

Classifier automatiquement des biens de consommation

Data & Analytics
Eric Blanvillain - 25-01-2022



Problématique

Mon rôle :

- La mission est de **réaliser une première 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.
- Il faut **analyser le jeu de données** en **réalisant un prétraitement** des images et des descriptions des produits, une **réduction de dimension**, puis un **clustering**. Les résultats du clustering seront présentés sous la forme d'une représentation en deux dimensions, qui illustrera le fait que les caractéristiques extraites permettent de regrouper des produits de même catégorie. Il faudra mettre en œuvre *a minima* un **algorithme de type SIFT / ORB / SURF**.
- Les représentations graphiques aideront à convaincre que cette approche de modélisation permettra bien de regrouper des produits de même catégorie.

Les points à aborder :

- Présentation de la base de donnée “Place de Marché” (image et texte)
- Nettoyage et exploration des catégories de produits sur “Place de Marché”
- Analyse des descriptions (vectorisation, topiques)
- Analyse des images, entraînement du modèle + VGC16
- Etude des prédictions (image et texte) - précision et erreurs

Compréhension des variables

Variables	Descriptions	Commentaires
uniq_id	Identifiant unique	Clé unique - ex : 59af3731b809a25f2bf99e99f645d8dd
crawl_timestamp	Horodatage	Format YYYY-MM-DD HH24:MI:SS +0000
product_url	URL d'accès au produit sur le site e-commerce Flickart	Clé unique
product_name	Nom du produit	Clé unique - Texte
product_category_tree	Arbre des catégories des produits	Notre cible - sous-catégories séparées par '>' - Texte - Données multiples
pid	Identifiant unique	Clé unique - ex : CAGEBTJBTNGGDZQZ
retail_price	Prix de consommation	prix en INR
discounted_price	Prix réduit	...
image	Image	Clé unique - Format : uniq_id.jpg ==> référence aux images du jeu de données
is_FK_Advantage_product	Produit avantageux FlipKart?	Booléen
description	Description du produit	Clé unique - Texte
product_rating	Évaluation du produit	Entre 1 et 5 avec 1 chiffre après la virgule ou 'No rating available'
overall_rating	Note moyenne globale d'évaluation	Entre 1 et 5 avec 1 chiffre après la virgule ou 'No rating available'
brand	Marque du produit	Valeurs manquantes - Texte
product_specifications	Spécification du produit	Propriétés techniques avec clé/valeur Données multiples

Caractéristiques des produits du Dataset



Caractéristiques :

Product_name = "Brillare Science Dandruff Control Shampoo & Intenso Creme Combo"

Brand = "Brillare Science"

Categorie_niveau_1 = "Beauty and Personal Care"

Categorie_niveau_2 = "Combos and Kits"

Categorie_niveau_3 = "Brillare Science Combos and Kits"

Description = "Specifications of Brillare Science Dandruff Control Shampoo..."

Retail_price = 450 INR



Caractéristiques :

Product_name = "Romex Ultimate Urban Analog Watch"

Brand = "Romex"

Categorie_niveau_1 = "Watches"

Categorie_niveau_2 = "Wrist Watches"

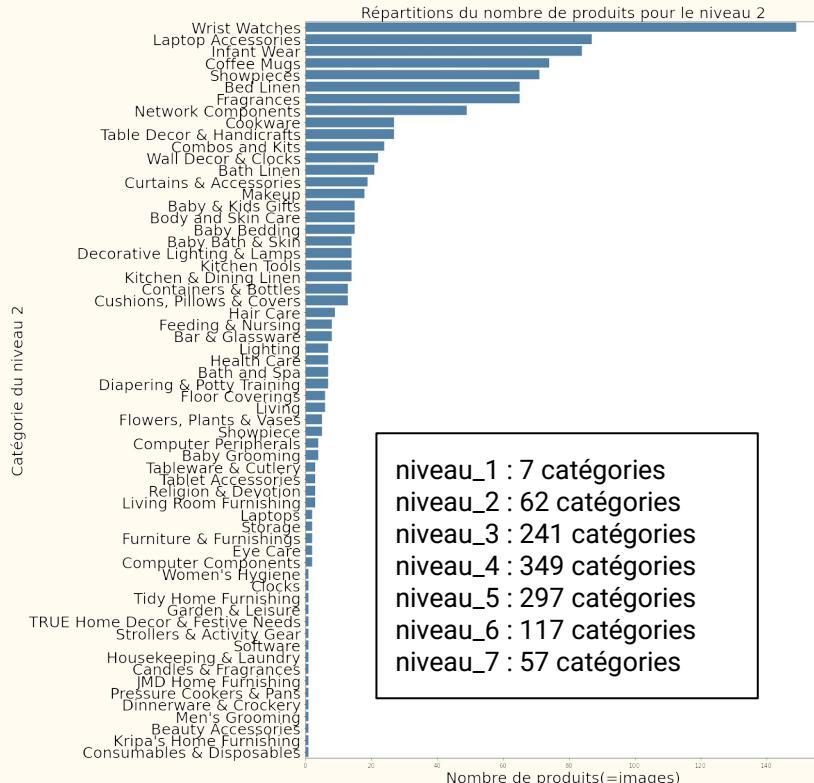
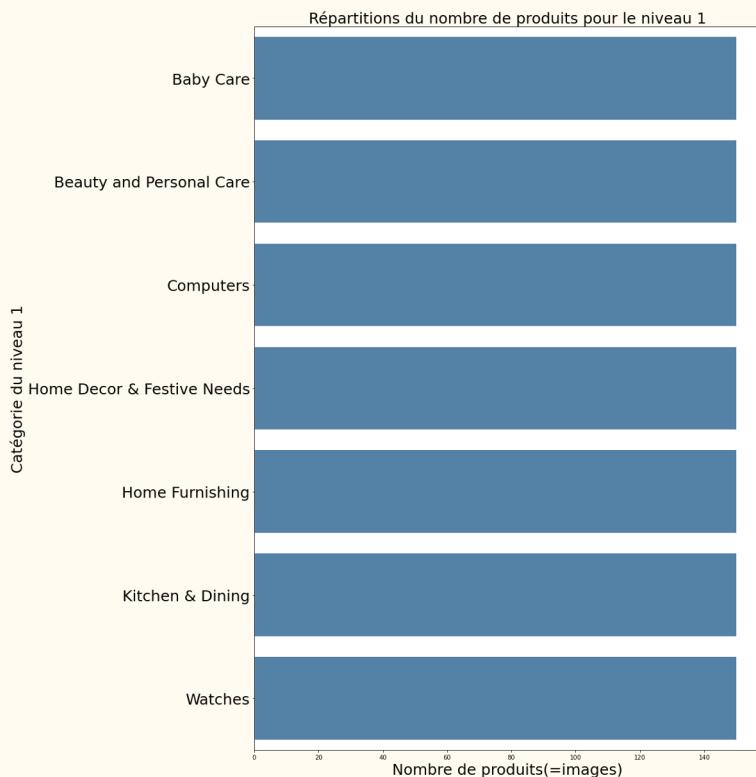
Categorie_niveau_3 = "Romex Wrist Watches"

Description = "Romex Ultimate Urban Analog Watch - For Boys, Men"

Retail_price = 1199 INR

Total = 1050 produits

Catégories les plus représentées (niveau 1 et 2)



niveau_1 : 7 catégories
niveau_2 : 62 catégories
niveau_3 : 241 catégories
niveau_4 : 349 catégories
niveau_5 : 297 catégories
niveau_6 : 117 catégories
niveau_7 : 57 catégories

PARTIE I

TRAITEMENT DU TEXT

Etape 1 - pré-traitement du texte (nltk)



Description pré-traitement :

Wallmantra Large Vinyl Sticker Sticker (Pack of 1) Price: Rs. 1,896 Bring home this exclusive Piece of Wall Art to give your home a refreshing look it deserves ! Wall Decals are the latest trend, sweeping the world of interior design, as a quick and easy way to personalise and transform your home. We at Wallmantra use only the highest quality premium self-adhesive vinyl for our products to ensure you receive the best quality product!

