Projet 6 : Classifiez des biens de consommation

Lancelot LECLERCQ

7 mars 2022

Sommaire

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 $\Longrightarrow$ 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

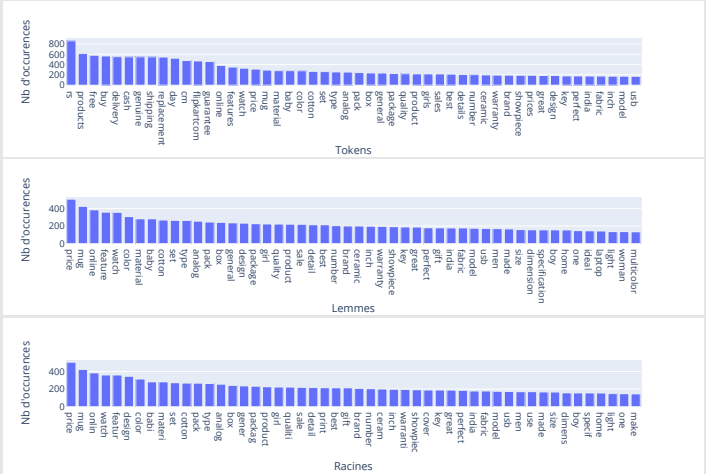
- Jeu d'images
  - Nous avons une image par objet

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

Classification des descriptions textuelles

Méthode : Nettoyage et création de ”bag of words”
Sélection des stopwords personalisés

- Nettoyage :
  - Retrait des chiffres et caractères spéciaux,
  - Retrait de la ponctuation,
  - Uniformisation de la casse
- Tokenisation
  - conservation des mots pertinents à partir de listes de ”stopwords”, des mots très récurrents à supprimer
- Lemmatisation
  - Similaire à la tokenisation avec la suppression des terminaisons des mots
  - Permet d’uniformiser les variations singulier/pluriel, masculin/féminin
- Racinisation (Stemmatisation)
  - Similaire à la lemmatisation avec conservation de la racine des mots
  - Permet d’uniformiser les variations de vocabulaires en regroupant les mots ayant les mêmes racines



Méthode : Nettoyage et création de ”bag of words”
Comparaison des différents traitements du texte

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

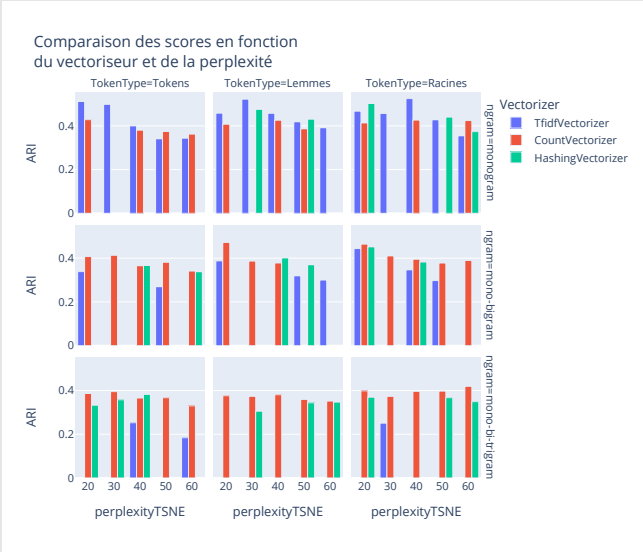
Modification	Contenu
Texte brut	Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price : Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100%
Tokenisation	polyester multicolor abstract eyelet door curtain height pack price curtain enhances 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

