Soutenance Projet OC P6 DS: Classifiez automatiquement des biens de consommation

Candidat: David CAPELLE
Mentor: Nicolas MICHEL
Evaluateur: T.Ibrahima Diop

25/11/2021

Formation 100% Pôle Emploi

Plan de la soutenance

- Problématique et présentation de la démarche
- Présentation du jeu de données, du pré-traitement, méthodes encodage sur textes/images
- Présentation des clustering selon la méthode d'encodage
  - Synthèse sur les métriques de qualité du clustering
  - Choix de la meilleure méthode
  - Conclusion sur la faisabilité d'un moteur de classification
- Illustration par la mise en œuvre d'une classification SVM
- Conclusion finale, améliorations et limites

Problématique du projet

- Souhait de l'entreprise « Place de marché » de réaliser une première étude de faisabilité d'un moteur de classification en se basant sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article .

- Présentation de la démarche sur le pré-traitement, la création des features sur les données texte et image, la réduction des dimensions et le clustering.

- Présentation du résultat du clustering sous la forme d'une représentation en 2 dimensions afin de visualiser les regroupements de clusters pour les comparer à la représentation 2D de la vérité terrain .

Démarche pour l'étude de faisabilité

- Pour chaque méthode d'encodage (texte et image) :
  - Réduction de dimensions PCA et/ou modélisation NMF.
  - Clustering K-means à partir de la sortie PCA/NMF.
  - Visualisation T-SNE 2D sur la vérité terrain (catégories d'articles) et sur les labels des clusters K-means.
  - Calcul de l'ARI score entre les labels vérité terrain / clusters, calcul de score de silhouette du clustering.
- Analyse des résultats du clustering :
  - Analyse de la similitude des regroupement d'articles entre la vérité terrrain et les clusters (ARI score)
  - Analyse du coefficient de silhouette du clustering et taille des clusters

  => choix de la meilleure méthode d'encodage des features pour l'étude

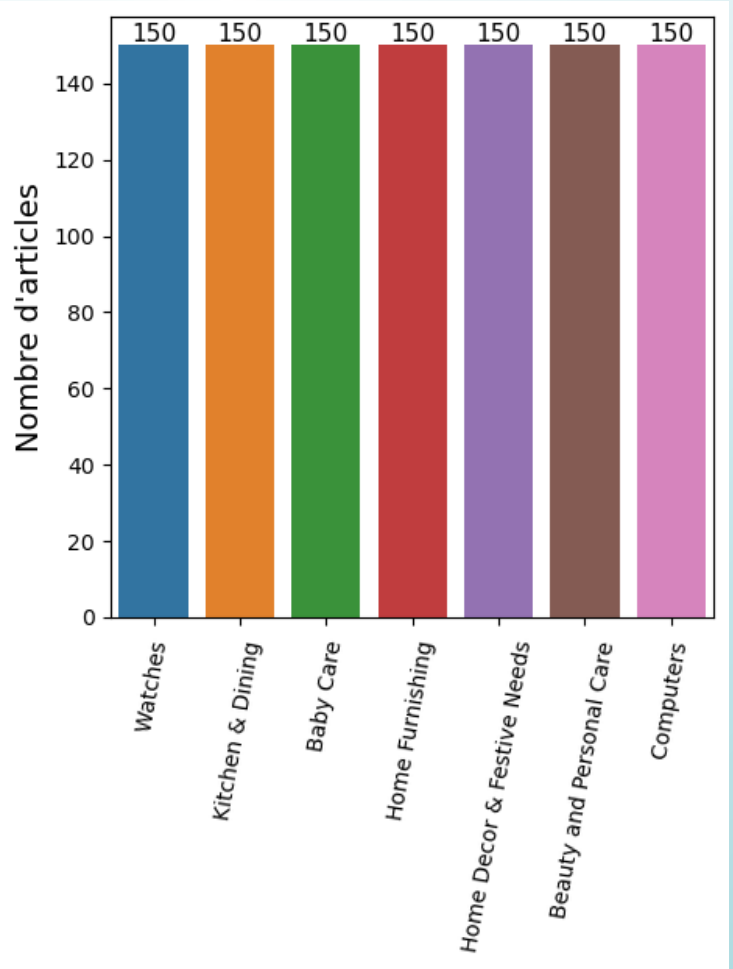
Présentation du jeu de données

- **Le jeu de données contient :**
  - 1050 lignes et 15 variables
  - correspondant aux articles vendus par l'entreprise avec leurs photos.

- **3 variables sont utilisées pour l'analyse :**
  - **product_category_tree :** arborescence complète décrivant la catégorie principale et sous-catégories des articles.
  - **description :** description de l'article ,sous forme d'une liste de mots décrivant l'article (corpus de textes). Cette variable fera l 'objet d'un pré-traitement pour extraction des features.
  - **image :** photos de l'articles

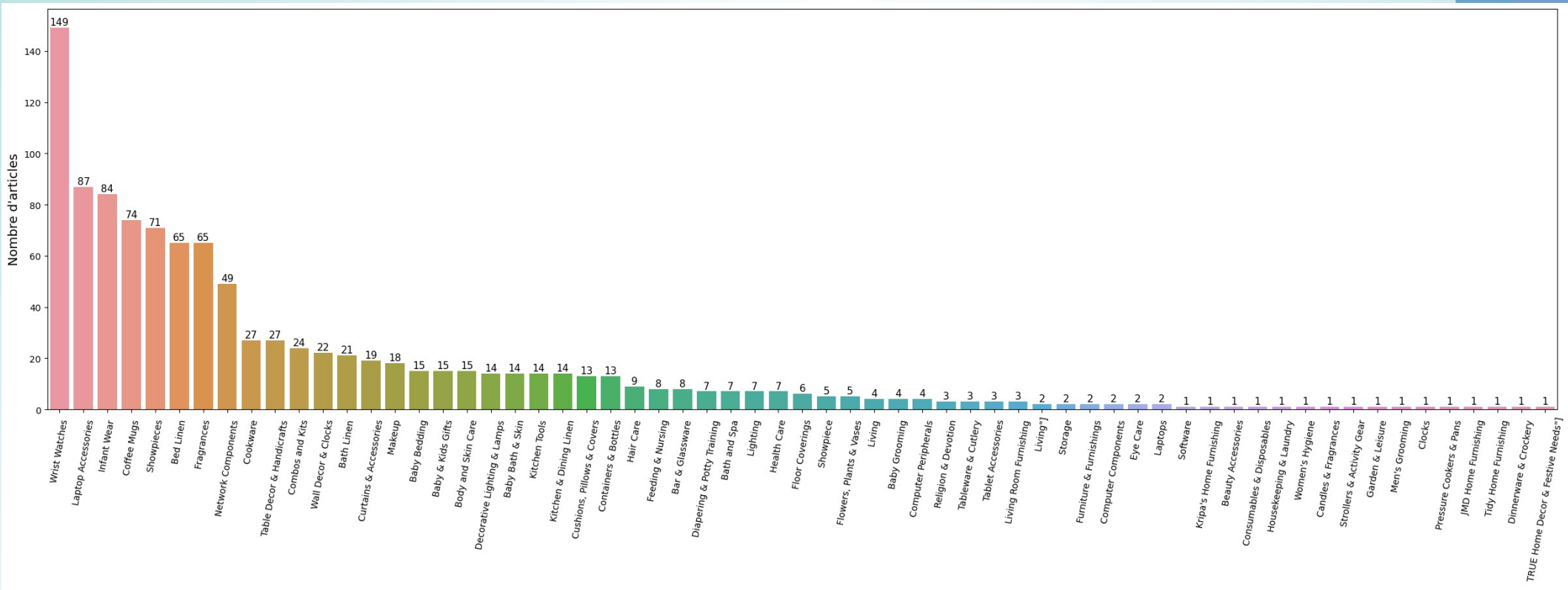
Analyse des données texte (1/2)

- **Le variable « product_category_tree » est décomposée en 2 variables :**
  - Variable « categ_princ » indiquant les 7 catégories principale d'un article



Analyse des données texte (2/2)

- **Le variable « product_category_tree » est décomposée en 2 variables :**
  - Variable « sous-categ » représentant les 63 sous-catégories d'articles de 2ème niveau.



Analyse des données image (1/3)

- Les images sont de tailles différentes en pixels
- Nécessité d'ajuster le contraste pour faciliter l'extraction des features
- On remarque que les images de certaines catégories d'articles sont facilement identifiables, mais d'autres non.
- Nécessité d'un prétraitement sur les images.

Analyse des données image (2/3)

- Images difficilement identifiables (catégorie « Home Decor & Festive Needs »)

Analyse des données image (3/3)

- Images facilement identifiables (catégorie « Watches »)

Texte : Pré-traitement et encodage du corpus de texte

- **Pré-traitements communs :**
  - Tokenisation du corpus de texte
  - Suppression des mots inutiles (« stopwords »)
  - Stemming
  - Lemmatisation
  - Limitation du nombre d'occurrences des mots, après lemmatisation

- **Encodage des features :**
  - Vectorisation Bag of Words (CounterVectorizer).

  => Taille vecteur encodage après pré-traitement = 1015

  - Vectorisation TF-IDF avec modèle N-gram (prise en compte des unigrammes et bigrammes).

  => Taille vecteur encodage après pré-traitement = 2346

  - Word / Sentence Embedding Doc2Vec

  => Taille vecteur encodage après pré-traitement = 1000.

Texte : Exemples de pré-traitement

Mise en minuscules du texte

- **Texte d'origine :**

'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstra
ct Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtai
n is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environm
ent romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a bright and

- **Mise en minuscules :**

'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain,elegance polyester multicolor abstra
ct eyelet door curtain (213 cm in height, pack of 2) price: rs. 899 this curtain enhances the look of the interiors.this curtai
n is made from 100% high quality polyester fabric.it features an eyelet style stitch with metal ring.it makes the room environm
ent romantic and loving.this curtain is ant- wrinkle and anti shrinkage and have elegant apparance.give your home a bright and

Texte : Exemples de pré-traitement

Suppression de caractères spéciaux / ponctuation

- **Texte en minuscules :**

'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain,elegance polyester multicolor abstra
ct eyelet door curtain (213 cm in height, pack of 2) price: rs. 899 this curtain enhances the look of the interiors.this curtai
n is made from 100% high quality polyester fabric.it features an eyelet style stitch with metal ring.it makes the room environm
ent romantic and loving.this curtain is ant- wrinkle and anti shrinkage and have elegant apparance.give your home a bright and

- **Suppression caractères spéciaux / ponctuations :**

'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstra
ct eyelet door curtain 213 cm in height pack of 2 price rs 899 this curtain enhances the look of the interiors this curtai
n is made from 100 high quality polyester fabric it features an eyelet style stitch with metal ring it makes the room environm
ent romantic and loving this curtain is ant wrinkle and anti shrinkage and have elegant apparance give your home a bright and

Texte : Exemples de pré-traitement

Suppression des mots contenant ou correspondant à des chiffres

- **Texte sans caractère spécial / ponctuation :**

```
'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstra
ct eyelet door curtain  213 cm in height  pack of 2  price  rs  899 this curtain enhances the look of the interiors this curtai
n is made from 100  high quality polyester fabric it features an eyelet style stitch with metal ring it makes the room environm
ent romantic and loving this curtain is ant  wrinkle and anti shrinkage and have elegant apparance give your home a bright and
```

- **Suppression des chiffres :**

```
'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstra
ct eyelet door curtain   cm in height  pack of   price  rs   this curtain enhances the look of the interiors this curtain is ma
de from   high quality polyester fabric it features an eyelet style stitch with metal ring it makes the room environment romant
ic and loving this curtain is ant  wrinkle and anti shrinkage and have elegant apparance give your home a bright and modernisti
```

Texte : Exemples de pré-traitement

Tokenisation / Suppression des mot inutiles

- **Texte tokenisé :**

```
'key features of elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstra
ct eyelet door curtain cm in height pack of price rs this curtain enhances the look of the interiors this curtain is made from
high quality polyester fabric it features an eyelet style stitch with metal ring it makes the room environment romantic and lov
ing this curtain is ant wrinkle and anti shrinkage and have elegant apparance give your home a bright and modernistic appeal wi
```

- **Texte sans la mots inutiles (« stopwords ») :**

```
'key features elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract
eyelet door curtain pack price curtain enhances look interiors curtain made high quality polyester fabric features eyelet style
stitch metal ring makes room environment romantic loving curtain ant wrinkle anti shrinkage elegant apparance give home bright
modernistic appeal designs surreal attention sure steal hearts contemporary eyelet valance curtains slide smoothly draw apart f
```

Texte : Exemples de pré-traitement

Stemming / Lemmatisation

- **Texte avec stemming :**

'key featur eleg polyest multicolor abstract eyelet door curtain floral curtain eleg polyest multicolor abstract eyelet door cu
rtain pack price curtain enhanc look interior curtain made high qualiti polyest fabric featur eyelet style stitch metal ring ma
ke room environ romant love curtain ant wrinkl anti shrinkag eleg appar give home bright modernist appeal design surreal attent
sure steal heart contemporari eyelet valanc curtain slide smoothli draw apart first thing morn welcom bright sun ray want wish

- **Texte avec lemmatisation :**

'key feature elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract e
yelet door curtain pack price curtain enhances look interior curtain make high quality polyester fabric feature eyelet style st
itch metal ring make room environment romantic love curtain ant wrinkle anti shrinkage elegant apparance give home bright moder
nistic appeal design surreal attention sure steal heart contemporary eyelet valance curtain slide smoothly draw apart first thi

Image : Pré-traitements et encodage des images

- **Bag of Visual Words avec algorithme SIFT**
  - Pré-traitement enregistrement des images en niveaux de gris,
  - Puis, réduction du bruit par filtrage gaussien,
  - Encodage des images avec SIFT

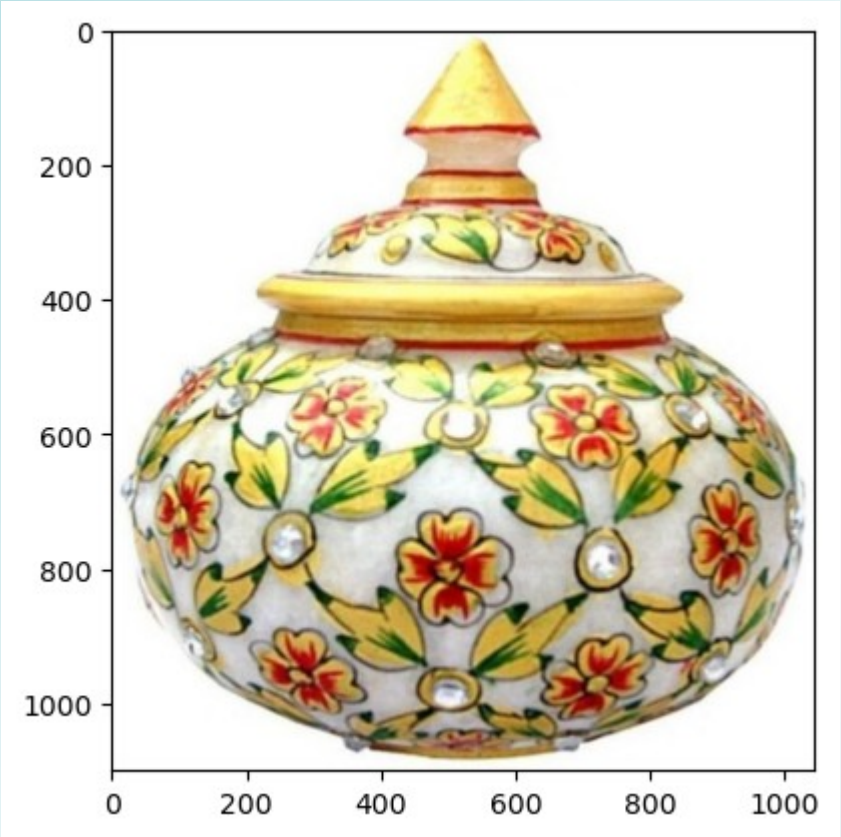
  => Création de 506398 descripteurs et création d'un mot visuel de dimension 712

- **Encodage à partir des réseaux de neurones CNN pré-entraîné VGG16 et ResNet50 :**
  - Pré-traitement redimensionnement des images couleur en 224*224 pixels,
  - Puis, amélioration du contraste
  - Encodage des images (suppression des 3 couches fully-connected qui permettent de classifier les images avec ImageNet pour VGG16).

