P6 CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

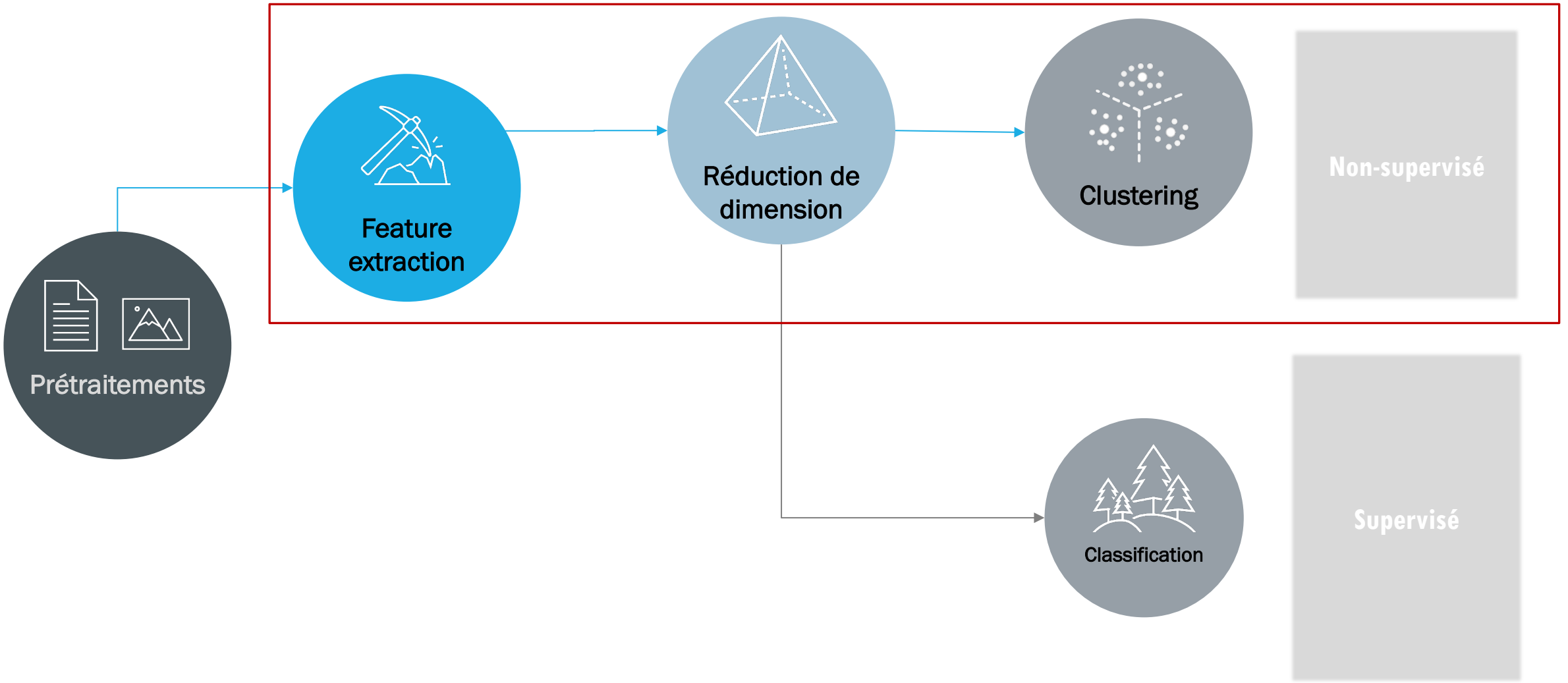
Formation Data Scientist
OpenClassrooms

CONTEXTE

OBJECTIFS

Problématiques :

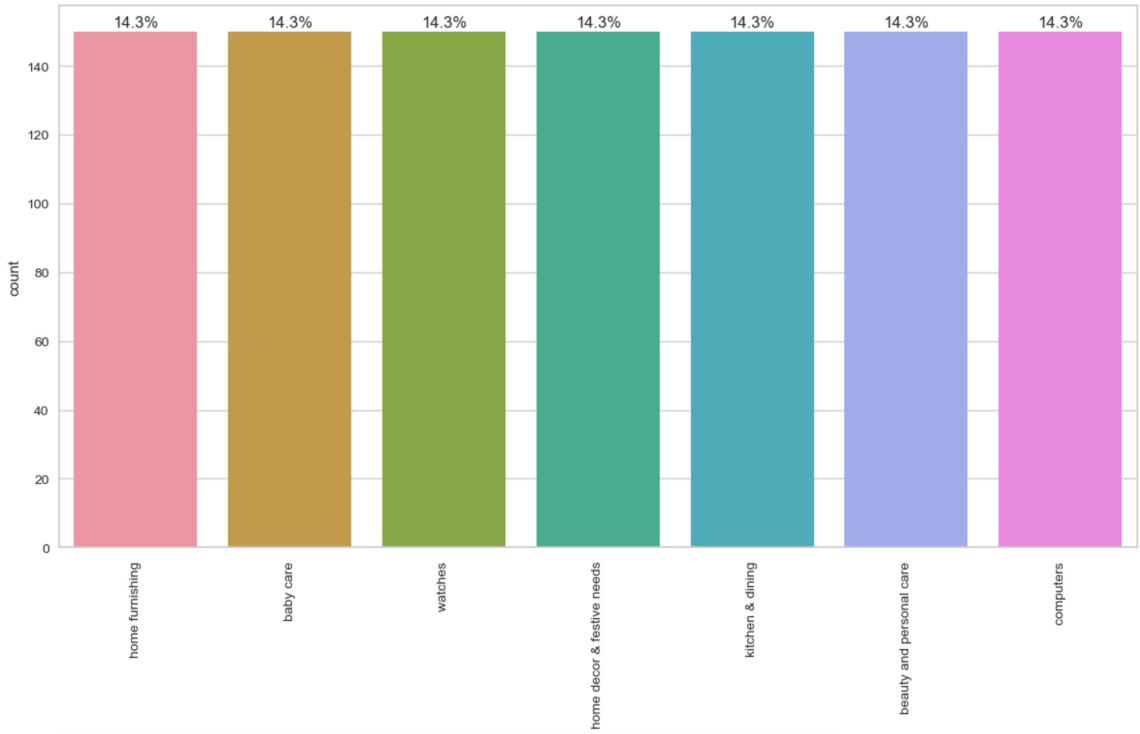
- **réaliser une première étude de faisabilité d'un moteur de classification** d'articles basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article.

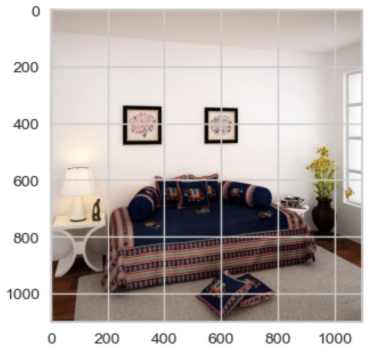


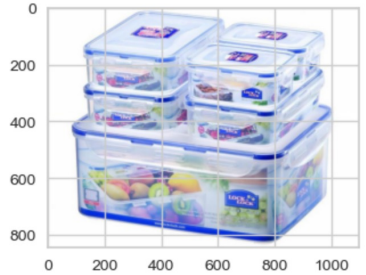
DATASET

- Une table de description
  - Séparation de la variable category_tree
  - Catégories : uniformément réparties
    - Outil d'évaluation des clusters ARI / accuracy

