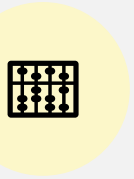
MOTEUR DE CLASSIFICATION

ETUDE DE FAISABILITE

SOMMAIRE

RAPPEL DE LA PROBLEMATIQUE



Contexte

Place de marché souhaite automatiser la catégorisation des articles proposés à la vente.

Aujourd'hui cette catégorisation est faîte manuellement.

Objectifs

Conduire une première étude sur la faisabilité d'un moteur de classification des articles en différentes catégories à partir d'une image et d'une description du produits.

Mission

Analyser le jeu de données en réalisant un prétraitement des descriptions des produits et des images, une réduction de dimension, puis un clustering.

Etudier les résultats du clustering pour tirer des conclusions préliminaire sur la faisabilité du projet.

SOMMAIRE

PRESENTATION DU JEU DE DONNEES

Fichier descriptif des produits

1050 entrées , une ligne par produit

#	Column
0	uniq_id
1	crawl_timestamp
2	product_url
3	product_name
4	product_category_tree
5	pid
6	retail_price
7	discounted_price
8	image
9	is_FK_Advantage_product
10	description
11	product_rating
12	overall_rating
13	brand
14	product_specifications

Images

Dossier contenant 1050 images de formes et tailles diverses

SOMMAIRE

DONNEES RETENUES AVANT PRETRAITEMENT

Features

Nous avons utilisé le nom des produits et la description des produits pour constituer l’ ensemble de données textuelles relatives à chaque produit.

Données

#	Column
0	uniq_id
1	crawl_timestamp
2	product_url
3	product_name
4	product_category_tree
5	pid
6	retail_price
7	discounted_price
8	image
9	is_FK_Advantage_product
10	description
11	product_rating
12	overall_rating
13	brand
14	product_specifications

Target

Nous avons utilisé la catégorie principale des produits pour déterminer le clustering.

Au total nous avons 7 catégories contenant chacune 150 produits.

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

APPROCHES BAG OF WORDS

Deux méthodes testées :

- N-gram count
- TF-IDF

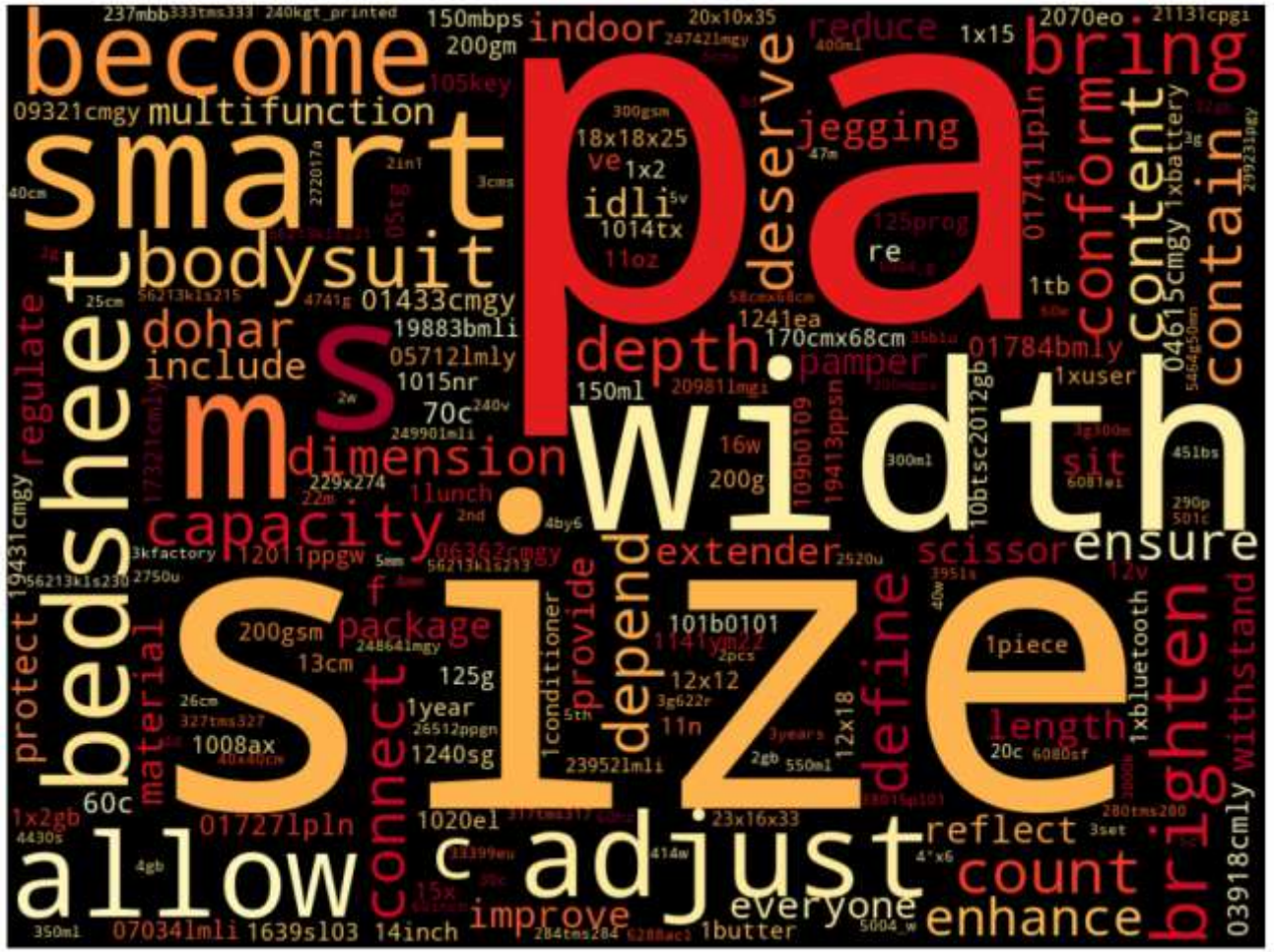
Deux approaches pour la modélisation du vocabulaire :

