Pour info, sur les slides 5, 9, 10, 11, 13 et 14 il y a des animations qui cachent le texte en dessous donc si possible plutôt regarder le ppt en mode diaporama pour voir le texte sur ces slides.

CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

SOMMAIRE

1. Rappel de la problématique
2. Présentation du jeu de données
3. Approches de modélisation du texte
4. Approches de modélisation des images
5. Association du texte et de l'image
6. Conclusion

RAPPEL DE LA PROBLÉMATIQUE

- L'entreprise ''Place de marché'' est une marketplace qui propose des articles à des acheteurs en postant une photo et une description.
- Besoin d'automatiser l'attribution d'une catégorie à un article.
- Etudier la faisabilité d'un moteur de classification avec un niveau de précision suffisant.

PRÉSENTATIC

- Jeu de données cor 50 articles.
- Les catégories sont i nne :
  - Exemple : ["Wat
- Manipulations pour gories par article.
- Certaines sous-caté u de données donc décis tude de faisabilité d'un mote

cat	sub_cat	uniq_id
Baby Care	Baby & Kids Gifts	15
	Baby Bath & Skin	14
	Baby Bedding	15
	Baby Grooming	4
	Diapering & Potty Training	7
	Feeding & Nursing	8
	Furniture & Furnishings	2
	Infant Wear	84
	Strollers & Activity Gear	1
Beauty and Personal Care	Bath and Spa	7
	Beauty Accessories	1
	Body and Skin Care	15
	Combos and Kits	24
	Eye Care	2
	Fragrances	65
	Hair Care	9
	Health Care	7
	Makeup	18
	Men's Grooming	1
	Women's Hygiene	1

PRÉSENTATION DU JEU DE DONNÉES

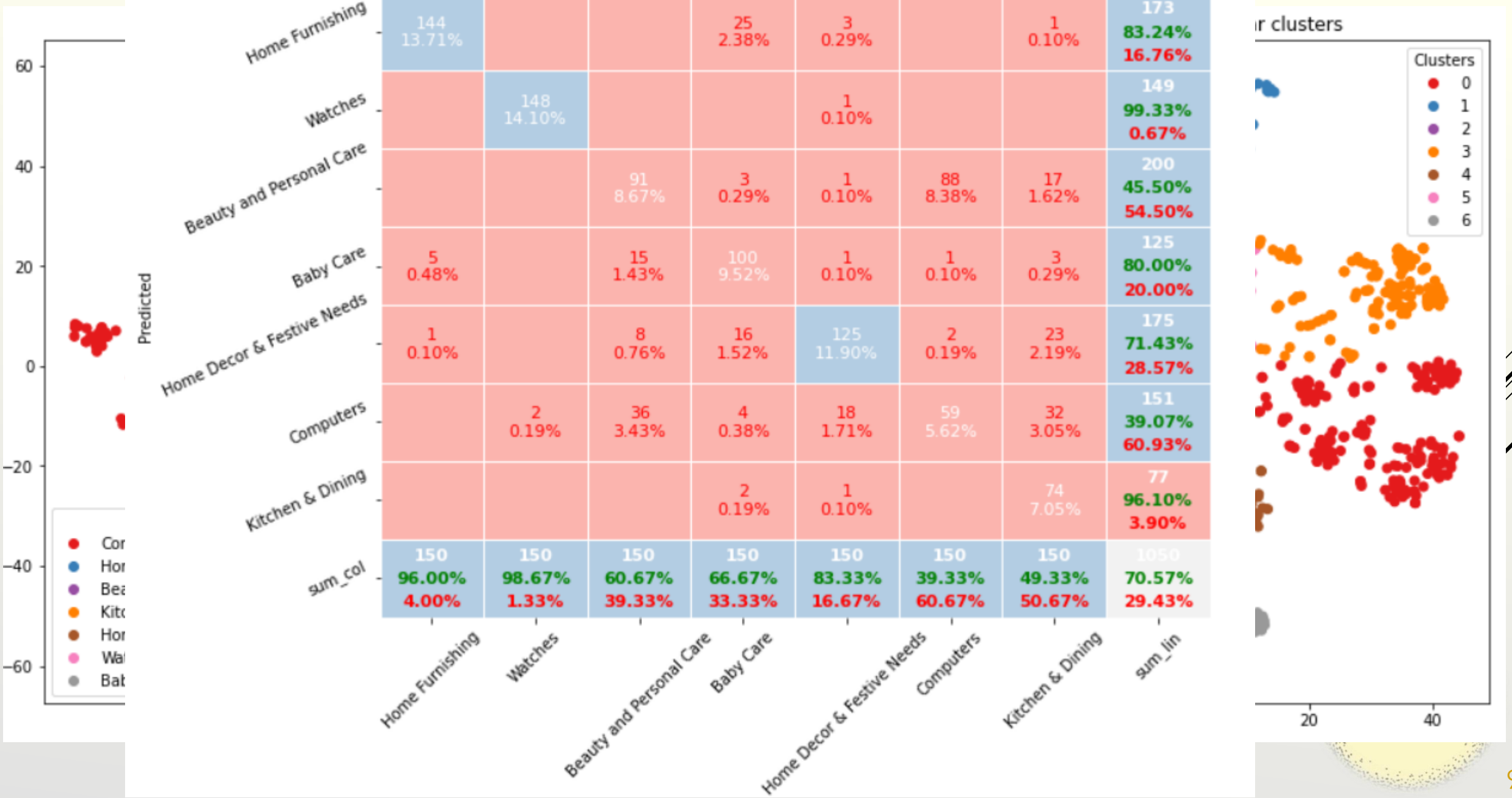
- 7 catégories principales parfaitement équilibrées :
  - Home Furnishing : 150 articles
  - Baby Care : 150 articles
  - Watches : 150 articles
  - Home Decor & Festive Needs : 150 articles
  - Kitchen & Dining : 150 articles
  - Beauty and Personal Care : 150 articles
  - Computers : 150 articles

