CLASSIFIER AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

MARWA EL HOURI

PROBLÉMATIQUE

- Objectif : Automatisation de l'attribution d'une catégorie aux produits a partir de leur description et/ou image
- Etude de faisabilité d'un moteur de classification des articles en différentes catégories
  - Prétraitement des données textes et images
  - Reduction de dimension
  - clustering

PLAN

- Présentation du jeu de données et détermination des catégories
- Prétraitement et extraction des features textes
- Prétraitement t extraction des features images
- Conclusion

PRÉSENTATION DU JEU DE DONNÉES

- Jeu de données de 1050 articles
- Taille (1050, 15)

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   object
 2   product_url             1050 non-null   object
 3   product_name            1050 non-null   object
 4   product_category_tree   1050 non-null   object
 5   pid                     1050 non-null   object
 6   retail_price            1049 non-null   float64
 7   discounted_price        1049 non-null   float64
 8   image                   1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null   bool
10   description             1050 non-null   object
11   product_rating          1050 non-null   object
12   overall_rating          1050 non-null   object
13   brand                   712 non-null    object
14   product_specifications  1049 non-null   object
dtypes: bool(1), float64(2), object(12)
```

PRÉSENTATION DU JEU DE DONNÉES

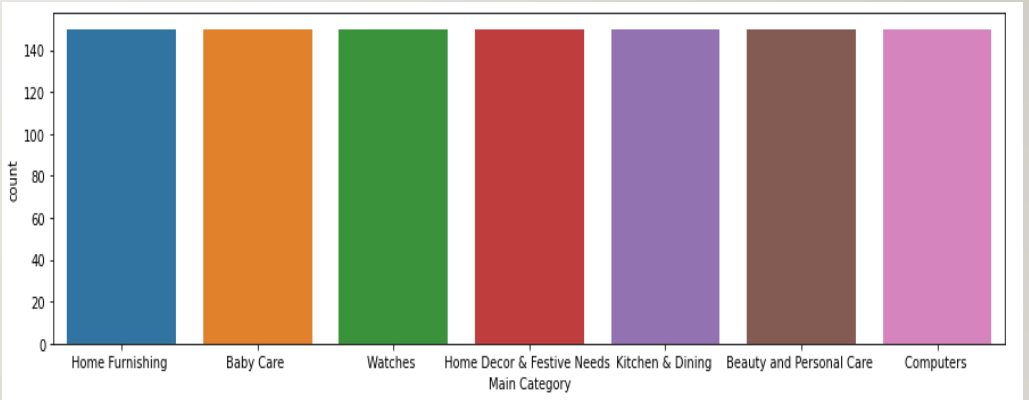
- Variables pertinentes
  - **product_category_tree** : Arbre des catégories et sous catégories des articles
  - **product_name** : nom du produit
  - **description** : description du produit
  - **image** : le nom du fichier image

	product_name	product_category_tree	description	image
0	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >>...	Key Features of Elegance Polyester Multicolor ...	55b85ea15a1536d46b7190ad6fff8ce7.jpg
1	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	Specifications of Sathiyas Cotton Bath Towel (...	7b72c92c2f6c40268628ec5f14c6d590.jpg
2	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	Key Features of Eurospa Cotton Terry Face Towe...	64d5d4a258243731dc7bbb1eef49ad74.jpg
3	SANTOSH ROYAL FASHION Cotton Printed King size...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	Key Features of SANTOSH ROYAL FASHION Cotton P...	d4684dcdc759dd9cdf41504698d737d8.jpg
4	Jaipur Print Cotton Floral King sized Double B...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	Key Features of Jaipur Print Cotton Floral Kin...	6325b6870c54cd47be6ebfbffa620ec7.jpg

EXPLORATION DES CATÉGORIES

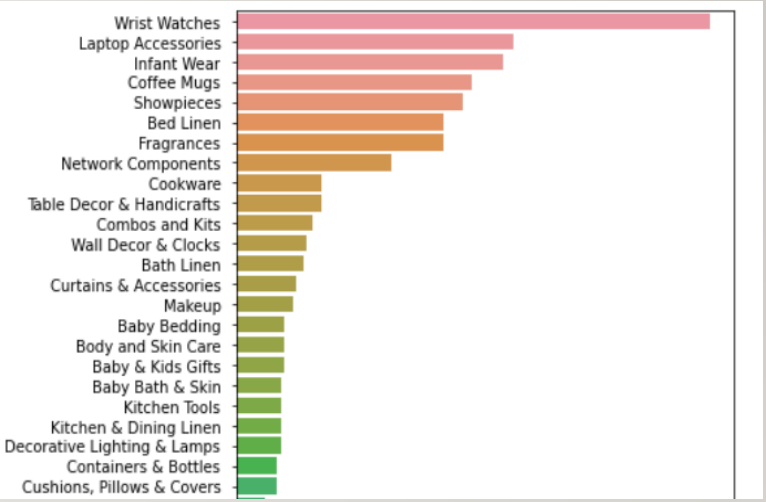
- 642 arbre de catégories uniques
  - Récupérer les catégories principales et les sous catégories
- Résultat : Distribution équilibrée des articles sur 7 catégories principales

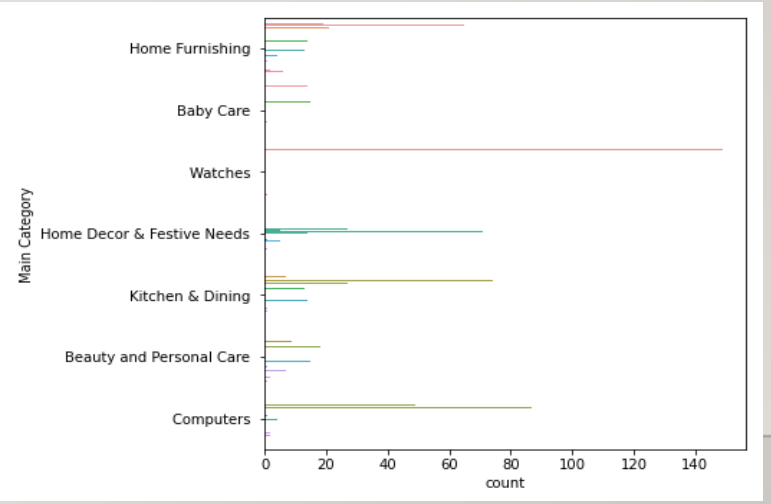
Home Furnishing	150
