OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Soutenance du Projet 6

Classifiez automatiquement des biens de consommation

Par
Elisée TCHANA

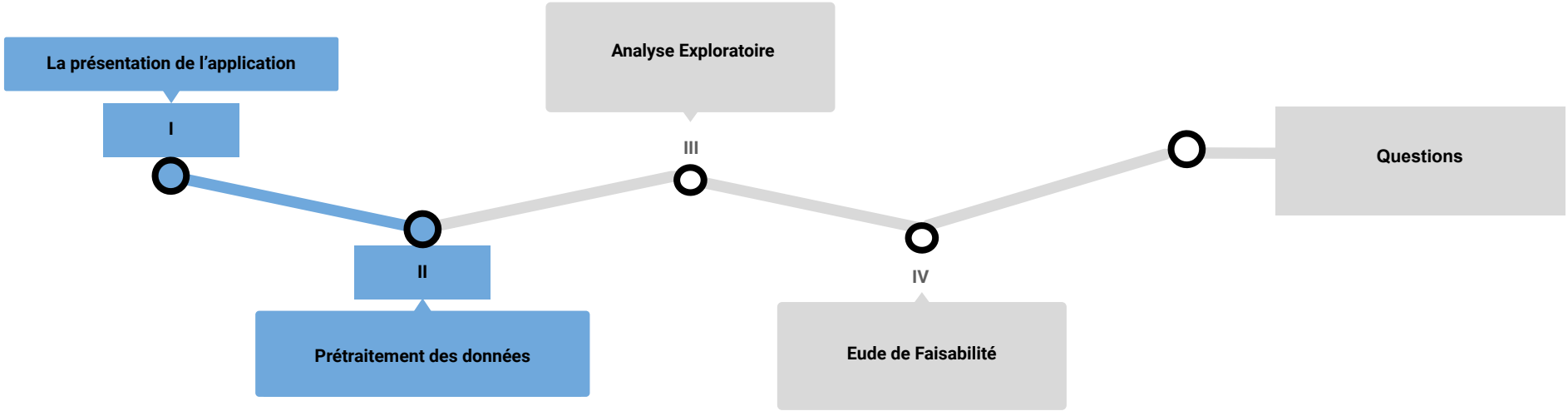
Mentor
Cyril MONTI

Août 2021

OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Plan de Soutenance

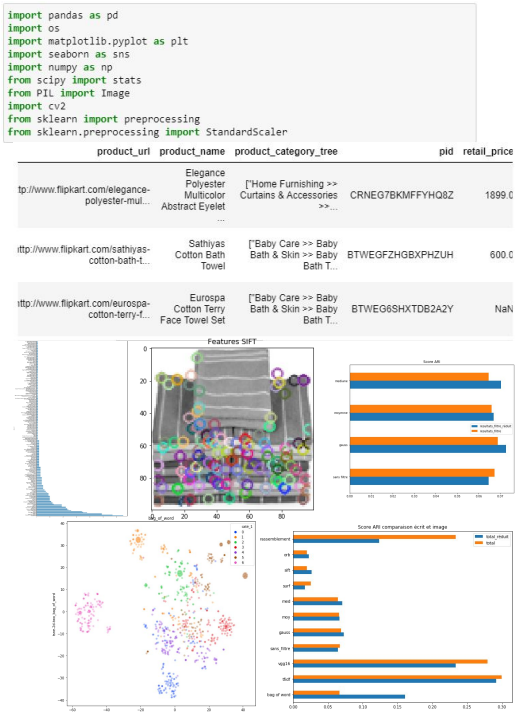


OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Data Analysis Workflow

Dataset

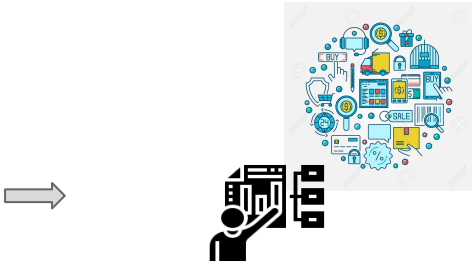


```
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy import stats
from PIL import Image
import cv2
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
```

product_url	product_name	product_category_tree	pid	retail_price
tp://www.flipkart.com/elegance-polyester-mul...	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >> ...	CRNEG7BKMFFYHQ8Z	1899.0
ttp://www.flipkart.com/sathiyas-cotton-bath-t...	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEGFZHGBXPHZUH	600.0
ttp://www.flipkart.com/eurospa-cotton-terry-f...	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEG6SHXTDB2A2Y	NaN

Features SIFT

Data Modelling & Analysis



Data Sharing

OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Objectifs du Projet

Outil de classification de produits

Objectifs

- Prétraiter des données texte pour obtenir un jeu de données exploitable
- Prétraiter des données image pour obtenir un jeu de données exploitable
- Représenter graphiquement des données à grandes dimensions.
- Mettre en œuvre des techniques de réduction de dimension

Informations contenues dans le jeu de données

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_price
0	55b85ea15a1536d46b7190ad6fff8ce7	2016-04-30 03:22:56 +0000	http://www.flipkart.com/elegance-polyester-mul...	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >>...	CRNEG7BKMFFYHQ8Z	1899.0
1	7b72c92c2f6c40268628ec5f14c6d590	2016-04-30 03:22:56 +0000	http://www.flipkart.com/sathiyas-cotton-bath-t...	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEGFZHGBXPHZUH	600.0
2	64d5d4a258243731dc7bbb1eef49ad74	2016-04-30 03:22:56 +0000	http://www.flipkart.com/eurospa-cotton-terry-f...	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEG6SHXTDB2A2Y	NaN

