CLASSIFIEZ AUTOMATIQUEMENT LES BIENS DE CONSOMMATION

DATA SCIENCE

SOMMAIRE

01

02

03

04

SOMMAIRE

SOMMAIRE

SOMMAIRE

SOMMAIRE

PRESENTATION

Contexte

- Classification manuelle des articles.
- Passage à grande échelle de la catégorisation.

PRESENTATION

Objectifs

- Montrer la possibilité d'une classification automatique des produits
- Elargir la gamme de produit via une API

Contraintes

- Automatisation de classification fondée sur l'image et la description
- Classification supervisée a partir de dataset image

PRESENTATION

Méthodologie

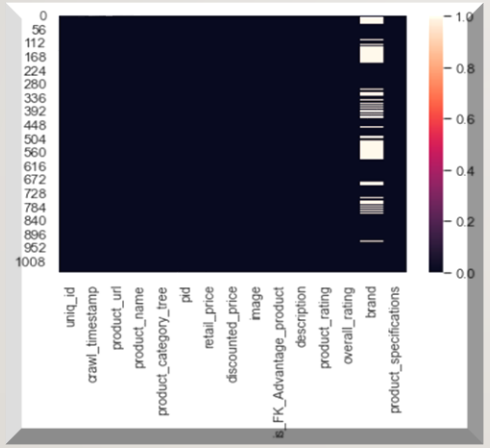
- Etudier la faisabilité de la classification textuelle
- Etudier la faisabilité de la classification visuelle
- Extraire et Analyser les produits de l'API

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

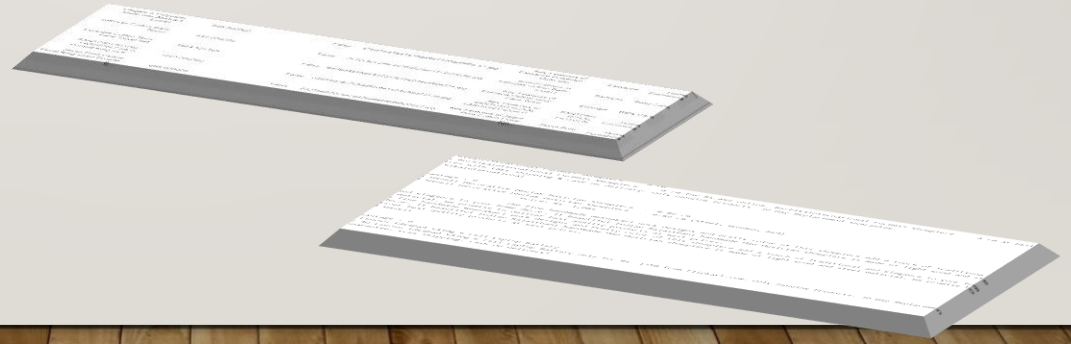
uniq_id	object
crawl_timestamp	object
product_url	object
product_name	object
product_category_tree	object
pid	object
retail_price	float64
discounted_price	float64
image	object
is_FK_Advantage_product	bool
description	object
product_rating	object
overall_rating	object
brand	object
product_specifications	object

dtype: object



Description des données :

- Une data frame (1.050, 15)
- 12 colonnes de type objet, 2 numériques et 1 booléen
- Valeurs manquantes au niveau de certaines variables
- Aucun doublon ni valeurs aberrantes



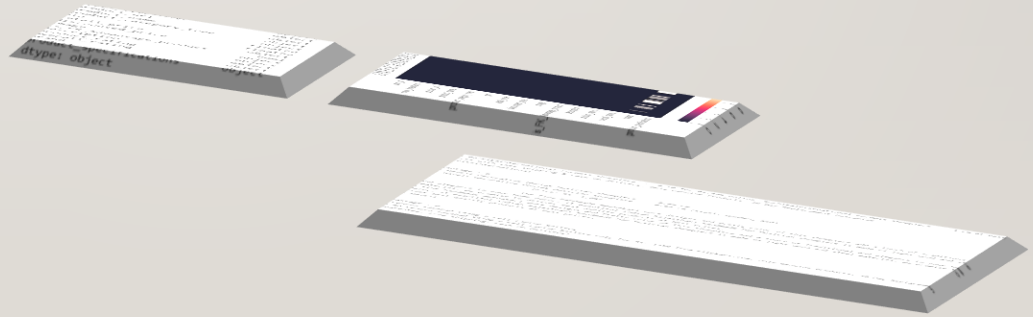
ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

	product_name	discounted_price	is_FK_Advantage_product	image	description	brand	category
0	Elegance Polyester Multicolor Abstract Eyelet ...	899.000000	False	55b85ea15a1536d46b7190ad6fff8ce7.jpg	Key Features of Elegance Polyester Multicolor ...	Elegance	Home Furnishing
1	Sathiyas Cotton Bath Towel	449.000000	False	7b72c92c2f6c40268628ec5f14c6d590.jpg	Specifications of Sathiyas Cotton Bath Towel (...	Sathiyas	Baby Care
2	Eurospa Cotton Terry Face Towel Set	1584.527169	False	64d5d4a258243731dc7bbb1eef49ad74.jpg	Key Features of Eurospa Cotton Terry Face Towe...	Eurospa	Baby Care
3	SANTOSH ROYAL FASHION Cotton Printed King size...	1299.000000	False	d4684dcdc759dd9cdf41504698d737d8.jpg	Key Features of SANTOSH ROYAL FASHION Cotton P...	SANTOSH ROYAL FASHION	Home Furnishing
4	Jaipur Print Cotton Floral King sized Double B...	698.000000	False	6325b6870c54cd47be6ebfbffa620ec7.jpg	Key Features of Jaipur Print Cotton Floral Kin...	Jaipur Print	Home Furnishing

