

Présentation du projet 6 : Classifiez automatiquement des biens de consommation



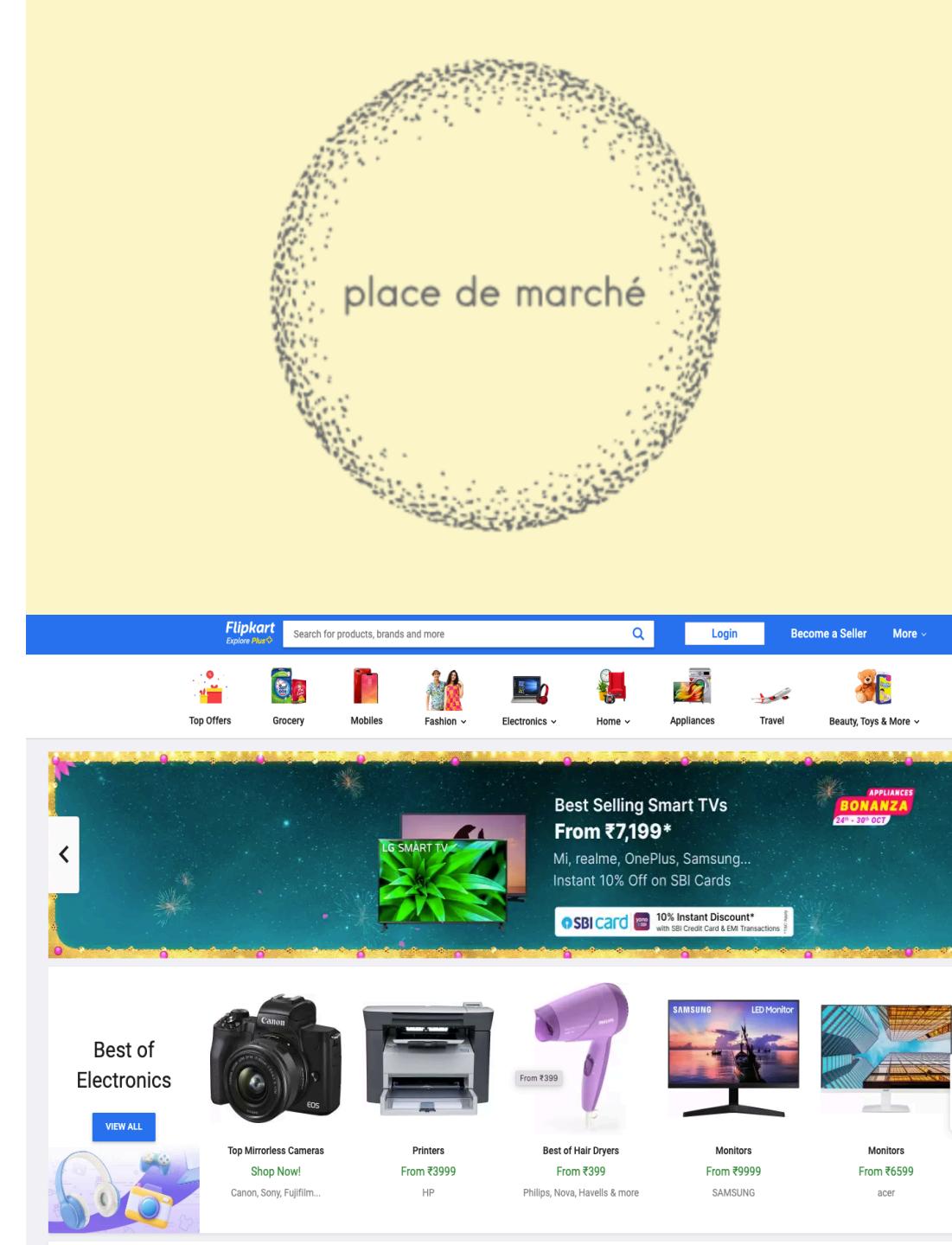
Préparé par :
MAGHLAZI Hanane

Projet 6 : Classifiez automatiquement des biens de consommation

- **Contexte**
- **Mission**
- **Démarche globale à suivre**
- **Exploration du jeu de données :**
 - Lecture des données
 - Vérification des données manquantes
- **Analyse exploratoires des données**
- **Traitement des données textuelles :**
 - Fonction de nettoyage texte
 - Bag of words
 - Tf-idf
 - Word2Vec
 - BERT
 - USE
- **Traitement des images :**
 - SIFT
 - ORB
 - CNN Transfer Learning
- **Approche supervisée**
- **Combinaison des données textuelles et images**
- **Conclusion**

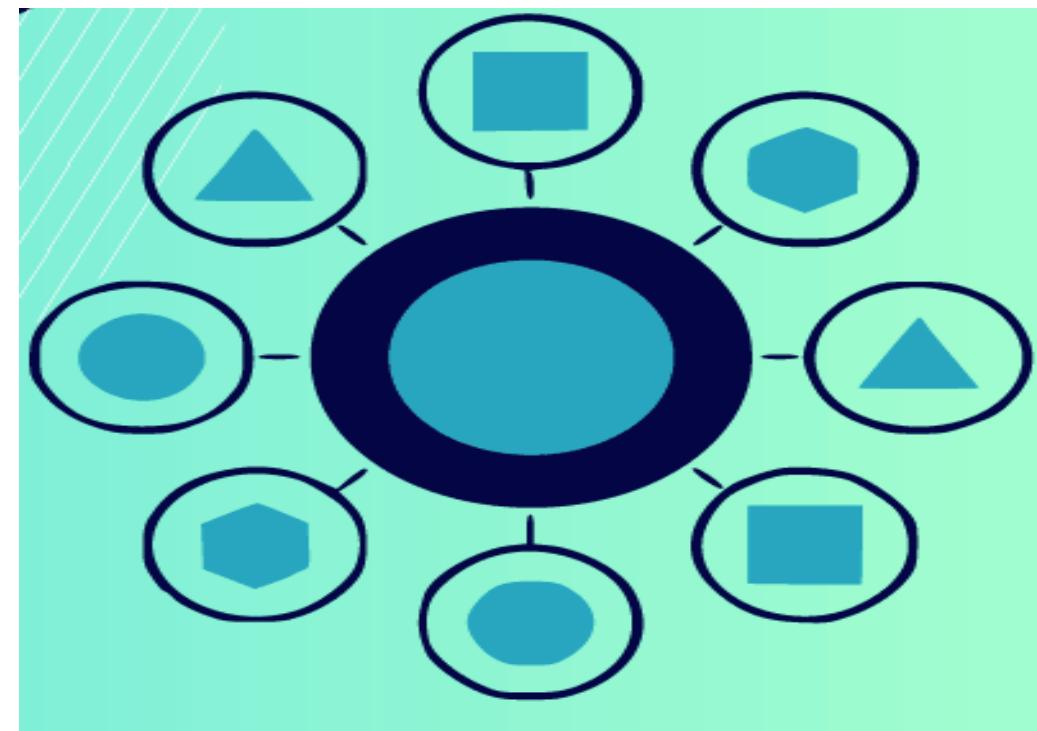
Contexte :

- Sur la **place de marché**, des vendeurs proposent des articles à des acheteurs en postant une **photo** et une **description**.
- L'attribution de la catégorie d'un article est effectuée **manuellement** par les vendeurs, et est donc **peu fiable**.
- Le volume des articles est pour l'instant très **petit** mais il est destiné à **s'accroître**.

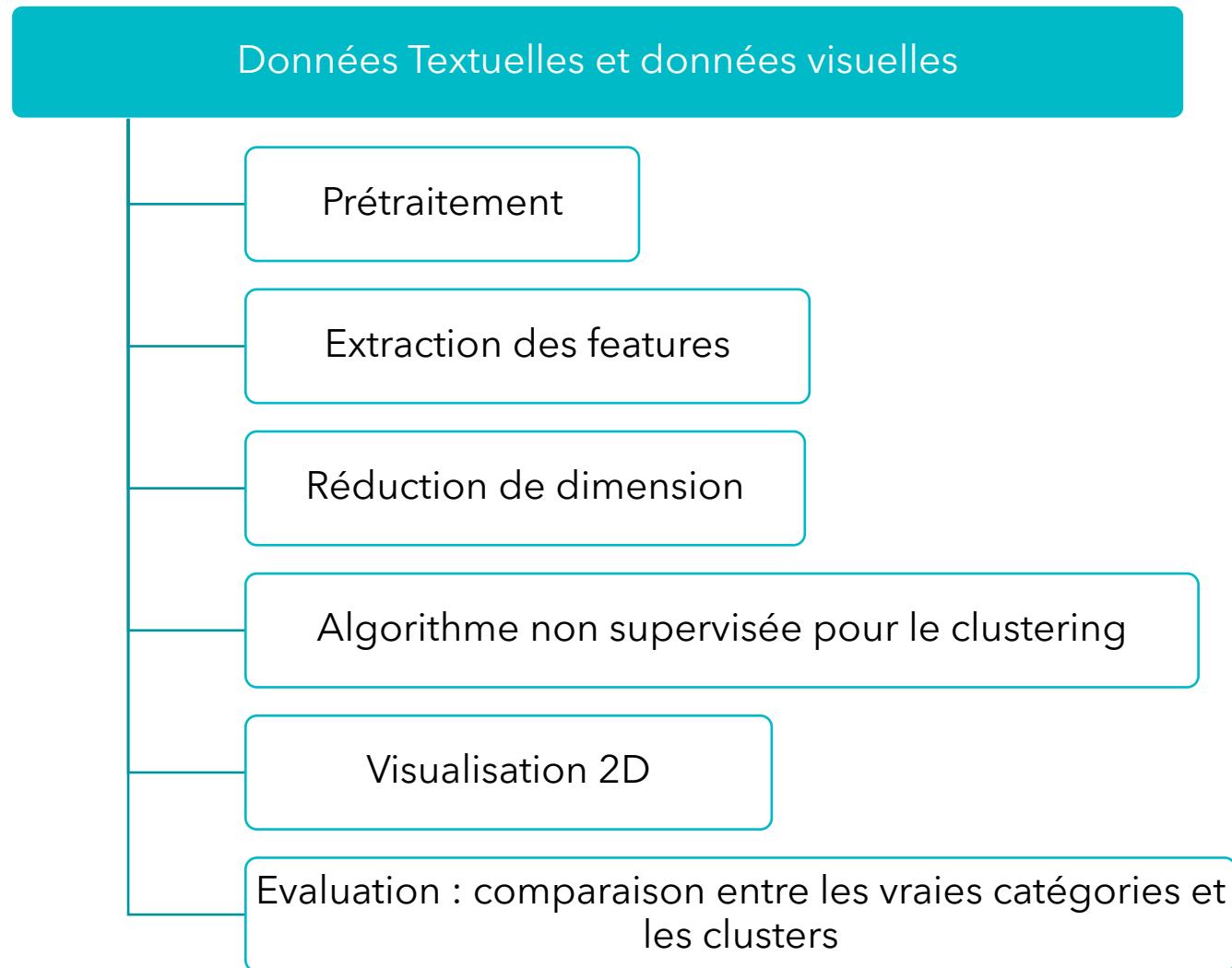


Mission :

- Réaliser une première étude de **faisabilité** **d'un moteur de classification** d'articles, basé sur une **image** et une **description**, pour l'automatisation de l'attribution de la catégorie de l'article.



Démarche globale à suivre :



Lecture du jeu de données:

- Présentation du jeu de données :

	Dimensions	NB_colonnes	NB_lignes	Total remplissage	Description
df	(1050, 15)	15	1050	15409	Fichier contenant des produits avec une description et image
<hr/>					
Data shape: (1050, 15)					
<hr/>					
Data types:					
object	12				types counts uniques nulls
float64	2				brand object 712 491 338
bool	1				crawl_timestamp object 1050 149 0
Name: types, dtype: int64					
<hr/>					
product_category_tree object 1050 642 0					
product_name object 1050 1050 0					
product_rating object 1050 27 0					
product_specifications object 1049 985 1					
product_url object 1050 1050 0					
retail_price float64 1049 355 1					
uniq_id object 1050 1050 0					

Données manquantes :

	Variable	nan	%nan
13	brand	338	32.1905%
6	retail_price	1	0.0952%
7	discounted_price	1	0.0952%
14	product_specifications	1	0.0952%
0	uniq_id	0	0.0000%
1	crawl_timestamp	0	0.0000%
2	product_url	0	0.0000%
3	product_name	0	0.0000%
4	product_category_tree	0	0.0000%
5	pid	0	0.0000%
8	image	0	0.0000%
9	is_FK_Advantage_product	0	0.0000%
10	description	0	0.0000%
11	product_rating	0	0.0000%
12	overall_rating	0	0.0000%

- Un dataframe bien **rempli**

Vérification des doublons :

```
La colonne uniq_id : duplicated 0
La colonne product_name : duplicated 0
La colonne product_category_tree : duplicated 408
La colonne retail_price : duplicated 695
La colonne discounted_price : duplicated 625
La colonne image : duplicated 0
La colonne is_FK_Advantage_product : duplicated 1048
La colonne description : duplicated 0
La colonne product_rating : duplicated 1023
La colonne brand : duplicated 559
La colonne product_specifications : duplicated 65
```

Données texte et image :

Texte : Nom, description et catégorie de l'article

Product_name:

Calibro SW-125 Analog-Digital Watch - For Men, Boys

Description: Calibro SW-125 Analog-Digital Watch - For Men, Boys
Price: Rs. 699
CALIBRO presents MTG Black Dial Round Watch...

