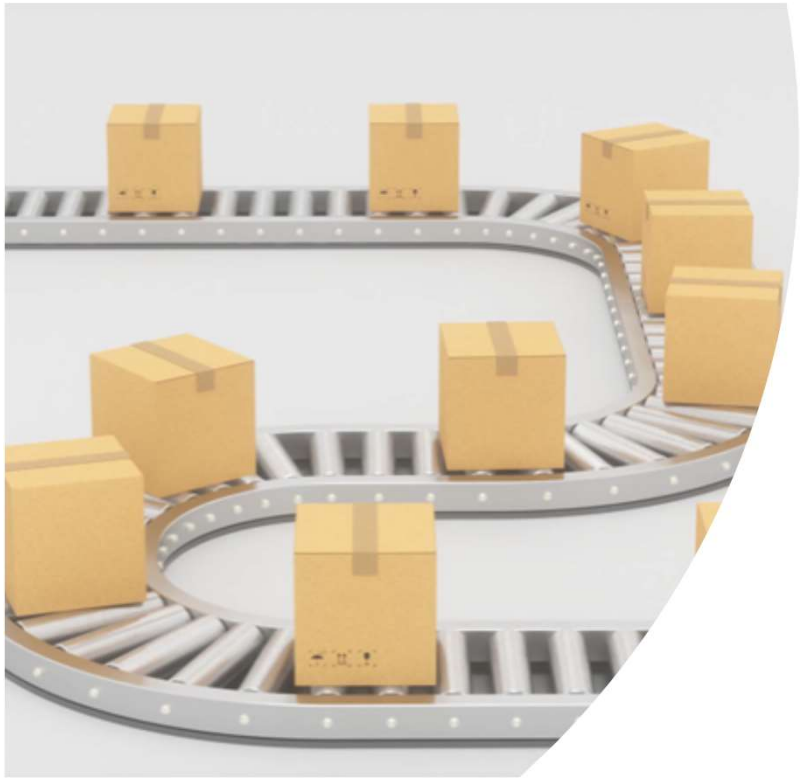



# Classification automatique de biens de consommation

Etude de faisabilité

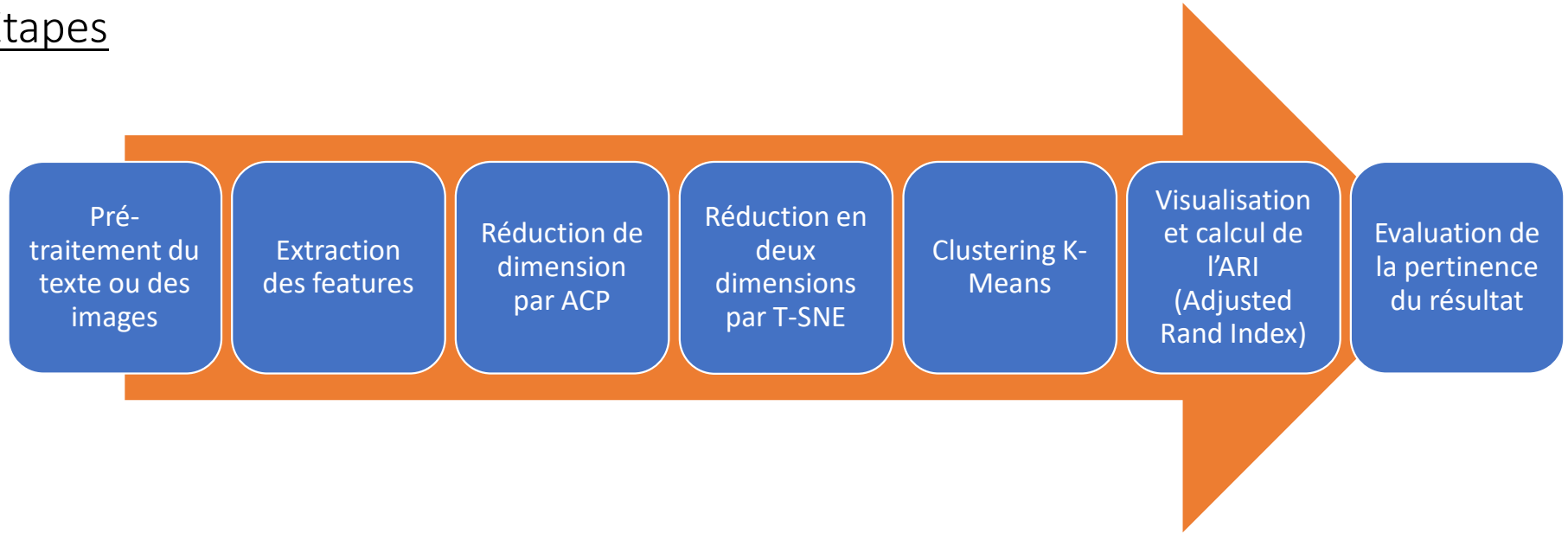
# Problématique

Automatisation de l'attribution de la catégorie d'un article

## Objectif

- Réalisation d'une première étude de faisabilité d'un moteur de classification d'articles, basé sur une image et une description

## Etapes



# Le jeu de données

## Sélection des colonnes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 16 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   uniq_id                   1050 non-null   object
 1   crawl_timestamp           1050 non-null   object
 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
 15  isin                      1050 non-null   bool
dtypes: bool(2), float64(2), object(12)
memory usage: 117.0+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   product_name              1050 non-null   object
 1   product_category_tree     1050 non-null   object
 2   image                     1050 non-null   object
 3   description               1050 non-null   object
 4   product_specifications    1050 non-null   object
dtypes: object(5)
memory usage: 41.1+ KB
```

- Extraction des catégories principales
- Enrichissement de la description par la spécification si besoin

# Jeu de données filtré

Le corpus de base (description)  contient 498561 caractères

| | product_name | product_category_tree | image | description | product_specifications |
|---|---|---|---|---|---|
| **0** | Elegance Polyester Multicolor Abstract Eyelet ... | ["Home Furnishing >> Curtains & Accessories >>... | 55b85ea15a1536d46b7190ad6fff8ce7.jpg | Key Features of Elegance Polyester Multicolor ... | {"product_specification"=>[{"key"=>"Brand", "v... |
| **1** | Sathiyas Cotton Bath Towel | ["Baby Care >> Baby Bath & Skin >> Baby Bath T... | 7b72c92c2f6c40268628ec5f14c6d590.jpg | Specifications of Sathiyas Cotton Bath Towel (... | {"product_specification"=>[{"key"=>"Machine Wa... |
| **2** | Eurospa Cotton Terry Face Towel Set | ["Baby Care >> Baby Bath & Skin >> Baby Bath T... | 64d5d4a258243731dc7bbb1eef49ad74.jpg | Key Features of Eurospa Cotton Terry Face Towe... | {"product_specification"=>[{"key"=>"Material",... |