  => Dimension du vecteur par image VGG16 = 25088

  => Dimension du vecteur par image ResNet50 = 100352

Image : Pré-traitements – Image d'origine / niveau de gris



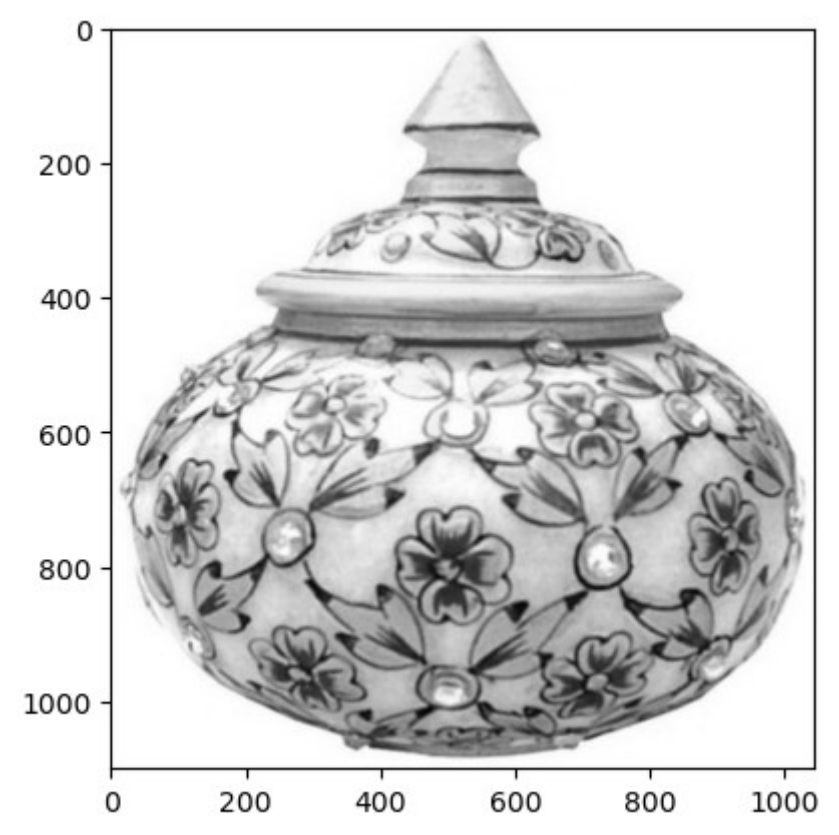
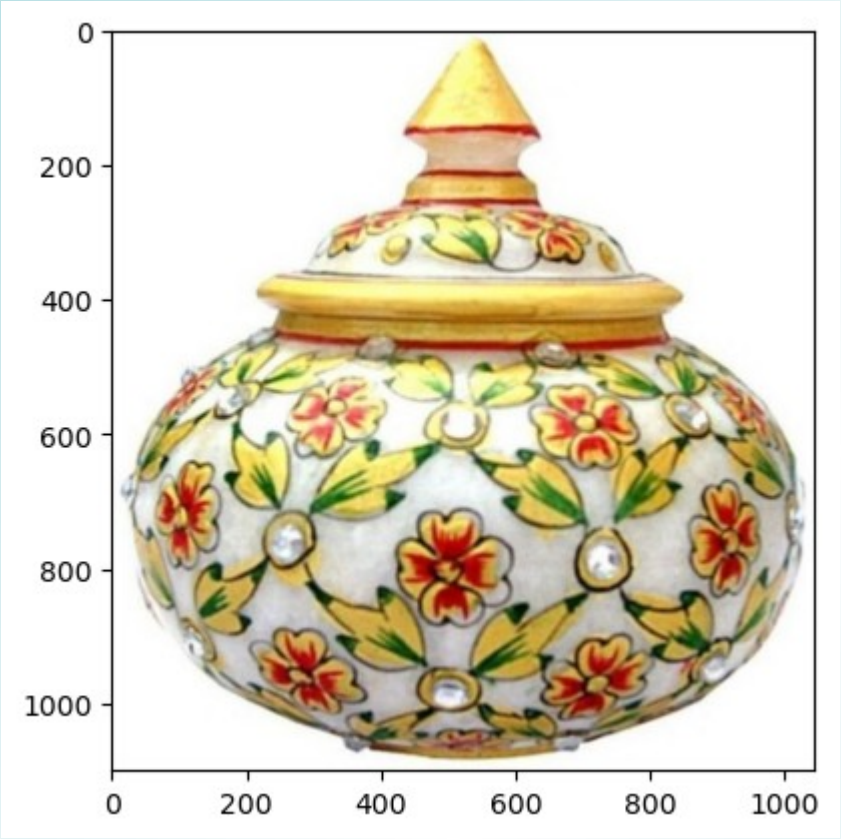
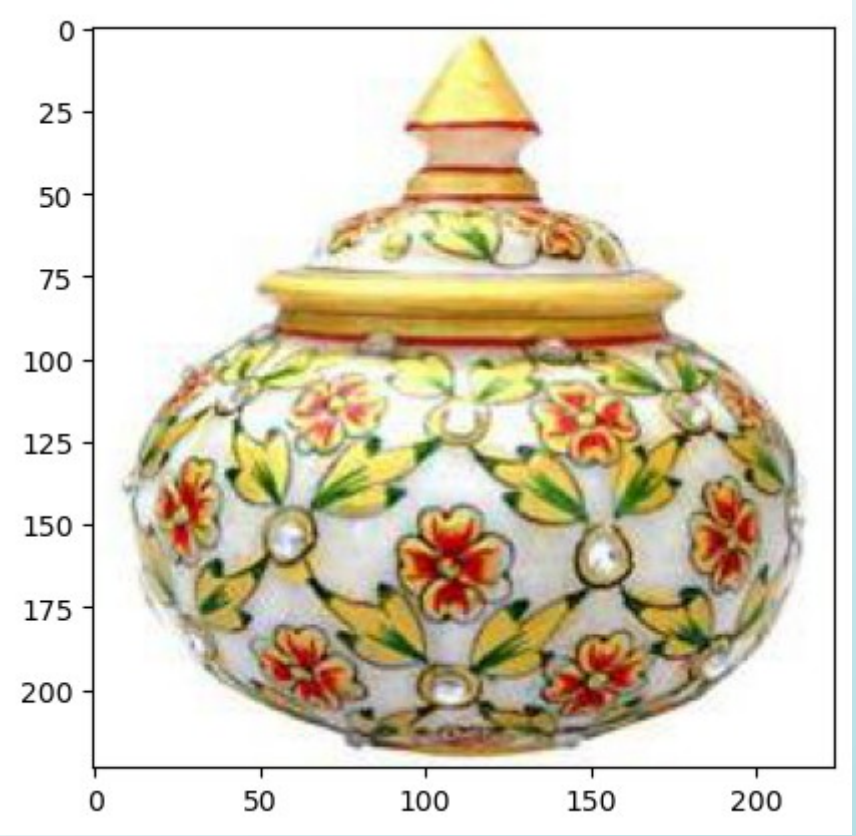


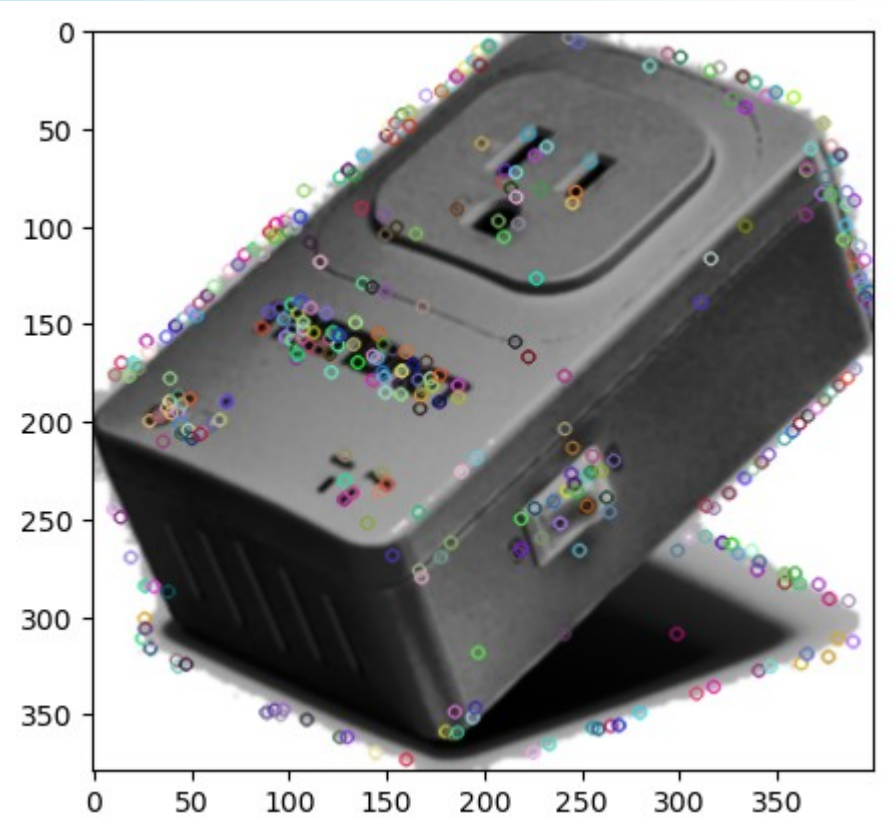
Image : Pré-traitements – Image d'origine / Redimensionnement 224*224 pixels





Passage d'une taille d'image de 1100x1044 pixels à 224x224 pixels en conservant le ratio hauteur / largeur.

Image : Encodage SIFT / Affichage des descripteurs



Affichage des 500 descripteurs de l’image en niveaux de gris

Etude de faisabilité – Textes (description)
Hypothèses d'encodage testées en entrée du clustering

- **Encodage textes avec Bag of Words (CounterVectorizer) :**
  - Bag of Words avec racinisation et PCA
  - Bag of Words avec lemmatisation et PCA
  - Bag of Words avec lemmatisation, nbre occurrences de mots réduit et PCA /NMF

- **Encodage textes avec vectorisation TF-IDF :**
  - TF-IDF avec racinisation et PCA
  - TF-IDF avec lemmatisation et PCA
  - TF-IDF avec lemmatisation, nbre d'occurrences de mots réduit et PCA
  - TF-IDF avec lemmatisation, nbre d'occurrences de mots réduit, modèle N-gram et PCA / NMF

- **Encodage Word / Sentence embedding Doc2Vec et PCA/NMF**

Etude de faisabilité - Images

Hypothèses d'encodage testées en entrée du clustering

- **Encodage images avec Bag of Visual Words - SIFT :**
  - Images en niveaux de gris, réduction du bruit par filtre gaussien, égalisation de l'histogramme
  - Réduction de dimensions PCA / NMF

- **Encodage avec un réseau de neurones pré-entraîné VGG16 / ResNet50 :**
  - Images redimensionnées en 224x224 pixels
  - Amélioration du contraste
  - Réduction de dimensions PCA / sans PCA / NMF

Etude de faisabilité – Textes + Images
Hypothèses d'encodage testées en entrée du clustering

- **Encodage textes + images TF-IDF et ResNet50 :**
  - Concaténation de 2 encodages textes et images **(1)** et **(2)**
  - **(1)** Vectorisation TF-IDF avec lemmatisation, limitation du nbre d'occurrences de mots, modèle N-gram
  - **(2)** extraction de features avec un réseau de neurones ResNet50
  - Modélisation et compression de features NMF

Etude de faisabilité – Textes - Bag of Words et PCA

Synthèse des résultats du clustering

- **Mauvais résultats sur le ARI score (entre 0,05 et 0,06)**

=> pas de similarité entre les catégories d'articles (vérité du terrain) et les clusters

- **Résultats moyens sur le score de silhouette du clustering (entre 0,27 et 0,28).**

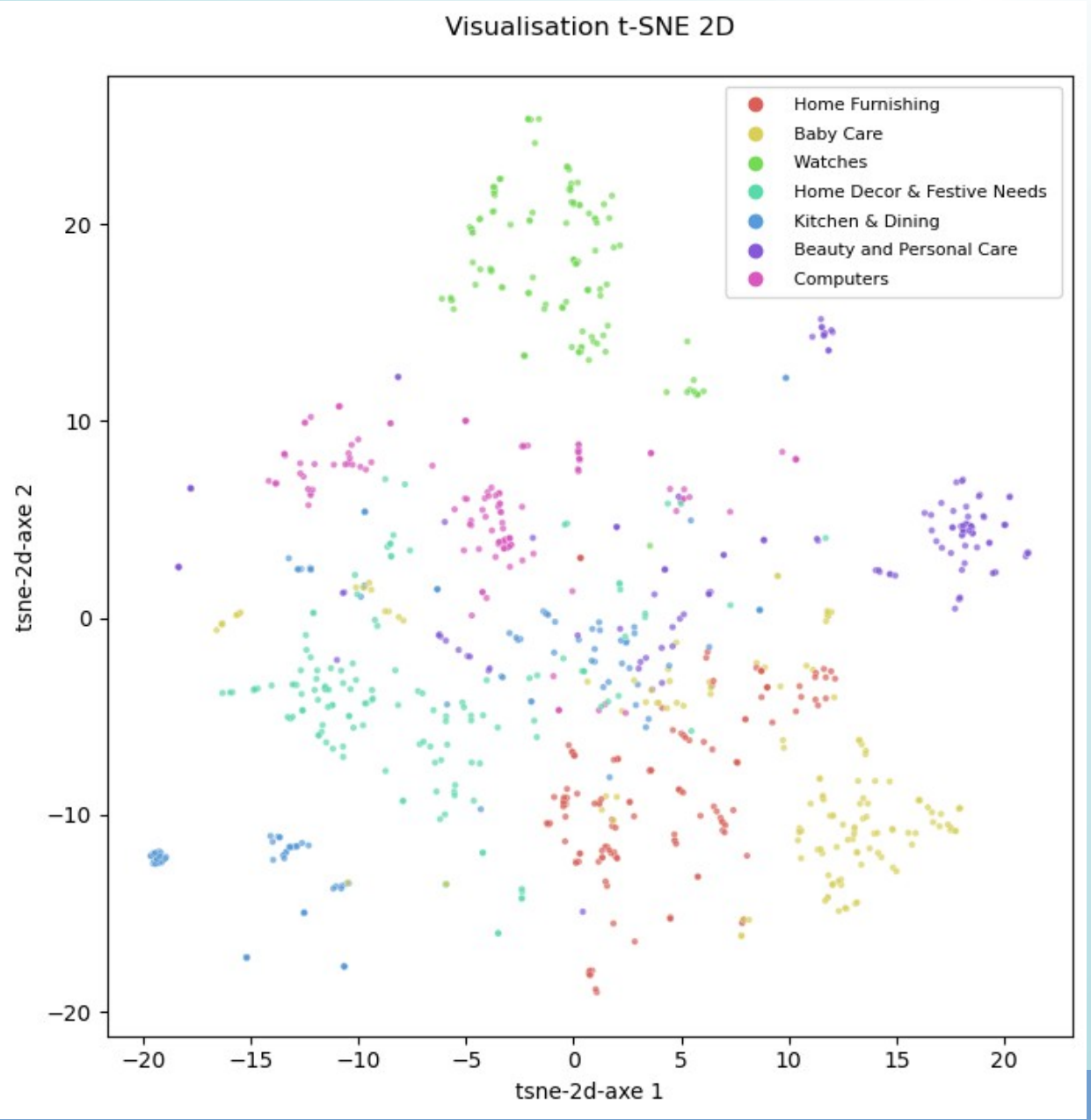
- La visualisation T-SNE 2D des labels de clusters :
  - montre des regroupements de données pas meilleurs que sur la visualisation T-SNE 2D de la vérité terrain.
  - des clusters disproportionnés en taille

- La visualisation des coefficient de silhouette par clusters montre des erreurs sur les clusters et des clusters non homogènes en taille

- Enfin, la projection de la matrice de confusion montre que les regroupements de clusters ne correspondent pas aux étiquettes de la vérité terrain.

