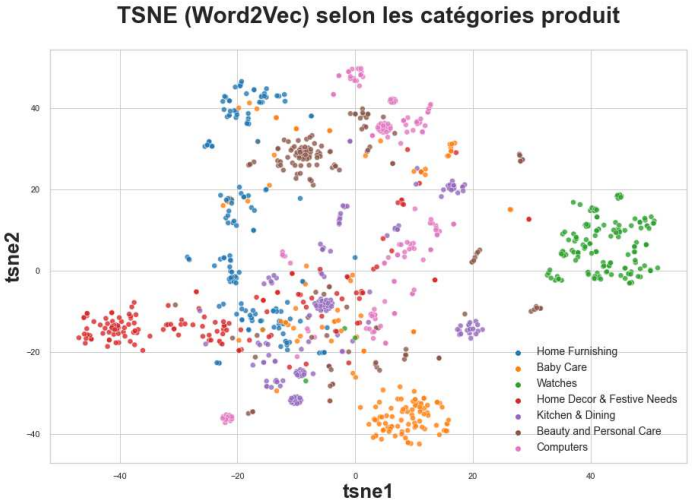
t-SNE

t-SNE, pour "t-distributed Stochastic Neighbor Embedding" est une technique qui permet de visualiser des données de (très) grandes dimensions, en effectuant un plongement (*embedding* en anglais) dans une variété de plus petite dimension, généralement 2 ou 3 pour pouvoir repérer des caractéristiques intéressantes du phénomène à modéliser.

PCA+T-SNE (TEXTES)







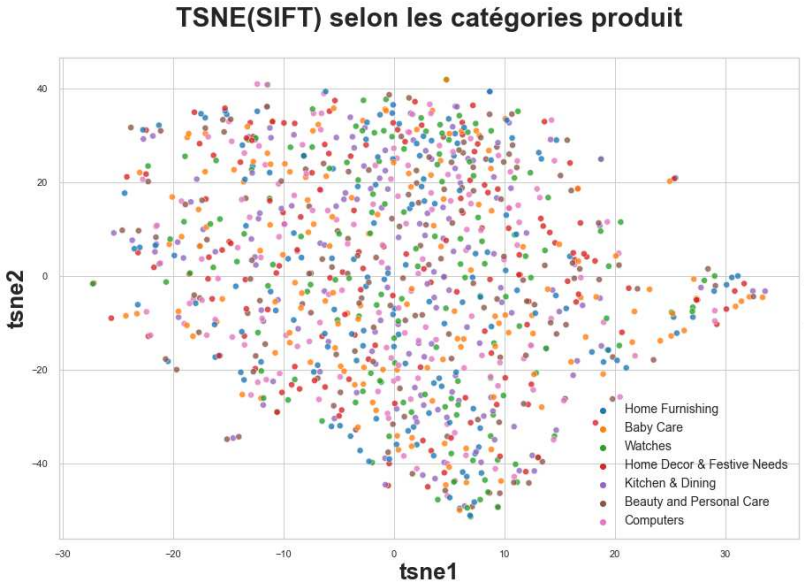
PCA (99% de la variance des données)

