

Projet 6 :

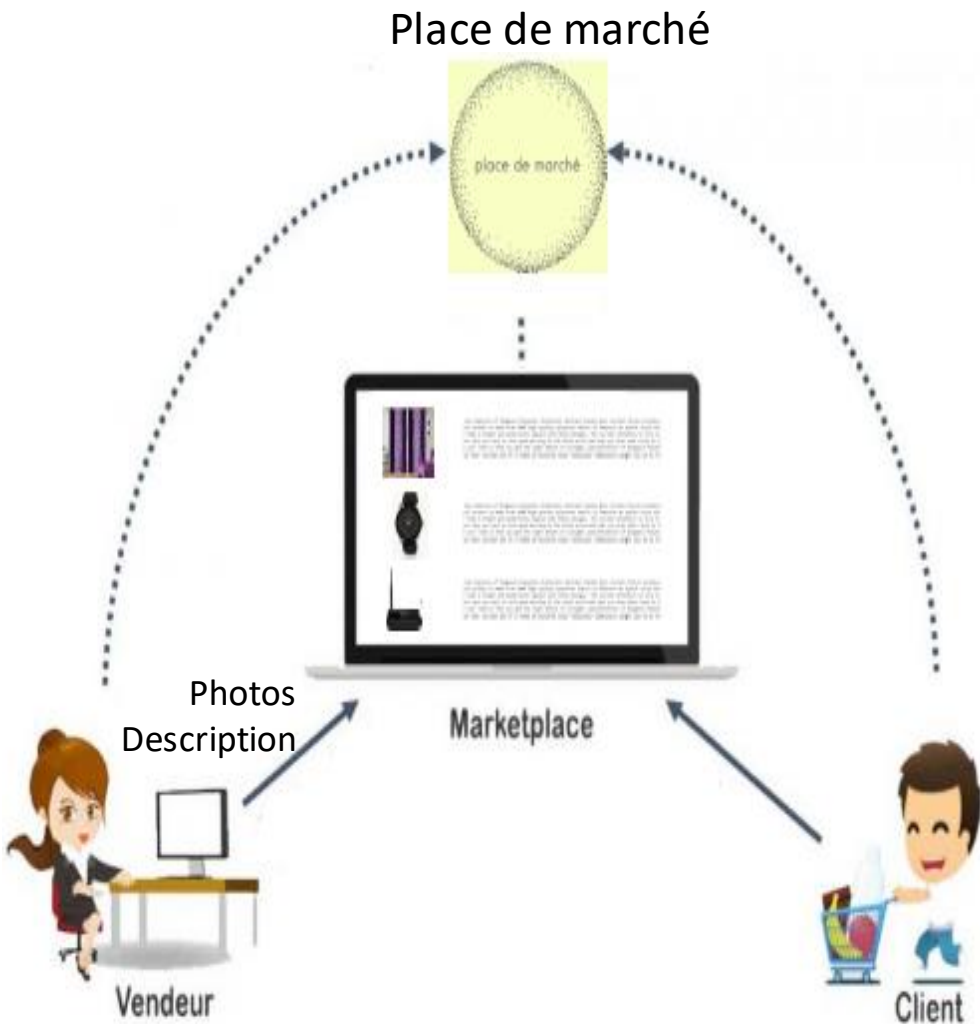
Classifiez automatiquement des biens de consommation

Fanjamalala Rajaonalison
12/2021

Présentation de l'étude

A vertical line is positioned to the right of the title. In the bottom right corner, there is a yellow right-angled triangle pointing towards the top right.

Contexte



Problématiques :

- catégorisation d'article manuelle par les vendeurs --> fiabilité?
- volume des articles très petit --> accroissement?
- facilité de l'expérience utilisateur des vendeurs et des acheteurs

⇒ **automatisation de la tâche**
(approche non supervisé)

Objectif : étudier la faisabilité d'un **moteur de classification** des articles

Dataset

15 colonnes :

- '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'

Selection de 3 features

'["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]'
label (Texte)

'55b85ea15a1536d46b7190ad6fff8ce7.jpg'
image



(Image)