Product_category_tree :

["Watches >> Wrist Watches >> Calibro Wrist Watches"]
WATECGPSDSFRHUSY

Image : une image par produit



Partie Texte

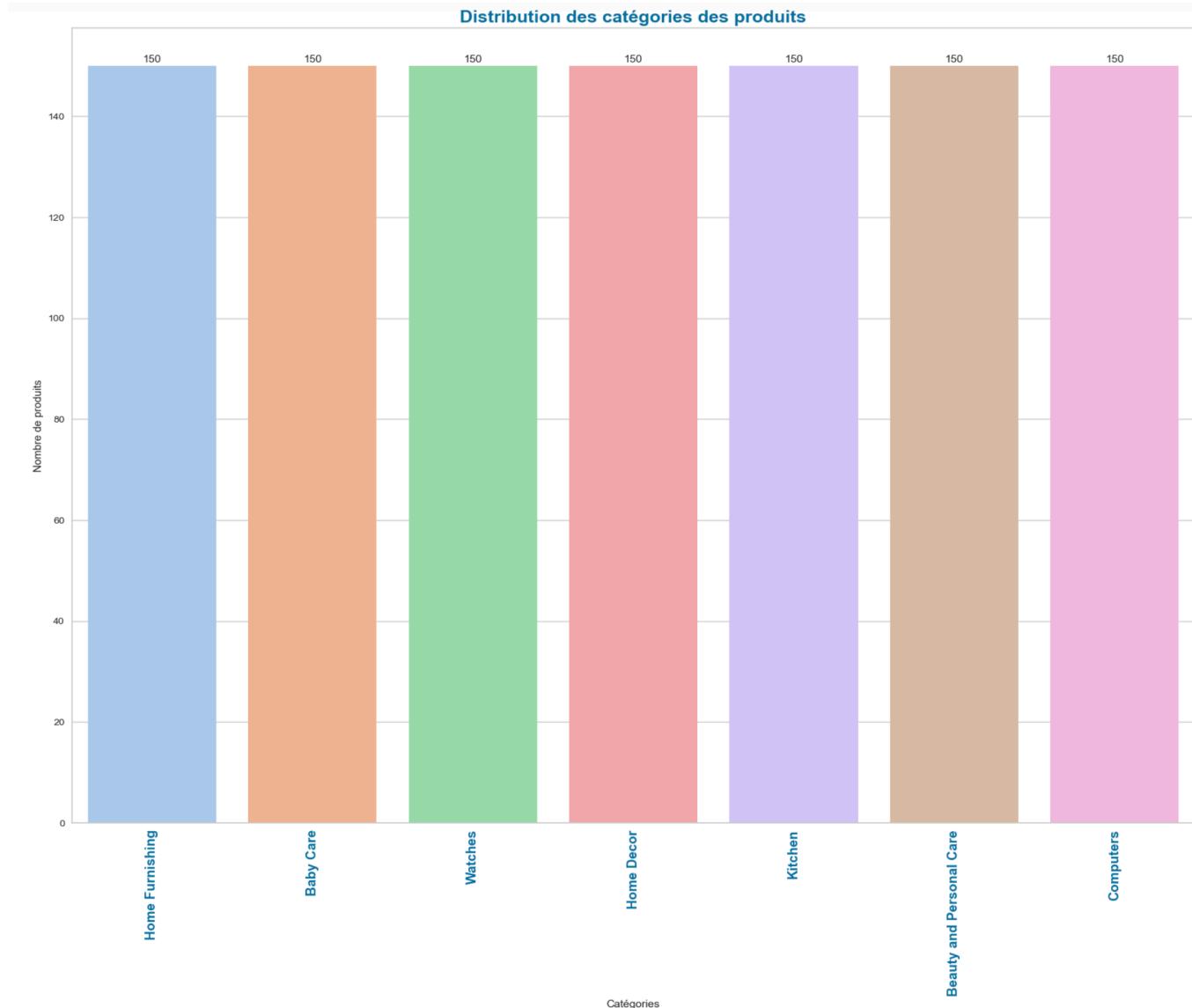
'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain , elegance polyester mult color abstract eyelet door curtain (213 cm in height , pack of 2) price : rs . 899 this curtain enhances the look of the interiors.this curtain is made from 100 % high quality polyester fabric.it features an eyelet style stitch with metal ring.it makes the room environment romantic and loving.this curtain is ant wrinkle and anti shrinkage and have elegant appearance.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 draw them apart first thing in the morning to welcome the bright sun rays you 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. , specifications of elegance polyester multicolor abstract eyelet door curtain (213 cm in height , pack of 2) general brand elegance designed for door type eyelet model name abstract polyester door curtain set of 2 model id duster25 color multicolor dimensions length 213 cm in the box number of contents in sales package pack of 2 sales package 2 curtains body & design material polyester',

'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 asvtwl322 color red , yellow , blue size medium dimensions length 30 inch width 60 inch in the box number of contents in sales package 3 sales package 3 bath towel',

'key features of eurospa cotton terry face towel set size : small height : 9 inch gsm : 360 , eurospa cotton terry face towel set (20 piece face towel set , assorted) price : rs . 299 eurospa brings to you an exclusively designed ,

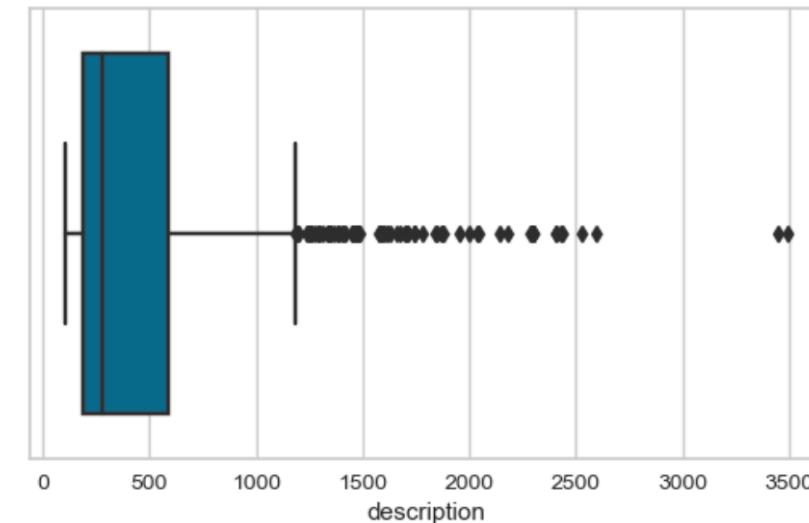
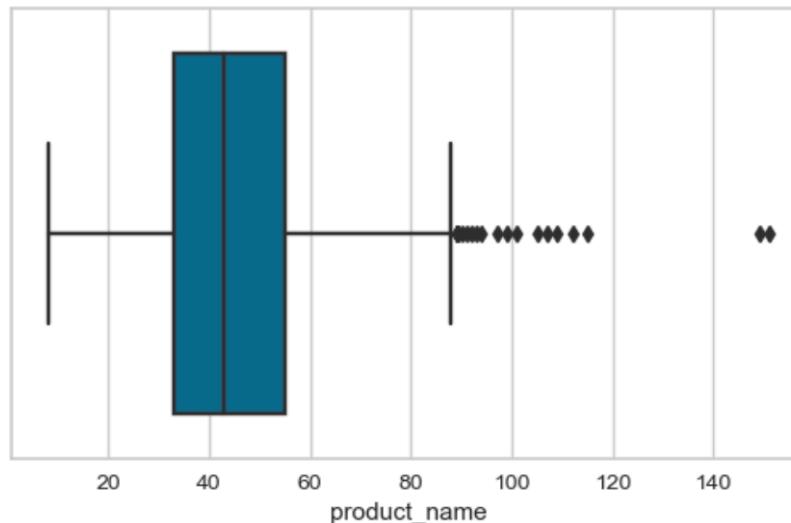
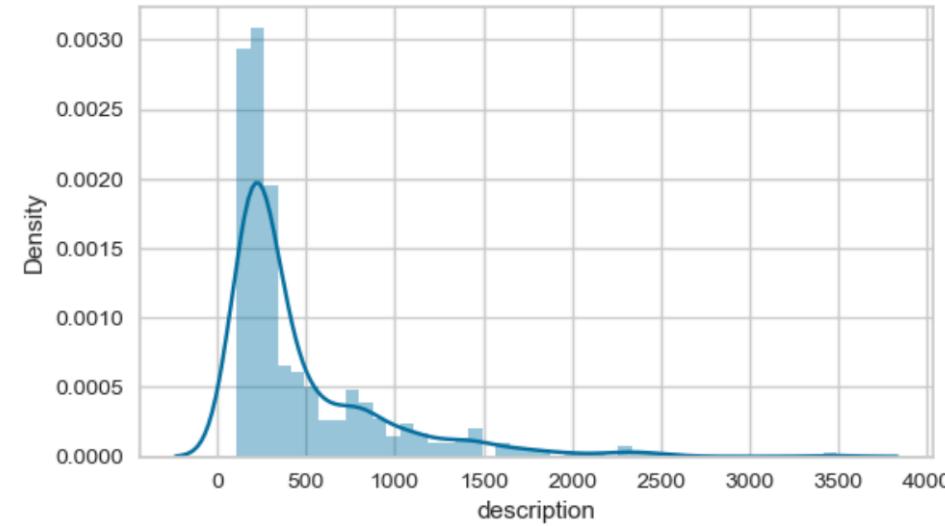
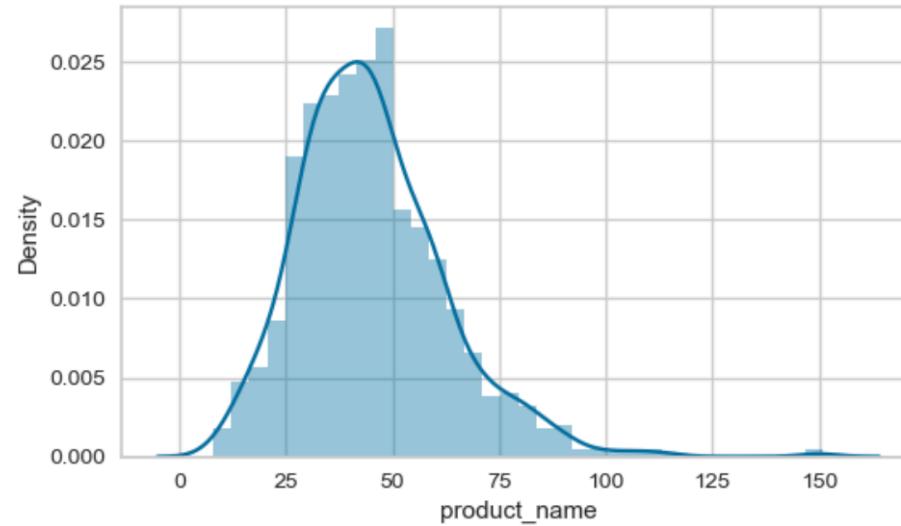
Analyse exploratoire : Texte

- La variable :
product_category_tree contient les niveaux de catégorie du produit.
- Pour cette étude on se contentera du **premier** niveau qui contient **7** catégories de produits bien réparties (150 articles par catégories).



Analyse exploratoire : Texte

Distribution des longueurs du nom et description du produit :



Analyse exploratoire : Texte

	_len_product_name	_len_description
count	1050.0000	1050.0000
mean	45.1019	473.8210
std	17.4973	457.9104
min	8.0000	109.0000
25%	33.0000	192.0000
50%	43.0000	278.0000
75%	55.0000	588.2500
max	151.0000	3490.0000

Analyse exploratoire : Texte

- **Nom** de produits pour chaque **catégorie** :

categorie: Kitchen

Prithish Friend Indeed Ceramic Mug...

categorie: Computers

Airtel B310s-927...

categorie: Home Decor

Tatvaarts Tribal Face Showpiece - 21.59 cm...

categorie: Home Furnishing

Rajasthan Crafts Abstract Single Quilts & Comforters Pink...

- **Description** pour chaque **catégorie** :

categorie: Kitchen

Buy Prithish Friend Indeed Ceramic Mug for Rs.225 online. Prithish Friend Indeed Ceramic Mug at best prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee....

categorie: Computers

Buy Airtel B310s-927 only for Rs. 2700 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!...

categorie: Home Decor

Tatvaarts Tribal Face Showpiece - 21.59 cm (Brass, Gold)

Price: Rs. 6,100

Analyse exploratoire : Texte

Watches



Kitchen



Computers



Baby Care



Beauty and Personal Care



Home Decor

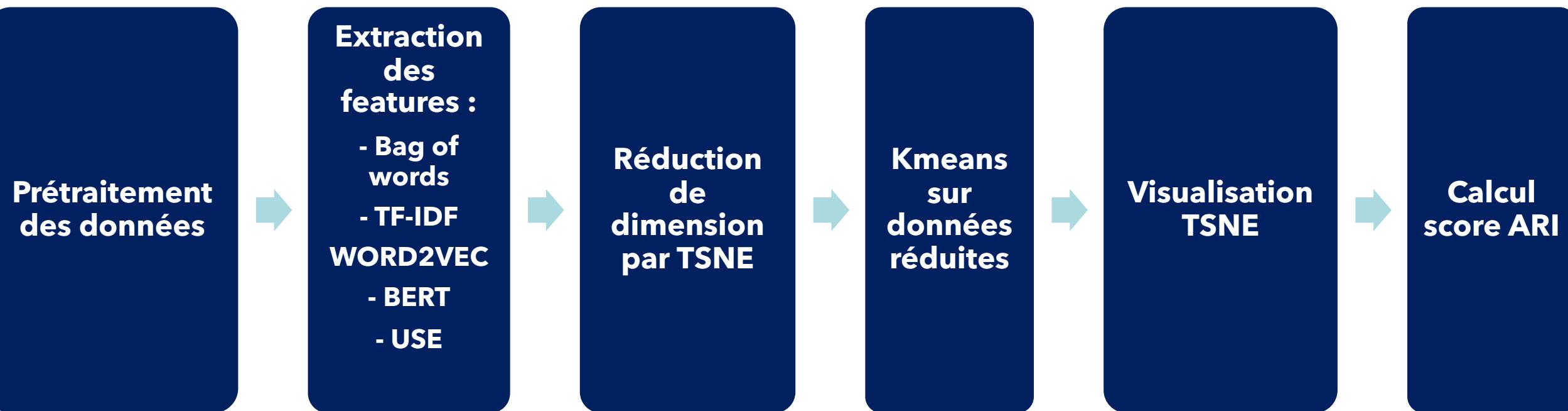


Home Furnishing



- _ On distingue bien des mots **clés significatifs** sur quelques **catégories**.
 - _ D'autres catégories contiennent des mots **non significatifs**.