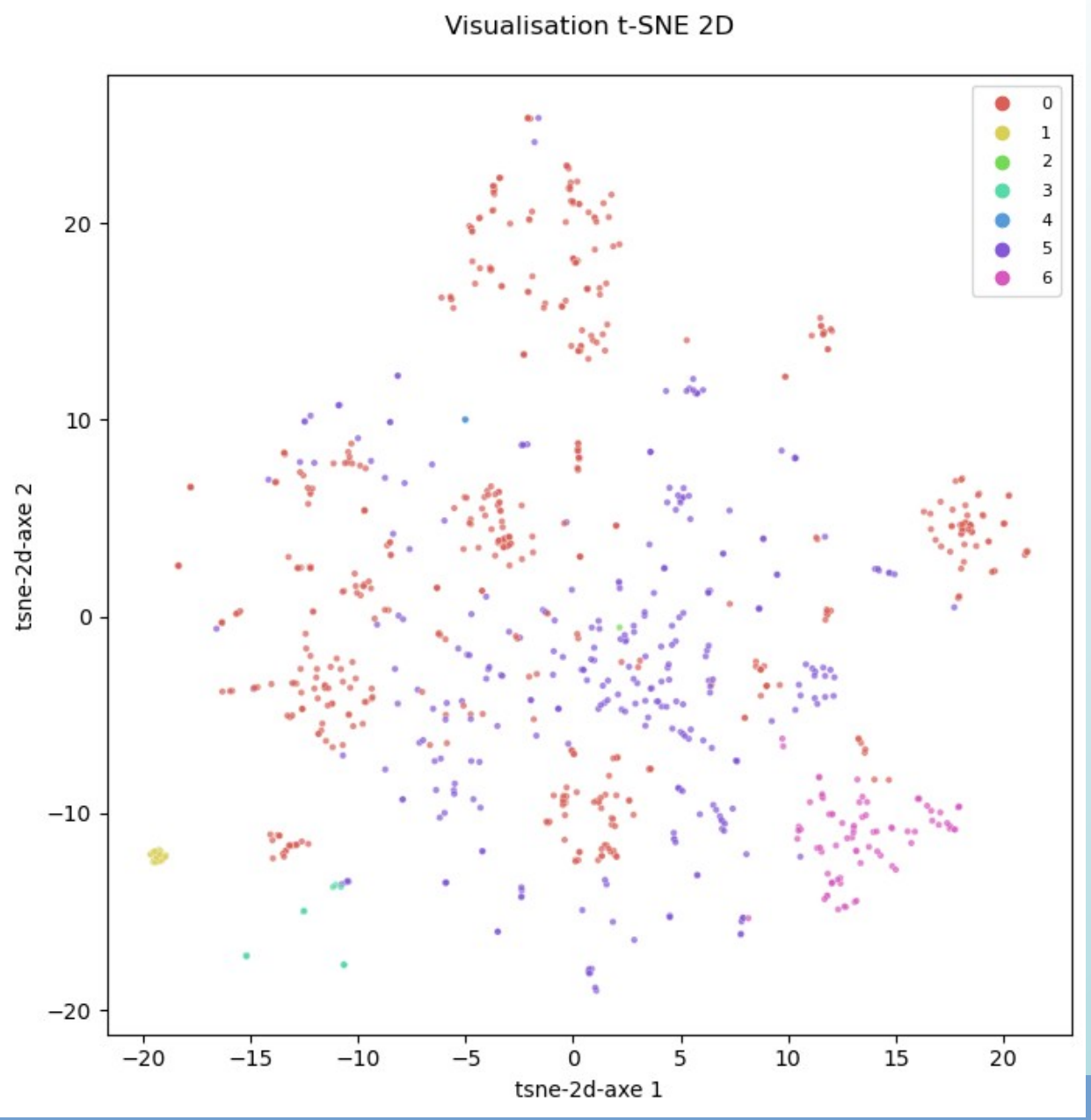
Etude de faisabilité – Textes – T-SNE sur vérité terrain

Cas Bag of Words avec lemmatisation et limitation occur. mots



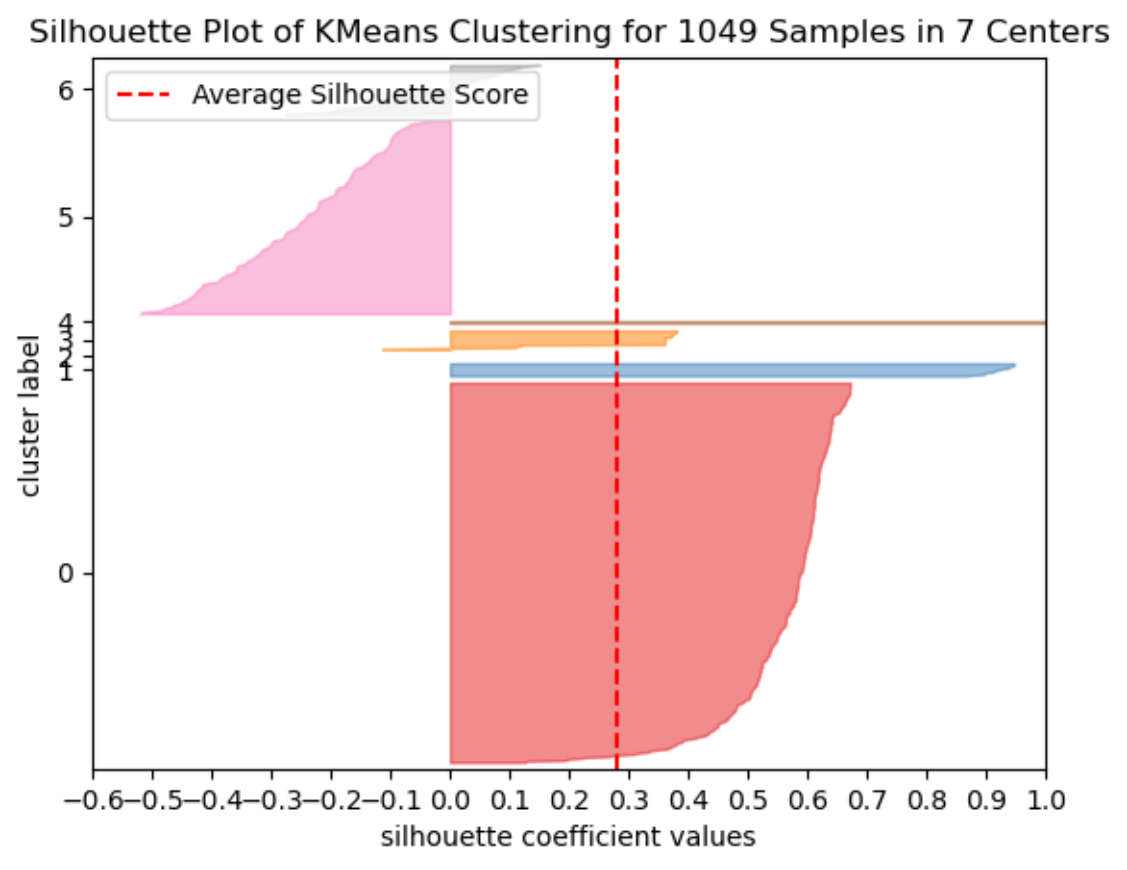
Etude de faisabilité – Textes – T-SNE clusters k-means

Cas Bag of Words avec lemmatisation et limitation occur. mots



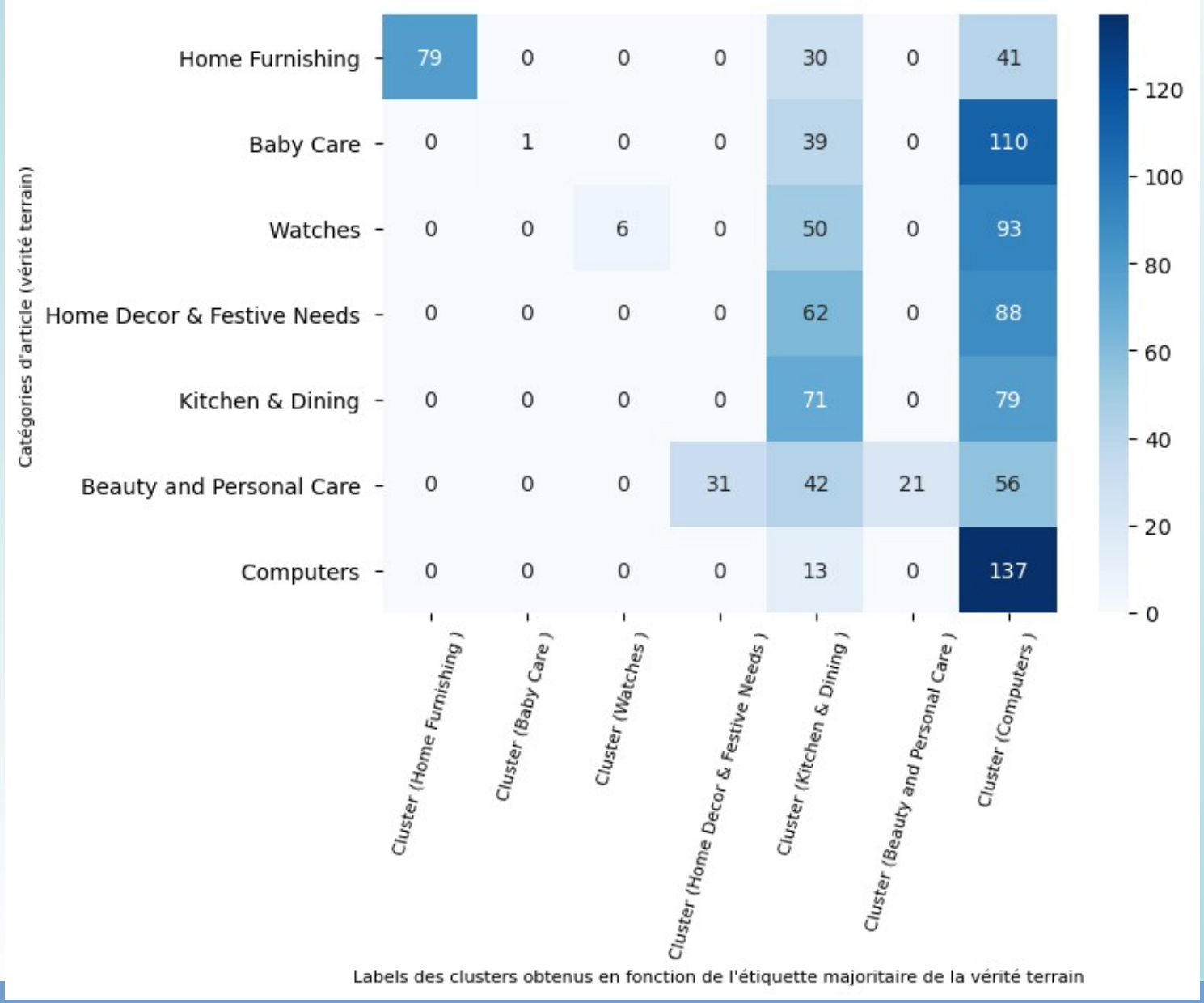
Etude de faisabilité – Textes – Coeff. de silhouette

Cas Bag of Words avec lemmatisation et limitation occur. mots



Etude de faisabilité – Textes – Matrice de confusion

Cas Bag of Words avec lemmatisation et limitation occur. mots



Etude de faisabilité – Textes - TF-IDF et PCA

Synthèse des résultats du clustering

- **Résultats médiocres sur le ARI score (entre 0,23 et 0,25)**

- **Résultats mauvais sur le score de silhouette du clustering (entre 0,05 et 0,1),** avec des erreurs sur données des clusters, clusters se chevauchant et/ou non denses.

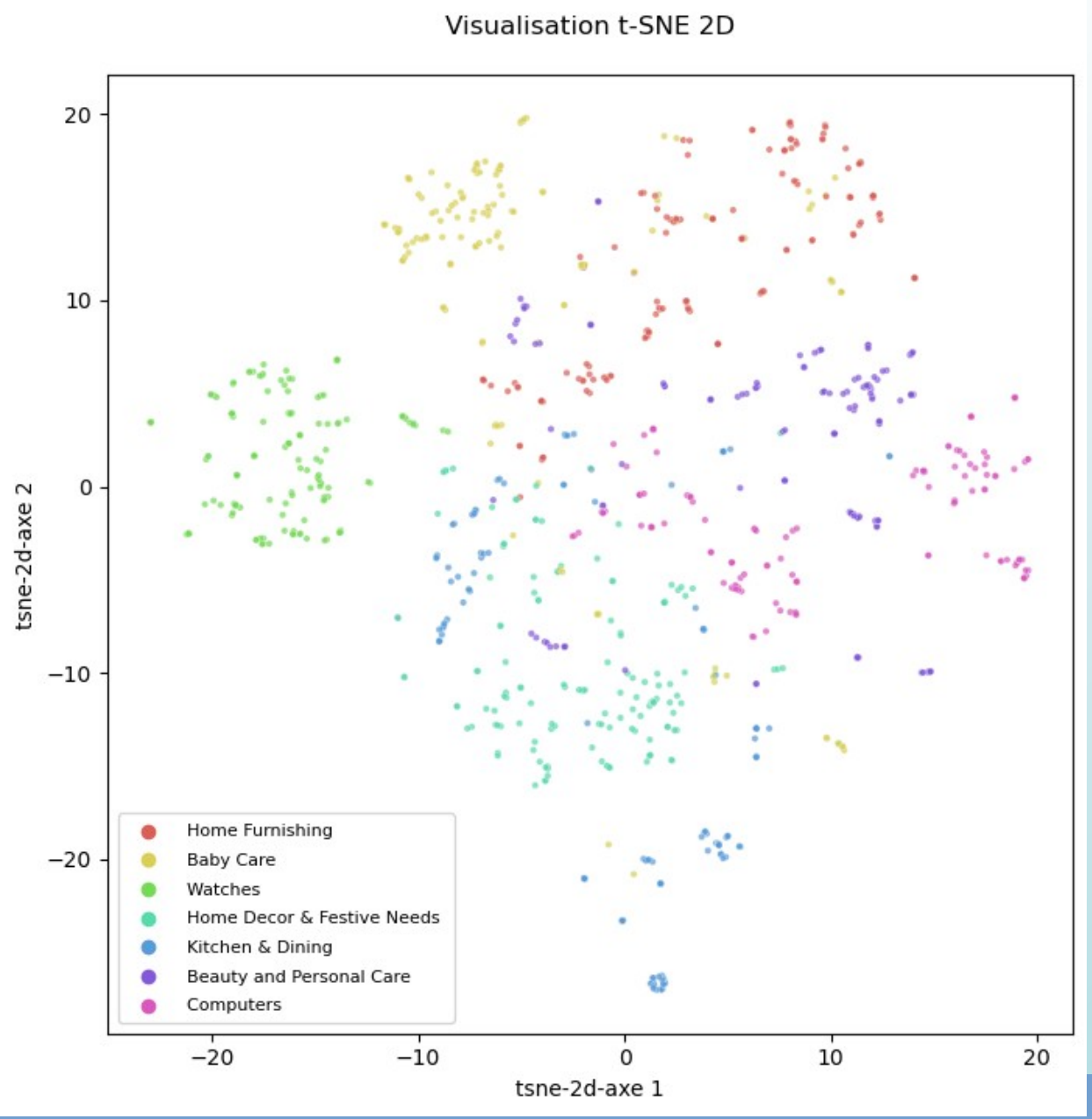
- La visualisation T-SNE 2D des labels de clusters :
  - montre un meilleur regroupement des données que sur la visualisation T-SNE 2D de la vérité terrain.
  - des clusters moins disproportionnés en taille que sur le clustering BOW

- La visualisation des coefficients de silhouette montre qu'il y a moins erreurs sur les clusters et clusters plus homogènes en taille (par rapport au clustering BOW).

- Enfin, la projection de la matrice de confusion montre que les regroupements de clusters ne correspondent pas aux étiquettes de la vérité terrain.

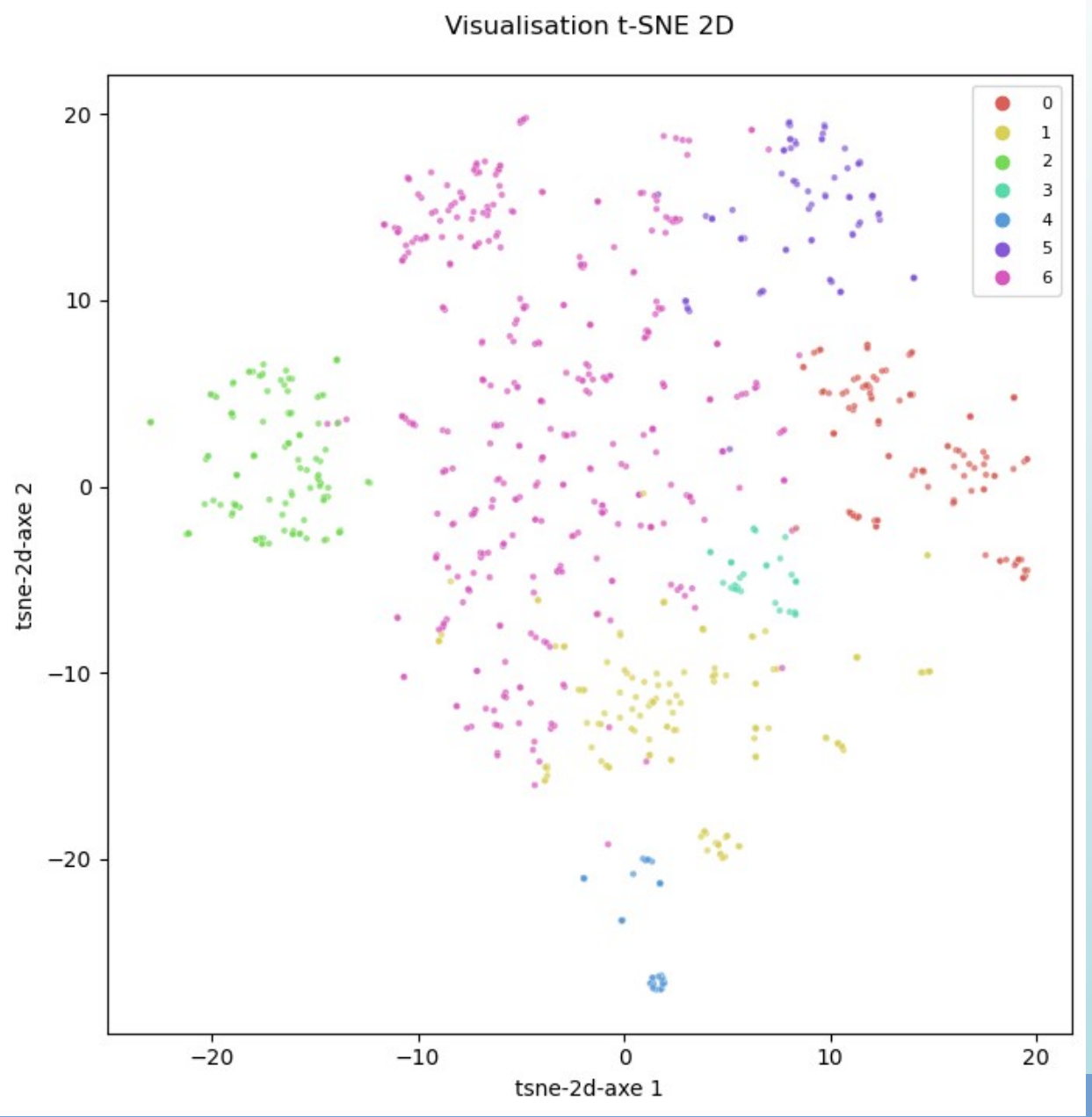
Etude de faisabilité – Textes – T-SNE sur vérité terrain