| **3** | SANTOSH ROYAL FASHION Cotton Printed King size... | ["Home Furnishing >> Bed Linen >> Bedsheets >>... | d4684dcdc759dd9cdf41504698d737d8.jpg | Key Features of SANTOSH ROYAL FASHION Cotton P... | {"product_specification"=>[{"key"=>"Brand", "v... |
| **4** | Jaipur Print Cotton Floral King sized Double B... | ["Home Furnishing >> Bed Linen >> Bedsheets >>... | 6325b6870c54cd47be6ebfbffa620ec7.jpg | Key Features of Jaipur Print Cotton Floral Kin... | {"product_specification"=>[{"key"=>"Machine Wa... |
| **...** | ... | ... | ... | ... | ... |
| **1045** | Oren Empower Extra Large Self Adhesive Sticker | ["Baby Care >> Baby & Kids Gifts >> Stickers >... | 958f54f4c46b53c8a0a9b8167d9140bc.jpg | Oren Empower Extra Large Self Adhesive Sticker... | {"product_specification"=>[{"key"=>"Number of ... |
| **1046** | Wallmantra Large Vinyl Sticker Sticker | ["Baby Care >> Baby & Kids Gifts >> Stickers >... | fd6cbcc22efb6b761bd564c28928483c.jpg | Wallmantra Large Vinyl Sticker Sticker (Pack o... | {"product_specification"=>[{"key"=>"Number of ... |
| **1047** | Uberlyfe Extra Large Pigmented Polyvinyl Films... | ["Baby Care >> Baby & Kids Gifts >> Stickers >... | 5912e037d12774bb73a2048f35a00009.jpg | Buy Uberlyfe Extra Large Pigmented Polyvinyl F... | {"product_specification"=>[{"key"=>"Number of ... |
| **1048** | Wallmantra Medium Vinyl Sticker Sticker | ["Baby Care >> Baby & Kids Gifts >> Stickers >... | c3edc504d1b4f0ba6224fa53a43a7ad6.jpg | Buy Wallmantra Medium Vinyl Sticker Sticker fo... | {"product_specification"=>[{"key"=>"Number of ... |
| **1049** | Uberlyfe Large Vinyl Sticker | ["Baby Care >> Baby & Kids Gifts >> Stickers >... | f2f027ad6a6df617c9f125173da71e44.jpg | Buy Uberlyfe Large Vinyl Sticker for Rs.595 on... | {"product_specification"=>[{"key"=>"Sales Pack... |

1050 rows × 5 columns

# Catégories

## Et spécifications

Il y a 7 catégories distinctes : Home Furnishing, Baby Care, Watches, Home Decor & Festive Needs, Kitchen & Dining, Beauty and Personal Care, Computers ainsi que 62 sous-catégories