- Unigram :

Vocabulaire de 5575 unigram après preprocessing.

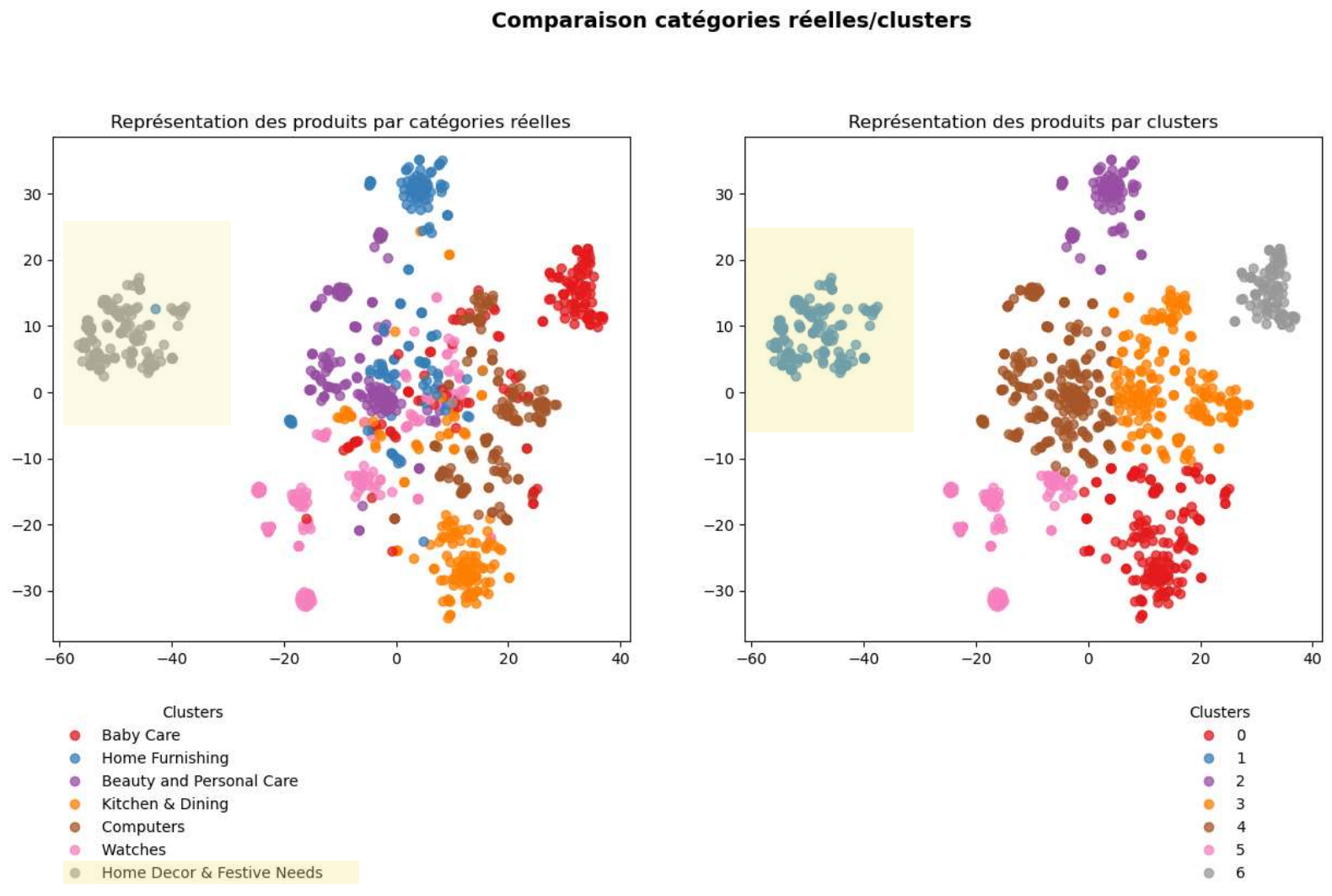
- Bigram :

Vocabulaire de 21602 après preprocessing.