Traitement des données:

- ❑ Sélection des variables pertinentes
- ❑ Remplacement de variables
- ❑ Traitement de valeurs manquantes



ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Récupération du Corpus :

```
advantage : 0
Ruchikainternational ruc0015 Showpiece  -  4 cm
Buy Ruchikainternational ruc0015 Showpiece  -  4 cm for Rs.499 online. Ruchikainternational ruc0015 Showpiece  -  4 cm at best
prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee.
Ruchikainternational


advantage : 0
Lal Haveli Decorative Dholak Musician Showpiece  -  8.89 cm
Lal Haveli Decorative Dholak Musician Showpiece  -  8.89 cm (Steel, Wooden, Red)
                        Price: Rs. 1,085

                       The fine handmade meenakari work designs and Pretty color of this showpiece add a touch of Traditiona
l and elegance to your home decor. It Beautiful Indian Rajasthani handmade Man Musician showpiece is made of light wood and ste
el material. We promise to deliver best quality products at best prices.
The fine handmade meenakari work designs and Pretty color of this showpiece add a touch of Traditional and elegance to your hom
e decor. It Beautiful Indian Rajasthani handmade Man Musician showpiece is made of light wood and steel material. We promise to
deliver best quality products at best prices.
Lal Haveli


advantage : 0
4D Lenovo IdeaPad Y430g 6 Cell Laptop Battery
Buy 4D Lenovo IdeaPad Y430g 6 Cell Laptop Battery only for Rs. 1350 from Flipkart.com. Only Genuine Products. 30 Day Replacemen
t Guarantee. Free Shipping. Cash On Delivery!
4D
```

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Natural Language Processing

- ❑ Text Cleaning :

Il consistera à :

- ✓ Mettre corpus en minuscule;
- ✓ Supprimer la ponctuation

- ❑ Text Preprocessing :

Il s'agira des opérations de :

- ✓ Tokennisation,
- ✓ Stopwords,
- ✓ Lemmatisation et
- ✓ Stemming

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Classification de texte

- Sentence embedding :

BoW, TF-IDF, Word2vec, BERT et USE.

- Détermination des clusters

- Visualisation des classes et Analyse

- Evaluation (ARI)