Baby Care	150
Watches	150
Home Decor & Festive Needs	150
Kitchen & Dining	150
Beauty and Personal Care	150
Computers	150



EXPLORATION DES CATÉGORIES

- Distribution des articles par sous-catégories
  - Déséquilibre par sous-catégorie
  - Déséquilibre dans la distribution des sous-catégories par catégorie principale
- Conclusion : Nous étudierons la faisabilité de la classification sur les catégories principales





EXPLORATION DES CATÉGORIES

- Label encoder pour la numérisation des Label afin de pouvoir calculer les ARI scores

Category	Target
Baby Care	0
Beauty and Personal Care	1
Computers	2
Home Decor & Festive Needs	3
Home Furnishing	4
Kitchen & Dining	5
Watches	6

PRÉTRAITEMENT ET EXTRACTION DES FEATURES TEXTES

- Préparation du texte
- Extraction des features
- Résultats et conclusion

PRÉPARATION DU TEXTE

- Préparation du texte
  1. Concaténation des variables `product_name` et `description`
  2. Nettoyage et tokenisation
     - Transformer le text en minuscule et enlever les espaces : `doc.lower().strip()`
     - Tokeniser
     - Supprimer les stopwords
     - Supprimer les ponctuations
     - Supprimer les mots d'une lettre
  3. Lemmatisation ou stemming et join

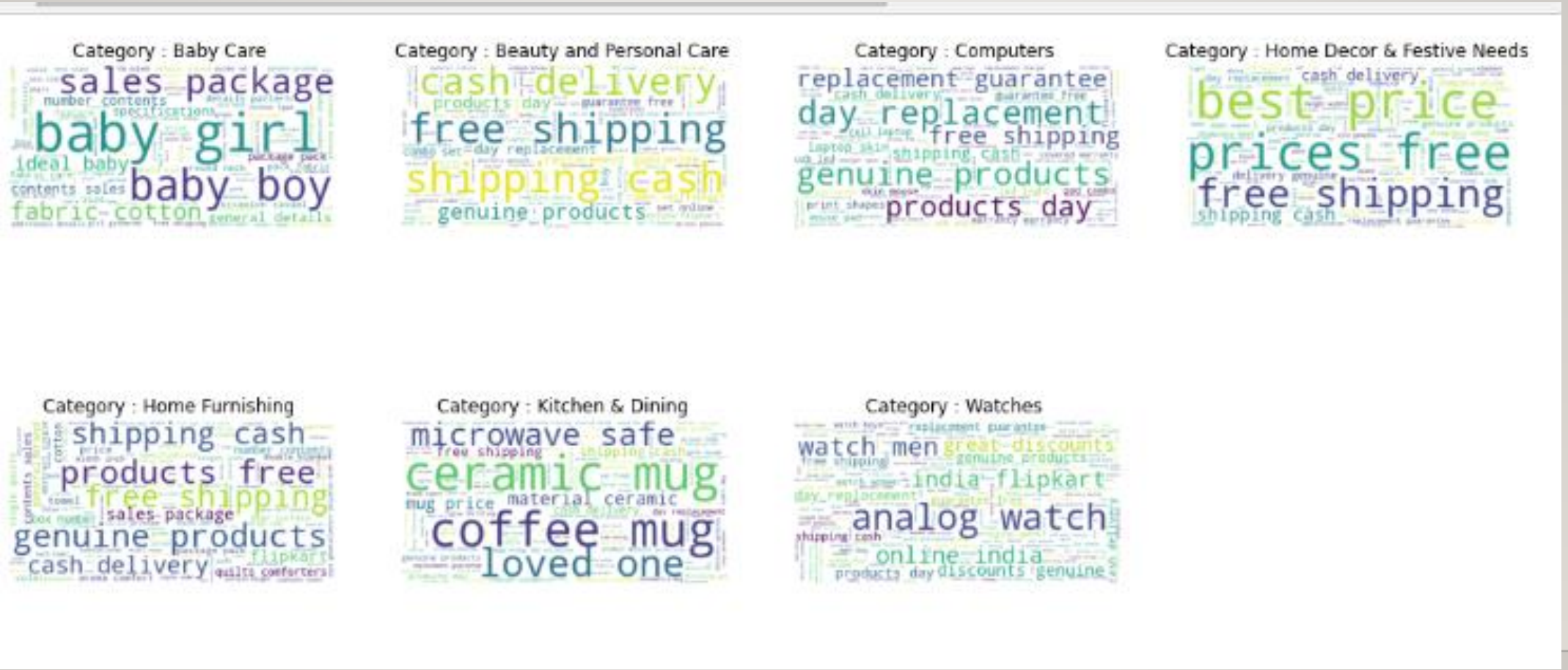
- Librairie : `nltk`
  - Lemmatisation/stemmer : `WordNetLemmatizer`, `PorterStemmer`
  - Tokeniser : `RegexpTokenizer`
  - Corpus : `stopwords`

PRÉPARATION DU TEXTE – NETTOYAGE 1

- Mots uniques : 6194
- Nombre total de mots : 61 149
- Longueur maximale du texte avant prétraitement : 643
- Longueur maximale du texte après prétraitement 365

tokenize_1	length_Text	length_tokenize_1	tokenize_1_lem	tokenize_1_stem	tokenize_1_dl
[elegance, polyester, multicolor, abstract, ey...	253	158	elegance polyester multicolor abstract eyelet ...	eleg polyest multicolor abstract eyelet door c...	elegance polyester multicolor abstract eyelet ...
[sathiyas, cotton, bath, towelspecifications, ...	87	65	sathiyas cotton bath towelspecifications sathi...	sathiya cotton bath towelspecif sathiya cotton...	sathiyas cotton bath towelspecifications sathi...
[eurospa, cotton, terry, face, towel, setkey, ...	257	159	eurospa cotton terry face towel setkey feature...	eurospa cotton terri face towel setkey featur ...	eurospa cotton terry face towel setkey feature...
[santosh, royal, fashion, cotton, printed, kin...	159	120	santosh royal fashion cotton printed king size...	santosh royal fashion cotton print king size d...	santosh royal fashion cotton printed king size...
[jaipur, print, cotton, floral, king, sized, d...	238	157	jaipur print cotton floral king sized double b...	jaipur print cotton floral king size doubl bed...	jaipur print cotton floral king sized double b...

WORDCLOUD PAR CATÉGORIES



FRÉQUENCE DES MOTS DANS LE CORPUS

- Conclusion : Exploration du corpus sans les mots fréquents

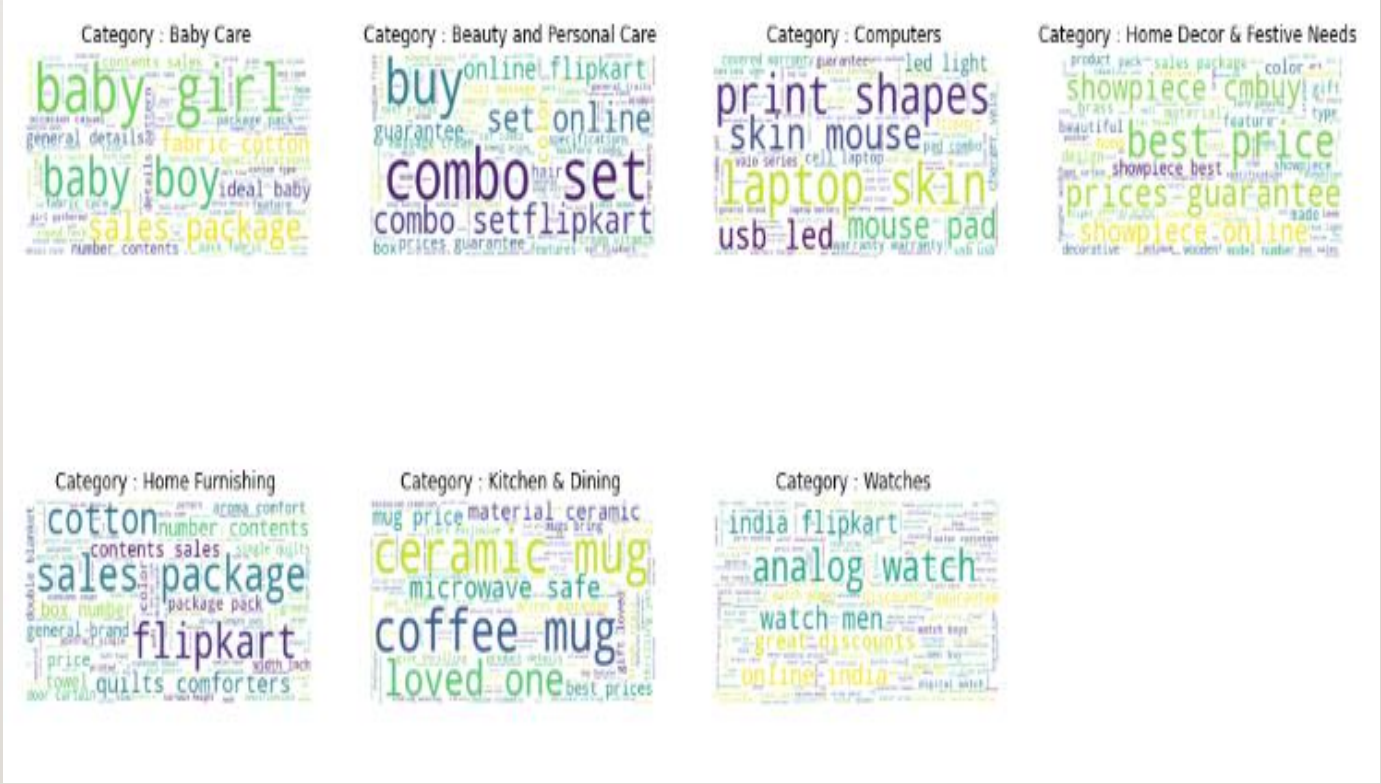
```
mots_frequents=freq[freq>500]
mots_frequents
```

