OPENCLASSROOMS

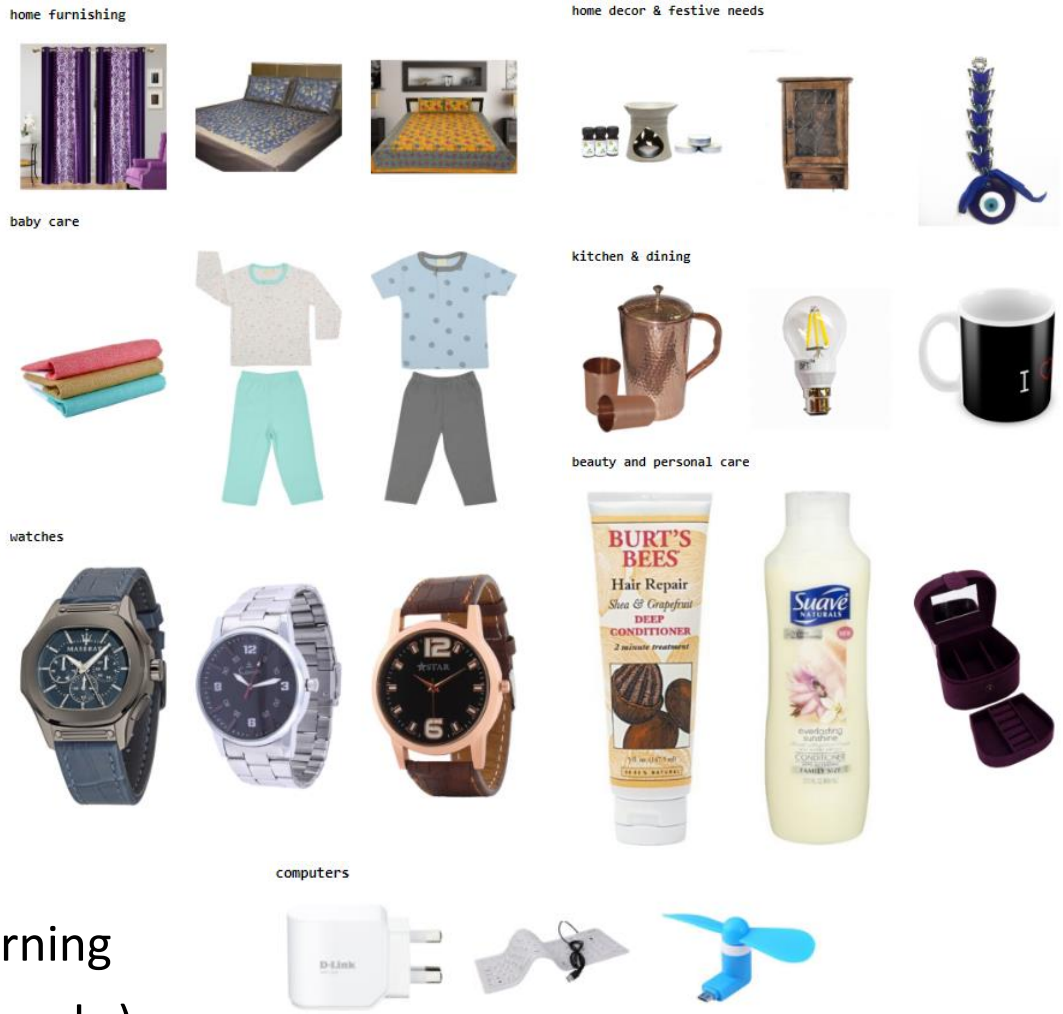
Project 6

Classifiez automatiquement des biens de consommation

Tetiana Lemishko

Sommaire:

- Mission et objectifs principaux
- Jeu de données
- Données textuelles:
  - Pré-traitement des données
    - Bag of words
    - Tf-idf
    - Word2Vec
    - BERT
    - USE
- Données visuelles:
  - Algorithme SIFT (Scale Invariant Feature Transform)
  - Algorithmes CNN (Transfer Learning with Convolutional Neural Networks)
- Conclusion



Mission et objectifs principaux

- "Place de marché" est une entreprise qui 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.
- L'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs, et est donc peu fiable.

Mission: 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.

Objectifs principaux:

- Analyser le jeu de données en réalisant un pré-traitement des descriptions des produits et des images, une réduction de dimension, puis un clustering.
- Ameliorer l'expérience des utilisateurs.
- Fiabiliser la catégorisation des articles.

Données textuelles

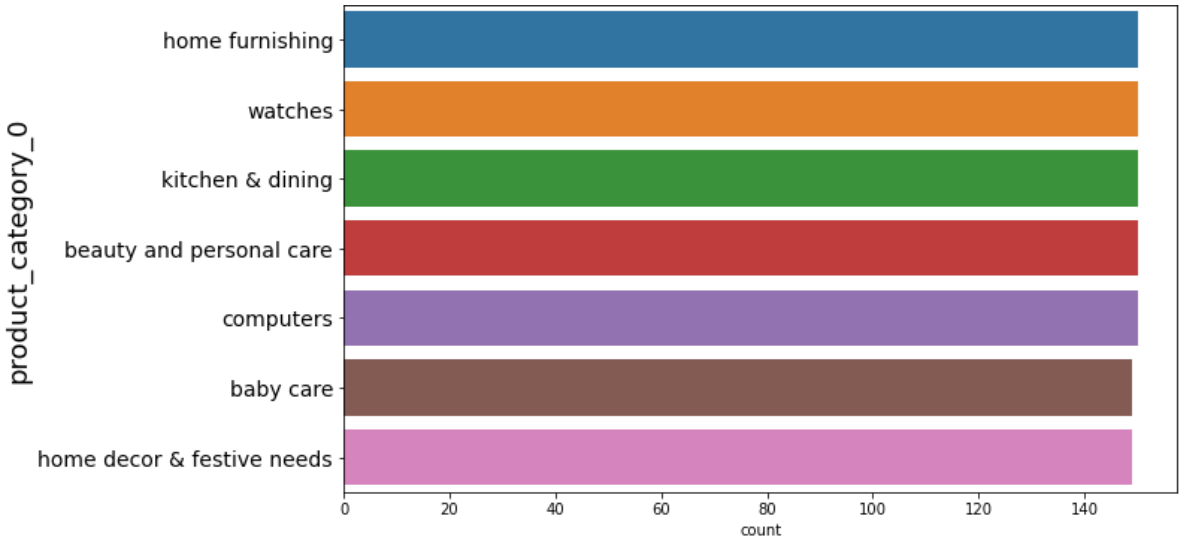
Jeu de données initiale

uniq_id	object
crawl_timestamp	object
product_url	object
product_name	object
product_category_tree	object
pid	object
retail_price	float64
discounted_price	float64
image	object
is_FK_Advantage_product	bool
description	object
product_rating	object
overall_rating	object
brand	object
product_specifications	object

1050 lignes

15 colonnes

Distribution du nombre de produits par categorie 0



Jeu de données de travail:

- uniq_id
- category(product_category_0)
- description(product_name + description)

Exemple (product_category_tree):

'["**Home Furnishing** >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]'

product_category_0

Données textuelles. Prétraitement des données

Étapes du traitement:

- tokenization (les mots et les caractères de ponctuation sont séparés et extraits en tant qu'unités individuelles)
- traitement de punctuation (les caractères de ponctuation sont remplacés par ' ')
- traitement des mots vides (élimination des mots non informatifs (p. ex., les articles) et les petits mots (< 3 caractères))
- traitement des majuscules (conversion des caractères majuscules en caractères minuscules)
- lemmatization (convertir les différentes formes fléchies d'un mot en une forme commune pour qu'ils puissent être regroupés et analysés comme un seul élément)

Exemple de description de l'article avant le traitement:

Elegance Polyester Multicolor Abstract Eyelet Door Curtain Elegance 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

nouvelles variables (avec le texte traité)

'sentence_bow_lem'
(tokenization +
traitement de punctuation +
traitement des mots vides +
traitement des majuscules +
lemmatization)

'description_tok'
(tokenization +
traitement des mots vides +
traitement de punctuation)