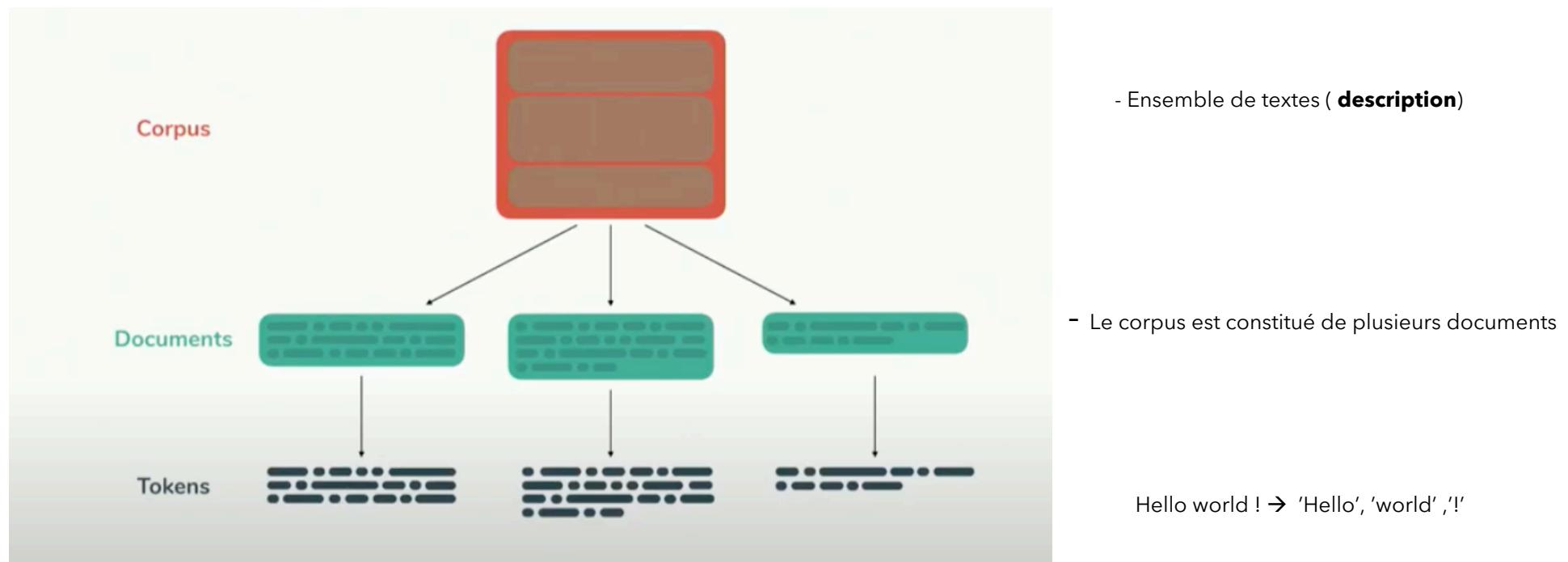
Etapes à suivre: Texte



Analyse exploratoire : Texte

Le **NLP**, ou Natural Language Processing, est une branche de l'intelligence artificielle qui permet aux machines de mieux comprendre le **langage humain**.

Il existe plusieurs techniques de NLP comme la traduction de texte, les chatbots, la classification de textes, le résumé d'un contenu ...



Analyse exploratoire : Texte

- **- Prétraitement des données :** Récupérer le corpus de textes qu'on pourra exploiter dans des algorithmes de machine Learning

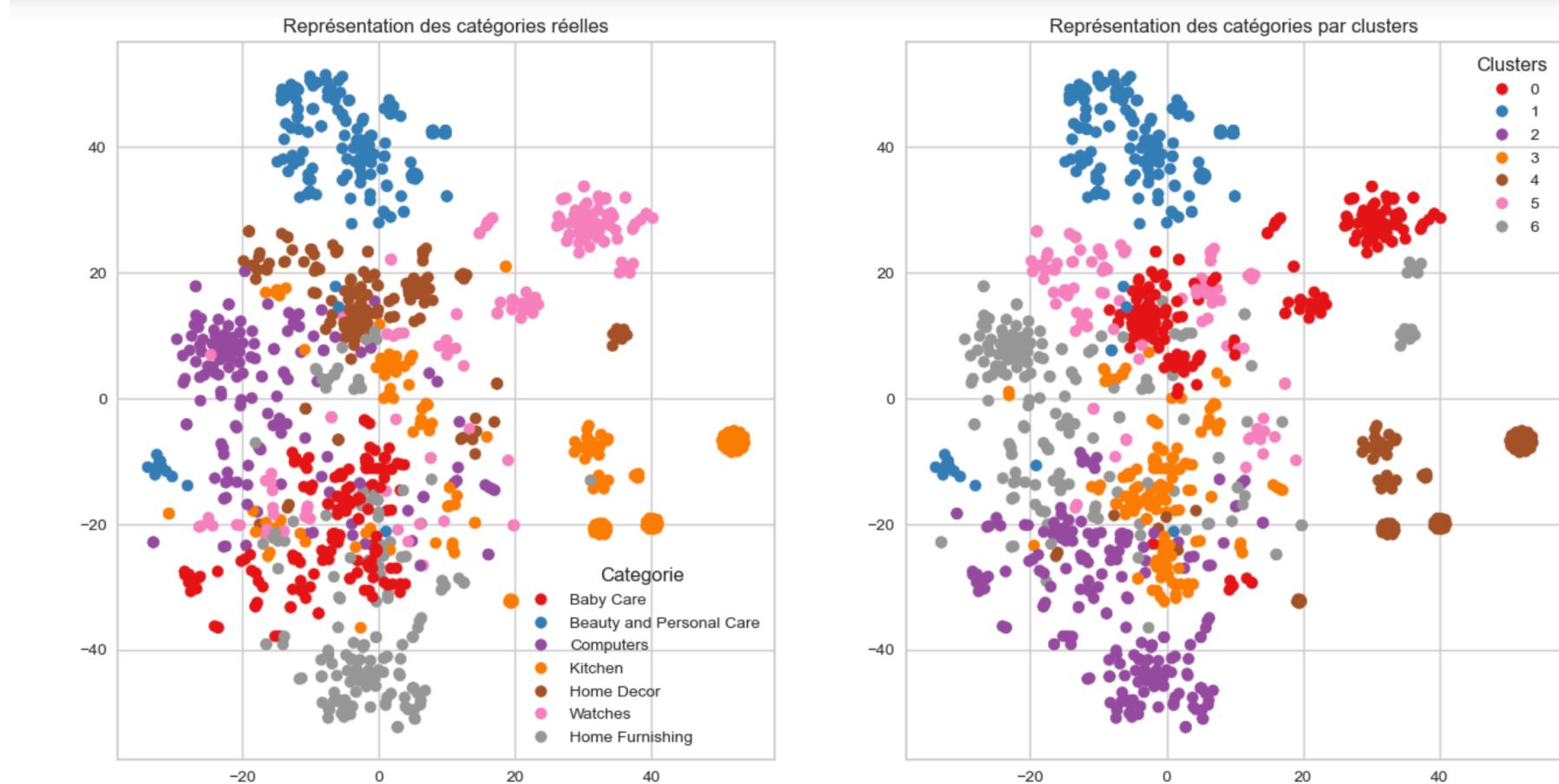


Natural Language Processing

- **bag-of-words** : On représente chaque document par un vecteur de la taille du vocabulaire et on utilisera la matrice composée de l'ensemble de ces N documents qui forment le corpus comme entrée de nos algorithmes. En gros utiliser les **fréquences d'apparition** des différents mots.

	bien	crêpes	fais	je	NLP	oui	python	va
Je fais du python	0	0	1	1	0	0	1	0
Python ? Oui, Python va bien avec le NLP	1	0	0	0	1	1	1	1
Je fais du NLP avec python	0	0	1	1	1	0	1	0
Je fais des crêpes	0	1	1	1	0	0	0	0

Natural Language Processing



ARI : 0.3743

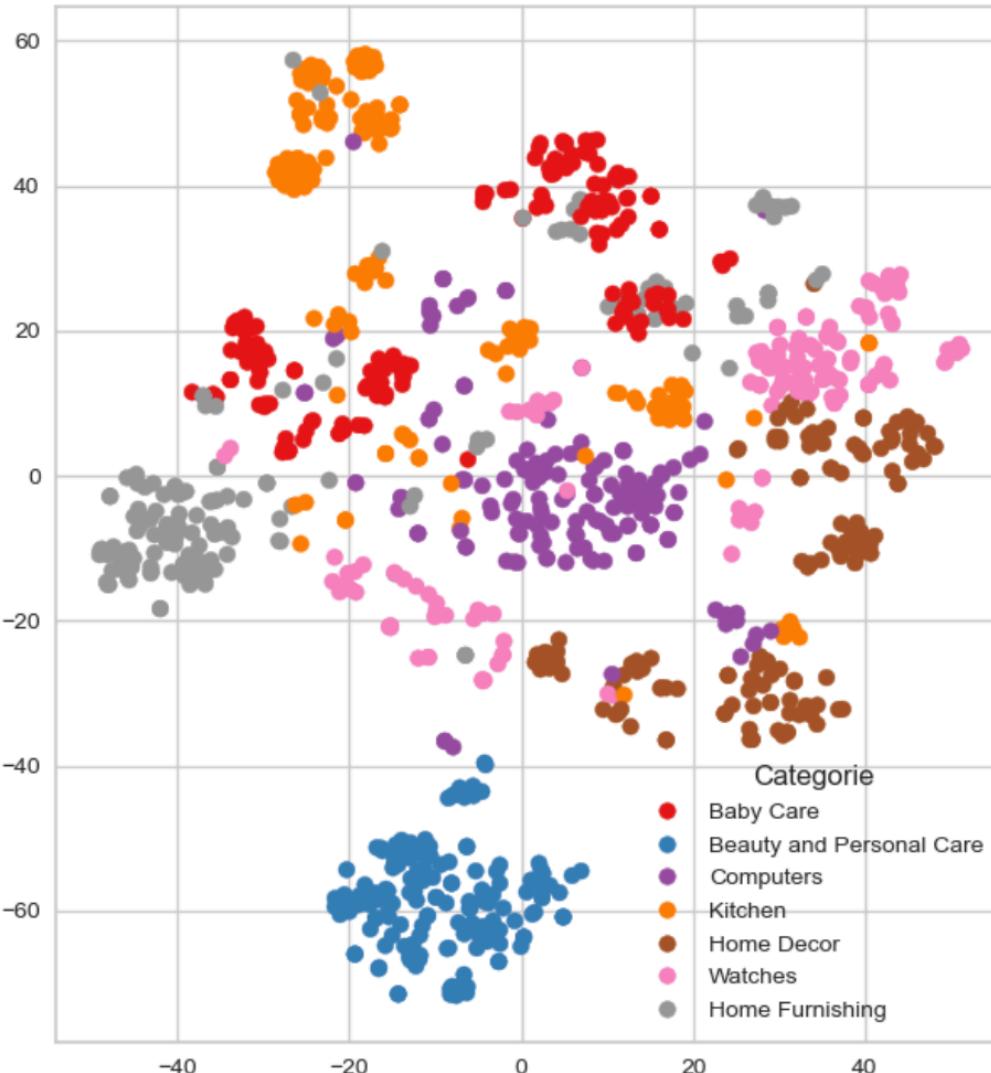
Natural Language Processing

- **TF-IDF** : pondérer cette fréquence par un indicateur **si ce mot est commun ou rare** dans tous les documents, ça permet de **réduire** l'importance des termes **présents** partout et faire ressortir les termes discriminants.