```
day            543
genuine        564
shipping       564
cash           564
delivery       567
replacement    568
free           622
products       633
dtype: int64
```

	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
free	46	111	112	83	77	59	134
products	39	128	95	92	77	68	134
genuine	34	101	94	76	74	51	134
shipping	34	101	94	76	74	51	134
cash	34	101	94	76	74	51	134
delivery	34	103	94	77	74	51	134
day	21	101	96	81	8	100	136
replacement	14	105	185	76	3	51	134

PRÉPARATION DU TEXTE – NETTOYAGE 2

- Tokenize_2 : enlever les mots fréquents du corpus
- Résultat :
  - Mots uniques : 6186
  - Nombre total de mots : 56 524
  - Longueur maximale du texte après prétraitement : 365
  - Longueur maximale du texte après prétraitement sans mots fréquents : 362



EXPLORATION DES DOUBLONS DANS LES CATÉGORIES

- Doublons : Les mots qui apparaissent dans plusieurs catégories
- Conclusion : Exploration du corpus sans les doublons

	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
cotton	210	NaN	NaN	6.0	138.0	NaN	NaN
com	32	159.0	68.0	NaN	74.0	6.0	134.0
online	26	83.0	26.0	78.0	NaN	49.0	134.0
flipkart	25	92.0	68.0	2.0	74.0	6.0	134.0
design	21	1.0	12.0	31.0	83.0	102.0	2.0
guarantee	14	101.0	96.0	76.0	NaN	50.0	134.0
skin	14	71.0	83.0	NaN	10.0	NaN	NaN
buy	8	79.0	4.0	6.0	4.0	9.0	134.0

PRÉPARATION DU TEXTE – NETTOYAGE 2

- Tokenize_3 : enlever les doublons du corpus
- Résultat :
  - Mots uniques : 6178
  - Nombre total de mots : 53 755
  - Longueur maximale du texte après prétraitement : 365
  - Longueur maximale du texte après prétraitement sans doublons : 360

PRÉTRAITEMENT DU TEXTE - SYNTHÈSE

- 3 corpus
  - Corpus1 : Tous les mots (sans stop words et ponctuation)
  - Corpus 2 : Sans les mots fréquents
  - Corpus 3 : Sans les doublons
- Pour chaque corpus
  - Stemming : `PorterStemmer`
  - Lemmatisation : `WordNetLemmatizer`
  - Join sans stemming ou lemmatisation pour le deep learning

TRAITEMENT DU TEXTE

- Préparation commune des traitement
- Méthodes de traitement
- Résultats et conclusion

PRÉPARATION COMMUNE DES TRAITEMENT

- Reduction de dimension :
  - PCA (en préservant 99% de la variance)
  - TSNE (en variant la perplexité)
- Clustering :
  - Kmeans (sur 7 clusters)
- Calcul de score
  - Silhouette score
  - ARI score
- Représentation graphique (catégories réelles vs clusters)
- Matrice de confusion
  - Correspondance des clusters
  - Rapport de classification (precision, recall et f1-score par catégorie)

MÉTHODES DE TRAITEMENT DE TEXTE

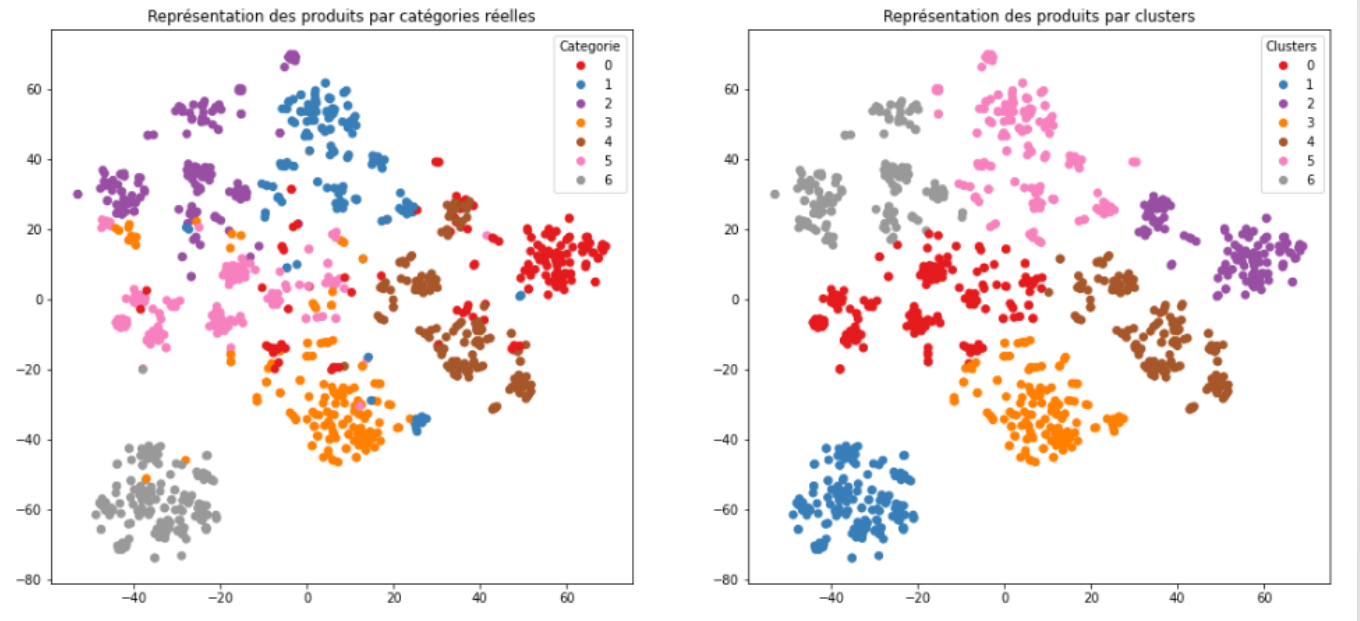
- Méthodes vectorielles
  - Bag of words
  - TF-IDF
  - Word2vec
- Résultats
  - Traitement sur les trois corpus
  - Avec lemmatisation ou avec stemming
  - Avec ou sans réduction de dimension

- Méthodes deep learning
  - BERT (HuggingFace et tensorflow)
  - USE (Universal sentence encoder)
- Résultats
  - Traitement sur les trois corpus
  - Avec ou sans réduction de dimension

MEILLEURS RÉSULTATS PAR MÉTHODE

Méthode	Meilleur corpus	Stem/Lem	PCA	ARI score	Silhouette score	Time(s)
TF-IDF	Corpus 2	Stem	Oui	0.6864	0.4843	38
USE	Corpus 3	-	Oui	0.6425	0.4815	33
Word2Vec	Corpus 3	Lem	Oui	0.5784	0.53171	24
BERT – Tensorflow	Corpus 1	-	Non	0.5462	0.4838	29
BERT - HuggingFace	Corpus 2	-	Non	0.5075	0.5079	25
Bag-of-words	Corpus 2	Lem	Non	0.45	0.44	42

I - TF-IDF

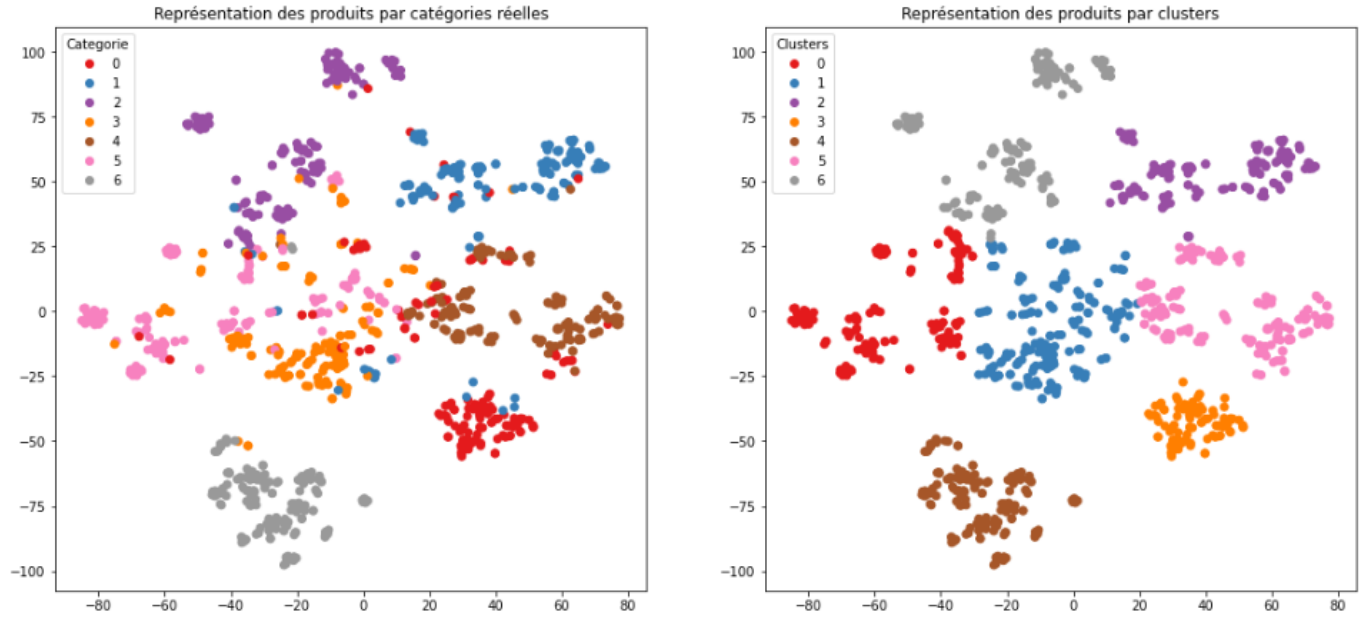