< Product Name >

Maserati Time R8851116001 Analog Watch - For Boys

< Description >

Maserati Time R8851116001 Analog Watch - For Boys - Buy Maserati Time R8851116001 Analog Watch - For Boys R8851116001 Online at Rs.24400 in India Only at Flipkart.com. - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!

< Product specifications >

{"product_specification"=>[{"key"=>"Chronograph", "value"=>"Yes"}, {"key"=>"Date Display", "value"=>"Yes, Day and Date Display"}, {"key"=>"Altimeter", "value"=>"No"}, {"key"=>"Barometer", "value"=>"No"}, {"key"=>"Alarm Clock", "value"=>"NO"}, {"key"=>"Compass", "value"=>"NO"}, {"key"=>"Calendar", "value"=>"No"}, {"key"=>"Luminous", "value"=>"No"}, {"key"=>"Type", "value"=>"Analog"}, {"key"=>"Style Code", "value"=>"R8851116001"}, {"key"=>"Ideal For", "value"=>"Boys"}, {"key"=>"Occasion", "value"=>"Party-Wedding"}, {"value"=>"1 Year Brand Warranty"}, {"key"=>"Dial Shape", "value"=>"Oval"}, {"key"=>"Strap Color", "value"=>"Green"}, {"key"=>"Scratch Resistant", "value"=>"No"}, {"key"=>"Water Resistant", "value"=>"Yes"}, {"key"=>"Dial Color", "value"=>"Grey"}]}

< Product category tree >

["Watches >> Wrist Watches >> Maserati Time Wrist Watches"]

Les éléments de `product_specifications` sont repris dans 742 descriptions (sur 1050 au total)

< Product Name >

Elegance Polyester Multicolor Abstract Eyelet Door Curtain

< Description >

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

< Product specifications >

{"product_specification"=>[{"key"=>"Brand", "value"=>"Elegance"}, {"key"=>"Designed For", "value"=>"Door"}, {"key"=>"Type", "value"=>"Eyelet"}, {"key"=>"Model Name", "value"=>"Abstract Polyester Door Curtain Set Of 2"}, {"key"=>"Model ID", "value"=>"Duster25"}, {"key"=>"Color", "value"=>"Multicolor"}, {"key"=>"Length", "value"=>"213 cm"}, {"key"=>"Number of Contents in Sales Package", "value"=>"Pack of 2"}, {"key"=>"Sales Package", "value"=>"2 Curtains"}, {"key"=>"Material", "value"=>"Polyester"}]}

< Product category tree >

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