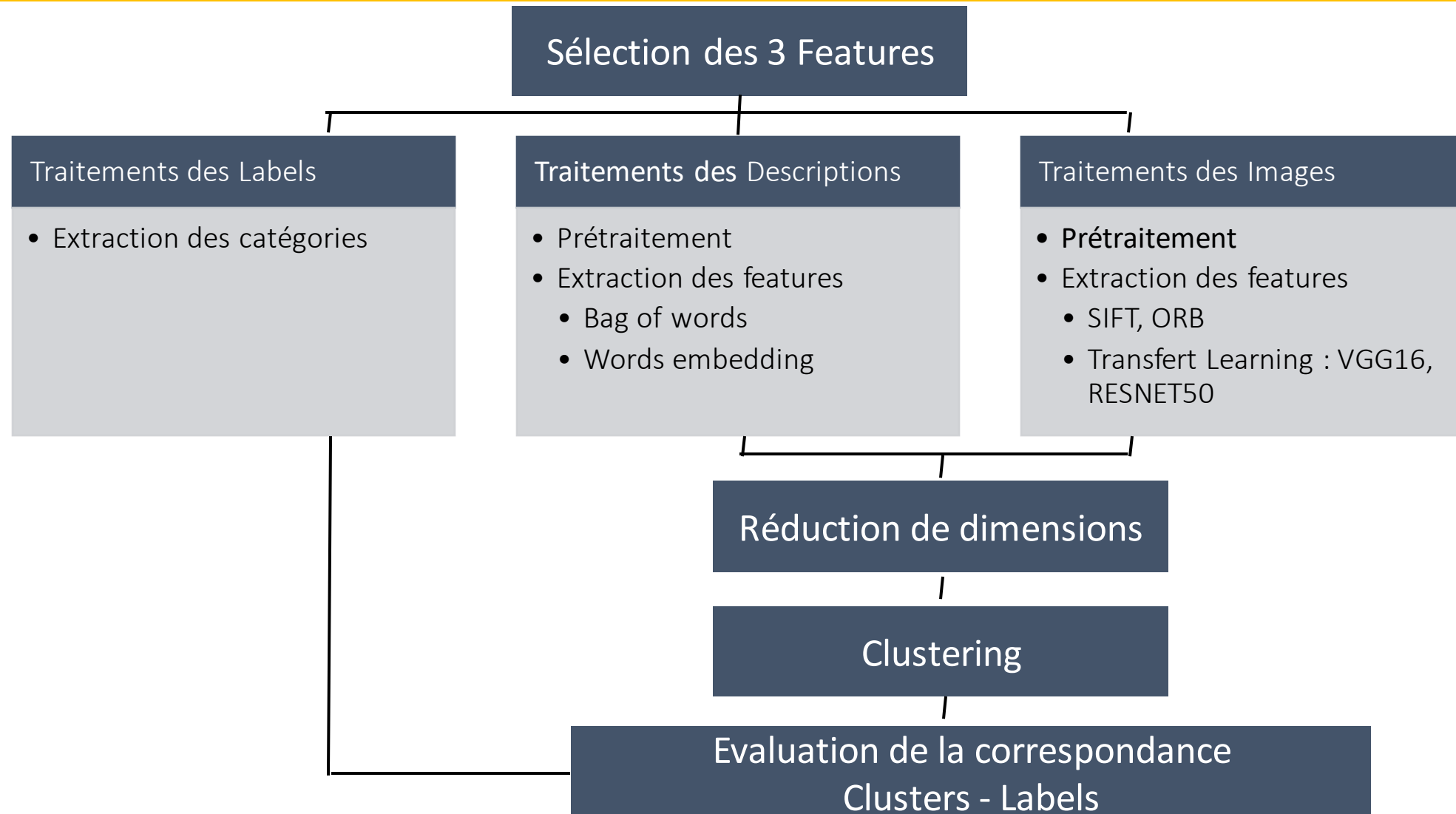
'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant appearance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sun light.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In The Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester'

(Texte)