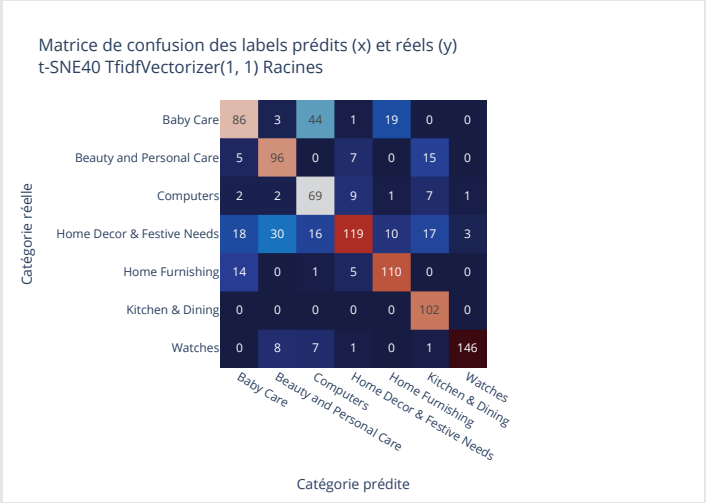
- Essais de trois vectoriseurs :
  - TfidfVectorizer
  - CountVectorizer
  - HashingVectorizer

- Essais de différents découpages :
  - mono-grams
  - mono et bigrams
  - mono, bi et trigrams

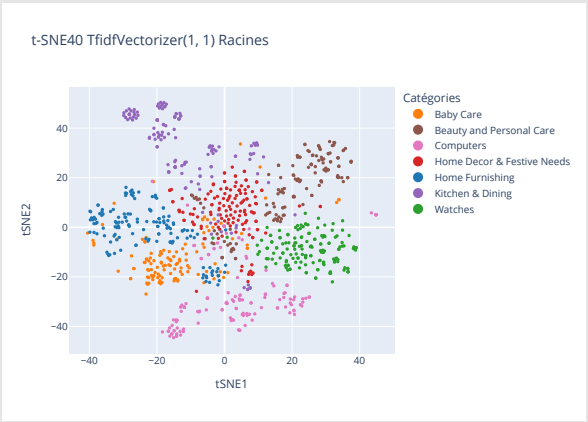


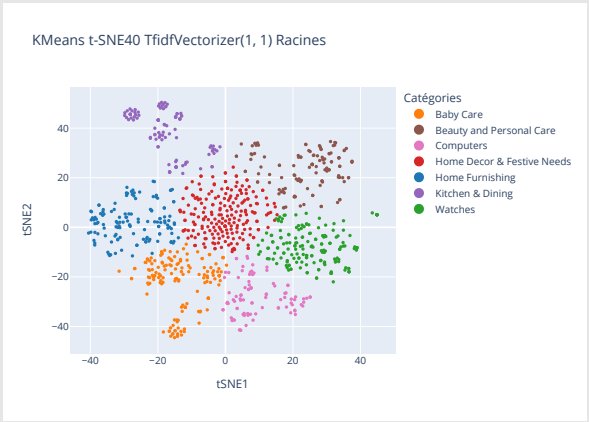
Meilleure classification

- Meilleur score en utilisant :
  - les racines des mots
  - des monograms
  - le vectoriseur tf-idf
- ARI = 0,525



