FLIPKART

NLP ET TRAITEMENT D'IMAGES

ETUDE DE FAISABILITÉ D'UN MOTEUR DE CLASSIFICATION

- Objectif : Automatiser l'attribution de la catégorie pour les articles vendus
- Echantillon : 1050 produits répartis en 7 classes de 150
- Méthodologie :
  - Partitionnement non supervisé en 7 classes à partir des données exploitables:
    - Product_name
    - Description
    - Image du produit

Exploitation des données textuelles : product_name

- Nom du produit composé de 2 à 27 mots
- Traitement:
  - Passage des mots en minuscules
  - Suppression de la ponctuation et des chiffres
  - Création d'une liste de mots
  - Suppression des mots les plus courants de la langue Anglaise
  - Stemming

➢ Corpus de 1941 mots

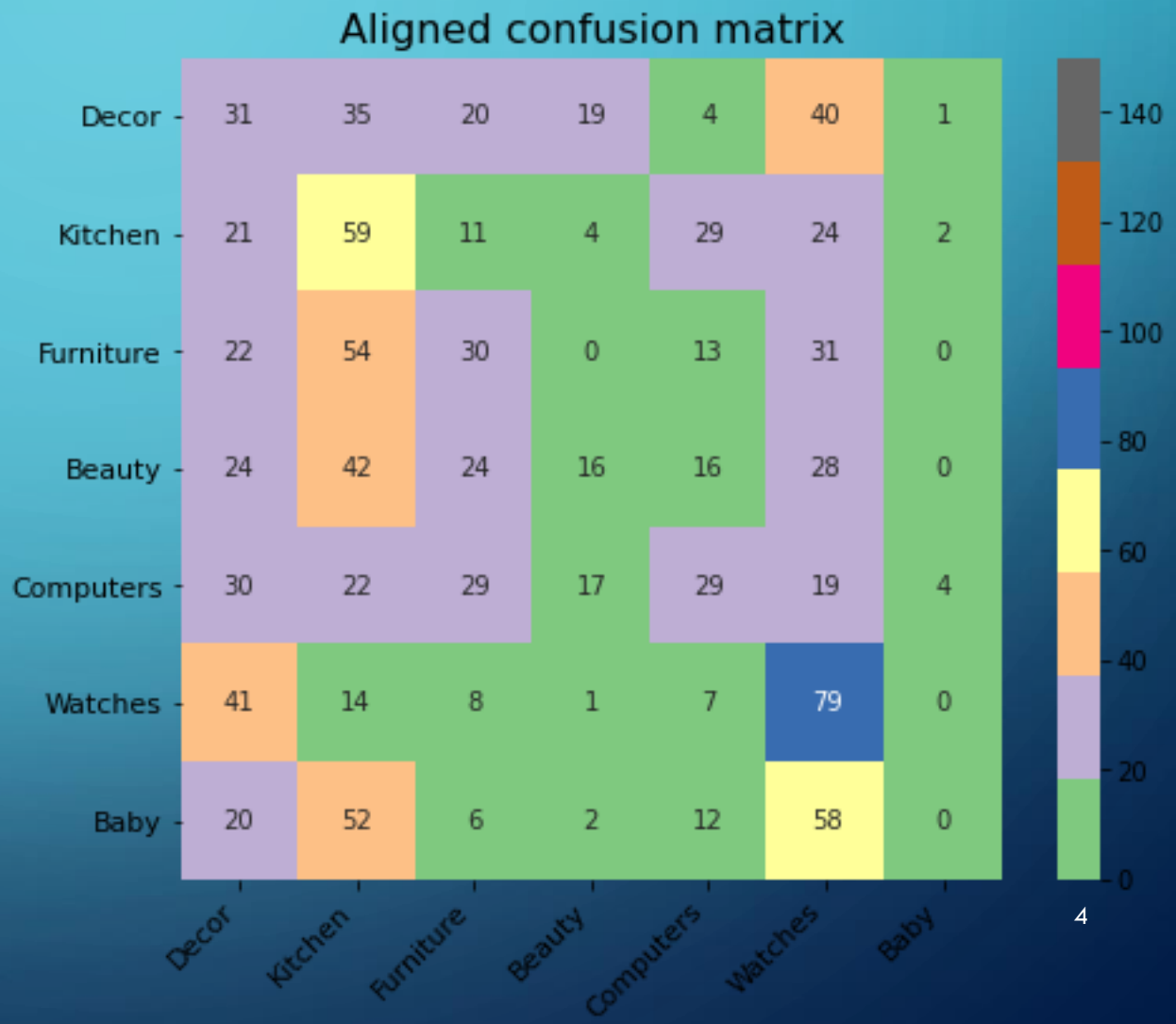
```
Original
236    Unique Design Mobile Stand Showpiece  -  15 cm
239          Ruchiworld Marble Pot Showpiece  -  4 cm
Name: product_name, dtype: object

Without stemming
236    cm showpiece mobile design unique stand
239         marble cm showpiece pot ruchiworld
Name: product_name, dtype: object

With stemming remove puctuation BEFORE tokenization
236    mobil cm showpiec uniqu design stand
239        marbl cm pot showpiec ruchiworld
Name: product_name, dtype: object

With stemming remove puctuation AFTER tokenization
236    mobil cm - 15 showpiec uniqu design stand
239         marbl cm - 4 pot showpiec ruchiworld
Name: product_name, dtype: object
```

Clusters sur les statistiques textuelles

Category	cluster	effectives
Decor	0	189
Kitchen	1	278
Furniture	6	128
Beauty	2	59
Computers	3	110
Watches	4	279
Baby	5	7

ARI : 0.03645444556585251

	precision	recall	f1-score	support
Decor	0.32	0.13	0.18	150
Kitchen	0.23	0.59	0.33	150
Furniture	0.23	0.20	0.22	150
Beauty	0.00	0.00	0.00	150
Computers	0.57	0.03	0.05	150
Watches	0.26	0.80	0.39	150
Baby	0.00	0.00	0.00	150
accuracy			0.25	1050
macro avg	0.23	0.25	0.17	1050
weighted avg	0.23	0.25	0.17	1050

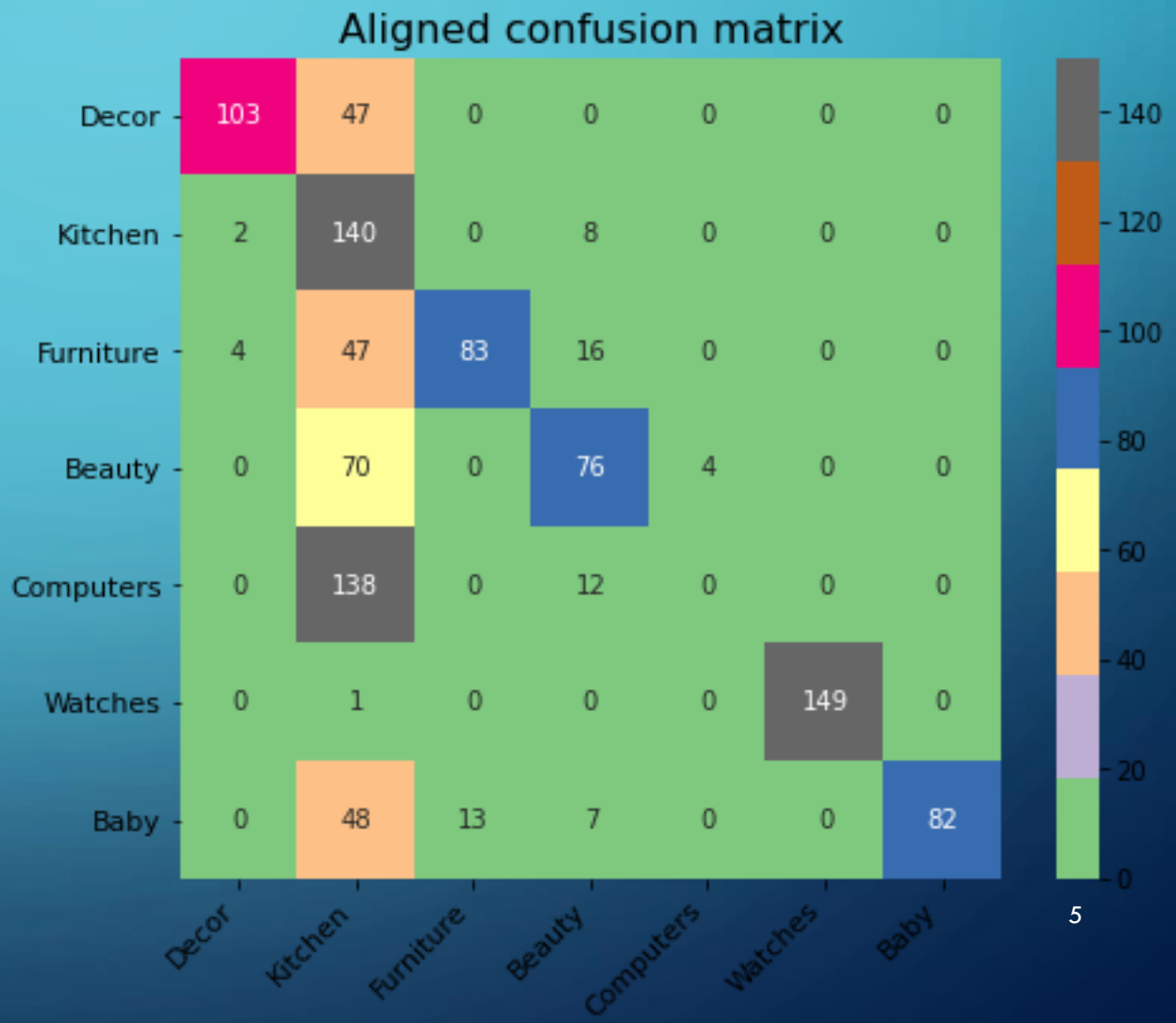


Bags of Words

Category	cluster	effectives
Decor	6	109
Kitchen	0	491
Furniture	5	98
Beauty	4	119
Computers	3	4
Watches	2	149
Baby	1	82

ARI : 0.31487230101378844

	precision	recall	f1-score	support
Decor	0.94	0.69	0.80	150
Kitchen	0.29	0.93	0.44	150
Furniture	0.86	0.55	0.67	150
Beauty	0.65	0.53	0.59	150
Computers	0.00	0.00	0.00	150
Watches	1.00	0.99	1.00	150
Baby	1.00	0.55	0.71	150
accuracy			0.61	1050
macro avg	0.68	0.61	0.60	1050
weighted avg	0.68	0.61	0.60	1050

