Projet 6 : Classifiez automatiquement des biens de consommation

Oumeima EL GHARBI

OpenClassrooms – Data Scientist

Soutenance : 27/11/2022

Plan

Introduction

- Problématique
- Analyse exploratoire

I. NLP

- Preprocessing
- Bag of words
- Word Embeddings

II. Images

- SIFT
- CNN Transfer Learning

Conclusion

- Faisabilité du moteur de classification
- Recommandations

Introduction

Problématique :

« Sur *la place de marché*, des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs, et est donc peu fiable. De plus, le volume des articles est pour l'instant très petit.

Pour rendre l'expérience utilisateur des vendeurs (faciliter la mise en ligne de nouveaux articles) et des acheteurs (faciliter la recherche de produits) la plus fluide possible, et dans l'optique d'un passage à l'échelle, il devient nécessaire d'automatiser cette tâche.

Donc, nous voulons étudier la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant.»

Implémentation :

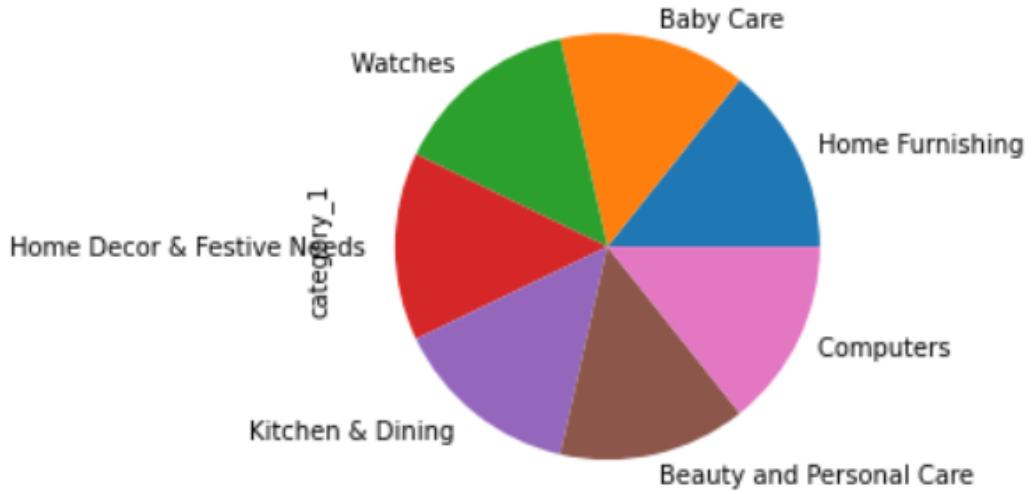
Cadre : NLP, Computer Vision, Clustering

Modèles testés :

- **Bag of words** : CountVectorizer, TF-IDF
- **Word embeddings** :
  - Word2Vec
  - BERT
  - USE
- **SIFT**
- **CNN**

Evaluation : score ARI et matrice de confusion.

Analyse exploratoire



1050 produits avec description et photo.

7 catégories : 150 produits par catégorie.

	category	target
0	Baby Care	0
1	Beauty and Personal Care	1
2	Computers	2
3	Home Decor & Festive Needs	3
4	Home Furnishing	4
5	Kitchen & Dining	5
6	Watches	6

	product_name	text	category	target
0	Elegance Polyester Multicolor Abstract Eyelet ...	Key Features of Elegance Polyester Multicolor ...	Home Furnishing	4
1	Sathiyas Cotton Bath Towel	Specifications of Sathiyas Cotton Bath Towel (...	Baby Care	0
2	Eurospa Cotton Terry Face Towel Set	Key Features of Eurospa Cotton Terry Face Towe...	Baby Care	0
3	SANTOSH ROYAL FASHION Cotton Printed King size...	Key Features of SANTOSH ROYAL FASHION Cotton P...	Home Furnishing	4
4	Jaipur Print Cotton Floral King sized Double B...	Key Features of Jaipur Print Cotton Floral Kin...	Home Furnishing	4

	product_name	image_path	category	target
0	Elegance Polyester Multicolor Abstract Eyelet ...	55b85ea15a1536d46b7190ad6fff8ce7.jpg	Home Furnishing	4
1	Sathiyas Cotton Bath Towel	7b72c92c2f6c40268628ec5f14c6d590.jpg	Baby Care	0
2	Eurospa Cotton Terry Face Towel Set	64d5d4a258243731dc7bbb1eef49ad74.jpg	Baby Care	0
3	SANTOSH ROYAL FASHION Cotton Printed King size...	d4684dcdc759dd9cdf41504698d737d8.jpg	Home Furnishing	4
4	Jaipur Print Cotton Floral King sized Double B...	6325b6870c54cd47be6ebfbffa620ec7.jpg	Home Furnishing	4

