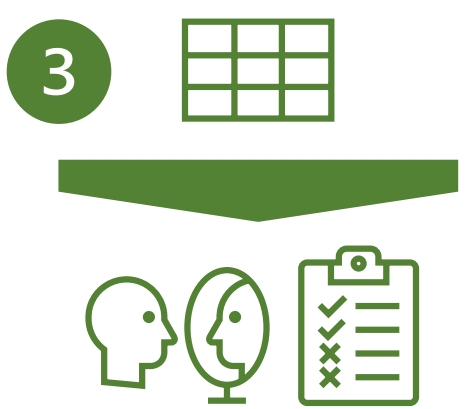
Classifiez automatiquement des biens de consommation



1

Analyse exploratoire

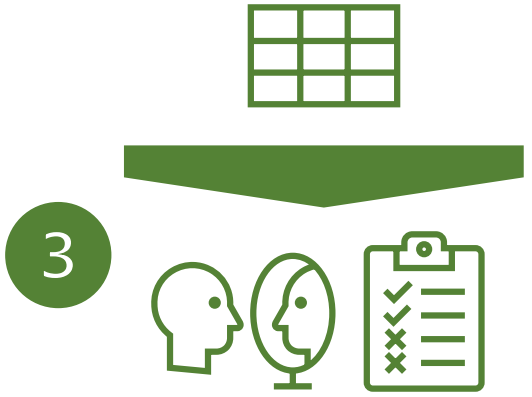
- Chargement DataFrame
- Séparation Catégories
- Tri features utiles
- Analyse des distributions

2

Pré-traitements / Clustering

- Image | gray, equalization, SIFT, VGG16
- Texte | token, stem, lem, freq, norm
- Texte | Bow, Word2Vec, BERT, USE
- Image & Text | Embedding

3



Validation faisabilité

- Réduction T-SNE
- Comparaison ARI Score
- Conclusion & préconisation

1 Analyse exploratoire

Chargement DataFrame Séparation Catégories

Chargement DataFrame
1050 individus x 15 champs

Features cibles :
description
image
cat_1 7 valeurs
label [0, 1, 2, 3, 4, 5, 6]

product_category_tree

["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Y..."]

["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]

cat_1 cat_2 cat_3 cat_4

description
image
cat_1
label
Beauty an ~~uniq_id~~
~~crawl_timestamp~~
~~product_url~~
~~product_name~~
~~product_category_tree~~
Home Decor ~~pid~~
~~retail_price~~
~~discounted_price~~
~~is_FK_Advantage_product~~
~~product_rating~~
~~overall_rating~~
~~brand~~
~~product_specifications~~

DataFrame final
1050 individus x 4 champs

2 Pré-traitements / Clustering

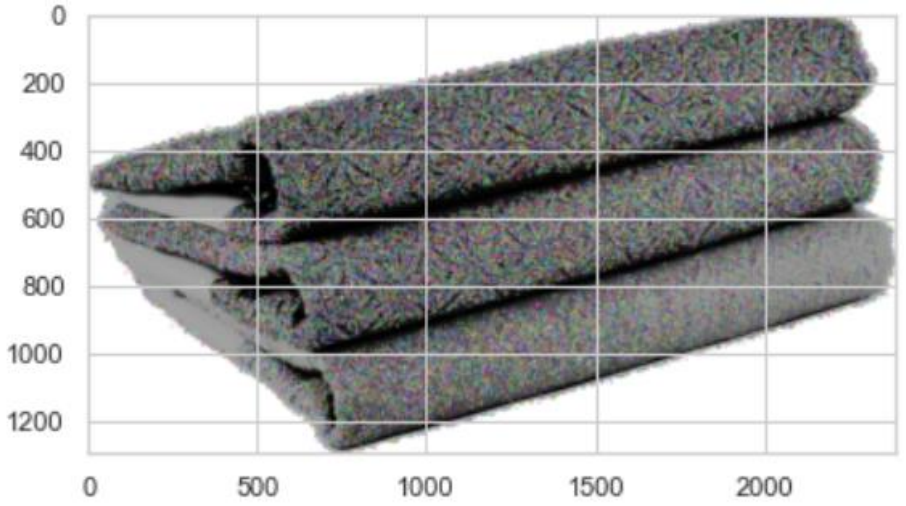
```
list_photos[0:3]
```

```
['55b85ea15a1536d46b7190ad6fff8ce7.jpg',
 '7b72c92c2f6c40268628ec5f14c6d590.jpg',
 '64d5d4a258243731dc7bbb1eef49ad74.jpg']
```

label

image

Image | gray, equalization préparation SIFT



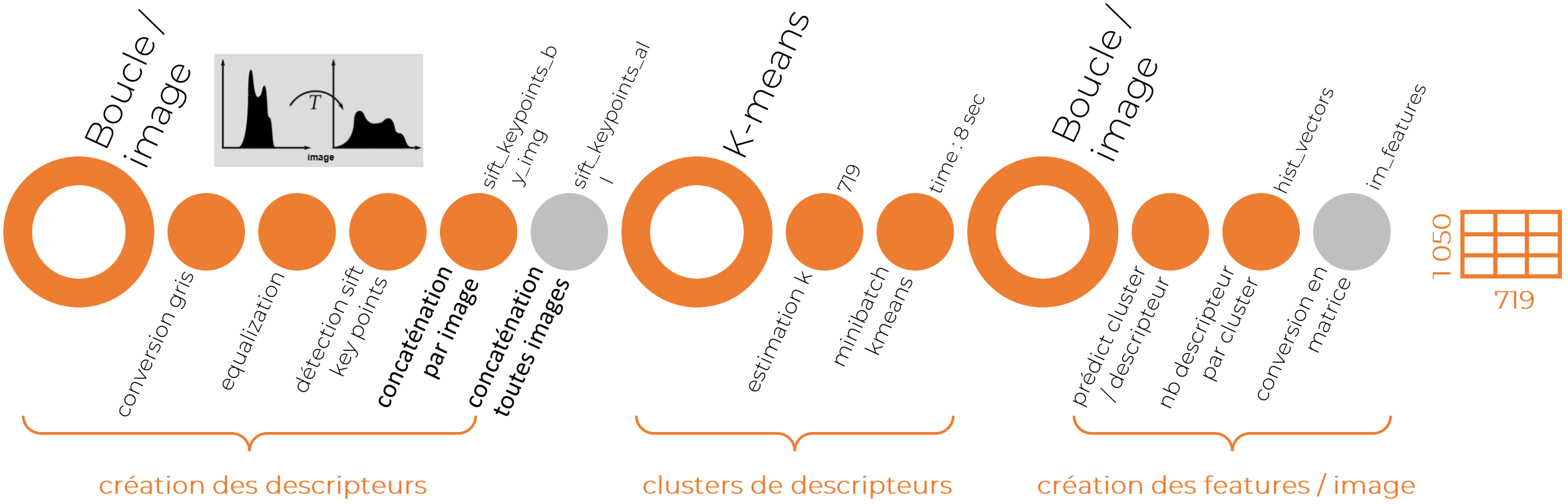
```
Descripteurs :  (47651, 128)

[[ 85. 113.  31. ...   0.   0.   0.]
 [ 29.  80.  73. ...   0.   0.   0.]
 [ 10.  67. 115. ...   0.   0.  22.]
 ...
 [  0.   0.   0. ...   0.   0.  13.]
 [ 40.   0.   0. ...   0.   0.  51.]
 [ 37.   1.   0. ...   0.   0.   0.]]
```

47 651 descripteurs

vecteur de longueur 128

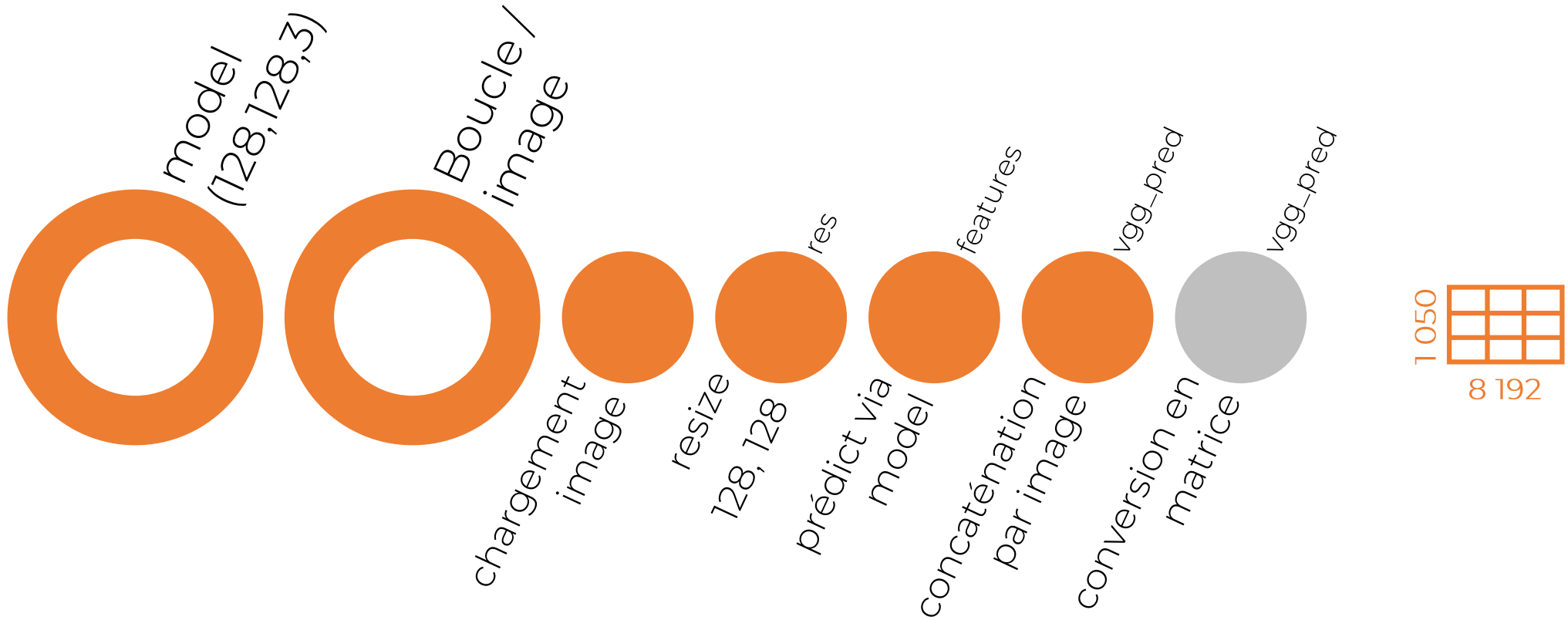
2 Pré-traitements / Clustering

Image | SIFT



2 Pré-traitements / Clustering

Image | VGG16



2 Pré-traitements / Clustering — Texte | tokenization & autres

```
test_text = """With just under a month until the transfer window closes, Cristiano Ronaldo's chances
of leaving Manchester United continue to diminish! He's made it clear : he wants to leave for a new
side (the Red Devils failed to qualify for the Champions League) but so far, they've refused every
offer made for the Portuguese star. #goodluckcr7"""
```

