

CLASSIFICATION AUTOMATIQUE DES BIENS DE CONSOMMATION





PARTIE I : INTRODUCTION

- 1) CONTEXTE GÉNÉRAL
 - 2) OBJECTIF DE L'ANALYSE
 - 3) MÉTHODE GÉNÉRALE
- 

CONTEXTE GÉNÉRAL

- **Contexte :**

L'entreprise "Place de marché" souhaite lancer une marketplace e-commerce. Pour cela, les vendeurs doivent attribuer une catégorie manuellement à leurs produits à partir d'une description et d'une photo. Afin de passer à une plus large échelle et faciliter le processus, il devient nécessaire d'automatiser cette tâche.



- **Mission :**

Réaliser une première étude de faisabilité d'un moteur de classification

1. Analyse de données textuelles (description) + Clustering
2. Analyse de données visuelles (image) + Clustering

- **Données :**

Export de la base de données contenant 1050 produits et leurs images associées

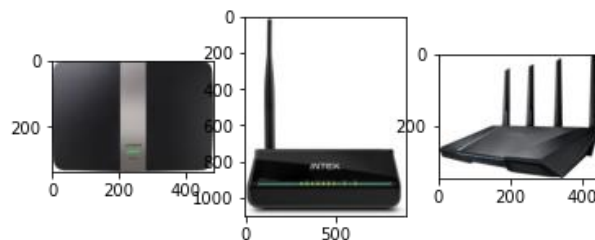
OBJECTIF DE L'ANALYSE :

Il existe 7 grandes catégories d'articles dans nos données (150 items par categories)

Watch



Computers



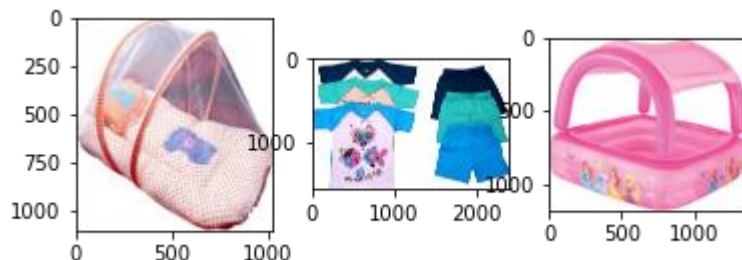
Kitchen & Dining



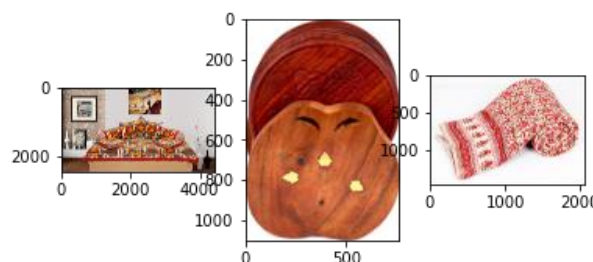
Home Decor & Festive Needs



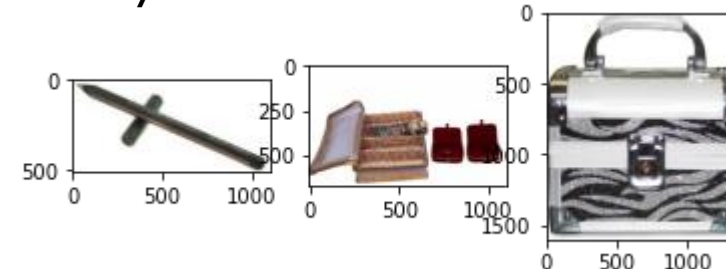
Baby Care



Home Furnishing

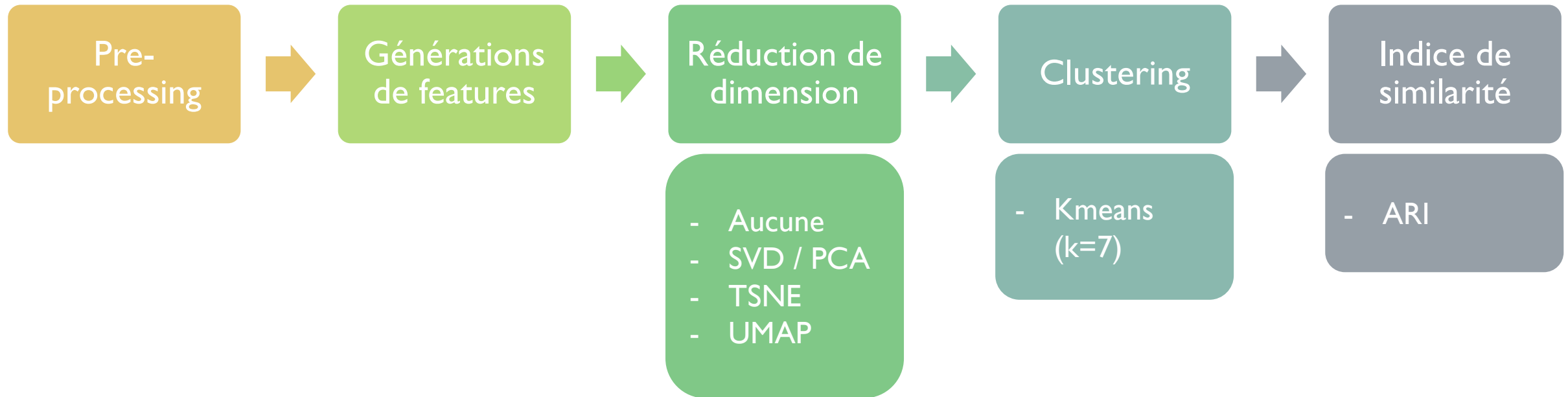


Beauty and Personal Care




- Peut on clusteriser les produits en 7 catégories uniquement par leur description + images ?

MÉTHODE GÉNÉRALE :





PARTIE II : ANALYSE DE TEXTE

- 1) PRE-PROCESSING
 - 2) MODÈLES DE BAG OF WORDS
 - 3) MODÈLES DE WORD/SENTENCE EMBEDDING
 - 4) RÉSULTATS
- 

PRE-PROCESSING



'Key Features of Elegance Polyester
Multicolor Abstract Eyelet Door
Curtain Floral'...

['key', 'features', 'of', 'elegance', 'polyester',
'multicolor', 'abstract', 'eyelet', 'door',
'curtain']

['key', 'features', 'elegance', 'polyester',
'multicolor', 'abstract', 'eyelet', 'door',
'curtain']

Stemming vs Lemmatization



Liste de mots non
informatifs du corpus (pre-
processed)

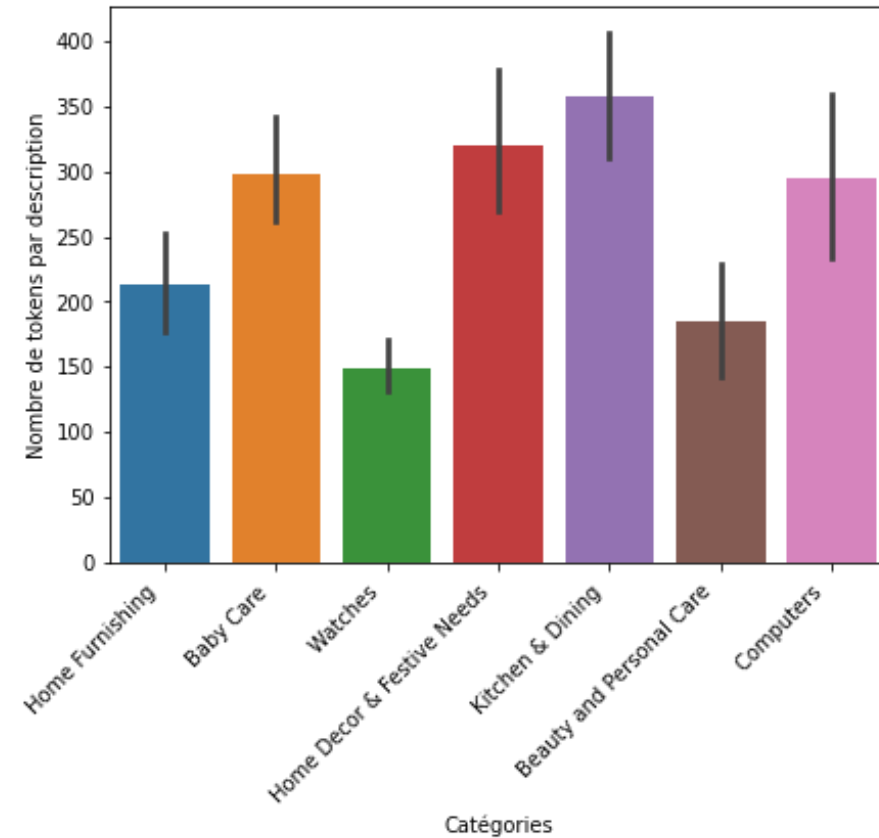
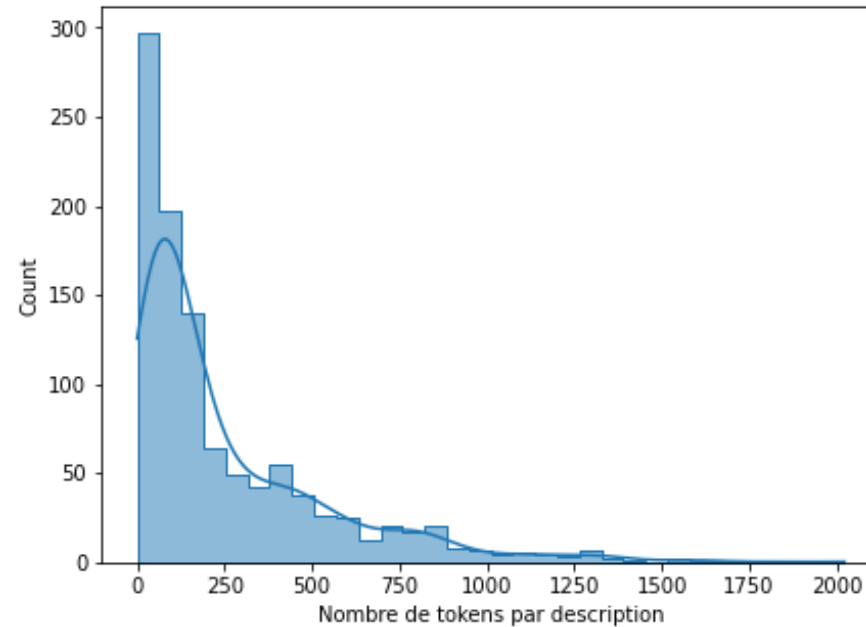
- Taille des mots :
< 3 lettres
- Récurrence des Mots :
> x (13) mots les plus
présents

