

Classifier automatiquement  
des biens de consommation

**Projet : 6**

**Présentée par: LYNDA HADJEMI**

**Parcours: Data science**

# Sommaire

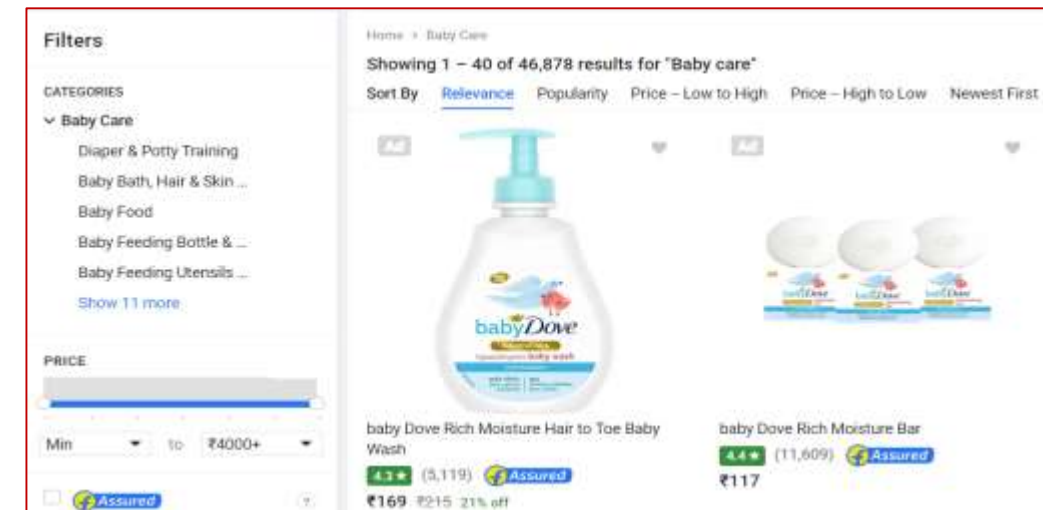
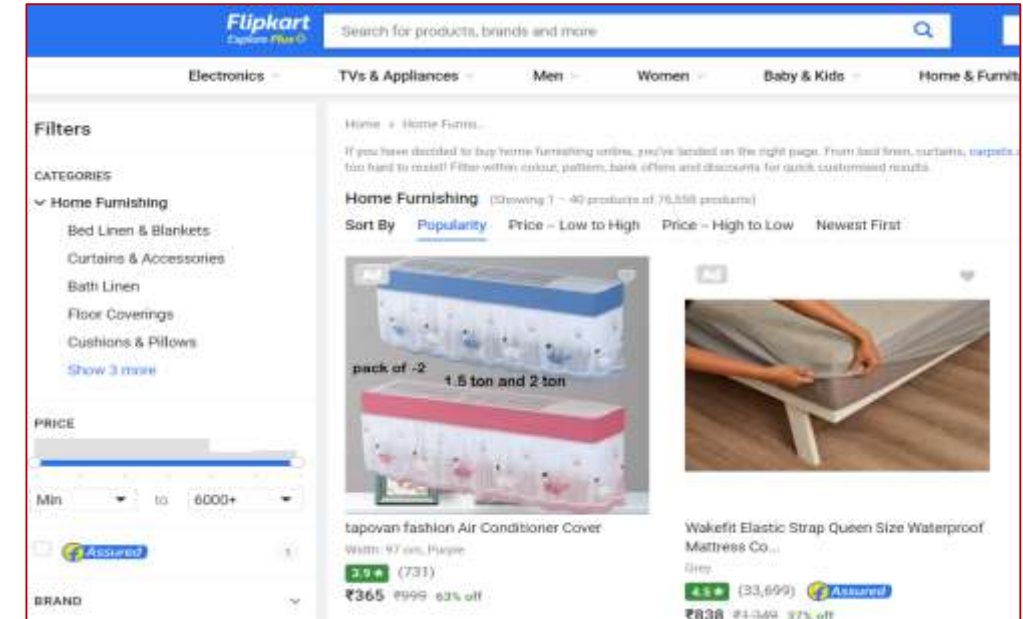
- Contexte et problématique .
- Jeu de données et analyse exploratoire
- Traitement des données textuelles
- Traitement des données images
- Classification non supervisée
- Classification supervisée
- Test de API
- Comparaison entre les modèles,
- Conclusion



- L'entreprise "Place de marché", souhaite lancer une marketplace e-commerce,
- Réaliser une étude de faisabilité d'un moteur de classification d'article basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article

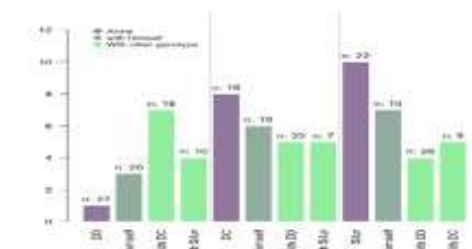
## Objectif:

- ❖ Assurer la fiabilité des articles avec une catégorisation précise.
- ❖ La fluidité de la recherche pour les clients et la facilité de mise en place de nouveau produit pour les vendeurs.



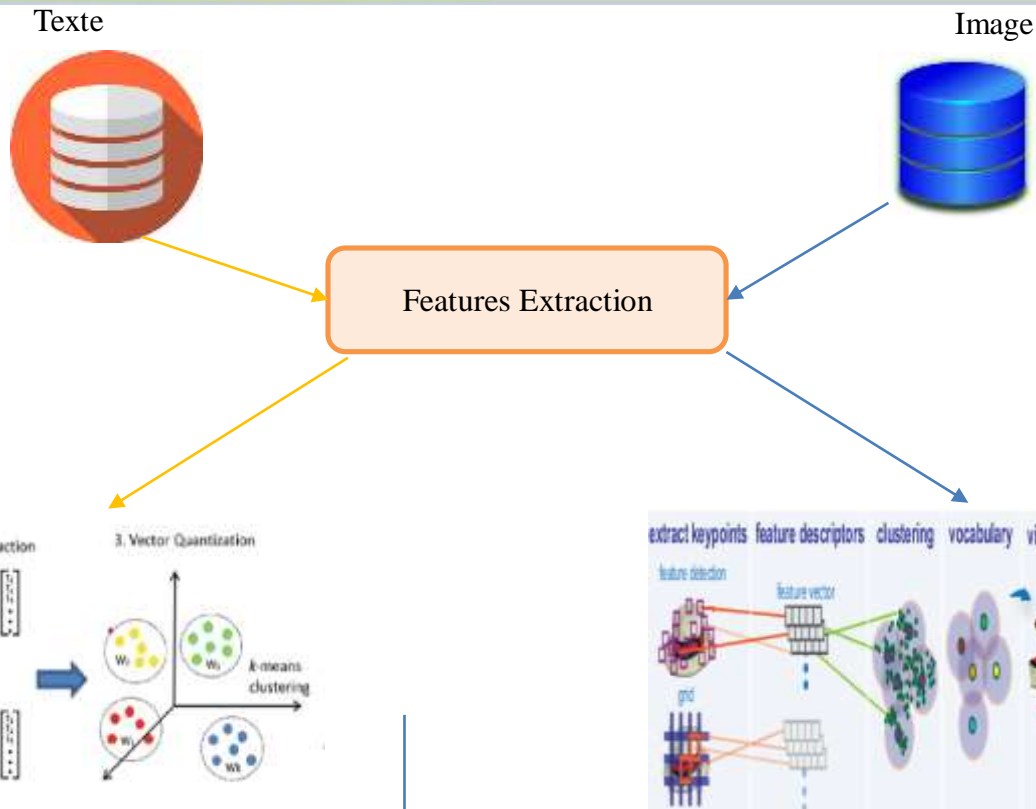
# Les démarches suivis:

- Traitement de deux types de données: images et textes.
- Réduction de dimension.
- Projections des données sur le plan 2D et 3D.
- Apprentissage Non supervisée.
- Apprentissage supervisée.
- Calculs des métriques et comparer les algorithmes.
- La collecte des données par l'Api.



	central	turns	background	chips	spells	coast	background
central	89.00	1.71	0.00	0.00	0.11	0.00	8.91
turns	0.00	70.16	0.04	0.00	0.00	0.21	28.33
background	2.60	0.04	81.44	0.00	0.00	0.00	32.20
chips	0.00	0.04	0.00	73.11	0.00	15.61	9.24
spells	1.50	1.69	0.23	0.00	66.61	0.52	16.45
coast	4.73	0.22	0.15	0.01	0.00	73.00	21.79
background	0.00	0.00	0.52	0.00	0.38	0.00	93.20

Calculs des métriques d'évaluations

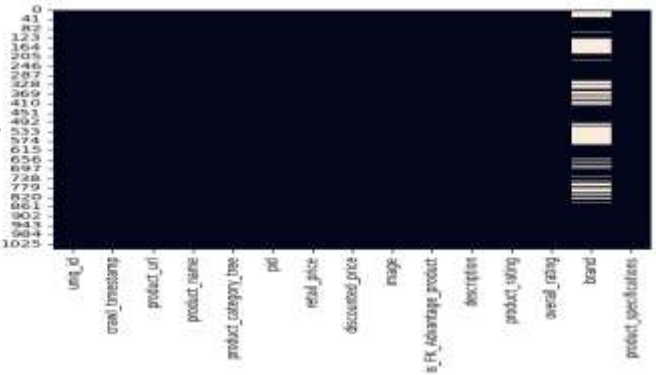




# Jeu de données:

- ❑ Base de données (e-commerce indien Flipkart)
- ❑ 1050 données et 15 colonnes.
- ❑ Données texte et images.
- ❑ Plusieurs niveaux des catégories.
- ❑ Equilibrée entre catégories.
- ❑ Suffisant pour une étude de faisabilité

Peu de données manquante dans la colonne Brand, environ 3%.



Product\_category\_tree

["Home Furnishing >> Curtains & Accessories >>...

Image



Description

Key Features of Elegance Polyester Multicolor

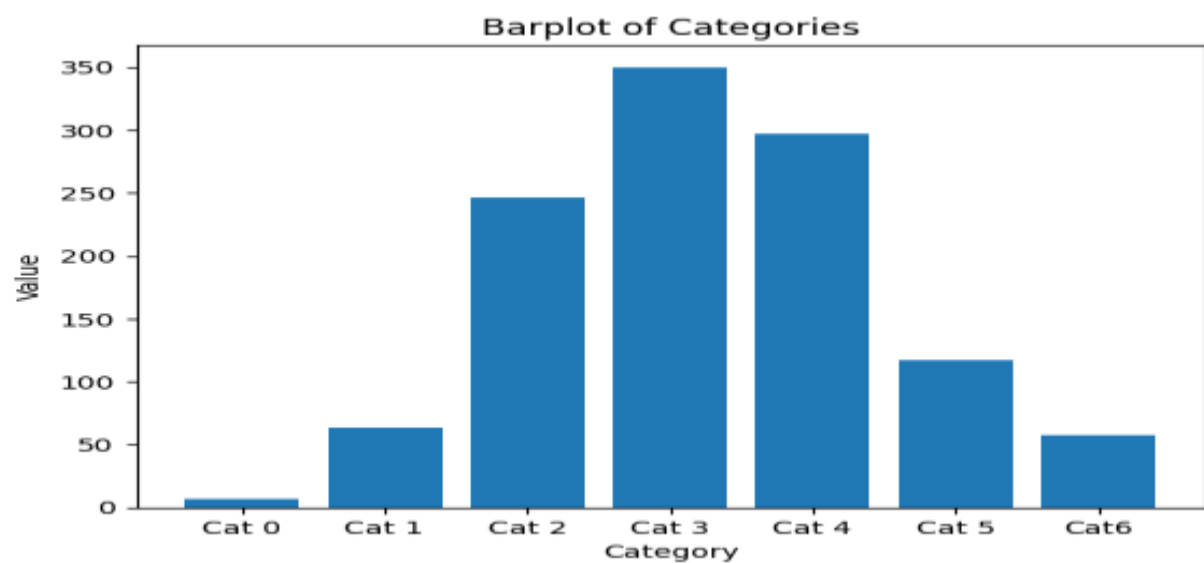
Product-name

Sathiyas Cotton Bath Towel





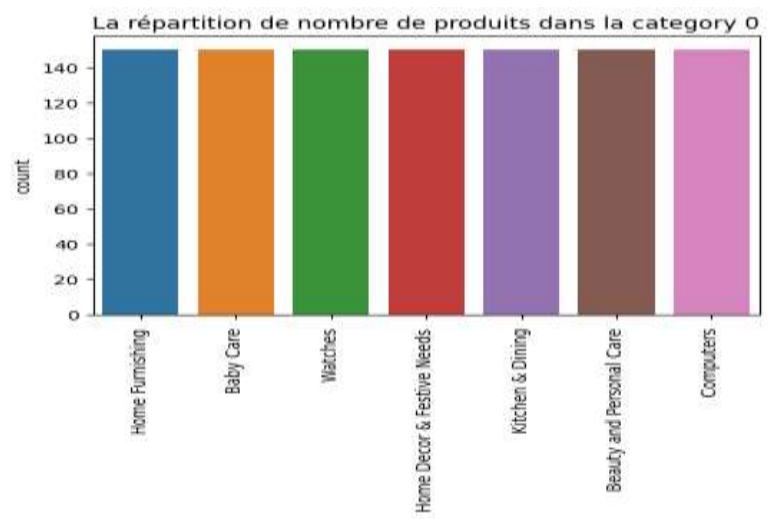
# Données cibles:



