Projet 6 :
Classifier automatiquement des biens de consommation

Ilham NOUMIR | Parcours Data Science | Date : 24/12/2021

Sommaire

1. Présentation de la problématique :

Description de l’organisme :
La place de marché est une plateforme e-commerce proposant des produits à la vente

Contexte :
Attribution manuelle de la catégorie d’un article par les vendeurs

Objectifs :

- Automatisation de l’attribution des catégories aux produits
- Réalisation d'une première étude de faisabilité d'un moteur de classification d'articles

1. Présentation de la problématique :

Tâches à effectuer :

- **Analyse du jeu de données**
- **Réalisation d'un prétraitement** des images et des descriptions des produits
- **Réduction de dimension.**
- **Clustering.**

Description du jeu de données :

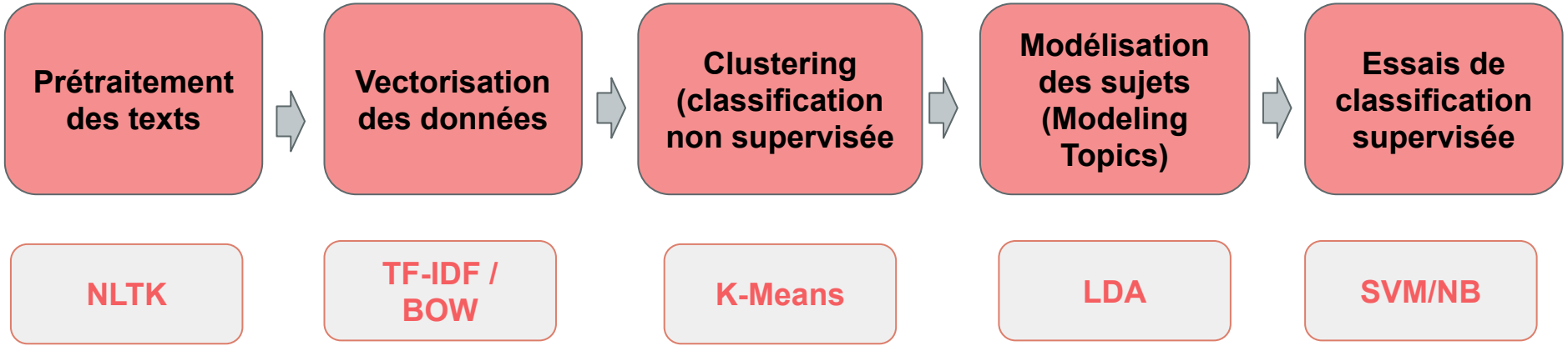
- ➔ 1050 observations pour chacune d'elle 15 informations

- ➔ Parmi les 15 informations , on trouve par exemple : l'identifiant unique du produit , le nom du produit, spécifications des produits…

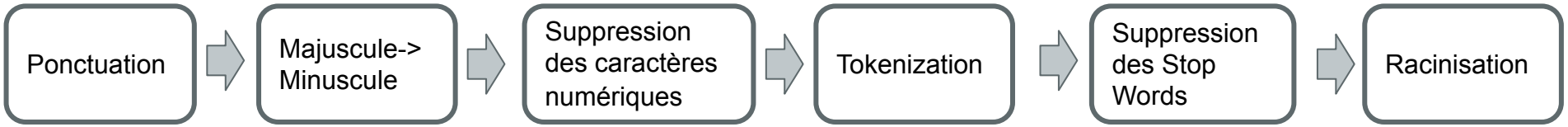
- ➔ On va s'intéresser à trois colonnes en particulier :
  - ◆ La description des produits
  - ◆ La catégorie des produits
  - ◆ L'image relative à chaque produit

Données Textuelles

Processus de traitement des données textuelles :



Prétraitement des données textuelles :