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

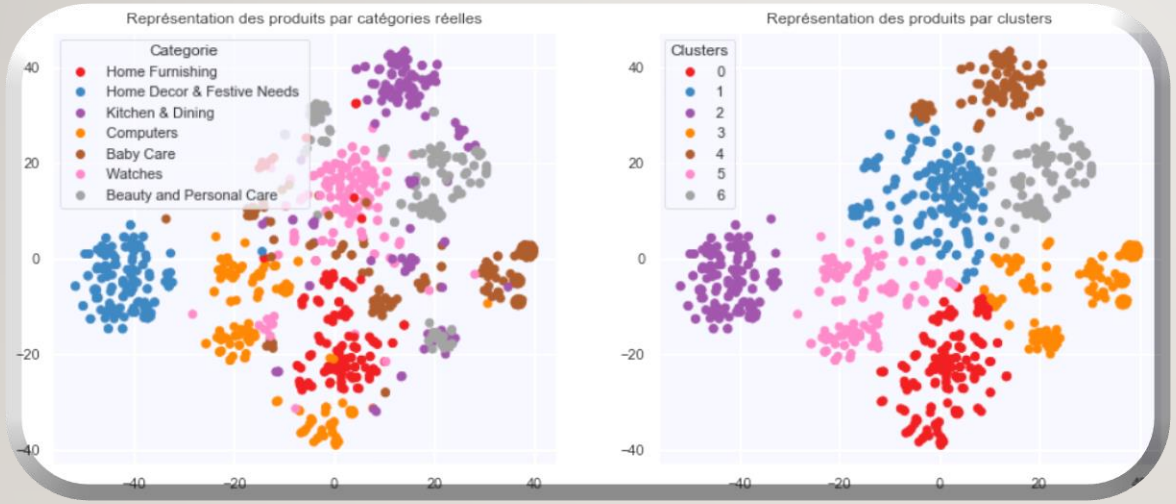
Classification de texte

BoW : 0,453

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Classification de texte

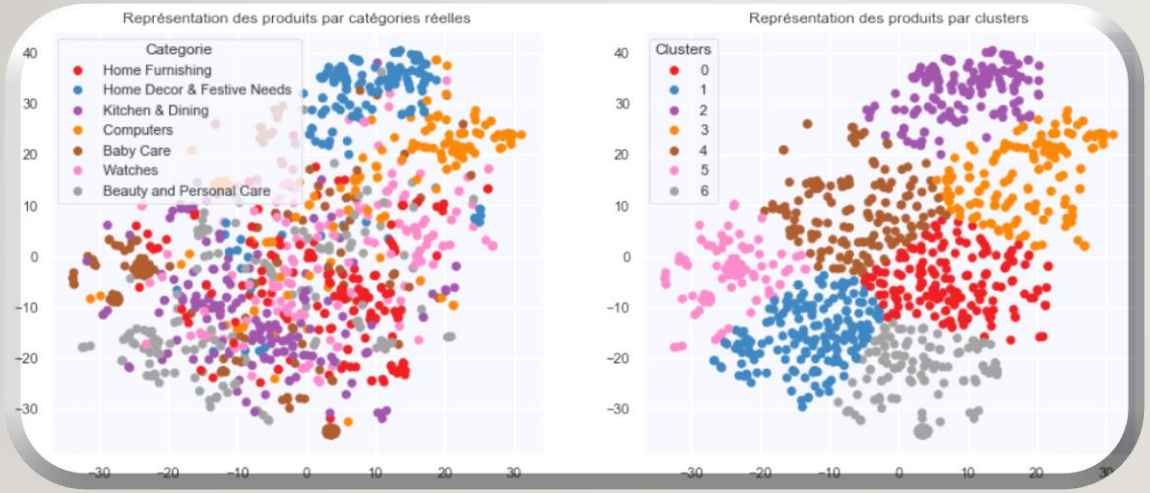


TF-IDF : 0,491

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Classification de texte



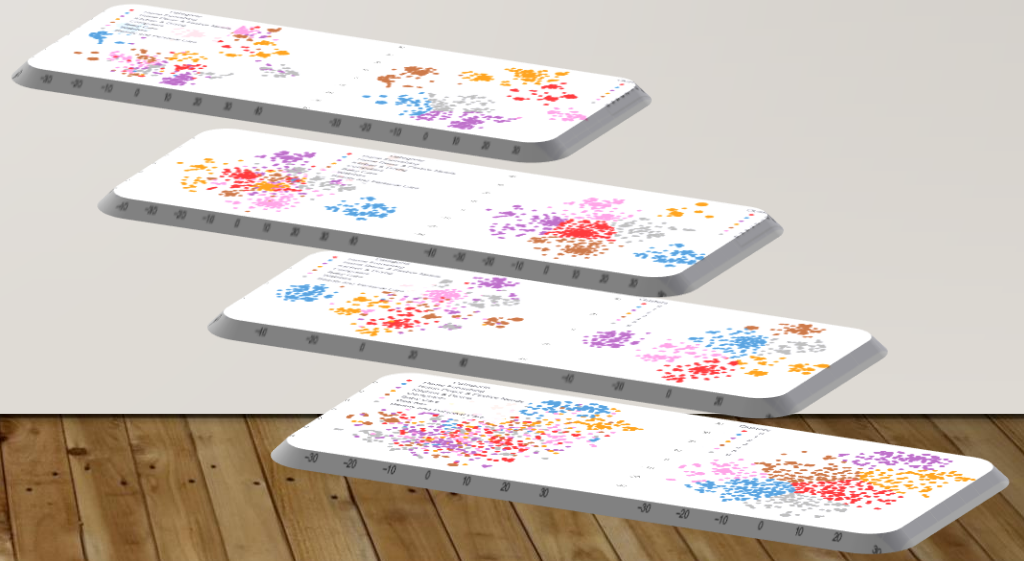
Word2Vec : 0,178

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

Classification de texte

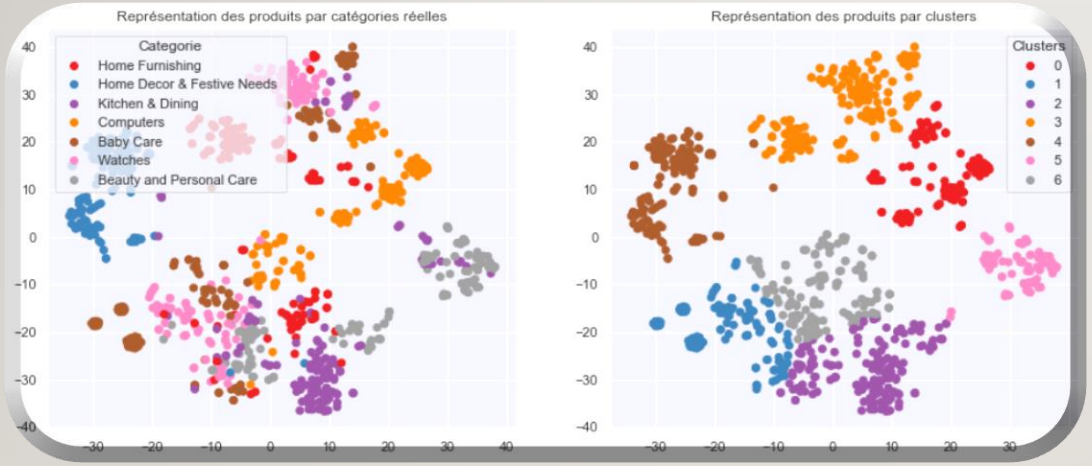
BERT : 0,348



ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse des données textuelles

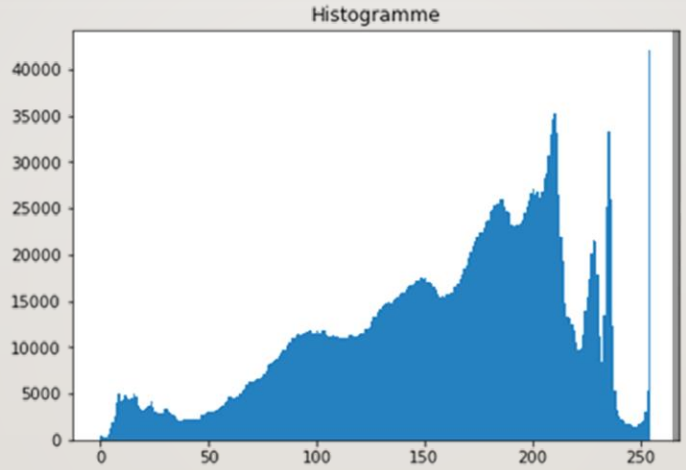
Classification de texte



USE : 0,365

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles



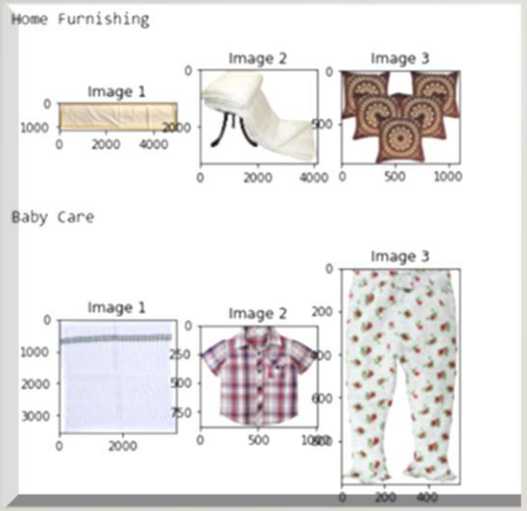
Description des données :

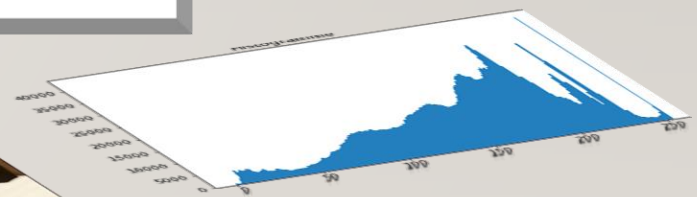
- Nous avons 1050 images.
- 07 catégories d'images
- 150 images par catégorie d'images

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Exemples d'images par catégories:





ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Segmentation des images

Nous allons utiliser deux algorithmes à savoir : SIFT et CNN, pour la détermination des features :

- ❑ Pré-traitement des images
- ❑ Extraction des features .

Ensuite, nous utiliserons l'algorithme K-Means pour la classification :

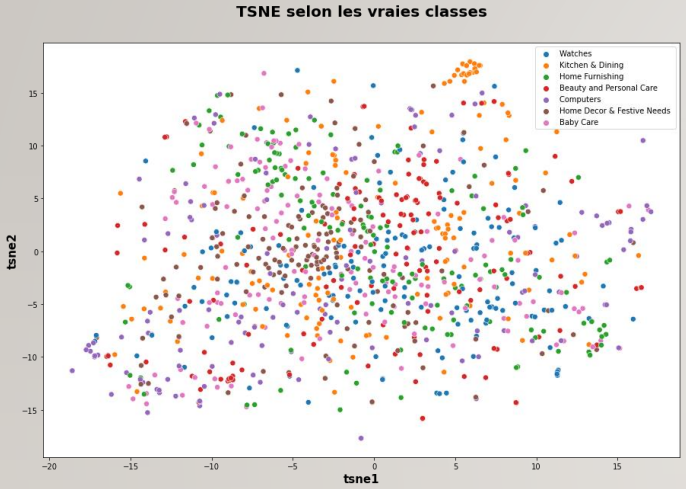
- ❑ Création des clusters
- ❑ Visualisation et analyse des classes

Enfin, nous évaluerons le modèle par le calcul du **Score ARI** et de **l'Accuracy**