APPROCHES DE MODÉLISATION DU TEXTE

- Extraction des features texte via différentes approches :
  - De type 'bag of words' avec un comptage simple de mots et avec Tf-idf
  - De type word embedding avec Word2Vec et Doc2Vec
  - De type word embedding avec BERT (Bidirectional Encoder Representations from Transformers)
  - De type sentence embedding avec USE (Universal Sentence Encoder)
- Réduction de dimension des features obtenus grâce à un t-sne (2 composantes).
- Clustering (k-means) sur les données réduites puis calcul de l'Adjusted Rand Index (ARI) par rapport aux vrais catégories.

APPROCHES DE MODÉLISATION DU TEXTE

- Prétraitements réalisés pour l'approche 'bag of words' et word embedding avec Word2Vec et Doc2Vec :
  - Mise du texte en minuscule
  - Suppression des ponctuations
  - Suppression des nombres
  - Tokenisation
  - Suppression des 'stopwords' anglais
  - Lemmatisation
  - Suppression des mots inférieurs ou égaux à 2 caractères


Confusion matrix
Predicted
Home Furnishing
Watches
Beauty and Personal Care
Baby Care
Home Decor & Festive Needs
Computers
Kitchen & Dining
sum_col
sum_lin
144 13.71%
25 2.38%
3 0.29%
1 0.10%
173 83.24% 16.76%
148 14.10%
1 0.10%
149 99.33% 0.67%
91 8.67%
3 0.29%
1 0.10%
88 8.38%
17 1.62%
200 45.50% 54.50%
5 0.48%
15 1.43%
100 9.52%
1 0.10%
1 0.10%
3 0.29%
125 80.00% 20.00%
1 0.10%
8 0.76%
16 1.52%
125 11.90%
2 0.19%
23 2.19%
175 71.43% 28.57%
2 0.19%
36 3.43%
4 0.38%
18 1.71%
59 5.62%
32 3.05%
151 39.07% 60.93%
2 0.19%
1 0.10%
74 7.05%
77 96.10% 3.90%
150 96.00% 4.00%
150 98.67% 1.33%
150 60.67% 39.33%
150 66.67% 33.33%
150 83.33% 16.67%
150 39.33% 60.67%
150 49.33% 50.67%
1050 70.57% 29.43%
Clusters
0
1
2
3
4
5
6

EXTE

Descriptions préprocessées de certains articles 'Computers' clusterisés en 'Beauty and Personal Care' :

'buy smart router genuine product day replacement guarantee free shipping cash delivery',
'buy genuine product day replacement guarantee free shipping cash delivery',
'buy wireless dual band router genuine product day replacement guarantee free shipping cash delivery',
'buy apple genuine product day replacement guarantee free shipping cash delivery',
'buy genuine product day replacement guarantee free shipping cash delivery',
'buy wireless range extender genuine product day replacement guarantee free shipping cash delivery',
'buy tew genuine product day replacement guarantee free shipping cash delivery',