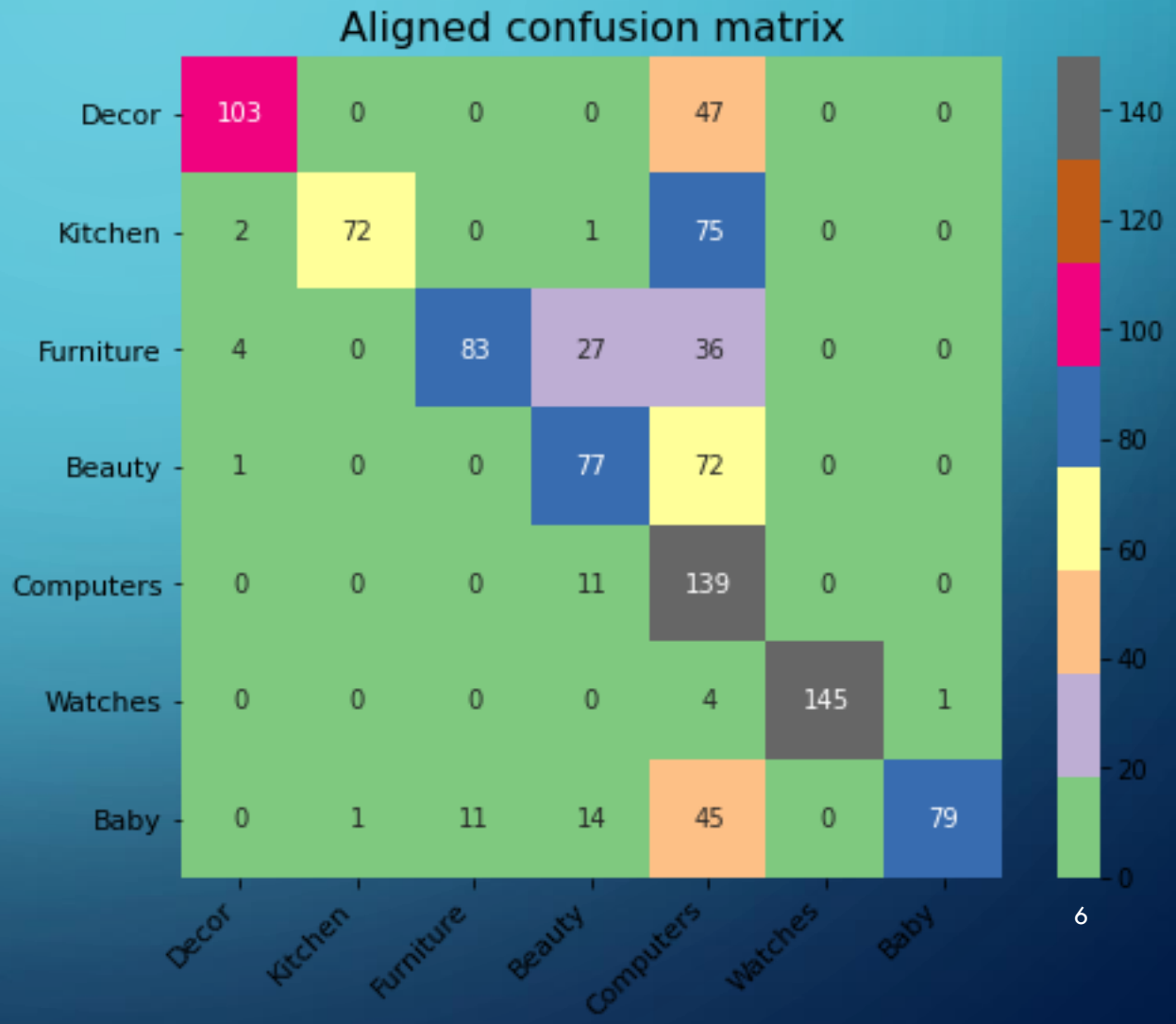


BoW et Term Frequency – Inverse Document Frequency

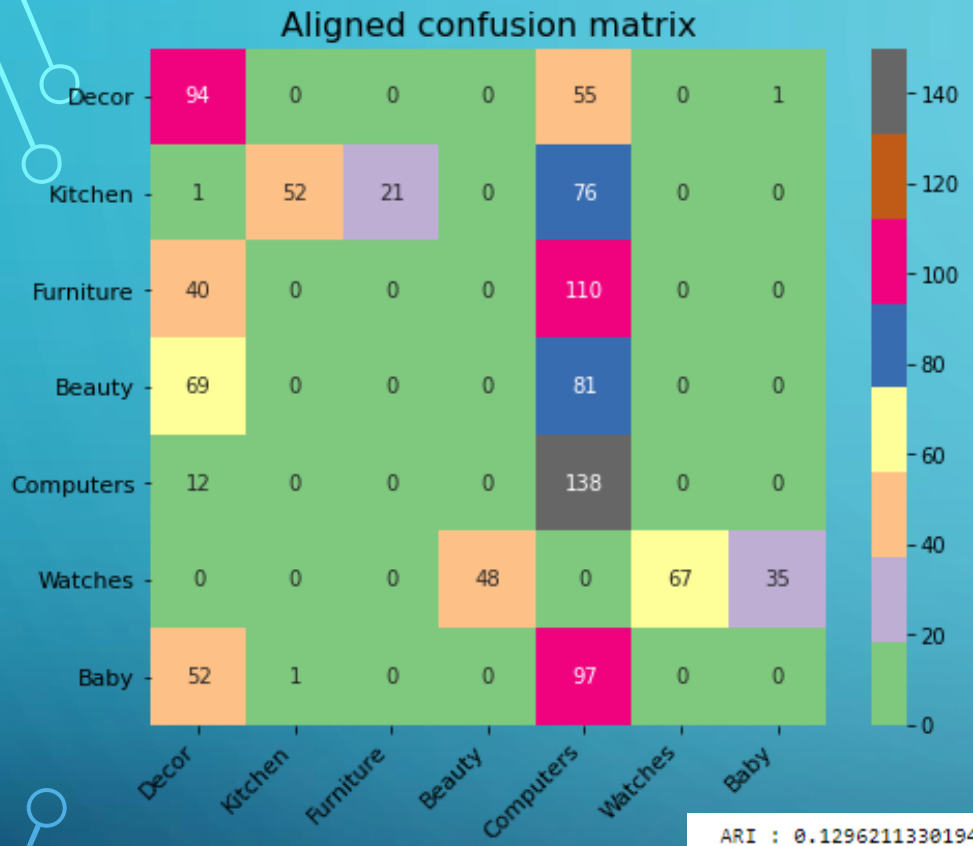
Category	cluster	effectives
Decor	4	110
Kitchen	5	73
Furniture	2	94
Beauty	3	130
Computers	1	418
Watches	0	145
Baby	6	80

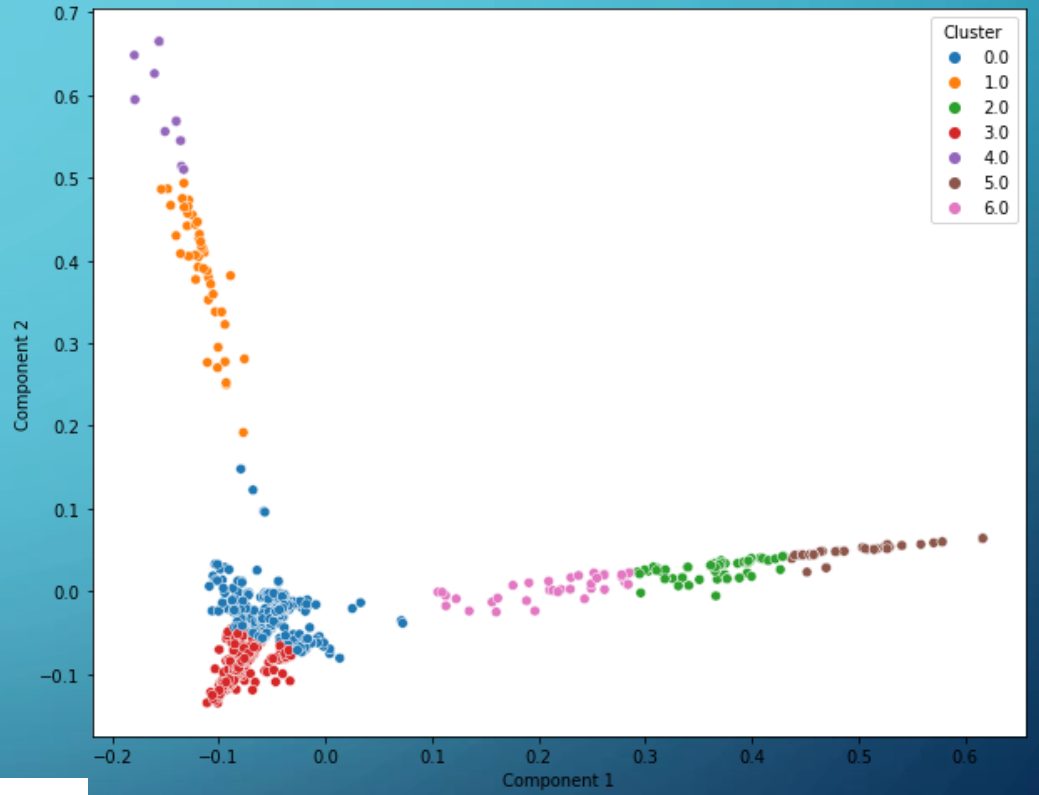
ARI : 0.34406698151953863

	precision	recall	f1-score	support
Decor	0.94	0.69	0.79	150
Kitchen	0.99	0.48	0.65	150
Furniture	0.88	0.55	0.68	150
Beauty	0.59	0.51	0.55	150
Computers	0.33	0.93	0.49	150
Watches	1.00	0.97	0.98	150
Baby	0.99	0.53	0.69	150
accuracy			0.66	1050
macro avg	0.82	0.66	0.69	1050
weighted avg	0.82	0.66	0.69	1050



BoW et TF-IDF réduction de dimensions ACP



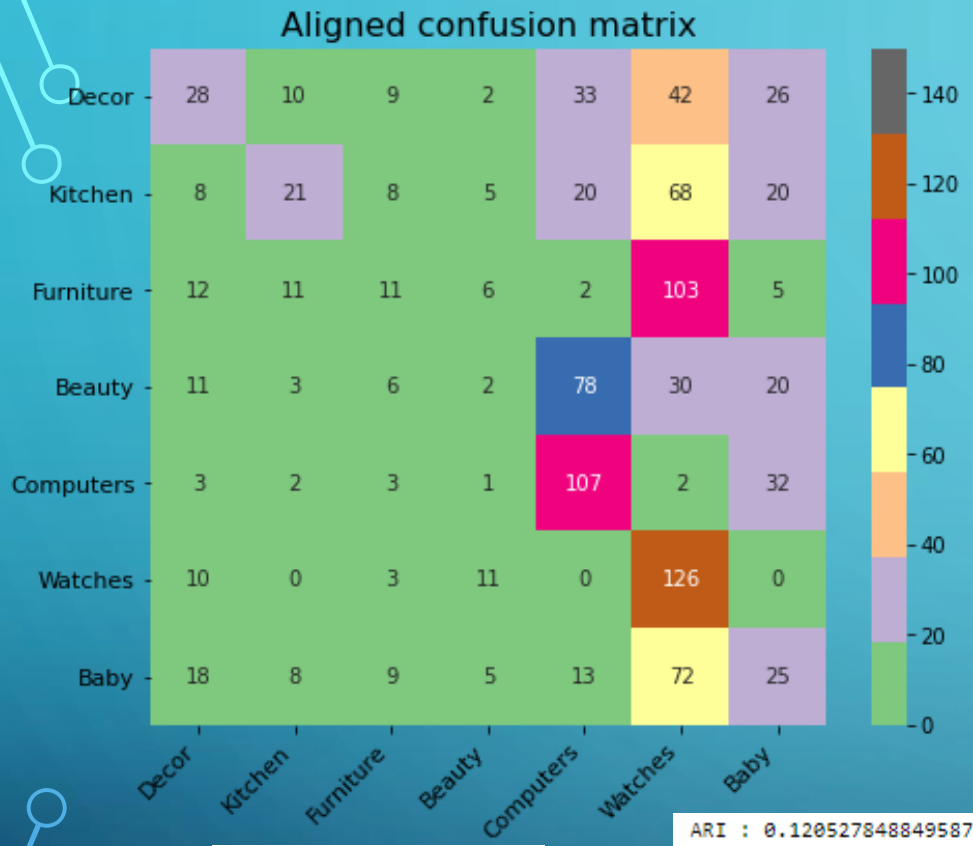


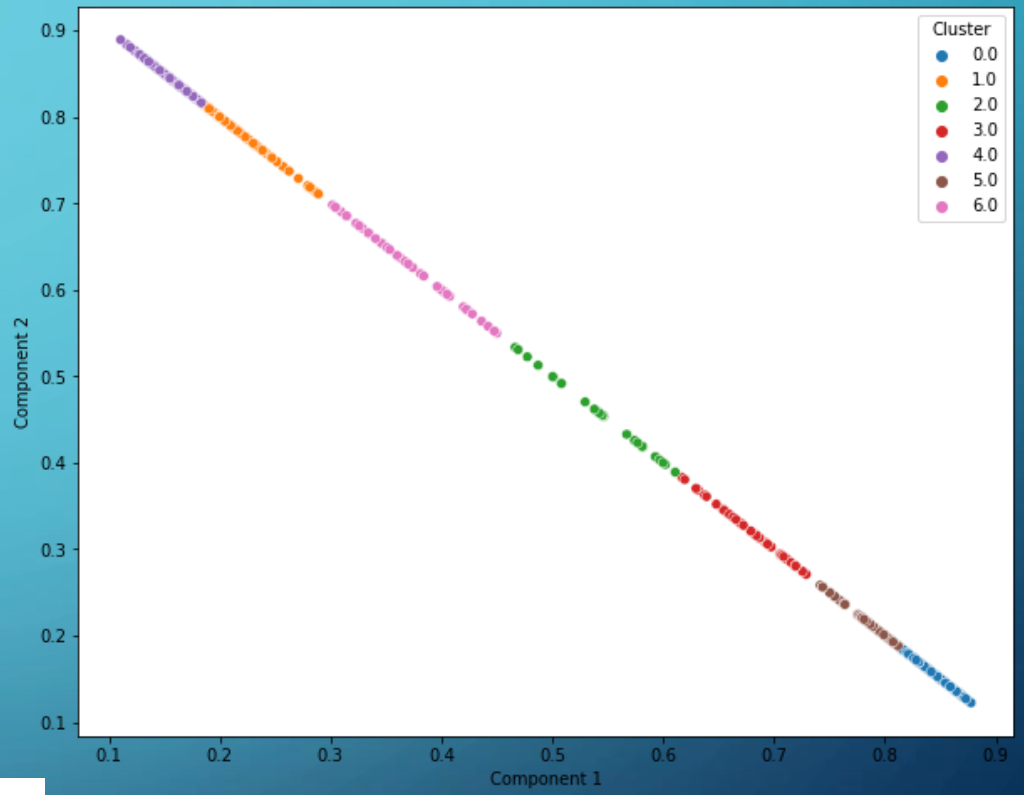
Category	cluster	effectives
Decor	3	268
Kitchen	1	53
Furniture	4	21
Beauty	5	48
Computers	0	557
Watches	2	67
Baby	6	36