Poids relatif des mots dans le vocabulaire

COMPTAGE SIMPLE - UNIGRAM

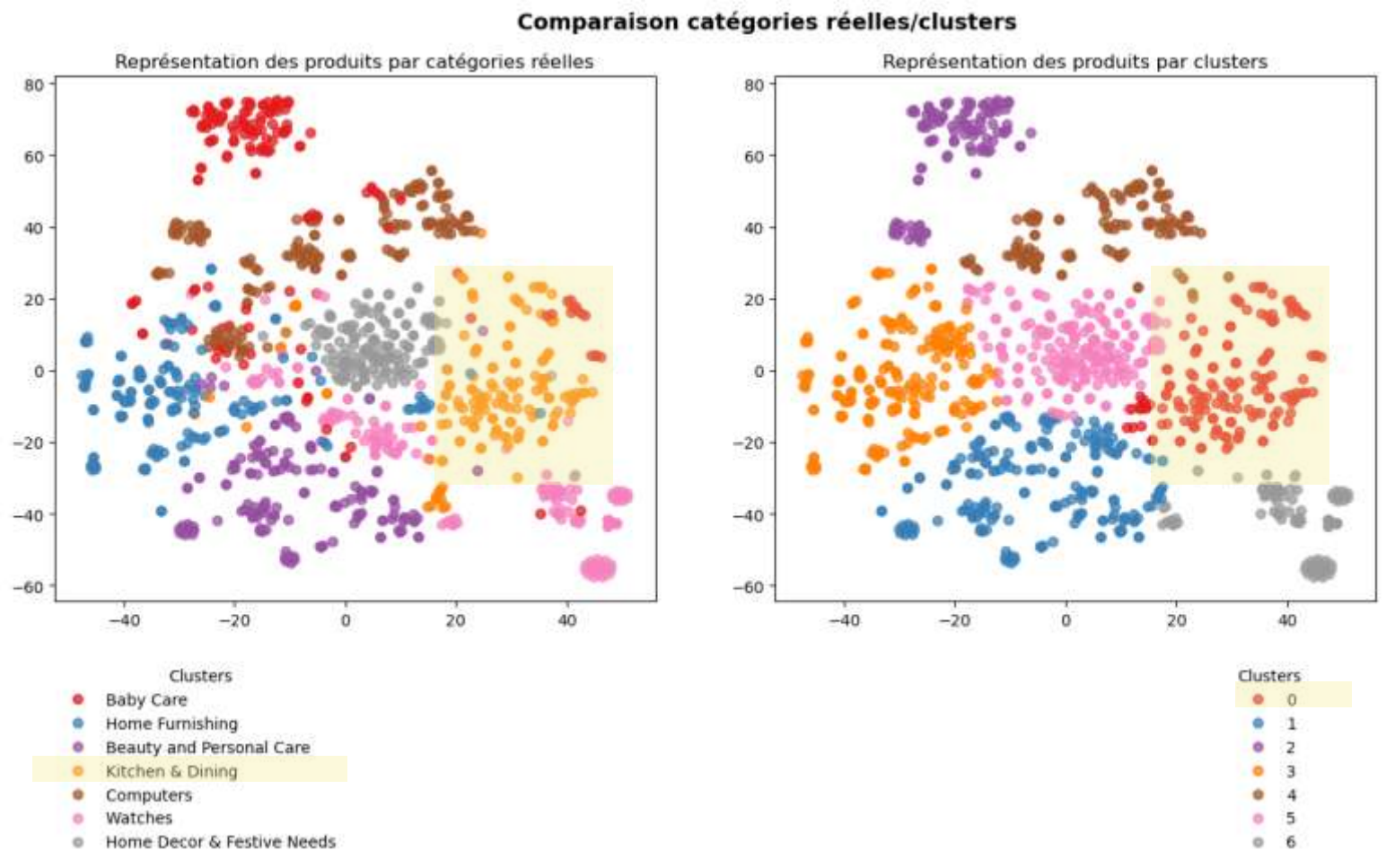


Produits par cluster

	Num prod_desc per cluster	%
4	239	22.76
3	213	20.29
0	170	16.19
1	150	14.29
5	101	9.62
2	94	8.95
6	83	7.90

