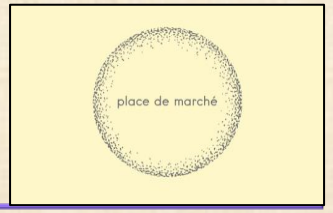
PROJET N°6 :

CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

Soutenance du P6: le 12/08/2022

Version notebook : **6.3.0**
Version Python : **3.8.8**
Version Pandas : **1.2.4**
Version Seaborn : **0.11.1**
Version Matplotlib: **3.3.4**

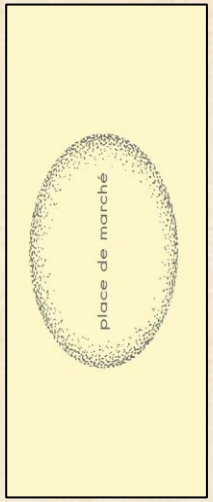
Plan



- Problématique & Présentation du jeu de données
- Prétraitement et modélisation partie NLP
- Prétraitement et modélisation partie Image
- Approche Combinée
- Conclusion: faisabilité et recommandations

Problématique & Présentation du jeu de données

Contexte:



- L’entreprise "Place de marché” souhaite lancer une marketplace e-commerce. Sur la "Place de marché", des vendeurs proposent des articles à des acheteurs en postant une photo et une description, sachant que tâche se faite manuellement par les vendeurs,
- Pour faciliter la mise en ligne de nouveaux articles nécessaire d'automatiser la tâche.

Mission:

- Etude 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.

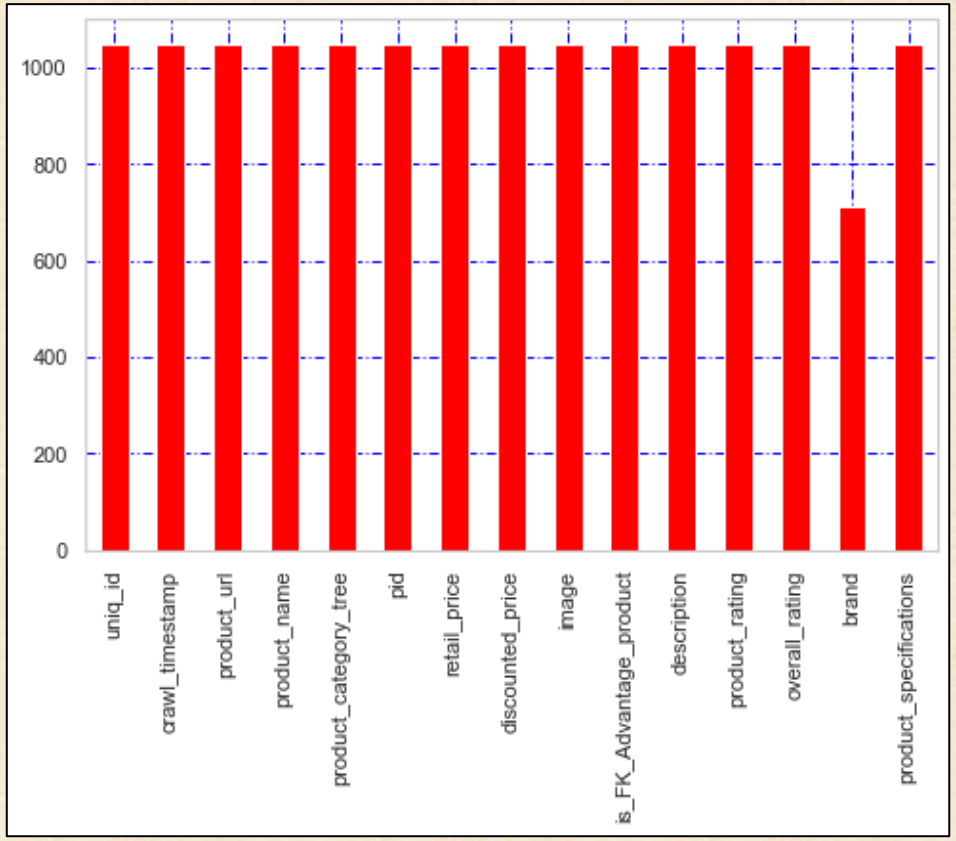
Problématique & Présentation du jeu de données

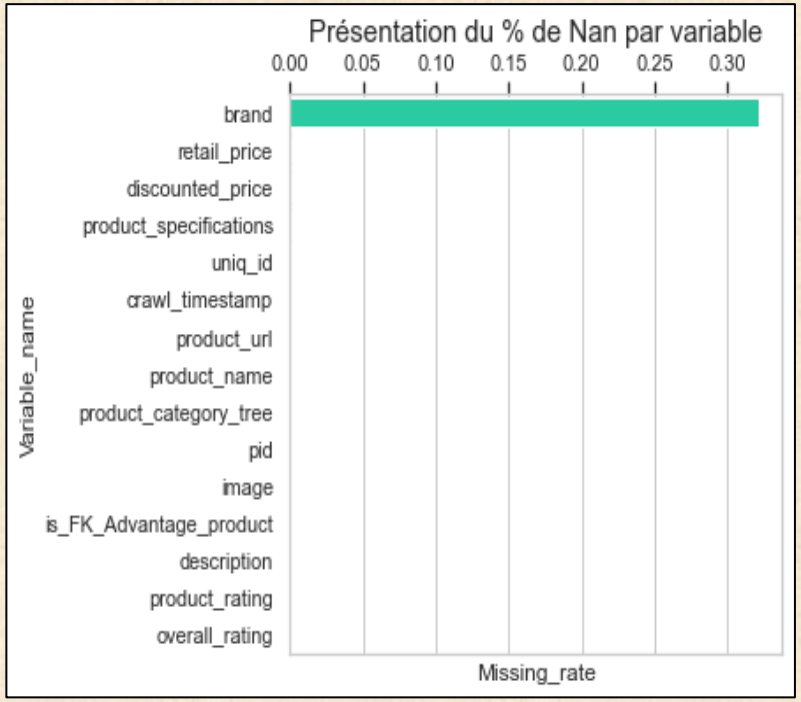
Présentation de données:

data

```
calc_inf(data, True)
```

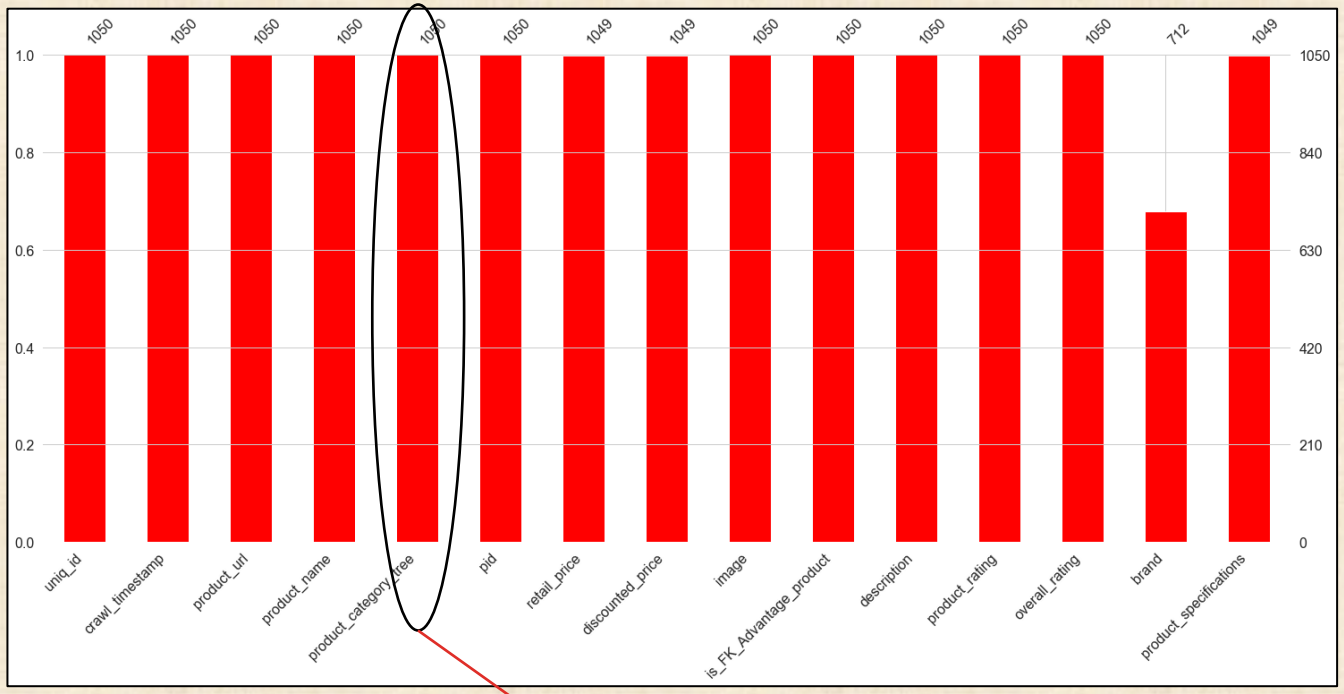
```
* Nombre de colonnes sans NaN -------: 11
* Nombre de colonnes NaN ------------: 0
* Nombre de colonnes mixtes----------: 4
* Nombre de ligne entièrement nulles : 0
* Nombre de ligne mixtes ----------- : 340
* Nombre de ligne sans NaN ----------: 710
* Nombre de lignes ----------------------: 1050
* Nombre de colonnes ---------------------: 15
* Nombre de cases ------------------------: 15750
* Nombre de valeurs nulles ---------------: 341
* Nombre de valeurs non nulles -----------: 15409
* le pourcentage des valeurs nulles ------: 2.2 %
* le pourcentage des valeurs non nulles --: 97.8 %
```





Problématique & Présentation du jeu de données

Présentation de Data-Set:



La variable ‘product_category_tree’ comporte plusieurs sous catégories de produits. Cette variable va nous aider à créer plusieurs sous catég qui nous aidera à tester la faisabilité d'un moteur de classification d'articles

```
0       ["Home Furnishing >> Curtains & Accessories >>...
1       ["Baby Care >> Baby Bath & Skin >> Baby Bath T...
2       ["Baby Care >> Baby Bath & Skin >> Baby Bath T...
3       ["Home Furnishing >> Bed Linen >> Bedsheets >>...
4       ["Home Furnishing >> Bed Linen >> Bedsheets >>...
```