"With just under a month until the transfer window closes, Cristiano Ronaldo's chances \nof leaving Manchester United continue to diminish! He's made it clear : he wants to leave for a new \nside (the Red Devils failed to qualify for the Champions Leagu e) but so far, they've refused every \noffer made for the Portuguese star. #goodluckcr7"

join lemma lower stop token

['With', 'just', 'under', 'a', 'month', 'until', 'the', 'transfer', 'window', 'closes', ',', 'Cristiano', 'Ronaldo', "'s", 'chances', 'of', 'leaving', 'Manchester', 'United', 'continue', 'to', 'diminish', '!', 'He', "'s", 'made', 'it', 'clear', ':', 'he', 'wants', 'to', 'leave', 'for', 'a', 'new', 'side', '(', 'the', 'Red', 'Devils', 'failed', 'to', 'qualify', 'for', 'the', 'Champions', 'League', ')', 'but', 'so', 'far', ',', 'they', "'ve", 'refused', 'every', 'offer', 'made', 'for', 'the', 'Portuguese', 'star', '.', 'goodluckcr7'] — len 65

['With', 'month', 'transfer', 'window', 'closes', 'Cristiano', 'Ronaldo', 'chances', 'leaving', 'Manchester', 'United', 'continue', 'diminish', 'made', 'clear', 'wants', 'leave', 'new', 'side', 'Red', 'Devils', 'failed', 'qualify', 'Champions', 'League', 'far', "'ve", 'refused', 'every', 'offer', 'made', 'Portuguese', 'star', 'goodluckcr7'] — 34

['with', 'month', 'transfer', 'window', 'closes', 'cristiano', 'ronaldo', 'chances', 'leaving', 'manchester', 'united', 'continue', 'diminish', 'made', 'clear', 'wants', 'leave', 'new', 'side', 'red', 'devils', 'failed', 'qualify', 'champions', 'league', 'far', "'ve", 'refused', 'every', 'offer', 'made', 'portuguese', 'star', 'goodluckcr7'] — 34

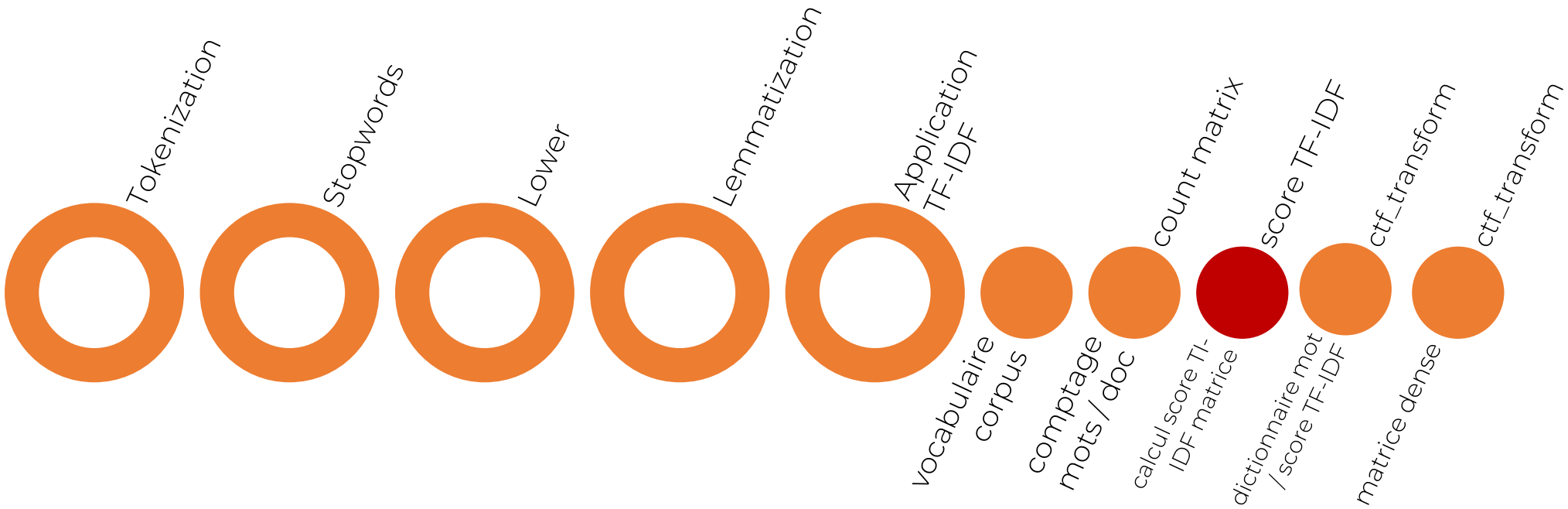
['with', 'month', 'transfer', 'window', 'close', 'cristiano', 'ronaldo', 'chance', 'leaving', 'manchester', 'united', 'continue', 'diminish', 'made', 'clear', 'want', 'leave', 'new', 'side', 'red', 'devil', 'failed', 'qualify', 'champion', 'league', 'far', "'ve", 'refused', 'every', 'offer', 'made', 'portuguese', 'star', 'goodluckcr7'] — 34

"with month transfer window close cristiano ronaldo chance leaving manchester united continue diminish made clear want leave ne w side red devil failed qualify champion league far 've refused every offer made portuguese star goodluckcr7" — 34