ARI : 0.12962113301944223

	precision	recall	f1-score	support
Decor	0.35	0.63	0.45	150
Kitchen	0.99	0.49	0.65	150
Furniture	0.00	0.00	0.00	150
Beauty	0.00	0.00	0.00	150
Computers	0.25	0.92	0.39	150
Watches	0.99	1.00	1.00	150
Baby	0.00	0.00	0.00	150
accuracy			0.43	1050
macro avg	0.37	0.43	0.36	1050
weighted avg	0.37	0.43	0.36	1050

BoW et TF-IDF réduction de dimensions Latent Dirichlet Allocation



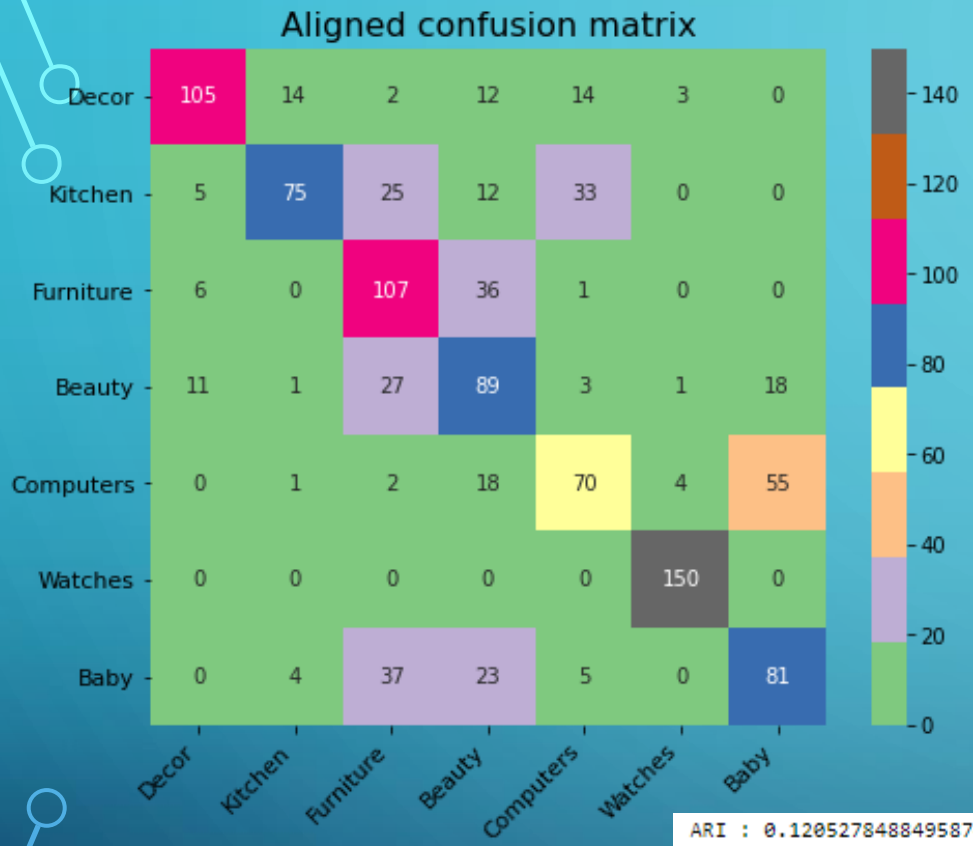


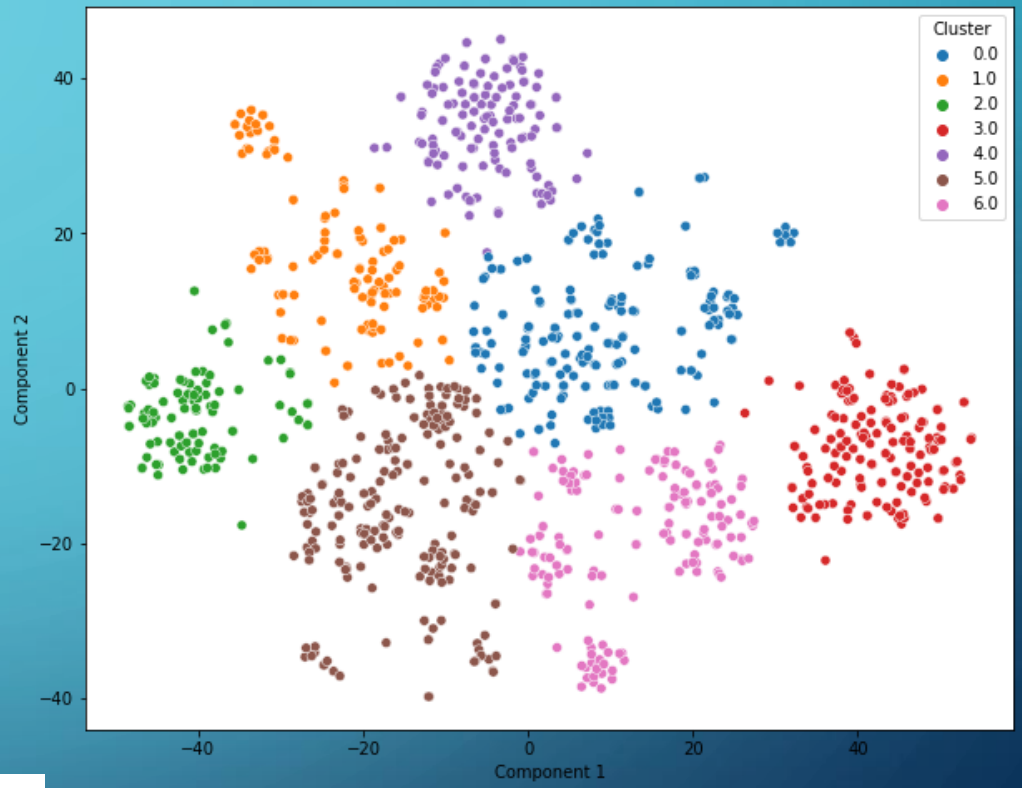
Category	cluster	effectives
Decor	6	90
Kitchen	3	55
Furniture	4	49
Beauty	2	32
Computers	1	253
Watches	0	443
Baby	5	128

ARI : 0.12052784884958742

	precision	recall	f1-score	support
Decor	0.31	0.19	0.23	150
Kitchen	0.38	0.14	0.20	150
Furniture	0.22	0.07	0.11	150
Beauty	0.00	0.00	0.00	150
Computers	0.36	0.93	0.52	150
Watches	0.29	0.91	0.44	150
Baby	0.00	0.00	0.00	150
accuracy			0.32	1050
macro avg	0.22	0.32	0.22	1050
weighted avg	0.22	0.32	0.22	1050

BoW et TF-IDF réduction de dimensions T-Sne



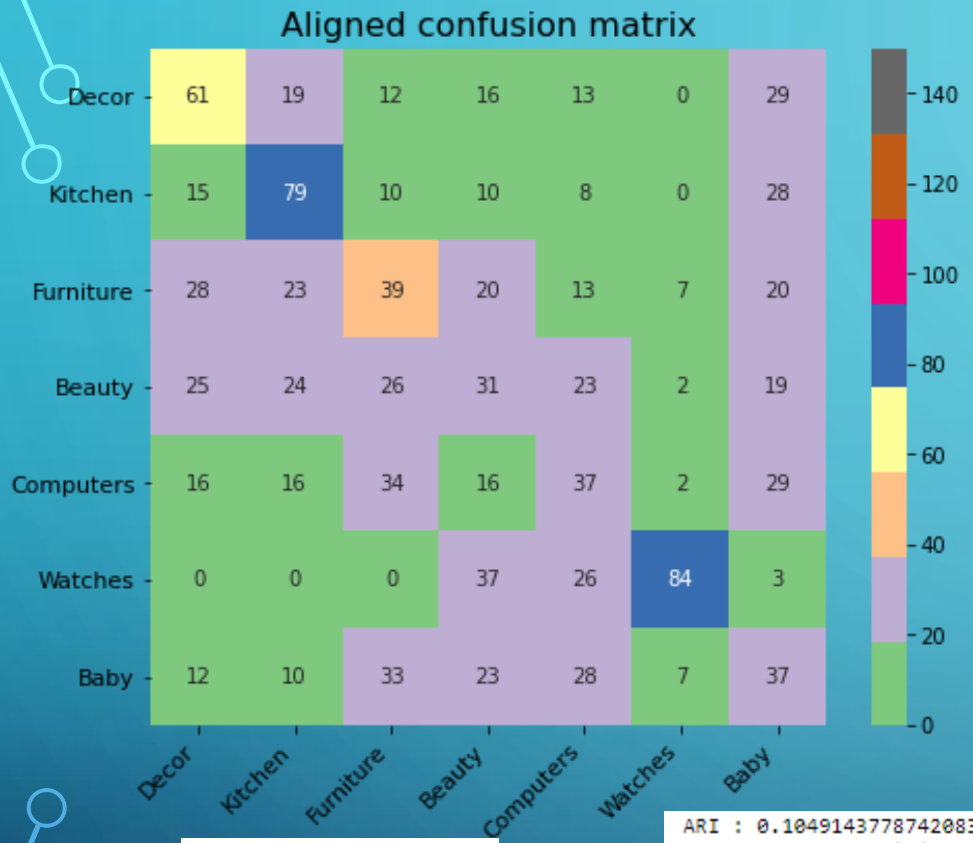


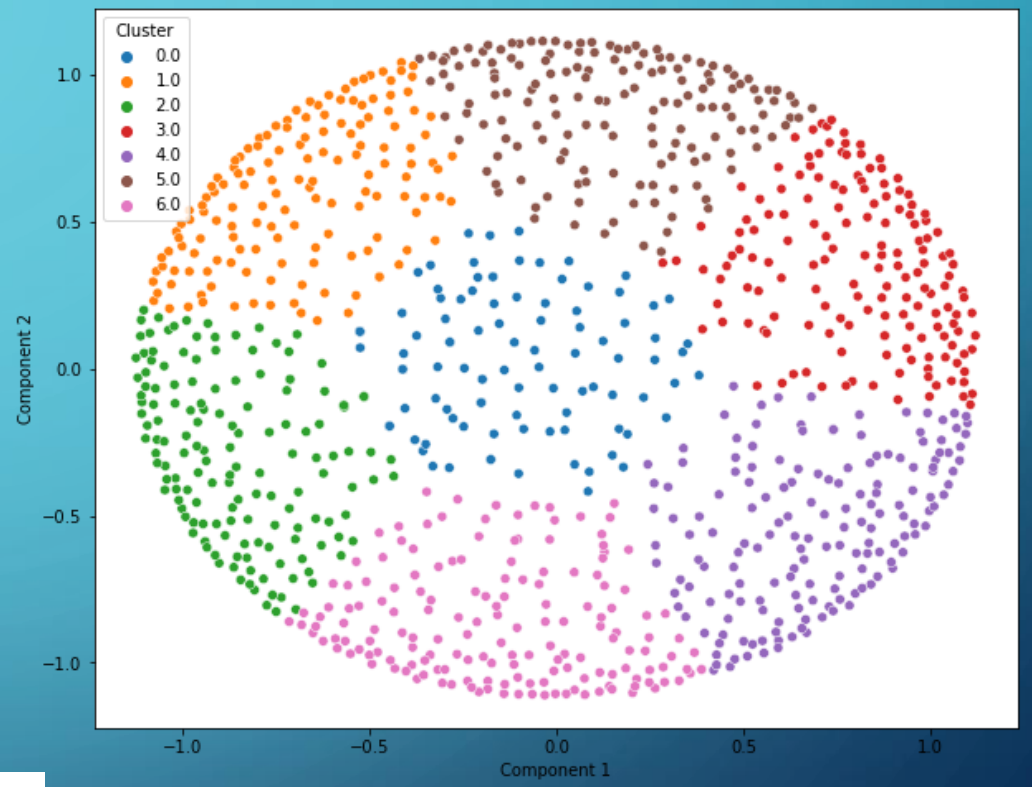
Category	cluster	effectives
Decor	4	127
Kitchen	2	95
Furniture	5	200
Beauty	0	190
Computers	1	126
Watches	3	158
Baby	6	154

ARI : 0.12052784884958742

	precision	recall	f1-score	support
Decor	0.31	0.19	0.23	150
Kitchen	0.38	0.14	0.20	150
Furniture	0.22	0.07	0.11	150
Beauty	0.00	0.00	0.00	150
Computers	0.36	0.93	0.52	150
Watches	0.29	0.91	0.44	150
Baby	0.00	0.00	0.00	150
accuracy			0.32	1050
macro avg	0.22	0.32	0.22	1050
weighted avg	0.22	0.32	0.22	1050