I) NLP

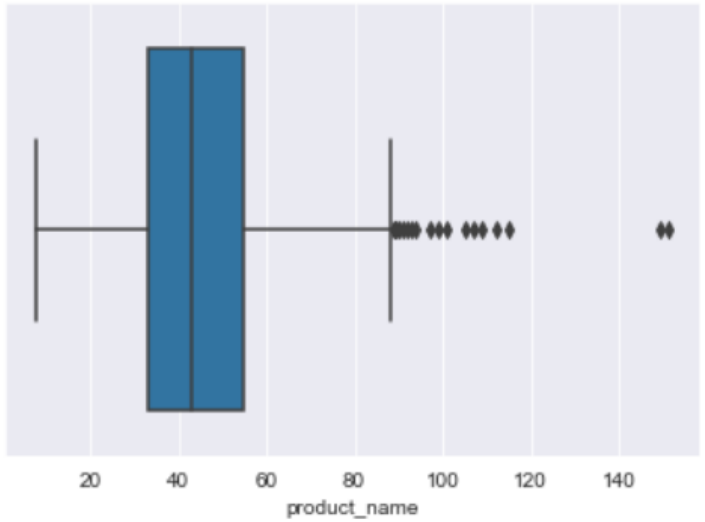
Preprocessing

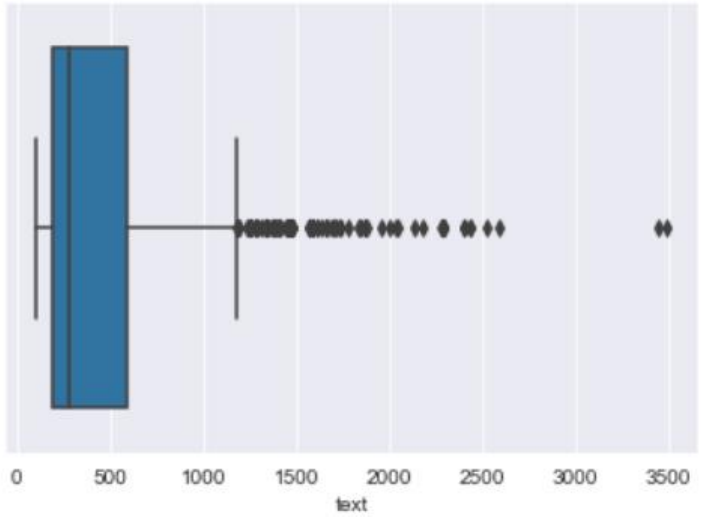
Bag of words

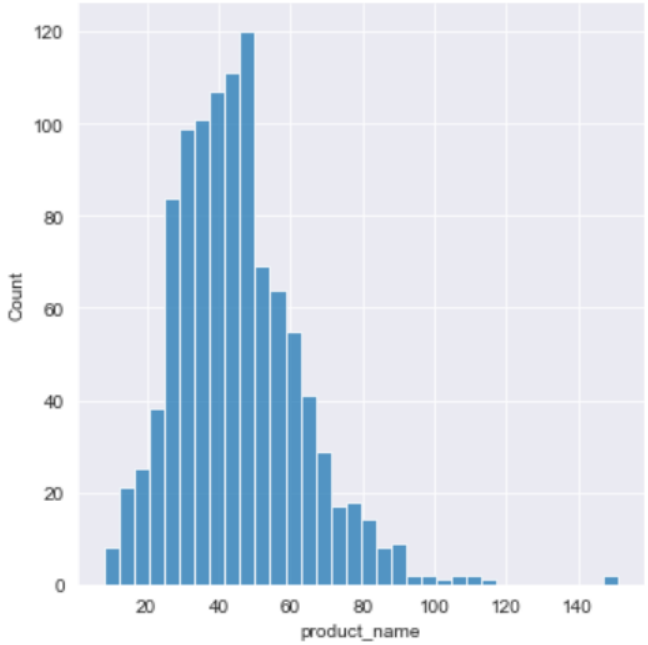
Word Embeddings

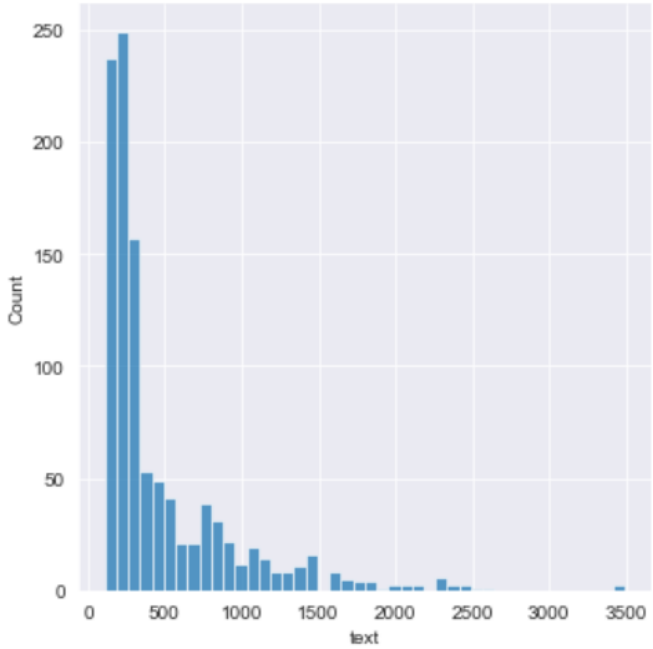
NLP
1) Preprocessing

Avant nettoyage et normalisation









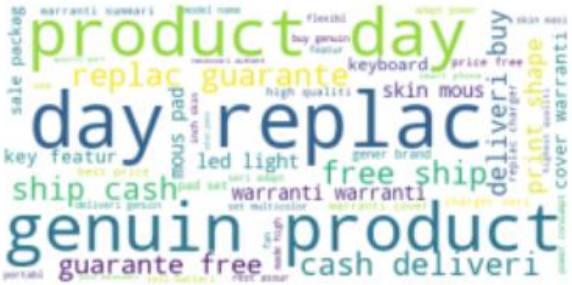
NLP
1) Preprocessing

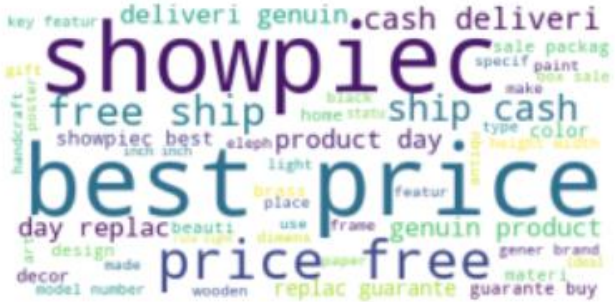
Après nettoyage et normalisation

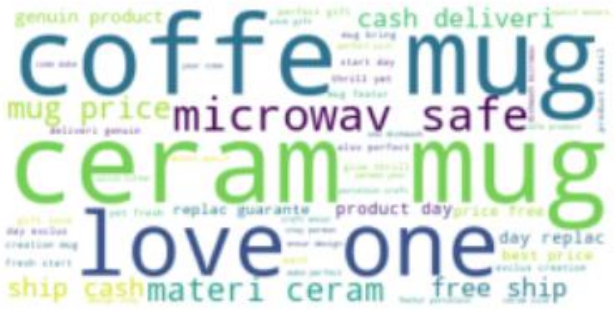
- Lower case
- Tokenization
- Stopwords
- Unique words : occurrence unique d'un mot
- Stemming and Lemmatization
- English words
- Duplicated words

- « text »

- We got better wordclouds when we deleted more duplicated words.
- However, we need to have non empty descriptions.