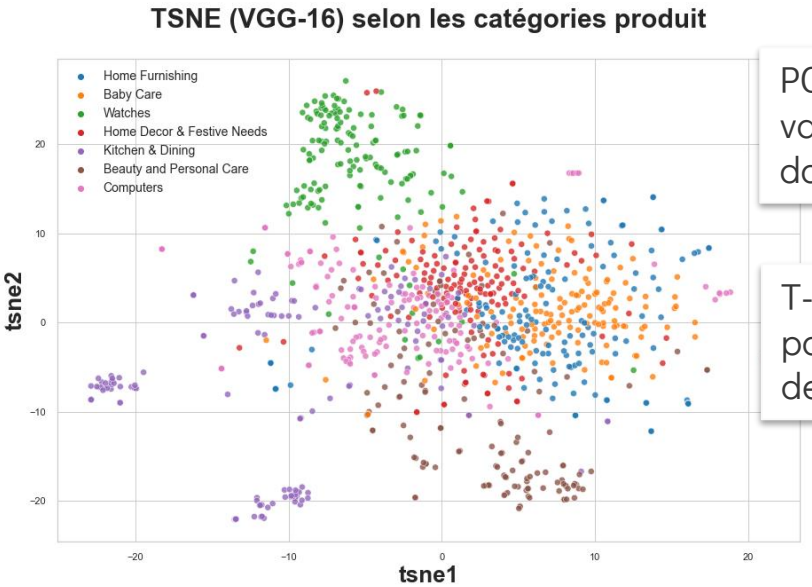
T-SNE 2D initialisé par les résultats de la PCA

➢On observe une sorte de regroupement des points appartenant à la même catégorie produit (surtout à partir des données TF-IDF et Word2Vec)

➢Avec TF-IDF et Word2Vec, les points appartenant à la catégorie *Watches* sont très bien regroupés et le groupe est bien séparé des autres groupes.

PCA+T-SNE (IMAGES)





PCA (99% de la variance des données)

T-SNE 2D initialisé par les résultats de la PCA

Le regroupement par catégories produits semble meilleur avec les features obtenus par VGG-16 qu'avec les features obtenus par SIFT

04

Clustering

ALGORITHME ET METRIQUES D'EVALUATION UTILISÉS

K-Means

K-means (k-moyennes) est un algorithme non supervisé de **clustering**, populaire en Machine Learning.
Il est « basé sur la distance ». Il permet de regrouper les observations du data set en **K** clusters distincts. Ainsi les données similaires (proches en terme de distance) se retrouveront dans un même cluster.

K = 7

Coefficient de silhouette

C'est une mesure utilisée pour calculer la qualité d'un clustering. Sa valeur est comprise entre -1 et 1