PRÉSENTATION DU CORPUS

Après pre-processing :

Nombre de mots (tokens) : 39567

Nombre de tokens uniques : 3910



MODÈLE DE BAG OF WORDS : COUNT VECTORIZER ET TF-IDF

■ Modèle Count Vectorizer

Phrase 1 : "Je suis à la maison"

Phrase 2 : "La maison est dans la prairie"

Phrase 3 : "Je suis à la plage"

	je	suis	à	la	maison	est	dans	prairie	plage
Phrase 1	1	1	1	1	1	0	0	0	0
Phrase 2	0	0	0	2	1	1	1	1	0
Phrase 3	1	1	1	1	0	0	0	0	1

■ Modèle TF-IDF

Phrase 1				0,2					
Phrase 2				0,3					
Phrase 3				0,2					

TF * IDF où:

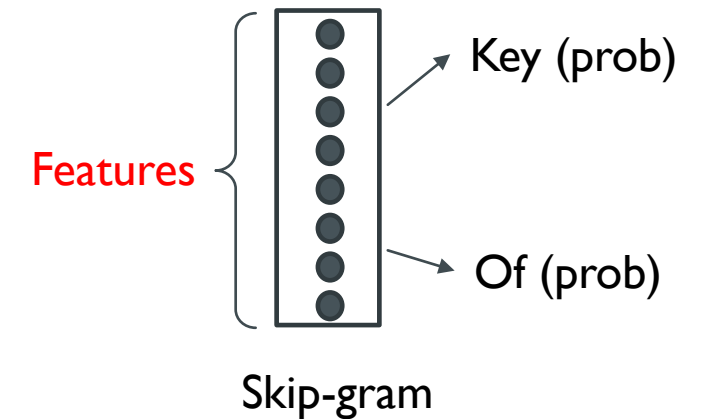
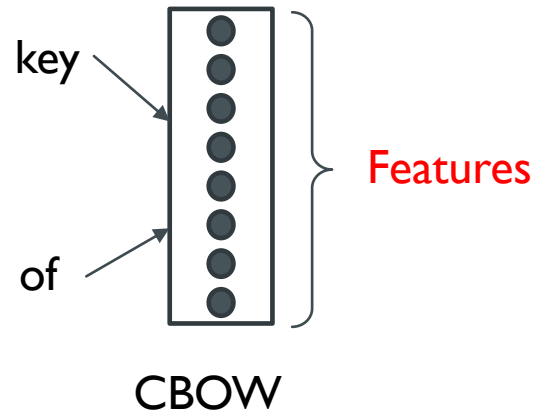
•TF = nombre de fois où le mot est dans le document / nombre de mots dans le document

•IDF = nombre de documents / nombre de documents où apparaît le mot

MODÈLE DE WORD EMBEDDING : WORD2VEC & FASTTEXT

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral

Modèle	Output
Word2vec	1200*
FastText	600*



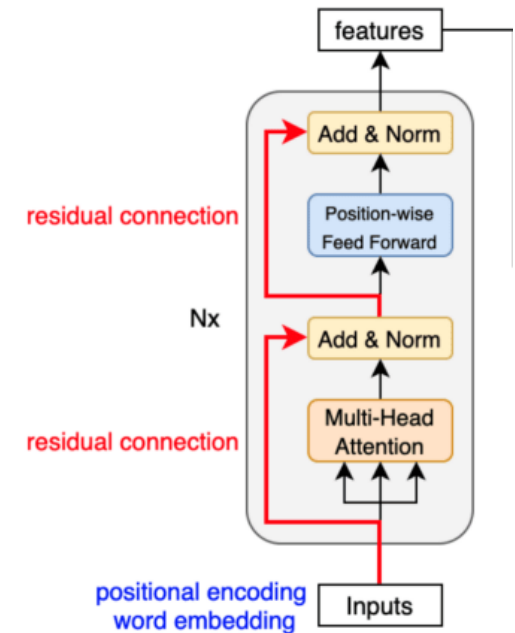
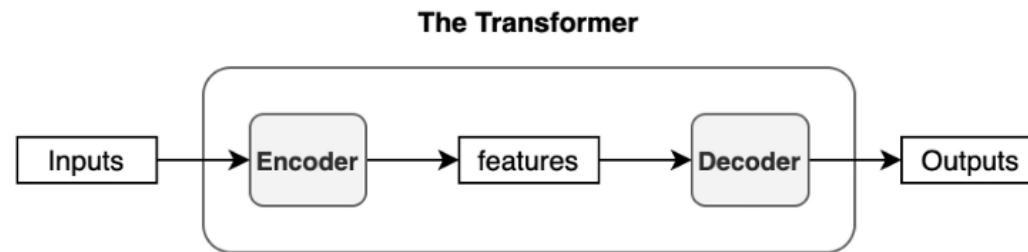
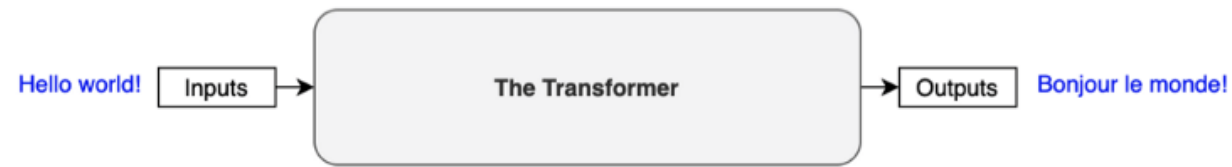
1. Création du modèle et entraînement sur tout le corpus.
 - Chaque mot transformé en vecteur (embedding)
2. Pour chaque document : Création de la matrice d'embedding des différents mots puis pooling (GlobalAverage)
 - Chaque document est un vecteur de dimension X

*Valeurs testées : 300, 600, 1200

Modèle : Devlin et al. *arXiv, 2019*

MODÈLE DE SENTENCE EMBEDDING : BERT

Modèle	Output
Word2vec	1200*
FastText	600*
BERT	768



Modèle :

