SOUTENANCE PROJET 6

Classifiez automatiquement des biens de consommation

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR
MONSIEUR PASCAL FARES

"l'IA et le traitement du langage naturel pourraient bien rendre Internet beaucoup plus utile qu'il ne l'est aujourd'hui"
Par Vint CERF

ORDRE DU JOUR

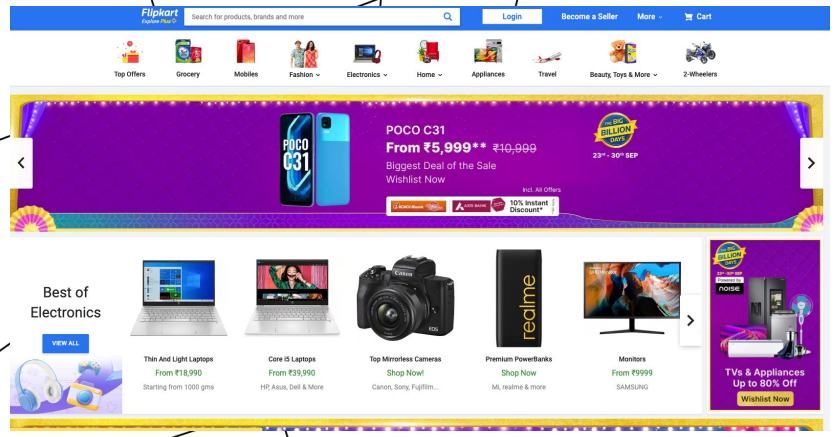
Ouverture de la problématique - 5 min

Explication des prétraitements et des résultats du clustering - 10 min

Conclusions et recommandations du moteur de classification - 5 min

Discussion - 5-10 min

OUVERTURE DE LA PROBLÉMATIQUE



Place de marché : marketplace e-commerce

Vendeurs : articles avec photo et description

Attribution manuelle : fastidieuse et peu fiable

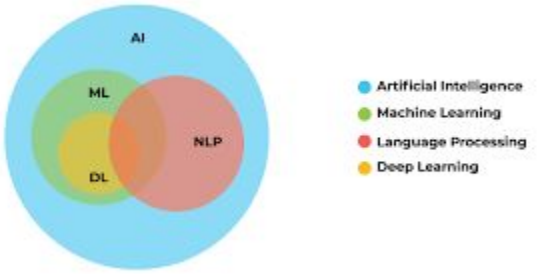
Perspective de passage à l'échelle

<u>Objectifs principaux:</u>

- Automatiser l'attribution d'une catégorie d'un article
- Etudier la faisabilité d'un moteur de classification

<u>Moyens pour y parvenir:</u>

- Mise à disposition de 2 datasets associés du site d'e-commerce flipkart.com :
  - données textuelles
  - données d'images



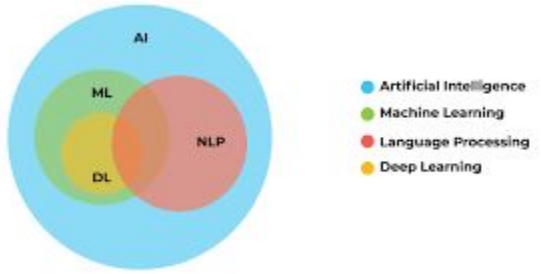
Données textuelles au format .csv

1 050 produits avec 15 variables descriptives

```
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   object
 2   product_url             1050 non-null   object
 3   product_name            1050 non-null   object
 4   product_category_tree   1050 non-null   object
 5   pid                     1050 non-null   object
 6   retail_price            1049 non-null   float64
 7   discounted_price        1049 non-null   float64
 8   image                   1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null   bool
 10  description             1050 non-null   object
 11  product_rating          1050 non-null   object
 12  overall_rating          1050 non-null   object
 13  brand                   712 non-null    object
 14  product_specifications  1049 non-null   object
dtypes: bool(1), float64(2), object(12)
```


```
raw_corpus[:1000]
```

```
'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstr
act Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curt
ain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room envir
onment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a bright
and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valanc
e curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wi
sh good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous
beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you '
```



Données textuelles au format .csv

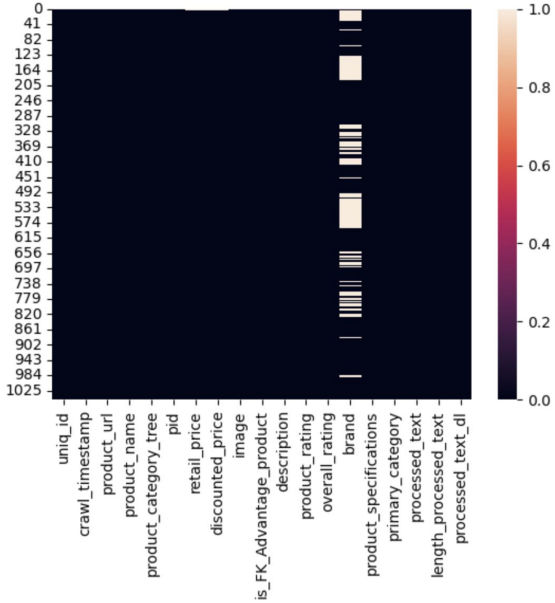
duplicates et données manquantes

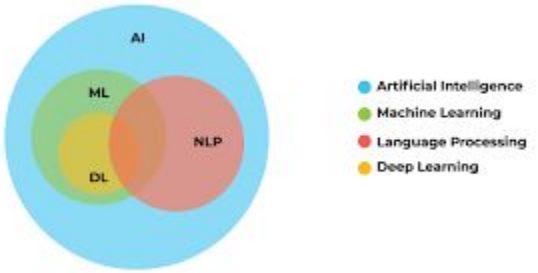
```
col : uniq_id -> duplicated : 0
col : crawl_timestamp -> duplicated : 901
col : product_url -> duplicated : 0
col : product_name -> duplicated : 0
col : product_category_tree -> duplicated : 408
col : pid -> duplicated : 0
col : retail_price -> duplicated : 695
col : discounted_price -> duplicated : 625
col : image -> duplicated : 0
col : is_FK_Advantage_product -> duplicated : 1048
col : description -> duplicated : 0
col : product_rating -> duplicated : 1023
col : overall_rating -> duplicated : 1023
col : brand -> duplicated : 559
col : product_specifications -> duplicated : 65
col : primary_category -> duplicated : 1043
col : processed_text -> duplicated : 68
col : length_processed_text -> duplicated : 899
col : processed_text_dl -> duplicated : 67
```

	781	101
uniq_id	9924fba9b2a738e5a141995952e73104	c7fcd0d983a51283e58b806f065cc920
crawl_timestamp	2015-12-12 11:46:53 +0000	2016-01-04 00:20:04 +0000
product_url	http://www.flipkart.com/perucci-pc-222-decker-...	http://www.flipkart.com/diviniti-mdf-color-rel...
product_name	Perucci PC-222 Decker Analog Watch - For Men	Diviniti MDF Color Religious Wall Hanging Size...
product_category_tree	["Watches >> Wrist Watches >> Perucci Wrist Wa...	["Home Decor & Festive Needs >> Table Decor & ...
pid	WATE6XU7GKU8KMGZ	SHIEBWNUB6QGHNGD
retail_price	1900.0	13400.0
discounted_price	290.0	13400.0
image	9924fba9b2a738e5a141995952e73104.jpg	c7fcd0d983a51283e58b806f065cc920.jpg
is_FK_Advantage_product	False	False
description	Perucci PC-222 Decker Analog Watch - For Men ...	Buy Diviniti MDF Color Religious Wall Hanging ...
product_rating	No rating available	No rating available
overall_rating	No rating available	No rating available
brand	NaN	Diviniti
product_specifications	{"product_specification"=>[{"key"=>"Chronograp...	{"product_specification"=>[{"key"=>"Brand", "v...
primary_category	Watches	Home Decor & Festive Needs
processed_text	perucci decker analog watch men buy perucci de...	buy diviniti mdf color wall hanging size radha...
length_processed_text	34	31
processed_text_dl	perucci decker analog watch for men buy perucc...	buy diviniti mdf color wall hanging size radha...

```
data.isna().sum()
```

```
uniq_id                     0
crawl_timestamp             0
product_url                 0
product_name                0
product_category_tree       0
pid                         0
retail_price                1
discounted_price            1
image                       0
is_FK_Advantage_product     0
description                 0
product_rating              0
overall_rating              0
brand                     338
product_specifications      1
primary_category            0
processed_text              0
length_processed_text       0
processed_text_dl           0
dtype: int64
```





