Projet 6 :

Classifiez automatiquement des biens de consommation

Nom : TRABIS

Prénom : Mohamed

Intitulé de formation : Data Scientist

Mentor: Mr. Christian NOUMSI

Table des matières

Introduction

Introduction

- **Contexte** :

  l’entreprise "**Place de marché**” souhaite lancer une marketplace e-commerce.

  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 et des acheteurs la plus fluide possible, **il devient nécessaire d'automatiser cette tâche.**

  Linda, lead data scientist, demande d'étudier la faisabilité d'un **moteur de classification** des articles en différentes catégories, avec un niveau de précision suffisant.

Introduction

- **Mission** :

  la mission consiste à réaliser une première étude de faisabilité d'un moteur de classification d'articles basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article.

  Analyser le jeu de données en réalisant un prétraitement des images et des descriptions des produits, une réduction de dimension, puis un clustering. Les résultats du clustering seront présentés sous la forme d'une représentation en deux dimensions, qui 'illustrera le fait que les caractéristiques extraites permettent de regrouper des produits de même catégorie.

  La représentation graphique va aider à convaincre Linda que cette approche de modélisation permettra bien de regrouper des produits de même catégorie.

Évaluation et découverte des données

Évaluation et découverte des données

- Le jeu de données est composé :

  - D'un fichier CSV « ***flipkart_com-ecommerce_sample_1050.csv*** » : 1050 lignes et 15 colonnes.
  - D'un dossier « ***Images*** » : 1050 images des produits mentionnés dans le fichier CSV.

- Ci-dessous les informations descriptives de notre base de données :

```
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   datetime64[ns, UTC]
 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), datetime64[ns, UTC](1), float64(2), object(11)
memory usage: 116.0+ KB
```

Prétraitement des données textuelles

Prétraitement des données textuelles - Catégorie

- Extraire les catégories de la colonne « *product_category_tree* » :

Splitter la colonne «*product_category_tree* »

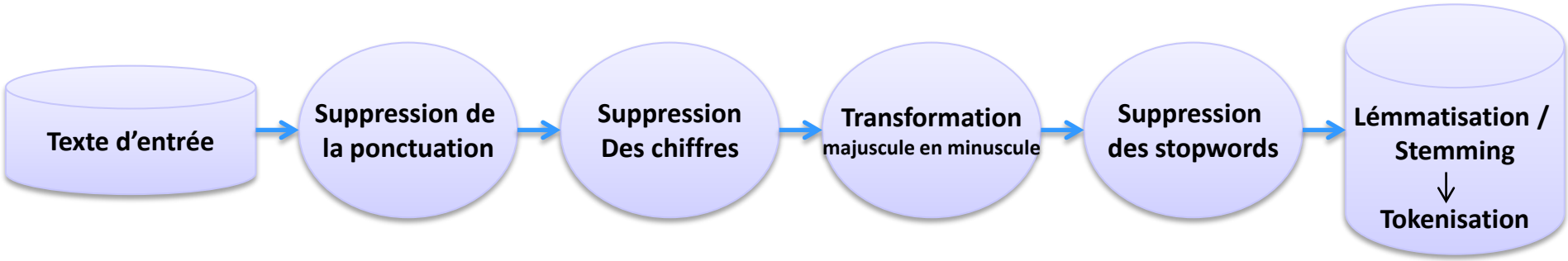
Extraire les catégories pricipales

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

La distribution des catégories

Prétraitement des données textuelles - Lémmatisation / Stemming

- Processus de nettoyage du texte de la colonne « *Description* » :



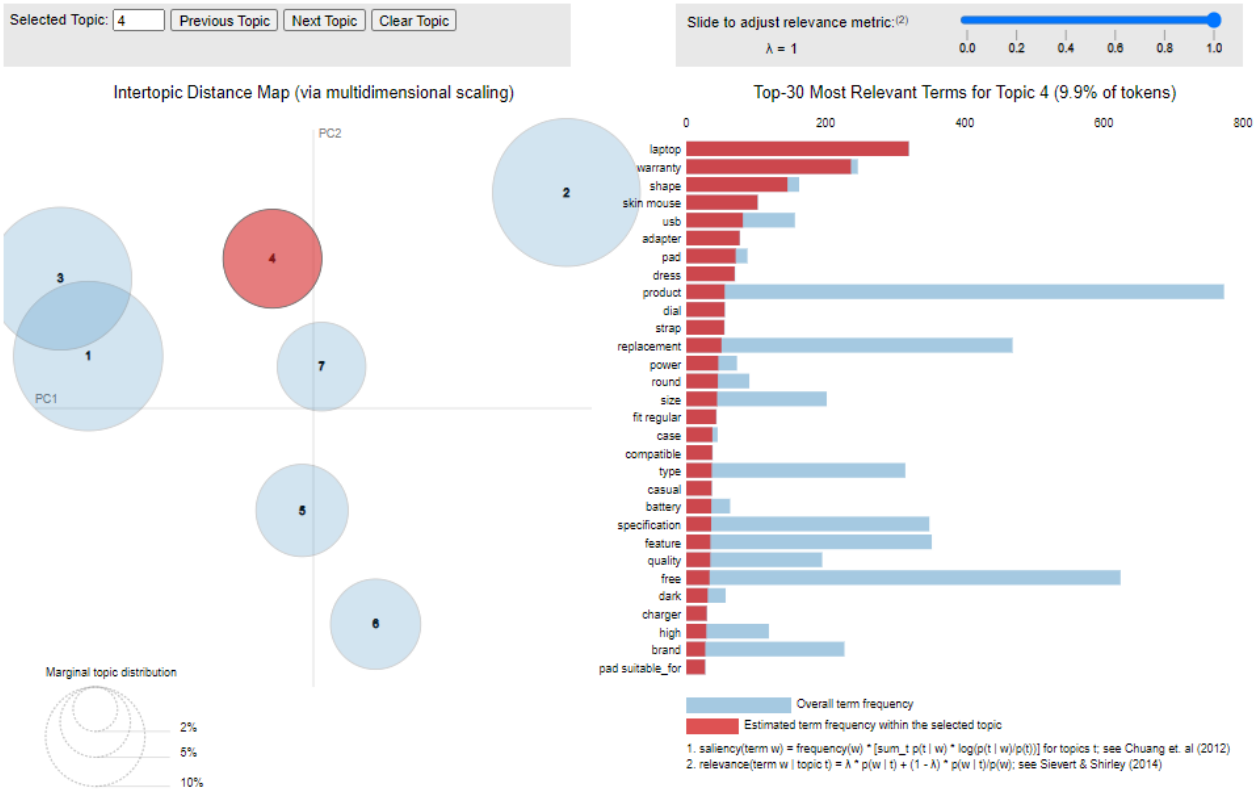
Description avant traitement
Skmei AD1057-Dark-Orange Sports Analog-Digital Watch - For Men, Boys - Buy Skmei AD1057-Dark-Orange Sports Analog-Digital Watch - For Men, Boys AD1057-Dark-Orange Online at Rs.1199 in India Only at Flipkart.com. Digital Chronograph, Alarm Watch, Light Function, Date & Month Display - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!