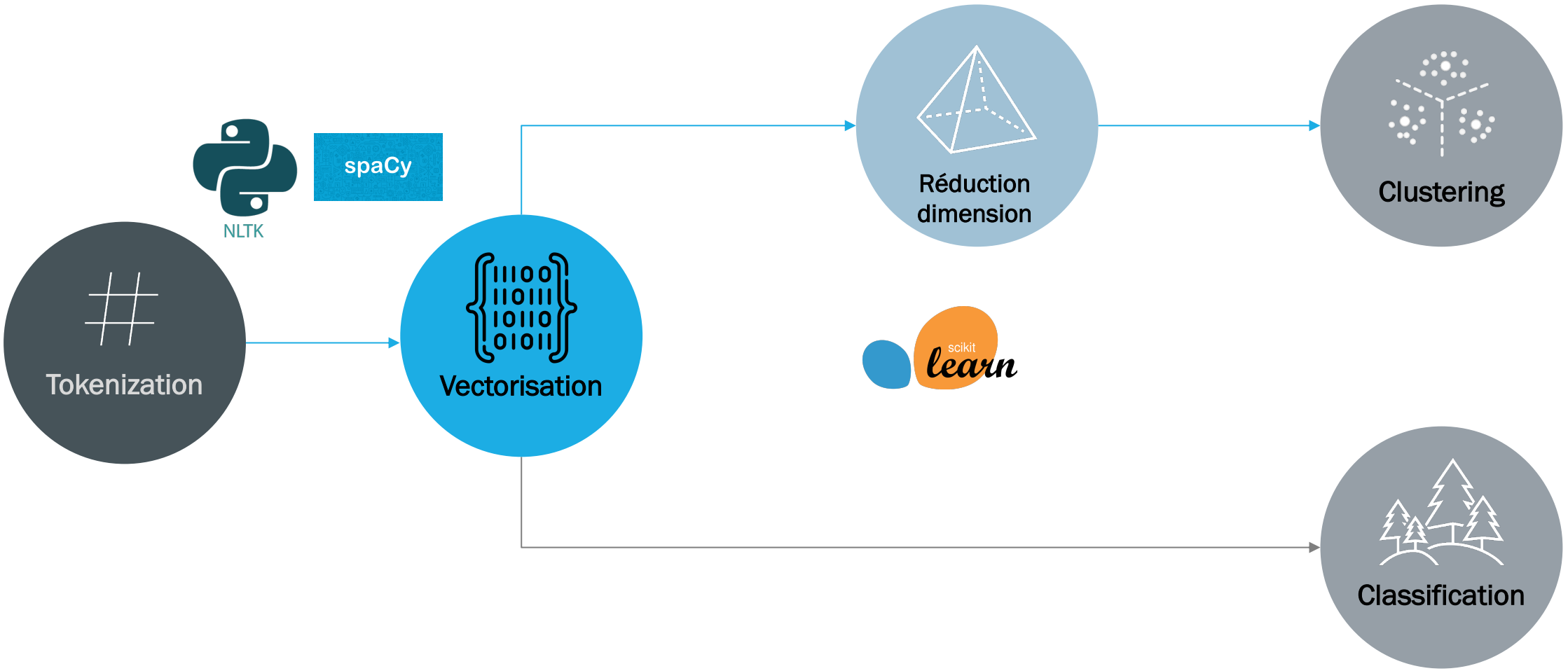
- 1050 images
  - Format .jpg
  - Tailles variables
  - RGB







NLP: MOTEUR DE CLASSIFICATION D'ARTICLES À PARTIR DE LEURS DESCRIPTIONS



PRÉ-TRAITEMENT

Minuscules → Ponctuation → URL → Nombre → Espace en trop → Lemmatisation → Tokenisation → Stop Words

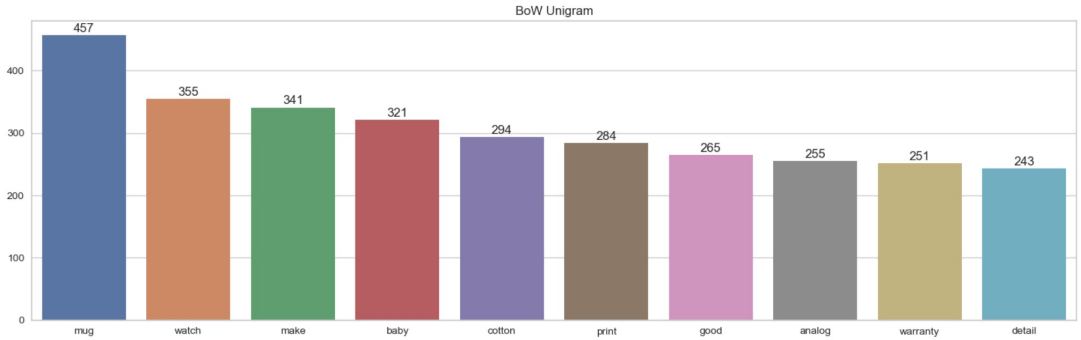
'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment 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 valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish 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 get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester'

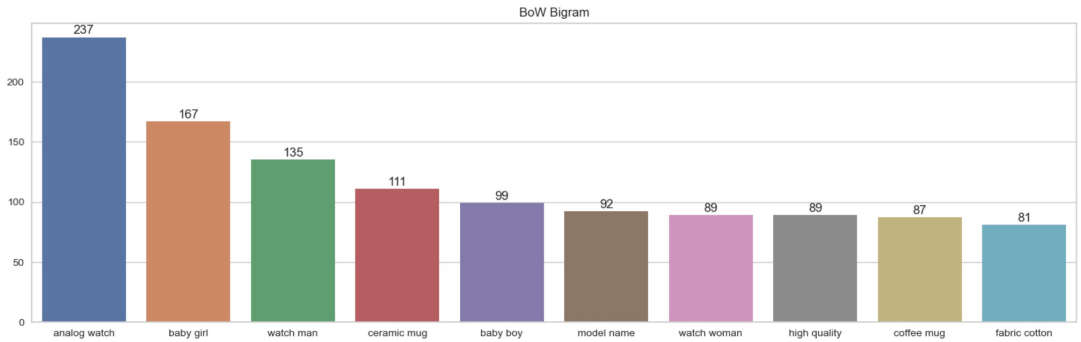
'elegance polyester multicolor eyelet door curtain floral curtain elegance polyester multicolor eyelet door curtain height curtain enhance look interior curtain make high quality polyester fabric eyelet style stitch metal ring make room environment romantic curtain ant wrinkle anti shrinkage elegant apparance give home bright modernistic appeal surreal attention sure steal heart contemporary eyelet valance curtain slide smoothly draw apart first thing morning welcome bright sun ray want wish good morning whole world draw close evening create special moment joyous beauty give soothe print bring home elegant curtain softly filter light room get right amount sunlight elegance polyester multicolor eyelet door curtain height elegance door eyelet model name polyester door curtain model duster multicolor length content curtain body polyester'

VECTORISATION

$n_{i,j}$: # apparition du mot i dans le doc j
L_j : # mot dans le doc j
N : # total de doc
f_i : # doc contenant i

Bag-of-Words

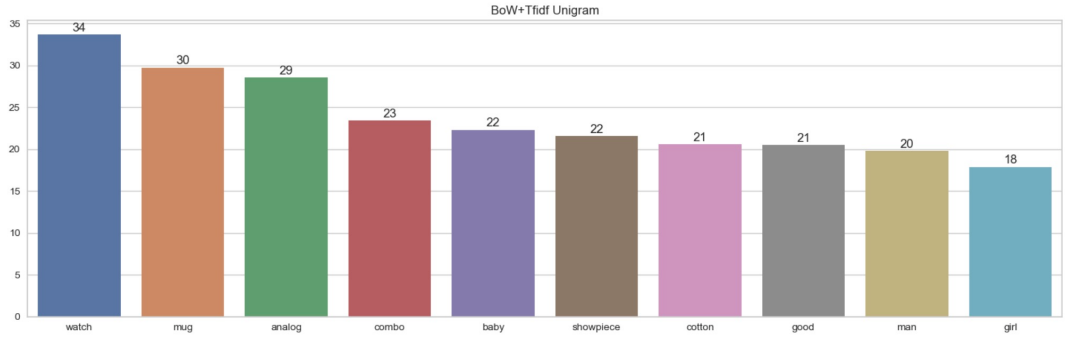


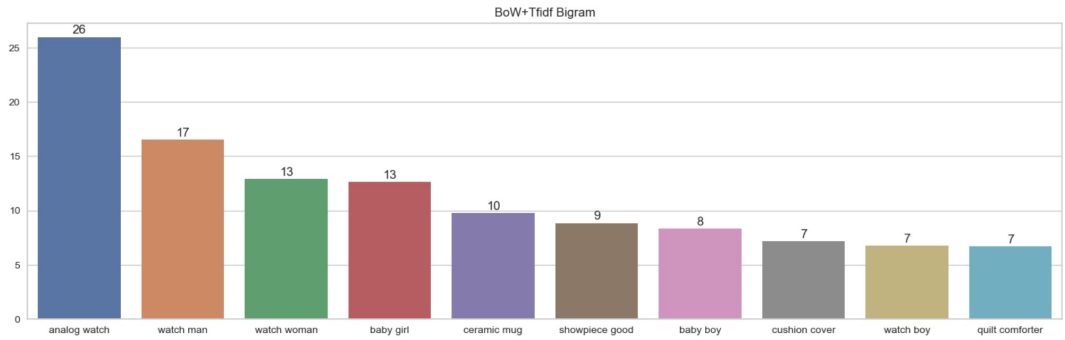


Avec TF-IDF

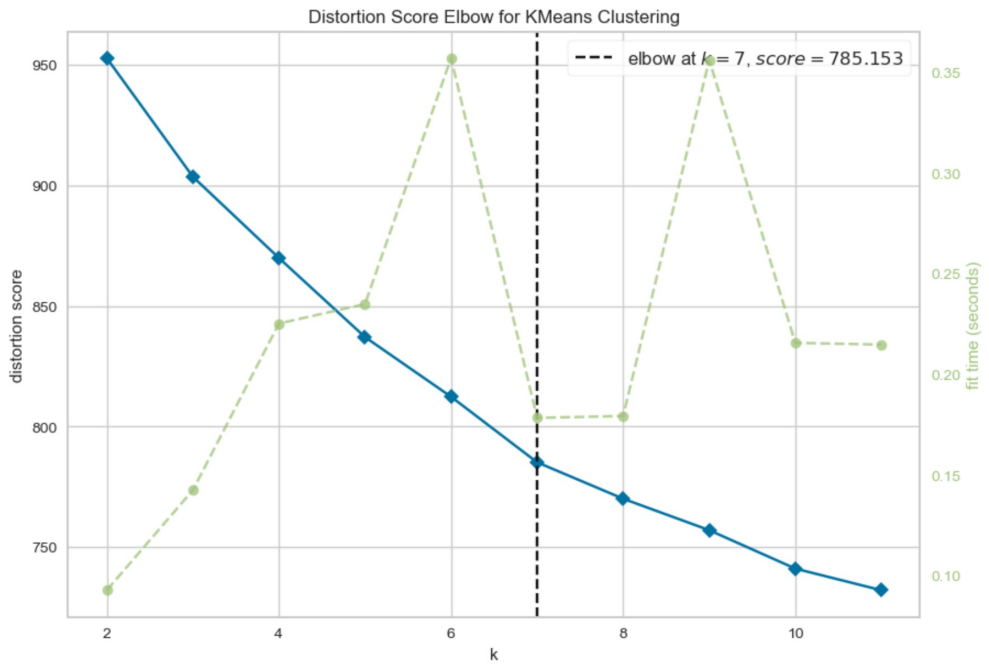
$$TF(i,j) \times IDF(i) = \frac{\log(1+n_{i,j})}{\log(L_j)} \times \log\left(\frac{N}{f_i}+1\right)$$