Exemple de description de l'article apres le traitement ('sentence_bow_lem'):

elegance polyester multicolor abstract eyelet door curtain elegance key feature elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract eyelet door curtain 213 height pack price 899 this curtain enhances look interior this curtain made 100 high quality polyester fabric feature eyelet style stitch metal ring make room environment romantic loving this curtain ant wrinkle anti shrinkage elegant apparance give home bright modernistic appeal design the surreal attention sure steal heart these 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 given soothing print bring home elegant curtain softly filter light room get right amount sunlight specification elegance polyester multicolor abstract eyelet door curtain 213 height pack general brand elegance designed for door type eyelet model name abstract polyester door curtain set model duster25 color multicolor dimension length 213 box number content sale package pack sale package curtain body design material polyester

Données textuelles. Comptage des mots plus caractéristiques par catégorie

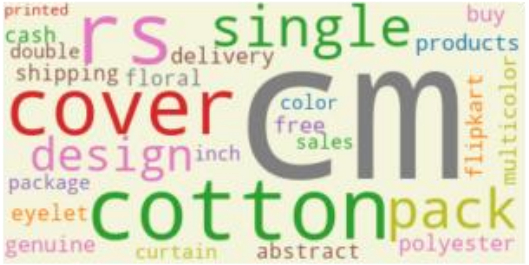
baby care

beauty and personal care

computers

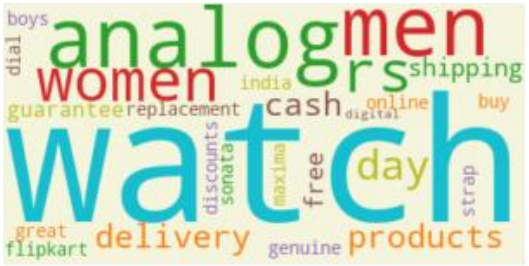
home decor & festive needs

home furnishing



kitchen & dining

watches



Données textuelles. Bag of words

Baf of words (BOW) - C'est un algorithme qui transforme le texte en vecteurs de longueur fixe. Ceci est possible en comptant le nombre de fois que le mot est présent dans un document.

1) Extraction de features (CountVectorizer function)

Exemple – 3 documents: {"Je suis à la maison", "La maison est dans la prairie", "Je suis à la plage"}

	je	*suis*	à	*la*	*maison*	*est*	*dans*	*prairie*	*plage*
phrase 1	1	1	1	1	1	0	0	0	0
phrase 2	0	0	0	2	1	1	1	1	0
phrase 3	1	1	1	1	0	0	0	0	1

2) Réduction de dimension

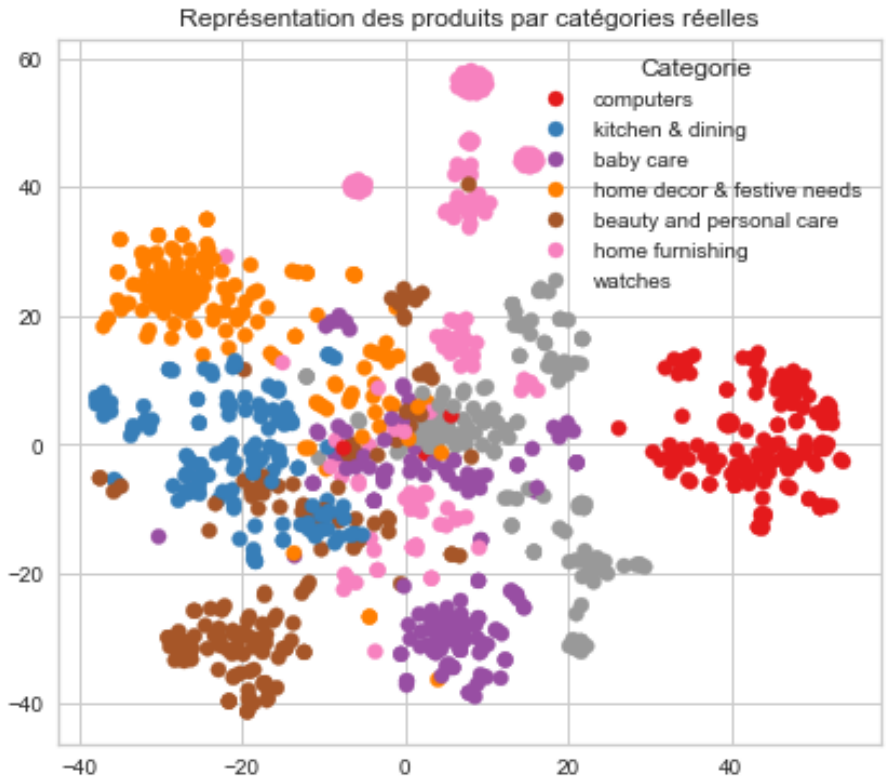
t-SNE (t-distributed stochastic neighbor embedding) est une technique qui permet de visualiser de données de grands dimensions, en effectuant un plongement (embedding) dans une variété de plus petites dimensions (2 ou 3) pour pouvoir distinguer des caractéristiques intéressantes.

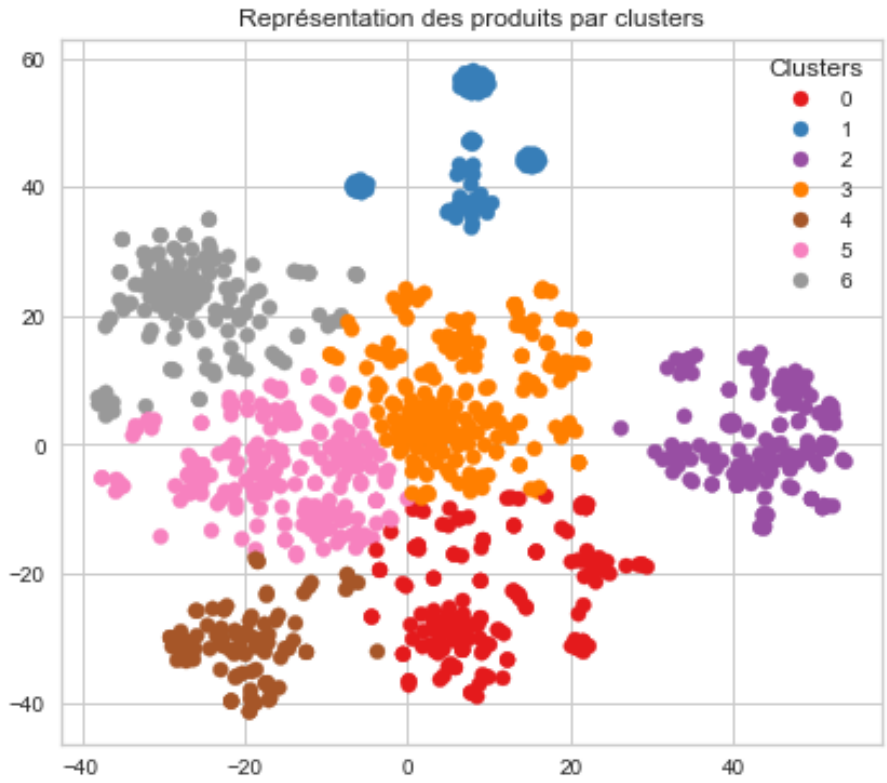
3) Clustering (k-means avec 7 clusters)

4) ARI score

Montre le degré de correspondance entre le partitionnement des catégories réelles et le clustering

Données textuelles. Bag of Words





ARI : 0.4329

Le clustering a bien séparé des catégories telles que ‘computers’, ‘home furnishing’ et ‘beauty and personal care’

Données textuelles. Tf-idf

Term Frequency (TF) = nombre de fois où le mot est dans le document / nombre de mots dans le document

Inverse Data Frequency (IDF) = nombre de documents / nombre de documents où apparaît le mot

1) Extraction de features (TfidfVectorizer function (TF * IDF))

Exemple – 3 documents: {"Je suis à **la** maison", "**La** maison est dans **la** prairie", "Je suis à **la** plage"}