1050 lignes : articles / produits

description

Méthodologie



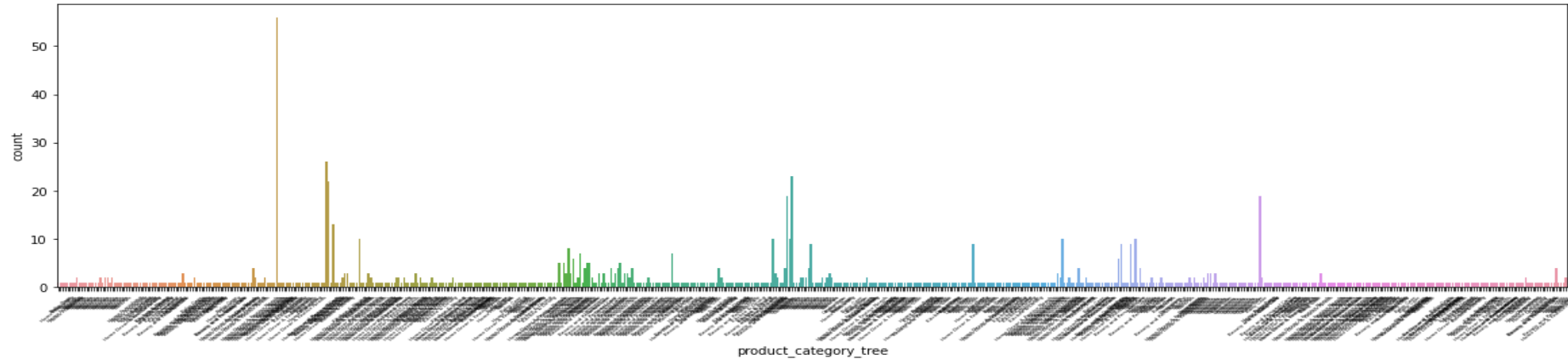
Traitements des Labels

A vertical line is positioned to the right of the title. A yellow triangle is located in the bottom right corner of the slide, pointing towards the top right.

Product_category_tree --> Labels

> arborescence complète

```
0 ["Home Furnishing >> Curtains & Accessories >>...  
1 ["Baby Care >> Baby Bath & Skin >> Baby Bath T...  
2 ["Baby Care >> Baby Bath & Skin >> Baby Bath T...  
3 ["Home Furnishing >> Bed Linen >> Bedsheets >>...  
4 ["Home Furnishing >> Bed Linen >> Bedsheets >>...  
...  
1045 ["Baby Care >> Baby & Kids Gifts >> Stickers >...  
1046 ["Baby Care >> Baby & Kids Gifts >> Stickers >...  
1047 ["Baby Care >> Baby & Kids Gifts >> Stickers >...  
1048 ["Baby Care >> Baby & Kids Gifts >> Stickers >...  
1049 ["Baby Care >> Baby & Kids Gifts >> Stickers >...
```

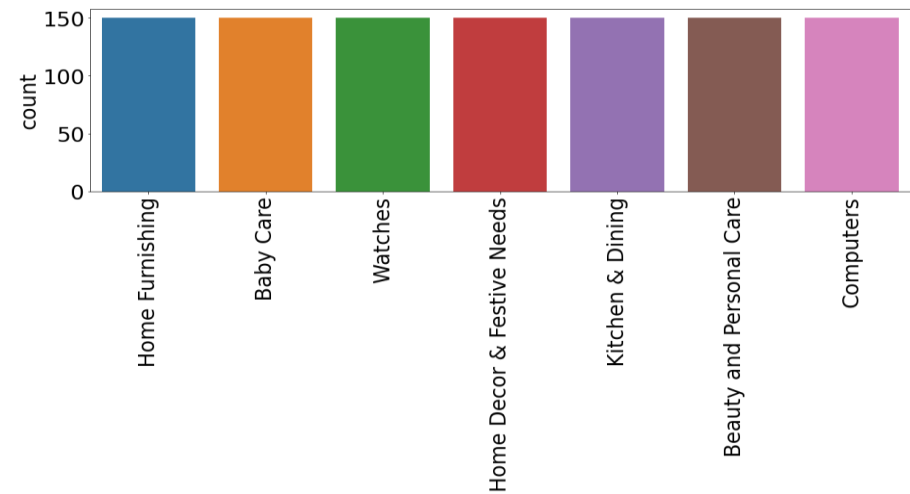


✗ 642 catégories

> premier élément de l'arborescence

```
'["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]'
```

Extraction du premier élément
(Traitement String)



✓ 7 catégories

Traitements des Descriptions

A vertical line is positioned to the right of the title. A yellow triangle is located in the bottom right corner of the slide, pointing towards the center.