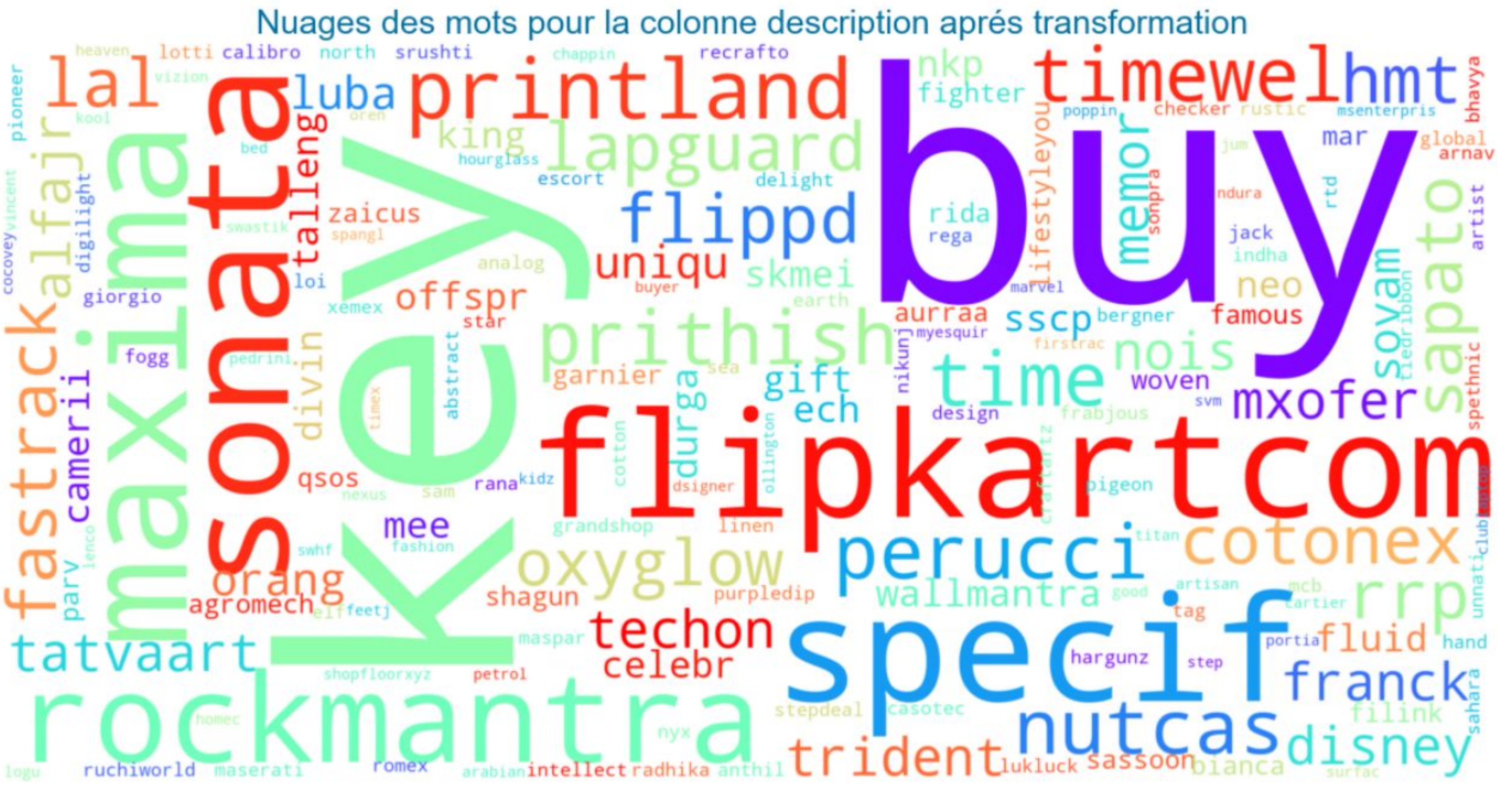
Description(Avant transformation)

```
Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance
ack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from
with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and
```

Description(Après transformation)

```
['key', 'featur', 'eleg', 'polyest', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain',
r', 'curtain', 'height', 'pack', 'price', 'curtain', 'enhanc', 'interiorsthi', 'curtain',
h', 'metal', 'ringit', 'room', 'environ', 'romant', 'lovingthi', 'curtain', 'ant', 'wrinkl
```

Les mots les plus fréquents dans le corpus :

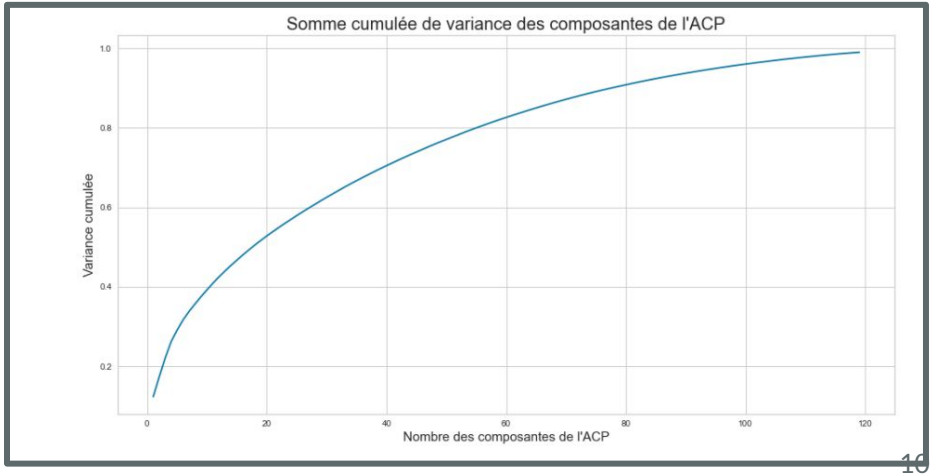


Vectorisation des données :

	abstract	add	addit	analog	babi	batteri	beauti	black	bleach	blue	...	warranti	wash	watch	water	wear	weight
0	0.433021	0.0	0.0	0.0	0.0	0.0	0.097107	0.0	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000
1	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.346331	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000
2	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.101620	0.000000	...	0.000000	0.250969	0.0	0.078353	0.0	0.081388
3	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.101890	0.000000	...	0.000000	0.083879	0.0	0.000000	0.0	0.000000
4	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.087768	0.000000	...	0.077186	0.000000	0.0	0.000000	0.0	0.070294

