OPENCLASSROOMS

Parcours Data Scientist

Projet n°6 – Classifiez automatiquement des biens de consommation

Objectifs

- Réaliser une étude de faisabilité d'un moteur de classification automatique des articles
- Utilisation de méthodes de réduction de dimension, clustering, algorithme d'extraction de features de textes et d'images e

Présentation du dataset

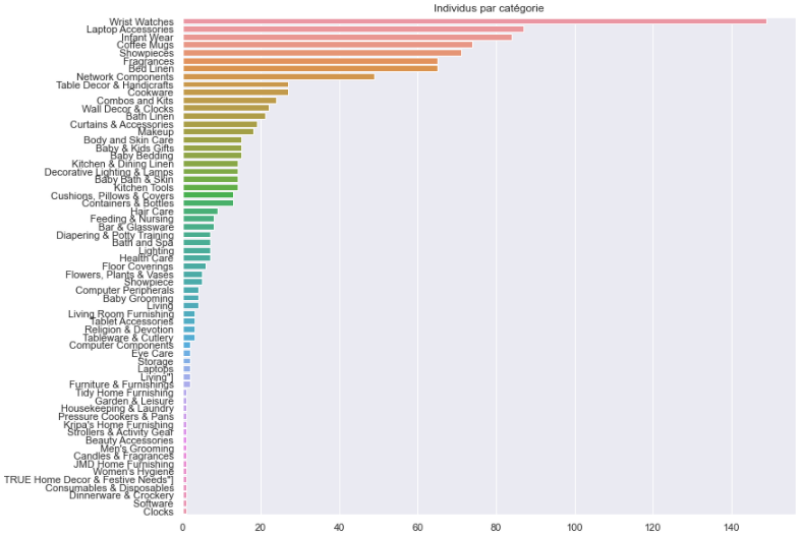
Dataset contenant les données relatives aux produits:

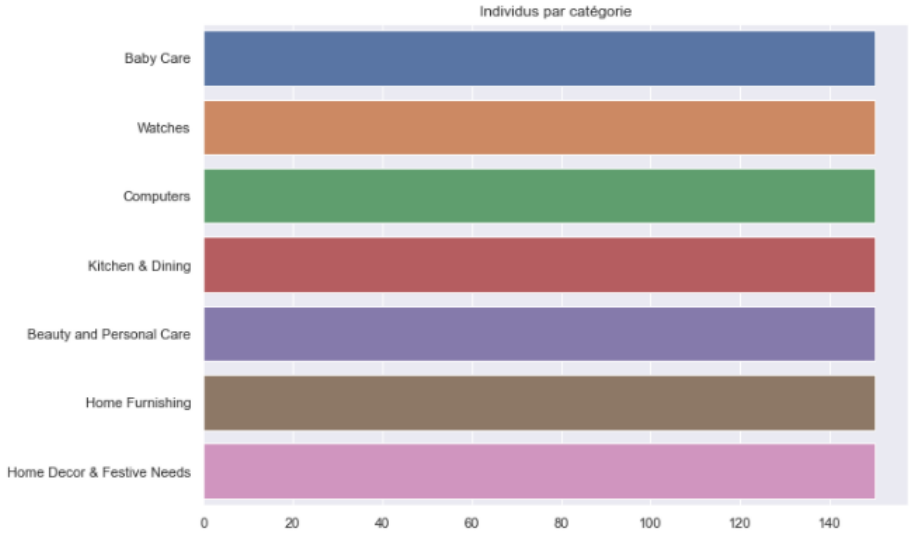
- ID unique
- Nom du produit, description
- Prix
- ID de l’image associée

Chaque produit est associée à une image/photo

Exploration

- Extraction de la catégorie de niveau 1 de chaque produit :
- Extraction de la catégorie principale de chaque produit :





Traitement des données Texte

- Nettoyage supplémentaire de la variable texte fusionnée

Variables considérées : (Spacy modul)