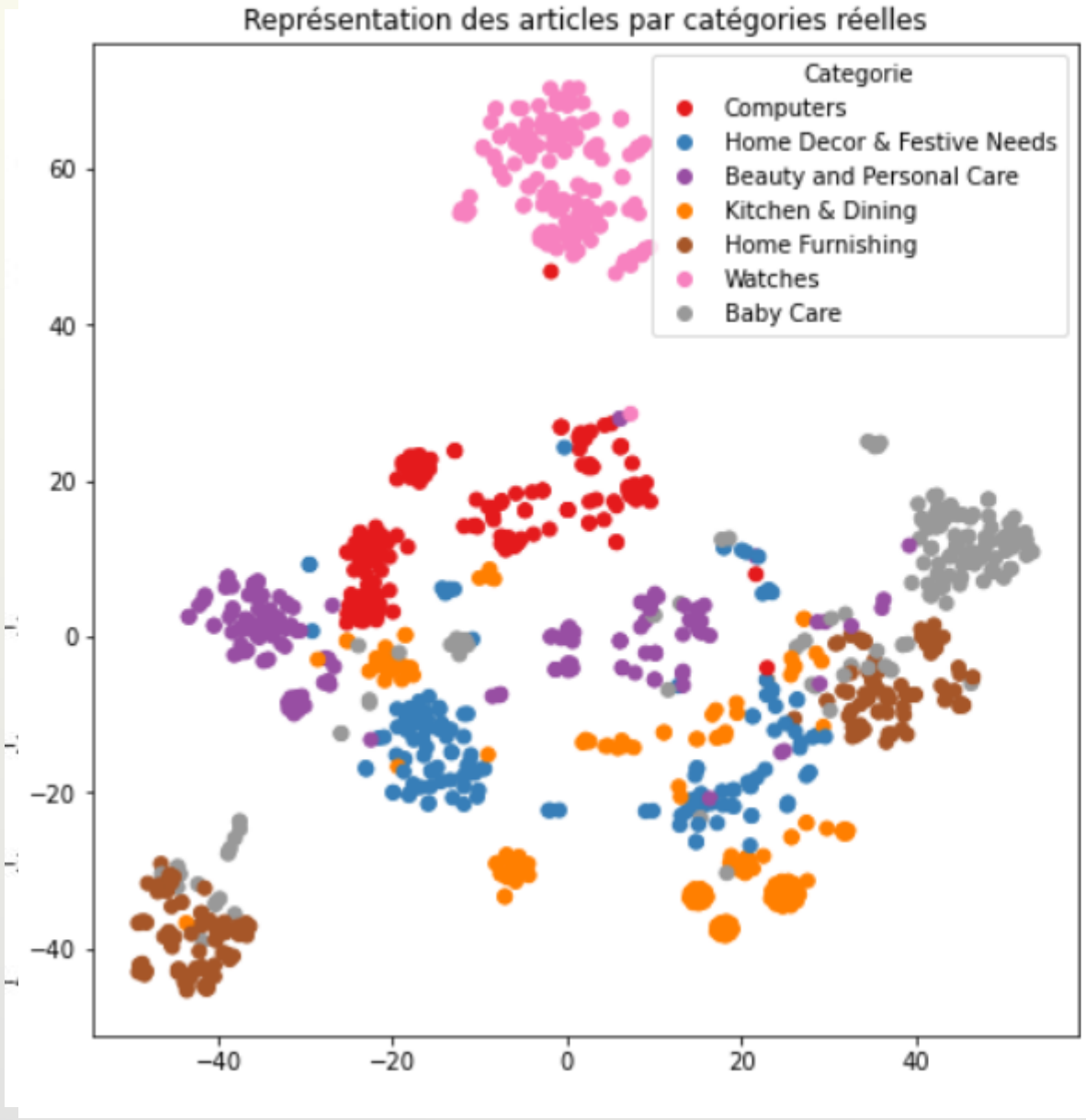
Descriptions préprocessées de certains articles 'Beauty and Personal Care' bien clusterisés en 'Beauty and Personal Care' :