```
Dimensions dataset avant réduction PCA :  (1050, 5673)
Dimensions dataset après réduction PCA :  (1050, 905)
Silhouette :  0.48435384
ARI :  0.6864 time :  38.0
```

```
Correspondance des clusters :  [5 6 0 3 4 1 2]
[[103   9   0   6  15  17   0]
 [  2 128   5  11   1   3   0]
 [  0  17 126   0   0   7   0]
 [  0   4  11 117   1  15   2]
 [ 19   0   0   1 130   0   0]
 [  1   3   8   2   0 136   0]
 [  0   0   0   0   0   1 149]]
```

	precision	recall	f1-score	support
0	0.82	0.69	0.75	150
1	0.80	0.85	0.82	150
2	0.84	0.84	0.84	150
3	0.85	0.78	0.82	150
4	0.88	0.87	0.88	150
5	0.76	0.91	0.83	150
6	0.99	0.99	0.99	150
accuracy			0.85	1050
macro avg	0.85	0.85	0.85	1050
weighted avg	0.85	0.85	0.85	1050

2- USE

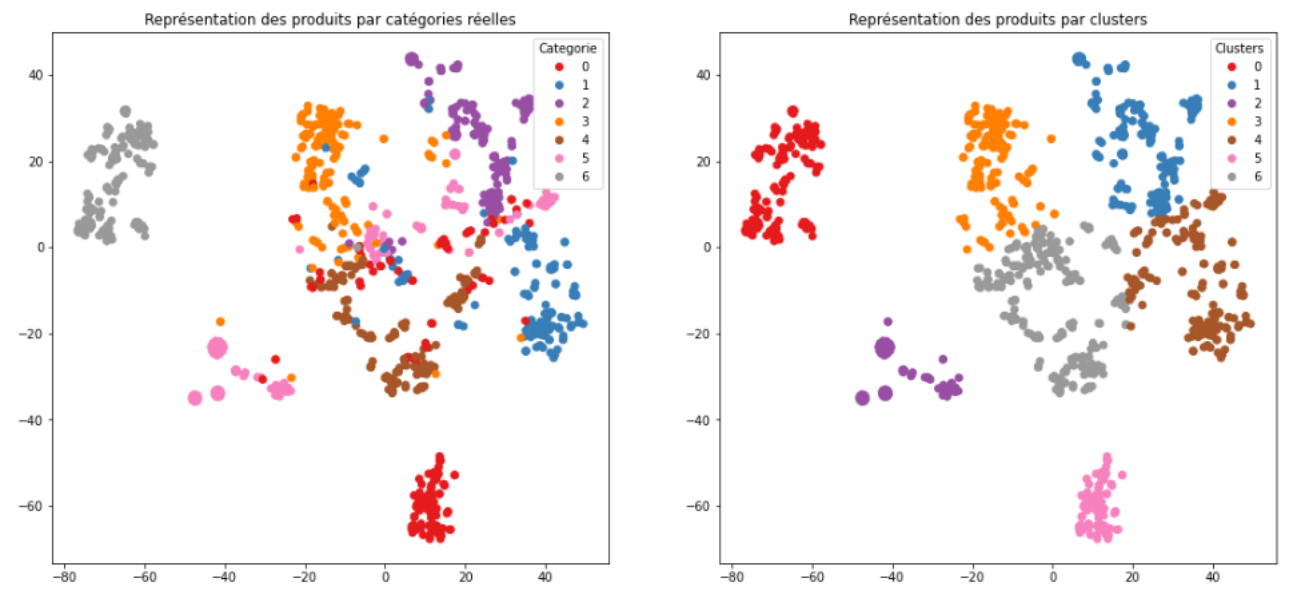



```
Dimensions dataset avant réduction PCA :  (1050, 512)
Dimensions dataset après réduction PCA :  (1050, 361)
Silhouette :  0.48147297
ARI :  0.6425 time :  33.0
```

