#6 Classifiez automatiquement des biens de consommation

Soutenance Emilie Groschêne le 03/03/2023
Evaluateur: Alexandre Gazagnes
Mentor: Lea Naccache

Sommaire

I. PROBLEMATIQUE

I. Présentation de la problématique

❑ Mission:

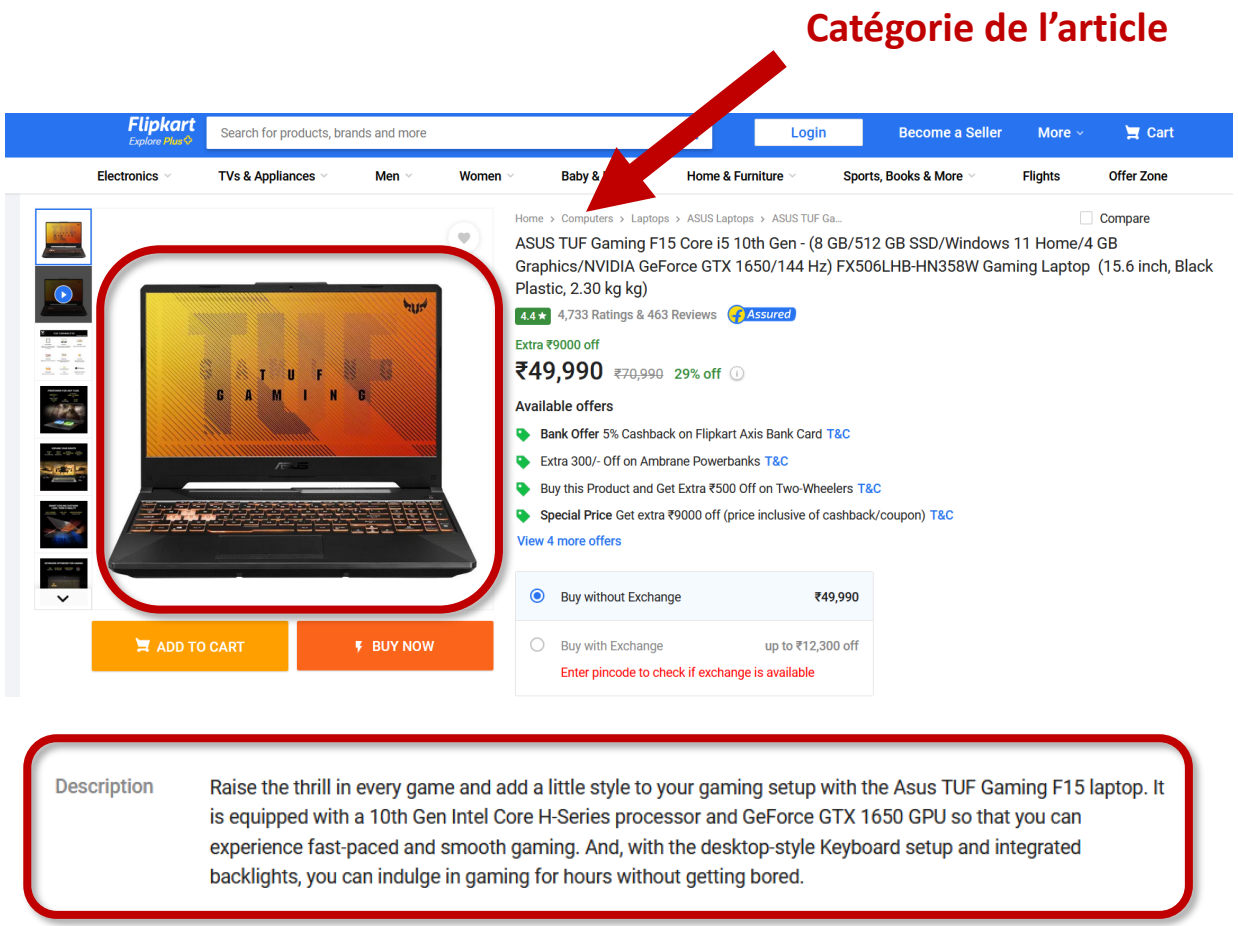
Réaliser une **étude de faisabilité** d'un moteur de classification d'articles, basé sur une **image** et une **description**, pour l'automatisation de l'attribution de la **catégorie** à l'article

❑ Objectifs :

Améliorer **l'expérience utilisateurs** (vendeurs et acheteurs)

Fiabiliser l'attribution des catégories aux articles

Perspective de **passage à l'échelle**



II. DONNEES

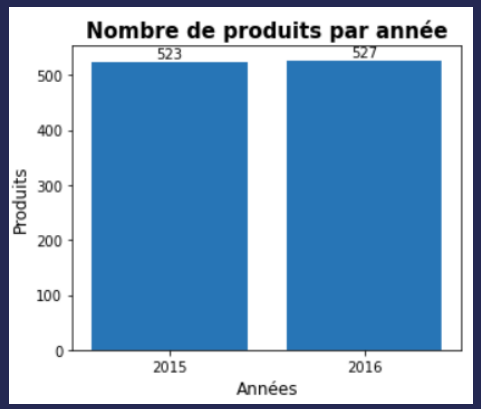
II. Présentation des données

❑ **Quelques statistiques:**

Nombre de lignes: 1 050
Nombre de colonnes: 15
% NaN dataset: 2,17%

Pas de doublon

Périodicité: 01/12/2015 au 26/06/2016



❑ Type des données:

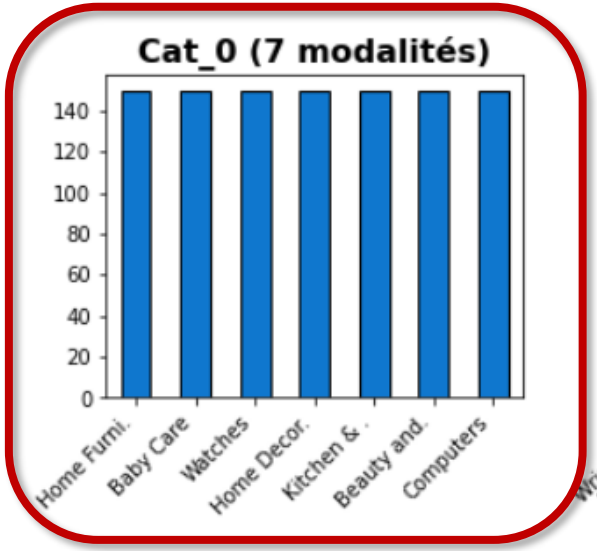
Données texte: nom du produit + description, en anglais, comportant des lettres et chiffres, longueur variable

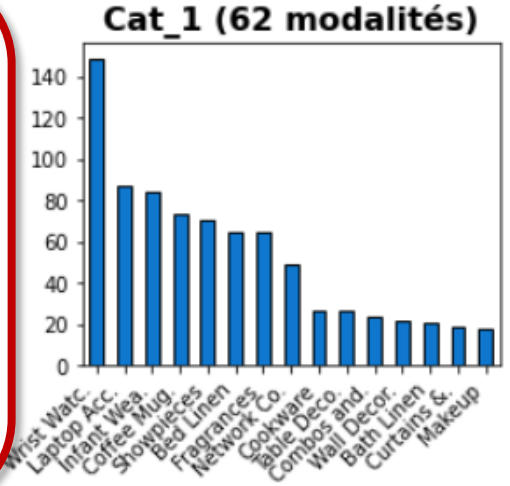
	product_name	description
75	Vitamins Embroidered Baby Girl's Denim Shorts	Specifications of Vitamins Embroidered Baby Girl's Denim Shorts General Details Ideal For Baby Girl's Occasion Casual Pattern Embroidered Shorts Details Type Denim Shorts Fabric Cotton Pockets Mitered Patch Pocket on Thigh Number of Contents in Sales Package Pack of 1 Fabric Care Gentle Machine Wash in Lukewarm Water, Do Not Bleach Additional Details Style Code 05TG-166-24-RAW RANI In the Box 1 SHORTS

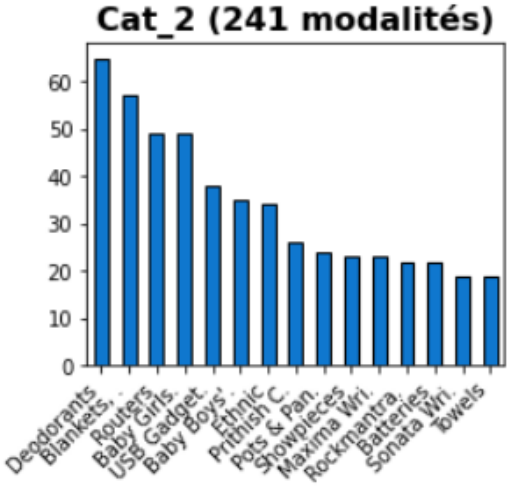
Données image: photo du produit, en couleur sur fond blanc au format jpg

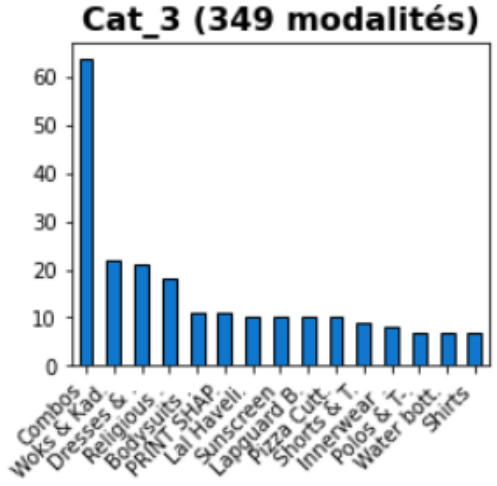
II. Présentation des données: les catégories

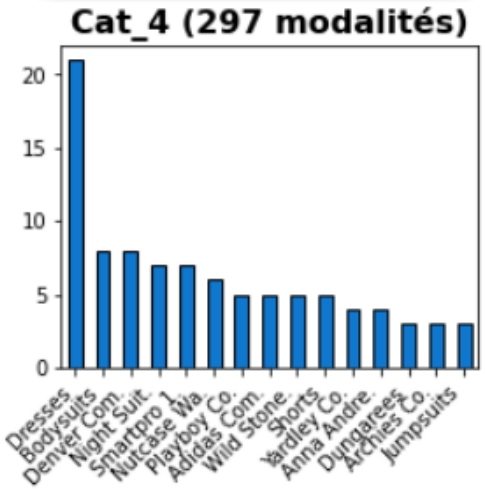
```
['["Kitchen & Dining >> Coffee Mugs >> Rockmantra Coffee Mugs"]',
 '["Kitchen & Dining >> Kitchen Tools >> Kitchen Implements >> Pizza Cutters >> King International Pizza Cutters"]']
```

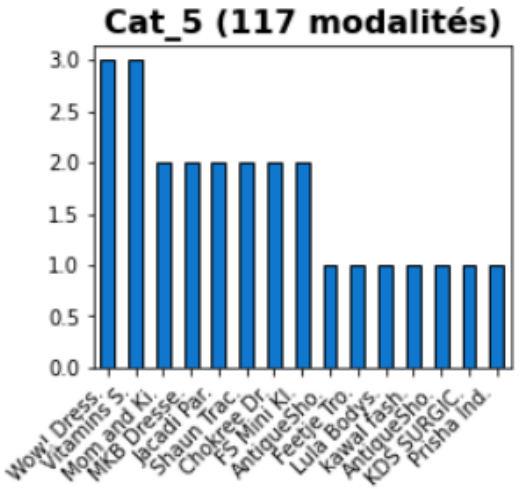


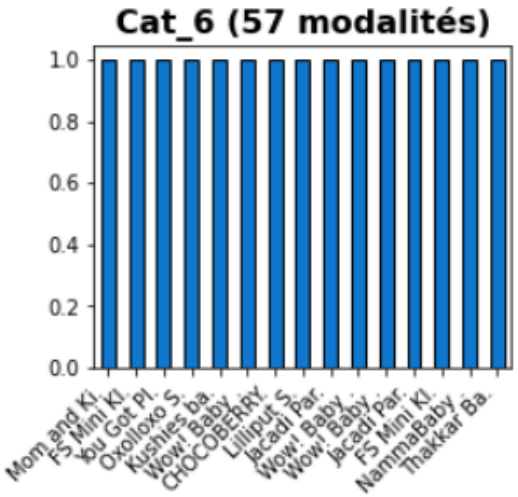












Catégories retenues:

