

# Projet 6 : Classifiez des biens de consommation

Lancelot LECLERCQ

7 mars 2022

# Sommaire

1. Introduction
2. Classification des descriptions textuelles
3. Classification des images
4. Classification de l'ensemble des données
5. Comparaison des différentes classifications
6. Conclusion

# Introduction

# Problématique

- L'entreprise Place de marché est un marketplace e-commerce
  - Vendeurs proposent des articles à des acheteurs en postant une photo et une description
  - Attribution de la catégorie d'un article effectuée manuellement par les vendeurs  $\Rightarrow$  peu fiable
- Objectif
  - Améliorer l'expérience utilisateur des vendeurs et des acheteurs
  - Automatisation de l'attribution d'une catégorie
  - Étude de la faisabilité d'un moteur de classification

# Jeu de données

- Jeu de données textuelles
  - Nom, prix, description, note, pour chaque objet

---

## Données

---

uniq\_id  
crawl\_timestamp  
product\_url  
product\_name  
product\_category\_tree  
pid  
retail\_price  
discounted\_price  
image  
is\_FK\_Advantage\_product  
description  
product\_rating  
overall\_rating  
brand  
product\_specifications

---

- Jeu d'images
  - Nous avons une image par objet



FIG. : Exemple d'image associée à un objet (ici des rideaux)

# Jeu de données

- Jeu de données textuelles
  - Nom, prix, description, note, pour chaque objet

---

## Données

---

uniq\_id  
crawl\_timestamp  
product\_url  
product\_name  
product\_category\_tree  
pid  
retail\_price  
discounted\_price  
image  
is\_FK\_Advantage\_product  
description  
product\_rating  
overall\_rating  
brand  
product\_specifications

---

- Jeu d'images
  - Nous avons une image par objet



FIG. : Exemple d'image associée à un objet (ici des rideaux)