2 Pré-traitements / Clustering Texte | Bow Count Vectorizer

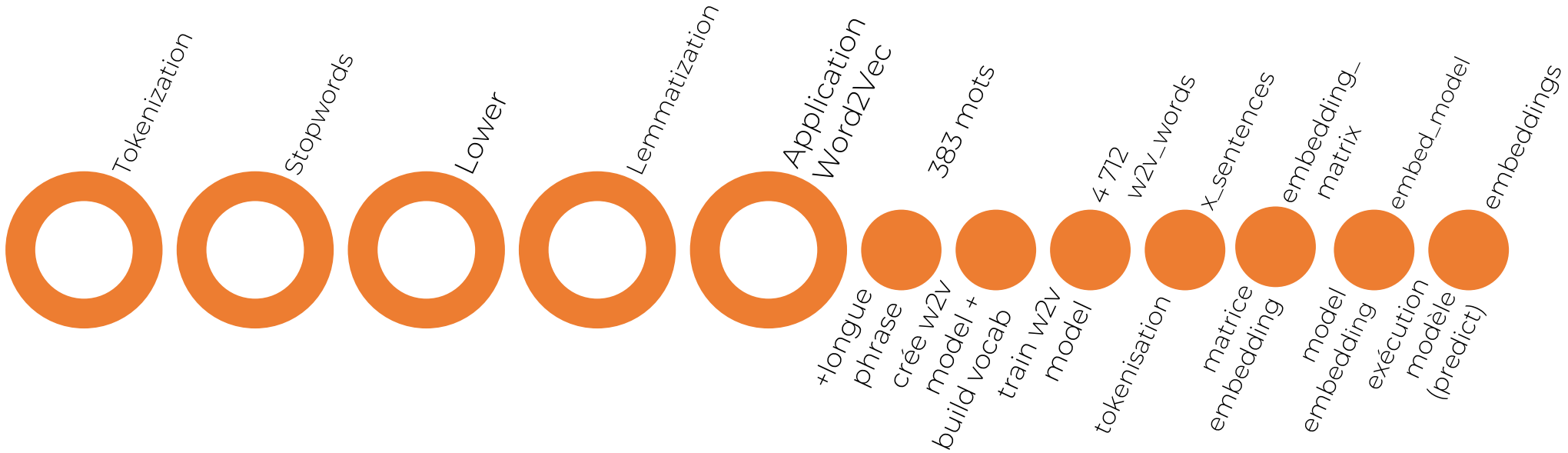
2 Pré-traitements / Clustering

Texte | Bow TF-idf



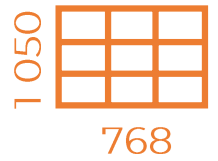
2 Pré-traitements / Clustering

Texte | Word2Vec



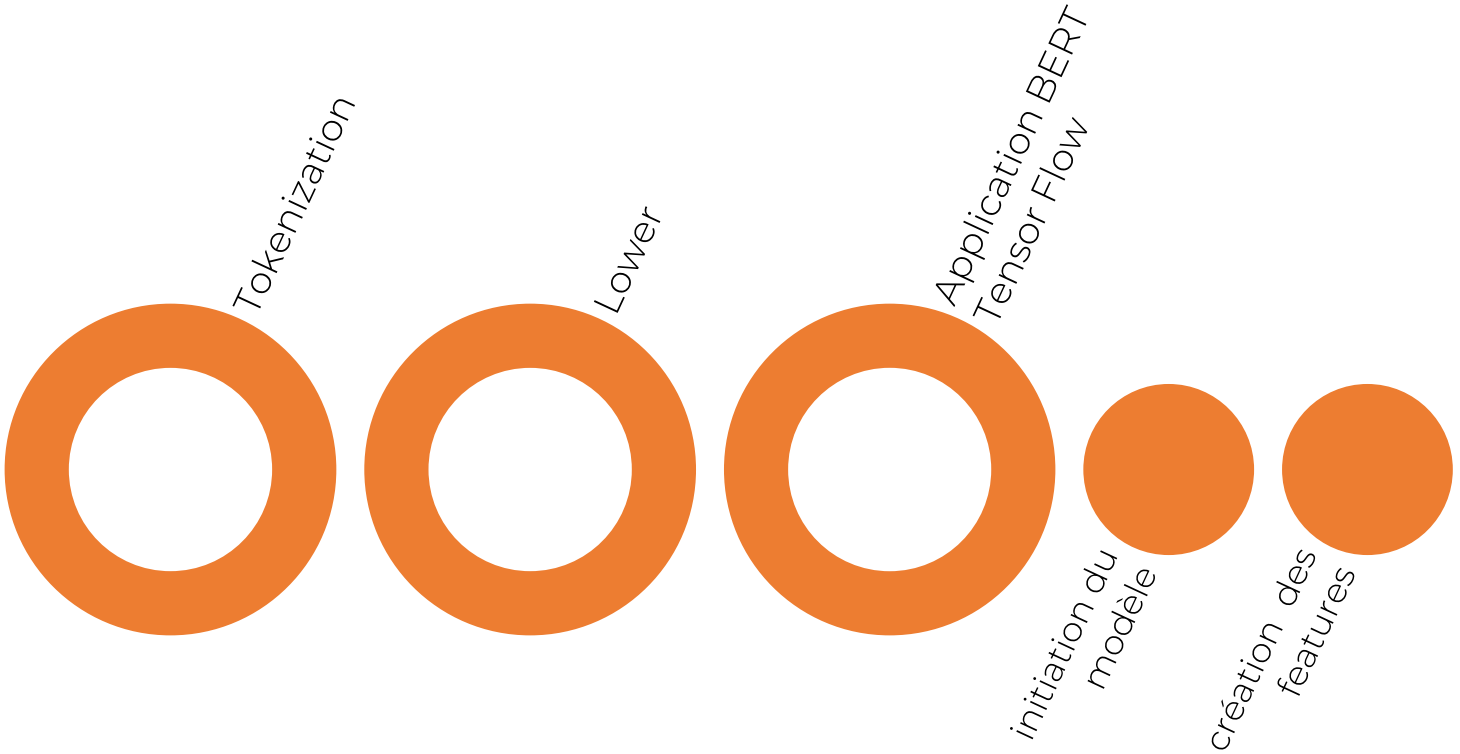
2 Pré-traitements / Clustering

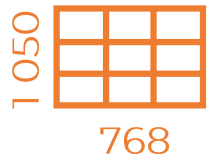
Texte | BERT HuggingFace



2 Pré-traitements / Clustering

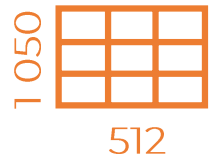
Texte | BERT TensorFlow





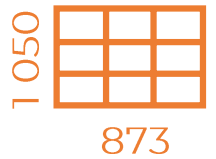
2 Pré-traitements / Clustering

Texte | USE



