

Projet 6 : Classification automatique des biens de consommation

Sommaire

1. Problématique
2. Exploration des données
3. Modèle de clustering et modèle supervisé de la partie texte
4. Modèle de clustering, modèle supervisé et approche CNN de la partie image
5. Conclusion

Problématique

- « Data Scientist » au sein de l'entreprise "Place de marché", qui souhaite lancer une marketplace e-commerce.
- Mission :
 - Réaliser une première étude de faisabilité d'un moteur de classification:
 - Analyser le jeu de données
 - Réaliser un prétraitement des images et des descriptions des produits
 - Réaliser une réduction de dimension
 - Réaliser un clustering

Exploration des données

- 1 dataset de 1050 individus et 15 variables ainsi qu'un dossier de 1050 images.
- 7 catégories majeures : Baby Care, Beauty and Personal Care, Computers, Home Decor & Festive Needs, Home Furnishing, Kitchen & Dining et Watches
- Chaque individu décrit 1 produit
- Chaque produit a 2 variables très intéressantes pour nous :
 - Une description et une photo qui seront traités

Prétraitement du texte

- Traitement avec NLTK :
 - Retrait ponctuation, digit, stopwords
 - Minuscule
 - Tokenisation
 - Lemmatisation/ racinisation
- Extraction des features :
 - Bag of words : Countvectorizer et TF-IDF

Prétraitement du texte

- Exemple :

----- Phrase de départ: -----

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

----- Exemple sans stopwords: -----

key features elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract eyelet door

----- Tokenization: -----

['key', 'features', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'floral', 'curtainelegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door']

----- Phrase racinisée: -----

key featur eleg polyest multicolor abstract eyelet door curtain floral curtaineleg polyest multicolor abstract eyelet door

----- Phrase lemmatisée: -----

key feature elegance polyester multicolor abstract eyelet door curtain floral polyester multicolor abstract eyelet door

Prétraitement du texte : Countvectorizer

- Un vecteur codé est renvoyé avec la longueur du vocabulaire entier et un nombre entier pour le nombre de fois où chaque mot est apparu dans le document.

	aa	able	abode	abrasion	absorbency	absorbent	absorbing	abstract	accent	access	...	york	youd	young	youth	youthful	youve	zero	zip	zipper
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

5 rows × 2435 columns

abstract	curtain	door	elegance	eyelet	feature	floral	key	multicolor	polyester
0	2	1	2	1	2	1	1	1	2

Prétraitement du texte : TF-IDF

- Cette mesure statistique permet d'évaluer l'importance d'un terme (et de le vectoriser en même temps) contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.

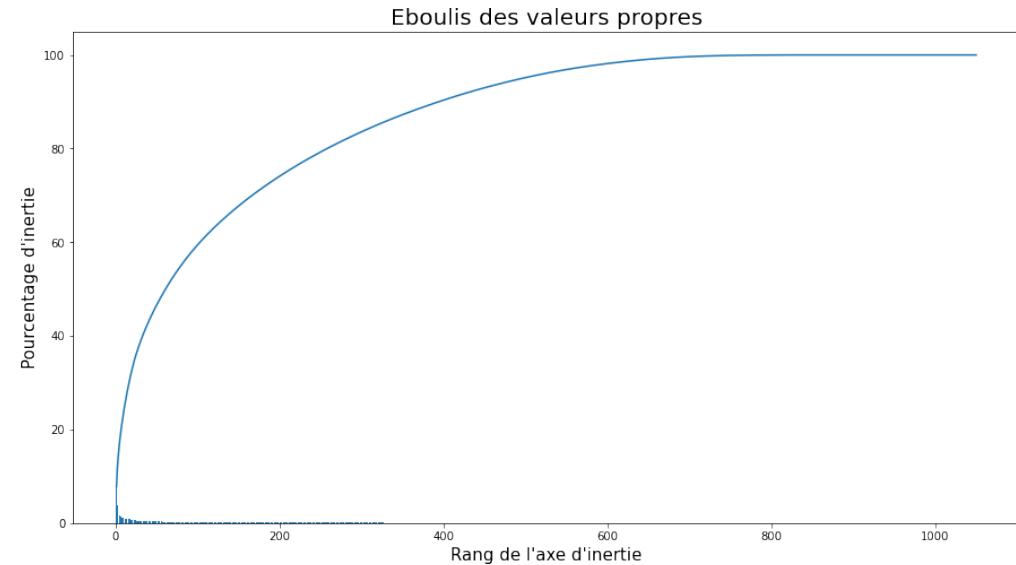
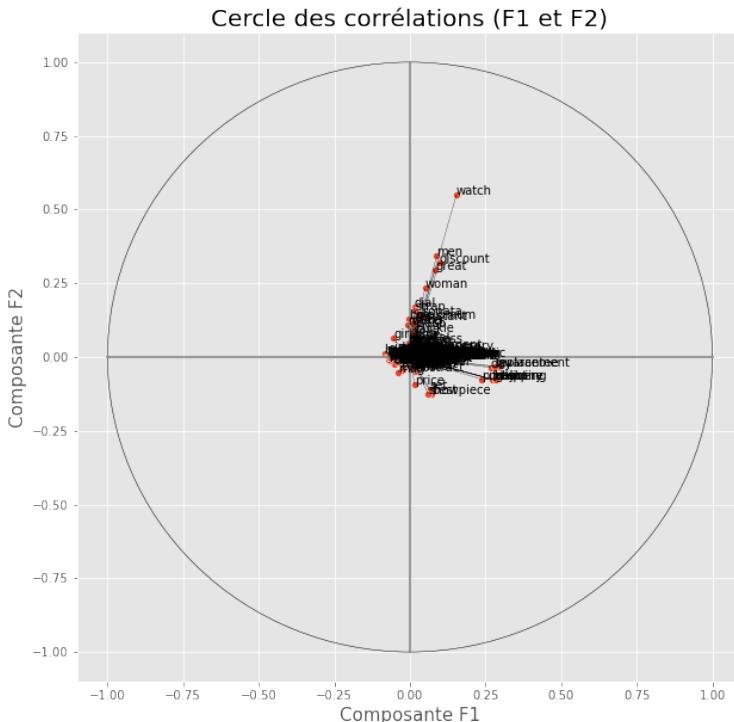
	aa	able	abode	abrasion	absorbency	absorbent	absorbing	abstract	accent	access	...	york	youd	young	youth	youthful	youve	zero	zip	zipper
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

5 rows × 2435 columns

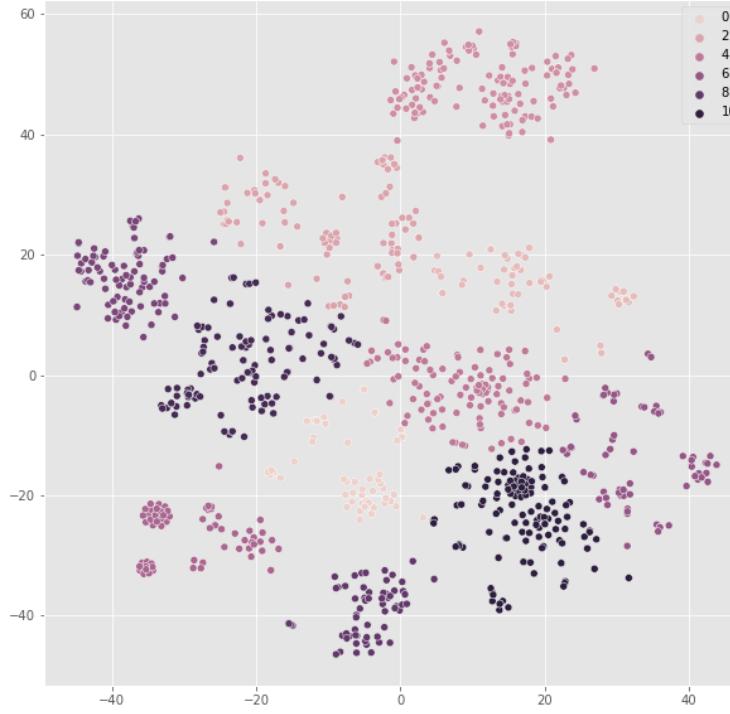
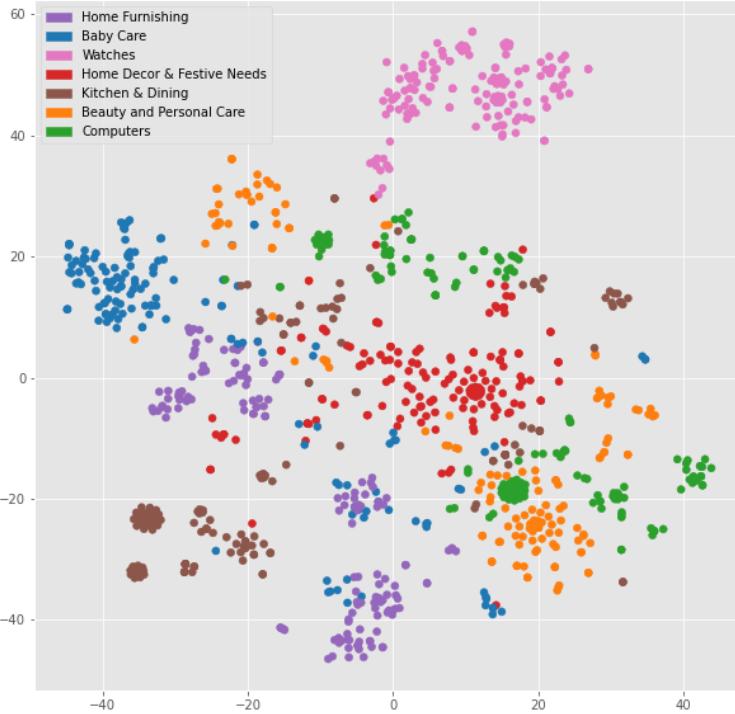
	abstract	curtain	door	elegance	eyelet	feature	floral	key	multicolor	polyester
0	0.385067	0.227427	0.385067	0.227427	0.385067	0.227427	0.227427	0.227427	0.385067	0.385067