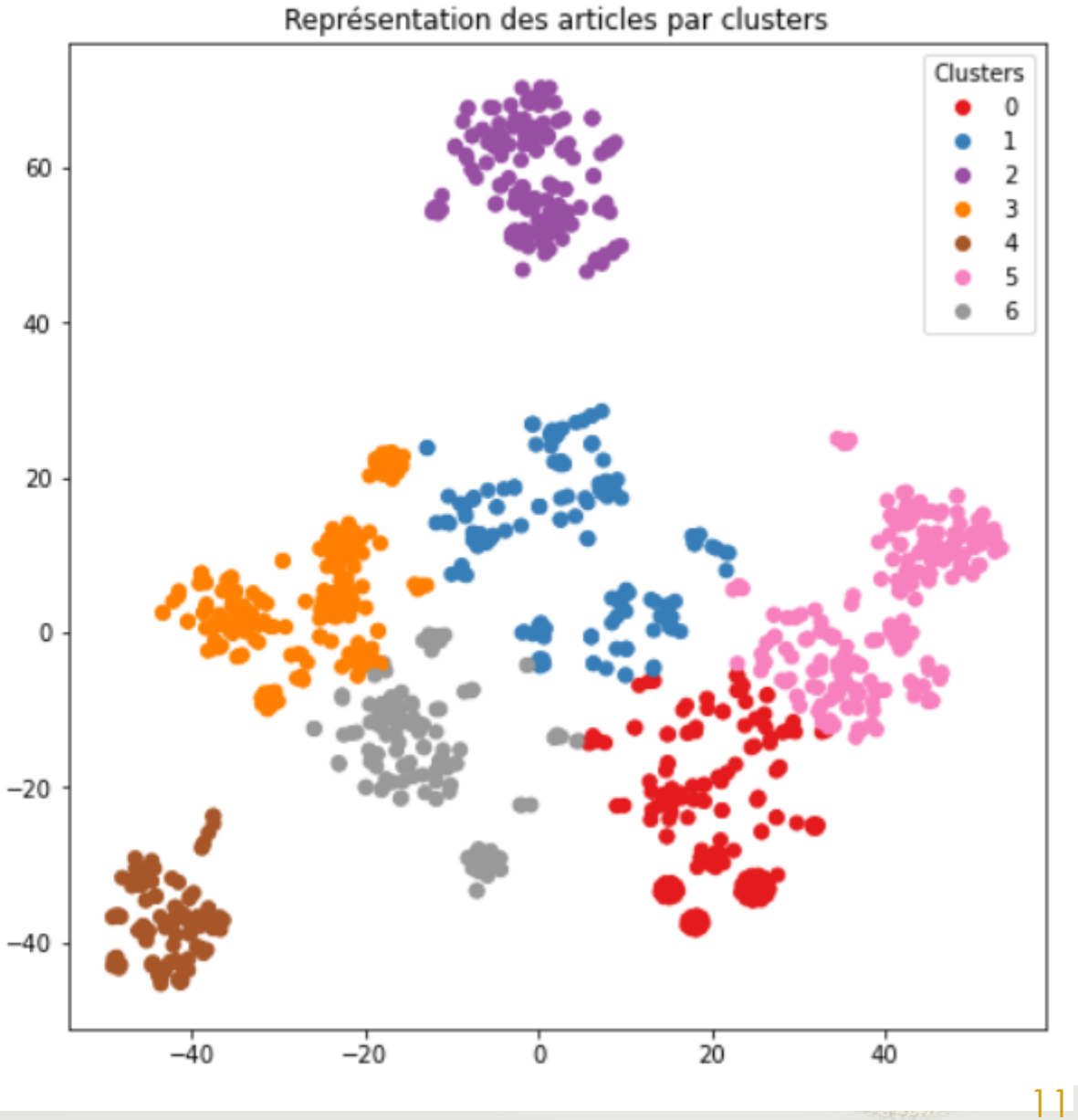
nd link box extender pl

array(['buy industry bangle four roll ring earing box vanity pouch industry bangle four roll ring earing box vanity pouch best price free shipping cash delivery genuine product day replacement guarantee',
'buy vanity pouch vanity pouch best price free shipping cash delivery genuine product day replacement guarantee',
'buy sally cross festival kit price genuine product day replacement guarantee free shipping cash delivery',
'buy equinox body fat analyzer genuine product day replacement guarantee free shipping cash delivery',
'buy wild stone red juice set genuine product day replacement guarantee free shipping cash delivery',
'buy ice drive dynamic pulse set genuine product day replacement guarantee free shipping cash delivery',
'vincent valentine set dark fire dark fire majesty deodorant set set price dark fire sparkle enchanting top note blend black currant star anise distinctive spiciness galanga root depth green pepper extract warmth base rich woody balsam cedar vanilla bean passionate royal majesty latest beautiful perfume anna young extravagant woman neo classical touch dark fire sparkle enchanting top note blend black currant star anise distinctive spiciness galanga root depth green pepper extract warmth base rich woody balsam cedar vanilla bean passionate royal majesty latest beautiful perfume anna young extravagant woman neo classical touch',
'buy black code gift set set genuine product day replacement guarantee free shipping cash delivery',
'buy wild stone smoke juice set genuine product day replacement guarantee free shipping cash delivery',
'buy set genuine product day replacement guarantee free shipping cash delivery'],
'',