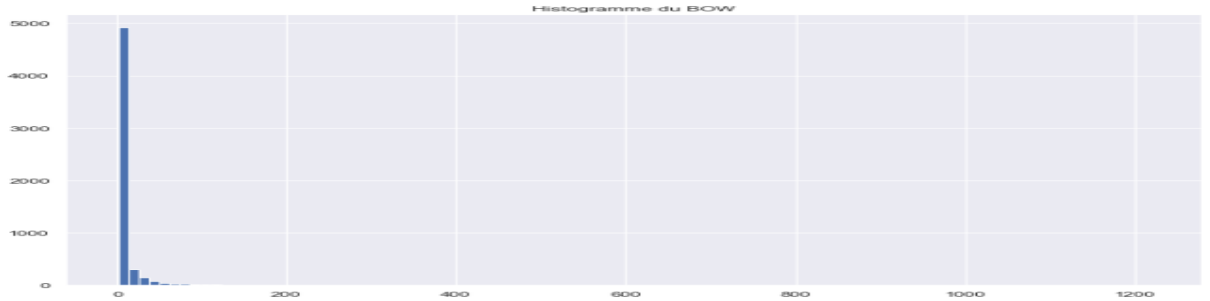
Description du produit (descriptions)

Nettoyage de ces variables

Tokenisation
Suppression des stopwords
Lemmatisation
Output Texte nettoyer (premiere ligne après la sortie

Nettoyage supplémentaire de la variable texte fusionnée

- Suppression de mots non pertinents
- Suppression de mots dépendemment de leur fréquence d'apparition
- Nouveau traitement par BOW et Tf-Idf



Métrique utilisé

World Cloud

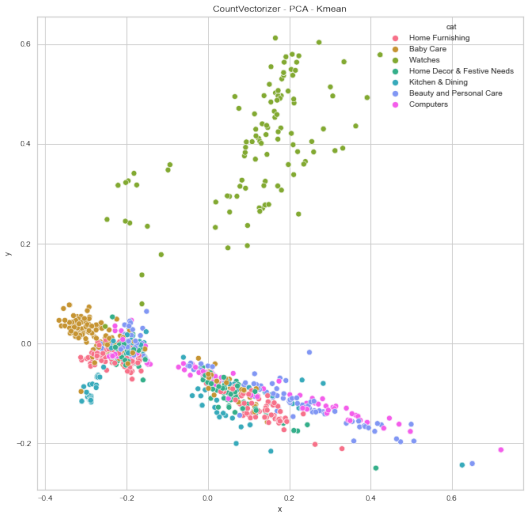

Flipkart
Analog Watch
Rs
Delivery
cm
Genuine Products
Free Shipping
Great Discounts
Products
Day
Shipping Cash
Online
Day Replacement
Replacement Guarantee
Digital Watch
Brand
Men Boys
Boys Girls
Cotton
Guarantee Free
Dimension
Sales Package
Key Features
Price Rs
Specifications
Coffee Mug
India
Cushion
Pillow Cover
Printed
Material Cotton
Bath Towel
Box Number
General Brand
Top Pyjama
Curtain
Ceramic Mug
Width inch
Copper
Type
Model Name
Towel Set
Color
inch cm
Boys Buy
Content
Material
AKUP
Light
Length inch
Pyjama Set