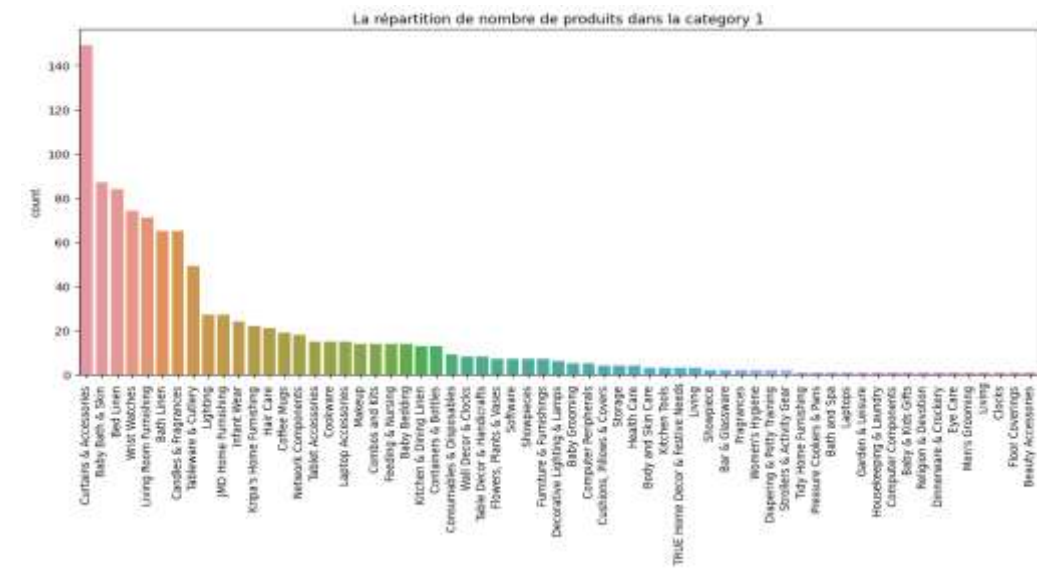
Nombres des produits dans chaque catégories

category	value
Cat 0	7
Cat 1	63
Cat 2	246
Cat 3	350
Cat 4	297
Cat 5	117
Cat 6	57

- 7 catégories
- 150 produits par catégories



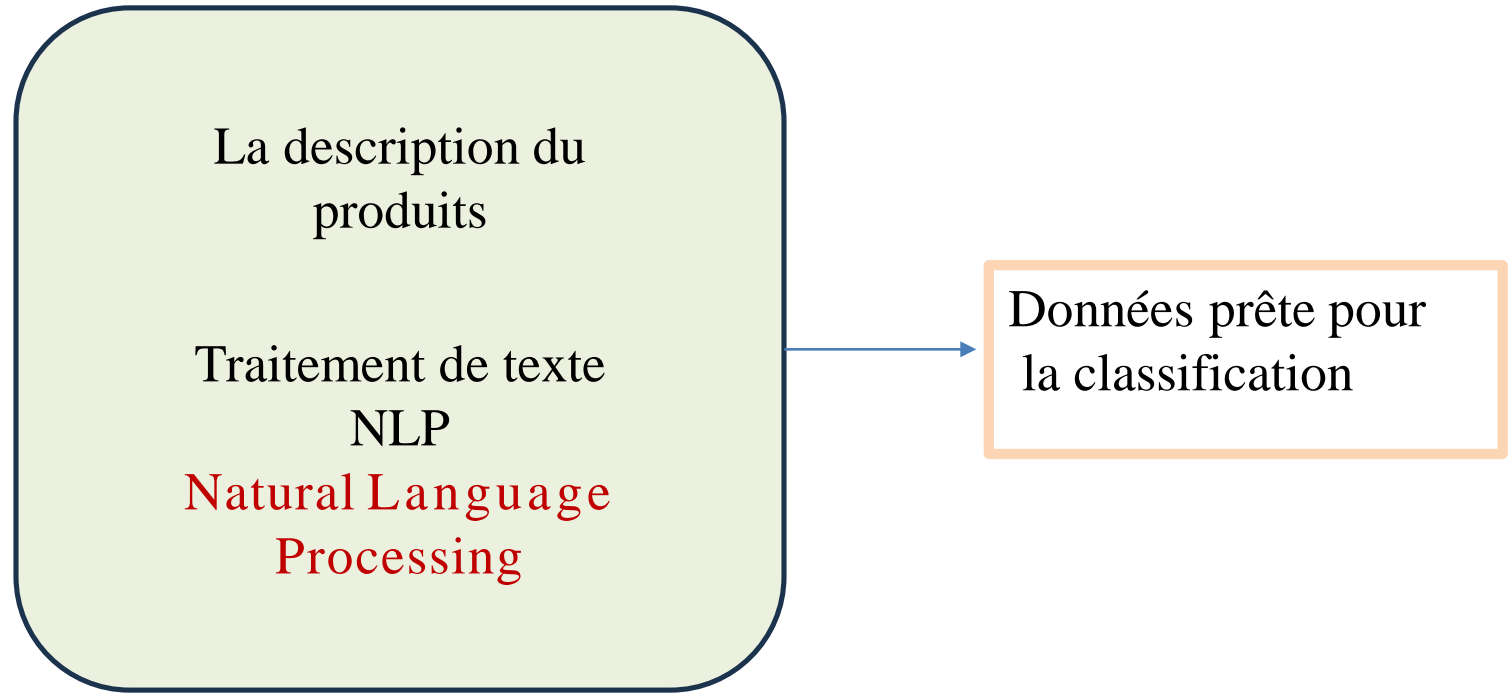
- 63 catégories différentes.
- Nombres de produits différents





## Etapes de NLP:

- Tokenisation NLTK.
- Stop-words.
- Mis en minuscule.
- Lemmatisation.
- Racinisation(Stemmatisation).

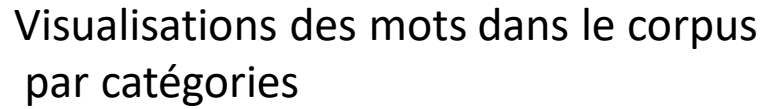




Exemple de traitement:

<b>Exemple de document dans le corpus</b>	<p>'Specifications' of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton bath towel Ideal For Men, Women, Boys, Girls Model ID Color Red, Yellow, Blue Size Mediam Dimensions Length 30 inch Width 60 inch In the Box Number of Contents in Sales Package 3 Sales Package 3 Bath Towel</p>
<b>Tokenisation avec NLTK</b>	<p>['Specifications', 'of', 'Sathiyas', 'Cotton', 'Bath', 'Towel', '(', '3', 'Bath', 'Towel', ',', 'Red', ',', 'Yellow', ',', 'Blue', ')', 'Bath', 'Towel', 'Features', 'Machine', 'Washable', 'Yes', 'Material', 'Cotton', 'Design', 'Self', 'Design', 'General', 'Brand', 'Sathiyas', 'Type', 'Bath', 'Towel', 'GSM', '500', 'Model', 'Name', 'Sathiyas', 'cotton', 'bath', 'towel', 'Ideal', 'For', 'Men', ',', 'Women', ',', 'Boys', ',', 'Girls', 'Model', 'ID', 'Color', 'Red', ',', 'Yellow', ',', 'Blue', 'Size', 'Mediam', 'Dimensions', 'Length', '30', 'inch', 'Width', '60', 'inch', 'In', 'the', 'Box', 'Number', 'of', 'Contents', 'in', 'Sales', 'Package', '3', 'Sales', 'Package', '3', 'Bath', 'Towel']</p>
<ul style="list-style-type: none"><li>▪ <b>Mis en minuscule.</b></li><li>▪ <b>suppression ponctuation.</b></li><li>▪ <b>Suppression des chiffres.</b></li></ul>	<p>[specifications, sathiyas, cotton, bath, towel, bath, towel, red, yellow, blue, bath, towel, features, machine, washable, yes, material, cotton, design, self, design, general, brand, sathiyas, type, bath, towel, gsm, model, name, sathiyas, cotton, bath, towel, ideal, men, women, boys, girls, model, id, color, red, yellow, blue, size, mediam, dimensions, length, inch, width, inch, box, number, contents, sales, package, sales, package, bath, towel]</p>
<b>Mis en phrase et lemmatisation</b>	<p>specification sathiyas cotton bath towel bath towel red yellow blue bath towel feature machine washable yes material cotton design self design general brand sathiyas type bath towel gsm model name sathiyas cotton bath towel ideal men woman boy girl model id asvtwl322 color red yellow blue size mediam dimension length inch width inch box number content sale package sale package bath towel</p>
<b>Mis en phrase et stemmatisation</b>	<p>specif sathiya cotton bath towel bath towel red yellow blue bath towel featur machin washabl ye materi cotton design self design gener brand sathiya type bath towel gsm model name sathiya cotton bath towel ideal men women boy girl model id asvtwl322 color red yellow blue size mediam dimens length inch width inch box number content sale packag sale packag bath towel</p>





Usage de mots pour la catégorie "Computers"

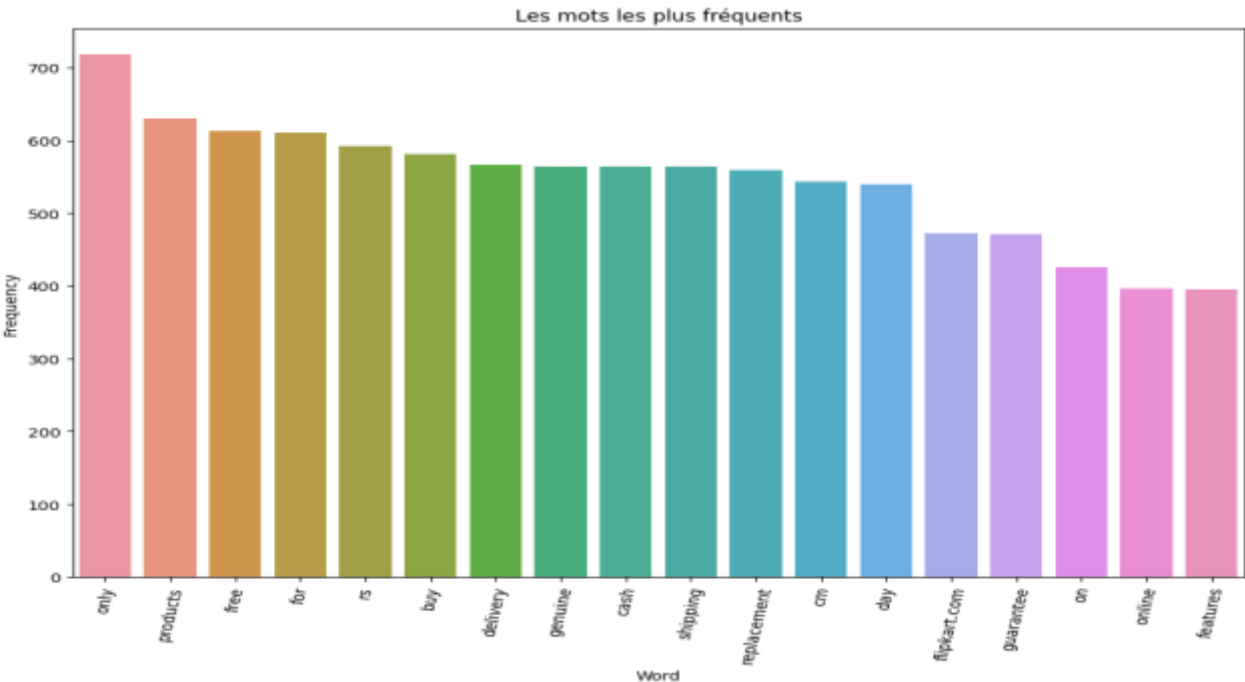
sale package charger vaio laptop battery  
 combo set skin mouse best price  
 9% charger lapguard apple macbook flexible  
 general brand keyboard high quality  
 led print shape  
 usb vaio cell laptop light mouse pad  
 covered warranty model name laptop adapter  
 warranty warranty  
 warranty summary 5v 9a pad combo

Word cloud visualization of product features for 'Home Furnishing'. The words are arranged in a circular pattern. The most prominent words are 'price', 'cotton', 'cushion', 'cover', 'package', 'sale', 'towel', 'color', 'multicolor', 'bath towel', 'design code', 'curtain height', 'specification', 'polyester', 'made', 'printed', 'width inch', 'general', 'brand', 'green', 'box number', 'floral', 'content', 'soft', 'blanket', 'style code', 'aroma', 'comfort', 'door', and 'curtain'. The word 'price' is highlighted with a red circle.

[illegible][illegible]



# Fréquence de mot:



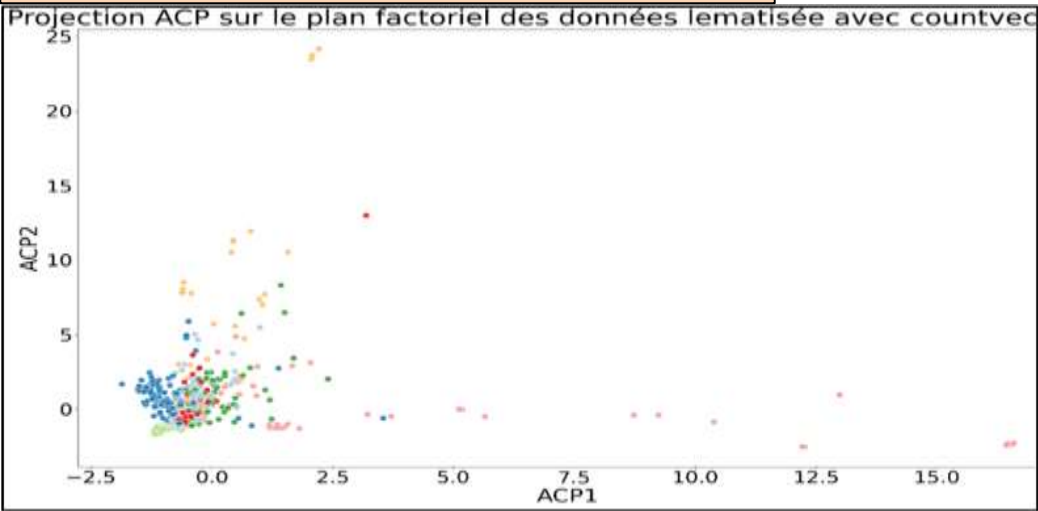
18 mot plus fréquent dans le corpus.