# Jeu de données final

## Avec les catégories principales et les descriptions enrichies

Le nouveau corpus, avec les descriptions enrichies, contient **734211** caractères

| | product_name | image | description | enriched_description | main_category |
|---|---|---|---|---|---|
| **0** | Elegance Polyester Multicolor Abstract Eyelet ... | 55b85ea15a1536d46b7190ad6fff8ce7.jpg | Key Features of Elegance Polyester Multicolor ... | Key Features of Elegance Polyester Multicolor ... | Home Furnishing |
| **1** | Sathiyas Cotton Bath Towel | 7b72c92c2f6c40268628ec5f14c6d590.jpg | Specifications of Sathiyas Cotton Bath Towel (... | Specifications of Sathiyas Cotton Bath Towel (... | Baby Care |
| **2** | Eurospa Cotton Terry Face Towel Set | 64d5d4a258243731dc7bbb1eef49ad74.jpg | Key Features of Eurospa Cotton Terry Face Towe... | Key Features of Eurospa Cotton Terry Face Towe... | Baby Care |
| **3** | SANTOSH ROYAL FASHION Cotton Printed King size... | d4684dcdc759dd9cdf41504698d737d8.jpg | Key Features of SANTOSH ROYAL FASHION Cotton P... | Key Features of SANTOSH ROYAL FASHION Cotton P... | Home Furnishing |
| **4** | Jaipur Print Cotton Floral King sized Double B... | 6325b6870c54cd47be6ebfbffa620ec7.jpg | Key Features of Jaipur Print Cotton Floral Kin... | Key Features of Jaipur Print Cotton Floral Kin... | Home Furnishing |
| **...** | ... | ... | ... | ... | ... |
| **1045** | Oren Empower Extra Large Self Adhesive Sticker | 958f54f4c46b53c8a0a9b8167d9140bc.jpg | Oren Empower Extra Large Self Adhesive Sticker... | Oren Empower Extra Large Self Adhesive Sticker... | Baby Care |
| **1046** | Wallmantra Large Vinyl Sticker Sticker | fd6cbcc22efb6b761bd564c28928483c.jpg | Wallmantra Large Vinyl Sticker Sticker (Pack o... | Wallmantra Large Vinyl Sticker Sticker (Pack o... | Baby Care |
| **1047** | Uberlyfe Extra Large Pigmented Polyvinyl Films... | 5912e037d12774bb73a2048f35a00009.jpg | Buy Uberlyfe Extra Large Pigmented Polyvinyl F... | Buy Uberlyfe Extra Large Pigmented Polyvinyl F... | Baby Care |
| **1048** | Wallmantra Medium Vinyl Sticker Sticker | c3edc504d1b4f0ba6224fa53a43a7ad6.jpg | Buy Wallmantra Medium Vinyl Sticker Sticker fo... | Buy Wallmantra Medium Vinyl Sticker Sticker fo... | Baby Care |