Description après lématisation
['skmei', 'dark', 'orange', 'sport', 'analog', 'digital', 'watch', 'men', 'boy', 'buy', 'skmei', 'dark', 'orange', 'sport', 'analog', 'digital', 'watch', 'men', 'boy', 'dark', 'orange', 'online', 'india', 'flipkart', 'com', 'digital', 'chronograph', 'alarm', 'watch', 'light', 'function', 'date', 'month', 'display', 'great', 'discount', 'genuine', 'product', 'day', 'replacement', 'guarantee', 'free', 'shipping', 'cash', 'delivery']

Description après stemming
['skmei', 'dark', 'orang', 'sport', 'analog', 'digit', 'watch', 'men', 'boy', 'buy', 'skmei', 'dark', 'orang', 'sport', 'analog', 'digit', 'watch', 'men', 'boy', 'dark', 'orang', 'onlin', 'india', 'flipkart', 'com', 'digit', 'chronograph', 'alarm', 'watch', 'light', 'function', 'date', 'month', 'display', 'great', 'discount', 'genuin', 'product', 'day', 'replac', 'guarante', 'free', 'ship', 'cash', 'deliveri']

Prétraitement des données textuelles - Latent Dirichlet Allocation (LDA)

- Ci-dessous la modélisation LDA pour classer notre corpus par thème :



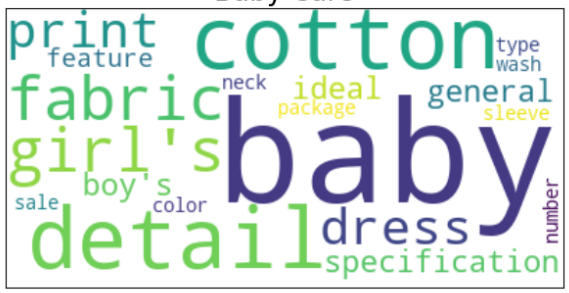
Méthode de modélisation	ARI score
Latent Dirichlet Allocation (LDA)	**0.088**

Topic 1	pack, design, color, use, feature, material, box, model, light…
Topic 2	free, buy, delivery, cash, shipping, genuine, product, flipkart, day…
Topic 3	print, baby, cotton, fabric, detail, pack, color, general, inch…
Topic 4	laptop, warranty, shape, skin mouse, usb, adapter, pad, dress..
Topic 5	mug, ceramic, coffee, perfect, sleeve, gift, design, strip, make…
Topic 6	home, wall, price, paper, apply, brass, like, durable, water…
Topic 7	skin, towel, quality, product, hair, trait, price, high, make, soap…

Prétraitement des données textuelles – Analyse TF-IDF

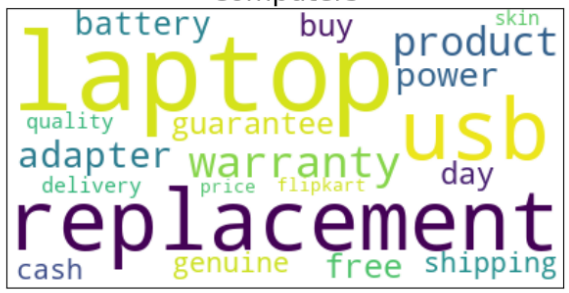
- Les mots-clés par catégorie:

Baby Care



Beauty and Personal Care

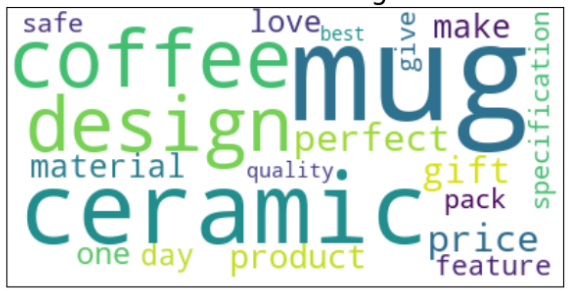
Computers



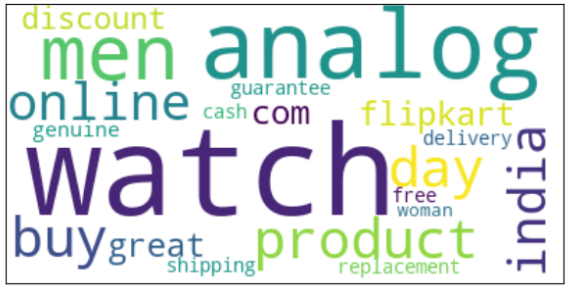
Home Decor Festive Needs

Home Furnishing

Kitchen Dining

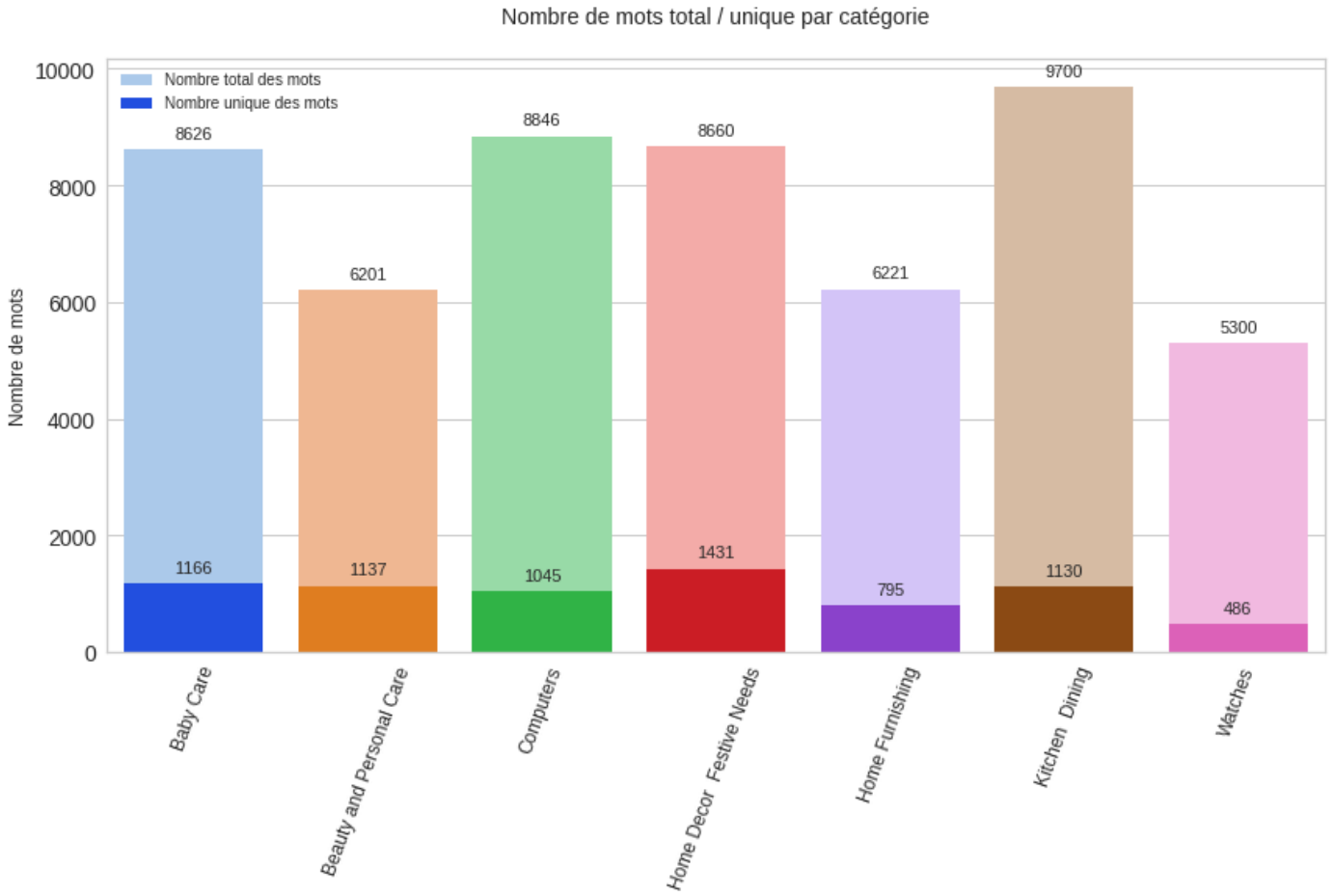


Watches



Prétraitement des données textuelles – Analyse TF-IDF

- Graphique du Nombre total de mots / unique par catégorie :