	Word	Frequency	%_frequency
4052	sand	3	0
2790	data	3	0
4075	sober	3	0
5115	jums	3	0
814	grapefruit	3	0
4752	st1025sl07	3	0
2808	seasons	3	0
5152	therapy	3	0
4008	ishita	3	0
4106	bristles	3	0
788	jmd	3	0
...			

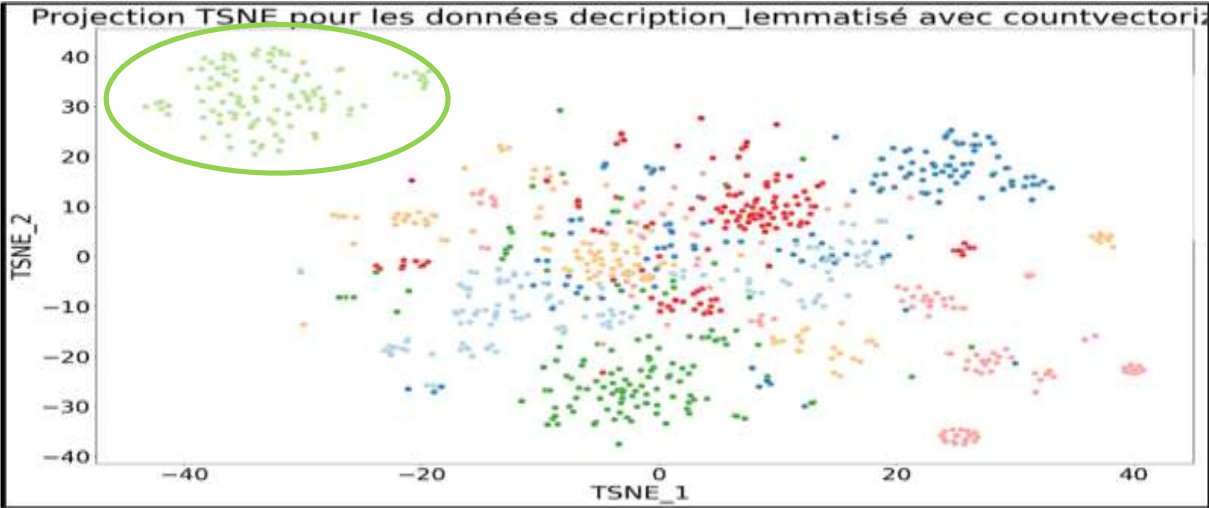
On trouve beaucoup de tokens qui sont rare dans le corpus, environ 4120 de tokens ont une fréquence  $\leq 0$



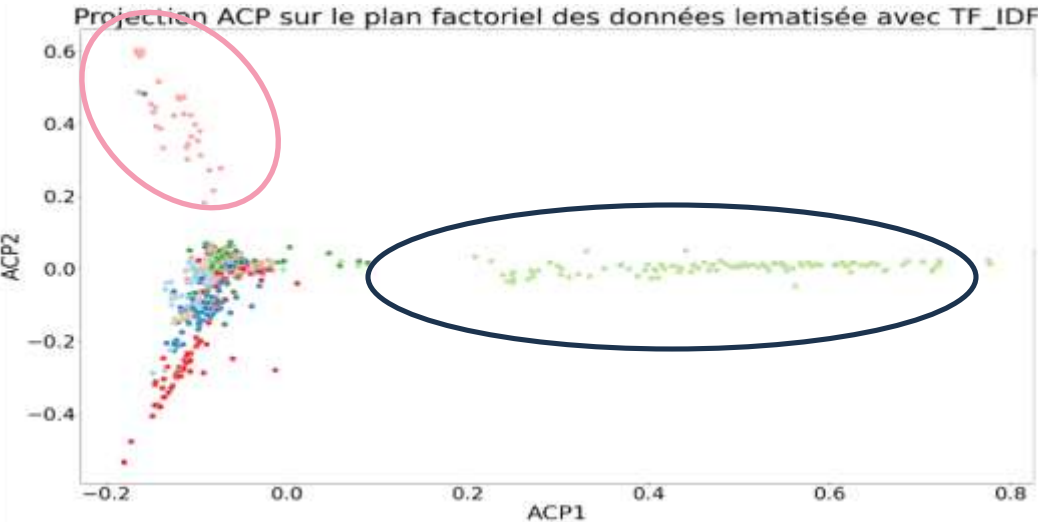
Combinaison: ACP + Countvectorizer



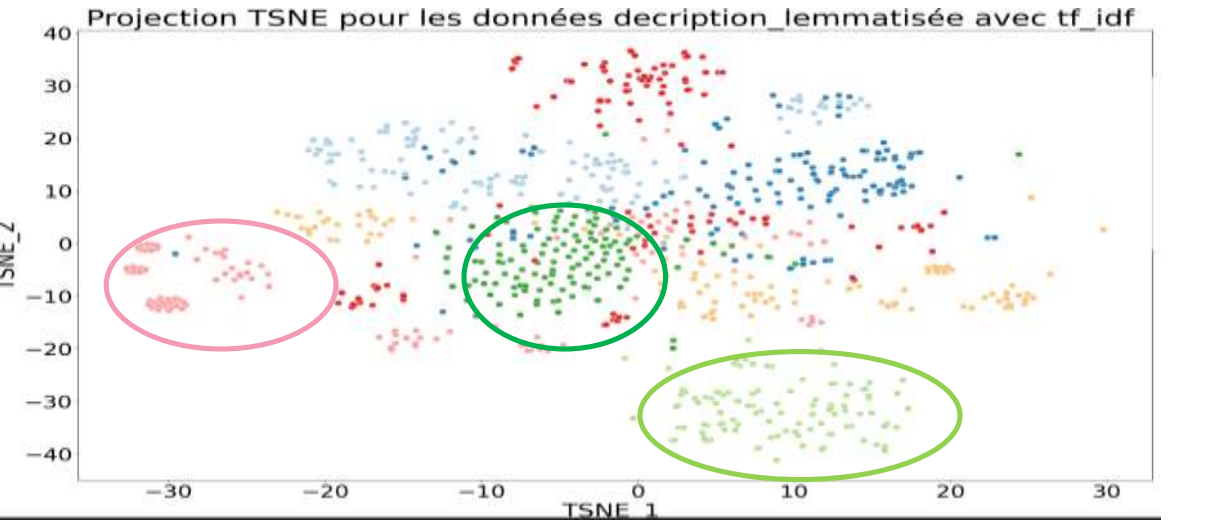
Countvectorizer + TSNE



TFIDF + PCA:

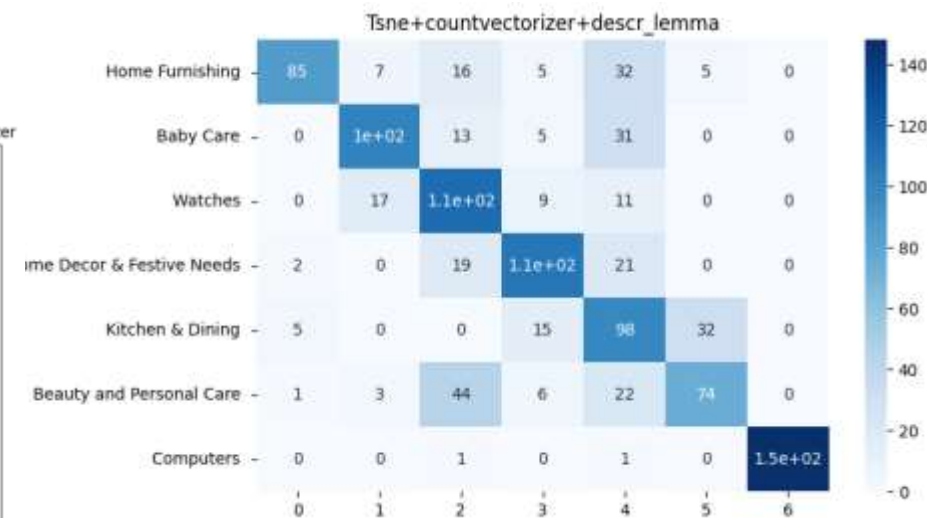
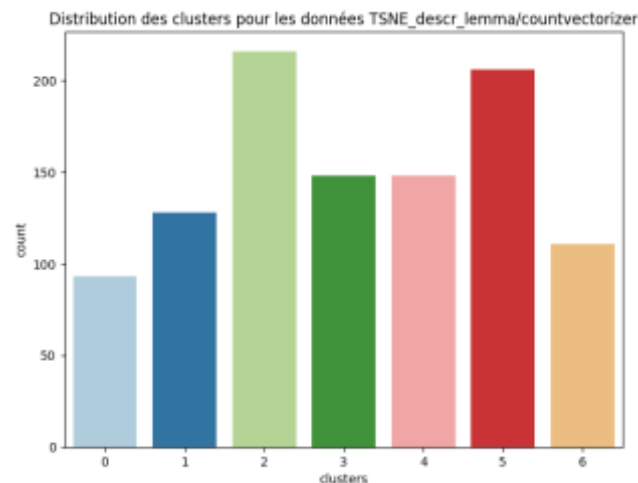
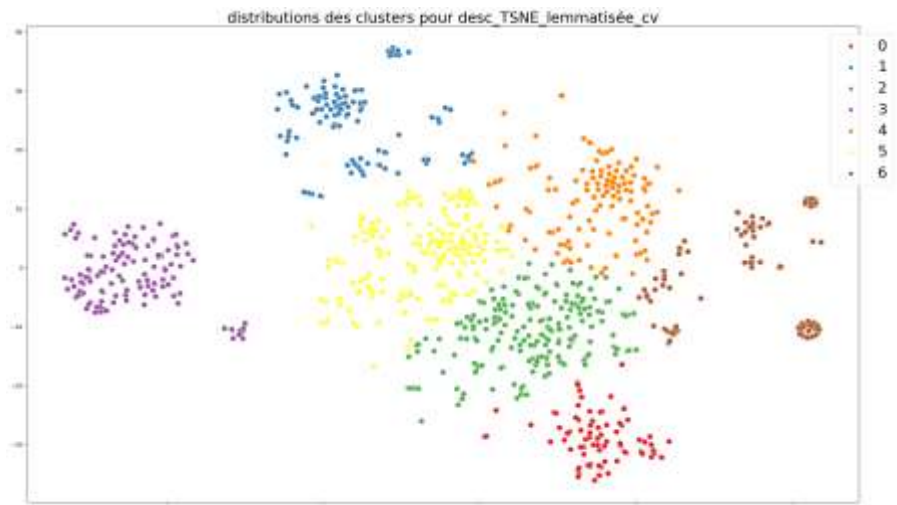


TFIDF + TSNE:





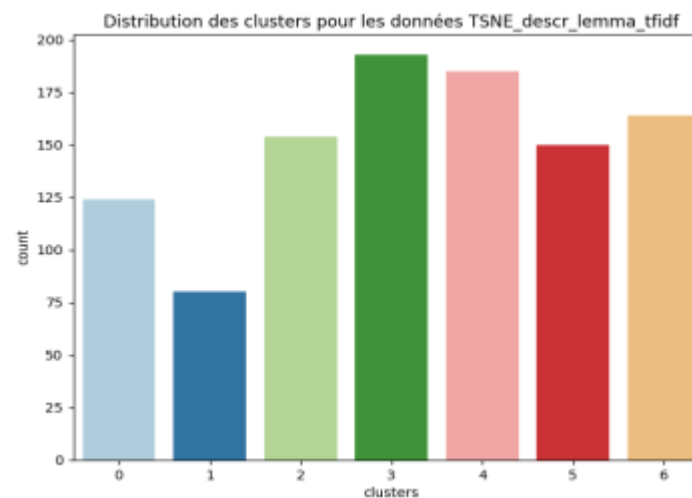
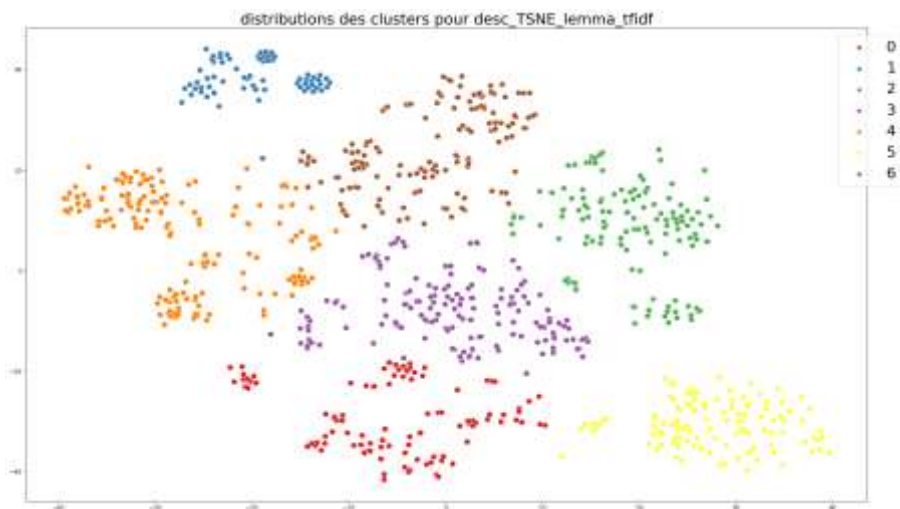
## Combinaison:KMEANS + TSNE+CountVectorizer