| **1049** | Uberlyfe Large Vinyl Sticker | f2f027ad6a6df617c9f125173da71e44.jpg | Buy Uberlyfe Large Vinyl Sticker for Rs.595 on... | Buy Uberlyfe Large Vinyl Sticker for Rs.595 on... | Baby Care |

1050 rows × 6 columns

# Le texte

# Pré-traitement du corpus

✓ Longueur du corpus : 734210 caractères

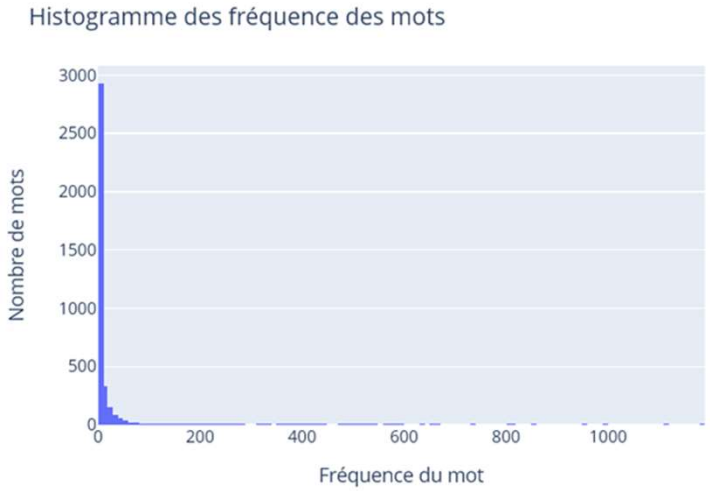
Pré-traitement basique

- passage en **minuscules** -> suppression des **caractères non_ascii** -> **tokenisation**

✓ Longueur du corpus : 734042 caractères, nombre de tokens total 120015 | uniques 7144

Complément pour pré-traitement avancé

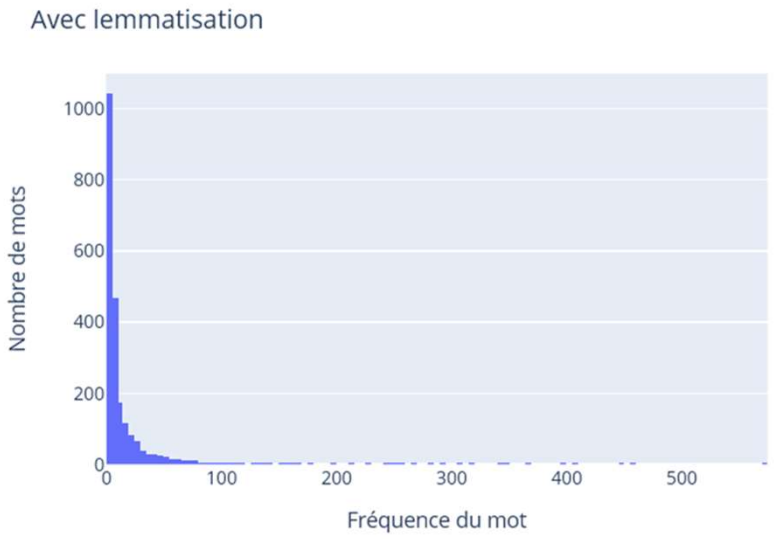
- suppression des **stopwords**
- suppression des **mots trop courts**
- conservation des mots avec des **caractères alphabétiques uniquement**
- conservation des **noms communs** et noms propres uniquement




✓ Longueur du corpus : 470572 caractères, nombre de tokens total 68062 | uniques 3775

# Pré-traitement des documents

- même étapes que pour le corpus précédemment et…
- suppression des mots **les moins fréquents**
- suppression des mots **les plus fréquents sans valeur descriptive** (liste adaptée avec une **approche métier**)
- **lemmatisation** ou racinisation (stemming)




Longueur finale du corpus (lemmatisation) : 233130 caractères (contre 734210 au tout début)

Les mots les plus fréquents dans le corpus :

| | |
|---|---|
| type | 1181 |
| specifications | 1112 |
| color | 997 |
| material | 951 |
| model | 855 |
| package | 816 |
| number | 814 |
| sales | 808 |
| warranty | 730 |
| brand | 662 |

dtype: int64

Les mots les plus rares dans le corpus :

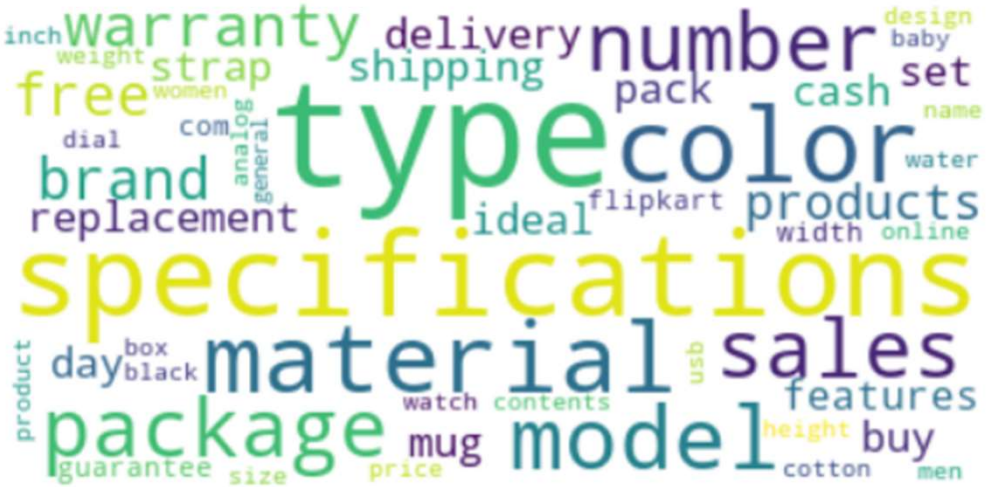
| | |
|---|---|
| straw | 1 |
| leakage | 1 |
| openings | 1 |
| alphabets | 1 |
| crush | 1 |
| motif | 1 |
| illustrations | 1 |
| word | 1 |
| bleech | 1 |
| canada | 1 |

dtype: int64

Obtenus à partir du corpus pré-traité comme précédemment décrit

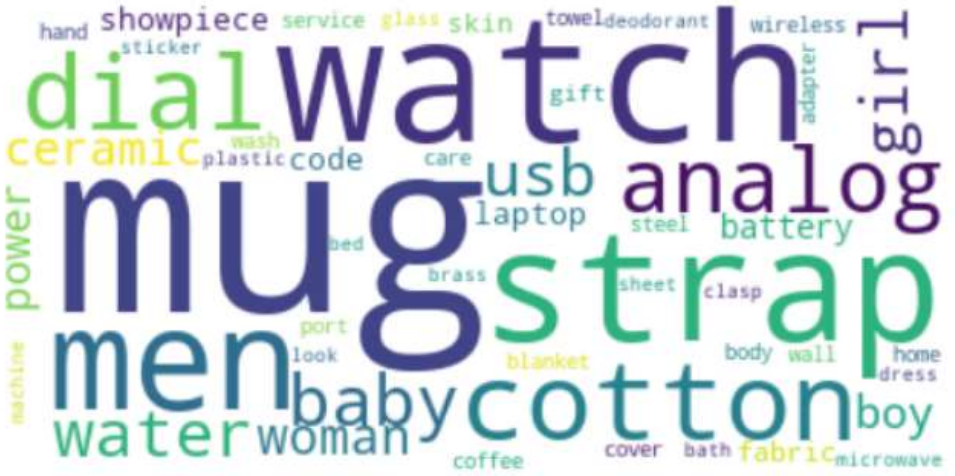
# Visualisation des nuages de mots

Le corpus initial, et le corpus obtenu à partir des documents traités