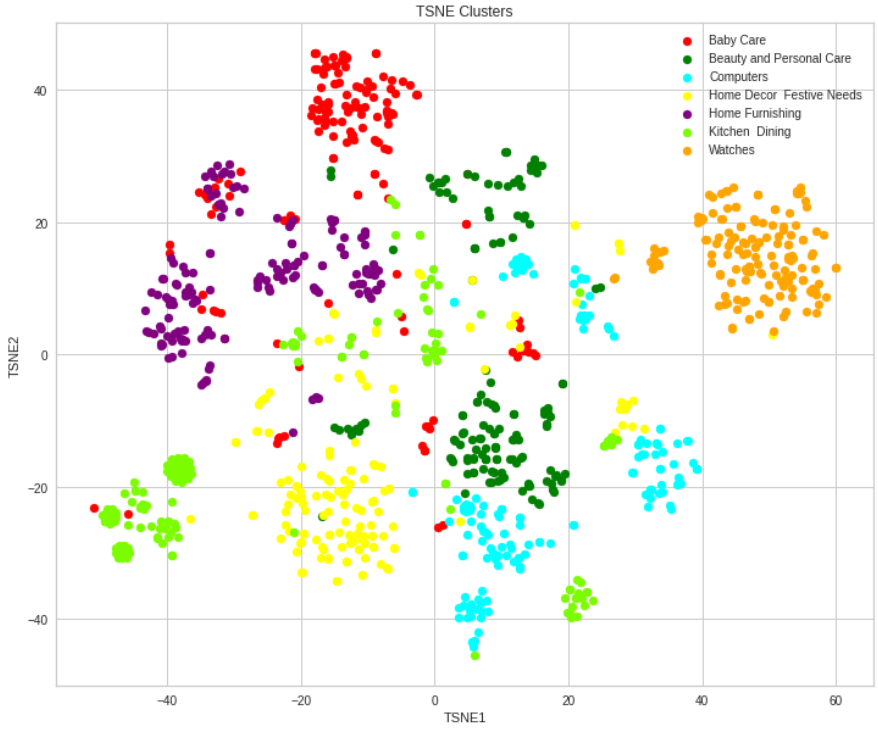
- Remarque : La catégorie « Ketchen Dinning » qui contient le plus de mots.

Prétraitement des données textuelles – Analyse TF-IDF

- Processus du traitement du corpus pour projeter nos données en 2D :

Initialiser le modèle « ***TfidfVectorizer, CountVectorizer…*** » → Entraîner / tranformer notre corpus → Réduire les dimensions (TSNE) → Projeter les données Par catégorie en 2D

- Ci-dessous la représentation 2D de notre corpus par catégorie :



Prétraitement des données textuelles - Segmentation K-Means

Lemmatisation

Réduction de dimension (TSNE)

Méthode d'extraction des features	ARI score
TfidfVectorizer	0.548
CountVectorizer	0.469
Word2Vec	0.216

Stemming

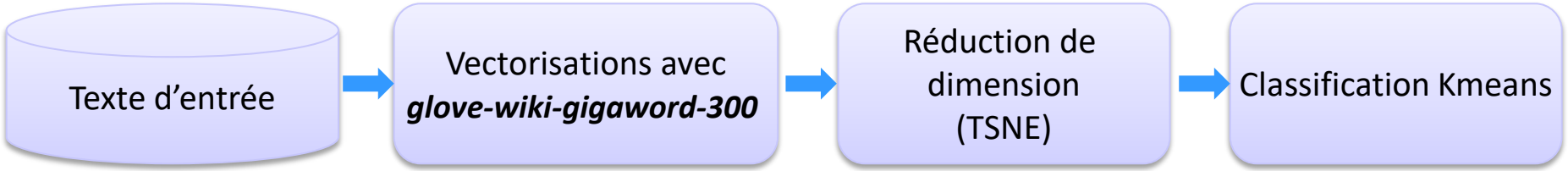
Réduction de dimension (TSNE)

Méthode d'extraction des features	ARI score
TfidfVectorizer	0.544
CountVectorizer	0.378
Word2Vec	0.218

Choix traitement texte : Lemmatisation
Choix d'extraction : TfidfVectorizer

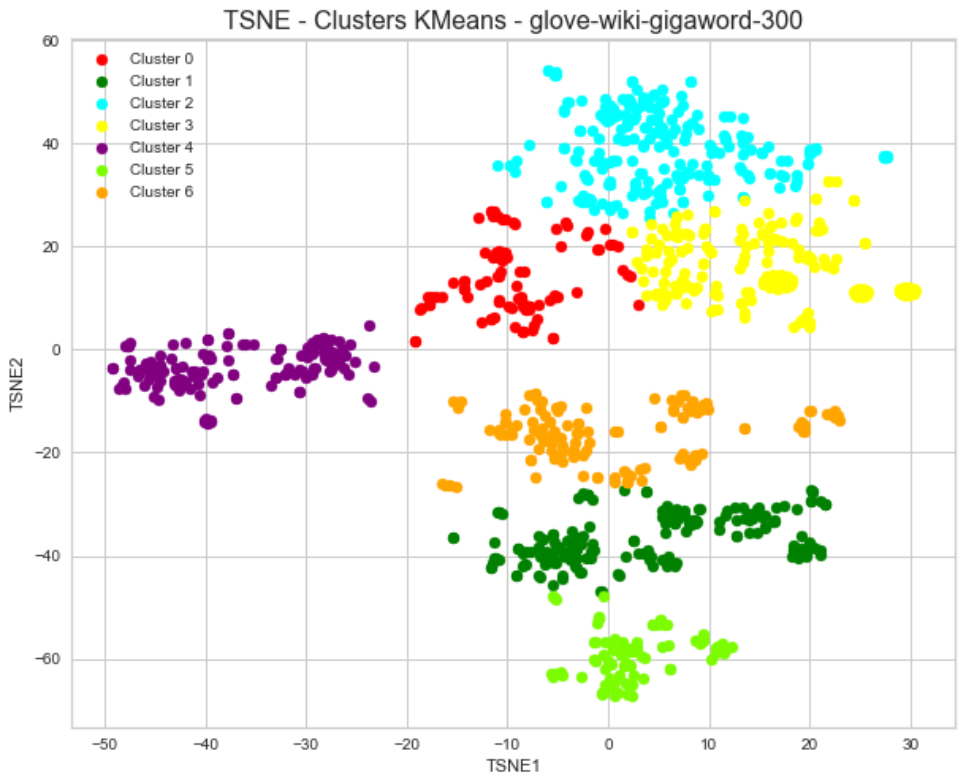
Prétraitement des données textuelles – Transfer Learning

