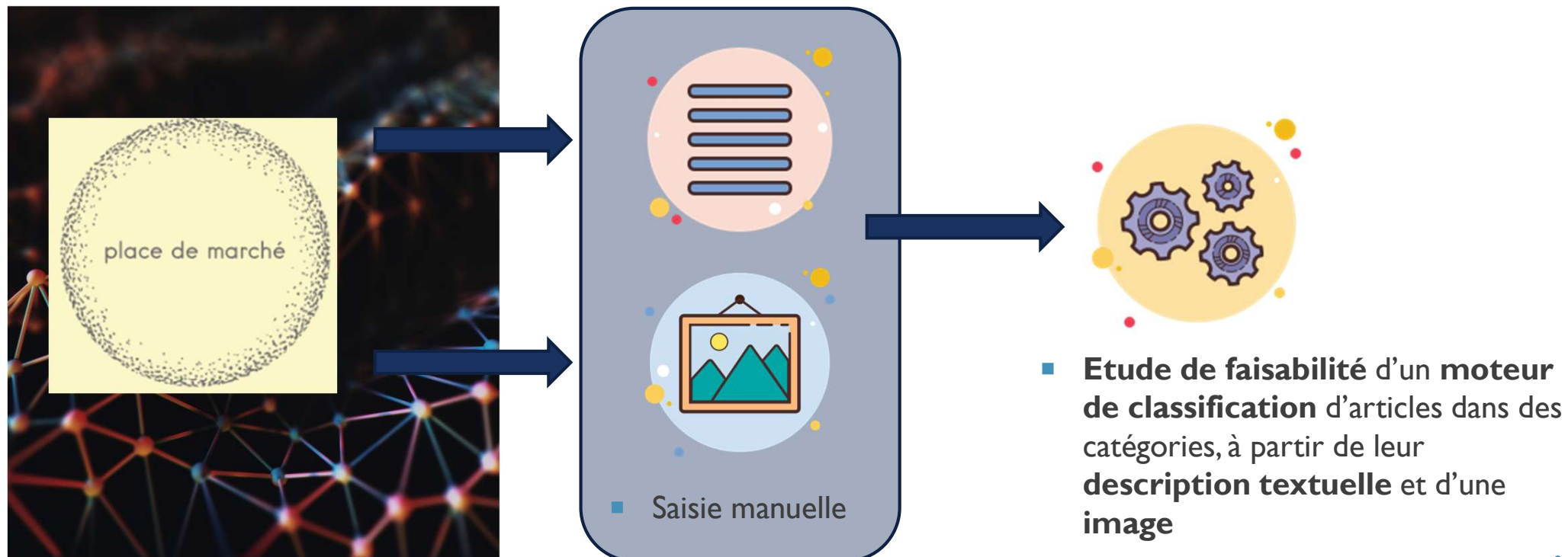




PROJET 6 : CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

LAURENT CAGNIART

PROBLÉMATIQUE



- Extrait de **1050 articles** référencés

- **7 catégories** de niveau 1 réparties équitablement

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_l1, dtype: int64



- **2 features principales :**
« product name » et
« description »

```
TAG Heuer CAU1116.BA0858 Formula 1 Analog Watc...
Scalabedding Cotton Striped King sized Double ...
Arabian Nights Soex Cranberry Assorted Hookah ...
Maxima 17321CMLY Gold Analog Watch - For Women
King Traders KI-BD-01 1 Kitchen Tool Set
Playboy Berlin Combo Set
Rastogi Handicrafts Showpiece - 20 cm
Piyo Piyo Four Stage Waterproof Bib
Fastrack 9913PP03 Tees Analog Watch - For Women
Timewel 1100-N1949_B Analog Watch - For Women
```