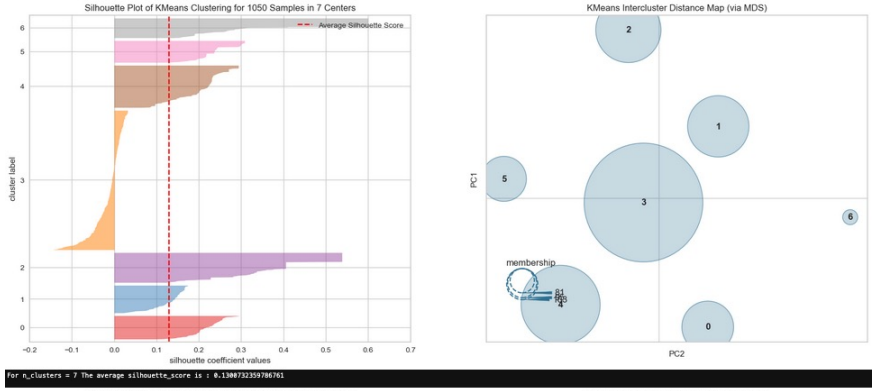
Importance du mot ***i*** dans : ↑ **doc *j*** ↓ **corpus**

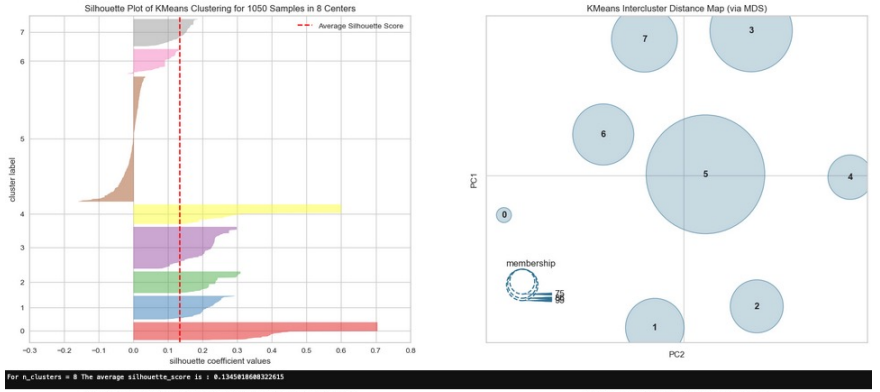


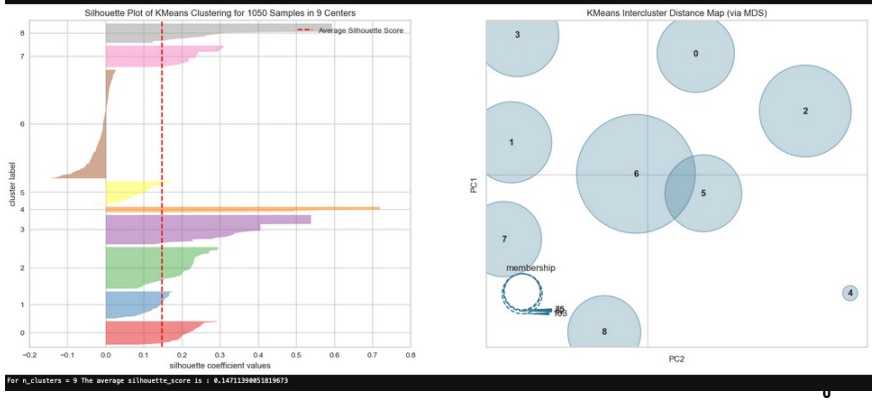


EVALUATION DU NOMBRE DE CLUSTERS







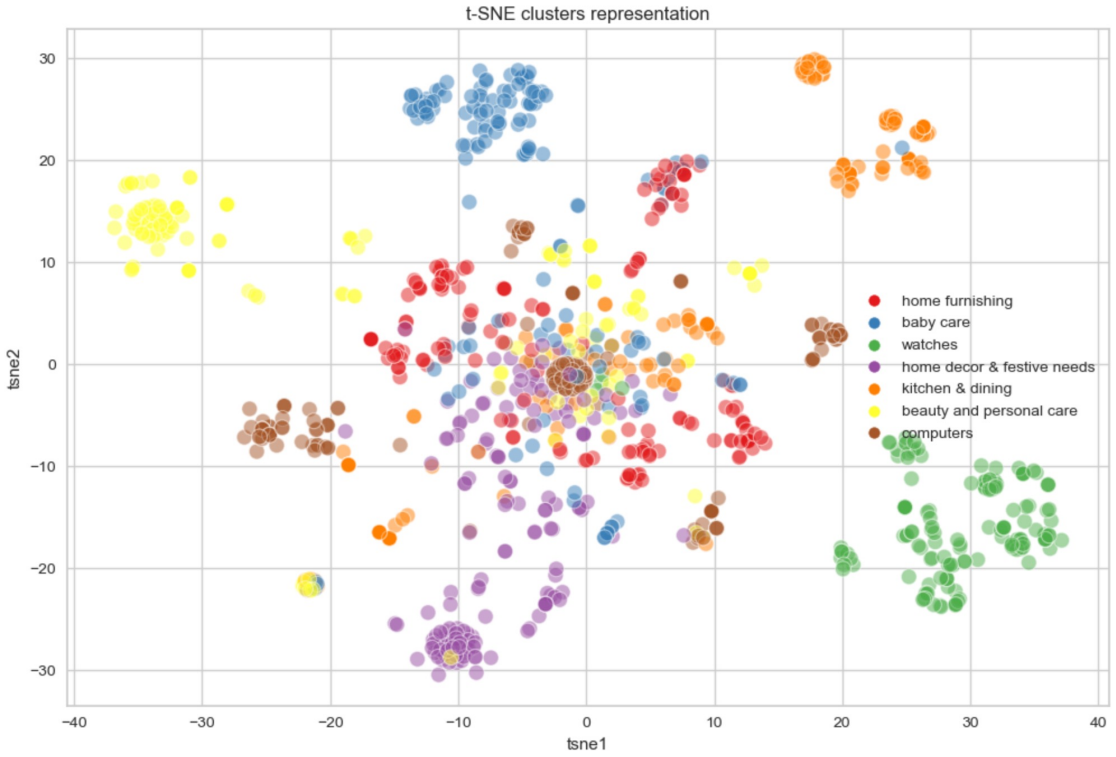


VISUALISATION DEUX DIMENSIONS (T-SNE)

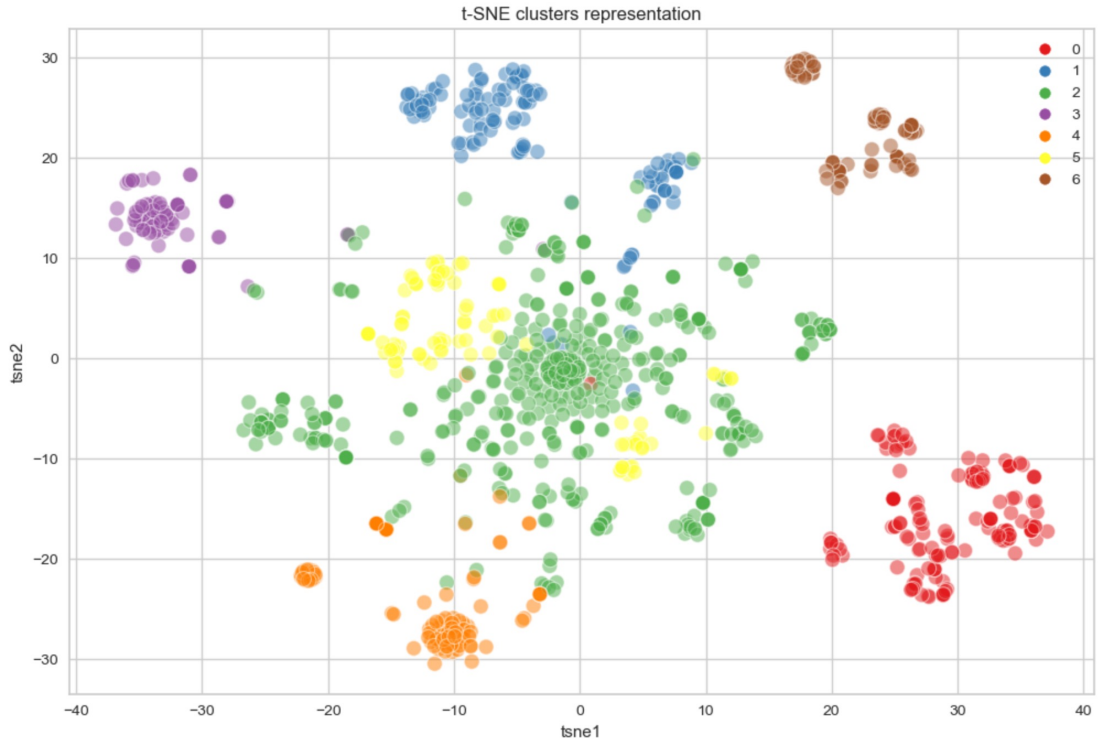
Dimension réduction par LSA (Latent Semantic Analysis)