- Le processus effectué pour le Transfer Learning :

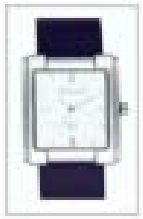


- Projection / ARI score :

Méthode d'extraction des features	ARI score
glove-wiki-gigaword-300	0.385



Prétraitement des images



Prétraitement des images – Exemple d’images par catégorie

Home Furnishing

Baby Care

Watches

Home Decor Festive Needs

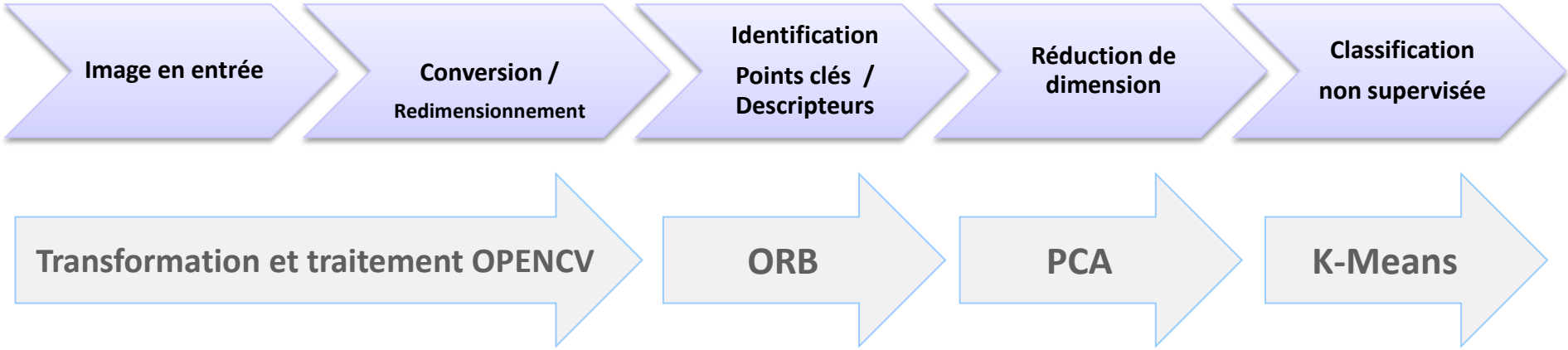
Kitchen Dining

Beauty and Personal Care

Computers

Prétraitement des images – Transformation / Classification

- Processus de transformation et classification des images :