ARI score: 0.4623

TF-IDF- UNIGRAM



Produits par cluster

	Num prod_desc per cluster	%
5	200	19.05
3	192	18.29
1	192	18.29
0	151	14.38
4	121	11.52
2	105	10.00
6	89	8.48

ARI score: 0.529

APPROCHE WORD EMBEDDING CLASSIQUE

Méthode testée :
Word2vec

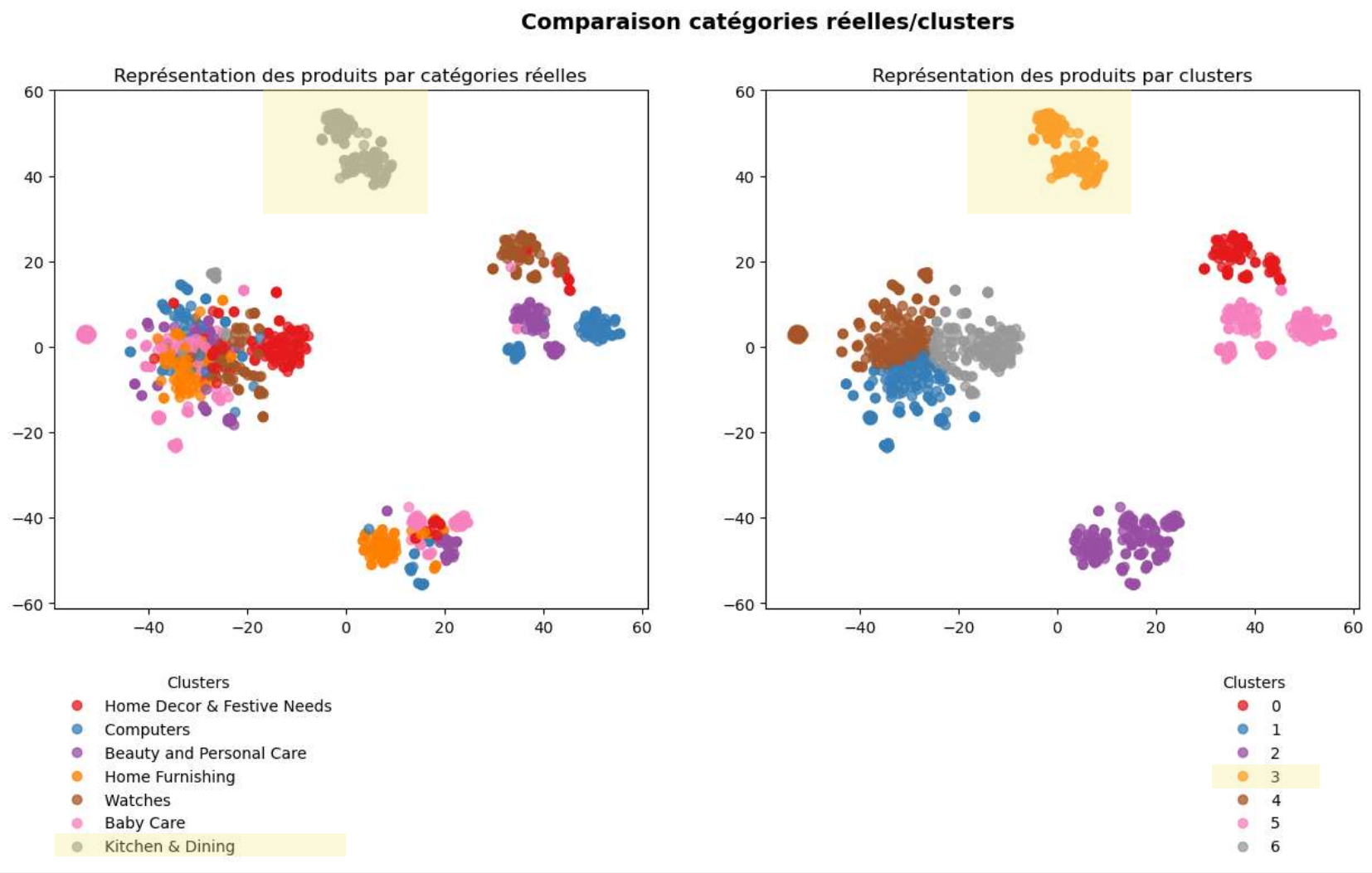
- Création d'un nouveau corpus réduit : 1858 mots

- Création d'un vecteur de word embedding pour chaque mot du corpus.

- Création d'un vecteur de word embedding pour chaque produit en prenant la moyenne des word embedding des mots du corpus contenus dans la description du produit.