Réduction de dimension ACP:

- 4347 variables → 496 composantes

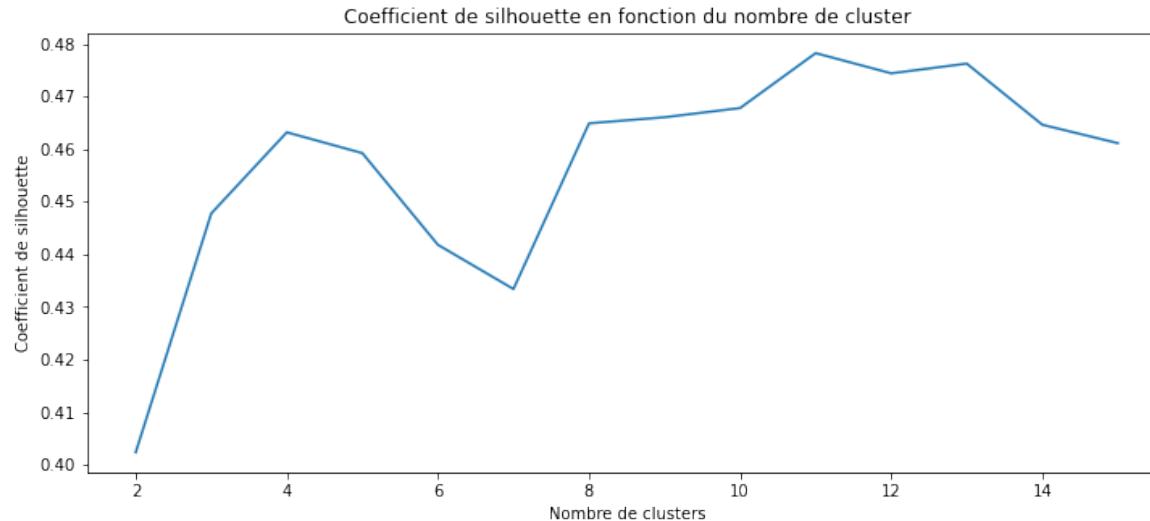


Réduction de dimension ACP:



Modèle de Clustering KMeans :

- Meilleur résultat : TF-IDF avec les données lemmatisées, après réduction via T-SNE.
 - 11 Clusters
 - Coefficient de silhouette : ~ 0,48
 - Indice de Rand ajusté pour KMeans: ~ 0.43

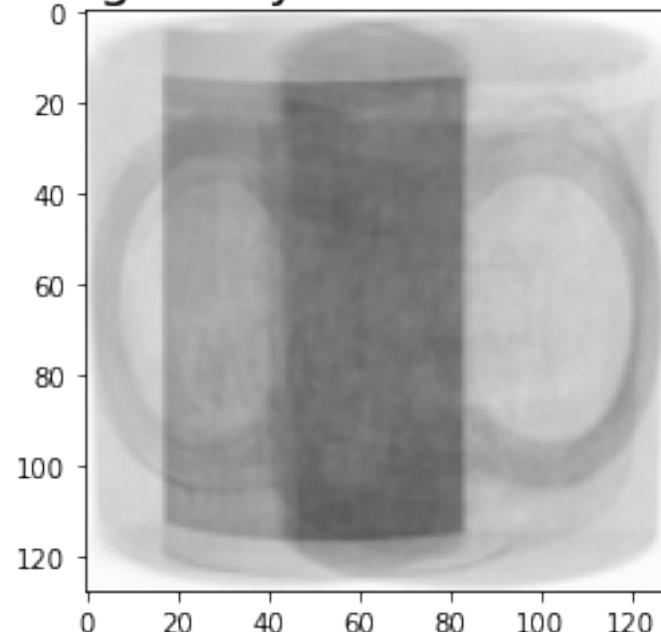


Modèle de Clustering KMeans :

- Nuage de mot et image moyenne pour chaque cluster :

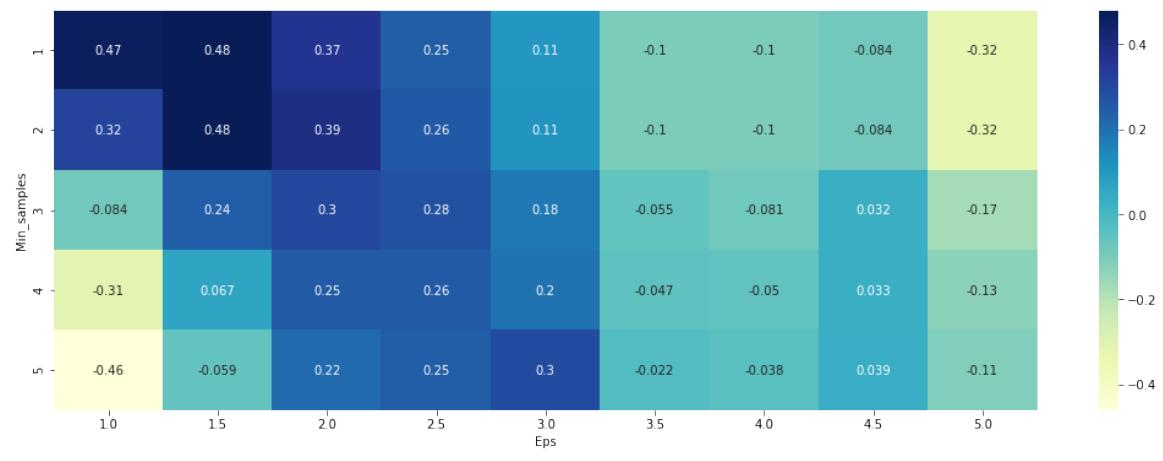
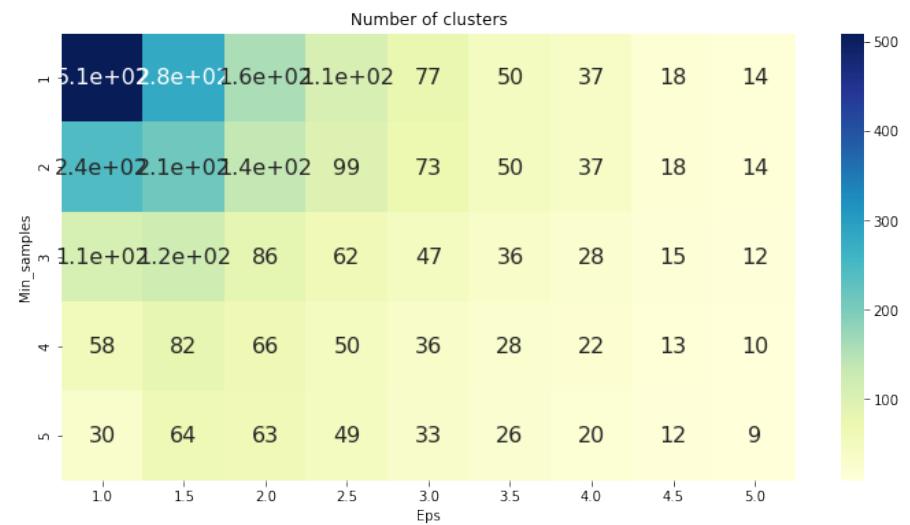