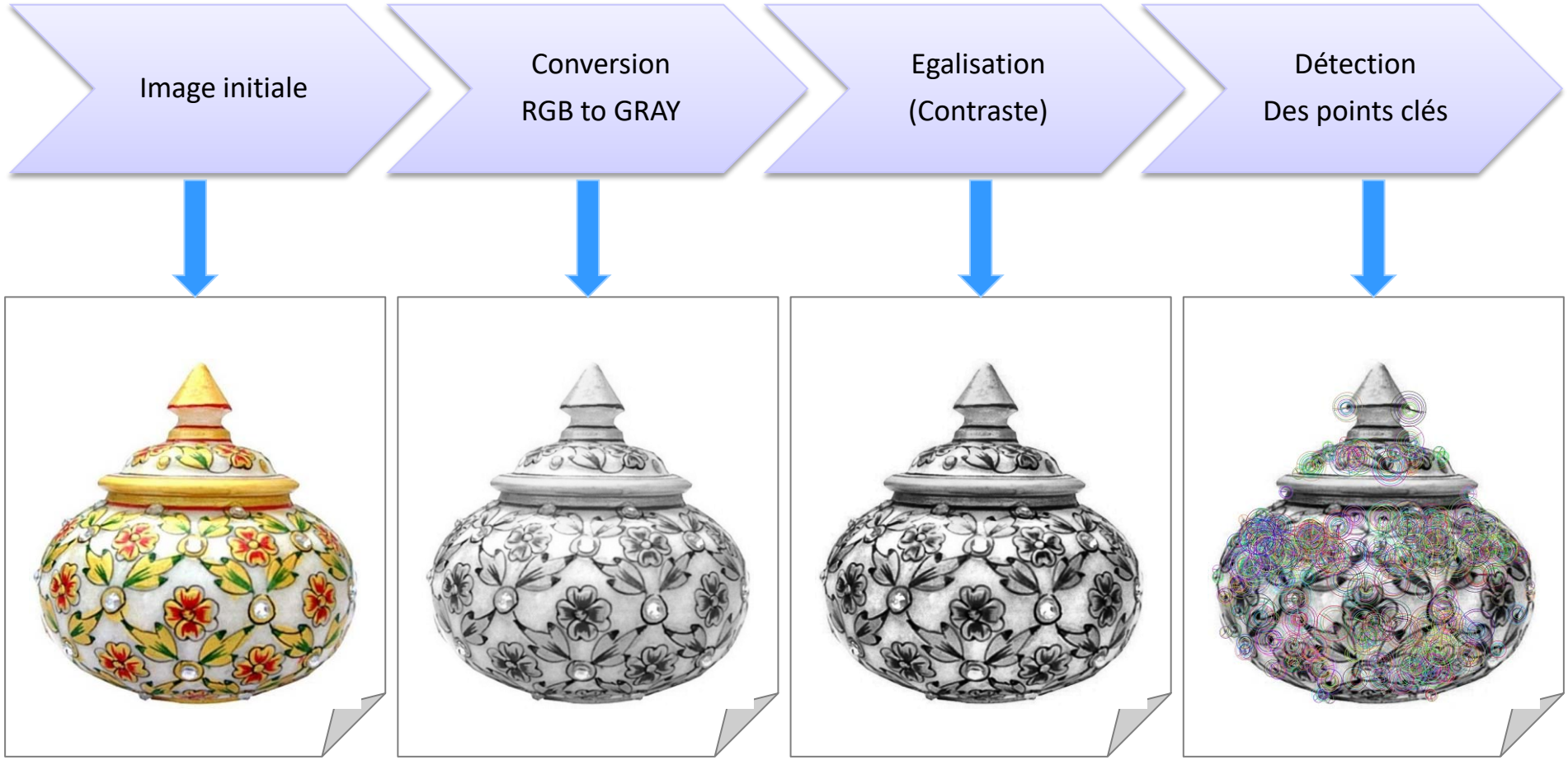


- **Remarque** :

  La classification supervisée n'est pas un choix judicieux pour ce projet, vu que le volume de données est très petit ce qui représente un risque de sur-apprentissage.

Prétraitement des images – Processus ORB (Exemple)

- Exemple de transformation et traitement d'une image avec ORB :



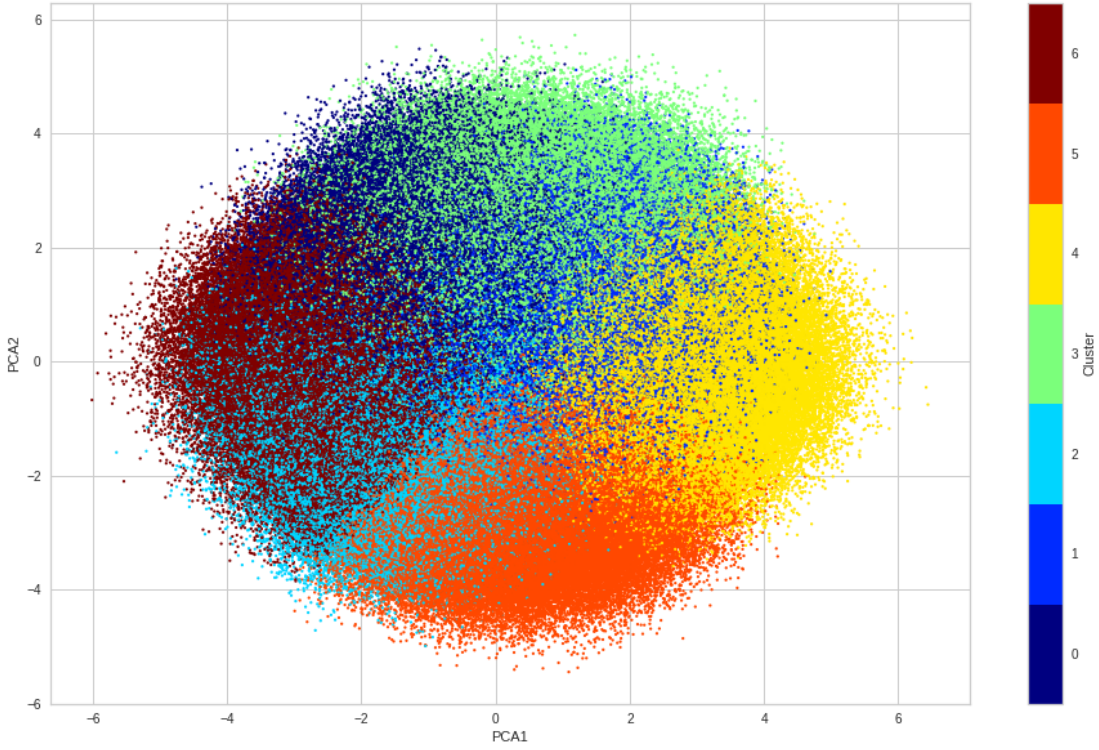
Prétraitement des images – Processus ORB

- Les étapes effectuées pour classer les images :

  - Créer les descripteurs avec ORB :
    - → ORB a généré 891903 descripteurs.
    - → Chaque descripteur a 32 dimensions.
  - Classifier les descripteurs avec K-Means (7 clusters).
  - Réduire les dimensions avec PCA pour visualiser les données.
  - Prédire les clusters des images.
  - Calculer la similarité avec l'ARI score.

Prétraitement des images – Processus ORB

- Ci-dessous la projection des données :