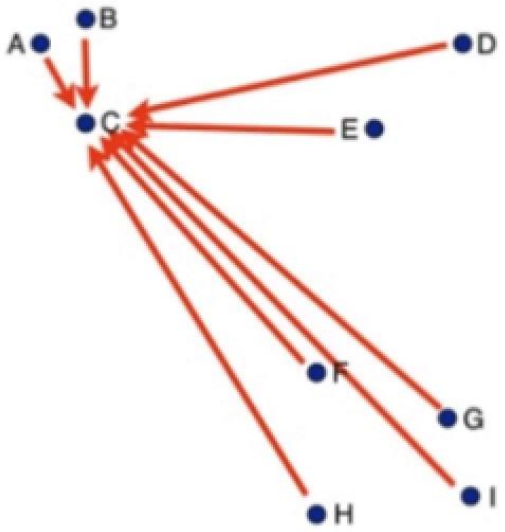
2 Pré-traitements / Clustering

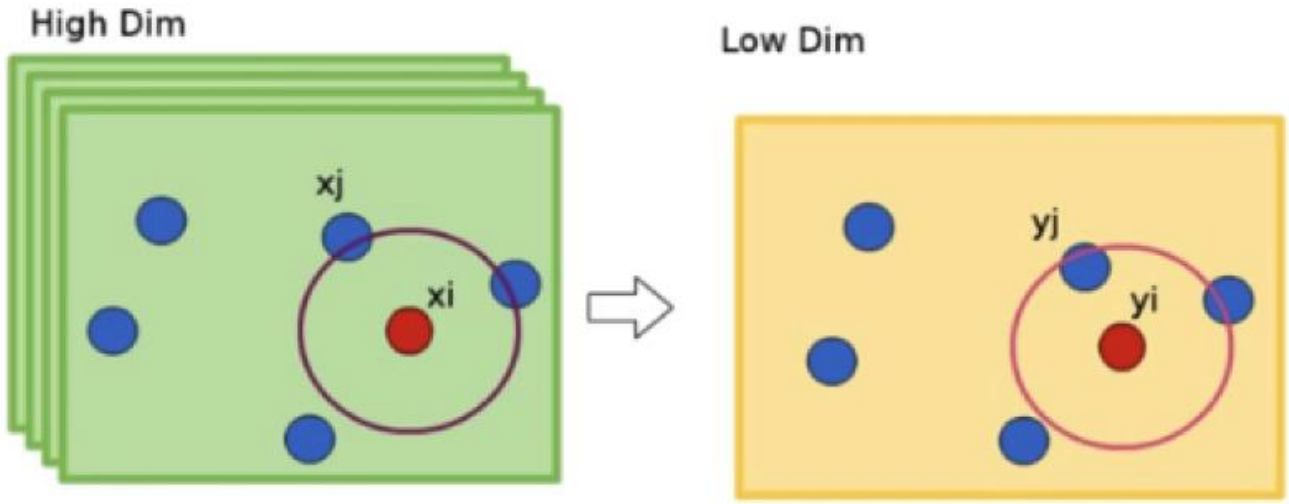
Image & Text | Embedding



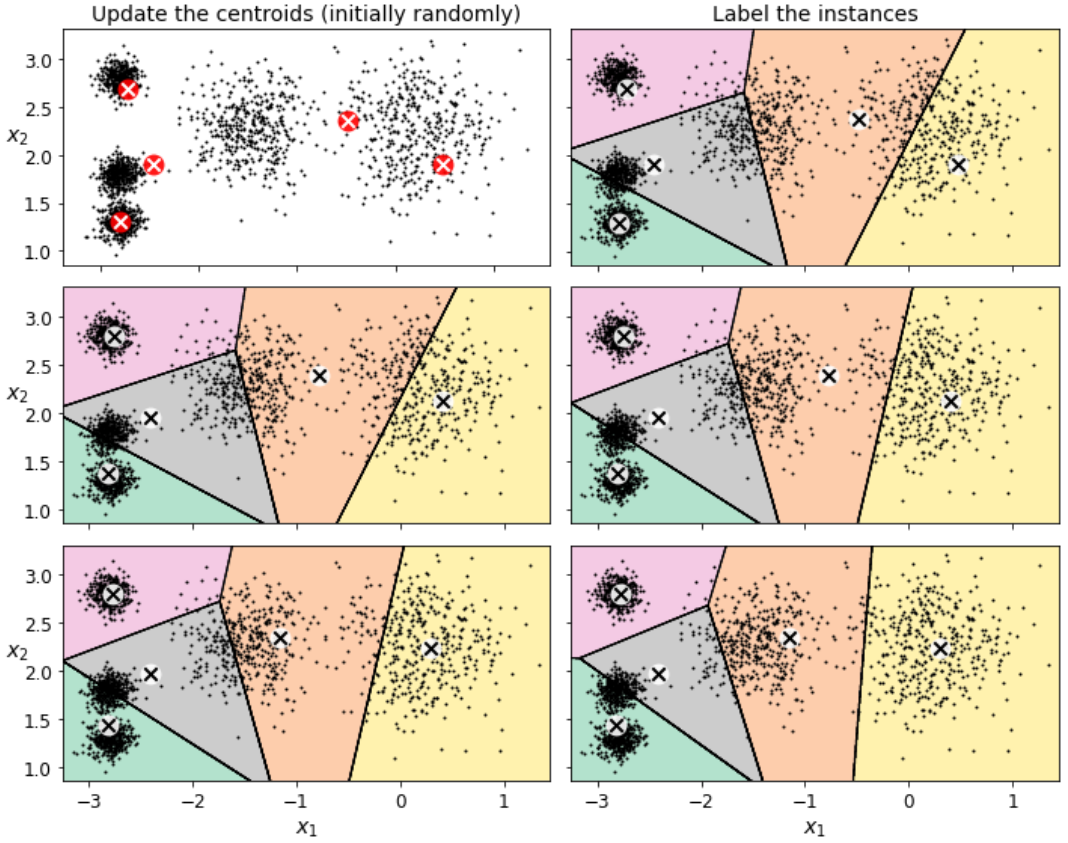
3 Validation faisabilité

t-distributed Stochastic Neighbor Embedding



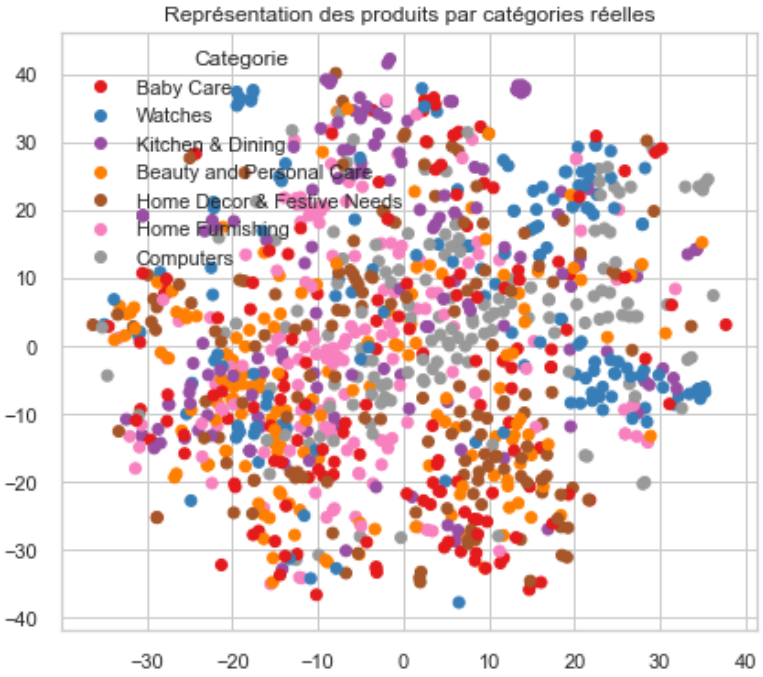


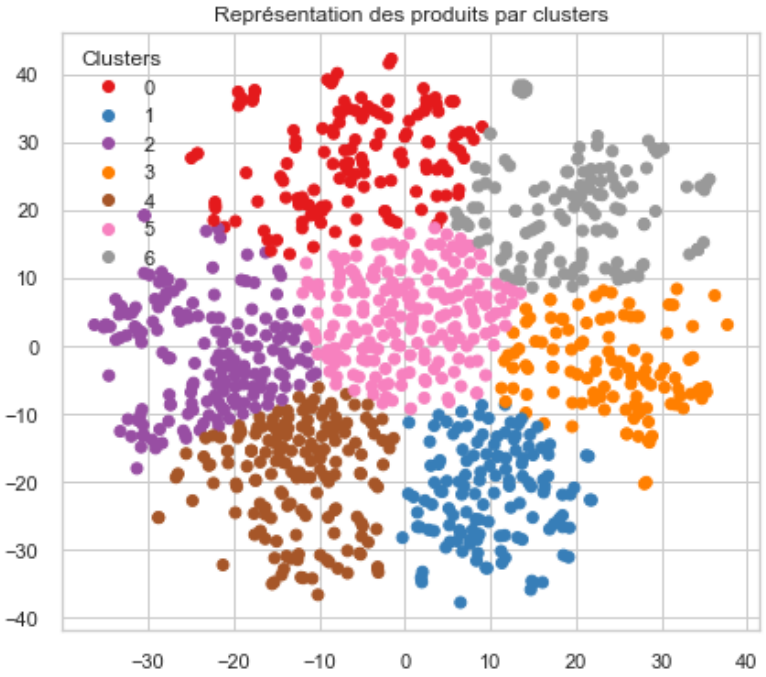
Réduction T-SNE & K-means 7 clusters

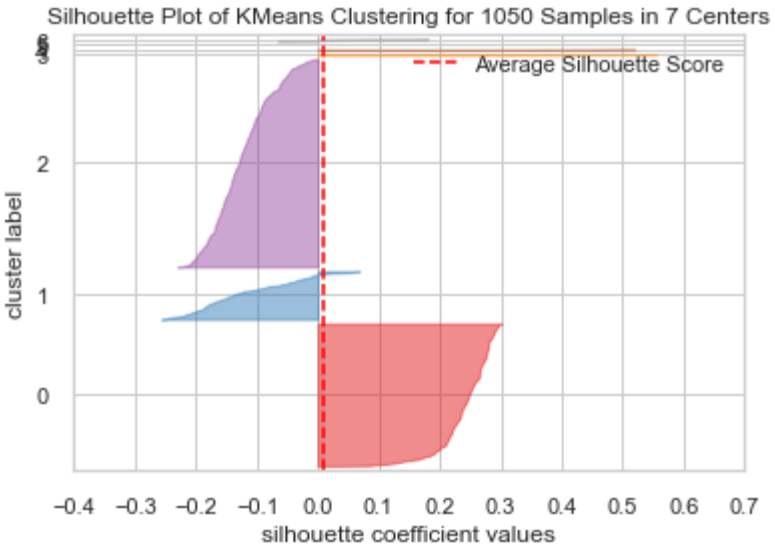


3 Validation faisabilité

Comparaison ARI Score
Image | SIFT





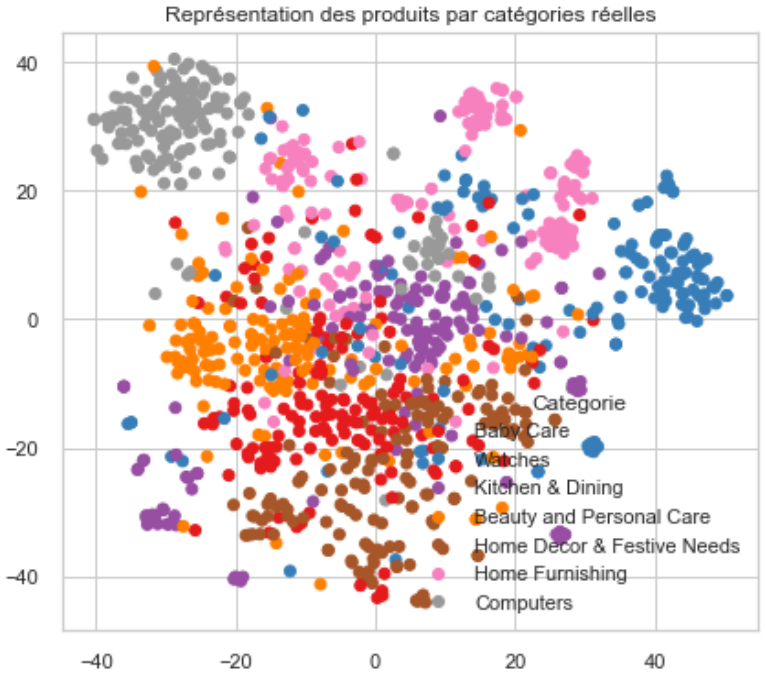


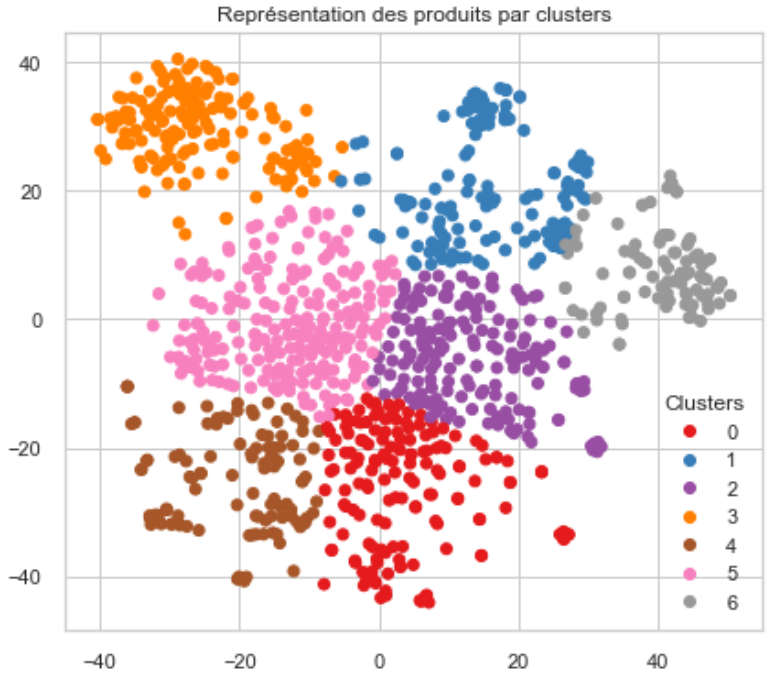
ARI : 0.0496