- Input = 213 dim
- Output = 163 dim
- 99% variance

Répartition des produits par catégorie



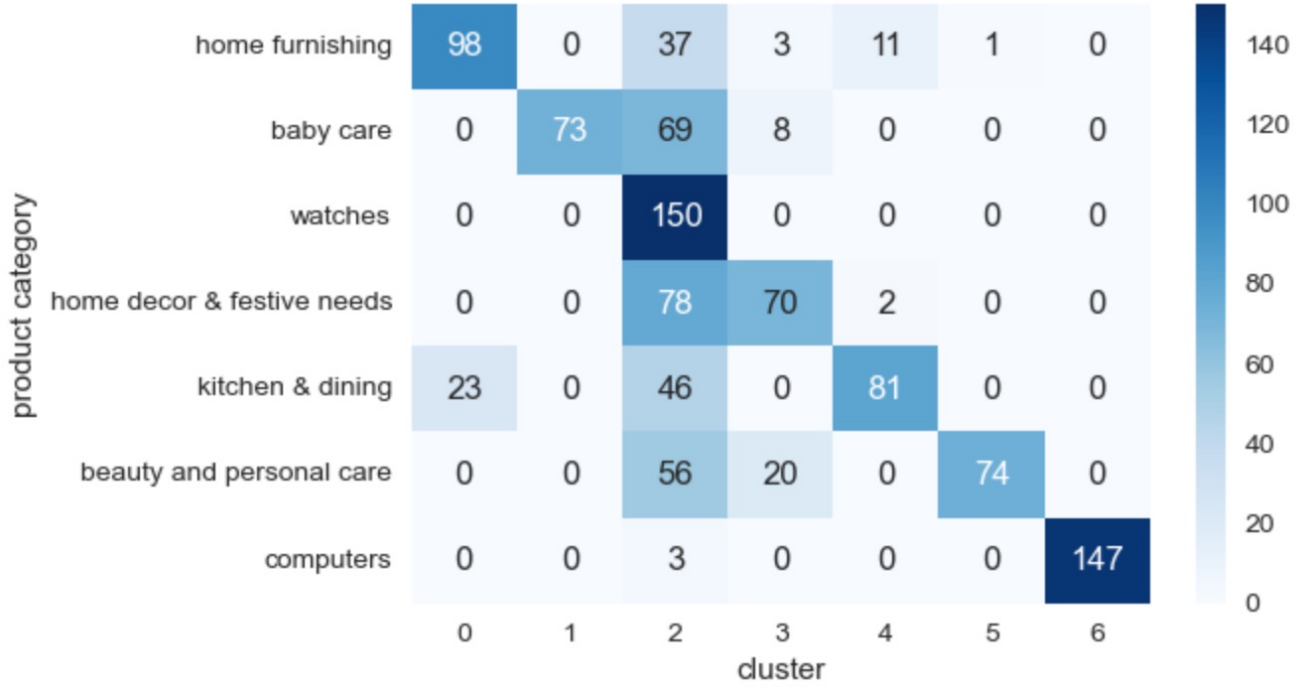
Répartition des produits par cluster



RÉSULTATS

Accuracy = 66% (après recalage)

ARI = 34%



Conclusion :

La classification à partir de la description des articles semble faisable.

- Il faudra néanmoins améliorer le traitement notamment avec un dictionnaire de stopword plus spécifique à la thématique.
- Il serait surement aussi nécessaire de définir de nouvelle catégorie de produit plus spécifique.

ESSAI DE CLASSIFICATION

But : prédire la faisabilité de classification en 7 categories

Algorithme : Multinomial Naives Bayes

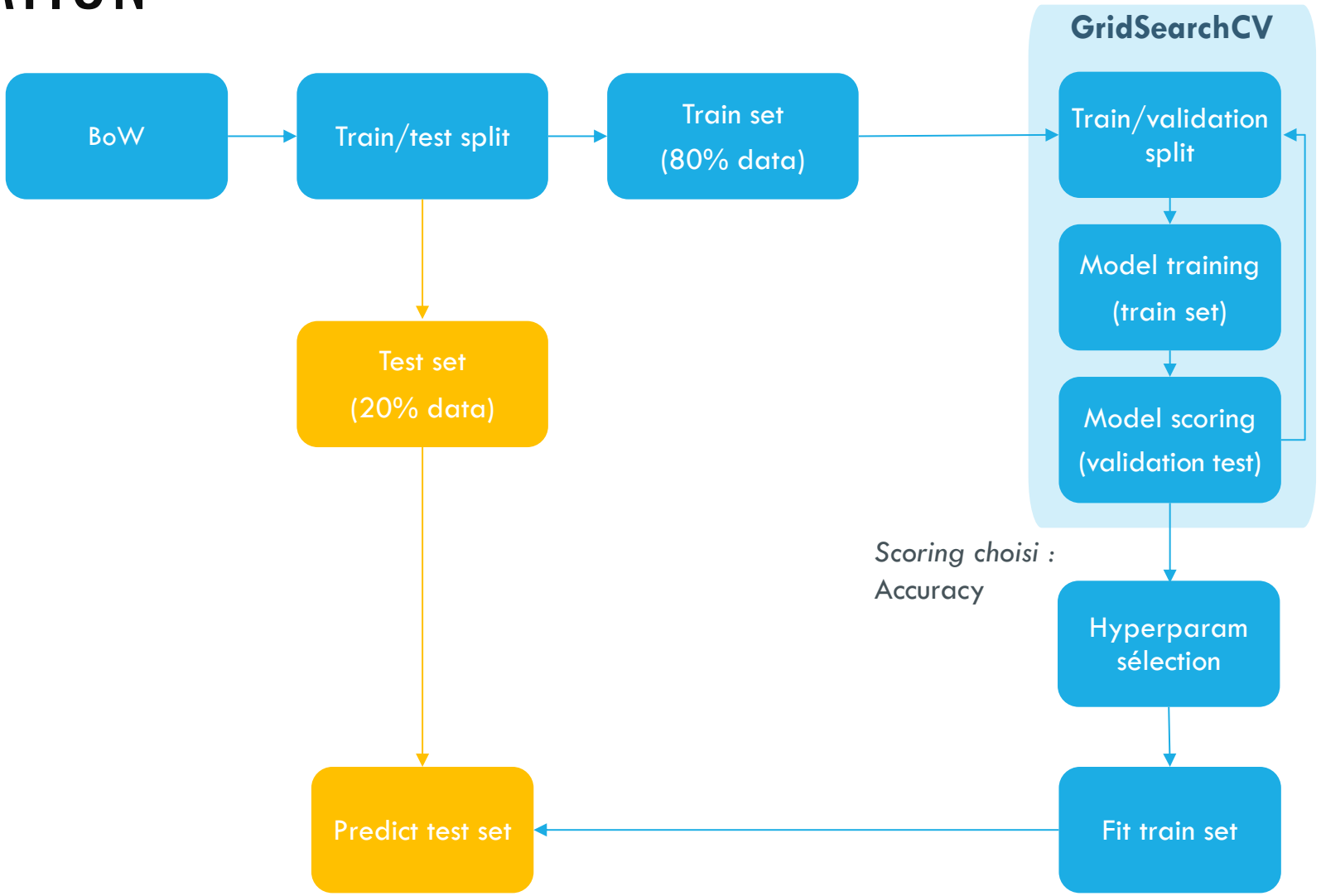
Hyperparameter :

- Utilisation de TF-iDF
- Régularisation l2
- Occurance minimal : 0
- Occurance maximal : 0.5
- Utilisation de Uni et bigramme