Données textuelles au format .csv

Catégories

```
data['product_category_tree'].nunique()
```

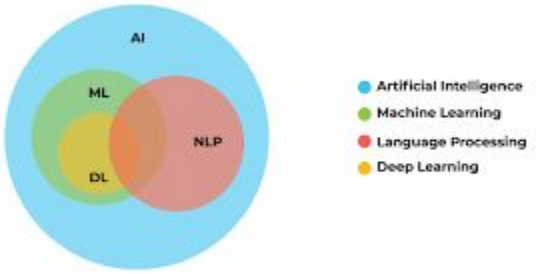
```
642
```

```
data['primary_category'].nunique()
```

```
7
```

```
data['primary_category'].unique()
```

```
array(['Home Furnishing', 'Baby Care', 'Watches',
       'Home Decor & Festive Needs', 'Kitchen & Dining',
       'Beauty and Personal Care', 'Computers'], dtype=object)
```

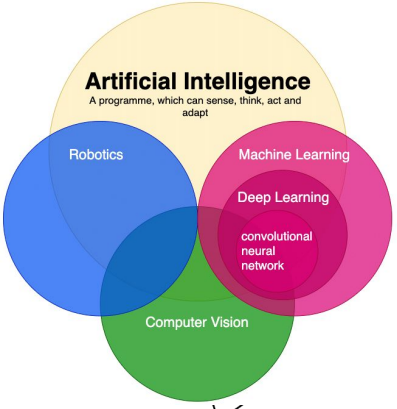


Données textuelles au format .csv

Répartition des 7 catégories

```
data.groupby(by='primary_category').count()
```

primary_category	uniq_id	crawl_timestamp	product_url	product_name
Baby Care	150	150	150	150
Beauty and Personal Care	150	150	150	150
Computers	150	150	150	150
Home Decor & Festive Needs	150	150	150	150
Home Furnishing	150	150	150	150
Kitchen & Dining	150	150	150	150
Watches	150	150	150	150



Données images au format .jpg

1 050 images (une image/produit)

Home Furnishing

Watches

Kitchen and dining

Baby care

Home decor and festive needs

Computers

Analyse des données textuelles

Méthodologie

1 Extraction des features

Transformation textuelle par des vecteurs. Nous utiliserons :
- Bag of words (Tf-idf)
- Word embedding (Word2vec)
- Modèles de langage (USE, BERT)

2 Réduction de dimensions

Réduction de ces features extraites via un t-SNE

3 Clustering

Clustering via KMeans (k = 7 catégories) sur la réduction du t-SNE

4 Analyse de similarité

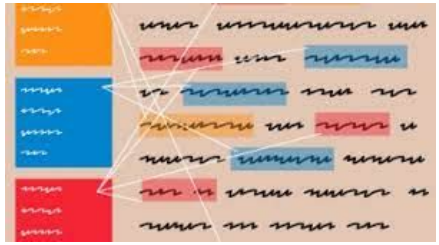
Comparaison entre valeur prédites par KMeans et les catégories réelles via ARI (Adjusted Rand Index)

Analyse des données textuelles - Bag of words

Prétraitement des données textuelles

Utilisation de NLTK pour filtrer les informations avec de la valeur dans le document :

- Texte en minuscule
- Tokenization
- suppression des stopwords
- suppression des mots rares (apparaisse 1 seul fois)
- suppression des mots trop court (au moins 3 lettres)
- suppression des caractères numériques
- Stemming ou lemmatization



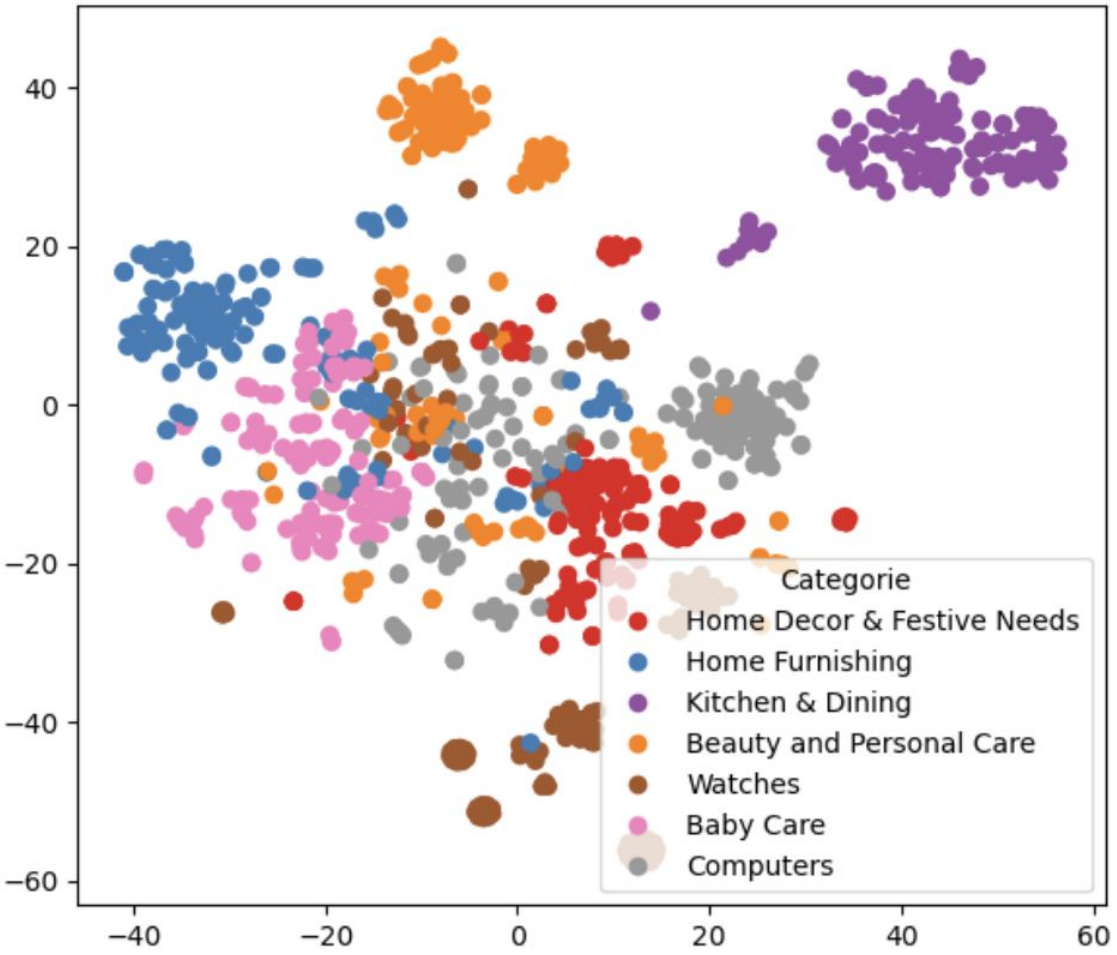
Réalisation d'un bag of words

2 méthodes utilisées :