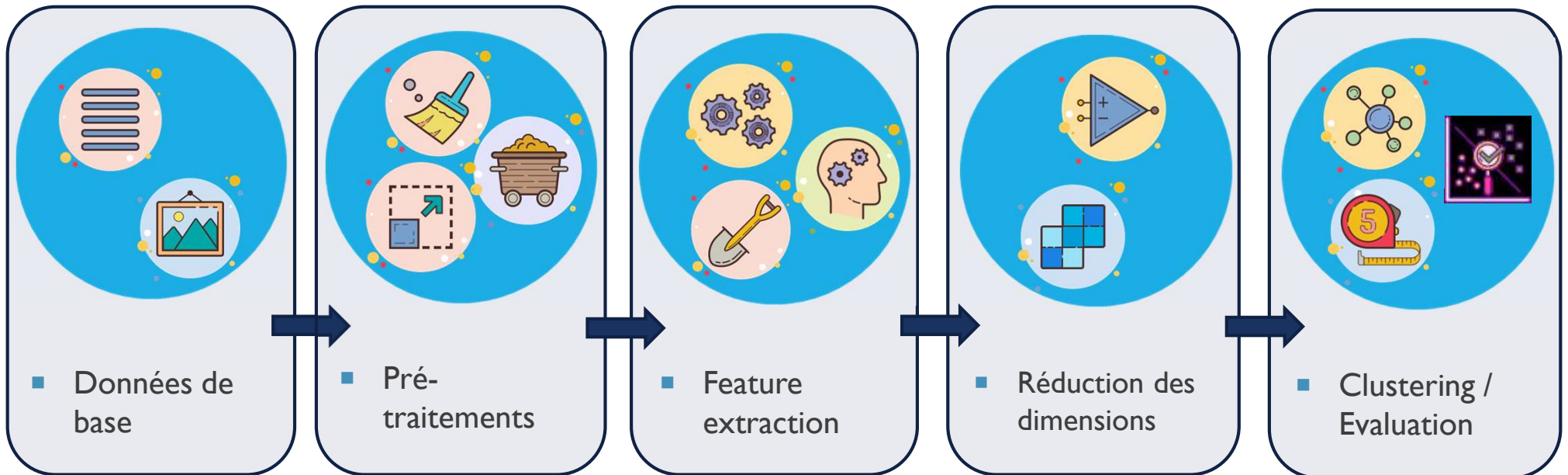


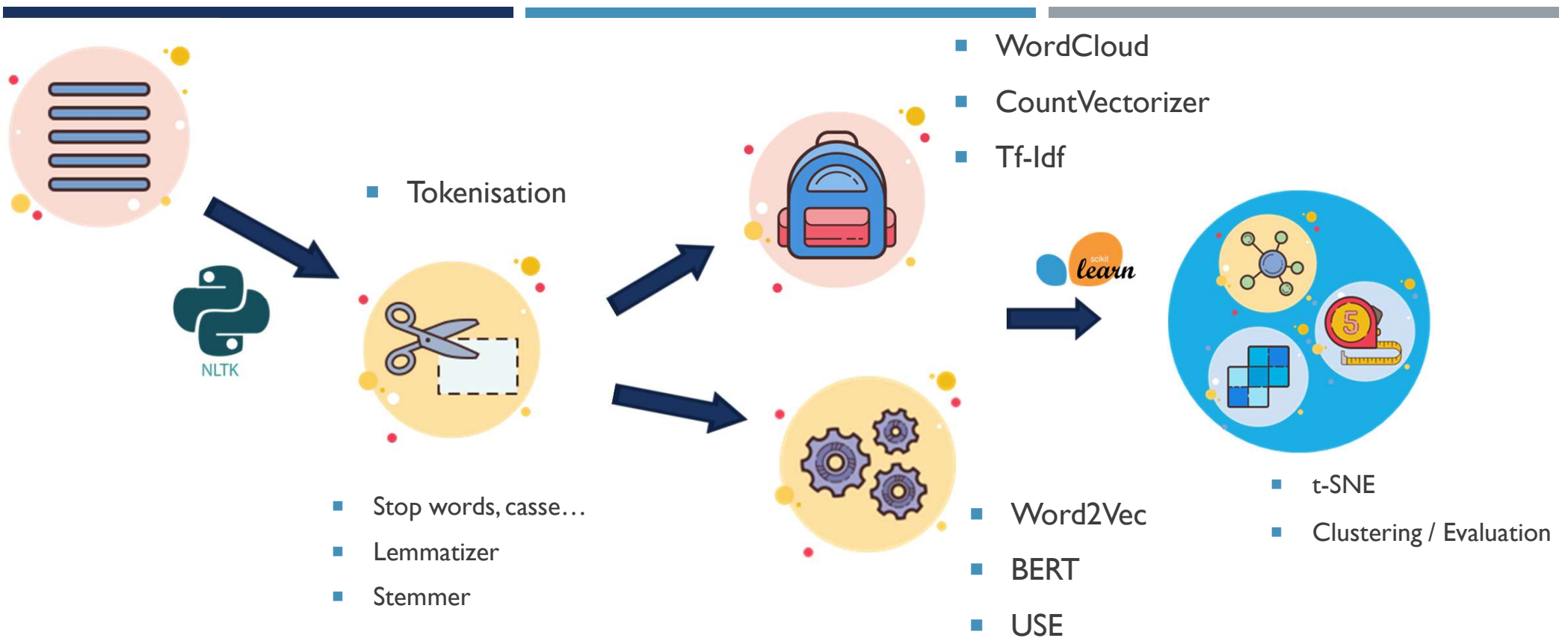
- **Format jpg**
- **Taille variable**



PRÉSENTATION DES DONNÉES

MÉTHODOLOGIE

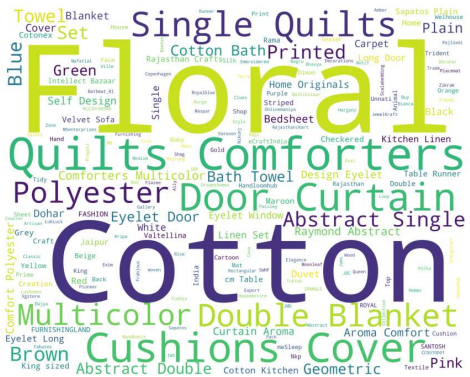




NATURAL LANGUAGE PROCESSING - NLP

WORDCLOUD + COUNTVECTORIZER