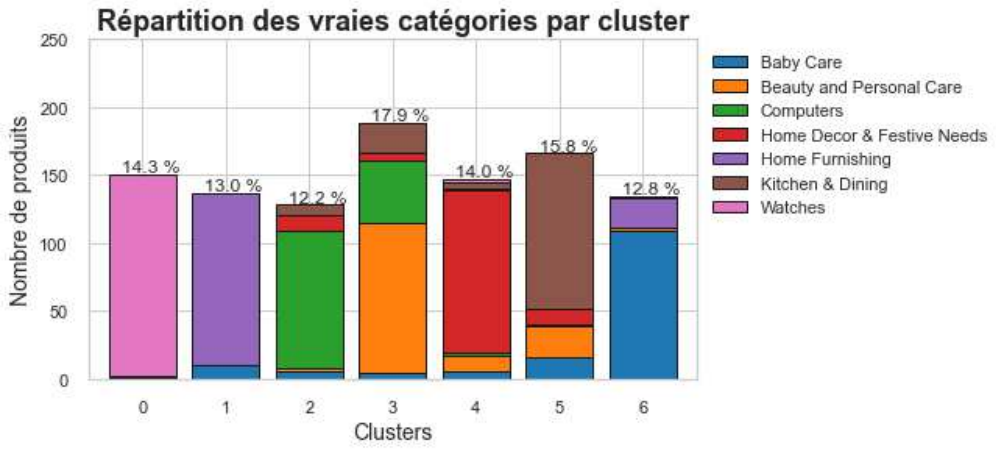
Score ARI

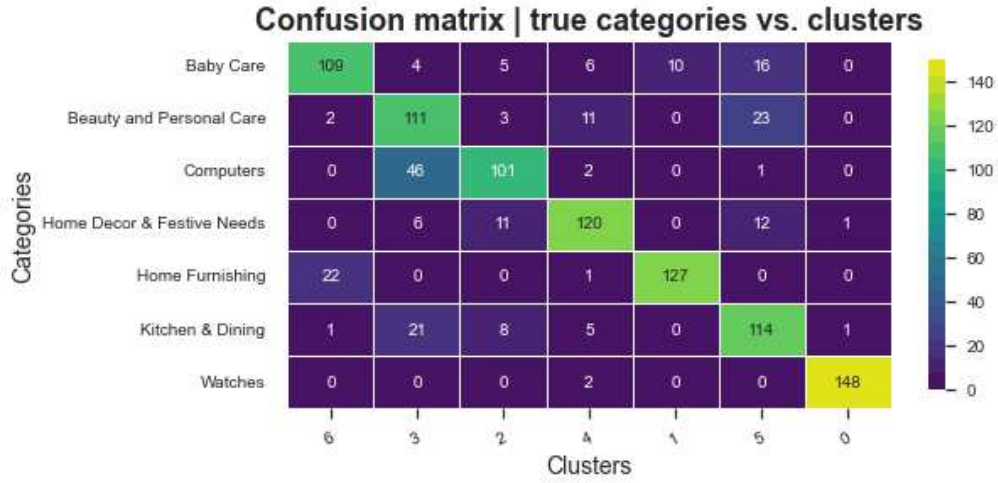
L'index de Rand Ajusté permet de comparer la similarité entre deux listes de labels (issues de deux partitionnements) [-1, 1]

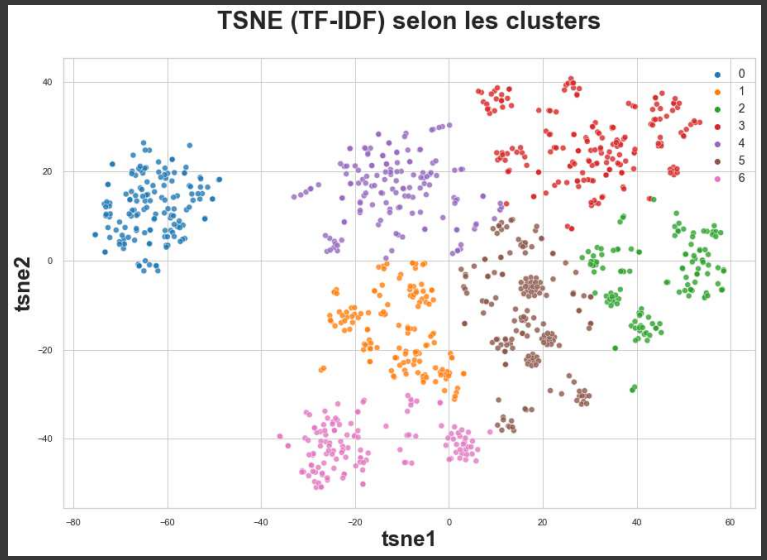
Données textes

Clustering (TF-IDF)