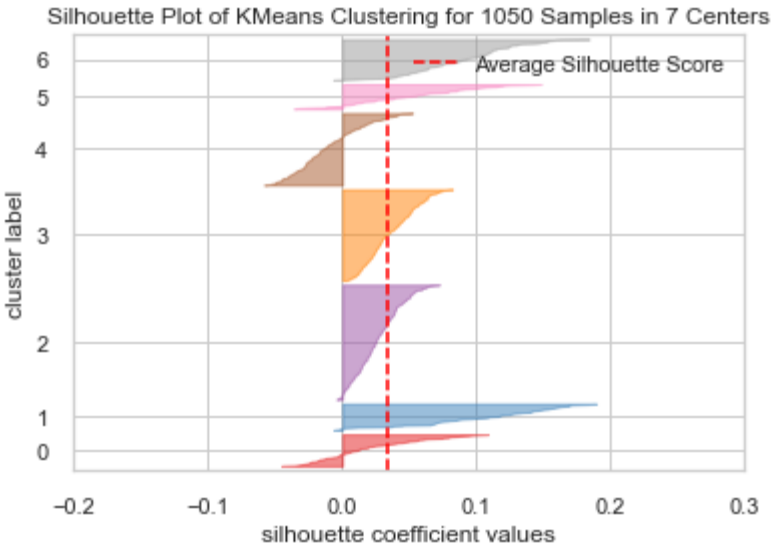
Silhouette : 0.0078

3 Validation faisabilité

Comparaison ARI Score Image | VGG16





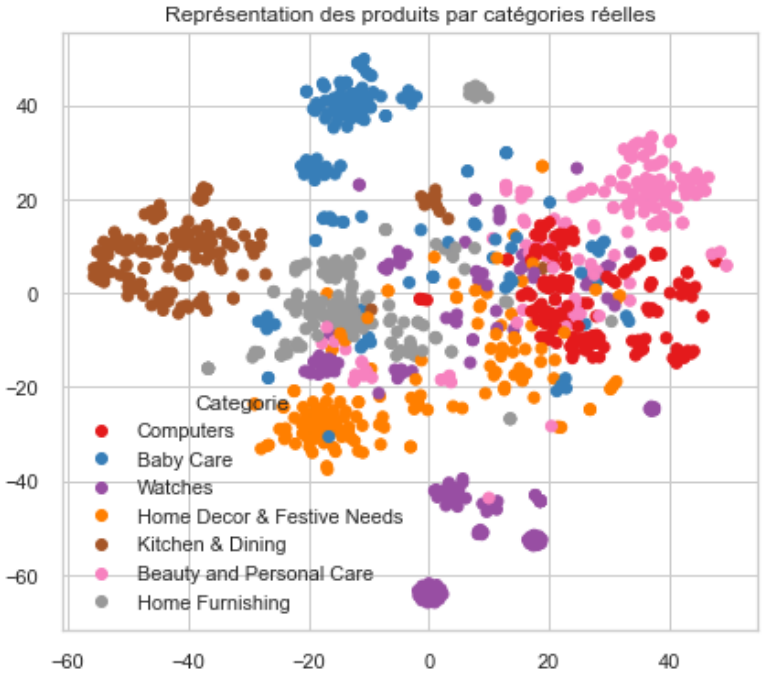


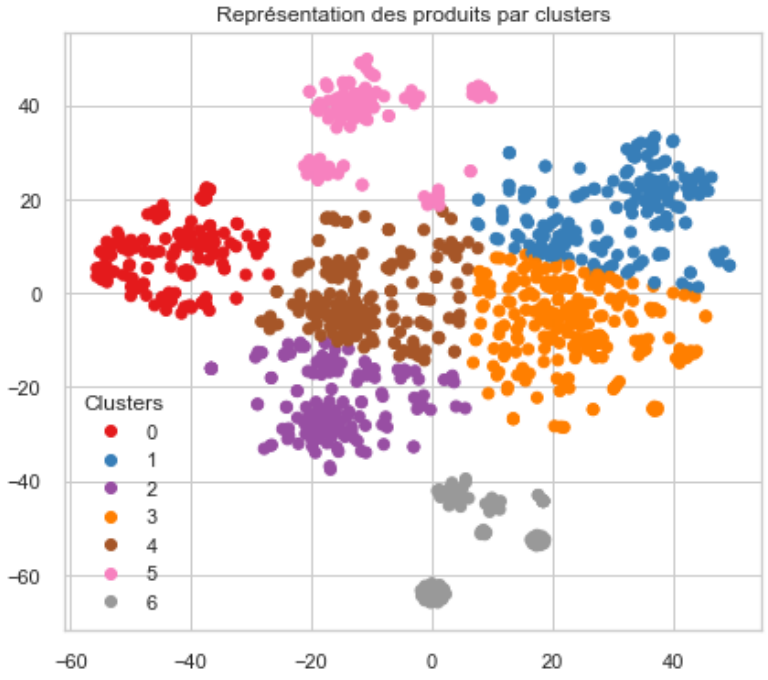
ARI : 0.2571

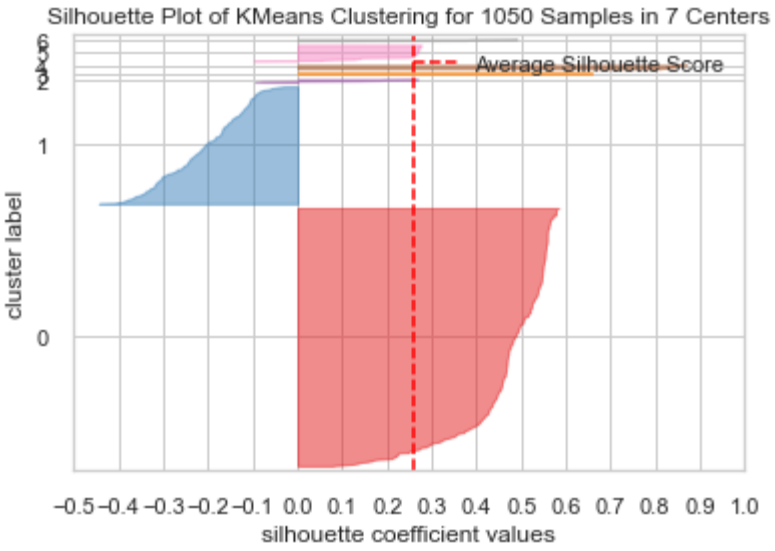
Silhouette : 0.034

3 Validation faisabilité

Comparaison ARI Score
Texte | BoW CountVectorizer





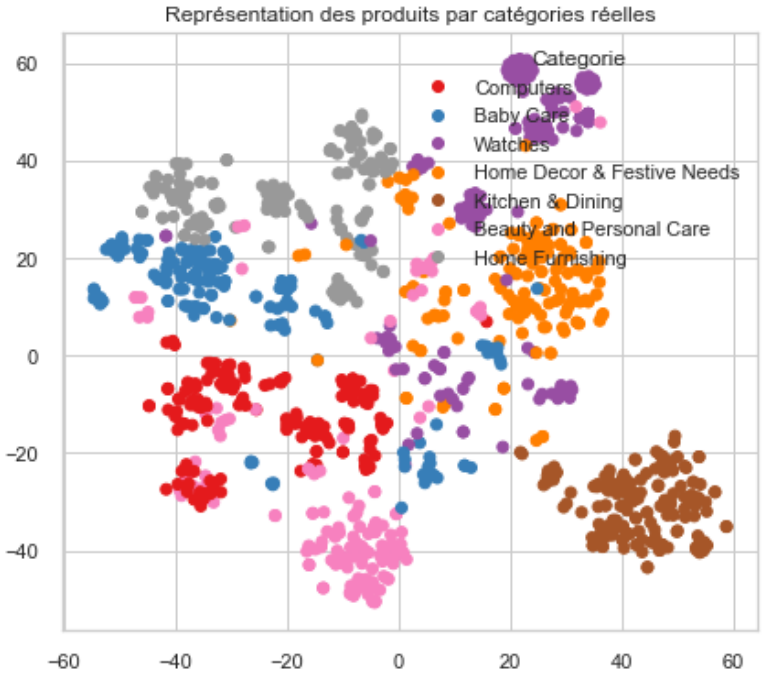


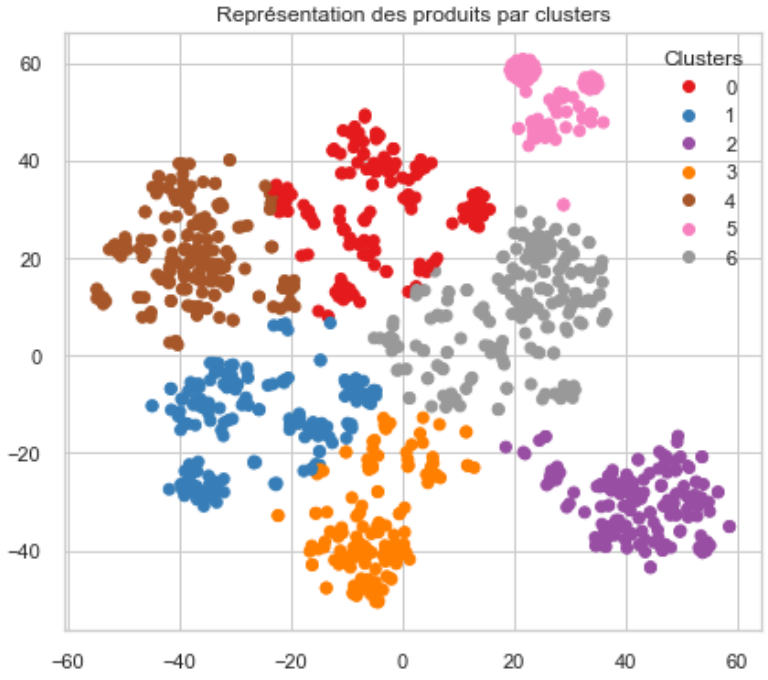
ARI : 0.4328

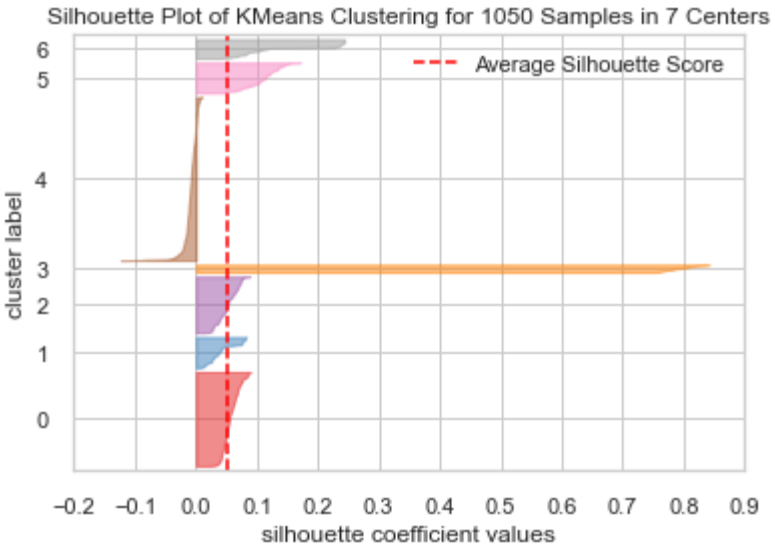
Silhouette : 0.2615

3 Validation faisabilité

Comparaison ARI Score
Texte | BoW Tf-Idf