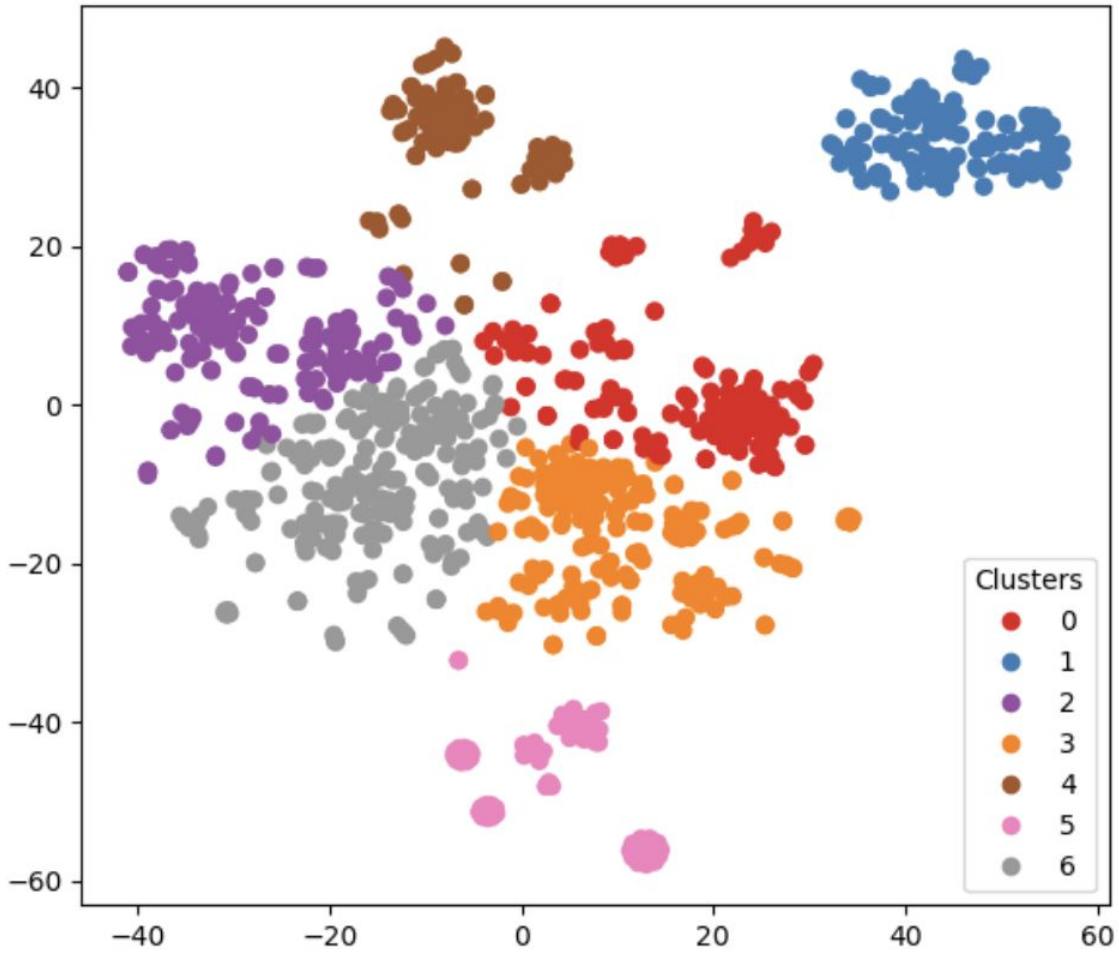
- CountVectorizer
- Tf-idf

CountVectorizer

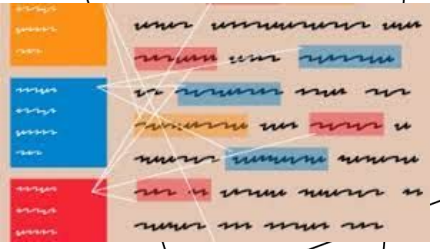
T-SNE des catégories réelles des produits