	⋯	*la*	⋯
phrase 1	⋯	0.2	⋯
phrase 2	⋯	0.3	⋯
phrase 3	⋯	0.2	⋯

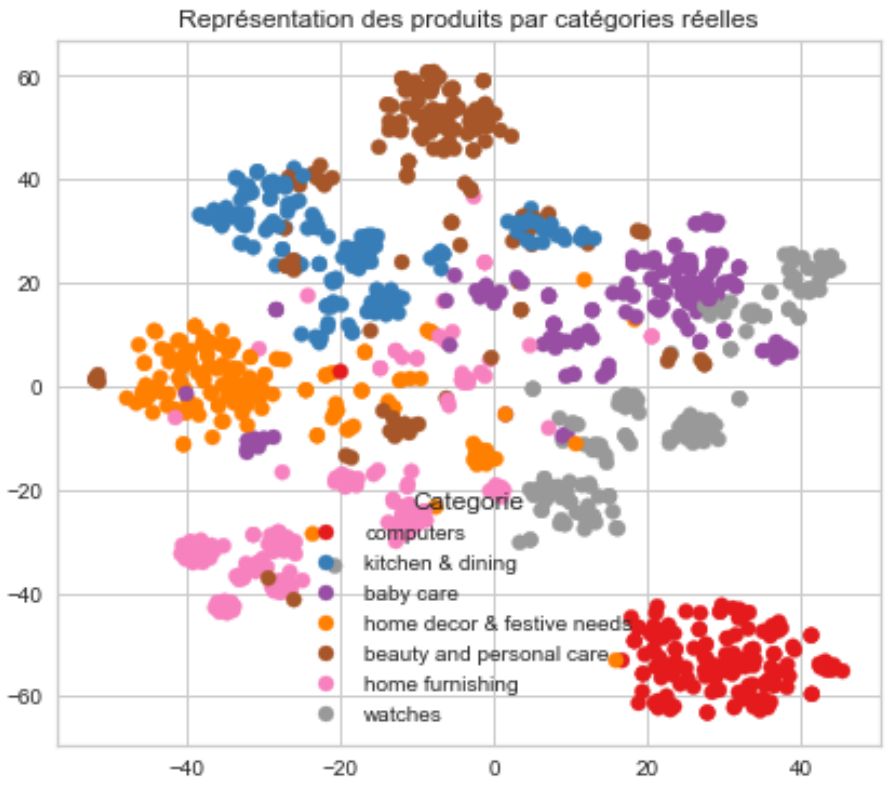
Phrase 1 (la): TF * IDF = (1/5)*(3/3) = 0.2

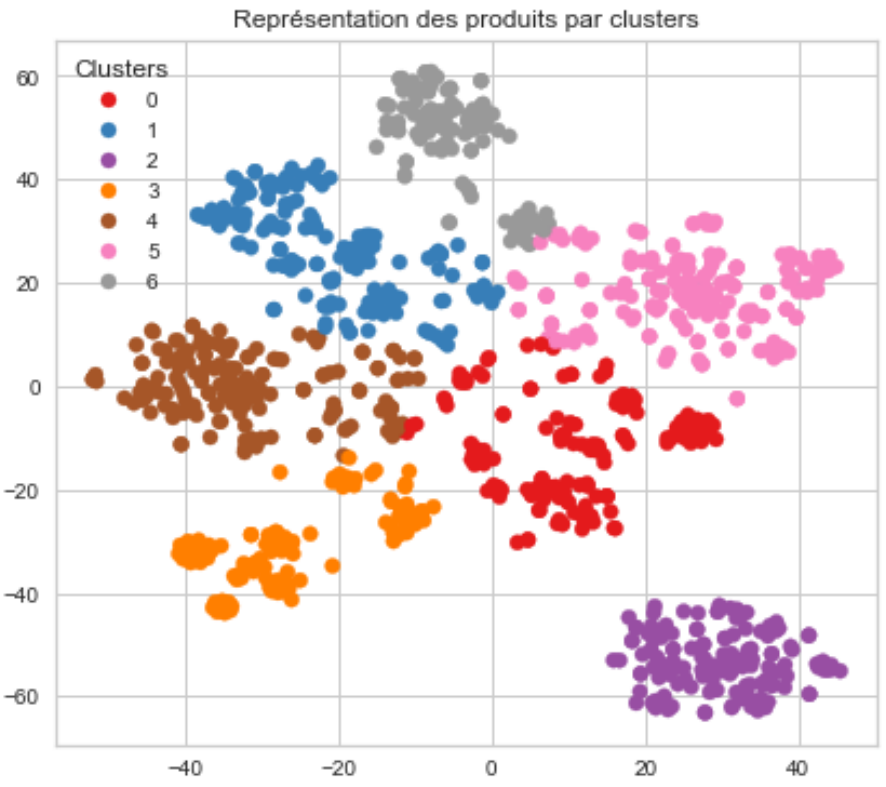
2) Réduction de dimension (t-SNE)

3) Clustering (k-means avec 7 clusters)

4) ARI score

Données textuelles. Tf-idf





ARI : 0.5927

Le clustering a bien séparé des catégories telles que ‘computers’ et ‘beauty and personal care’