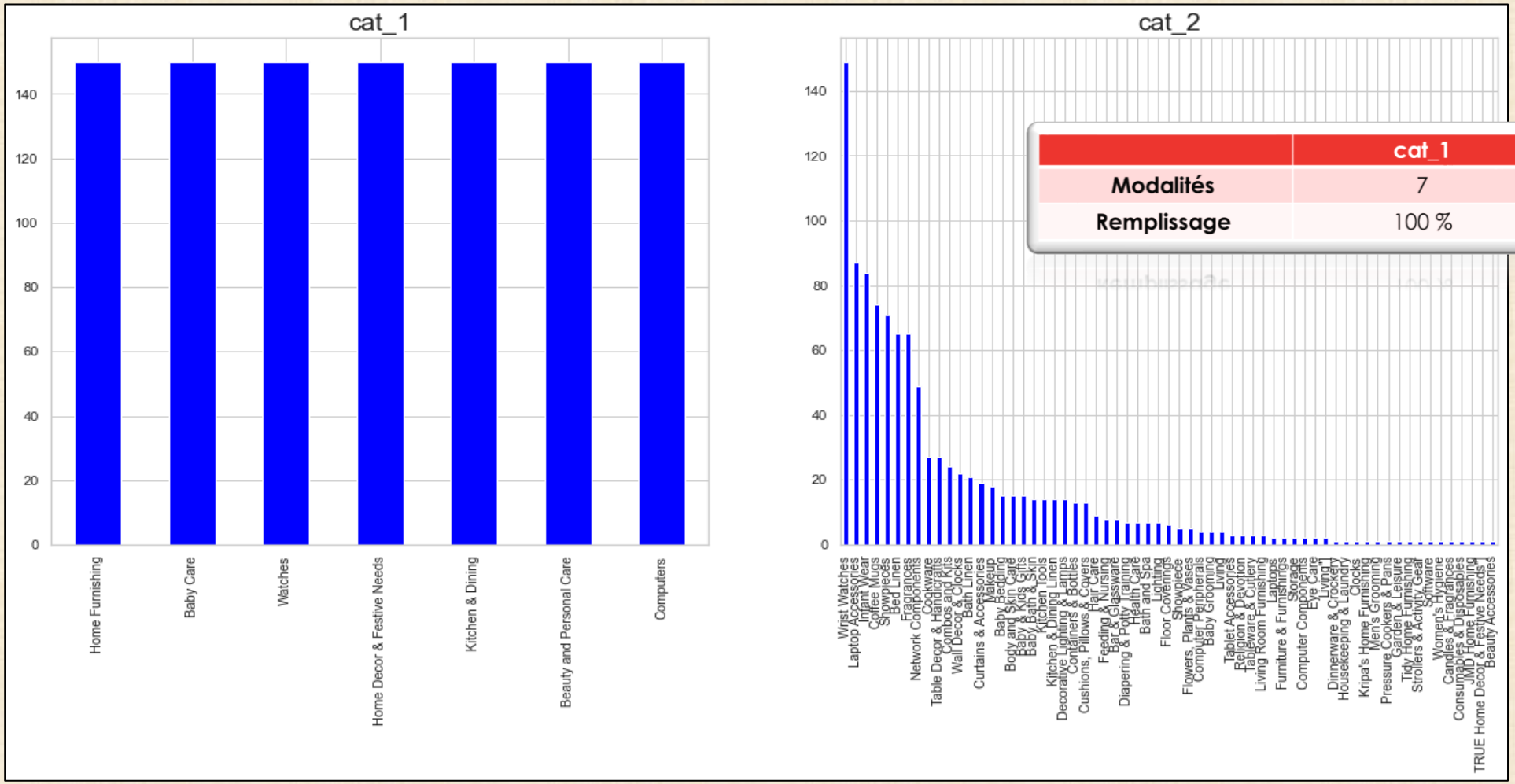
```
def clean(item):
    a  = item.split('>>')[0].split('"')[1].strip()
    return a
def clean_2(item):
    b  = item.split('>>')[1].strip()
    return b
```

```
1  data['cat_1'].value_counts(), data['cat_2'].value_counts()
```

```
(Home Furnishing                150
 Baby Care                      150
 Watches                        150
 Home Decor & Festive Needs     150
 Kitchen & Dining               150
 Beauty and Personal Care       150
 Computers                      150
 Name: cat_1, dtype: int64,
 Wrist Watches                          149
 Laptop Accessories                      87
 Infant Wear                             84
 Coffee Mugs                             74
 Showpieces                              71
                                       ...
 Candles & Fragrances                     1
 Consumables & Disposables                1
 JMD Home Furnishing                      1
 TRUE Home Decor & Festive Needs"]        1
 Beauty Accessories                       1
 Name: cat_2, Length: 63, dtype: int64)
```

Problématique & Présentation du jeu de données

Présentation de sous catég:



	cat_1	cat_2
Modalités	7	62
Remplissage	100 %	100 %

Prétraitement et modélisation partie NLP

Les étapes de sélection et de prétraitement de texte:

Prétraitement + Modélisation

Sélection Features

uniq_id
product_name
discounted_price
image
description
cat_1
cat_2

Fonctions secondaires

* tokenizer_fct
* clean_text
* stop_word_filter_fct
* lower_start_fct
* lemma_fct

Fonctions principales

1. Prétraitement de texte pour le BOW (CountVectorizer, TF-IDF)
……..
2. Prétraitement de texte pour le deep learning model

1. Encodage CVT, TF-IDF, par RN BERT, USE…..
2. Réduction de…….. Dimension .TSNE….
3. Clustering avec…….. l'algo K-Means
4. Visualisation……..
5. Calcul de score……
ARI

TEXTE

Analyse de faisabilité

Prétraitement et modélisation partie NLP

Exemple de prétraitement de texte

Exemple d'un texte pour le bag of words avec lemmatisation:

- word tokens
- stop word
- lower case
- lemmatisation

```
1 data_mdl['sentence_bow_lem'].iloc[0]
```

```
'key feature elegance polyester multicolor abstract eyelet door curtain flo
ral curtain elegance polyester multicolor abstract eyelet door curtain heig
ht pack price this curtain enhances look interior this curtain made high qu
ality polyester fabric feature eyelet style stitch metal ring make room env
ironment romantic loving this curtain ant wrinkle anti shrinkage elegant ap
parance give home bright modernistic appeal design the surreal attention su
re steal heart these contemporary eyelet valance curtain slide smoothly dra
w apart first thing morning welcome bright sun ray want wish good morning w
hole world draw close evening create special moment joyous beauty given soo
thing print bring home elegant curtain softly filter light room get right a
mount sunlight specification elegance polyester multicolor abstract eyelet
door curtain height pack general brand elegance designed for door type eyel
et model name abstract polyester door curtain set model color multicolor di
mension length box number content sale package pack sale package curtain bo
dy design material polyester'
```

Exemple d'un texte pour le Deep learning (USE et BERT):

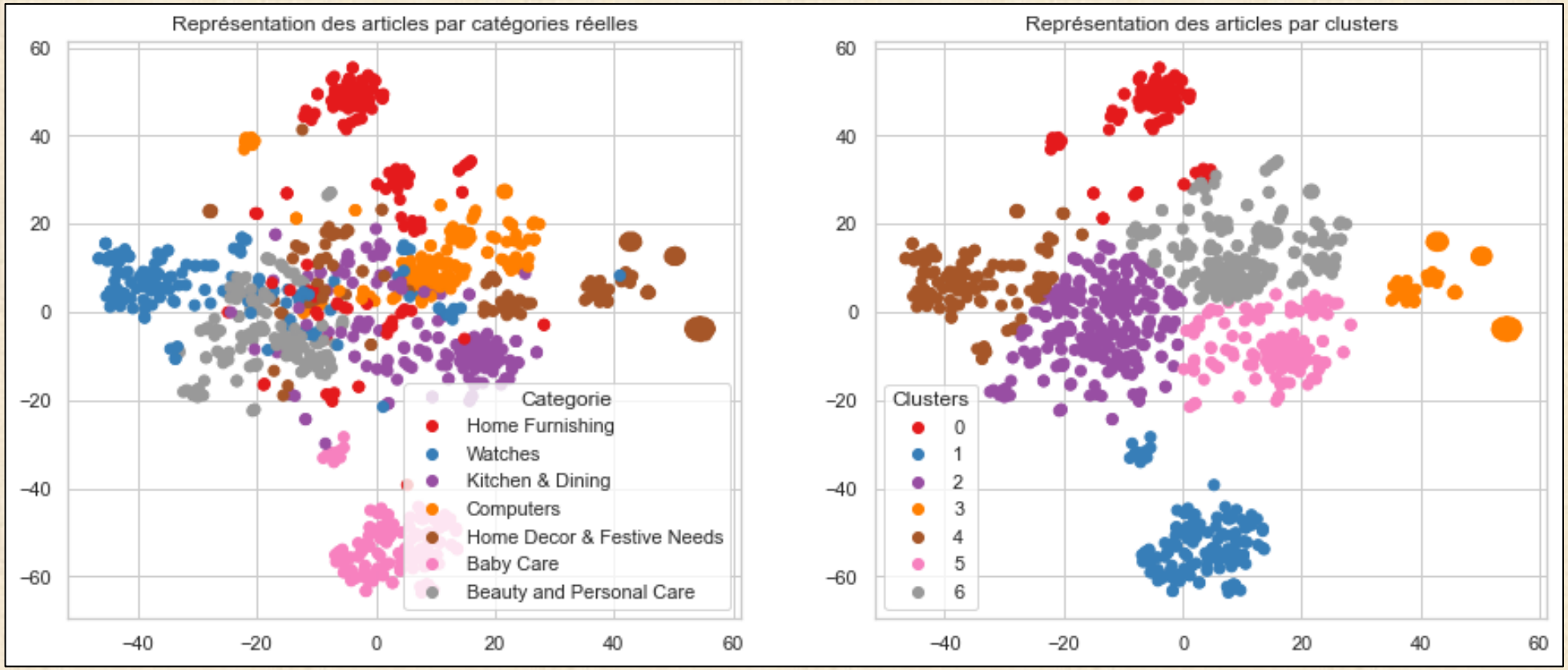
- Word tokens + lower case
- sans lemmatisation, sans enlever les stop word

```
2 data_mdl['sentence_dl'].iloc[0]
```

```
'key features of elegance polyester multicolor abstract eyelet door curtain
floral curtain , elegance polyester multicolor abstract eyelet door curtain
( 213 cm in height , pack of 2 ) price : rs . 899 this curtain enhances the
look of the interiors.this curtain is made from 100 % high quality polyeste
r fabric.it features an eyelet style stitch with metal ring.it makes the ro
om environment romantic and loving.this curtain is ant- wrinkle and anti sh
rinkage and have elegant apparance.give your home a bright and modernistic
appeal with these designs . the surreal attention is sure to steal hearts .
these contemporary eyelet and valance curtains slide smoothly so when you d
raw them apart first thing in the morning to welcome the bright sun rays yo
u want to wish good morning to the whole world and when you draw them close
in the evening , you create the most special moments of joyous beauty given
by the soothing prints . bring home the elegant curtain that softly filters
light in your room so that you get the right amount of sunlight. , specific
ations of elegance polyester multicolor abstract eyelet door curtain ( 213
cm in height , pack of 2 ) general brand elegance designed for door type ey
elet model name abstract polyester door curtain set of 2 model id duster25
color multicolor dimensions length 213 cm in the box number of co
sales package pack of 2 sales package 2 curtains body & design ma
yester'
```