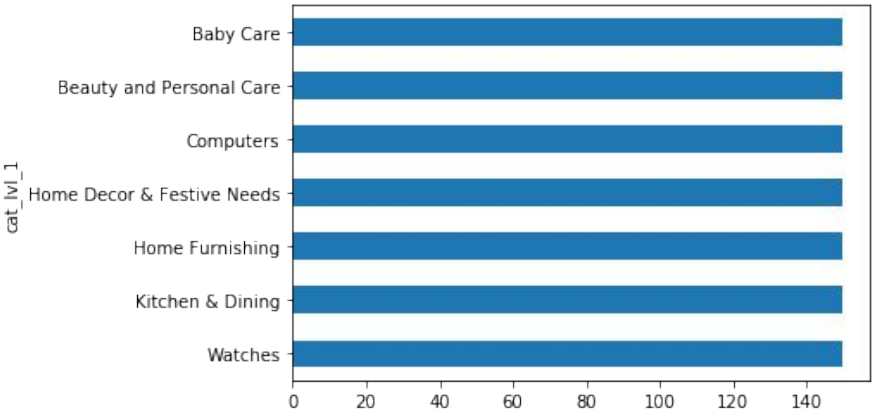
```
df.shape
```

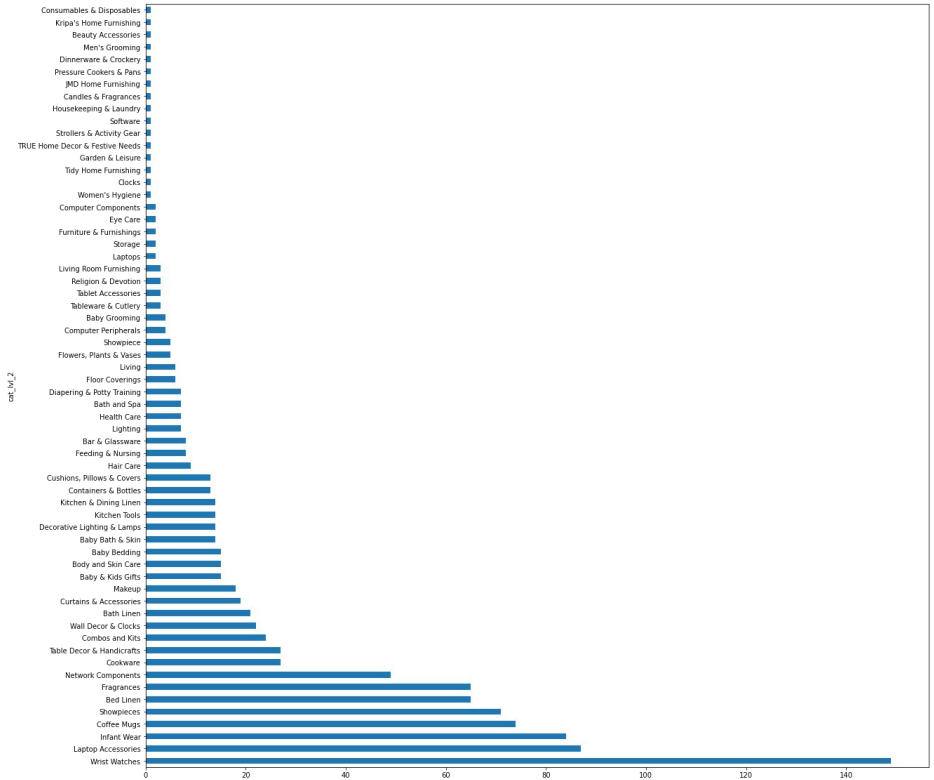
```
(1050, 15)
```

- **Informations produits**
- **Informations tarifaires**
- **Notes produits**
- **Images produits**

Informations contenues dans le jeu de données

Catégories





OPENCLASSROOMS

Traitement des images

CentraleSupélec | université PARIS-SACLAY

Méthodes

Filtres images

1. Sans filtre (transformation seulement en matrice noir et blanc)
2. Moyenne (application d'un filtre de moyenne)
3. Gauss (limite le temps de montée et de descente)
4. Médiane (application d'un filtre médiane)

BAG of visuel words

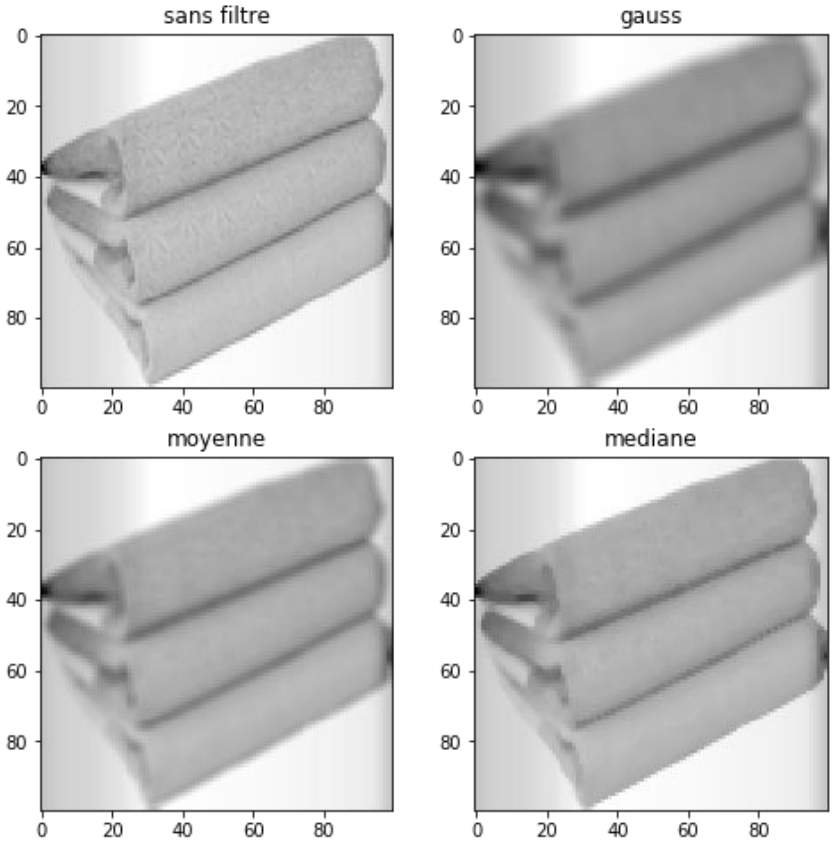
1. SIFT (utilise la moyenne gaussienne)
2. SURF (filtre en boite / utilise les carrés)
3. ORB (fusion pour trouver points clés et descripteur)