Corpus Word2vec après preprocessing

Word2Vec

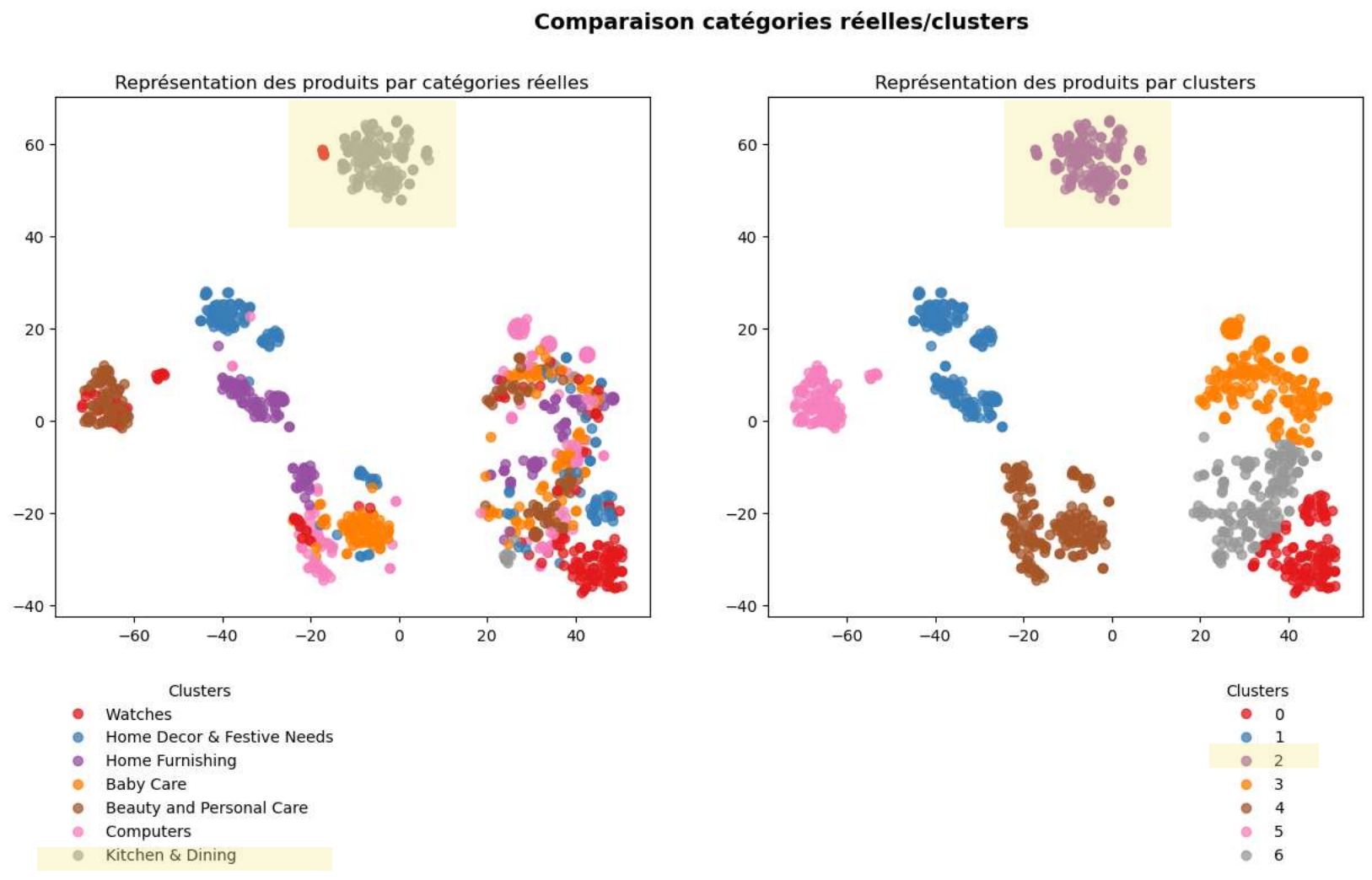


Produits par cluster

	Num prod_desc per cluster	%
2	185	17.62
4	177	16.86
6	162	15.43
5	153	14.57
1	146	13.90
3	135	12.86
0	92	8.76

