

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)

Classification des descriptions textuelles

Sélection des stopwords personnalisés

-
- Tokens**
- | Token | Nb d'occurrences |
|--------------|------------------|
| usb | 800 |
| model | 650 |
| inch | 600 |
| fabric | 550 |
| india | 500 |
| perfect | 450 |
| key | 400 |
| design | 350 |
| great | 300 |
| prices | 250 |
| showpiece | 200 |
| brand | 150 |
| warranty | 100 |
| ceramic | 50 |
| number | 40 |
| details | 30 |
| best | 20 |
| sales | 10 |
| gifts | 5 |
| product | 5 |
| quality | 5 |
| package | 5 |
| general | 5 |
| box | 5 |
| pack | 5 |
| analog | 5 |
| type | 5 |
| set | 5 |
| cotton | 5 |
| color | 5 |
| baby | 5 |
| material | 5 |
| mug | 5 |
| price | 5 |
| watch | 5 |
| features | 5 |
| online | 5 |
| guarantee | 5 |
| flipkart.com | 5 |
| cm | 5 |
| day | 5 |
| replacement | 5 |
| shipping | 5 |
| genuine | 5 |
| cash | 5 |
| delivery | 5 |
| buy | 5 |
| free | 5 |
| products | 5 |
- Lemmes**
- | Lemme | Nb d'occurrences |
|------------|------------------|
| multicolor | 450 |
| woman | 400 |
| light | 350 |
| laptop | 300 |
| ideal | 250 |
| one | 200 |
| home | 150 |
| boy | 100 |
| dimension | 50 |
| size | 40 |
| made | 30 |
| men | 20 |
| usb | 10 |
| model | 5 |
| fabric | 5 |
| india | 5 |
| gift | 5 |
| perfect | 5 |
| great | 5 |
| key | 5 |
| showpiece | 5 |
| warranty | 5 |
| inch | 5 |
| ceramic | 5 |
| brand | 5 |
| number | 5 |
| best | 5 |
| detail | 5 |
| sale | 5 |
| product | 5 |
| quality | 5 |
| girl | 5 |
| package | 5 |
| design | 5 |
| general | 5 |
| box | 5 |
| pack | 5 |
| analog | 5 |
| type | 5 |
| set | 5 |
| cotton | 5 |
| baby | 5 |
| material | 5 |
| color | 5 |
| watch | 5 |
| feature | 5 |
| online | 5 |
| mug | 5 |
| price | 5 |
- Racines**
- | Racine | Nb d'occurrences |
|-----------|------------------|
| make | 450 |
| one | 400 |
| light | 350 |
| home | 300 |
| specific | 250 |
| boy | 200 |
| size | 150 |
| dimens | 100 |
| made | 50 |
| use | 40 |
| men | 30 |
| usb | 20 |
| model | 10 |
| fabric | 5 |
| india | 5 |
| perfect | 5 |
| great | 5 |
| key | 5 |
| cover | 5 |
| showpiece | 5 |
| warrant | 5 |
| inch | 5 |
| ceram | 5 |
| number | 5 |
| brand | 5 |
| gift | 5 |
| best | 5 |
| print | 5 |
| detail | 5 |
| sale | 5 |
| qualiti | 5 |
| product | 5 |
| packag | 5 |
| gener | 5 |
| box | 5 |
| analog | 5 |
| type | 5 |
| pack | 5 |
| cotton | 5 |
| set | 5 |
| materi | 5 |
| babl | 5 |
| color | 5 |
| design | 5 |
| featur | 5 |
| watch | 5 |
| onlin | 5 |
| mug | 5 |
| price | 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

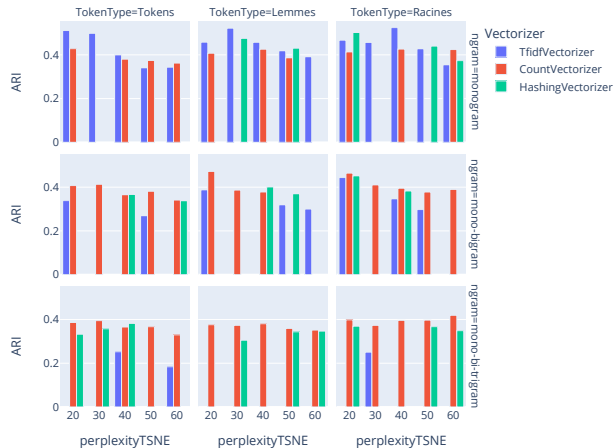
- Essais de trois vectoriseurs :

- 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

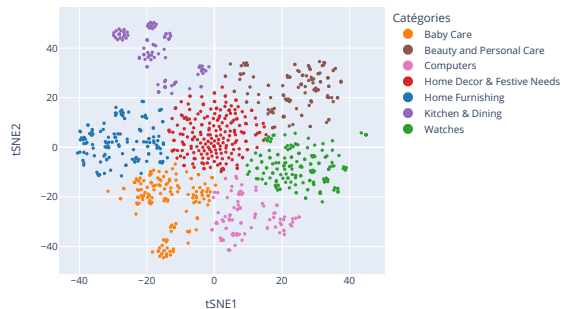
Catégorie réelle	Baby Care	86	3	44	1	19	0	0
	Beauty and Personal Care	5	96	0	7	0	15	0
	Computers	2	2	69	9	1	7	1
	Home Decor & Festive Needs	18	30	16	119	10	17	3
	Home Furnishing	14	0	1	5	110	0	0
	Kitchen & Dining	0	0	0	0	0	102	0
	Watches	0	8	7	1	0	1	146
		Baby Care	Beauty and Personal Care	Computers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