Transformers : Vaswani et al. *arXiv, 2017*

BERT : Devlin et al. *arXiv, 2019*

Src images : <https://kikaben.com/transformers-encoder-decoder/>

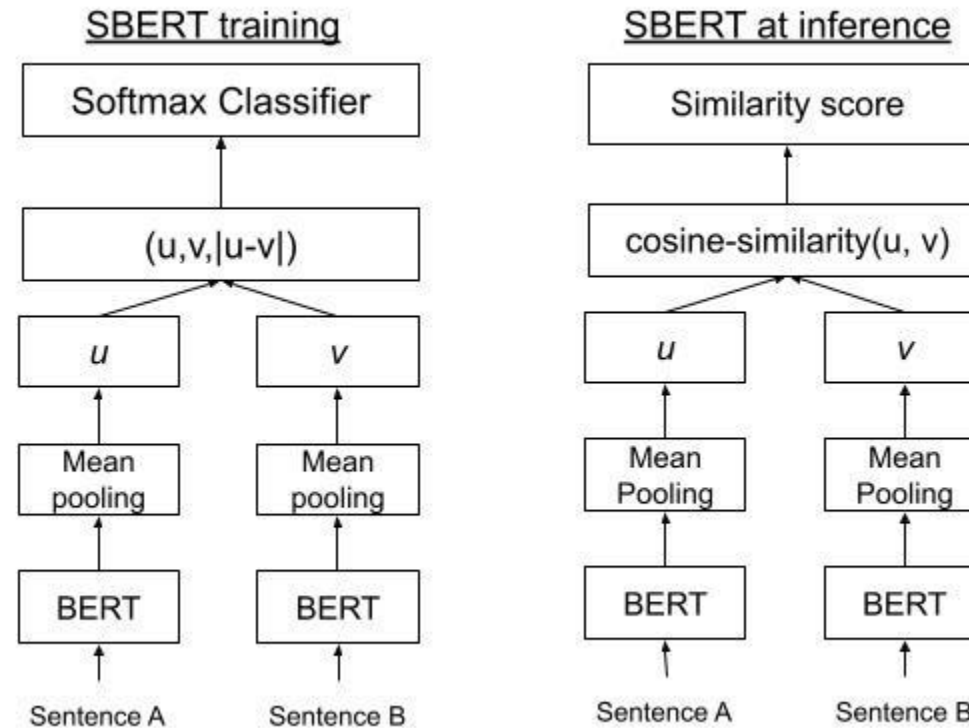
Bidirectional Encoder Representations from Transformers

MODÈLE DE SENTENCE EMBEDDING : SBERT

Sentence - BERT

Modèle	Output
Word2vec	1200*
FastText	600*
BERT	768
SBERT	768

- Modèle entraîné sur deux BERT en parallèle (Siamese Networks)

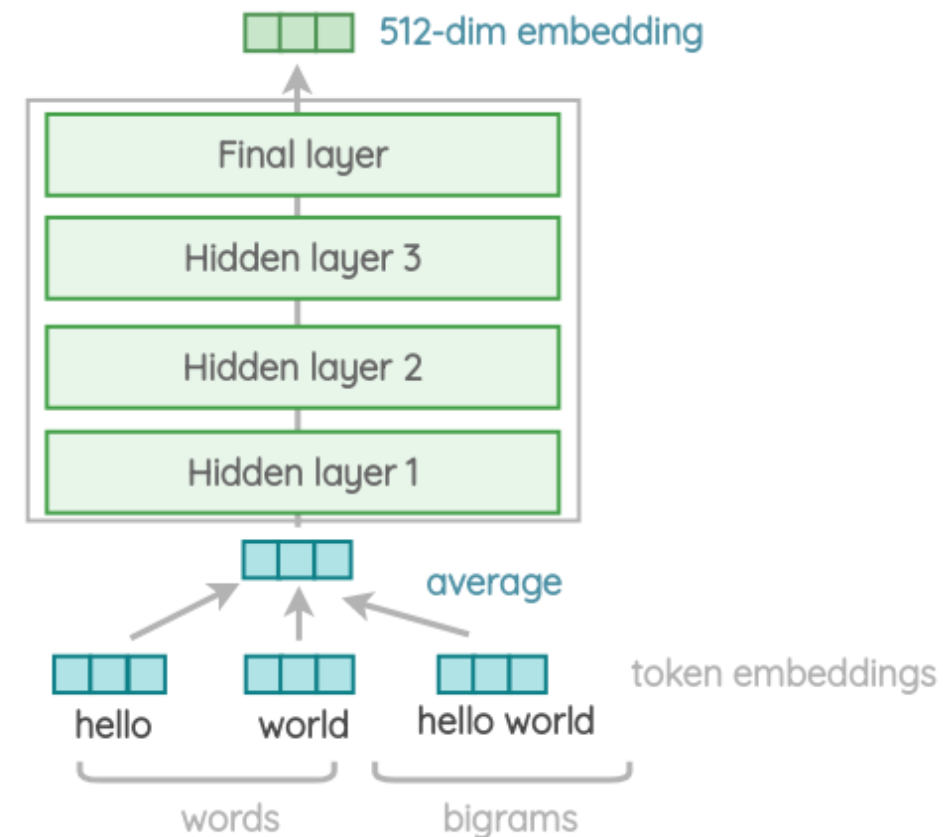


Modèle : Reimer & Gurevych. *arXiv*, 2019
Src image : Saketh Kotamraju,
<https://towardsdatascience.com/an-intuitive-explanation-of-sentence-bert-1984d144a868>

MODÈLE DE SENTENCE EMBEDDING : USE

Modèle	Output
Word2vec	1200*
FastText	600*
BERT	768
SBERT	768
USE	512

Universal Sentence Embedding



Modèle :

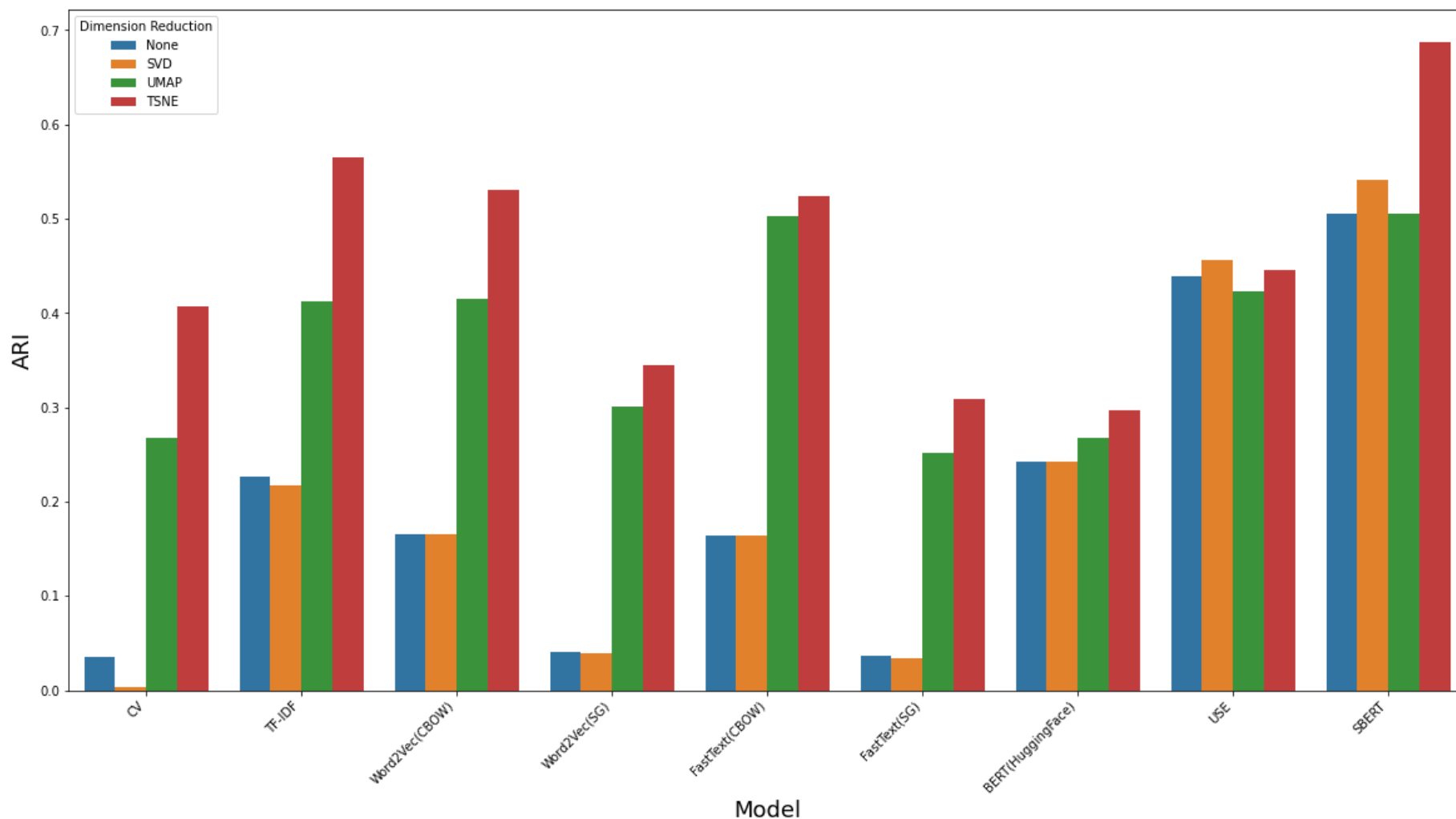
Iyyer et al. *Proceedings of ACL/IJCNLP, 2015*