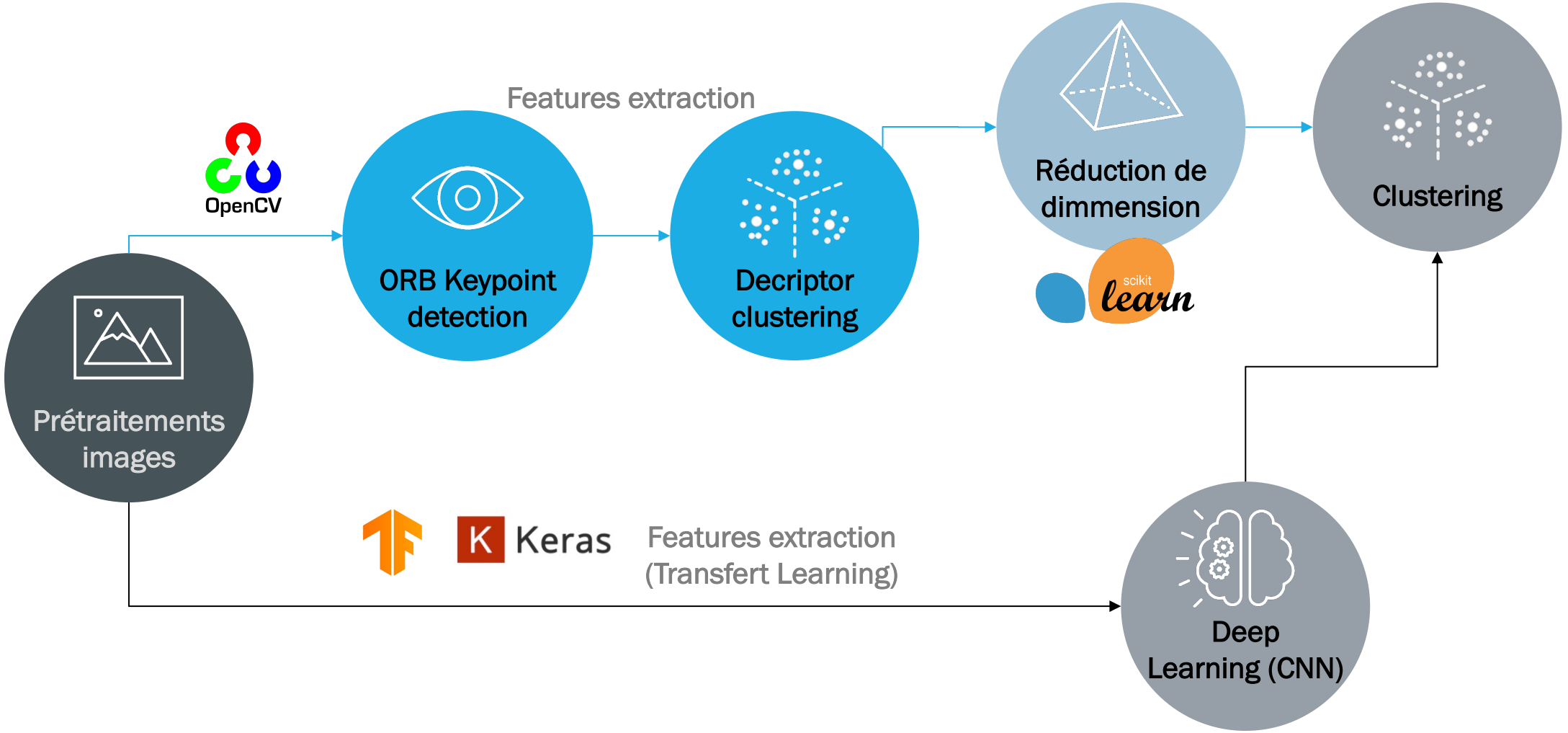
Temps : 87 secondes

Accuracy sur le jeu de test : 92%

ARI sur le jeu de test : 83%

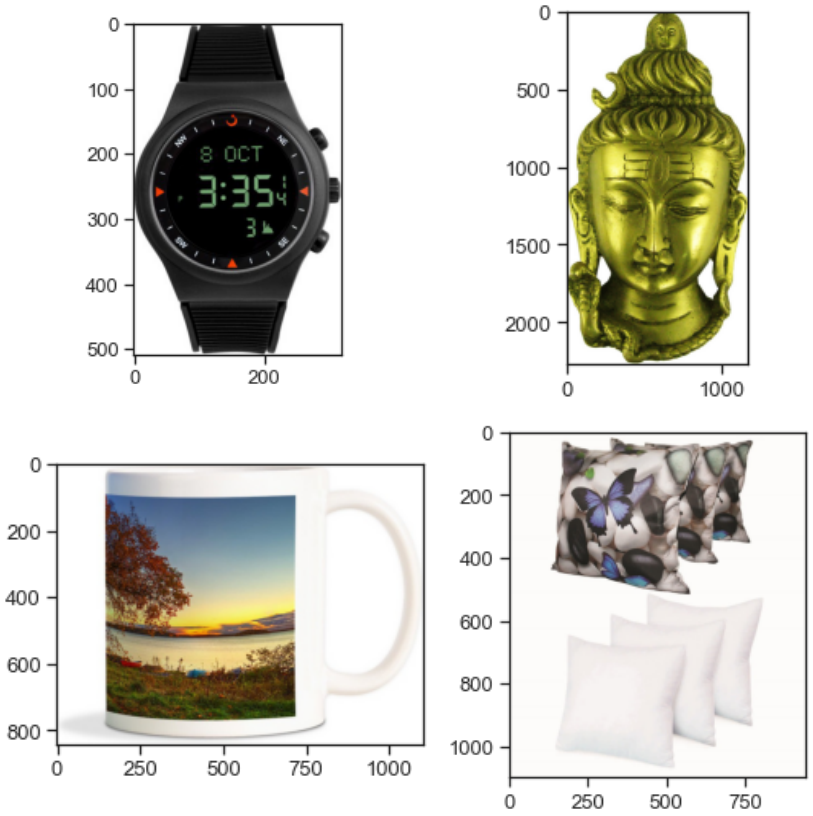


CV: MOTEUR DE CLASSIFICATION D'ARTICLES À PARTIR DE LEURS IMAGES



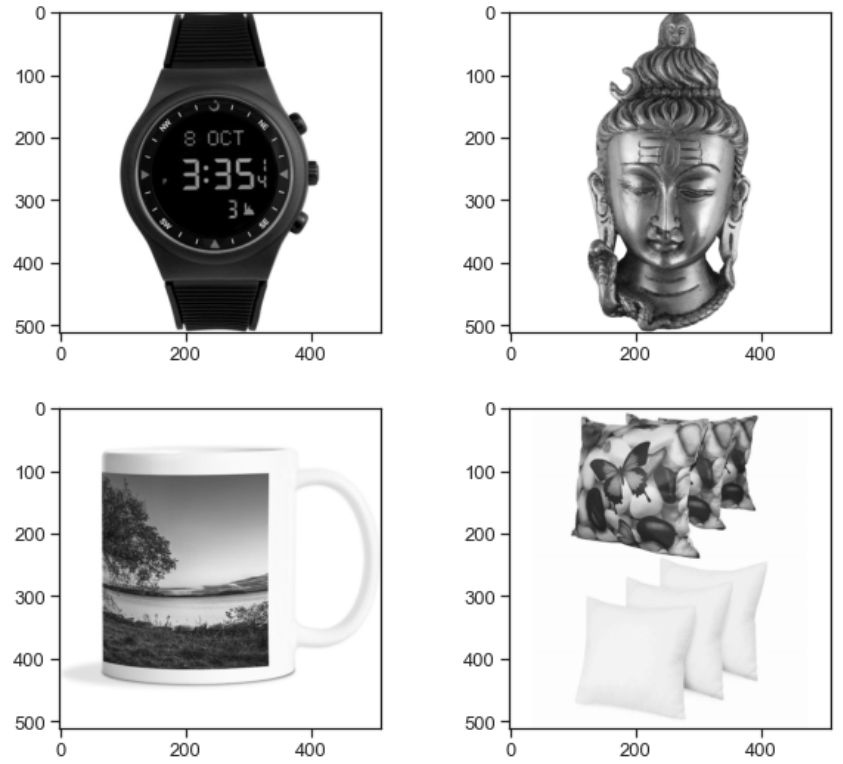
PRÉTRAITEMENT DES IMAGES

RGB et taille variable



OpenCV

BW et taille = 512 x 512 px

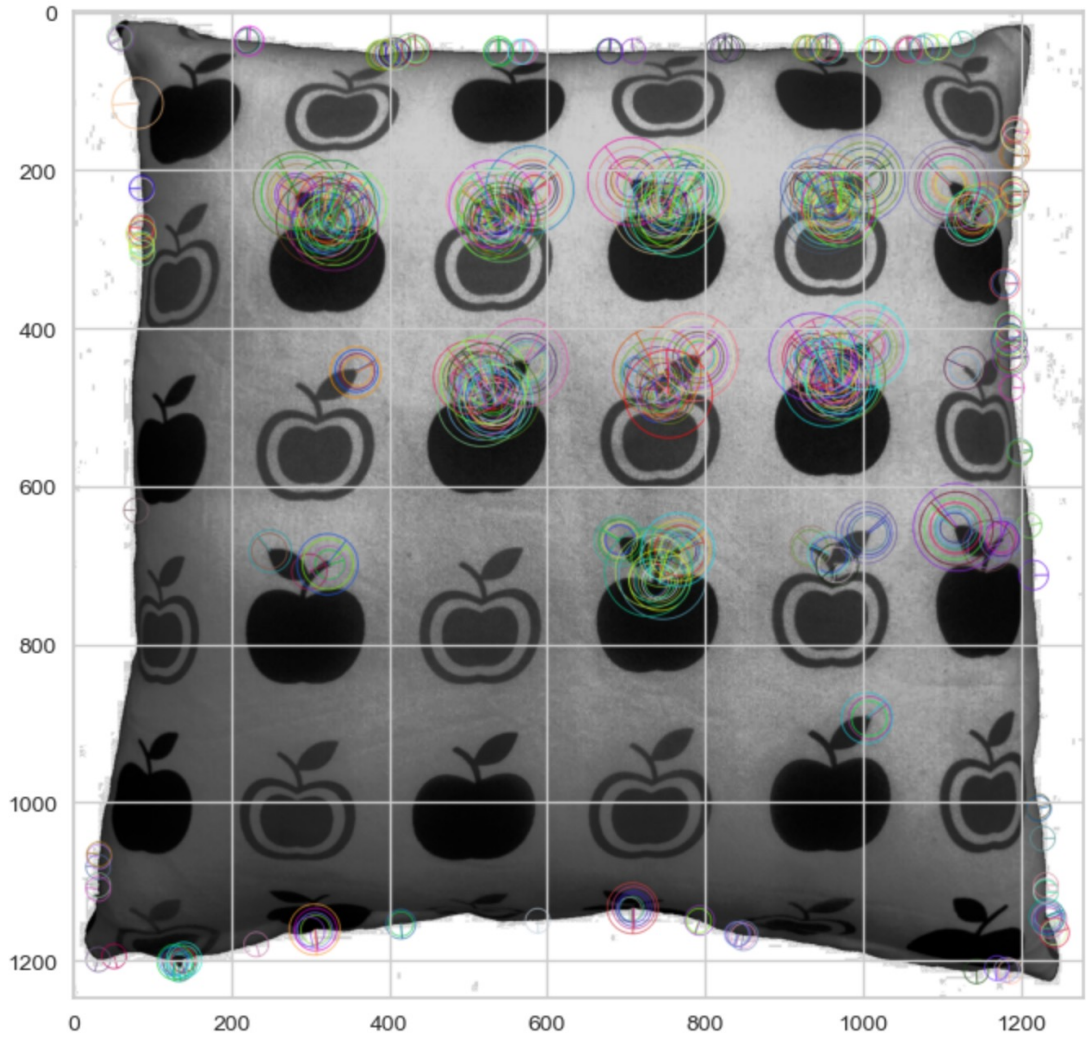


KEYPOINTS DETECTION USING ORIENTED FAST AND ROTATED BRIEF (ORB)

Par image :

- 500 key-points (max)