Cas TF-IDF, lemmatisation, limitation occur. mots, modèle N-gram



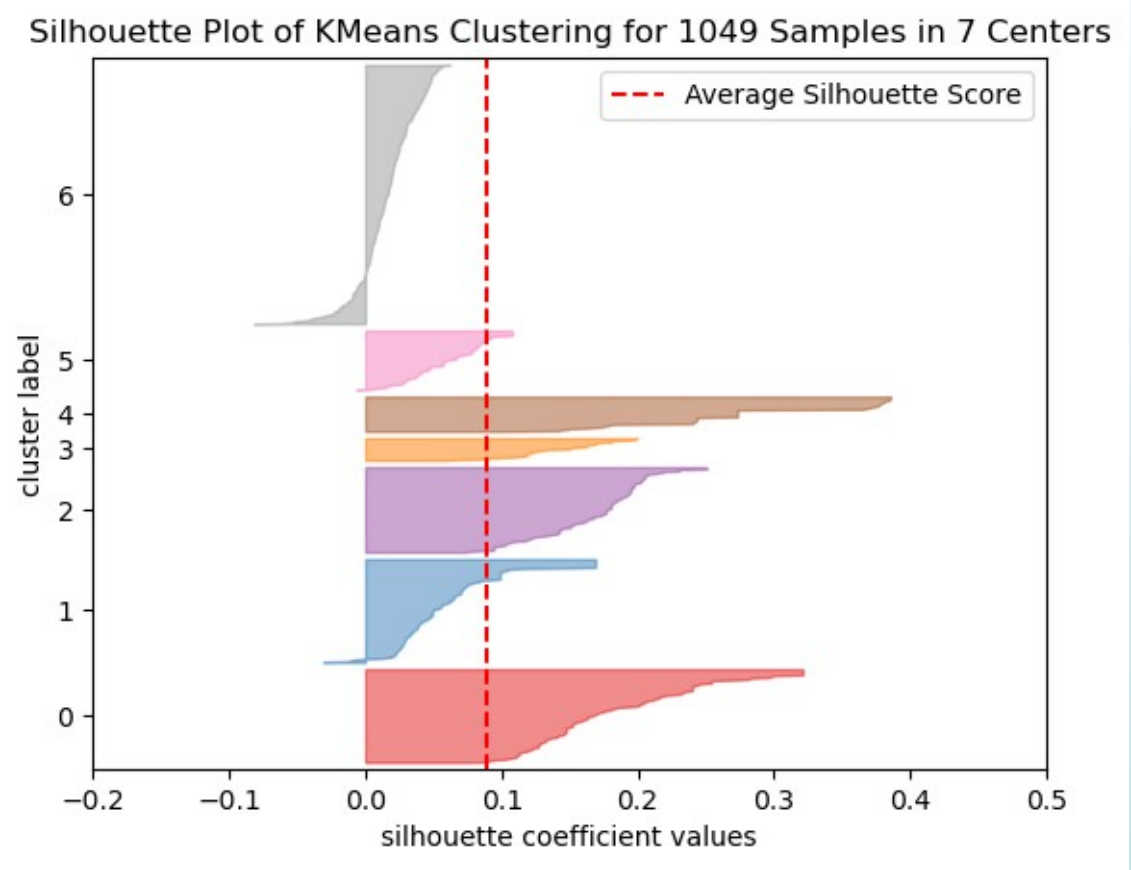
Etude de faisabilité – Textes – T-SNE clusters k-means

Cas TF-IDF, lemmatisation, limitation occur. mots, modèle N-gram



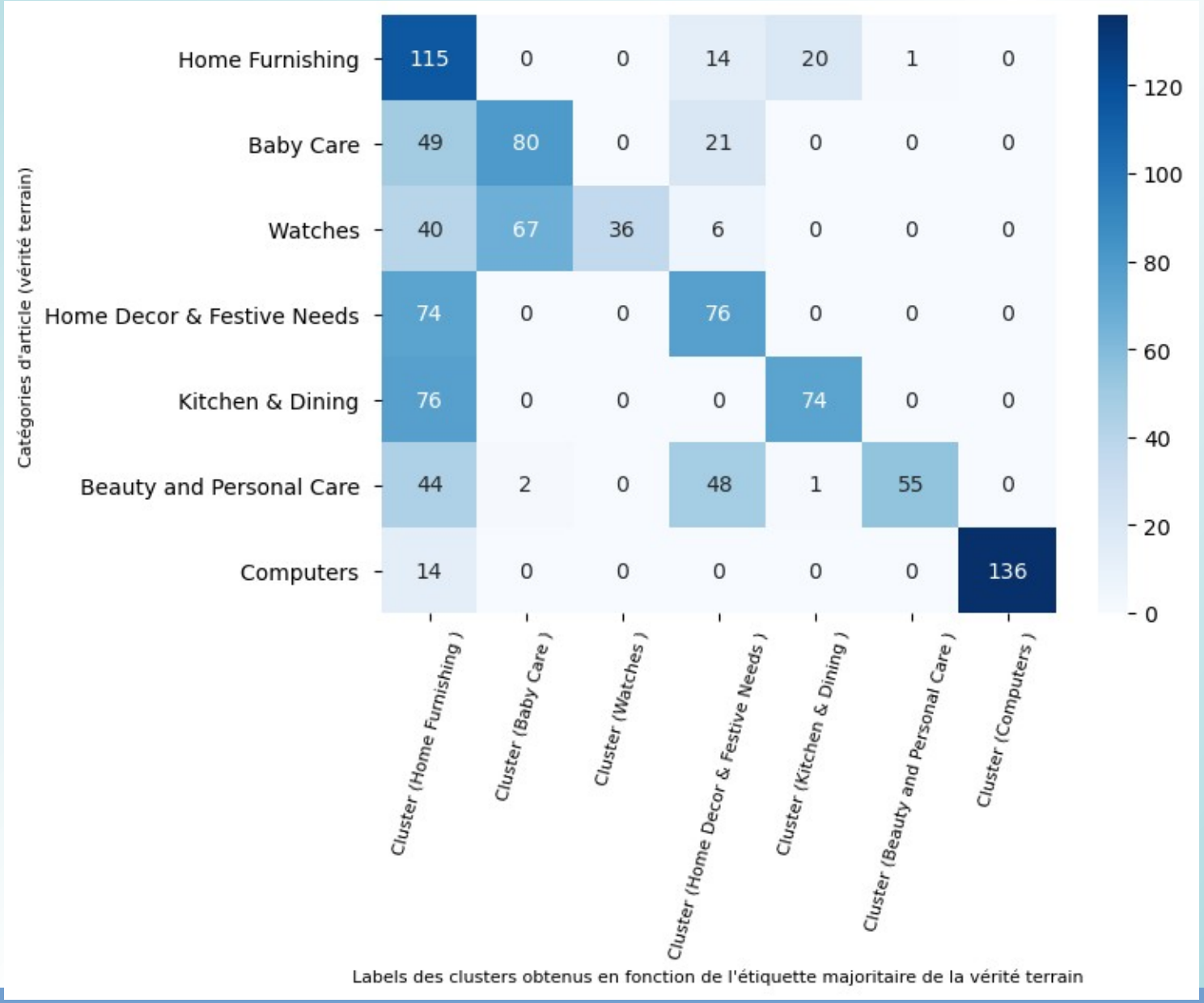
Etude de faisabilité – Textes – Coeff. de silhouette

Cas TF-IDF, lemmatisation, limitation occur. mots, modèle N-gram



Etude de faisabilité – Textes – Matrice de confusion

Cas TF-IDF, lemmatisation, limitation occur. mots, modèle N-gram



Etude de faisabilité – Images - Encodage SIFT et PCA

Synthèse des résultats du clustering

- **Résultats très mauvais sur le ARI score (0,02)**

- **Résultats très mauvais sur le score de silhouette du clustering (0,07),** indiquant des erreurs sur les données appartenant aux clusters, avec des clusters se chevauchant et non denses.

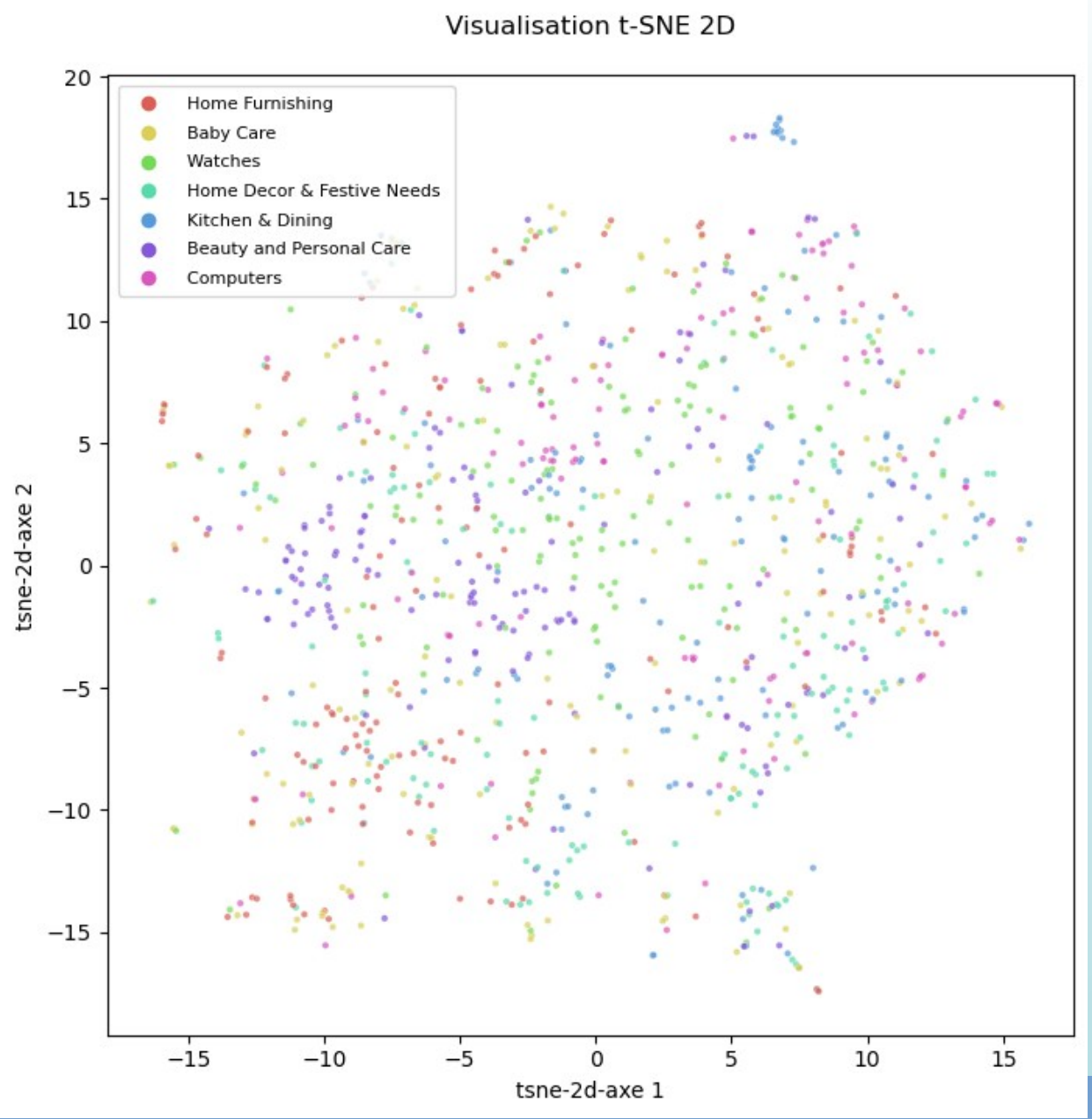
- La visualisation T-SNE 2D des labels de clusters :
  - montre des meilleurs regroupements de données, plus séparés que sur la visualisation T-SNE 2D de la vérité terrain.
  - des clusters très disproportionnés en taille

- La visualisation des coefficient de silhouette par clusters montre qu'il y a beaucoup d'erreurs sur les clusters et des clusters très petits en nombre d'articles.

- Enfin, la projection de la matrice de confusion montre que les regroupements de clusters ne correspondent pas aux étiquettes de la vérité terrain.

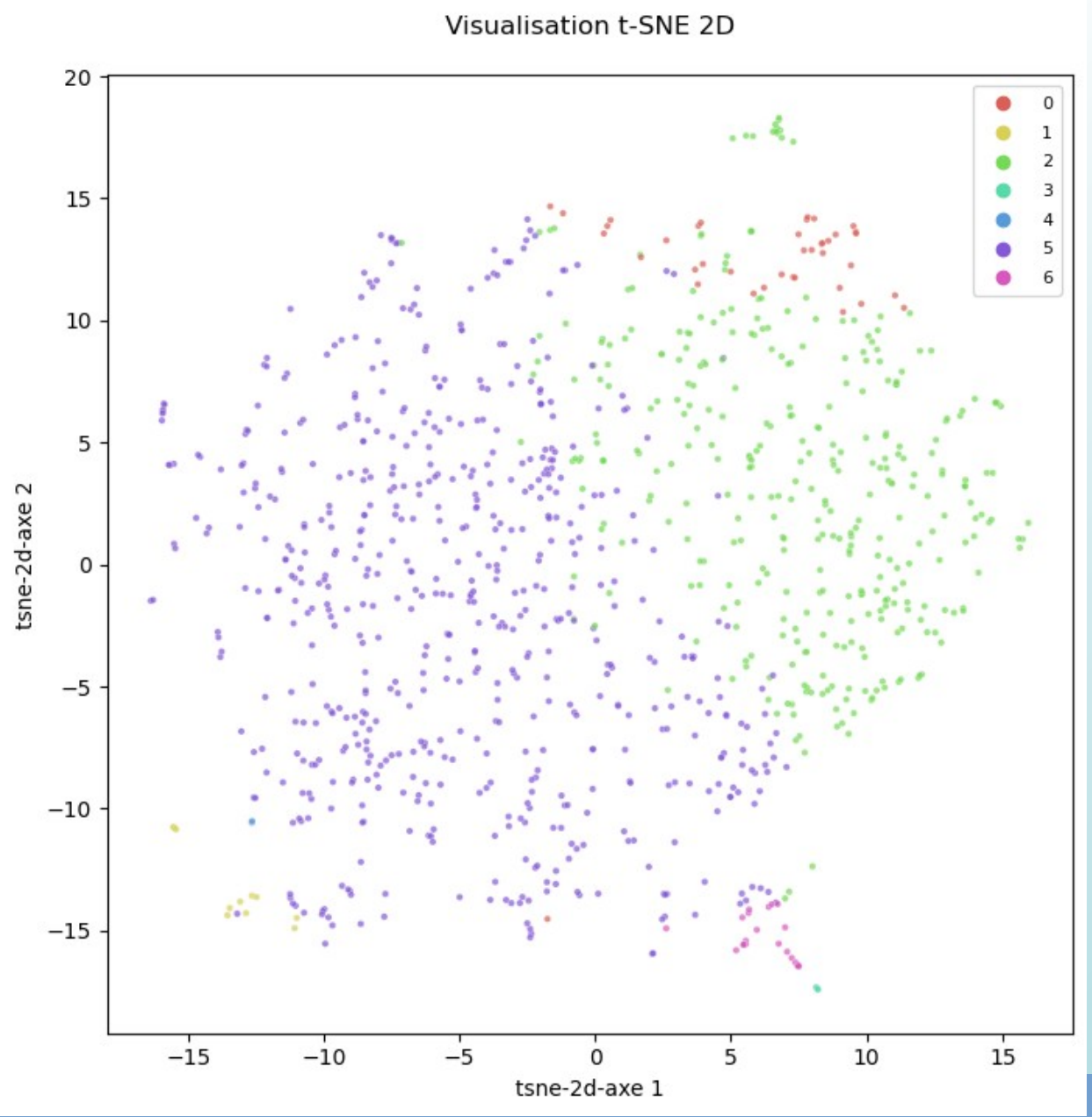
Etude de faisabilité – Images – T-SNE sur vérité terrain