Prétraitement

1. Mise en forme

changement des majuscules
en minuscule

2. Tokenization

découpage de la description en
un ensemble de mots d'intérêts

3. Nettoyage

suppression des stop words
et des mots < 3 caractères

4. Normalisation

lemmatization

'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain, Elegance 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% high quality polyester fabric. It features an eyelet style stitch with Metal Ring. It makes the room environment romantic and loving. This curtain is ant- wrinkle and anti shrinkage and have elegant apparence. Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight., Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester'

'key feature elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract eyelet door curtain height pack price curtain enhance look interiors curtain make high quality polyester fabric feature eyelet style stitch metal ring make room environment romantic love curtain ant wrinkle anti shrinkage elegant apparence give home bright modernistic appeal design surreal attention sure steal hearts contemporary eyelet valance curtain slide smoothly draw apart first thing morning welcome bright sun ray want wish good morning whole world draw close even create special moments joyous beauty give soothe print bring home elegant curtain softly filter light room get right amount sunlight specifications elegance polyester multicolor abstract eyelet door curtain height pack general brand elegance design door type eyelet model name abstract polyester door curtain set model duster color multicolor dimension length box number content sales package pack sales package curtain body design material polyester'

Extraction de features : Bag of Words

TF-IDF (Term-Frequency - Inverse Document Frequency)

Conversion de l'ensemble de mots prétraités en une matrice de features TF-IDF

Matrice de poids où pour chaque mot : $\text{poids} = \text{frequence du n-gram} \times \text{idf}(\text{n-gramme})$

	aaa	aapno	aari	aarika	abide	abilities	ability	abkl	able	abrasions	...	zikrak	zinc	zingalalaa	zip	zipexterior	zipper	zone	zoom	zora	zyxel
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

> Dimensions TF-IDF dataset: (1050, 4314)

PCA

Features décorréliées entre elles, réduction de leur dimension tout en gardant un niveau de variance expliquée élevé (99%)

> Dimensions TF-IDF dataset après réduction PCA : (1050, 791)

Extraction de features : Word embedding

word2vec

Représentation des mots dans un espace avec une forme de similarité entre eux et de dimension inférieure, dans lesquels le sens des mots les rapproche dans cet espace
Prise en compte du contexte linguistique des mots.

	0	1	2	3	4	5	6	7	8	9	...	502	503	504	505	
0	-0.053409	-0.053755	0.005996	0.051688	-0.013113	0.048114	-0.024108	0.049812	-0.035168	-0.011691	...	-0.021008	-0.053990	-0.053589	0.053384	0
1	-0.048651	-0.049992	-0.008068	0.044625	0.051083	0.059246	0.051602	-0.041411	-0.036145	-0.046453	...	0.017058	-0.059093	-0.046367	-0.048076	-0
2	-0.053827	-0.050586	-0.013539	0.053693	0.051161	0.053597	0.049655	0.030850	-0.046326	0.024244	...	0.036250	-0.053930	-0.048848	0.029378	0

> Dimensions word2vec dataset: (1050, 512)

PCA

Features décorréliées entre elles, réduction de leur dimension tout en gardant un niveau de variance expliquée élevé (99%)

> Dimensions word2vec dataset après réduction PCA : (1050, 318)

Traitements des Images

A vertical line is positioned to the right of the title. A yellow triangle is located in the bottom right corner of the slide, pointing towards the top right.

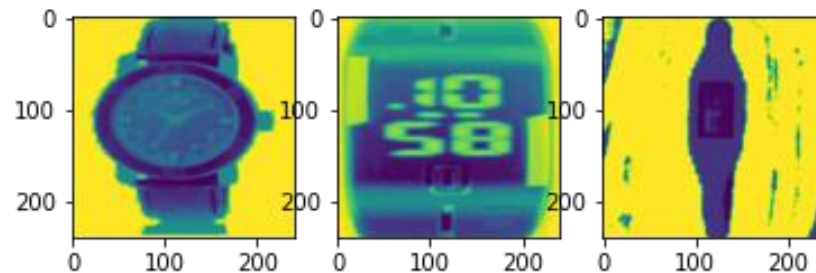
Prétraitement

1. Conversion en gris

2. Redimensionnement

3. Egalisation de l'histogramme

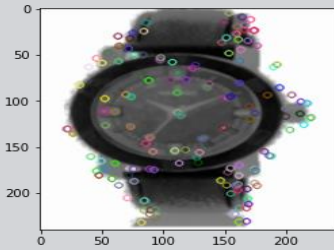
4. Réduction des bruits



Extraction de features : Bag of Visual Words

SIFT

- Extraction de descripteurs



Descripteurs : (151, 128)

Descripteurs Totaux : (303338, 128)

- Création de cluster de descripteurs

MiniBatchKMeans, $n = \sqrt{303338} = 551$

Clusters_Descripteurs : (1050, 551)