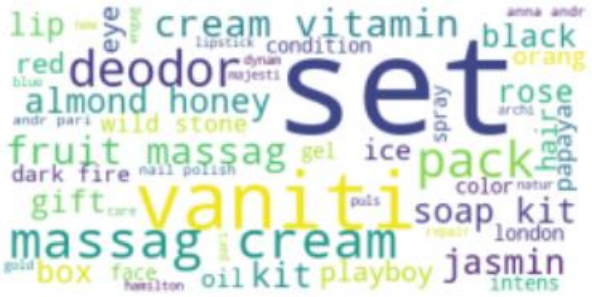


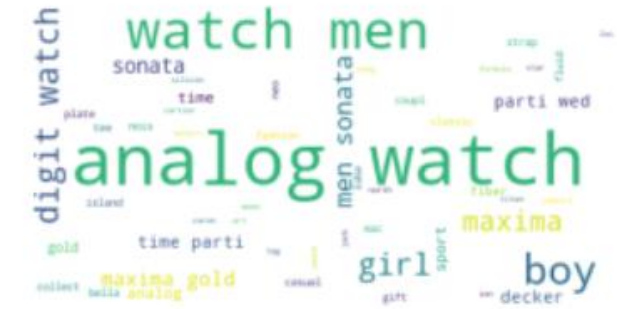
NLP
1) Preprocessing

Après nettoyage et normalisation

- Lower case
- Tokenization
- Stopwords
- Unique words : occurrence unique d'un mot
- Stemming and Lemmatization
- English words
- Duplicated words

- « product _name »





NLP
1) Preprocessing

__Before__
Mahadev Handicrafts Cotton Cartoon Double Bedsheet

__After__
handicraft cotton cartoon doubl

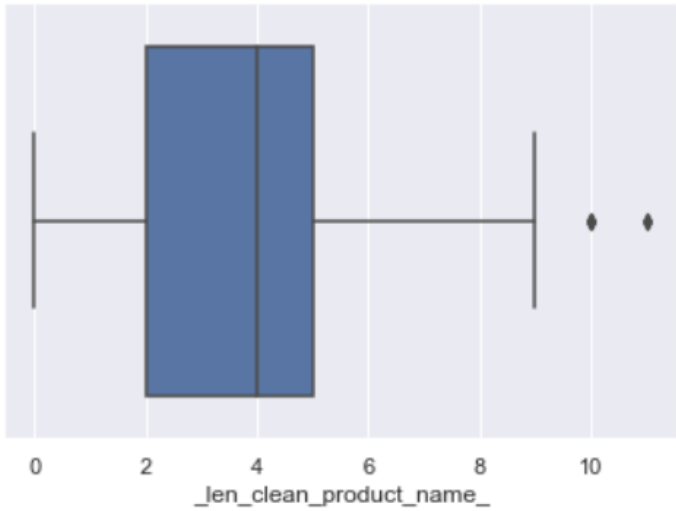
__Before__
LITTLE FEETZ Baby Girl's Solid Top & Skirt Set

__After__
babi girl solid top set

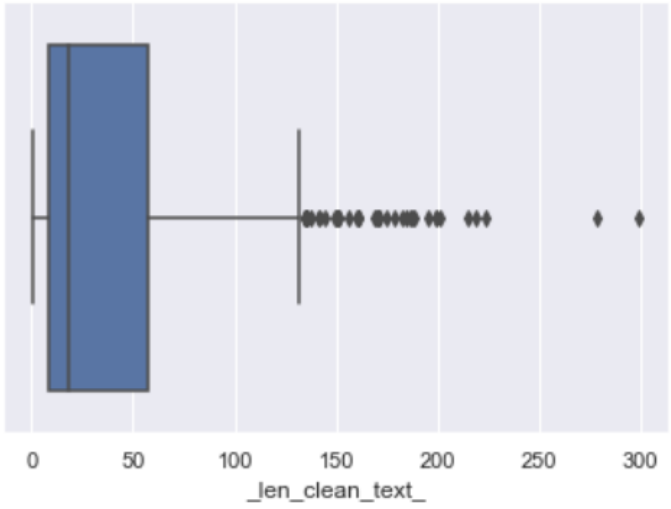
__Before__
Lollipop Lane Tiddly Wink Safari Bath Set

__After__
lollipop bath set

« Product name » nettoyé



« Text » nettoyé



__Before__
Key Features of Kandyfloss Baby Boy's, Baby Girl's Romper Fabric: COTTON Brand Color: RED,Specifications of Kandyfloss Baby Boy's, Baby Girl's Romper Top Details Number of Contents in Sales Package Pack of 3 Fabric COTTON Type Romper Neck ENVELOPE NECK General Details Pattern Geometric Print Ideal For Baby Boy's, Baby Girl's In the Box 3 Romper

__After__
key featur babi boy babi girl romper fabric brand color red specif babi boy babi girl romper top detail number content sale packag pack fabric type romper neck envelop neck gener detail pattern geometr print ideal babi boy babi girl box romper

NLP
2) Bag of words

- 1-gram : comptage du nombre d'occurrence de chaque mot du vocabulaire
- TF-IDF

- Expériences :

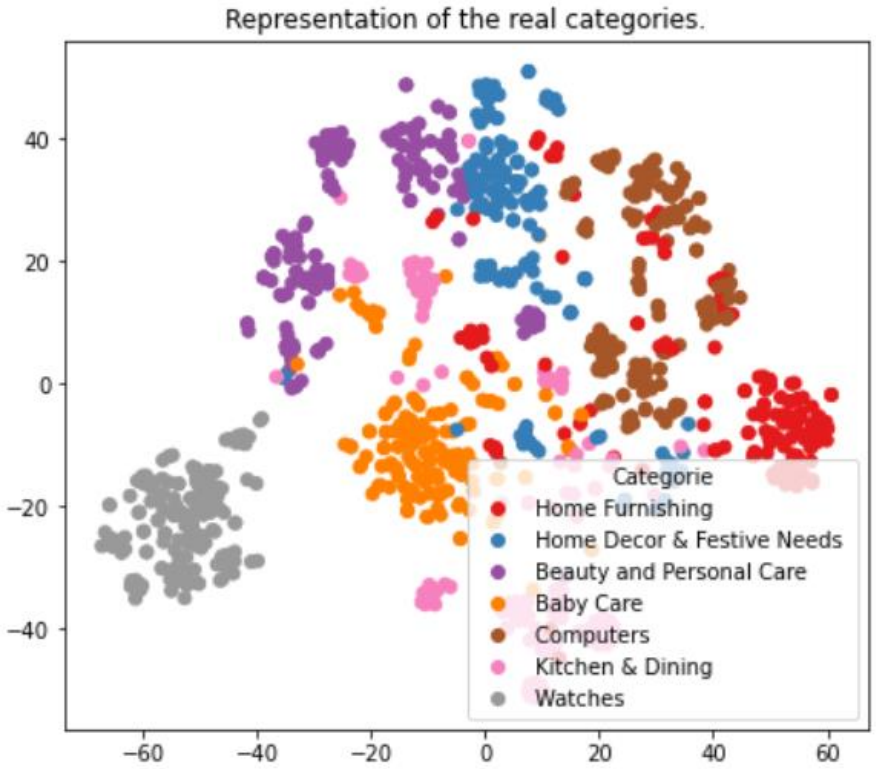
- 1) fit / transform « text » : ARI correct
- 2) fit / transform « product_name » et « text » : ARI correct
- 3) fit « product_name », transform « text » : ARI faible