Cas encodage SIFT et PCA



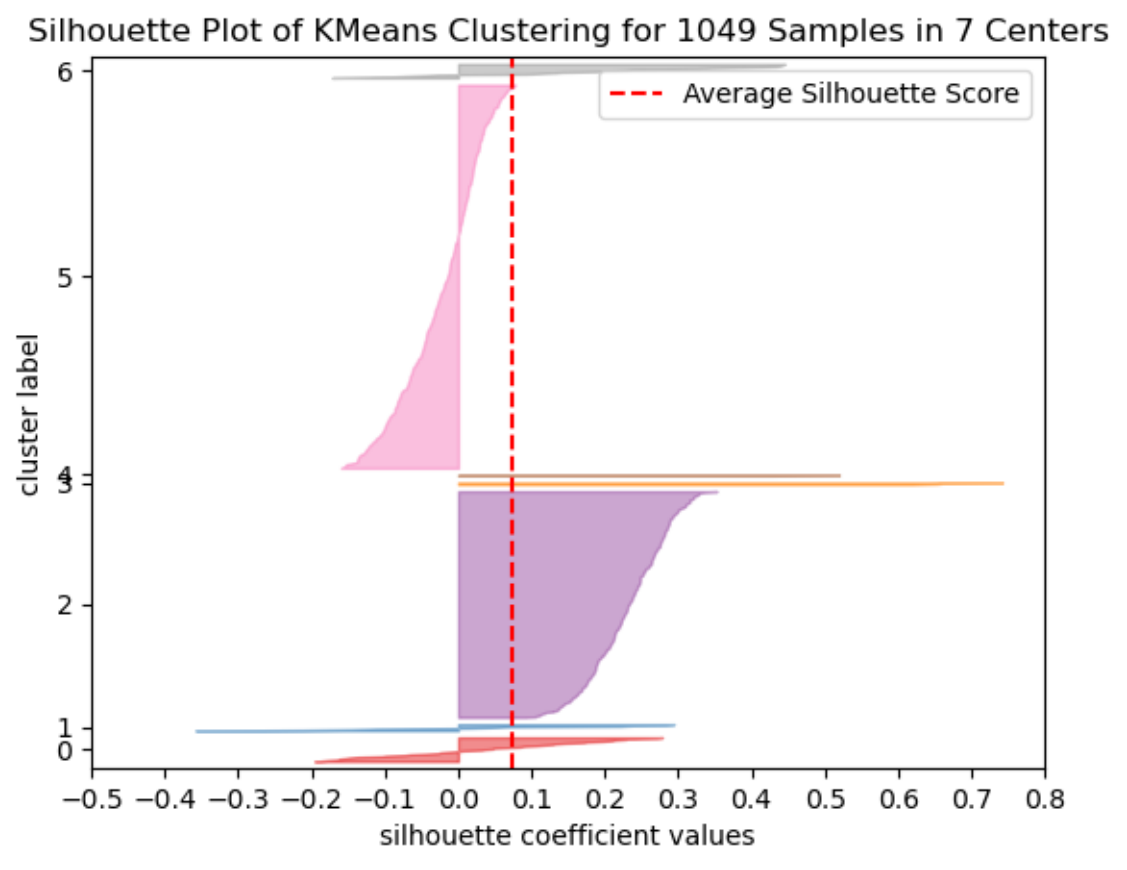
Etude de faisabilité – Images – T-SNE clusters k-means

Cas encodage SIFT et PCA



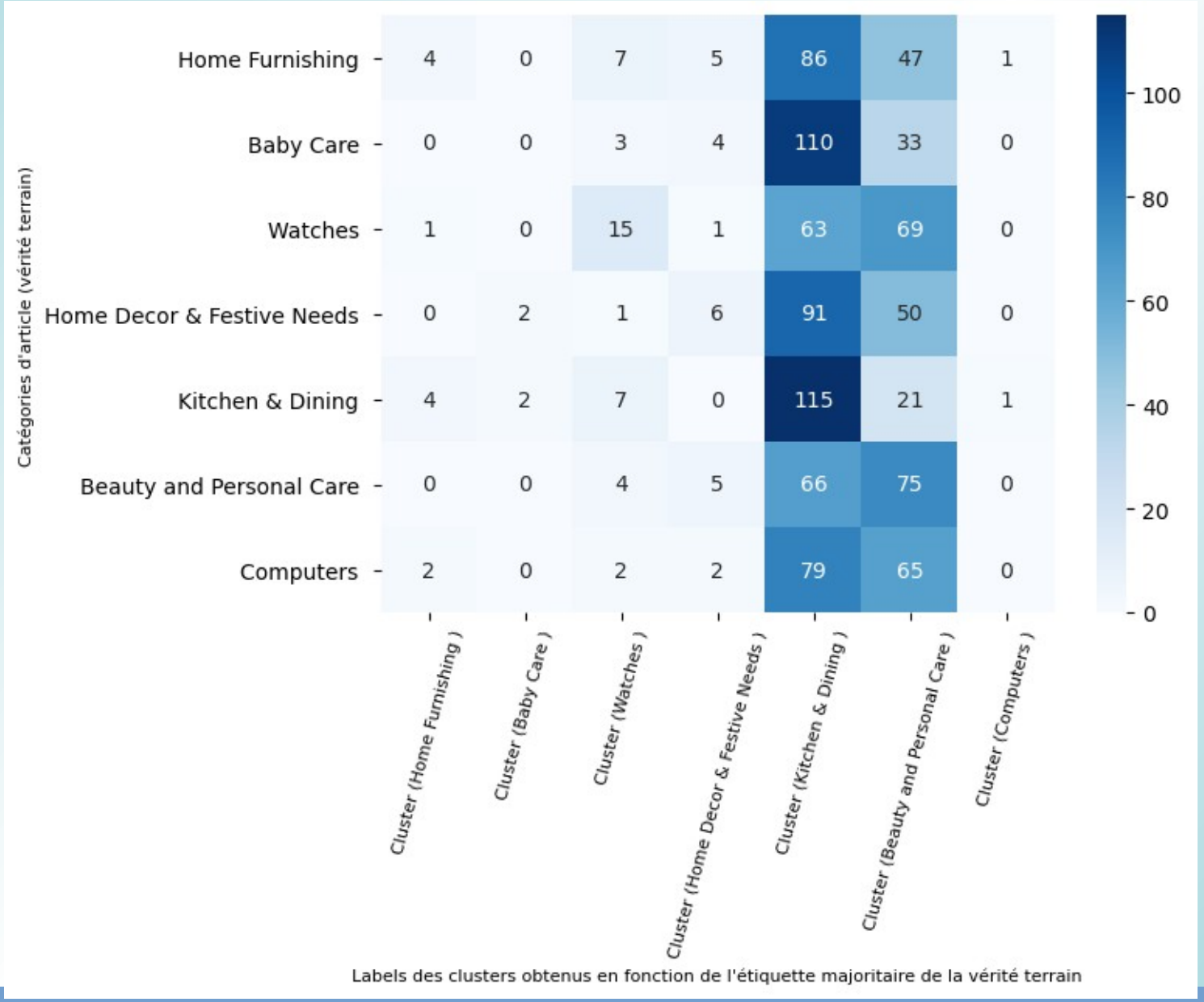
Etude de faisabilité – Images – Coeff. de silhouette

Cas encodage SIFT et PCA



Etude de faisabilité – Images – Matrice de confusion

Cas encodage SIFT et PCA



Etude de faisabilité – Images - Réseaux neurones CNN et PCA

Synthèse des résultats du clustering

- **ARI score : résultat très mauvais pour le réseau de neurones pré-entraîné VGG16 (0,09) et moyen (0,34) pour ResNet50.**

- **Score de silhouette du clustering : résultats médiocres pour VGG16 (0,13), et mauvais pour ResNet50 (0,02) avec** des erreurs sur les données appartenant aux clusters, des clusters se chevauchant et non denses.

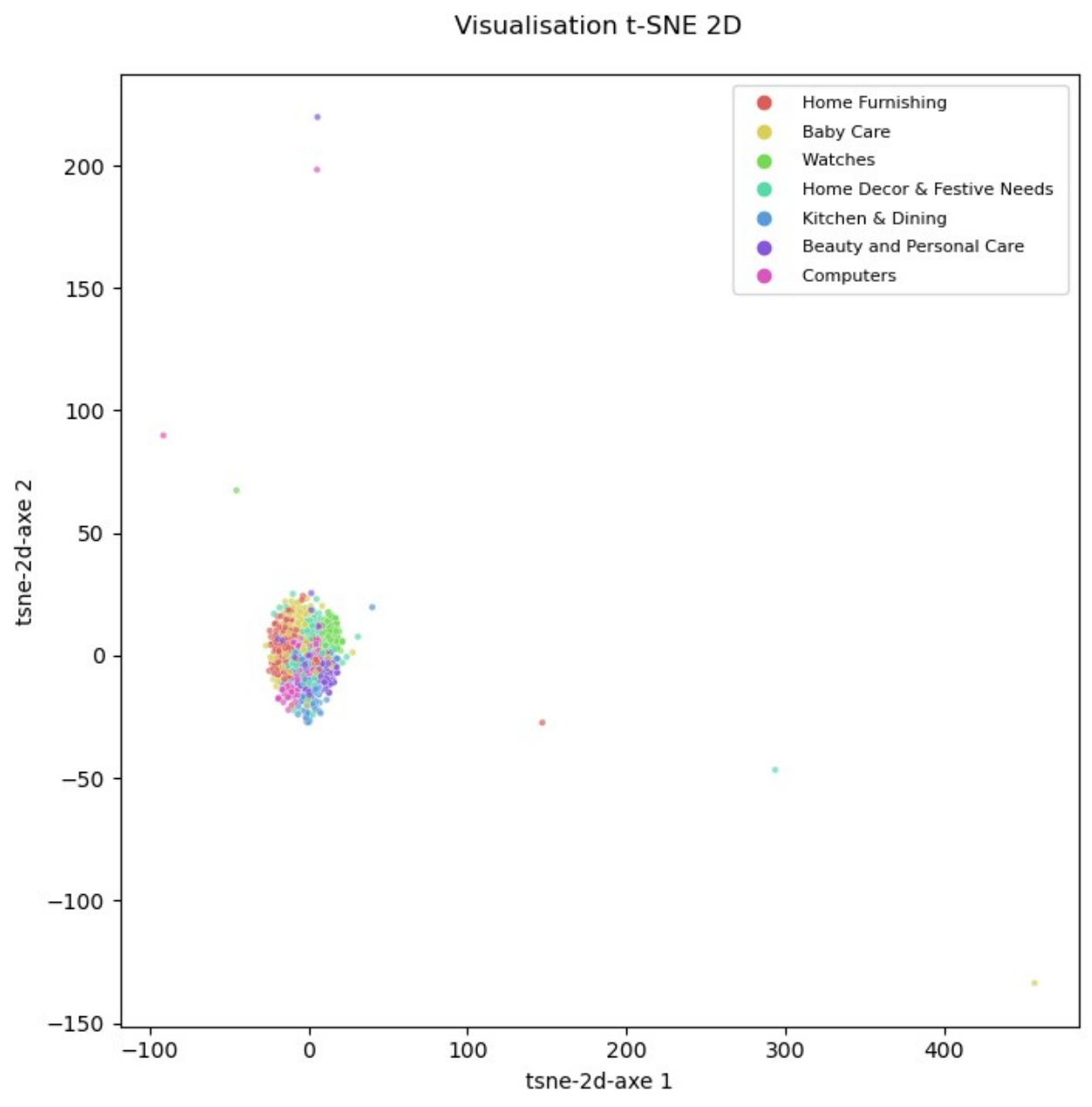
- La visualisation T-SNE 2D des labels de clusters :
  - montre des groupements peu ou pas séparés aussi bien sur la visualisation T-SNE 2D de la vérité terrain / labels des clusters.

- La visualisation des coefficient de silhouette par clusters montre qu'il y a beaucoup d'erreurs sur les clusters.

- Pour ResNet50, la projection de la matrice de confusion montre que les regroupements de clusters semblent correspondre partiellement, avec un taux d'erreurs important, aux étiquettes de la vérité terrain (4 catégories sur 7).

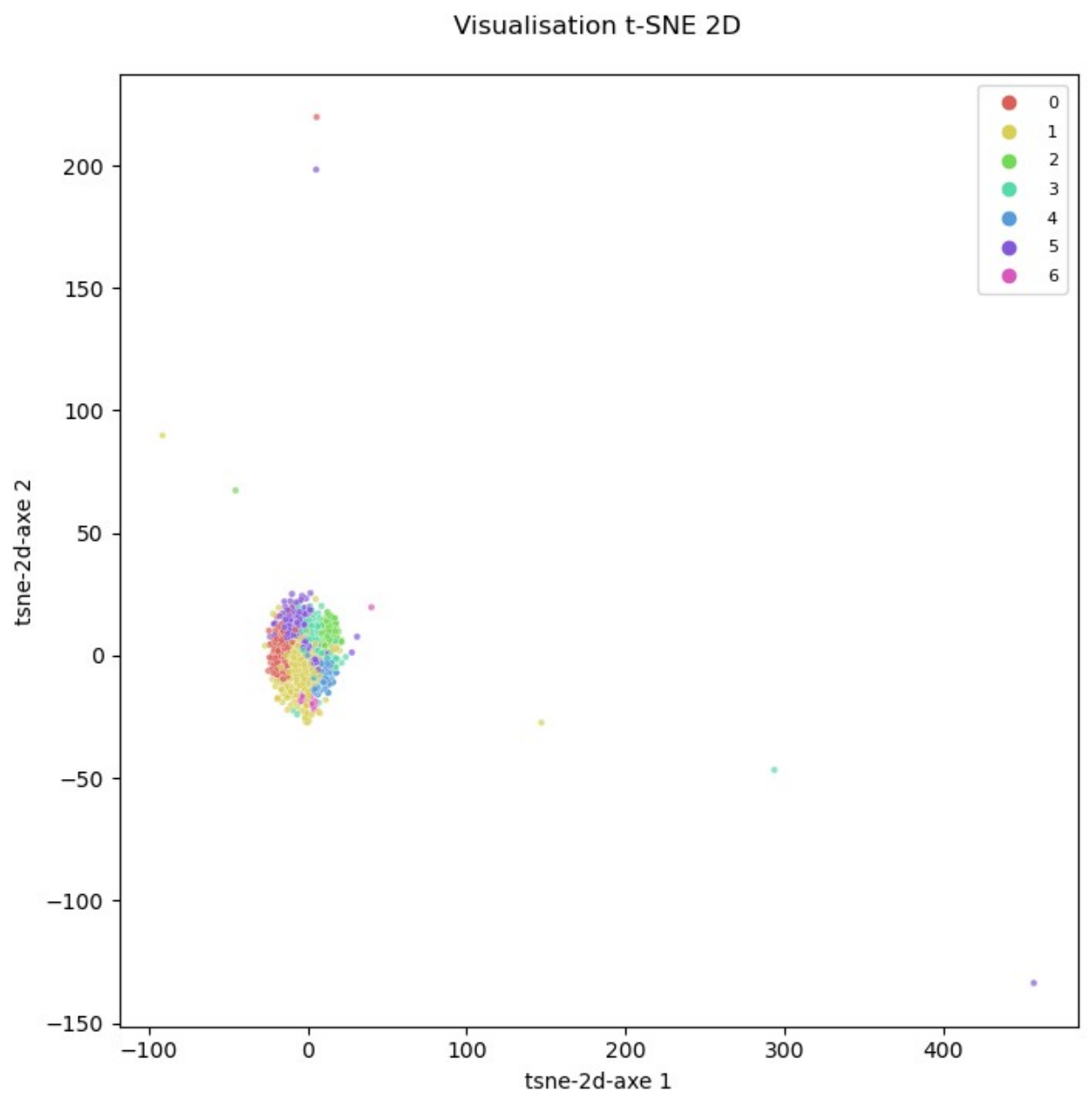
Etude de faisabilité – Images – T-SNE sur vérité terrain