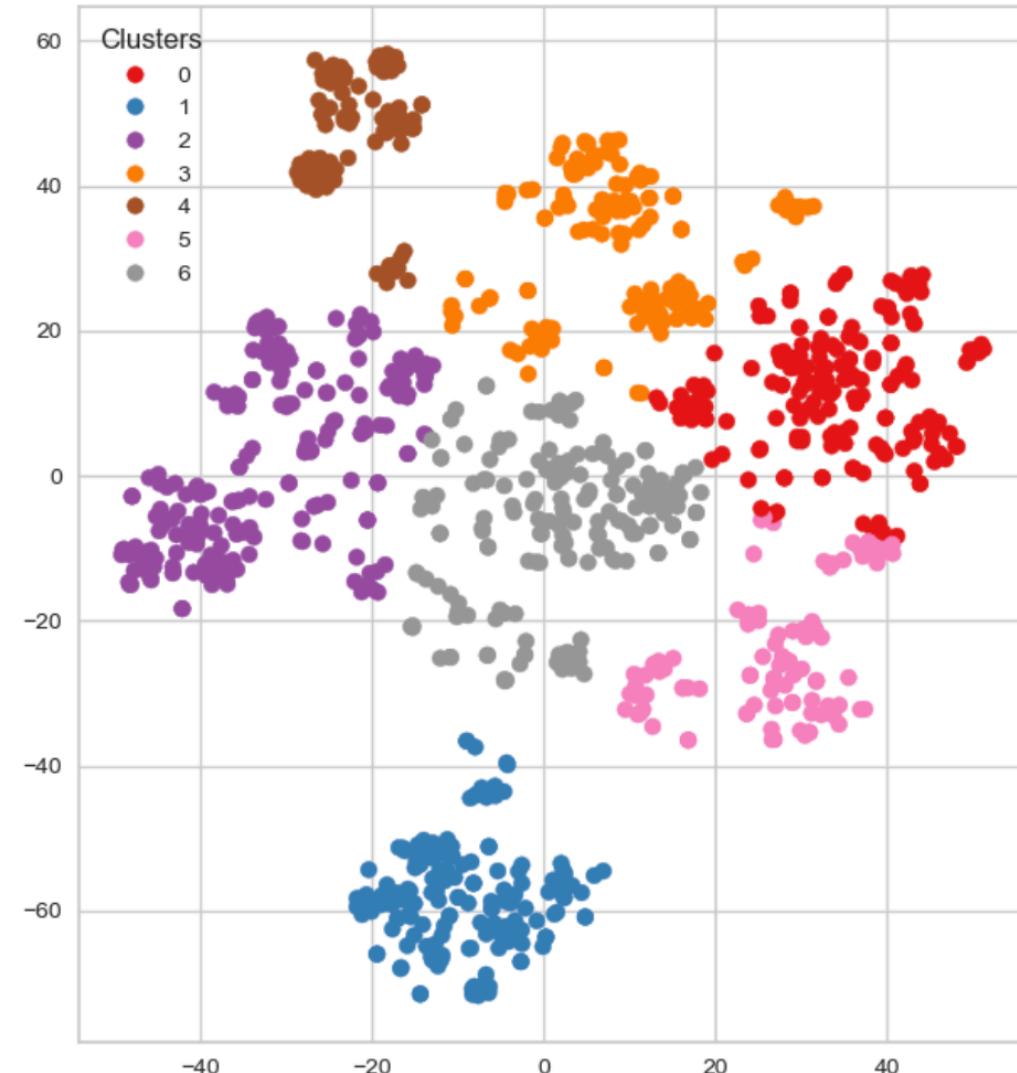
$w_{t,d} = TF_{t,d} \times \log \frac{N}{n_t}$	bien	crêpes	fais	je	NLP	oui	python	va
Term frequency ↘ proportionnel à la fréquence du terme dans le document	0	0	0.33	0.33	0	0	0.33	0
Inverse document frequency inversement proportionnel à la fréquence d'un terme dans le corpus	1	0	0	0	0.5	1	0.66	1
Je fais du python	0	0	0.33	0.33	0.5	0	0.33	0
Python ? Oui, Python va bien avec le NLP	0	1	0.33	0.33	0	0	0	0
Je fais du NLP avec python	0	0	0.33	0.33	0.5	0	0.33	0
Je fais des crêpes	0	0	0.33	0.33	0	0	0	0

Natural Language Processing

Représentation des catégories réelles



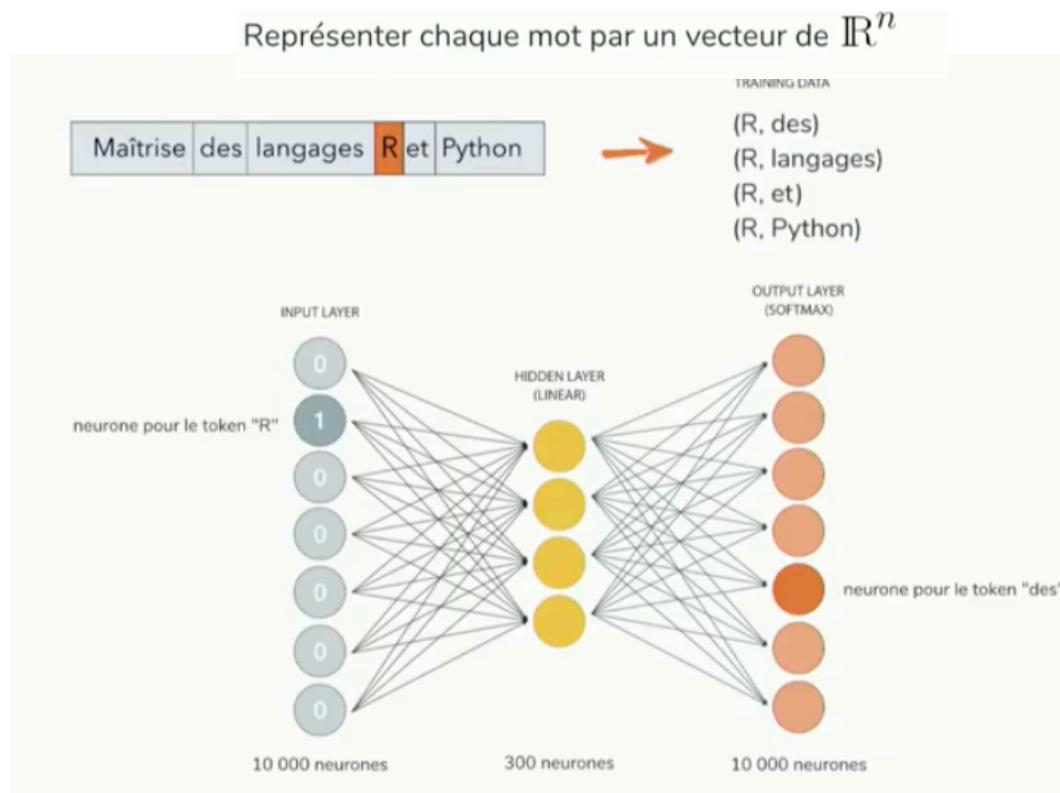
Représentation des catégories par clusters



ARI : 0.4418

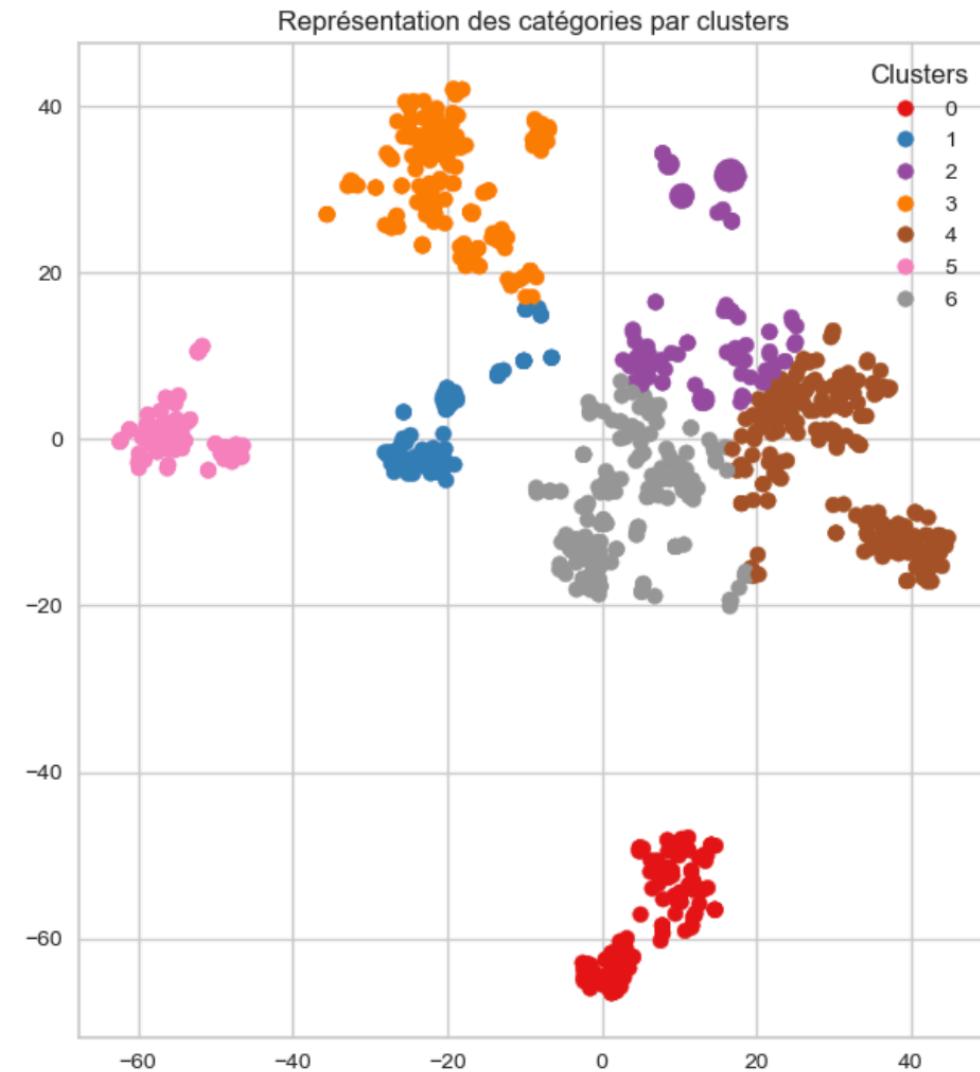
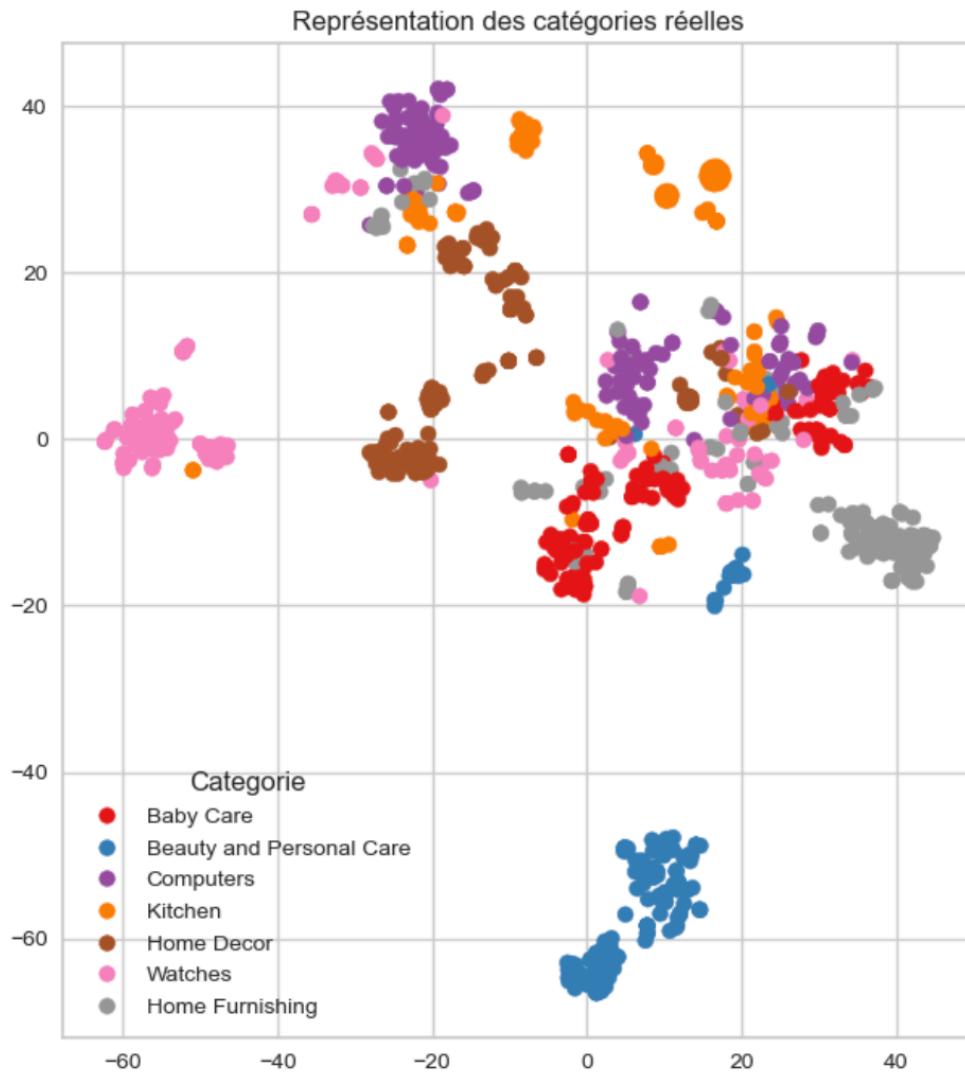
Natural Language Processing

- **Word2Vec**: C'est un algorithme de **word embedding**, il a été développé par une équipe de recherche de Google. Il repose sur des réseaux de neurones qui cherchent à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches.