ARI =0,44

Accuracy = 69%

## Combinaison :KLEANS + TSNE + TFIDF



ARI: 0,48

accuracy = 73%



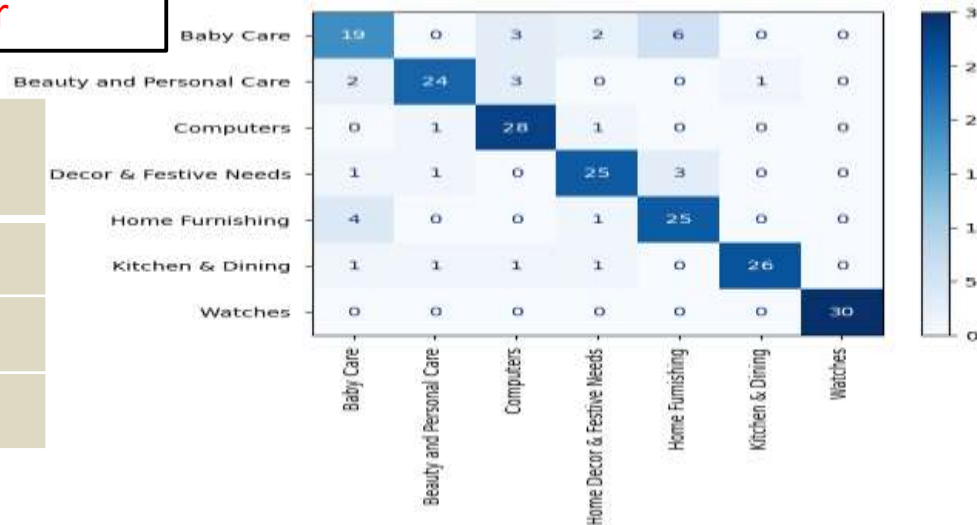
# Classification Supervisée



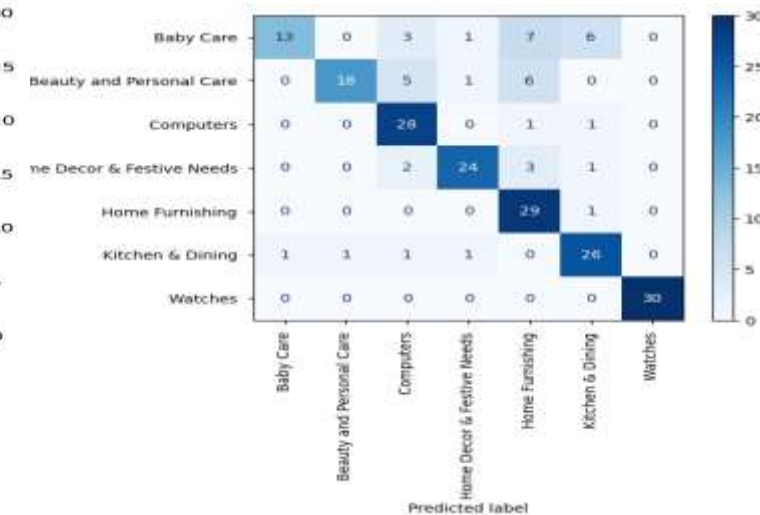
## Combinaison: TSNE + Countvectorizer

Modèles	KNeighborsClassifier	SVM
Accuracy	91%	80%
Train_score	91%	80%
Test_score	84%	80%

Modèle KNeighbors



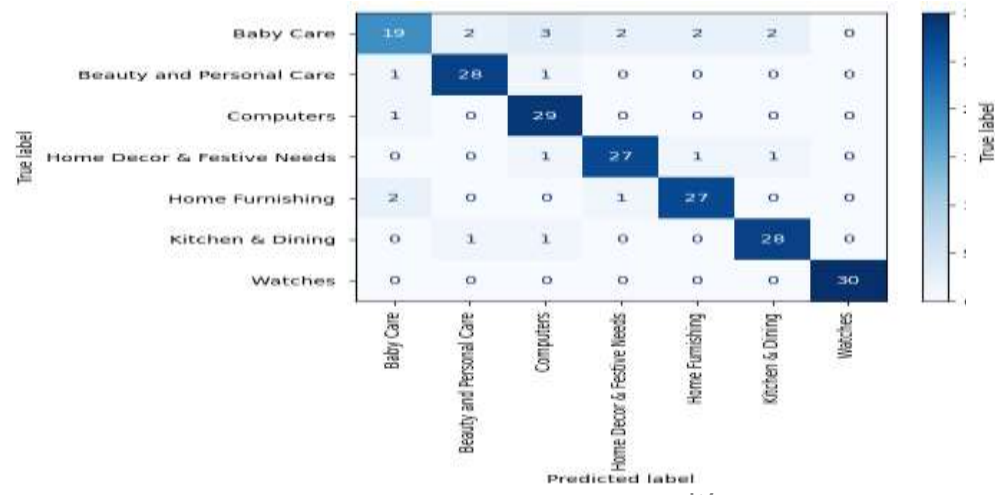
Modèle SVM



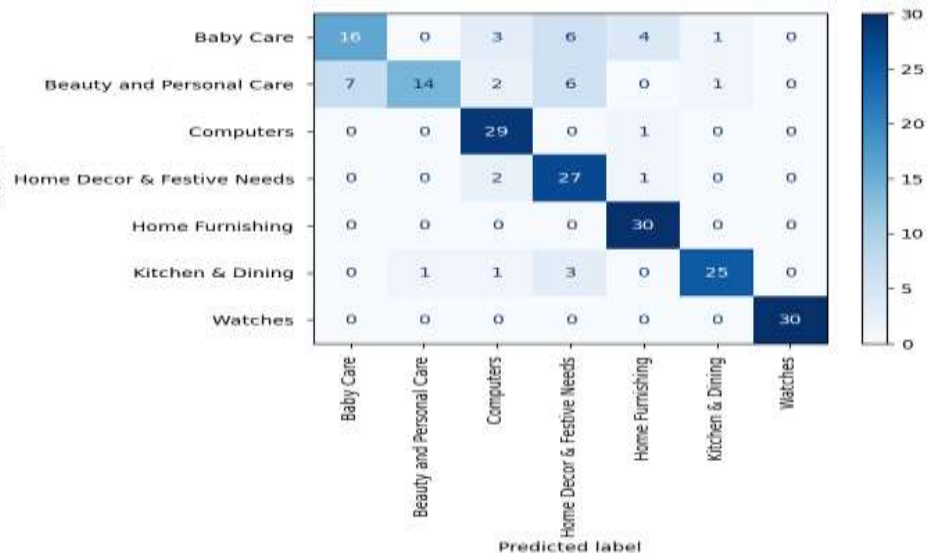
## Combinaison: TSNE +TFIDF

Modèles	KNeighborsClassifier	SVM
Accuracy	94%	83%
Train_score	94%	82%
Test_score	89%	81%

Modèles kneighbors



Modèle SVM





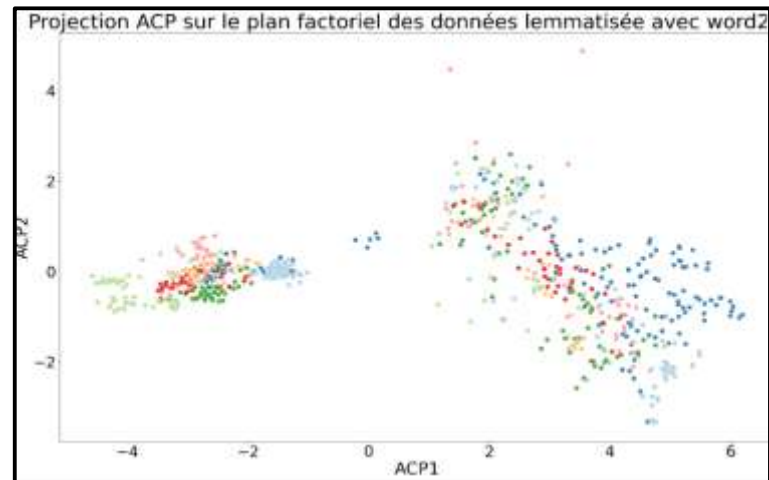
# Modèles embeddings



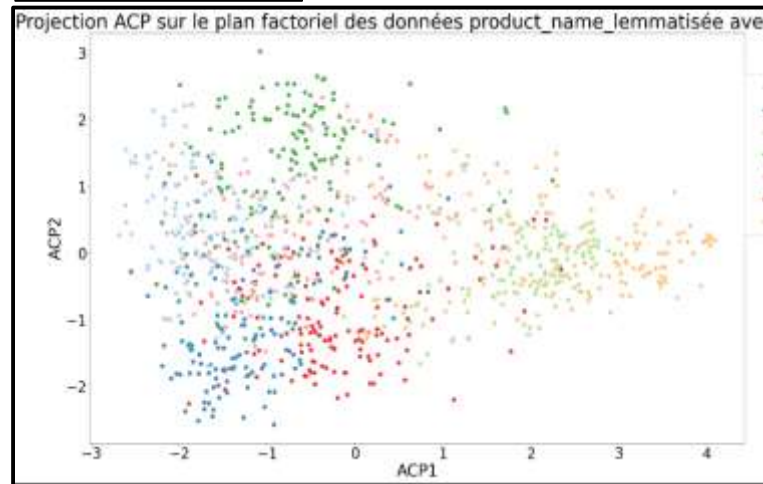


# Features embedding et projections

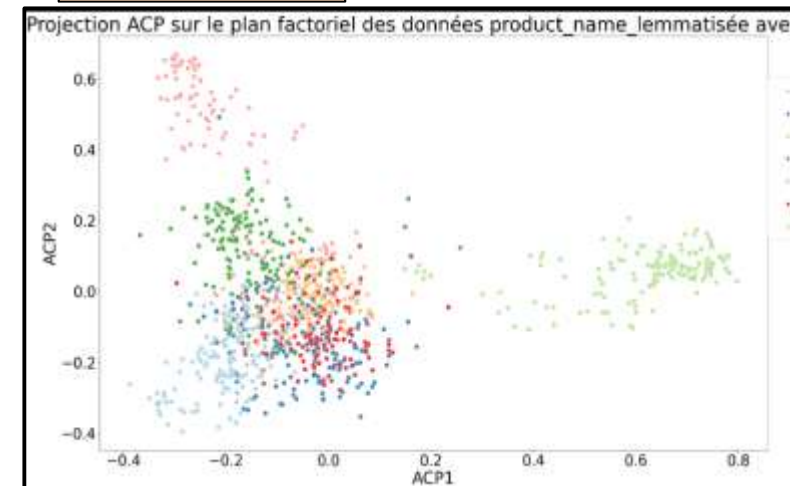
Combinaison: ACP + word2vec



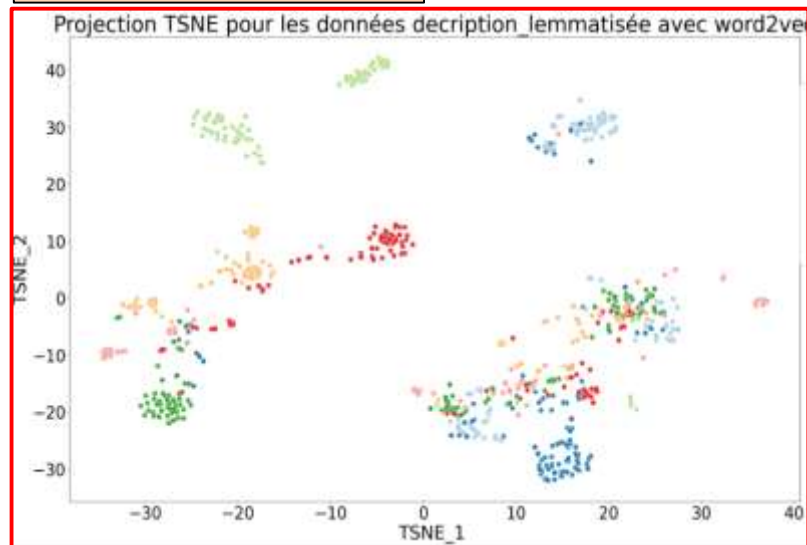
ACP + Bert



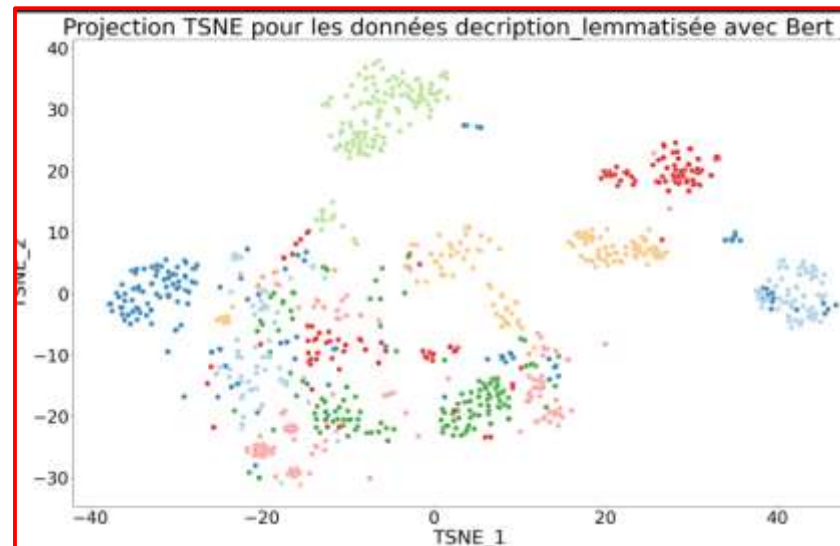
ACP+ USE



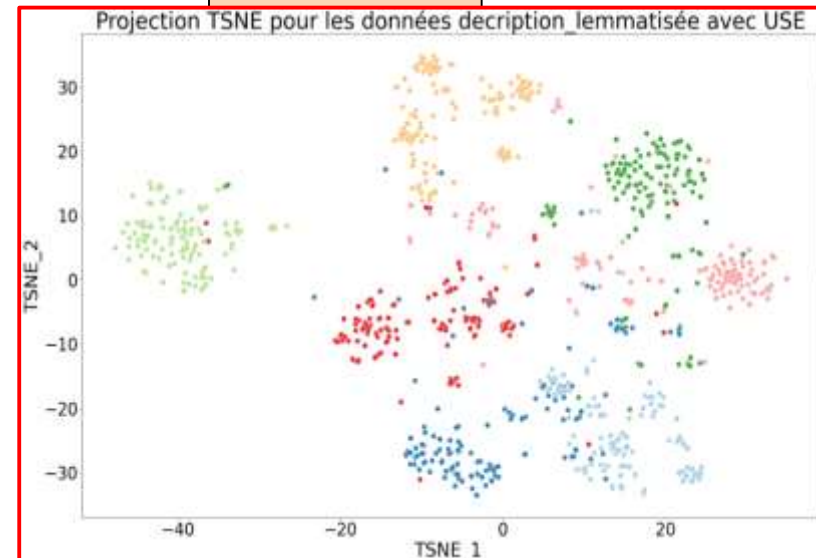
TSNE + word2vec



Bert+ TSNE:



TSNE+ USE

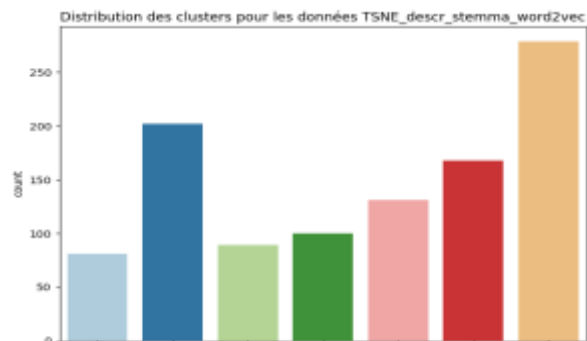
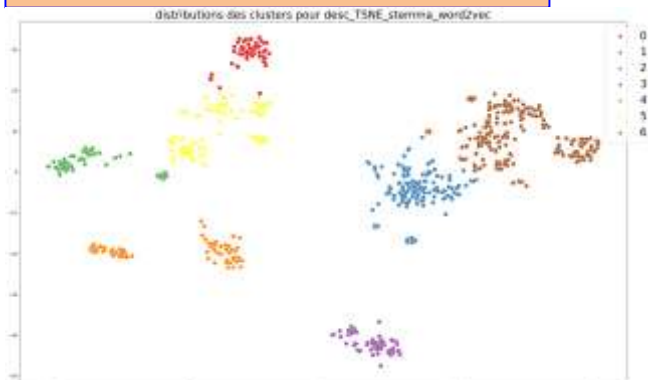




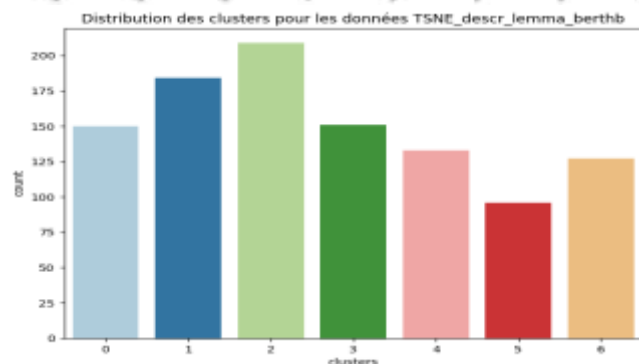
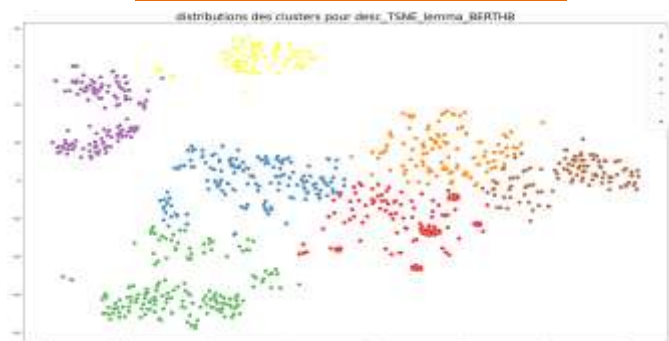
# Classification Non Supervisée KMEANS



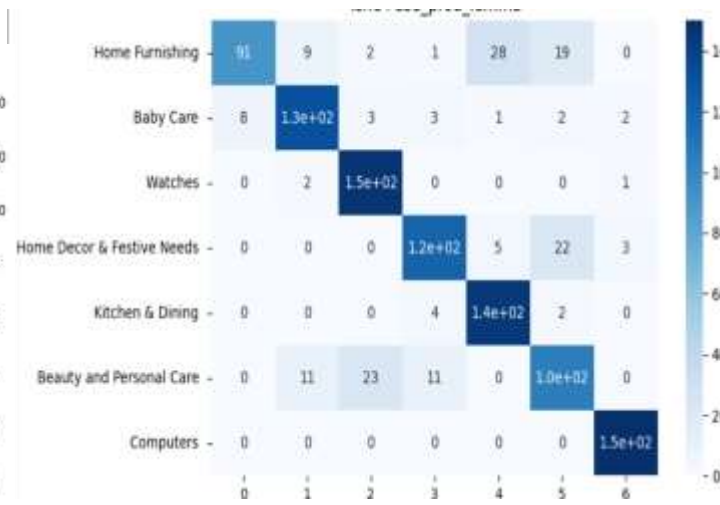
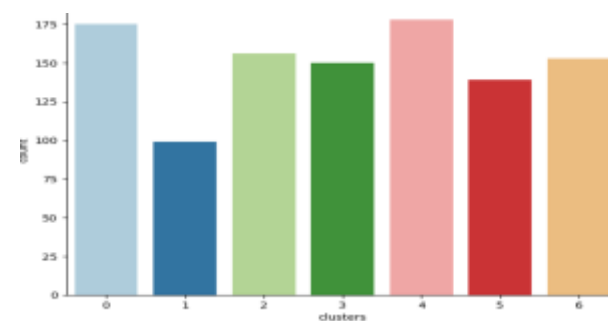
## KMEANS+TSNE+WORD2VEC



## KMEANS+TSNE+BERT



## KMEANS+TSNE+USE

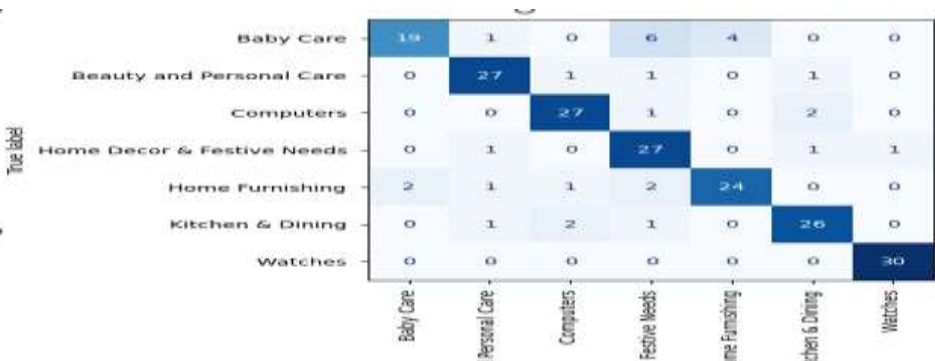


Les scores obtenus

Modél	ARI	Accuracy
w2vec	26%	55%
Bert	41%	62%
Use	52%	80%

# Classification Supervisée

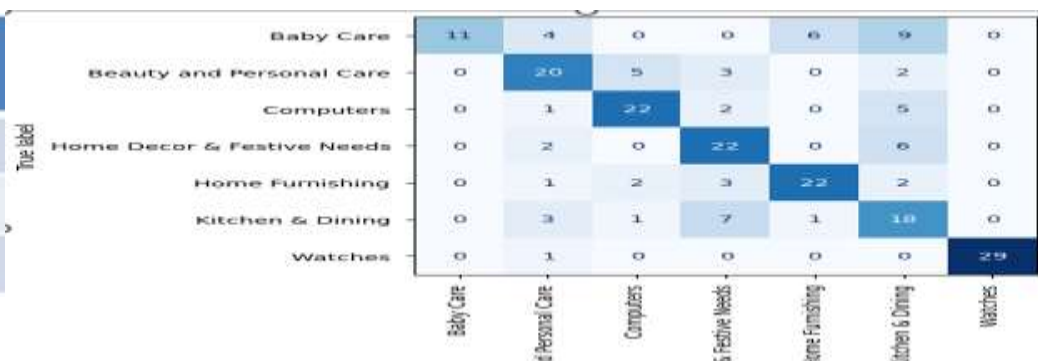
Modèles kneighboors



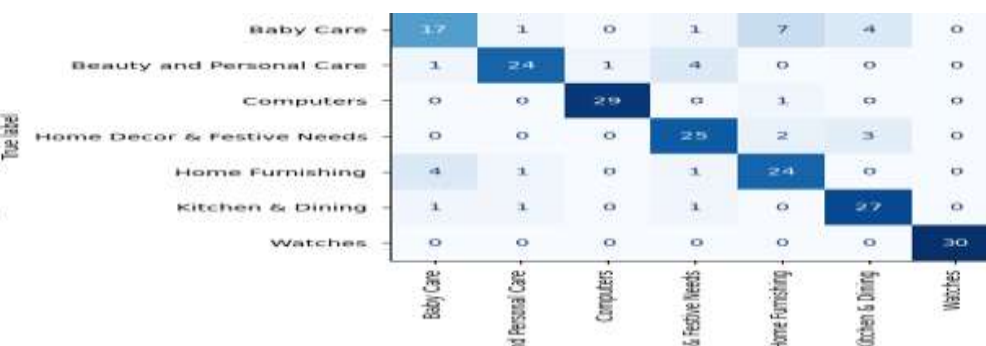
Modèle Word2Vec

Modél	kneighboors	SVM
Acuracy	91%	71%
Train score	90%	70%
Test score	85%	68%

Modèle SVM



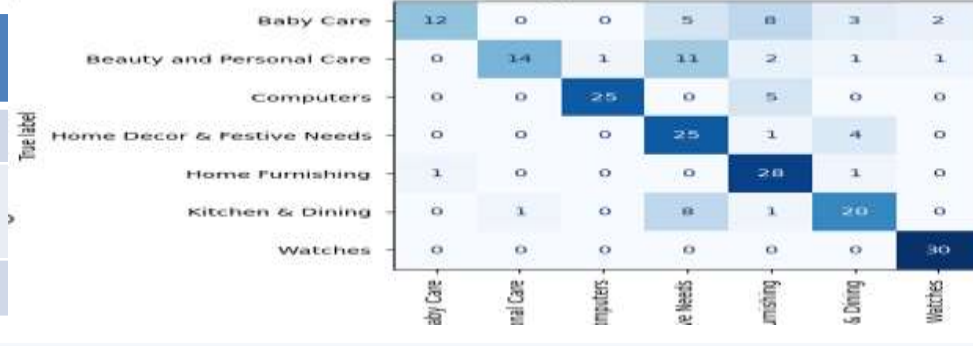
Modèles kneighboors



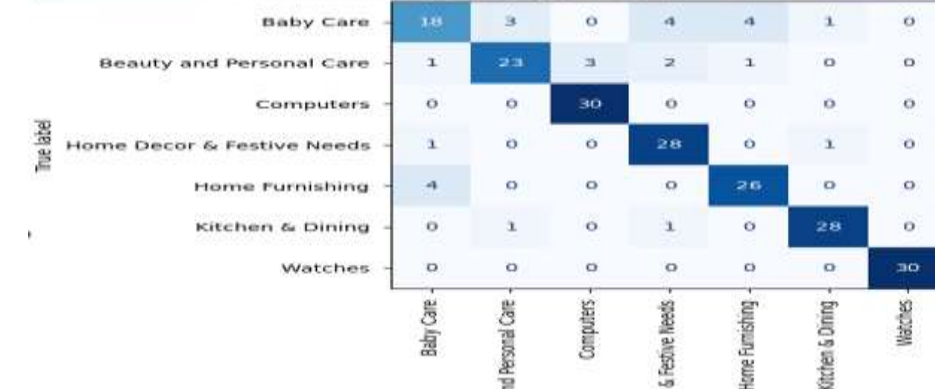
Modèle de Bert

Modél	kneighboors	SVM
Acuracy	91%	76%
Train score	91%	75%
Test score	83%	73%

Modèle SVM



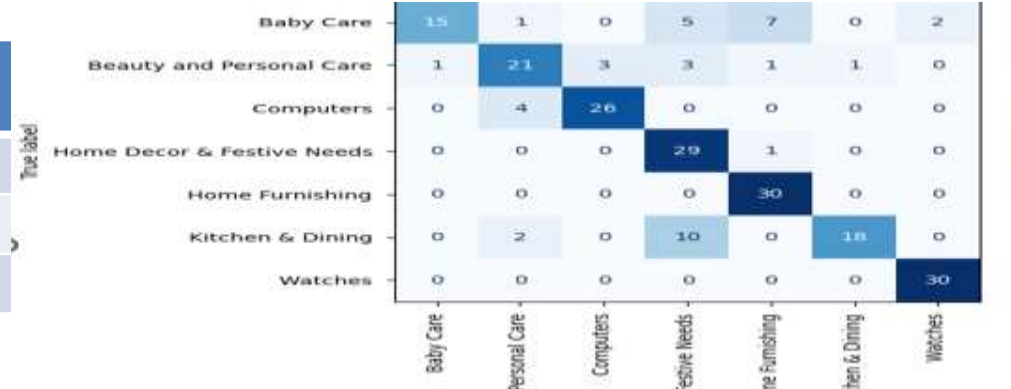
Modèles kneighboors



Modèle de USE

Modél	kneighboors	SVM
Acuracy	94%	84%
Train score	93%	83%
Test score	88%	80%

Modèle SVM





# Partie images


## Procédure de traitements:

Données images



Prétraitement



Bags of visual words  
SIFT 

CNN transfert learning  
VGG16 

Data augmentation

Classification

Classification non  
supervisée

&

Classification supervisée





## Analyse exploratoire :

### SIFT:

Des algorithmes pour la détection et la description des caractéristiques/features dans les images.