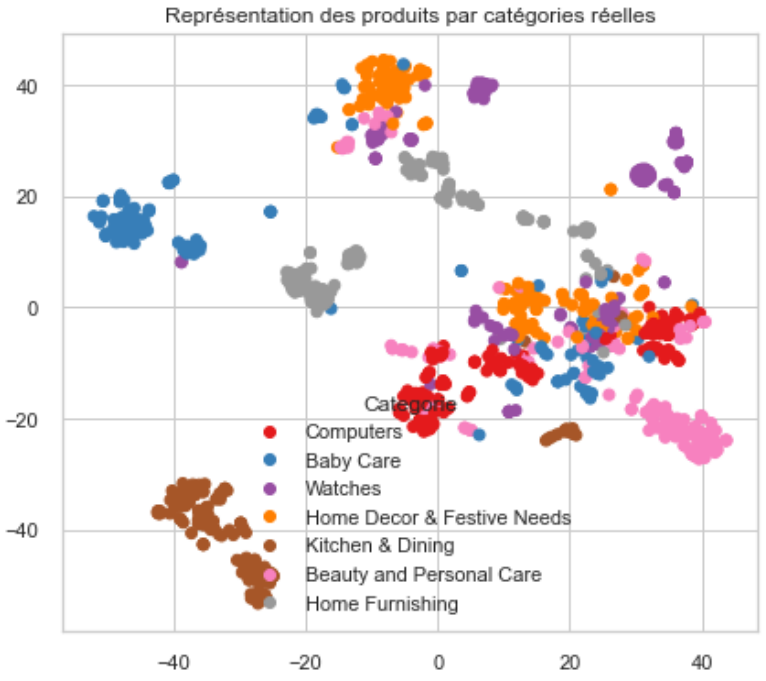


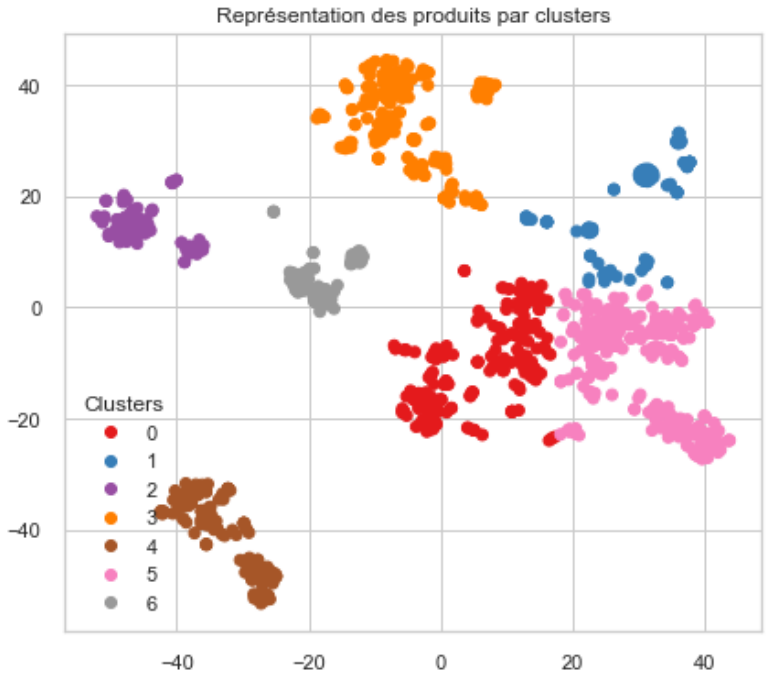
ARI : 0.5274

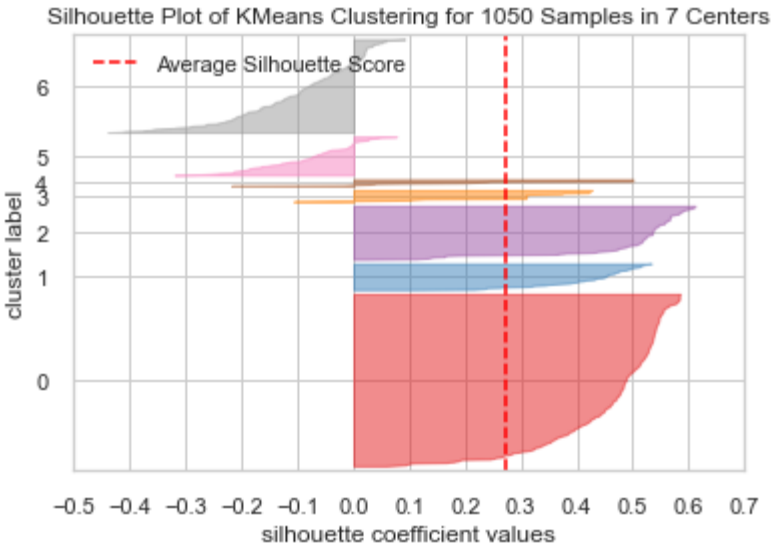
Silhouette : 0.0517

3 Validation faisabilité

Comparaison ARI Score
Texte | Word2Vec





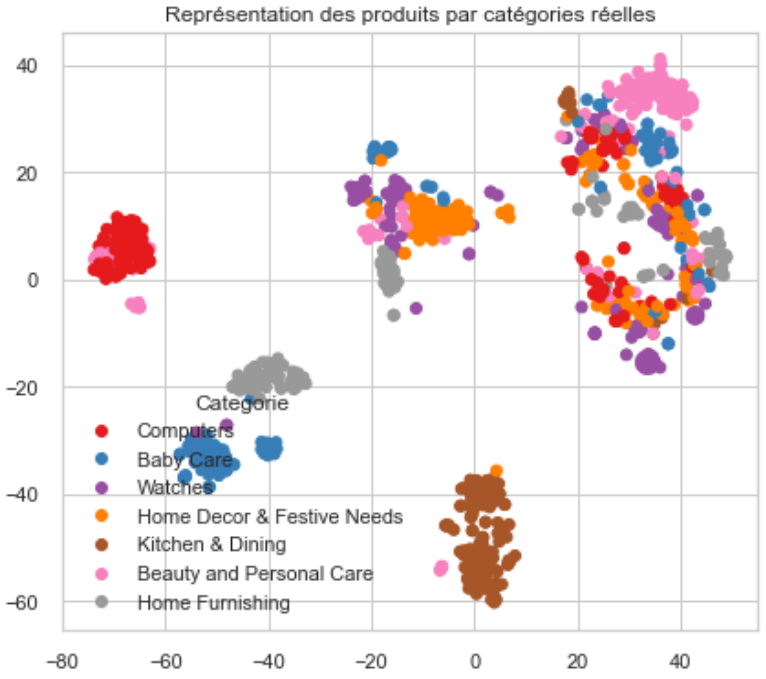


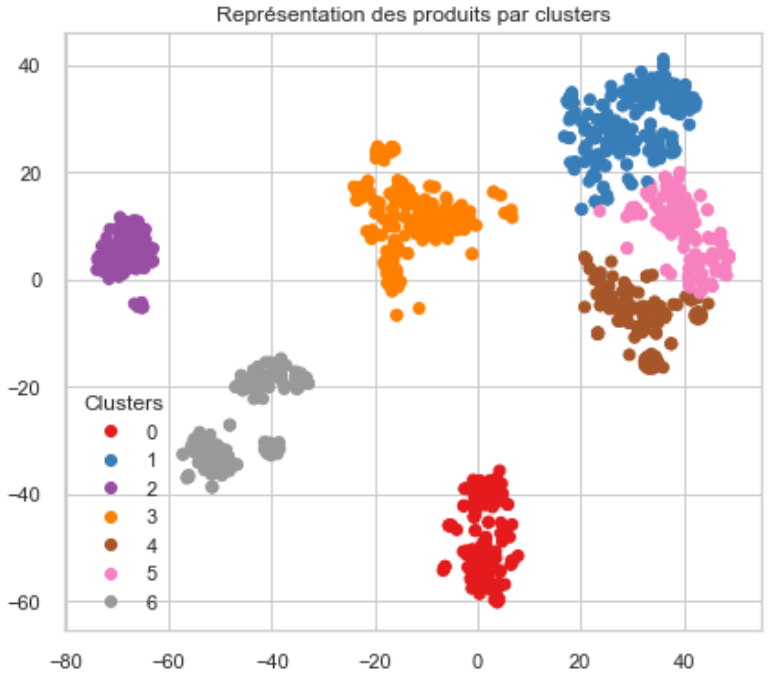
ARI : 0.3215

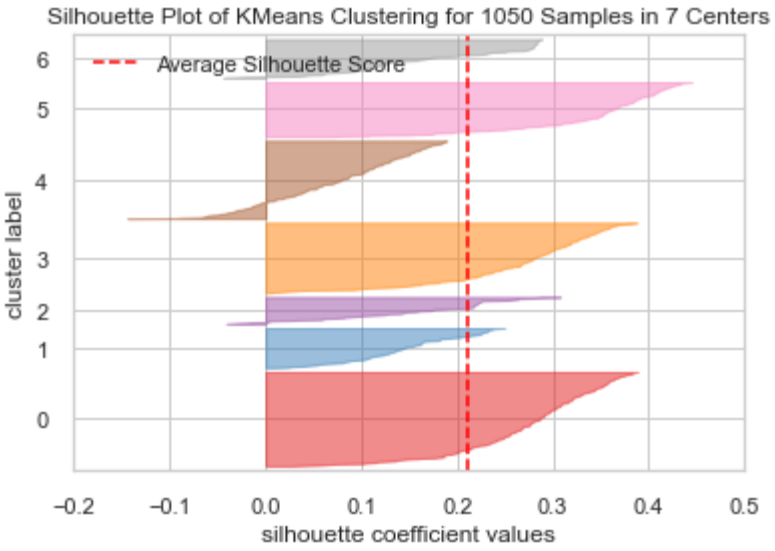
Silhouette : 0.2717

3 Validation faisabilité

Comparaison ARI Score
Texte | BERT HuggingFace





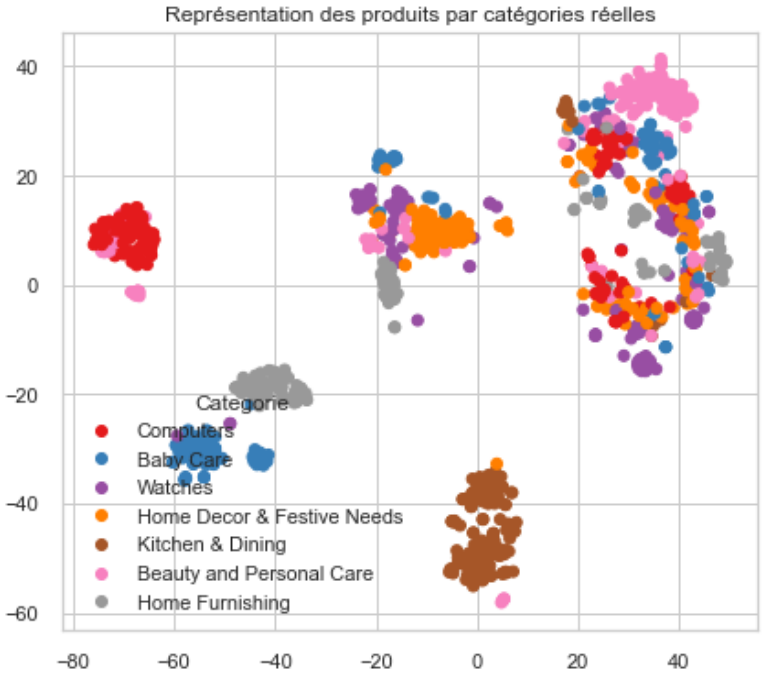


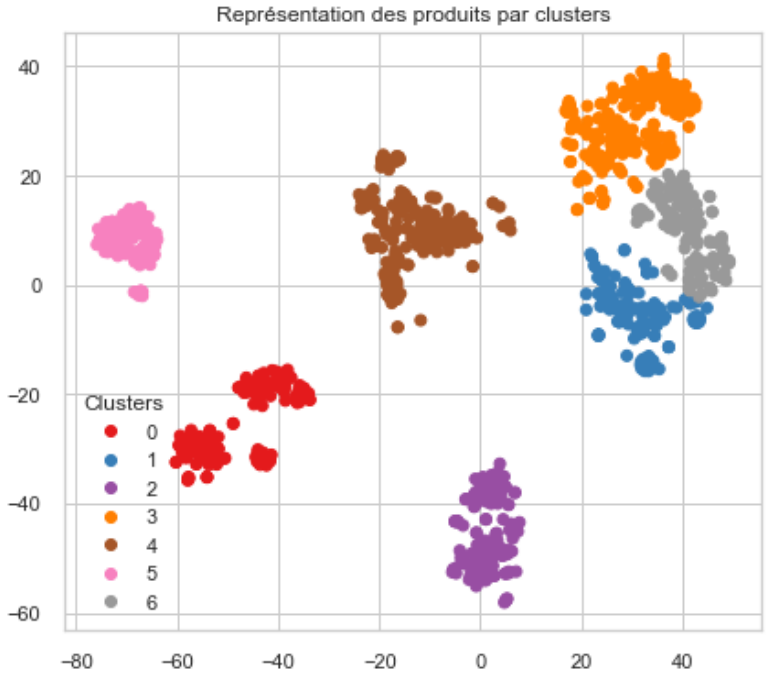
ARI : 0.293

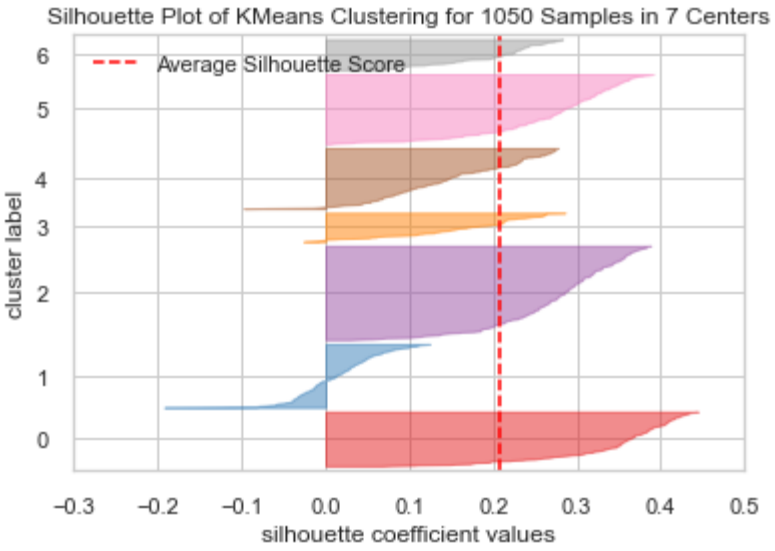