BoW et TF-IDF réduction de dimensions MDS



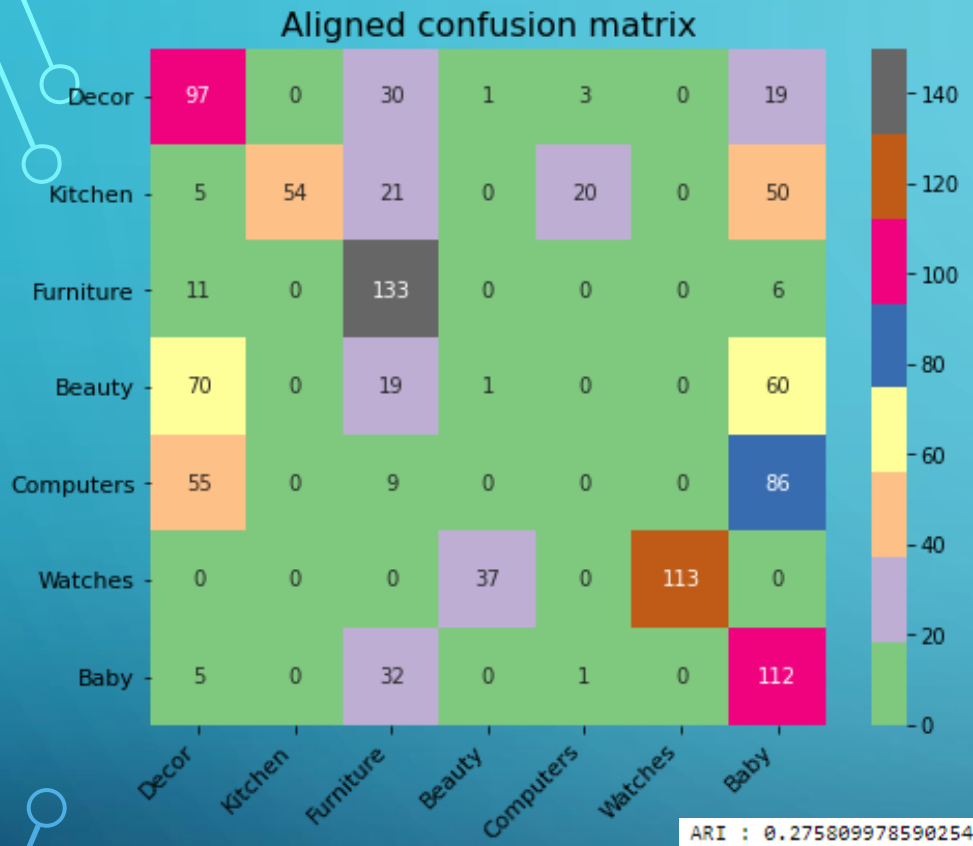


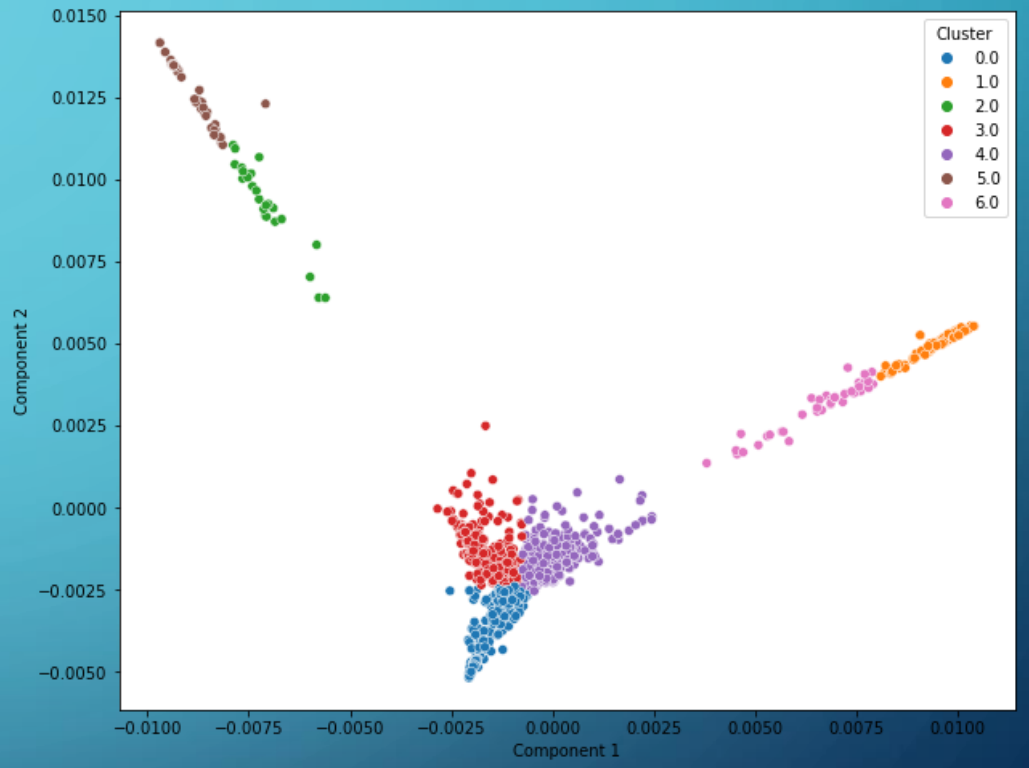
Category	cluster	effectives
Decor	3	157
Kitchen	4	171
Furniture	5	154
Beauty	2	153
Computers	1	148
Watches	0	102
Baby	6	165

ARI : 0.1049143778742083

	precision	recall	f1-score	support
Decor	0.39	0.41	0.40	150
Kitchen	0.46	0.53	0.49	150
Furniture	0.25	0.26	0.26	150
Beauty	0.00	0.00	0.00	150
Computers	0.25	0.25	0.25	150
Watches	0.47	0.81	0.60	150
Baby	0.22	0.25	0.23	150
accuracy			0.36	1050
macro avg	0.29	0.36	0.32	1050
weighted avg	0.29	0.36	0.32	1050

BoW et TF-IDF réduction de dimensions Spectral



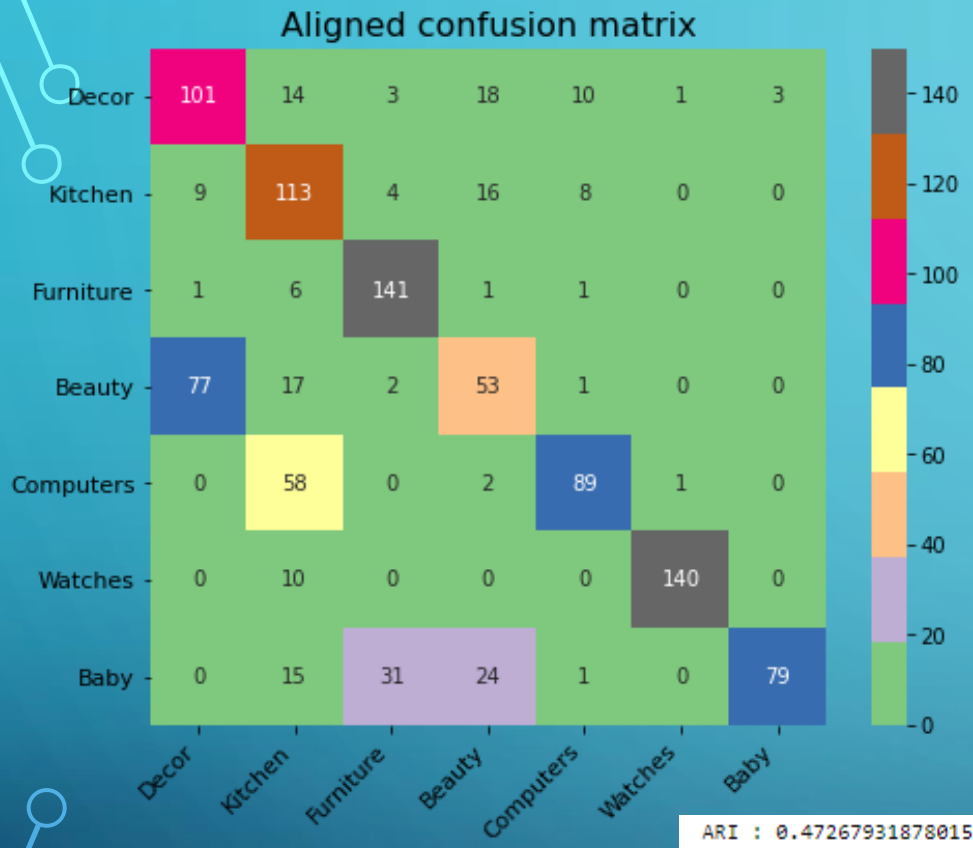


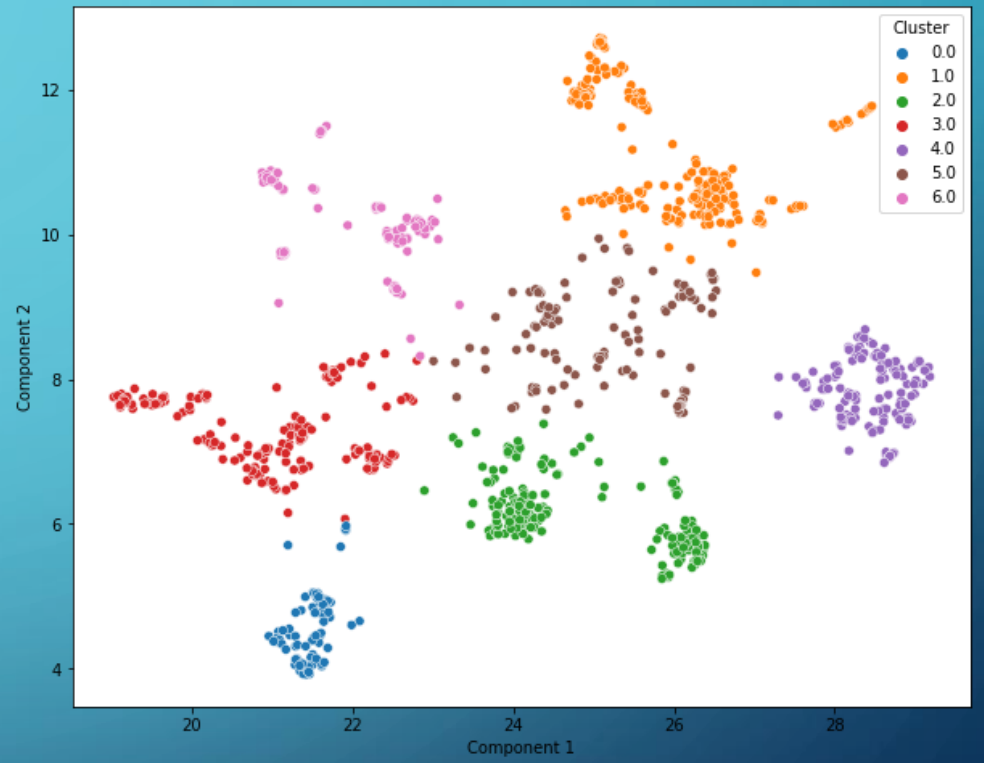
Category	cluster	effectives
Decor	0	243
Kitchen	5	54
Furniture	3	244
Beauty	6	39
Computers	2	24
Watches	1	113
Baby	4	333

ARI : 0.27580997859025425

	precision	recall	f1-score	support
Decor	0.40	0.65	0.49	150
Kitchen	0.95	0.49	0.65	150
Furniture	0.55	0.89	0.68	150
Beauty	0.00	0.00	0.00	150
Computers	0.00	0.00	0.00	150
Watches	0.99	1.00	0.99	150
Baby	0.34	0.75	0.46	150
accuracy			0.54	1050
macro avg	0.46	0.54	0.47	1050
weighted avg	0.46	0.54	0.47	1050

BoW et TF-IDF réduction de dimensions Umap





Category	cluster	effectives
Decor	2	188
Kitchen	1	233
Furniture	3	181
Beauty	5	114
Computers	6	110
Watches	4	142
Baby	0	82