UN VOCABULAIRE ASSEZ SPÉCIFIQUE PAR CATÉGORIE

- Home Furniture



- Kitchen & Dining



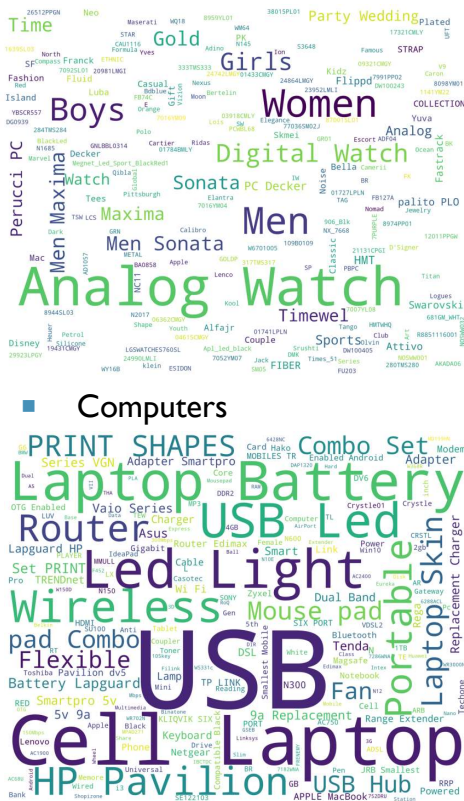
- Baby Care



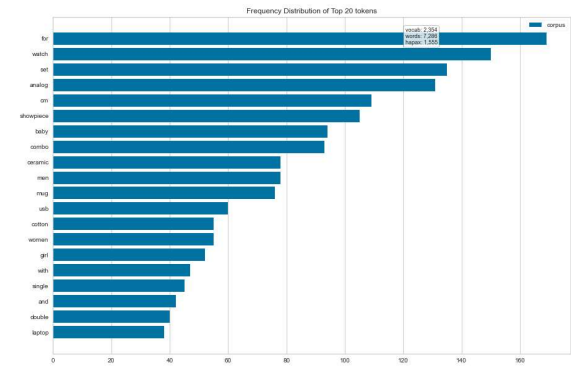
- Beauty and Personal Care



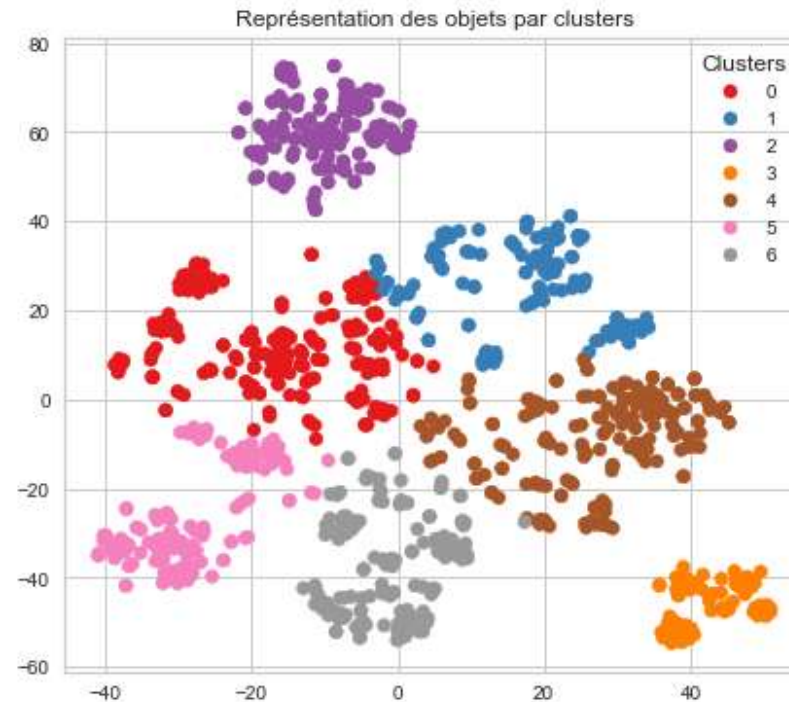
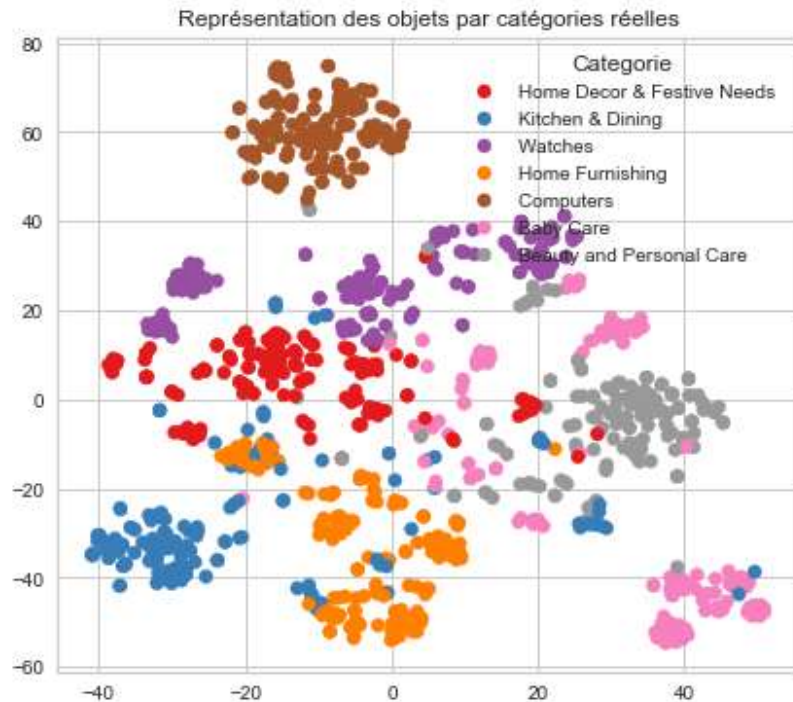
- Watch