Transfer learning

1. VGG16

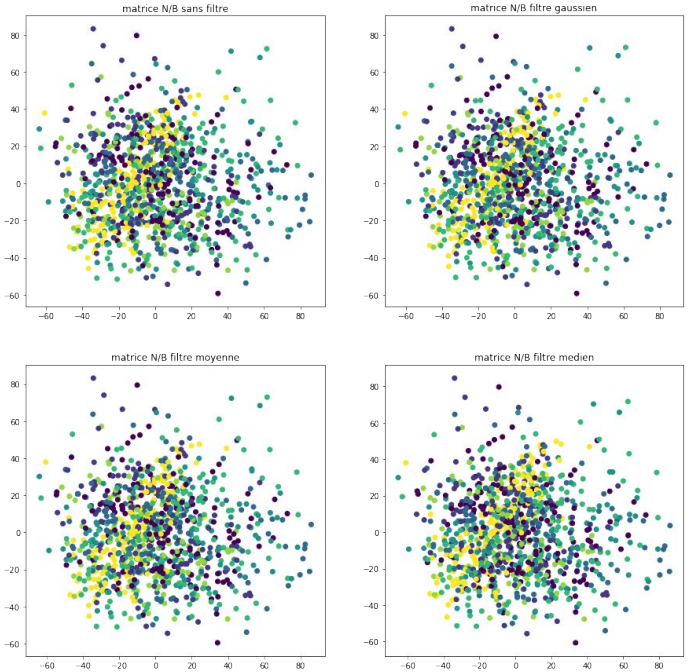
Traitement des images

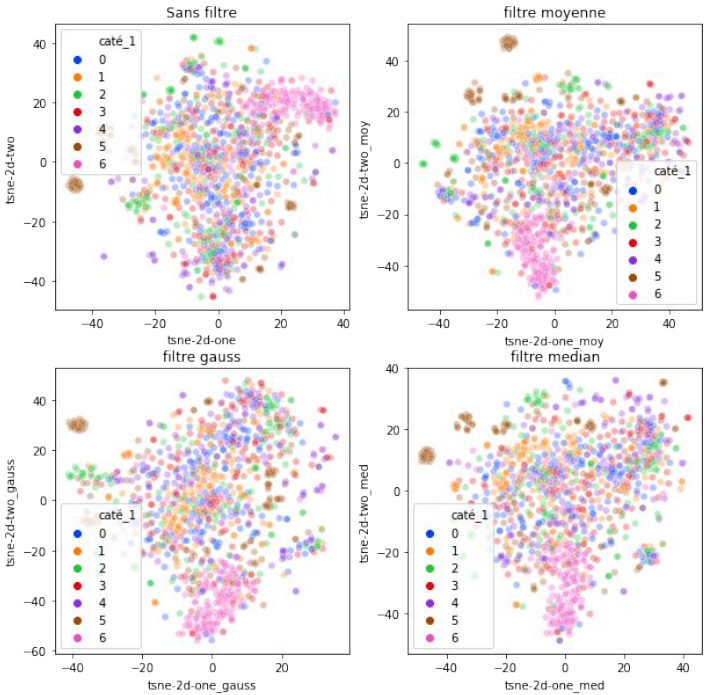
Filtre D'images



Traitement des images: Visualisation

Filtre D'images



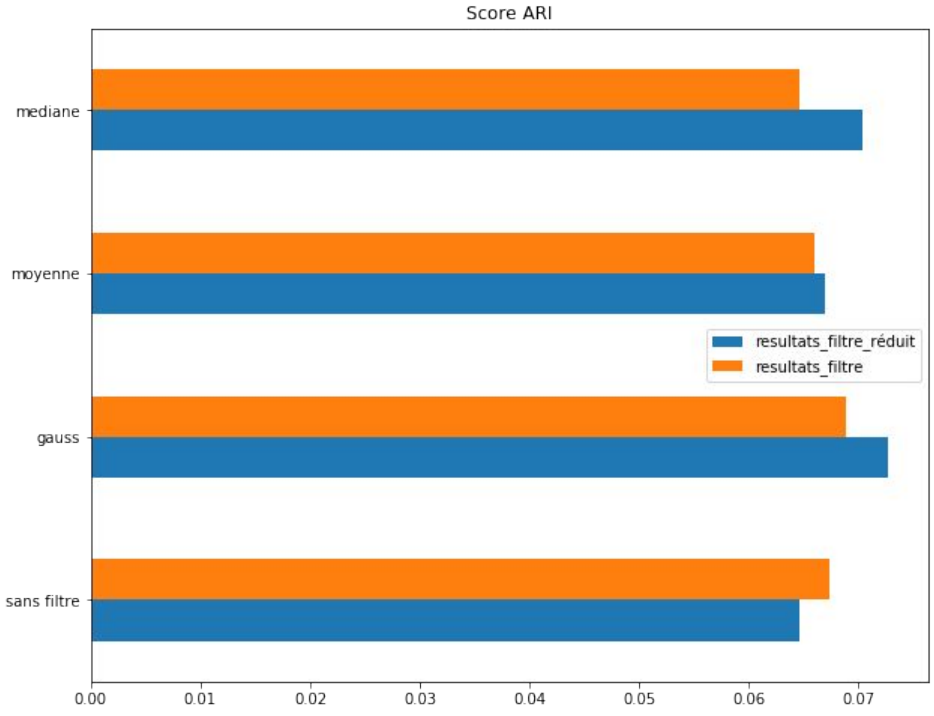


Pca (n_components=0,8)