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

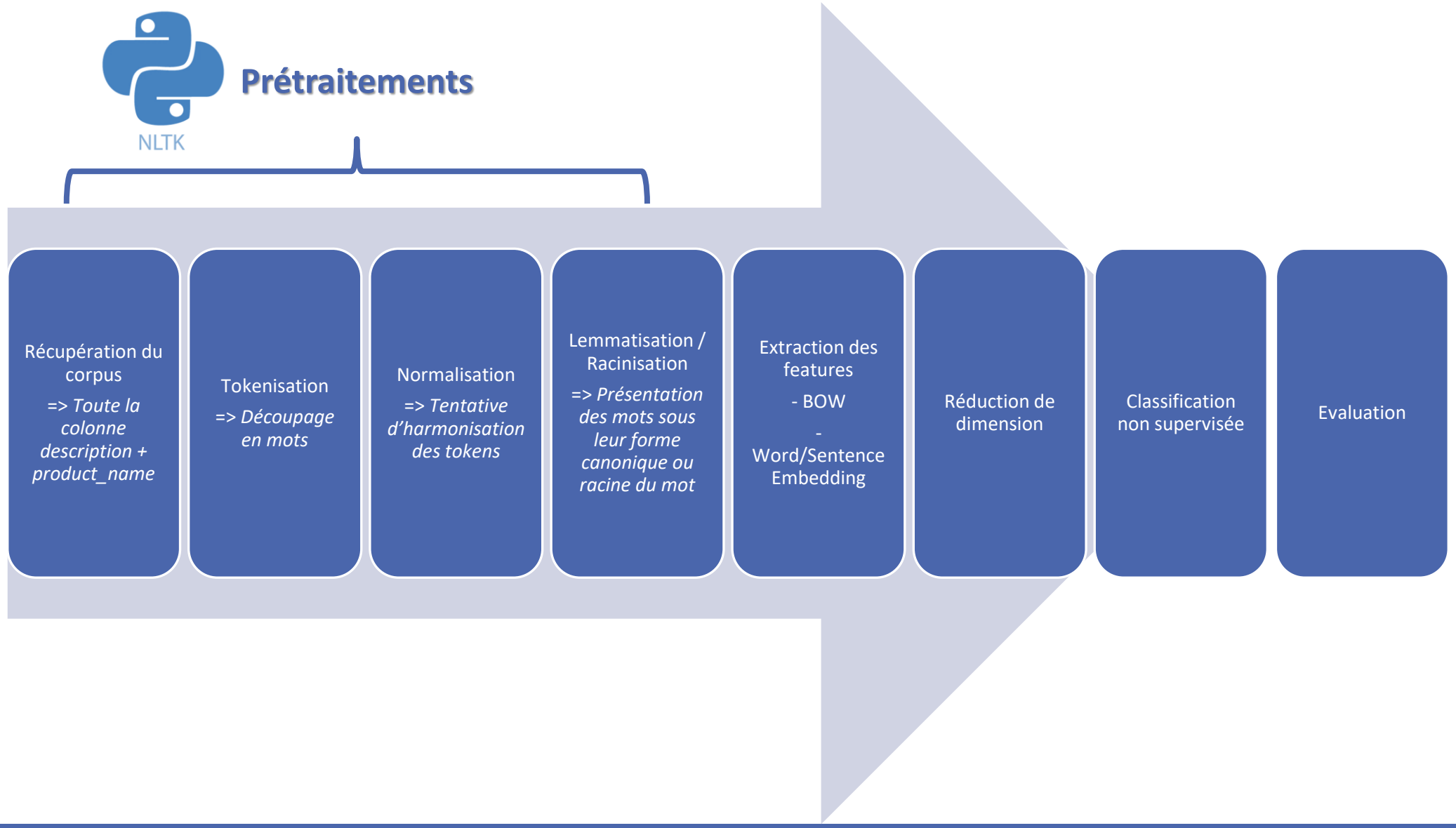
II. Présentation des données: sélection des features

- **uniq_id et product_name** => identifiants uniques du produit
- **Cat_0** => catégorie la plus équilibrée (150 produits par catégorie)
- **image** => nom de l'image associée au produit présente dans le dossier Images
- **description** => pour choisir la catégorie à laquelle affecter le produit

- **crawl_timestamp** => ne permet pas d'identifier un produit
- **product_url** => le lien ne fonctionne pas et cette rubrique est constituée des variables product_name et pid
- **pid** => pas d'information sur cette variable et non précise
- **brand** => marque du produit pouvant aider à son identification, déjà intégrée dans la description
- **retail_price et discounted_price** => pourrait permettre de classer un produit en fonction de son prix mais ce n'est pas le travail demandé ici
- **is_FK_Advantage_product** => ne permet pas d'identifier un produit
- **product_rating et overall_rating** => ne permet pas d'identifier un produit
- **product_specifications** => redondant avec la colonne description et moins complet

III. PRETRAITEMENTS DONNEES TEXTUELLES ET RESULTAT DU CLUSTERING

III. Prétraitements données textuelles - Pipeline



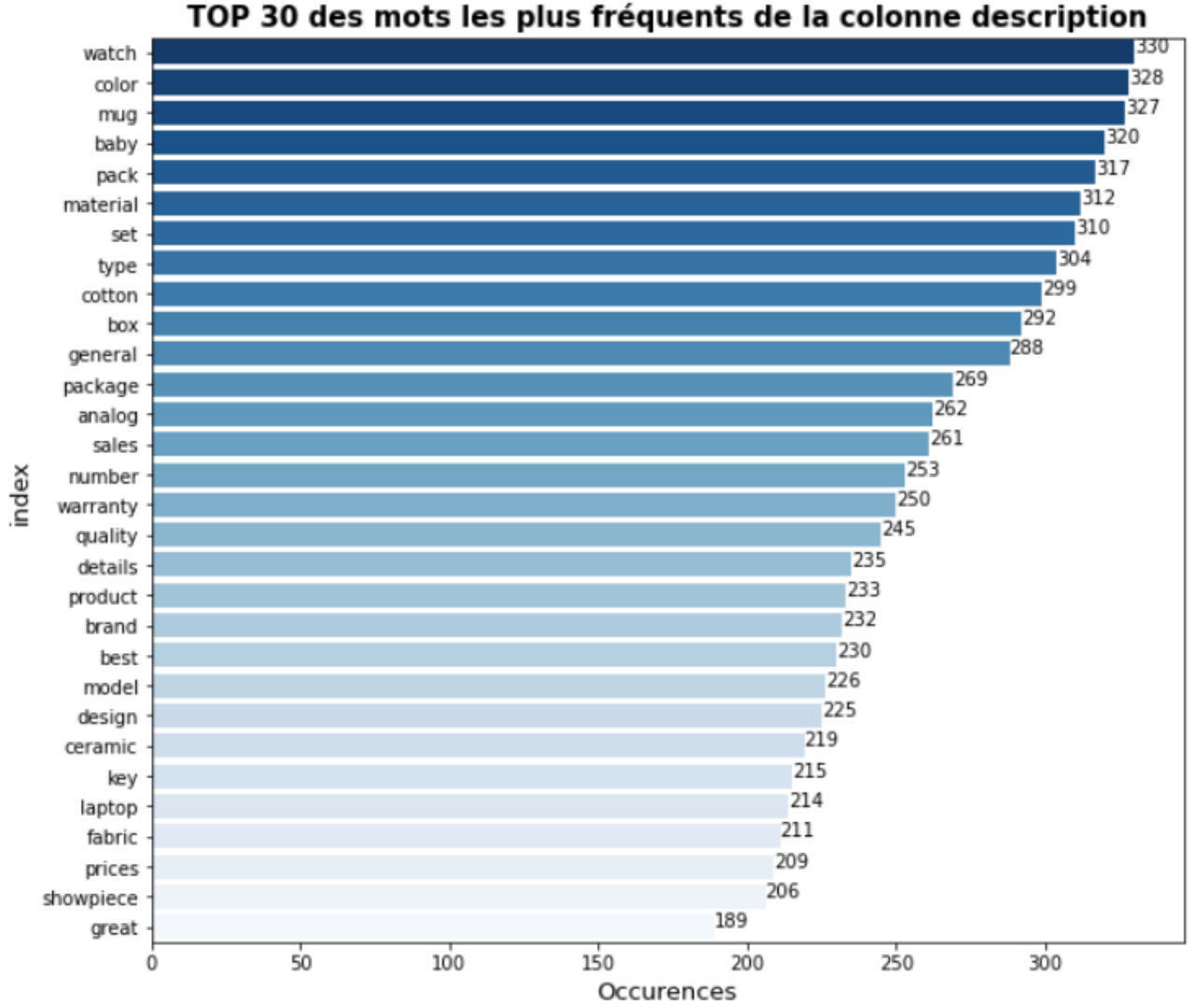
III. Prétraitements données textuelles – fréquence des mots

Plus fréquents: n'apportent **pas de valeur informative**. Seront considérés comme des **stopwords**:

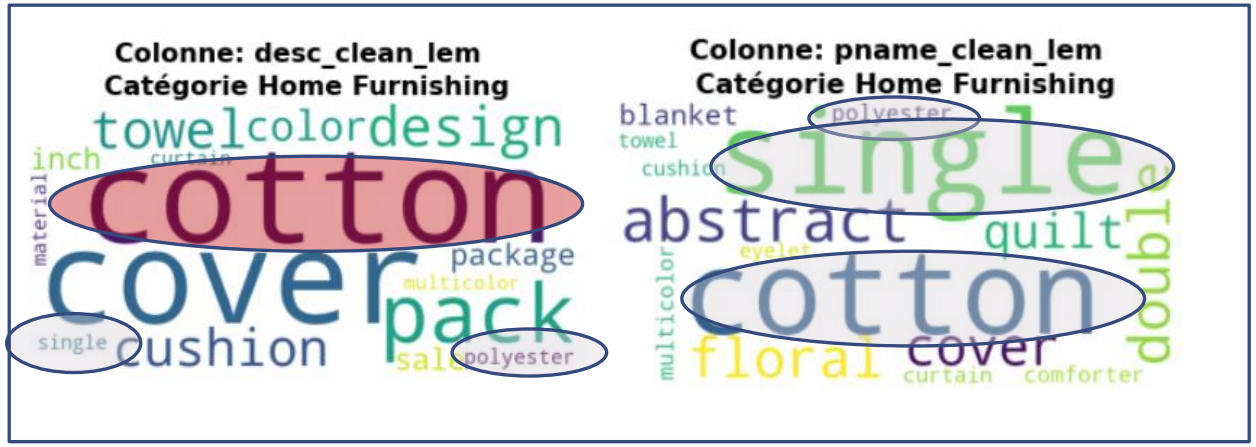
Rs, products, free, buy, delivery, cash, shipping, genuine, replacement, cm, day, flipkart, com …

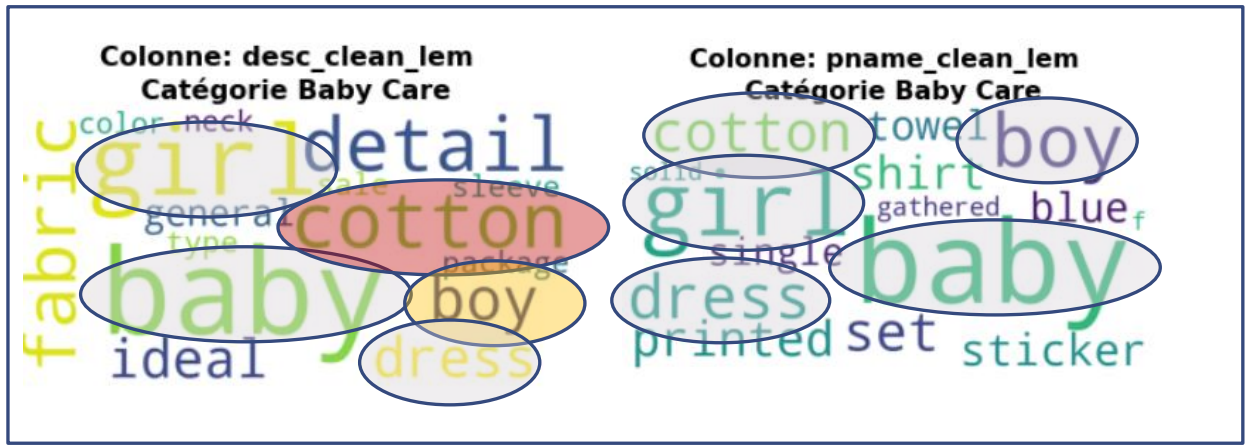
Moins fréquents: seront également **écartés**:

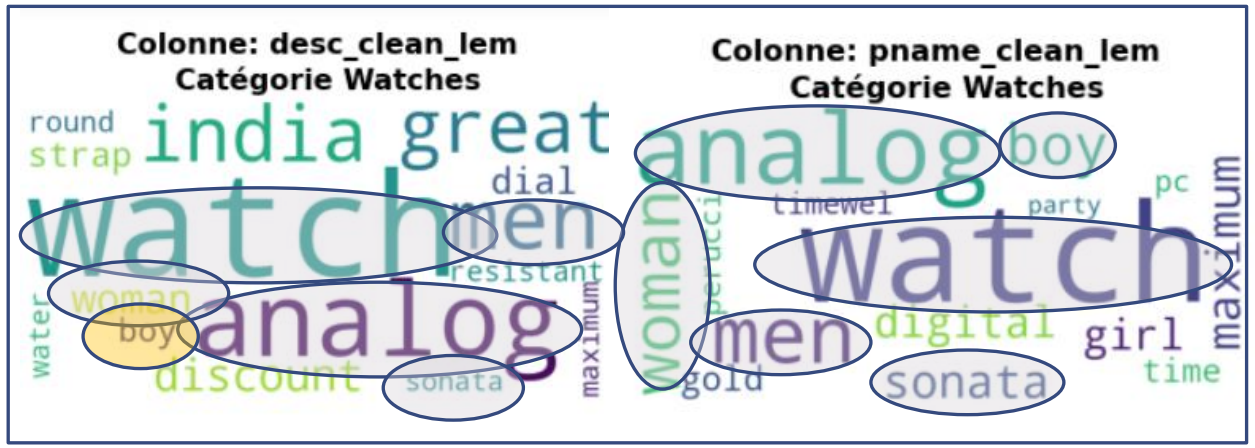
Dropping, mx, acer, structure, covering, liabilities, opportunity, loss, derivatives, woman, ingress …