Cer et al. *arXiv, 2018*

Src image :

Amit Chaudhary, <https://amitniss.com/2020/06/universal-sentence-encoder/>

RÉSULTATS DES DIFFÉRENTS MODÈLES

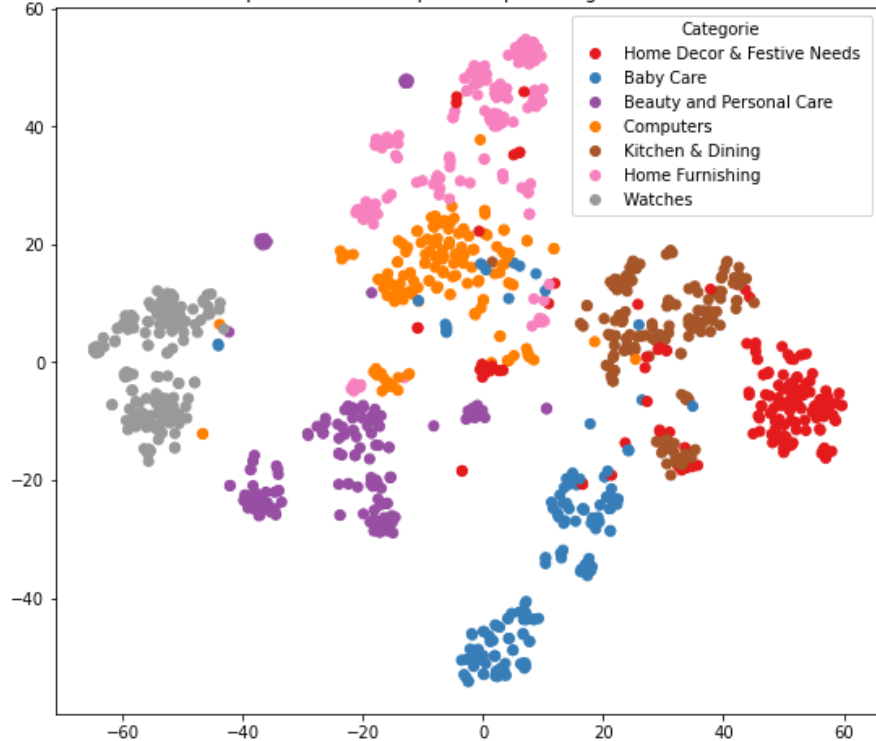


Meilleur modèle :
Reduction par TSNE +SBERT

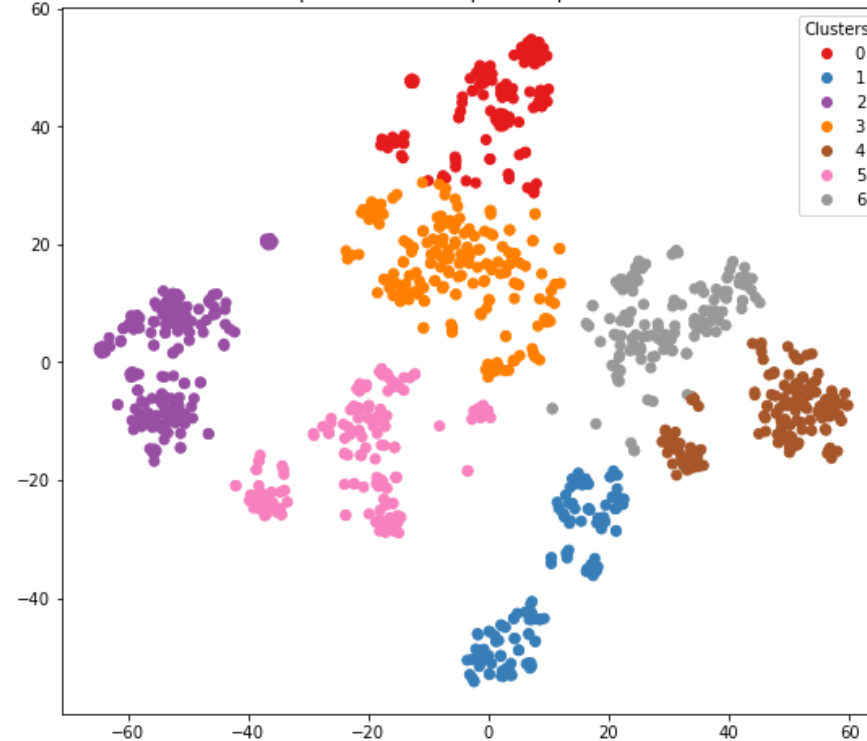
ARI : 0.69

REPRÉSENTATION DU MEILLEUR MODÈLE : SBERT ET TSNE

Représentation des produits par catégories réelles



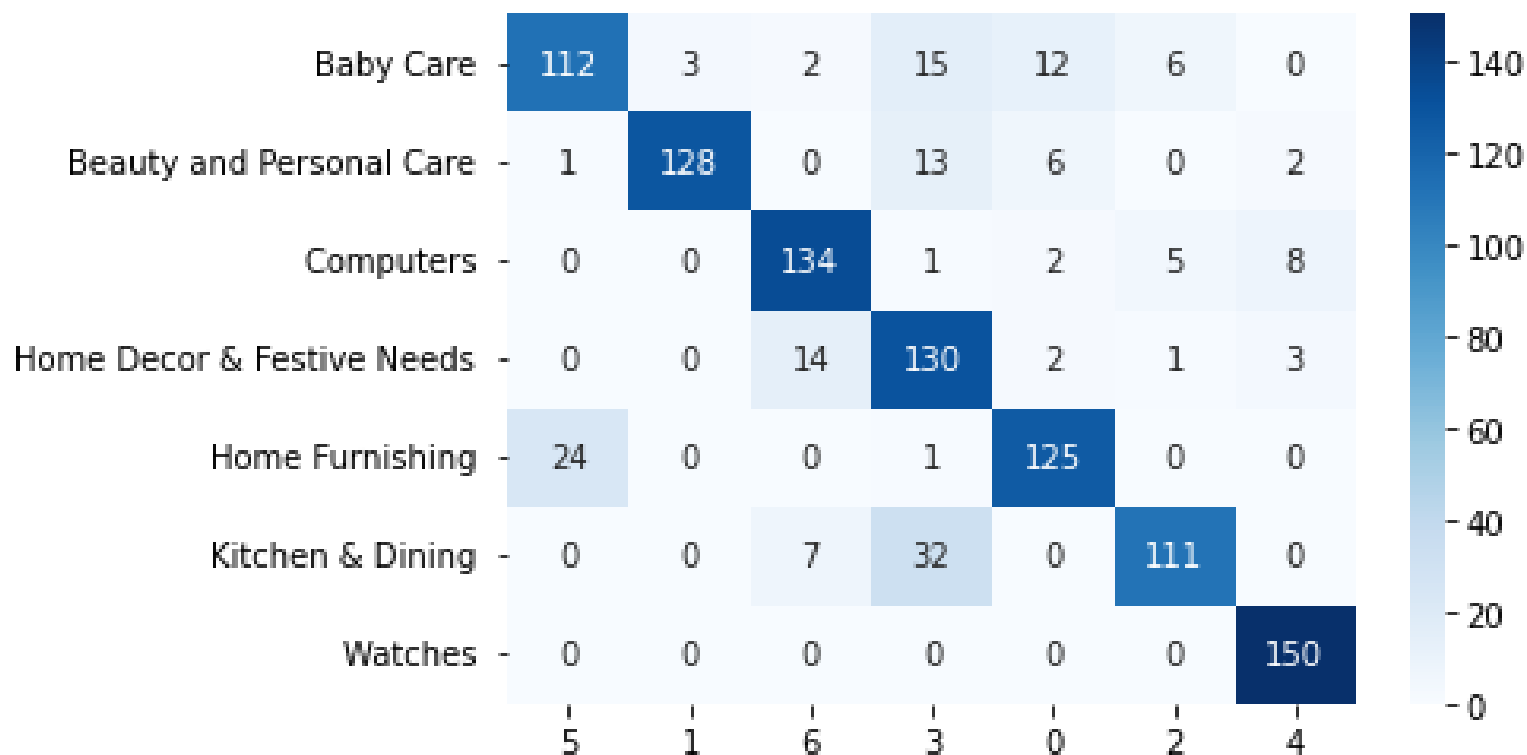
Représentation des produits par clusters



Meilleur modèle :
Reduction par TSNE + SBERT

ARI : 0.69

MATRICE DE CONFUSION



Erreurs principales :

- Kitchen & dining (74%)
- Baby Care (75%)