Traitement des données Texte

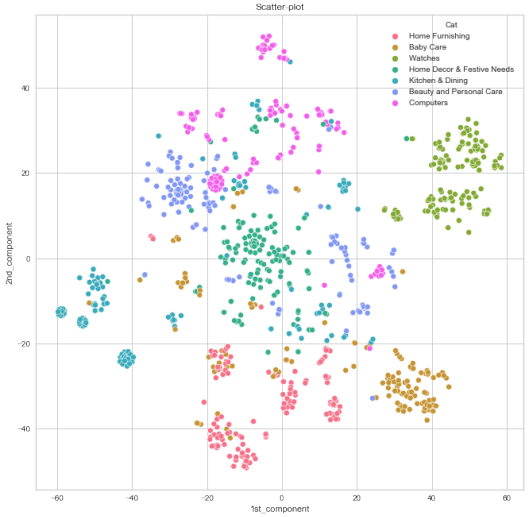
- TFIDF

	Adjusted Rand Score	
TFIDF	0.20	

PCA



TSNE

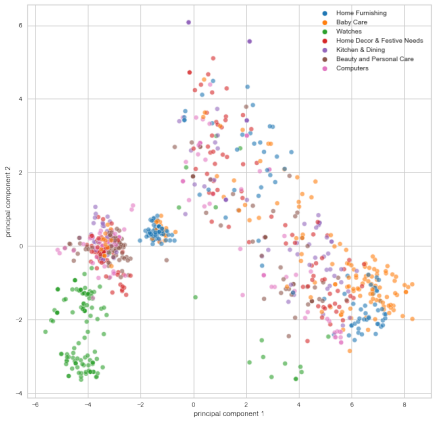


Traitement des données Texte

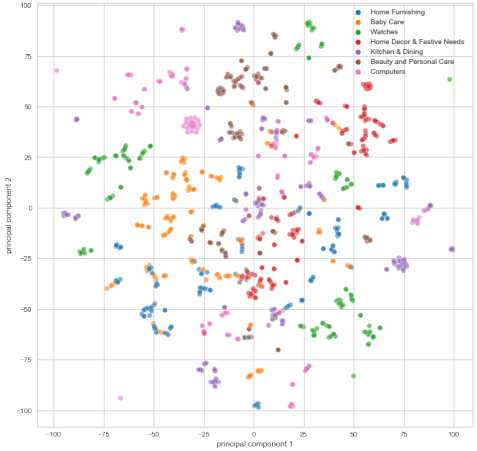
- Word 2 Vect

	Adjusted Rand Score	
W2V	0.18	

PCA



TSNE



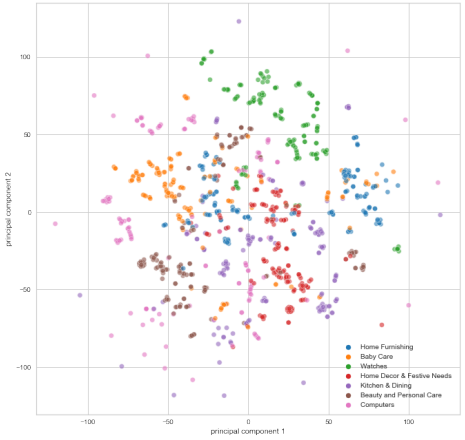
Traitement des données Texte

- BERT

	Adjusted Rand Score	
BERT	0.19	

PCA

TSNE

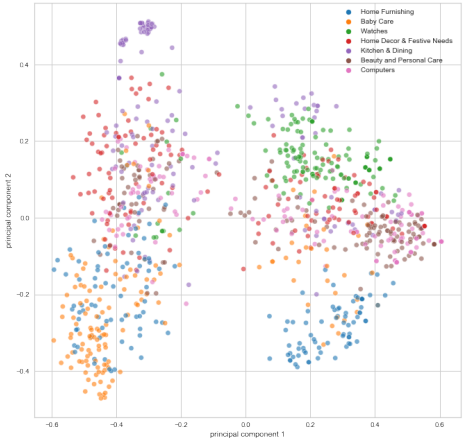


Traitement des données Texte

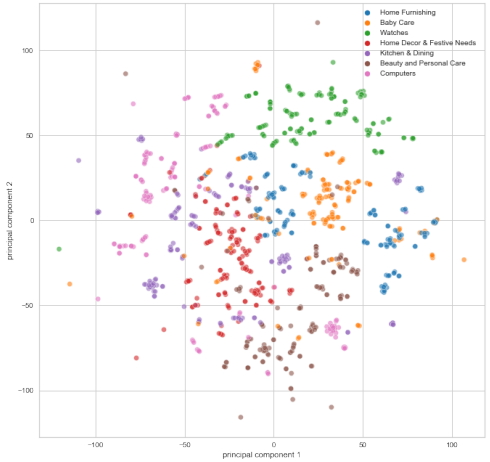
- USE

	Adjusted Rand Score	
USE	0.32	

PCA

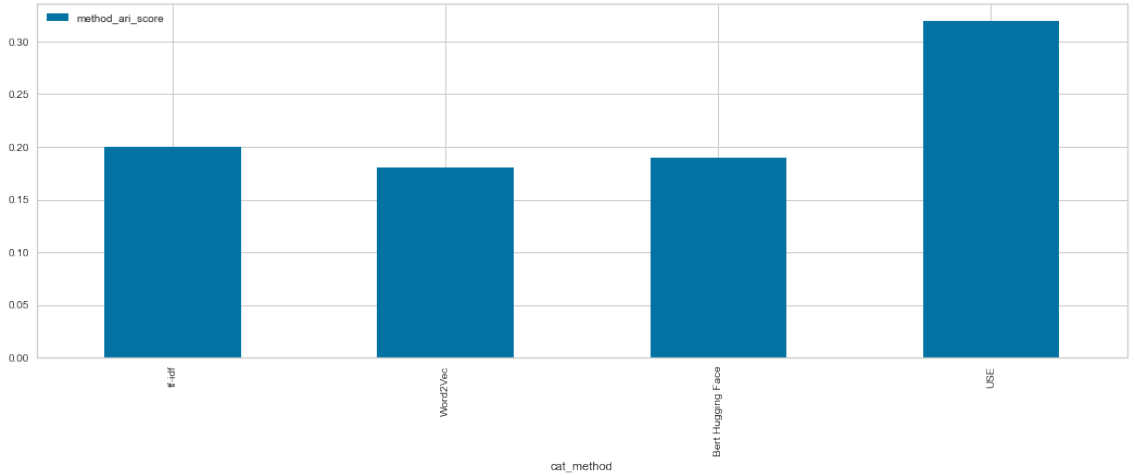


TNSE



Traitement des données Texte

	cat_method	method_ari_score
0	tf-idf	0.20
1	Word2Vec	0.18
2	Bert Hugging Face	0.19
3	USE	0.32



Traitement des données Image