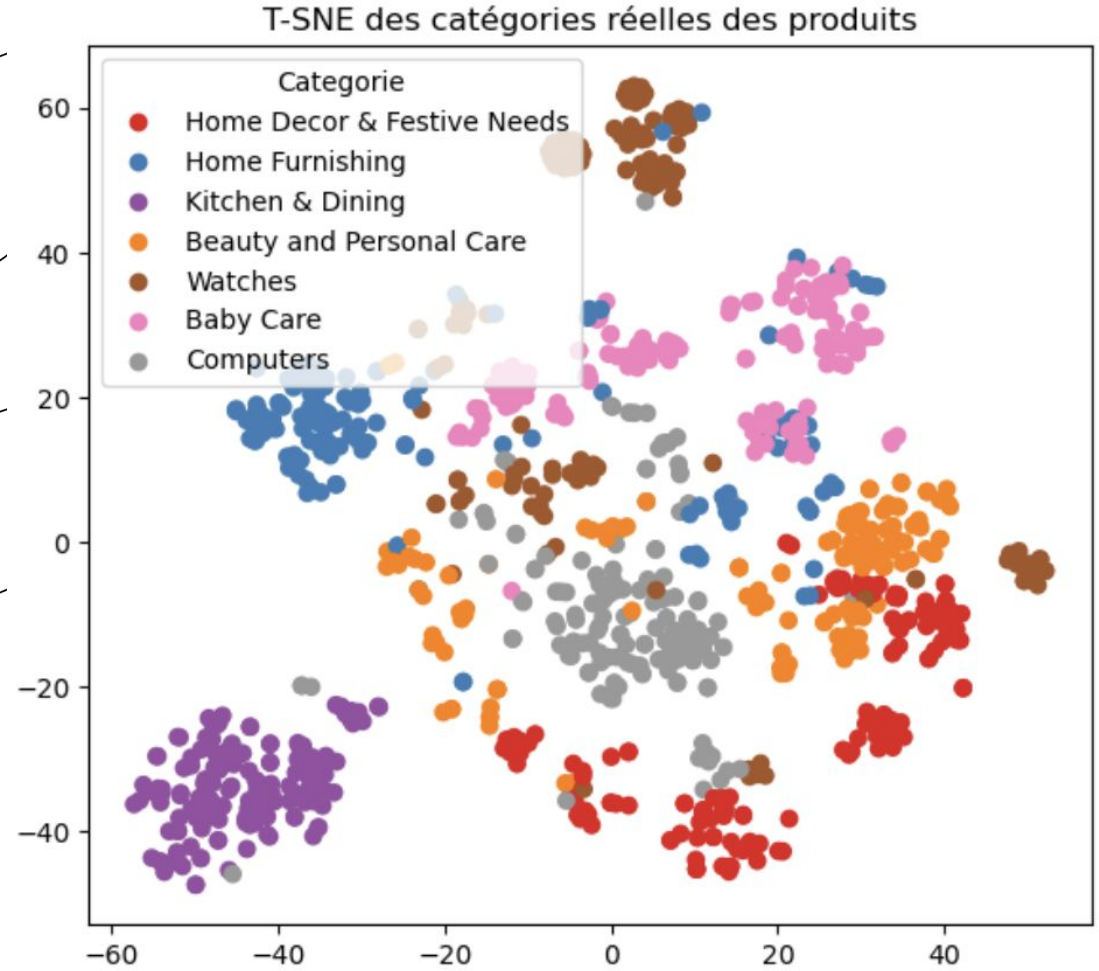
T-SNE des clustering des produits

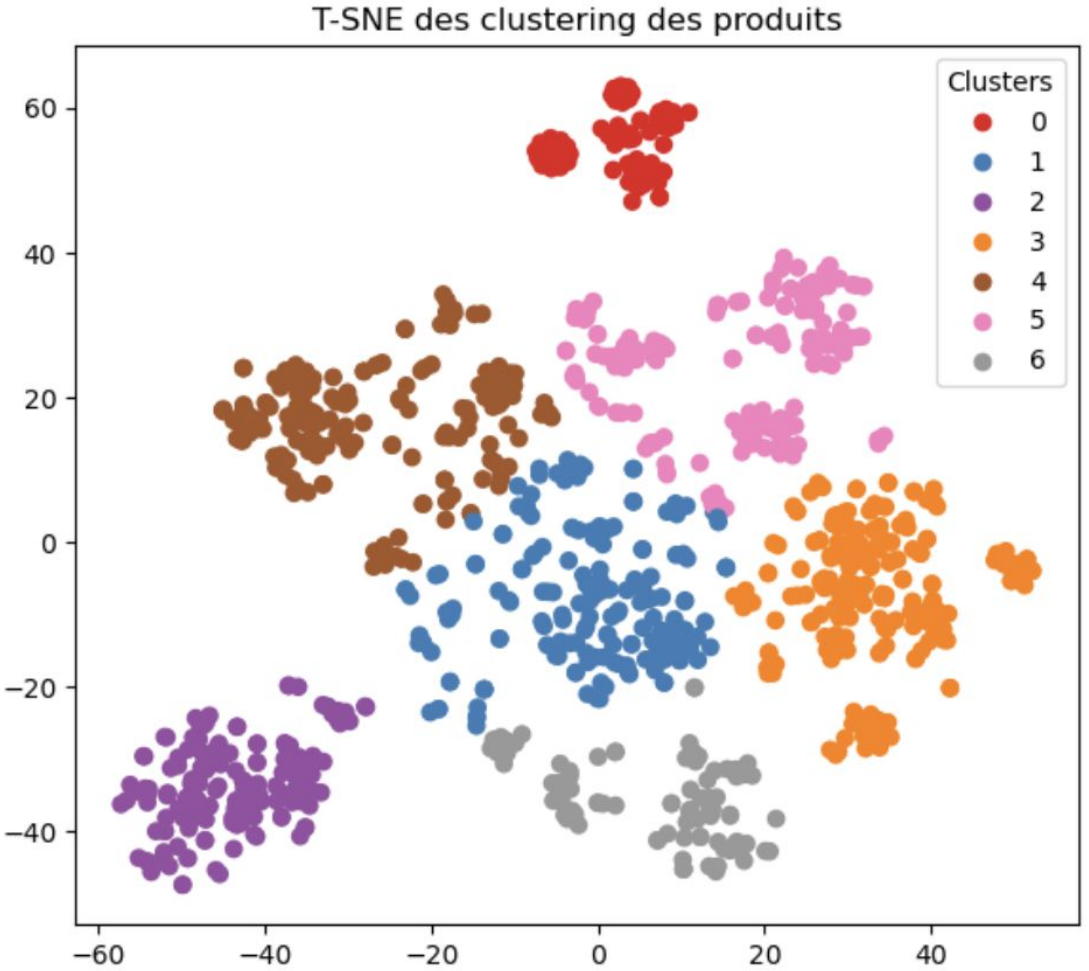


ARI Score 0,41



Tf-idf
(term frequency-inverse document frequency)

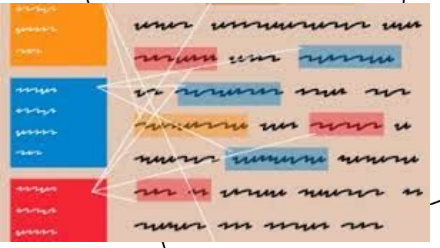




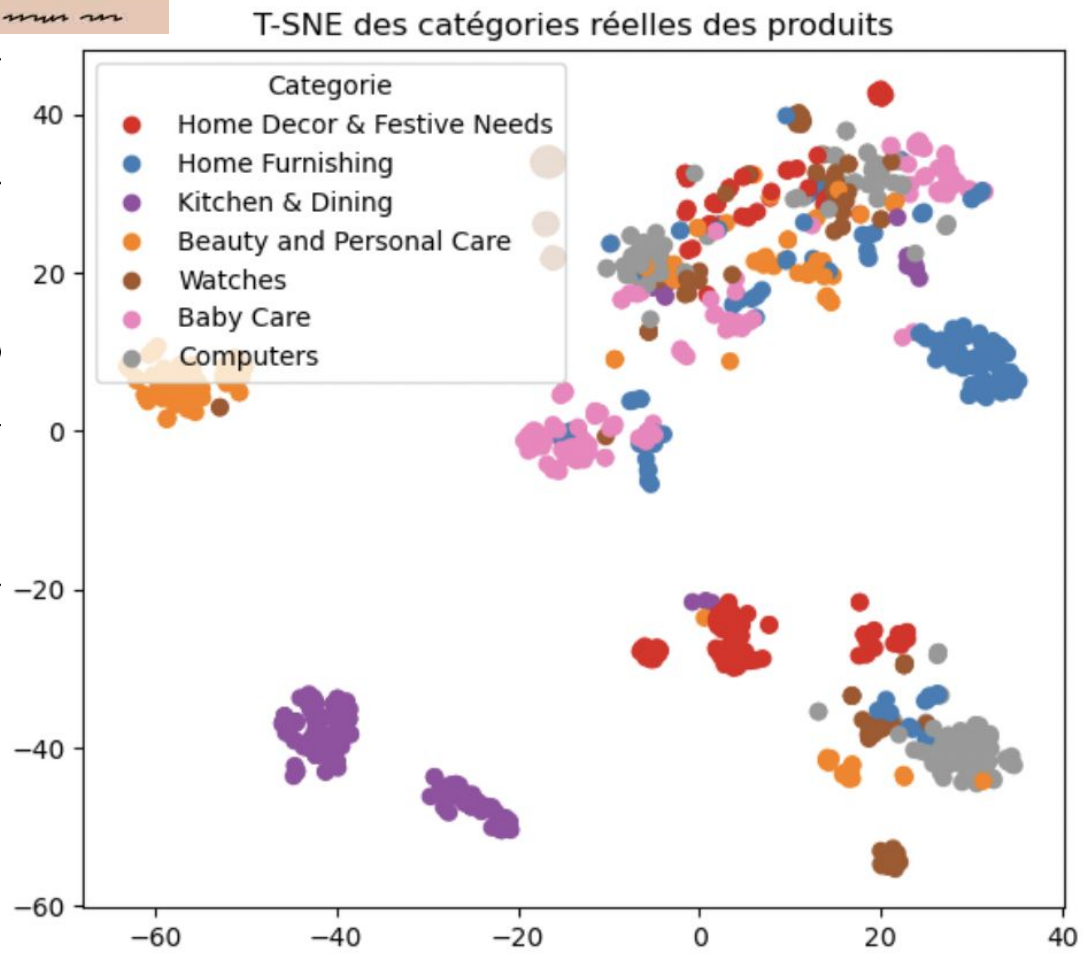
ARI Score 0,46

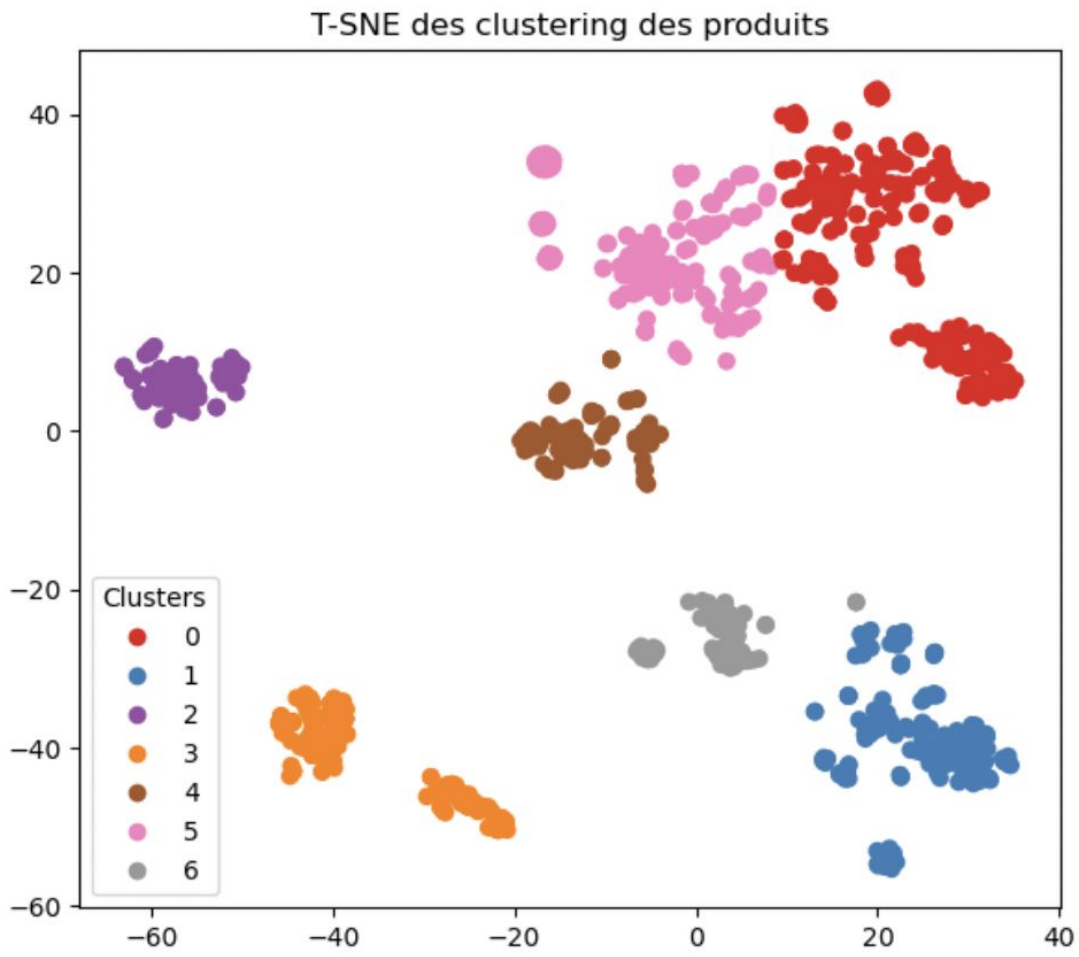
Analyse des données textuelles - word embedding