ARI score: 0. 3111

WORD EMBEDDING – BERT

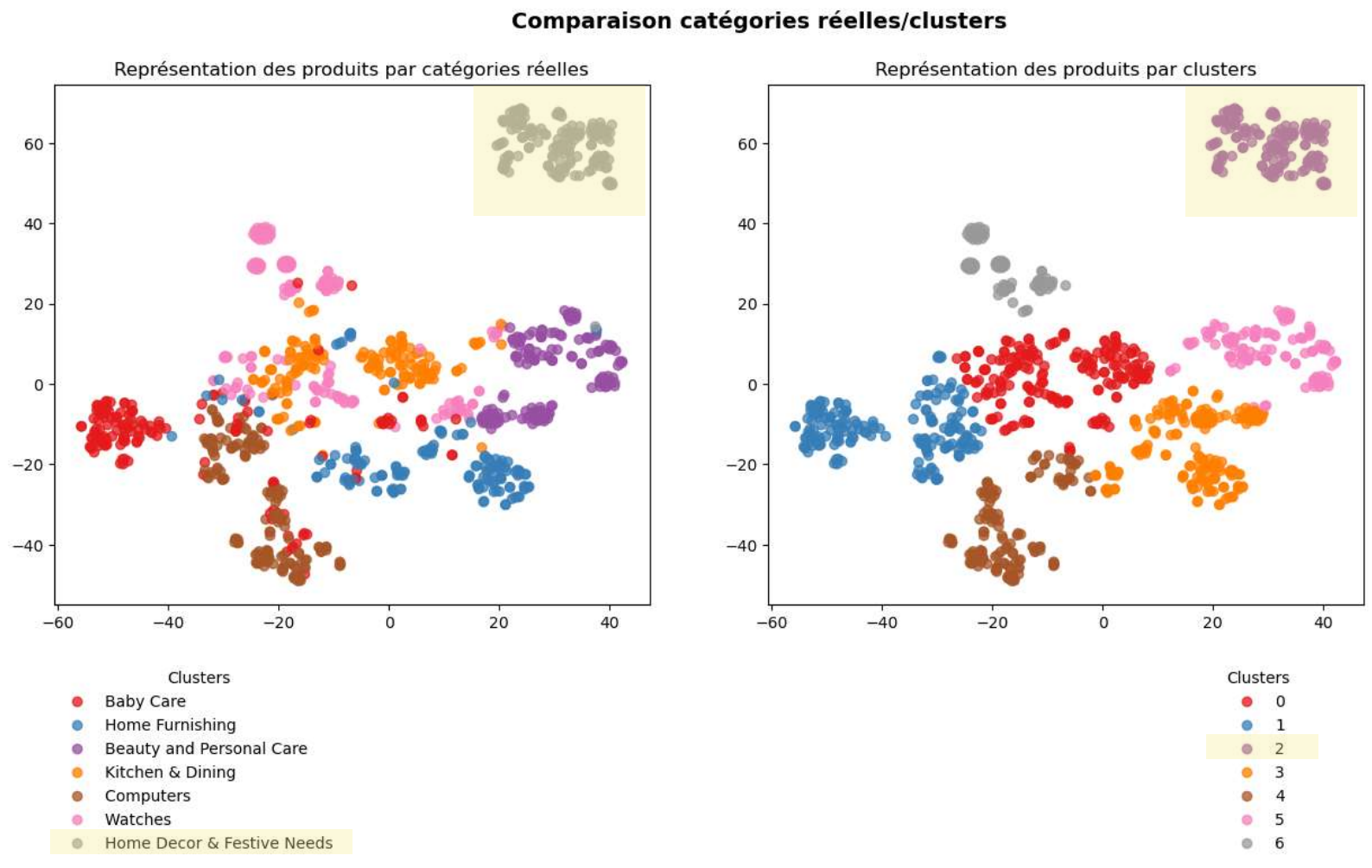


Produits par cluster

	Num prod_desc per cluster	%
3	190	18.10
4	186	17.71
6	166	15.81
1	153	14.57
2	140	13.33
0	121	11.52
5	94	8.95

ARI score: 0. 3081

WORD EMBEDDING – USE



Produits par cluster

	Num prod_desc per cluster	%
0	196	18.67
1	190	18.10
3	176	16.76
2	149	14.19
4	136	12.95
5	123	11.71
6	80	7.62