Image moyenne du cluster5



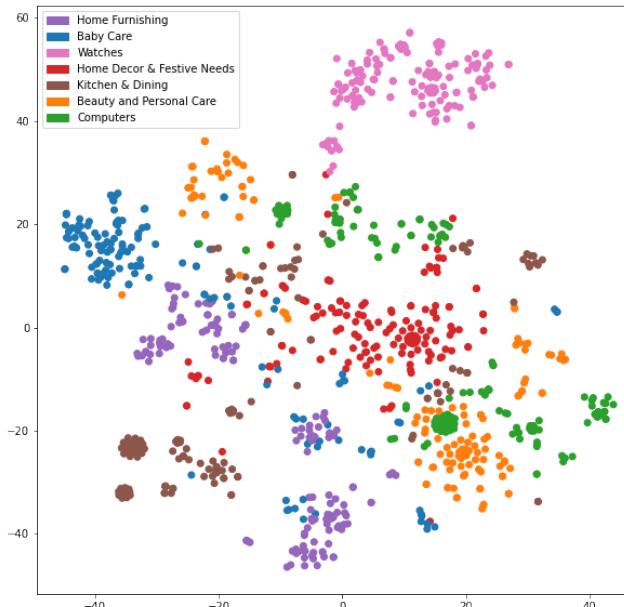
Modèle de Clustering DBScan :

- Meilleur résultat :
 - 280 Clusters
 - Coefficient de silhouette : $\sim 0,48$



Modèle de Clustering KPrototypes :

- Meilleur résultat : TF-IDF avec les données lemmatisées, après réduction via T-SNE.
 - Indice de Rand ajusté pour KPrototypes: ~0.44



Modèle de Classification supervisé :

- Modèle Naïve de Bayes :
 - La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classificateurs linéaires. En termes simples, un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques.
 - Précision: 83.81 %
 - Précision Train: 86.67 %
 - RMSE Test: 1.05
- Modèle SVC :
 - L'objectif d'un SVC linéaire (classificateur de vecteur de support, modèle supervisé) est de s'adapter aux données que l'on va fournir, en renvoyant un hyperplan qui divise ou catégorise nos données au mieux.
 - Précision: 93.02 %
 - Précision Train: 93.33 %
 - RMSE Test: 0.73

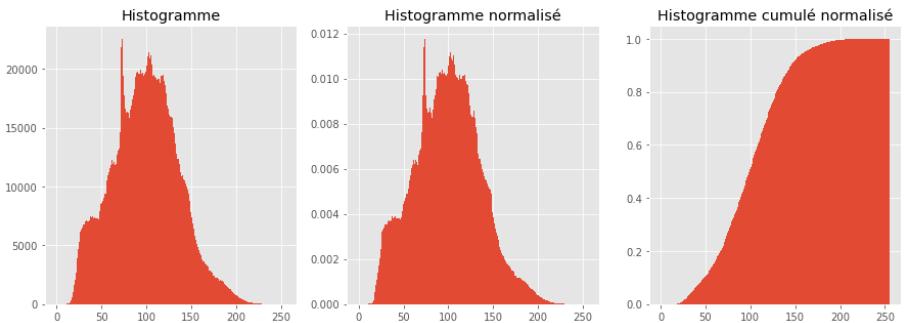
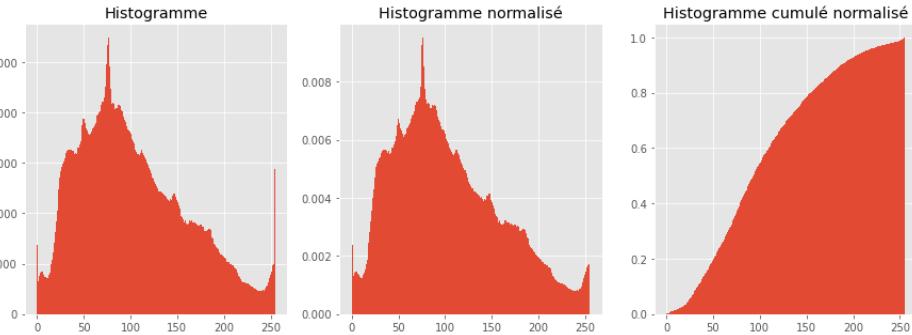
Conclusion de l'analyse texte

- Indice de Rand ajusté : ~ 0,44 (en utilisant K-Prototype, pour le clustering non supervisée)
- RMSE : ~ 0,73 (en utilisant SVC, pour la classification supervisée)
- Avec l'analyse et le prétraitement textuelle effectué le résultat n'est pas concluant pour permettre un clustering vraiment intéressant, cependant nous pouvons voir que l'approche supervisée est plus intéressante, nous donc nous pencher plutôt sur cette dernière.

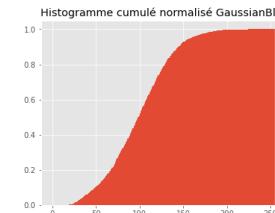
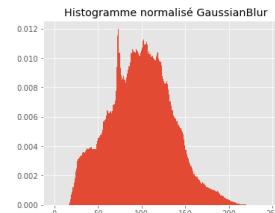
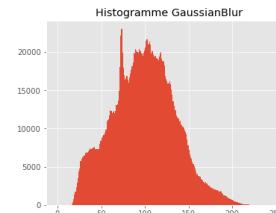
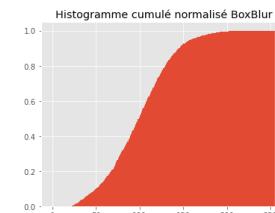
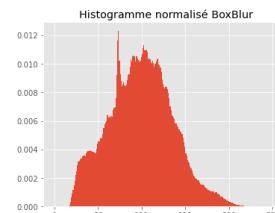
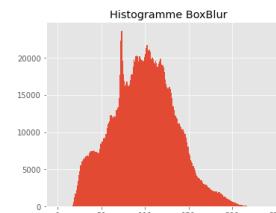
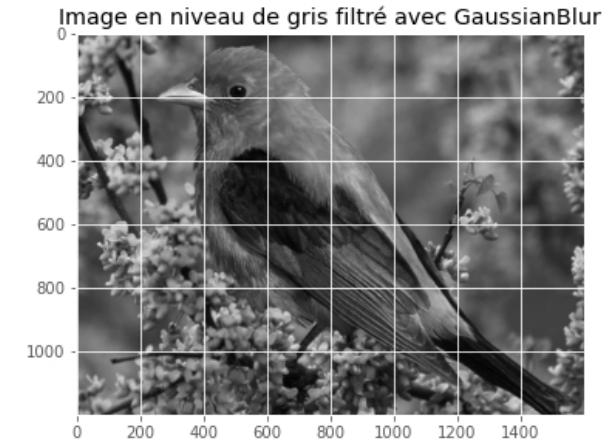
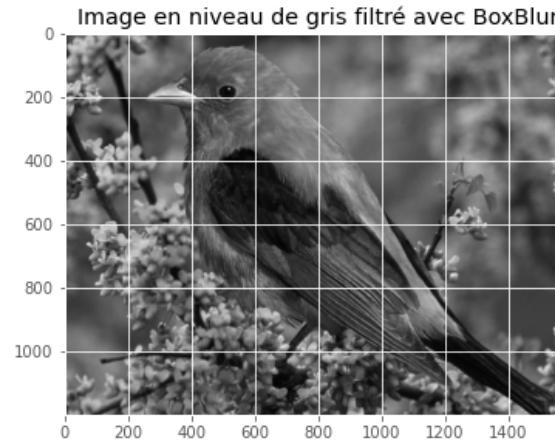
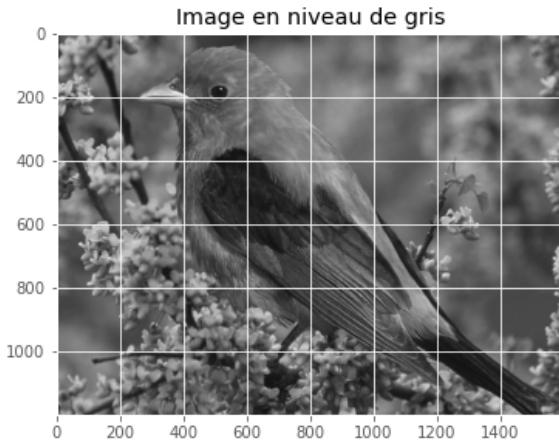
Prétraitement des images