Silhouette : 0.2114

3 Validation faisabilité

Comparaison ARI Score
Texte | BERT TensorFlow





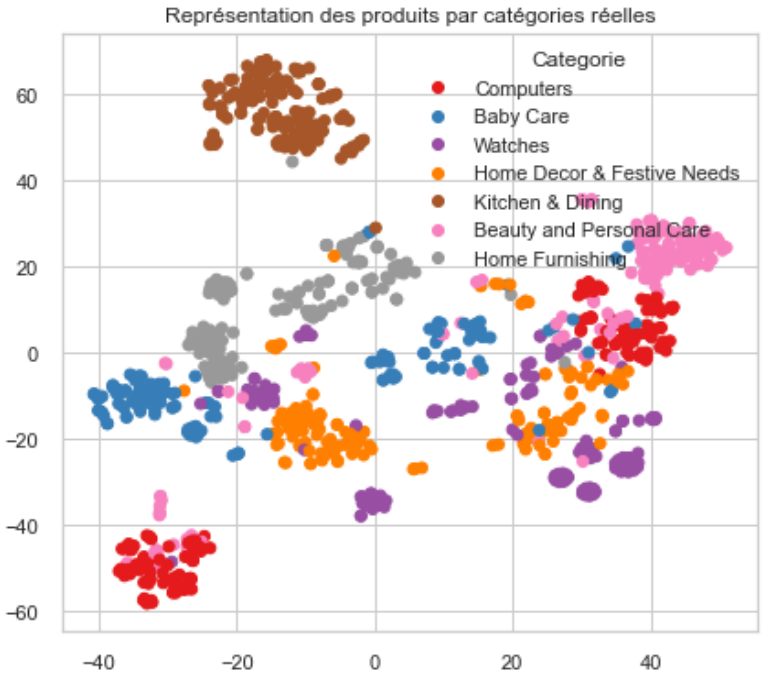


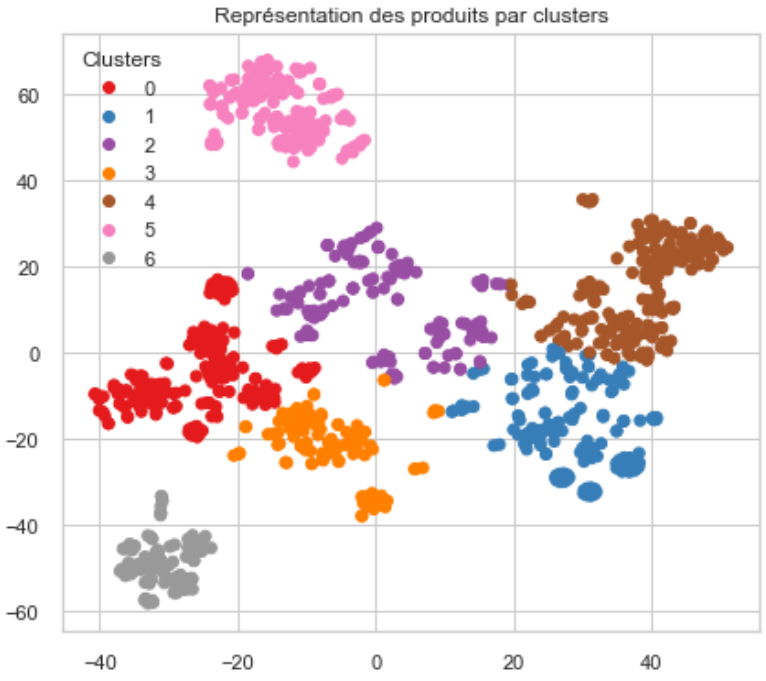
ARI : 0.2935

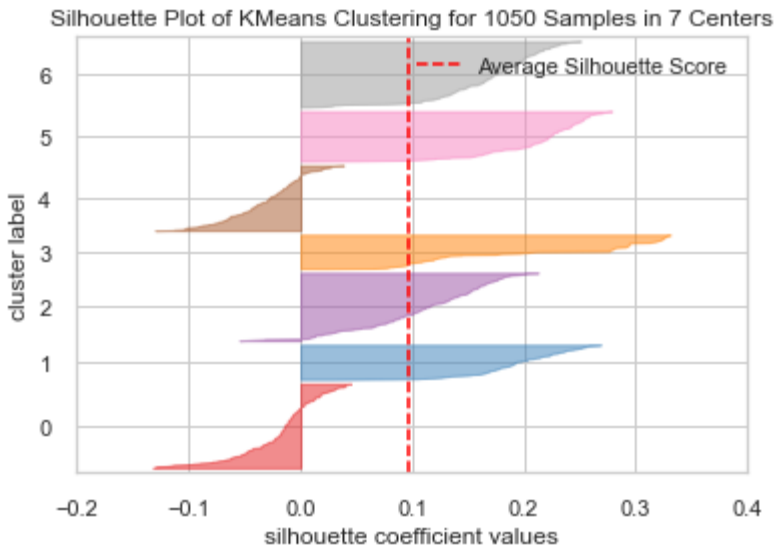
Silhouette : 0.2071

3 Validation faisabilité

Comparaison ARI Score
Texte | USE







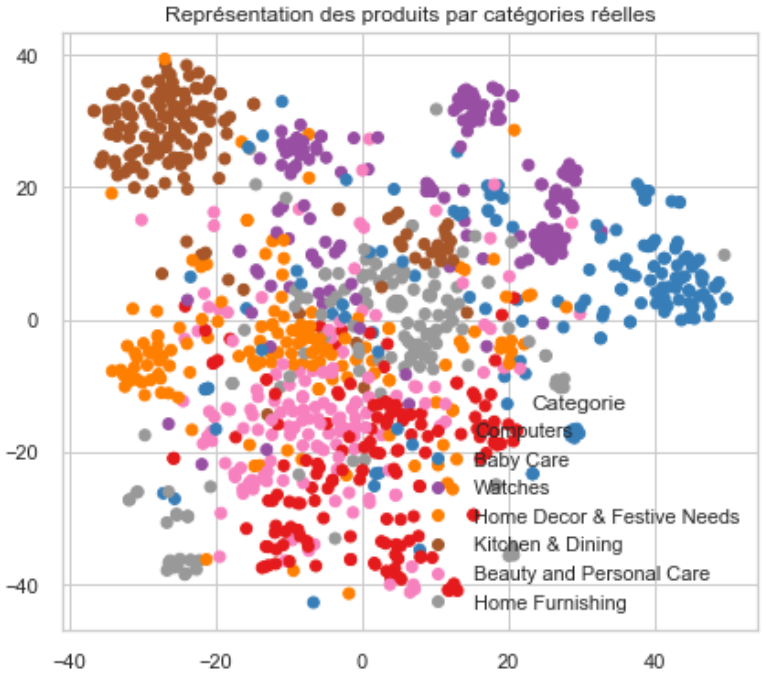
ARI : 0.4336

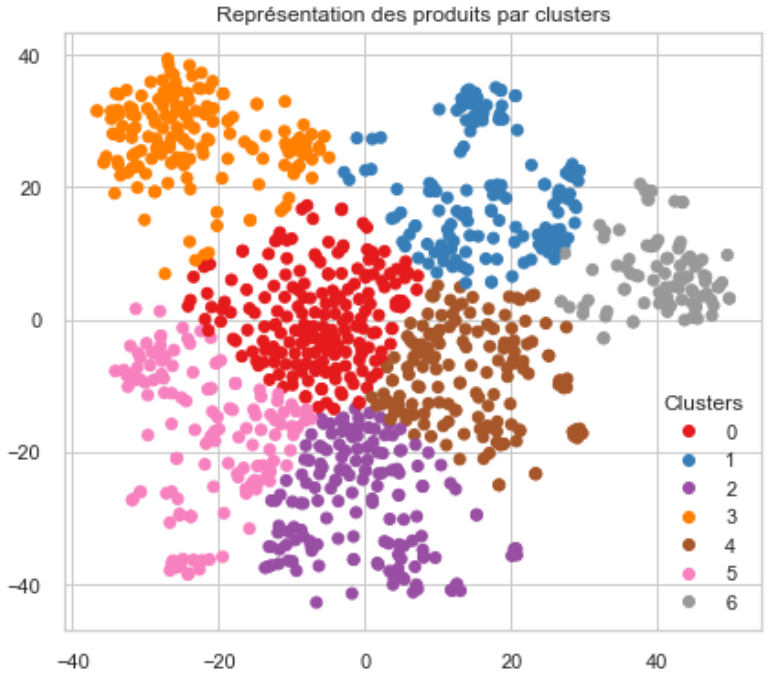
Silhouette : 0.0973

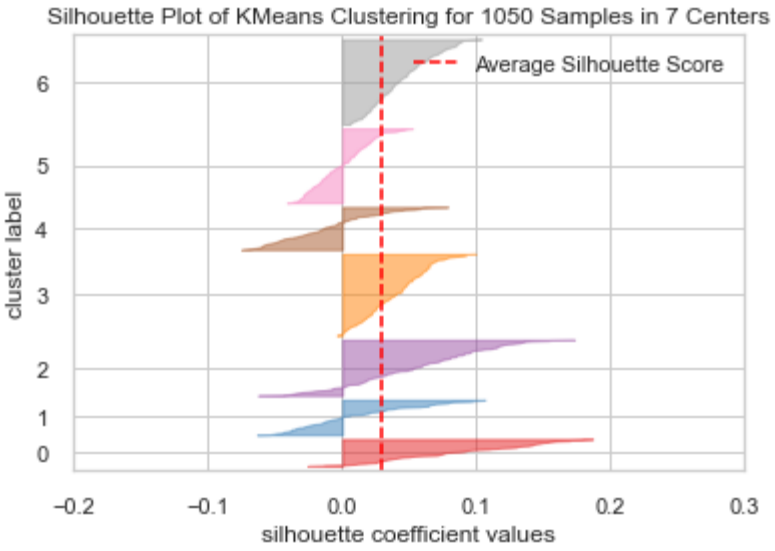
3 Validation faisabilité

Comparaison ARI Score

Texte + Image | BoW Tf-IdF + VGG16