arché

‘word2v USE entrainé

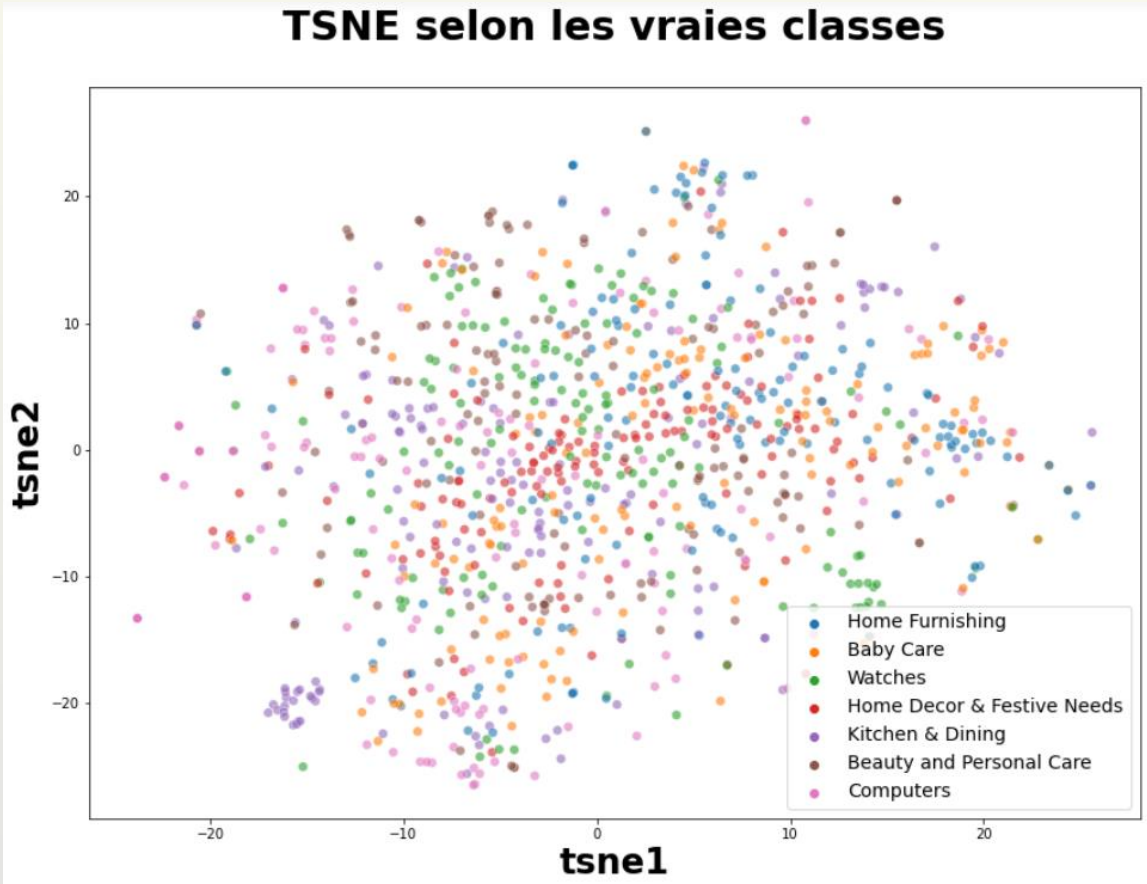


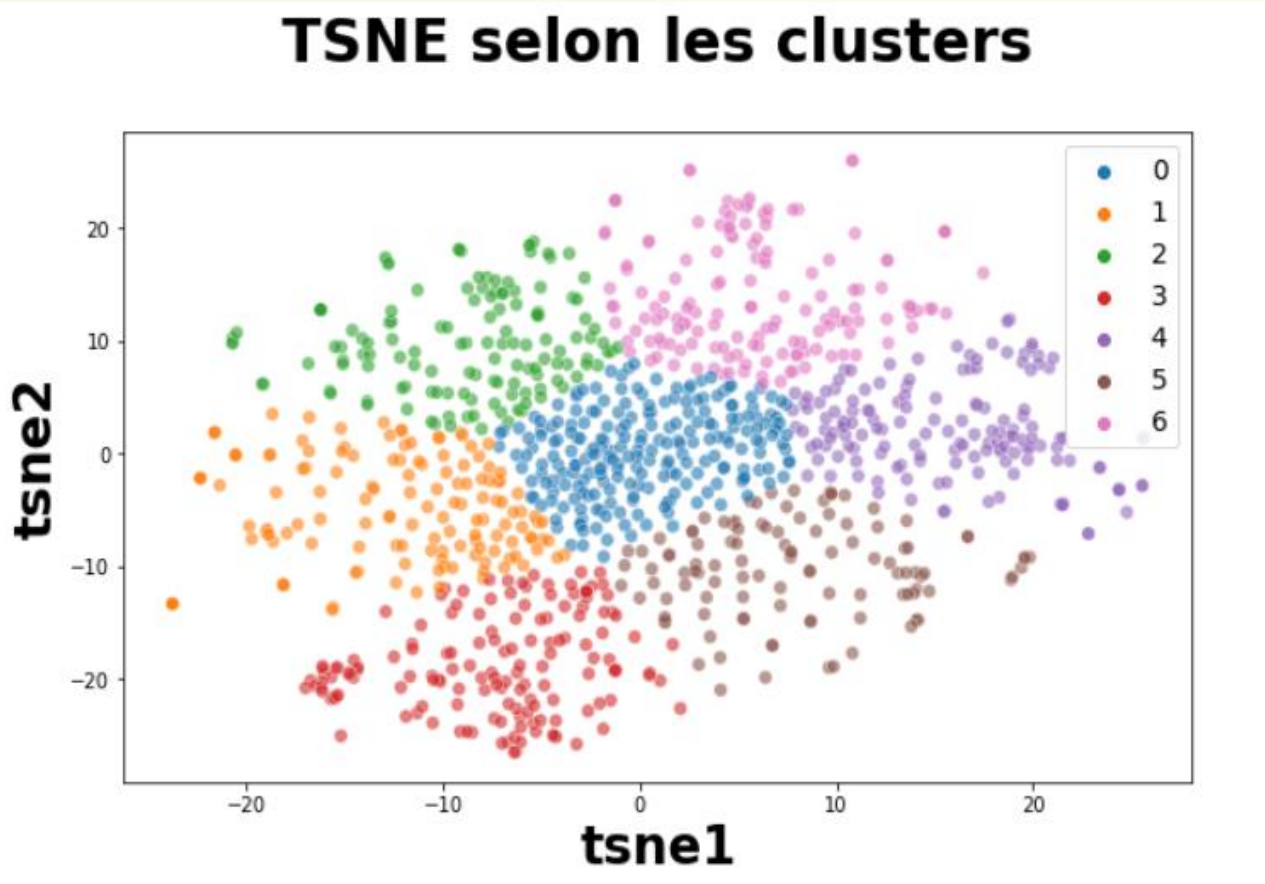


APPROCHES DE MODÉLISATION DES IMAGES

- Extraction des features image via différentes approches :
  - Un algorithme de type SIFT
  - Des algorithmes de type CNN Transfer Learning
    - VGG16
    - EfficientNet
    - ResNet50
- Réduction de dimension des features obtenus grâce à un t-sne (2 composantes).
- Clustering (k-means) sur les données réduites puis calcul de l'Adjusted Rand Index (ARI) par rapport aux vrais catégories.