Sans filtre	Gauss	Moyenne	Médiane
10000□197	10000□ 58	10000□ 93	10000□ 126

Traitement des images: Visualisation

Filtre D'images

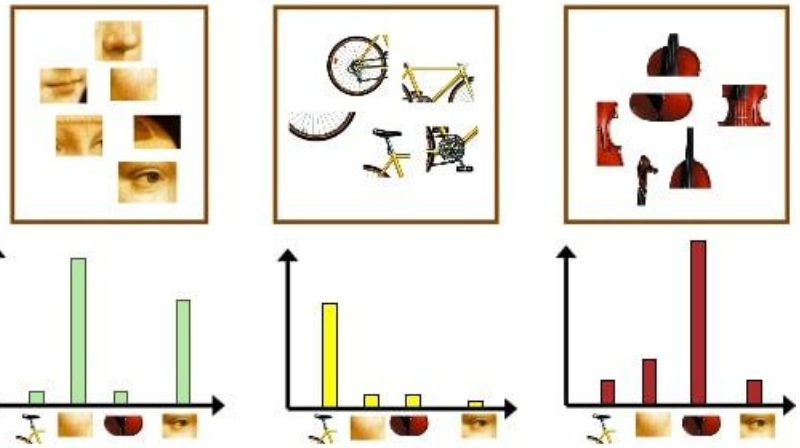


Traitement des images: Bag of visuel words

Filtre D'images

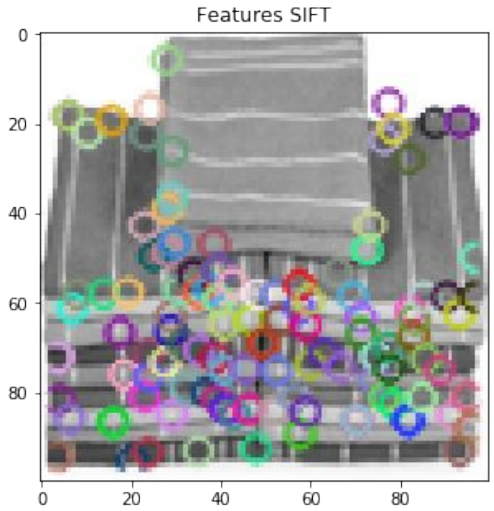
Bag of *visual words*

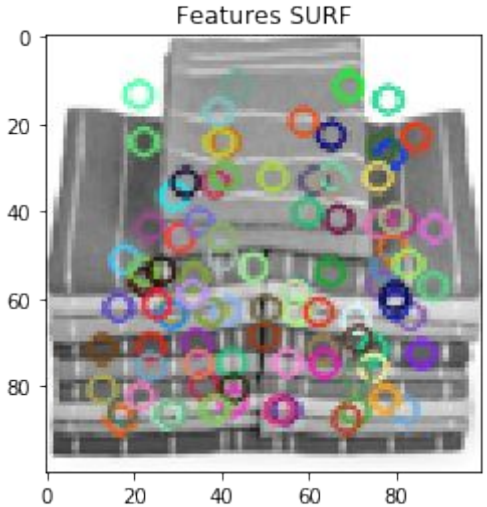
- Image patches
- BoW histogram
- Codewords