- Créer un couple d'entraînement(entrée et sortie)
- Prédire les sorties à partir des entrées
- C'est un réseau de neurone à 1 couche d'entrée, 1 couche intermédiaire et 1 de sortie

Natural Language Processing



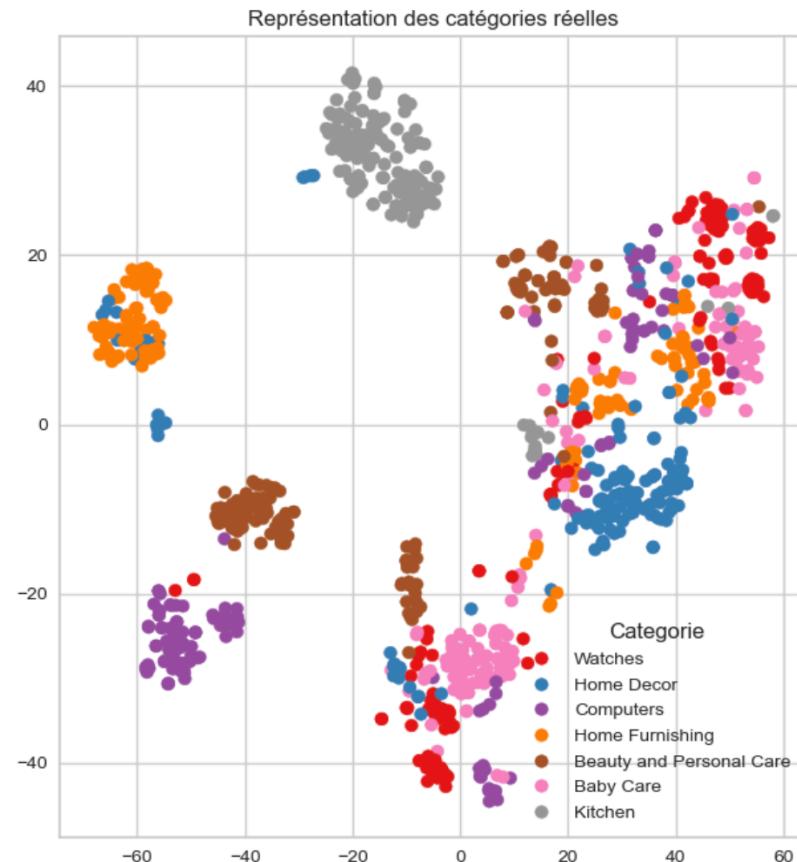
Natural Language Processing

- **BERT** : est un modèle d'apprentissage en profondeur. Il signifie représentations d'encodeur bidirectionnel pour transformateurs. Il a été préformé sur Wikipedia et BooksCorpus et nécessite (uniquement) un réglage fin spécifique à la tâche. IL utilise une nouvelle technique appelée Masked LM (MLM) : il masque aléatoirement des mots dans la phrase, puis il essaie de les prédire.
- Le masquage signifie que le modèle regarde dans les deux sens et qu'il utilise le contexte complet

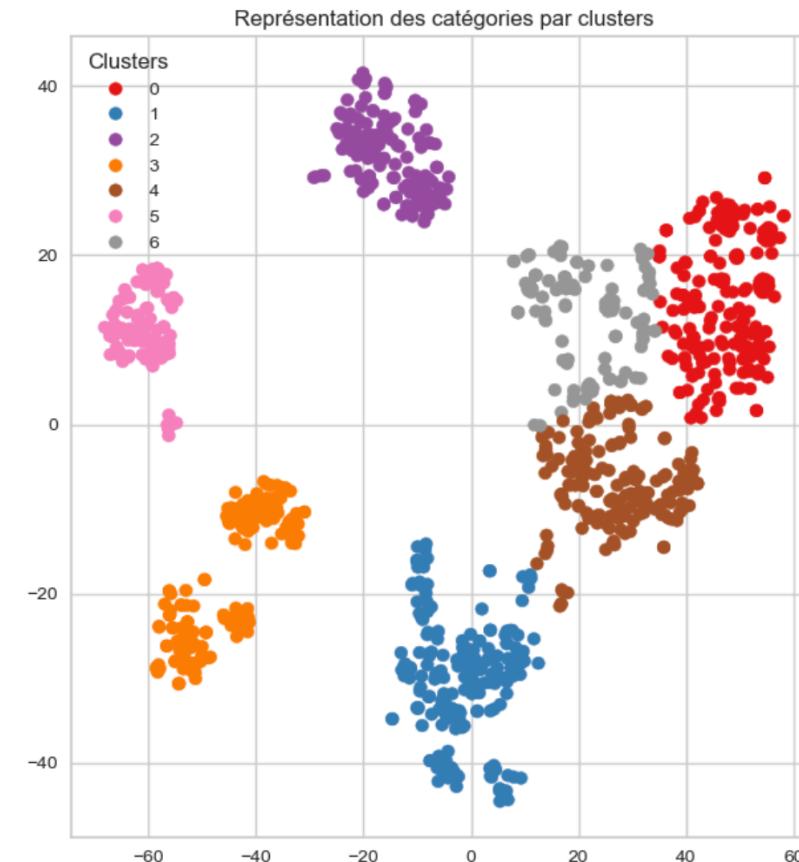


Natural Language Processing :

- **Bert-base-uncased :**

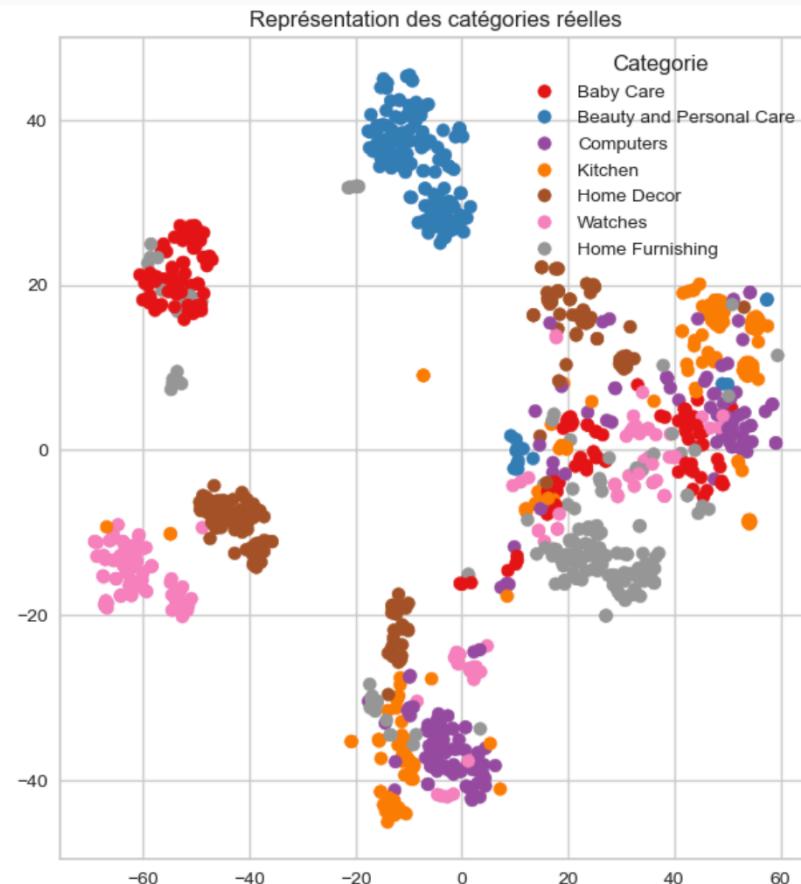


ARI : 0.3304

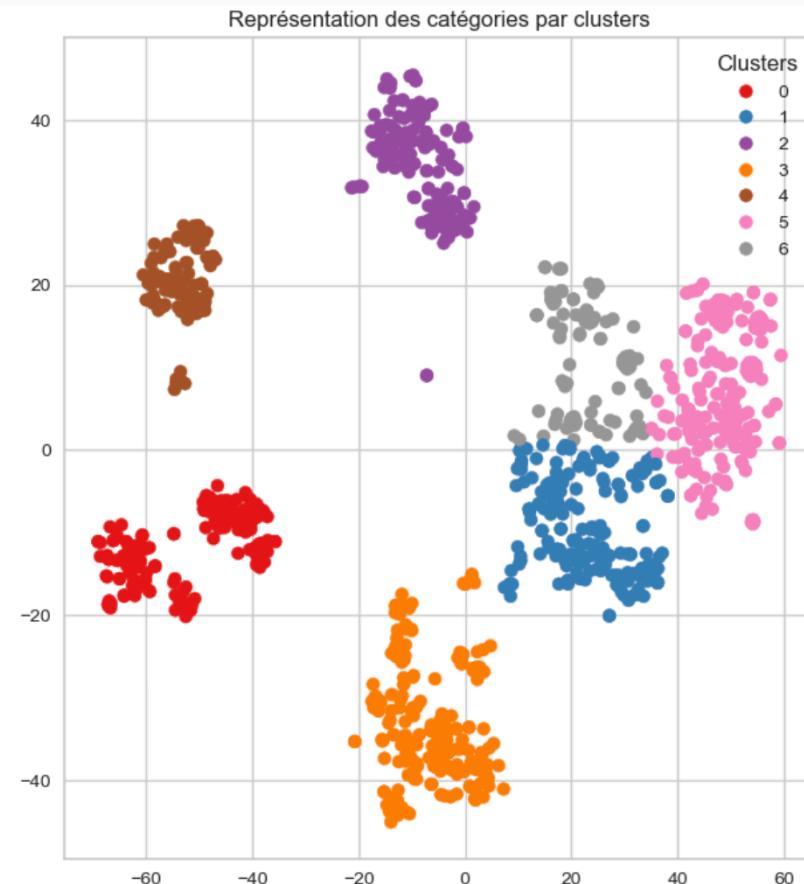


Natural Language Processing :

- BERT hub Tensorflow :

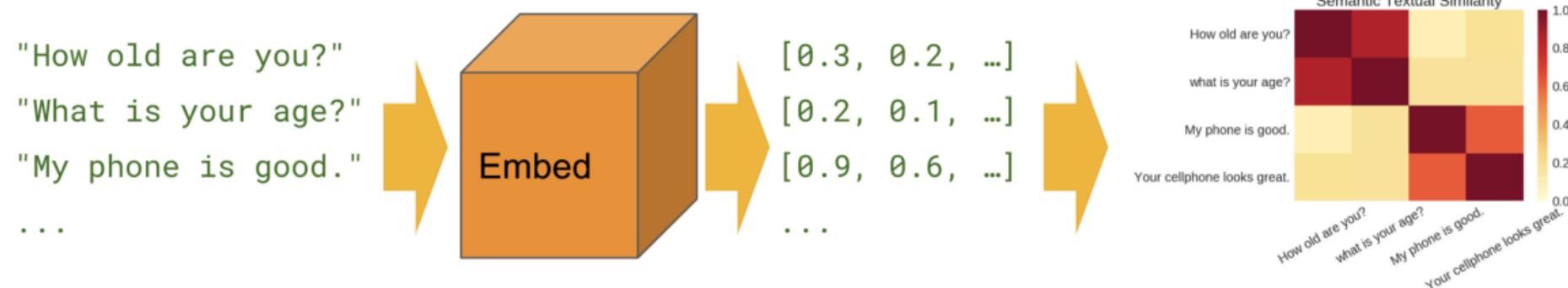


ARI : 0.3221

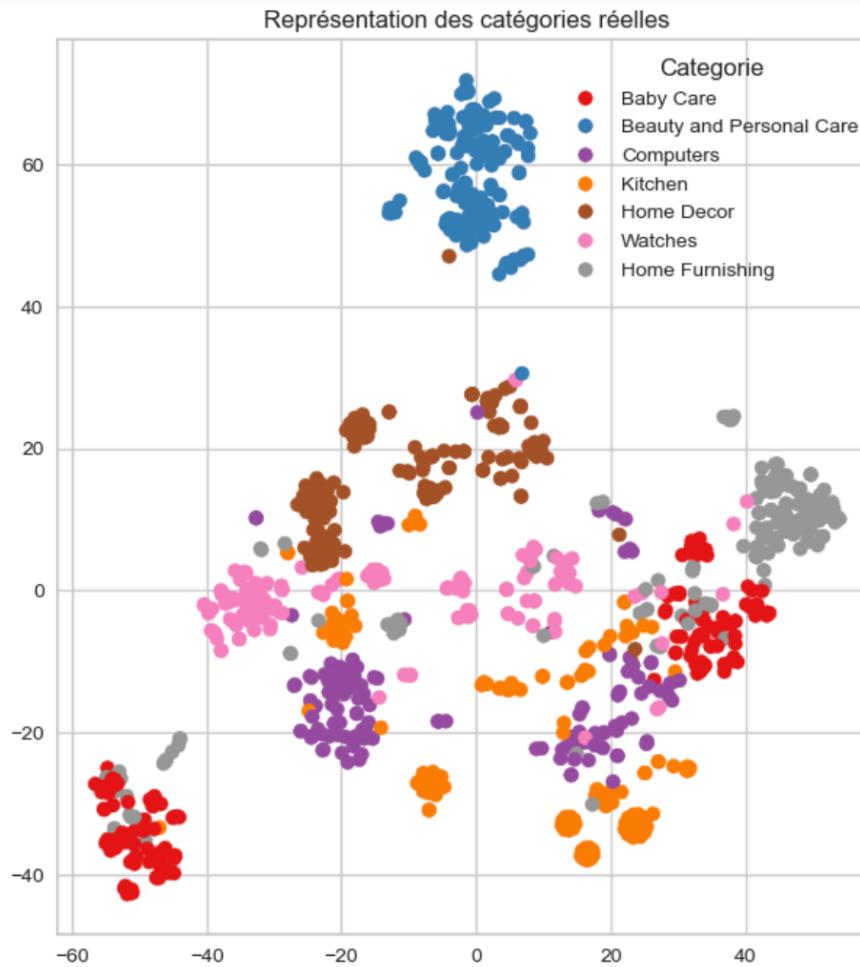


Natural Language Processing :