3 étapes de traitement :

1. Normalisation (lowercase, stopwords, ponctuation)
2. Tokenization ("série de mots clefs")
3. Lemmatisation / Racinisation (stemming)

Description post-traitement :

wallmantra large vinyl sticker sticker pack bring home exclusive piece wall art give home refreshing
look wall decal latest trend world interior design quick easy way transform home wallmantra use
highest quality premium self adhesive vinyl ensure best quality product

Etape 2 - feature extraction (word embedding / tf-idf)

Vectorisation (tf-idf) :



mots	wallmantra	sticker	wall	...	trend	decal	quick
tfidf	0.323178	0.323178	0.323178	...	0.036978	0.036242	0.033699

Une transformation tf-idf (term frequency-inverse document frequency) permet:

- de pondérer les fréquences d'apparition des termes par leur nombre d'occurrences dans l'ensemble des documents.
- La pondération tf-idf permet de contrebalancer l'importance d'un mot utilisé très fréquemment dans tous les documents du corpus par rapport aux termes plus spécifiques à certains documents.

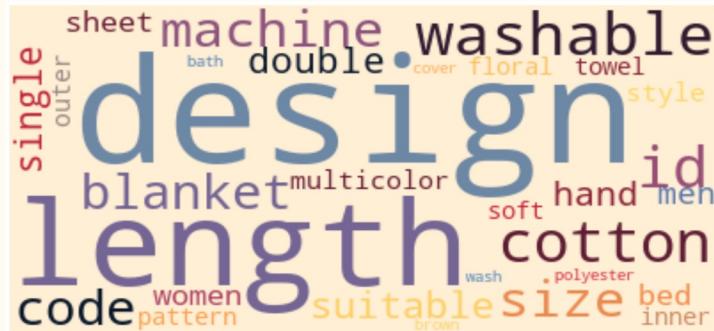
TF-IDF est un produit de deux parties :

- TF (Term Frequency) - Elle est définie comme le nombre de fois qu'un mot apparaît dans une phrase donnée.
- IDF (Inverse Document Frequency) - Il est défini comme le logarithme à la base e du nombre total de documents divisé par les documents dans lesquels le mot apparaît.

-> Cette transformation permet de produire des premiers éléments d'analyses, au premier titre duquel apparaissent les nuages de mots (wordcloud) qui représentent la fréquence relative des termes et qui donnent une première idée des significations contenues dans le corpus.

Etape 2 - Extraction des caractéristiques (word embedding)

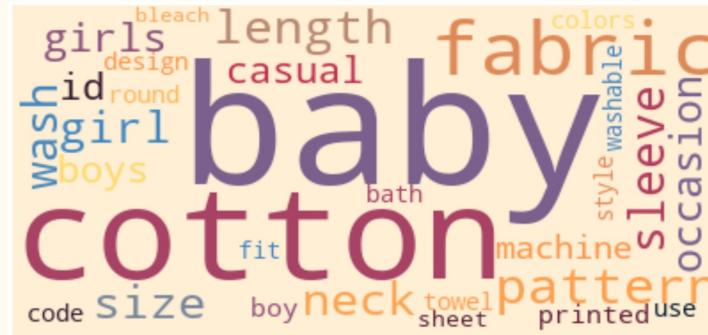
Mots les plus fréquents de la catégorie : Home Furnishing



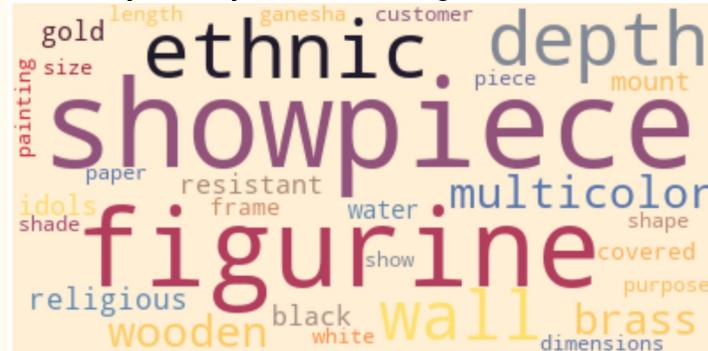
Mots les plus fréquents de la catégorie : Watches



Mots les plus fréquents de la catégorie : Baby Care



Mots les plus fréquents de la catégorie : Home Decor & Festive Needs



Etape 2 - Extraction des caractéristiques (topic LDA)



LDA topiques "abstraits" (Latent Dirichlet Allocation) : > non-supervisé

Topic #0: shipping cash genuine delivery flipkart buy guarantee replacement

Topic #1: rockmantra mug ceramic permanent stay thrilling ensuring porcelain crafting

Topic #2: sleeve detail boy 's fit shirt baby regular fabric casual

Topic #3: watch men analog perucci discount india great decker timewel

Topic #4: mug coffee printland perfect ceramic presented tea/coffee coffee/tea fantastic wardrobe

Topic #5: showpiece cm price online statue polyresin

Topic #6: quilt comforter single floral flipkart multicolor genuine cash shipping delivery

...

Topic #61: pyjama top girl set 's baby detail printed neck suit

topic_lda	cat_1	cat_1_count
8	Watches	96
	Home Furnishing	25
	Baby Care	3

topic_lda	cat_1	cat_1_count
9	Home Furnishing	8
	Watches	1

cat_1	topic_lda
Watches	8
	14
	16
	17
	9
	21
	22
	30
	35
	40

Etape 3 - Classification supervisé



Evaluation du “test set”:

	Multinomial Naive Bayes	SVC	Logistic Regression	SVC post LDA
Accuracy	0.55133	0.86312	0.74905	0.63498
Précision	0.17083	0.62084	0.38503	0.39972
Recall	0.16948	0.60191	0.37813	0.37573
F1 Score	0.14591	0.59015	0.36463	0.35996