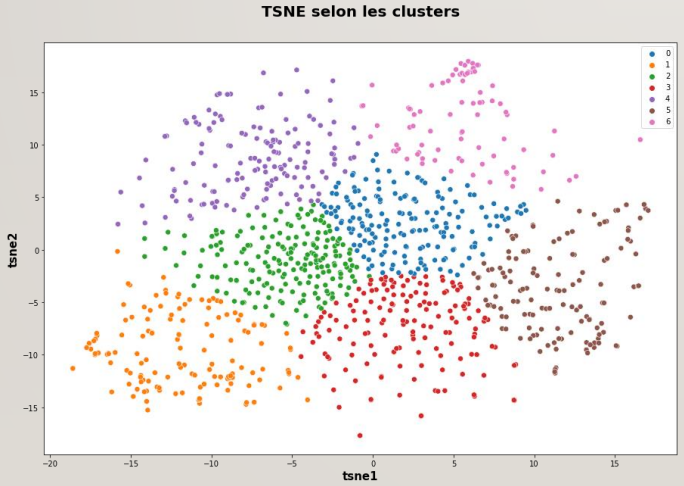
ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Segmentation des images



SIFT : 0,045 (ARI)

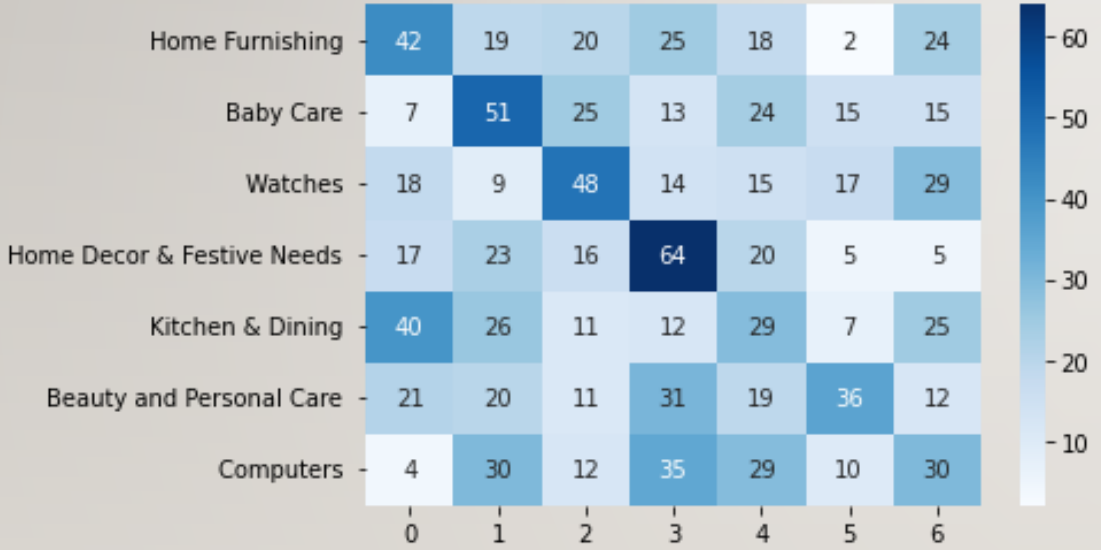


ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Segmentation des images

SIFT : 0,29 (Accuracy)



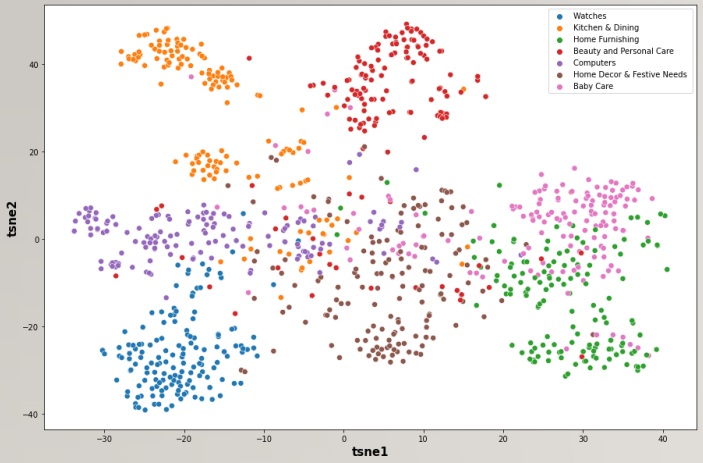
Matrice de confusion

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

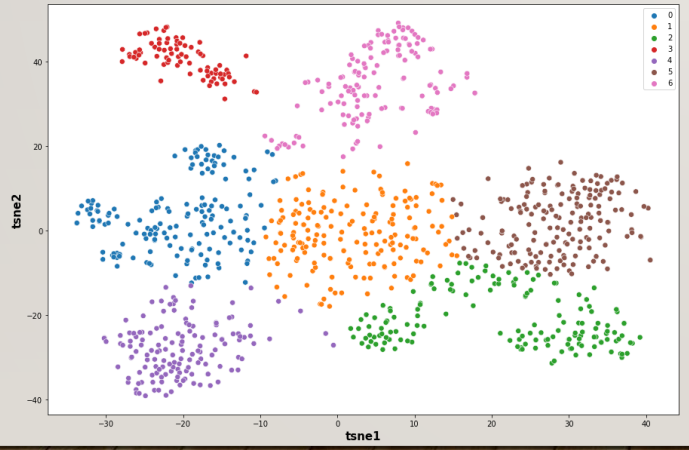
Segmentation des images

TSNE selon les vraies classes



CNN : 0,45 (ARI)