- Home Decor & Festive Needs



■ T-SNE et clustering



- Paramètres :
- Max_df = 0,9
- Min_df = 2
- Feature : « name + desc »



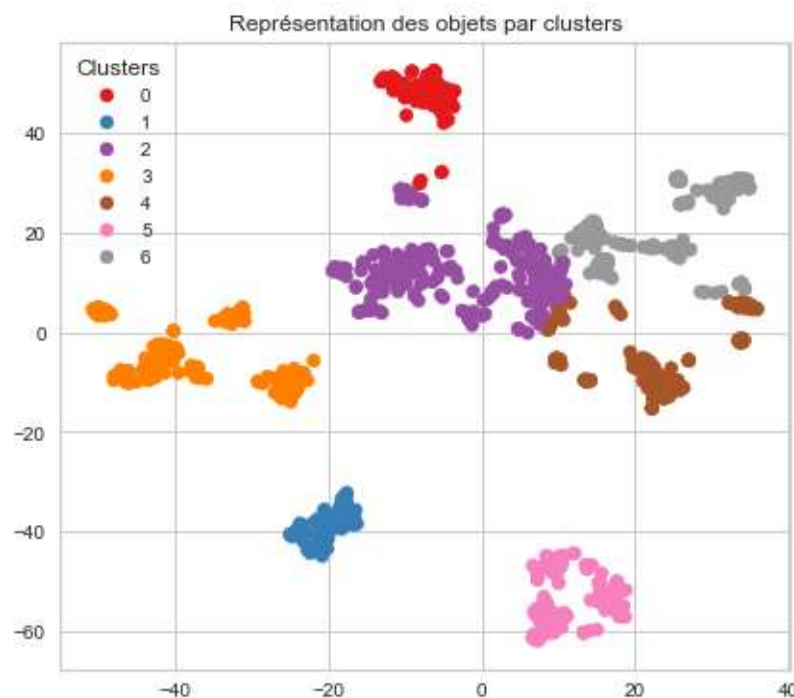
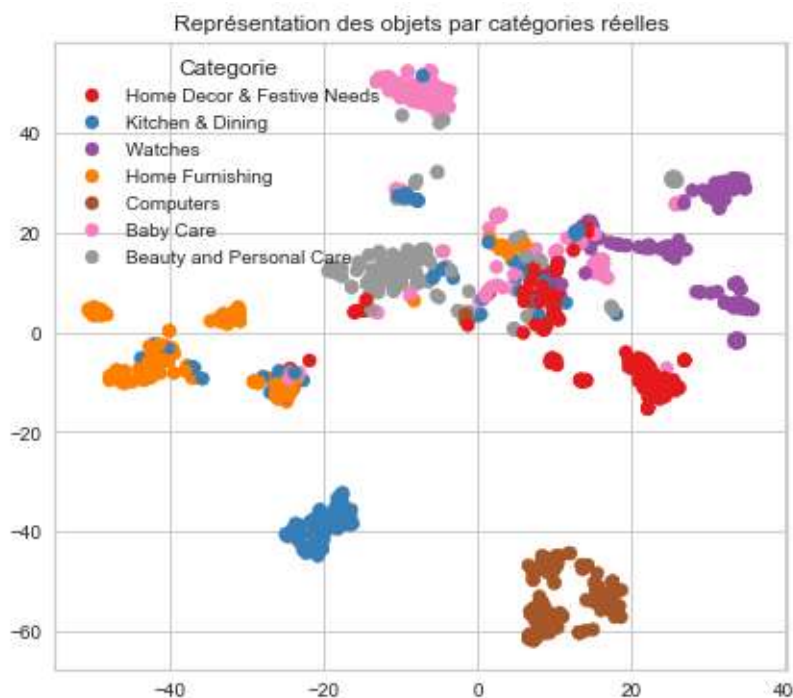
■ ARI =
0,5637