Données textuelles. Word2Vec

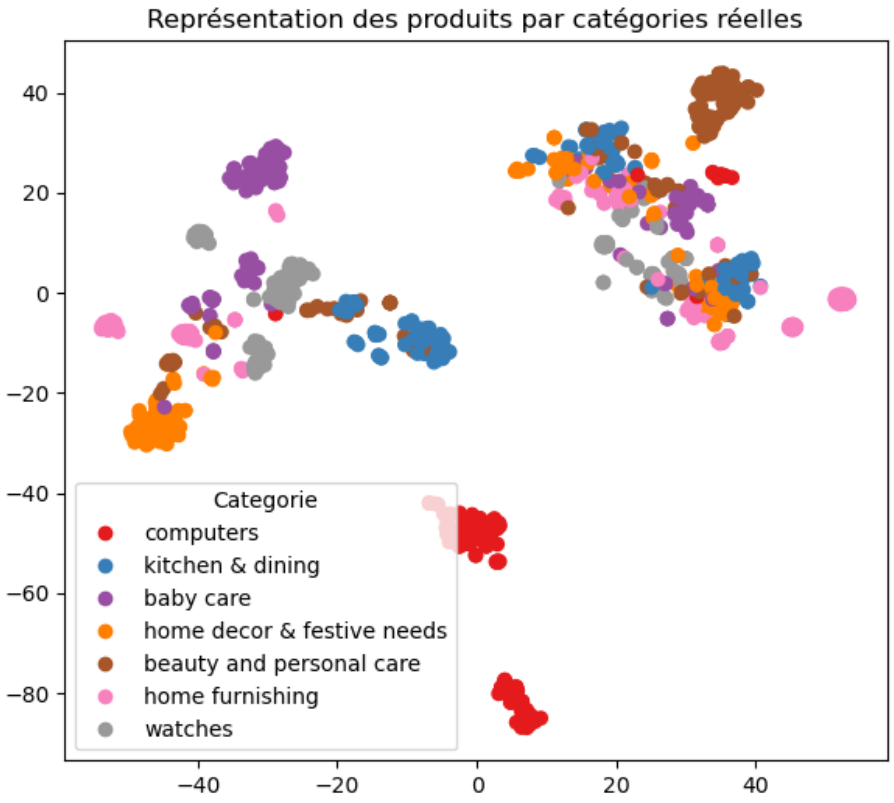
1) Extraction de features

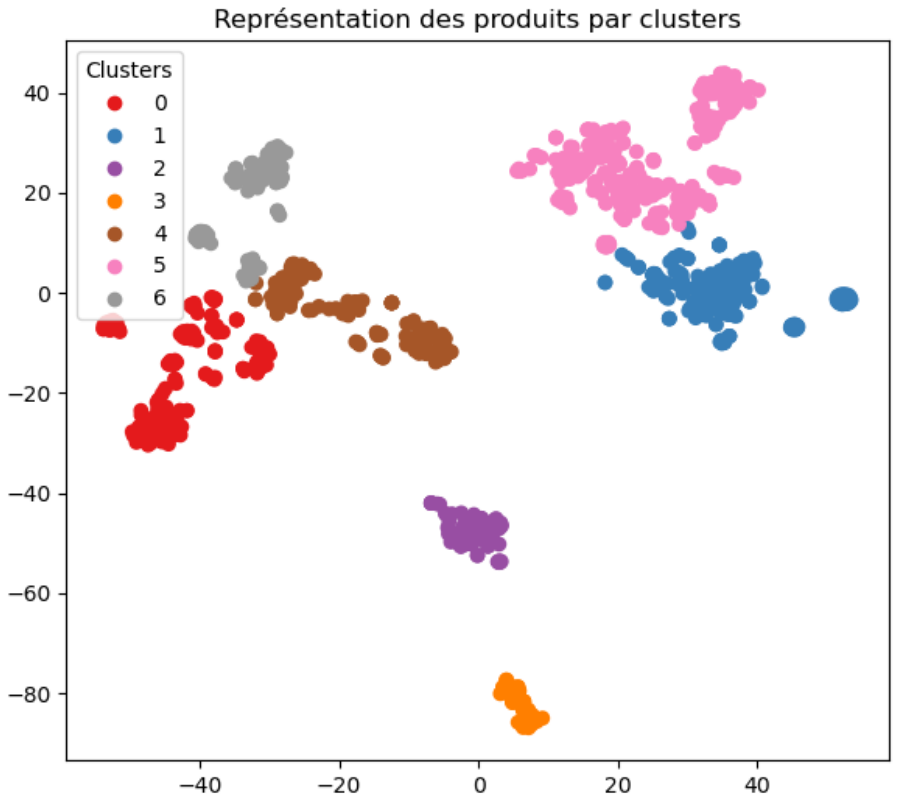
Convertit les mots en vecteurs. Reconnaît les mots de sens similaire (p.ex, "big" ~ "large“, "small" ~ "tiny“).

2) Réduction de dimension (t-SNE)

3) Clustering (k-means avec 7 clusters)

4) ARI score





ARI : 0.1938

Le clustering a bien séparé des catégories telles que ‘baby care’ et ‘home decor & festive needs’

Données textuelles. BERT (Bidirectional Encoder Representations from Transformers)

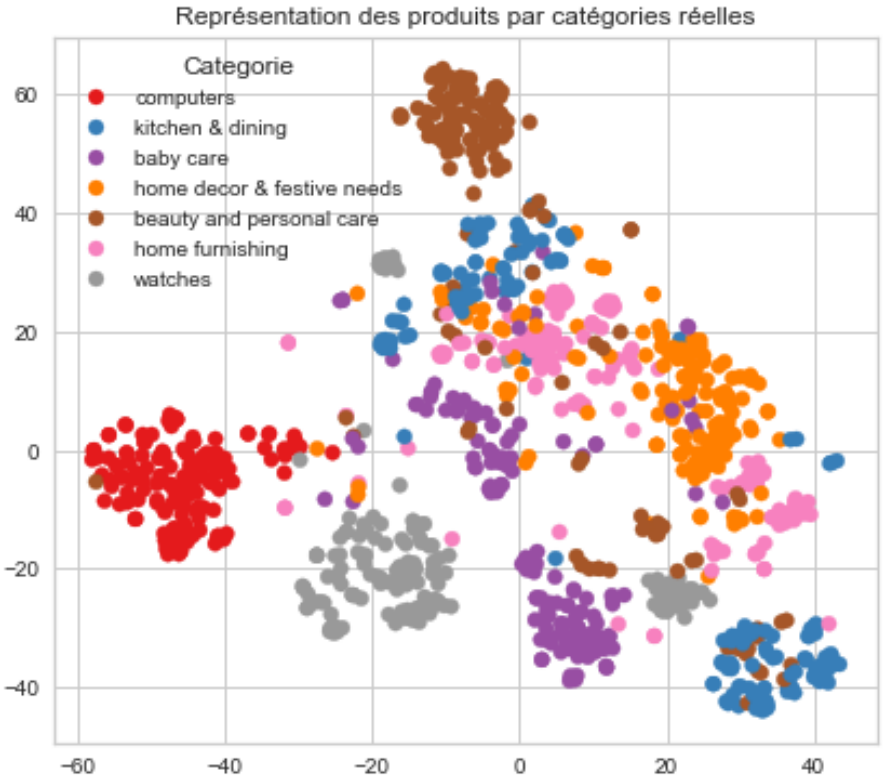
1) Extraction de features (base model (uncased))

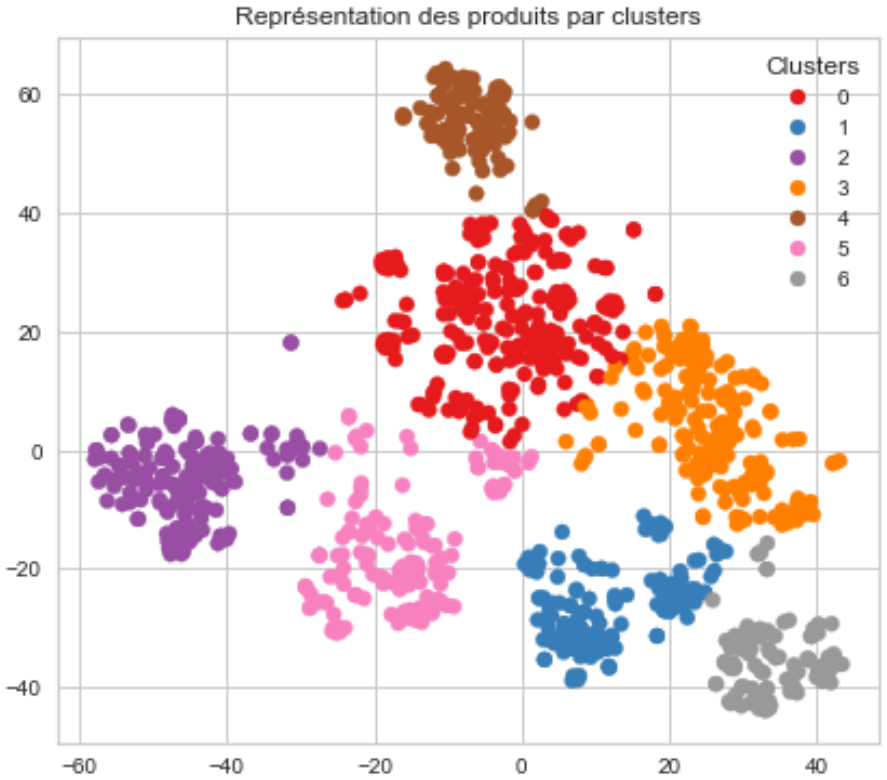
Ce modèle est pré-entraîné sur des données textuelles brutes. L'algorithme apprend les relations contextuelles entre les mots dans une phrase/un texte . On génère les 'embeddings' qui capturent en quelque sorte les principales caractéristiques des textes.

