Réalisé par: Maidoumi Mohamed

04/02/2022

Projet 6: Classifiez automatiquement des biens de consommation.

Sommaire:


Missions:

Problèmatique:

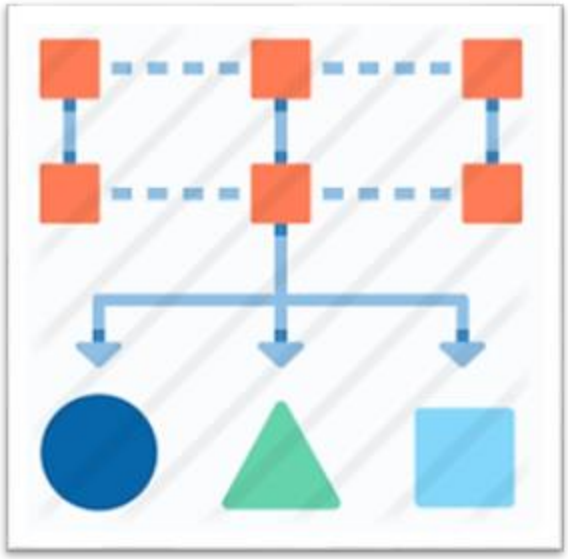
- Flipkart.com une **plateforme d'e-commerce** proposant des produits à la vente
- Les données des produits issus de la base FlipKart incluent des **descriptions textuelles** et des **images.**
- **L'attribution manuelle des catégories** : fastidieuse et peu fiable.
- Les catégories déjà renseignées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

⇨ Est-il faisable une classification selon les catégories à grand échelle ?

Mission du projet :
Etudier la faisabilité d'une automatisation de la classification des produits à partir de leur nom, description, et d'une photo.

Cahier des charges :

- Travail sur une **base de données limitée** de 1050 produits
- Obtenir une classification pertinente des produits de manière **non-supervisée**
- Niveau de **précision** suffisant et à quantifier
- Fournir une **représentation 2D** des données pour illustrer les résultats



Données:

Dimension Data
=
(1050,15)

Données visuelles

Image.jpg

Deux types de données :

- **Textuelles** : descriptions et noms des article, de longueurs variables, en anglais.
- **Visuelles** : une image par produit, isolé sur fond blanc, résolution variables.

Données textuelles

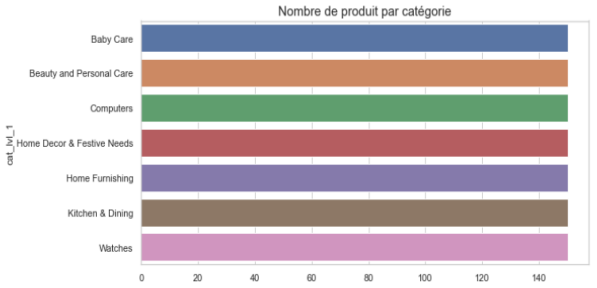
product_name
612 League Baby Boy's Checkered Casual Shirt

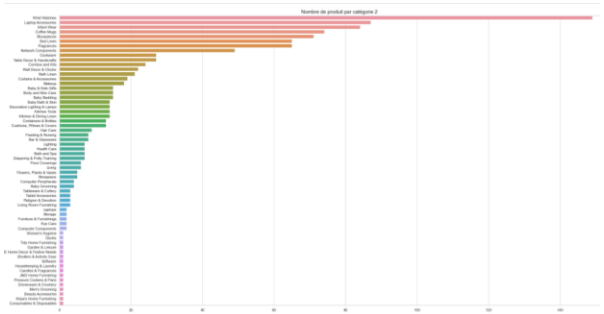
description
Specifications of 612 League Baby Boy's Checkered Casual Shirt General Details Pattern Checkered Occasion Casual Ideal For Baby Boy's Shirt Details Sleeve Half Sleeve Number of Contents in Sales Package Pack of 1 Brand Fit Regular Fabric 100% COTTON Fit Regular Additional Details Style Code BLS00S380001B Fabric Care ENZYME WASH

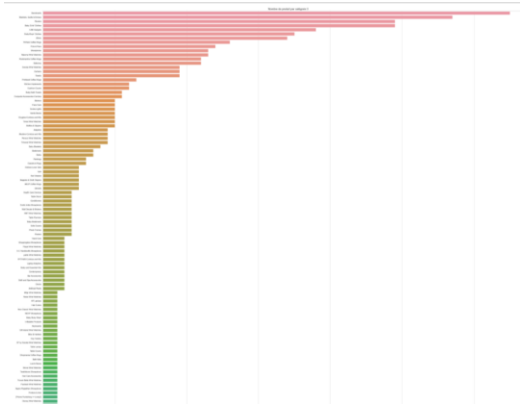
Choix de catégorie cible:

* Choix du premier niveau, présentant des effectifs homogènes (150 articles par catégories).