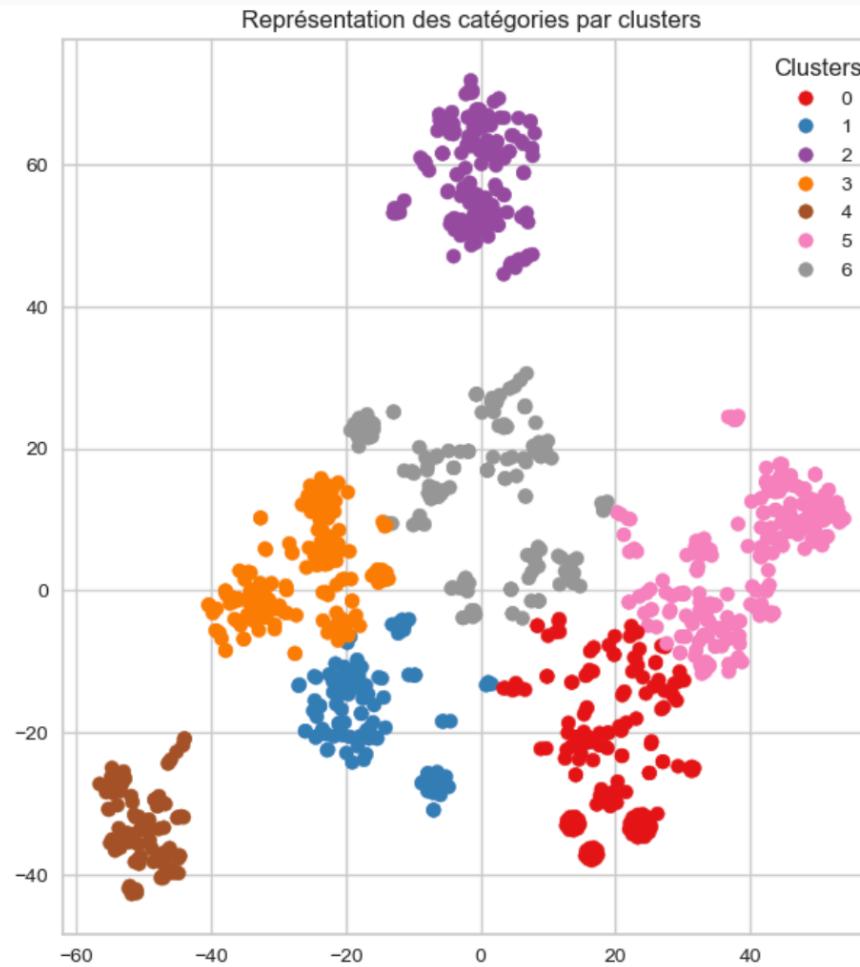
- **USE - Universal Sentence Encoder** : encode le texte dans des vecteurs de grande dimension qui peuvent être utilisés pour la classification de texte, la similarité sémantique, le regroupement et d'autres tâches en langage naturel. L'encodeur de phrase universel pré entraîné est disponible publiquement dans Tensorflow-hub.
- C'est un modèle capable de traiter du texte en 16 langues et de produire des intégrations adaptées aux tâches de similarité sémantique du texte.



Natural Language Processing :



ARI : 0.4401



Partie Image

watches



Home Furnishing



Computers



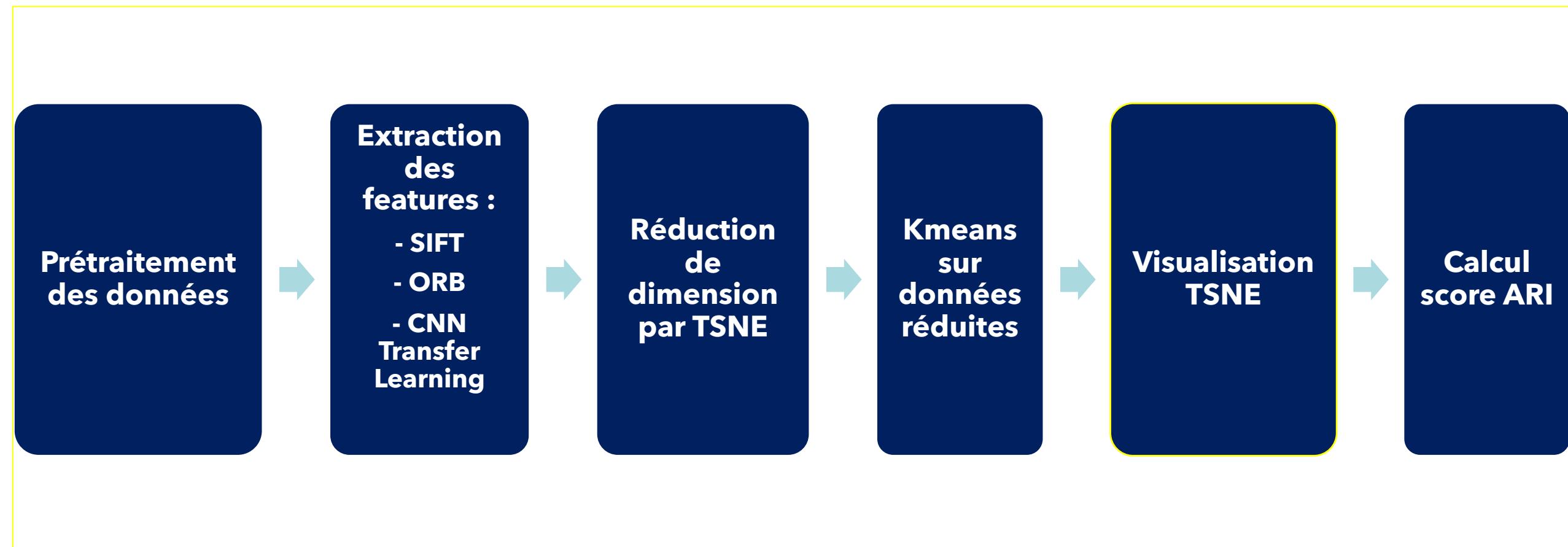
Baby Care



Home Furnishing



Démarche à suivre: Image



Prétraitement des données image



Image Dimension : (1120, 1431, 3)
Image Height : 1120
Image Width : 1431
Number of Channels : 3

- La taille des images est exprimée par **pixels**.

Prétraitement des images :

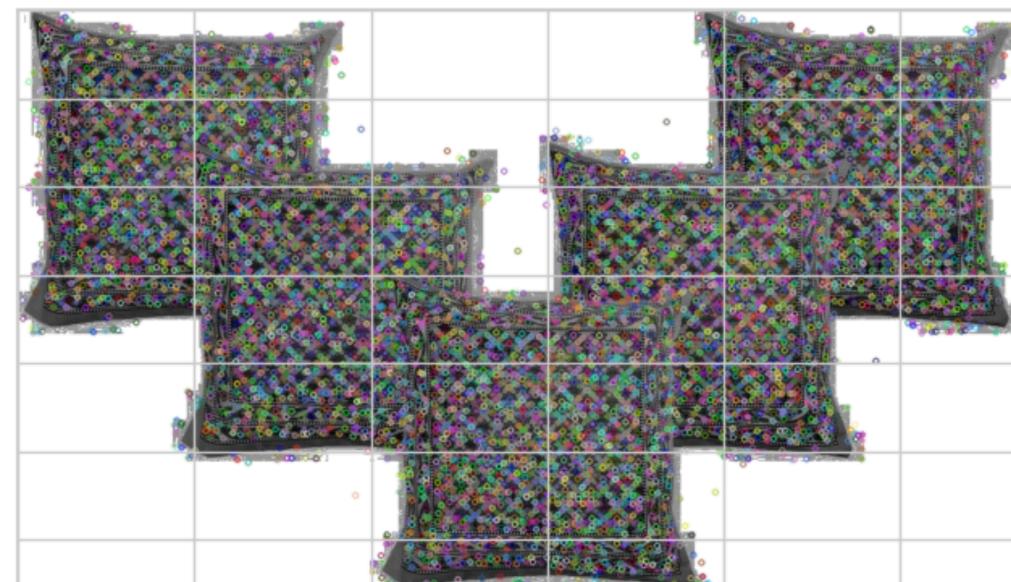
- Noir et blanc
- Redimensionnement de l'image
- L'égalisation d'histogrammes est une technique permettant de réajuster le contraste d'une image

Démarche générale pour l'algorithme SIFT et ORB



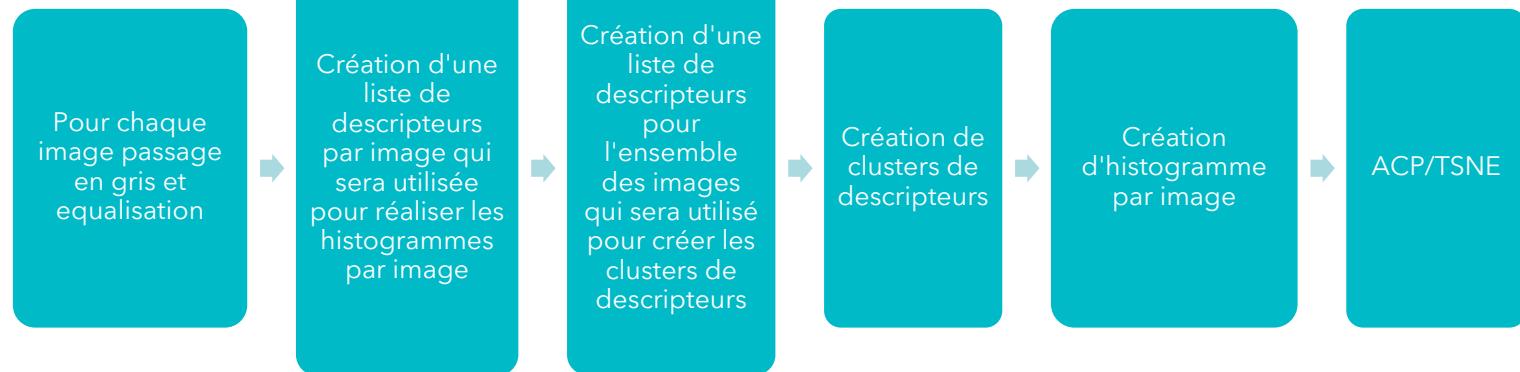
Données visuelles

- **SIFT** : l'algorithme SIFT (Scale-invariant feature transform). Cette méthode, permet d'extraire des features (ou points d'intérêt) de l'image et de calculer leurs descripteurs.



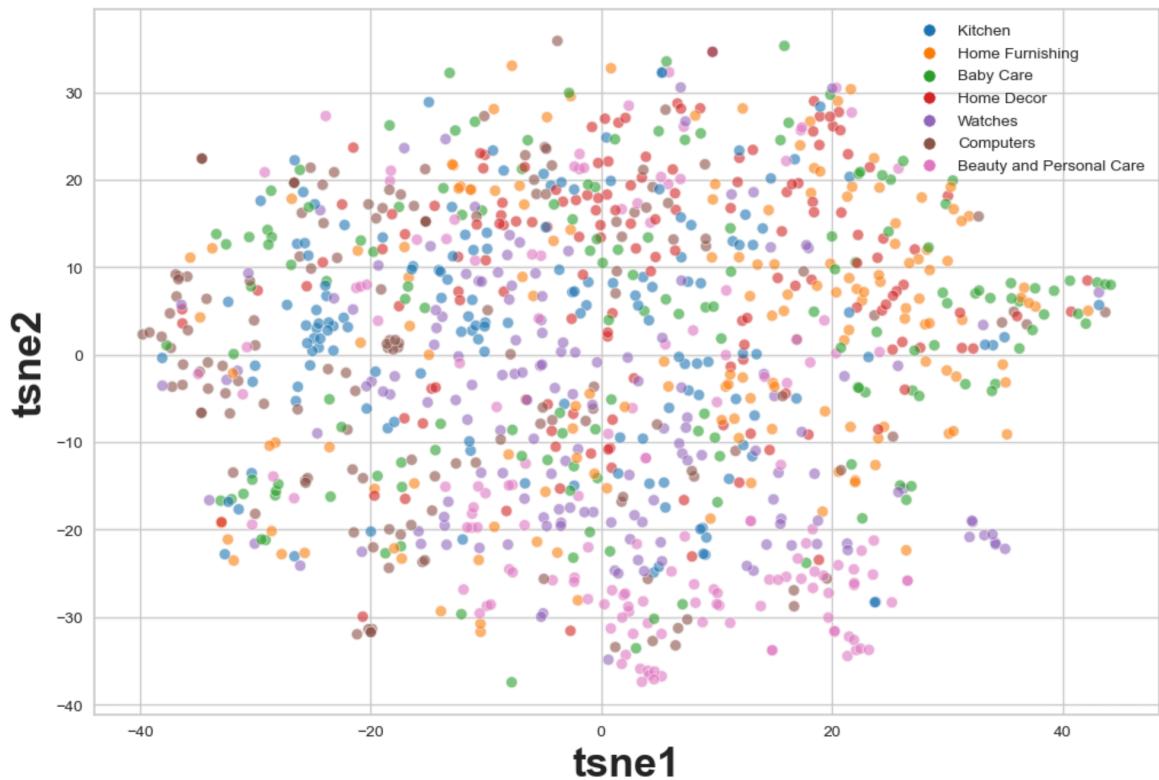
Descripteurs : (11002, 128)