- Traitement avec PIL :
 - Conversion en niveau de gris
 - Réduction du bruit avec BoxBlur et GaussianBlur
 - Correction de la luminosité avec Autocontrast et Equalize
 - Redimensionnement (à partir de l'approche CNN)
- Extraction des features :
 - Sift
 - ORB

Prétraitement des images : Conversion en niveau de gris



Prétraitement des images : Réduction du bruit



Prétraitement des images : Correction de la luminosité avec Autocontrast

Image en niveau de gris sans transformation

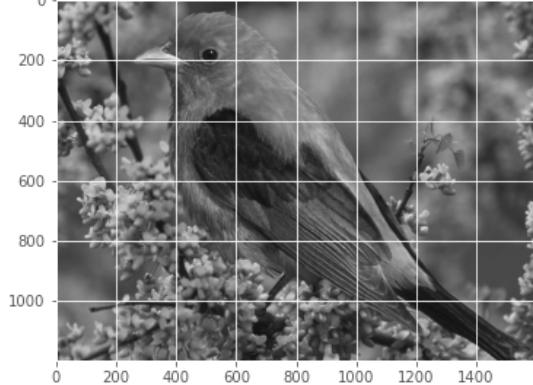


Image en niveau de gris filtré avec BoxBlur

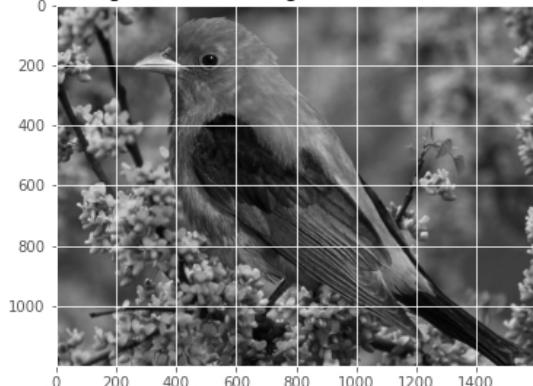


Image en niveau de gris

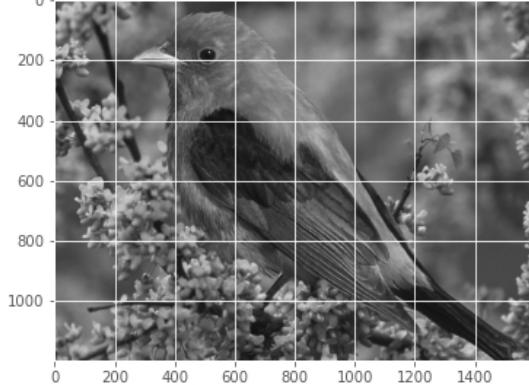
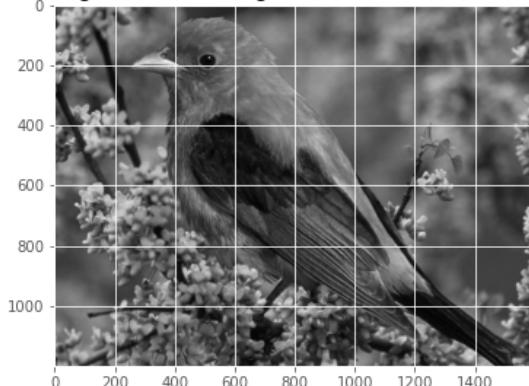
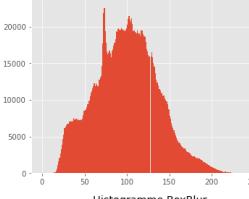


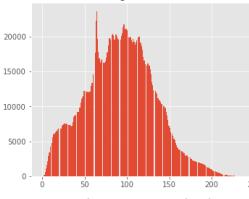
Image en niveau de gris filtré avec GaussianBlur



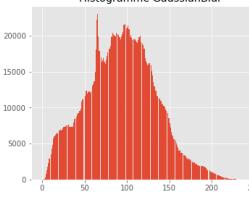
Histogramme



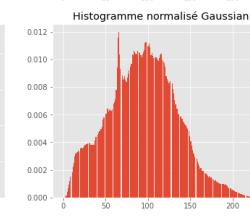
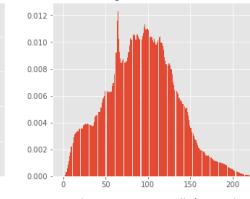
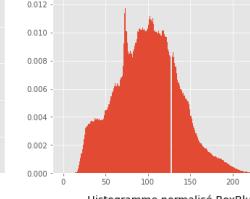
Histogramme BoxBlur



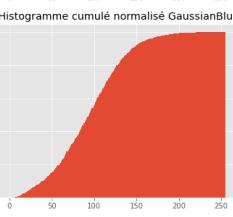
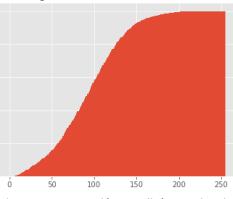
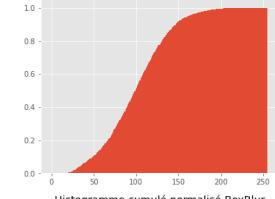
Histogramme GaussianBlur



Histogramme normalisé



Histogramme cumulé normalisé



Prétraitement des images : Correction de la luminosité avec Equalize

Image en niveau de gris sans transformation

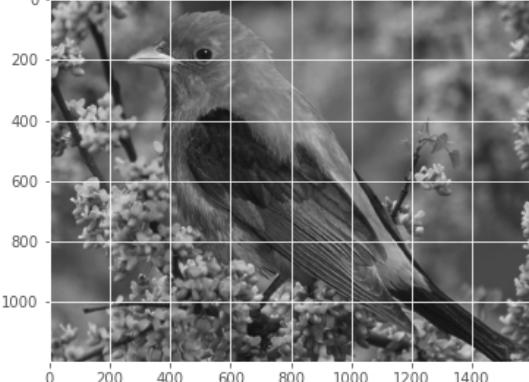


Image en niveau de gris

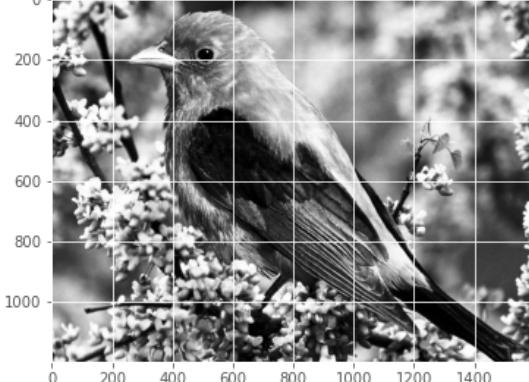


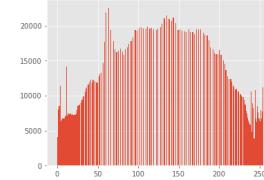
Image en niveau de gris filtré avec BoxBlur



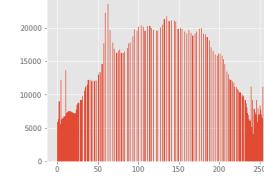
Image en niveau de gris filtré avec GaussianBlur