* 7 catégories :Furnishing, Baby, Watches, Decor, Kitchen, Beauty, Computers.







Modélisation NLP:

Occurrence Globale des mots

0	product	863
1	free	618
2	buy	583
3	delivery	567
4	genuine	564
5	cash	564
6	price	561
7	replacement	559
8	day	553
9	guarantee	473

Analyse textuelle:

Démarche global:

Preprocessing NLP

- Pré-traitement des données textes (tokenisation, lemmatisation, supprimer la ponctuation, les stops words…).

Feature engenering NLP

- Extraction des features Bag of word (vectorisation des mots par une méthode tfidf)
- Exploration non supervisé des topics latent (LDA, NMF).

Modélisation NLP:

- Réduction de dimention (2D par la méthode T-SNE)
- Partitionnement non supervisé
- Evaluation de la correspondance des clusters avec les catégories « vraies »

Avant pré-traitement

description

Specifications of 612 League Baby Boy's Checkered Casual Shirt General Details Pattern Checkered Occasion Casual Ideal For Baby Boy's Shirt Details Sleeve Half Sleeve Number of Contents in Sales Package Pack of 1 Brand Fit Regular Fabric 100% COTTON Fit Regular Additional Details Style Code BLS00S380001B Fabric Care ENZYME WASH

Après pré-traitement

description

league baby casual shirt general occasion casual ideal baby boy shirt half sleeve number pack brand fit regular fabric cotton fit regular additional style code bls fabric care wash

Préparation des données données textuelles

In [249]:
```
# tokenization, lemmatizing and removing stop_word using spacy
import spacy
import nltk
# Stopwords and single letters
# english_sw = nltk.corpus.stopwords.words('english')
# single_let_sw = list(string.ascii_lowercase)
#sw = list( set(english_sw) )
nlp = spacy.load( "en_core_web_sm", disable=['parser', 'ner'] )
nlp.Defaults.stop_words.add("cm")
nlp.Defaults.stop_words.add("r")
```

In [250]:
```
def lemmatize(text):
    doc = nlp(text) # chargement ??
    tokens=[ token.lemma_.strip() for token in doc if not (token.is_stop or token.is_punct or token.like_num) ]
    return ' '.join(tokens).lower()
data['description'] = data['description'].apply(lambda x: lemmatize(x).replace("\r\n",""))
```

- Les opérations de traitement de text :
  - Elimination des caractères non-alphabétiques
  - Mise en minuscule
  - Elimination des stopwords et des lettres isolées
  - Lemmatisation
  - Extraction des seuls noms et adjectifs

Croisement topic de LDA et vraie catégorie:

topic_lda	cat_lvl_1
0	Watches
	Kitchen & Dining
	Home Furnishing
	Home Decor & Festive Needs
	Computers
	Beauty and Personal Care
	Baby Care
1	Kitchen & Dining
	Home Decor & Festive Needs
	Computers
	Beauty and Personal Care
	Baby Care
2	Watches
	Kitchen & Dining
	Home Furnishing
	Home Decor & Festive Needs
	Beauty and Personal Care

topic_nmf	cat_lvl_1
0	Kitchen & Dining
	Home Furnishing
	Computers
	Beauty and Personal Care
	Baby Care
1	Watches
	Beauty and Personal Care
2	Kitchen & Dining
	Home Furnishing
	Home Decor & Festive Needs
	Baby Care
3	Home Furnishing
	Beauty and Personal Care
	Baby Care
4	Kitchen & Dining
	Home Decor & Festive Needs
	Computers
	Beauty and Personal Care
	Baby Care
5	Home Furnishing
	Baby Care

Résultats NMF relaviment plus homogène par rapport à LDA.

Exploration non supervisé des topics detectés selon 7 catégories (NMF et LDA).