```
Correspondance des clusters :  [5 3 1 0 6 4 2]
[[ 88   7   1  24  27   3   0]
 [  5 129   2  11   1   2   0]
 [  0   0 136   2   0  12   0]
 [  0   1   9 108   1  28   3]
 [  0   1   0   8 141   0   0]
 [  0   0   7  37   0 106   0]
 [  0   0   0   2   0   0 148]]
```

	precision	recall	f1-score	support
0	0.95	0.59	0.72	150
1	0.93	0.86	0.90	150
2	0.88	0.91	0.89	150
3	0.56	0.72	0.63	150
4	0.83	0.94	0.88	150
5	0.70	0.71	0.70	150
6	0.98	0.99	0.98	150
accuracy			0.82	1050
macro avg	0.83	0.82	0.82	1050
weighted avg	0.83	0.82	0.82	1050

3 - WORD2VEC

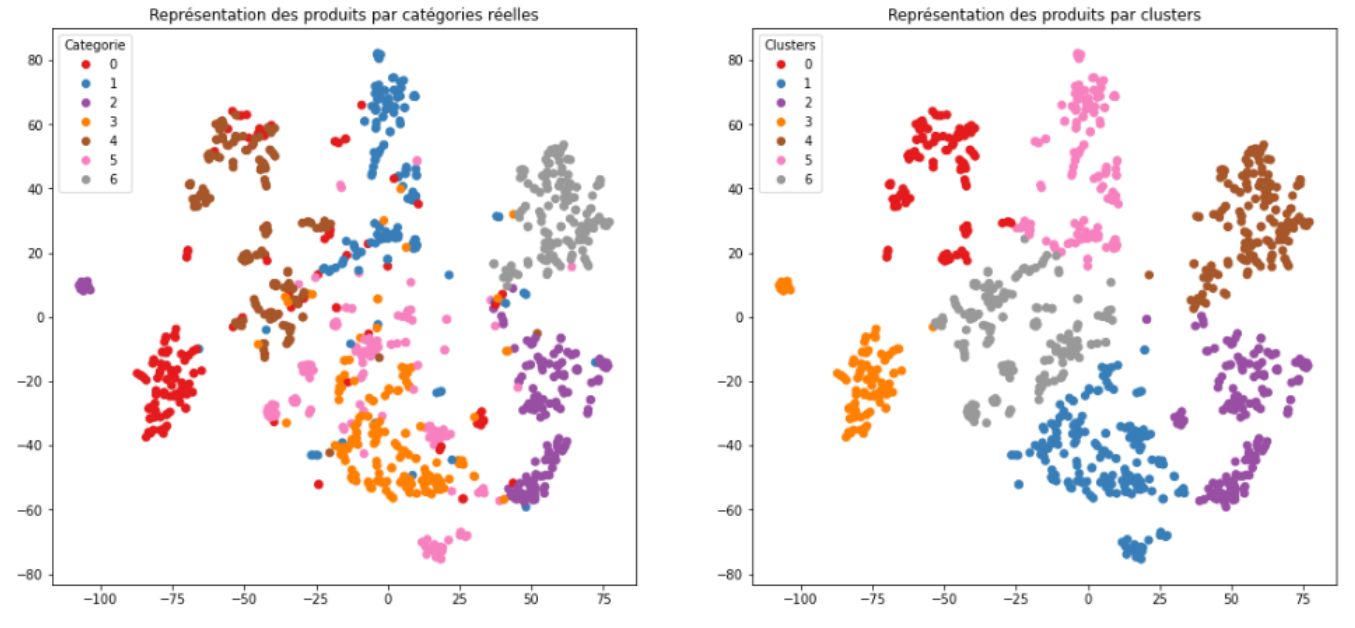



```
Dimensions dataset avant réduction PCA :  (1050, 300)
Dimensions dataset après réduction PCA :  (1050, 51)
Silhouette :  0.53171396
ARI :  0.5784 time :  24.0
```

```
Correspondance des clusters :  [6 2 5 3 1 0 4]
[[ 84  22   2   3  37   2   0]
 [  0 120   4   9  17   0   0]
 [  0   2 143   0   5   0   0]
 [  0   3   9 124  10   4   0]
 [  0  16   0   1 133   0   0]
 [  0  23  22   5  26  74   0]
 [  0   0   0   0   1   0 149]]
```

	precision	recall	f1-score	support
0	1.00	0.56	0.72	150
1	0.65	0.80	0.71	150
2	0.79	0.95	0.87	150
3	0.87	0.83	0.85	150
4	0.58	0.89	0.70	150
5	0.93	0.49	0.64	150
6	1.00	0.99	1.00	150
accuracy			0.79	1050
macro avg	0.83	0.79	0.78	1050
weighted avg	0.83	0.79	0.78	1050

4 - BERT - TENSORFLOW

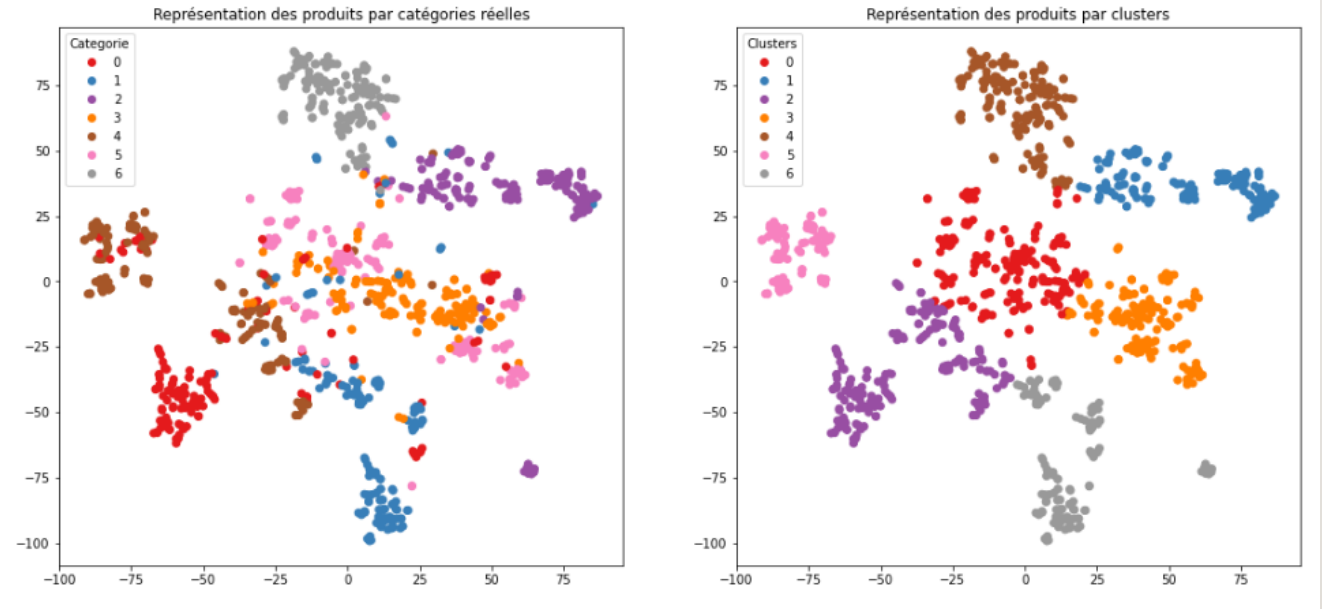