## Classification des descriptions textuelles

## Sélection des stopwords personnalisés

- 
- The figure consists of three bar charts, each representing a different level of linguistic analysis: Tokens, Lemmas, and Racines (stems). Each chart shows the frequency of various linguistic units in a corpus.
- Tokens:** This chart shows the frequency of individual words. The y-axis ranges from 0 to 800. The x-axis lists 50 tokens. The most frequent tokens are 'usb' (approx. 850), 'model' (approx. 650), 'inch' (approx. 600), 'fabric' (approx. 550), 'india' (approx. 500), 'perfect' (approx. 450), 'key' (approx. 400), 'design' (approx. 350), 'great' (approx. 300), 'prices' (approx. 250), 'showpiece' (approx. 200), 'brand' (approx. 150), 'warranty' (approx. 100), 'ceramic' (approx. 50), 'number' (approx. 40), 'details' (approx. 30), 'best' (approx. 20), 'sales' (approx. 10), 'gifts' (approx. 5), 'product' (approx. 5), 'quality' (approx. 5), 'package' (approx. 5), 'general' (approx. 5), 'box' (approx. 5), 'pack' (approx. 5), 'analog' (approx. 5), 'type' (approx. 5), 'set' (approx. 5), 'cotton' (approx. 5), 'color' (approx. 5), 'baby' (approx. 5), 'material' (approx. 5), 'mug' (approx. 5), 'price' (approx. 5), 'watch' (approx. 5), 'features' (approx. 5), 'online' (approx. 5), 'guarantee' (approx. 5), 'flipflop' (approx. 5), 'cm' (approx. 5), 'day' (approx. 5), 'replacement' (approx. 5), 'shipping' (approx. 5), 'genuine' (approx. 5), 'cash' (approx. 5), 'delivery' (approx. 5), 'buy' (approx. 5), 'free' (approx. 5), and 'products' (approx. 5).
- Lemmas:** This chart shows the frequency of word forms. The y-axis ranges from 0 to 400. The x-axis lists 50 lemmas. The most frequent lemmas are 'multicolor' (approx. 450), 'woman' (approx. 400), 'light' (approx. 350), 'laptop' (approx. 300), 'ideal' (approx. 250), 'one' (approx. 200), 'home' (approx. 150), 'boy' (approx. 100), 'dimension' (approx. 50), 'size' (approx. 40), 'made' (approx. 30), 'men' (approx. 20), 'usb' (approx. 10), 'model' (approx. 5), 'fabric' (approx. 5), 'india' (approx. 5), 'gift' (approx. 5), 'perfect' (approx. 5), 'great' (approx. 5), 'key' (approx. 5), 'showpiece' (approx. 5), 'warranty' (approx. 5), 'inch' (approx. 5), 'ceramic' (approx. 5), 'brand' (approx. 5), 'number' (approx. 5), 'best' (approx. 5), 'detail' (approx. 5), 'sale' (approx. 5), 'product' (approx. 5), 'quality' (approx. 5), 'girl' (approx. 5), 'package' (approx. 5), 'design' (approx. 5), 'general' (approx. 5), 'box' (approx. 5), 'pack' (approx. 5), 'analog' (approx. 5), 'type' (approx. 5), 'set' (approx. 5), 'cotton' (approx. 5), 'baby' (approx. 5), 'material' (approx. 5), 'color' (approx. 5), 'watch' (approx. 5), 'feature' (approx. 5), 'online' (approx. 5), 'mug' (approx. 5), and 'price' (approx. 5).
- Racines:** This chart shows the frequency of word stems. The y-axis ranges from 0 to 400. The x-axis lists 50 racines. The most frequent racines are 'make' (approx. 450), 'one' (approx. 400), 'light' (approx. 350), 'home' (approx. 300), 'specific' (approx. 250), 'boy' (approx. 200), 'size' (approx. 150), 'dimens' (approx. 100), 'made' (approx. 50), 'use' (approx. 40), 'men' (approx. 30), 'usb' (approx. 20), 'model' (approx. 10), 'fabric' (approx. 5), 'india' (approx. 5), 'perfect' (approx. 5), 'great' (approx. 5), 'key' (approx. 5), 'cover' (approx. 5), 'showpiece' (approx. 5), 'warrant' (approx. 5), 'inch' (approx. 5), 'ceram' (approx. 5), 'number' (approx. 5), 'brand' (approx. 5), 'gift' (approx. 5), 'best' (approx. 5), 'print' (approx. 5), 'detail' (approx. 5), 'sale' (approx. 5), 'qualiti' (approx. 5), 'product' (approx. 5), 'packag' (approx. 5), 'gener' (approx. 5), 'box' (approx. 5), 'analog' (approx. 5), 'type' (approx. 5), 'pack' (approx. 5), 'cotton' (approx. 5), 'set' (approx. 5), 'materi' (approx. 5), 'babl' (approx. 5), 'color' (approx. 5), 'design' (approx. 5), 'featur' (approx. 5), 'watch' (approx. 5), 'onlin' (approx. 5), 'mug' (approx. 5), and 'price' (approx. 5).



# Méthode : Nettoyage et création de "bag of words"

## Comparaison des différents traitements du texte

TAB. : Tableau comparatif des différents procédés de nettoyage du texte

Modification	Contenu
Texte brut	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%
Tokenisation	polyester multicolor abstract eyelet door curtain height pack price curtain enhances look curtain made
Lemmatisation	polyester multicolor abstract eyelet door curtain height pack curtain enhances look curtain made
Racinisation	polyest multicolor abstract eyelet door curtain height pack curtain enhanc look curtain made