PARTIE III : ANALYSE D'IMAGE

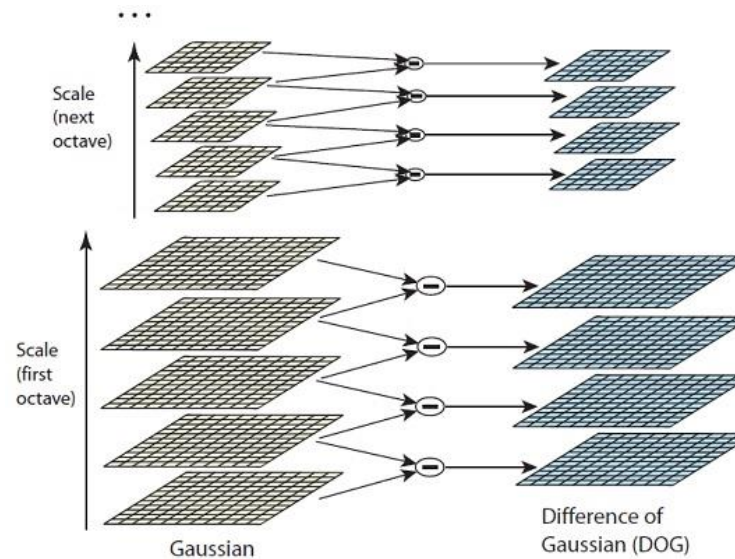
- 1) PRE-PROCESSING
 - 2) MODÈLE SIFT
 - 3) MODÈLES DE RÉSEAUX DE NEURONES CONVOLUTIFS
 - 4) RÉSULTATS
- 

ETAPE I : PRE-PROCESSING

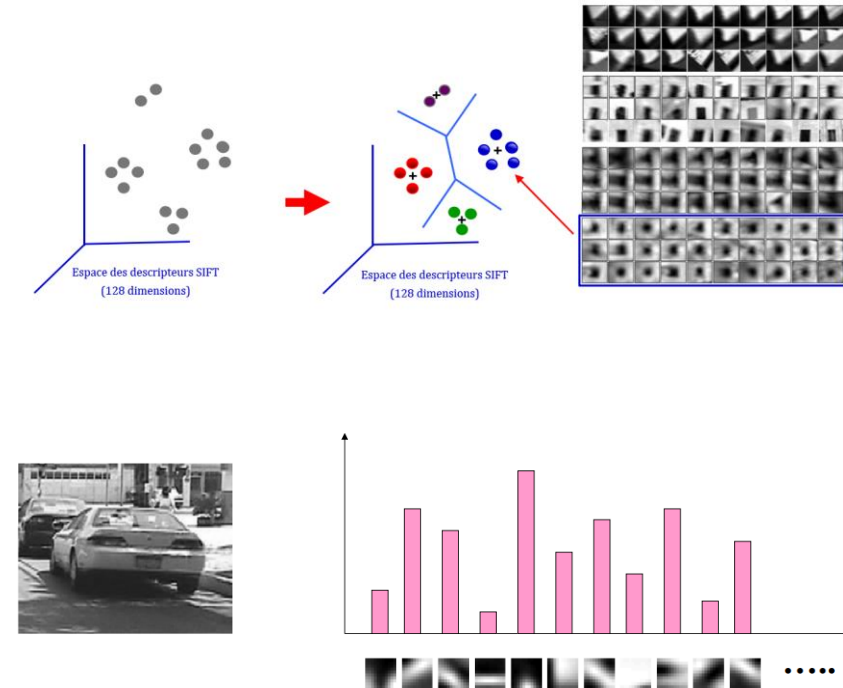


MODEL SIFT : SCALE-INVARIANT FEATURE TRANSFORM

- 1) Création des descripteurs de chaque images
- 2) Détermination de cluster de descripteurs (résumer l'info de milliers de descripteurs)
- 3) Création des Features :
 - prédiction des clusters de chaque descripteur
 - création d'un histogramme = comptage du nombre de descripteurs par cluster



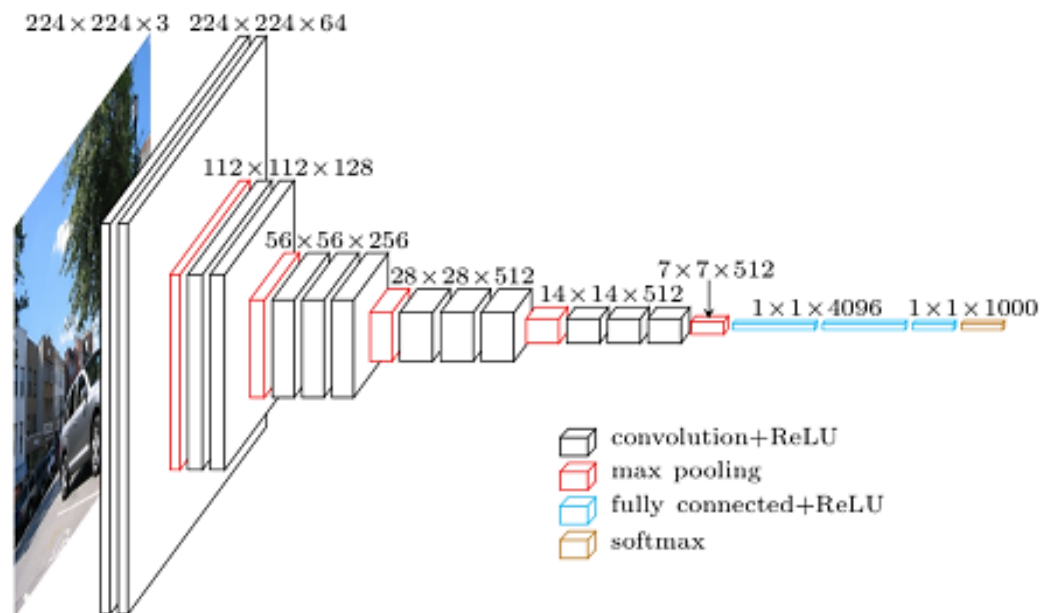
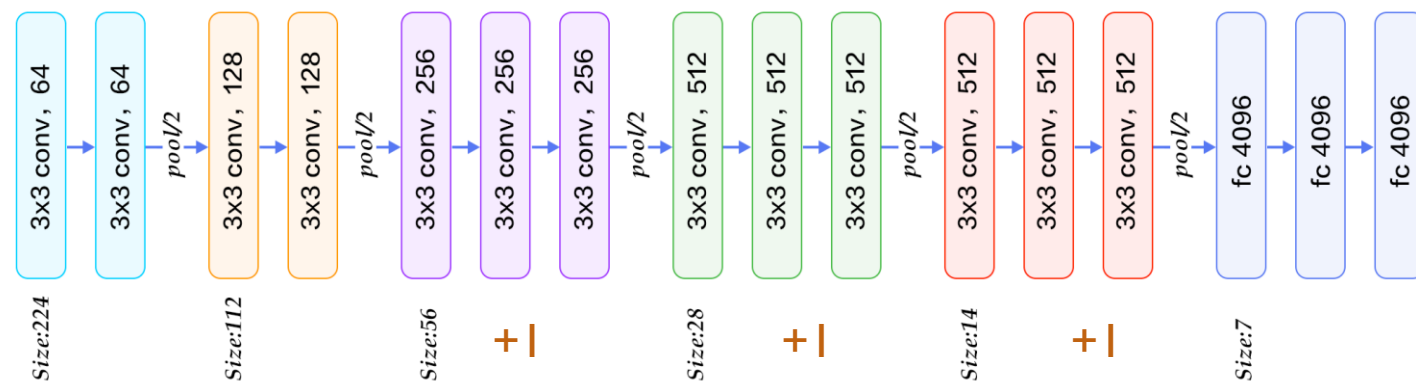
La technique de LoG permet de déterminer des maxima et minima.
source : Lowe, 2004



source : <https://openclassrooms.com/fr/courses/4470531-classez-et-segmentez-des-donnees-visuelles/5072281-utilisez-ces-features-pour-classifier-des-images#/id/r-5144451>