Création des **descripteurs** de chaque image :

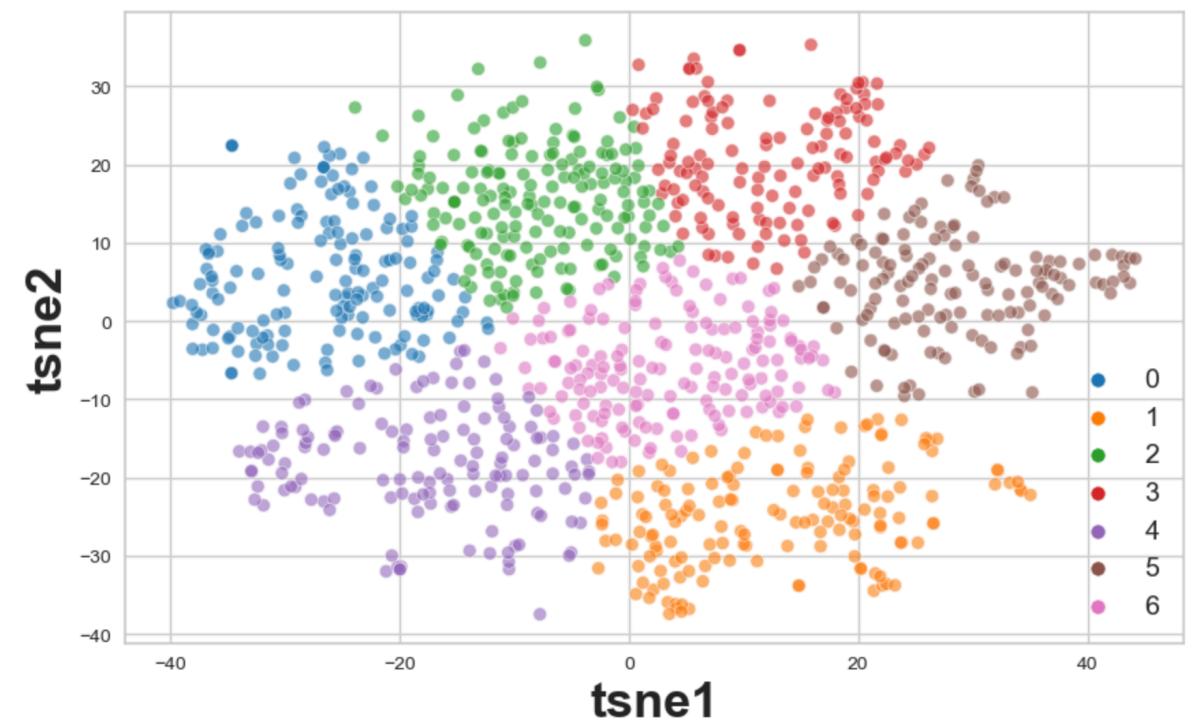


Données visuelles

TSNE selon les vraies classes



TSNE selon les clusters



ARI : 0.062

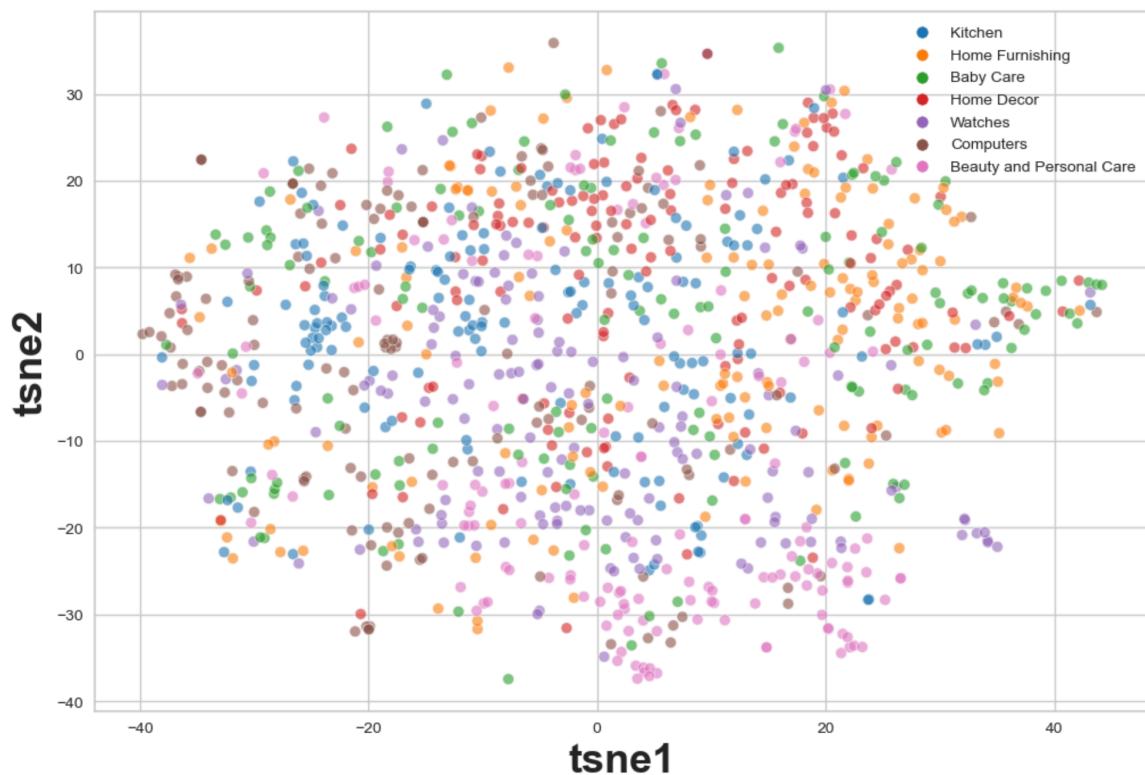
Données visuelles

- **ORB** : est une fusion du détecteur de points clés FAST et du descripteur BRIEF avec quelques fonctionnalités supplémentaires pour améliorer les performances. FAST correspond aux caractéristiques du test de segment accéléré utilisé pour détecter les caractéristiques de l'image fournie. Il utilise également une pyramide pour produire des caractéristiques multi-échelles.

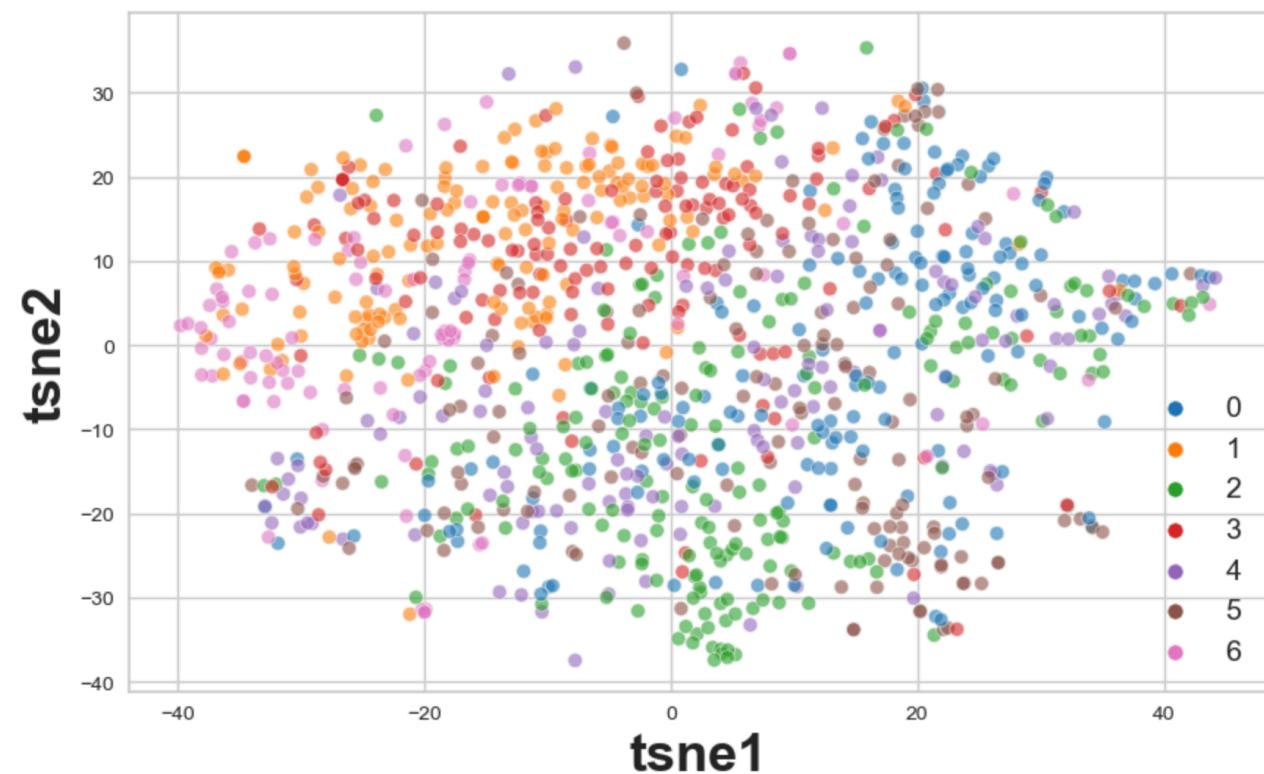


Données visuelles

TSNE selon les vraies classes



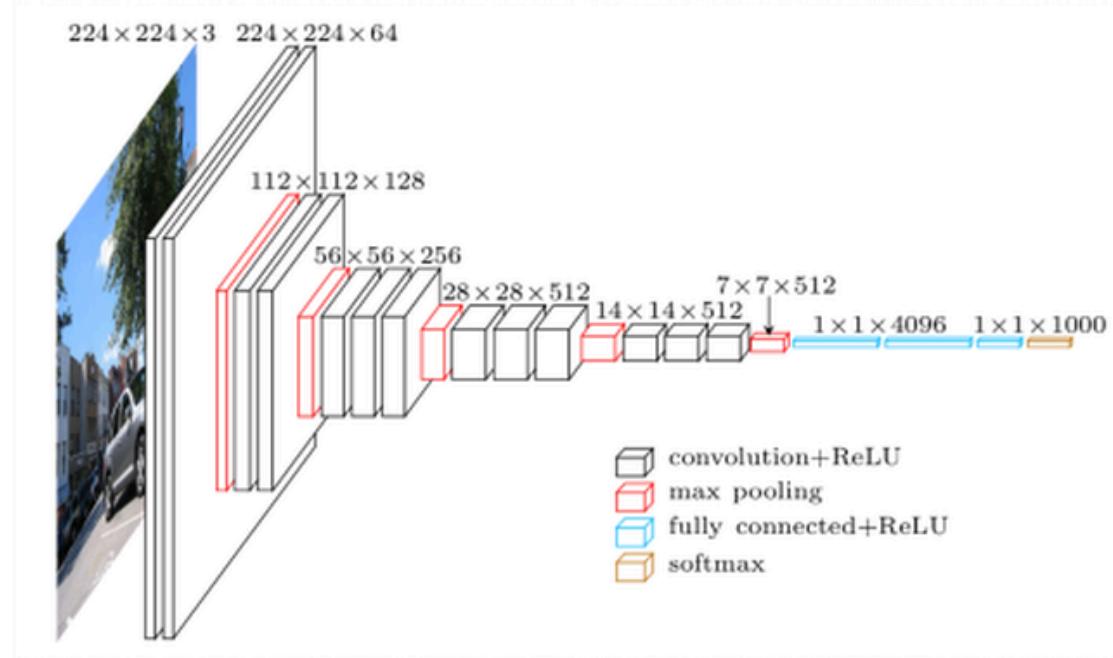
TSNE selon les clusters



ARI : 0.031

Données visuelles

- CNN Transfer Learning : Les CNN sont des réseaux de neurones spécialement conçus pour traiter des images en entrée
- **VGG-16:** une version du réseau de neurones convolutif



Représentation 3D de l'architecture de VGG-16

13 couches de convolution et 3 *fully-connected*

Données visuelles

- Utilisation du VGG16 en tant que Pre-Trained Model as Feature Extractor Preprocessor