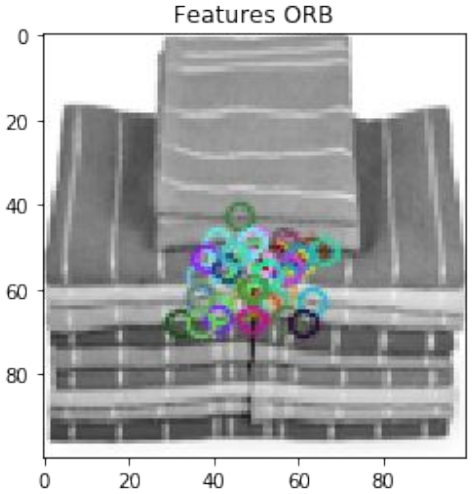


Traitement des images: Bag of visuel words

Filtre D'images

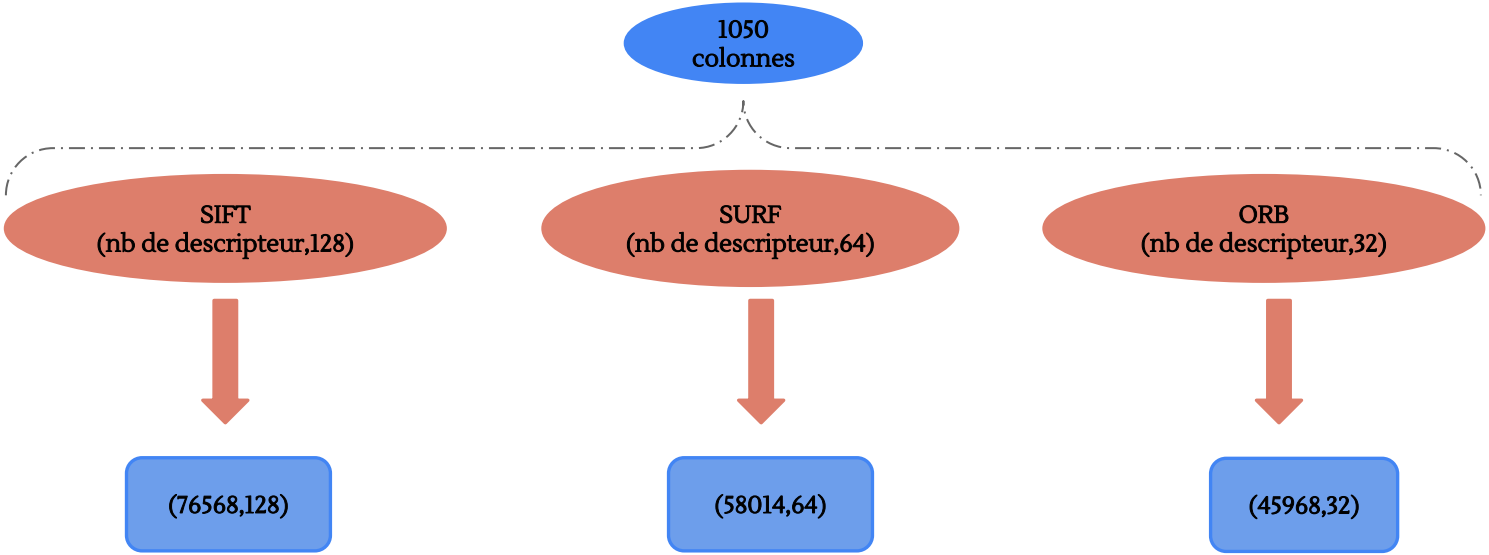






Traitement des images: Bag of visuel words

Image Classification



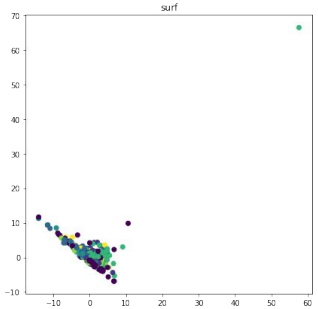
OPENCLASSROOMS

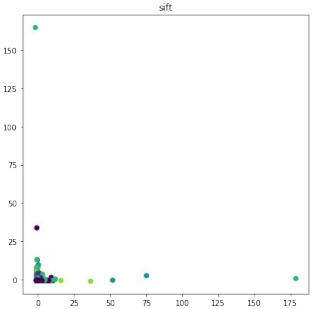
CentraleSupélec | université PARIS-SACLAY

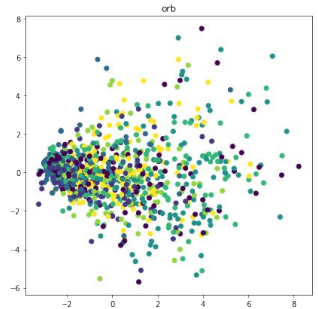
Traitement des images: Kmeans

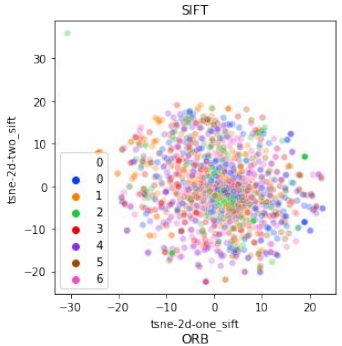
Kmeans

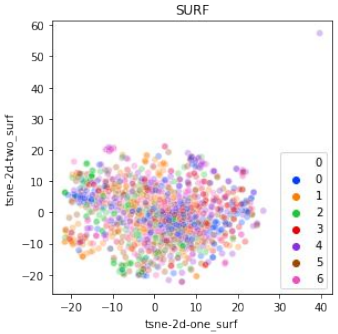
Nb de cluster = racine de la taille matrice → L'implémenter à une matrice de 0

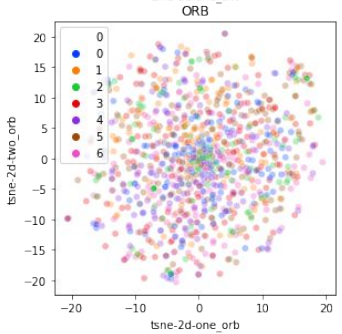










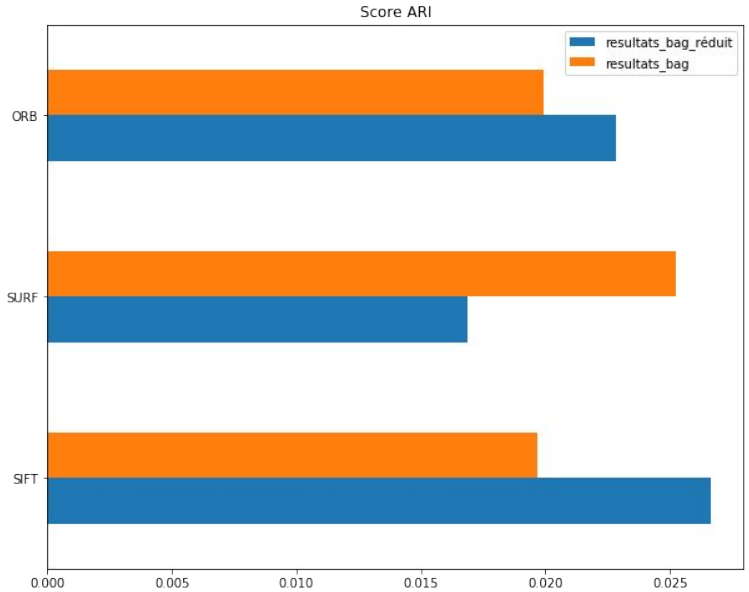


Réduction de dimension → Pca (n_components=0,8)