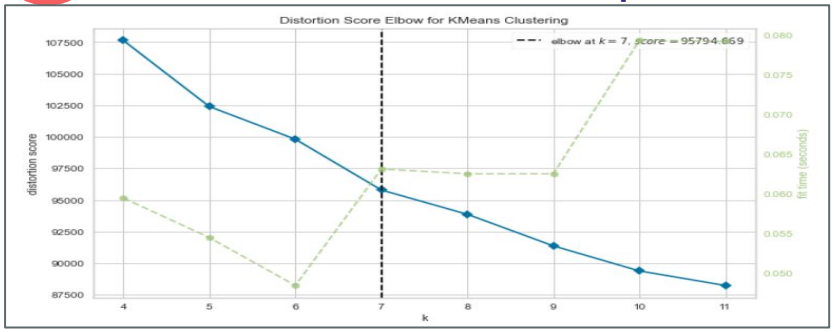
Réduction des dimensions :

```
pca = PCA(n_components= 0.99)
# Entraînement des données :
pca.fit(X_scaled)
```



Clustering K-Means :

1 Choix du nombre de cluster optimal

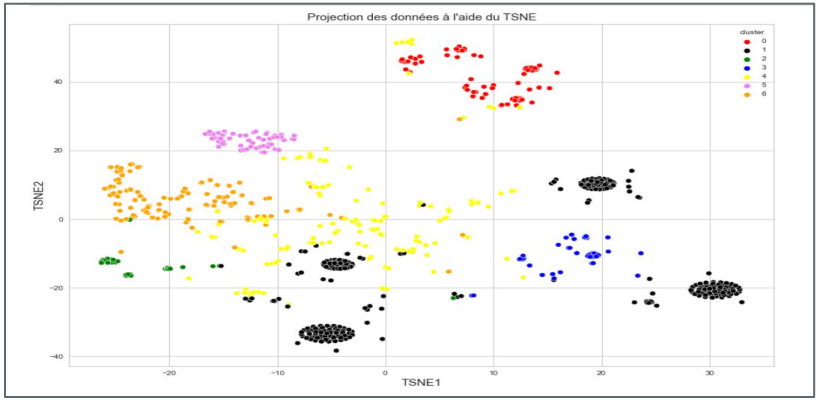


2 Entraînement du modèle avec un nombre de cluster de 7

```
temps1 = time.time()
my_clust = cluster.KMeans(n_clusters=7)
my_clust.fit(X_pca)

duration1 = time.time() - temps1
```

3 Projection des données à l'aide du T-SNE

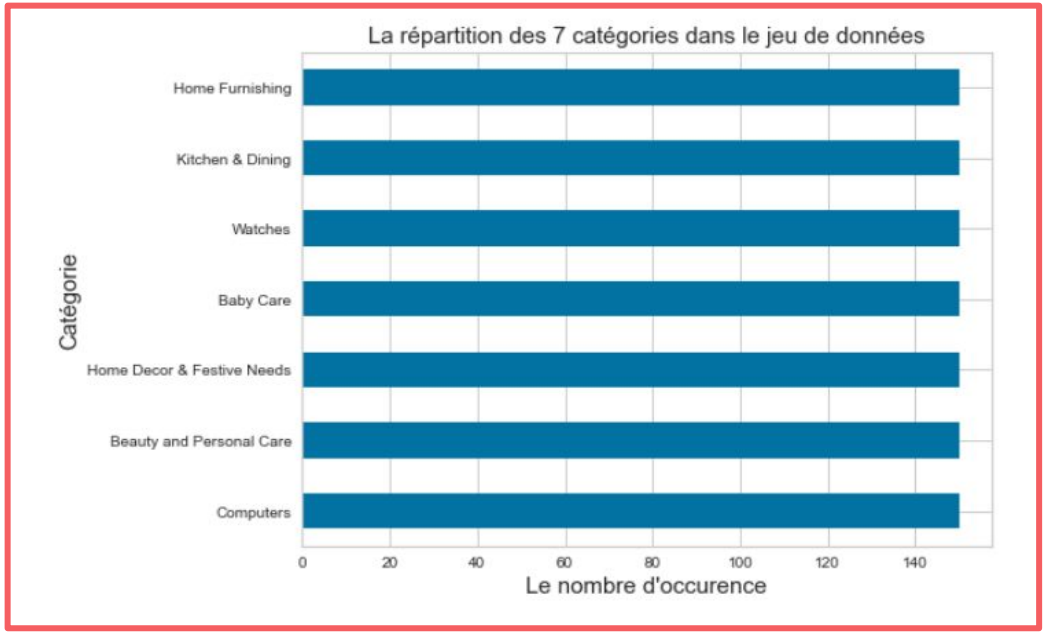


4 Calcul des métriques

Type de données	ARI	Silhouette	Durée
Textuelles	0.25	0.17	0.074

Essais avec la classification supervisée :

Création d'une nouvelle colonne **Cat1** à partir de la colonne "**product_category_tree**" avec 7 catégories :