01	Word2vec	• Basé sur les Google news • Vecteur statique • Réseau de neurone avec une couche caché • Vecteur retenu représente les poids de la couche cachée
02	USE (Universal Sentence Encoder)	• Basé sur Wikipédia • Réseau de neurone basé sur l'architecture Transformers • Vecteur dynamique (se base sur les phrases)
03	BERT (Bidirectional Encoder Representations from Transformers	• Basé sur Wikipédia • Réseau de neurone basé sur l'architecture Transformers • Vecteur dynamique (se base sur les phrases) • Modèle bidirectionnel (mots avant et après la cible)

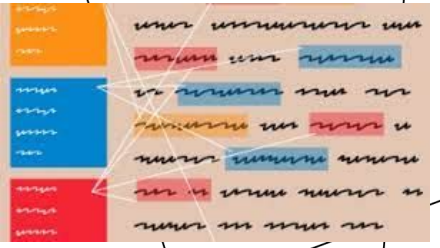


Word2vec

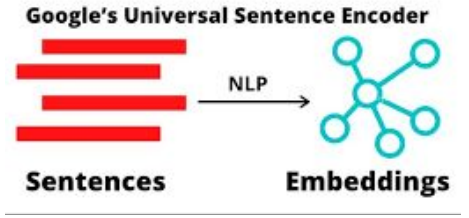