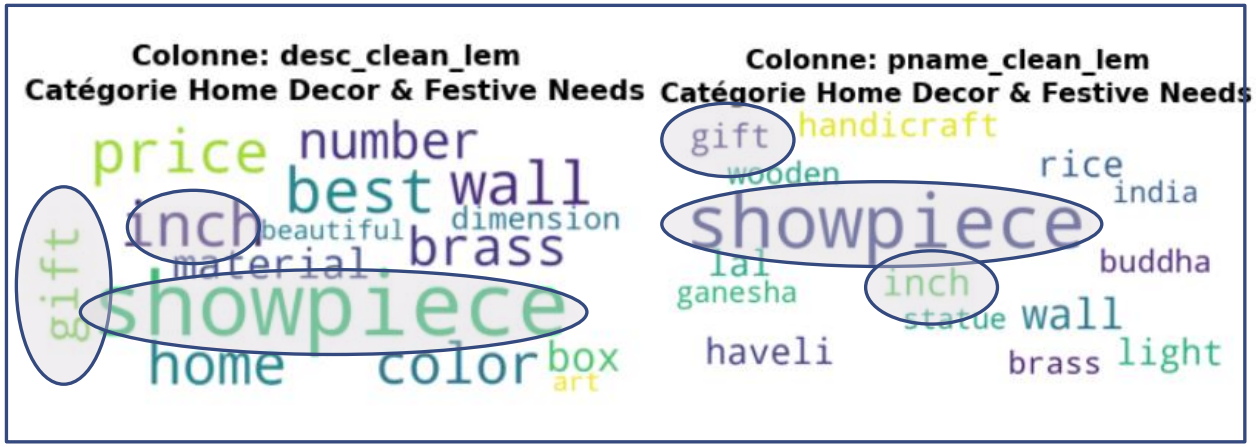


III. Prétraitements données textuelles – fréquence des mots

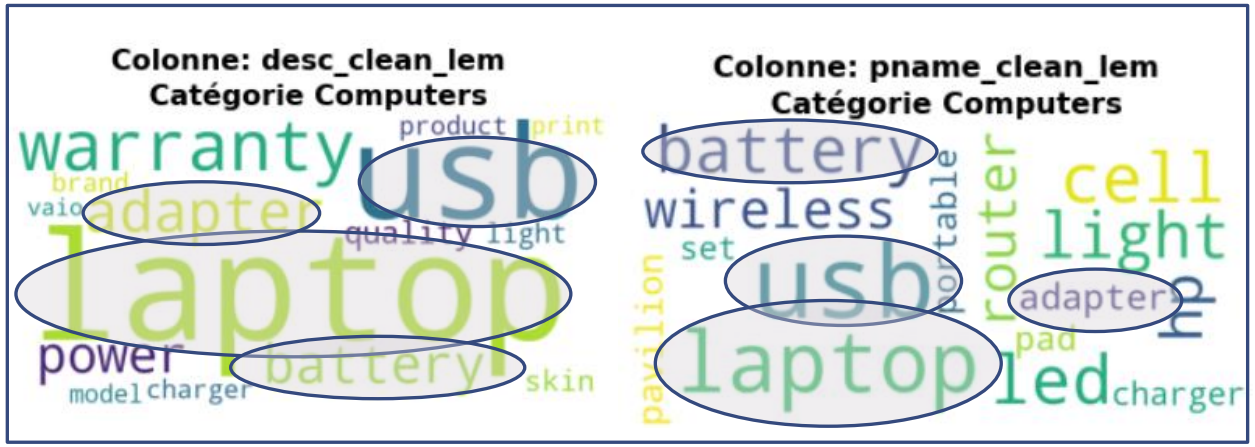


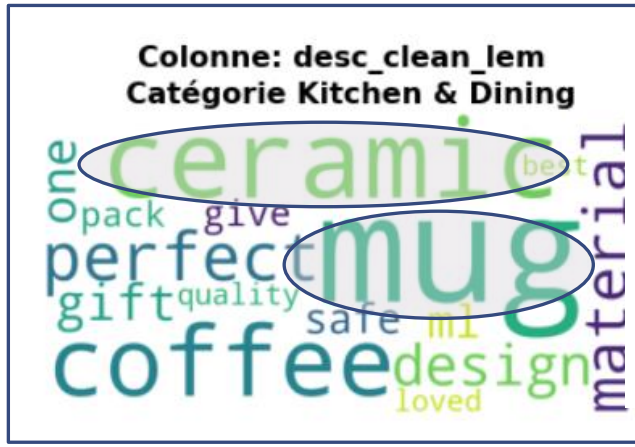


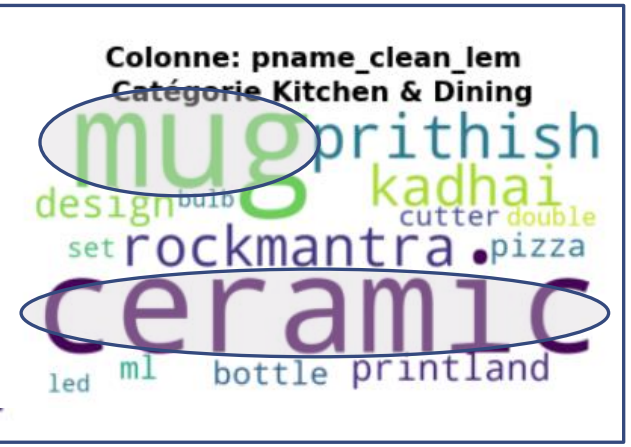


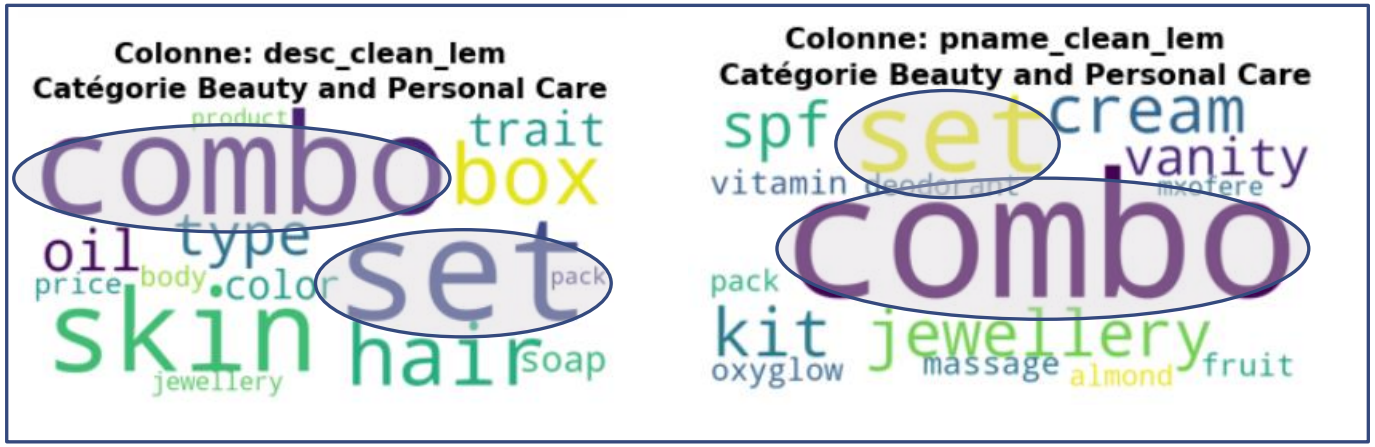


III. Prétraitements données textuelles – fréquence des mots





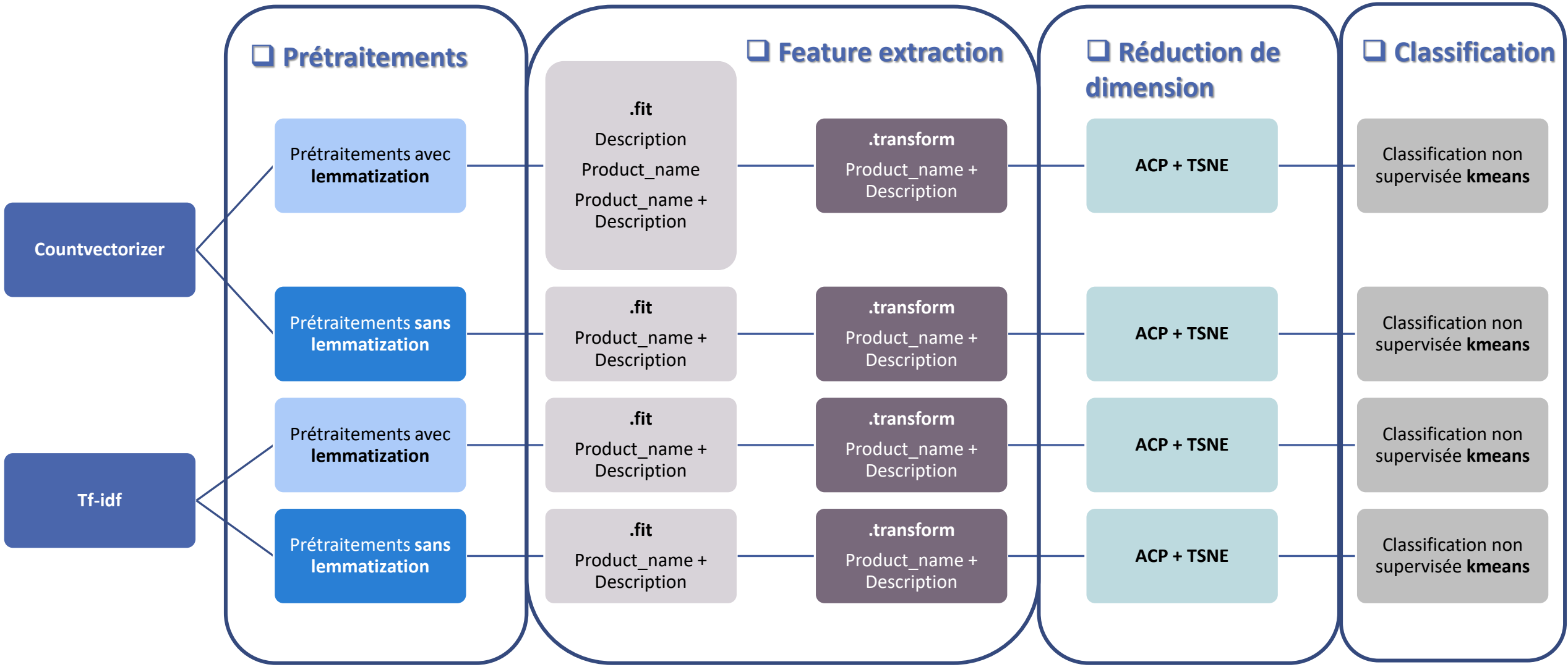




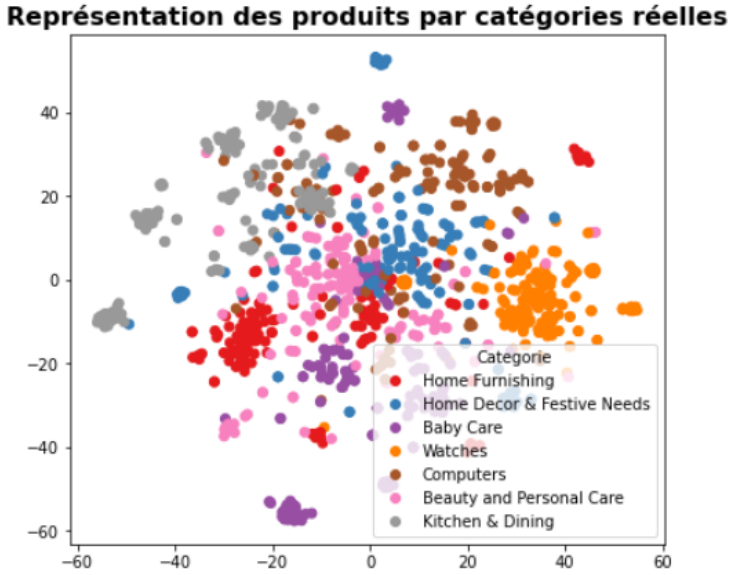
III. Prétraitements données textuelles - exemples

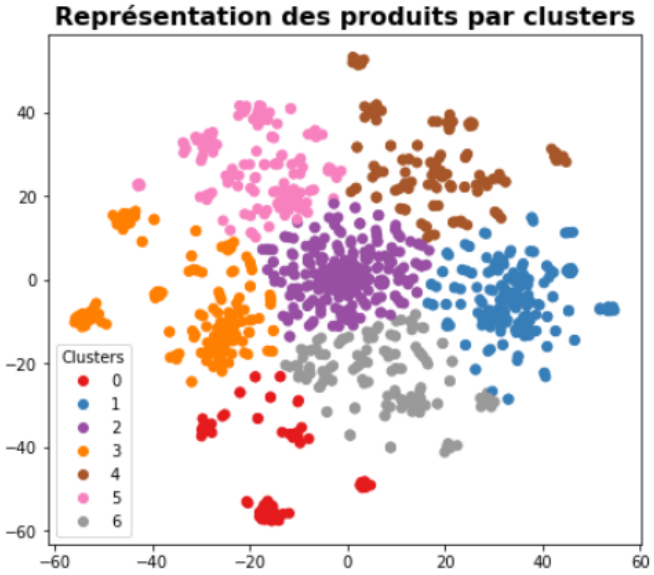
Prétraitements	Exemples	Nb de tokens	Nb de tokens uniques
Corpus	'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric. It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.'		
Tokenization	['Key', 'Features', 'of', 'Elegance', 'Polyester', 'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain', 'Floral', 'Curtain', 'Elegance', 'Polyester', 'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain', '213', 'cm', 'in', 'Height', 'Pack', 'of', '2', 'Price', 'Rs', '899', 'This', 'curtain', 'enhances', 'the', 'look', 'of', 'the', 'interiors', 'This', 'curtain', 'is', 'made', 'from', '100', 'high', 'quality', 'polyester', 'fabric', 'It', 'features', 'an', 'eyelet', 'style', 'stitch', 'with', 'Metal', 'Ring', 'It', 'makes', 'the', 'room', 'environment', 'romantic', 'and', 'loving']	81 563	7670
Normalization	['key', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'floral', 'curtain', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'curtain', 'enhances', 'look', 'interiors', 'curtain', 'made', 'high', 'quality', 'polyester', 'fabric', 'eyelet', 'style', 'stitch', 'metal', 'ring', 'makes', 'room', 'environment']	44 050	3 427
Lemmatization	['key', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'floral', 'curtain', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'curtain', 'enhances', 'look', 'interior', 'curtain', 'made', 'high', 'quality', 'polyester', 'fabric’ , 'eyelet', 'style', 'stitch', 'metal', 'ring', 'make', 'room', 'environment']	44 050	3 155

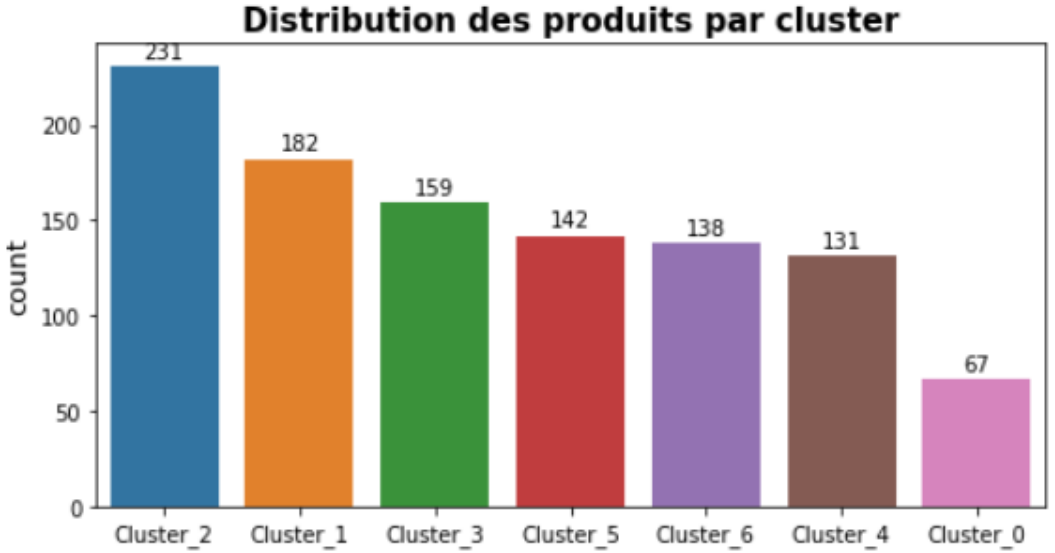
III. Prétraitements données textuelles – Approches de type Bag of Word



III. Prétraitements données textuelles – Approches de type Bag of Word - Countvectorizer



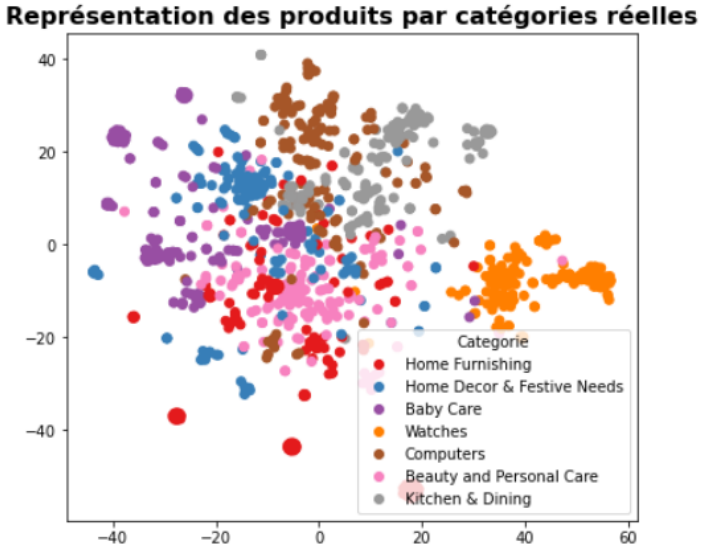


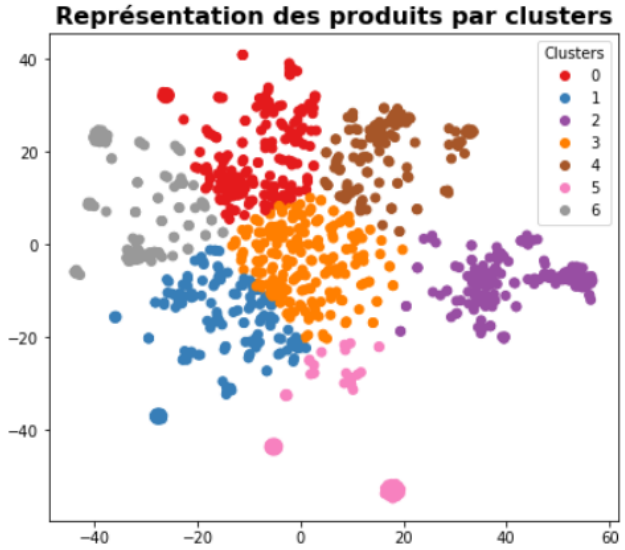


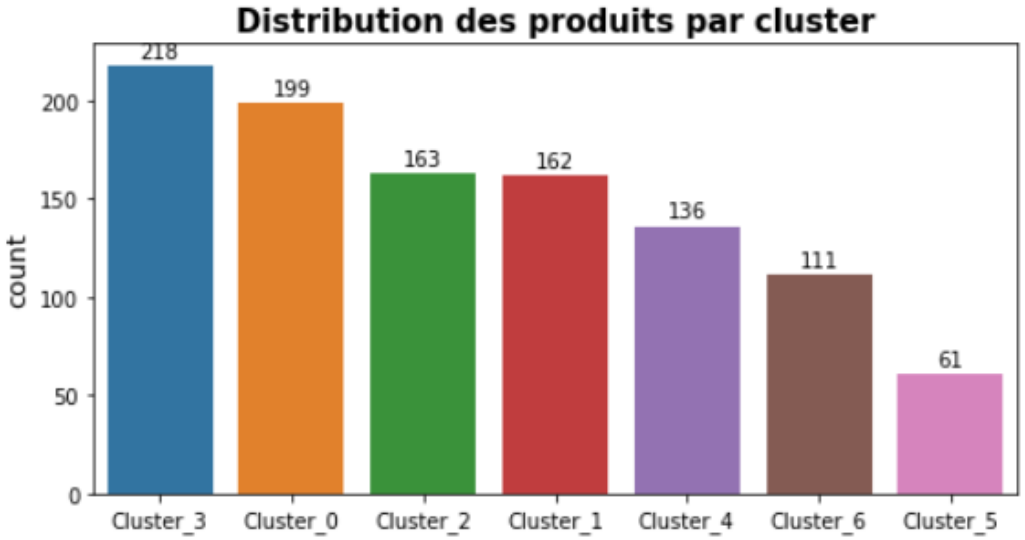
ARI : 0.3064

Meilleur modèle: données lemmatisées (extraction features colonne product_name)