- Utilisation de l'algorithme ORB
  - Passage de l'image en niveau de gris
  -Egalisations des histogrammes (pour éclaircir le contraste)
  - Détection des keypoints
  - Identification des features

  - Création des Visual Bag of Words
  - Création des histogrammes pour chaque image

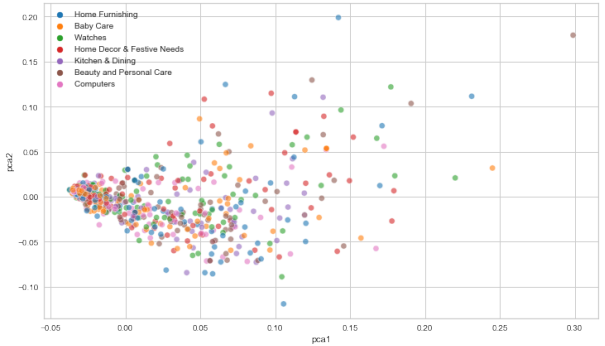
Exemple avec une image

Traitement des données Image

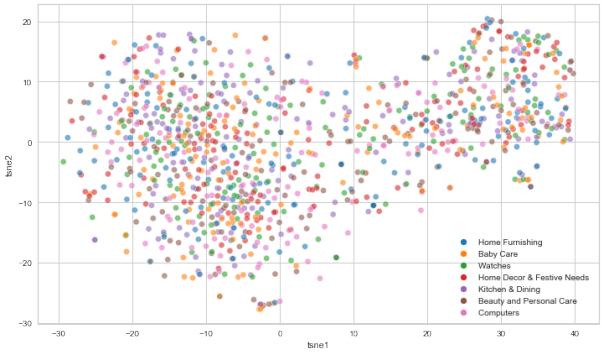
- Résultats du traitement avec ORB

	Adjusted Rand Score	
ORB	0.0005	

PCA



TSNE

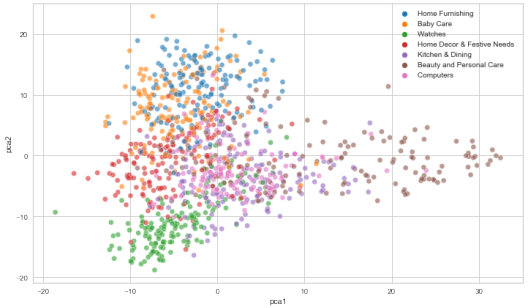


Traitement des données Image

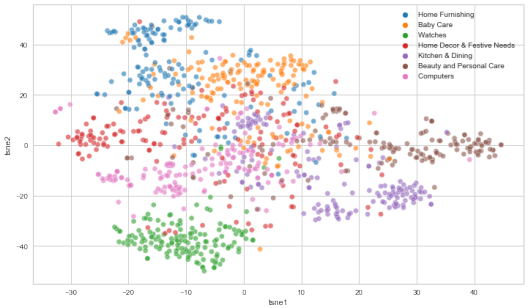
- Transfer Learning via réseau CNN
  - Feature extraction (modèle EfficientNetB0)
  - Suppression de la dernière couche
  - Représentation des images à partir des features déjà apprises