- Silhouette = 0.48
- ARI = 0.59
- 220 erreurs de catégorisation

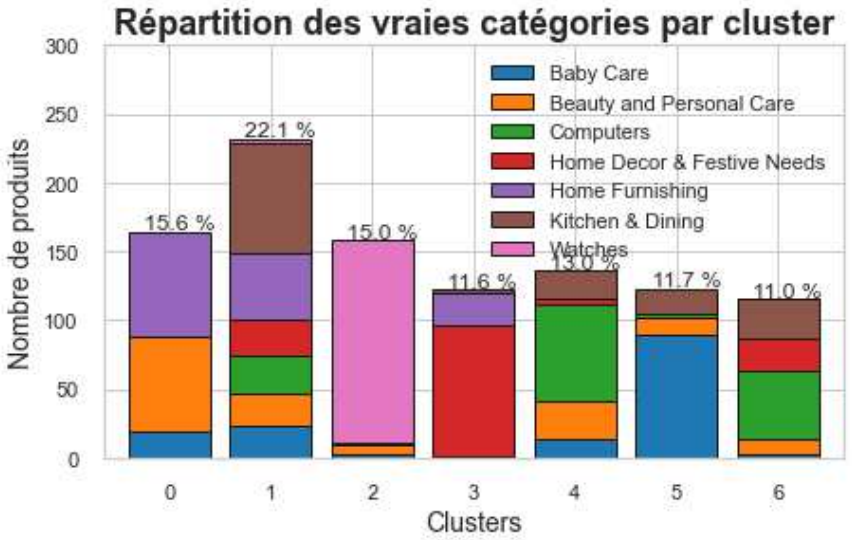


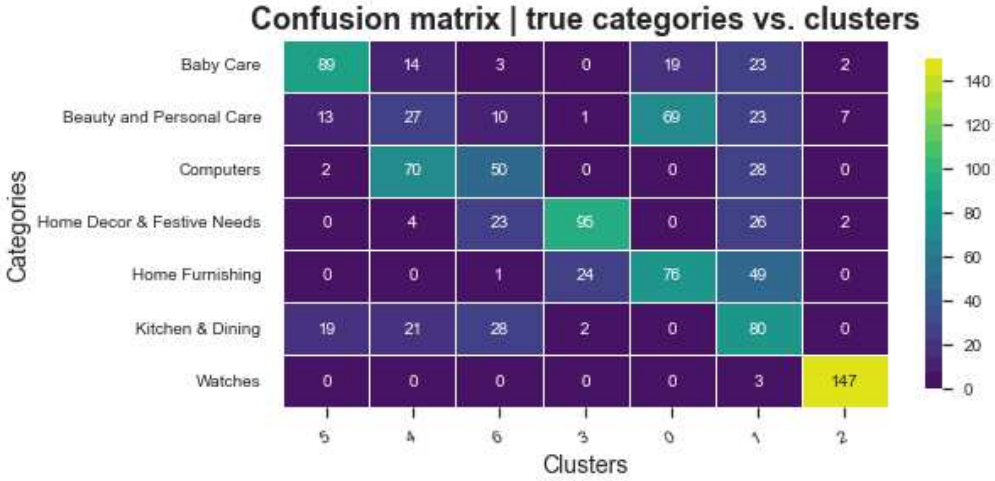


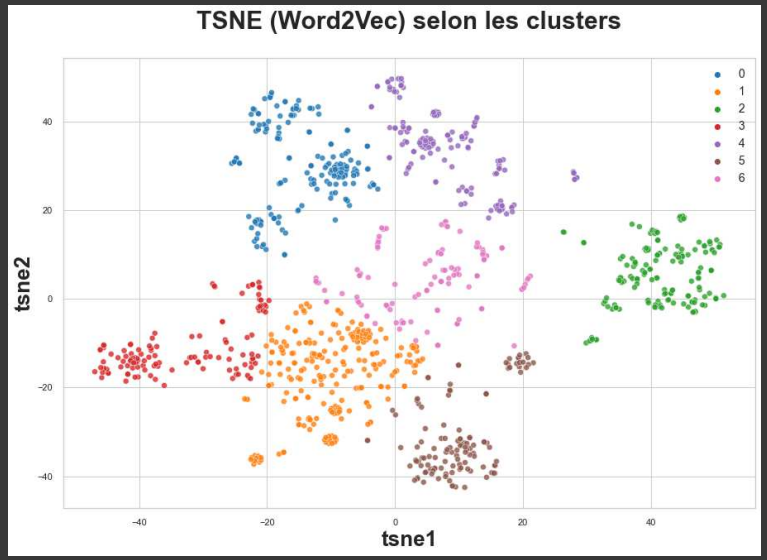


CLUSTERING (W2V)