III. Prétraitements données textuelles – Approches de type Bag of Word – Tf-idf



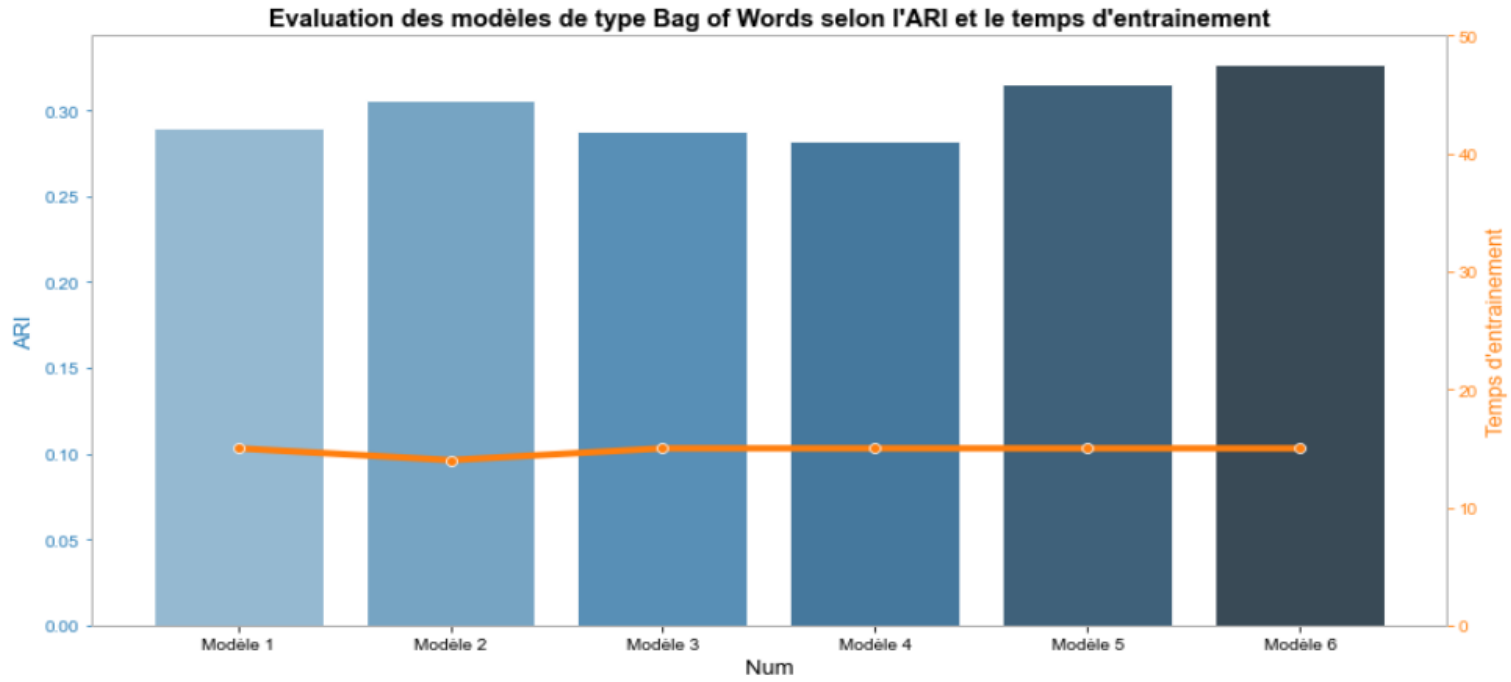




ARI : 0.3273

❑ **Données non lemmatisées (extraction features colonne product_name + description)**

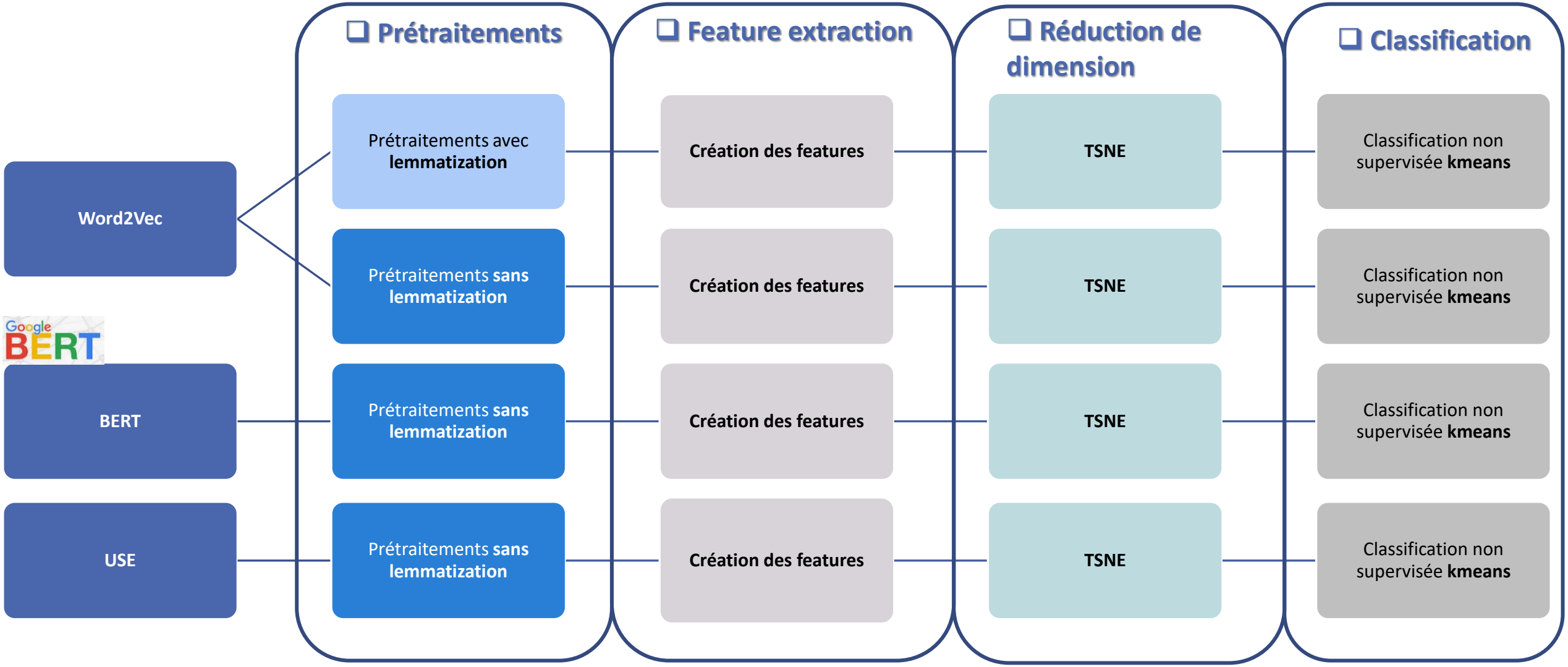
III. Prétraitements données textuelles – Approches de type Bag of Word – Evaluation des modèles



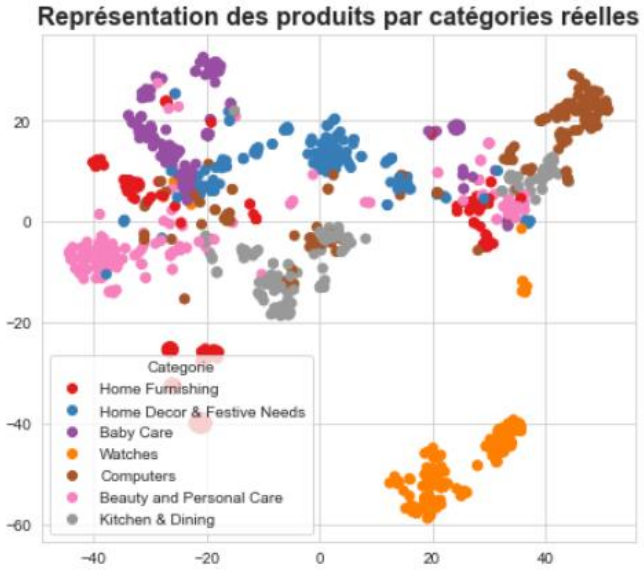
	Num	Modèle	Preprocessing	.fit	.transform	ARI	Temps d'entrainement
0	Modèle 1	Countvectorizer	Lemmatization	desc_clean_lem	pname_desc_clean_lem	0.2898	15.0
1	Modèle 2	Countvectorizer	Lemmatization	pname_clean_lem	pname_desc_clean_lem	0.3064	14.0
2	Modèle 3	Countvectorizer	Lemmatization	pname_desc_clean_lem	pname_desc_clean_lem	0.2883	15.0
3	Modèle 4	Countvectorizer	No Lemmatization	pname_desc_clean	pname_desc_clean	0.2826	15.0
4	Modèle 5	Tf-idf	Lemmatization	pname_desc_clean_lem	pname_desc_clean_lem	0.3155	15.0
5	Modèle 6	Tf-idf	No Lemmatization	pname_desc_clean	pname_desc_clean	0.3273	15.0

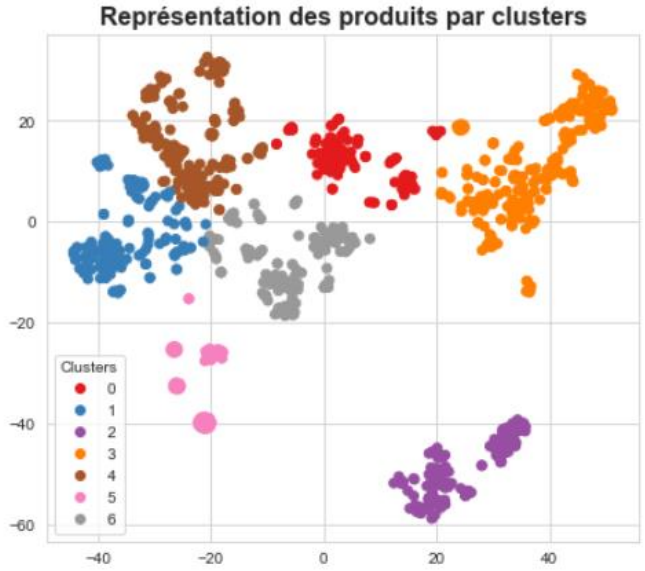
Target ARI: 0.4
=> Modèles de type BOW non retenus

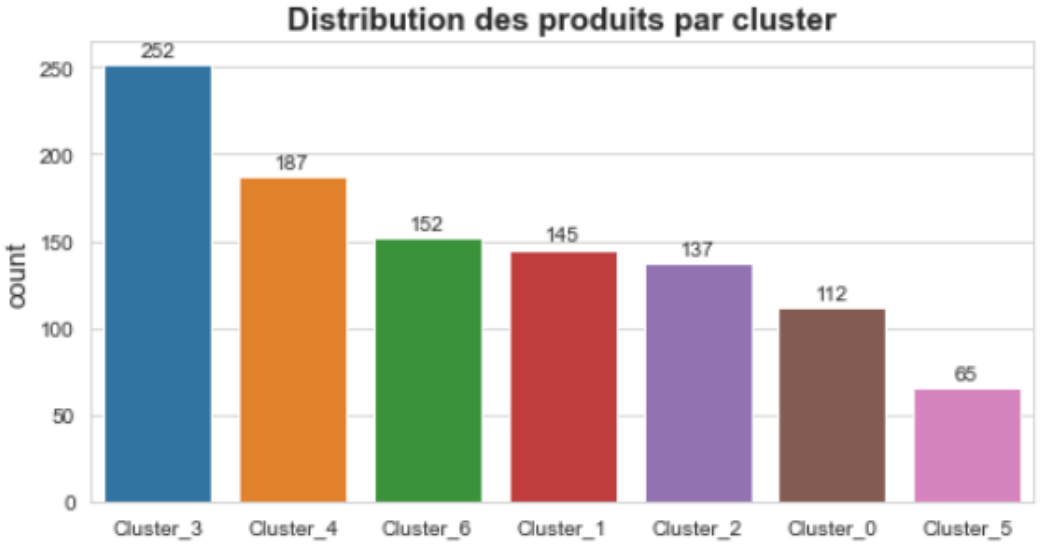
III. Prétraitements données textuelles – Approches de type Word/Sentence embedding