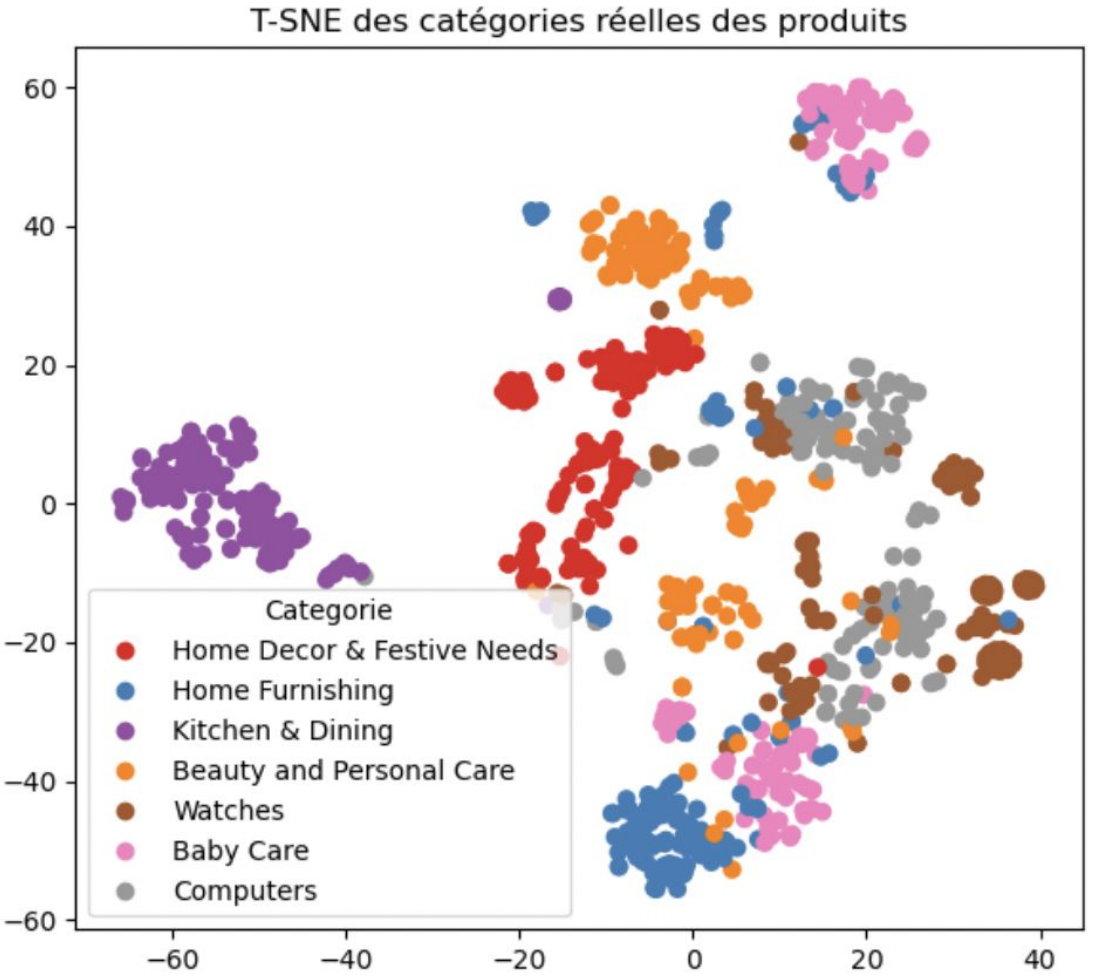


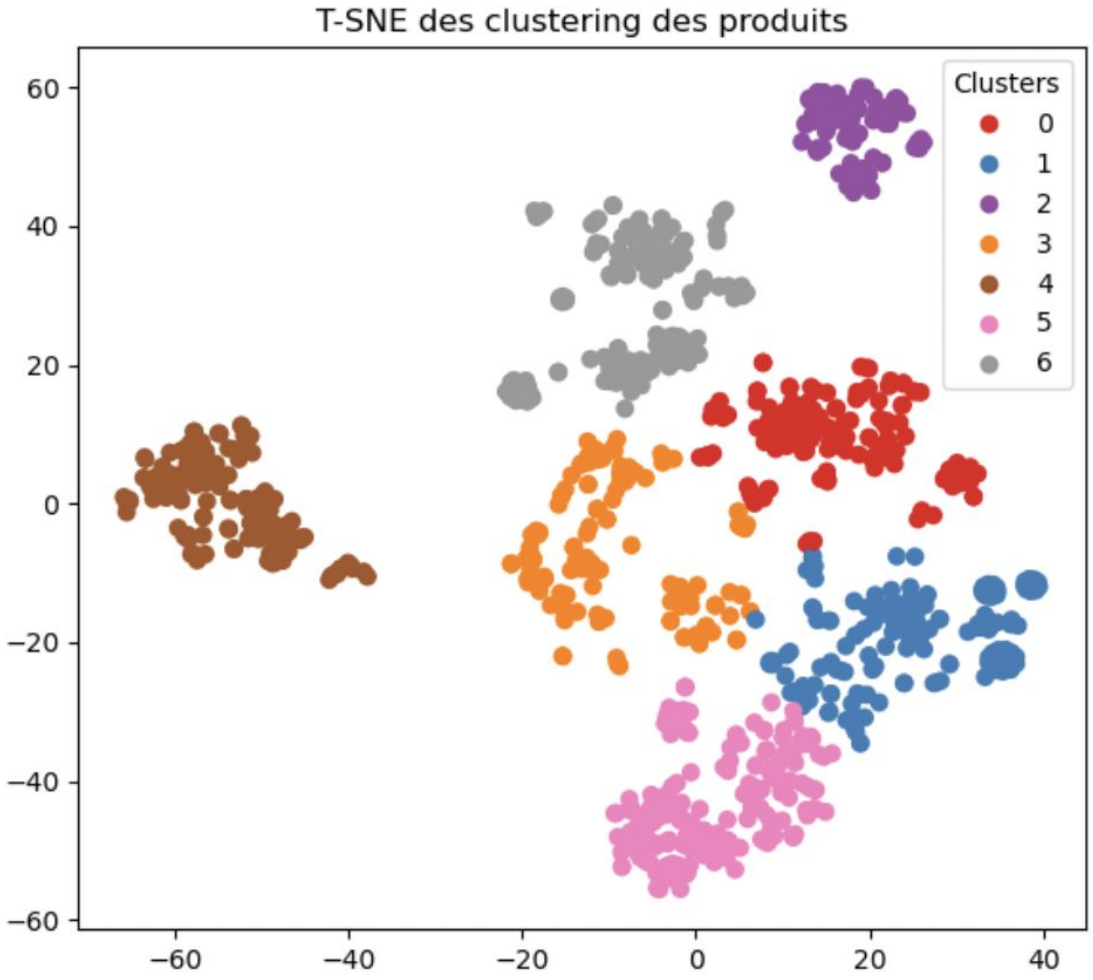
ARI Score 0,27



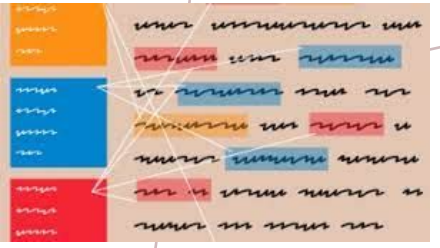
USE
(Universal Sentence Encoder)

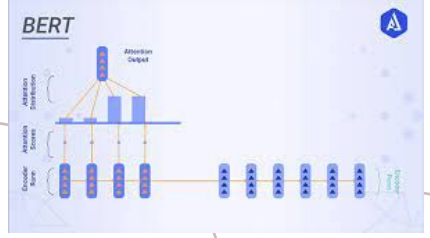




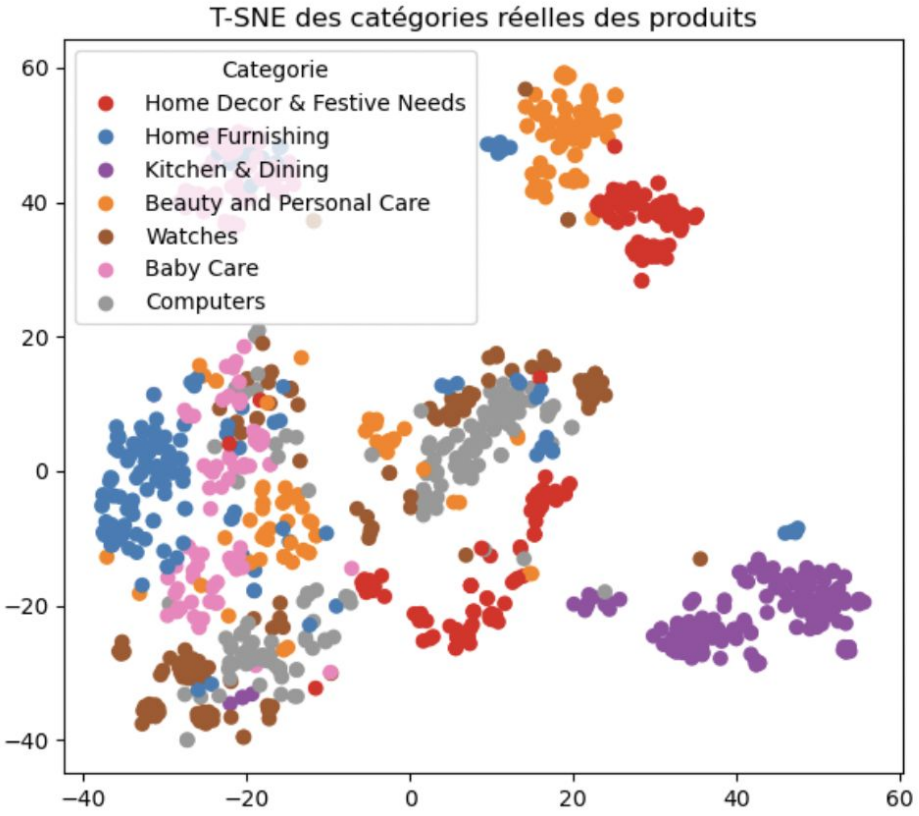


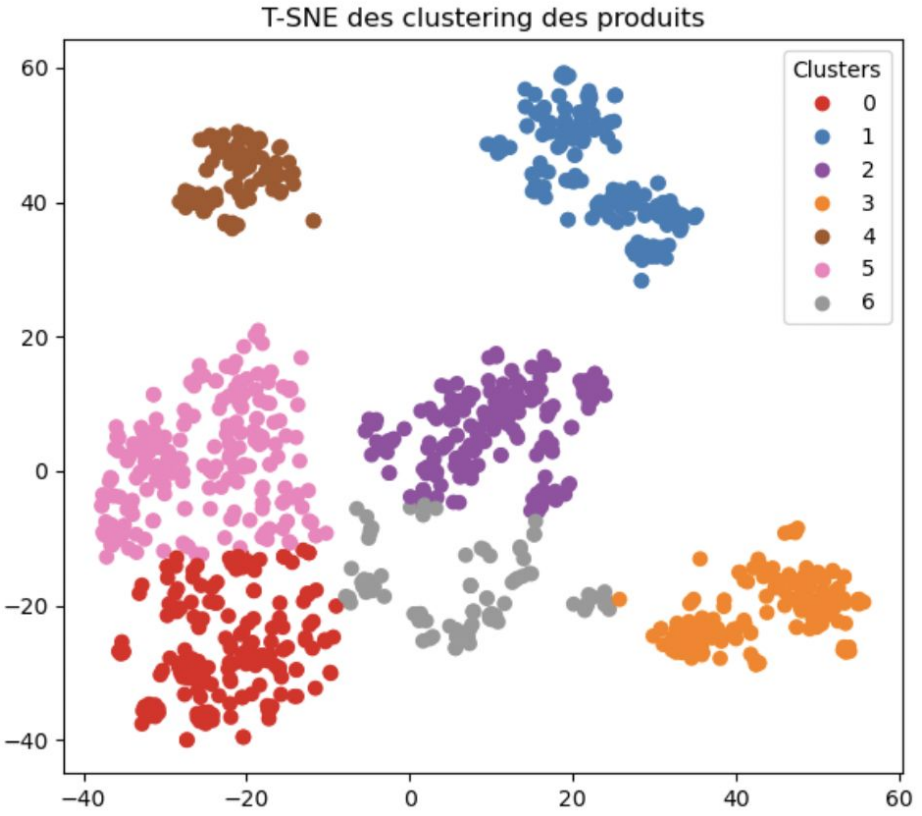
ARI Score 0,42





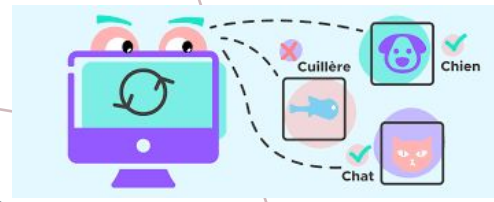
BERT
(Bidirectional Encoder Representations from Transformers)





ARI Score 0,32

Analyse des données images



Méthodologie

1 Extraction des features

Transformation des images par des vecteurs. Nous utiliserons :
- SIFT et application d'un bag of visual words
- Réseau de neurones pré-entrainé (CNN)