TSNE selon les clusters

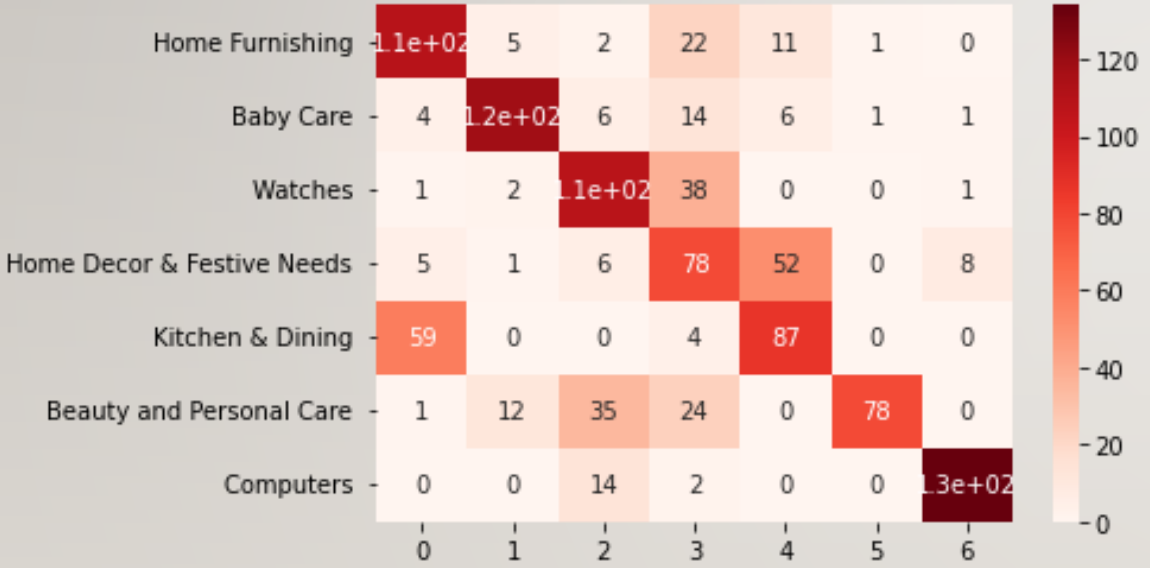


ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Segmentation des images

CNN : 0,68 (Accuracy)



Matrice de confusion

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Classification Supervisée des images

Nous allons utiliser deux approches à savoir :

- Une approche simple par préparation initiale de l'ensemble des images avant classification supervisée
- Une approche par data augmentation, permettant facilement la data augmentation

Nous allons utiliser la démarche suivante :

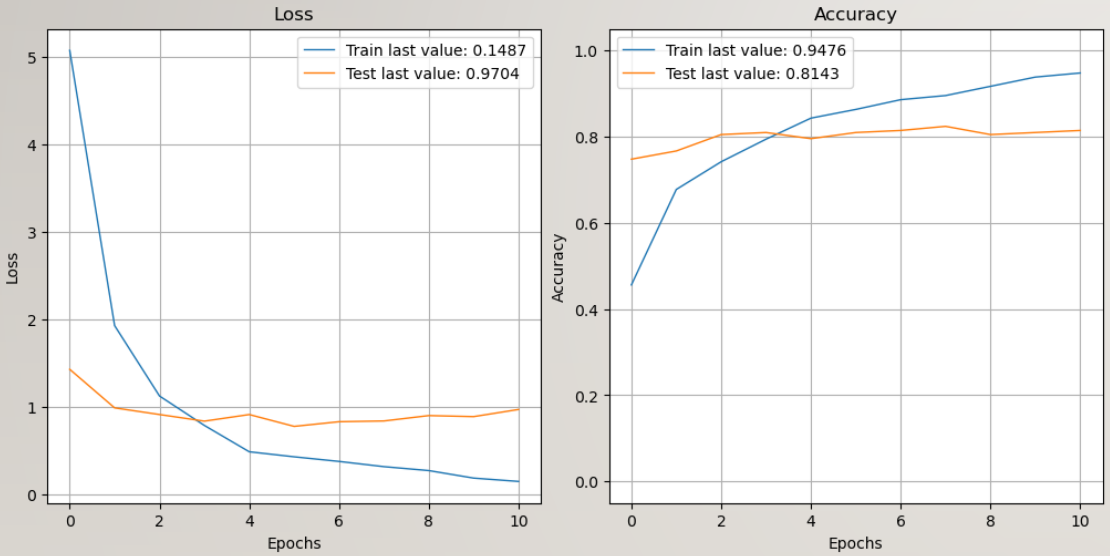
- ❑ Préparation des données
- ❑ Création du modèle
- ❑ Entrainement des données
- ❑ Evaluation et Analyse

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Classification Supervisée des images

Approche simple : **0,707** (ARI)