ARI : 0.4726793187801535

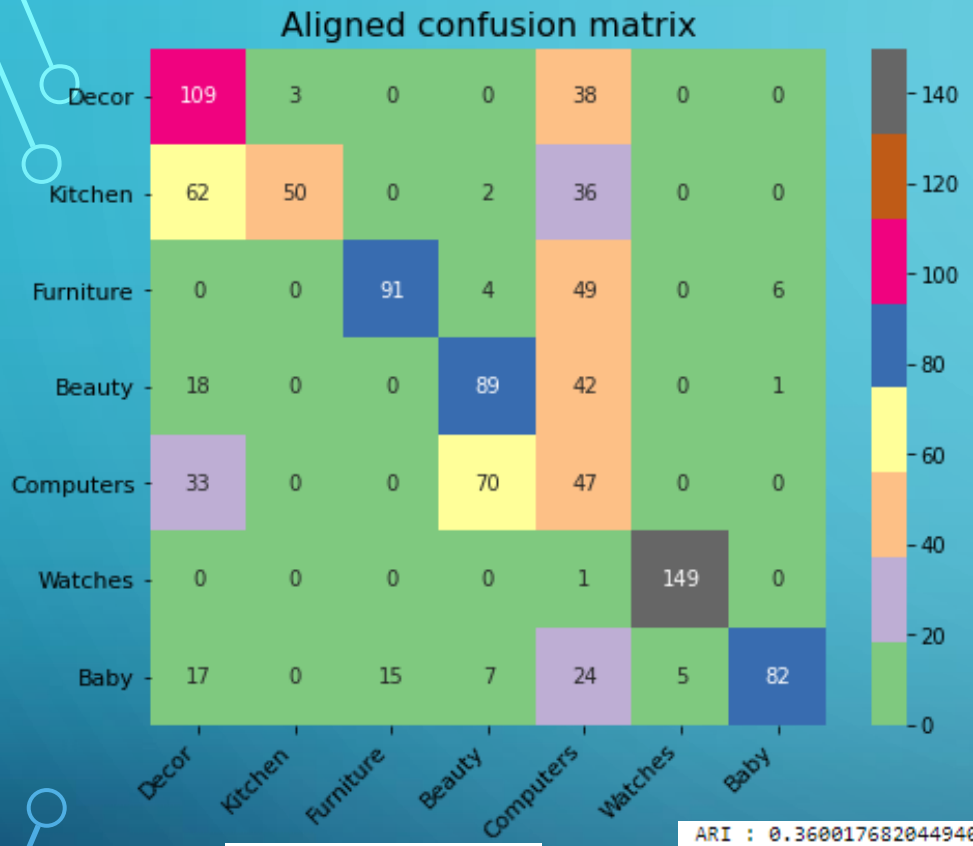
	precision	recall	f1-score	support
Decor	0.54	0.67	0.60	150
Kitchen	0.48	0.75	0.59	150
Furniture	0.78	0.94	0.85	150
Beauty	0.46	0.35	0.40	150
Computers	0.81	0.59	0.68	150
Watches	0.99	0.93	0.96	150
Baby	0.96	0.53	0.68	150
accuracy			0.68	1050
macro avg	0.72	0.68	0.68	1050
weighted avg	0.72	0.68	0.68	1050

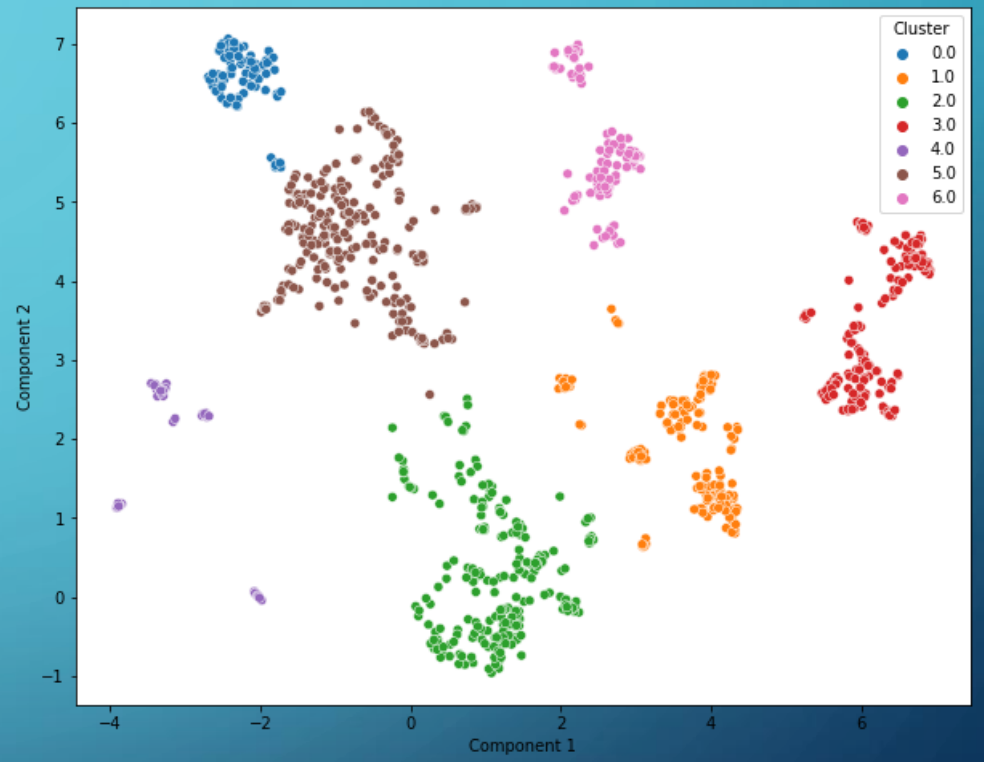
Exploitation des données textuelles : description

- Nom du produit composé de 18 à 589 mots
- Traitement:
  - Passage des mots en minuscules
  - Suppression de la ponctuation et des chiffres
  - Création d'une liste de mots
  - Suppression des mots les plus courants de la langue Anglaise
  - Stemming

➢ Corpus de 4595 mots

BoW et TF-IDF réduction de dimensions Umap



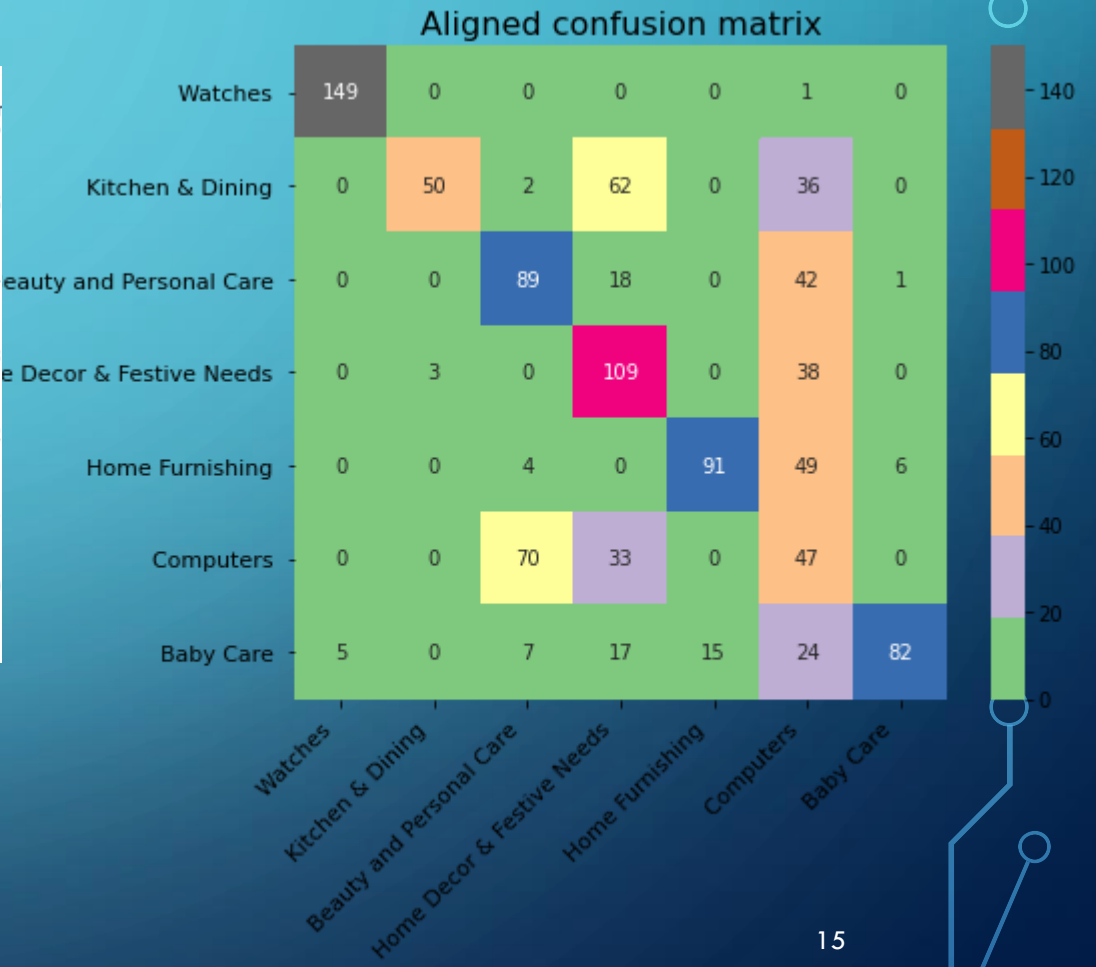


Category	cluster	effectives
Decor	2	239
Kitchen	4	53
Furniture	6	106
Beauty	1	172
Computers	5	237
Watches	3	154
Baby	0	89

ARI : 0.3600176820449404

	precision	recall	f1-score	support
Decor	0.46	0.73	0.56	150
Kitchen	0.94	0.33	0.49	150
Furniture	0.41	0.93	0.57	150
Beauty	0.52	0.59	0.55	150
Computers	0.00	0.00	0.00	150
Watches	0.97	0.99	0.98	150
Baby	0.92	0.55	0.69	150
accuracy			0.59	1050
macro avg	0.60	0.59	0.55	1050
weighted avg	0.60	0.59	0.55	1050

7 CLUSTERS SUR BOW + TF-IDF + UMAP



ARI : 0.3600176820449404

	precision	recall	f1-score	support
Watches	0.97	0.99	0.98	150
Kitchen & Dining	0.94	0.33	0.49	150
Beauty and Personal Care	0.52	0.59	0.55	150
Home Decor & Festive Needs	0.46	0.73	0.56	150
Home Furnishing	0.41	0.93	0.57	150
Computers	0.00	0.00	0.00	150
Baby Care	0.92	0.55	0.69	150
accuracy			0.59	1050
macro avg	0.60	0.59	0.55	1050
weighted avg	0.60	0.59	0.55	1050

Clustering sur description : BoW + TFiDF + UMAP 2 composantes

Scores

product_name

title	precision	recall	f1_score	ARI
BoW	67.79%	60.67%	59.95%	31.49%
Text statistics	23.02%	24.86%	16.63%	3.65%
BoW + TFiDF	81.69%	66.48%	68.97%	34.41%
BoW + TFiDF + PCA 2 components	36.83%	43.33%	35.55%	12.96%
BoW + LatentDirichletAllocation	67.80%	60.86%	62.16%	32.78%
BoW + TFiDF + LatentDirichletAllocation	22.44%	32.00%	21.58%	12.05%
BoW + TFiDF + T-SNE 2 composantes	66.44%	64.48%	64.56%	40.66%
BoW + TFiDF + MDS 2 composantes	29.32%	35.62%	31.81%	10.49%
BoW + TFiDF + Spectral 2 composantes	45.95%	53.90%	46.79%	27.58%
BoW + TFiDF + UMAP 2 composantes	71.78%	68.19%	68.08%	47.27%

description

title	precision	recall	f1_score	ARI
BoW	38.71%	30.86%	24.59%	5.93%
Text statistics	20.61%	24.19%	17.53%	2.27%
BoW + TFiDF	53.88%	47.33%	45.92%	23.33%
BoW + TFiDF + PCA	32.76%	41.62%	36.14%	15.47%
BoW + LatentDirichletAllocation	38.07%	40.86%	35.72%	16.67%
BoW + TFiDF + LatentDirichletAllocation	21.80%	30.48%	21.91%	9.98%
BoW + TFiDF + T-SNE 2 composantes	52.47%	58.48%	53.76%	35.13%
BoW + TFiDF + MDS 2 composantes	32.18%	35.81%	33.16%	16.05%
BoW + TFiDF + Spectral 2 composantes	63.96%	53.14%	55.69%	27.37%
BoW + TFiDF + UMAP 2 composantes	60.20%	58.95%	54.86%	36.00%

Exploitation des images

- Chargement de l'image en niveaux de gris
- Application d'un filtre adaptatif d'amélioration des contrastes
- Descripteurs
  - Création des descripteurs de chaque image
  - Sur l'ensemble des descripteurs on regroupe ceux qui sont semblables pour obtenir les Visual Words
  - Pour chaque image comptage des Visual Words