```
Silhouette :  0.48381656
ARI :  0.5462 time :  29.0
```

```
Correspondance des clusters :  [4 3 2 0 6 1 5]
[[ 85  13   9   6  19  16   2]
 [  1 115   2  11   0  13   8]
 [ 11   0 137   0   0   0   2]
 [  0   3   4 120   0  21   2]
 [  0   7   1   1  94  47   0]
 [  0   4   5  57   0  82   2]
 [  0   0   0   0   0   0 150]]
```

	precision	recall	f1-score	support
0	0.88	0.57	0.69	150
1	0.81	0.77	0.79	150
2	0.87	0.91	0.89	150
3	0.62	0.80	0.70	150
4	0.83	0.63	0.71	150
5	0.46	0.55	0.50	150
6	0.90	1.00	0.95	150
accuracy			0.75	1050
macro avg	0.77	0.75	0.75	1050
weighted avg	0.77	0.75	0.75	1050

5 - BERT - HUGGINGFACE

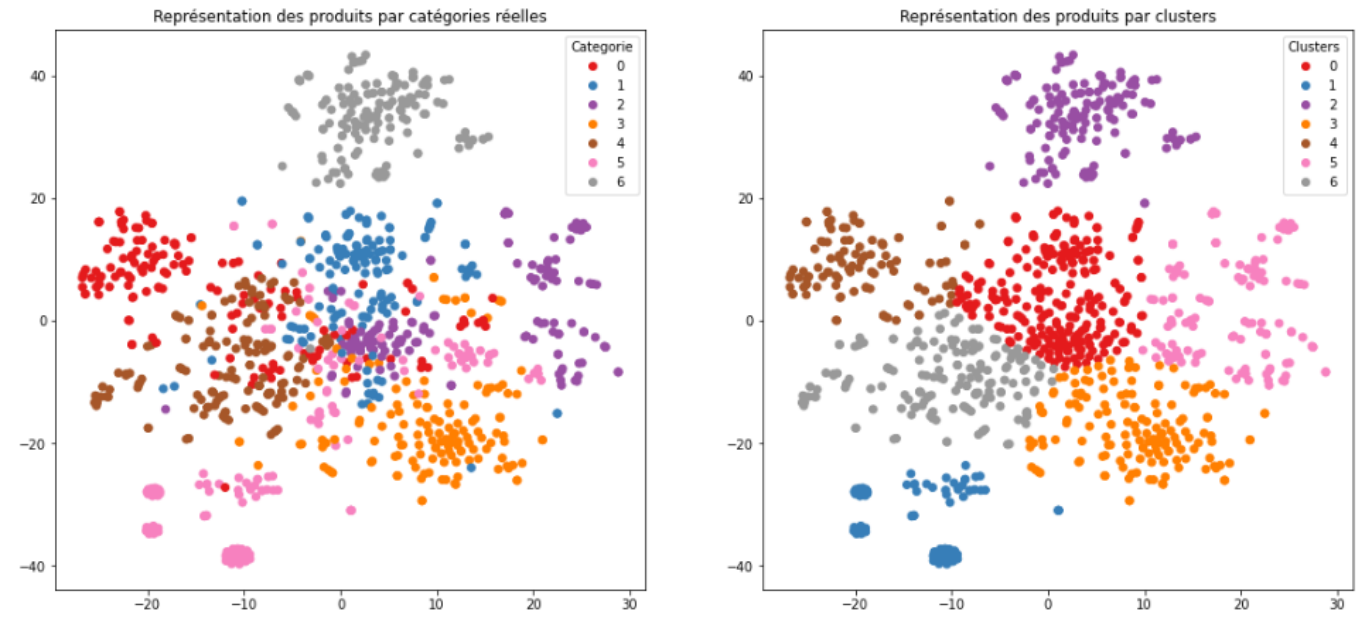



```
Silhouette :  0.50786966
ARI :  0.5075 time :  25.0
```

```
Correspondance des clusters :  [5 2 0 3 6 4 1]
[[102   9   0  13  14  10   2]
 [ 18 107   2   4   0  12   7]
 [  0  11 131   5   0   0   3]
 [  2   3   0  76   0  67   2]
 [ 66   0   1   1  74   8   0]
 [  2   2   0  50   0  94   2]
 [  0   0   0   0   0   1 149]]
```

	precision	recall	f1-score	support
0	0.54	0.68	0.60	150
1	0.81	0.71	0.76	150
2	0.98	0.87	0.92	150
3	0.51	0.51	0.51	150
4	0.84	0.49	0.62	150
5	0.49	0.63	0.55	150
6	0.90	0.99	0.95	150
accuracy			0.70	1050
macro avg	0.72	0.70	0.70	1050
weighted avg	0.72	0.70	0.70	1050

6 - BAG-OF-WORDS