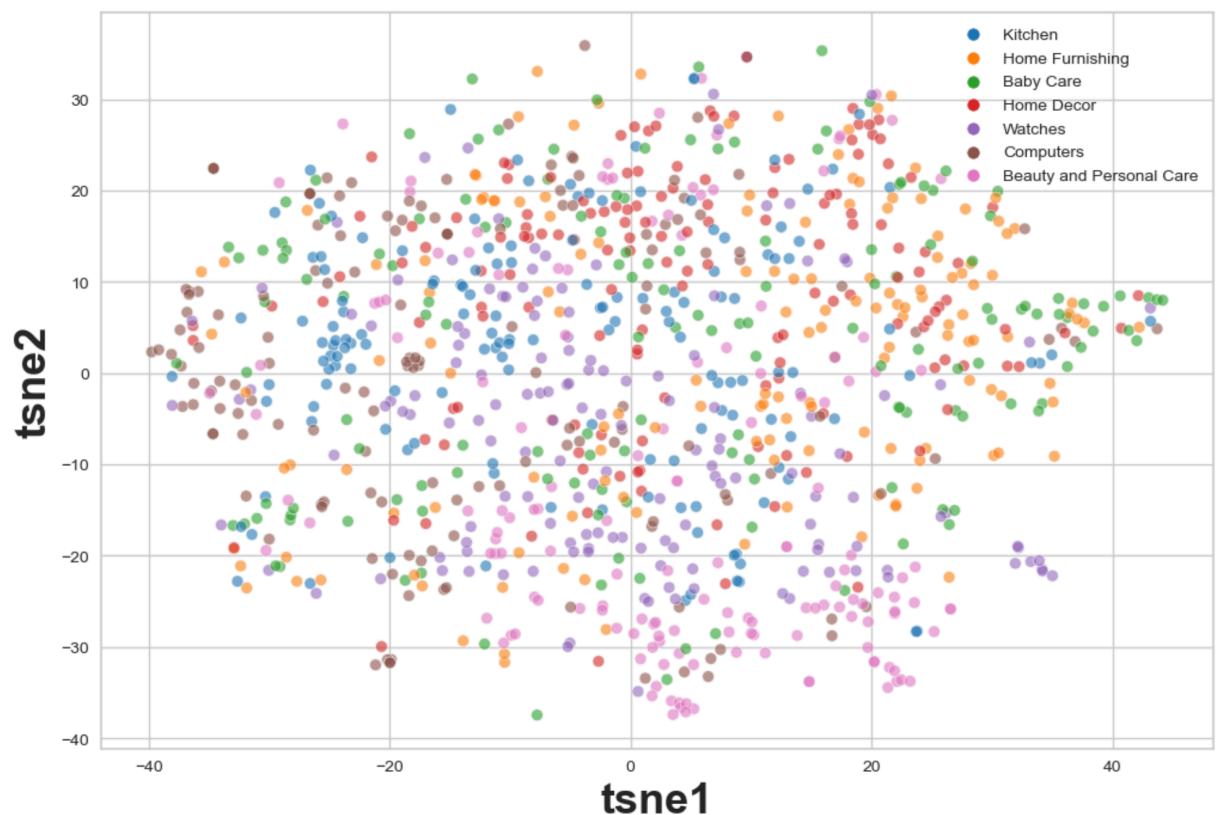
```
# remove the output layer
model = Model(inputs=model.inputs, outputs=model.layers[-2].output)

def extract_features(file, model):
    # load the image as a 224x224 array and convert in gray

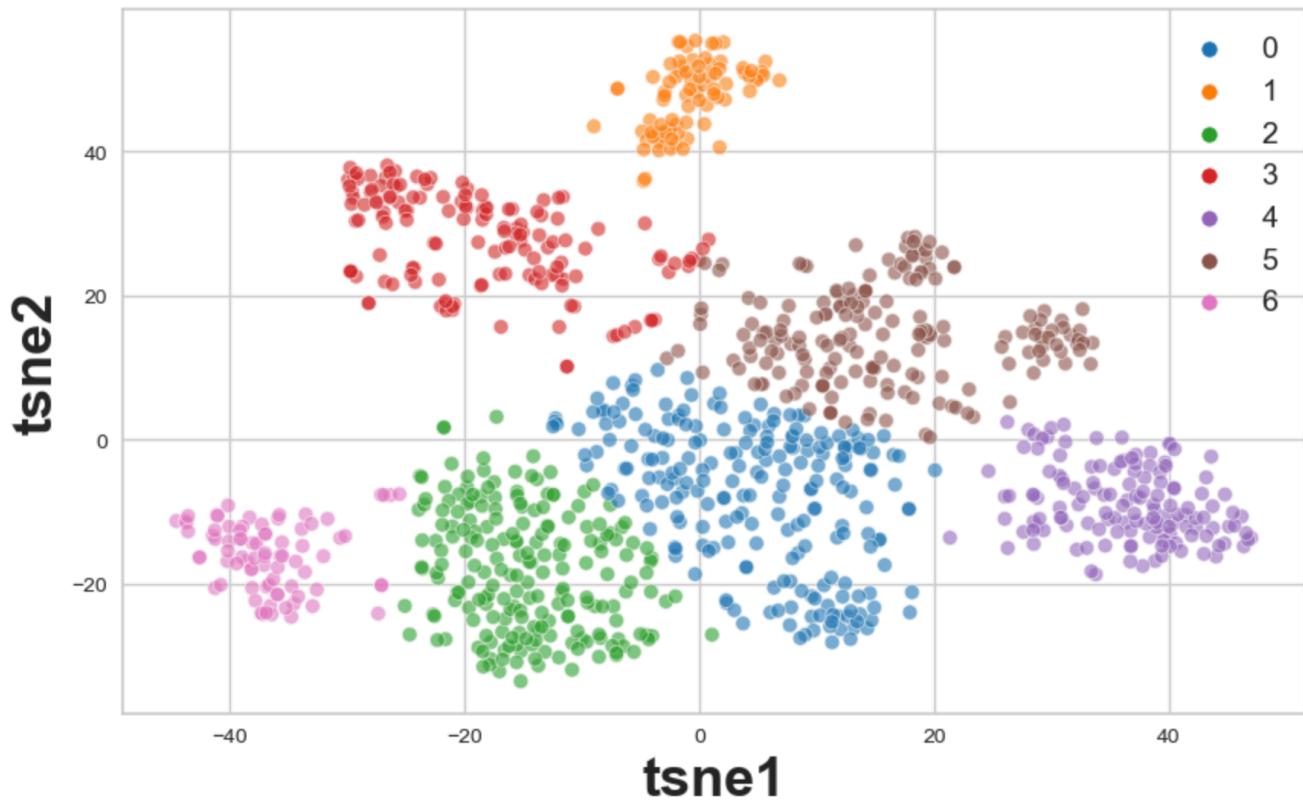
    img = load_img(path + list_photos[image_cnn], target_size=(224, 224))
    # convert the image pixels to a numpy array
    img = np.array(img)
    # reshape the data for the model reshape(num_of_samples, dim 1, dim 2, channels)
    reshaped_img = img.reshape((1, img.shape[0], img.shape[1], img.shape[2]))
    # prepare image for model
    imgx = preprocess_input(reshaped_img)
    # get the feature vector
    features = model.predict(imgx, use_multiprocessing=True)
    return features
```

Données visuelles

TSNE selon les vraies classes



TSNE selon les clusters



ARI : 0.474

Bilan des résultats

Texte	Image
Bag of words : 0.374	SIFT : 0.062
TF_IDF : 0.441	ORB : 0.031
WORD2VEC : 0.360	VGG16 : 0.474
BERT : 0.322	
USE : 0.440	

Combinaison des features texte et image

- Extraction de features texte avec l'algorithme **Tf-IDF** et extraction de features images par **VGG16**
- **Combinaison** des features **texte + image**
- Réduction par **ACP**
- **Kmeans** sur données réduites
- Calcul score **ARI**

```
6]: 1 | cls = cluster.KMeans(n_clusters=7, n_init=100, random_state=42)
      2 | cls.fit(df_tf_cnn2)
      3 | ARI = np.round(metrics.adjusted_rand_score(labels, cls.labels_), 4)
      4 | print("ARI : ", ARI)
```

ARI : 0.4774

Approche supervisée

- Données textuelles :

Classification SVM avec features extraites à partir du **TF-IDF**

	precision	recall	f1-score	support
Beauty and Personal Care	0.95	0.86	0.90	44
	0.94	0.96	0.95	47
	0.98	1.00	0.99	45
	0.87	0.94	0.90	35
	0.96	0.96	0.96	53
	0.98	0.98	0.98	43
	1.00	0.98	0.99	48
accuracy			0.96	315
macro avg	0.95	0.95	0.95	315
weighted avg	0.96	0.96	0.96	315

```
: 1 print(f1_score(y_test, y_pred, average="micro"))
```

0.9555555555555556



Approche supervisée

- **Données visuelles :**

```
NBCLASSES = 7

def create_model():
    vgg = VGG16(input_shape=(224, 224, 3), weights="imagenet", include_top=False)

    # Freeze existing VGG already trained weights
    # On entraîne seulement le nouveau classifieur et on ne ré-entraîne pas les autres couches :
    for layer in vgg.layers:
        layer.trainable = False

    # get the VGG output
    out = vgg.output

    # Add new dense layer at the end
    x = Flatten()(out)
    x = Dense(NBCLASSES, activation="softmax")(x)

    model = Model(inputs=vgg.input, outputs=x)

    model.compile(
        loss="SparseCategoricalCrossentropy",
        optimizer="adam",
        metrics=[ "accuracy" ],
    )

    model.summary()

    return model

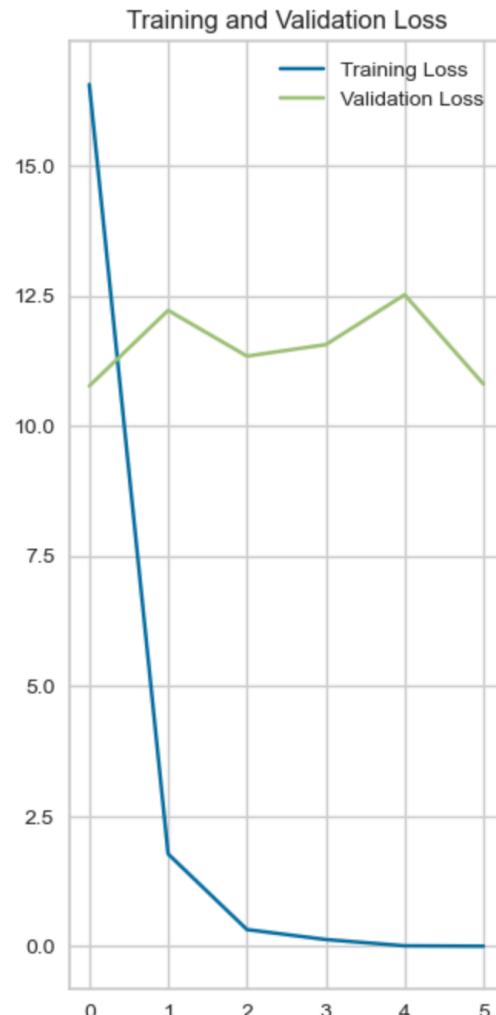
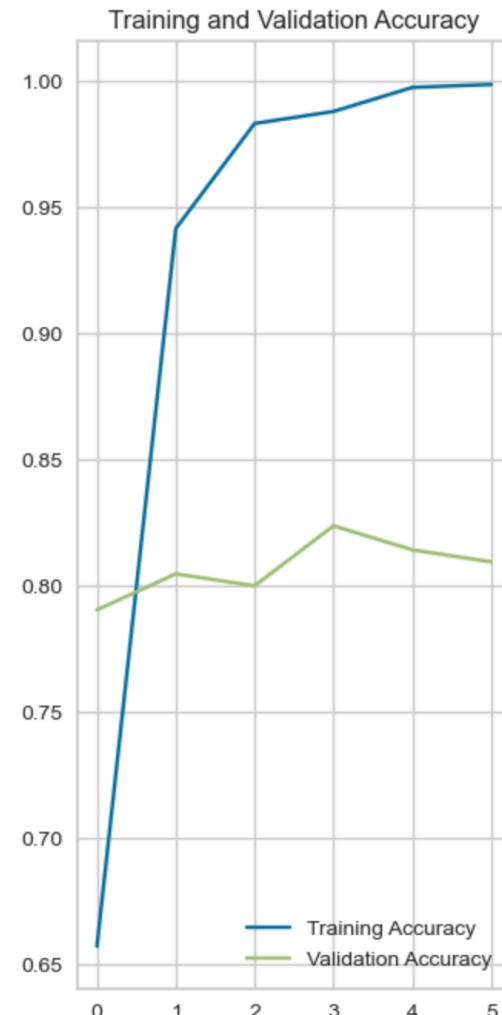
mymodel = create_model()
```

- On remplace la dernière couche *fully-connected* par le nouveau classifieur, et on fixe les paramètres de certaines couches du réseau pré-entraîné. Ainsi, en plus du classifieur, on entraîne sur les nouvelles images les couches non-fixées.

Approche supervisée

```
Epoch 1/6
27/27 [=====] - 75s 3s/step - loss: 16.5653 - accuracy: 0.6571 - val_loss: 10.7715 - val_accuracy: 0.7905
Epoch 2/6
27/27 [=====] - 87s 3s/step - loss: 1.7778 - accuracy: 0.9417 - val_loss: 12.2238 - val_accuracy: 0.8048
Epoch 3/6
27/27 [=====] - 88s 3s/step - loss: 0.3247 - accuracy: 0.9833 - val_loss: 11.3444 - val_accuracy: 0.8000
Epoch 4/6
27/27 [=====] - 89s 3s/step - loss: 0.1335 - accuracy: 0.9881 - val_loss: 11.5655 - val_accuracy: 0.8238
Epoch 5/6
27/27 [=====] - 89s 3s/step - loss: 0.0135 - accuracy: 0.9976 - val_loss: 12.5248 - val_accuracy: 0.8143
Epoch 6/6
27/27 [=====] - 89s 3s/step - loss: 0.0061 - accuracy: 0.9988 - val_loss: 10.8074 - val_accuracy: 0.8095
```

Approche supervisée



Conclusion

- Cette première étude de faisabilité d'un moteur de classification n'a pas été concluante vu les résultats des scores ARI.
- La faisabilité de la classification automatique peut être amenée à changer et donner de meilleurs résultats si la qualité des images changeait .
- Une approche supervisée peut-être envisagée en vue des résultats satisfaisants obtenus.



categorie: Kitchen

Buy Prithish Friend Indeed Ceramic Mug for Rs.225 online. Prithish Friend Indeed Ceramic Mug at best prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee....

categorie: Computers

Buy Airtel B310s-927 only for Rs. 2700 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!...

categorie: Home Decor

Tatvaarts Tribal Face Showpiece - 21.59 cm (Brass, Gold)

Price: Rs. 6,100

Discussion :



FIN

• MERCI