	Adjusted Rand Score	
Feature extraction	0.34	

PCA

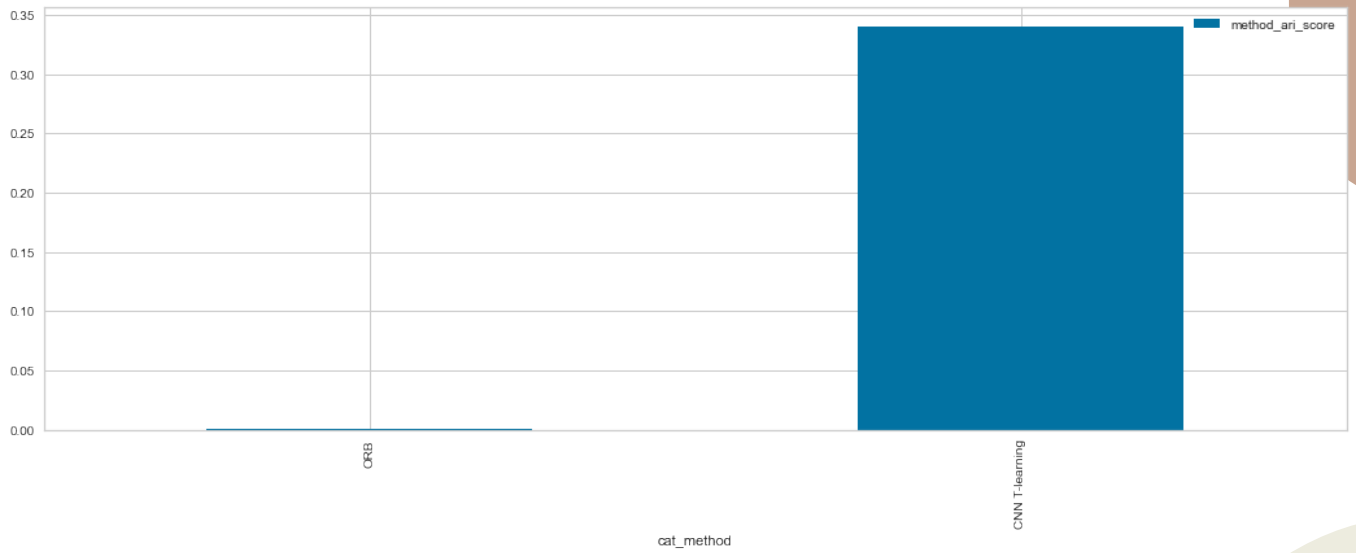


TNSE



Traitement des données IMG

	cat_method	method_ari_score
0	ORB	0.0005
1	CNN T-learning	0.3400

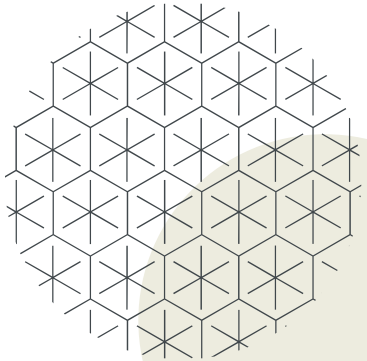


Conclusion

- Faisabilité du moteur de classification démontrée
  - Les visualisations TSNE sont équivoques
  - Les résultats de modélisation les confirment
- Selon le type de modèle utilisé (supervisé ou non), les données les plus pertinentes sont différentes

Limites et perspectives

- Dataset assez faiblement fourni, les résultats sont à confirmer
- Détection d'objet pour préciser la classification (nécessite davantage de ressources)



Merci!