2 Réduction de dimensions 1

Réduction de ces features extraites via ACP/PCA en gardant 99% de la variance

3 Réduction de dimensions 2

Réduction des composantes principales de l'ACP/PCA à 2 dimensions via un t-SNE

4 Clustering

Clustering via KMeans (k = 7 catégories) sur la réduction du t-SNE

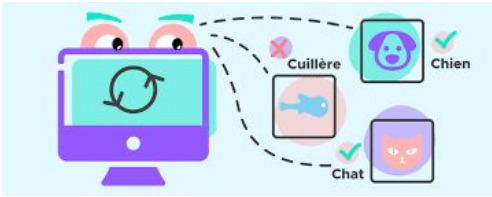
5 Analyse de similarité

Comparaison entre valeur prédites par KMeans et les catégories réelles via ARI (Adjusted Rand Index)

Analyse des données textuelles - Bag of words

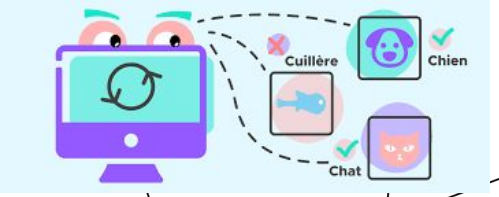
Générateurs des descripteurs SIFT

- Scale-invariant feature transform (SIFT) : algorithme permettant de définir les points d'intérêts d'une image (descripteurs)
- Ces points correspondent à des bords ou coins d'une image : zones autour desquelles ont observe de fortes variations d'intensité ou de couleur des pixels, qui indiquent donc la jonction entre des objets différents sur l'image
- Les descripteurs constituent des vecteurs qui décrivent le voisinage de la feature à laquelle ils sont associés
- Les descripteurs du SIFT ont l'avantage d'être invariant ts par rotation, par changement d'échelle et par exposition

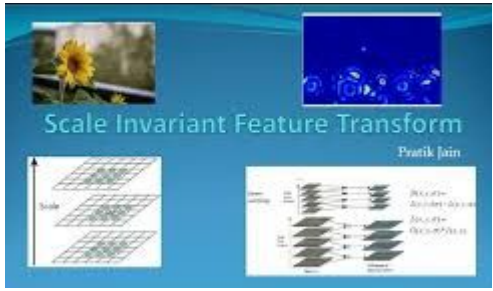


Réalisation d'un bag of visual words

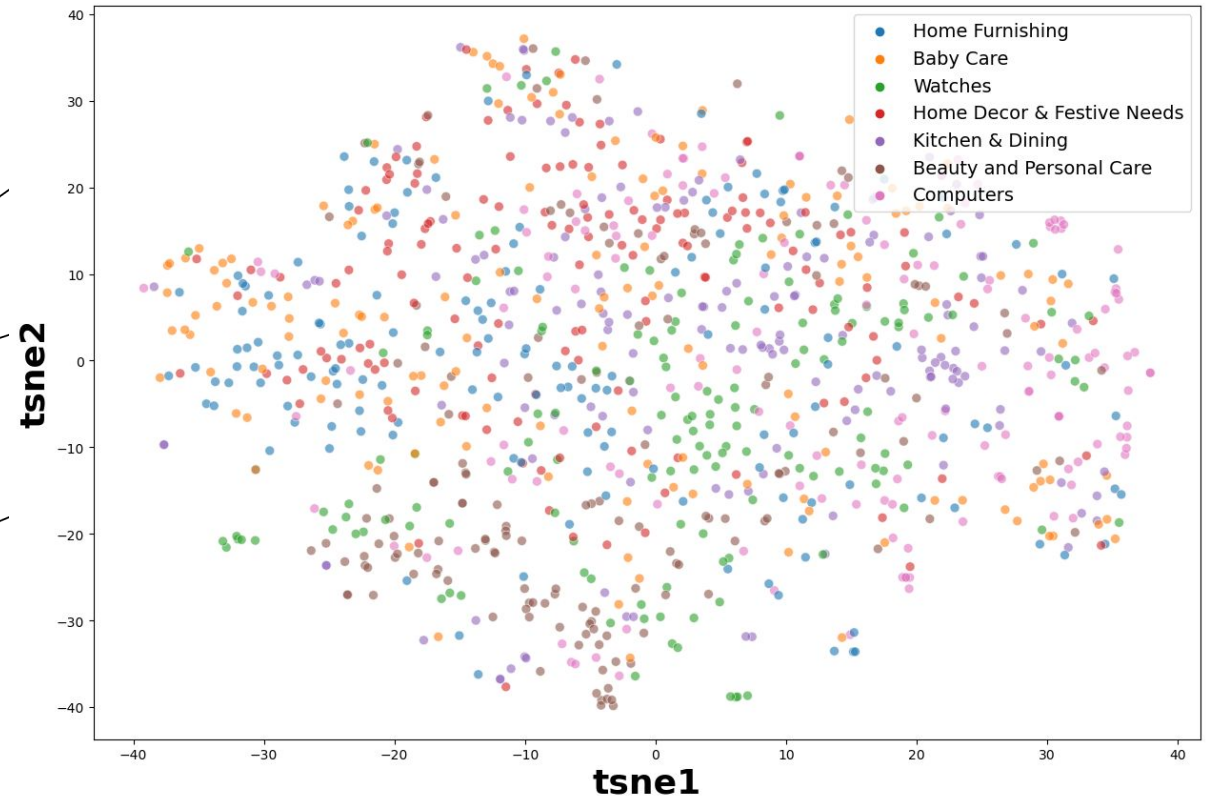
- La taille du vecteur de chacun des descripteur est identique, en revanche le nombre de descripteurs varie pour chacune des images, il n'est donc pas possible d'utiliser directement les descripteurs comme features pour une classification
- Afin de pallier cela, nous avons appliqué un « bag of visual words »
  - Clustering des descripteurs via un KMeans (k = racine carrée du nombre total de descripteurs)
  - Pour chaque image nous allons déterminer le nombre de descripteurs par cluster: chaque image disposera donc d'un vecteur de taille k avec pour valeurs le nombre d'occurrences pour chacun des clusters
- Ces vecteurs seront nos features finales auxquelles nous appliqueront les étapes décrites précédemment afin d'obtenir un ARI



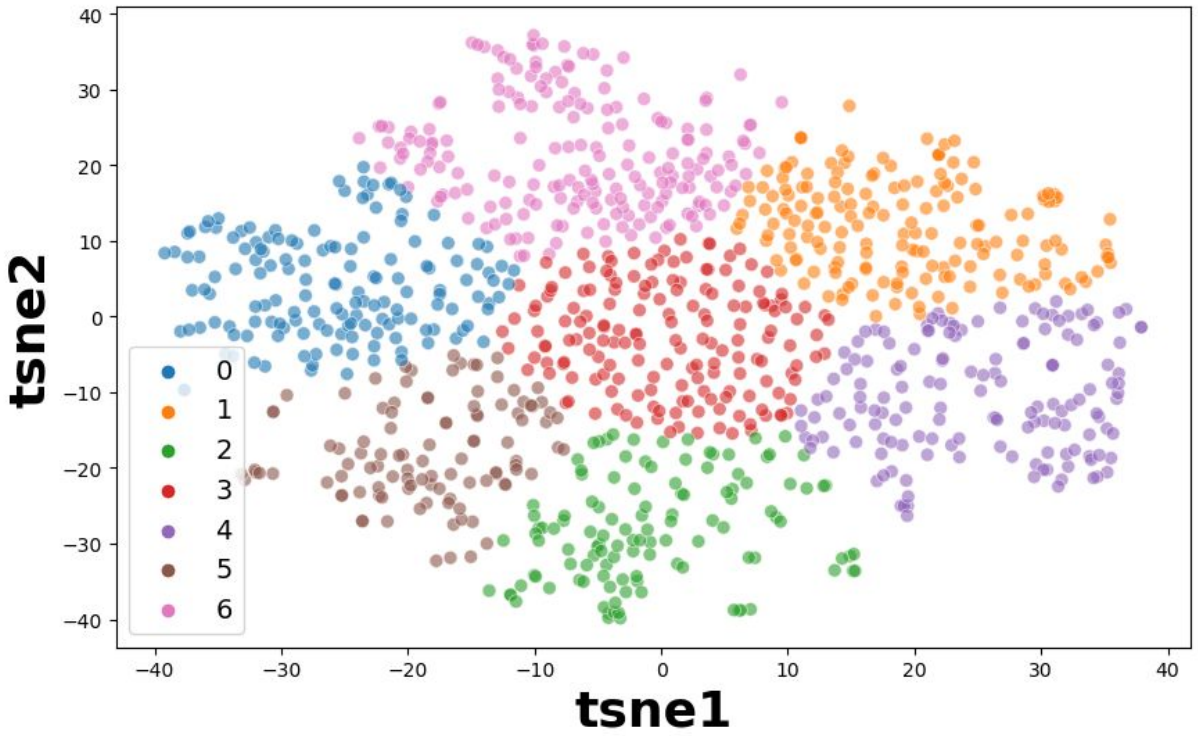
SIFT
(Scale-invariant feature transform)