Prétraitement et modélisation partie NLP

Création du Bag Of Words (CountVectorizer): T-SNE + K-Means + Score ARI




```
CountVectorizer :
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
ARI :  0.4662 time :  17.0
--------------------------------------------------------------------------------
```

Prétraitement et modélisation partie NLP

Création du Bag Of Words (CountVectorizer + LDA):

Test de LDA sur le BOW réalisé par CountVectorizer avec un nombre de topics = 7 en fonction de cat_1, ainsi 10 mots par topic

```
Topic 0:
laptop skin baby warranty specification type set print general shape
Topic 1:
design pack inch feature wall color sticker material home quality
Topic 2:
product free delivery buy shipping cash genuine day replacement guarantee
Topic 3:
usb battery bowl box singing crystal quality product fan jewellery
Topic 4:
polyester eyelet curtain aroma door window sstudio comfort pant gown
Topic 5:
bowl brow mask terracotta frame handmade bengal ornamental village diviniti
Topic 6:
mug ceramic coffee perfect gift one design material tea loved
```

Classifier SVC sur jeu de données LDA

```
Topic 0:
hair conditioner tigi trait type infused goddess head colour oil
Topic 1:
cotton cotonex pack design yellow set green package sale glove
Topic 2:
mug coffee tea perfect one ceramic printland made fantastic enjoy
Topic 3:
free product delivery genuine shipping cash buy day replacement guarantee
Topic 4:
intex inflatable air chair dress kid gathered detail baby cotton
Topic 5:
dhol admiration beyoutiful repouss shop copper drawing surity exclusive man
Topic 6:
baby cotton detail fabric color specification girl general boy type
Topic 7:
com flipkart day shipping buy cash delivery genuine product guarantee
Topic 10:
```

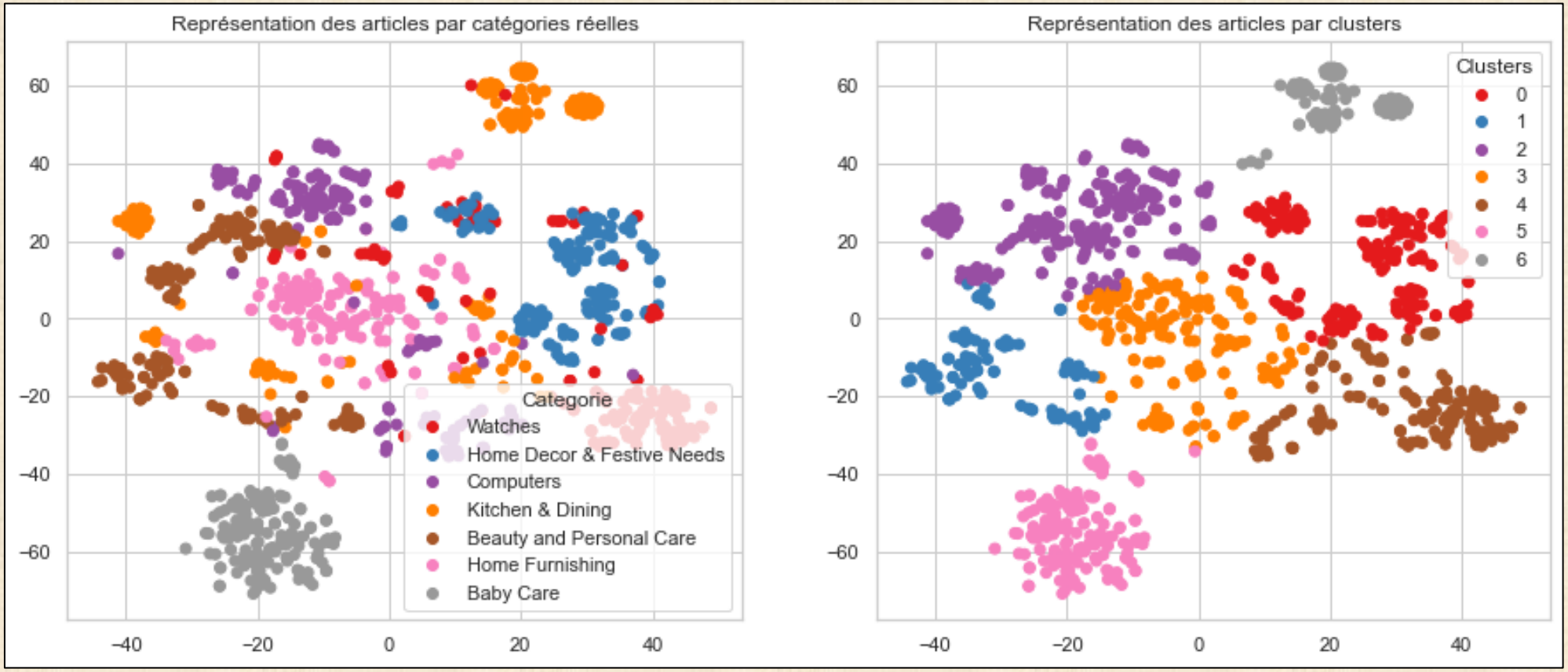
Test du classifieur SVC classification supervisée + un BOW réalisé par CountVectorizer avec un nombre de topics = 62 en fonction de sous cat_2 et 10 mots par topic, en dessous le score d'accuracy sur le jeu de donnée d'entrainement de test

```
1 train_txt.shape, test_txt.shape
((787, 10), (263, 10))
```

```
accuracy sur jeu train :  0.4269377382465057
accuracy sur jeu test :  0.376425855513308
```

Prétraitement et modélisation partie NLP

Création du bag of words (Tf-IDF): T-SNE + K-Means + Score ARI




```
Tf-idf :
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
ARI :  0.4491 time :  13.0
--------------------------------------------------------------------------------
```

Prétraitement et modélisation partie NLP

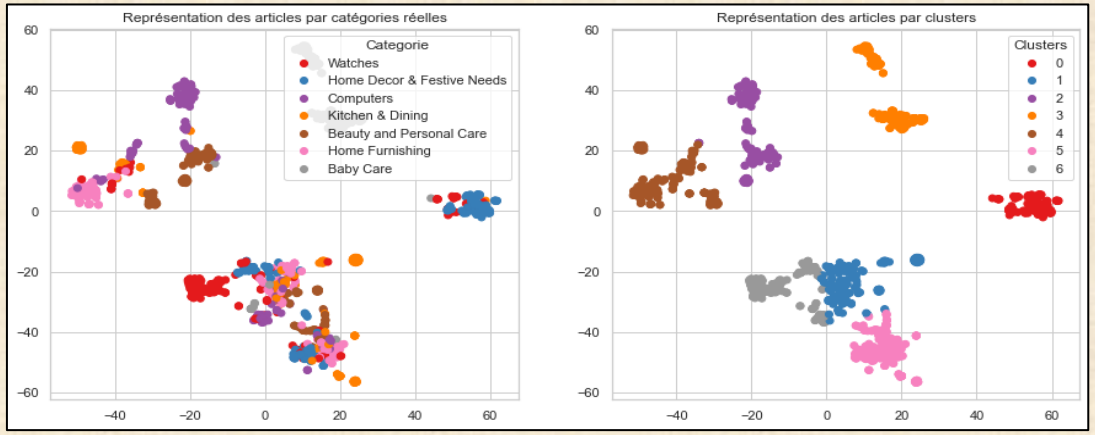
Création du bag of words (Tf-IDF + NMF): Negative Matrix Factorisation

```python
from sklearn.decomposition import NMF

# NMF is able to use tf-idf

no_topics = 7

# Run NMF
nmf = NMF(n_components=no_topics, random_state=1,
          alpha=.1, l1_ratio=.5, init='nndsvd')
nmf.fit(tfidf_fit_trans)

no_top_words = 10
display_topics(nmf, ctf.get_feature_names_out(),
               no_top_words)
```

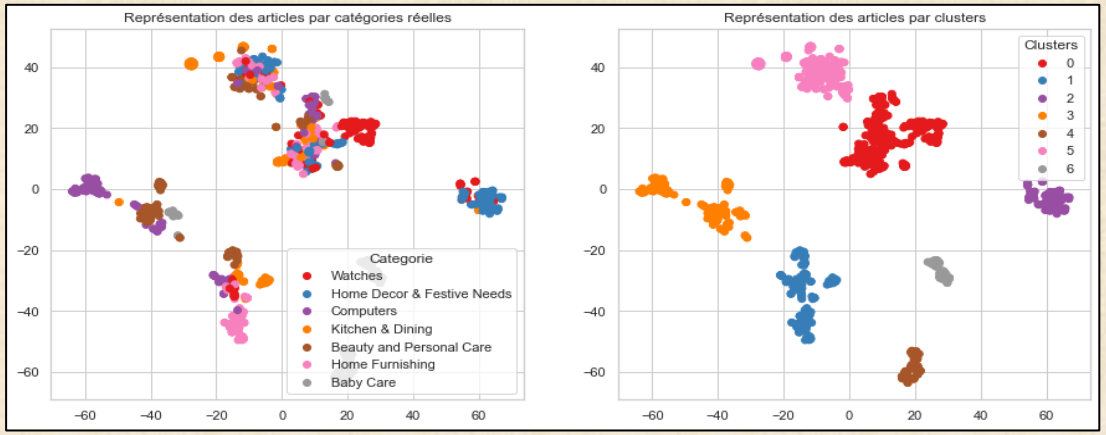
```
Topic 0:

com flipkart cash shipping genuine delivery guarantee buy free replacement

Topic 1:

watch analog men discount india great woman sonata maximum dial

Topic 2:

mug coffee ceramic perfect tea prithish printland one gift get

Topic 3:

baby girl detail cotton fabric dress sleeve boy neck pack

Topic 4:

rockmantra mug ceramic crafting porcelain permanent thrilling ensuring stay start

Topic 5:

laptop battery cell skin warranty pavilion lapguard adapter shape mouse

Topic 6:

abstract blanket double single com flipkart quilt comforter buy shipping
```