2) Réduction de dimension (t-SNE)

3) Clustering (k-means avec 7 clusters)

4) ARI score





ARI : 0.4145

Le clustering a bien séparé des catégories telles que ‘beauty and personal care’ et ‘kitchen & dining’

Données textuelles. USE - Universal Sentence Encoder

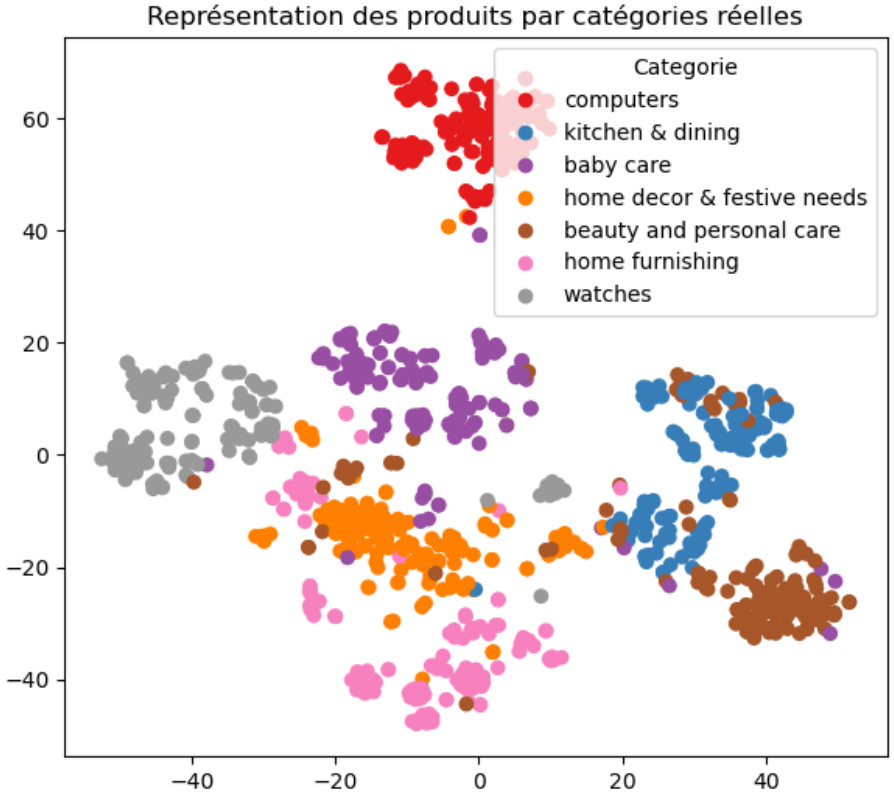
1) Extraction de features

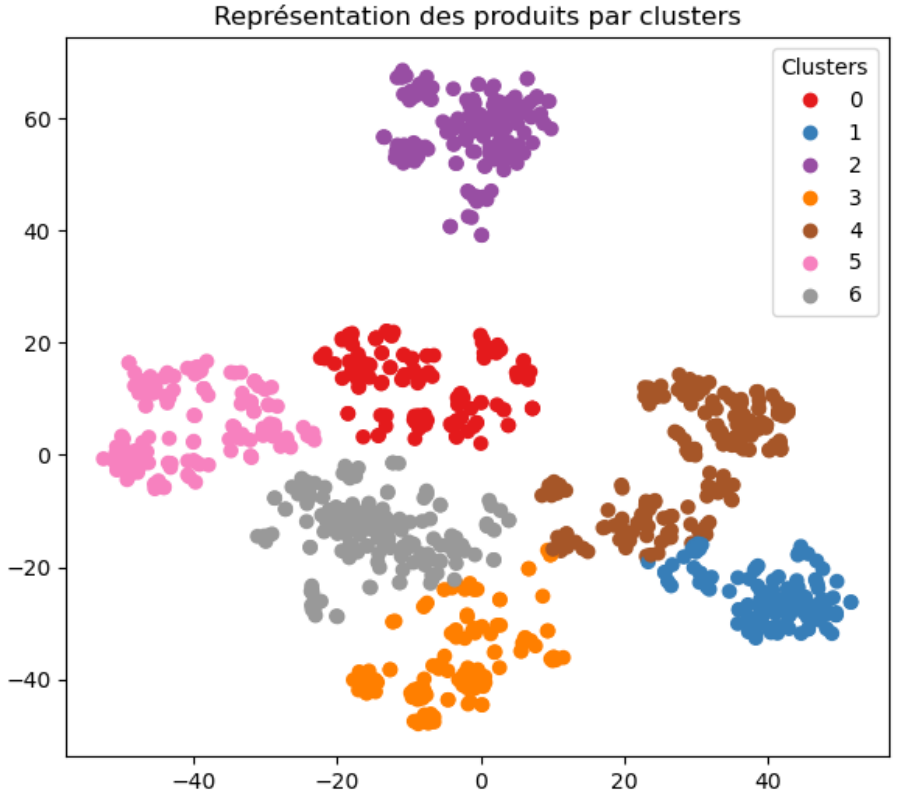
Contrairement aux techniques de word embedding dans lesquelles on représent le mot dans des vecteurs, dans Sentence Embeddings, la phrase ou le texte entier ainsi que ses informations sémantiques sont mappés dans des vecteurs de nombres réels. Cette technique permet de comprendre et de traiter les informations utiles d'un texte entier, qui peuvent ensuite être utilisées pour mieux comprendre le contexte ou le sens de la phrase.

2) Réduction de dimension (t-SNE)

4) ARI score

3) Clustering (k-means avec 7 clusters)





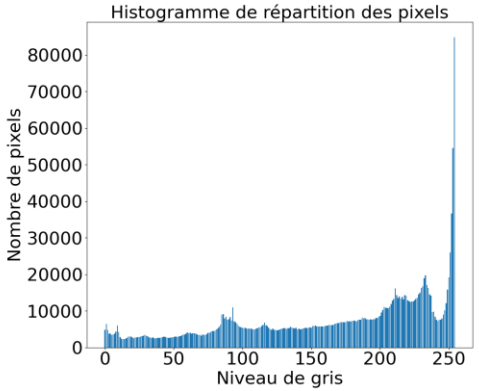
ARI : 0.6632

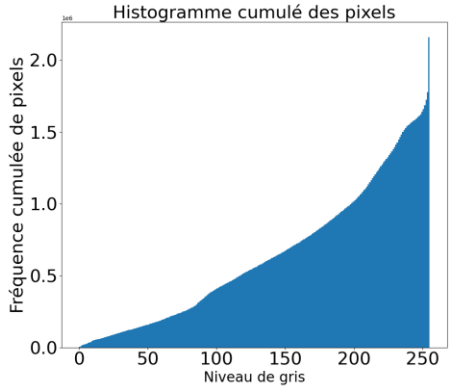
Le clustering a bien séparé la catégorie ‘computers’

Données visuelles. Prétraitement des images

On va utiliser les librairies PIL (Python Imaging Library) et OpenCV (Open Compute Vision)

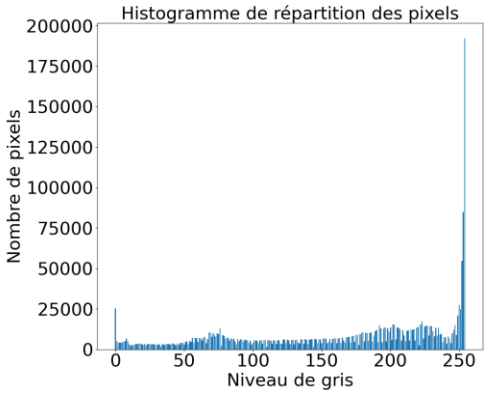
Exemple (Image de deodorant):

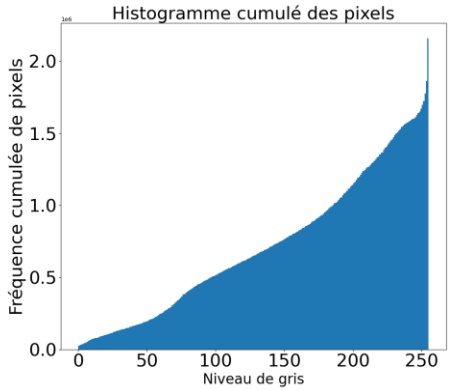




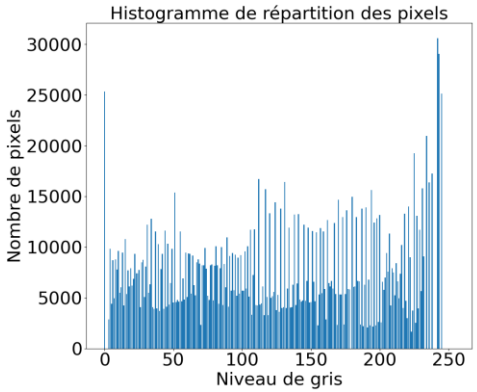
5 étapes de traitement:

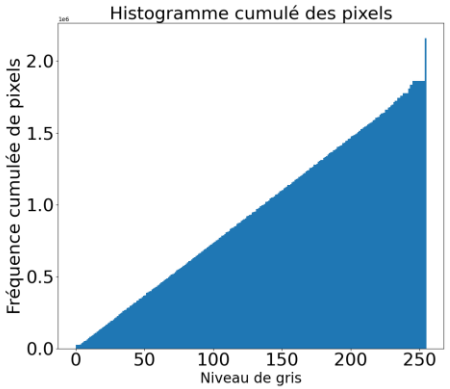
1) Correction d'exposition





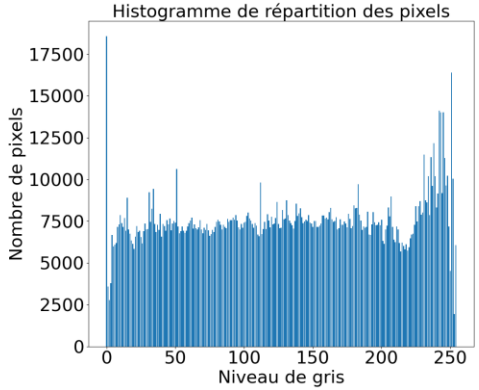
2) Correction du contraste

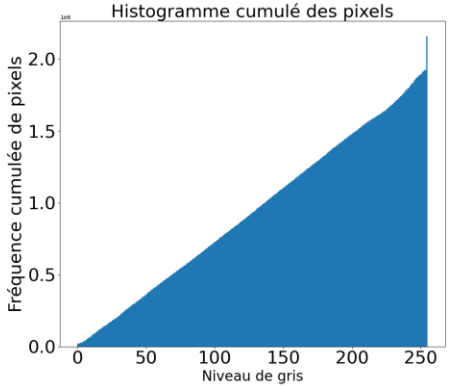




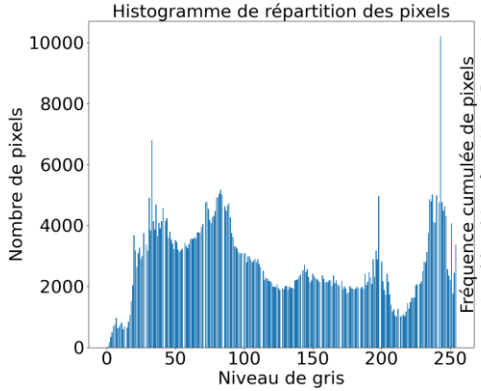
Données visuelles. Prétraitement des images

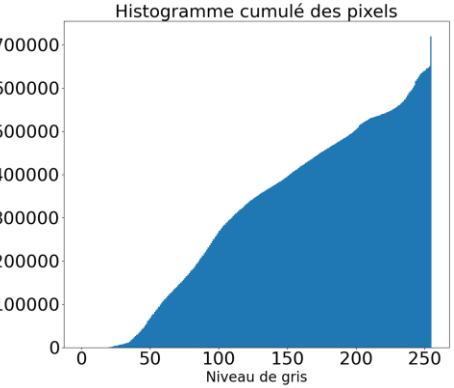
3) Filtre ou réduction de bruit



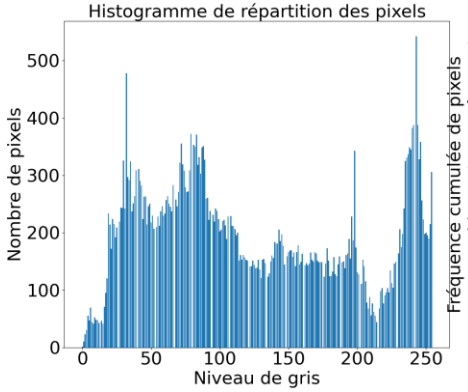


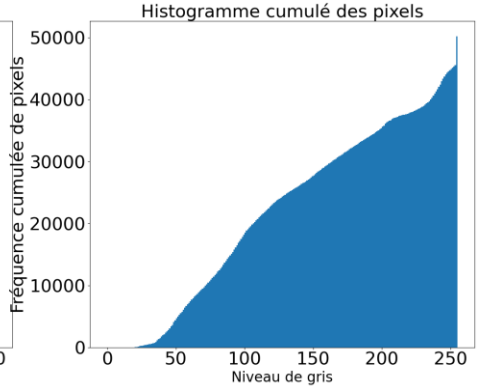
4) Conversion en niveau de gris





5) Redimensionnement (en 224*224)





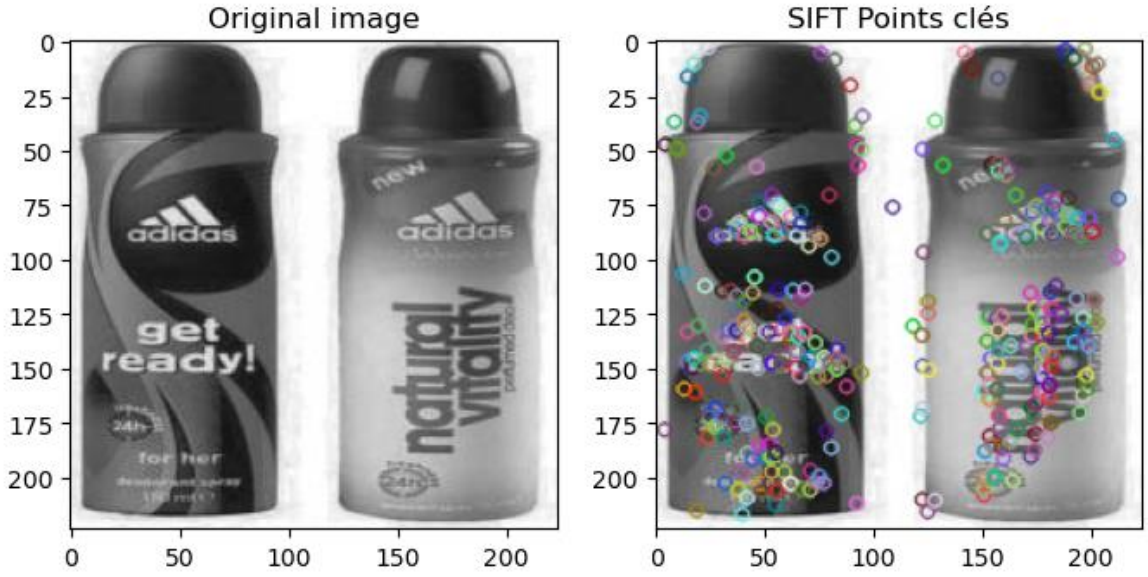
Données visuelles. SIFT

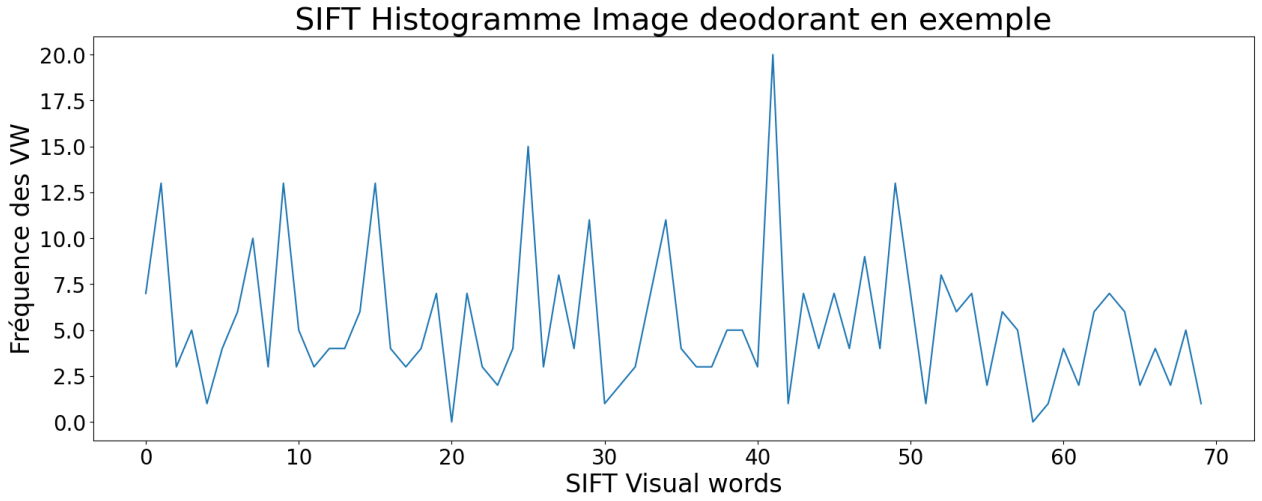
1) Extraction de features

Récupération des descripteurs de l'image

Clustering de l'ensemble de descripteurs et identification des centers (utilisées comme vocabulaire du dictionnair visuel)