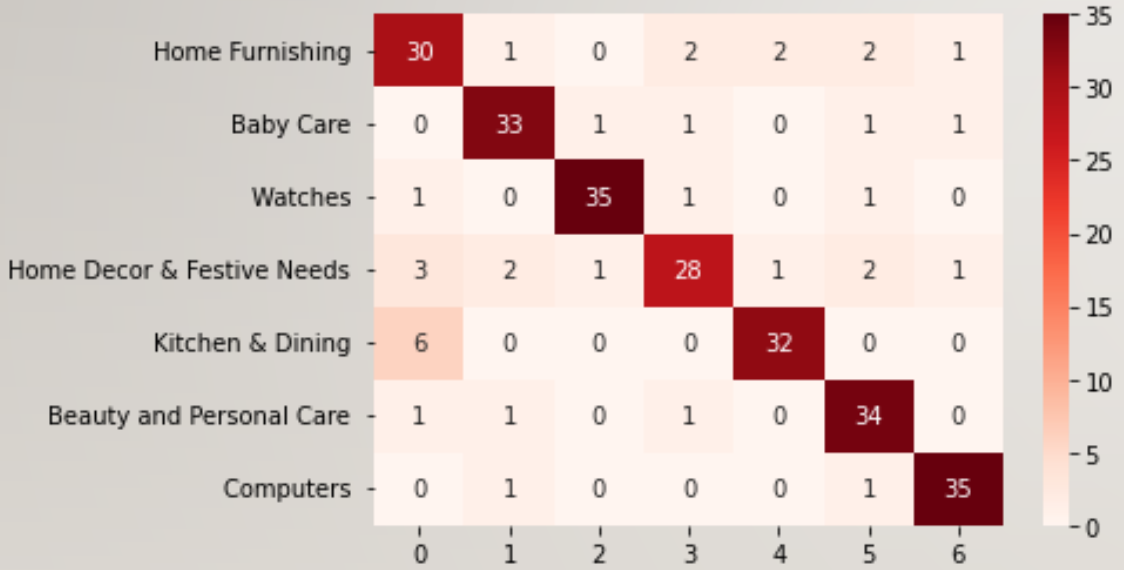


ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Classification Supervisée des images

Approche simple : **0,86** (Accuracy)



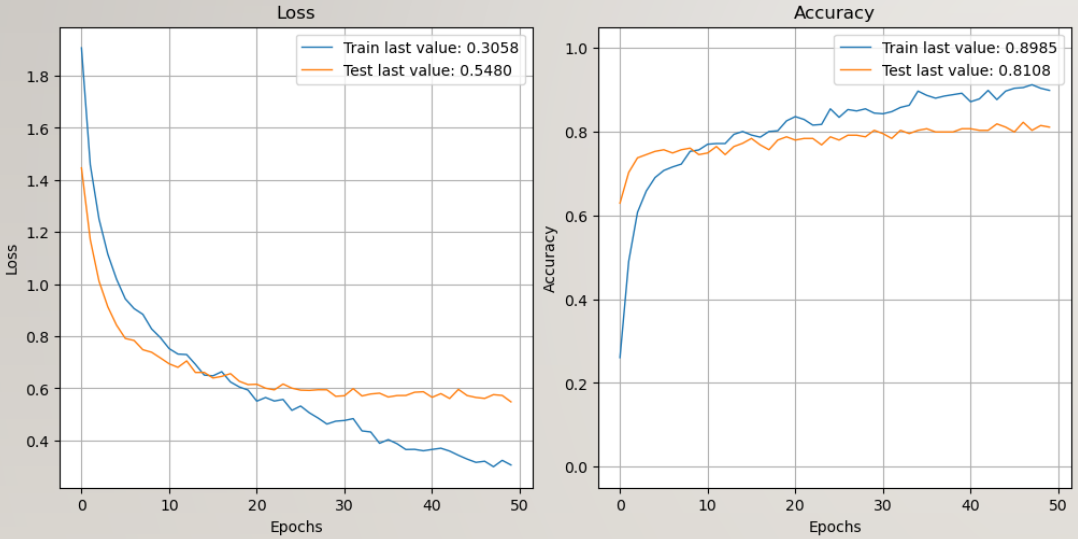
Matrice de confusion

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Classification Supervisée des images

Approche par data augmentation : 0,610 (ARI)