III. Prétraitements données textuelles – Approches de type Word/Sentence embedding – Word2Vec





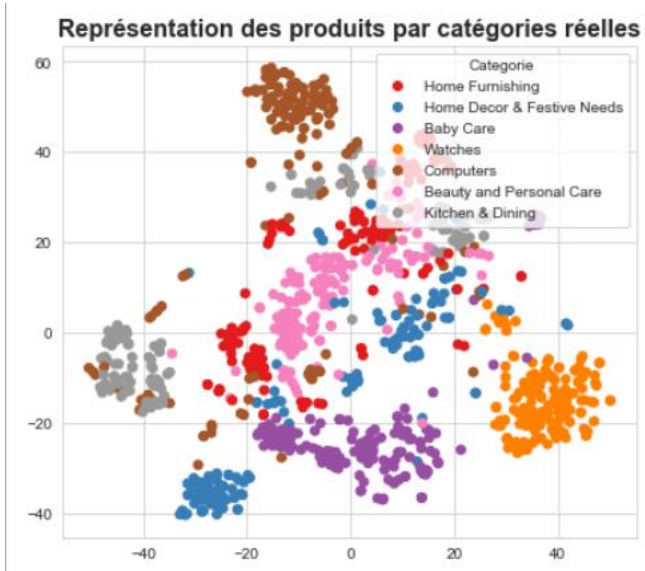


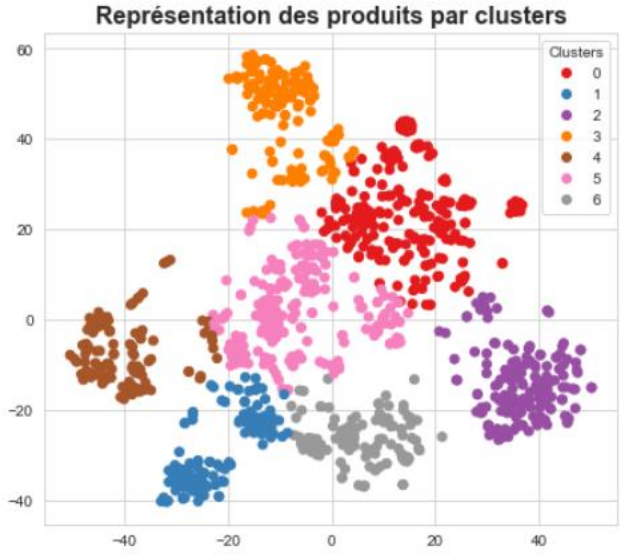
ARI : 0.427

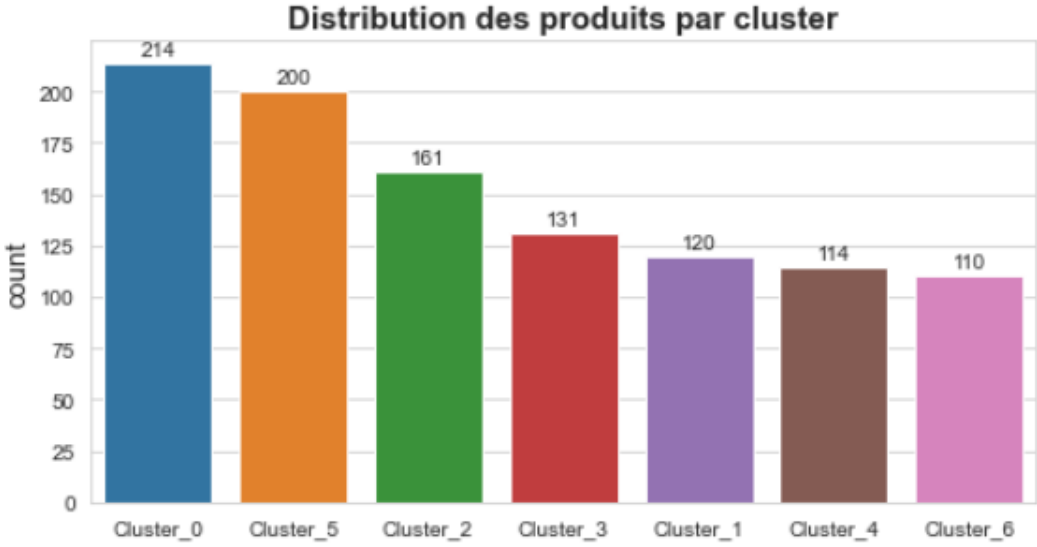
❑ **Données lemmatisées (extraction features colonne product_name + description et vector_size = 300)**

III. Prétraitements données textuelles – Approches de type Word/Sentence embedding – BERT

❑ Données non lemmatisées



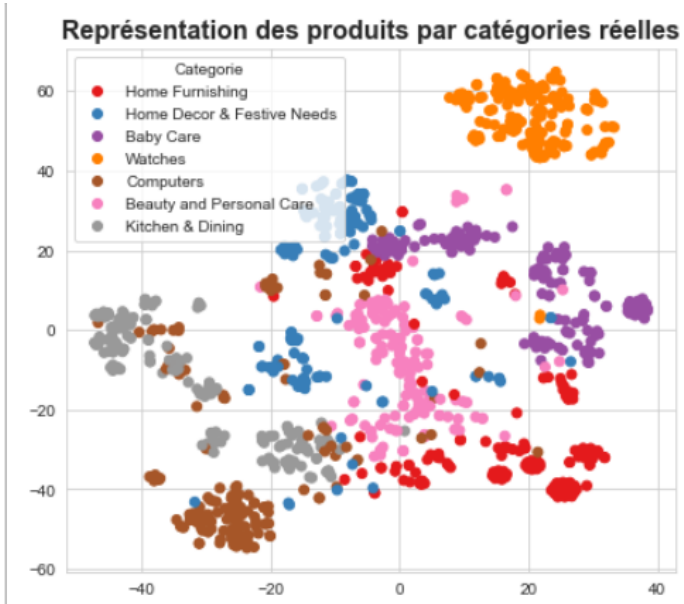


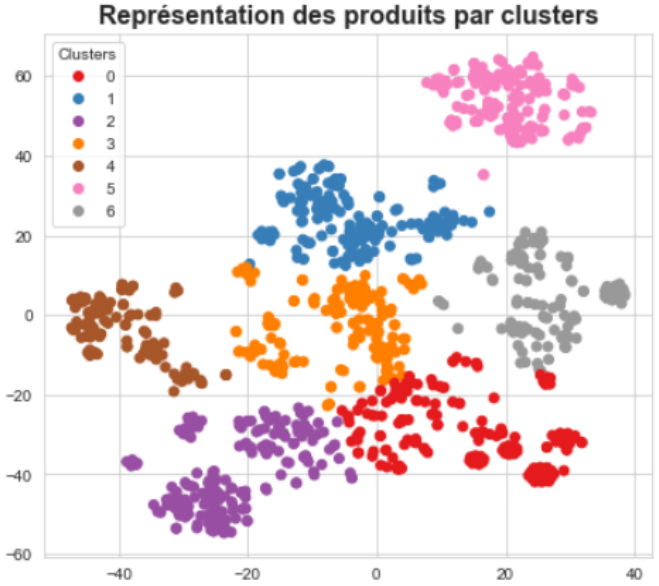


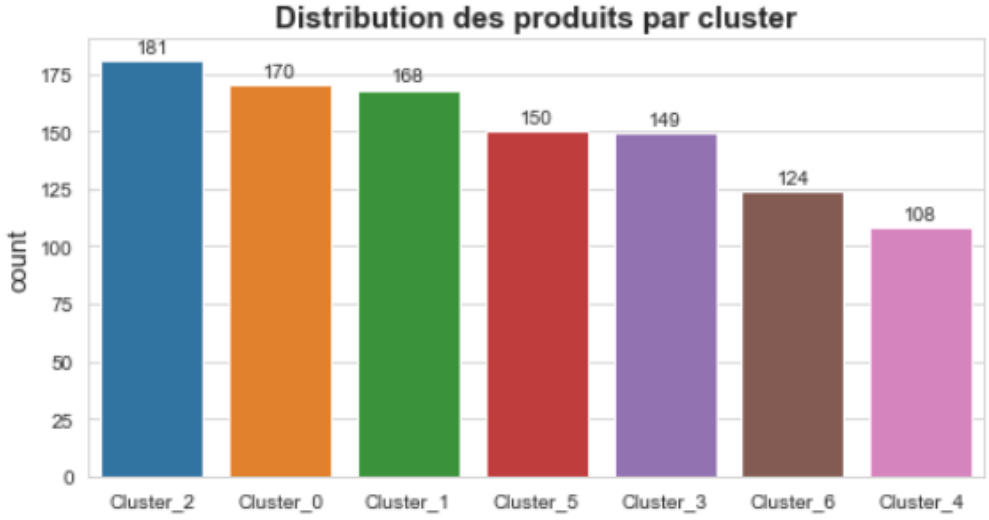
ARI : 0.4106

III. Prétraitements données textuelles – Approches de type Word/Sentence embedding – USE

❑ Données non lemmatisées

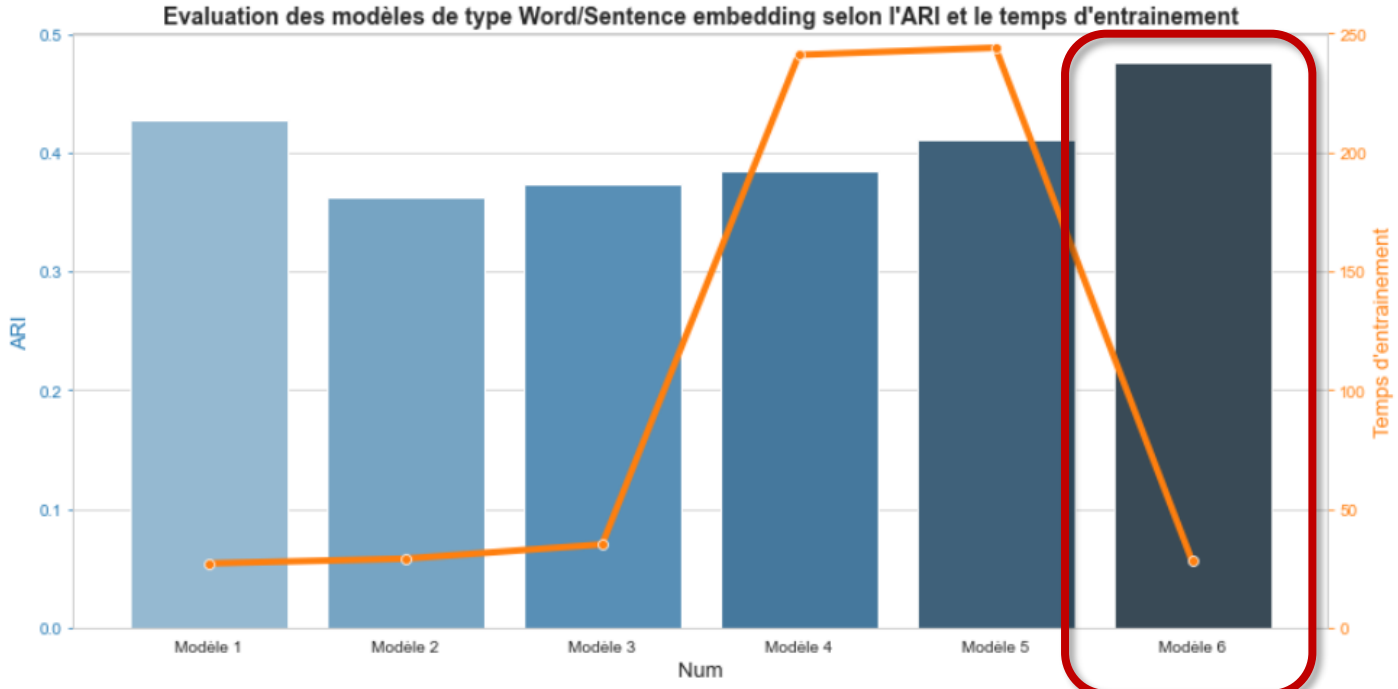






ARI : 0.4765

III. Prétraitements données textuelles – Approches de type Word/Sentence embedding – Evaluation des modèles

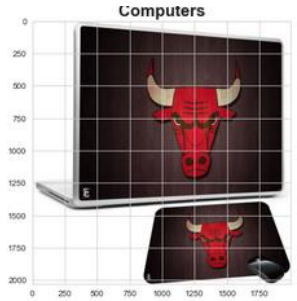


	Num	Modèle	Preprocessing	ARI	Temps d'entrainement
0	Modèle 1	Word2Vec size=300	Lemmatization	0.4270	27.0
1	Modèle 2	Word2Vec size=300	No Lemmatization	0.3631	29.0
2	Modèle 3	Word2Vec size=600	Lemmatization	0.3736	35.0
3	Modèle 4	Bert_HF	No Lemmatization	0.3843	241.0
4	Modèle 5	Bert_TF	No Lemmatization	0.4106	244.0
5	Modèle 6	USE	No Lemmatization	0.4765	28.0