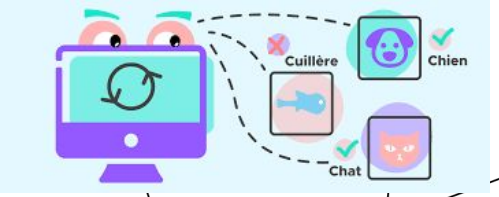
T-SNE des catégories réelles



T-SNE avec clusters

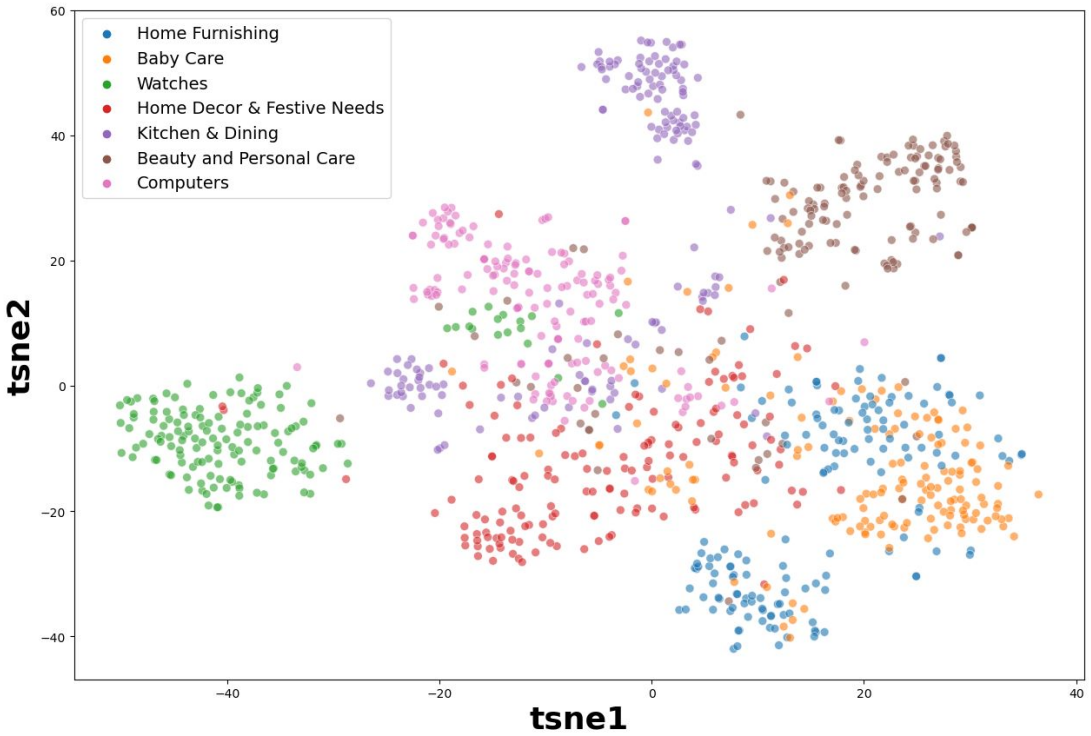


ARI Score 0,05

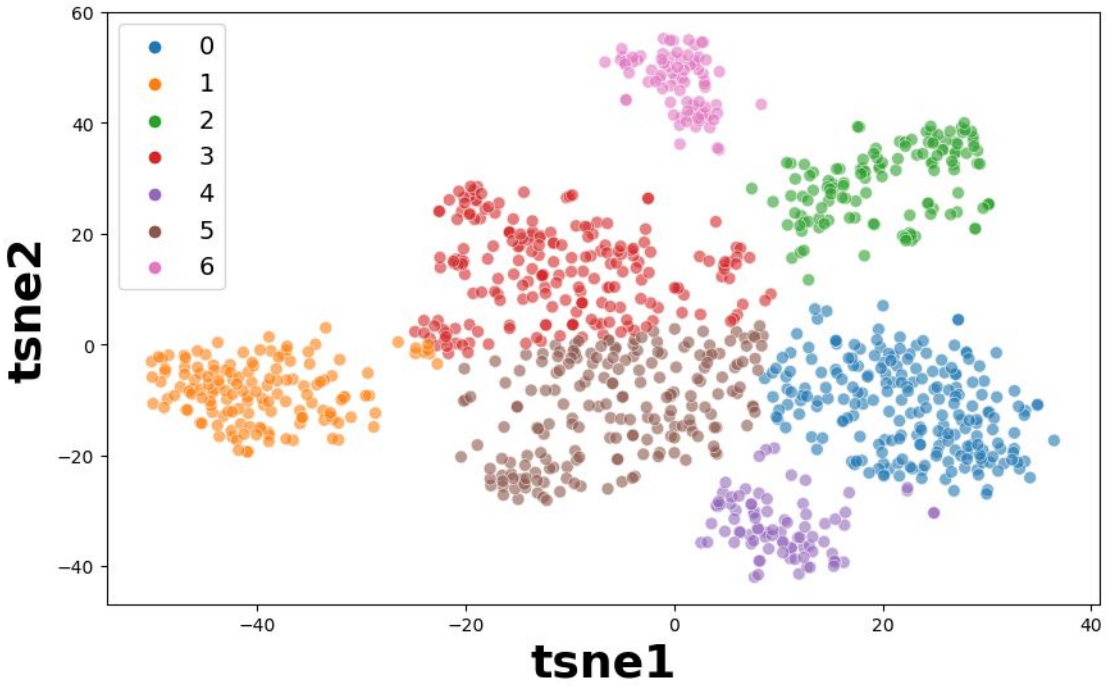


CNN
(Convolutional Neural Network)- VGG16

T-SNE avec les classes réelles



T-SNE avec clusters



ARI Score 0,47

Classement ARI des différentes méthodes

Type de méthode	Structure	ARI score
CountVectorizer	Bag of words	0,41
Tf-idf	Bag of words	0,46
Word2vec	Word embedding statique	0,27
USE	Sentence embedding	0,42
BERT	Word embedding dynamique	0,32
SIFT	Descripteurs	0,05
CNN - VGG16	Réseau de neurones convolutif CNN	0,47

Tf-idf

CNN avec VGG16

CONCLUSION

- L'étude de faisabilité du moteur de classification donne des résultats encourageants

- Nous obtenons des ARI supérieurs à 0,45 aussi bien pour l'analyse du texte que d'images

- Nous pouvons considérer la mise en œuvre d'un moteur de classification automatique avec des algorithmes supervisés entraînés sur notre jeu de données et optimisés

- Il est probable que même après entraînement et optimisation, nous n'obtiendrons pas de classifications parfaites et des erreurs de classifications seront présentes

- L'analyse a été réalisée sur la base des catégories larges de produits. L'utilisation des sous-catégories n'est pas réaliste.

AXES D'AMÉLIORATIONS DE L'ÉTUDE

1. Coupler l'analyse NLP et CV
2. Faire du fine-tuning
3. Discuter avec les métiers pour reconfigurer les catégories/sous catégories en utilisant la Classification non supervisé via Kmeans puis supervisé
4. Utiliser la LDA (Latent Dirichlet Allocation) pour la réduction de dimension
5. Utiliser SpaCy, Glove ou FastText pour les données textuelles
6. Utiliser ORB / SURF, VGG19 - RESNET50 pour les données images
7. Utiliser les algorithmes GAN (Generative adversarial networks)
8. Augmenter la taille du dataset d'apprentissage

MERCI POUR VOTRE ÉCOUTE ET ATTENTION