MODÈLES DE RÉSEAUX DE NEURONES CONVOLUTIFS : VGG16 & VGG19

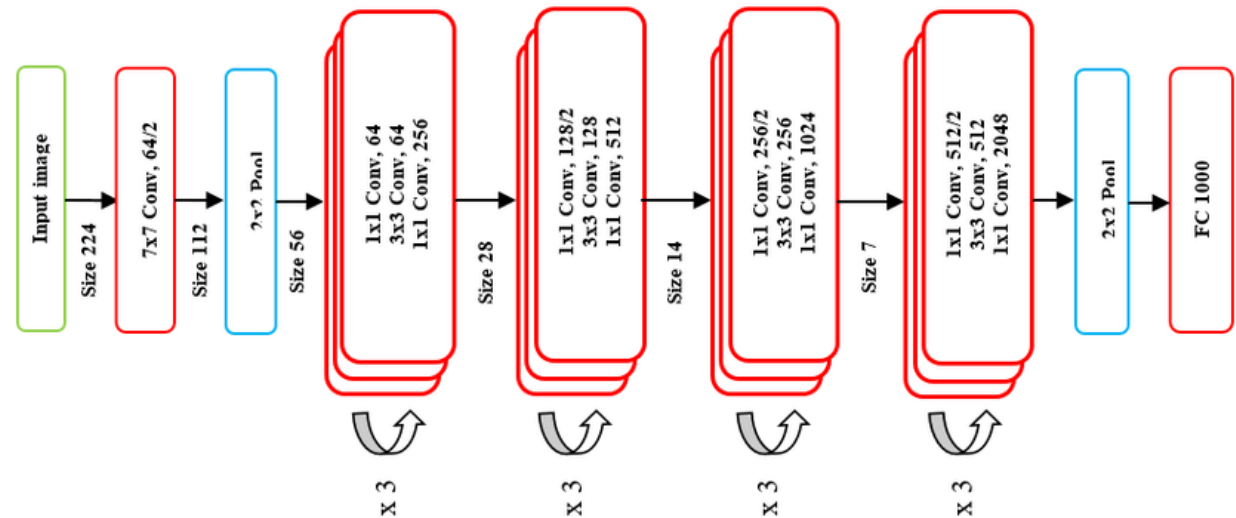
Modèle	Paramètres	Output
VGG16	138,357,544	4096
VGG19	143,667,240	4096



Modèle : Simonyan & Zisserman. *Conference paper at ICLR, 2015*
 Src image : Nash et al. *Materials Degradation (Nature), 2018*

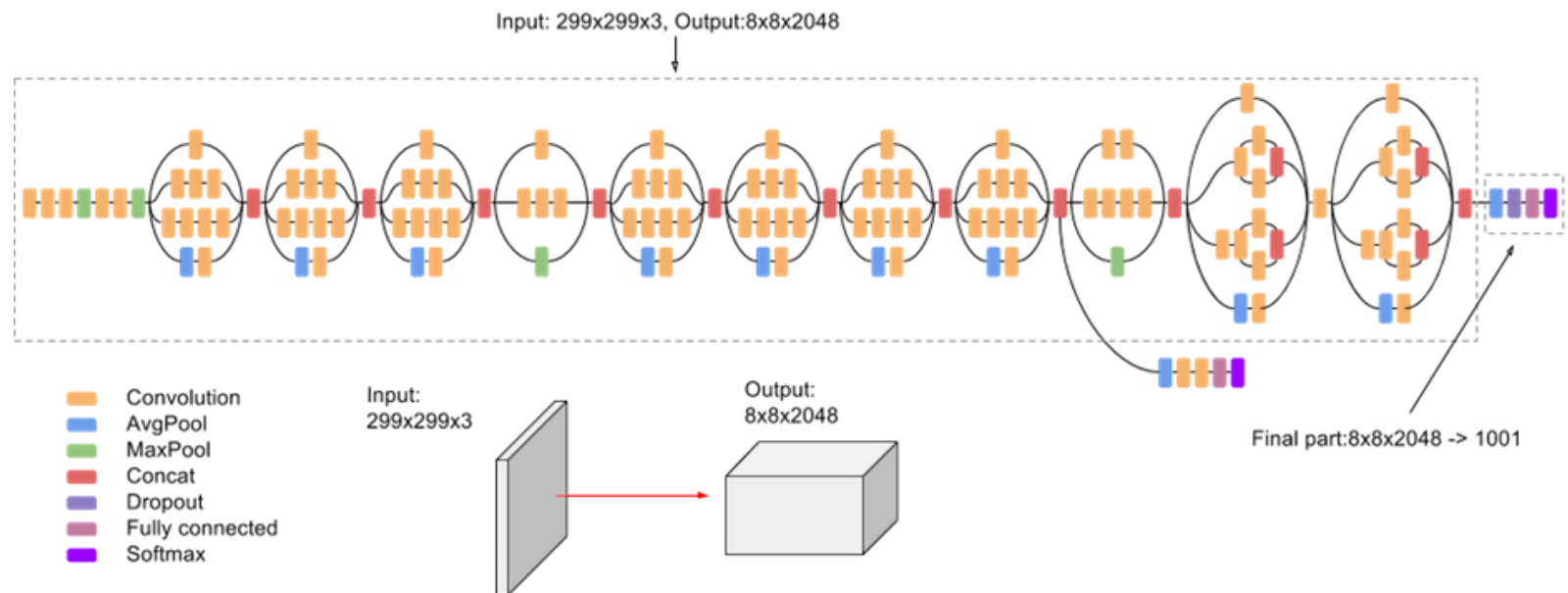
MODÈLES DE RÉSEAUX DE NEURONES CONVOLUTIFS : RESNET50

Modèle	Paramètres	Output
VGG16	138,357,544	4096
VGG19	143,667,240	4096
ResNet50	25,636,712	2048



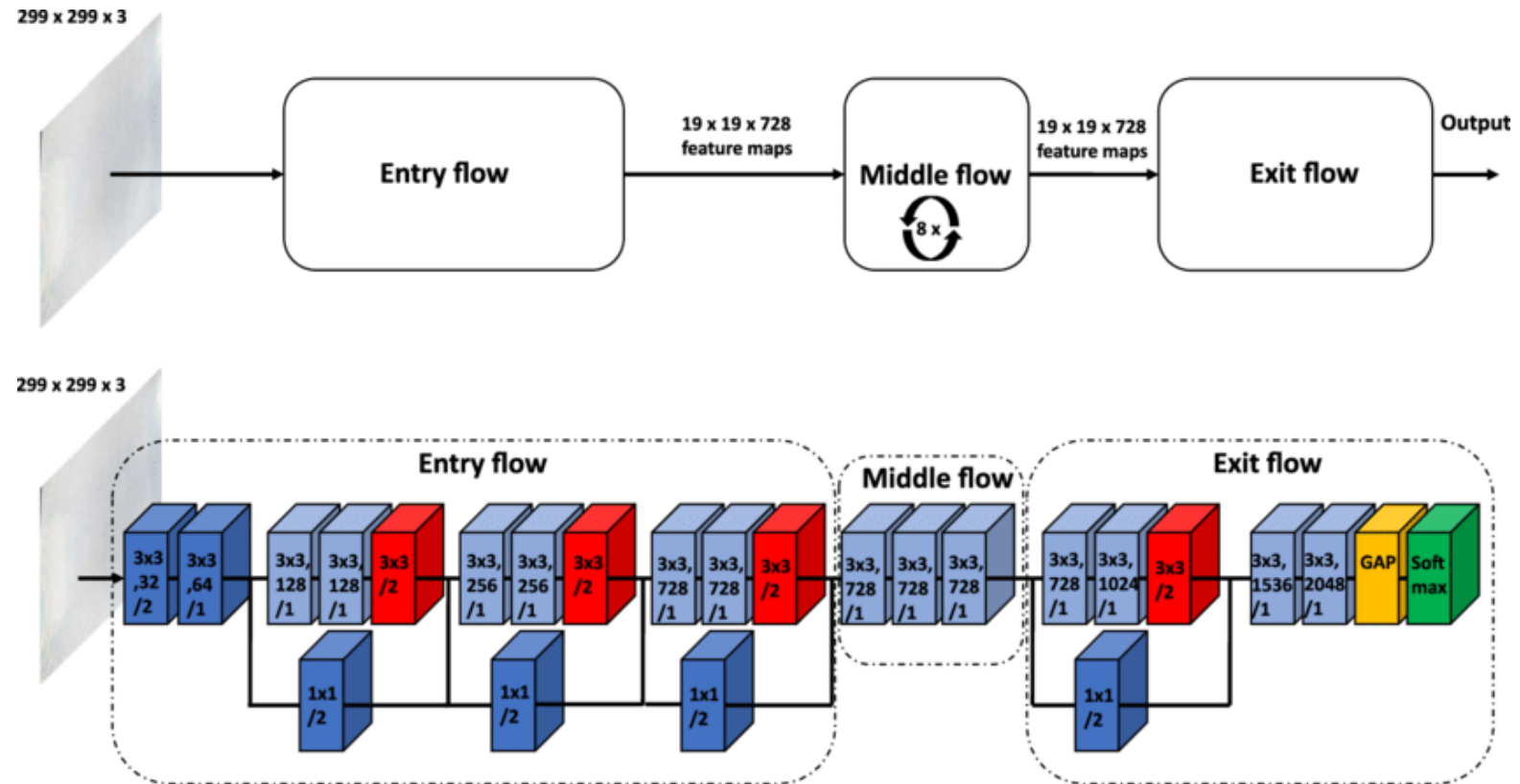
MODÈLES DE RÉSEAUX DE NEURONES CONVOLUTIFS : INCEPTIONV3

Modèle	Paramètres	Output
VGG16	138,357,544	4096
VGG19	143,667,240	4096
ResNet50	25,636,712	2048
InceptionV3	23,851,784	2048