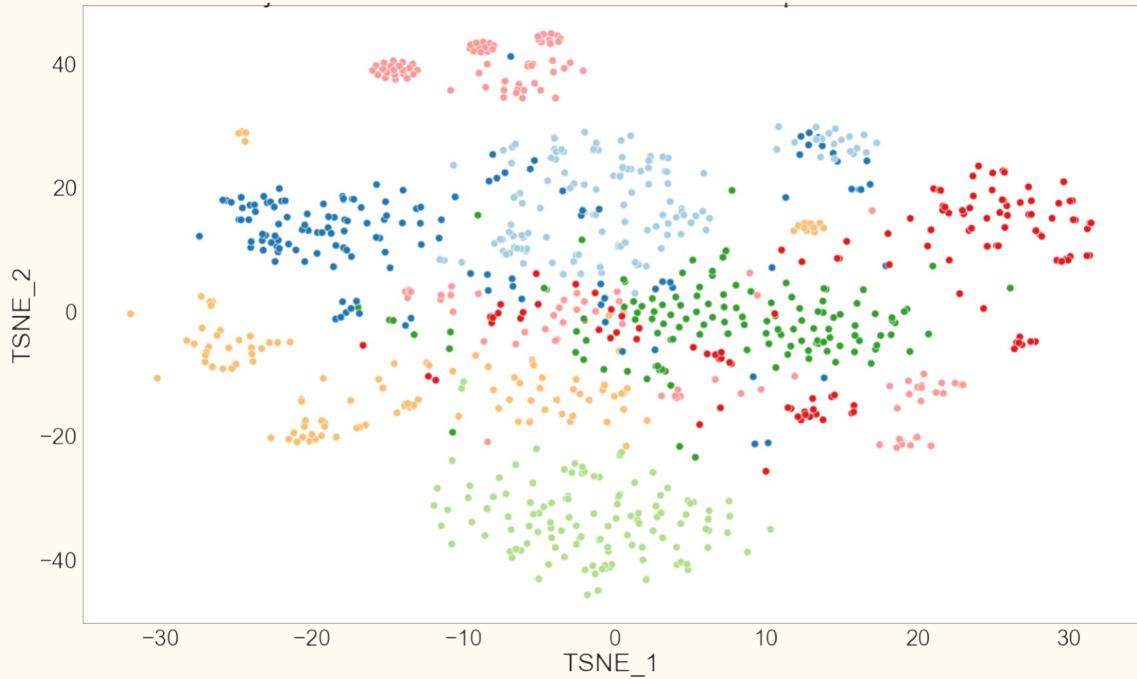
Conclusion :

=> Premiers résultats encourageants sur le classifieur SVC : 86 % d'accuracy sur le jeu de test, le jeu de données complet étant de seulement 1050 individus

train set = 787 produits
test set = 263 produits

Etape 4 - Evaluation des clusters

Projection t-SNE du clustering SVC sur tf-idf de Description :

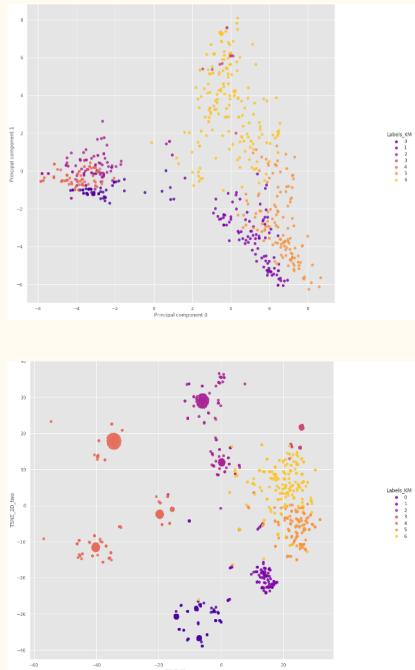


- Home Furnishing
- Baby Care
- Watches
- Home Decor & Festive Needs
- Kitchen & Dining
- Beauty and Personal Care
- Computers

		Matrice de confusion							
Vraie catégorie	Prédiction	Furnishing	Baby	Watches	Decor	Kitchens	Beauty	Computers	Accuracy %
		47	1	0	0	2	0	0	
Baby	Furnishing	4	36	0	0	5	3	0	75
	Baby	0	0	51	0	1	0	0	
Decor	Furnishing	5	0	0	27	12	0	0	61
	Decor	2	0	0	0	38	1	0	
Kitchen	Furnishing	1	1	0	0	9	25	0	69
	Kitchen	2	0	0	0	16	1	25	
Beauty	Furnishing	0	0	0	0	0	0	0	94
	Beauty	0	0	0	0	0	0	0	
Computers	Furnishing	0	0	0	0	0	0	0	98
	Computers	0	0	0	0	0	0	0	

Etape 5 - Classification non supervisé (Kmeans / GMM)

Classification Kmeans avec visualisation PCA / t-SNE



Classification GMM avec visualisation PCA / t-SNE



PCA

t-SNE

silhouette_score : 0.24157502799498745
ari_score : 0.2533991046239121

silhouette_score : 0.21443578605070604
ari_score : 0.18156841400946622

PARTIE II

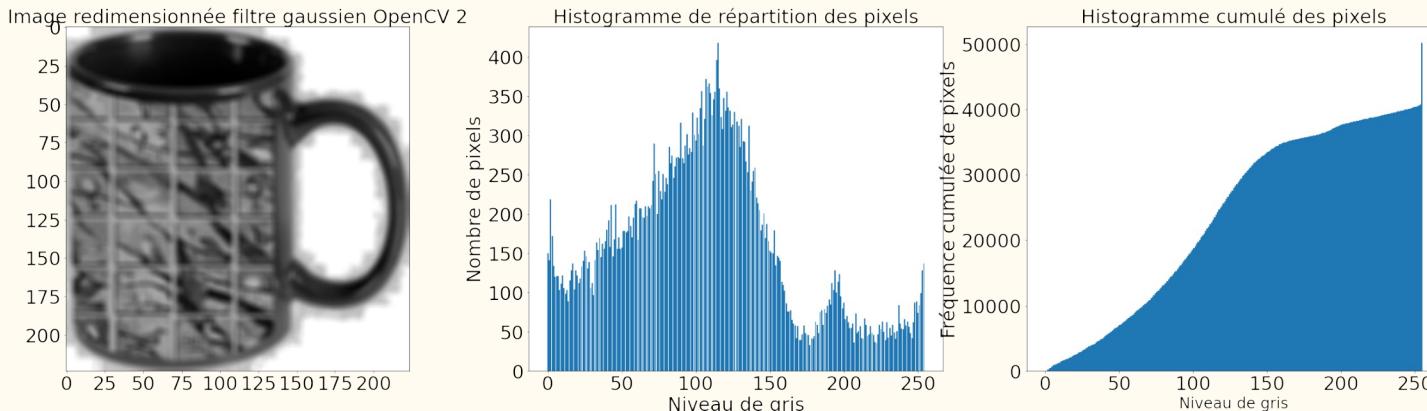
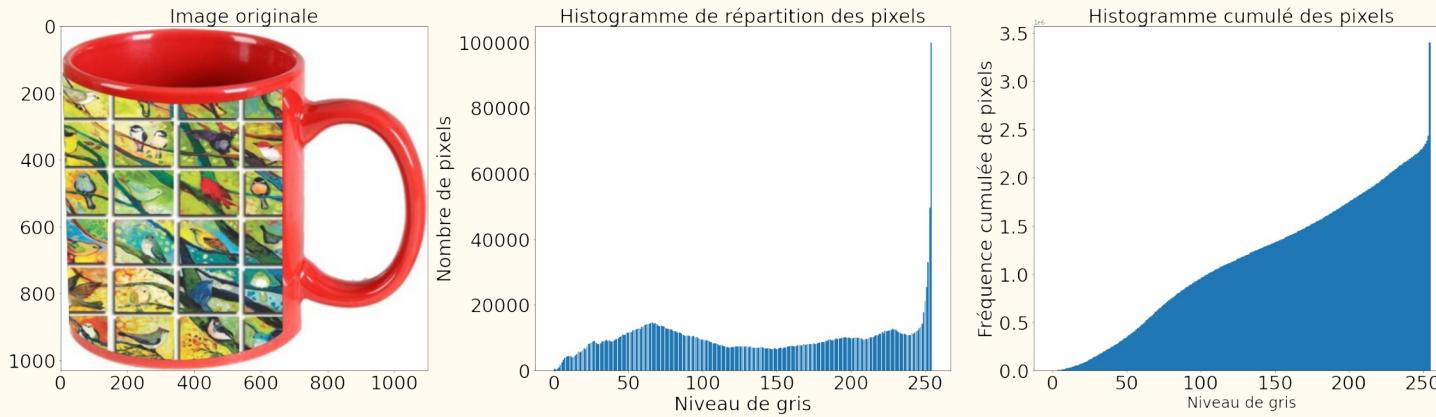
TRAITEMENT DES IMAGES