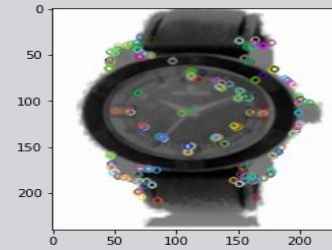
- Réduction de dimensions

PCA, variance expliquée = 99%

Clusters_Descripteurs_pca : (1050, 450)

ORB

- Extraction de descripteurs



Descripteurs : (271, 32)

Descripteurs Totaux : (479850, 32)

- Création de cluster de descripteurs

MiniBatchKMeans, $n = \sqrt{479850} = 693$

Clusters_Descripteurs : (1050, 693)

- Réduction de dimensions

PCA, variance expliquée = 99%

Clusters_Descripteurs_pca : (1050, 548)

Extraction de features : Transfer Learning

VGG-16

- DESCRIPTION : réseau de neurones convolutifs de **16 couches**
 - PRE-ENTRAINEMENT : sur plus d'un million d'images à partir de la base de données ImageNet
 - ENTREE : image couleurs de dimensions 224x224
 - PRINCIPE : classification de l'image dans l'une des 1000 classes de ImageNet
 - SORTIE : vecteur des probabilités d'appartenance à chacune des classes
- VGG_Dataset : (1050, 25088)
- PCA : réduction de dimensions

VGG_Dataset_PCA : (1050, 939)

RESNET50

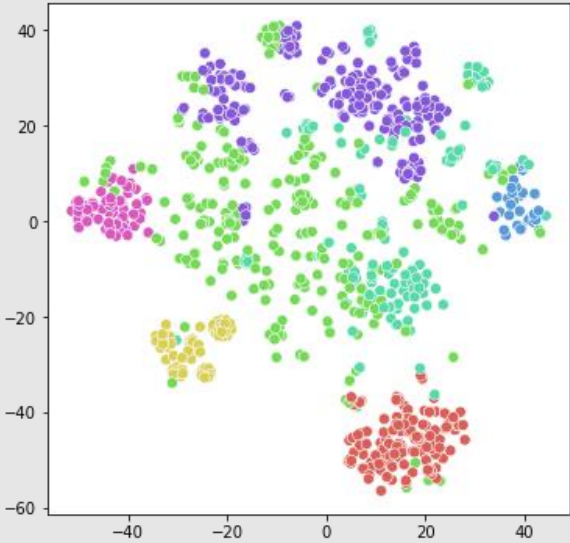
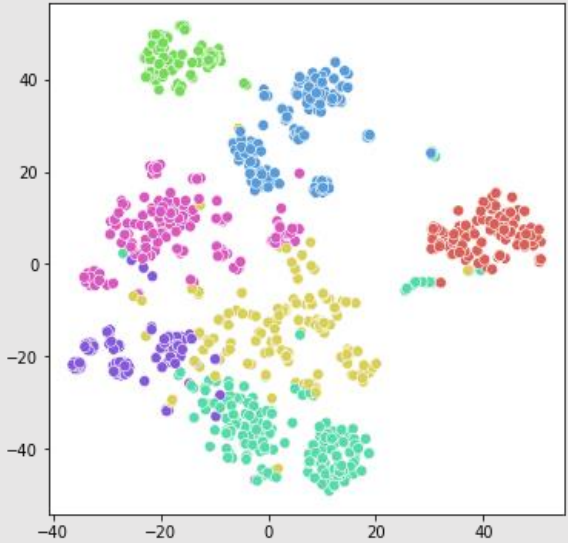
- DESCRIPTION : réseau de neurones convolutifs de **50 couches**
 - PRE-ENTRAINEMENT : sur plus d'un million d'images à partir de la base de données ImageNet
 - ENTREE: image couleurs de dimensions 224x224
 - PRINCIPE : classification de l'image dans l'une des 1000 classes de ImageNet
 - SORTIE: vecteur des probabilités d'appartenance à chacune des classes
- RESNET_Dataset : (1050, 131072)
- PCA : réduction de dimensions

RESNET_Dataset_PCA : (1050, 976)