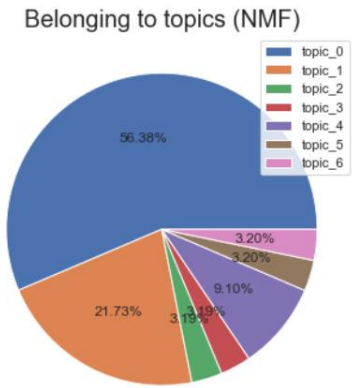
LDA

Topic 0:
baby cotton pack details girl fabric number general box color
Topic 1:
usb light power warranty adapter glass inch denver rice lights
Topic 2:
showpiece good rs price online guarantee day replacement products genuine
Topic 3:
watch analog men guarantee india replacement rs online day women
Topic 4:
com flipkart genuine cash shipping delivery products free buy guarantee
Topic 5:
mug hair bring ml bottle wall design coffee nutcase gift
Topic 6:
battery rockmantra mug cell ceramic product year craft safe design

NMF

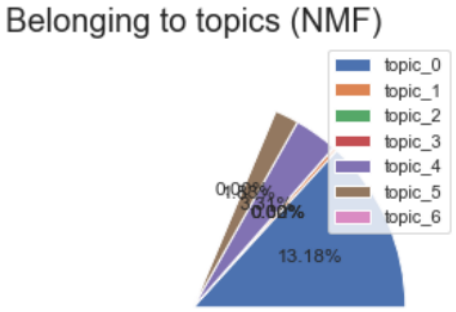
Topic #0: com flipkart combo set guarantee replacement genuine cash shipping delivery
Topic #1: watch analog men india great women discounts rs sonata online
Topic #2: mug ceramic rockmantra coffee perfect gift love safe creation microwave
Topic #3: baby details girl fabric boy cotton neck sleeve dress shirt
Topic #4: showpiece good rs online guarantee replacement day price genuine shipping
Topic #5: abstract single blanket com flipkart quilts comforters genuine shipping cash
Topic #6: warranty laptop inch skin color pack box model features type

Elegance Polyester Multicolor Abstract Eyelet Door Curtain



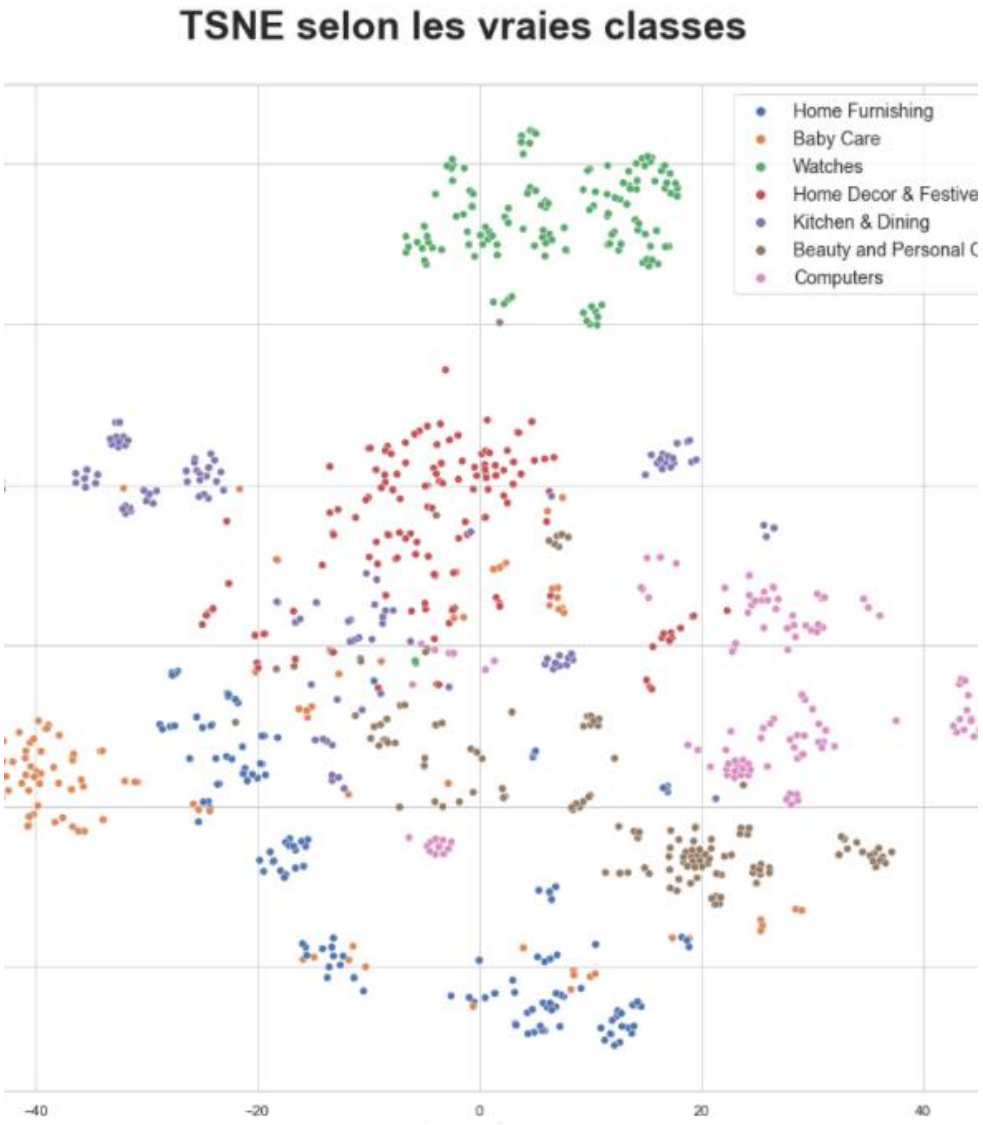
Sathiyas Cotton Bath Towel

```
C:\Users\maido\AppData\Local\Temp/ipykernel_1
lize if the sum is less than 1 but this behav
period the default value will be normalize=Tr
  plt.pie(frequencies,
```