TF-IDF :
DES RÉSULTATS SATISFAISANTS SERVANT DE BASE DE COMPARAISON

WORD2VEC

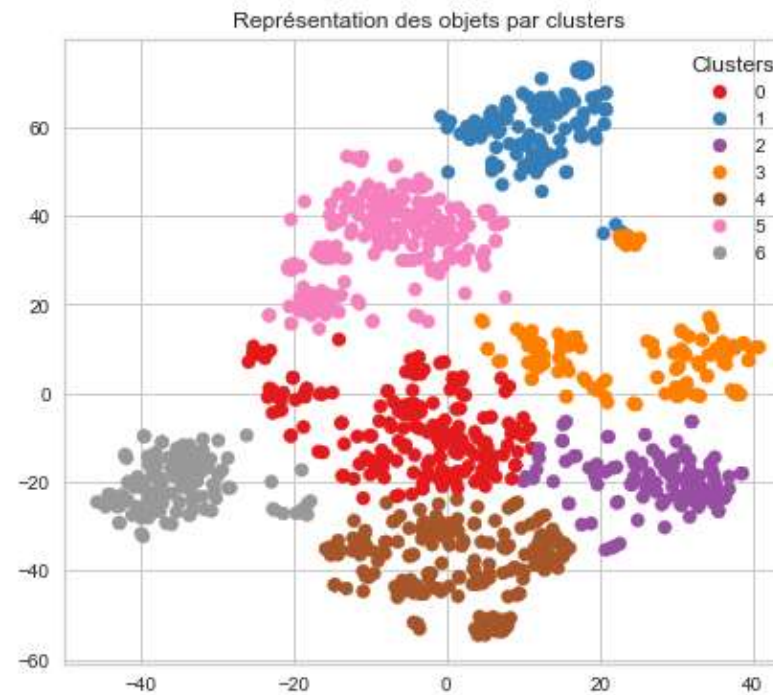
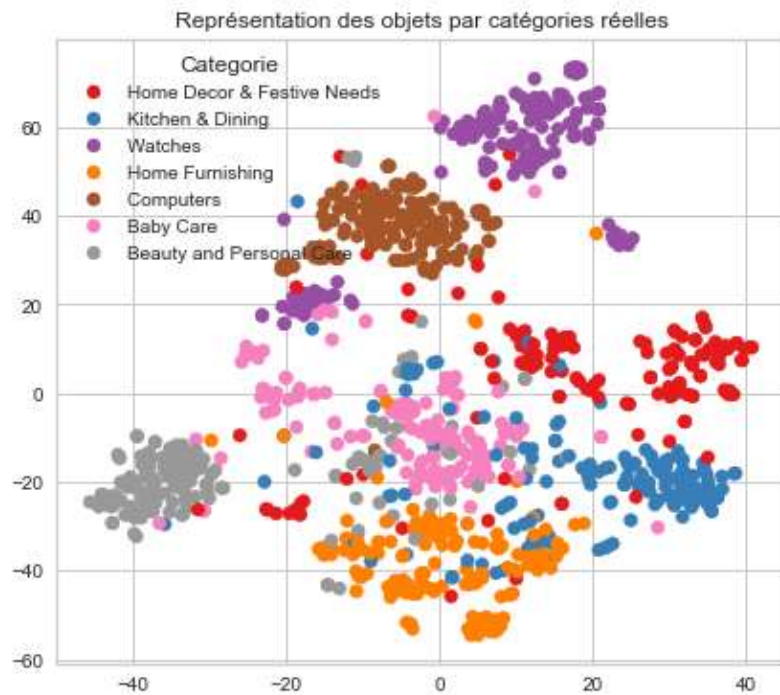
UN 1^{ER} MODELE DE WORD EMBEDDING MOINS PERFORMANT

■ T-SNE et clustering



■ ARI =
0,5002

■ T-SNE et clustering



- Paramètres : 🤗
- Max_length = 30
 - Batch_size = 10
 - Feature : « name »