- Outils de computer vision (détection d'objets)

### Manipulation d'images:

- Filtres, transformations, couleurs, affichages...

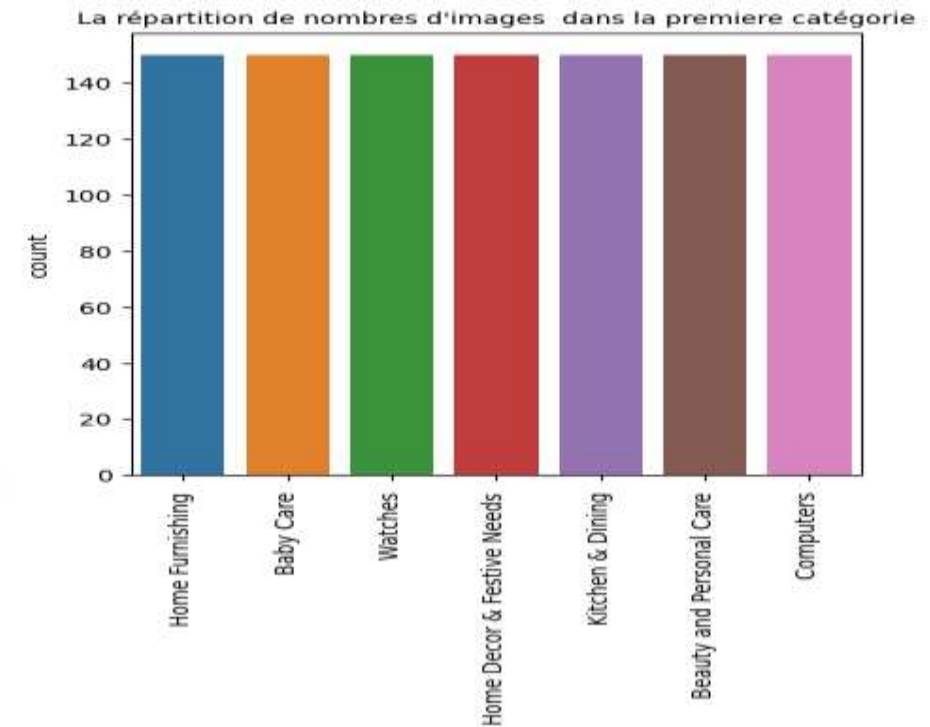
Computer Vision:

- Détection des features, objets....

Visualisation d'images:



Nombre d'images par catégorie



## Traitement d'images:

Image originale

En Gris

Egalisation

Keypoints  
Descripteurs



621 points clés  
décrits par  
descripteurs  
SIFT, vecteur de  
longueur 128





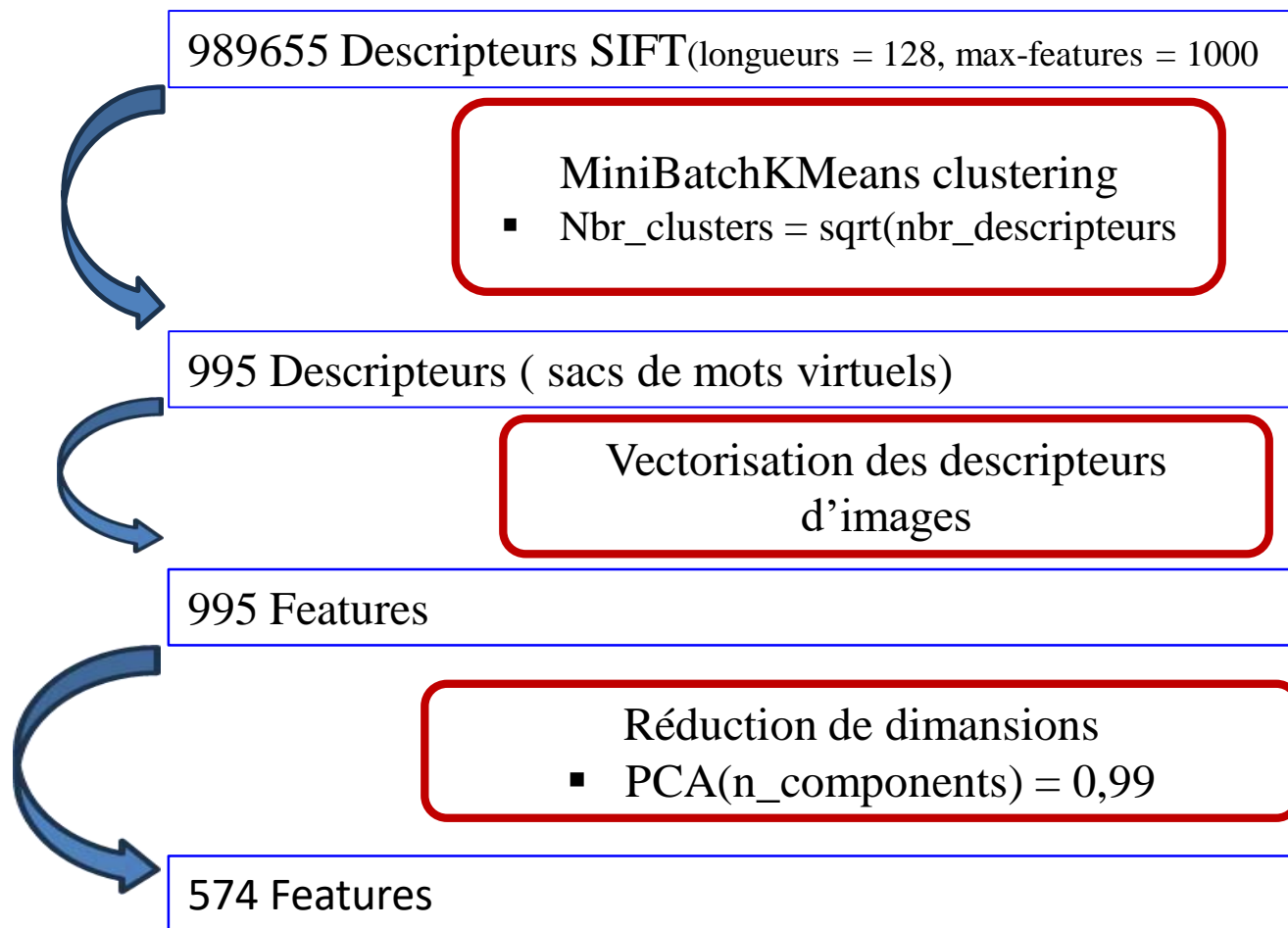
## Traitement d'images

### Création de features:

Pour le jeux de données image , nous allons crée des features pour chaque images, le résumé montres les étapes suivis

### MiniBatchKMeans:

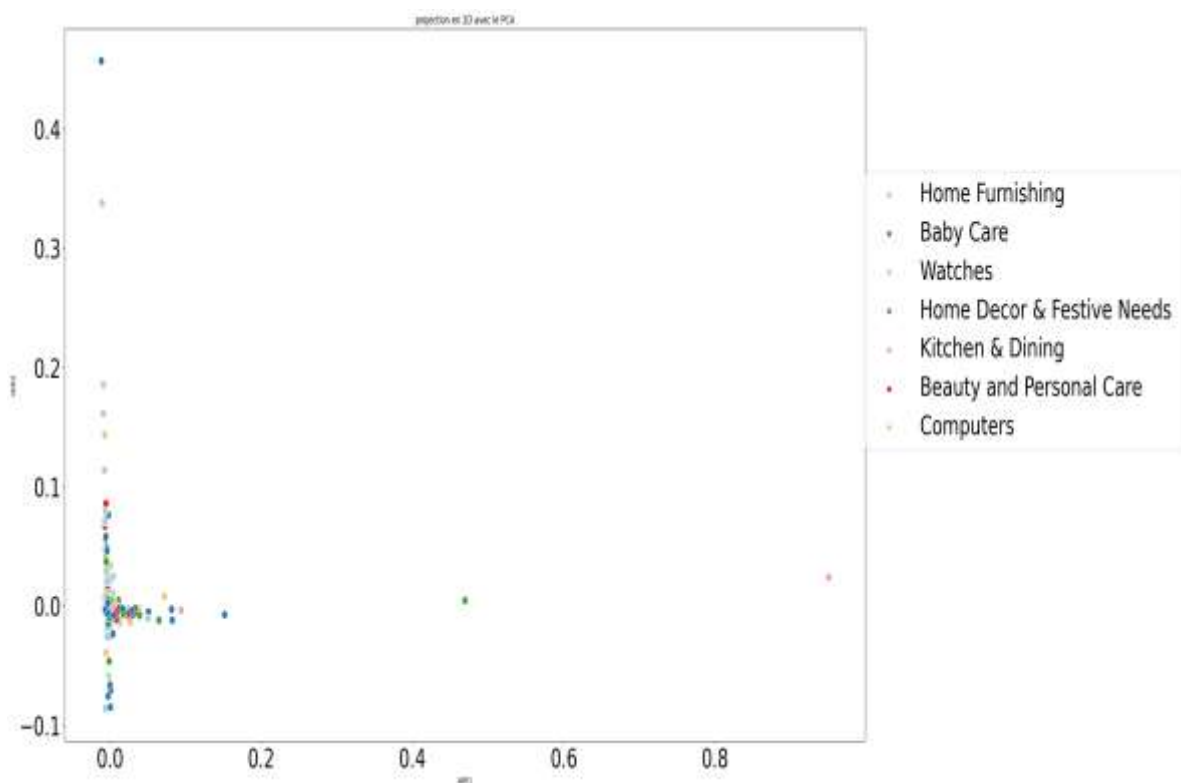
MiniBatchKMeans est un algorithme de regroupement (clustering) utilisé pour regrouper des points de données en clusters