Cas réseaux de neurones pré-entraîné ResNet50 et PCA



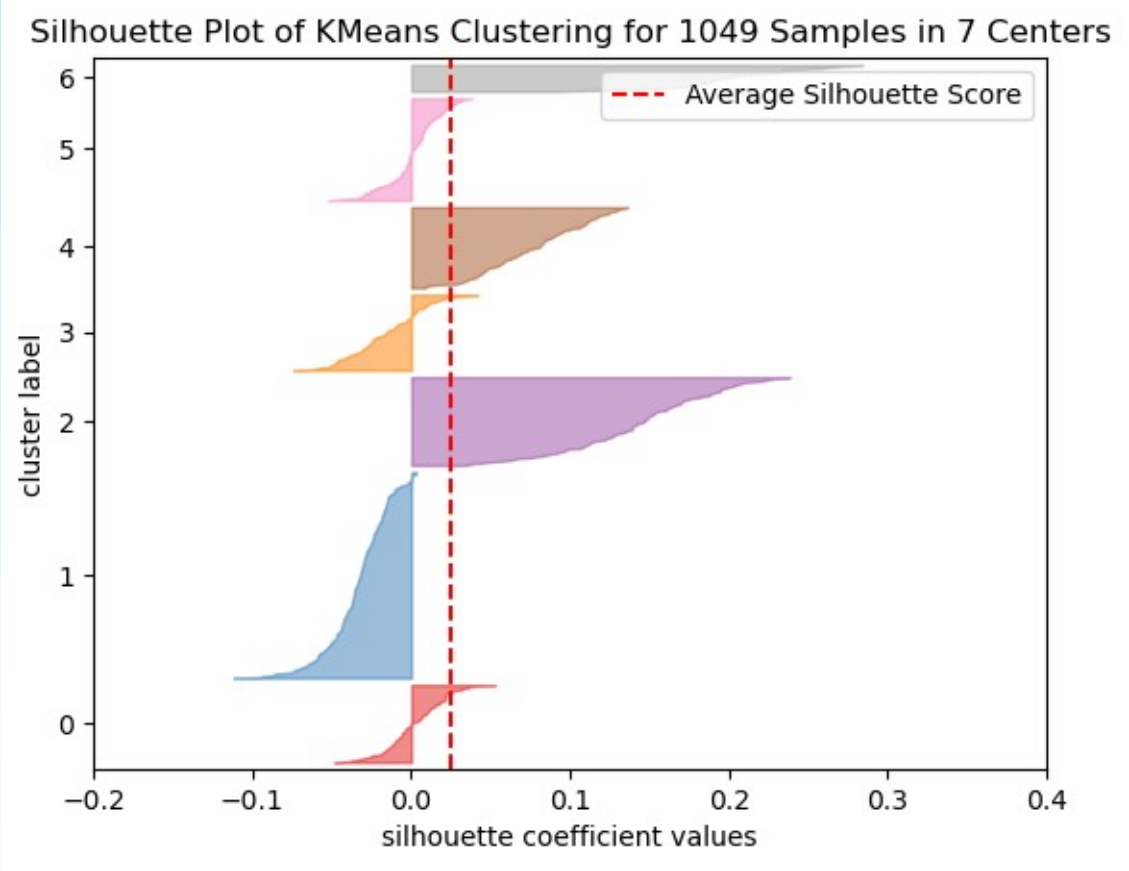
Etude de faisabilité – Images – T-SNE clusters k-means

Cas encodage réseaux de neurones ResNet50 et PCA



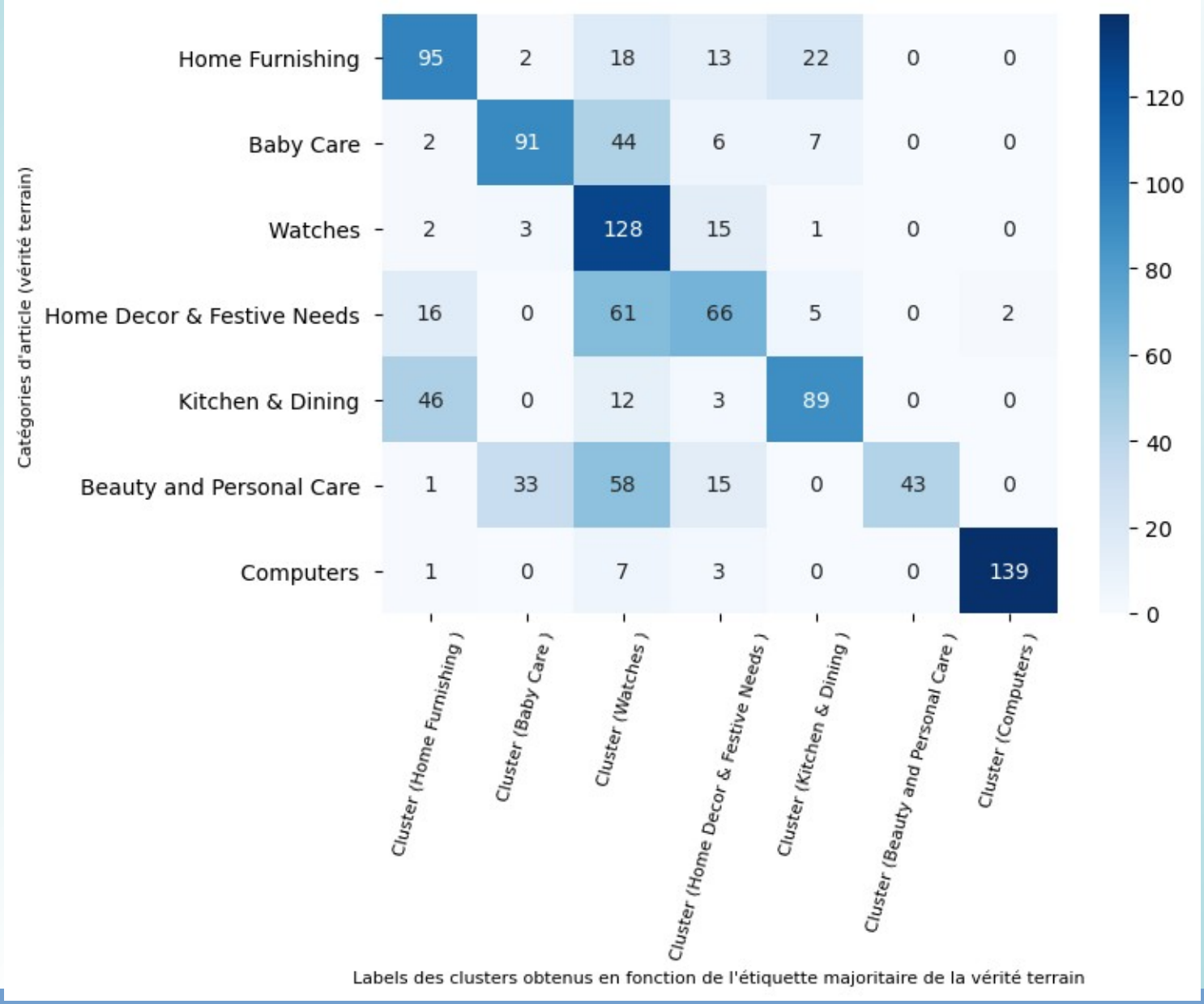
Etude de faisabilité – Images – Coeff. de silhouette

Cas encodage réseaux de neurones ResNet50 et PCA



Etude de faisabilité – Images – Matrice de confusion

Cas encodage réseaux de neurones ResNet50 et PCA



Etude de faisabilité – Images / textes - NMF

Synthèse des résultats du clustering

- Utilisation NMF pour les encodages images SIFT, réseaux de neurones VGG16 / ResNet50

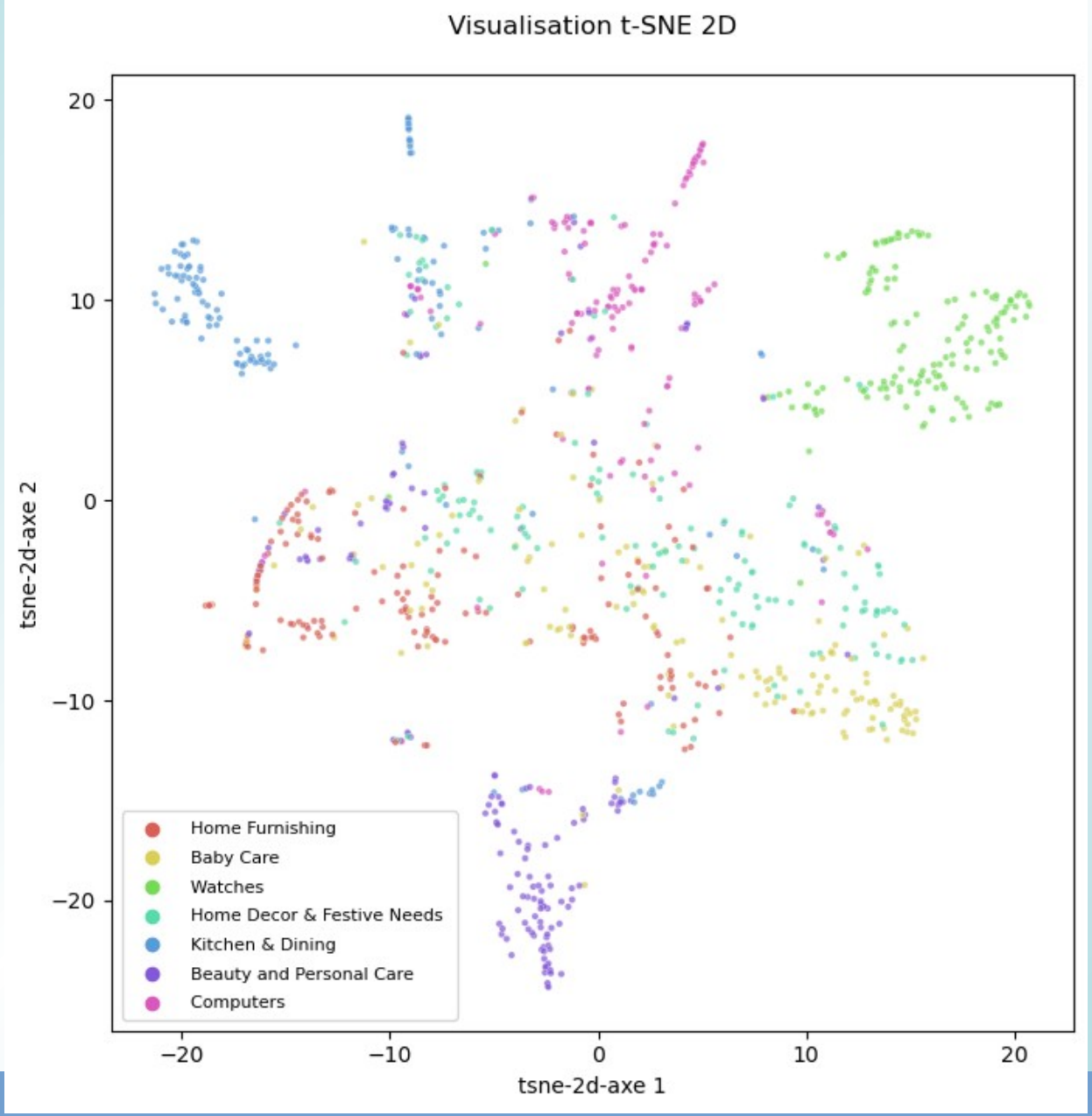
  et pour les meilleurs encodages texte (BOW lemmatisation et limitation occurrences de mots, Tf-Idf avec lemma, min_df et modèle N-gram)

- **ARI score :**
  - pour les encodages texte, résultats identiques à PCA,
  - pour les encodages image, sensible amélioration pour VGG16 uniquement.

- **Score de silhouette du clustering :**
  - amélioration nette pour les encodages texte/images (score de 0,31 à 0,63)
  - moins d'erreurs sur les données appartenant aux clusters,
  - moins de chevauchement entre clusters et clusters plus denses.

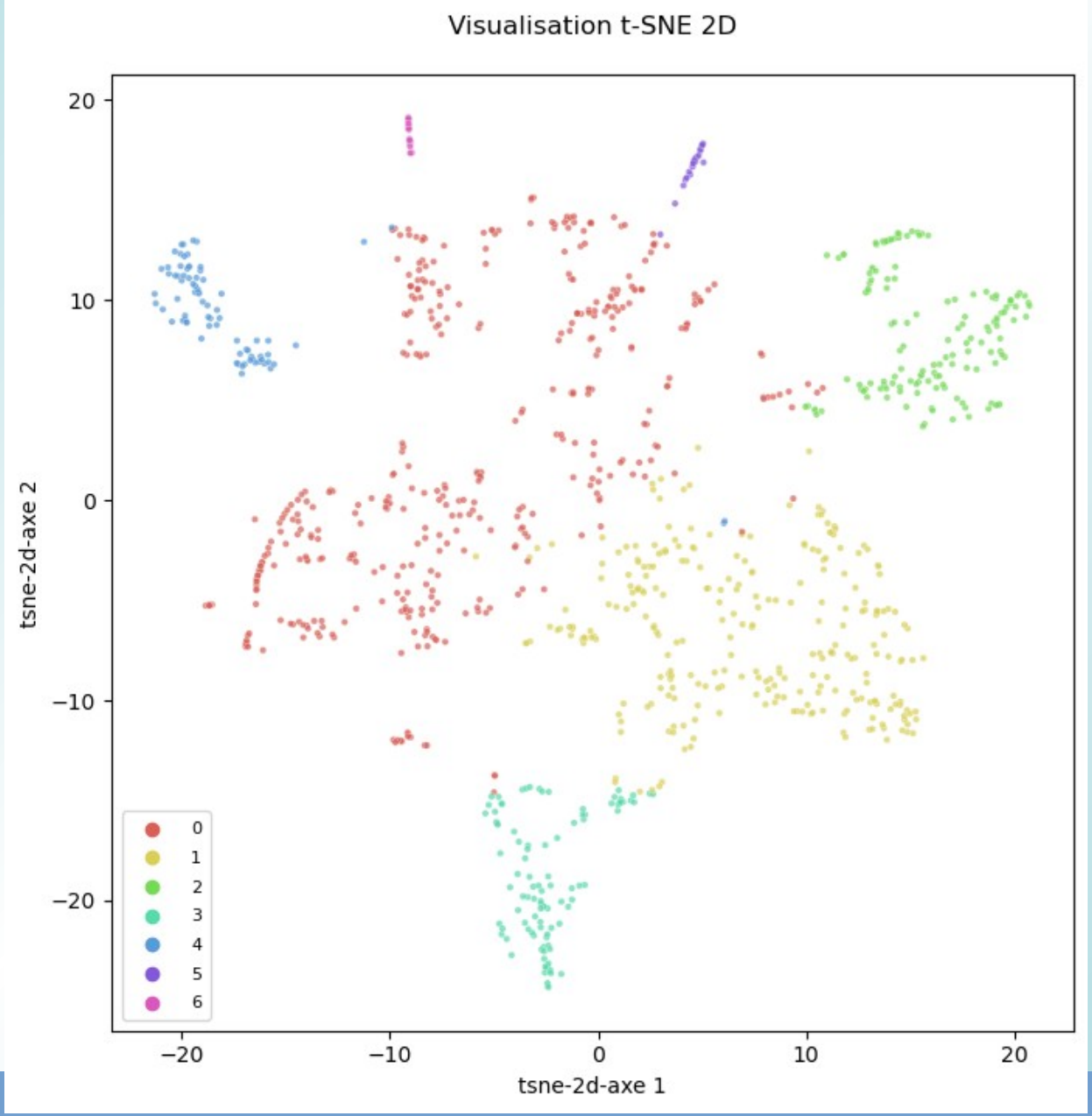
Etude de faisabilité – Images – T-SNE sur vérité terrain

Cas réseaux de neurones pré-entraîné ResNet50 et NMF



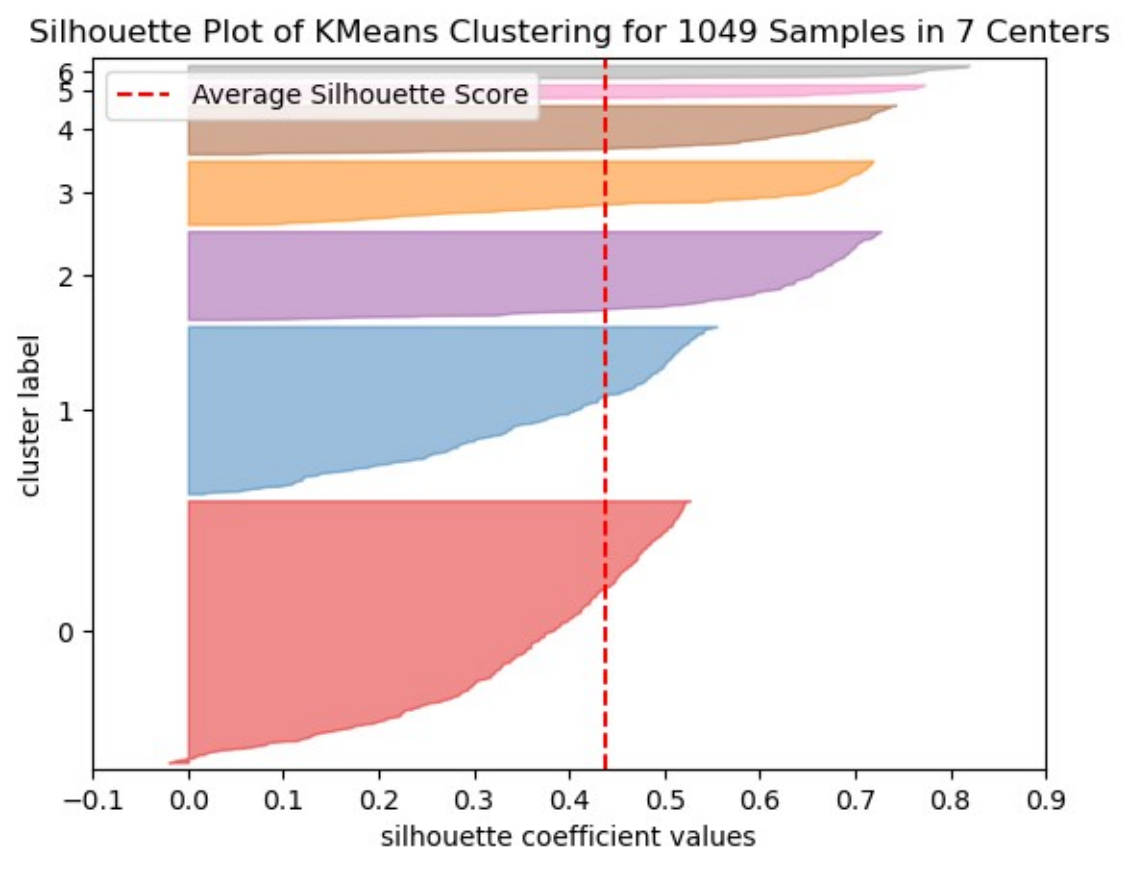
Etude de faisabilité – Images – T-SNE clusters k-means