Traitement des documents

Nuage de mots, avec lemmatisation




Les mots les plus fréquents dans le corpus :

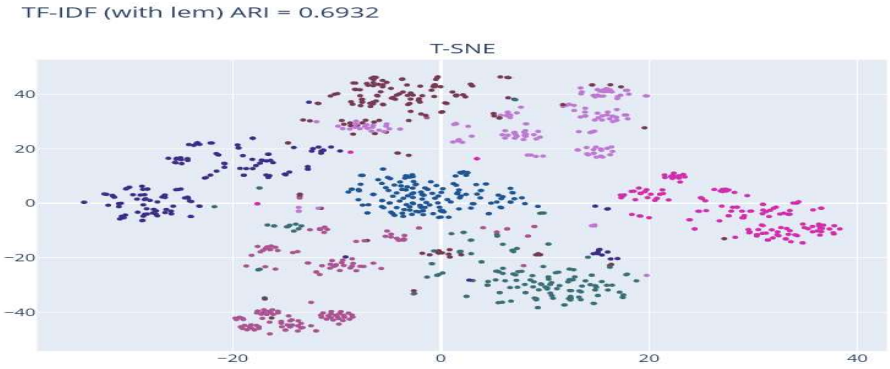
| | |
|---|---|
| type | 1181 |
| specifications | 1112 |
| color | 997 |
| material | 951 |
| model | 855 |
| package | 816 |
| number | 814 |
| sales | 808 |
| warranty | 730 |
| brand | 662 |

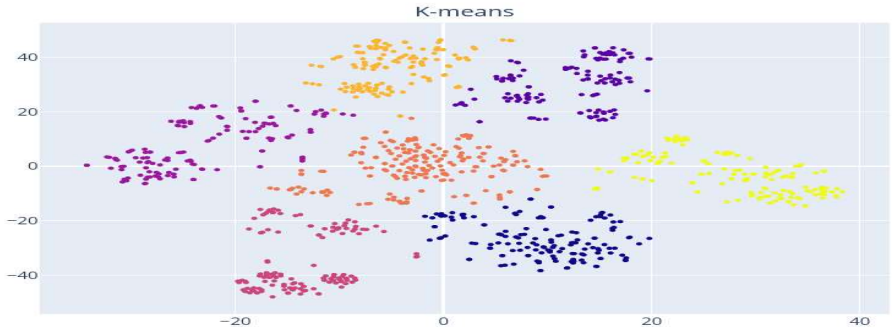
# Récapitulatif des résultats obtenus pour le texte

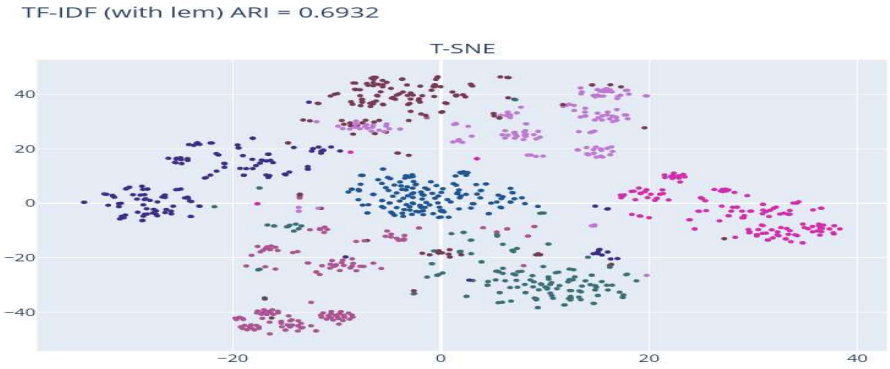
| Approche | Pré-traitement | ARI |
|---|---|---|
| Term Frequency (**Bag of words,comptage simple**) | Avancé + stemming | 0.5705 |
| | Avancé + lemmatisation | 0.4312 |
| Term Frequency – Inverse Document Frequency (**Fréquence de mots pondérée**) | Avancé + stemming | 0.6845 |
| | Avancé + lemmatisation | **0.6932** |
| Word2Vec (**Word/Sentence embedding ,** Continuous Bag of Words) | Avancé + lemmatisation | 0.5593 |
| BERT (Bidirectional Encoder Representations from Transformers) **RNN (Recurrent Neural Network), Word/Sentence embedding , ) Hugging-Face, modèle pré-entraîné bert-base-uncased** | Basique | 0.5562 |
| USE (Universal Sentence Encoder) : RNN**, Word/Sentence embedding , modèle pré-entraîné** https://tfhub.dev/google/universal-sentence-encoder/4 | Basique | **0.6849** |

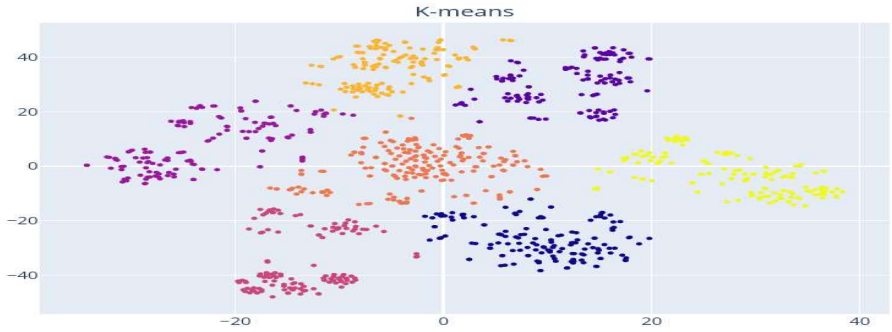
# TF-IDF, avec lemmatisation

Après extraction des features, ACP, T-SNE, K-means et visualisation



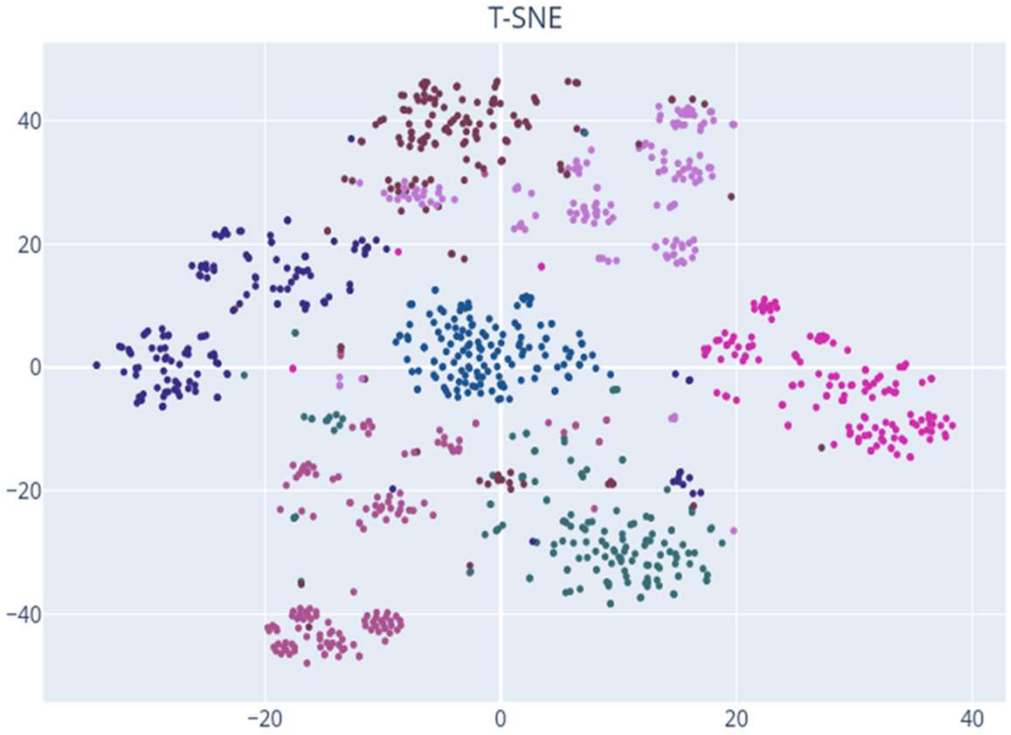



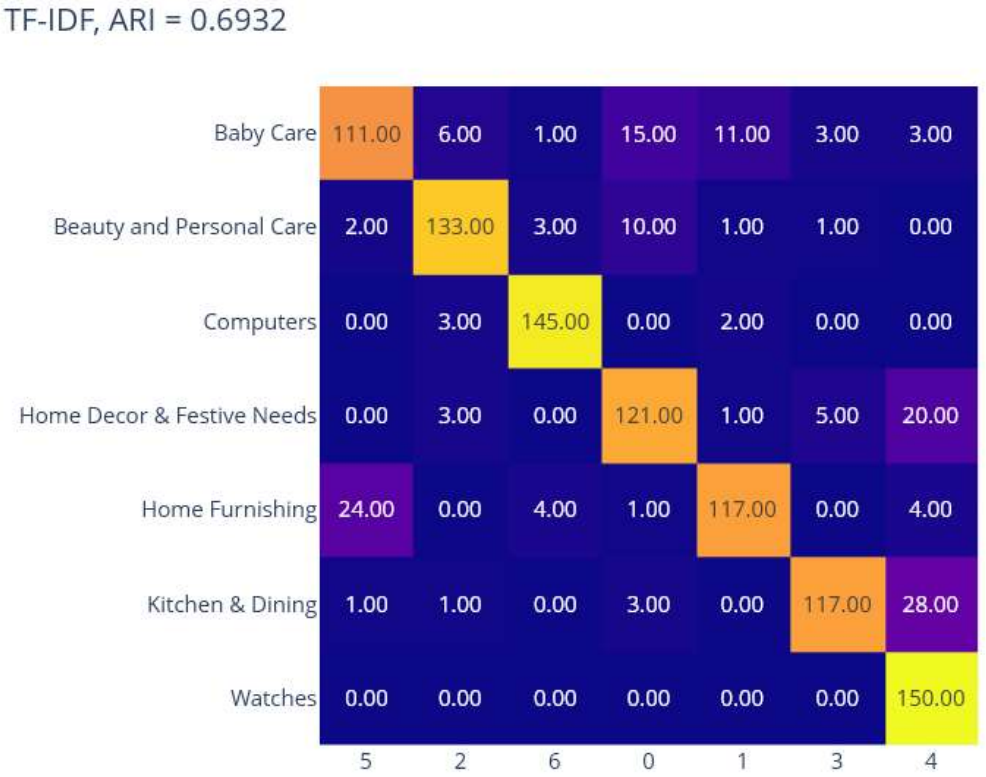



# TF-IDF

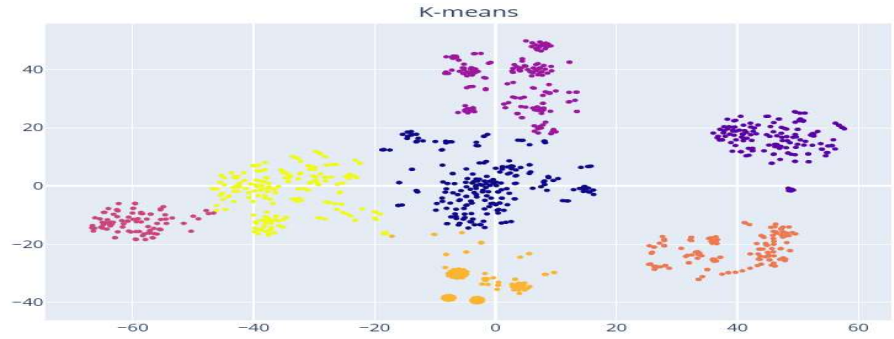
Visualisation et correspondance des clusters avec les catégories initiales

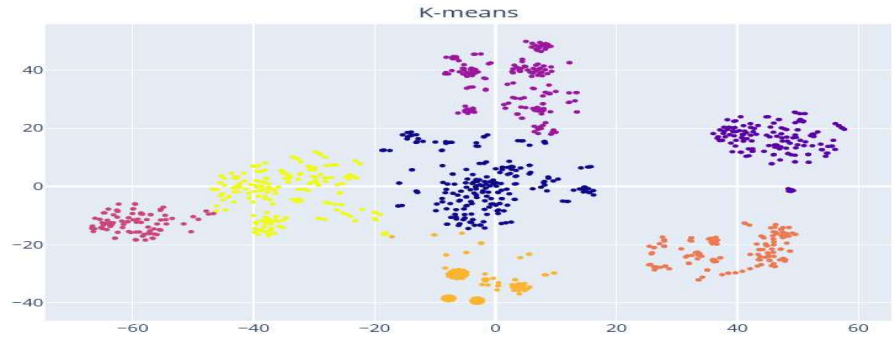



# USE (Universal Sentence Encoder)

Après extraction des features, ACP, T-SNE, K-means et visualisation

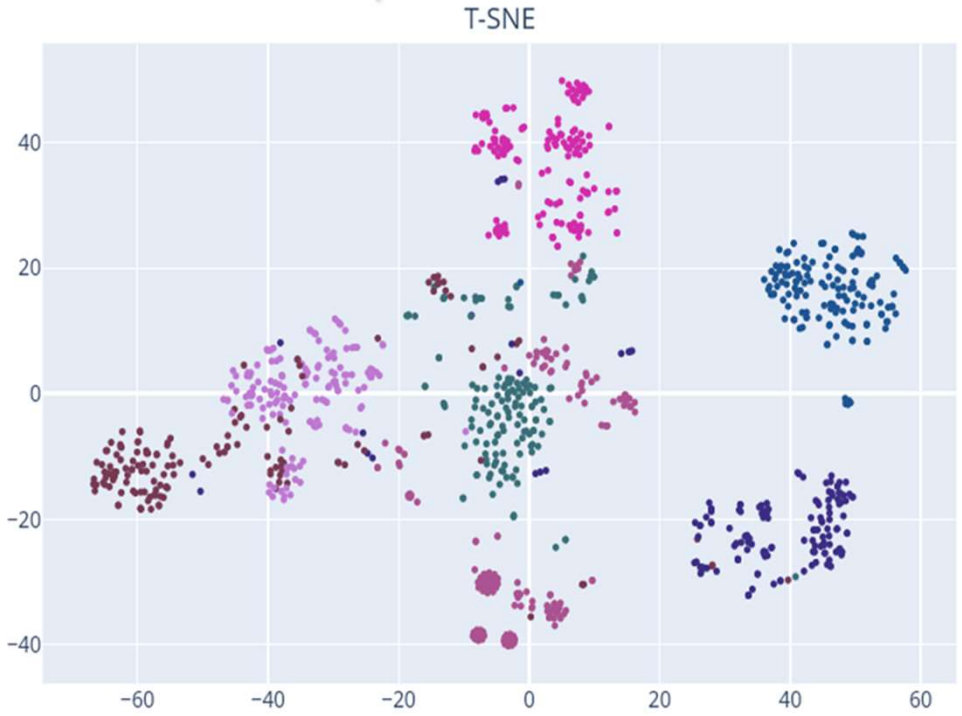



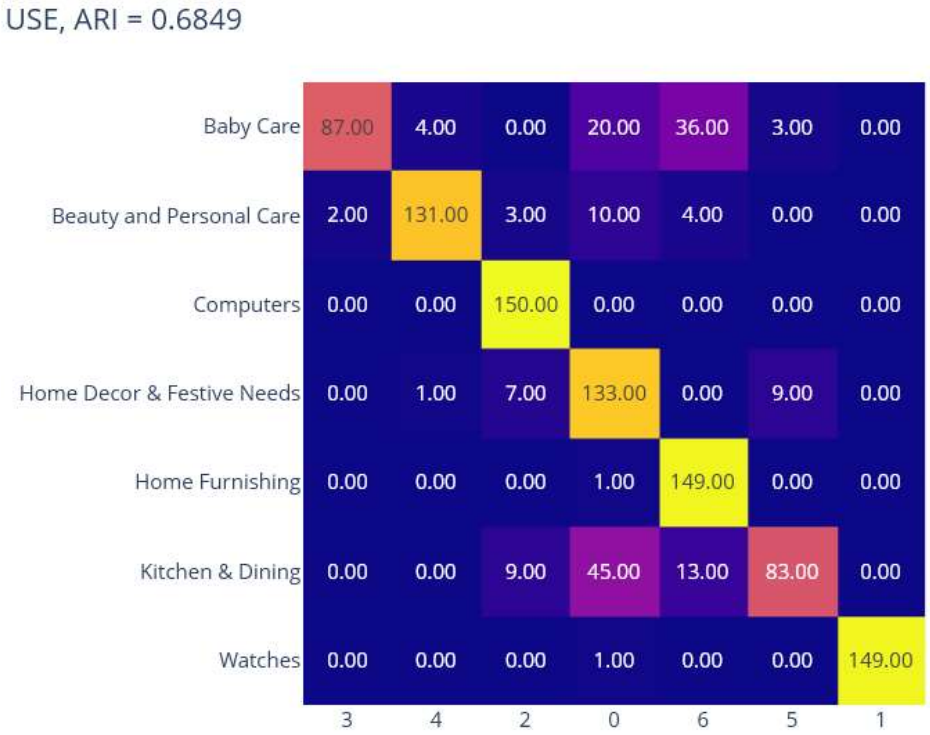




# USE

Visualisation et correspondance des clusters avec les catégories initiales

# Les images

# Pré-traitement des images

## Et réduction de dimensions du dataset

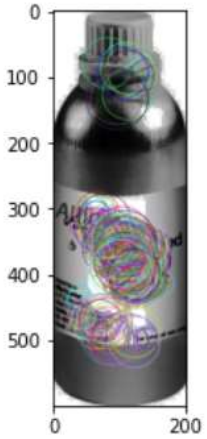
**Pour SIFT et ORB**

Redimensionnement (ratio conservé, max 600 px), niveaux de gris, égalisation des histogrammes