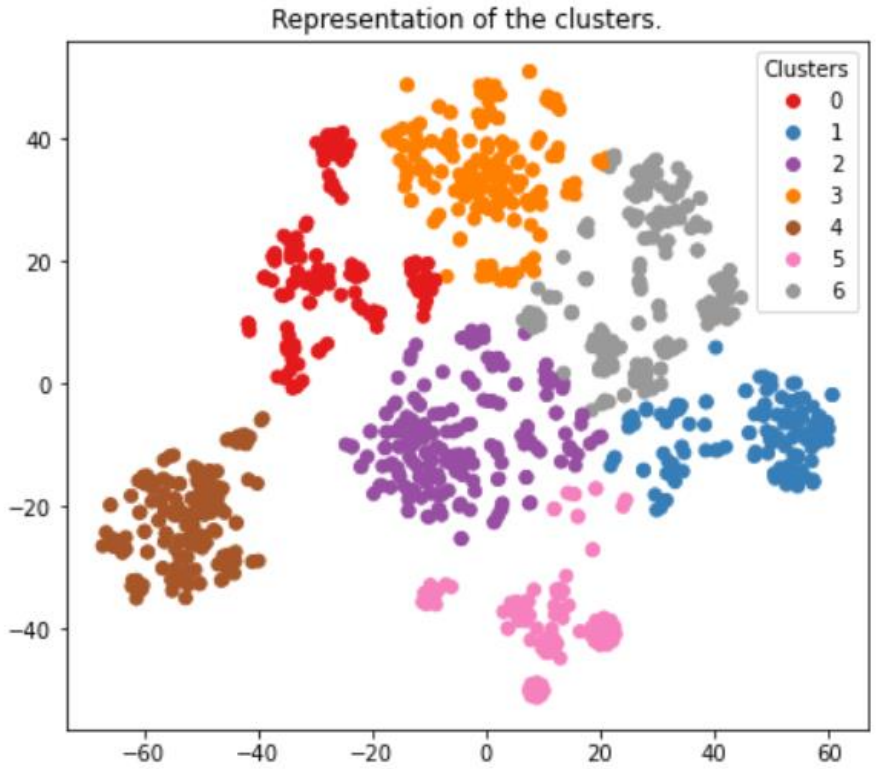
NLP
2) Bag of words

```
CountVectorizer :
-----------------
ARI : 0.399, time : 16.0 seconds.

Tf-idf :
--------
ARI : 0.5567, time : 15.0 seconds.
```

- Expériences :
- 1) fit / transform « text »





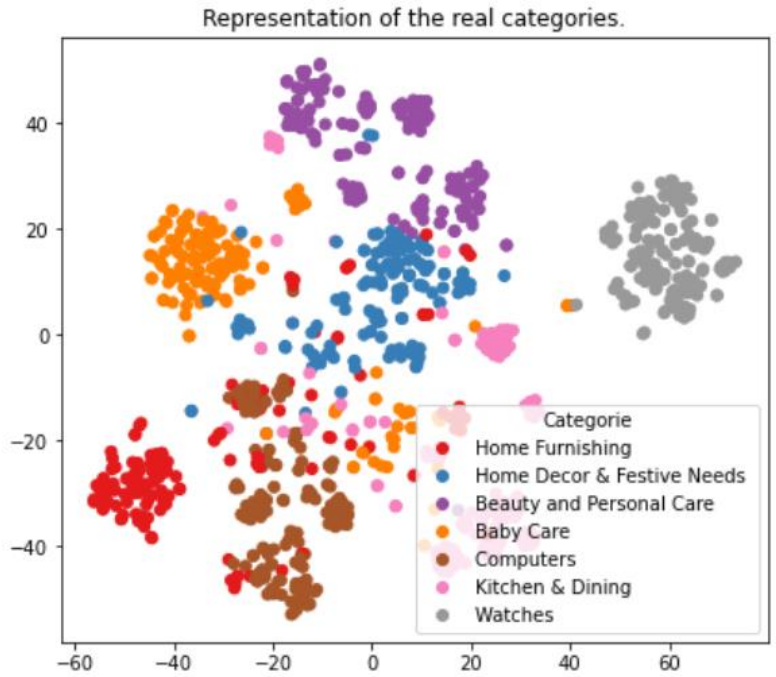

```
ARI :  0.5567
```

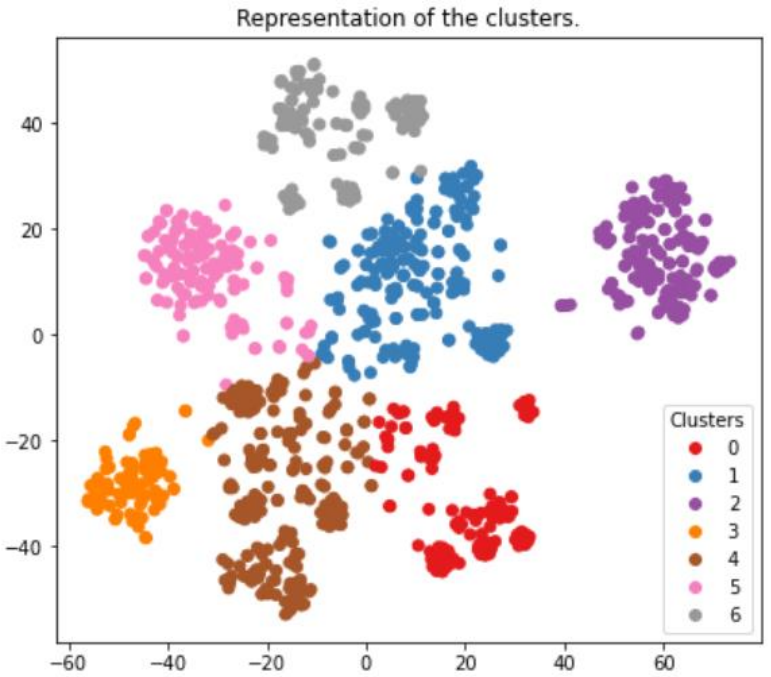
NLP
2) Bag of words

```
CountVectorizer :
-----------------
ARI : 0.4025, time : 18.0 seconds.

Tf-idf :
--------
ARI : 0.5585, time : 15.0 seconds.
```

- Expériences :
- 2) fit / transform « product_name » et « text »