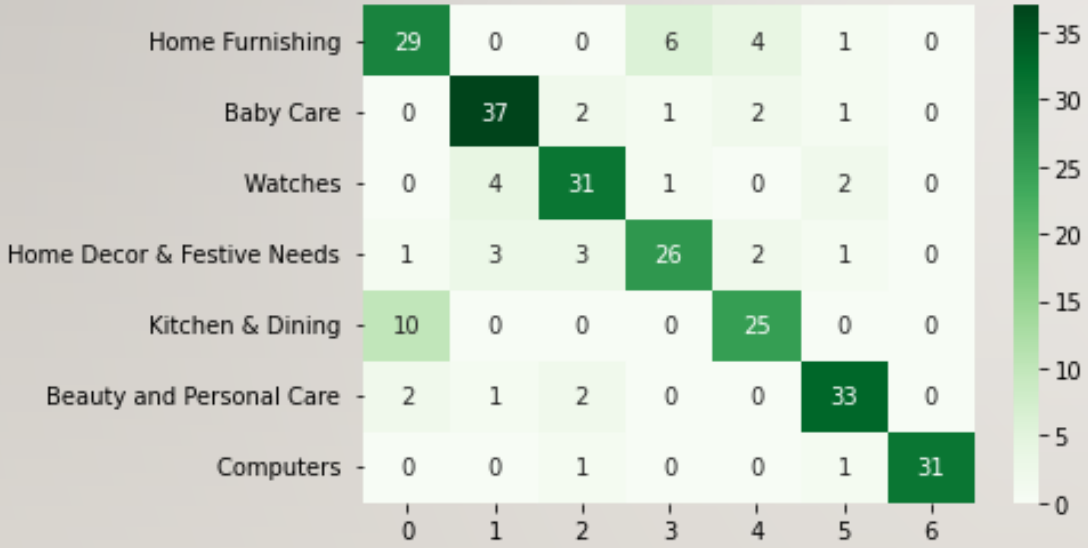


ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Analyse et segmentation des données visuelles

Classification Supervisée des images

Approche par data augmentation : **0,88** (Accuracy)



Matrice de confusion

ETUDE DE FAISABILITE DE LA CLASSIFICATION AUTOMATIQUE DES PRODUITS

Evaluation des algorithmes

Tableau récapitulatif des mesures de performance

	CLASSIFICATION DE TEXTE	SEGMENTATION DES IMAGES
	ARI	Accuracy
BOW	0,453	
TF-IDF	0,491	
WORD2VEC	0,178	
BERT	0,348	
USE	0,365	
SIFT	0,045	0,13
CNN	0,45	0,68
APPROCHE SIMPLE	0,707	0,86
DATA AUGMENTATION	0,61	0,88

- Les algorithmes de classification de texte ont des scores relativement faibles ;
- Les algorithmes de classification des images sont plus performants sauf le SIFT.

EXTRACTION DE PRODUITS DE L'API

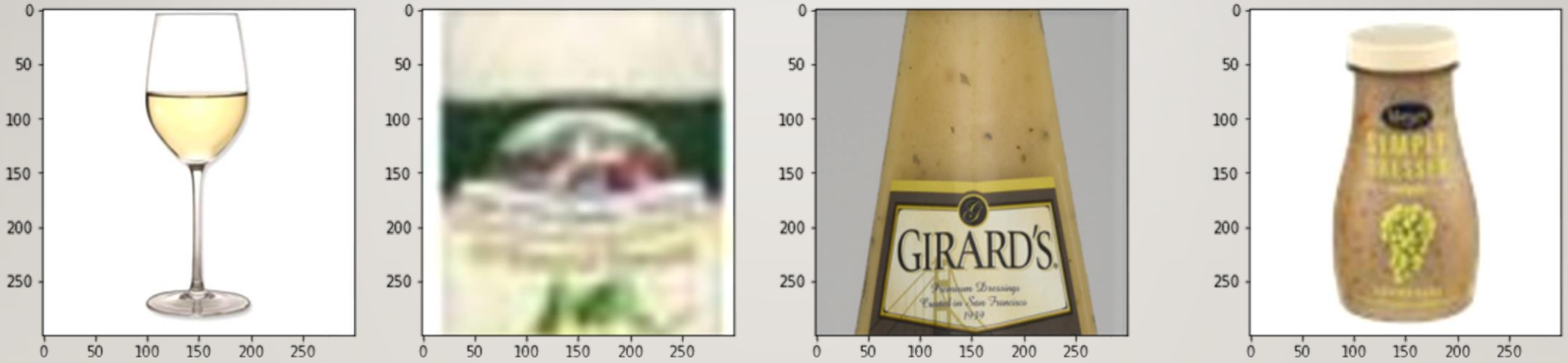
Dataframe des produits extraits:

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN

- Nous avons une liste de 20 produits décrits avec 12 variables
- Les données manipulées ne sont pas des données à caractère personnel; c'est-à-dire les informations s'y trouvant ne se rapportent à des personnes physiques.

EXTRACTION DE PRODUITS DE L'API

Images des produits extraits:



- Les données de L'API ne contient pas beaucoup d'images descriptif des produits
- Quatre (4) produits ont des images

CONCLUSION

La faisabilité de la classification automatique s'avère possible; l'algorithme TF-IDF de la classification textuelle a un score ARI de 0,49 et les algorithmes de classification des images donnent de très bon score.

Toutefois , les produits extraits de l'API ont peu d'images (1/5). Donc, la catégorisation des produits issus de l'API se fera de façon textuelle; alors que les algorithmes de classification textuelle sont moins performants que ceux des images .

Le Règlement général sur la protection des données (RGPD) ne s'appliquera pas à nos données extraites de l'API car ne traitant de données à caractère personnel.

OPENCLASSROOM