- Silhouette = 0.45
- ARI = 0.34
- 485 erreurs de catégorisation



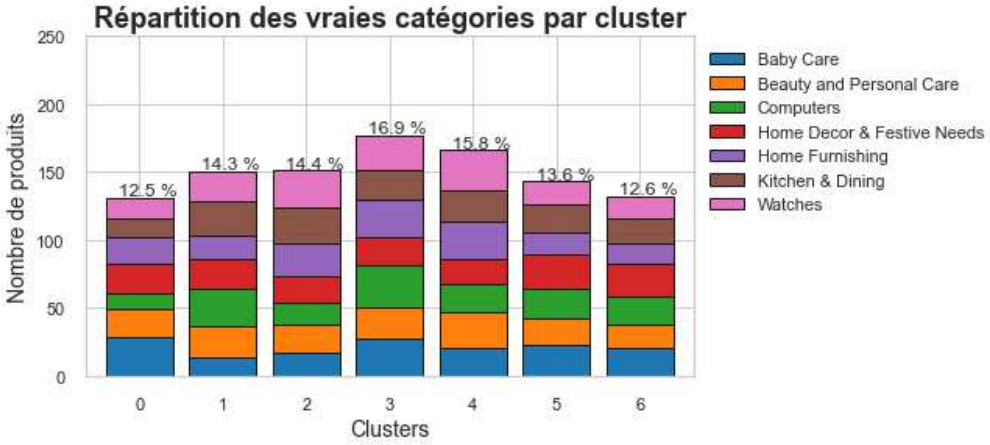


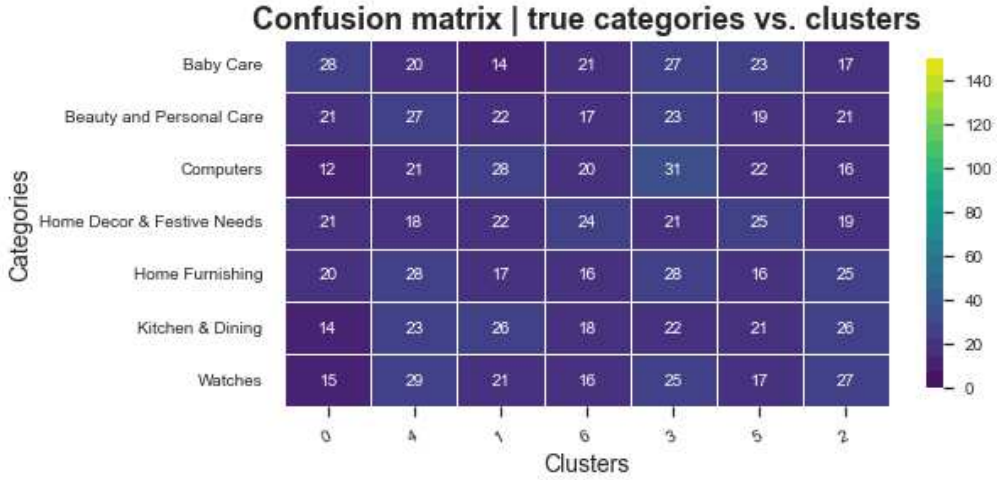