```
Silhouette :  0.44255793
ARI :  0.4591 time :  42.0
```

```
Correspondance des clusters :  [1 5 6 3 0 2 4]
[[ 87  27   8   6  21   1   0]
 [  6 114  10  13   5   0   2]
 [  0  62  77   8   3   0   0]
 [  1  10  11 117  10   1   0]
 [ 12  21   0   0 117   0   0]
 [  4  14  27  11  15  79   0]
 [  0   0   0   0   1   0 149]]
```

	precision	recall	f1-score	support
0	0.79	0.58	0.67	150
1	0.46	0.76	0.57	150
2	0.58	0.51	0.54	150
3	0.75	0.78	0.77	150
4	0.68	0.78	0.73	150
5	0.98	0.53	0.68	150
6	0.99	0.99	0.99	150
accuracy			0.70	1050
macro avg	0.75	0.70	0.71	1050
weighted avg	0.75	0.70	0.71	1050

CONCLUSION - COMPARAISON DES MEILLEURS SCORE AVEC LES RÉSULTATS DE TF-IDF

Catégorie	Meilleure méthode	F1-score (TF-IDF score)	Precision (TF-IDF score)	Recall (TF-IDF score)
1 – Beauty and Personal Care	USE	0.90 (0.82)	0.93 (0.80)	0.86 (0.85)
2 – Computers	USE	0.89 (0.84)	0.88 (0.84)	0.91 (0.84)
3 – Home Decor & Festive Needs	W2V	0.85 (0.82)	0.87 (0.85)	0.83 (0.78)
6 - Watches	W2V	1 (0.99)	0.99 (0.99)	1 (0.99)
0 – Baby Care	TF - IDF	0,75	0.82	0.69
4 – Home Furnishing	TF-IDF	0.88	0.88	0.87
5 – Kitchen & Dining	TF-IDF	0.83	0.76	0.91

CONCLUSION

- TF-IDF sur le corpus sans les mots fréquents avec stemming donnent les meilleurs résultats pour la reconnaissance des catégories a partir des descriptions
- Différentes méthodes permettent la reconnaissance de différentes catégories
  - La catégorie 0 (Baby Care) est la plus difficile a reconnaitre (Les méthodes Word2Vec et USE donnent de très bons scores en terme de précision mais les scores de recall sont assez bas)
  - La catégorie 6 (Watches) est très bien reconnu par toutes les méthode de traitements de texte (le Word2Vec permet une reconnaissance complète de cette catégorie)

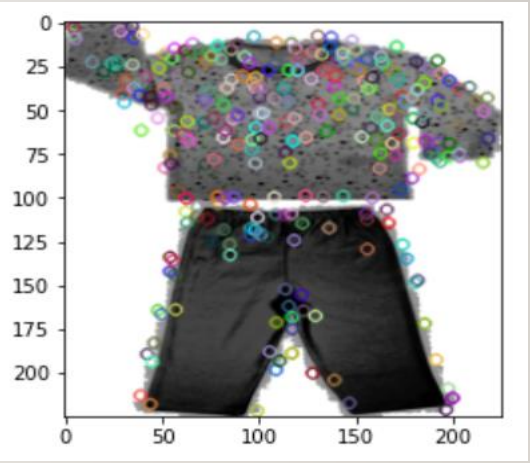
IMAGE

- Présentation des données images
- Extraction des features par SIFT
- Extraction des feature par CNN (transfer learning)
- Résultat et conclusion
- Combinaison de traitement texte et image

EXPLORATION DES IMAGES

PRÉTRAITEMENT POUR SIFT

- Prétraitement
  - GaussianBlur : Suppression du bruit de l'image
  - Histogram equalization : Amélioration du contraste dans l'image
  - Resize : Redimensionnement de la taille des images
- Génération des descripteurs par image
  - sift.detectAndCompute()



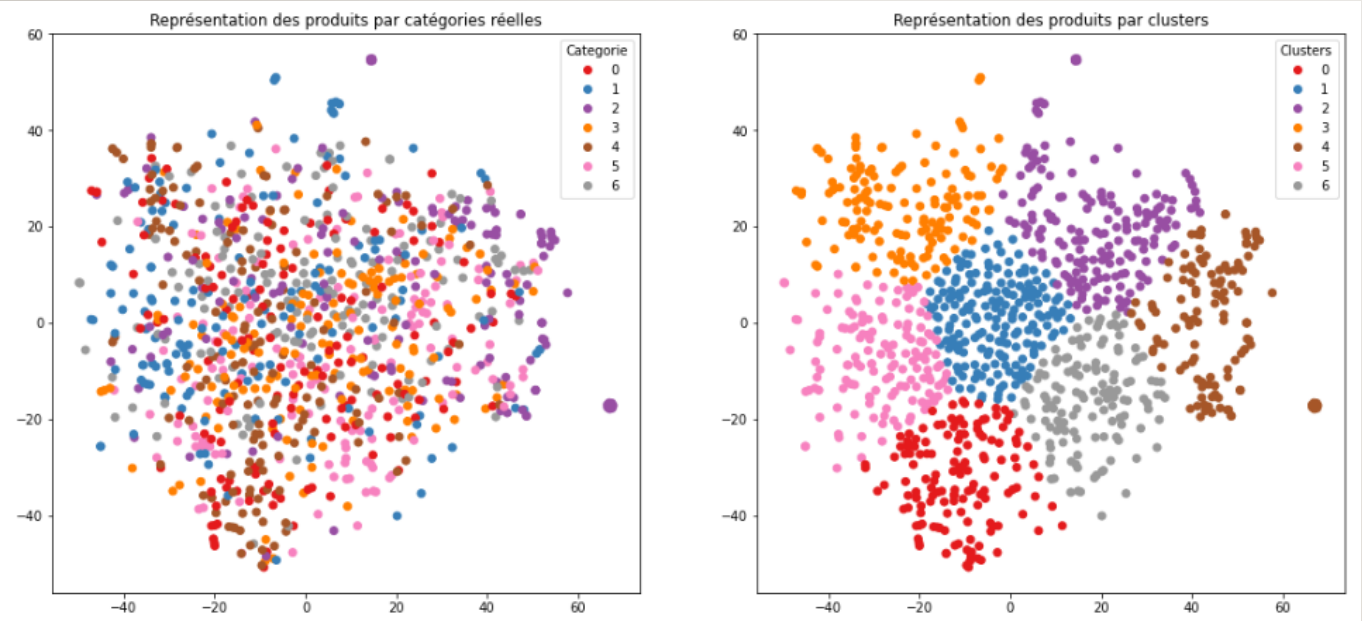

```
Descripteurs :  (303, 128)

[[  5.   4.   6. ...   0.   0.   0.]
 [  1.   0.   5. ...   0.   0.   0.]
 [ 21.  10.   3. ...   0.   0.   0.]
 ...
 [  6.  23.  30. ...   0.   1.   2.]
 [ 21.  18.  17. ...   0.   0.   0.]
 [  0.   0.  15. ...   0.  10. 138.]]
```

PRÉTRAITEMENT POUR SIFT

- Création des clusters de descripteurs
  - MiniBatchKMeans (k=sqrt(nbre de descripteurs))
- Création des features image
  - Construction de l'histogramme de chaque image a partir des cluster de ses descripteurs

RÉSULTAT EXTRACTION DES FEATURES PAR SIFT




```
Dimensions dataset avant réduction PCA :  (1050, 518)
Dimensions dataset après réduction PCA :  (1050, 425)
Silhouette :  0.3564449
ARI :  0.0415 time :  31.0
```

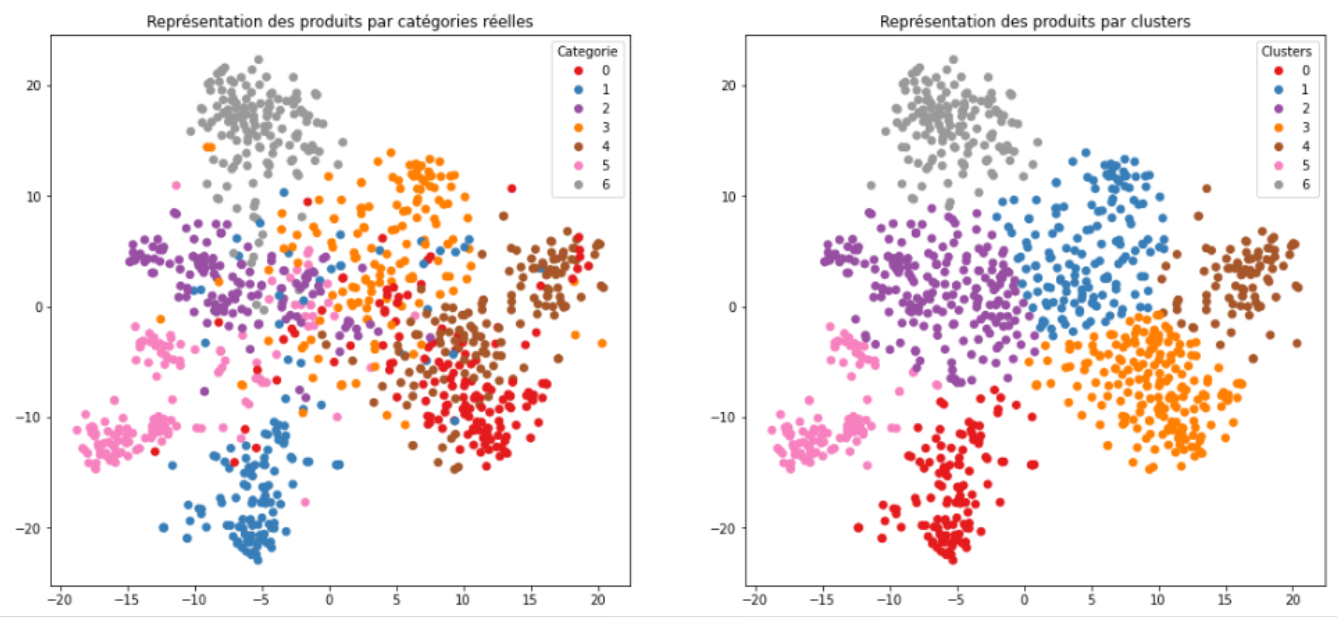
```
Correspondance des clusters :  [4 3 6 4 2 1 5]
[[ 0 13 10 29 58 24 16]
 [ 0 48  7 23 37 11 24]
 [ 0 11 56 13 26 12 32]
 [ 0 22 12 38 26 25 27]
 [ 0 15  2 23 85 13 12]
 [ 0 18 21 23 29 36 23]
 [ 0 18 15 35 31 12 39]]
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	150
1	0.33	0.32	0.33	150
2	0.46	0.37	0.41	150
3	0.21	0.25	0.23	150
4	0.29	0.57	0.38	150
5	0.27	0.24	0.25	150
6	0.23	0.26	0.24	150
accuracy			0.29	1050
macro avg	0.25	0.29	0.26	1050
weighted avg	0.25	0.29	0.26	1050