Prétraitement et modélisation partie NLP

Modèle d'embedding(Word2Vec): Embedding




```
Word2Vec Embedding :
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
ARI :  0.3023 time :  10.0
--------------------------------------------------------------------------------
```




```
Word2Vec Embedding :
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
ARI :  0.1961 time :  9.0
--------------------------------------------------------------------------------
```

```
print(model_vectors.similarity('package', 'lot'),'\n************')
print(model_vectors.similarity('girl', 'boy'),'\n************')
print(model_vectors.similarity('yellow', 'red'),'\n************')
```

```
0.08573361
************
0.8860109
************
0.6878723
************
```

```
model_vectors.most_similar('green')
```

```
[('yellow', 0.877837061882019),
 ('bucket', 0.8738964796066284),
 ('pink', 0.8515307307243347),
 ('grey', 0.8319422602653503),
 ('white', 0.819561779499054),
 ('giftsthatwow', 0.8166391253471375),
 ('cactus', 0.8129180073738098),
 ('mini', 0.8124147653579712),
 ('blue', 0.799816906452179),
 ('skyblue', 0.7961534261703491)]
```

```
print(model_vectors.similarity('package', 'lot'),'\n************')
print(model_vectors.similarity('girl', 'boy'),'\n************')
print(model_vectors.similarity('yellow', 'red'),'\n************')
```

```
0.0870305
************
0.9291192
************
0.6948403
************
```

```
model_vectors.most_similar('green')
```

```
[('white', 0.9115871787071228),
 ('yellow', 0.8791596293449402),
 ('rice', 0.8592472672462463),
 ('grey', 0.8552006483078003),
 ('pink', 0.8524134755134583),
 ('indigocart', 0.8492087125778198),
 ('multicolor', 0.8409568667411804),
 ('digilight', 0.8156141638755798),
 ('blue', 0.8143661618232727),
 ('clips', 0.8085724711418152)]
```

Le modèle d'embedding fonction mieux avec les données volumineuses c'est pour cela le score qu'on obtenu est mauvais par rapport aux données textuelles prétraitées vis à vis des données brutes

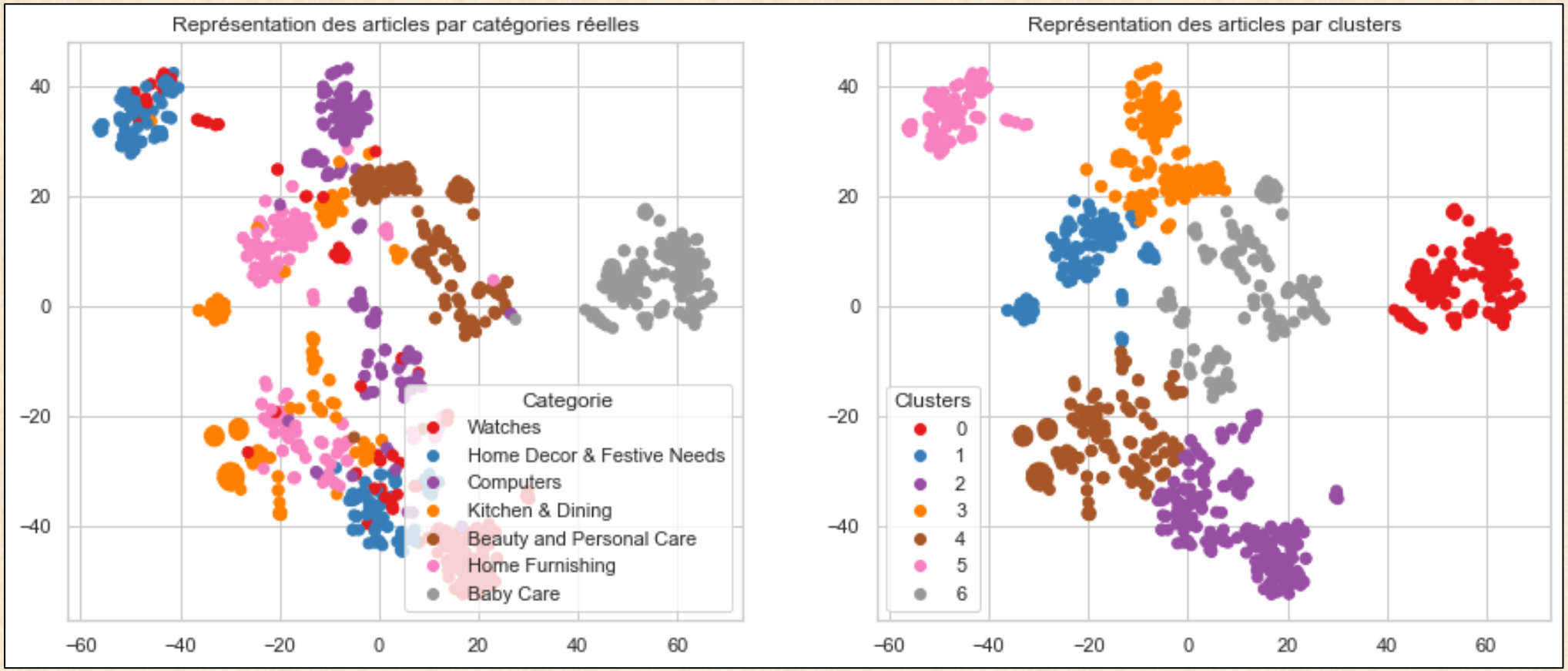
Prétraitement et modélisation partie NLP

Encodage avec BERT hub Tensorflow:


```
BERT hub Tensorflow :
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
ARI :  0.3099 time :  18.0
----------------------------------------------------------------------------------------------------
```

Prétraitement et modélisation partie NLP

USE - Universal Sentence Encoder




```
USE - Universal Sentence Encoder :
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
ARI :  0.4491 time :  25.0
----------------------------------------------------------------------------------------------------
```

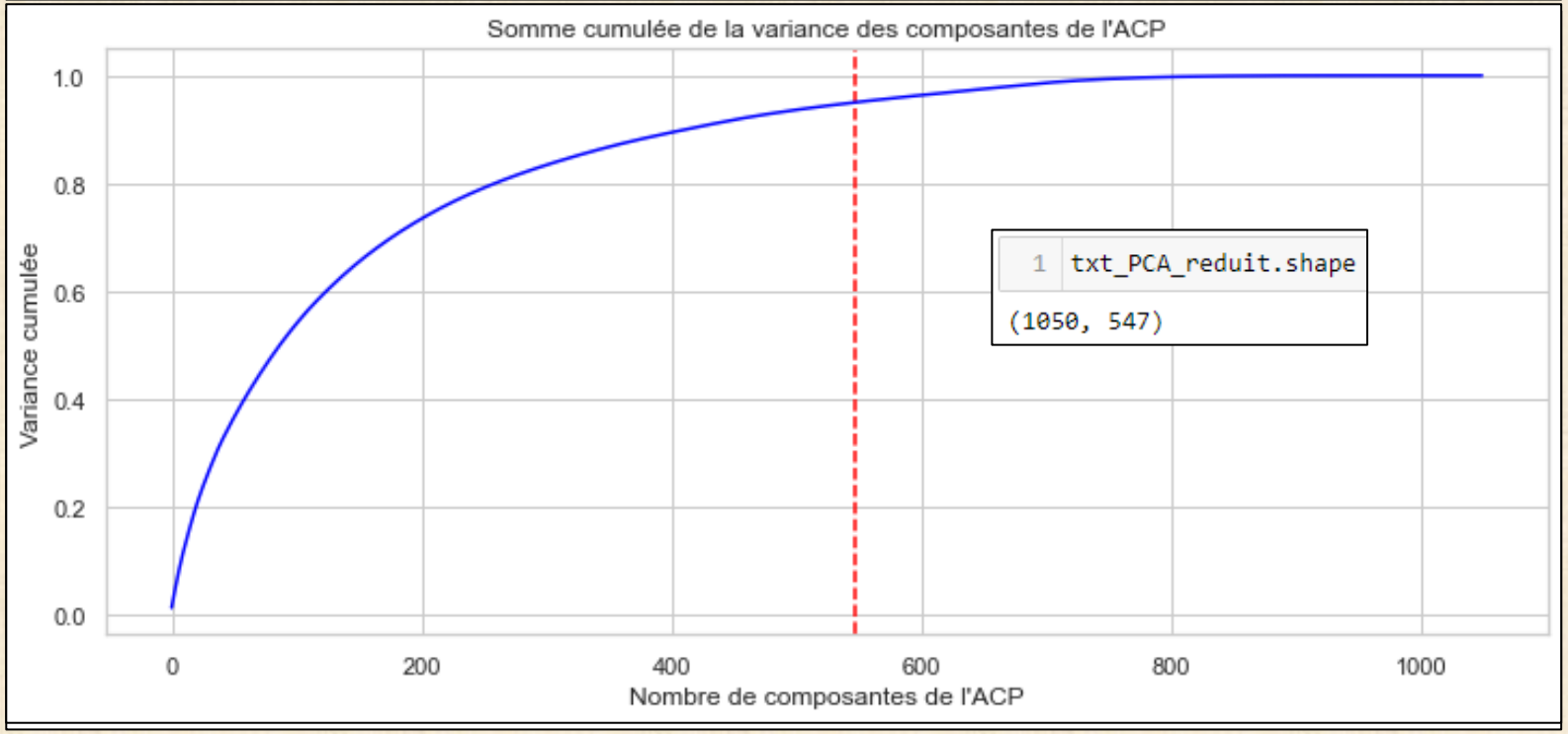
Prétraitement et modélisation partie NLP

Réduction de dimension ACP de données textuelles basé sur les données de CountVectorizer:

Les données réduites seront utilisé pour la partie approche combinée

```
1 txt_PCA_reduit, PCA_txt = PCA_features_reduction(np.asarray(X_), var_threshold=0.95, verbose=True)
```