- 512 x 32 descripteurs

Total : 500k x 32 descripteurs

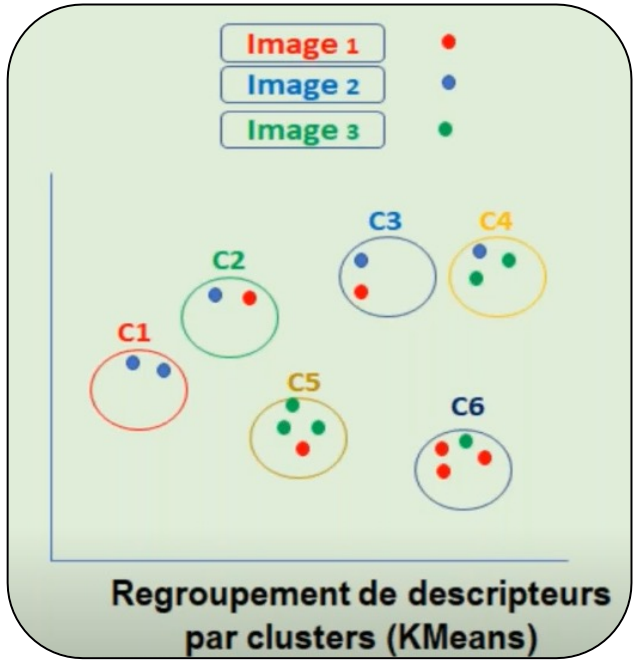


CRÉATION DES FEATURES PAR IMAGE

Clustering des descripteurs = création des features

MiniBatch Kmeans

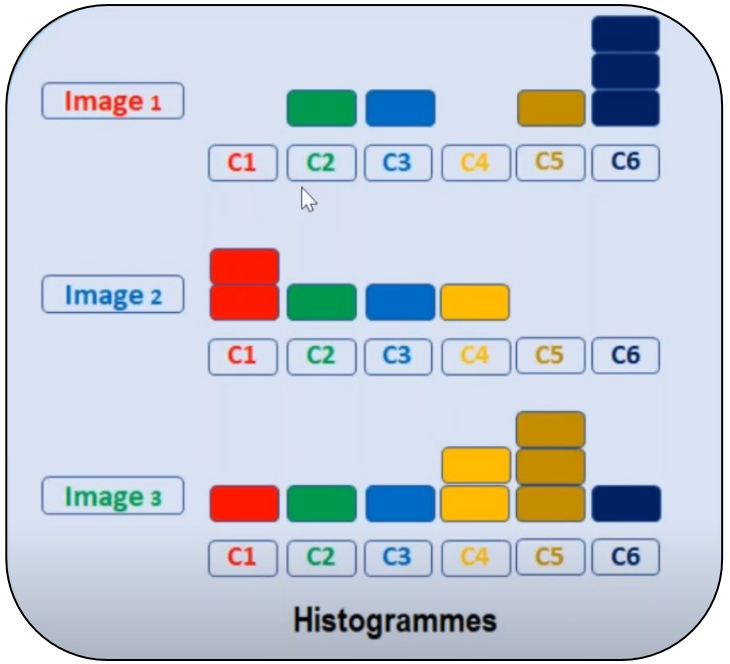
$\sqrt{520k} = 721$ features



Histogrammes = remplissage features (Bag of Visual Words)

Combien de fois une feature est présente sur l'image

1 / image avec toutes les features et leurs comptage

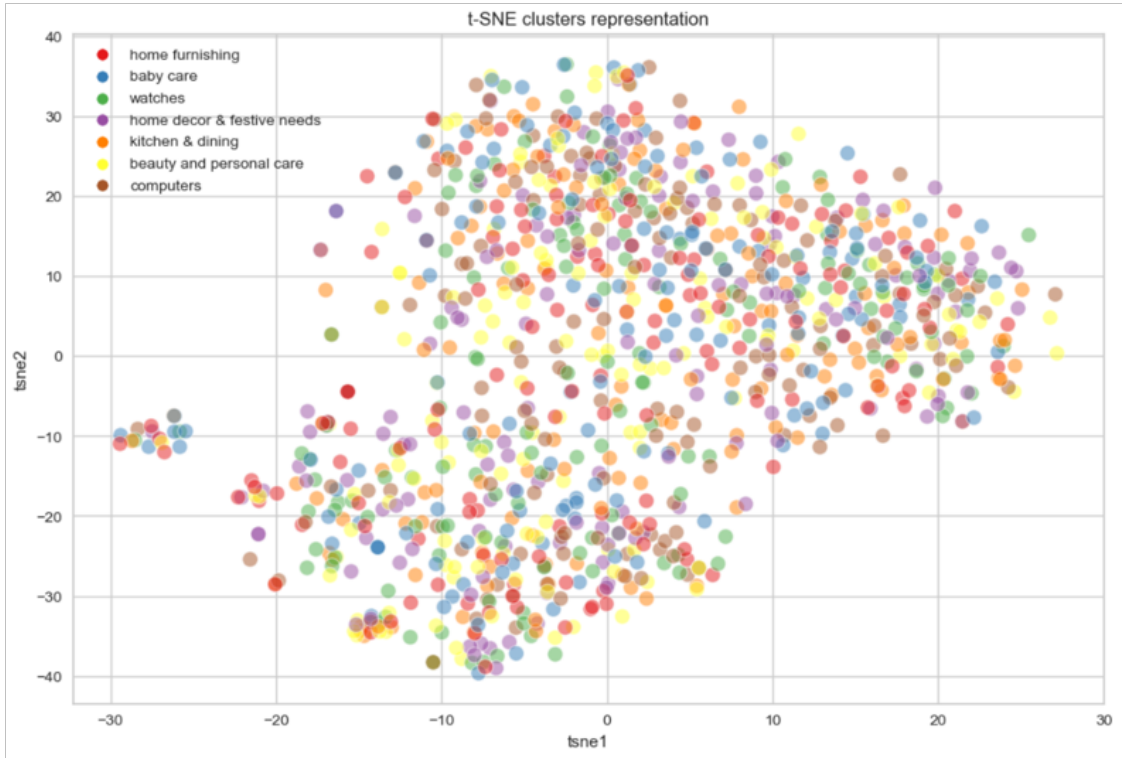


VISUALISATION DEUX DIMENSIONS (T-SNE)