EXTRACTION DES FEATURES IMAGE PAR TRANSFER LEARNING

- Model CNN sans la couche de classification
  - `model = VGG16(weights="imagenet", input_shape=(224, 224, 3))`
- Prétraitement des images avant CNN
  - Charger les images en taille (224,224)
  - Convertir en un tableau numpy
  - Convertir en collection d'images et faire le prétraitement pour VGG16
- Regroupement des résultats et reconversion en tableau numpy

EXTRACTION DES FEATURES IMAGE PAR TRANSFER LEARNING




```
Dimensions dataset avant réduction PCA :  (1050, 4096)
Dimensions dataset après réduction PCA :  (1050, 803)
Silhouette :  0.46024776
ARI :  0.5374 time :  49.0
```

```
Correspondance des clusters :  [1 3 2 0 4 5 6]
[[111   3   7  17  11   1   0]
 [  4 118  13  12   1   1   1]
 [  1   2 134  11   0   2   0]
 [ 12   1  16 111   5   2   3]
 [ 67   0   2   5  76   0   0]
 [  1   6  36   2   0 104   1]
 [  0   0  11   1   0   0 138]]
```

	precision	recall	f1-score	support
0	0.57	0.74	0.64	150
1	0.91	0.79	0.84	150
2	0.61	0.89	0.73	150
3	0.70	0.74	0.72	150
4	0.82	0.51	0.63	150
5	0.95	0.69	0.80	150
6	0.97	0.92	0.94	150
accuracy			0.75	1050
macro avg	0.79	0.75	0.76	1050
weighted avg	0.79	0.75	0.76	1050

CONCLUSION
RÉSULTATS TEXTE ET IMAGE

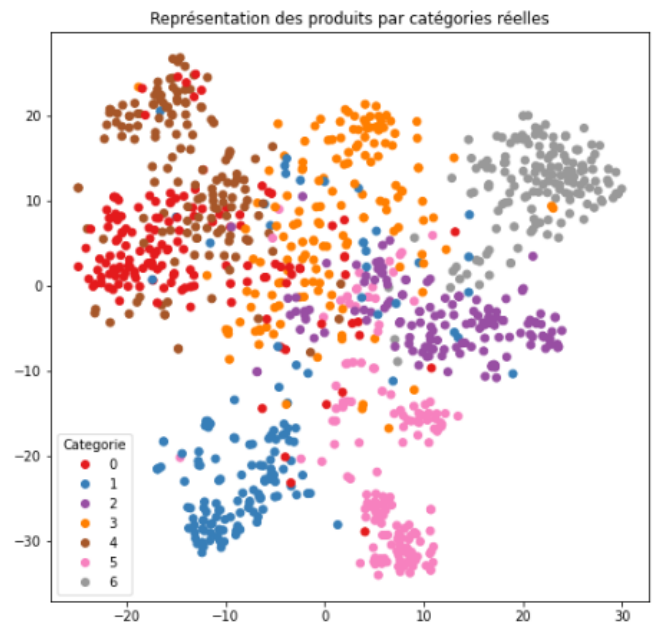
- Le traitement du texte donne de meilleurs résultats que le traitement d'image pour la reconnaissance de la catégorie de l'article.
- La méthode de transfer learning pour le traitement d'image donne des bon résultats notamment dans la reconnaissance des catégories
  - 1 (Beauty and Personal Care),
  - 5 (Kitchen & Dining)
  - 6 (Watches)
- La méthode TF-IDF sur le corpus sans les mots fréquents avec stemming donne le meilleur résultat dans le traitement des caractéristiques des articles.
  - Il permet une bonne reconnaissance de la majorité des catégories
  - Seule la catégorie 0 (Baby Care) est moyennement reconnaissable

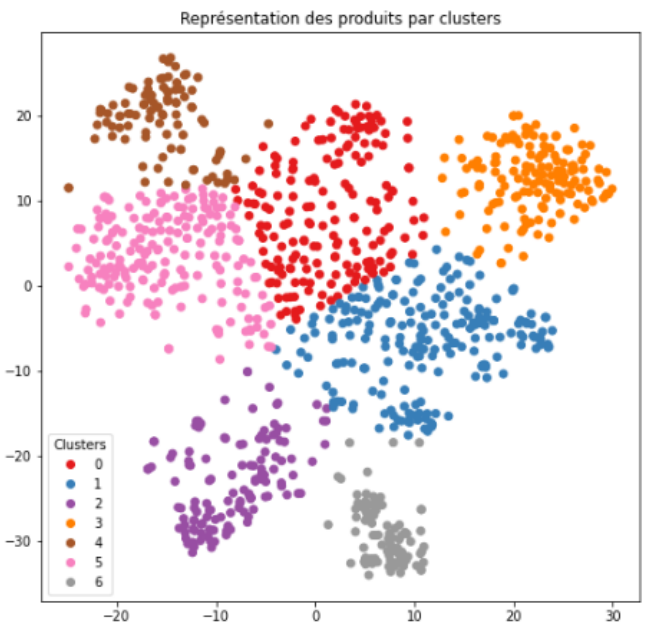
CONCLUSION
COMBINAISON DES FEATURES TEXTE ET IMAGE

- La combinaison des features texte (TF-IDF) et images (CNN) donne des résultats moins satisfaisants

Méthode	ARI score
TF-IDF (texte)	0.68
CNN (image)	0.54
TF-IDF + CNN	0.49

CONCLUSION
COMBINAISON DES FEATURES TEXTE ET IMAGE






```
Dimensions dataset avant réduction PCA :  (1050, 9220)
Dimensions dataset après réduction PCA :  (1050, 803)
Silhouette :  0.45112062
ARI :  0.4915 time :  51.0
```

```
Correspondance des clusters :  [3 2 1 6 4 0 5]
[[114   4   7  15   8   1   1]
 [  4 118  12  13   1   1   1]
 [  1   2 123  23   0   0   1]
 [ 23   1  12 109   2   0   3]
 [ 56   0   2   8  84   0   0]
 [  0   7  50  12   0  81   0]
 [  0   0  12   1   0   0 137]]
```

	precision	recall	f1-score	support
0	0.58	0.76	0.66	150
1	0.89	0.79	0.84	150
2	0.56	0.82	0.67	150
3	0.60	0.73	0.66	150
4	0.88	0.56	0.69	150
5	0.98	0.54	0.70	150
6	0.96	0.91	0.94	150
accuracy			0.73	1050
macro avg	0.78	0.73	0.73	1050
weighted avg	0.78	0.73	0.73	1050

CONCLUSION

- La combinaison des features texte et image ne permet pas d'améliorer les résultats.
- Une classification automatique supervisée est faisable mais en utilisant de préférence la description du produit avec un traitement du texte par la méthode TF-IDF.
- Il serait intéressant de tester une classification supervisée sur la combinaison des features texte et image pour tester l'effet de l'ajout des features images sur une classification supervisee.

Merci!