# Comparaison : validation croisée

## - Essais pour les 3 types de nettoyages

- tokenisation
- lemmatisation
- racinisation

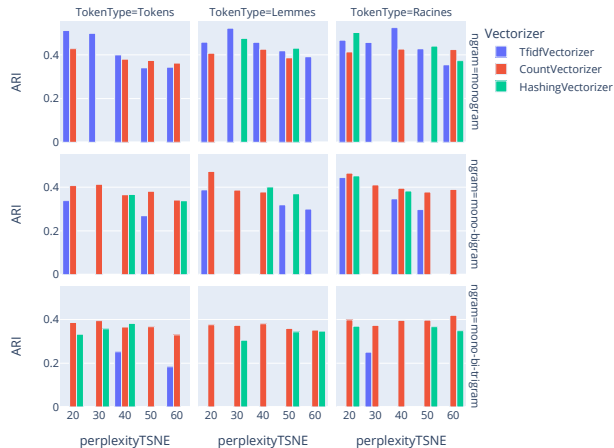
## - Essais de trois vectoriseurs :

- TfidfVectorizer
- CountVectorizer
- HashingVectorizer

## - Essais de différents découpages :

- mono-grams
- mono et bigrams
- mono, bi et trigrams

Comparaison des scores en fonction du vectoriseur et de la perplexité



# Meilleure classification

- Meilleur score en utilisant :

- les racines des mots
- des monograms
- le vectoriseur tf-idf

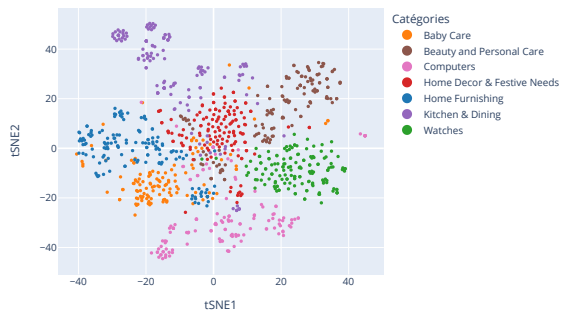
- ARI = 0,525

Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE40 TfidfVectorizer(1, 1) Racines

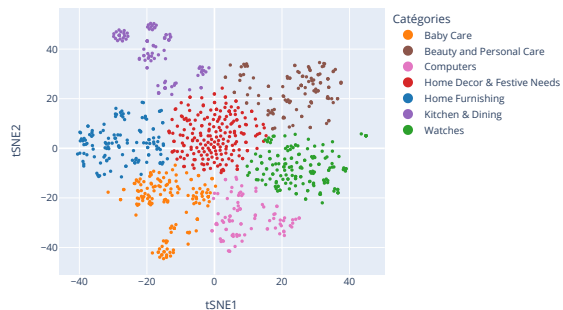
Catégorie réelle	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	86	3	44	1	19	0	0
	5	96	0	7	0	15	0
	2	2	69	9	1	7	1
	18	30	16	119	10	17	3
	14	0	1	5	110	0	0
	0	0	0	0	0	102	0
Catégorie prédite	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	0	8	7	1	0	1	146

# Visualisation de la meilleure classification

t-SNE40 TfidfVectorizer(1, 1) Racines



KMeans t-SNE40 TfidfVectorizer(1, 1) Racines

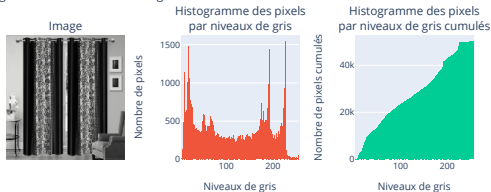


# Classification des images

# Méthode : Nettoyage

## Traitement des images pour les algorithmes SIFT et ORB

Image convertie en niveaux de gris



- ← Utilisation de niveaux de gris
- ✓ Égalisation de l'histogramme
- ↓ Filtration de l'image

Image égalisée par CLAHE

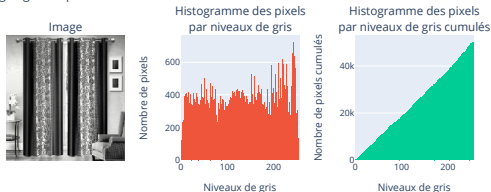
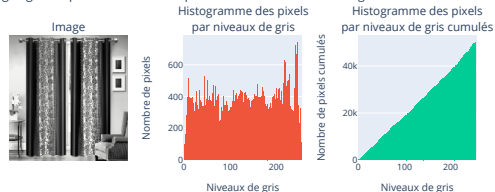