Cas encodage réseaux de neurones ResNet50 et NMF



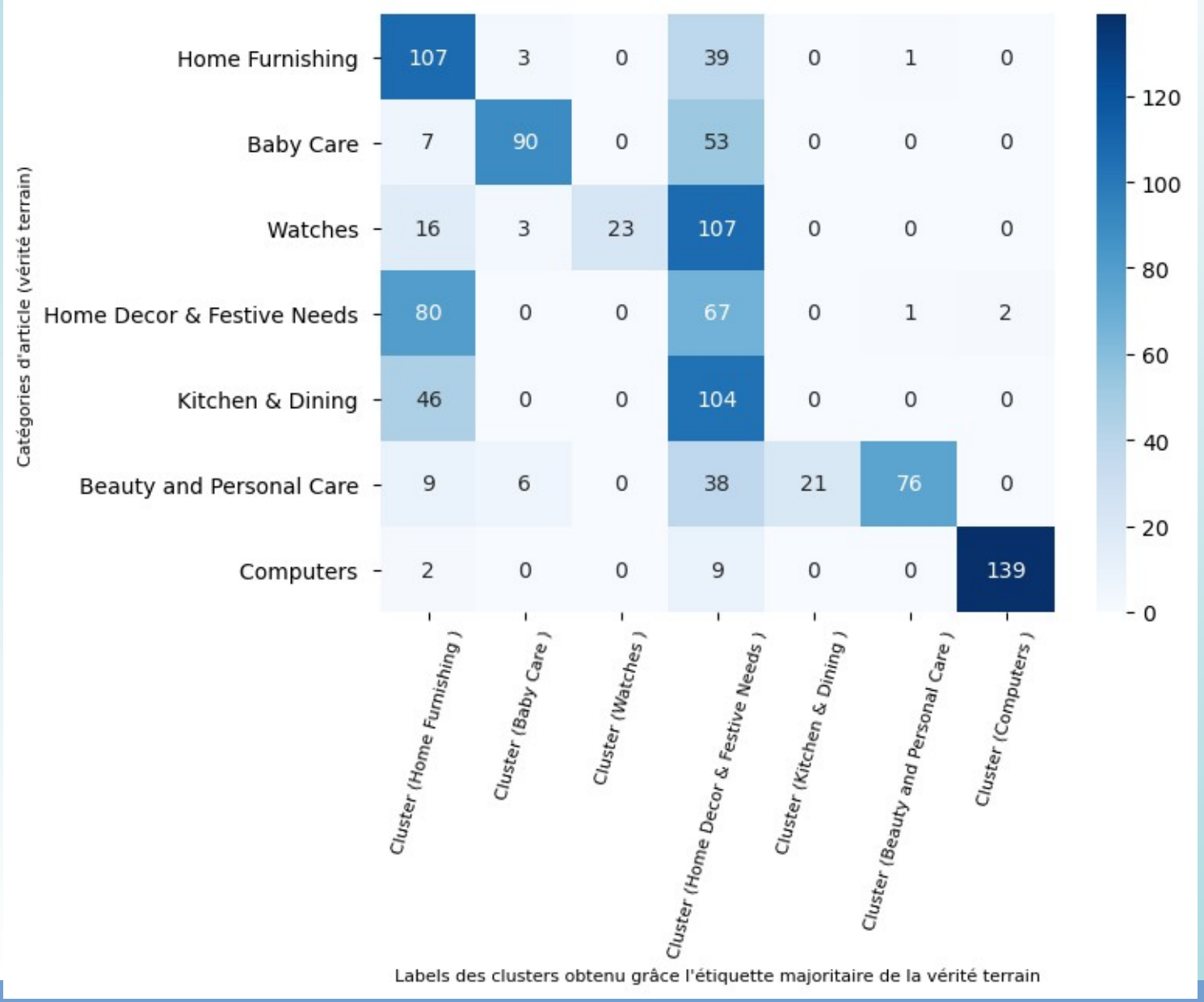
Etude de faisabilité – Images – Coeff. de silhouette

Cas encodage réseaux de neurones ResNet50 et NMF



Etude de faisabilité – Images – Matrice de confusion

Cas encodage réseaux de neurones ResNet50 et NMF



Etude de faisabilité – Images + textes - PCA / NMF

Synthèse des résultats du clustering

- Comparaison encodage texte TF-IDF + encodage image réseaux de neurones ResNet50 avec PCA et NMF

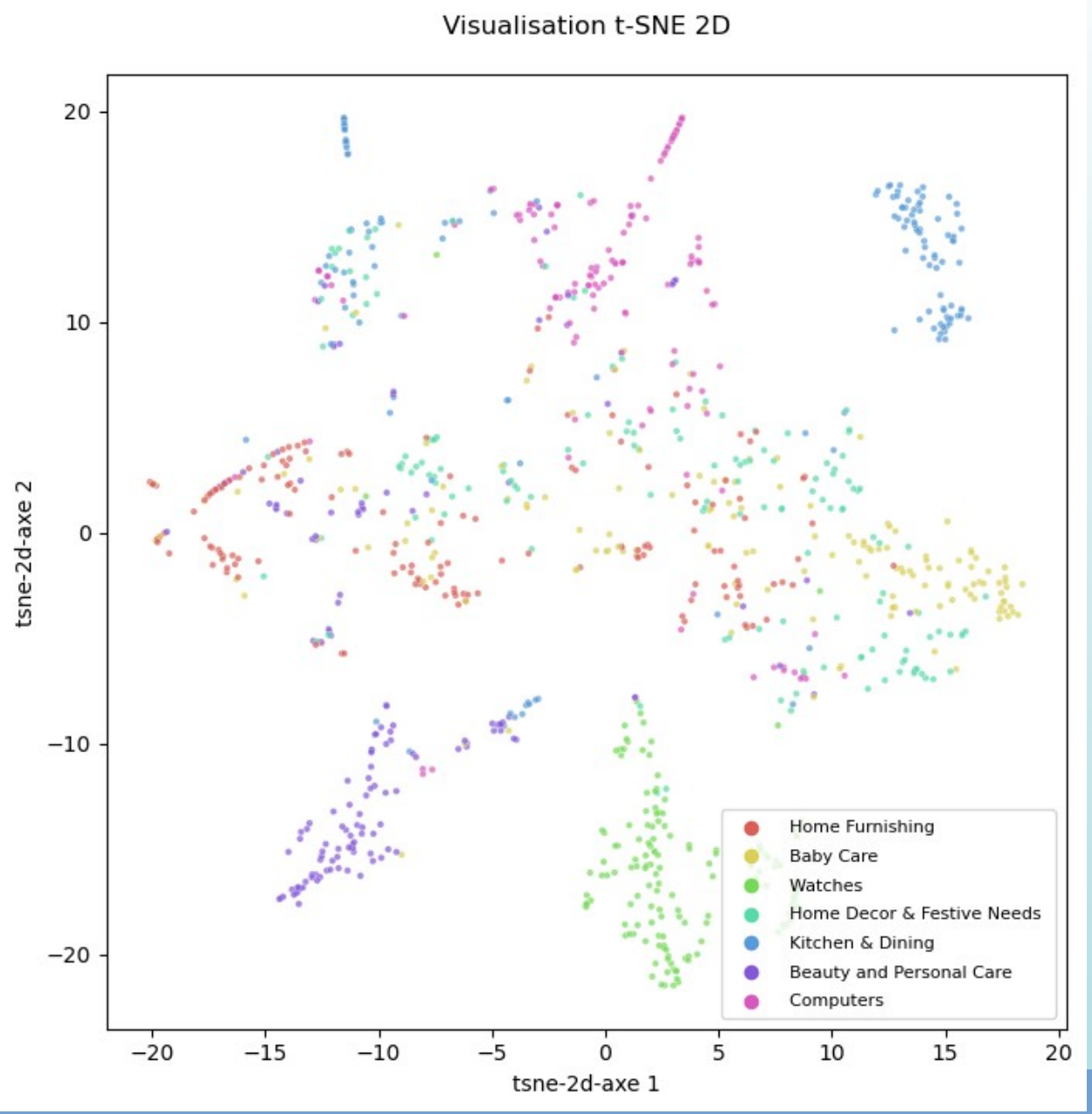
- **Résultats :**
  - Avec PCA, résultats inférieurs aux encodages texte et image seuls.
  - Av**ec** NMF, résultats identiques à l'encodage images seul

    => ARI score = 0,26

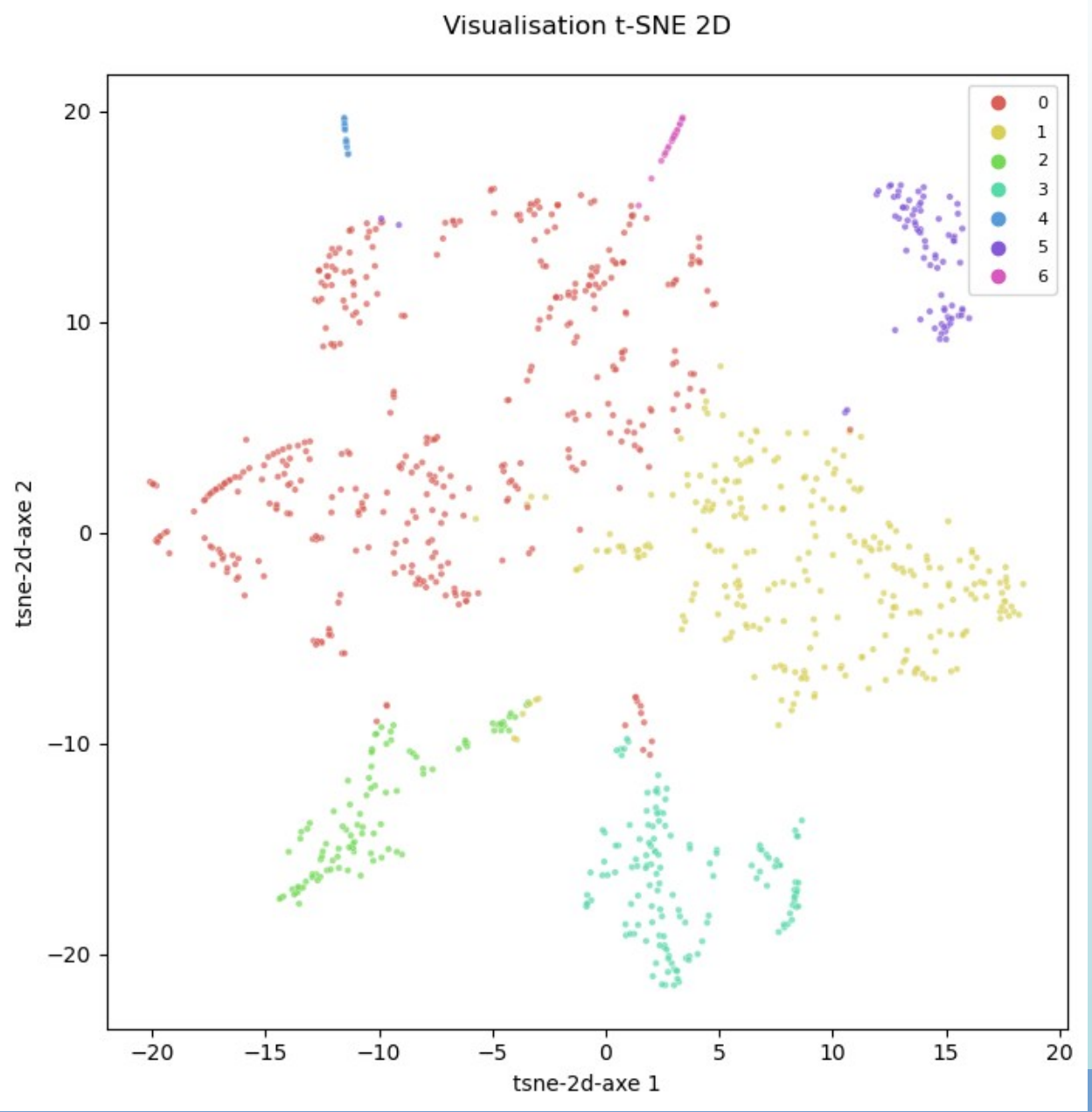
    => Score silhouette = 0,44
  - L'ARI score est inférieur à celui de l'encodage texte seul

- **Qualité de clustering et similitude des labels vérité terrain / clusters :**
  - Peu d'erreurs sur les clusters, clusters plus homogènes en taille
  - On peut dégager des similitudes entre plusieurs catégories d'articles et les labels de clusters.

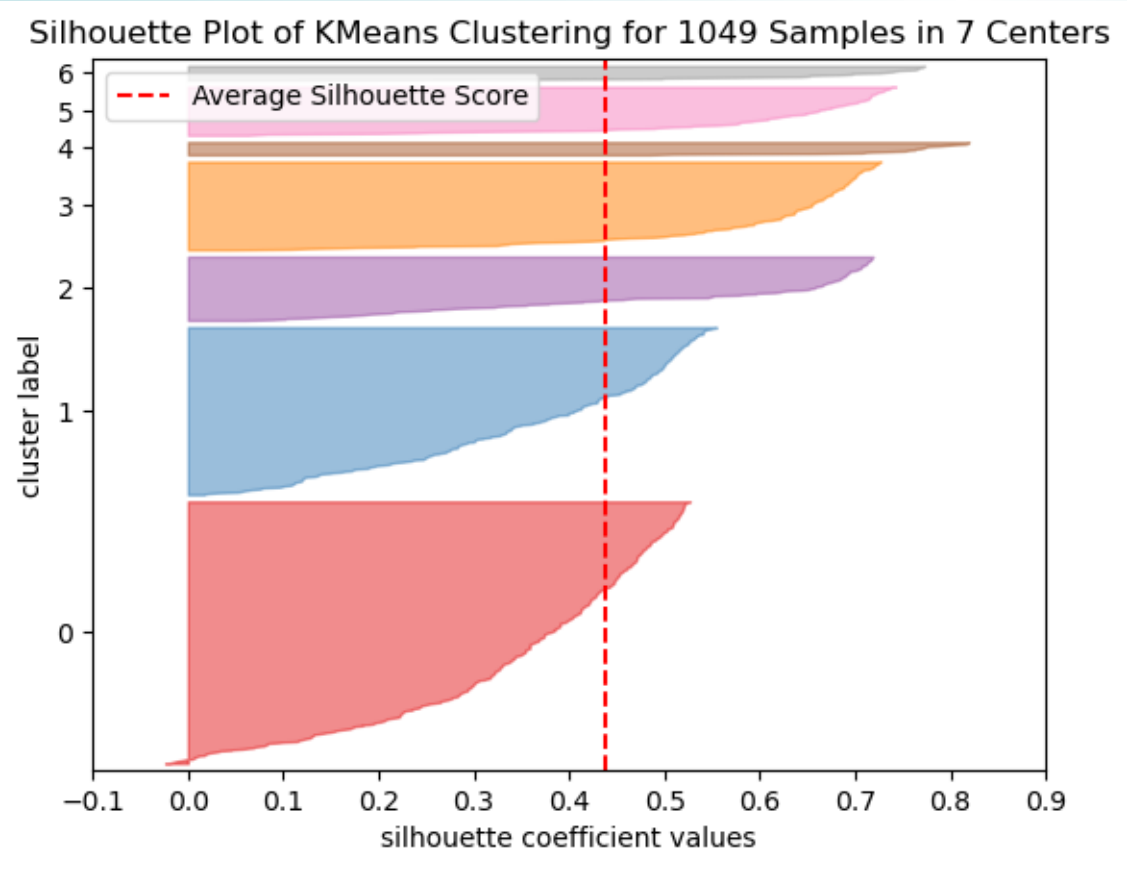
Etude de faisabilité – Images + textes
T-SNE sur vérité terrain
Cas TF-IDF + réseaux de neurones pré-entraîné ResNet50 et NMF