(SIFT / ORB)

Extrait d'images



Etape 1 - Fonction de prétraitement des images



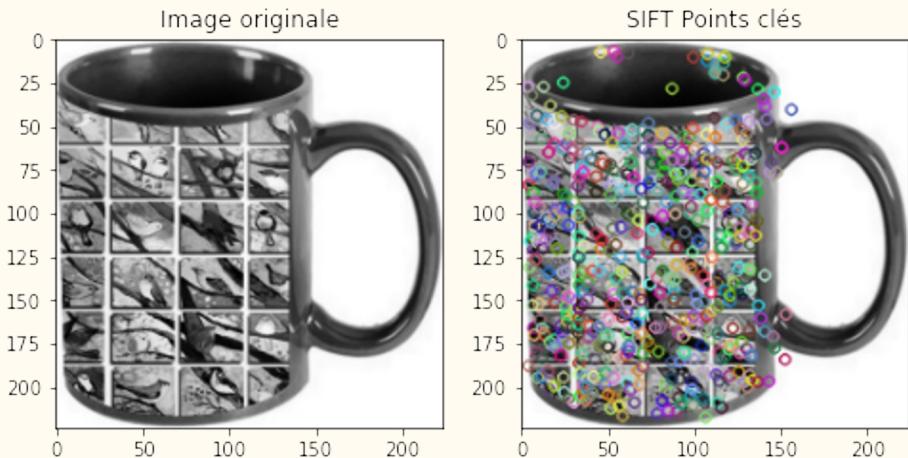
Etape 1 - Fonction de prétraitement des images



3 étapes de traitement :

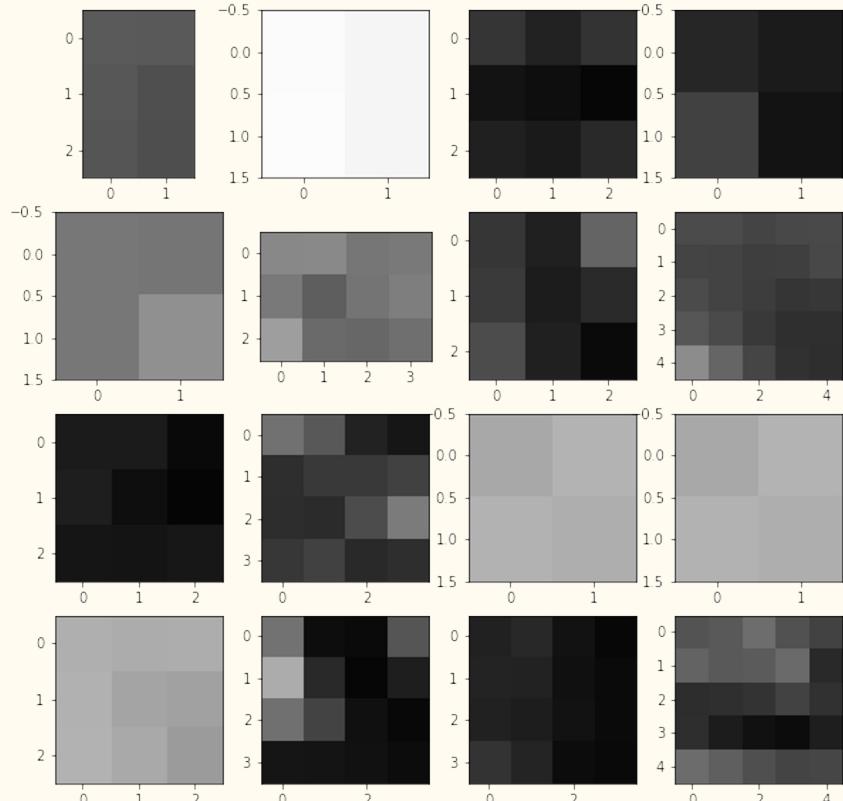
- filtre gris
- flou gaussien
- redimensionnement

Etape 2 - Extraction des caractéristiques (SIFT)

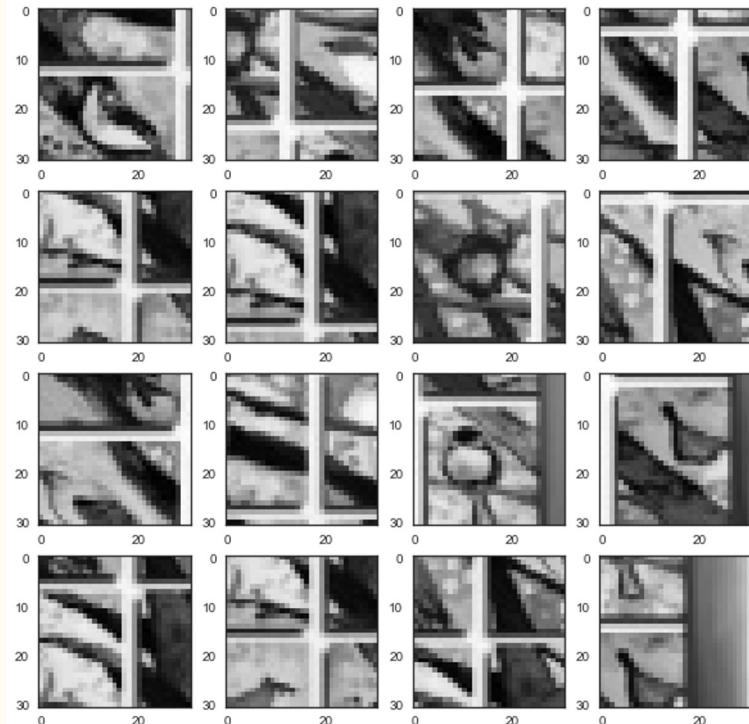
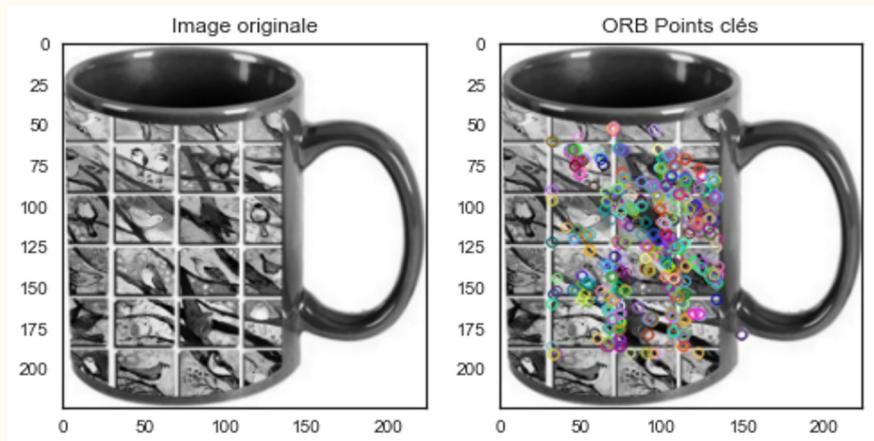


SIFT (Scale Invariant Feature Transform):

- Méthode, développée en 1999
- Permet d'extraire des features (ou points d'intérêt) de l'image et de calculer leurs descripteurs.
- L'algorithme SIFT se divise en plusieurs étapes :
 - Détection : création de l'espace des échelles, calcul des "DoG" (Différence of Gaussian), localisation des points d'intérêt.
 - Description : assignation d'orientation, création des descripteurs.