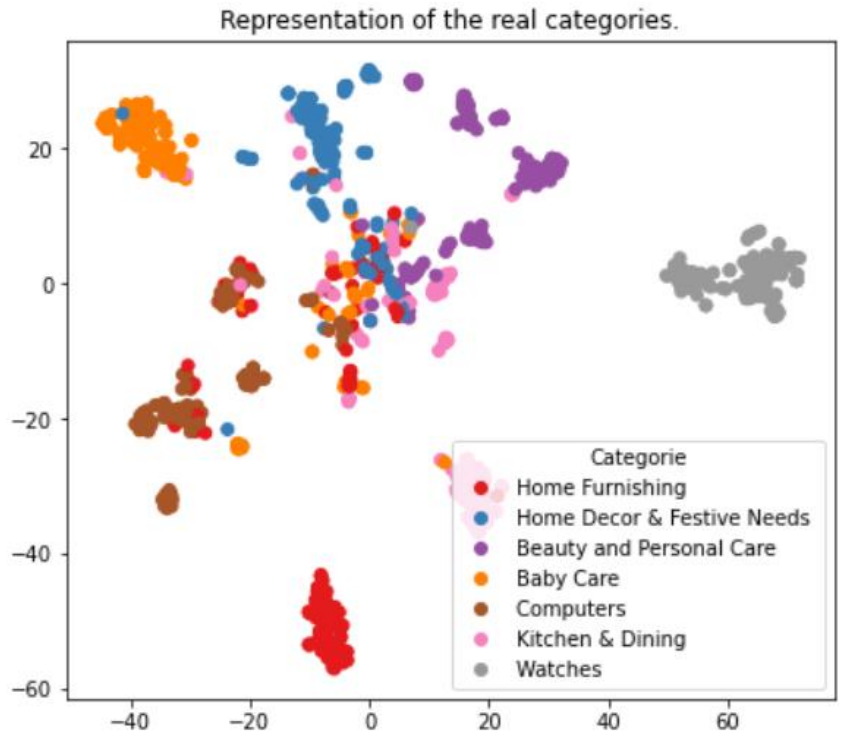




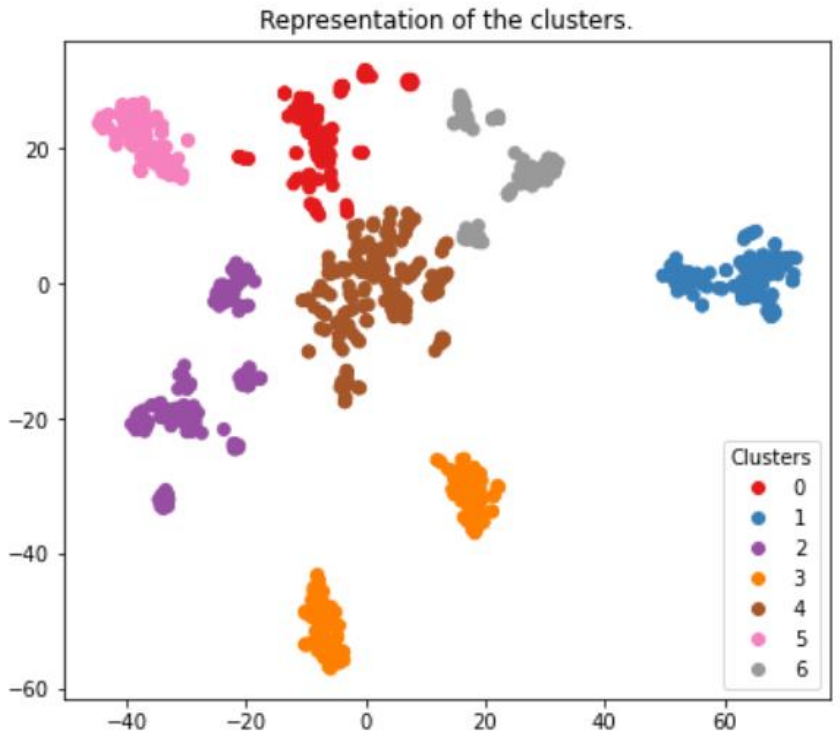
ARI : 0.5585

NLP
3) Word embeddings : Word2Vec

« product_name »