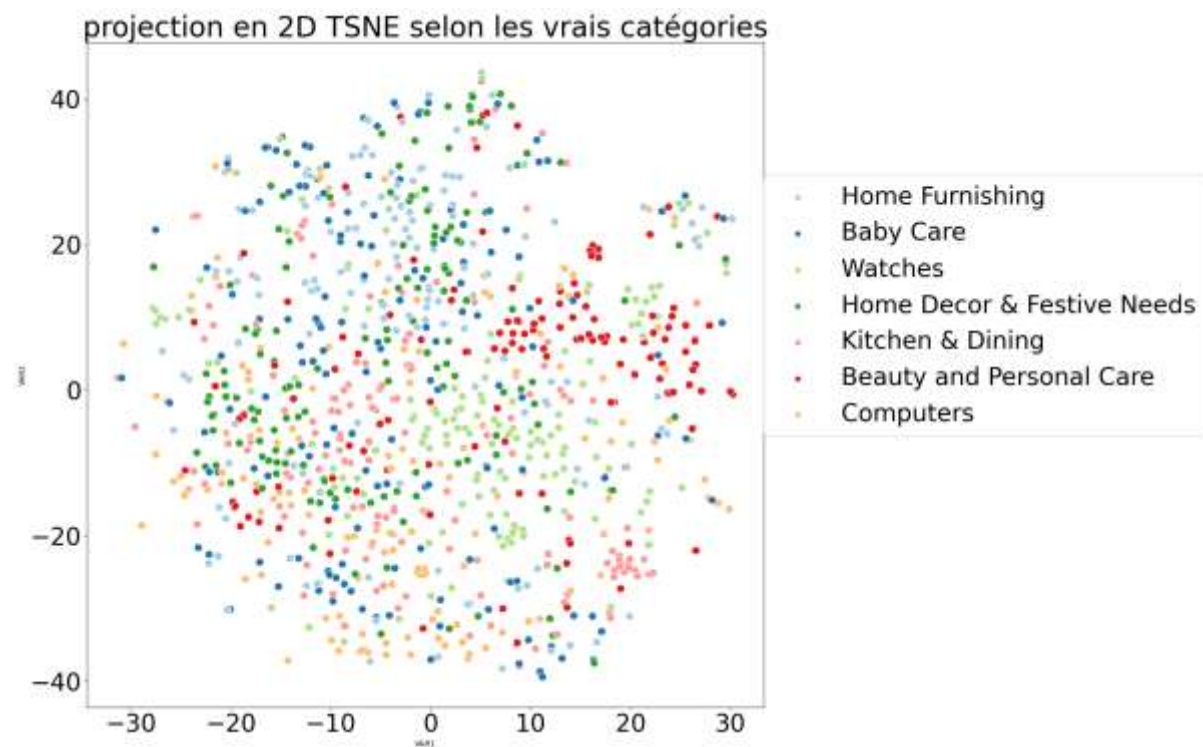


## Projection en 2D:

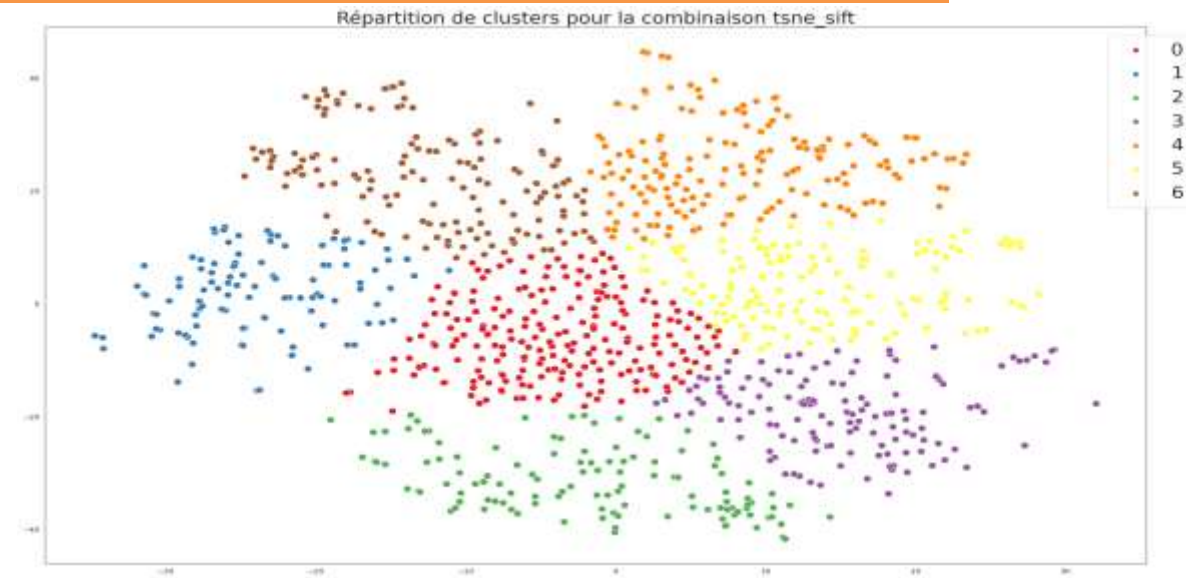
PCA + sift



TSNE+ Sift



## Combinaison: KMEANS + TSNE + SIFT



## Résultats de la combinaison: SIFT + TSNE:

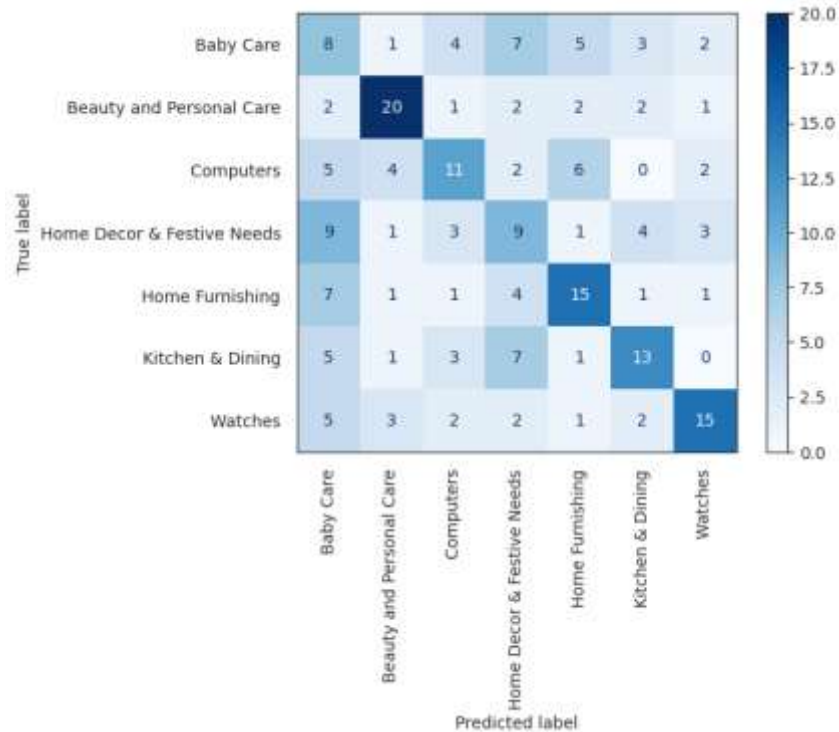
- On arrive pas à distinguer les clusters,
- Pas de regroupement en fonction des catégories
- Trop de mélanges dans les catégories.
- Des scores très faible:
- ACCURACY = 0,18 ARI = 0,05

True label	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
Baby Care	27	6	18	19	43	16	21
Beauty and Personal Care	23	51	12	8	6	35	15
Computers	35	8	26	39	11	22	9
Home Decor & Festive Needs	54	4	8	13	29	10	32
Home Furnishing	14	7	14	14	52	6	43
Kitchen & Dining	42	6	37	28	13	11	13
Watches	24	31	10	19	10	44	12
Predicted label	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches

ARI = 0,05

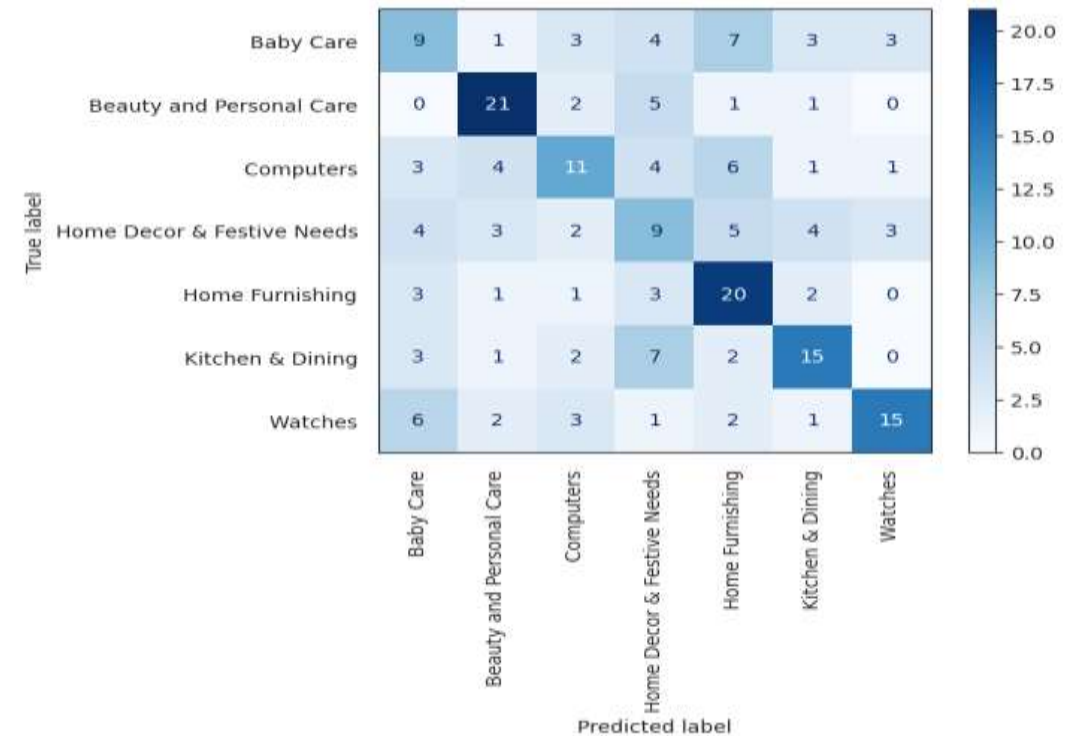
Accuracy = 0,18%

## KNeighborsClassifier\_sift\_tsne



Accuracy = 43%  
 Train\_score = 0,57  
 Test\_score = 0,473

## RandomForestClassifier\_sift\_tsne



Accuracy = 48%  
 Train\_score = 0,98  
 Test\_score = 0,47

Avec l'apprentissage supervisée on est bien arrivée a distinguer les catégories cependant avec beaucoup d'erreur et de mélanges.

- Le moteur de classification basé sur les descripteurs SIFT est beaucoup difficile à trouver les catégories pour les produits.



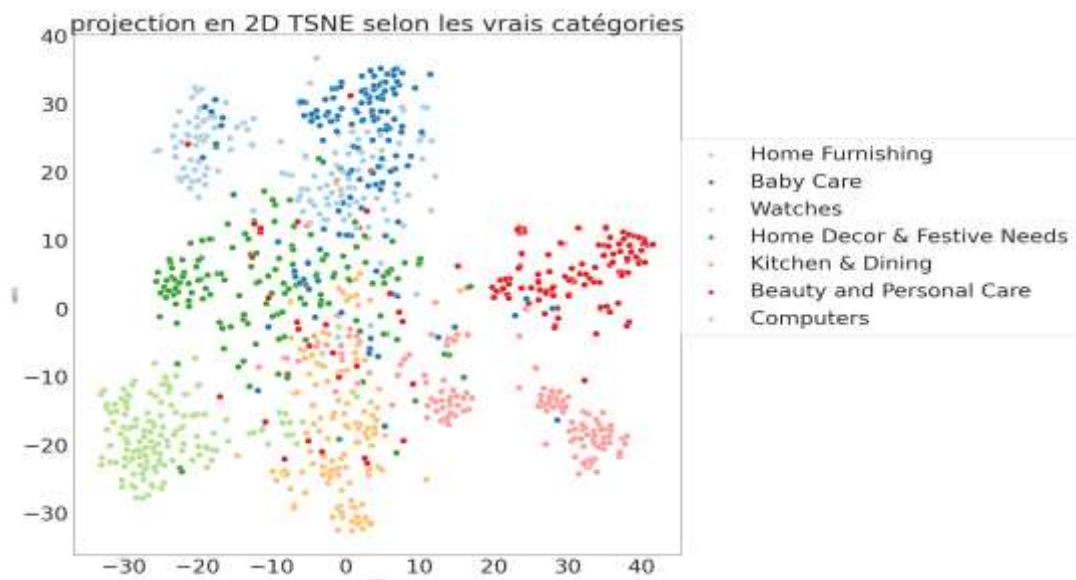
## Prétraitement des images:

-VGG16 est un réseau de neurones convolutifs (CNN) profond

## Architecture:

-16 Couche dont 13 sont des couche de convolution et 3 des couches connectés.

## Visualisation en 2D TSNE:



## Préparation des images (prétraitements)

Redimensionnement(224\*224)  
Conversion en tableau numpy  
Normalisation par **preprocess\_input**  
Conversion en tenseurs(tableau multidimensionnels)

## Chargement du modèle VGG16

Utiliser les images préparer comme entré  
de modèle vgg16,

## Reduction de dimenssion PCA

Pca(n\_comppnents = 0,99)

4096 à 803

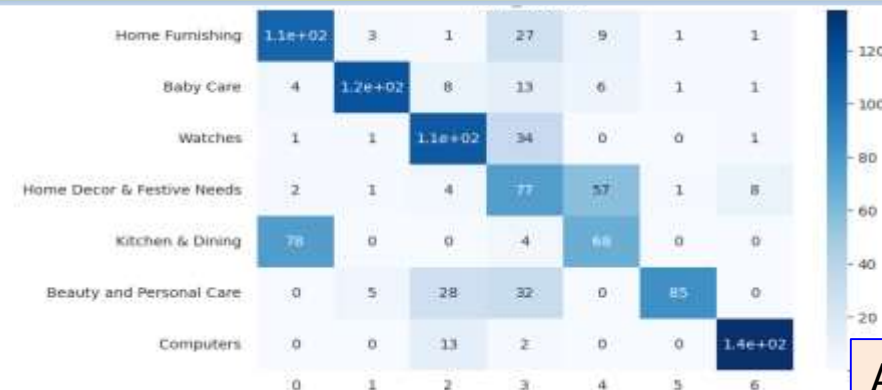
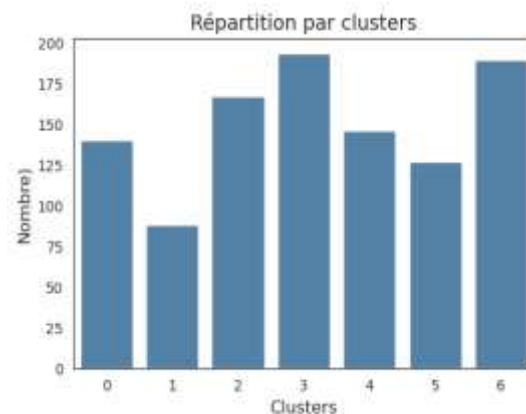
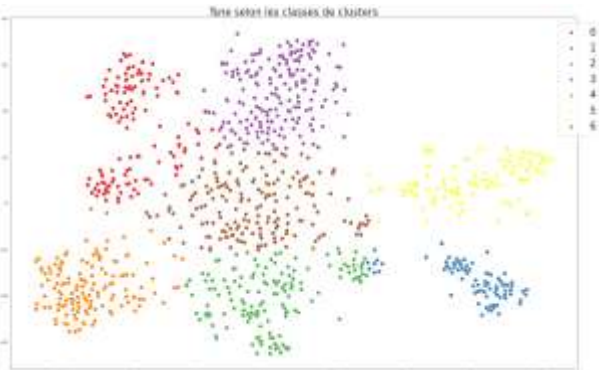