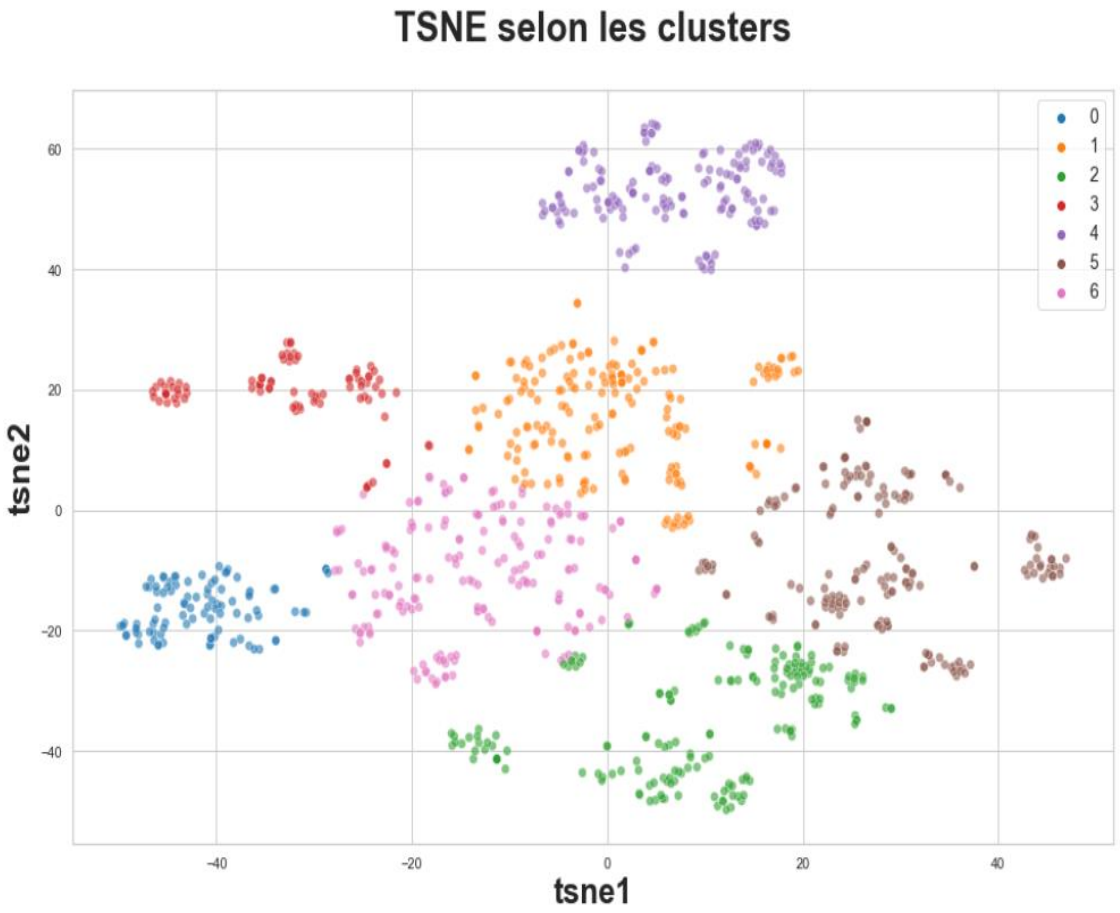


Visualisation 2D des vraies catégorie après pca (90% expliquée) et TSNE:

Résultat peu homogéne en terme de séparabilité(watche , baby care, Beaty , personal care relaviement mieux séparé



Classification non supervisé sel on 7 catégories des produits:



ARI= 43%


Modélisation
par la
méthode Sift:

Image par catégories:

Création d'un vecteur de descripteurs Sift par image (array of array) et un vecteur de tout les descripteurs de la base de donnée (array)

Création de clusters de descripteurs (nb de clusters racine carrée du nombre des descripteurs).