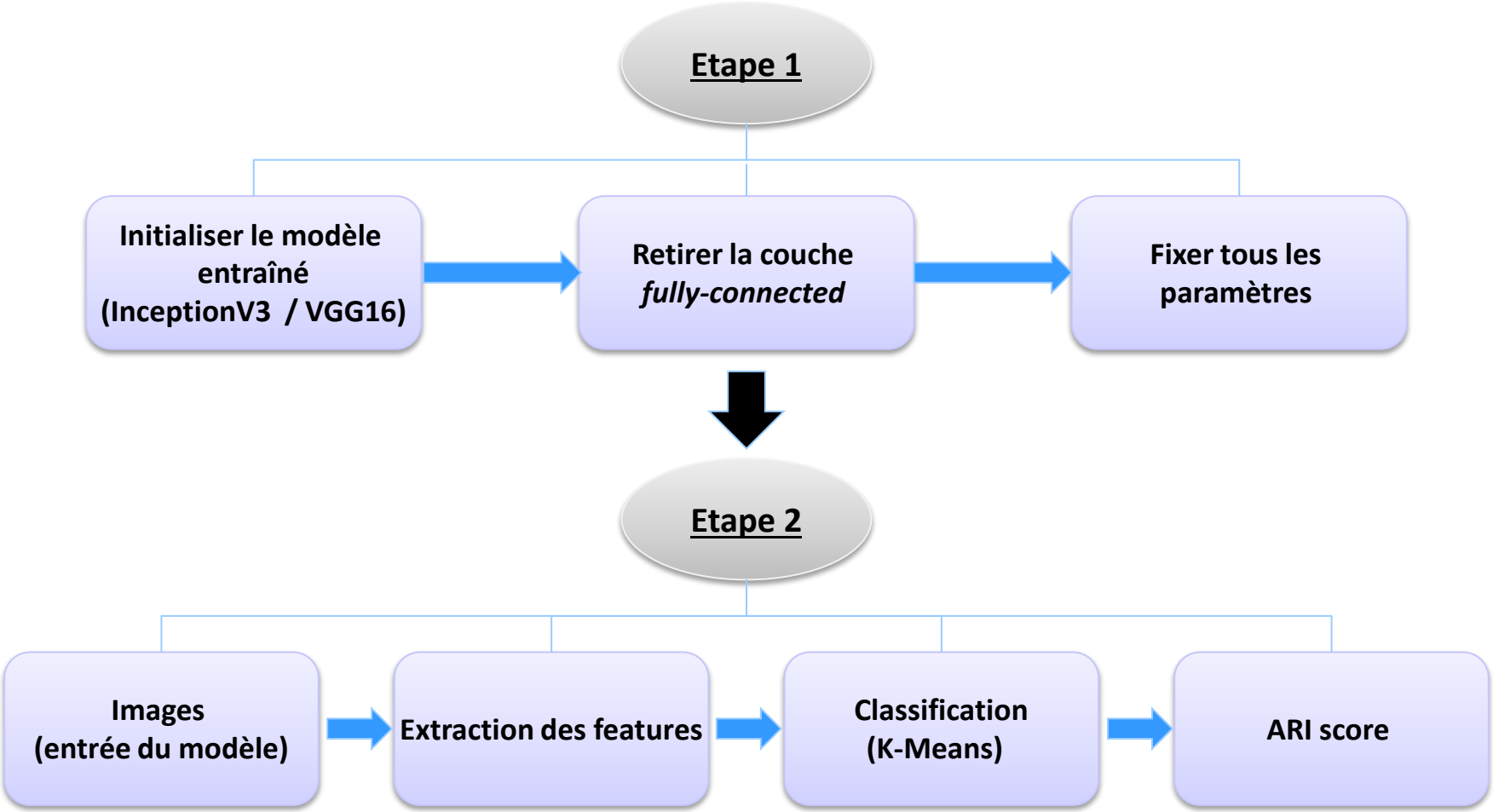


- ARI score :

Méthode d'extraction des features	ARI score
ORB	0.018

Prétraitement des images – Transfer Learning (Non supervisé)

- Le processus du Transfer Learning s'effectué en deux étapes :



Prétraitement des images – Résultat

Transfer Learning

Extraction des features

Méthode d'extraction	ARI score
InceptionV3	0.247
VGG16	0.093

ORB

Extraction des features

Méthode d'extraction	ARI score
ORB	0.018

Choix d'extraction : InceptionV3

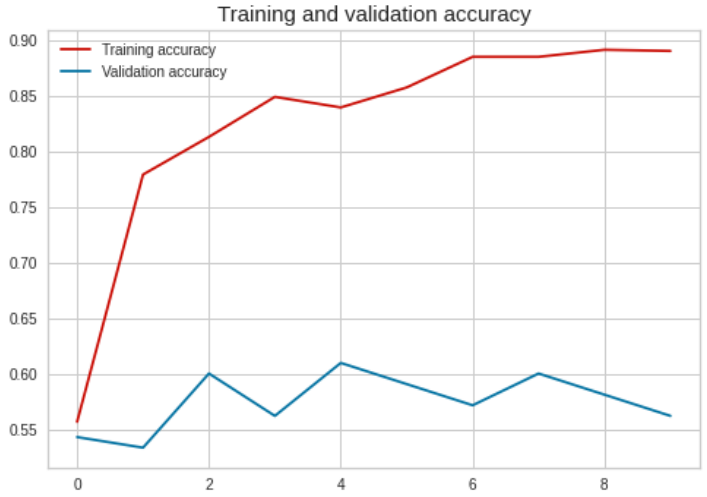
Prétraitement des images – Classification supervisée