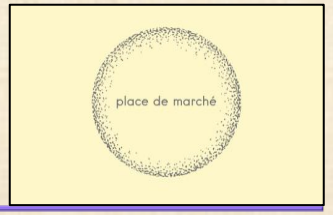
```
Initial number of features: 4381
Number of selected features: 547
Cumulative explained variance:  95.01%
```




```
1 txt_PCA_reduit.shape
```

```
(1050, 547)
```

Prétraitement et modélisation partie Image



Les étapes de prétraitement et modélisation d'image:

Prétraitement + Modélisation

Fonction intégrale de prétraitement

* le passage en gris
* De l'Auto contrast
* L'égalisation de l'histogramme
* Lissage pour eliminer le bruit
* Resize en 224*224

Fonctions d'extraction de Features

* Récupération la liste des image prétraitées
* Génération de Key-point, descripteur totale, et la matrice des descripteur descripteurs_array
* Création des Features à partir des descripteurs

Réduction dimensionnelle

Réduction dimensionnelle ACP

Classification non supervisée

Classification non supervisée des donnée réduites K-Means

Comparaison de la somme des inerties en fonction du nombre de groupes (clusters)

Coefficient de silhouette moyen en fonction du nombre de groupes (clusters)

IMAGE

Analyse de faisabilité

Exemple de prétraitement d'image

Prétraitement et modélisation partie Image

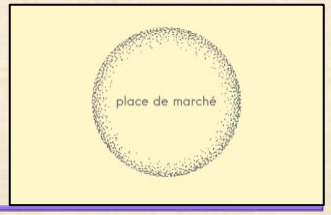
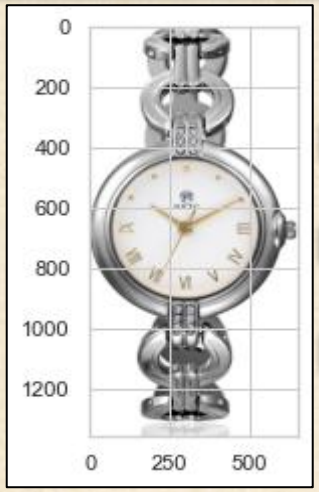
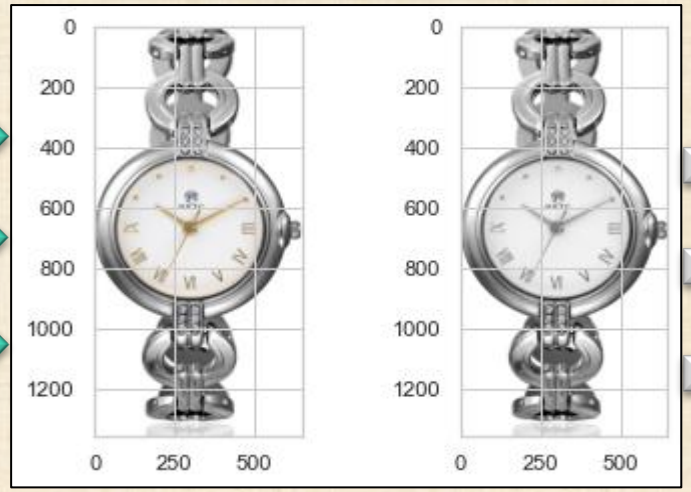


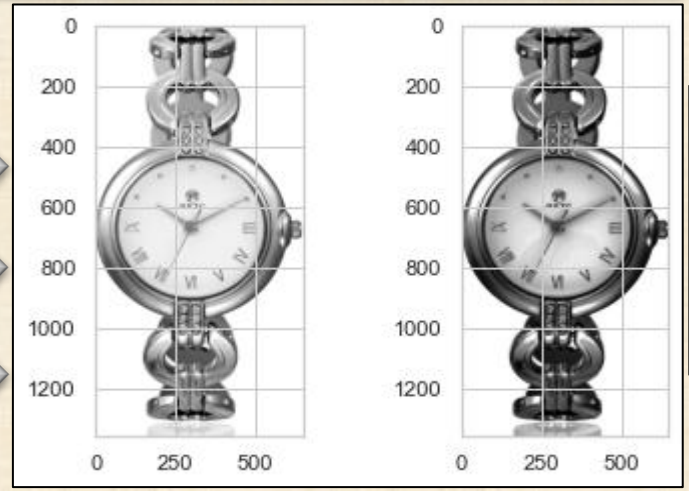
Image d'entrée RGB

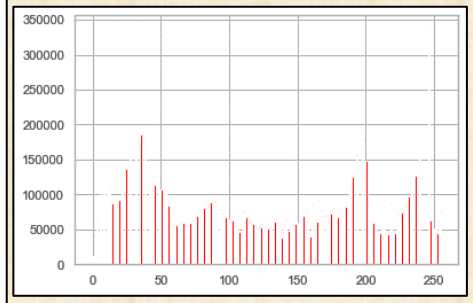


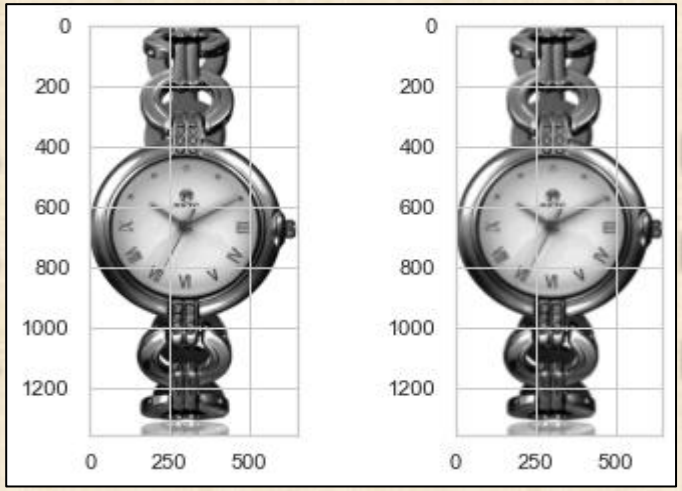
Passage en gray scale



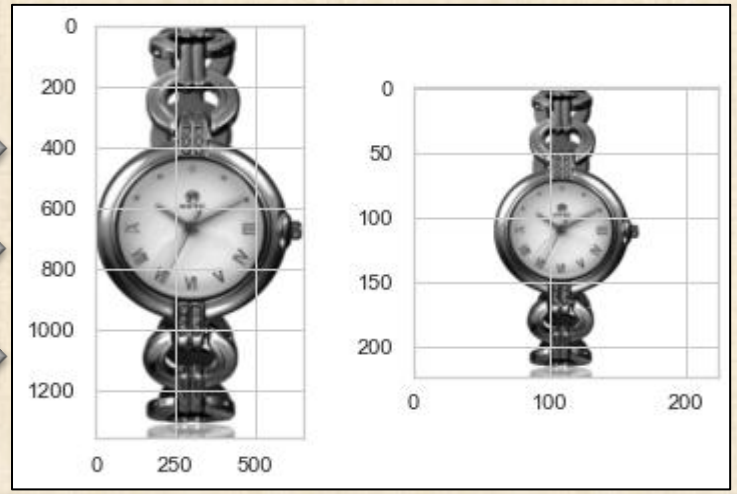
Contrast + égalisation d'histogramme







Filtre Gaussian Blur

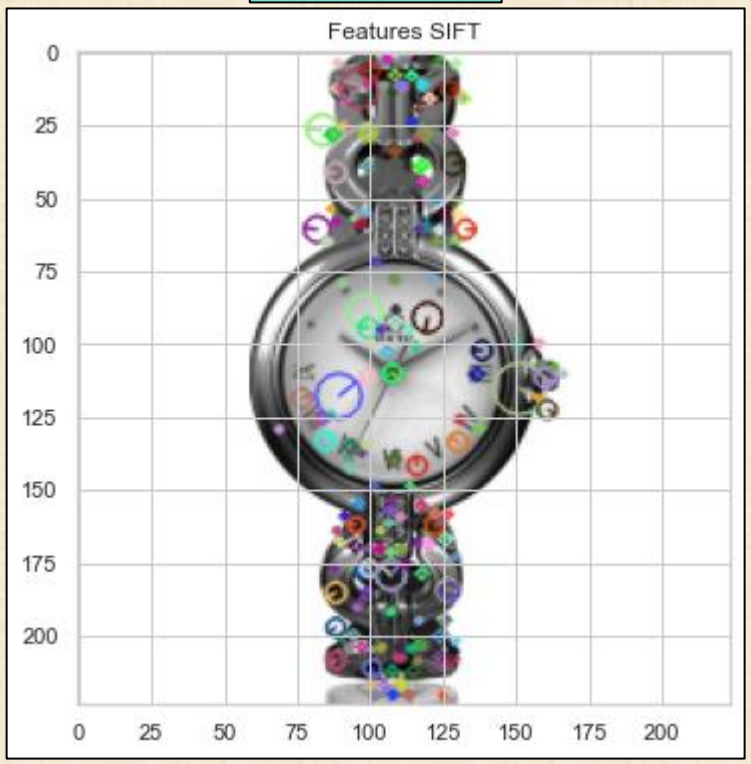


Redimensionnement de l'image 224*224

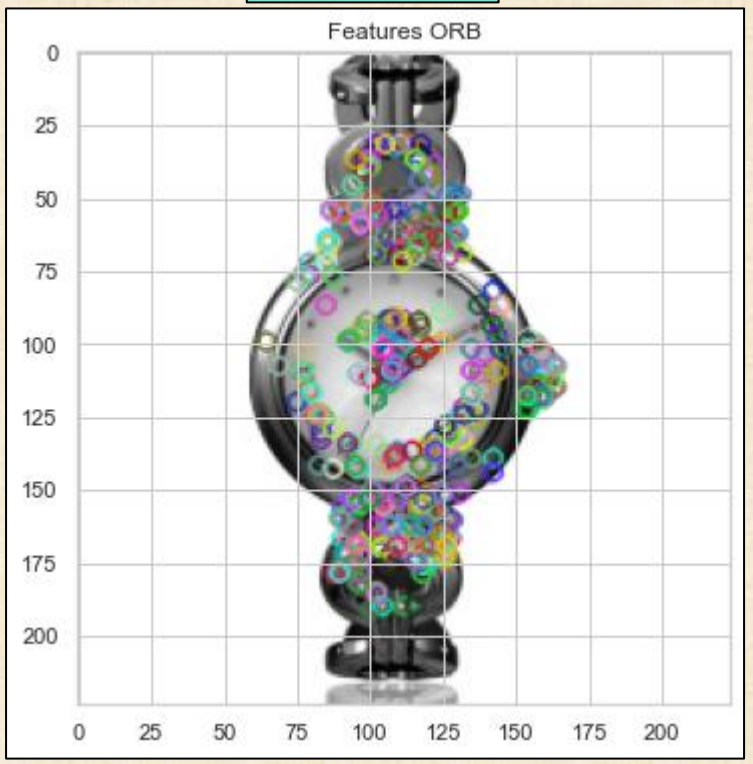
Prétraitement et modélisation partie Image