Support Vector Machine

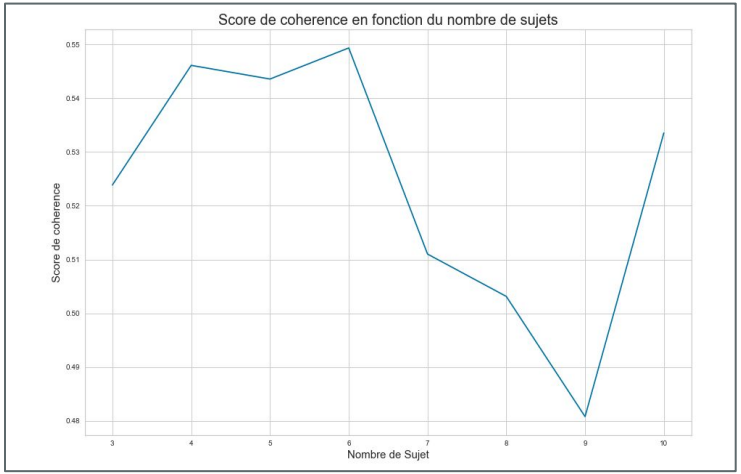
95,05 % d'accuracy sur le jeu de test

Naive Bayes

90,11% d'accuracy sur le jeu de test

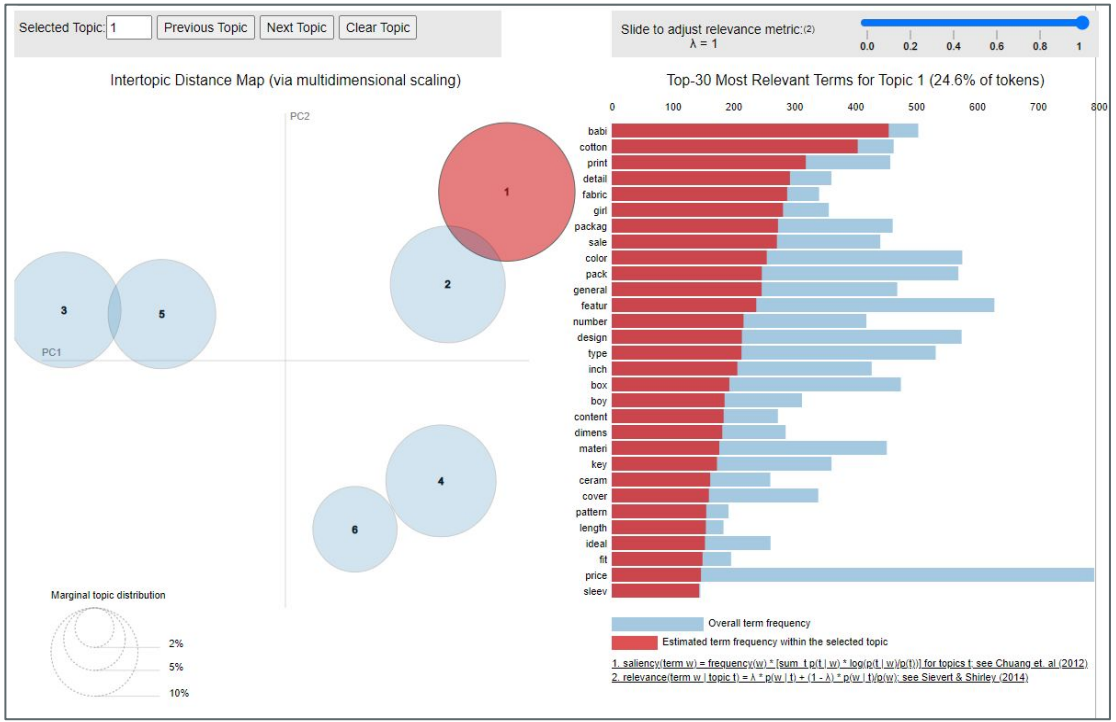
Modélisation des sujets: LDA (Latent Dirichlet Allocation)

Score de cohérence pour le choix optimal du nombre de topics



Nombre optimal égale à 6

Exemple des mots présents par topics



Essais avec d'autres méthodes de traitement de textes :

Méthode de Racinisation

Lemmatisation vs Stemming

Métriques

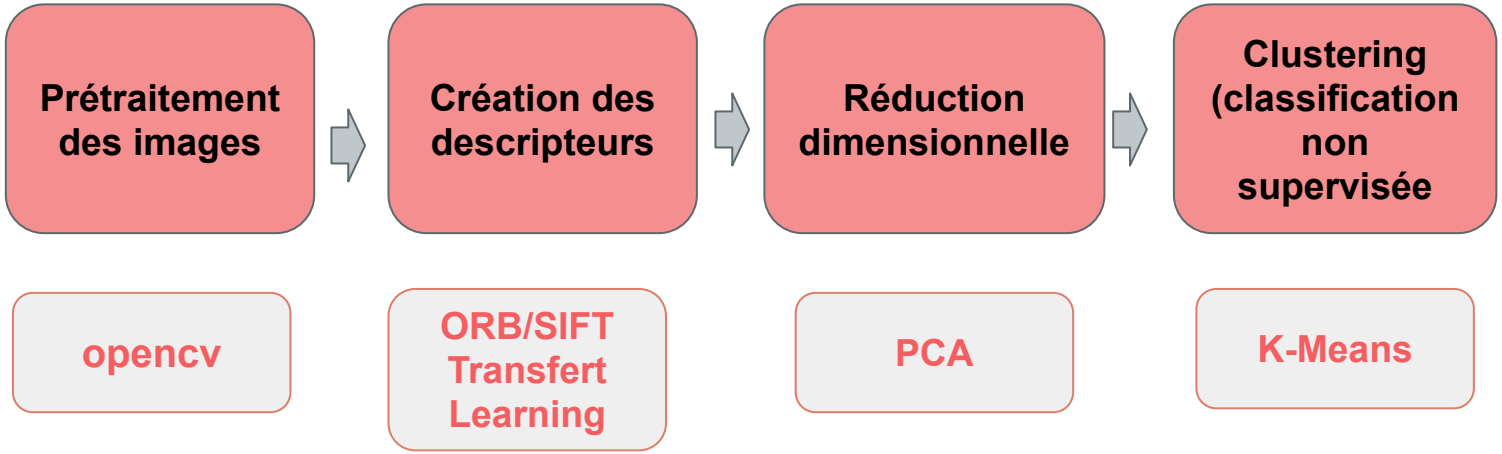
Type de traitement	ARI	Silhouette	durée
Racinisation avec Stemming	0.256816	0.176610	0.074425
Racinisation avec Lemmatisation	0.173225	0.152421	0.067436