Processus de classification supervisée

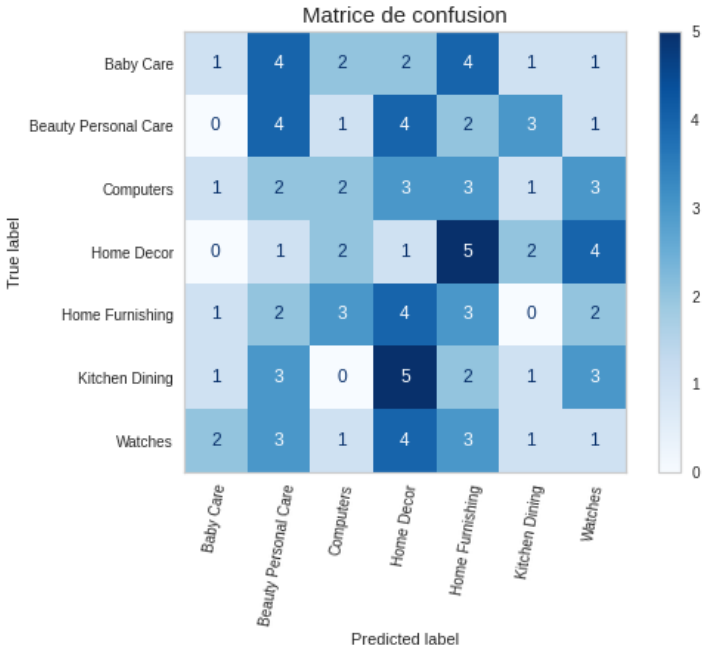
1

- Création des répertoires Train / Test
- Initialisation InceptionV3
- Remplacement de la dernière couche
- Augmentation de données
- Entrainement du modèle
- Evaluation

2



3



Combinaison des données textuelles et visuelles

Combinaison des données textuelles et visuelles

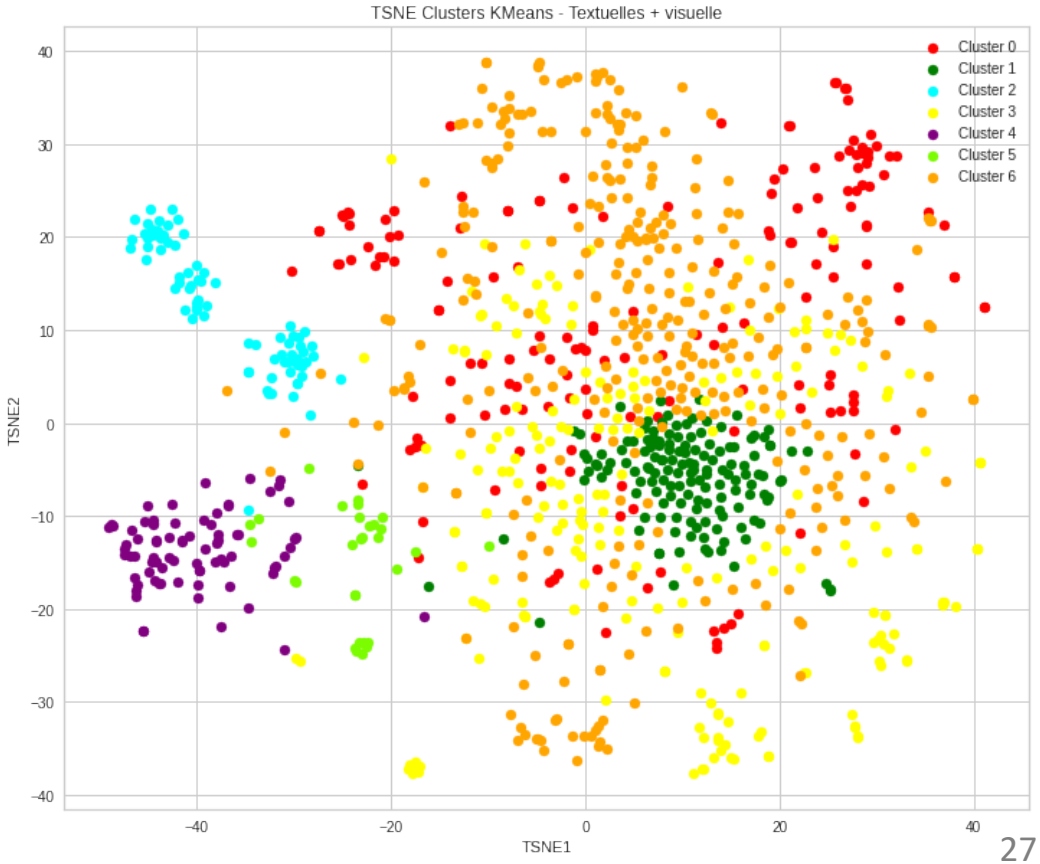
- Nous allons effectuer une combinaison des données générées par :

  - Les données textuelle : **TfidfVectorizer**
  - Les données visuelle : Transfer Learning **InceptionV3**

Données textuelles
TfidfVectorizer

Données visuelle
TL - InceptionV3

Combinaison des features	ARI score
TfidfVectorizer + InceptionV3	0.323



Conclusions

Conclusions

- La partie textuelle :
  - Traitement du texte : La lemmatisation est le meilleur choix.
  - Extraction des features : ***TfidfVectorizer*** a donné des résultats acceptables (54,8%)
  - Réduction de dimension : Le TSNE améliore nettement l'ARI score.

- La partie visuelle :
  - le Transfer Learning ***InceptionV3*** a donné le meilleur score (24,7%), ce score reste néanmoins insuffisant pour classer les images.

- La combinaison des données textuelles et visuelles :
  - Cette combinaison n'améliore pas le score (32%) par rapport au score du traitement textuelle.

Conclusions

- Etude de faisabilité :

  La faisabilité d’un moteur de classification d'articles en utilisant plusieurs approches n’a pas donné de résultats concluants.

- Les recommandations pour sa création éventuelle :

  - Améliorer les descriptions des produits en utilisant des mots clés.
  - Favoriser le Transfer Learning pour le traitement des images.
  - Scinder les catégories avec un score faible en plusieurs sous catégories.
  - Favoriser l’apprentissage supervisé si le volume des données est important.