Création des features (histogramme normalisé des clusters par image)

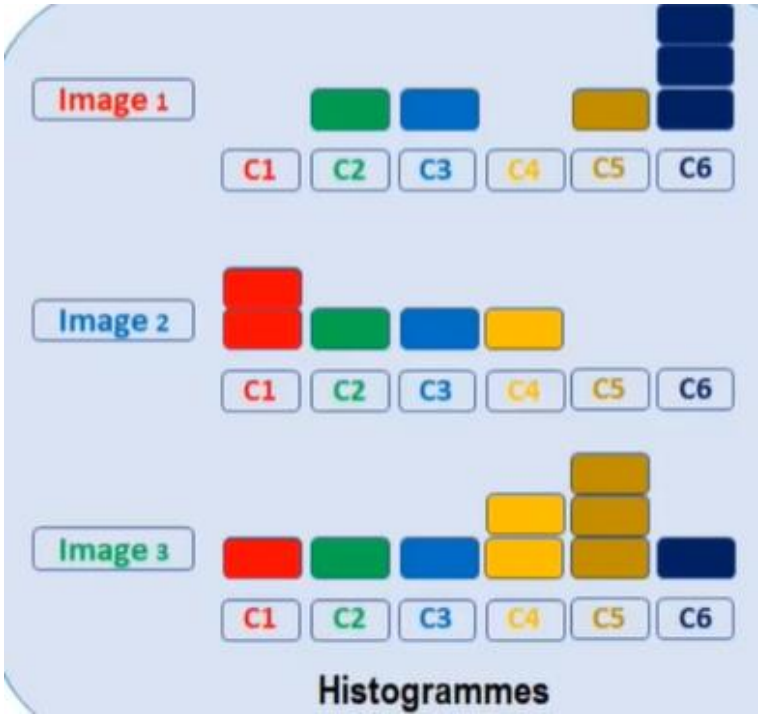
Réduction de dimension PCA/TSNE.

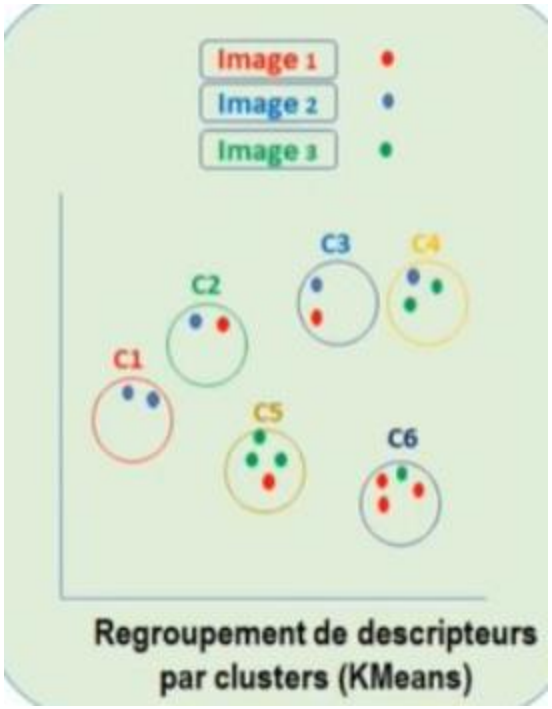
Visualisation 2D TSNE selon les vraies classes.

Clustering à partir des deux composente principales.

Analyse de similarité ARI

Démarche Générale:

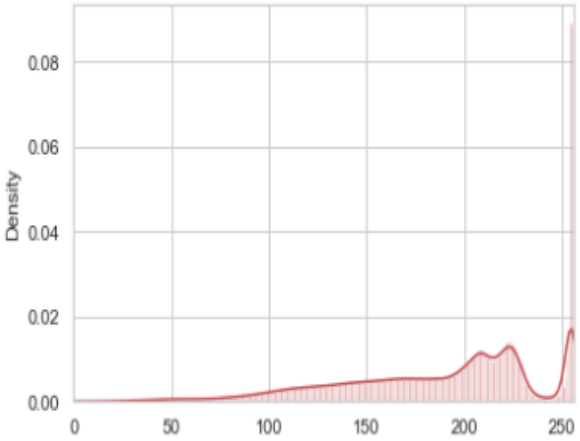




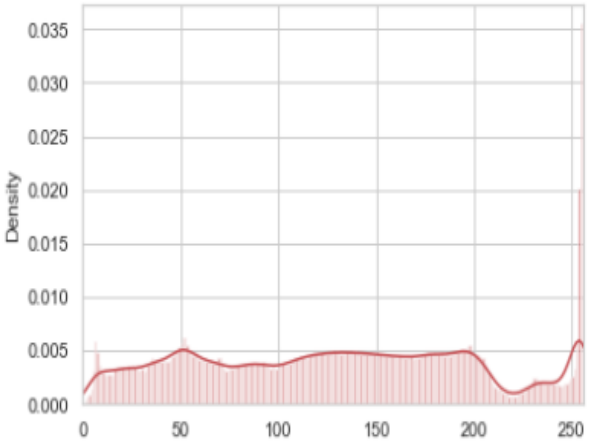
Bag of visual word et extraction de features:

Filtrage et égalisation des images avant modélisation:

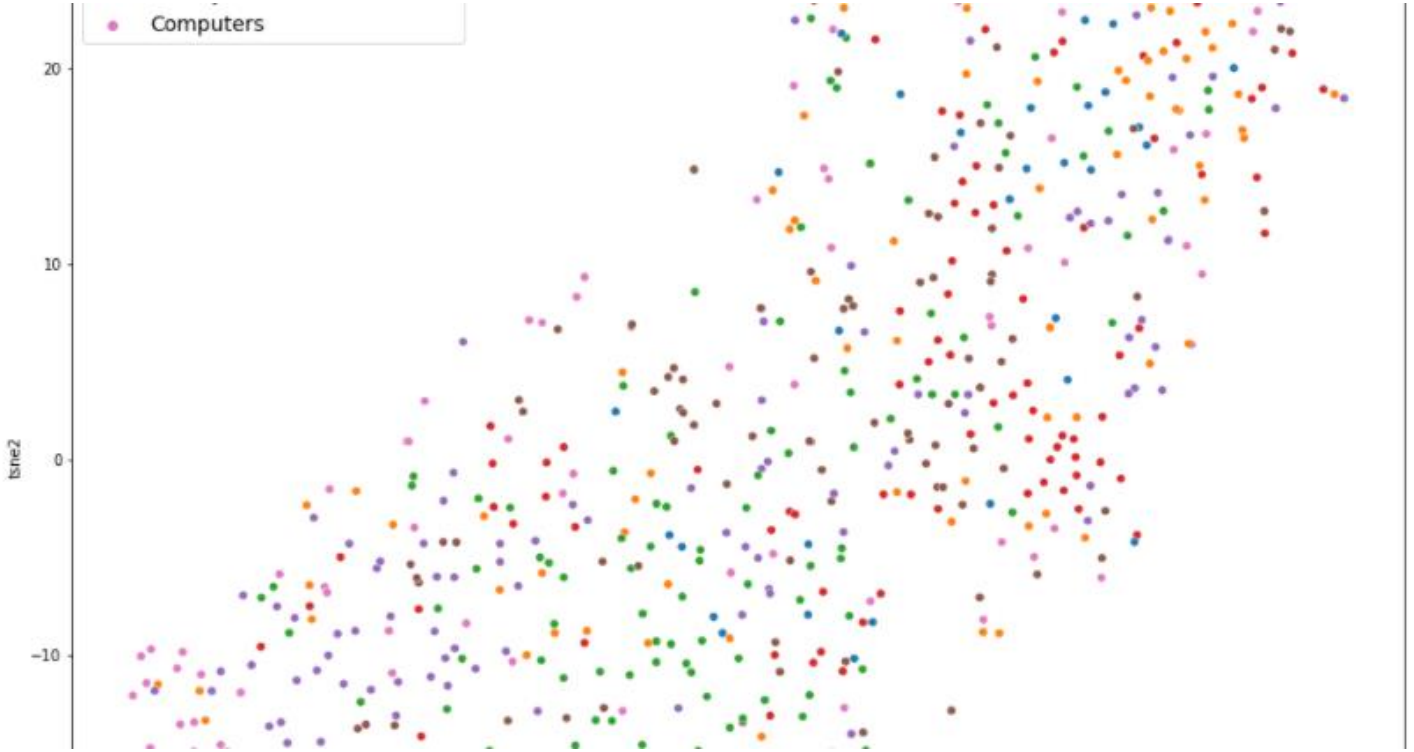
Histograme avant et après traitement:



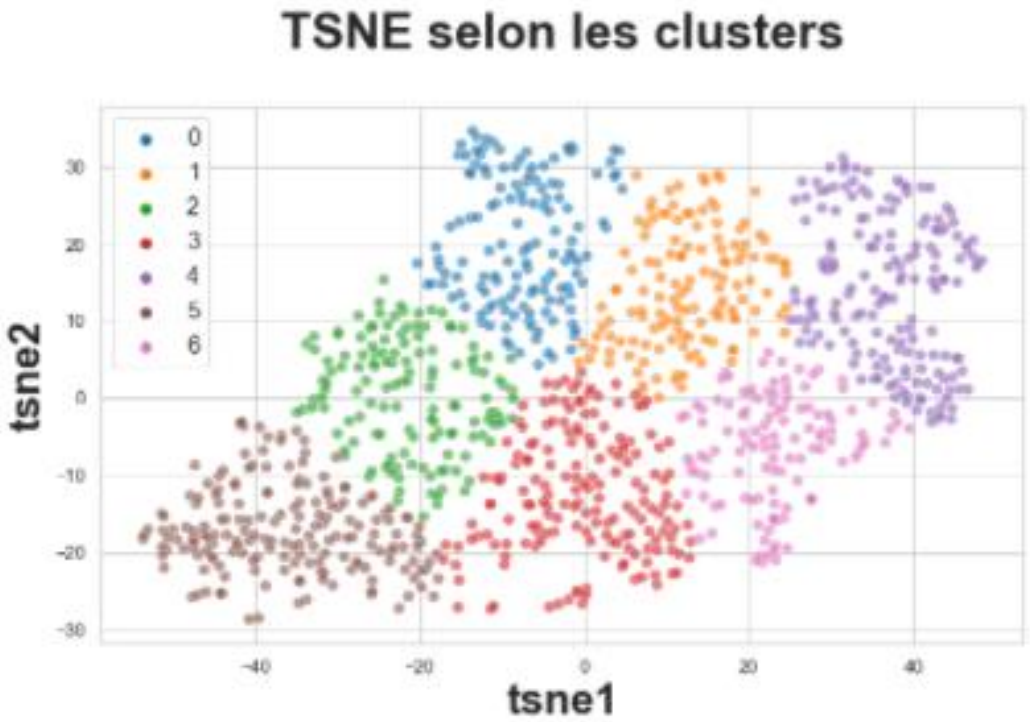

```
C:\Users\maido\anaconda3\lib\site-packages\seaborn\distri
d will be removed in a future version. Please adapt your
xibility) or `histplot` (an axes-level function for histo
  warnings.warn(msg, FutureWarning)
```



- ❑ Taille de l'image (112, 224, 448, 896)
- ❑ Egalisation des histogrammes (opencv)
- ❑ Application de filtre Gaussien.
- ❑ Image en gris

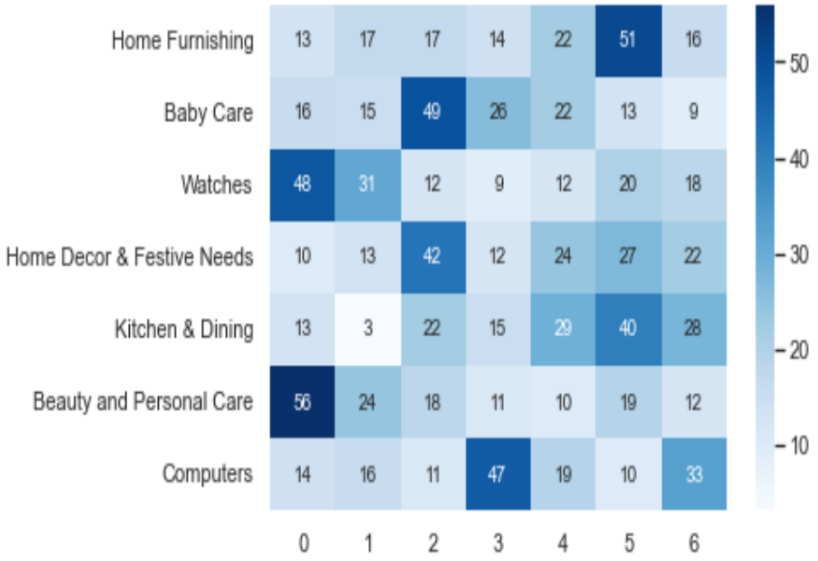


Représentation 2D des descripteurs sift selon les vrais catégorie:



Representation 2D des déscripteurs Sift selon le clustering Kmeans:

Matrice de confusion et résultat du clustering:




```
Correspondance des clusters:  [6, 3, 4, 2, 0, 5, 1]
[[13 17 17 14 22 51 16]
 [16 15 49 26 22 13  9]
 [48 31 12  9 12 20 18]
 [10 13 42 12 24 27 22]
 [13  3 22 15 29 40 28]
 [56 24 18 11 10 19 12]
 [14 16 11 47 19 10 33]]
```

	precision	recall	f1-score	support
0	0.08	0.09	0.08	150
1	0.13	0.10	0.11	150
2	0.07	0.08	0.07	150
3	0.09	0.08	0.08	150
4	0.21	0.19	0.20	150
5	0.11	0.13	0.12	150
6	0.24	0.22	0.23	150
accuracy			0.13	1050
macro avg	0.13	0.13	0.13	1050
weighted avg	0.13	0.13	0.13	1050

Modélisation par Transfer Learning CNN

Démarche Globale:

Séparation les images en training set et test set.

Prétraitement des images pour une compatibilité avec l'input VGG16 (process_input).

One hote Encoding de la variable catégorie pour l'implémentation.

Implémentation des couches du réseau de noronne ou importation du réseau VGG16 en supprimant la dernière couche fully connected et insertion de la nouvelle couche en adéquation avec le problème.

Compiler le réseau

Fiter le réseau sur les données training.

Predir les données test et calculer l'accuracy avec les vraies catégories.

Classification selon un réseau:

```
[72]:
model = Sequential()

model.add(Conv2D(32, kernel_size=(3,3), padding='same', activation='relu', input_shape=(224,224,3,)))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Conv2D(32, kernel_size=(3,3), padding='same', activation='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Flatten())
model.add(Dense(ohe.categories_[0].shape[0], activation='softmax'))
model.compile(loss='mean_squared_error', optimizer='sgd')
```

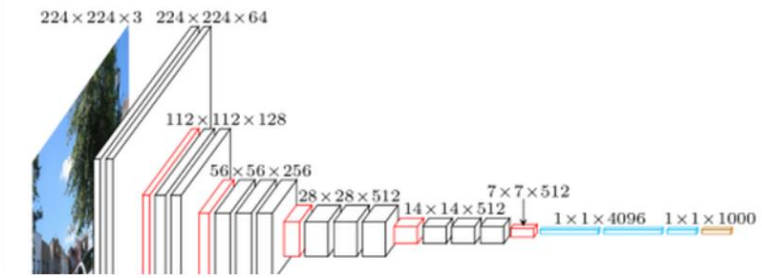
```
[73]:
model.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_2 (Conv2D)            (None, 224, 224, 32)      896

max_pooling2d_2 (MaxPooling  (None, 112, 112, 32)      0
2D)

conv2d_3 (Conv2D)            (None, 112, 112, 32)      9248

max_pooling2d_3 (MaxPooling  (None, 56, 56, 32)        0
2D)

flatten_1 (Flatten)          (None, 100352)            0

dense_7 (Dense)              (None, 7)                 702471

=================================================================
Total params: 712,615
Trainable params: 712,615
Non-trainable params: 0
_________________________________________________________________
```

Accuracy prédiction = 14%

Classification avec un réseau VGG16 préntrainer ImageNet:

Architecture de VGG-16




```
: 1 vgg_transfer.summary()
```

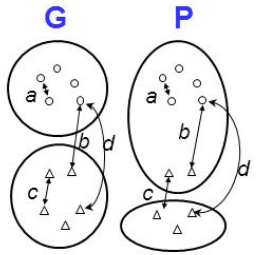
```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vgg16 (Functional)          (None, 7, 7, 512)         14714688

 flatten_2 (Flatten)         (None, 25088)             0

 dense_10 (Dense)            (None, 7)                 175623

=================================================================
Total params: 14,890,311
Trainable params: 175,623
Non-trainable params: 14,714,688
```

Accuracy prédiction = 83%

Conclusion:

Questions

ARI score (à maximiser)



Agreement: a, d
Disagreement: b, c

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

L'*Adjusted Rand Index* (ARI) est la normalisation de l'indice de Rand (RI) qui permet de comparer deux partitions de nombres de classes différentes.

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

- RI : indice de Rand, proportion de paires de points qui sont groupés de la même façon dans les deux partitions.
- E(RI) : espérance de l'indice de Rand (pour une partition aléatoire)
- max(RI) : indice de Rand maximal qui pourrait être obtenu étant donné le nombre de classes distincts