Test de l'algo SIFT et ORB

SIFT



ORB

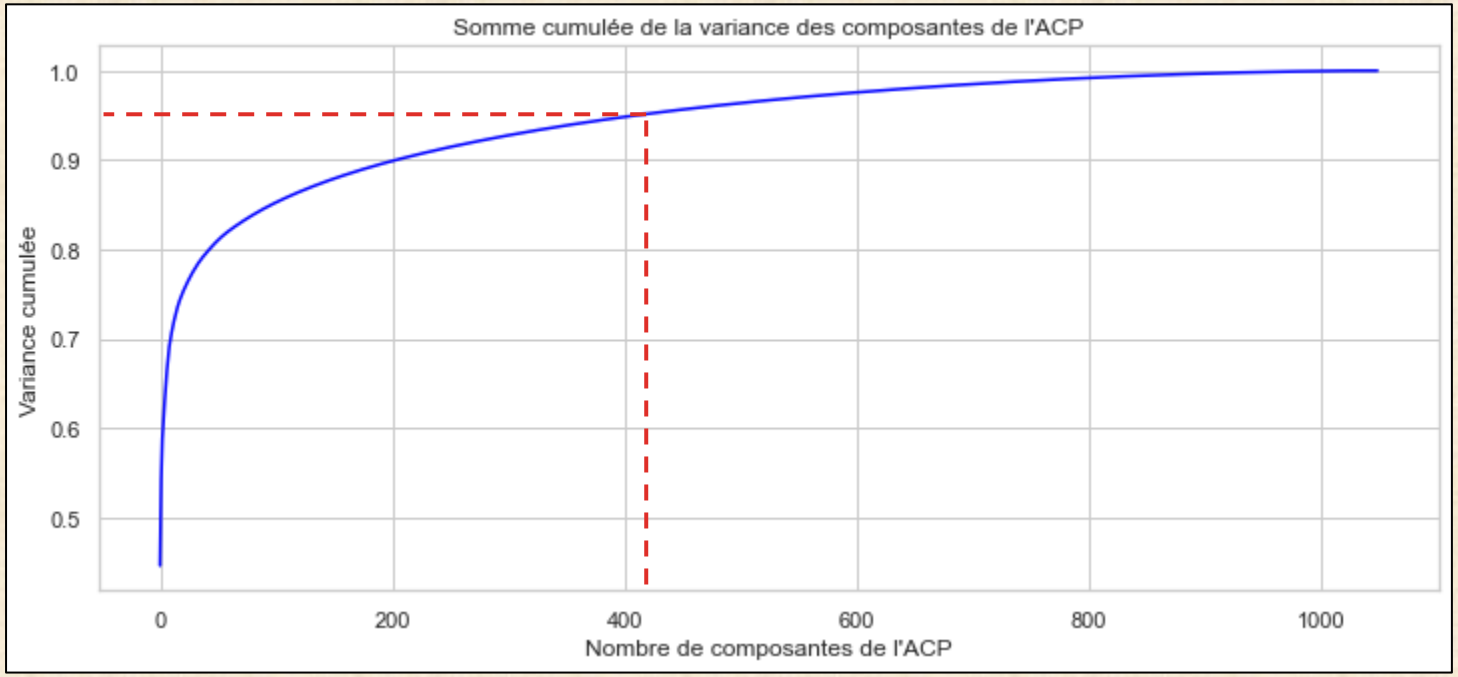


Prétraitement et modélisation partie Image

Extraction des Features:

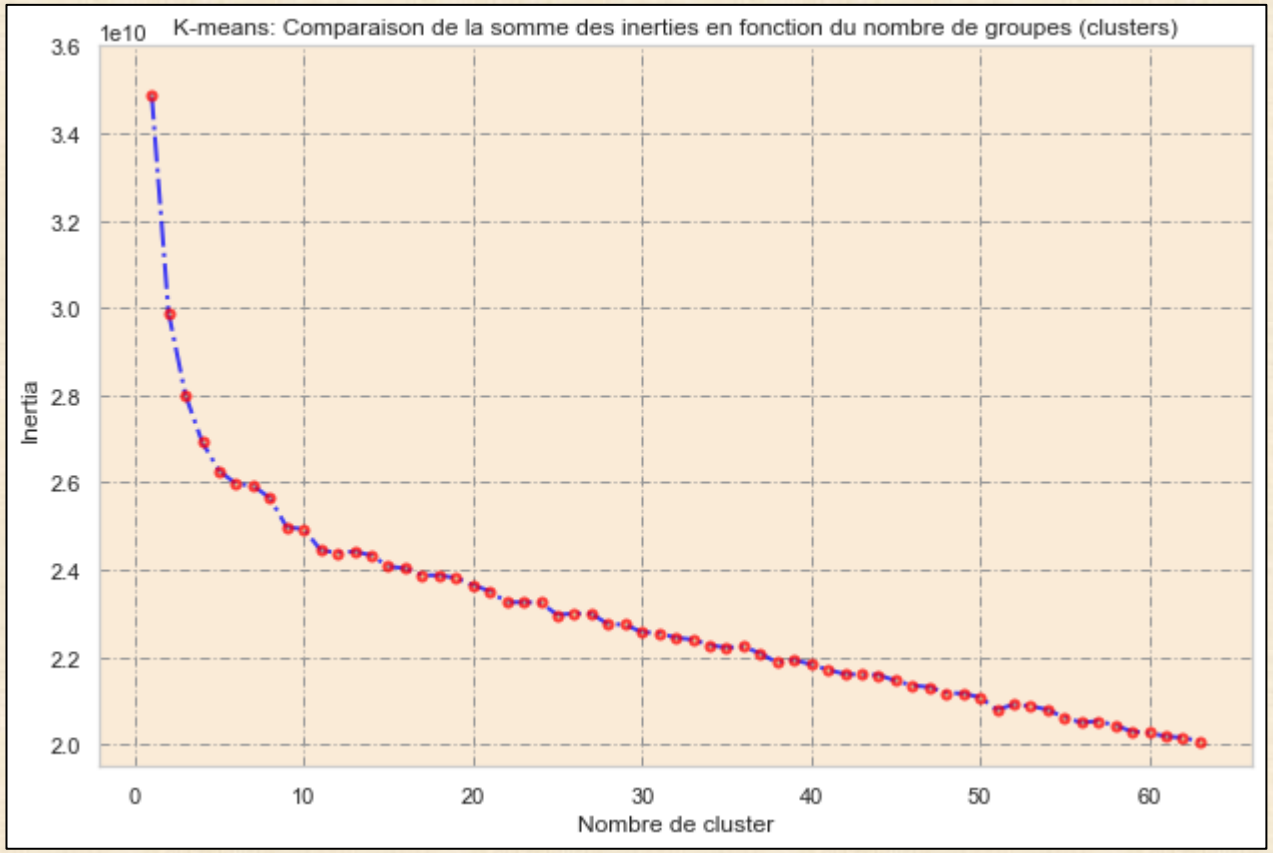
Prétraitement d'image → Extraction features **SIFT**
Funct: **extraction_features(**ListOfImages**)** → Matrice de descripteurs **descripteurs_array** → **ACP** → **Classification non supervisée (Clustering**

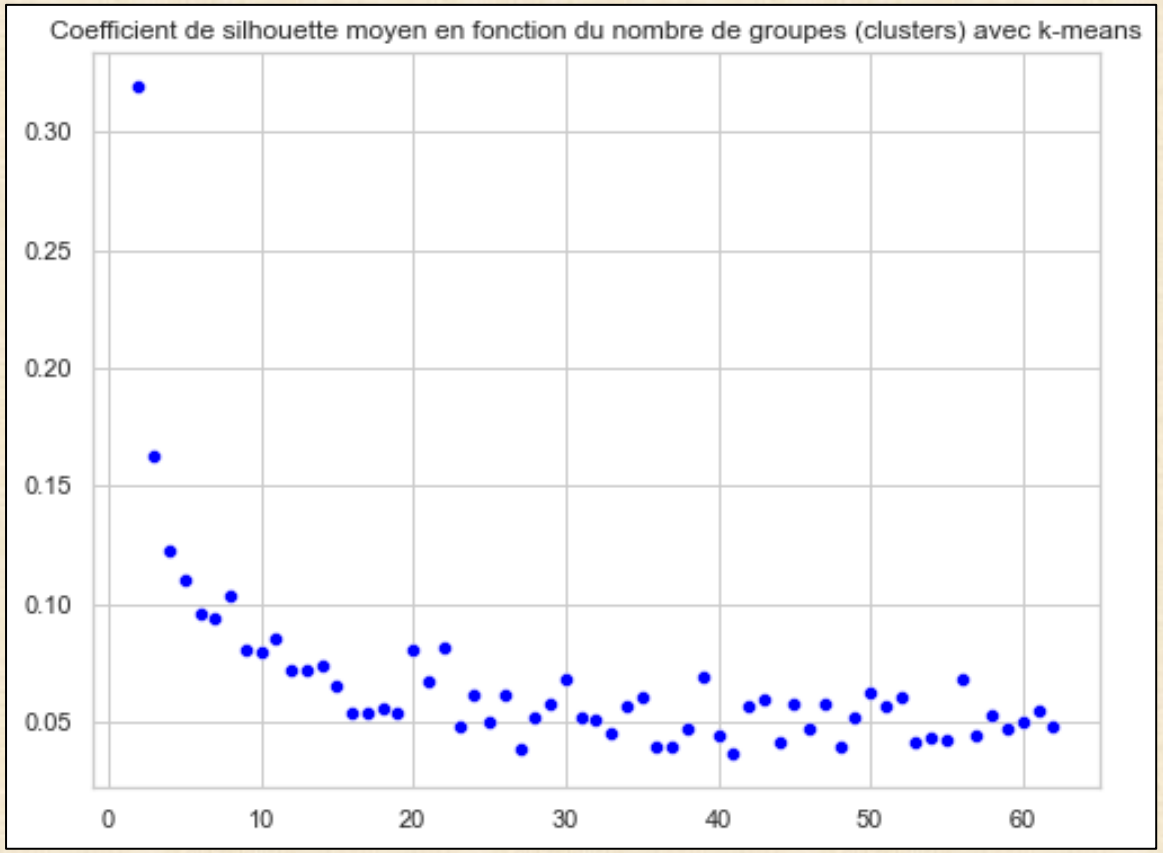
```
4  desc_PCA_reduit, pca = PCA_features_reduction(descripteurs_array, var_threshold=0.95, verbose=True)

Initial number of features: 257024
Number of selected features: 411
```