Visualisation de la meilleure classification



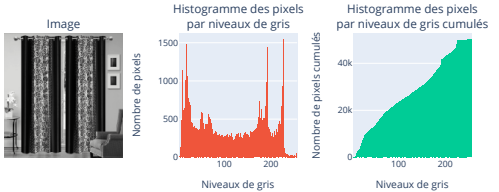


Classification des images

Méthode : Nettoyage

Traitement des images pour les algorithmes SIFT et ORB

Image convertie en niveaux de gris



← Utilisation de niveaux de gris

↙ Équalisation de l'histogramme

↓ Filtration de l'image

Image égalisée par CLAHE

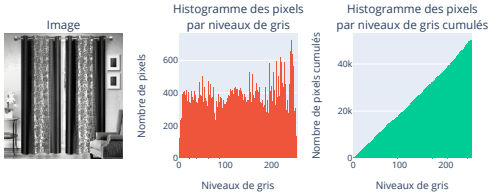
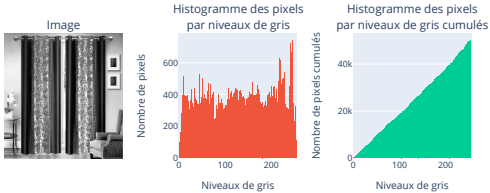
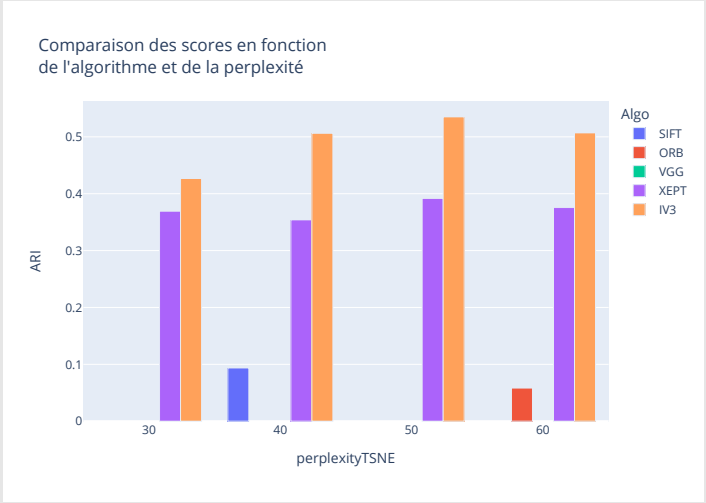


Image égalisée par CLAHE et filtrée par Non-local Means Denoising



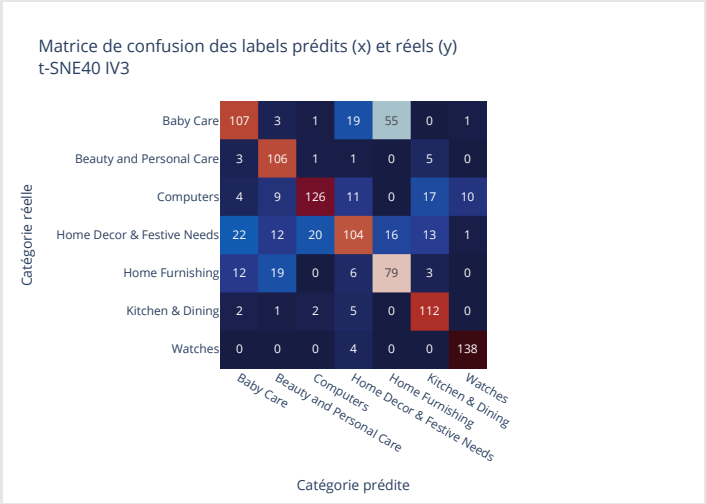
Comparaison

- Essais avec 5 algorithmes :
  - SIFT
  - ORB
  - VGG16
  - Xeption
  - InceptionV3