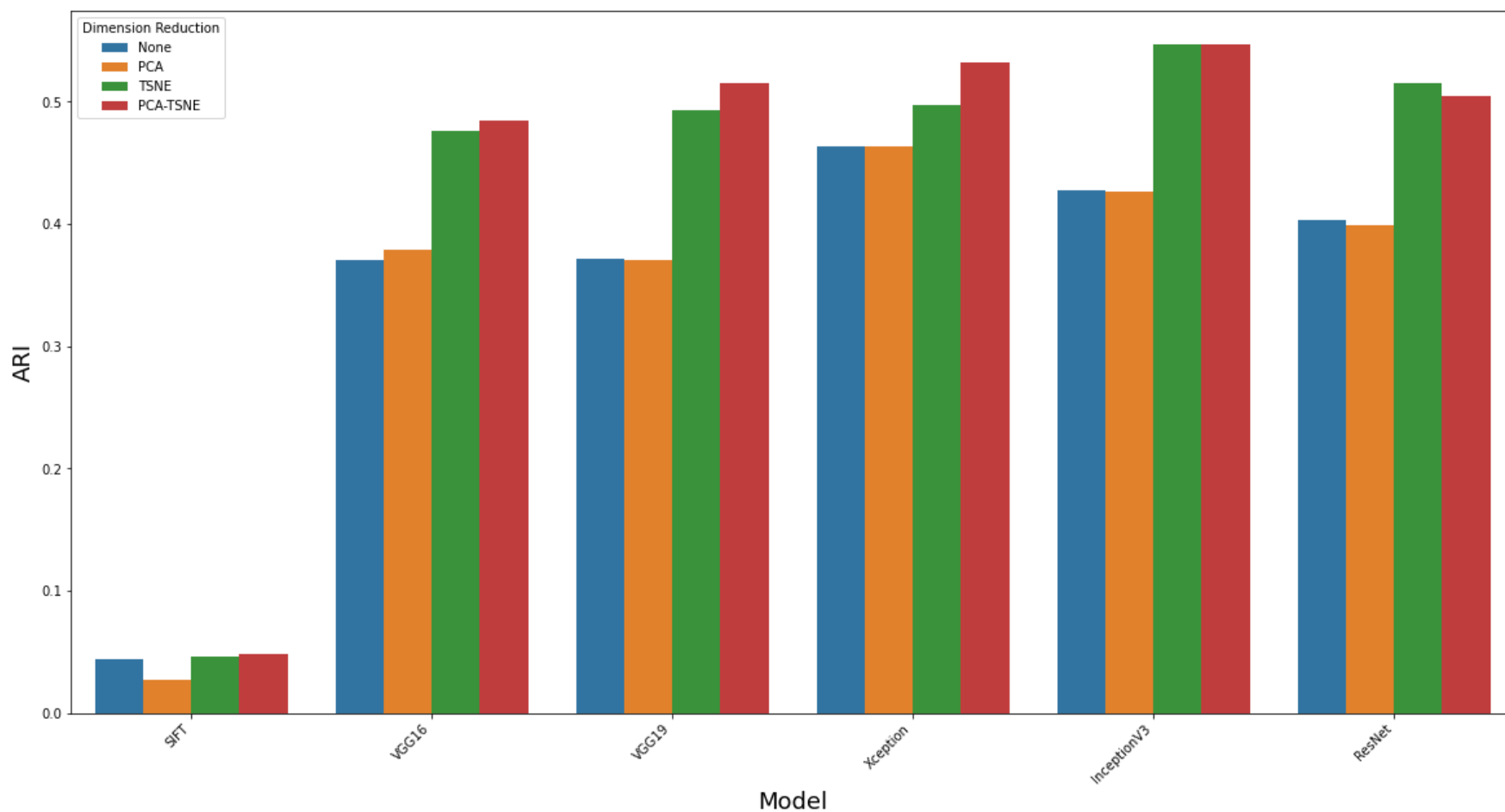


MODÈLES DE RÉSEAUX DE NEURONES CONVOLUTIFS : XCEPTION

Modèle	Paramètres	Output
VGG16	138,357,544	4096
VGG19	143,667,240	4096
ResNet50	25,636,712	2048
InceptionV3	23,851,784	2048
Xception	22,910,480	2048