Etape 2 - Extraction des caractéristiques (ORB)

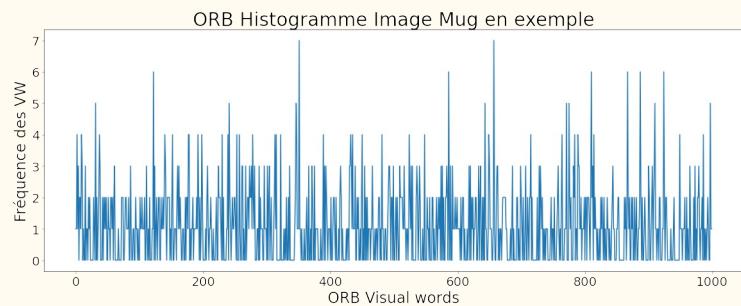
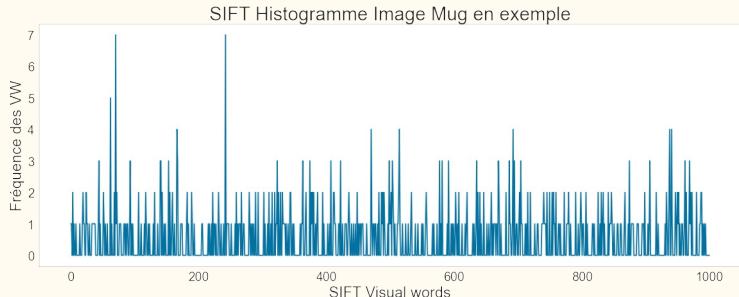


ORB (Oriented FAST and Rotated BRIEF):

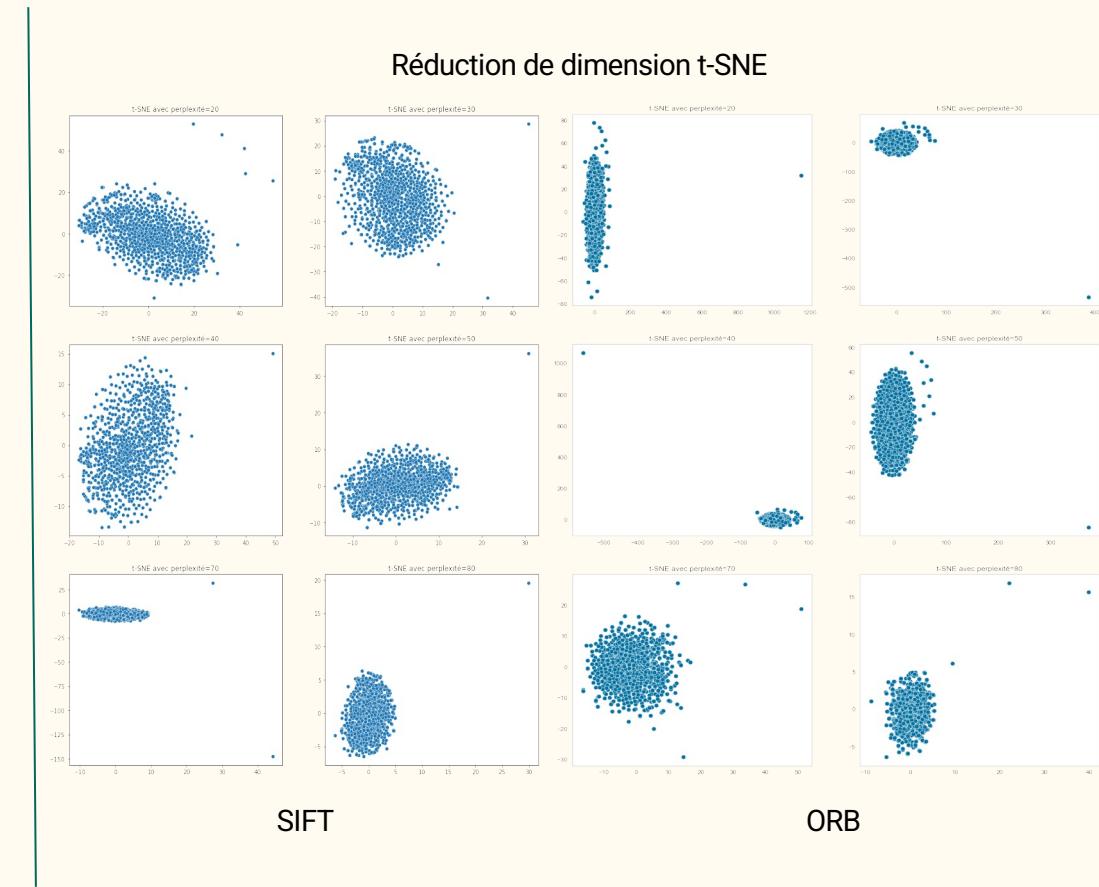
- S'appuie sur le détecteur de points clés FAST et le descripteur BRIEF.
- Les principales contributions d'ORB sont les suivantes :
 - L'ajout d'une composante d'orientation rapide et précise à FAST.
 - Le calcul efficace des caractéristiques BRIEF orientées
 - L'analyse de la variance et de la corrélation des caractéristiques BRIEF orientées
 - Une méthode d'apprentissage pour la dé-corrélation des caractéristiques BRIEF sous invariance rotationnelle, conduisant à de meilleures performances dans les applications de type "nearest-neighbor".

Etape 3 - Histogramme et Réduction des dimensions

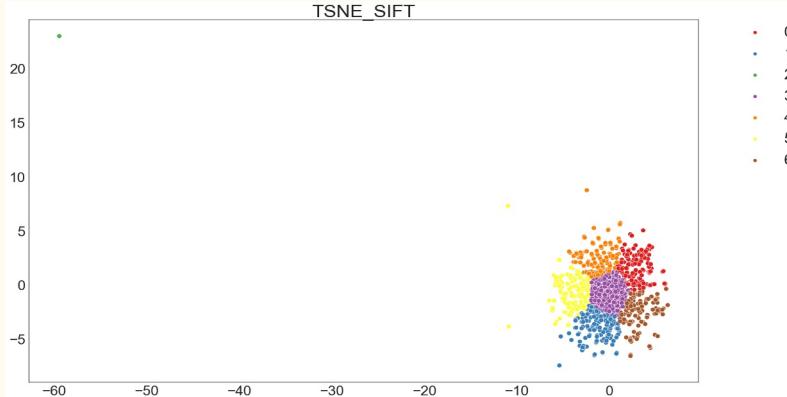
Histogramme des 'Visual Words'