SIFT	SURF	ORB
276 □ 126	240 □ 126	214 □ 113

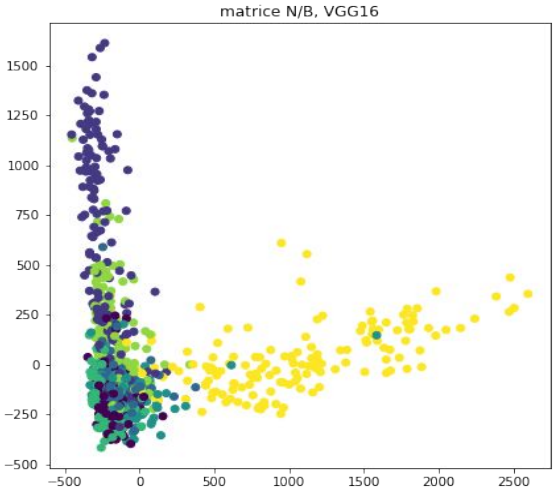
Traitement des images: Kmeans Score ARI

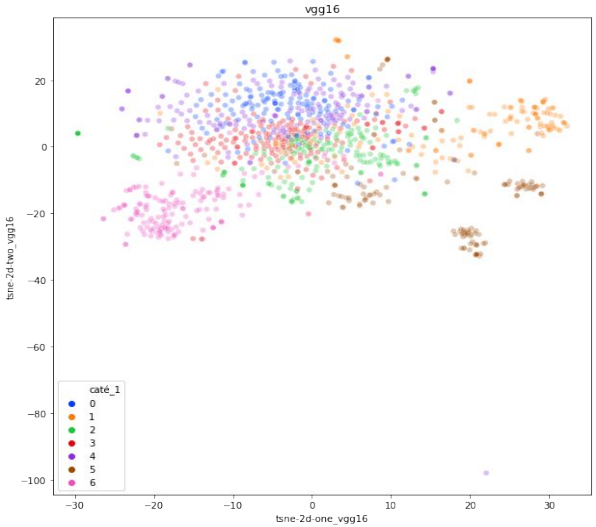
Image Classification

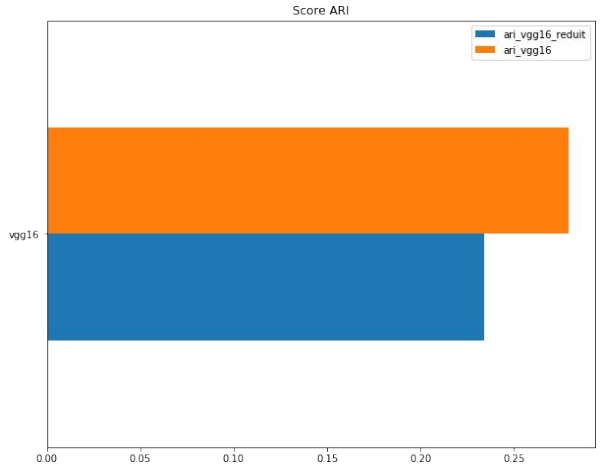


Traitement des images: Transfer Learning

Image Classification



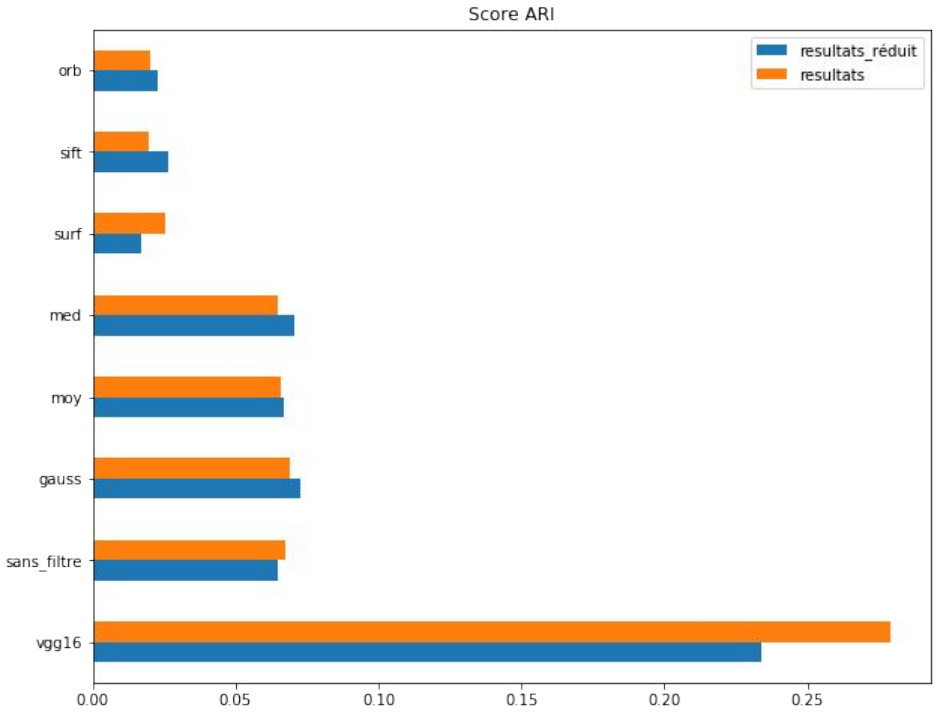




Réduction de dimension → Pca (n_components=0,8) → vgg16 25088 □ 477

Traitement des images

Conclusion



Traitement de text

Processing

TOKENISER → LEMMATISER → STOPWORDS

Réduction de dimension → Pca (n_components=0,8)