# Apprentissage non supervisée



## KMEANS+ VGG16:

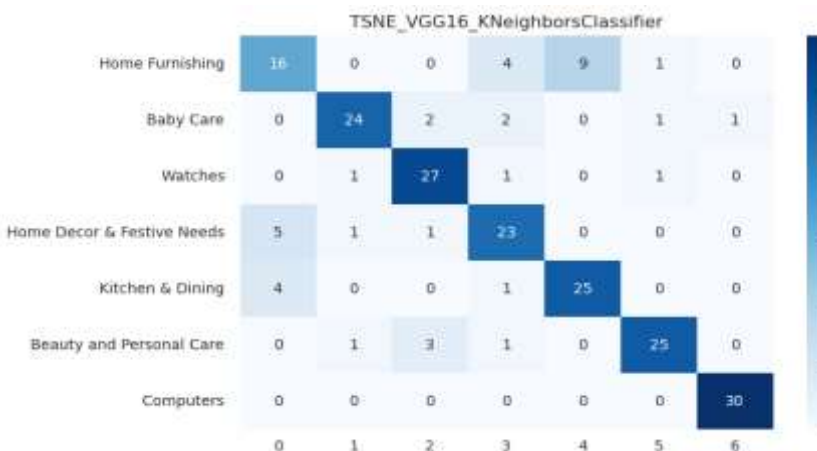


ARI :0,46

Accuracy : 67%

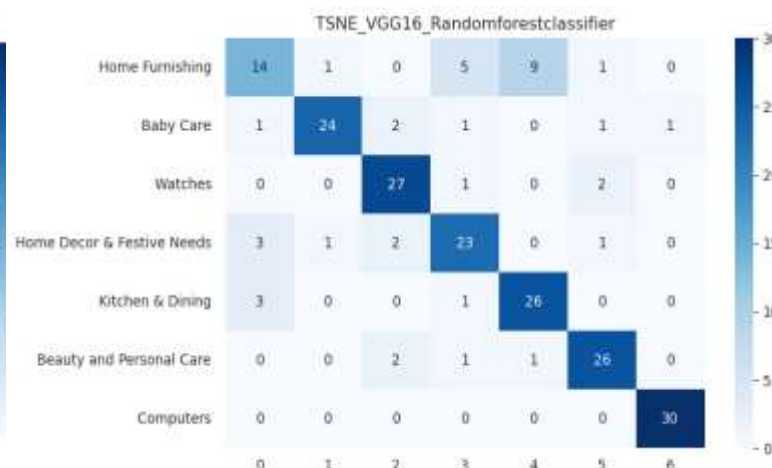
# Apprentissage supervisée méthodes classique

## TSNE\_VGG16\_KNeighborsClassifier



Accuracy =81%

## TSNE\_VGG16\_RandomForestClassifier



Accuracy =82%

## TSNE\_VGG16\_SVM



Accuracy =81%





# Réseau de neurones et classification supervisée

Entraînement avec le modèle VGG16:

Algorithme d'optimisation utilisé: Adam.

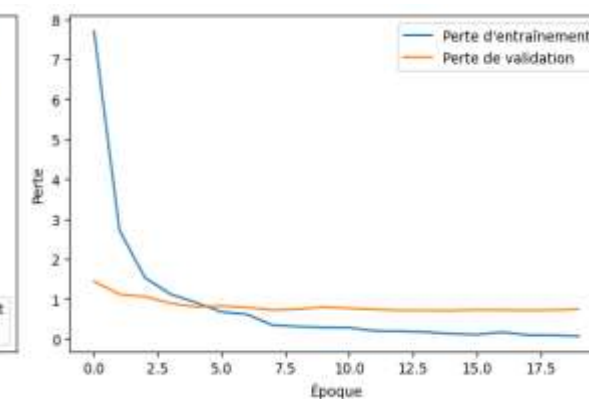
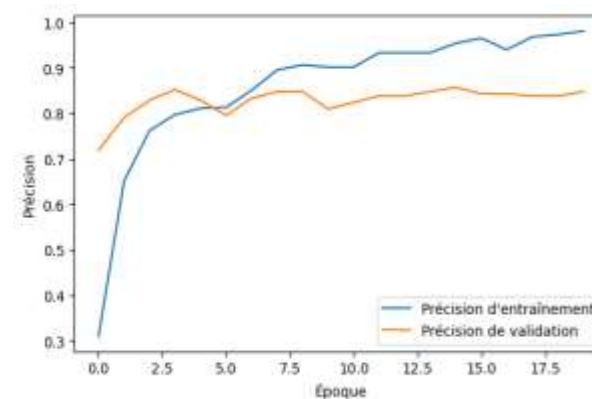
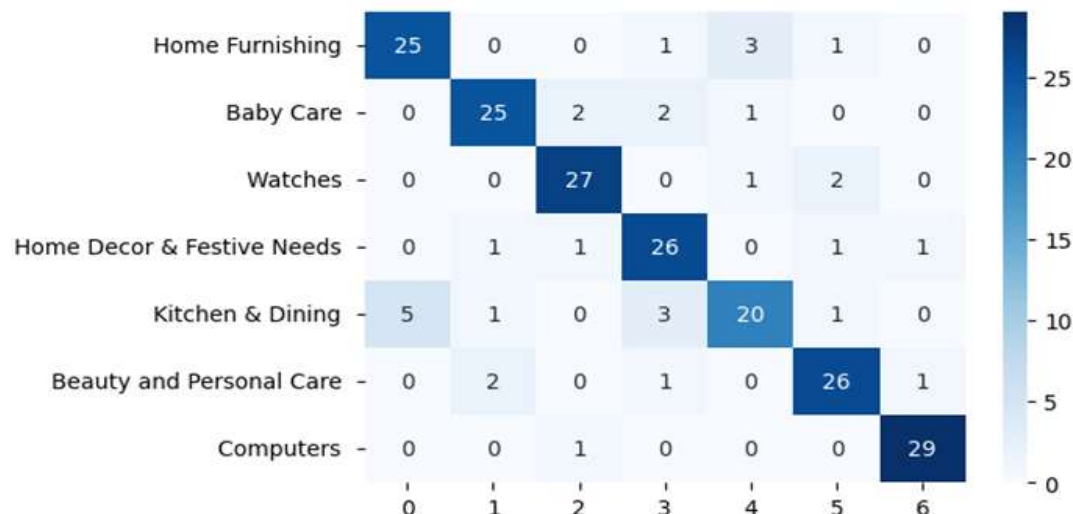
Fonction de perte: `categorical_crossentropy`.



Partitionnement de l'ensemble de données

Diviser l'ensemble de données en ensembles d'apprentissage, de validation et de test pour évaluer les performances du modèle

Matrice de confusion



Accuracy : 85%  
Train\_score:0,85  
Test\_score : 0,81



## Approche image data generator avec data augmentation

- Ajouter en amont une couche d'augmentation(rotation, orizontal)
- Remplacement de la derniere couche par:
  - Couche Flatten.
  - Couche Dense avec activation Relu
  - Couche décision Dense softmax.

Entrainement sur 75%des données initiales avec une validation de 25%.

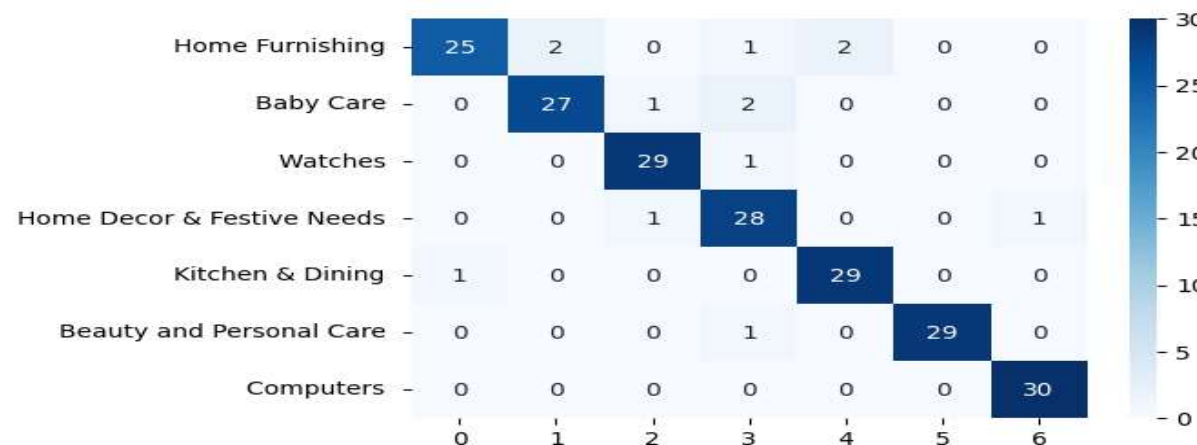
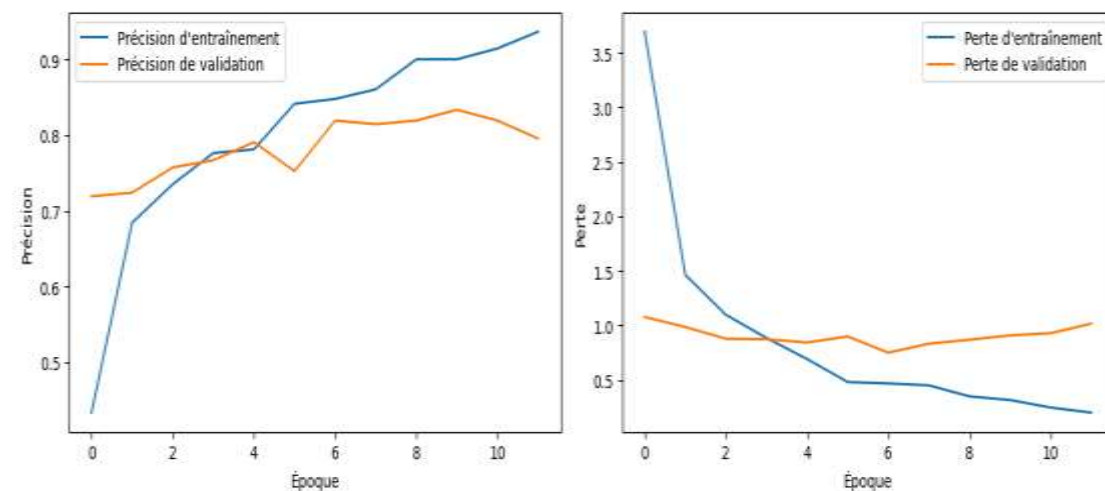
- Test sur 25%des données initiales (240)

Avec cette approche on trouve un **accuracy** très important est à **94%**

Validation\_score: 81%

Test\_score : 78%

## Graphique perte et précision





## l'API Edamam Food and Grocery Database:

Une interface de programmation qui donne accès à une base de données riche en informations sur les aliments et les recettes.

### Objectif :

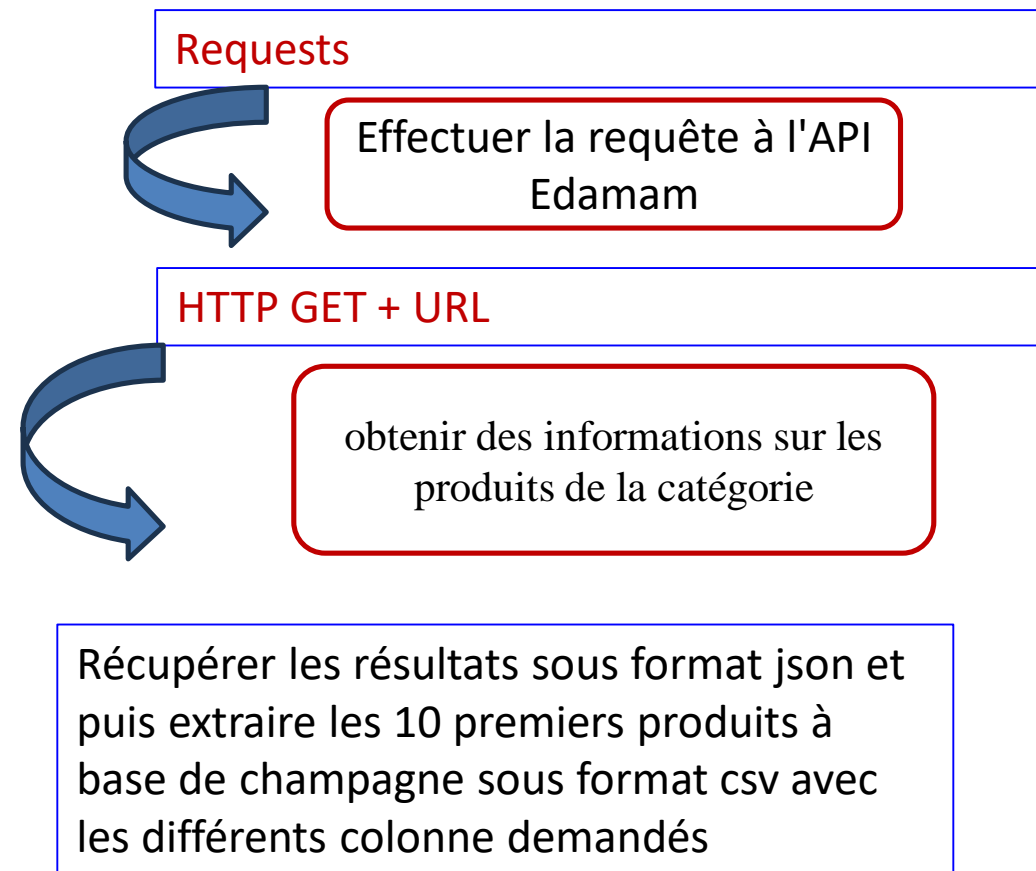
- Elargir la gamme de produits, en particulier dans l'épicerie fine.
- Tester la collecte des produits à base de champagne.
- Extraction des 10 premiers produits dans un fichier csv, contenant pour

chaque produits doit contenir les données suivantes:

- \* foodId.
- \* label.
- \* foodContentsLabel.
- \*category
- \* Image

### 5 principes de RGPD:

- Principe de sécurité et de confidentialité
- Droits des personnes
- Principe d'une durée de conservation limitée
- Principe de proportionnalité et de pertinence
- Principe de finalité





## Conclusion:

- Etude de faisabilité pour les données textuels et image est validé.
- Pour les images la faisabilités avec les méthodes récentes sont très efficaces,
- Les descripteurs sift ne sont pas adaptés :
  - Pas de clusters évidents
  - Pas de regroupement en fonction des catégories.
  - Des scores très faibles