| | Nombre de descripteurs | Première réduction de dimensions (k-means) | ACP |
|---|---|---|---|
| SIFT | **1299247** | **1140** | **883** |
| ORB | **523598** | **724** | **689** |

Keypoints SIFT

Keypoints ORB




**Pour VGG16**

Redimensionnement en 224 x 224 avec ajout de padding blanc, niveaux de gris, égalisation des histogrammes


Dimensions du dataset en sortie de VGG16 : (1050, 4096)

Dimensions du dataset après ACP : (1050, 1020)

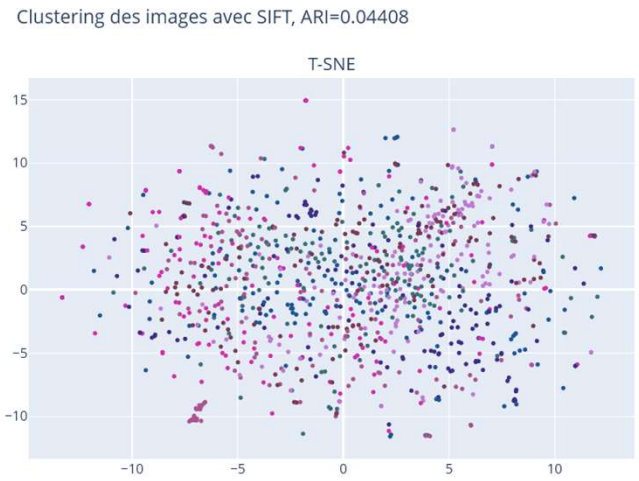

# Récapitulatif des résultats obtenus pour les images

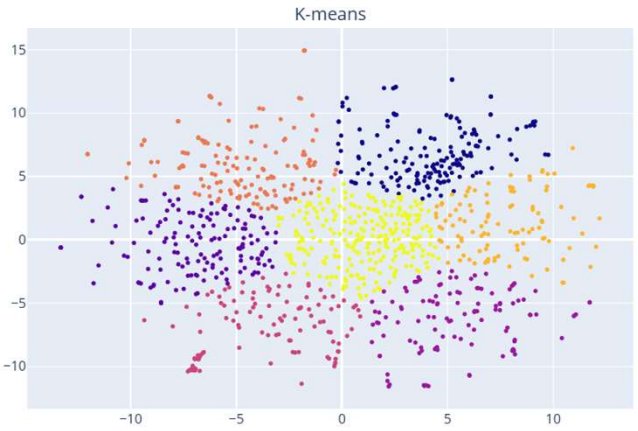
Enorme différence entre les approches « visual words » et le CNN (VGG16)

| Approche | Pré-traitement des images | ARI |
|---|---|---|
| Bag-of-features (visual words) : SIFT (Scale-invariant feature transform) | Redimensionnement (ratio conservé, max 600 px), niveaux de gris, égalisation des histogrammes | 0.0440 |
| Bag-of-features (visual words) : ORB (Oriented FAST and Rotated BRIEF) | Redimensionnement (ratio conservé, max 600 px), niveaux de gris, égalisation des histogrammes | 0.0495 |
| CNN (Convolutional neural network) pré-entraîné (VGG16) | Redimensionnement en 224 x 224 avec ajout de padding blanc, niveaux de gris, égalisation des histogrammes | **0.5107** |

# Visualisation SIFT et ORB

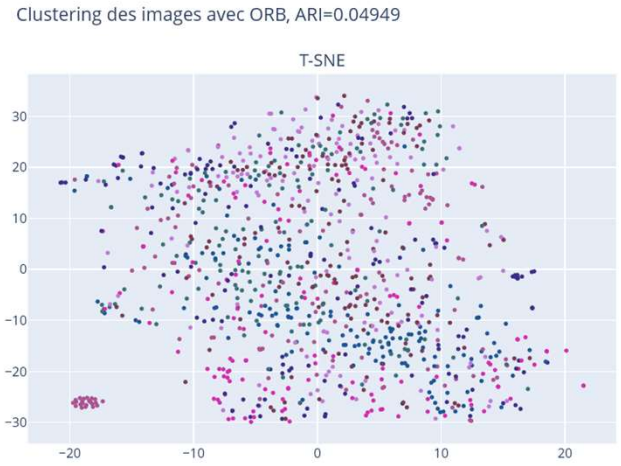
Ces approches ne sont clairement pas satisfaisantes

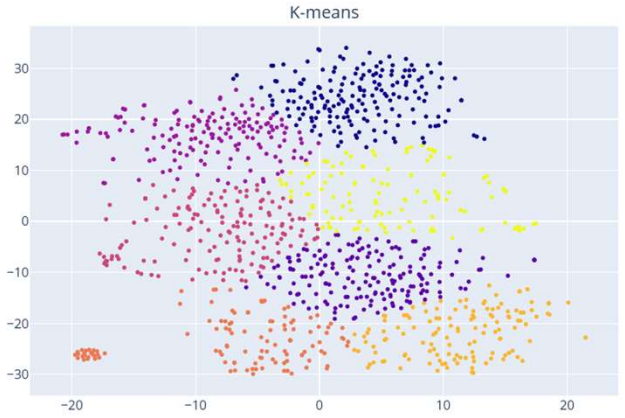



SIFT

ORB

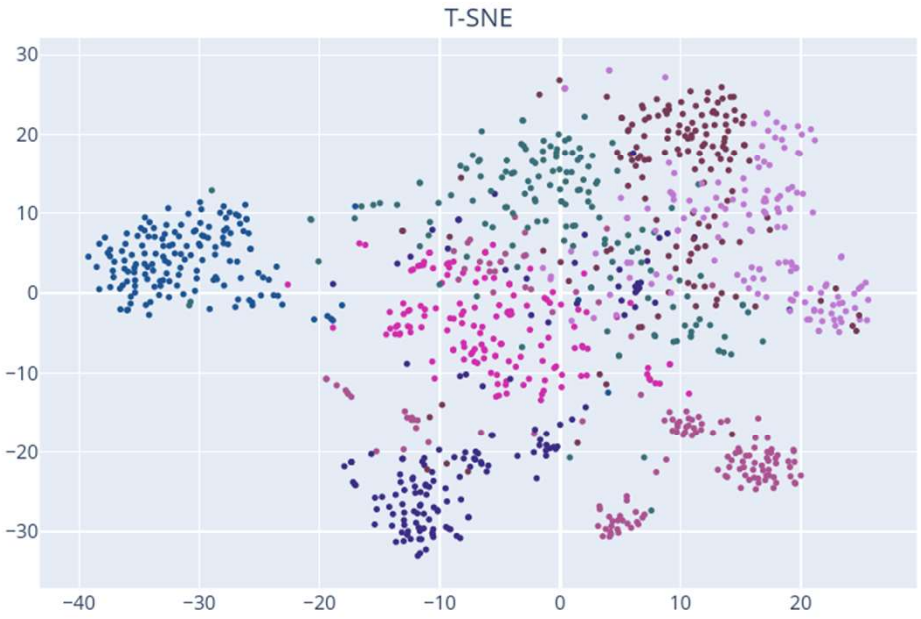


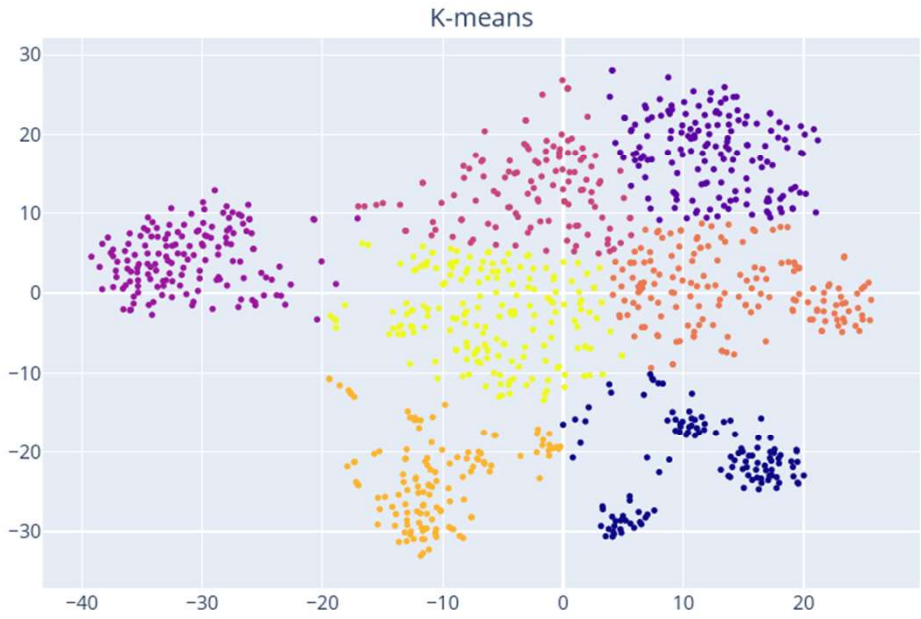



# CNN – VGG16

Après génération des features, ACP, T-SNE, K-means et visualisation

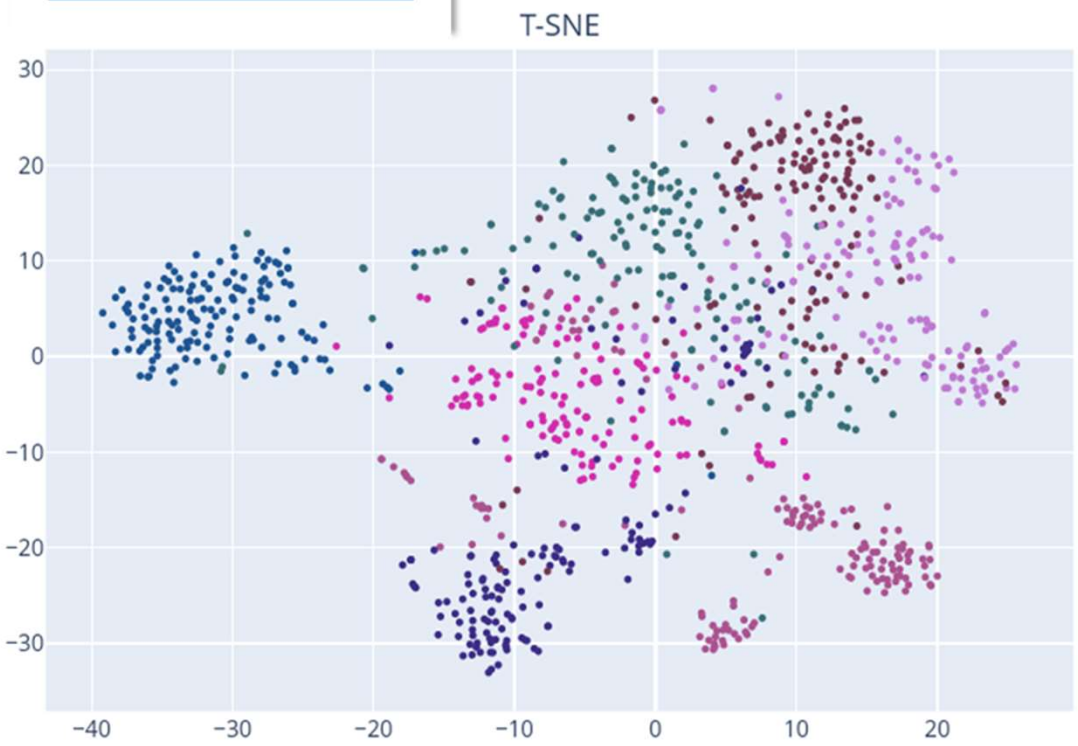
VGG16, ARI=0.5107

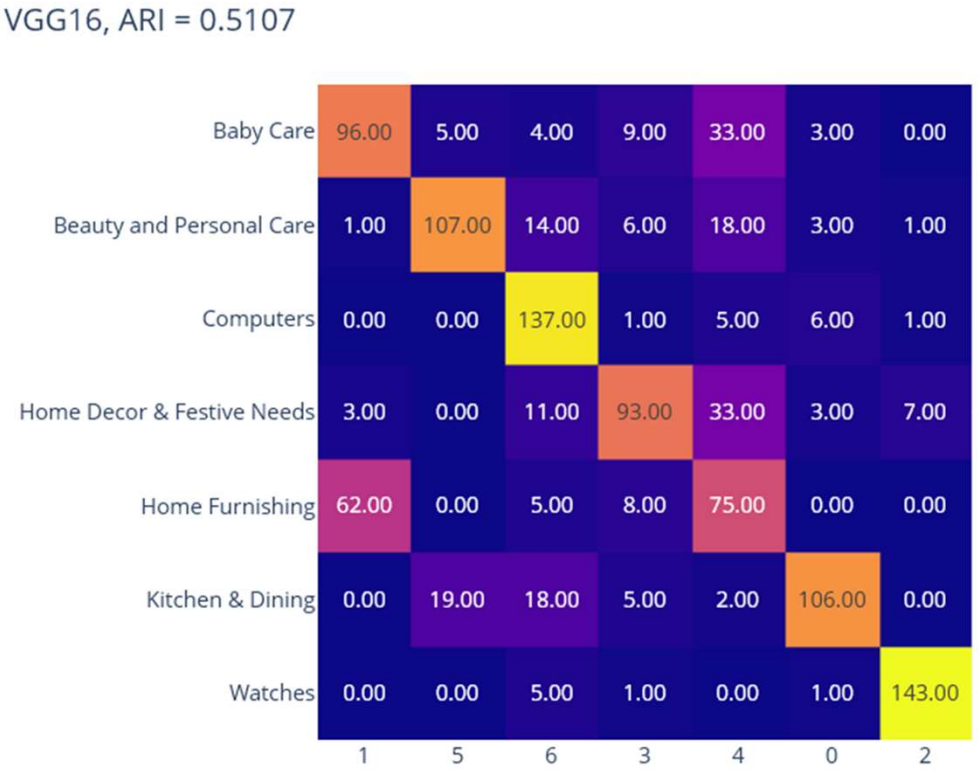


# CNN – VGG16

Visualisation et correspondance des clusters avec les catégories initiales

# Conclusion sur la faisabilité

Et recommandations

# Faisabilité

## Combinaison des résultats et verdict : POSITIF

On considère la liste des résultats des correspondances catégories/clusters pour les approches TF_IDF, Word2vec, Bert, Use et VGG16. Le résultat combiné est correct si au moins une approche donne une correspondance avec la catégorie réelle.