Prétraitement et modélisation partie Image

Clustering »K-Means « après réduction ACP:

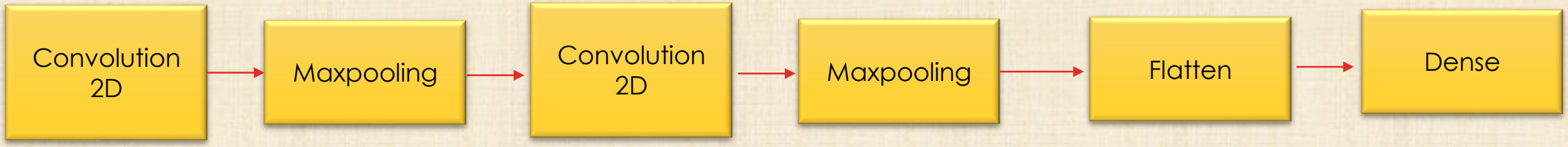





```
Le coefficient de silhouette est maximal pour 2 groupes (score = 0.3192656782116323)
```

Prétraitement et modélisation partie Image

Réseau de neurones: construction du réseau de neurones



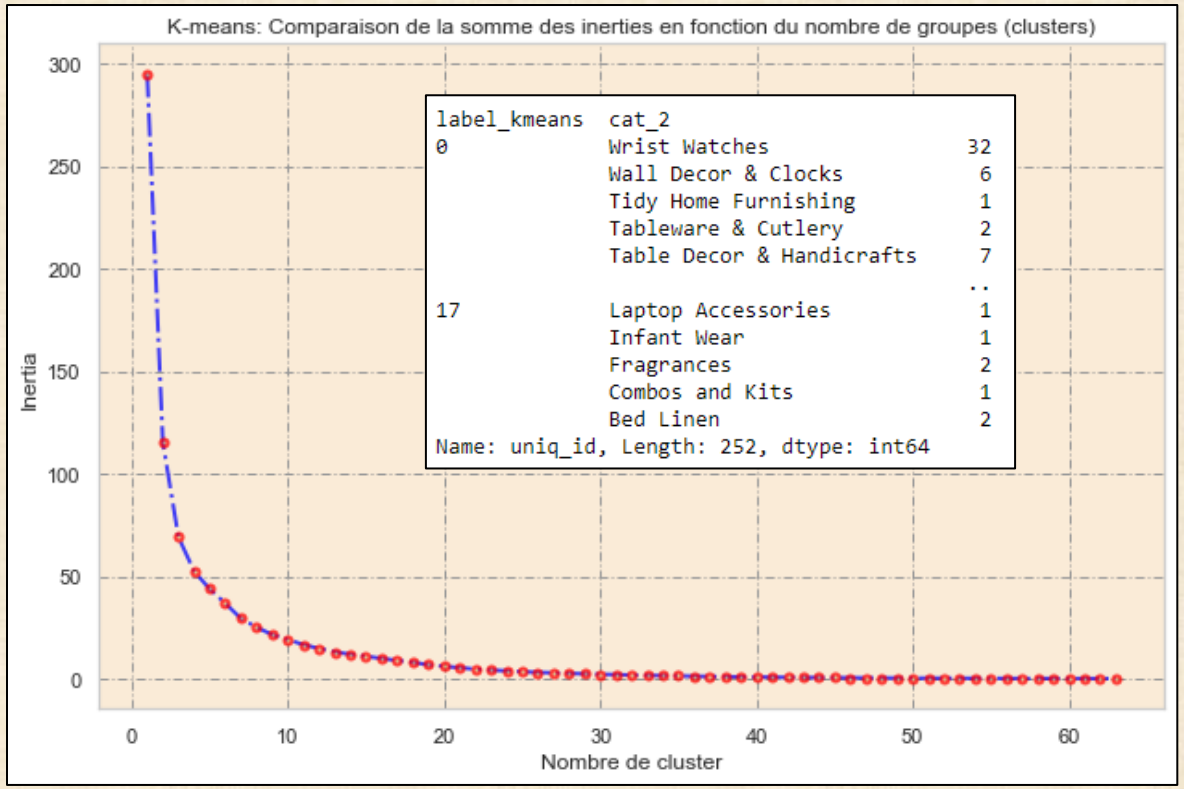
Entrainement supervisé: le score obtenu est de 0.05%, ce score n'est pas satisfaisant

```
1  accuracy_score(np.argmax(predictions_test, axis=1).reshape(-1,1),
2                 np.argmax(test_array_cats, axis=1).reshape(-1,1))
```

```
0.049429657794676805
```

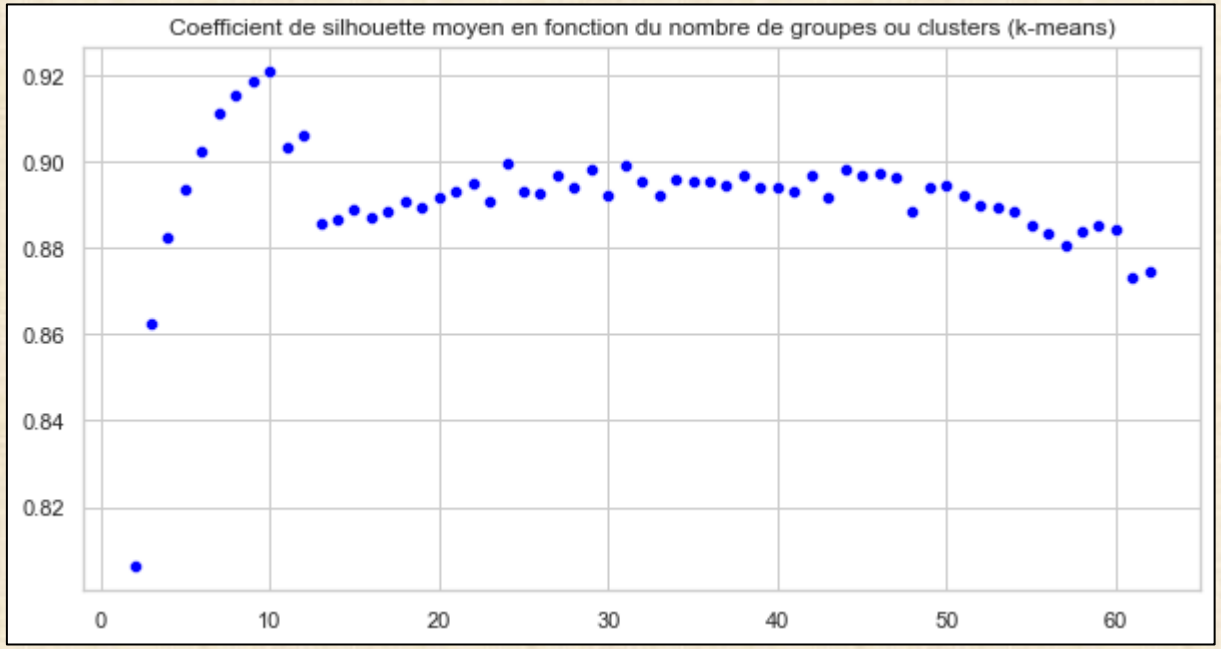
Prétraitement et modélisation partie Image

Classification non supervisée: pour accéder aux Features on enlève les deux couches du réseau de neurones



label_kmeans	cat_2	
0	Wrist Watches	32
	Wall Decor & Clocks	6
	Tidy Home Furnishing	1
	Tableware & Cutlery	2
	Table Decor & Handicrafts	7
		..
17	Laptop Accessories	1
	Infant Wear	1
	Fragrances	2
	Combos and Kits	1
	Bed Linen	2

Name: uniq_id, Length: 252, dtype: int64




```
for key, value in silhouettes_kmeans.items():
    if value == max(silhouettes_kmeans.values()):
        print('Le coefficient de silhouette est maximal pour {} clusters (score = {})'.format(key, value))
        nb_clusters_optimal = key
```

```
Le coefficient de silhouette est maximal pour 10 clusters (score = 0.9208968877792358)
```

Prétraitement et modélisation partie Image

Transfert learning:

Transfert learning

↓

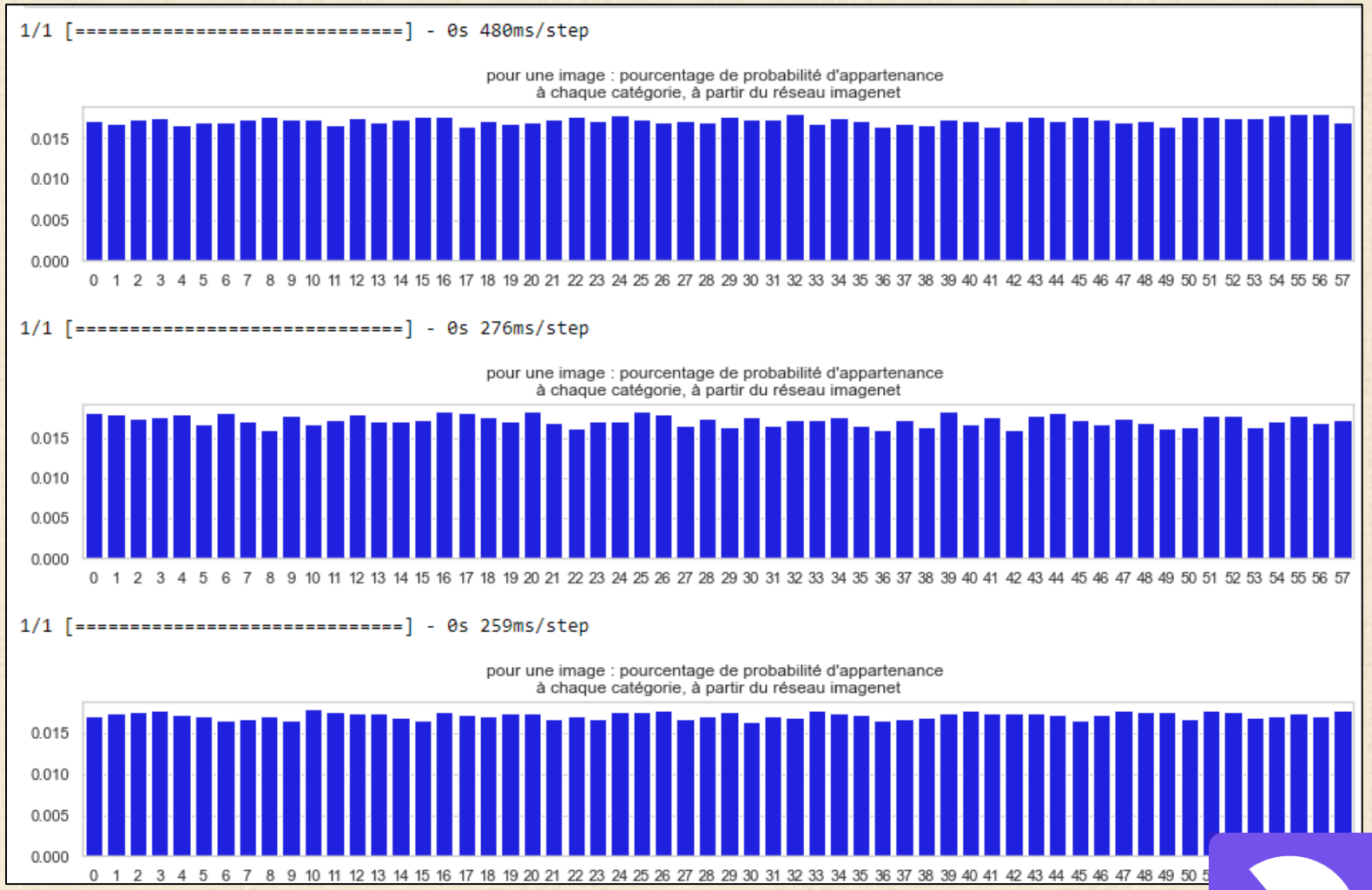
Substitution dernières couches par couche Dense