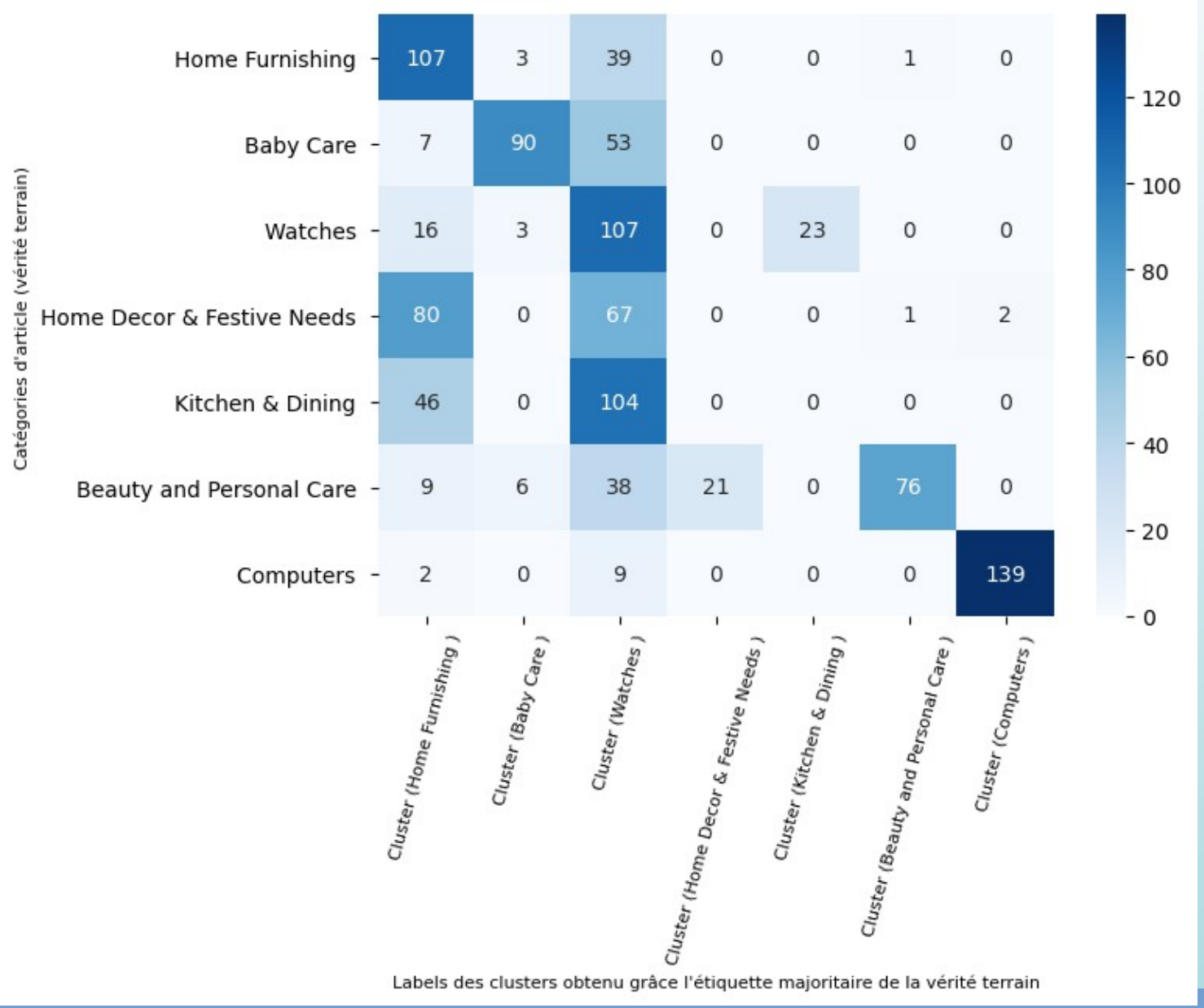
Etude de faisabilité – Images + textes
T-SNE clusters k-means
Cas TF-IDF + encodage réseaux de neurones ResNet50 et NMF



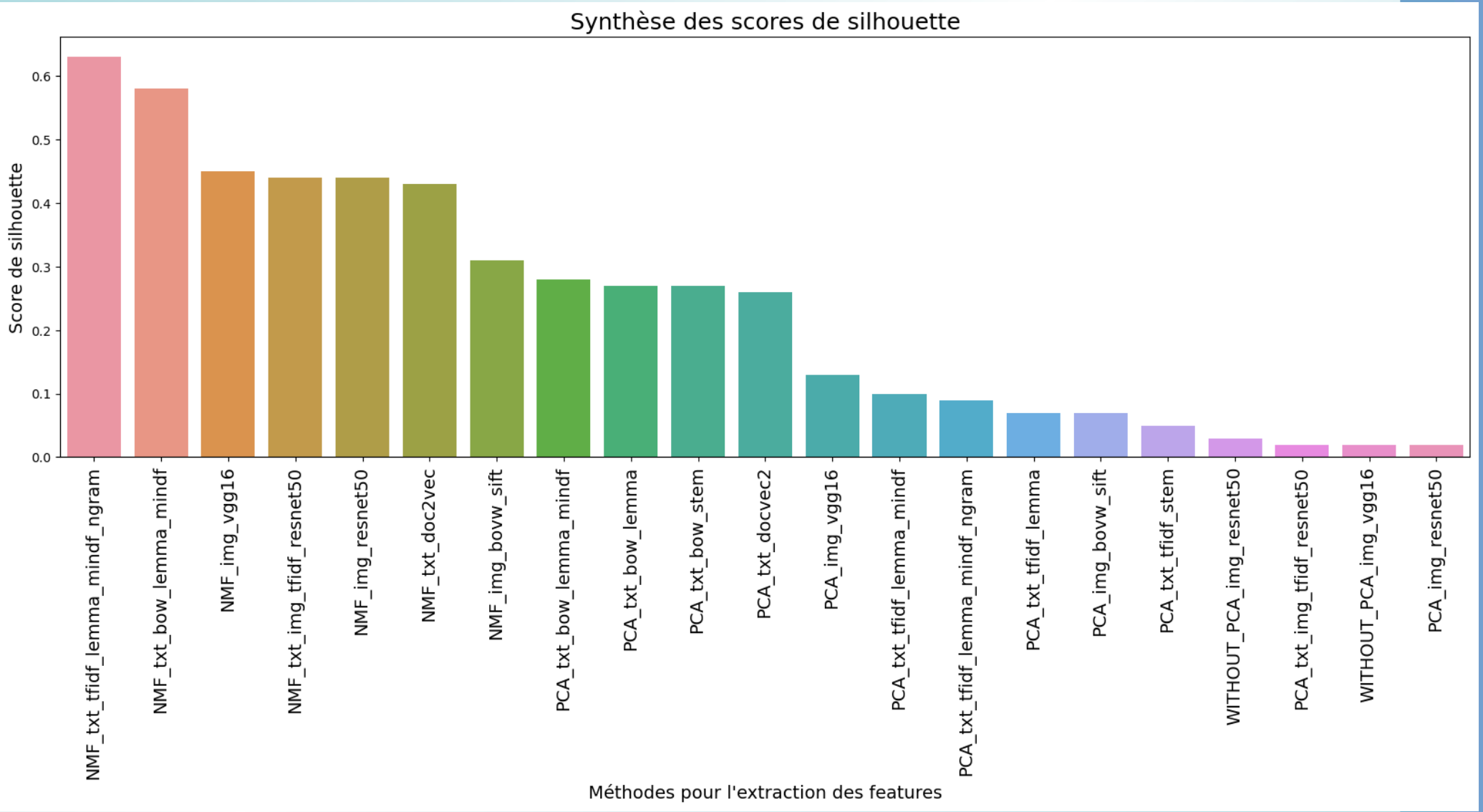
Etude de faisabilité – Images + textes
Coeff. de silhouette
Cas TF-IDF + encodage réseaux de neurones ResNet50 et NMF



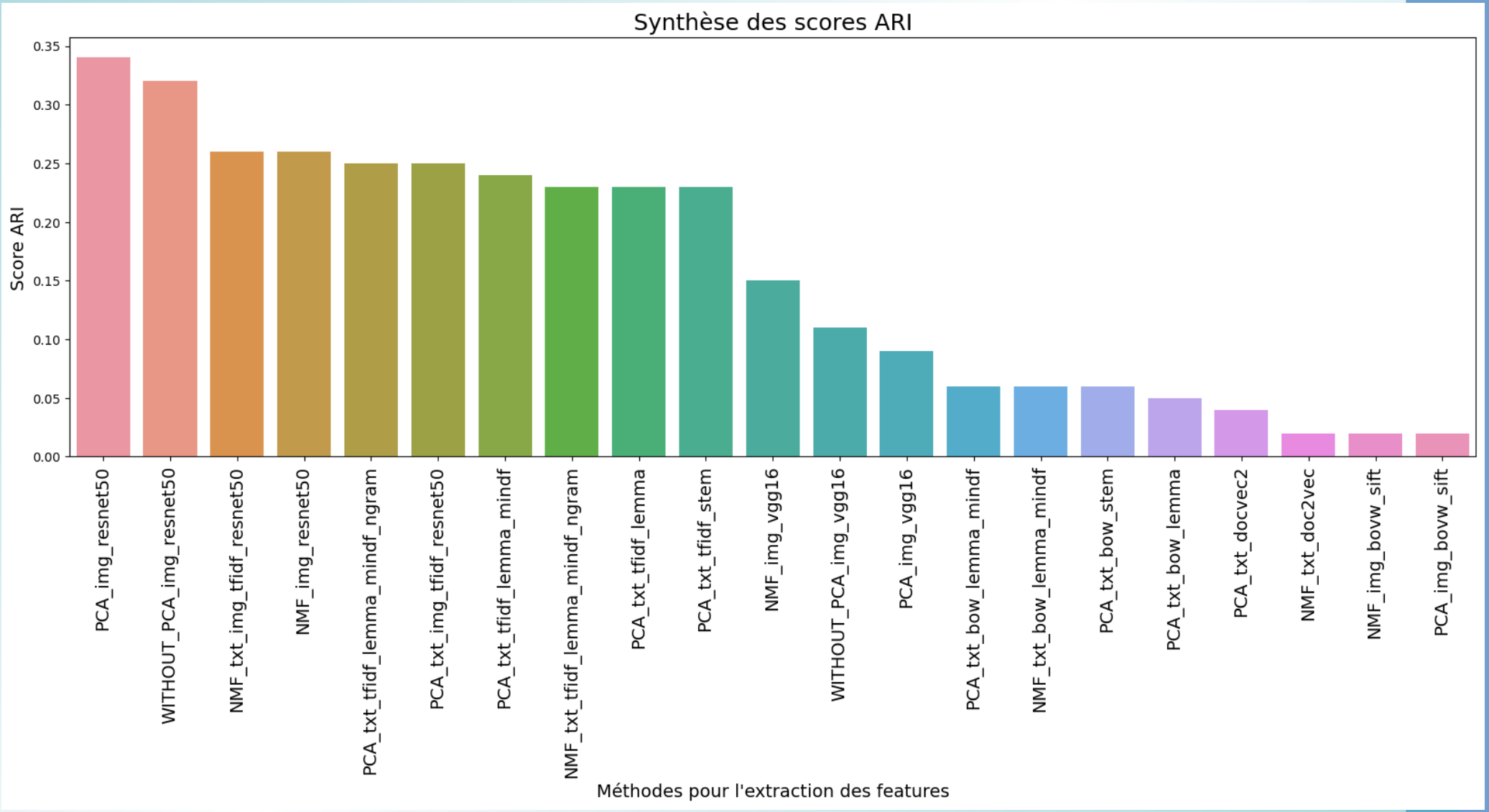
Etude de faisabilité – Images + textes
Matrice de confusion
Cas TF-IDF + encodage réseaux de neurones ResNet50 et NMF



Conclusion sur l'étude de faisabilité
Synthèse de scores de silhouette sur le clustering



Cas de classification SVM – Matrice de confusion



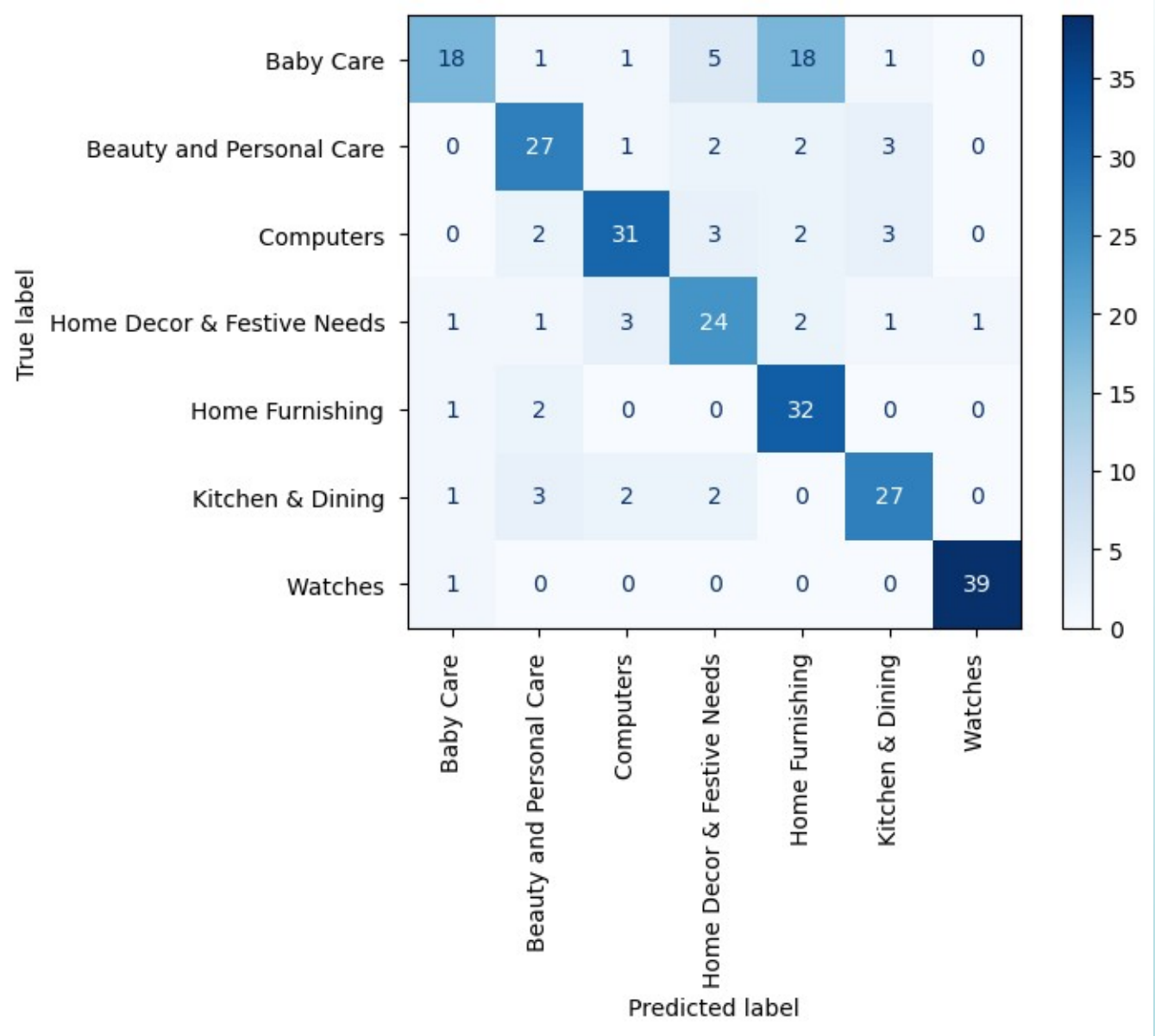
Cas de classification SVM

Encodage textes seuls, images seules, textes + images et NMF

- **Résultats encodage textes seuls :**
  - TF-IDF avec lemmatisation, min_df, et modèle N-gram et NMF.
  - Accuracy moyen = 0,57
  - Beaucoup d'erreurs sur les prédictions.

- **Résultats encodage images seules / encodage textes + images**
  - Réseau de neurones ResNet50 pré-entraîné et NMF
  - Accuracy = 0,75
  - Erreurs de prédiction limitées, sauf pour la catégorie « Baby Care ».
  - Même précision pour encodage texte+images que pour encodage images seules

Cas de classification SVM
Matrice de confusion
Cas TF-IDF + encodage réseaux de neurones ResNet50 et NMF



Conclusion finale – Améliorations - Limites

- **Faiblesse de la PCA par rapport à NMF :**
  - PCA pas vraiment adapté au traitement des textes et images.
  - NMF plus efficace pour la reconnaissance de mots et vocabulaire, le traitement des images.

- **L'étude de faisabilité montre qu'il est possible d 'envisager un moteur de classification à partir de l'encodage texte seul, de l'encodage image seul et de l'encodage texte + image.**
  - La combinaison texte + image améliore sensiblement la similitude des regroupements par rapport à un encodage texte seul.

- **Limites :**
  - Les meilleurs clustering arrivent à fournir des regroupements distincts et bien répartis, mais qui ne correspondent pas à nos catégories initiales (ARI moyen).

- **Améliorations :**
  - Augmenter la volume de données pour améliorer les performances du clustering / classification.
  - Expérimenter des méthodes Word Embedding plus performantes (FastText, ELMo, Google Universal Sentence,…).