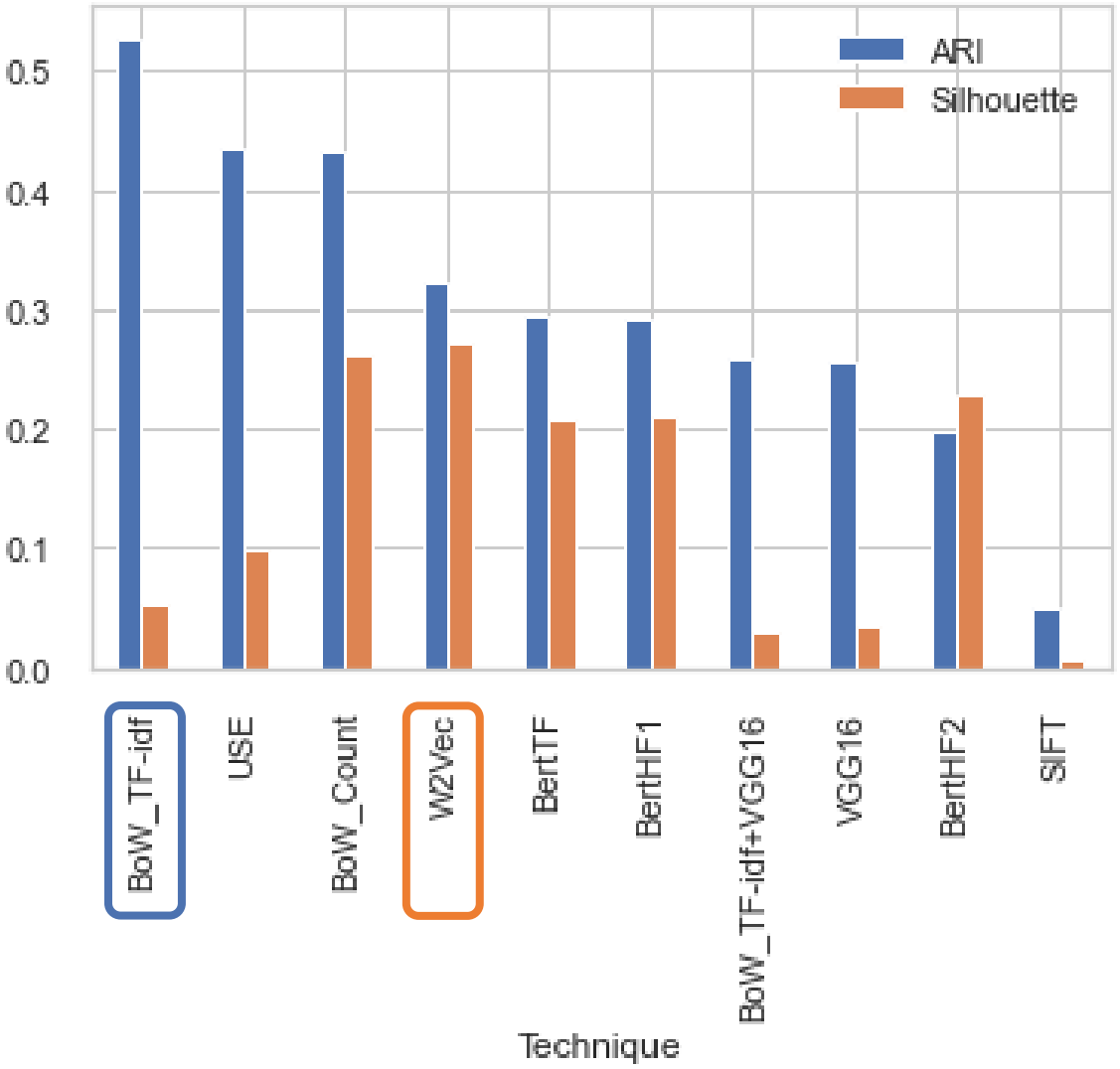




ARI : 0.2582

Silhouette : 0.0295

3 Validation faisabilité



Pas de modèle idéal

Conclusion & préconisations

	Type	Technique	ACP	ARI	Silhouette
3	Texte	BoW_TF-idf	Yes	0.5274	0.0517
8	Texte	USE	Yes	0.4336	0.0973
2	Texte	BoW_Count	Yes	0.4328	0.2615
4	Texte	W2Vec	Yes	0.3215	0.2717
7	Texte	BertTF	Yes	0.2935	0.2071
5	Texte	BertHF1	Yes	0.293	0.2114
9	Txt+Img	BoW_TF-idf+VGG16	Yes	0.2582	0.0295
1	Image	VGG16	Yes	0.2571	0.034
6	Texte	BertHF2	Yes	0.1971	0.2272
0	Image	SIFT	Yes	0.0496	0.0078

Stacker / entrainer les modèles

Travailler sur cat_2

Questions ?

Merci !