Meilleure classification

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



Visualisation de la meilleure classification

Classification de l'ensemble des données

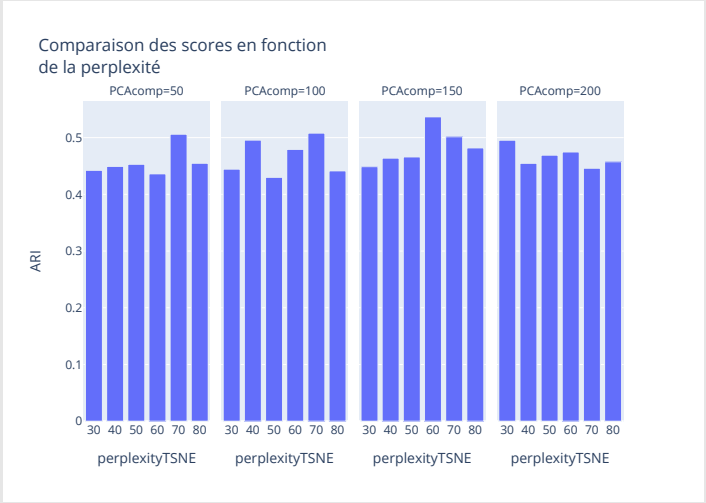
Méthode

- 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

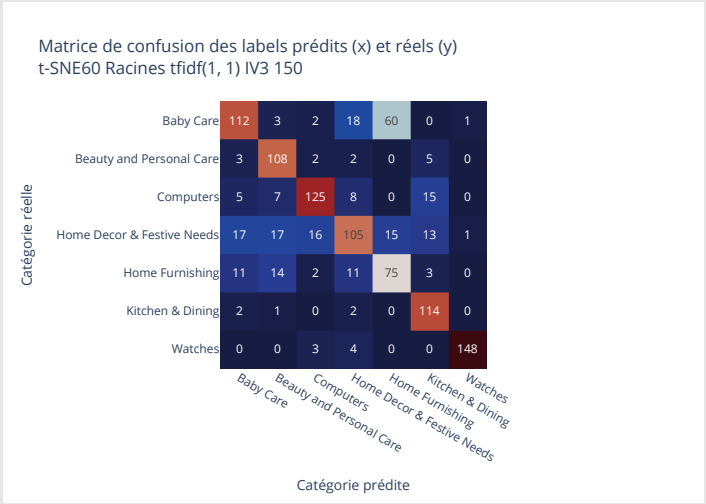
Comparaison

- Essais pour différentes valeurs de composantes utilisée pour la PCA

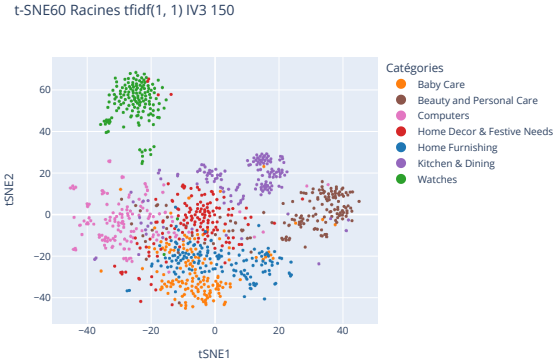


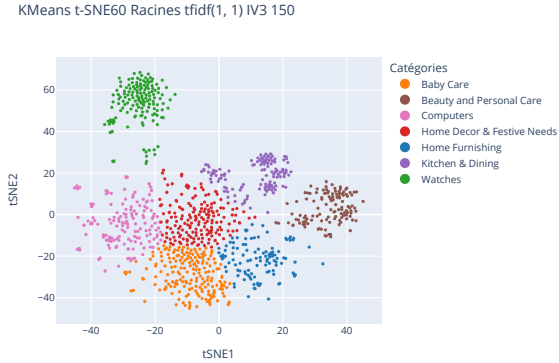
Meilleure classification

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



Visualisation de la meilleure classification

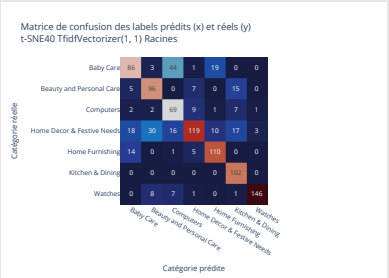




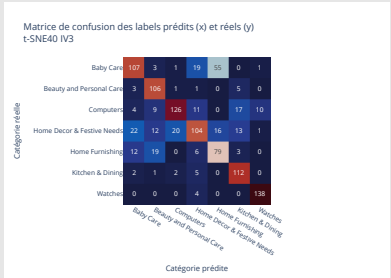
Comparaison des différentes classifications

Comparaison des différentes classifications
Matrices de corrélations

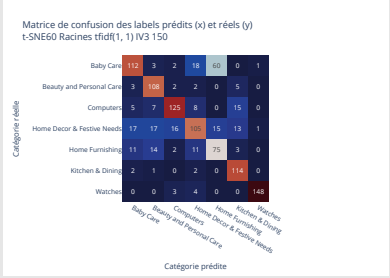
Classification à partir des données textuelles uniquement



Classification à partir des images uniquements



Classification à partir de toutes les données



- 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 classifier 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 evite 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'afiner le classement

- Il pourrait aussi utiliser des mots clés dans les descriptions qui permettraient de positionner sans ambiguité certains produits difficiles à classer avec les images

- Il peut aussi revoir les catégories afin d'améliorer le classement automatique