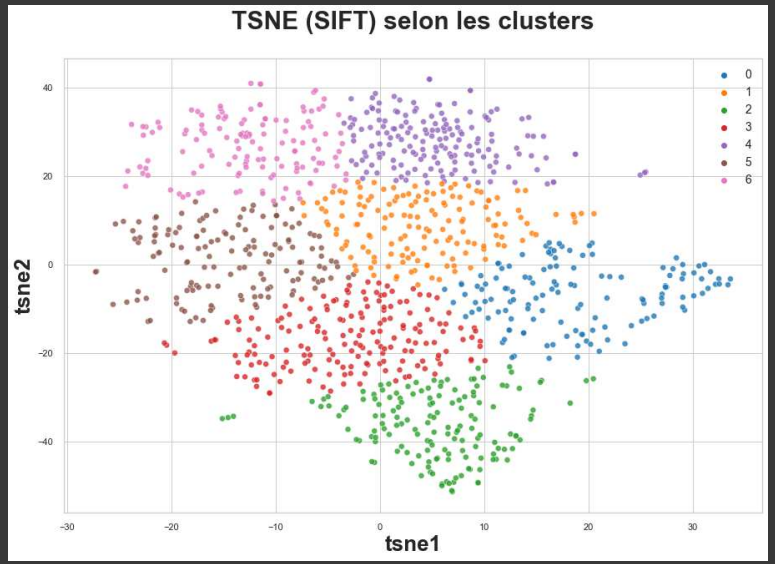
Données images

CLUSTERING (SIFT)

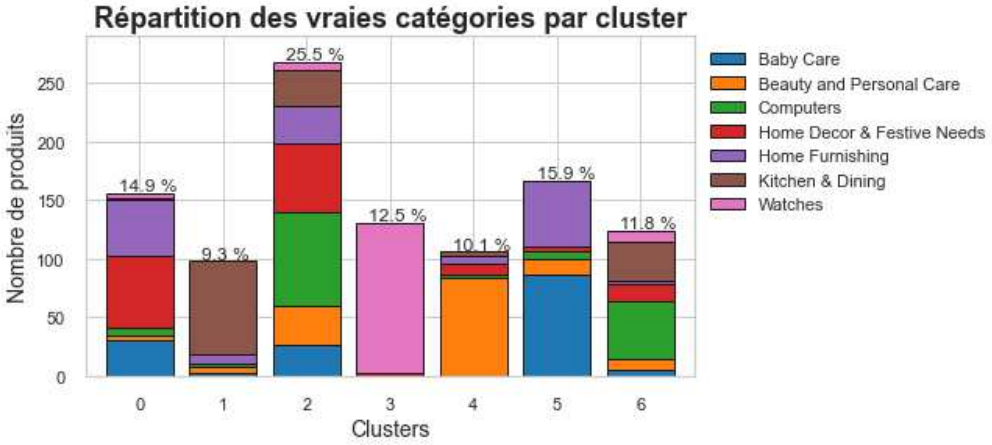
- Silhouette = 0.37
- ARI = -0.00019
- 867 erreurs de catégorisation