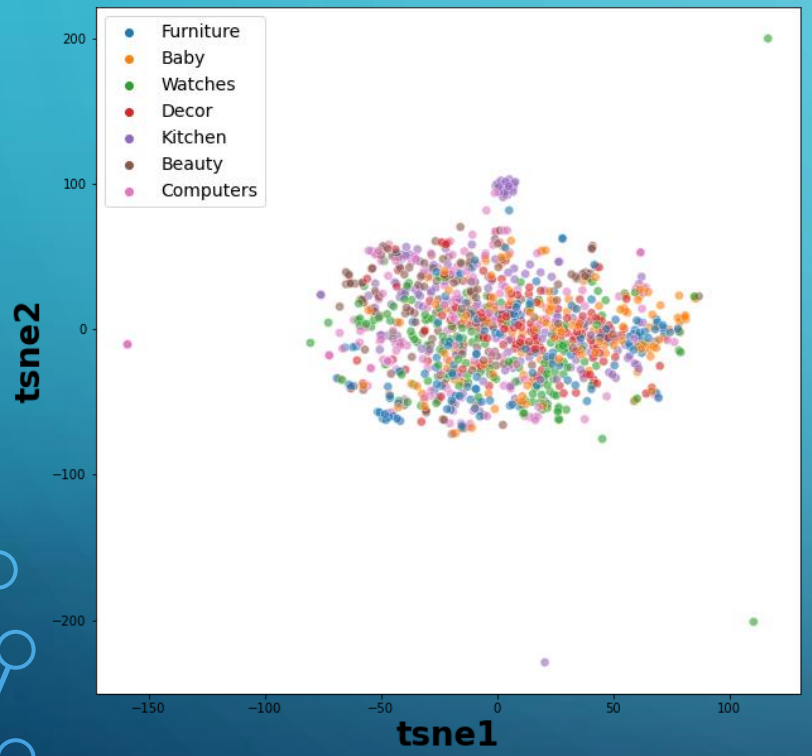
➢ Features crées à partir des descripteurs

Images descripteurs SIFT

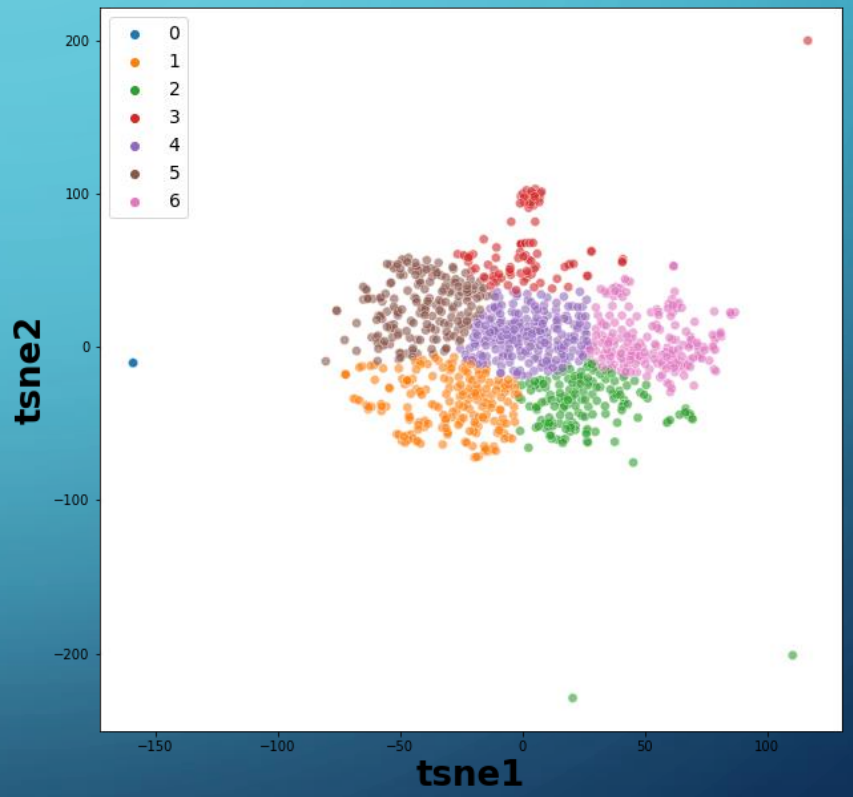
Descripteurs SIFT pour toutes les images

- Pour chaque image un descripteur sous forme d'un vecteur de longueur 128
- Sur l'ensmble des images 549 281 descripteurs
- Création de 741 visual words
- Création de l'histogramme de chaque image
- Acp en gardant les composantes principales expliquant 99% de la variance: reste 562 composantes

TSNE selon les vraies classes



TSNE selon les clusters

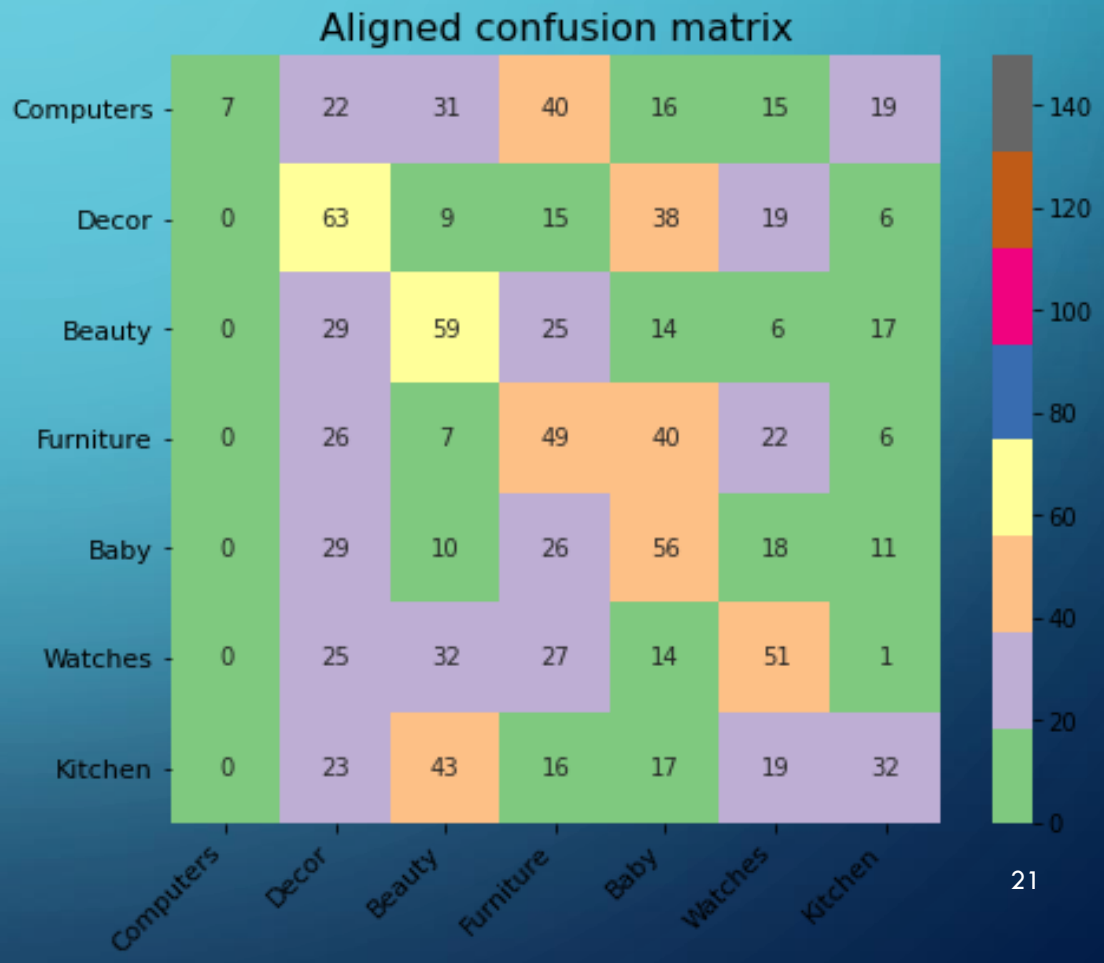


Clustering sur les Bags of Visual Words issus des descripteurs SIFT

Category	cluster	effectives
Computers	0	7
Decor	4	217
Beauty	5	191
Furniture	1	198
Baby	6	195
Watches	2	150
Kitchen	3	92

ARI : 0.04984062766276105

	precision	recall	f1-score	support
Computers	1.00	0.05	0.09	150
Decor	0.29	0.42	0.34	150
Beauty	0.31	0.39	0.35	150
Furniture	0.25	0.33	0.28	150
Baby	0.29	0.37	0.32	150
Watches	0.34	0.34	0.34	150
Kitchen	0.35	0.21	0.26	150
accuracy			0.30	1050
macro avg	0.40	0.30	0.28	1050
weighted avg	0.40	0.30	0.28	1050

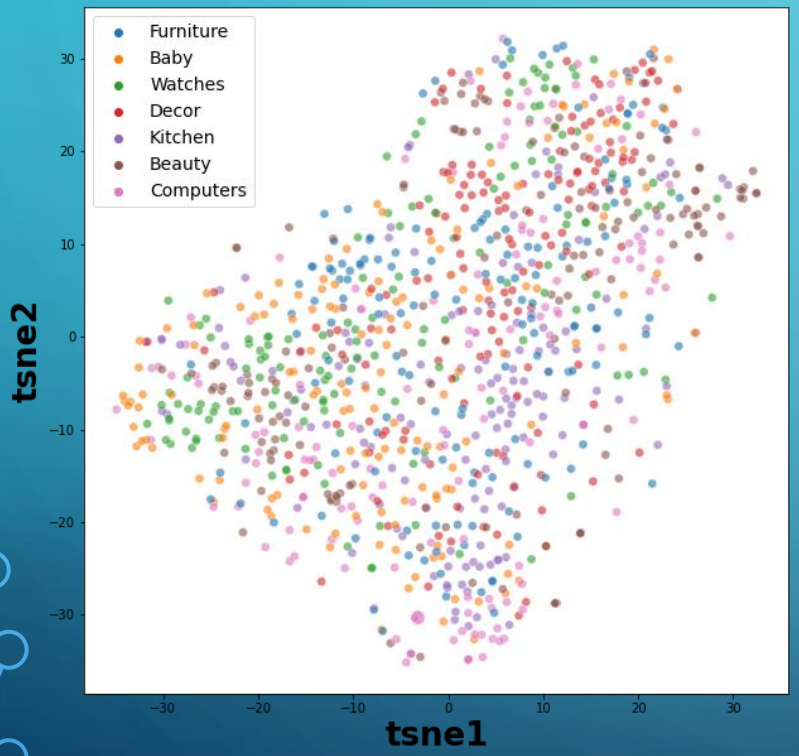


DESCRIPTEURS ORB

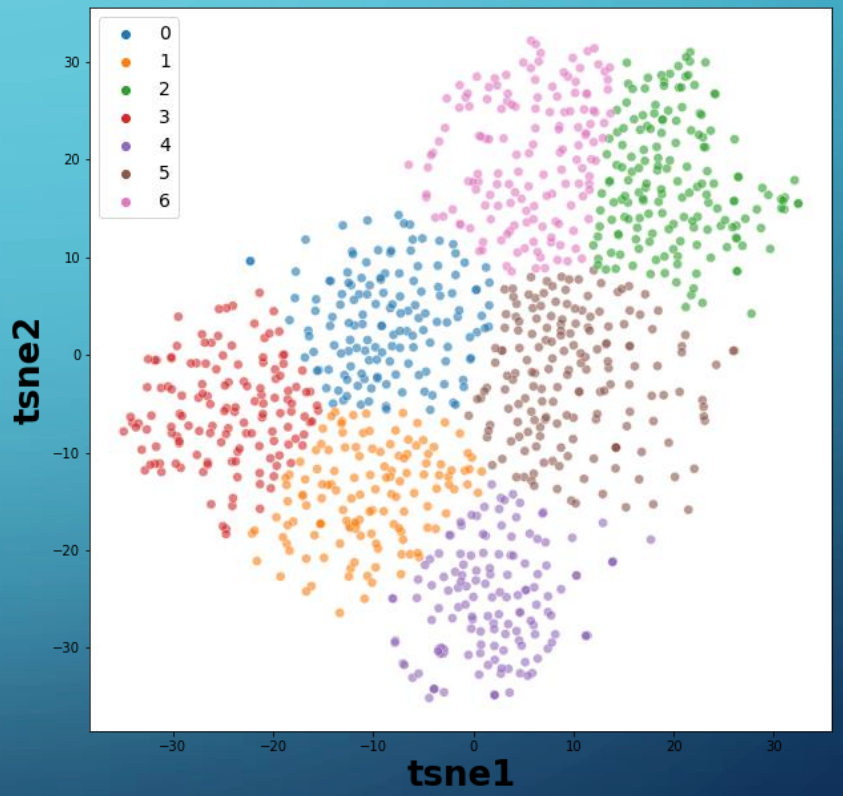
Descripteurs SIFT pour toutes les images

- Pour chaque image un descripteur sous forme d'un vecteur de longueur 32
- Sur l'ensmble des images 1 033 634 descripteurs
- Création de 1017 visual words
- Création de l'histogramme de chaque image
- Acp en gardant les composantes principales expliquant 99% de la variance: reste 675 composantes