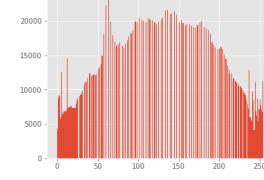
Histogramme



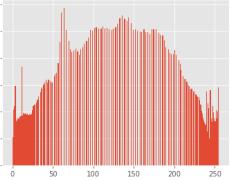
Histogramme BoxBlur



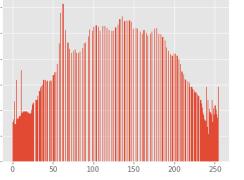
Histogramme GaussianBlur



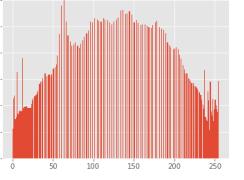
Histogramme normalisé



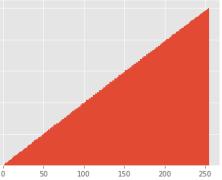
Histogramme normalisé BoxBlur



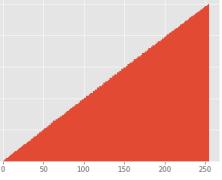
Histogramme normalisé GaussianBlur



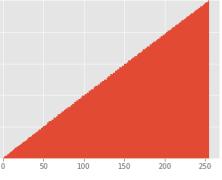
Histogramme cumulé normalisé



Histogramme cumulé normalisé BoxBlur

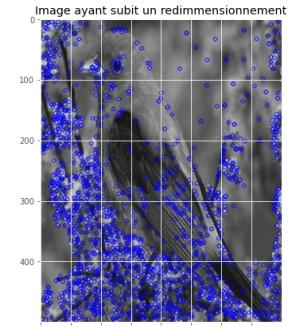
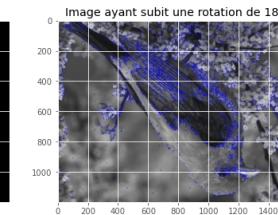
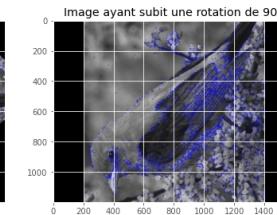
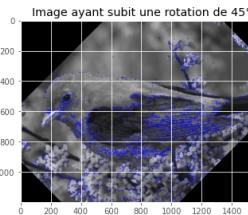
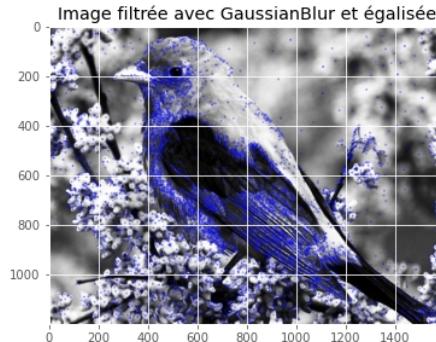
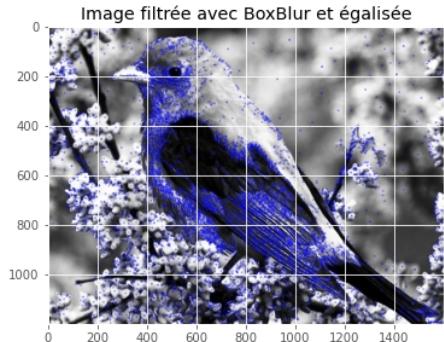
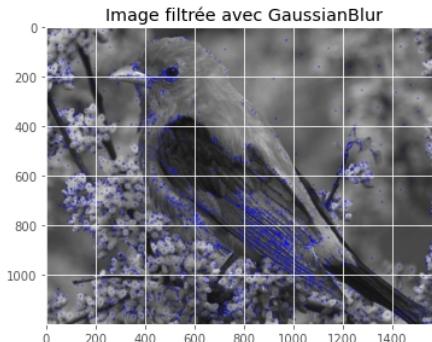
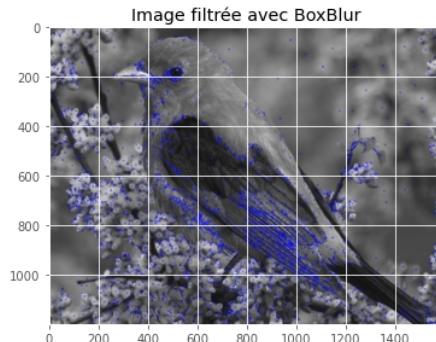
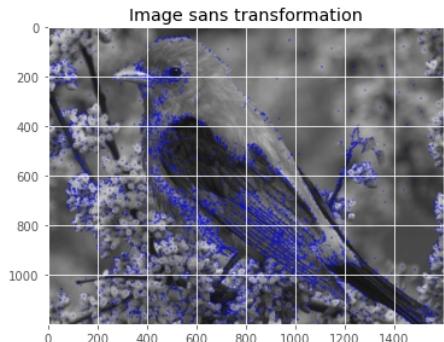


Histogramme cumulé normalisé GaussianBlur

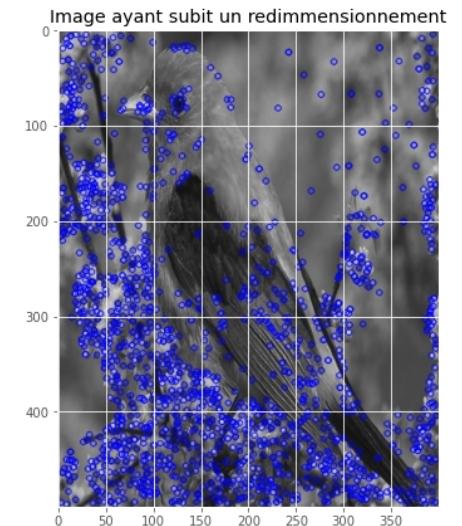
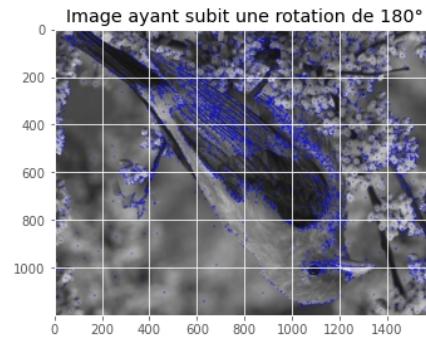
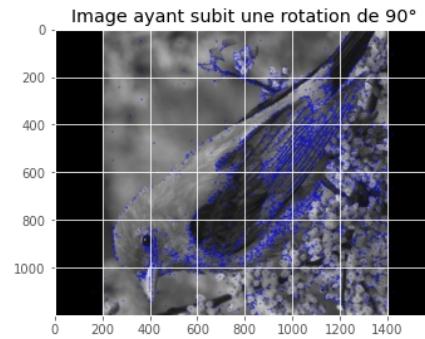
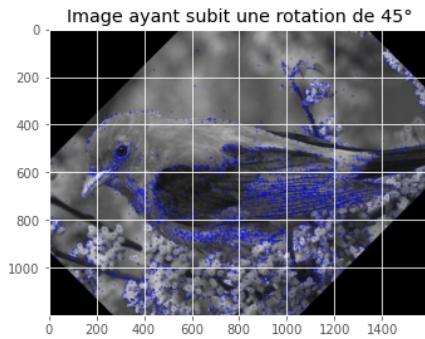


Extraction des features: SIFT

- SIFT est un algorithme utilisé dans le domaine de la vision par ordinateur pour détecter et identifier les éléments similaires entre différentes images numériques.