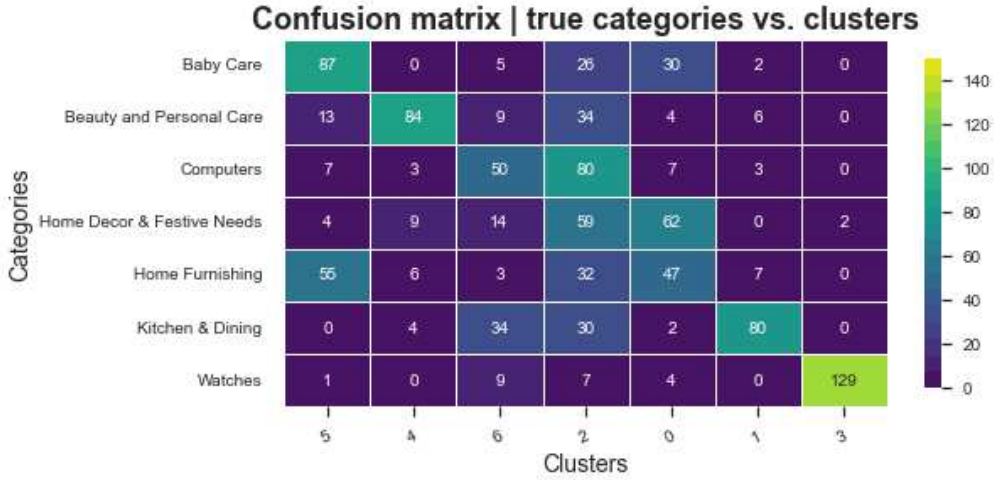




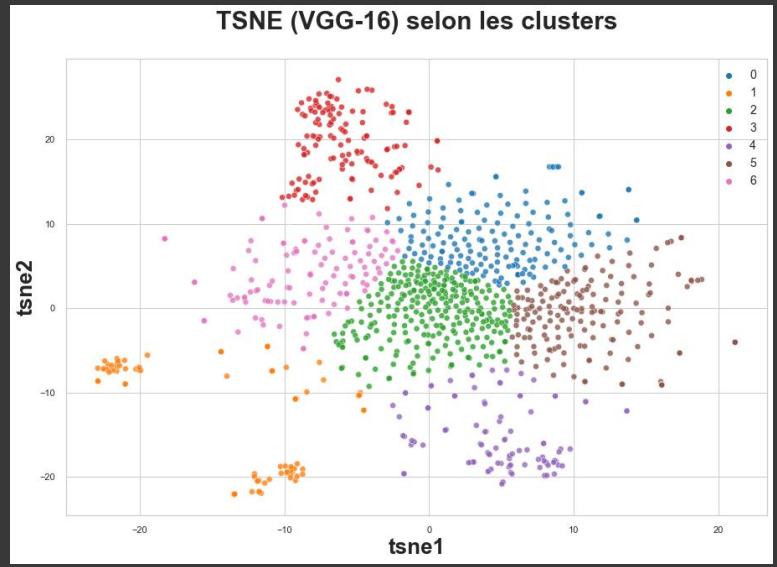


Clustering (VGG-16)





- Silhouette = 0.38
- ARI = 0.28
- 514 erreurs de catégorisation



05

Conclusion

Conclusion

- Analyse des données textuelles et visuelles

- Extraction de *features* en utilisant des techniques adaptées
  - **Textes (NLTK)** : BoW, TF-IDF, Word2Vec
  - **Images (PIL, Open-CV)** : SIFT, Transfer Learning

- Réduction de dimension
  - PCA (99% de variance des données)
  - T-SNE (2 Dimensions)

- Clustering
  - KMeans
  - 7 clusters

Conclusion sur la faisabilité du moteur de classification

FEATURES	ARI	SILHOUETTE	NB ERREURS
TF-IDF	0.59	0.48	220
Word2Vec	0.34	0.45	485
VGG-16	0.28	0.38	514
SIFT	-0.00019	0.37	867

- ✓ L'utilisation des données texte (features TF-IDF) permet une meilleur catégorisation des produits par rapport aux données image.
- ✓ La classification avec les données images est améliorée en utilisant un algorithme de type CNN (Transfer Learning).

Nous validons la faisabilité de la mise en œuvre d'un moteur de classification automatique des produits

Perspectives (recommandations)

- Enrichir la base de données produits
- Des produits équilibrés sur les sous-catégories devraient permettre d'avoir un modèle plus fin
- Combiner les modèles image + texte
- Pour guider la classification non-supervisée (produits ambigus), on pourrait envisager de choisir des mots-clés (*tags*) qui permettraient de faciliter le clustering

ANNEXE

EXPLORATION DES TOPICS

Mise en lumière des sujets abordés dans les textes.

LDA
(Latent Dirichlet Allocation)

NMF
(Negative Matrix Factorization)

Est-il possible d'utiliser ces méthodes pour faire ressortir les catégories ?