TSNE selon les vraies classes



TSNE selon les clusters

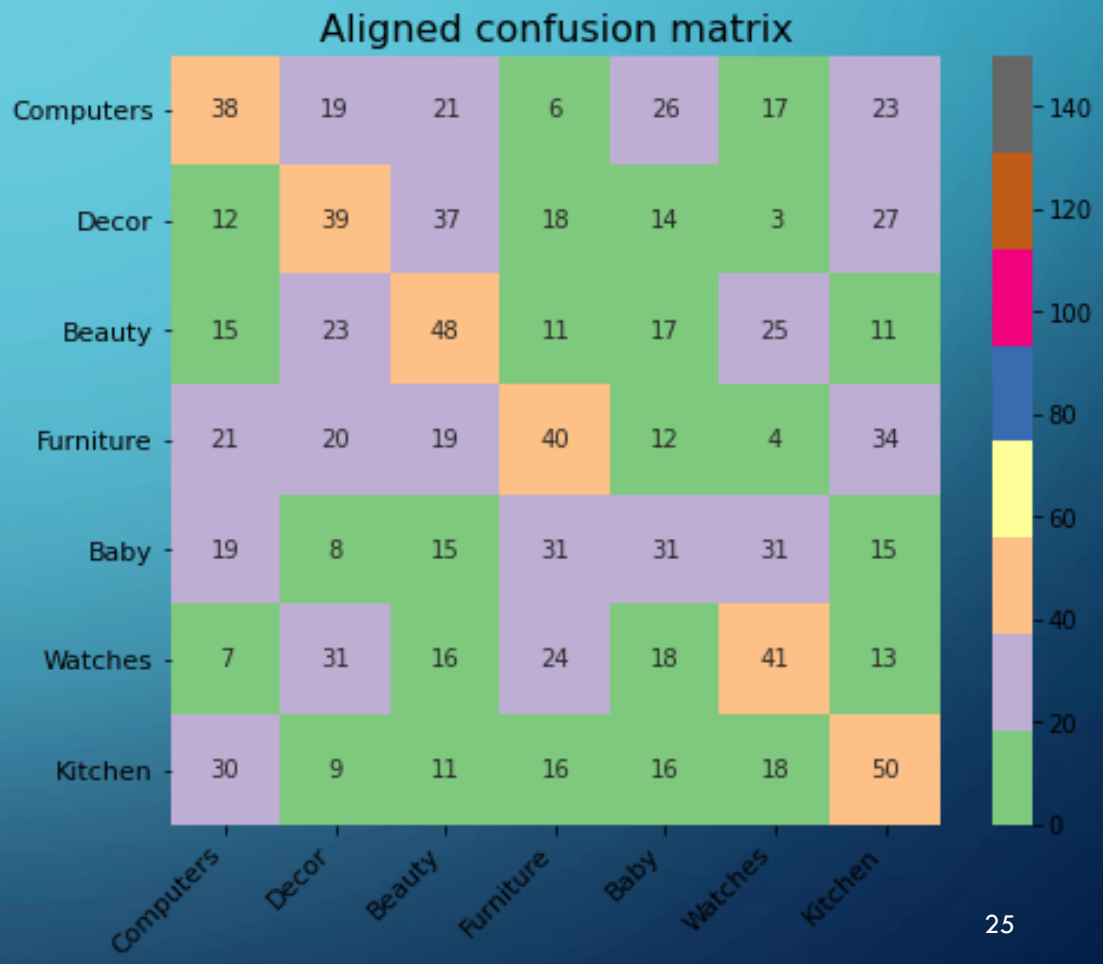


Clustering sur les Bags of Visual Words issus des descripteurs ORB

Category	cluster	effectives
Computers	4	142
Decor	6	149
Beauty	2	167
Furniture	0	146
Baby	1	134
Watches	3	139
Kitchen	5	173

ARI : 0.03830128474861062

	precision	recall	f1-score	support
Computers	0.27	0.25	0.26	150
Decor	0.26	0.26	0.26	150
Beauty	0.29	0.32	0.30	150
Furniture	0.27	0.27	0.27	150
Baby	0.23	0.21	0.22	150
Watches	0.29	0.27	0.28	150
Kitchen	0.29	0.33	0.31	150
accuracy			0.27	1050
macro avg	0.27	0.27	0.27	1050
weighted avg	0.27	0.27	0.27	1050



TRANSFERT LEARNING

Word embedding transfert learning modèle BERT product_name et description

Utilisation du modèle pré-entrainé BERT pour encoder les variables:

- encodage des variables avec le pré-processing inférant à ce modèle
- Pour chaque phrase on obtient un vecteur de longueur 768

title	precision	recall	f1_score	ARI
product_name BERT enbedding	26.51%	34.10%	29.05%	11.23%
product_name BERT embedding + PCA(99%)	26.30%	33.90%	28.89%	11.25%
description BERT embedding	26.41%	32.86%	27.92%	8.85%
description BERT embedding + PCA(99%)	26.49%	32.86%	27.95%	8.88%
both BERT embedding	25.14%	35.71%	28.59%	11.10%
both BERT embedding + PCA(99%)	20.47%	35.52%	25.59%	10.85%

Feature extraction images transfert learning modèles VGG16 et RESNET50

Utilisation des modèles pré-entrainé pour encoder les images:

- On enlève la couche dense des modèles pour obtenir des vecteurs pour chaque image
- On effectue un prétraitement des images
- Pour chaque imgae on obtient un encodage

VGG16

- Taille des images 224 x 224
- Taille de l'encodage 4096

RESNET50

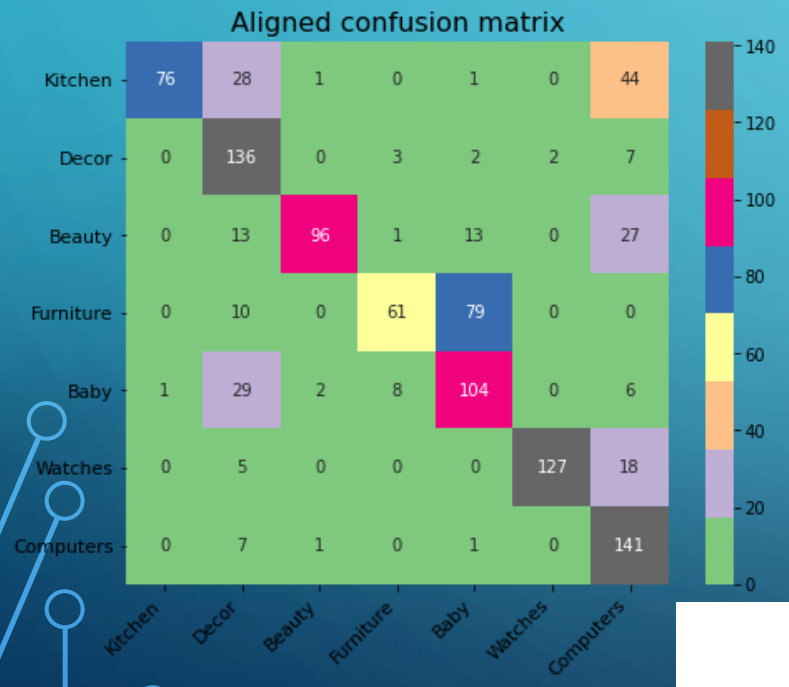
- Taille des images 224 x 224
- Taille de l'encodage 2048

VGG16

Category	cluster	effectives
Kitchen	3	77
Decor	5	228
Beauty	2	100
Furniture	4	73
Baby	0	200
Watches	6	129
Computers	1	243

ARI : 0.466329330294151

	precision	recall	f1-score	support
Kitchen	0.99	0.51	0.67	150
Decor	0.60	0.91	0.72	150
Beauty	0.96	0.64	0.77	150
Furniture	0.84	0.41	0.55	150
Baby	0.52	0.69	0.59	150
Watches	0.98	0.85	0.91	150
Computers	0.58	0.94	0.72	150
accuracy			0.71	1050
macro avg	0.78	0.71	0.70	1050
weighted avg	0.78	0.71	0.70	1050

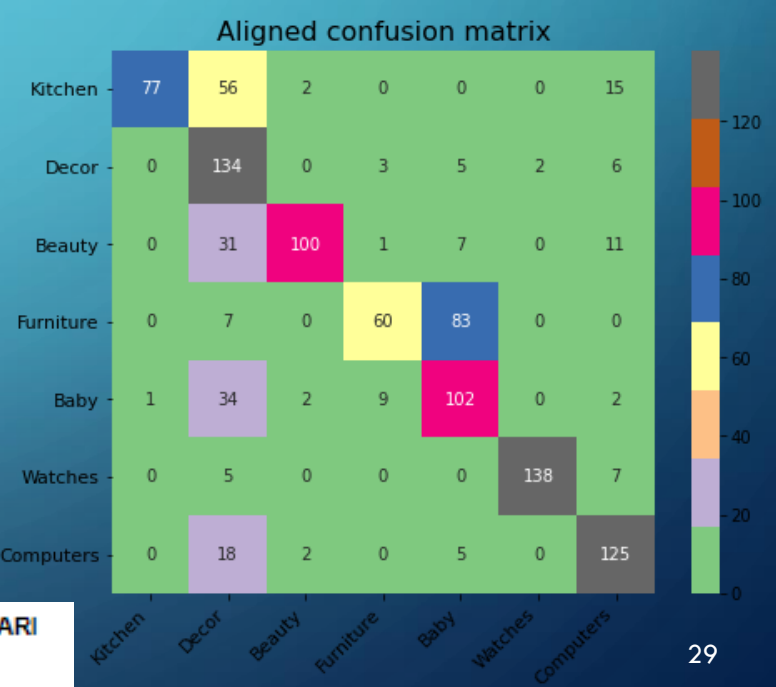


RESNET50

Category	cluster	effectives
Kitchen	5	78
Decor	3	285
Beauty	4	106
Furniture	6	73
Baby	1	202
Watches	0	140
Computers	2	166

ARI : 0.4613154559389895

	precision	recall	f1-score	support
Kitchen	0.99	0.51	0.68	150
Decor	0.47	0.89	0.62	150
Beauty	0.94	0.67	0.78	150
Furniture	0.82	0.40	0.54	150
Baby	0.50	0.68	0.58	150
Watches	0.99	0.92	0.95	150
Computers	0.75	0.83	0.79	150
accuracy			0.70	1050
macro avg	0.78	0.70	0.70	1050
weighted avg	0.78	0.70	0.70	1050

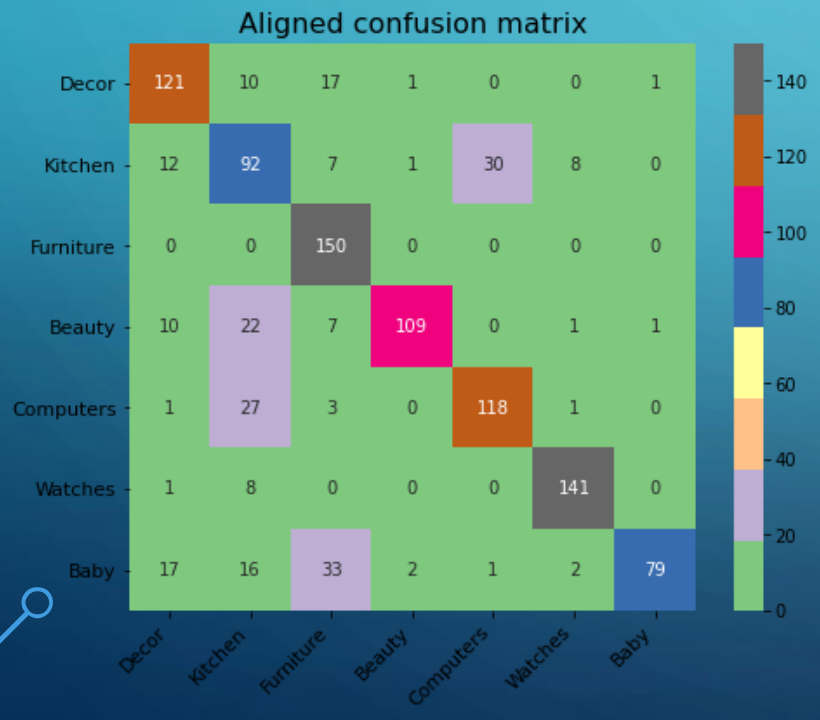


title	precision	recall	f1_score	ARI
features extracted by VGG16	78.06%	70.57%	70.38%	46.63%
features extracted by ResNet50	78.09%	70.10%	70.48%	46.13%