Image égalisée par CLAHE et filtrée par Non-local Means Denoising

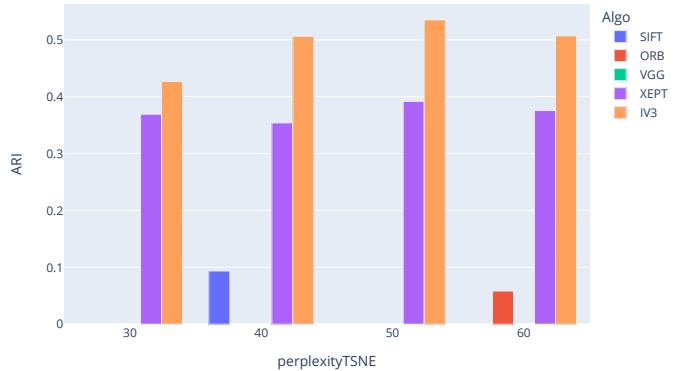


# Comparaison

## - Essais avec 5 algorithmes :

- SIFT
- ORB
- VGG16
- Xception
- InceptionV3

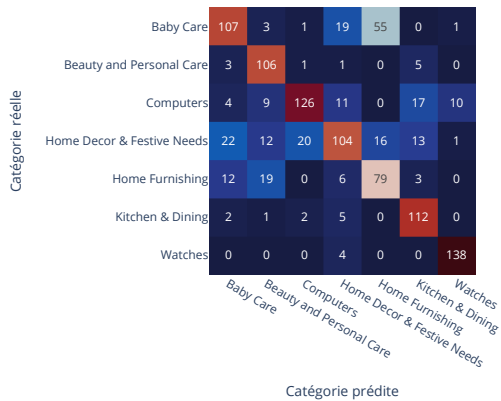
Comparaison des scores en fonction  
de l'algorithme et de la perplexité



# Meilleure classification

- Meilleure classification avec : InceptionV3
- ARI = 0,535

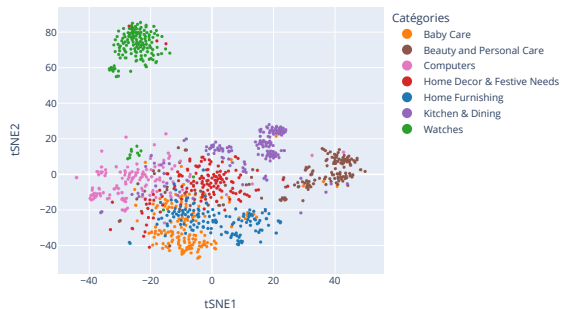
Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE40 IV3



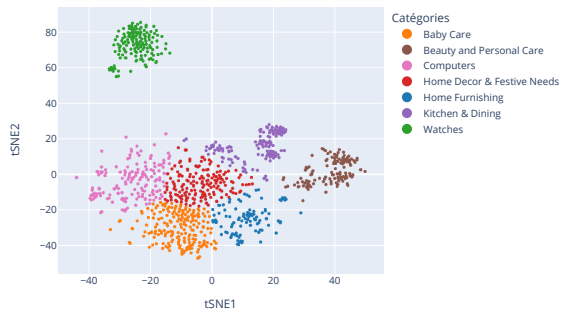


# Visualisation de la meilleure classification

t-SNE50 IV3



KMeans t-SNE50 IV3



## Classification de l'ensemble des données

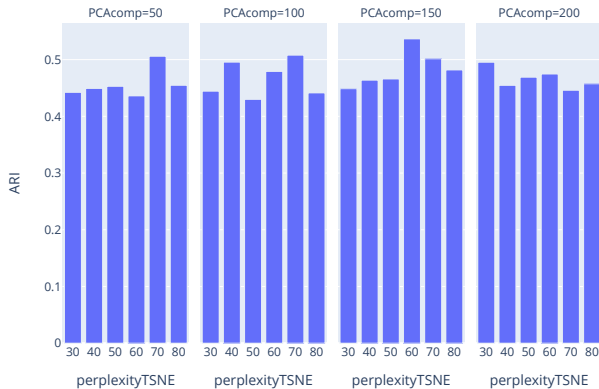
# Méthode

- Utilisation de la matrice "bag of words" pour la meilleure classification du texte (racines) avec la matrice "bag of visual words" pour la meilleure classification des images (InceptionV3)
- Essais avec plusieurs valeurs de composantes utilisées pour la PCA afin d'affiner la classification