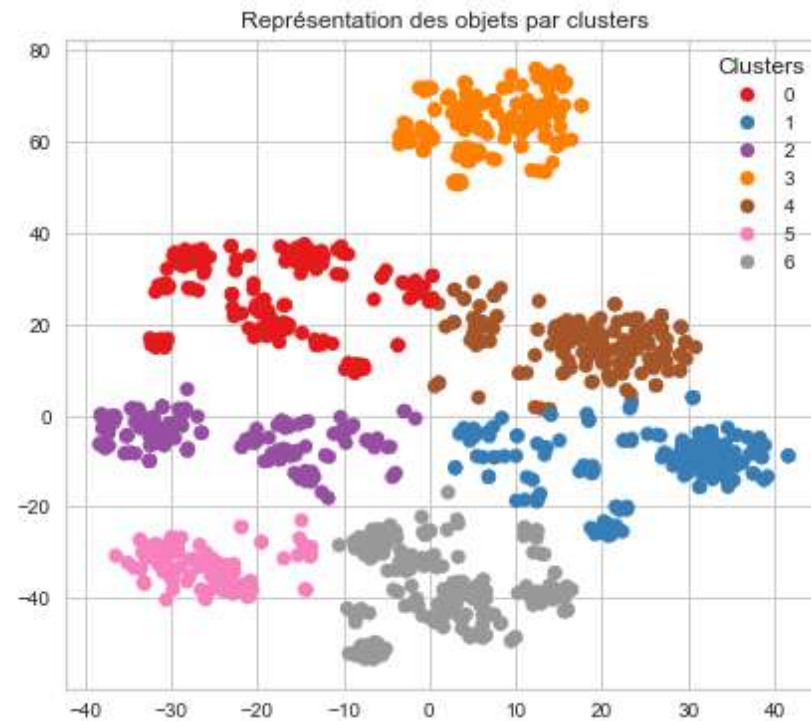
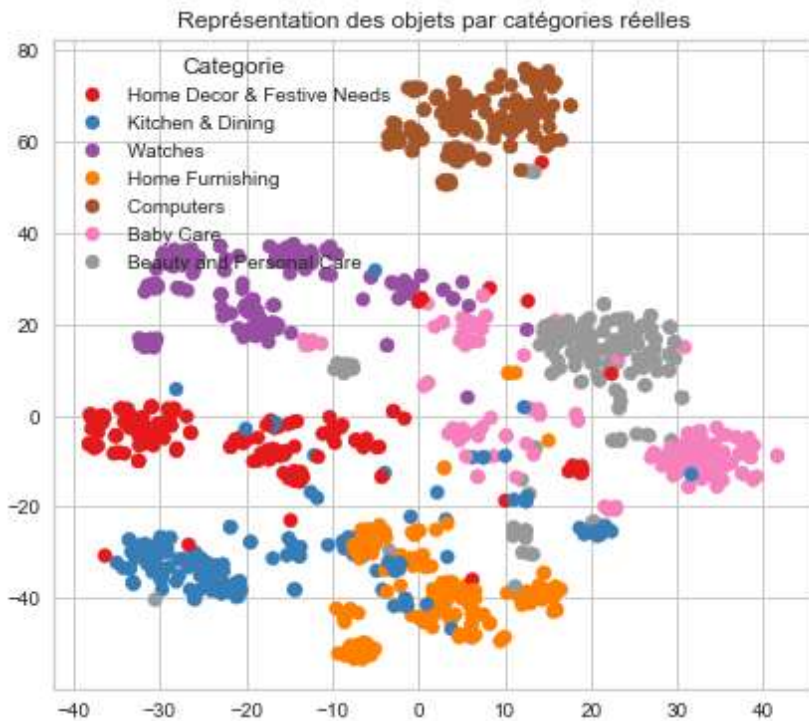


■ ARI =
0,603

BERT :
DES RÉSULTATS EN AMÉLIORATION VIA HUGGINGFACE

USE – UNIVERSAL SENTENCE ENCODER LE MODÈLE PRÉSENTANT LES MEILLEURS RÉSULTATS

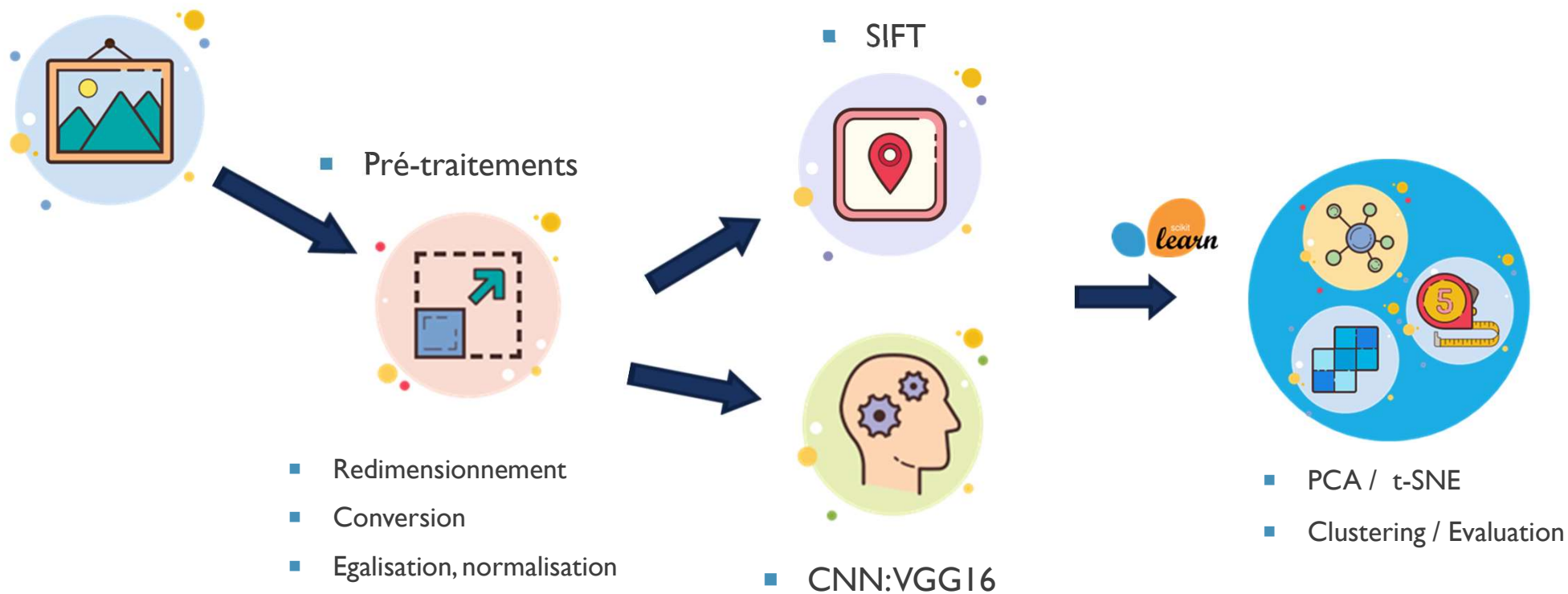
■ T-SNE et clustering



- Paramètres :
- Batch_size = 15
 - Feature : « name »