Optimisation des hypers paramètres de countvectorizer, tf-idf et Umap

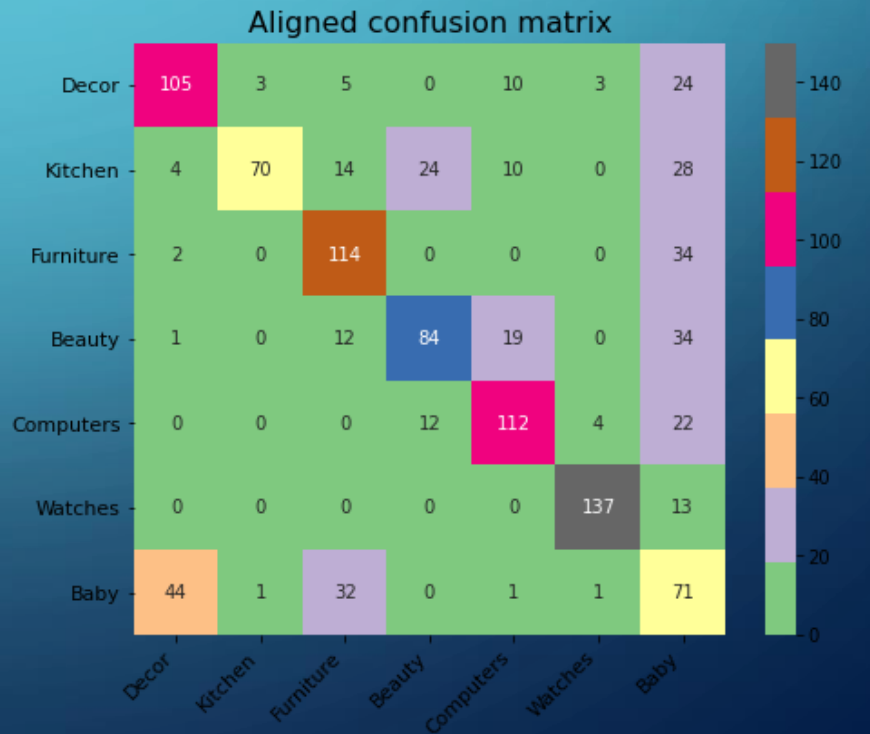
product_name

title	precision	recall	f1_score	ARI
product_name prep pipeline(BoW + TFiDF)	81.69%	66.48%	68.97%	34.41%
product_name prep pipeline(BoW + TFiDF) optimisé	81.69%	66.48%	68.97%	34.41%
product_name prep pipeline(BoW + TFiDF +Umap2)	71.78%	68.19%	68.08%	47.27%
product_name prep pipeline(BoW + TFiDF +Umap) optimisé	80.25%	77.14%	77.03%	55.87%



description

title	precision	recall	f1_score	ARI
description prep pipeline(BoW + TFiDF)	53.88%	47.33%	45.92%	23.33%
description prep pipeline(BoW + TFiDF) optimisé	50.22%	52.29%	49.55%	28.82%
description prep pipeline(BoW + TFiDF +Umap2)	60.20%	58.95%	54.86%	36.00%
description prep pipeline(BoW + TFiDF +Umap) optimisé	70.84%	66.00%	66.84%	40.42%



Essais avec combinaisons des différentes features extraites de product_name – description – images

Méthodologie:

- Séparation des données en jeux d'entrainement et de test (classes é quilibrées)
- Entrainement de l'algorithme de partitionnement sur train
- Prédictions sur test

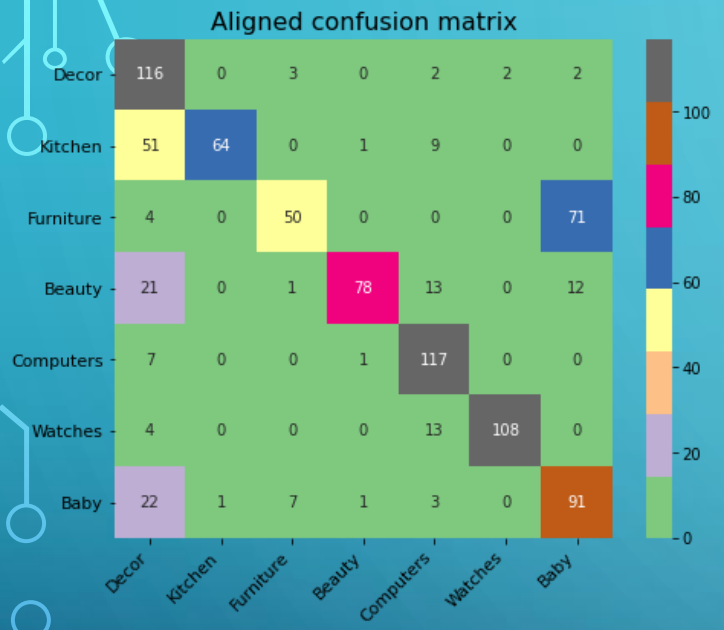
Entrainement 875 items ,125 par catégorie

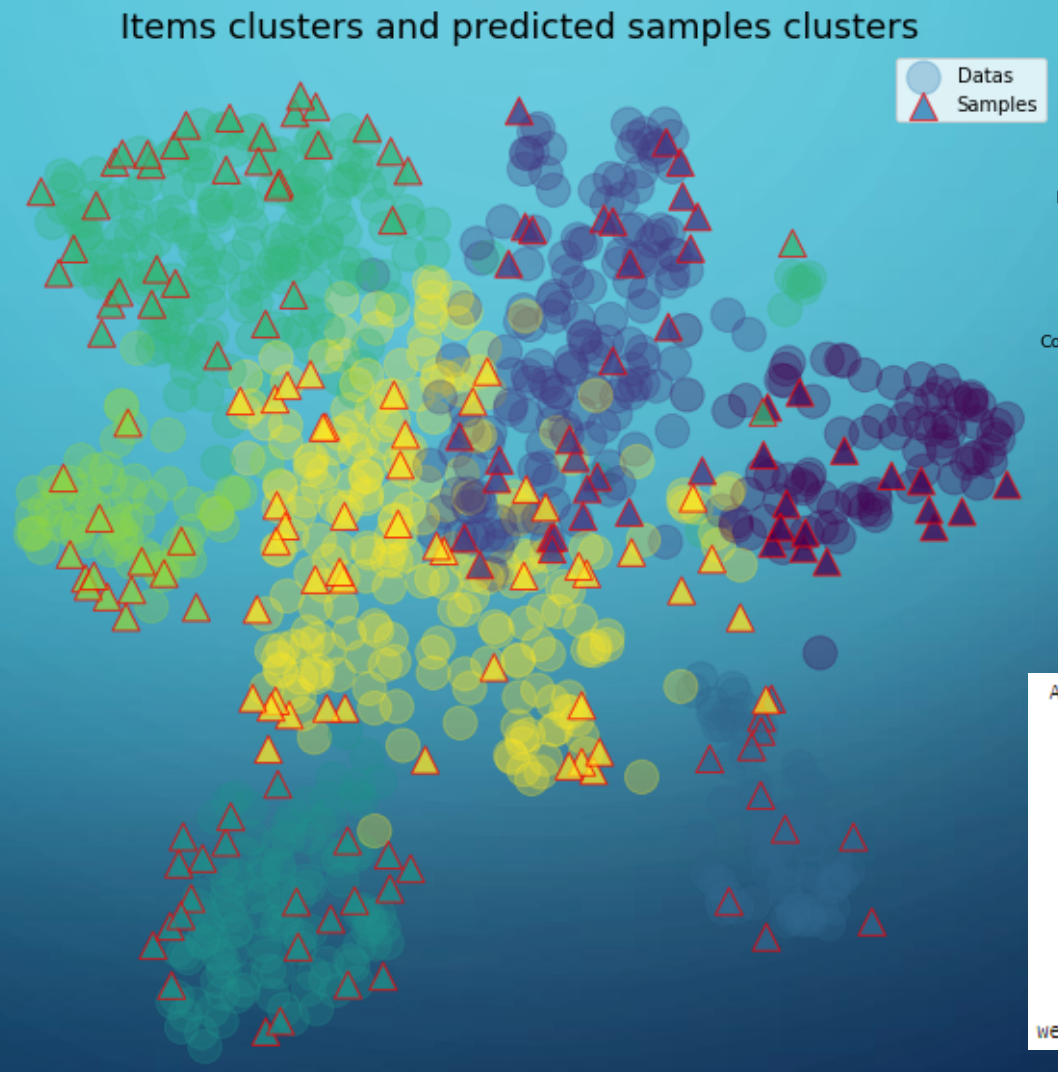
title	precision	recall	f1_score	ARI
BoW + BoVW_SIFT	80.88%	58.29%	60.15%	24.60%
BoW + BoVW_ORB	82.79%	66.51%	69.41%	32.92%
BoW + VGG16	78.43%	70.29%	70.18%	46.48%
BoW + RESNET50	77.87%	70.86%	71.15%	47.75%
Umap + BoVW_SIFT	76.90%	72.11%	72.24%	46.76%
Umap + BoVW_ORB	76.90%	72.11%	72.24%	46.76%
Umap + VGG16	78.43%	70.29%	70.18%	46.48%
Umap + RESNET50	78.46%	70.86%	70.97%	48.18%
BERT + BoVW_SIFT	24.88%	35.20%	28.14%	10.58%
BERT + BoVW_ORB	24.88%	35.20%	28.14%	10.58%
BERT + VGG16	78.54%	69.71%	69.86%	45.36%
BERT + RESNET50	72.48%	64.57%	66.12%	41.08%
All extracted features	78.88%	70.51%	70.77%	46.69%
All features pca optimisé	78.96%	71.31%	71.21%	48.73%

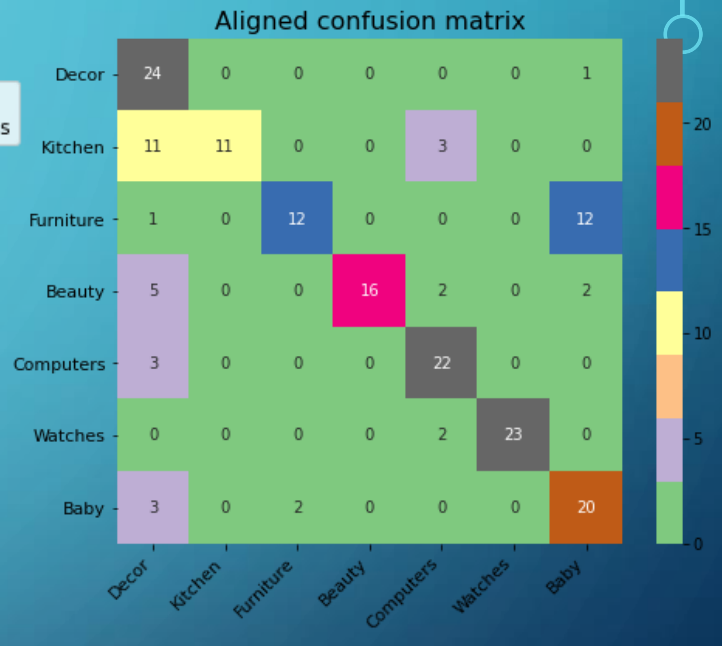
Validation 175 items, 25 par catégorie

title	precision	recall	f1_score	ARI
BoW + BoVW_SIFT	79.83%	56.00%	57.13%	23.54%
BoW + BoVW_ORB	82.90%	64.57%	67.55%	29.66%
BoW + VGG16	81.15%	73.14%	72.89%	49.03%
BoW + RESNET50	80.09%	72.57%	72.89%	47.47%
Umap + BoVW_SIFT	79.64%	76.00%	75.40%	53.17%
Umap + BoVW_ORB	79.64%	76.00%	75.40%	53.17%
Umap + VGG16	81.15%	73.14%	72.89%	49.03%
Umap + RESNET50	79.71%	72.00%	72.14%	46.94%
BERT + BoVW_SIFT	19.90%	34.86%	25.01%	9.54%
BERT + BoVW_ORB	19.90%	34.86%	25.01%	9.54%
BERT + VGG16	81.45%	72.00%	72.36%	46.88%
BERT + RESNET50	77.05%	68.00%	69.33%	42.59%
All extracted features	81.90%	72.57%	72.91%	48.18%
All features pca optimisé	81.40%	73.14%	73.05%	49.01%

Sur toutes les features extraites avec optimisation des paramètres de la réduction et du clusterer







ARI : 0.4872624096158117

	precision	recall	f1-score	support
Decor	0.52	0.93	0.66	125
Kitchen	0.98	0.51	0.67	125
Furniture	0.82	0.40	0.54	125
Beauty	0.96	0.62	0.76	125
Computers	0.75	0.94	0.83	125
Watches	0.98	0.86	0.92	125
Baby	0.52	0.73	0.60	125
accuracy			0.71	875
macro avg	0.79	0.71	0.71	875
weighted avg	0.79	0.71	0.71	875

ARI : 0.49014464902485266

	precision	recall	f1-score	support
Decor	0.51	0.96	0.67	25
Kitchen	1.00	0.44	0.61	25
Furniture	0.86	0.48	0.62	25
Beauty	1.00	0.64	0.78	25
Computers	0.76	0.88	0.81	25
Watches	1.00	0.92	0.96	25
Baby	0.57	0.80	0.67	25
accuracy			0.73	175
macro avg	0.81	0.73	0.73	175
weighted avg	0.81	0.73	0.73	175

CONCLUSION

- On arrive à segmenter de manière non supervisée les produit présents sur le site avec un bon taux de réussite sur l'échantillon donné.
- On peut donc automatiser l'affectation à une catégorie pour les objets proposés

- Améliorations:
  - Plus de données
  - Mise en place d'un modèle de classification à partir de toutes, ou partie, des features extraites des variables product_name, description et des images

- fin