Extraction des features: SIFT

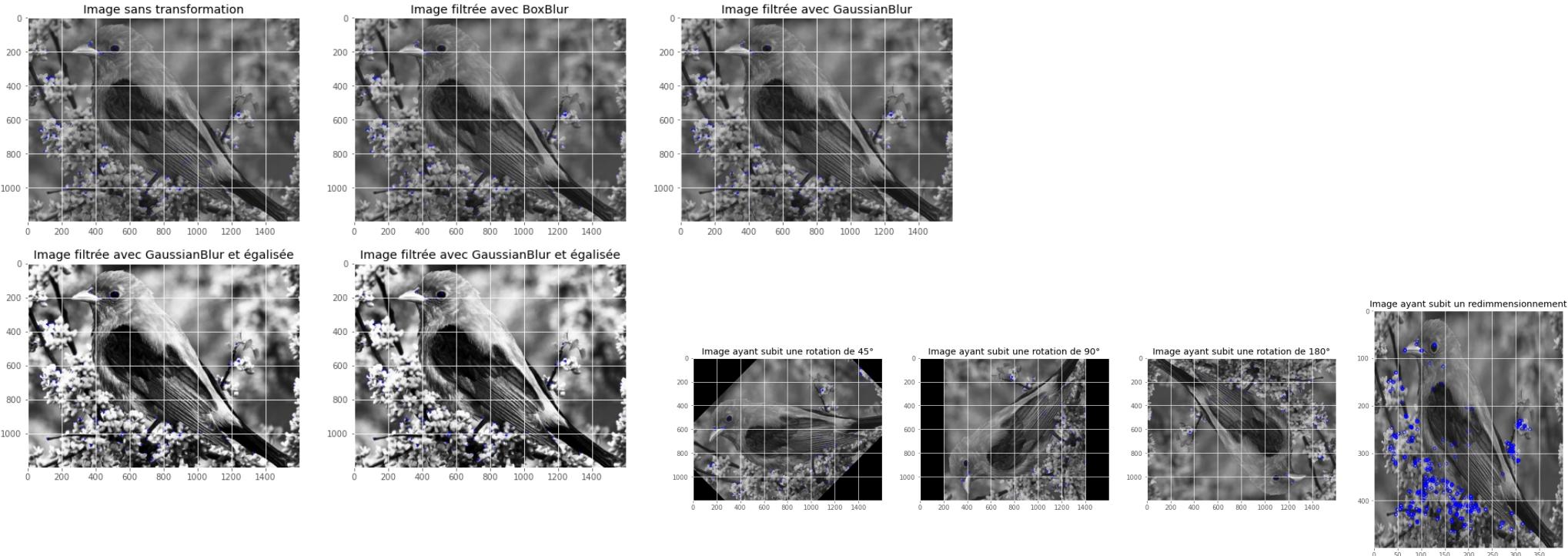


	0	1	2	3	4	5	6	7	8	9	...	118	119	120	121	122	123	124	125	126	127
0	64.0	4.0	0.0	0.0	5.0	9.0	21.0	43.0	34.0	5.0	...	4.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	1.0	...	9.0	17.0	31.0	4.0	0.0	0.0	0.0	0.0	9.0	73.0
2	13.0	6.0	19.0	92.0	16.0	0.0	0.0	1.0	39.0	4.0	...	53.0	33.0	0.0	0.0	0.0	0.0	32.0	33.0	47.0	22.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.0	2.0	...	67.0	23.0	0.0	22.0	78.0	21.0	38.0	14.0	4.0	0.0
4	3.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	68.0	65.0	...	3.0	86.0	20.0	13.0	4.0	1.0	0.0	19.0	108.0	

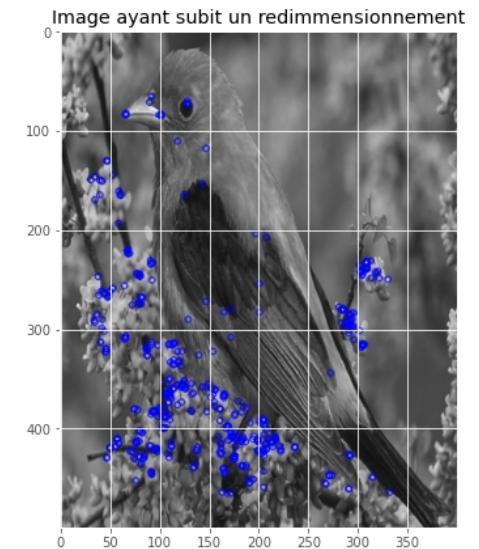
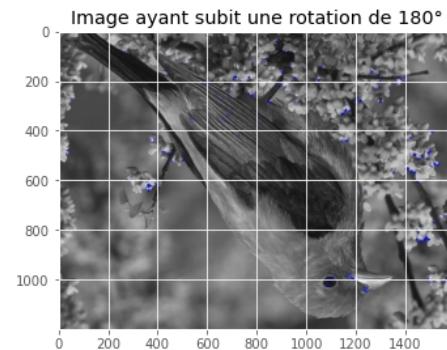
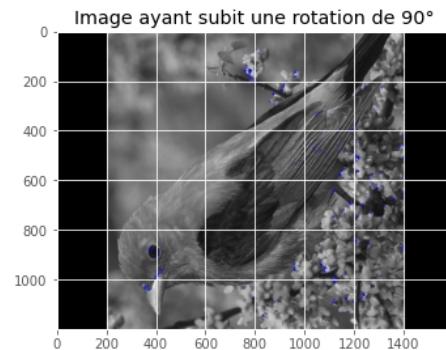
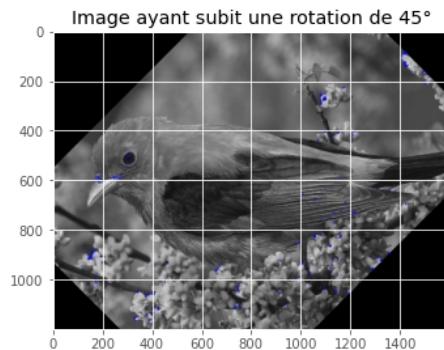
5 rows × 128 columns

Extraction des features: ORB

- ORB fonctionne aussi bien que SIFT dans la tâche de détection de caractéristiques (et est meilleur que SURF) tout en étant presque deux ordres de grandeur plus rapide.



Extraction des features: ORB

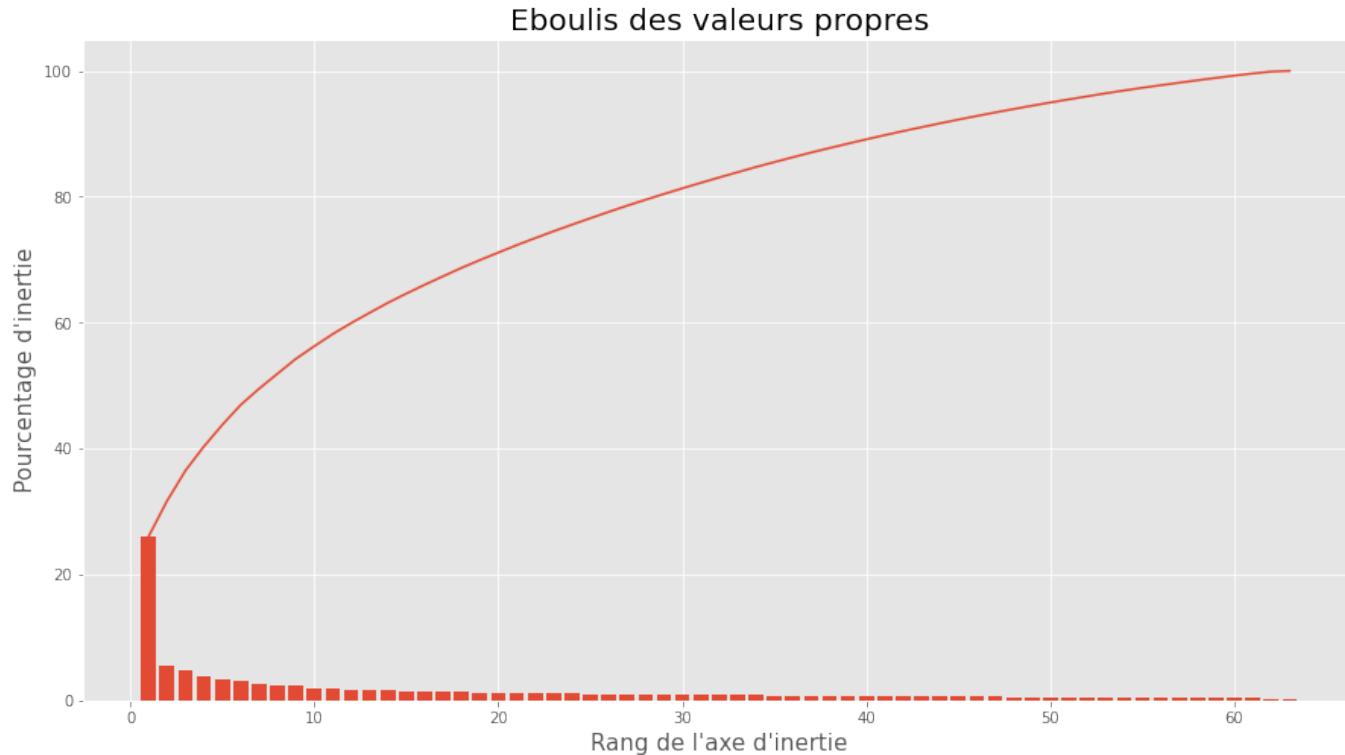


	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	93	1	142	31	129	58	101	155	118	151	...	255	40	199	231	85	5	234	242	234	0
1	50	233	236	17	221	213	252	207	107	117	...	183	241	63	95	157	231	199	20	222	222
2	185	143	67	204	244	22	78	119	69	66	...	37	175	50	237	166	0	111	254	248	45
3	96	253	22	62	121	141	163	230	190	156	...	218	117	189	54	31	106	72	0	43	154
4	176	97	14	93	194	34	36	77	54	0	...	195	92	53	42	154	0	33	16	234	130