↓

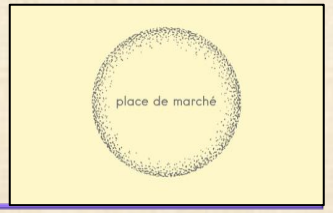
Préparation des données:
Redimensionnement
One-hot-encoding catégories

↓

Entraînement du réseau avec probabilité d'appartenance à chaque catégorie



Approche Combinée



Fusion des données textuelles et images

IMAGE

Données Images

Fusion des données

TEXTE

Données textuelles

Réduction dimensionnelle

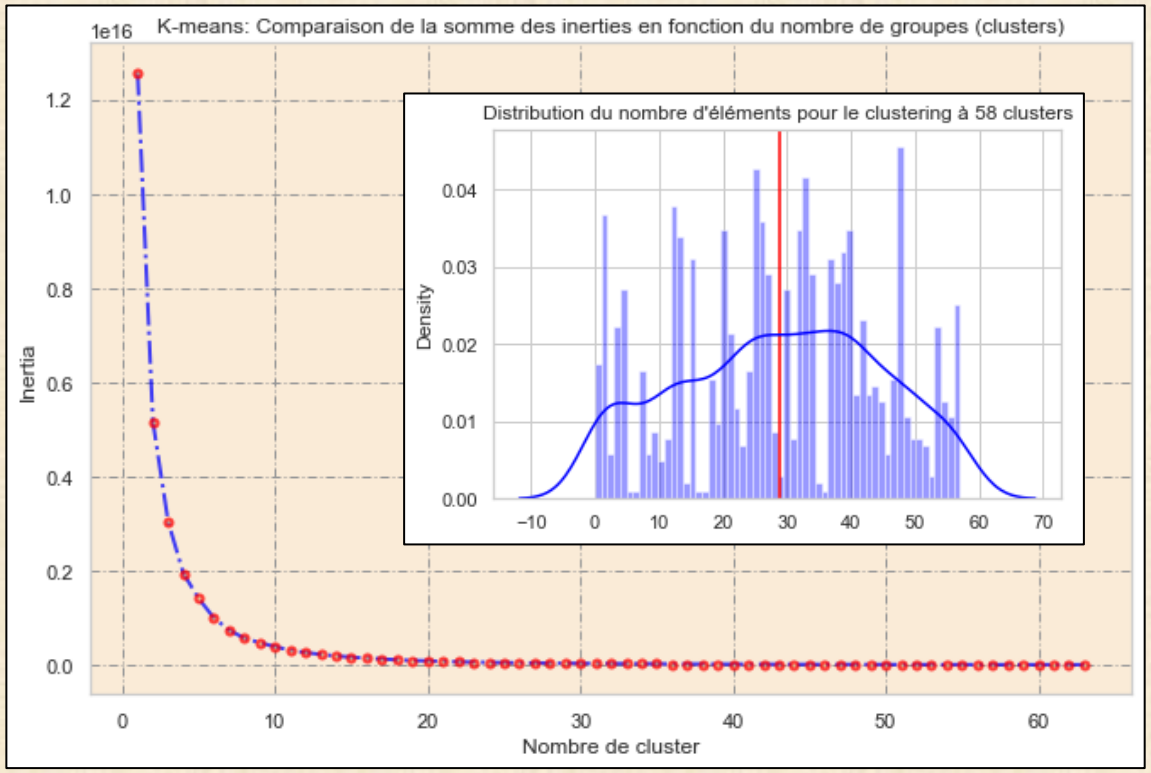
Partitionnement non supervisé

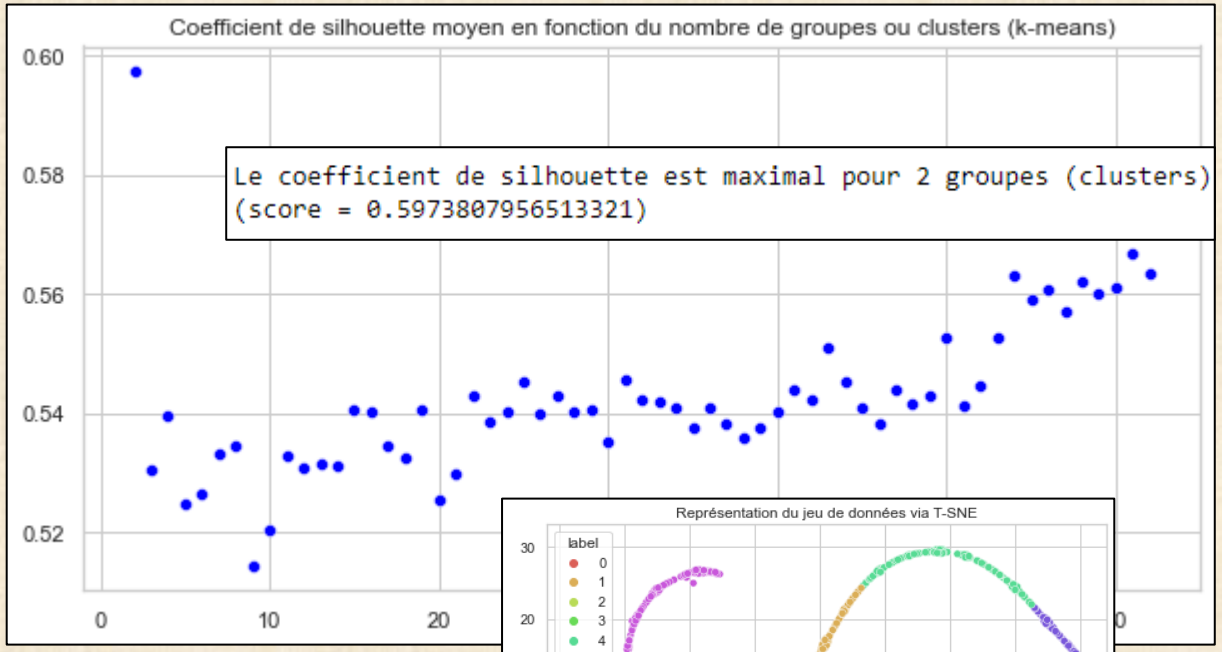
Evaluation de la correspondance des clusters avec les catégories « vraies »

Analyse

Approche combiné

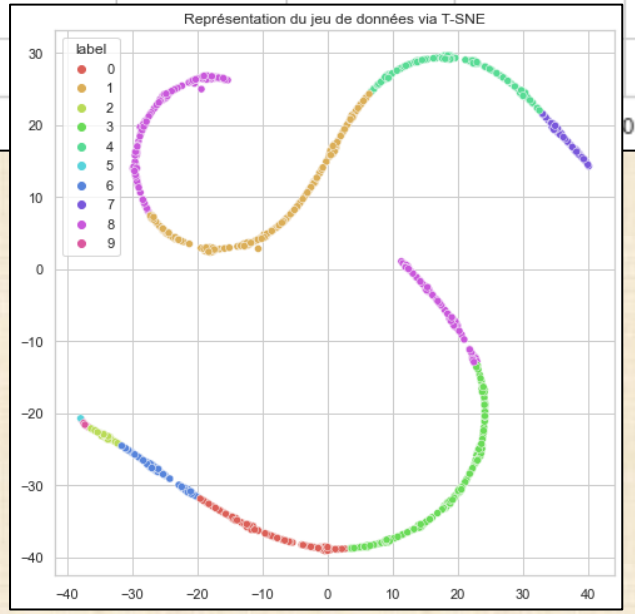
Clustering des données fusionnée






```
1 jeu_categoriel.shape
```

```
(1050, 548)
```



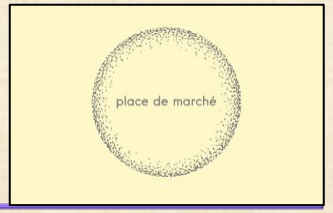

```
1 print('NLP          : ', X_NLP.shape)
2 print('Descripteurs : ', X_descripteurs.T.shape)
3 print('CNN          : ', X_CNN.shape)
```

```
NLP          :  (1050, 547)
Descripteurs :  (1050, 6)
CNN          :  (1050, 58)
```

```
1 X_combined.shape
```

```
(1050, 612)
```

Conclusion



- Possibilité de prédire les catégories grâce aux textes descriptifs et aux images

- Revoir les parties prétraitement d'image et texte pour améliorer les résultats.

Perspectives:

- Plus la taille du jeu de données est grande plus l'apprentissage des algos est mieux
- Créer un Data-Set mieux labélisé
- La faisabilité de la classification automatique peut s'améliorer si la qualité des descriptions (vocabulaire du e-commerce) ou des images(la netteté) sont bons,