# Comparaison

- Essais pour différentes valeurs de composantes utilisée pour la PCA

Comparaison des scores en fonction de la perplexité



# Meilleure classification

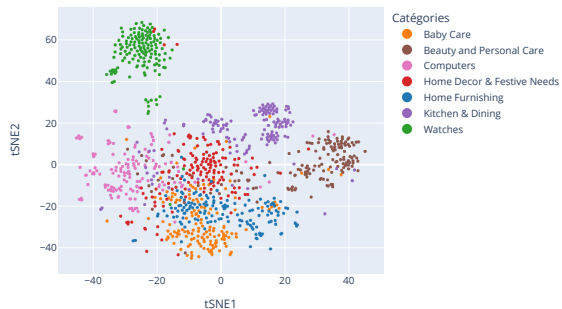
- Meilleure classification avec 150 composantes
- ARI = 0,536

Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE60 Racines tfidf(1, 1) IV3 150

Catégorie réelle	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	112	3	2	18	60	0	1
	3	108	2	2	0	5	0
	5	7	125	8	0	15	0
	17	17	16	105	15	13	1
	11	14	2	11	75	3	0
	2	1	0	2	0	114	0
Catégorie prédite	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	0	0	3	4	0	0	148

# Visualisation de la meilleure classification

t-SNE60 Racines tfidf(1, 1) IV3 150



KMeans t-SNE60 Racines tfidf(1, 1) IV3 150



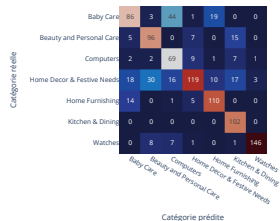
## Comparaison des différentes classifications

# Comparaison des différentes classifications

## Matrices de corrélations

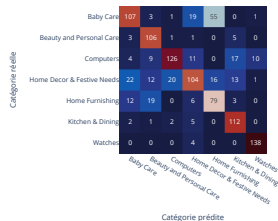
### Classification à partir des données textuelles uniquement

Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE40 TfIdVectorizer(1, 1) Racines



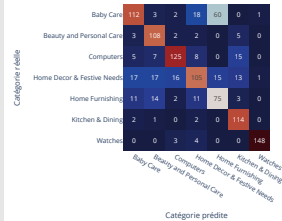
### Classification à partir des images uniquement

Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE40 IV3



### Classification à partir de toutes les données

Matrice de confusion des labels prédits (x) et réels (y)  
t-SNE60 Racines tfidf(1, 1) IV3 150



- Les images permettent d'éviter les classifications aberrantes comme la confusion entre produits informatiques et produits pour bébé avec le texte uniquement
- Confusion entre les fournitures de maison et les produits pour bébé avec les données d'image uniquement qui se répercute lors de l'utilisation des données complètes
- Difficulté à classer les produits de décoration et de fête (Home Decor & Festive Needs)



## Conclusion

- Classification des données séparées moyenne
  - ARI légèrement supérieure à 0,5 dans les deux cas
- La combinaison des deux types de données n'améliore pas significativement le score mais évite les classements aberrants (mélanges entre informatique et produits pour bébé avec le texte uniquement)
- Il pourrai être intéressant d'effectuer une classification supervisée à partir des catégories déjà assignées afin d'affiner le classement
- Il pourrait aussi utiliser des mots clés dans les descriptions qui permettraient de positionner sans ambiguïté certains produits difficiles à classer avec les images
- Il peut aussi revoir les catégories afin d'améliorer le classement automatique