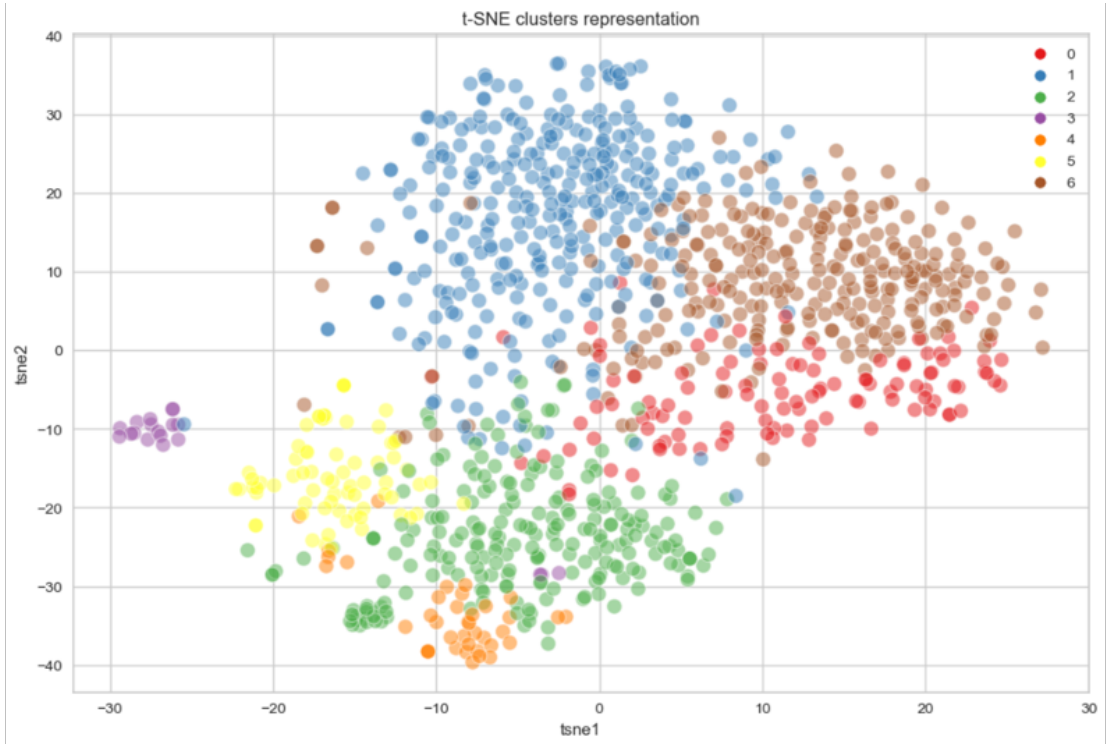
Dimension réduction par ACP

- Input = 721 dim
- Output = 580 dim
- 99% variance

Répartition des produits par catégorie



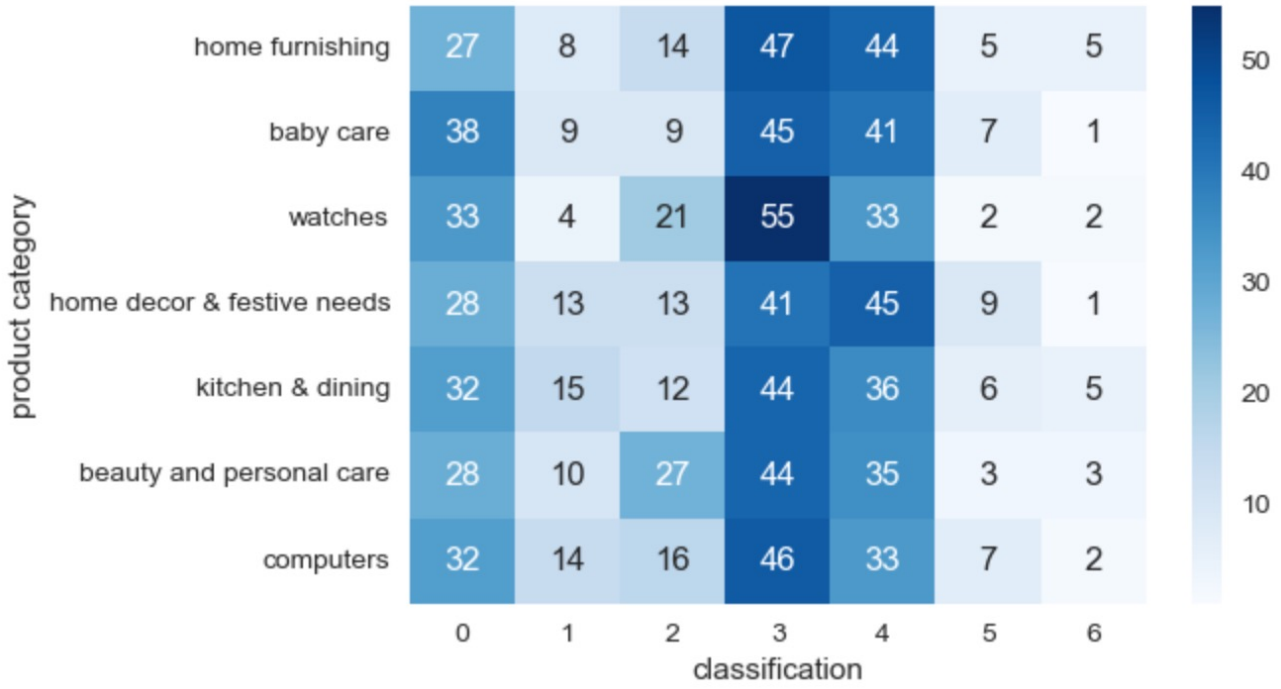
Répartition des produits par cluster



RÉSULTATS

Accuracy = 13% (après recalage)

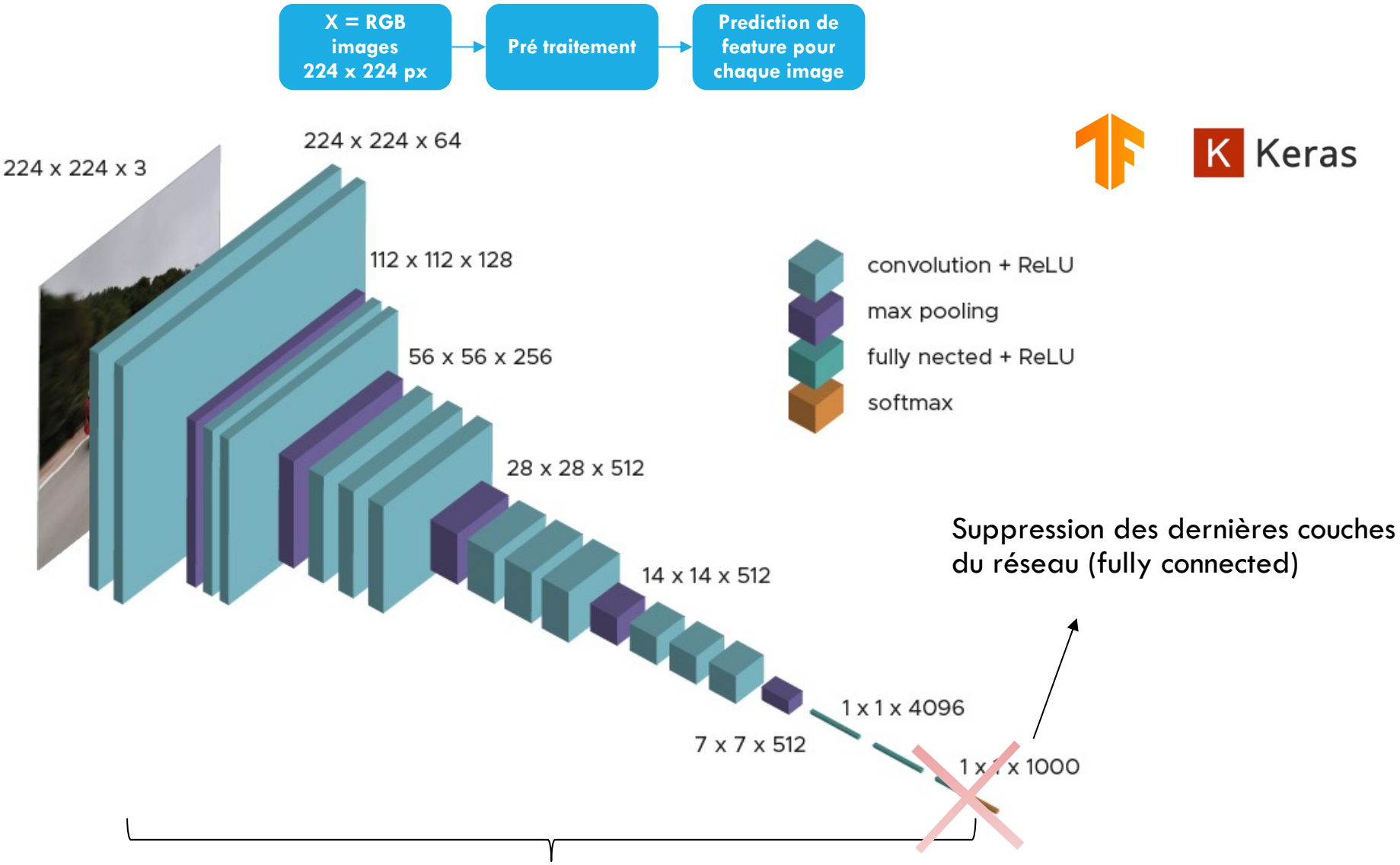
ARI = 0.015%



Conclusion :
La classification à partir de features extraitent grâce à l'algorithme ORB ne semble pas faisable.