Construction de l'histogramme de l'image (Bag of Visual Words)



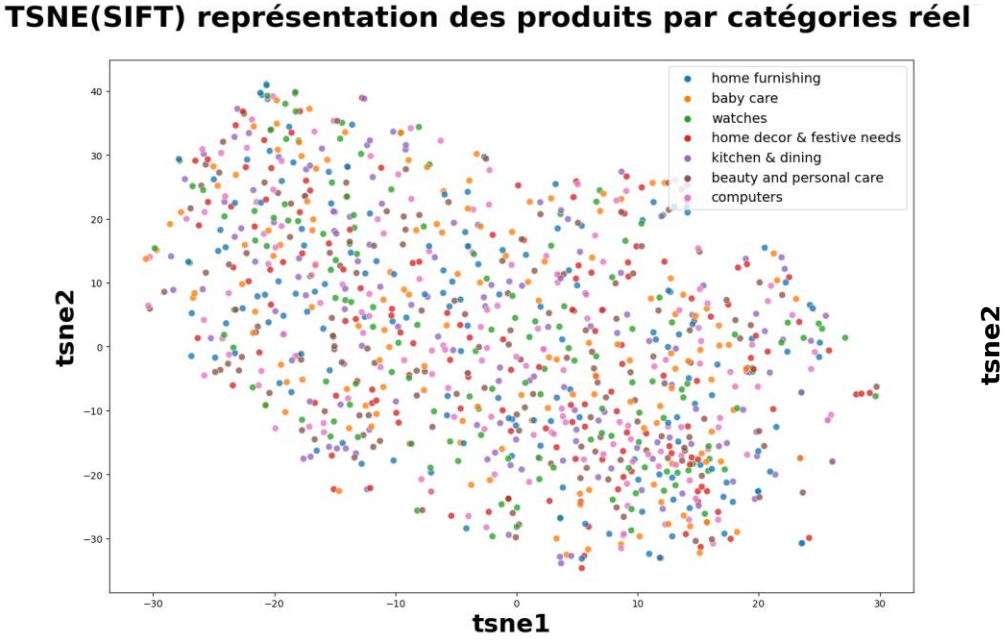


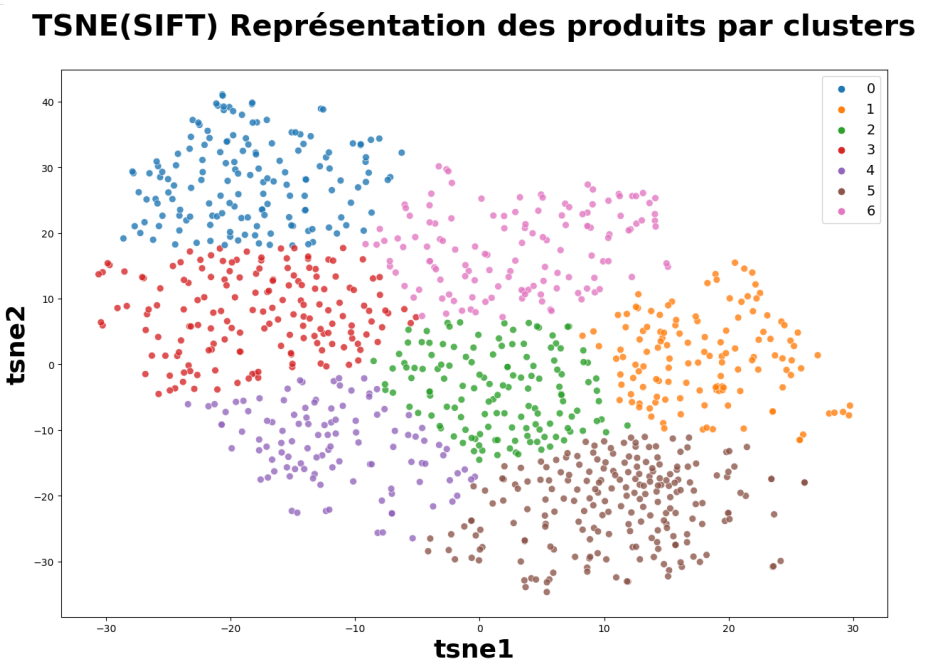
Données visuelles. SIFT

2) Réduction de dimension (t-SNE)

3) Clustering (k-means avec 7 clusters)

4) ARI score





Le score ARI est très bas et la separation des clusters n'est pas bonne.

La classification avec les données images pourrait certainement être améliorée en utilisant un algorithme de type Transfer Learning with Convolutional Neural Networks (CNN)

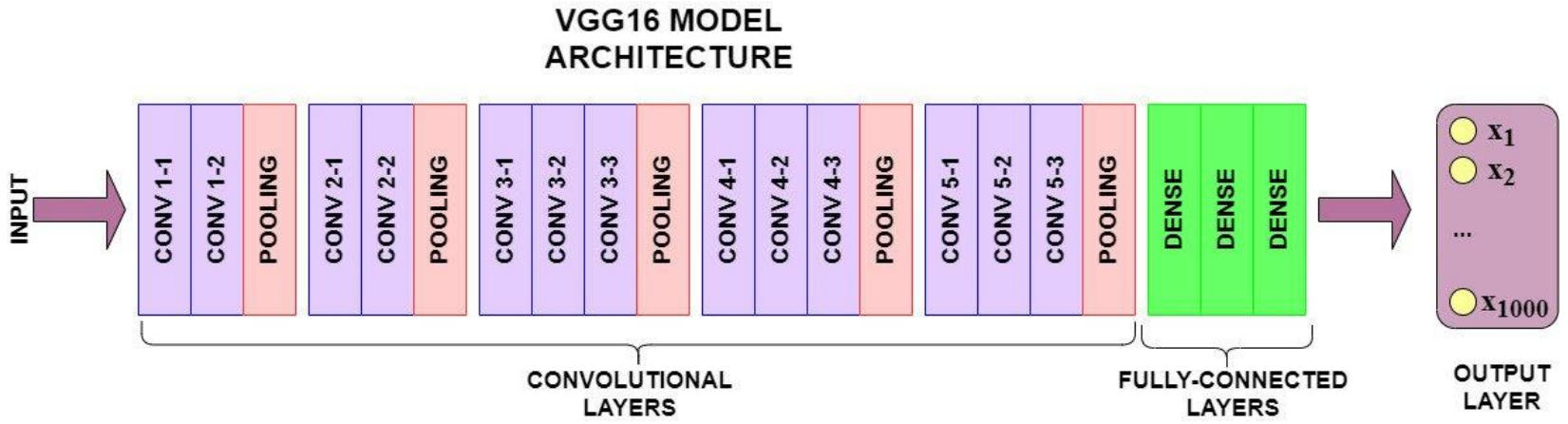
ARI : 0.0010

Données visuelles. Transfer learning algorithme (VGG-16)

Transfer learning est une méthode qui utilise un modèle qui a été pré-entraîné sur des données à grande echelle. La tâche de Transfer Learning algorithme est de former un convolutional neural network (**CNN**) . CNN est une sorte de réseau neuronal multicouche, conçu pour identifier des objets directement à partir d'images en pixels.