vgg16
25088 □ 477

Traitement de text

Processing

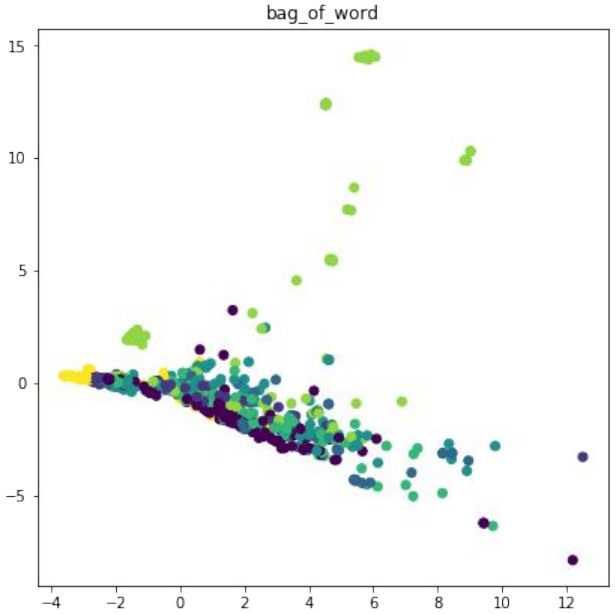
TOKENISER → LEMMATISER → STOPWORDS

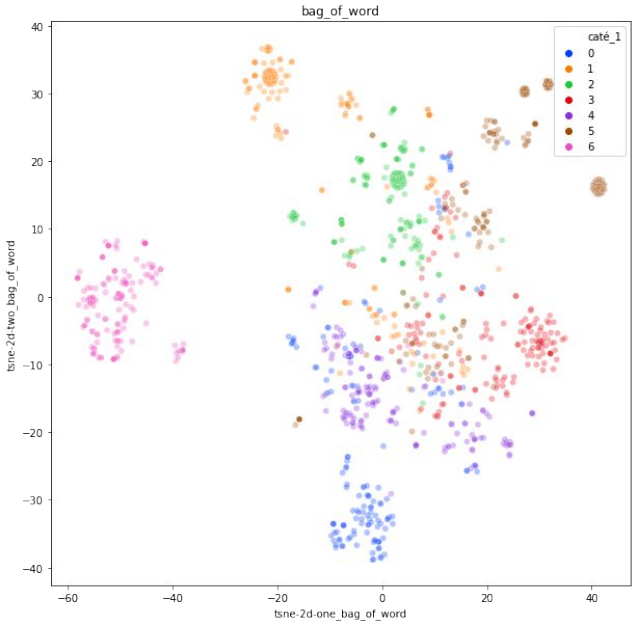
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398
0	0	0	0	1	0	0	0	0	0	0	0	5	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	...	0	0	0	0	1	0	0	1	0	0	0	0	0	0	2	1	0
2	0	0	1	2	0	0	0	0	0	0	0	0	0	1	1	0	0	...	1	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0
3	0	3	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	...	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0
4	4	4	0	2	0	0	0	0	0	0	0	4	0	0	0	0	0	...	1	0	3	0	3	0	0	0	0	0	0	0	0	0	0	1	0

Bag of Words

Traitement de text: Visualisation

Processing



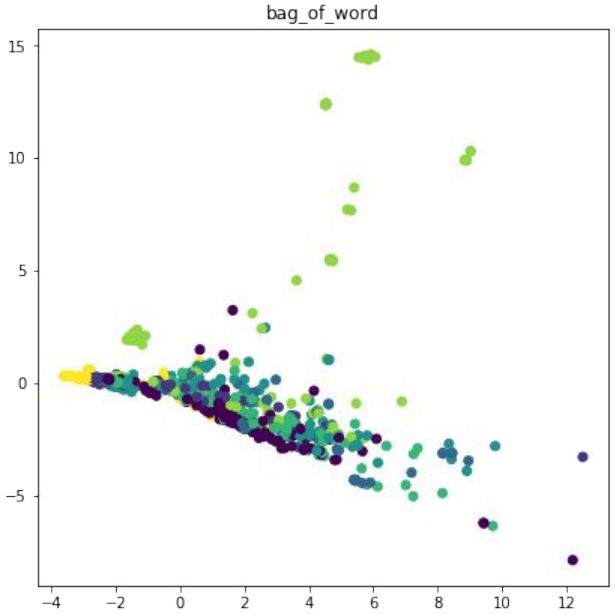


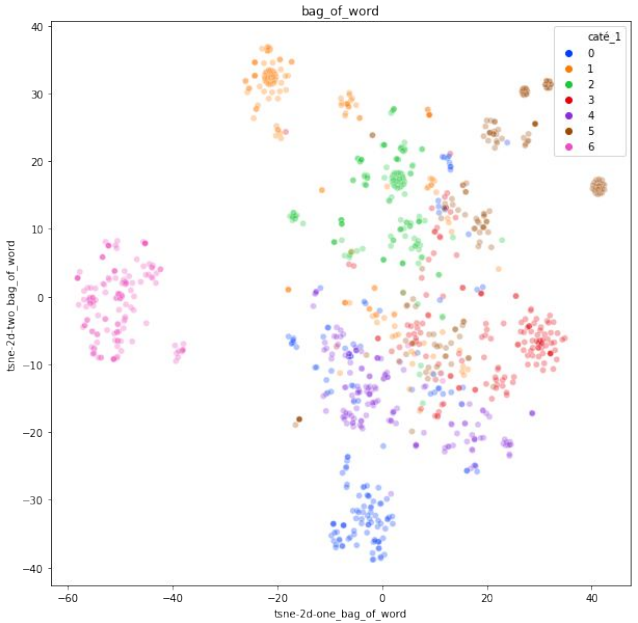
Réduction de dimension → Pca (n_components=0,8)

Bag of words 399□ 96

Traitement de text: Visualisation

Processing





Réduction de dimension → Pca (n_components=0,8)