Target ARI: 0.4
=> Modèles de type Word / Sentence embedding retenus

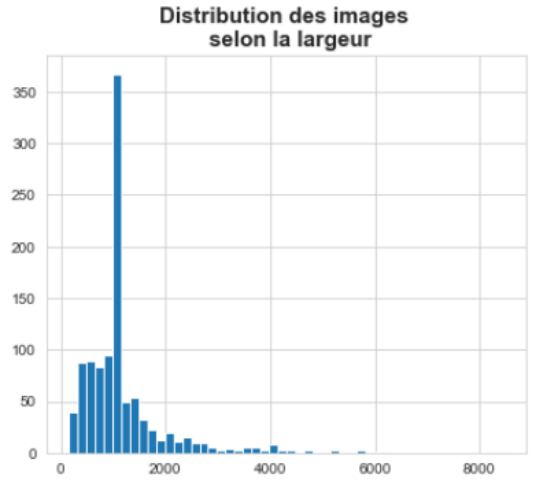
IV. PRETRAITEMENTS DONNEES IMAGES ET RESULTAT DU CLUSTERING

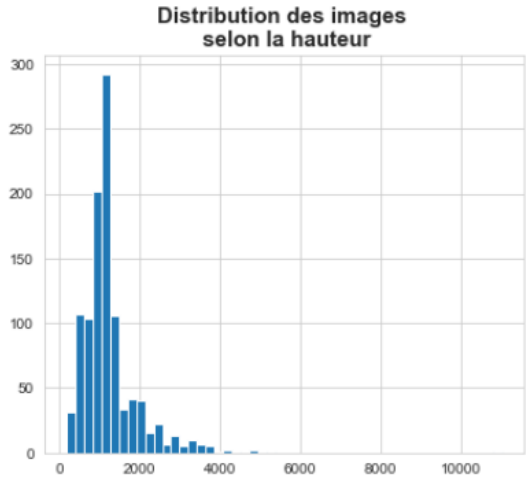
III. Prétraitements données images – Analyse exploratoire

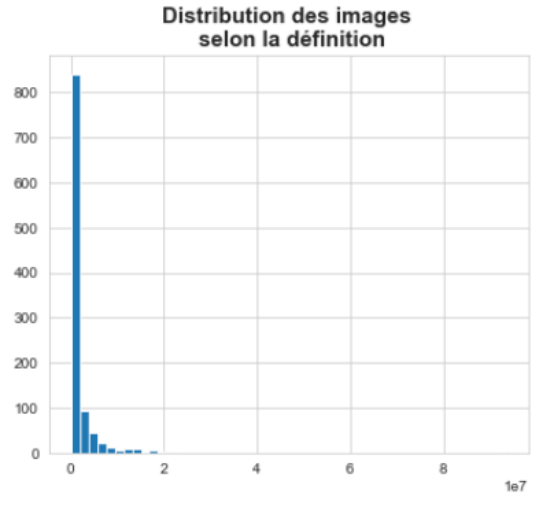


Cat_0	image
Baby Care	150
Beauty and Personal Care	150
Computers	150
Home Decor & Festive Needs	150
Home Furnishing	150
Kitchen & Dining	150
Watches	150

- 150 images par catégorie
- Images en **couleur**
- **Définitions** différentes

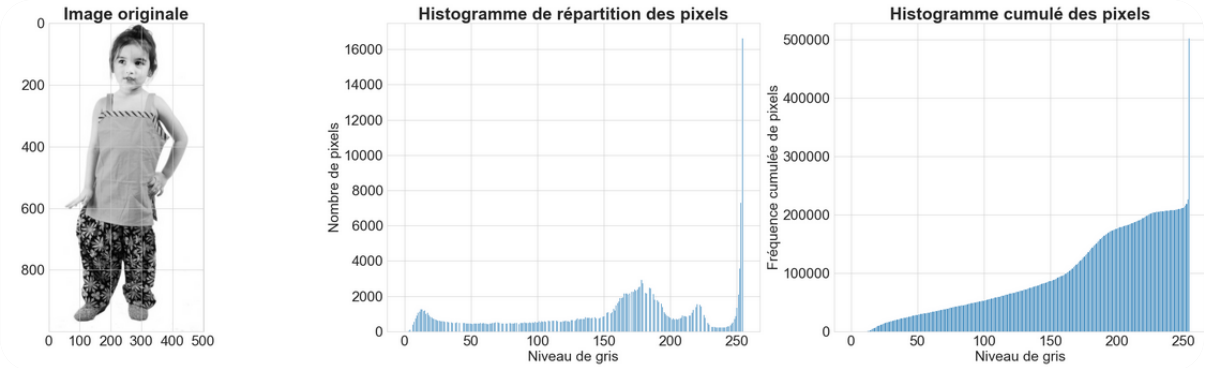




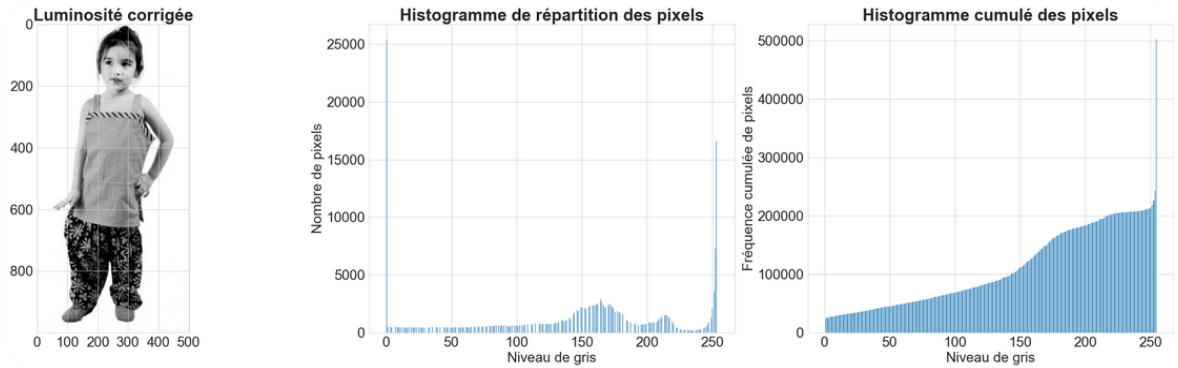


III. Prétraitements données images - Exemples

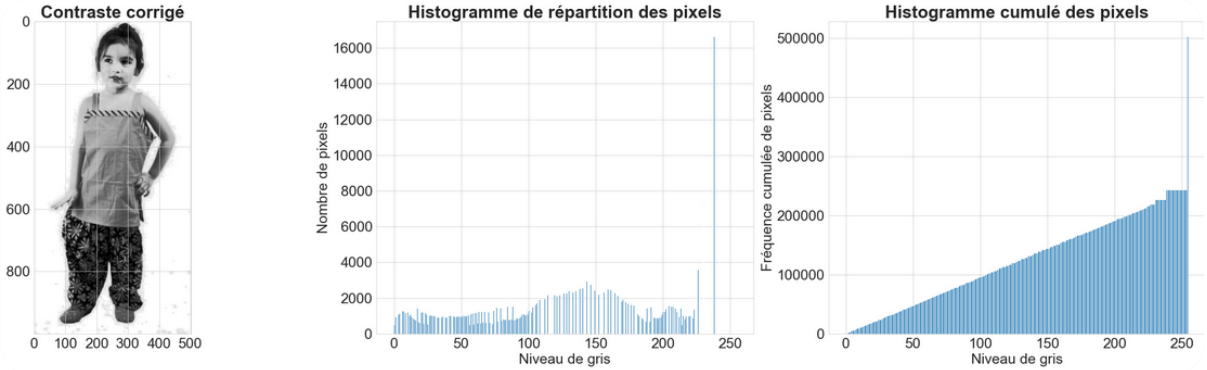
❑ Passage en gris:



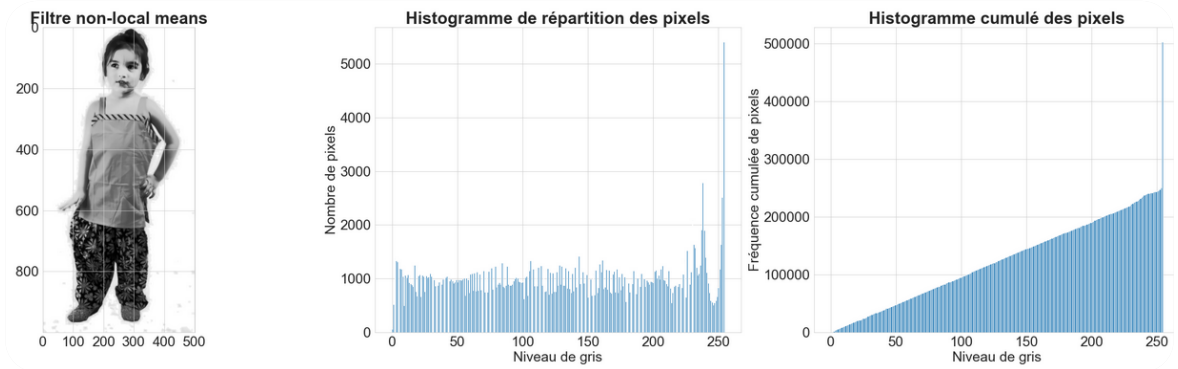
❑ Correction de la luminosité (étirement d'histogramme):



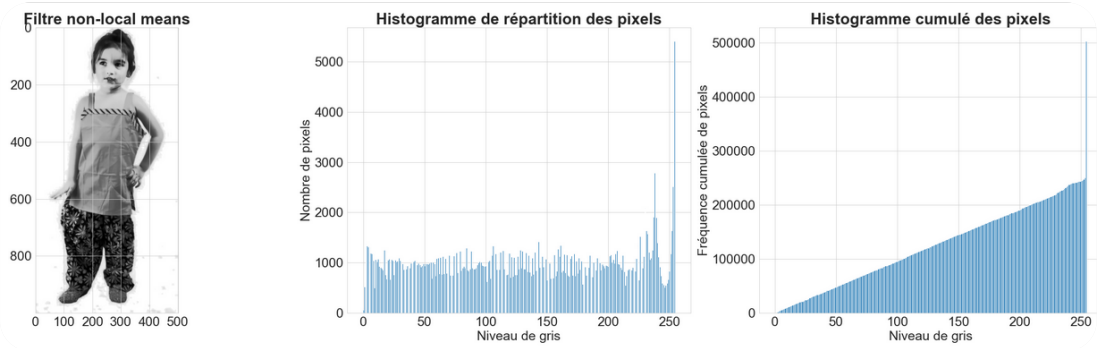
❑ Correction du contraste (égalisation d'histogramme):



❑ Elimination du bruit:



III. Prétraitements données images - Pipeline



Redimension nement 224 * 224

Conversion de l'image en gris

Correction du contraste

Elimination du bruit à l'aide du filtre non-local-means

Extraction des features

- SIFT
- VGG-16

Réduction de dimension

Classification non supervisée

Evaluation

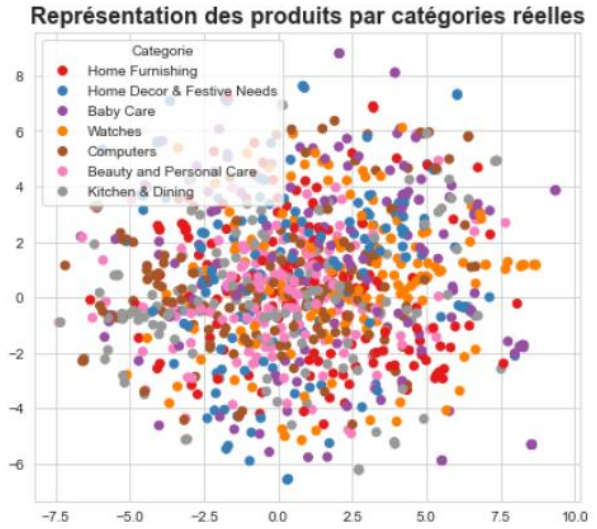
pillow

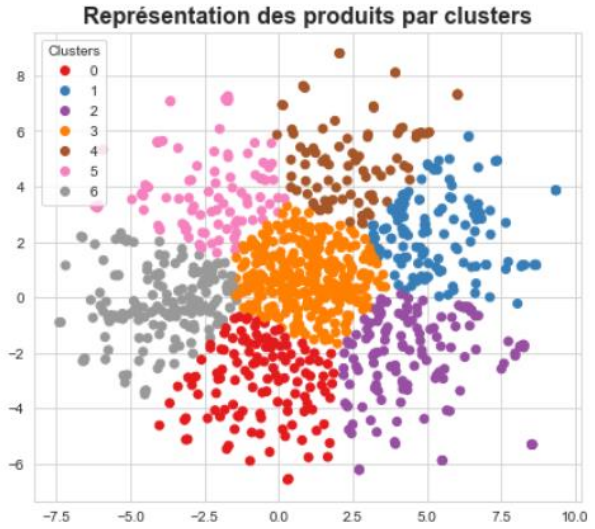
Prétraitements

OpenCV

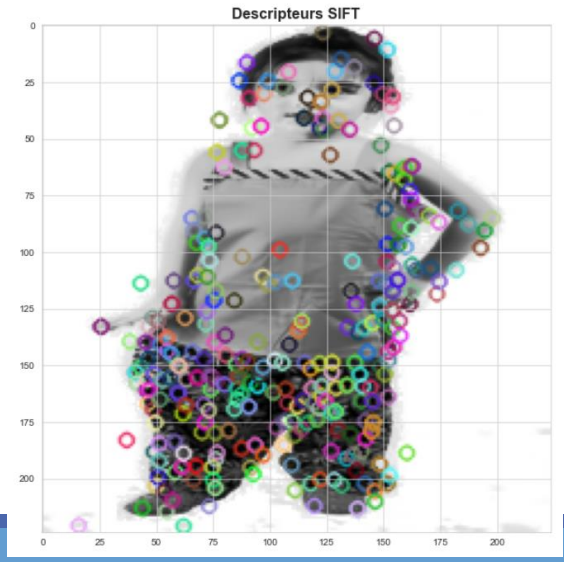
III. Prétraitements données images – SIFT