Choix final

Stemming

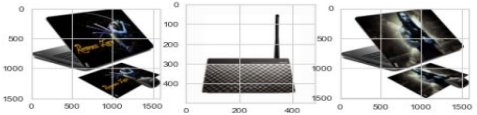
Données Visuelles

Processus de traitement des images :

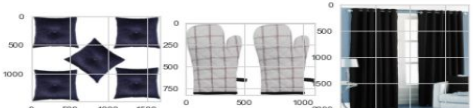


Images par catégorie :

Computers :

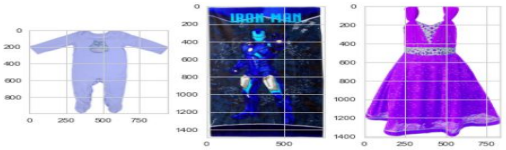


Home_funshing :

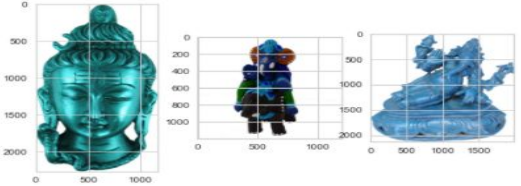


Kitchen :

Baby_ Care:



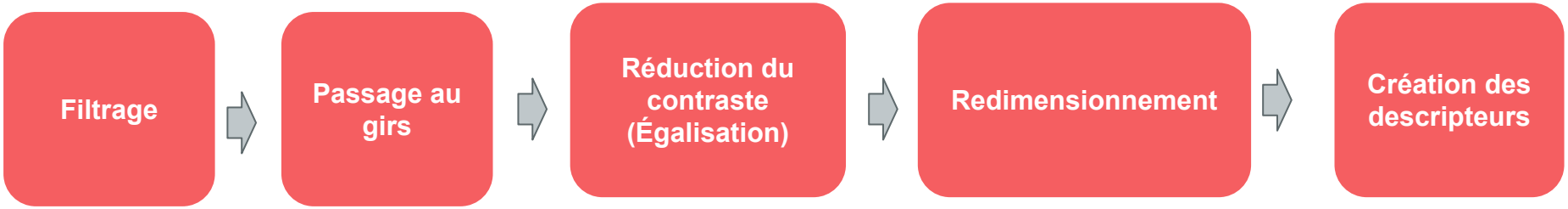
Home_Decor:



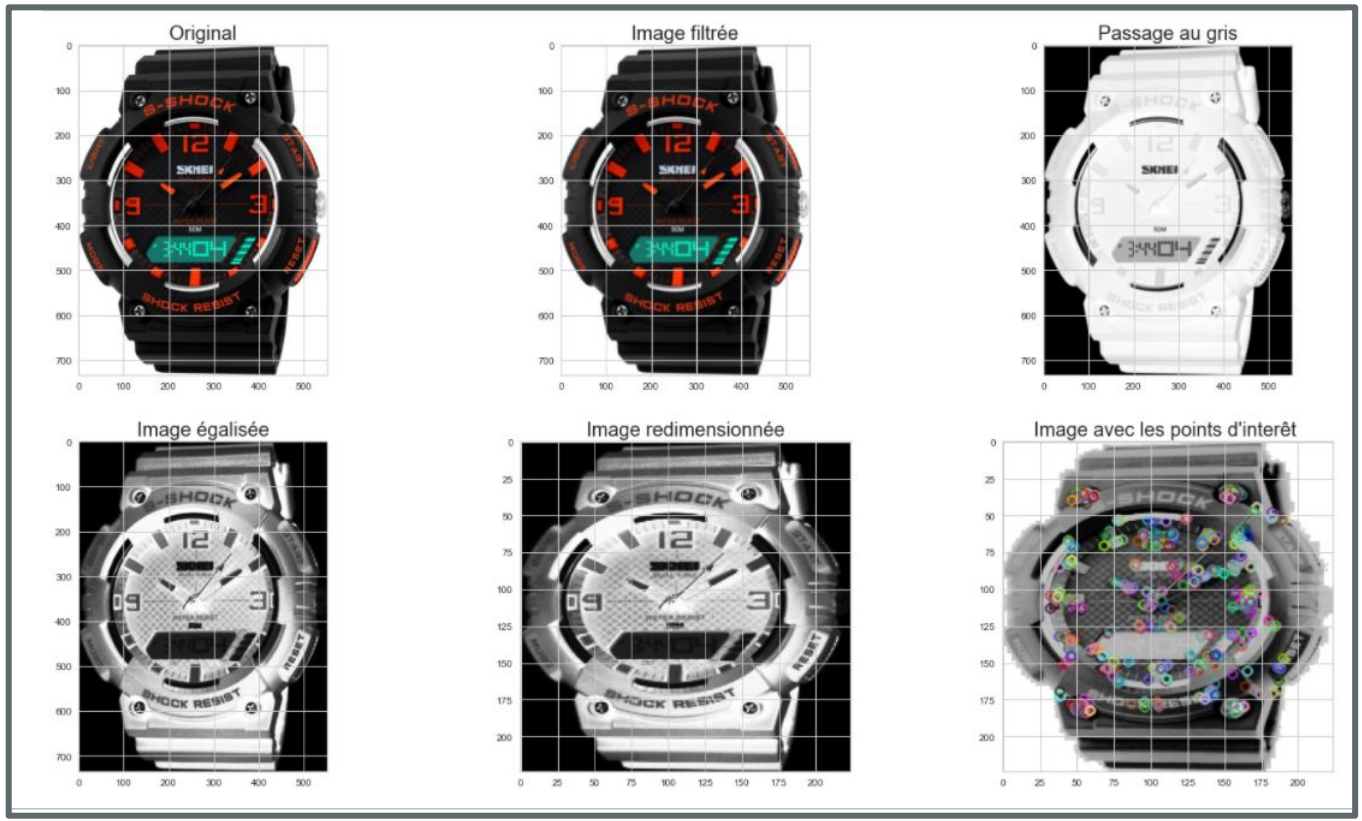
Personal_care:

Watches :

Processus de traitement des images :



Processus de traitement des images (Exemple)



Traitement des images :

1

Création des clusters de descripteurs :

A partir des descripteurs générés (1537554, 32)

Chaque descripteurs est un vecteur de longueur 32

Le nombre de cluster K est la racine carré des nombres de descripteurs

2

Création des features des images :

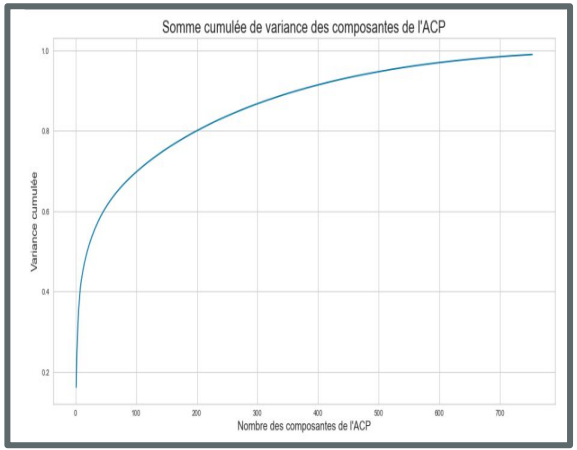
Pour chaque image on regroupe les descripteurs par cluster

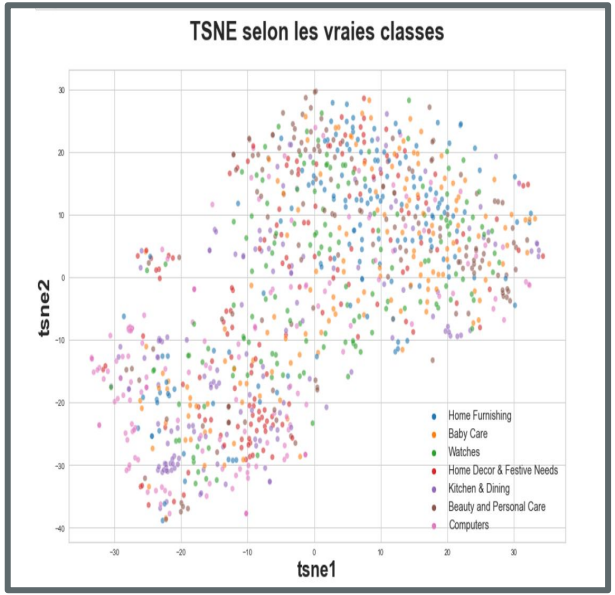
3

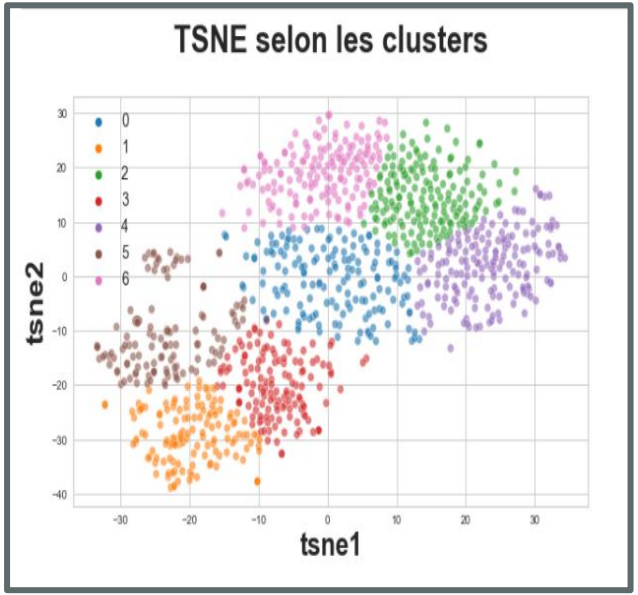
Réduction dimensionnelle :