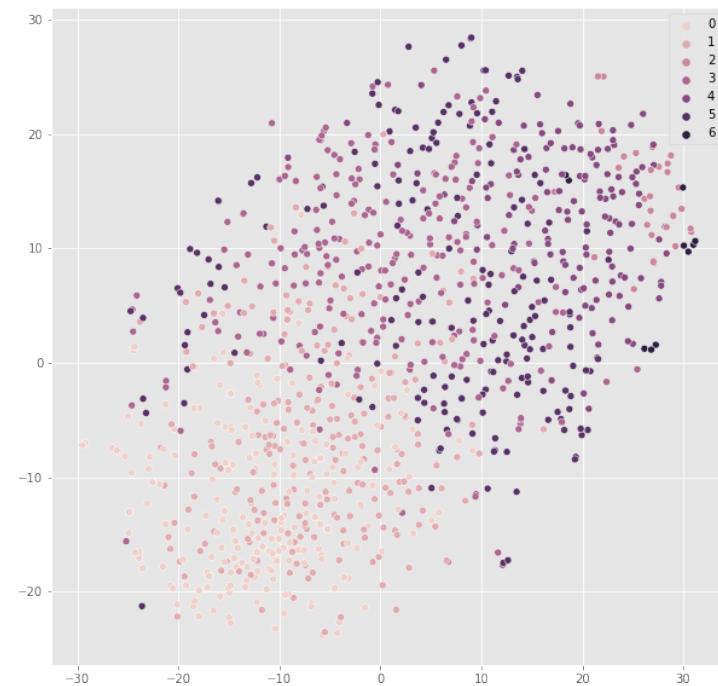
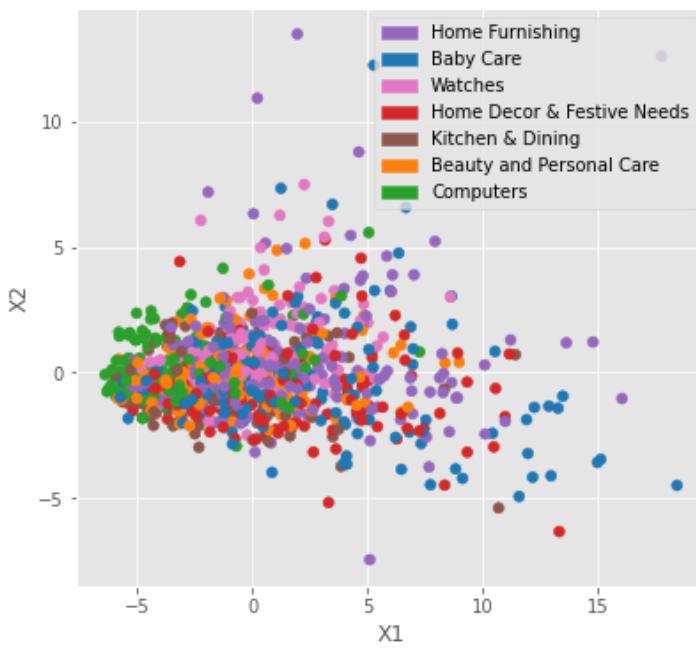
5 rows × 32 columns

Réduction de dimension ACP:

- 63 variables → 51 composantes

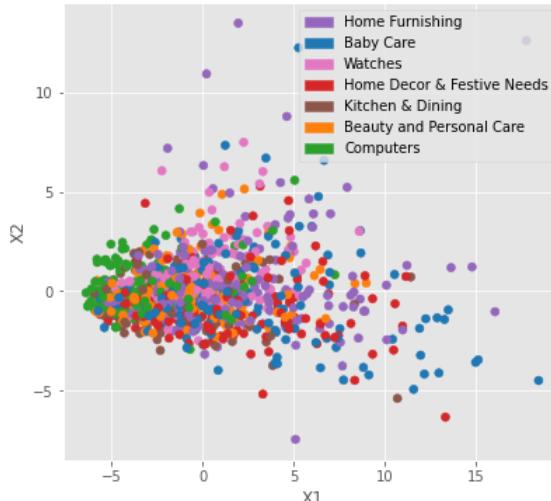


Réduction de dimension ACP:



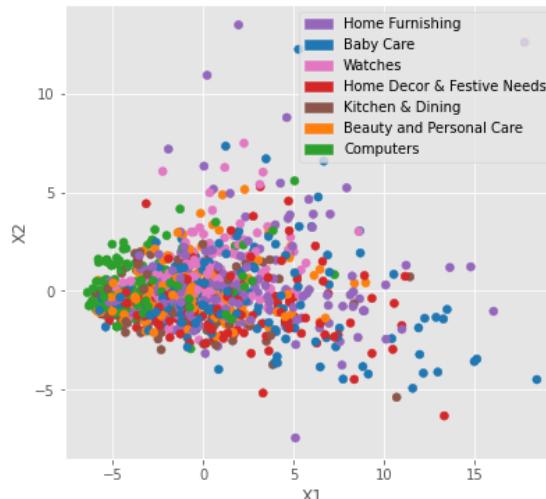
Modèle de Clustering KMeans non supervisé:

- Meilleur résultat :
 - 4 Clusters
 - Coefficient de silhouette : ~ 0,08
 - Indice de Rand ajusté pour KMeans: ~ 0.05



Modèle de Clustering KMeans semi-supervisé :

- Meilleur résultat :
 - 4 Clusters
 - Coefficient de silhouette : ~ 0,11
 - Indice de Rand ajusté pour KMeans: ~ 0.232



Modèle de Classification supervisé :

- Modèle SVC :
 - L'objectif d'un SVC linéaire (classificateur de vecteur de support, modèle supervisé) est de s'adapter aux données que l'on va fournir, en renvoyant un hyperplan qui divise ou catégorise nos données au mieux.
 - La prédiction de SVC est de 32.38 %
 - RMSE : 2.54
- Modèle Random Forest :
 - Les forêts aléatoires ou forêts de décision aléatoire sont une méthode d'apprentissage d'ensemble pour la classification, la régression et d'autres tâches qui fonctionne en construisant une multitude d'arbres de décision au moment de la formation. Pour les tâches de classification, la sortie de la forêt aléatoire est la classe sélectionnée par la plupart des arbres.
 - RMSE : 1.93

Conclusion de l'analyse visuelle

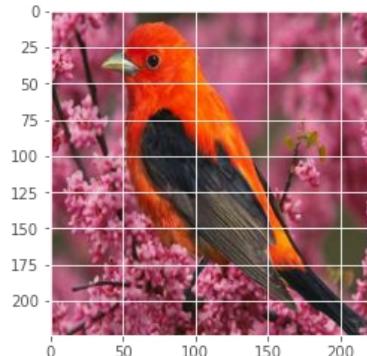
- Indice de Rand ajusté: 0.227 (en utilisant KMeans semi supervisé, pour le clustering non supervisée)
- RMSE : 1,93 (en utilisant Random Forest, pour la classification supervisée)
- Avec l'analyse et le prétraitement visuelle effectué le résultat n'est pas concluant pour permettre un clustering visuelle vraiment intéressant.

Approche CNN pour les images

- Utilisation du transfert learning :
 - Redimensionnement des images à la taille d'entrée du réseau de neurones VGG16 (224x224)
 - Fine Tuning
 - Compilation et entraînement du « nouveau » réseau de neurones
- Le Transfer Learning, désigne l'ensemble des méthodes qui permettent de transférer les connaissances acquises à partir de la résolution de problèmes donnés pour traiter un autre problème.
- Dans notre cas nous utiliserons le **CNN VGG16** entraîné sur la base **ImageNet**.
- ImageNet est une gigantesque base de données de plus de 14 millions d'images labellisées réparties dans plus de 1000 classes.
- Les CNN désignent une sous-catégorie de réseaux de neurones et sont à ce jour un des modèles de classification d'images réputés être les plus performant.

Transfert learning pour les images

- Fine tuning, pourquoi ?
- Le fine tuning, en général, signifie apporter de petits ajustements à un processus pour obtenir le résultat ou les performances souhaités.