Dans le cas où aucune approche ne correspond à la catégorie réelle, on lui attribute le label “Wrong Guess”. On obtient ainsi un ARI de 0.9 !

La classification étudiée est donc tout à fait possible, pour peu que les descriptions et illustrations soient bien choisies, et que les pré-traitements et modèles soient optimisés si besoin.

# Recommandations, à partir d'un exemple

## La qualité de la description et le choix de l'illustration sont très importants

'Key Features of Relaxfeel Floral Single Dohar White Pack of1 Color: White Height: 10 cm Width:30 cm,Relaxfeel Floral Single Dohar White (AC Dohar) Price: Rs. 499 A Good Collection Of Dohars Which Are Good Compaingns For Summers & Winters As Well. High Quality Gurantee Extremely Soft And Warm. Perfect For Everyday Use. These Dohars Gives Luxarious And Elegant Look, Easy To Carry During Travelling. Easy Storage. All Season Dohars, Can Be Used In AC Rooms Also.,Specifications of Relaxfeel Floral Single Dohar White (AC Dohar) General Brand Relaxfeel Type Dohar Hand Washable Yes Model ID single dohar Color White Design floral Machine Washable Yes Suitable For A/C ROOMS Inner Material Polycotton Model Name single dohar Ideal For Baby Boys and Baby Girls Outer Material PLOYESTER Size Single Dimensions Weight 700 g Length 3 inch / 10 cm Width 11 inch / 30 cm Depth 30 cm In the Box Number of Contents in Sales Package 1'

La destination du bien doit être plus « franche » et détaillée

| | 904 |
|---|---|
| product_name | Relaxfeel Floral Single Dohar White |
| image | a541b3aba326d7749b4c086c3cea9273.jpg |
| main_category | Baby Care |
| tf_idf_predicted | Home Furnishing |
| word2vec_predicted | Home Furnishing |
| bert_predicted | Beauty and Personal Care |
| use_predicted | Home Furnishing |
| vgg16_predicted | Home Furnishing |
| predicted | Wrong Guess |

où l'aurions-nous classé en lisant sa description et en regardant l'image ? 😉

La photographie doit être en adéquation avec la destination du bien en plus de sa description

# Merci de votre attention ☺