Utilisation de la PCA et visualisation avec T-SNE

Images : Résultat du clustering



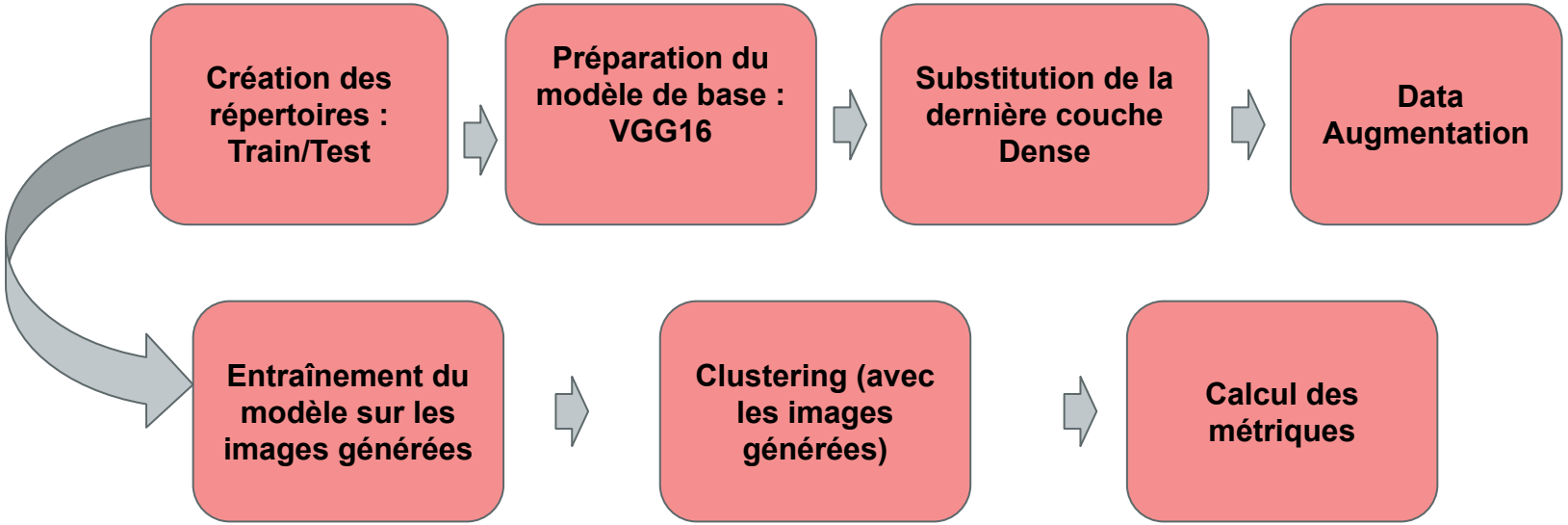




Métriques

Type de données	ARI	Silhouette	Durée
Visuelles	**0.03**	0.36	0.08

Transfert learning : Processus de mise en oeuvre



Images : Résultat du clustering

Type de données	ARI	Silhouette	Durée
Visuelles avec TL	**0.0025**	0.89	0.07

Essais avec d'autres méthodes de traitement des images:

Méthode d'extraction des features

ORB vs SIFT vs TL

Métriques

Type de traitement	ARI	Silhouette	durée
Image avec ORB	0.036384	0.368908	1.554144e+02
Image avec SIFT	0.063851	0.343059	4.541167e+02
Image avec transfert learning	0.002157	0.890340	1.640121e+09