Clustering

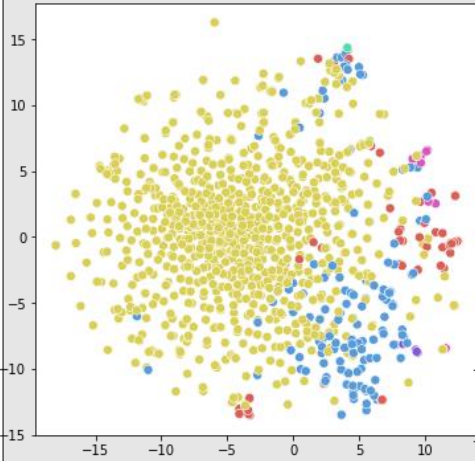
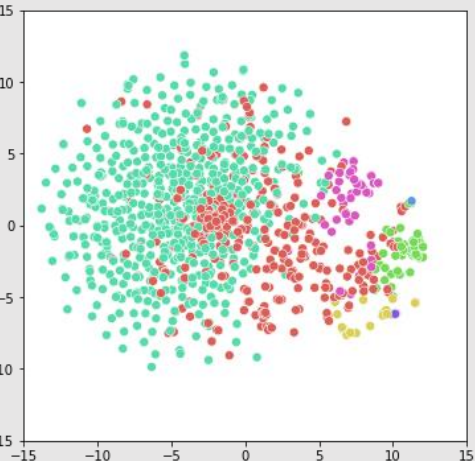
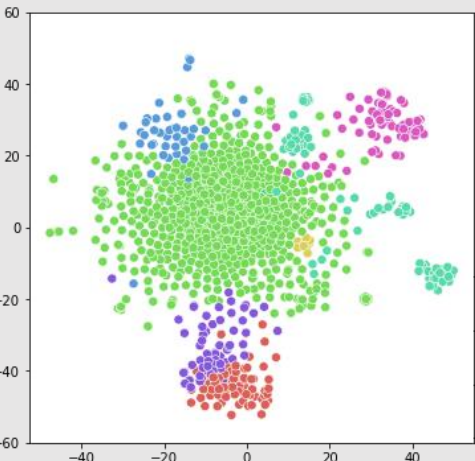
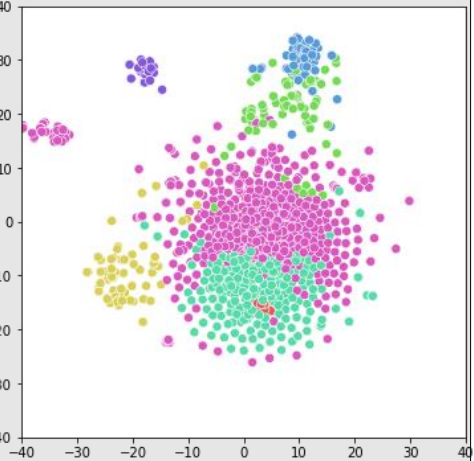
A vertical line is positioned to the right of the word "Clustering". In the bottom right corner of the slide, there is a yellow right-angled triangle pointing towards the top-left.

Clustering des Descriptions

	K-MEANS avec n-clusters = 7	
	TF-IFD	word2vec
Score Silhouette	0.060	0.108
Score D. Bouldin	4.375	2.702
Diagramme TSNE		
ARI	0.281	0.325



Clustering des Images

	K-MEANS avec n-clusters = 7			
	SIFT	ORB	VGG-16	RESNET50
Score Silhouette	0.169	0.026	-0.018	0.018
Score D. Bouldin	2.058	2.438	4.253	4.105
Diagramme TSNE				
ARI	0.011	0.028	0.133	0.214



Conclusion de l'étude

A vertical line is positioned to the right of the text. In the bottom right corner of the slide, there is a yellow right-angled triangle pointing towards the top right.

Résultats

Descriptions

- Moteur de classification faisable mais avec des scores pas très bons

-> Améliorations ?

- Correspondance des clusters avec Labels faible

-> Fiabilité des Labels ?

-> Améliorations ?

Images

- Moteur de classification faisable mais avec des scores pas très bons

-> Améliorations ?

- Correspondance des clusters avec Labels faible

-> Fiabilité des Labels ?

-> Améliorations ?

Améliorations

Amélioration des prétraitements de textes

-> suppression des mots les plus fréquemment utilisés, limitation de vocabulaire

Amélioration des prétraitements des images

-> préservation du ration longueur/largeur, amélioration des contrastes

Prise en compte simultanée des Descriptions & Images

-> concaténation des features avec les meilleurs scores

Essais d'autres modèles de clustering

-> DBSCAN, clustering hiérarchique, ...

Optimisation des modèles

-> GridSearchCV

Approche supervisée

MERCI

Questions



Incompréhensions