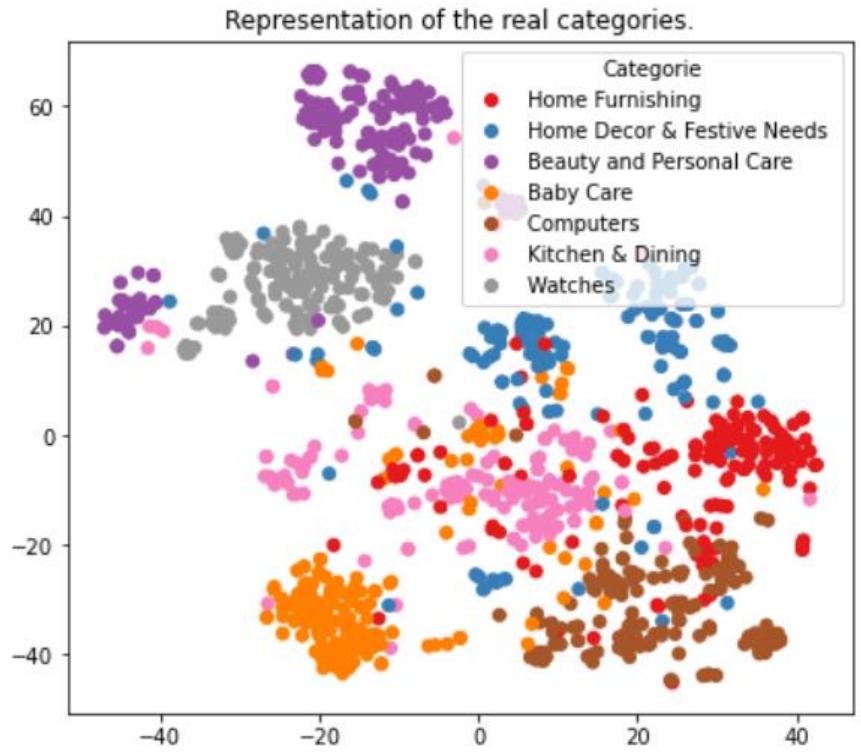


ARI : 0.5056

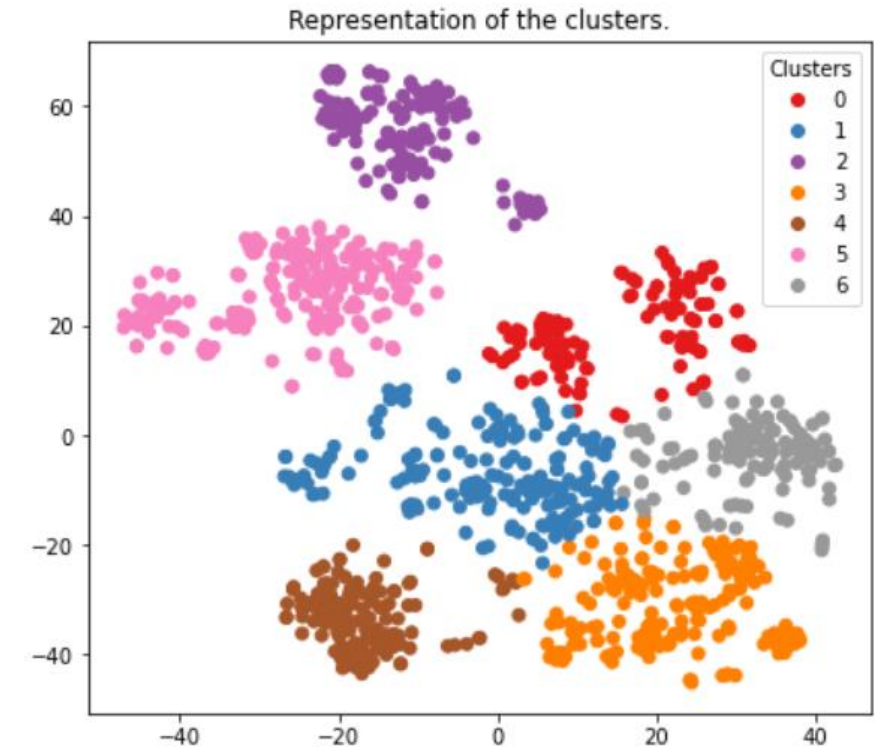


NLP
3) Word embeddings : BERT Hugging Face

« product_name »
Modèle BERT pré-entraîné : 'bert-base-uncased'

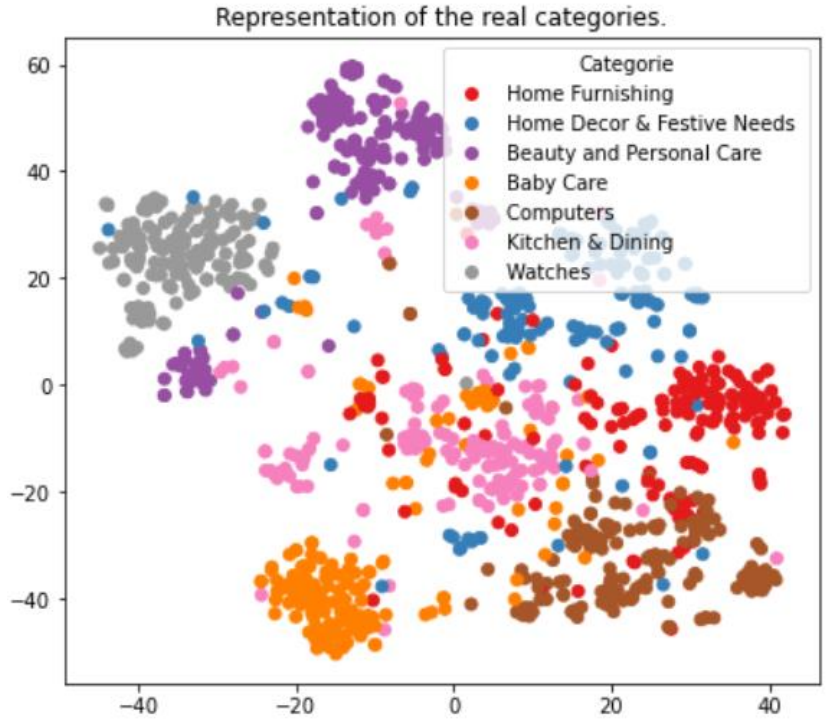


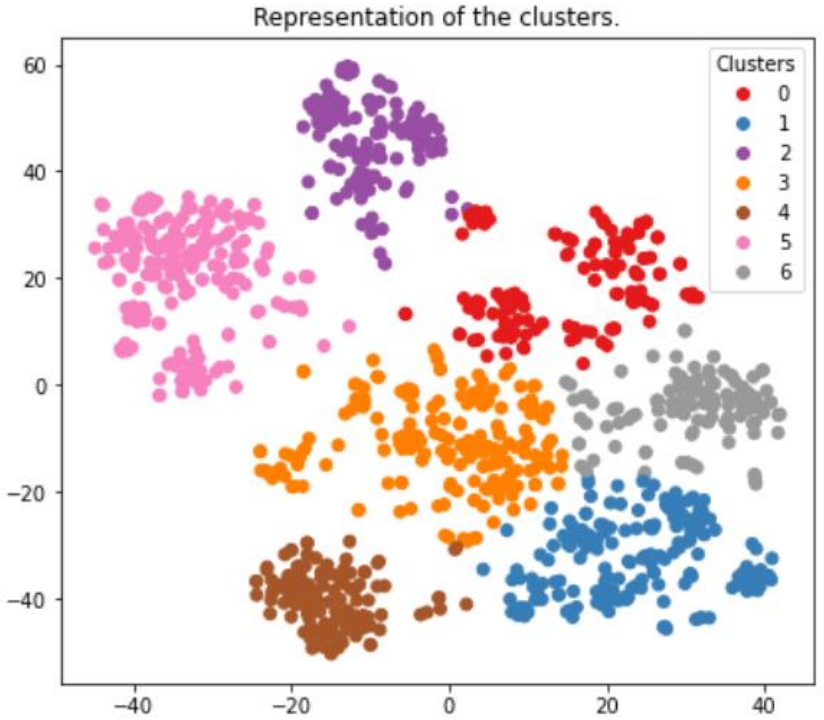
ARI : 0.6187



NLP
3) Word embeddings : BERT TF

« product_name »
Modèle BERT pré-entraîné : ‘bert_en_uncased_L-12_H-768_A-12/4'