RÉSULTATS DES DIFFÉRENTS MODÈLES

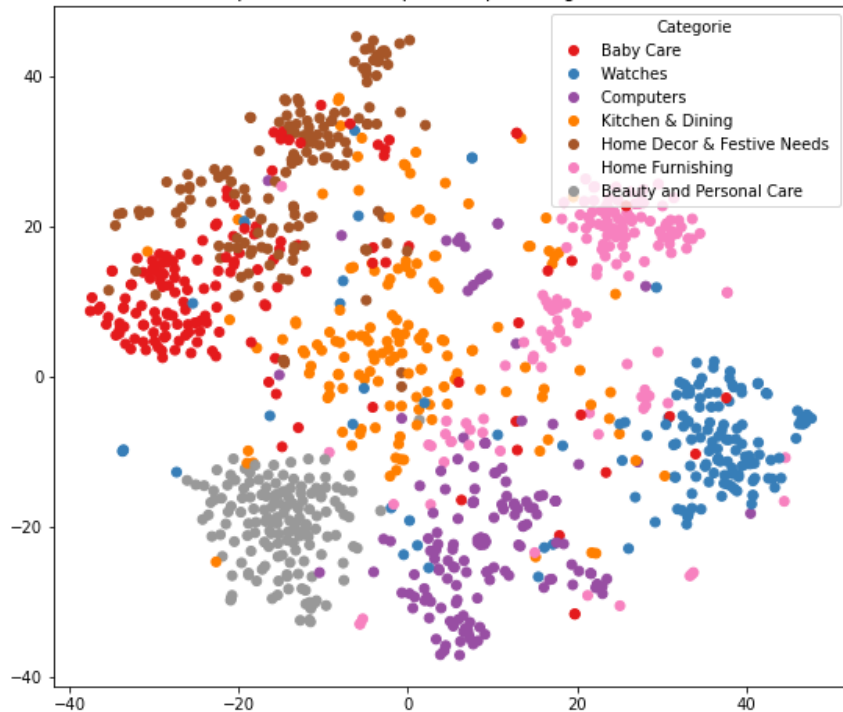


Meilleur modèle :
Reduction par TSNE + InceptionV3

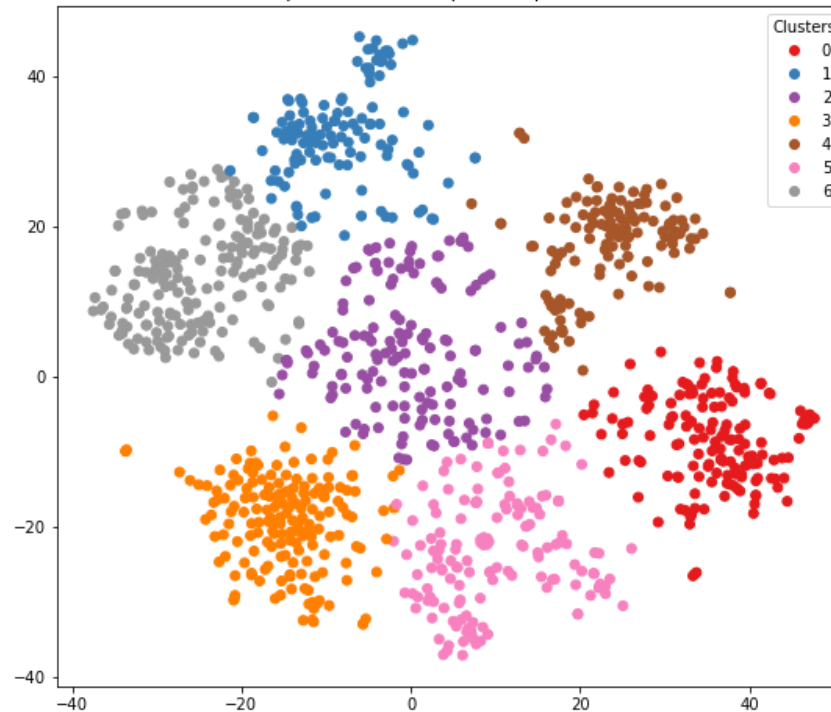
ARI : 0.547

REPRÉSENTATION DU MEILLEUR MODÈLE : INCEPTIONV3 ET TSNE

Représentation des produits par catégories réelles



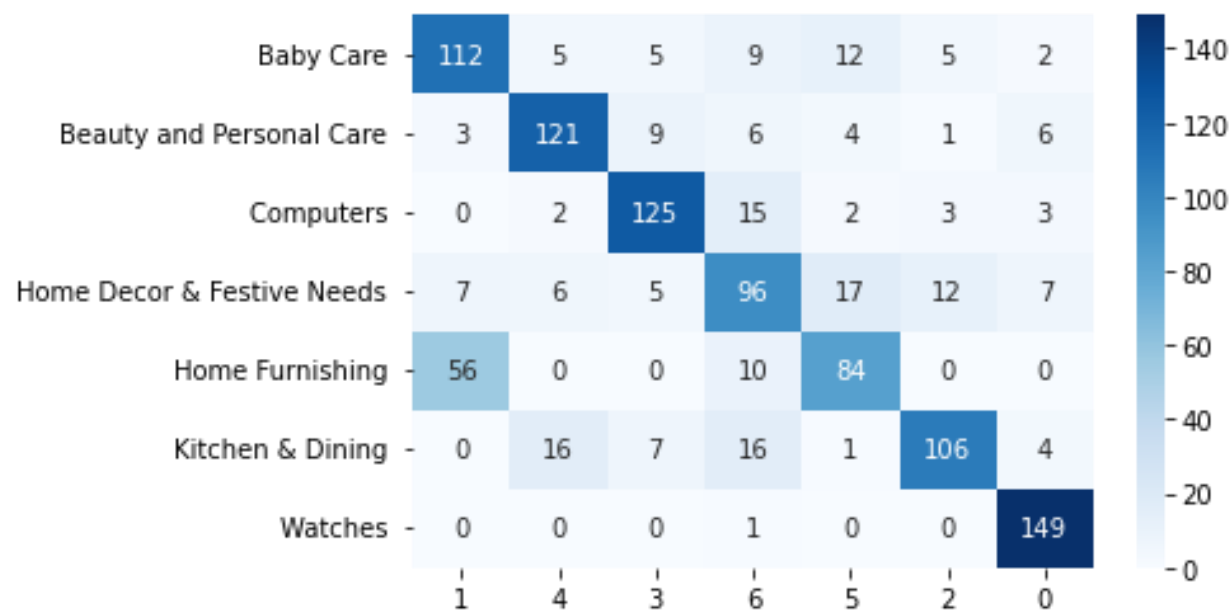
Représentation des produits par clusters



Meilleur modèle :
Réduction par TSNE + InceptionV3

ARI : 0.547

MATRICE DE CONFUSION

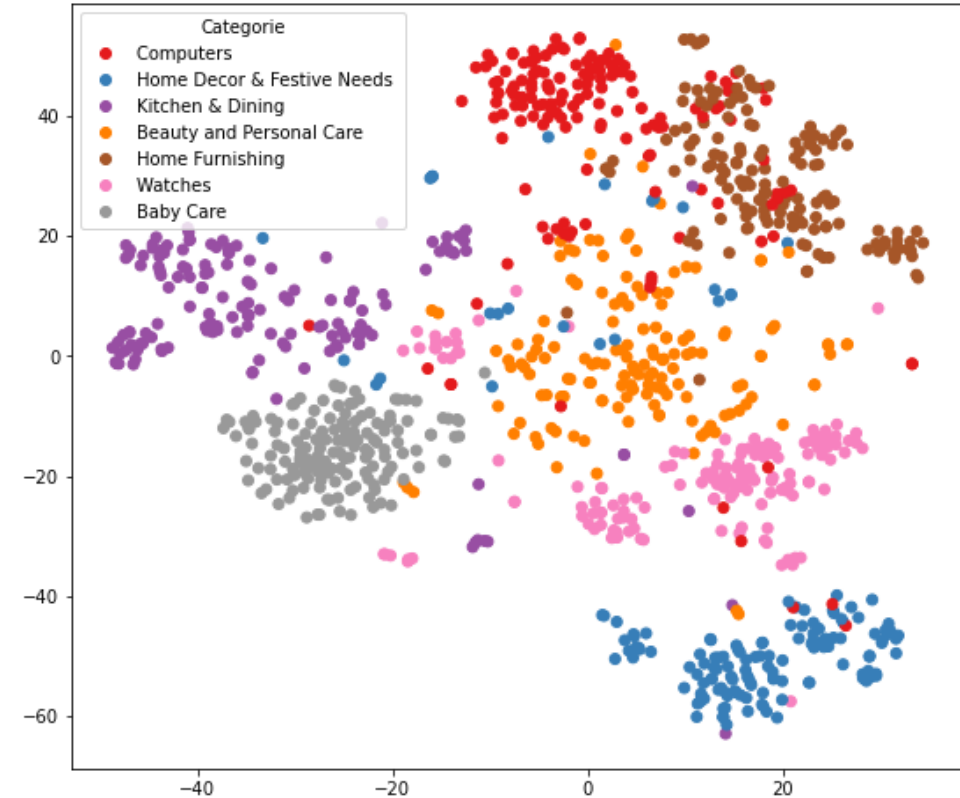


Erreurs principales :

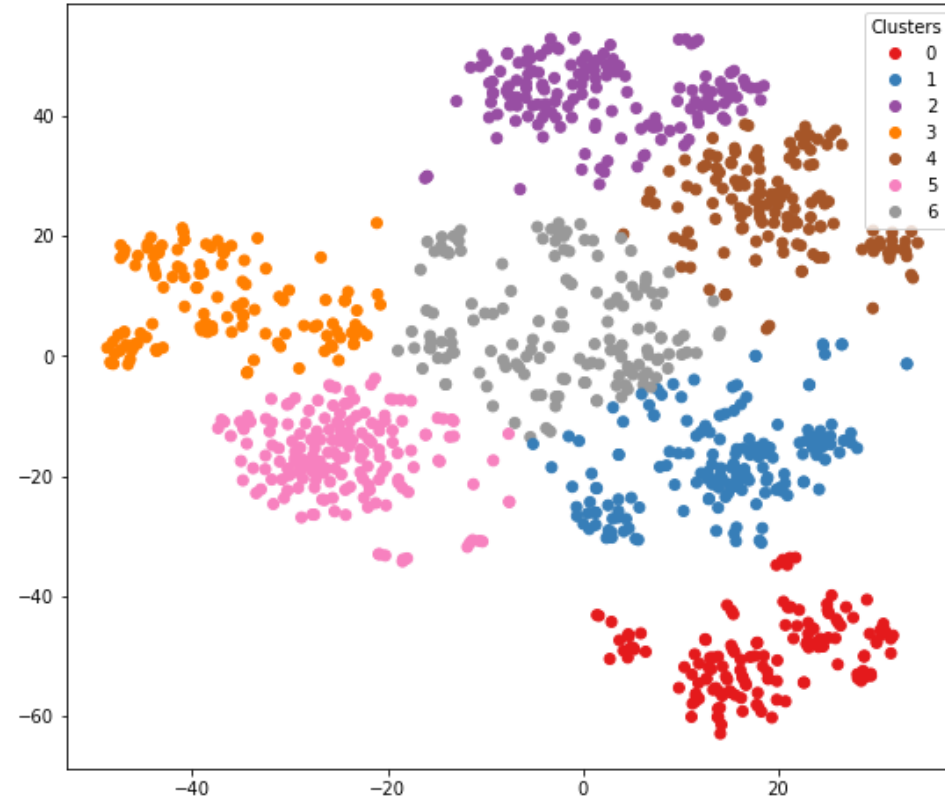
- Home Furnishing (56%)
- Home Decor & Festive Needs (64%)

COMBINAISON DES FEATURES TEXTES (SBERT) + IMAGES (INCEPTIONV3)

Représentation des produits par catégories réelles



Représentation des produits par clusters



Reduction par TSNE

ARI : 0,61

CONCLUSION

- Le meilleurs modèle pour l'analyse de texte est le S-BERT (ARI 0,69)
- Le meilleur modèle pour la classification d'image est le modèle InceptionV3 (ARI 0,55)
- Une combinaison des features des deux approches n'est pas plus performante
- Les données permettent une segmentation non supervisée

MOTEUR DE CLASSIFICATION

■ Rapide :

- Embedding par le modèle le plus efficace
- Application d'un algorithme de classification

Features sur Texte : Reduction par TSNE + SBERT

Classification : Random Forest

Accuracy : 0,94 (paramètres de base)

■ Lente :

- Sélection des meilleurs modèles de réseaux de neurones (Textes et/ou Images)
- Utilisation d'un classifieur directement dans le modèle
- Fine tuning partiel (entraînement de certaines couches sur nos données)