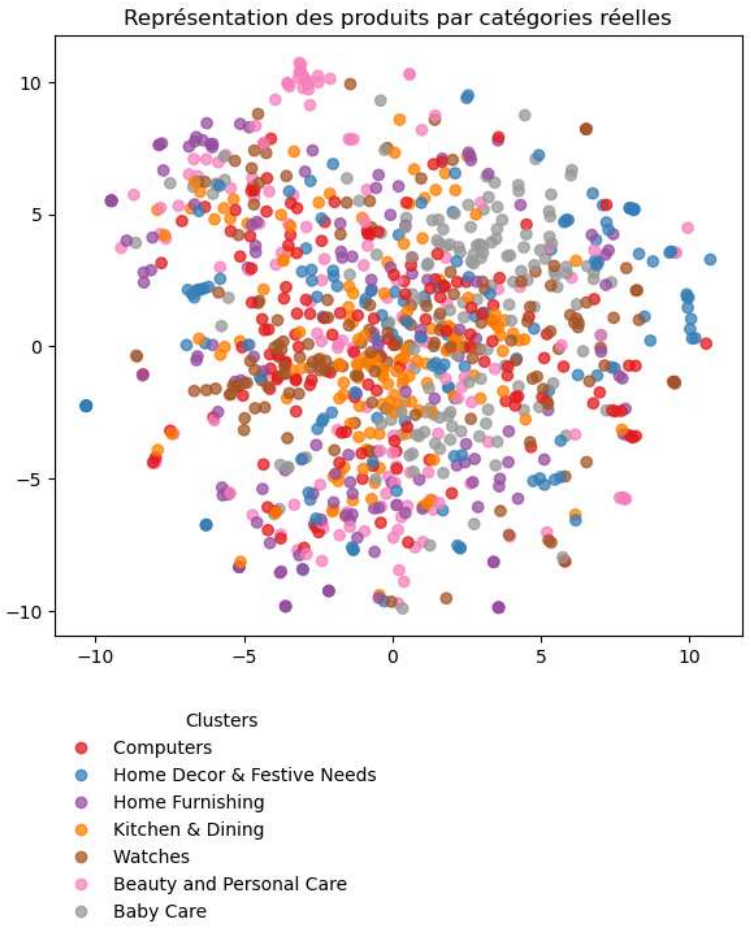
ARI score: 0. 5074

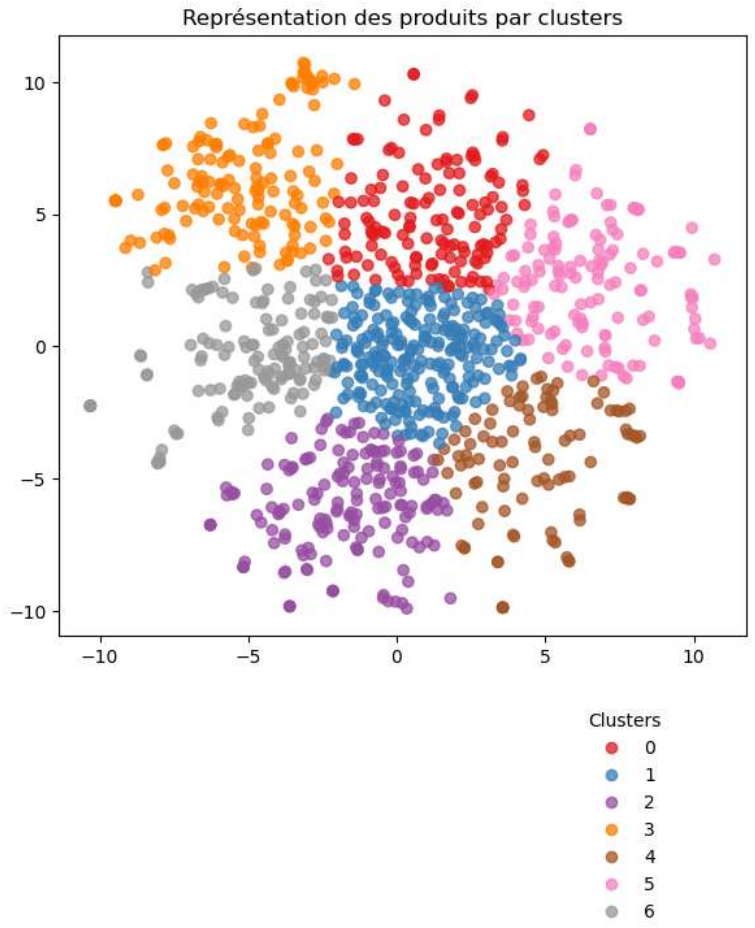
SYNTHESE DES APPROCHES SUR LES DONNEES TEXTUELLES

	COMPTAGE SIMPLE UNIGRAM	TF-IDF- UNIGRAM	Word2Vec	BERT	USE
Preprocessing	Custom tokenizer	Custom tokenizer	Custom tokenizer	Bert Tokenizer	Pas de prétraitement
ARI	0.4623	0.529	0. 3111	0. 3081	0. 5074
Catégorie la mieux classée	Home Decor & Festive Needs	Kitchen & Dining	Kitchen & Dining	Kitchen & Dining	Home Decor & Festive Needs

SOMMAIRE

EXTRACTION DES FEATURES - SIFT





Produits par cluster

	Num prod_desc per cluster	%
1	234	22.29
2	162	15.43
3	143	13.62
6	141	13.43
0	137	13.05
5	134	12.76
4	99	9.43

ARI score: 0.0509

EXTRACTION DES FEATURES - CNN TRANSFERT LEARNING VGG16

Input

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0

+

Output

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 4096)	102764544
dense_1 (Dense)	(None, 4096)	16781312