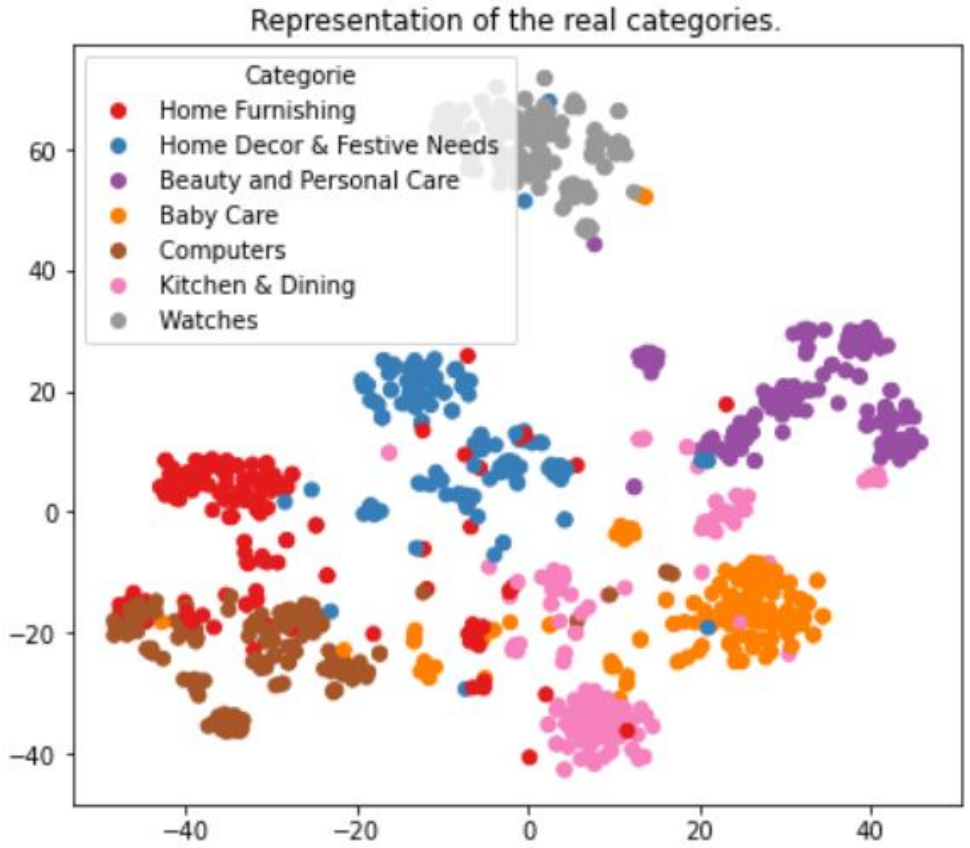




ARI : 0.5928

NLP
3) Word embeddings : USE

« product_name »



ARI : 0.7094

II) Images

SIFT

CNN Transfer Learning

Beauty and Personal Care

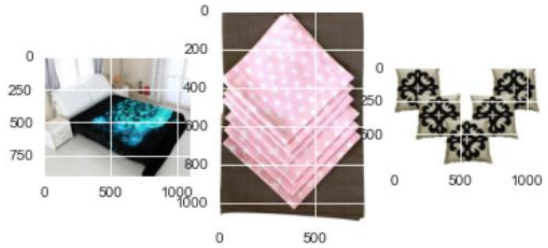
For this category, we have 150 images.

Computers

For this category, we have 150 images.

Home Furnishing

For this category, we have 150 images.



Baby Care

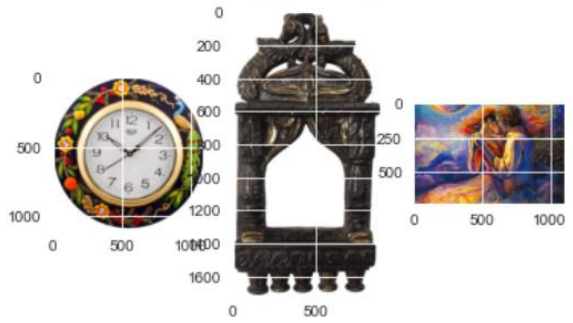
For this category, we have 150 images.

Watches

For this category, we have 150 images.

Home Decor & Festive Needs

For this category, we have 150 images.