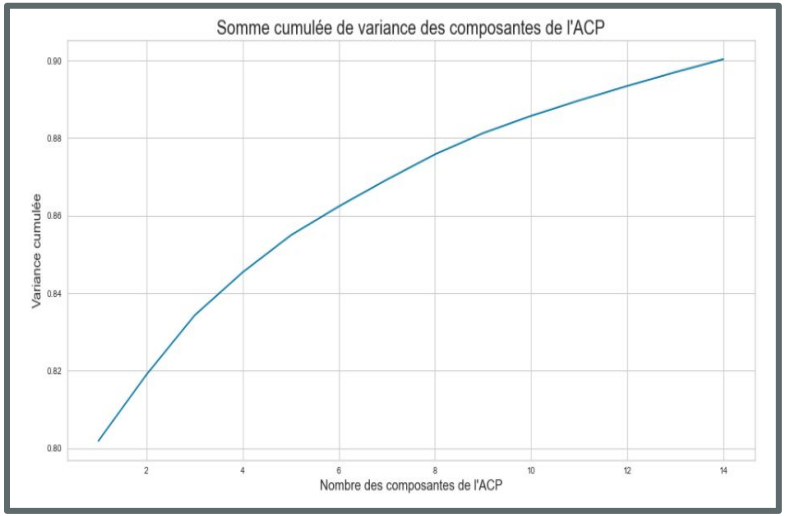
Choix final

ORB

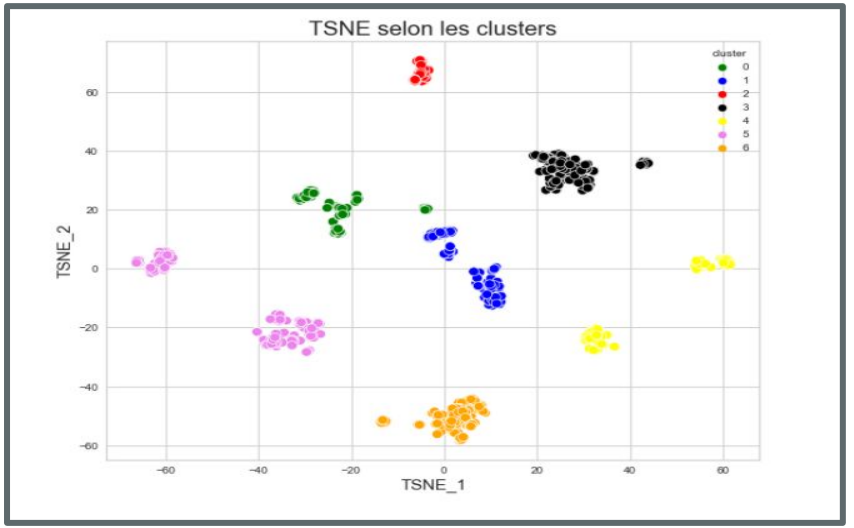
Données
Textuelles + Visuelles

Regroupement des images et des texts :

Réduction dimensionnelle



Projection avec T-SNE

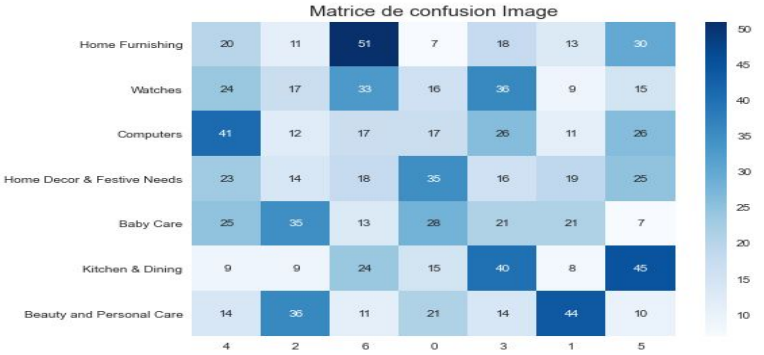


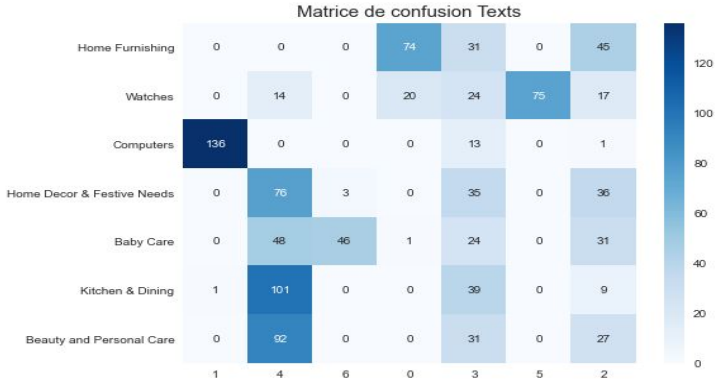
Métriques

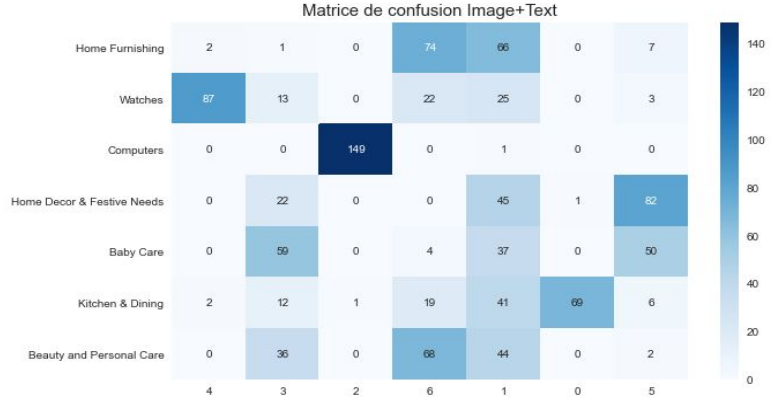
Type de données	ARI	Silhouette	Durée
Visuelles+Textuelles	0.32	0.60	0.10

Comparaison des modèles selon les types de données choisis:

Type de données	ARI	Silhouette	durée
Textuelles	0.256816	0.176610	0.074425
Visuelles	0.036384	0.368908	0.087871
Visuelles + Textuelles	0.321538	0.600863	0.103920







Conclusion :

- ➔ **Les données textuelles** : Le stemming qui donnent des meilleures résultats
- ➔ **Les données visuelles** : l'ORB qui donnent des résultats acceptables
- ➔ **La combinaison des données textuelles et visuelles** améliorent les performances de façon considérable mais insuffisante

Etude de faisabilité :

Le projet de moteur de classification automatique reste à voir car l'ARI dans la meilleure approche donne un score 0.32. On peut envisager :

- ➔ Apprentissage supervisé
- ➔ Elargissement de la catégorie des produits(Au lieu de 7 catégories)
- ➔ Augmentation de la qualité de la description (Utilisation des mots clés)

Perspectives :

- ➔ Entraînement d'un CNN et comparaison des résultats avec les autres approches
- ➔ Etude de la possibilité d'utiliser un transfert learning pour les données textuelles
- ➔ Utilisation de Word2vec pour le traitement de textes

Merci pour votre attention