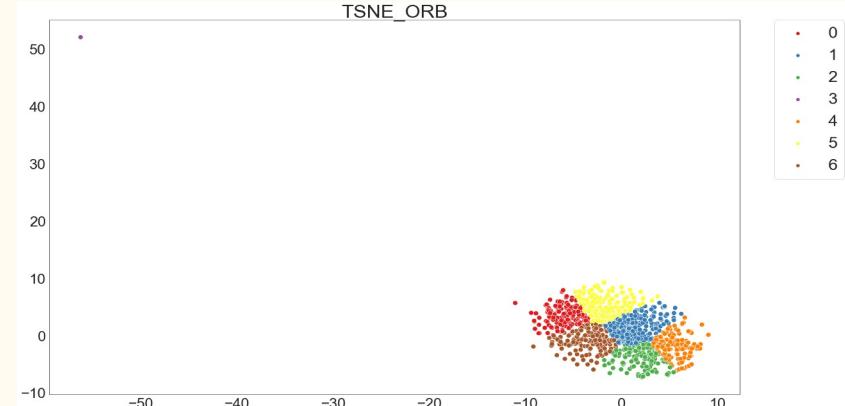
Réduction de dimension t-SNE



Etape 4 - Clustering SIFT/ORB et précision du clustering



	TSNE_SIFT							
Baby Care	73	30	34	33	42	34	34	
Beauty and Personal Care	8	33	20	23	20	16	26	
Computers	22	51	71	40	24	70	64	
Home Decor & Festive Needs	29	12	0	49	32	18	5	
Home Furnishing	18	24	25	5	32	12	21	
Kitchen & Dining	0	0	0	0	0	0	0	
Watches	0	0	0	0	0	0	0	
	0	1	2	3	4	5	6	



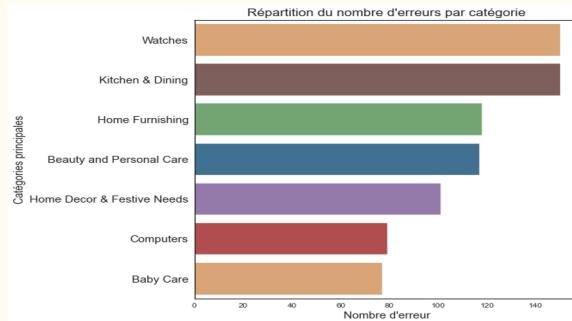
	TSNE_ORB							
Baby Care	35	26	12	23	31	16	24	
Beauty and Personal Care	0	0	0	0	0	0	0	
Computers	20	20	57	22	7	17	7	
Home Decor & Festive Needs	0	0	0	0	0	0	0	
Home Furnishing	53	41	14	54	73	31	36	
Kitchen & Dining	10	25	32	23	5	35	20	
Watches	32	38	35	28	34	51	63	
	0	1	2	3	4	5	6	

'Précision sur trainset : 24.57%'

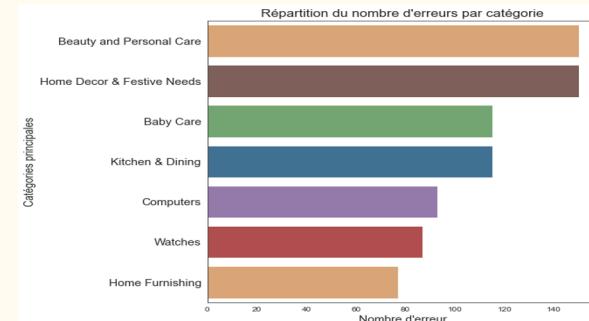
'Précision sur trainset: 25.05%'

Etape 5 - Répartition des erreurs

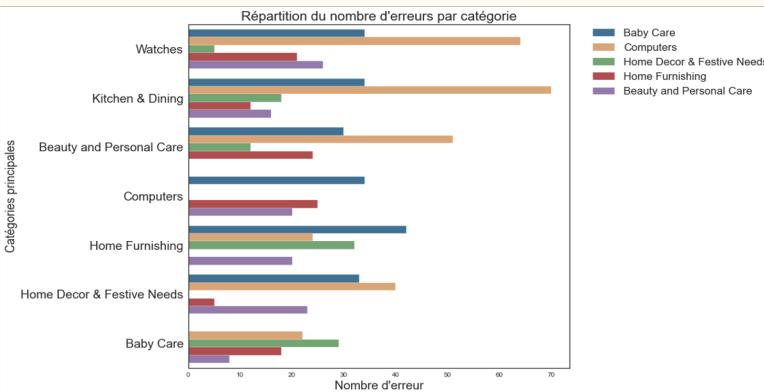
SIFT



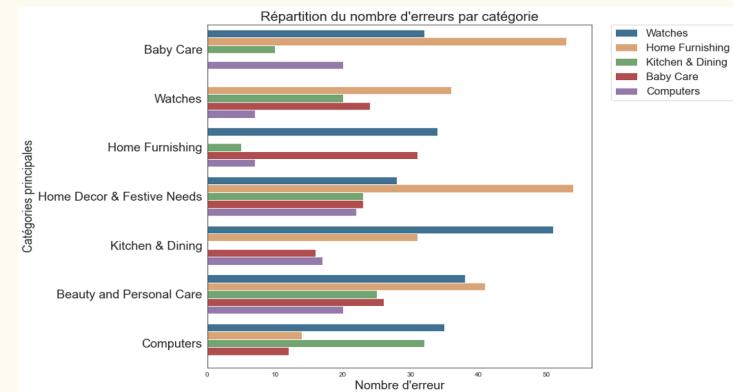
ORB



Répartition du nombre d'erreurs par catégorie



Répartition du nombre d'erreurs par catégorie



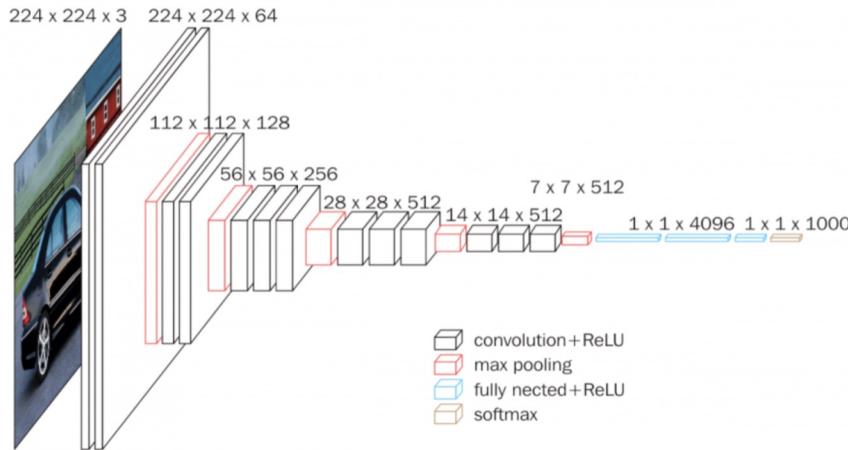
PARTIE III

TRAITEMENT DES IMAGES

(CNN / VGG16)