Kitchen & Dining

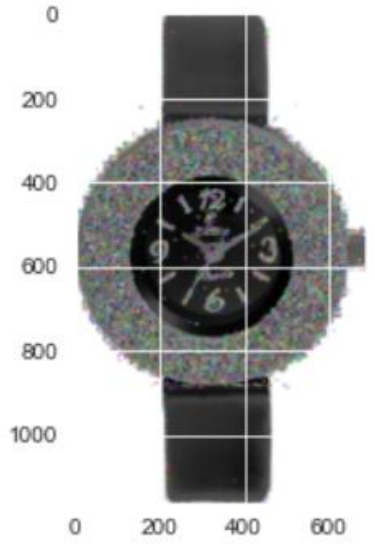
For this category, we have 150 images

Images
1) SIFT

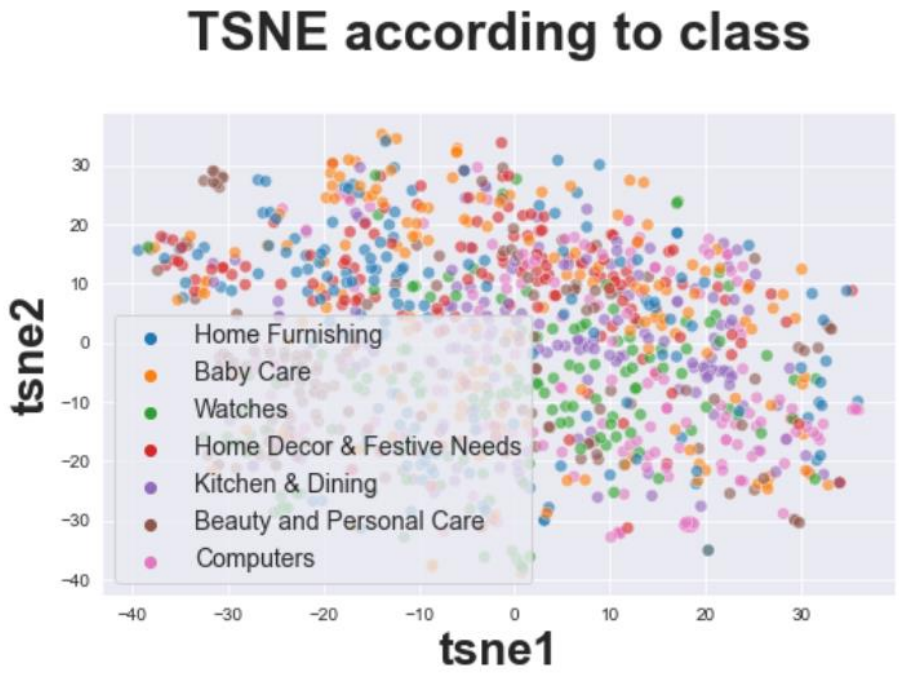
Descripteurs SIFT
Bag of visual words

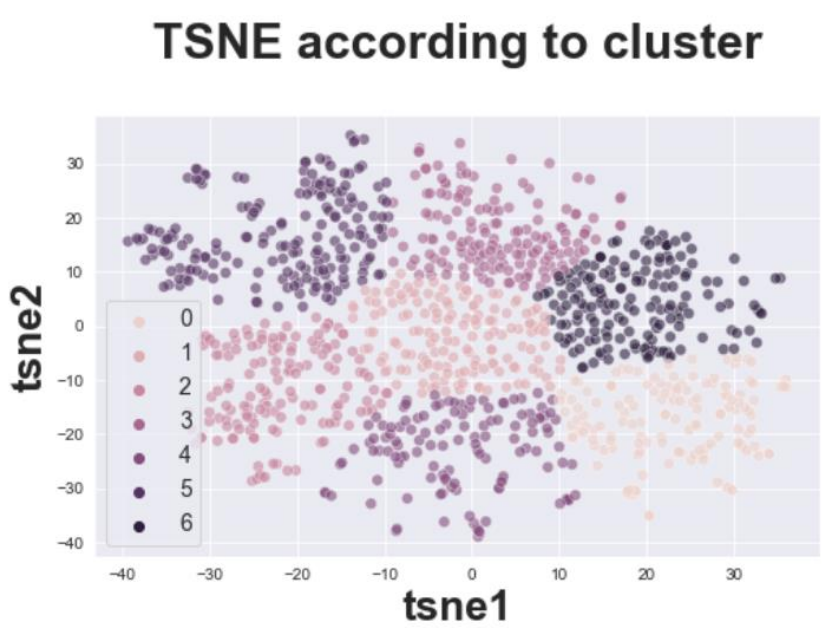
ACP
t-SNE
Kmeans clustering

ARI faible : SIFT ne parvient pas à décrire les différentes catégories



```
SIFT keypoints :  (4150, 128)
```



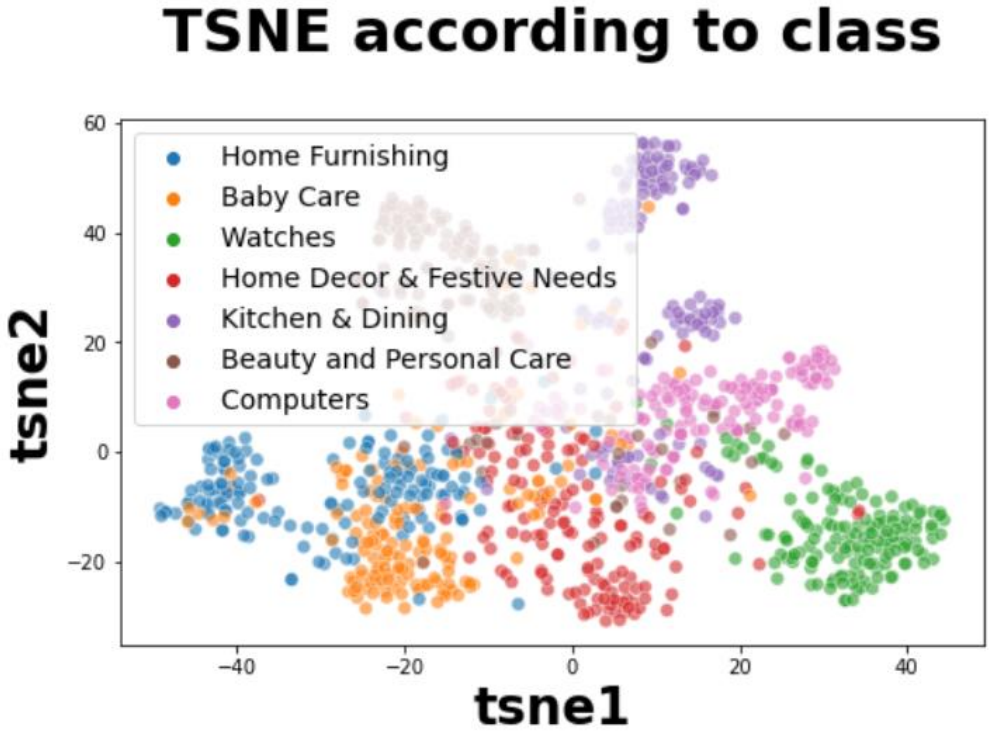


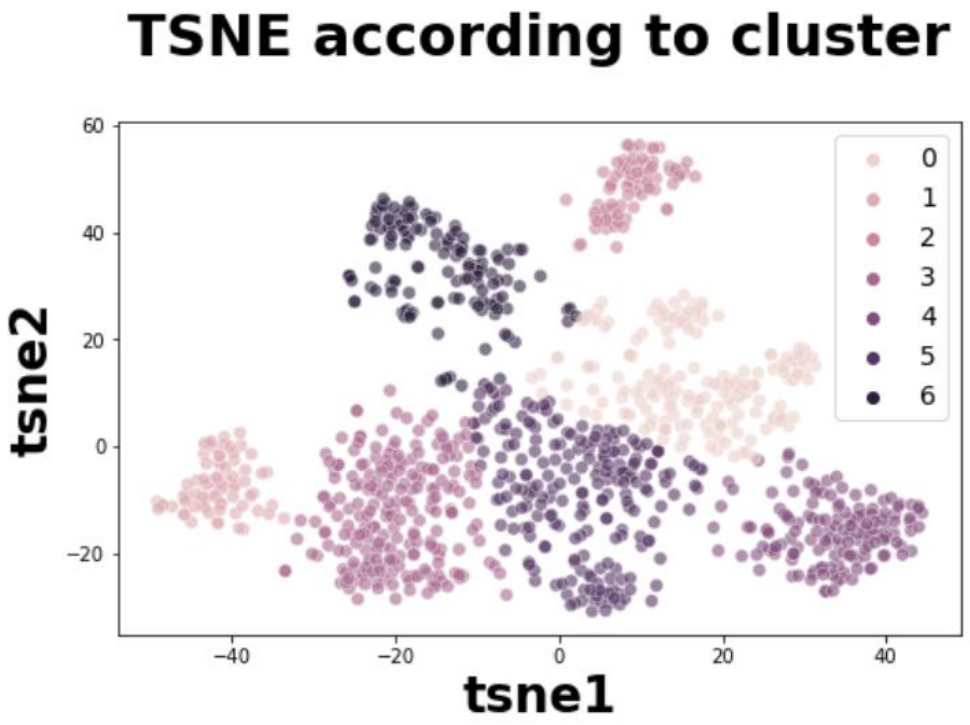

```
ARI :  0.05777523226425486
```

Images
2) CNN Transfer Learning

CNN : VGG16 pré-entraîné sans les deux dernières couches
Extraction des features

ACP
t-SNE
Kmeans clustering





ARI : 0.466

Conclusion

- **Faisabilité d'un moteur de classification**

- Classification des descriptions des produits
- Bon résultats pour les scores ARI : utilisation du nom du produit
- Bonne distinction des différentes catégories par t-SNE

- Classification des images des produits
- Bon résultats pour les scores ARI avec CNN Transfer Learning
- Distinction des différentes catégories par t-SNE pour le CNN mais non pas avec l'algorithme SIFT

- **Recommandations**
- Utiliser à la fois les images et le texte pour le moteur de classification
- Catégoriser plus précisément les produits : 7 catégories ne suffiront peut-être pas pour avoir une bonne précision de la classification

Merci !