-> Essai d'extraction de features grâce au transfert learning sur VGG16.

DEEP LEARNING AVEC UN CNN PRÉ-ENTRAINÉ : VGG16 (KERAS)

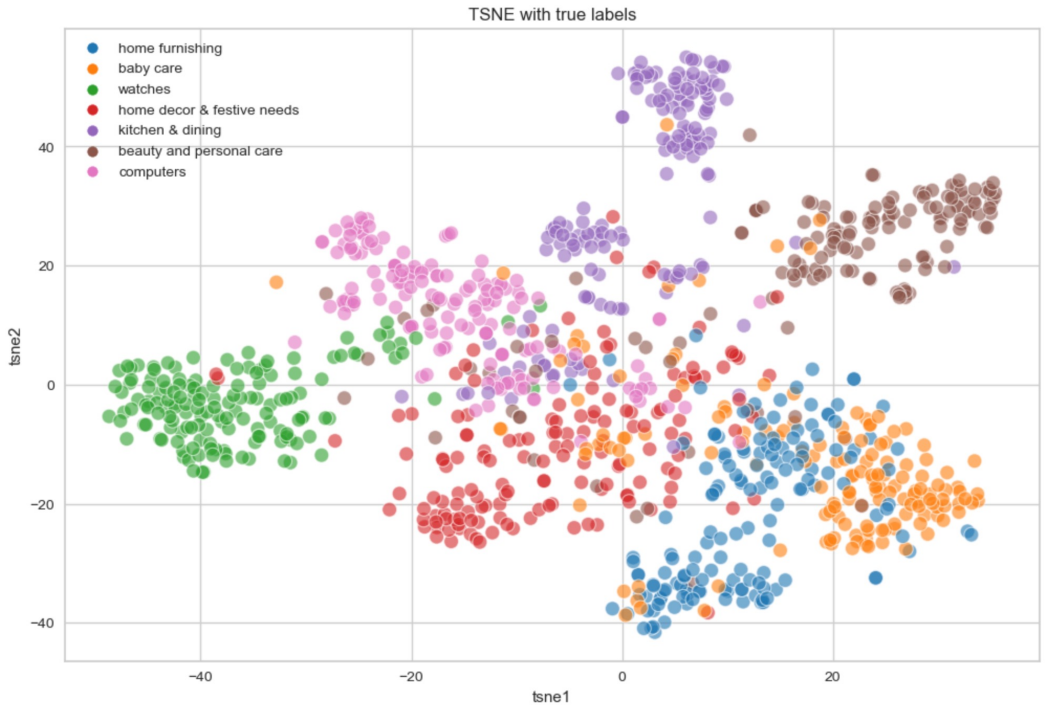


VISUALISATION DEUX DIMENSIONS (T-SNE)

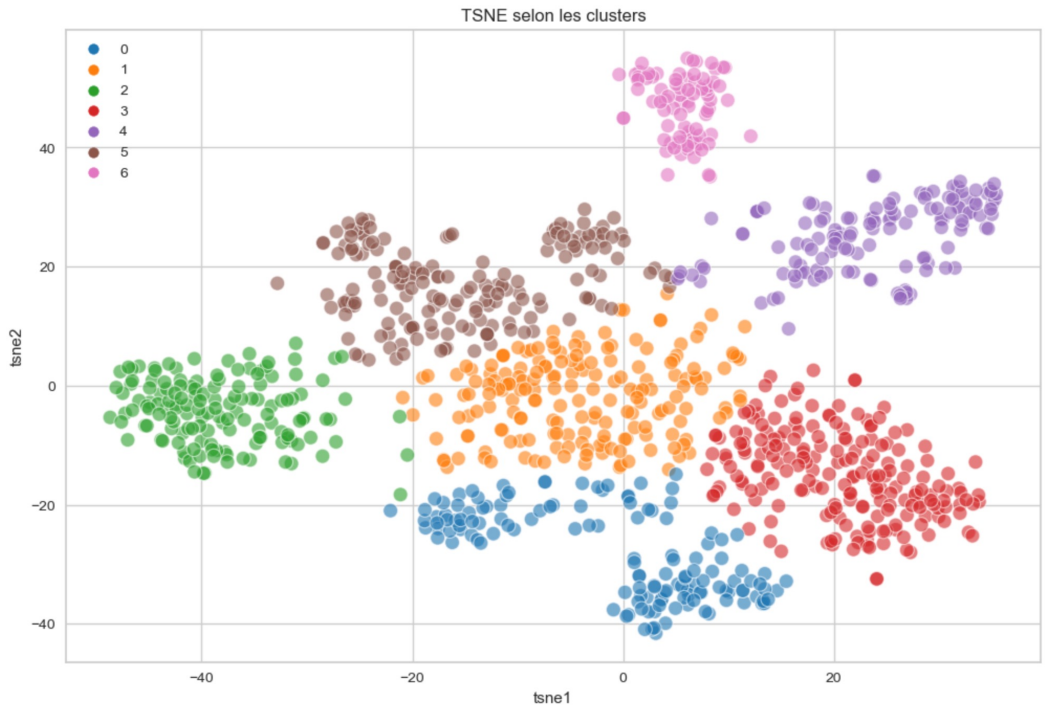
Dimension réduction par ACP

- Input = 4096 dim
- Output = 803 dim
- 99% variance

Répartition des produits par catégorie



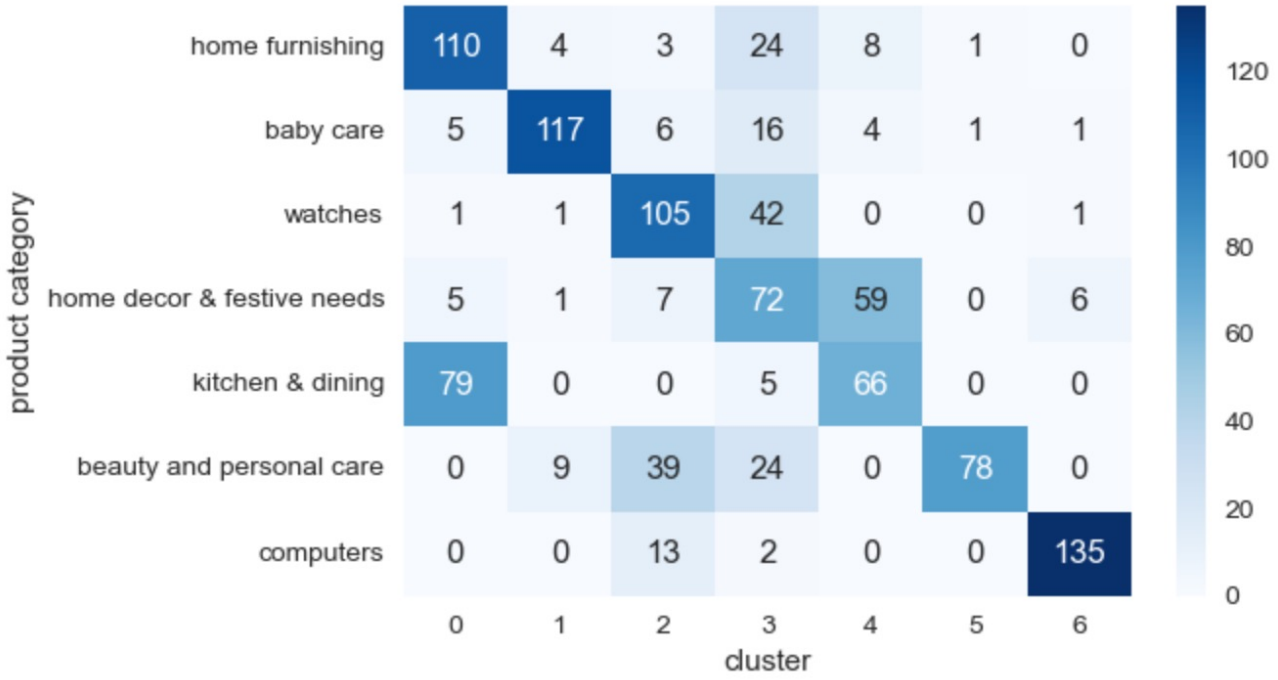
Répartition des produits par cluster



RÉSULTATS

Accuracy = 65% (après recalage)

ARI = 45%



Conclusion :
La classification à partir de features extraitent grâce au réseau de neuronne VGG16 pré entrainé permettent un bon clustering des produits.

Merci pour votre attention.