Transfer Learning & Entraînement avec VGG-16

VGG-16

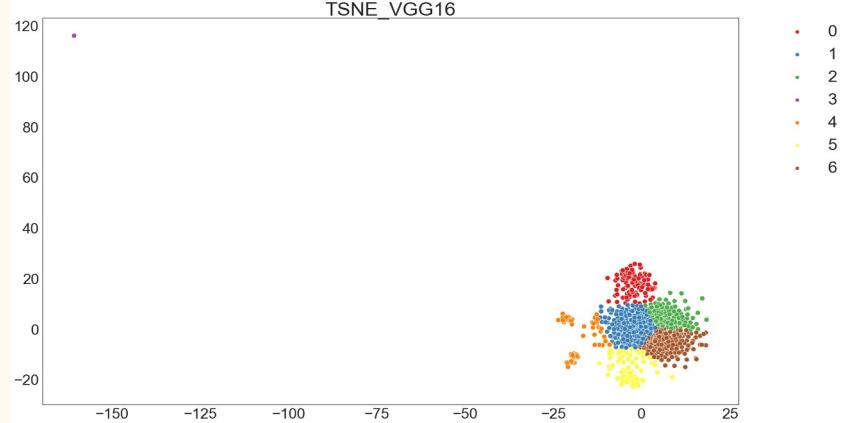


VGG-16 est une version du réseau de neurones convolutif VGG-Net

- VGG-16 est constitué de plusieurs couches, dont 13 couches de convolution et 3 fully-connected :
- Il prend en entrée une image en couleurs de taille 224 × 224 px et la classifie dans une des 1000 classes :
 - Il renvoie donc un vecteur de taille [1,7] qui contient les probabilités d'appartenance à chacune des classes.

Model: "vgg16"		
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[None, None, None, 3]	0
block1_conv1 (Conv2D)	(None, None, None, 64)	1792
block1_conv2 (Conv2D)	(None, None, None, 64)	36928
block1_pool (MaxPooling2D)	(None, None, None, 64)	0
block2_conv1 (Conv2D)	(None, None, None, 128)	73856
block2_conv2 (Conv2D)	(None, None, None, 128)	147584
block2_pool (MaxPooling2D)	(None, None, None, 128)	0
block3_conv1 (Conv2D)	(None, None, None, 256)	295168
block3_conv2 (Conv2D)	(None, None, None, 256)	590080
block3_conv3 (Conv2D)	(None, None, None, 256)	590080
block3_pool (MaxPooling2D)	(None, None, None, 256)	0
block4_conv1 (Conv2D)	(None, None, None, 512)	1180160
block4_conv2 (Conv2D)	(None, None, None, 512)	2359808
block4_conv3 (Conv2D)	(None, None, None, 512)	2359808
block4_pool (MaxPooling2D)	(None, None, None, 512)	0
block5_conv1 (Conv2D)	(None, None, None, 512)	2359808
block5_conv2 (Conv2D)	(None, None, None, 512)	2359808
block5_conv3 (Conv2D)	(None, None, None, 512)	2359808
block5_pool (MaxPooling2D)	(None, None, None, 512)	0
Total params: 14,714,688		
Trainable params: 14,714,688		
Non-trainable params: 0		

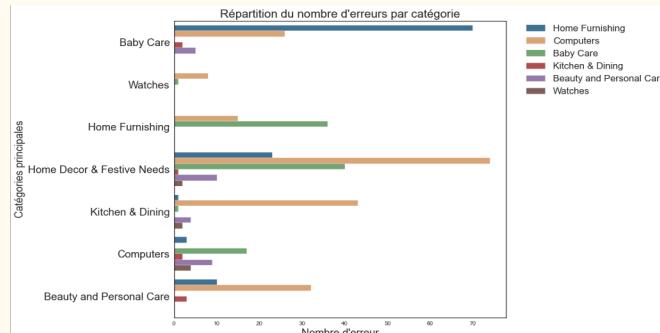
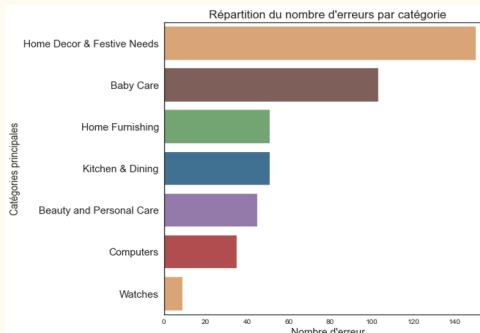
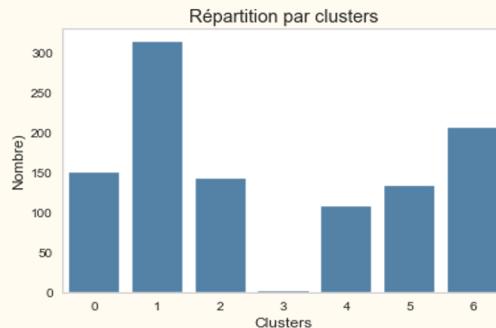
VGG16 - Réduction des dimensions et qualité de la prediction



TSNE_VGG16						
Baby Care	47	0	17	40	36	1
Beauty and Personal Care	5	105	9	10	0	4
Computers	26	32	115	74	15	43
Home Decor & Festive Needs	0	0	0	0	0	0
Home Furnishing	70	10	3	23	99	1
Kitchen & Dining	2	3	2	1	0	0
Watches	0	0	4	2	0	2
	0	1	2	3	4	5
						6

→ 'Précision: 57.71%'

→ "Nombre d'erreurs : 444"



Conclusions

La catégorisation automatique est-elle possible ?

- Il est possible de catégoriser automatiquement les produits dans les 7 catégories principales, mais certaines catégories sont moins évidentes à catégoriser que d'autres
- On pourrait imaginer un système de recommandation plutôt qu'un système de catégorisation automatique

Comment améliorer la qualité des résultats ?

- Nous recommandons d'inciter les vendeurs à améliorer la qualité des descriptions ainsi que celle des images, cela permettra de plus d'améliorer la qualité de notre dataset d'entraînement et ainsi améliorer nos recommandations
- Il pourrait être bénéfique de demander plusieurs images par produits si le taux de certitude du modèle est faible
- Il sera conseillé de privilégier une méthode combinée (image + texte) afin d'optimiser la fiabilité du modèle et ainsi de la recommandation au vendeur