LDA (BOW)

Topic	% Home Furnishing	% Baby Care	% Watches	% Home Decor & Festive Needs	% Kitchen & Dining	% Beauty and Personal Care	% Computers
0.0	1.111111	52.222222	7.222222	3.888889	5.555556	13.888889	16.111111
1.0	25.000000	6.250000	0.000000	25.000000	31.250000	6.250000	6.250000
2.0	0.000000	0.000000	0.000000	0.000000	0.000000	85.714286	14.285714
3.0	0.000000	7.692308	23.076923	0.000000	15.384615	17.948718	35.897436
4.0	32.773109	7.983193	0.000000	24.369748	29.411765	3.361345	2.100840
5.0	12.406015	5.827068	24.060150	14.661654	8.834586	18.796992	15.413534
6.0	0.000000	5.263158	0.000000	7.894737	31.578947	7.894737	47.368421

NMF (BOW)

Topic	% Home Furnishing	% Baby Care	% Watches	% Home Decor & Festive Needs	% Kitchen & Dining	% Beauty and Personal Care	% Computers
0.0	29.302326	9.302326	0.465116	31.162791	15.813953	8.837209	5.116279
1.0	0.000000	1.388889	0.000000	0.000000	98.611111	0.000000	0.000000
2.0	17.452830	8.254717	0.000000	18.396226	9.198113	24.528302	22.169811
3.0	0.000000	0.000000	0.000000	3.030303	6.060606	9.090909	81.818182
4.0	9.615385	86.538462	0.000000	0.000000	0.000000	3.846154	0.000000
5.0	6.000000	8.000000	0.000000	8.000000	8.000000	36.000000	34.000000
6.0	0.000000	0.000000	98.026316	0.000000	0.000000	1.315789	0.657895

Topics vs Catégories

Quelques associations possibles :

LDA

- topic #0 : Baby Care (52%)
- topic #2 : Beauty and Personal Care (86%)
- topic #6 : Computers (47%)

NMF

- topic #1 : Kitchen & Dining (99%)
- topic #3 : Computers (82%)
- topic #4 : Baby Care (87%)
- topic #6 : Watches (98%)

LDA (TF-IDF)

Topic	% Home Furnishing	% Baby Care	% Watches	% Home Decor & Festive Needs	% Kitchen & Dining	% Beauty and Personal Care	% Computers
0.0	1.680672	7.563025	60.924370	2.941176	8.403361	14.705882	3.781513
1.0	0.000000	0.000000	0.000000	50.000000	0.000000	50.000000	0.000000
2.0	0.000000	0.000000	0.000000	4.347826	86.956522	4.347826	4.347826
3.0	16.300940	4.702194	0.000000	12.852665	12.852665	22.570533	30.721003
4.0	22.488038	27.990431	1.196172	22.966507	8.373206	7.894737	9.090909
5.0	0.000000	0.000000	0.000000	30.000000	10.000000	20.000000	40.000000
6.0	0.000000	0.000000	0.000000	0.000000	86.842105	13.157895	0.000000

NMF (TF-IDF)

Topic	% Home Furnishing	% Baby Care	% Watches	% Home Decor & Festive Needs	% Kitchen & Dining	% Beauty and Personal Care	% Computers
0.0	0.000000	1.104972	82.320442	3.867403	4.972376	4.972376	2.762431
1.0	0.000000	3.076923	0.000000	10.769231	86.153846	0.000000	0.000000
2.0	23.762376	51.980198	0.000000	11.881188	5.940594	5.940594	0.495050
3.0	5.586592	7.821229	0.000000	22.905028	11.452514	31.843575	20.391061
4.0	0.000000	0.000000	0.000000	25.581395	69.767442	4.651163	0.000000
5.0	81.000000	9.000000	0.000000	9.000000	0.000000	1.000000	0.000000
6.0	0.990099	3.960396	0.990099	9.900990	1.980198	11.881188	70.297030

Topics vs Catégories

Quelques associations possibles :

LDA

- topic #0 : Watches (61%)
- topic #2 : Kitchen & Dining (87%)
- topic #5 : Computers (40%)
- topic #6 : Kitchen & Dining (87%)

NMF

- topic 0 : Watches (82%)
- topic 1 : Kitchen & Dining (86%)
- topic 2 : Baby Care (52%)
- topic 4 : Kitchen & Dining (70%)
- topic 5 : Home Furnishing (81%)
- topic 6 : Computers (70%)

MERCI

Des questions ?