Top 3 : [('lorikeet', 0.7676914), ('worm_fence', 0.05068266), ('robin', 0.024736932)]

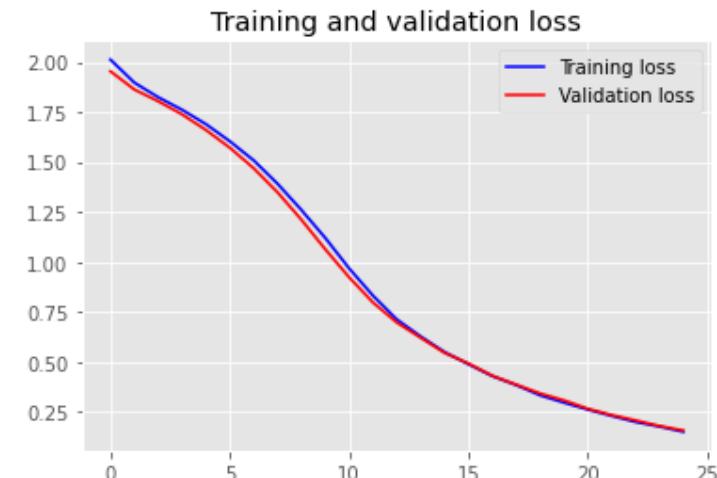
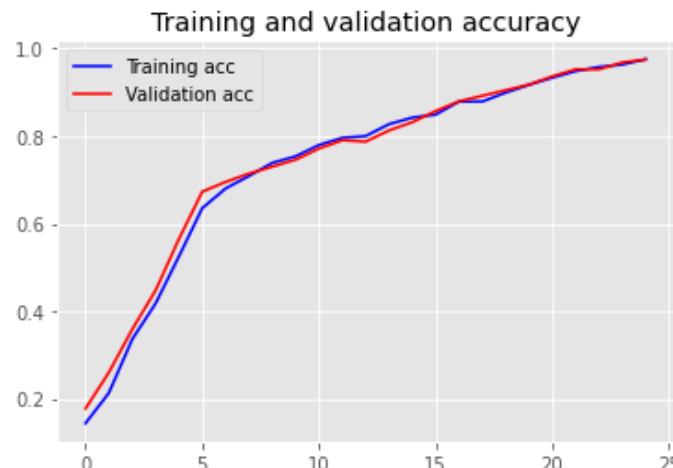
- Catégorie à notre disposition :
 - Baby Care, Beauty and Personal Care, Computers, Home Decor & Festive Needs, Home Furnishing, Kitchen & Dining et Watches

Approche CNN pour les images

- Fine tuning, lequel?
- Il y a 3 types principaux de Fine tuning :
 - **Fine-tuning total :** On remplace la dernière couche fully-connected du réseau pré-entraîné par un classifieur adapté au nouveau problème (SVM, régression logistique...) et initialisé de manière aléatoire. Toutes les couches sont ensuite entraînées sur les nouvelles images.
Doit être utilisée lorsque la nouvelle collection d'images est grande.
 - **Extraction des features :** Cette stratégie consiste à se servir des features du réseau pré-entraîné pour représenter les images du nouveau problème. Pour cela, on retire la dernière couche fully-connected et on fixe tous les autres paramètres.
Doit être utilisée lorsque la nouvelle collection d'images est petite et similaire aux images de pré-entraînement.
 - **Fine-tuning partiel :** Il s'agit d'un mélange des stratégies #1 et #2 : on remplace à nouveau la dernière couche fully-connected par le nouveau classifieur initialisé aléatoirement, et on fixe les paramètres de certaines couches du réseau pré-entraîné.
On utilise cette stratégie lorsque la nouvelle collection d'images est petite mais très différente des images du pré-entraînement.

Approche CNN pour les images

- Compilation et entraînement :
- 0,7 pour train, 0,3 test sur catégorie du site



Approche CNN pour les images

- Compilation et entraînement :
- Résultat :
- Nombre d'image dans chaque catégorie
 - Catégorie: Home Furnishing , nombre: 33 soit 73.3 % de bonne classification.
 - Catégorie: Baby Care , nombre: 22 soit 48.9 % de bonne classification.
 - Catégorie: Watches , nombre: 42 soit 93.3 % de bonne classification.
 - Catégorie: Home Decor & Festive Needs , nombre: 29 soit 64.4 % de bonne classification.
 - Catégorie: Kitchen & Dining , nombre: 14 soit 31.1 % de bonne classification.
 - Catégorie: Beauty and Personal Care , nombre: 16 soit 35.6 % de bonne classification.
 - Catégorie: Computers , nombre: 22 soit 48.9 % de bonne classification.

Approche CNN pour les images

- Compilation et entraînement :
- Résultat mauvaise classification :

Image de catégorie Baby Care identifiée comme Home Furnishing



Image de catégorie Watches identifiée comme Personal Care

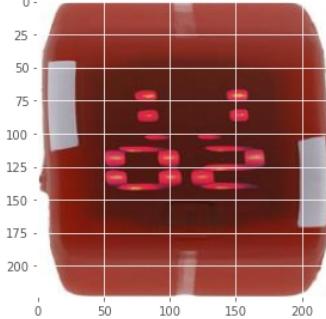


Image de catégorie Home Furnishing identifiée comme Baby Care

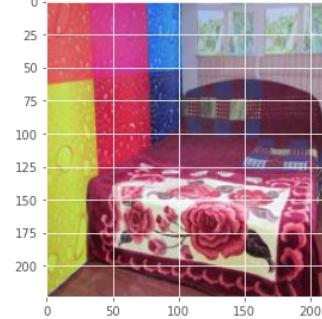


Image de catégorie Watches identifiée comme Computer

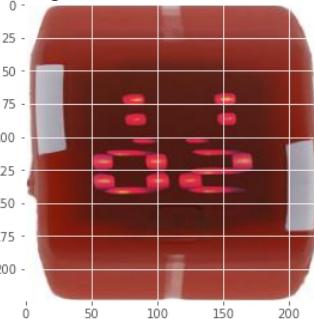


Image de catégorie Kitchen & Dining identifiée comme Watches



Approche CNN pour les images

- Compilation et entraînement :
- Résultat bonne classification :

Image de catégorie Kitchen & Dining identifiée comme telle

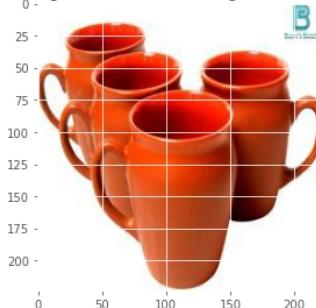


Image de catégorie Baby Care identifiée comme telle

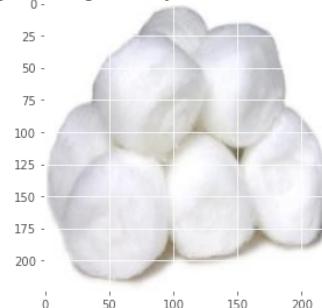


Image de catégorie Beauty and Personal Care identifiée comme telle



Image de catégorie Computers identifiée comme telle

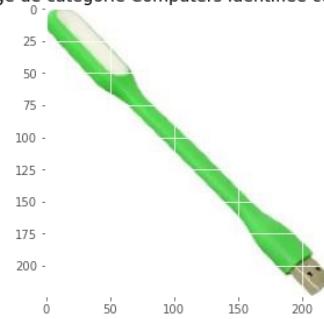


Image de catégorie Watches identifiée comme telle



Aller plus loin

- Si nous obtenions de bons résultats pour le texte et les images nous pourrions regarder la combinaison des 2 approches. Cependant ce n'est pas le cas.
- Nous pourrions pour cela utiliser des modèles supervisés comme Random Forest ou encore le transfert learning en utilisant la description des produits (plus précisément le résultat des différents cluster pour chaque catégorie) comme label.

Conclusion

- Un système de classification qui fonctionne bien n'est pas envisageable si on utilise seulement des modèles de clustering classique vue nos résultats. Nous pouvons rapidement voir les limites dans ce travail.
- En utilisant des algorithmes mieux adaptés, K-Prototype pour le texte ou en utilisant le transfert learning pour les images, nous obtiendrons de meilleurs résultats bien que des erreurs subsistes encore, notamment pour le texte.
- Une meilleure catégorisation métier serait intéressante aussi, comme on a pu voir avec les images.
- Volume de données faible ce qui limite les performances.



Merci de votre attention