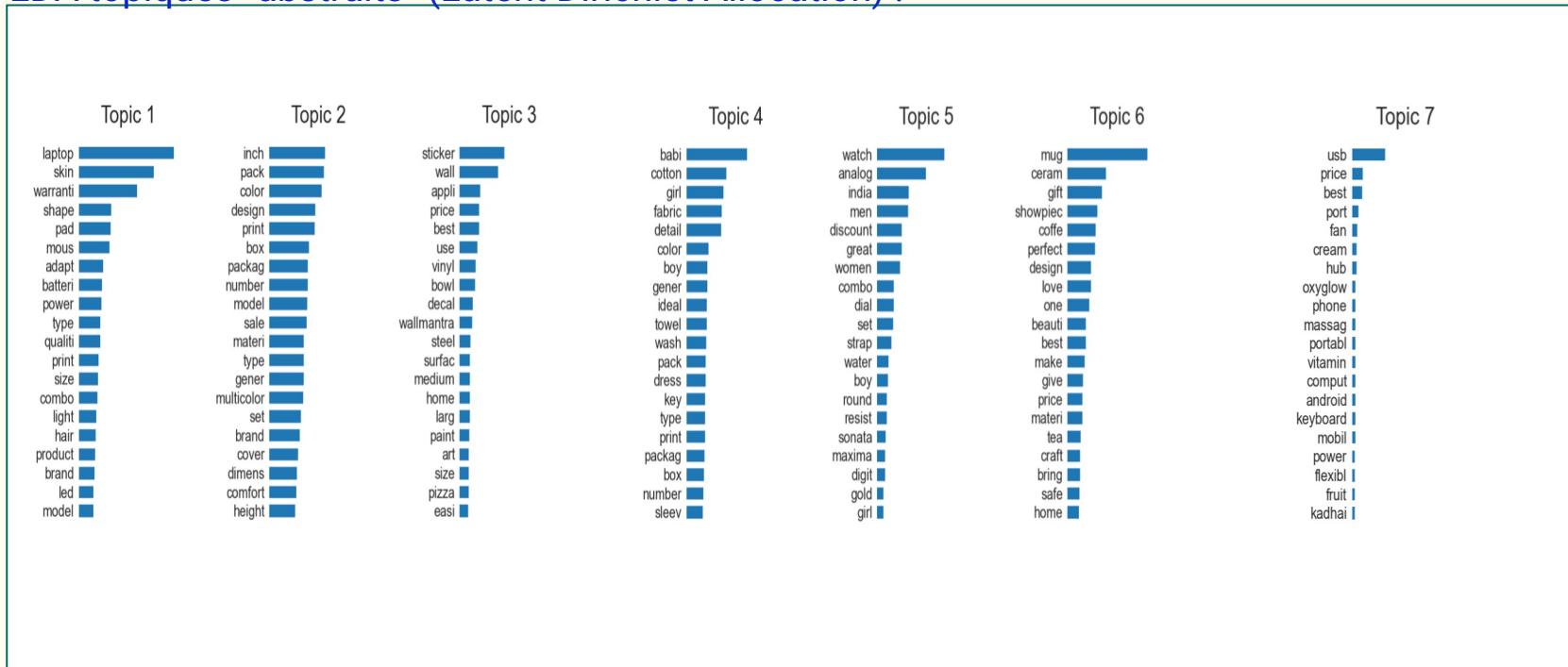
Bibliographie

- Chaudhary, Mukesh. "TF-IDF Vectorizer Scikit-Learn." Medium. Medium, January 28, 2021. <https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a>.
- Kulshrestha, Ria. "Latent Dirichlet Allocation(Lda)." Medium. Towards Data Science, September 28, 2020. <https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2>.
- Real Python. "Natural Language Processing with Python's NLTK Package." Real Python. Real Python, April 21, 2021. <https://realpython.com/nltk-nlp-python/>.
- "Convolutional Neural Networks Cheatsheet Star." CS 230 - Convolutional Neural Networks Cheatsheet. <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks>.

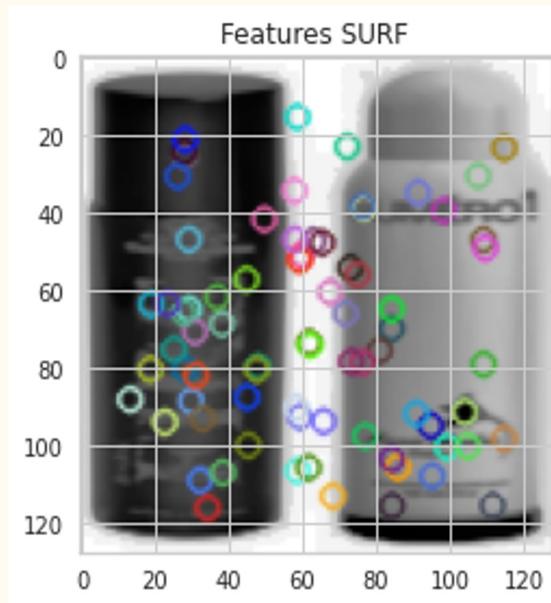
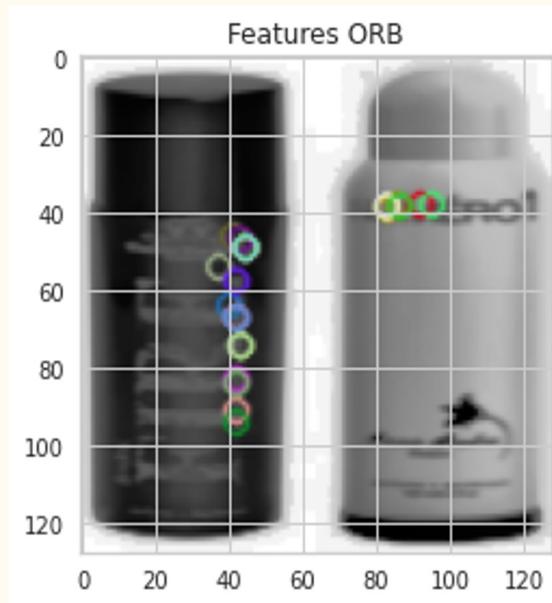
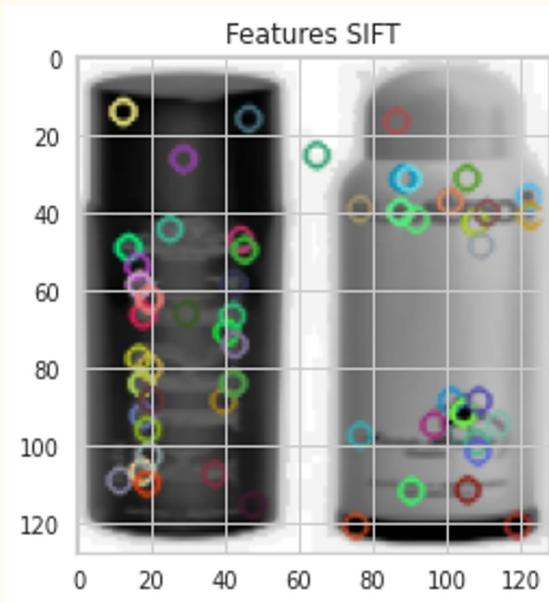
ANNEXES

Etape 2 - Extraction des caractéristiques (topic LDA)

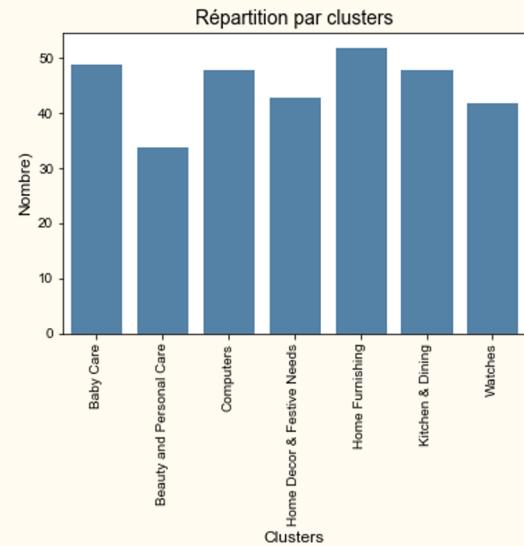
LDA topiques “abstraits” (Latent Dirichlet Allocation) :



Extraction des caractéristiques



Approche combinée



TSNE_VGG16							
	0	1	2	3	4	5	6
Baby Care	21	0	0	6	8	1	0
Beauty and Personal Care	3	32	5	0	2	1	0
Computers	1	0	35	3	1	2	1
Home Decor & Festive Needs	6	0	4	25	3	0	1
Home Furnishing	18	1	0	6	38	1	0
Kitchen & Dining	0	1	3	2	0	40	0
Watches	0	0	1	1	0	3	40
	0	1	2	3	4	5	6

'Précision: 73.1%'