Visualisation de la meilleure classification

t-SNE40 TfidfVectorizer(1, 1) Racines



KMeans t-SNE40 TfidfVectorizer(1, 1) Racines

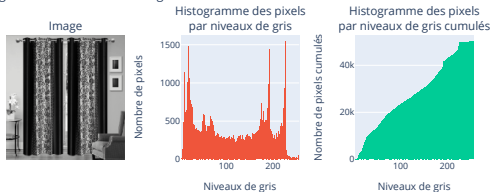


Classification des images

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

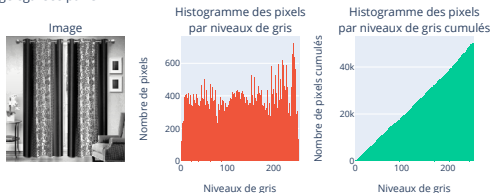
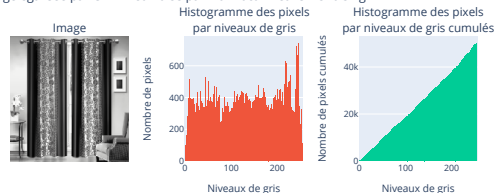


Image égalisée par CLAHE et filtrée par Non-local Means Denoising

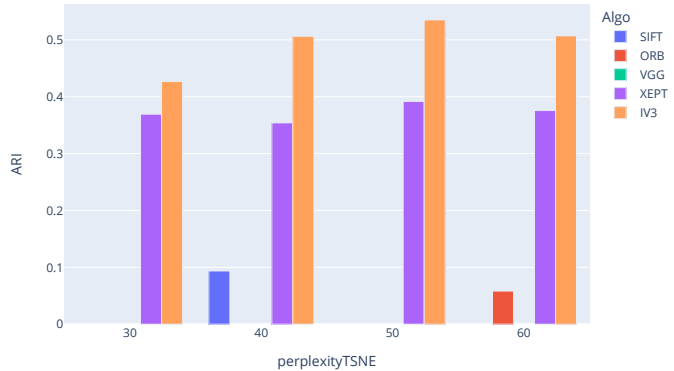


Comparaison

- Essais avec 5 algorithmes :

- SIFT
- ORB
- VGG16
- Xception
- InceptionV3

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



Meilleure classification

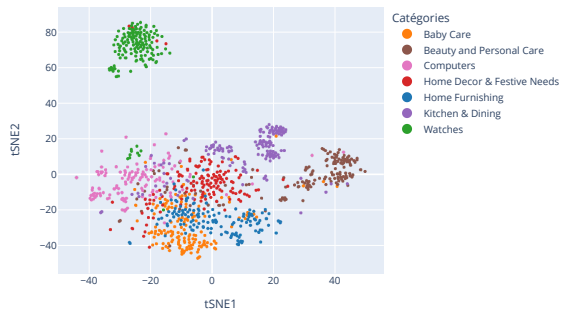
- Meilleure classification avec : InceptionV3
- ARI = 0,535

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

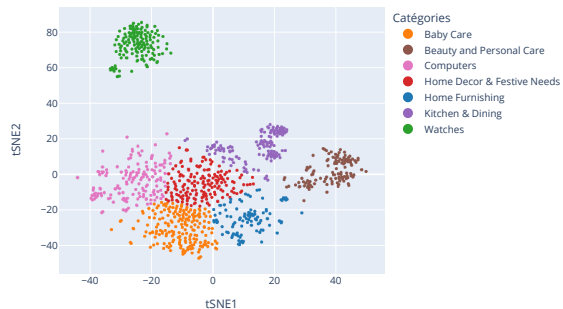
Catégorie réelle							
	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	107	3	1	19	55	0	1
	3	106	1	1	0	5	0
	4	9	126	11	0	17	10
	22	12	20	104	16	13	1
	12	19	0	6	79	3	0
Catégorie prédite	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
	0	0	0	4	0	0	138

Visualisation de la meilleure classification

t-SNE50 IV3



KMeans t-SNE50 IV3



Classification de l'ensemble des données

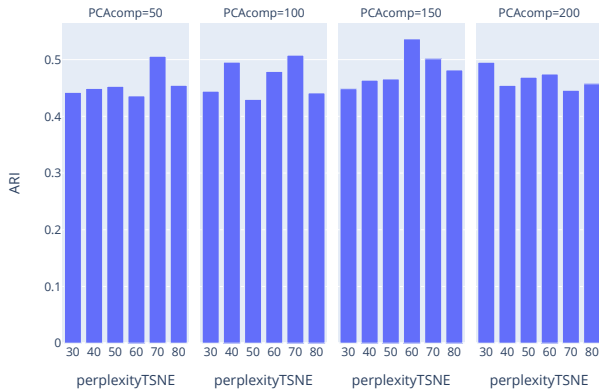
Méthode

- 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



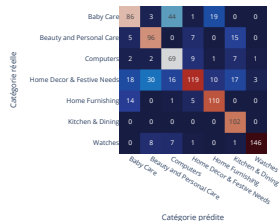
Comparaison des différentes classifications

Comparaison des différentes classifications

Matrices de corrélations

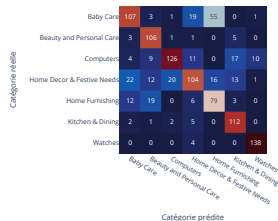
Classification à partir des données textuelles uniquement

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



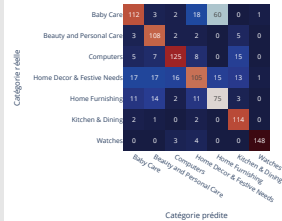
Classification à partir des images uniquement

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