Récupération des features vector

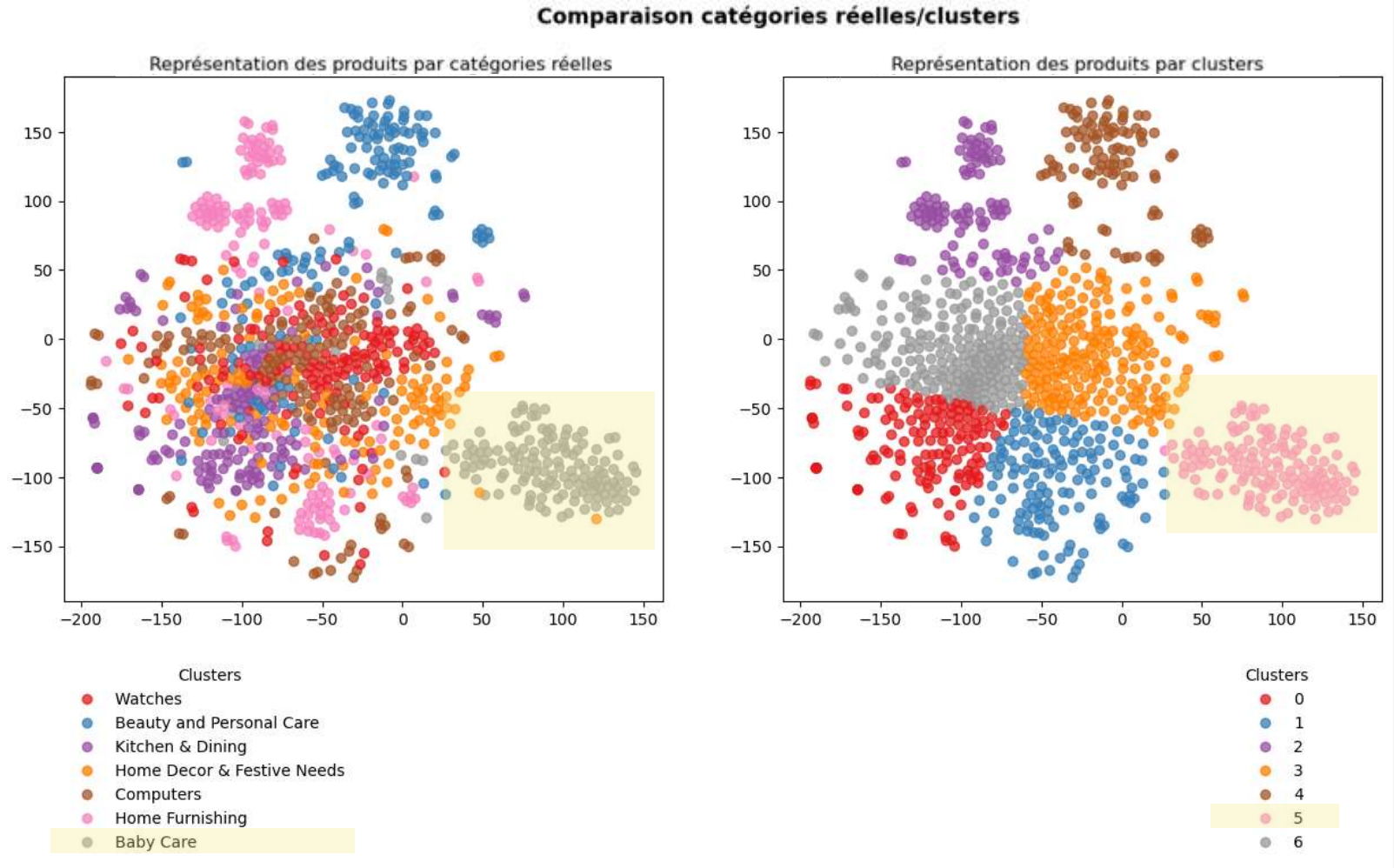
Suppression du layer prediction

=

New Model

Total params: 134,260,544

CNN TRANSFER LEARNING



Produits par cluster

	Num prod_desc per cluster	%
3	210	20.00
6	205	19.52
0	145	13.81
1	142	13.52
5	139	13.24
2	107	10.19
4	102	9.71

ARI score: 0.2886

SOMMAIRE

CONCLUSION

FAISABILITE

Conclusion positive sur la faisabilité du projet de classification

RESULTATS SUR LE DONNEES TEXTUELLES

- TF-IDF et Universal sentence encoder ont donné les meilleurs résultats sur le jeu de données.
- Universal sentence encoder présente un meilleur potentiel à long terme pour la mise en oeuvre de la classification des produits basée sur les descriptions de produits.

RESULTATS SUR LES IMAGES

- Un algorithme de type CNN semble beaucoup mieux adapté à la mise en oeuvre de la classification des catégorie de produit.