■ ARI =
0,6676



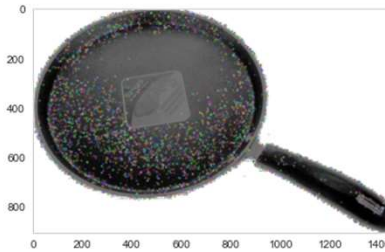
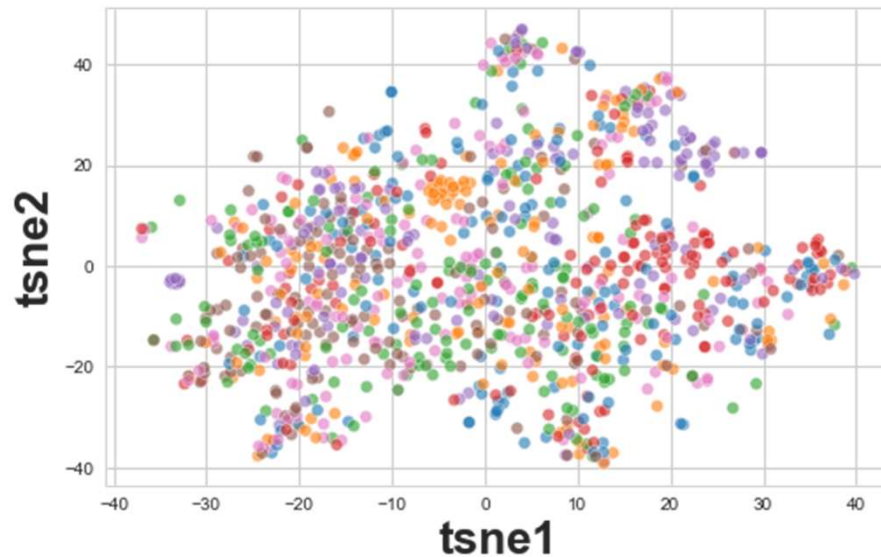
COMPUTER VISION- CV

SIFT

DES RÉSULTATS NON SATISFAISANTS

- T-SNE et clustering

TSNE selon les vraies classes



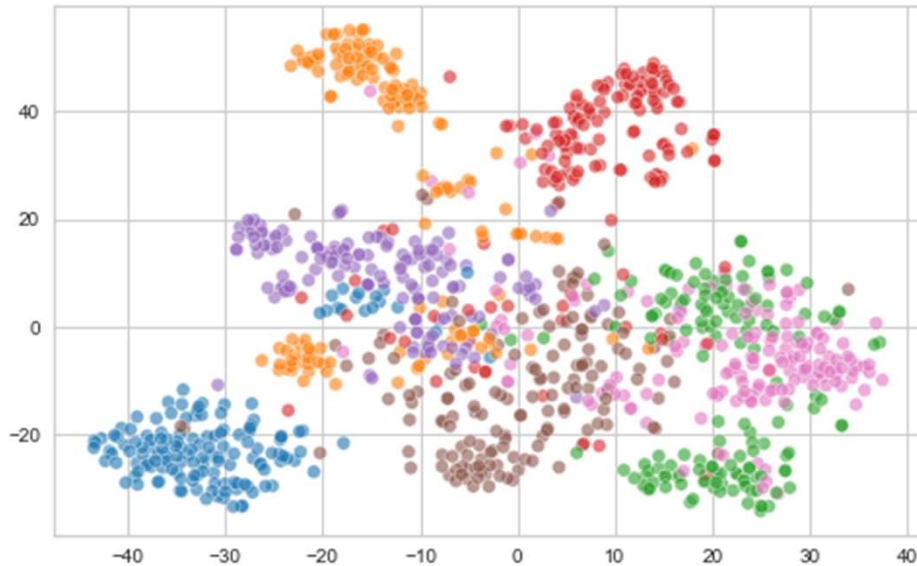
- Watches
- Kitchen & Dining
- Home Furnishing
- Beauty and Personal Care
- Computers
- Home Decor & Festive Needs
- Baby Care

- Paramètres :
- Nb descripteurs = (105381, 128)
- PCA (99%) :
- Avant : (1050, 325)
- Après : (1050, 281)



- **ARI = 0,0452**


■ T-SNE et clustering



- Watches
- Kitchen & Dining
- Home Furnishing
- Beauty and Personal Care
- Computers
- Home Decor & Festive Needs
- Baby Care