Bag of words 399□ 96

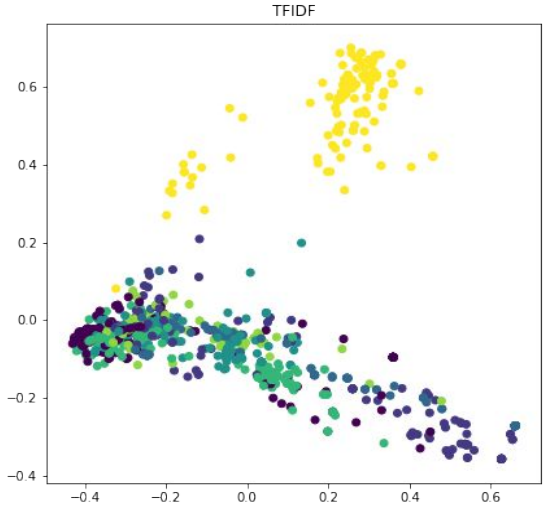
OPENCLASSROOMS

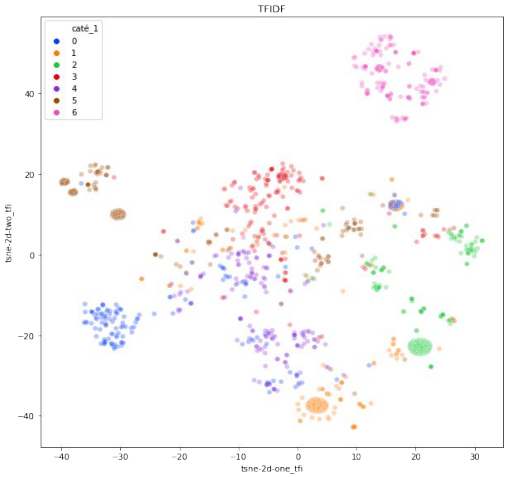
CentraleSupélec | université PARIS-SACLAY

Traitement de text: Visualisation

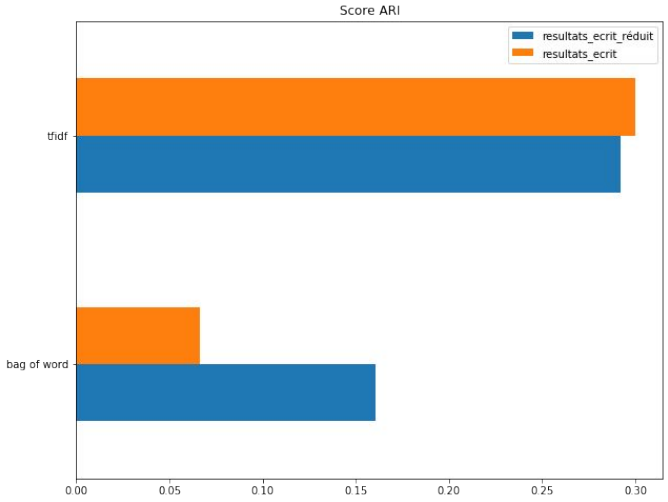
Processing TFIDF

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	157	158	159	16
0	0.000000	0.000000	0.000000	0.097922	0.0	0.0	0.401820	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.06216
1	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.368411	0.066265	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.09265
2	0.000000	0.000000	0.092114	0.166937	0.0	0.0	0.000000	0.093868	0.070240	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.285367	...	0.0	0.0	0.0	0.05299
3	0.000000	0.136861	0.000000	0.000000	0.0	0.0	0.191618	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.04941
4	0.275326	0.147125	0.000000	0.125497	0.0	0.0	0.205989	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.03983



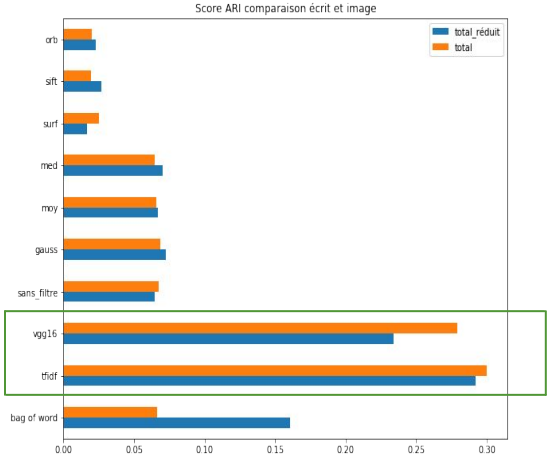


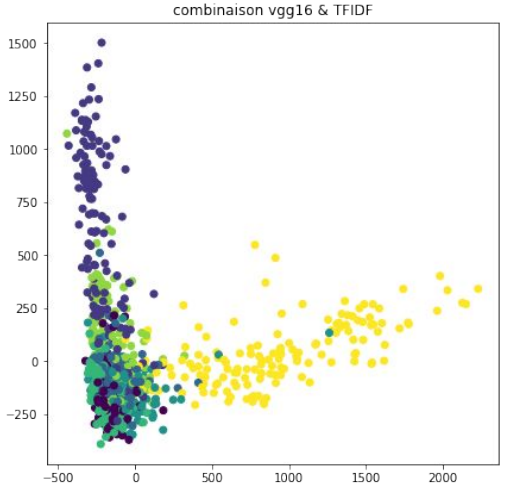
Comparison

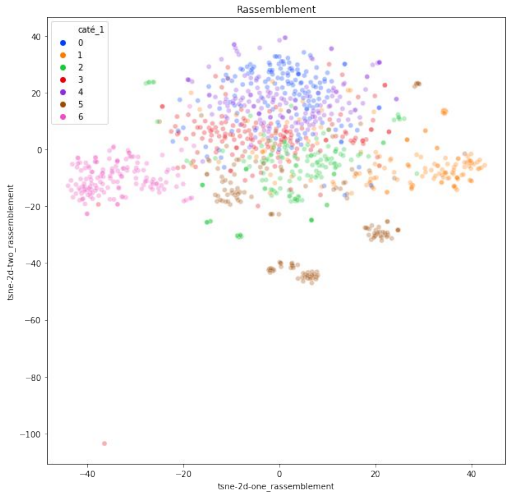


Réduction de dimension → Pca (n_components=0,8) → TFIDF 174□ 73

Visualisation traitement : Value Image & Text







Réduction de dimension → Pca (n_components=0,8)

Rassemblement 550□ 298

Conclusion

Méthodes utilisées

Pour le traitement des images plusieurs algorithmes :

1. Moyenne
2. Médiane
3. Gauss
4. SIFT
5. SURF
6. ORB
7. VGG16

Pour le traitement de texte

1. Bag of words
2. TFIDF (Term frequency-inverse document frequency)

Deux méthodes utilisées pour visualiser les données de grandes échelles:

1. PCA
2. TSNE

Merci pour votre attention !!