1) Extraction de features

VGG-16 st le modèle de plusieurs niveaux:



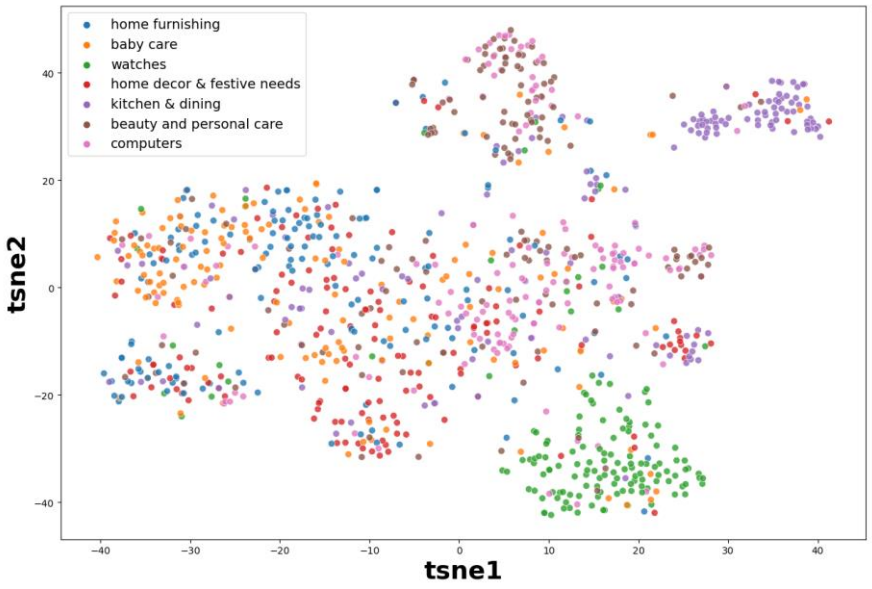
2) Réduction de dimension (t-SNE)

3) Clustering (k-means avec 7 clusters)

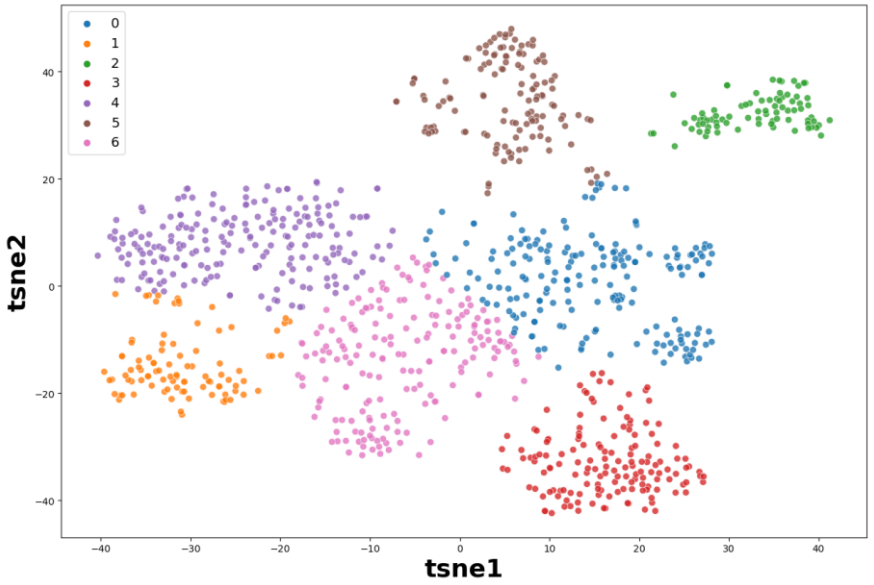
4) ARI score

Données visuelles. Transfer learning algorithme (VGG-16)

TSNE (VGG-16) représentation des produits par catégories réelles



TSNE (VGG-16) Représentation des produits par clusters



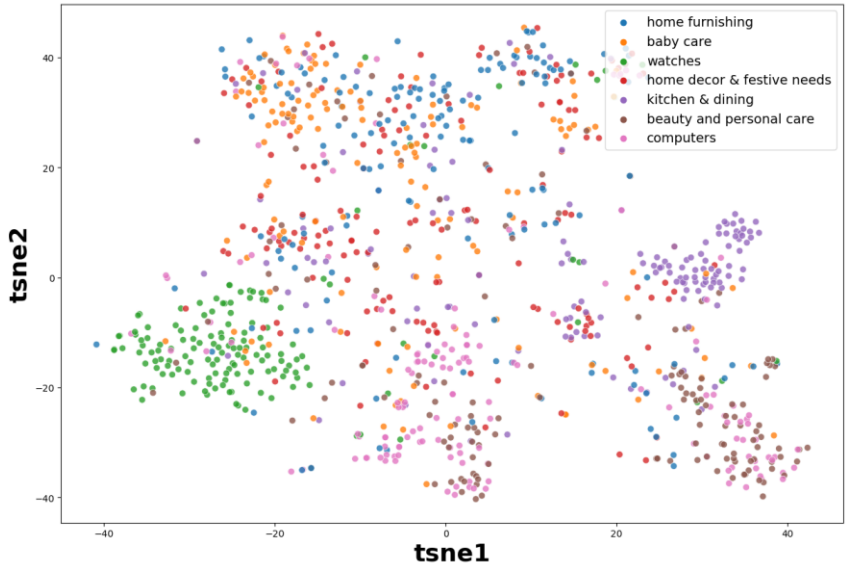
ARI : 0.4884

Le score ARI a augmenté. Le clustering a bien séparé la catégorie ‘kitchen & dining’

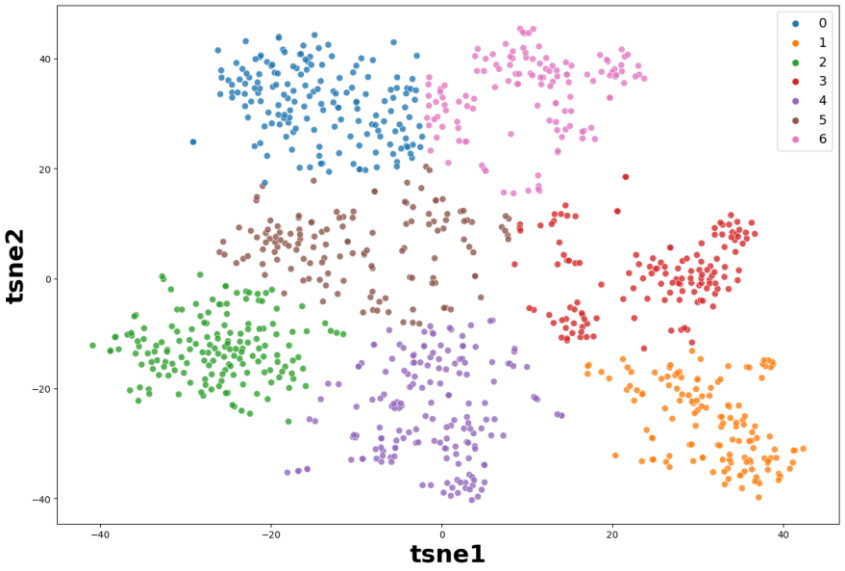
Données visuelles. Transfer learning algorithme (ResNet50)

ResNet50 est un autre algorithme de Transfer Learning

TSNE (ResNet50) représentation des produits par catégories réelles



TSNE (ResNet50) Représentation des produits par clusters



ARI : 0.5571

Le score ARI a encore augmenté. Le clustering a bien séparé les catégories ‘kitchen & dining’, ‘beauty and personal care’

Conclusions:

• On peut obtenir une meilleure caractérisation des données textuelles en choisissant un algorithme pré-entrainé (USE).

• On peut obtenir une meilleure caractérisation des données d'image en choisissant un algorithme pré-entrainé (ResNet50).

• En général, l'utilisation des algorithms de clustering des données textuelles montre une meilleure performance par rapport aux données visuelles.

• On peut améliorer les performances des features textuelles en adaptant la liste des mots vides au vocabulaire du e-commerce.

• Même si les résultats restent à améliorer, on est en mesure de montrer qu'il est possible de prédire les catégories grâce aux éléments textuels descriptifs ou grâce aux image chargées par le vendeur.

	Features	ARI
text	**Bag of Words**	0.4329
text	**Tf-idf**	0.5927
text	**Word2Vec**	0.1938
text	**BERT**	0.4145
text	**USE**	0.6632
image	**SIFT**	0.0010
image	**Transfer learning VGG-16**	0.4884
image	**Transfer learning ResNet50**	0.5571