APPROCHES DE MODÉLISATION DES IMAGES

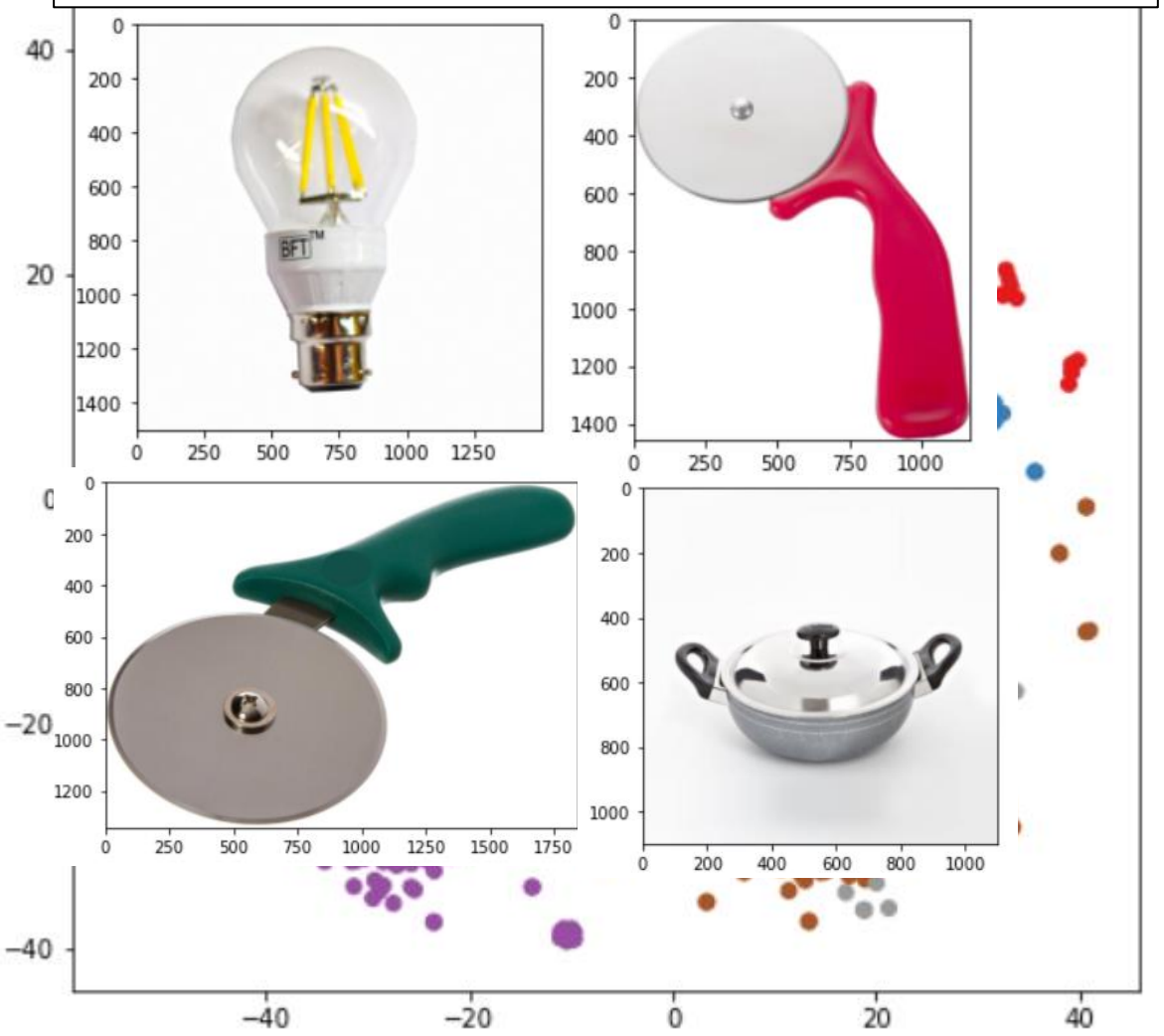




ResNet50

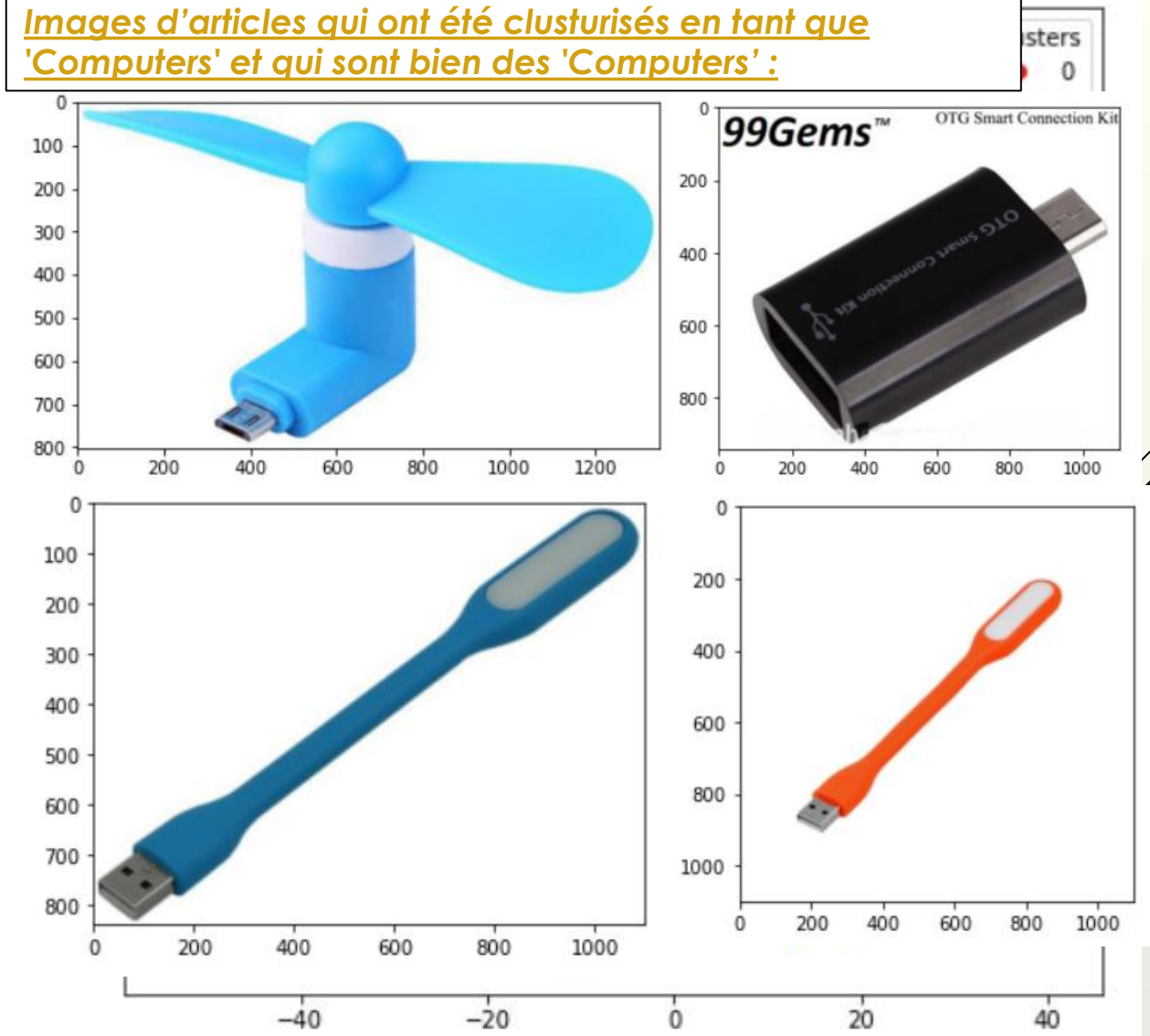
Représentation des articles par catégories réelles

Images d'articles qui ont été clusturisés en tant que 'Computers' alors que ce sont des 'Kitchen & Dining' :



Représentation des articles par clusters

Images d'articles qui ont été clusturisés en tant que 'Computers' et qui sont bien des 'Computers' :



ASSOCIATION DU TEXTE ET DE L'IMAGE

- Association des approches ayant obtenus le meilleur score ARI pour le texte avec celle ayant obtenus le meilleur score ARI pour les images.

- Deux méthodes testées :
  - Agrégation des 2 composantes du t-sne obtenus pour le texte et les images puis réalisation d'un clustering (k-means)
  - Agrégation directement des features extraites pour le texte et les images puis réalisation d'un t-sne et d'un clustering (k-means)

- Meilleurs scores ARI obtenus pour le texte avec l'approche Tf-idf et USE.

- Meilleurs scores ARI obtenus pour les images avec les algorithmes EfficientNet et ResNet50.

ASSOCIATION DU TEXTE ET DE L'IMAGE

- Rappel des scores ARI obtenus :
  - Tf-idf = 0,53 ; USE = 0,43
  - EfficientNet = 0,36 ; ResNet50 = 0,43

- Résultats des associations :

t-sne	EfficientNet	ResNet50
Tf-idf	0,53	0,55
USE	0,46	0,53

Features puis t-sne	EfficientNet	ResNet50
Tf-idf	0,36	0,42
USE	0,35	0,42

CONCLUSION

- Les caractéristiques extraites pour le texte et les images permettent de regrouper des produits de même catégorie.

- Certaines catégories sont très bien clusterisées ('Watches', 'Home Furnishing'), d'autres sont plus compliquées à différencier ('Computers').

- Cependant en appliquant un modèle de classification aux images par exemple, on obtient tout de même de bons résultats sur l'ensemble du jeu de test (86% d'accuracy pour EfficientNet par exemple, non présenté dans ces slides mais présenté dans le notebook).