Watches	107	3	5	14	20	1	0
Kitchen & Dining	8	117	12	9	2	1	1
Home Furnishing	3	1	126	19	0	0	1
Beauty and Personal Care	21	1	10	111	4	0	3
Computers	73	0	0	4	73	0	0
Home Decor & Festive Needs	2	5	35	19	0	78	11
Baby Care	0	0	16	1	0	0	133
	0	1	2	3	4	5	6

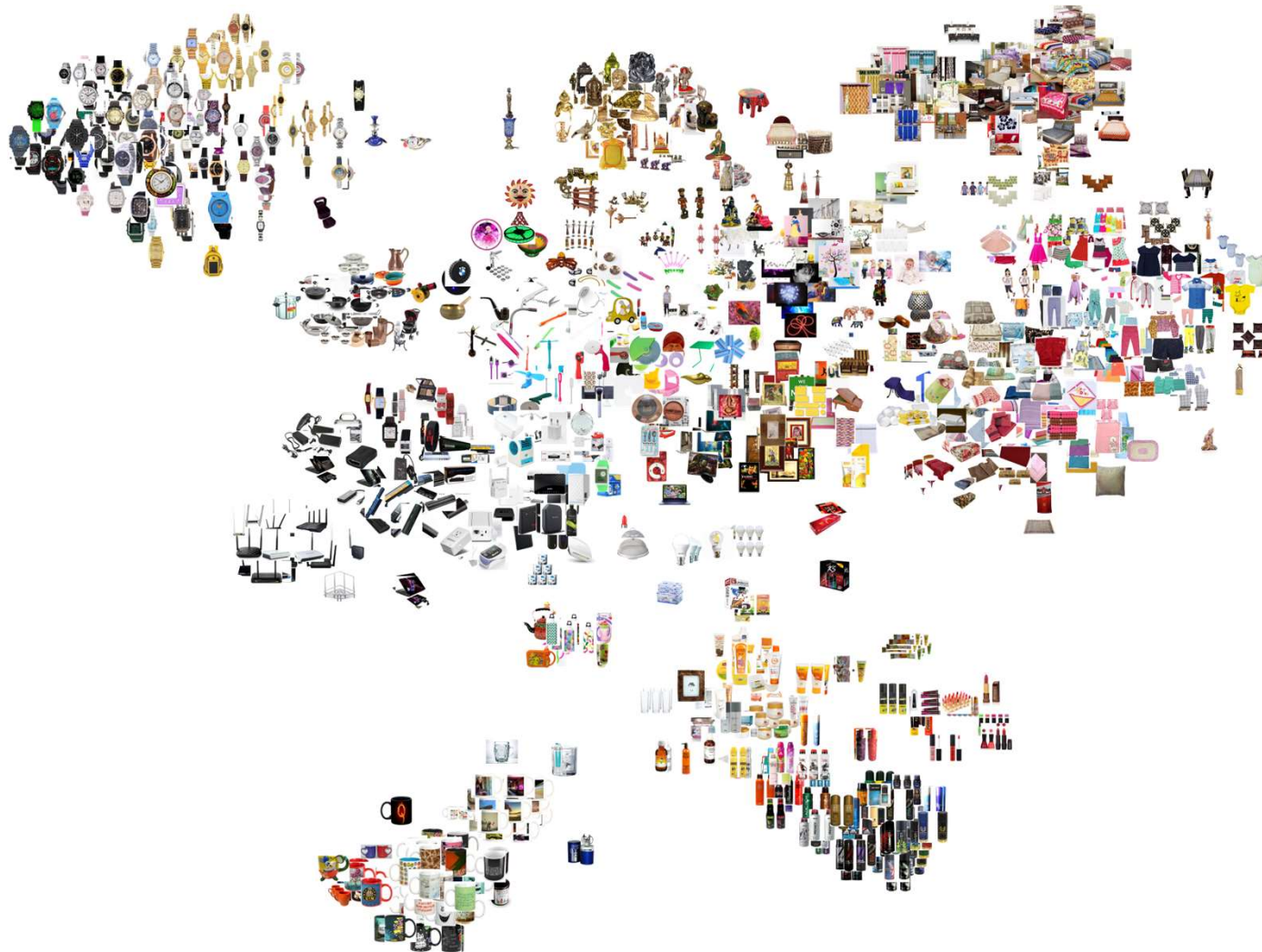
K

- Paramètres :
- Img size= (224, 224) 
- Pipeline preprocess VGG16
- PCA (99%) :
- Avant : (1050, 4096)
- Après : (1050, 803)



■ **ARI = 0,4728**

**CNN – VGG16 :
DES RÉSULTATS PROMETTEURS**



VGG16 VISUALISATION DU CLUSTERING PAR IMAGE

CONCLUSION

Des approches NLP et CV prometteuses : USE et CNN

Etude de faisabilité concluante pour le développement d'un moteur de classification supervisée

Croisement des 2 approches
Agrandissement de la base d'apprentissage (dont data augmentation)
Affiner le transfer learning avec le modèle VGG16



MERCI DE VOTRE ATTENTION

LAURENT CAGNIART

MARQUEUR ICON BY ICONS8