- Création des **descripteurs** de chaque image
- **Regroupement des descripteurs** en clusters pour diminuer le temps de calcul
- Création des **histogrammes** de clusters
- **Réduction de dimension** (ACP + TSNE)
- Classification non supervisée **kmeans** avec 7 clusters

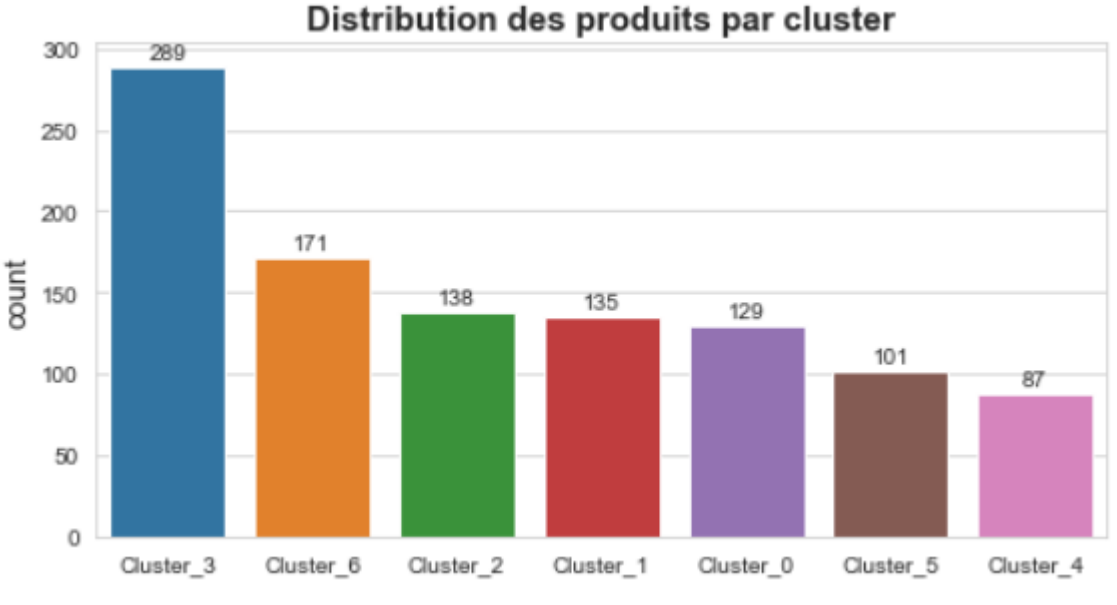





```
ARI :  0.0405
```

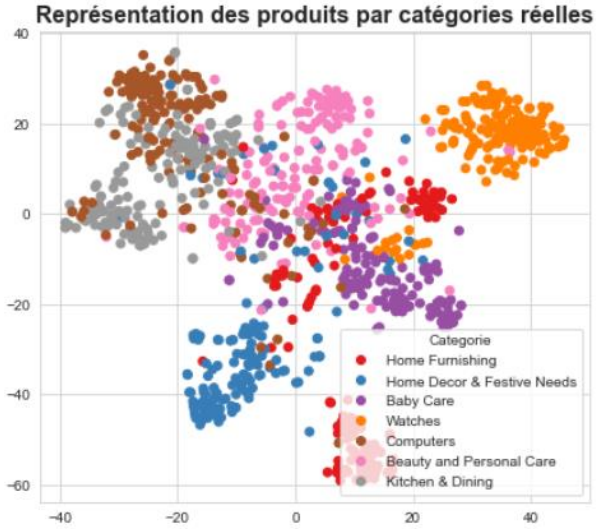


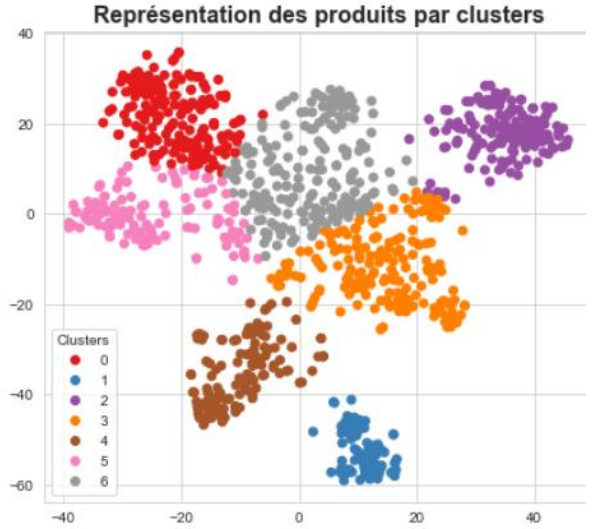

```
Descripteurs :  (339, 128)

[[ 88.  44.  21. ...   0.   0.   5.]
 [  0.   0.   0. ...   0.   1.  28.]
 [  0.   0.   0. ...   0.   0.  10.]
 ...
 [ 74.  38.  16. ...   0.   0.  19.]
 [  3.  73.  87. ...   0.   0.   6.]
 [ 85. 131.   0. ...   0.   0.   2.]]
```



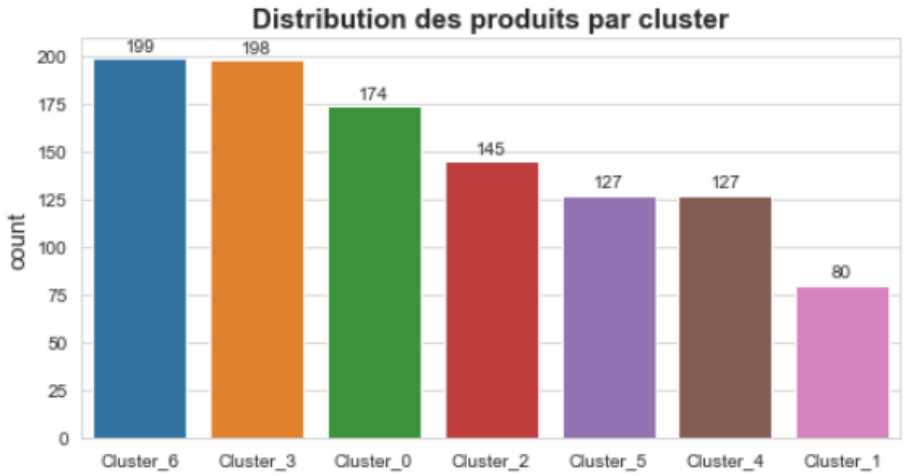
III. Prétraitements données images – VGG-16

- Utilisation d'un **réseau de neurones à convolution existant et pré-entrainé**
- On **utilise les connaissances acquises** par le réseau de neurone lors de la résolution d'un problème pour en résoudre un autre plus ou moins similaire
- La **première couche** de convolution apprend des **features simples** (contours, coin…)
- Plus les **couches** sont **hautes** plus les **features** apprises sont **complexes** (elles se composent des features plus simples des couches précédentes)
- On **retire la dernière couche fully-connected** (la couche de classification)

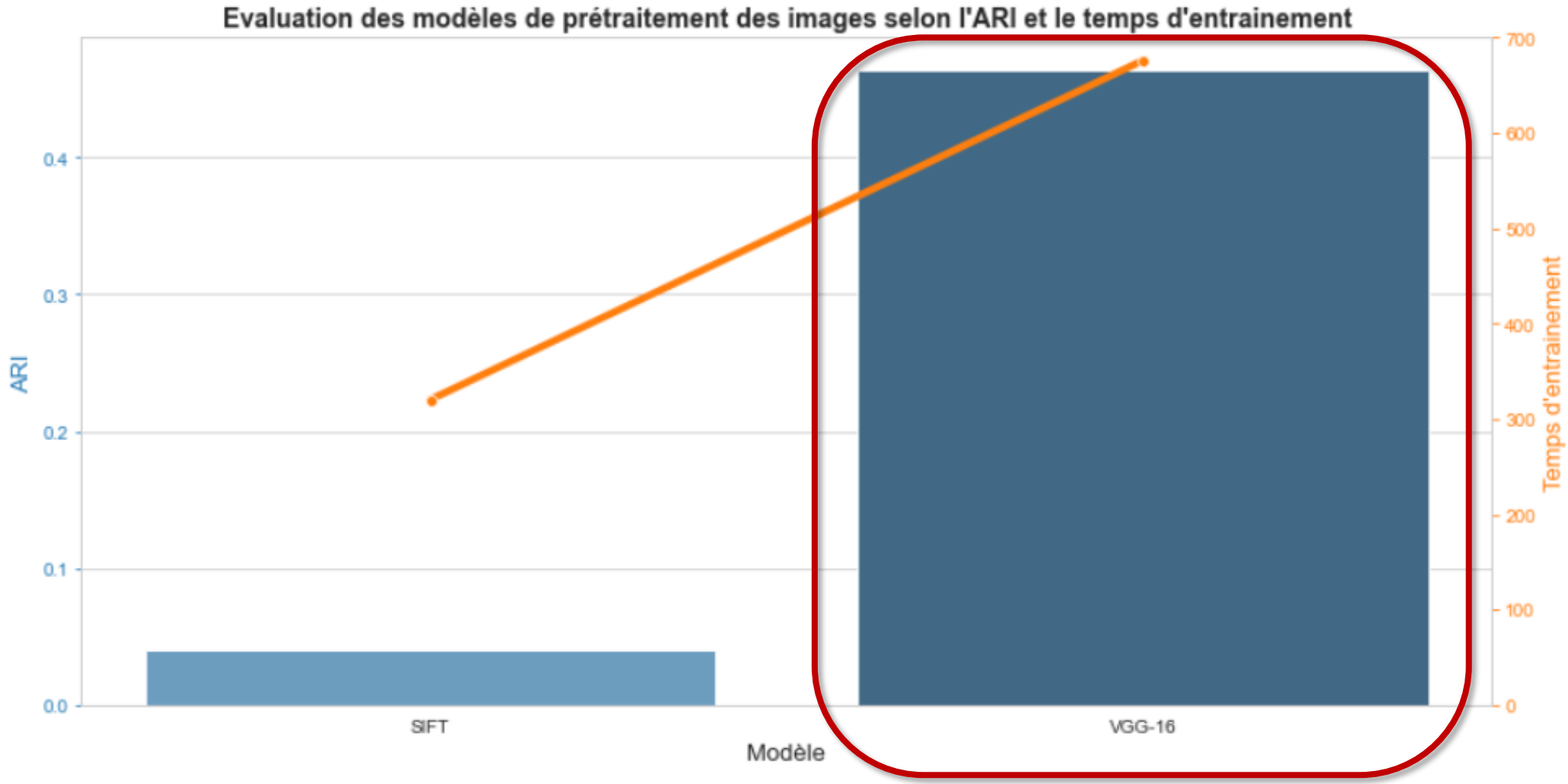




ARI : 0.4644



III. Prétraitements données images – Evaluation des modèles



V. CONCLUSIONS

V. Conclusions

L'étude de faisabilité est validée:

- ARI de 0,48 pour les données textuelles avec USE
- ARI de 0,47 pour les données images avec VGG-16

=> Scores prometteurs car très peu d'optimisation

Pistes à envisager pour l'implémentation:

- Ajout de produits (plus d'images et de texte) => point d'attention car split du jeu de données)
- Autres preprocessings des données à tester (racinisation, stopwords, chiffres conservés, rotation d'une image etc)
- Optimisation des hyperparamètres (GridSearchCV par exemple)
- Classification supervisée

MERCI