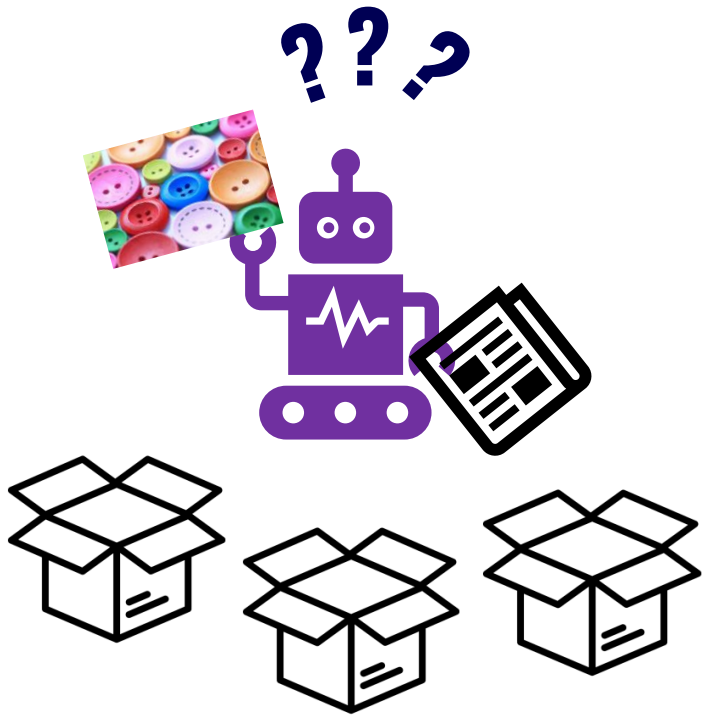
Parcours Data Scientist

Projet N°6 :
Classifiez automatiquement des biens de consommation



Daniel CHASTANET

Sommaire

Non-Supervisé

Rappel de la problématique

Base de données

Natural language processing

Computeur vision

Conclusion

Semi-Supervisé

Rappel de la problématique

Data Scientist au sein de l'entreprise "**Place de marché**", qui souhaite lancer une marketplace e-commerce.

Automatisation de la catégorisation d'article proposé à la vente :
→ **une photo**
→ **une description**.

Linda, lead data scientist, nous demande d'étudier la faisabilité d'un **moteur de classification** avec un niveau de précision suffisant.

Réaliser une première étude de faisabilité
Convaincre Linda avec de beaux graphiques

La base de données

Taille : (1050, 15)
Duplicata : 0

Base de donnée très « propre »

Currently unavailable

« None »

ps manquante, tout le reste est là

	attributs	valeurs_manquantes	type
0	uniq_id	0	object
1	crawl_timestamp	0	object
2	product_url	0	object
3	product_name	0	object
4	product_category_tree	0	object
5	pid	0	object
6	retail_price	1	float64
7	discounted_price	1	float64
8	image	0	object
9	is_FK_Advantage_product	0	bool
10	description	0	object
11	product_rating	0	object
12	overall_rating	0	object
13	brand	338	object
14	product_specifications	1	object

La base de données

product_category_tree
["Home Decor & Festive Needs >> Decorative Lighting & Lamps >> Floor Lamp >> Nutcase Floor Lamp >> Nutcase Multicolor Column Floor Lamp (31 cm)"]

product_name	description	brand
Nutcase Multicolor Column Floor Lamp	Key Features of Nutcase Multicolor Column Floor Lamp Designer Lamp Bulb Included Cool & Quirky Design Sturdy Structure,Nutcase Multicolor Column Floor Lamp (31 cm) Price: Rs. 1,299 Nutcase brings you for the first time in India - Design...	Nutcase

image
e488005c7fb68747d3458c7d73760bae.jpg

NLP (Non supervisé)

Pré traitement du texte

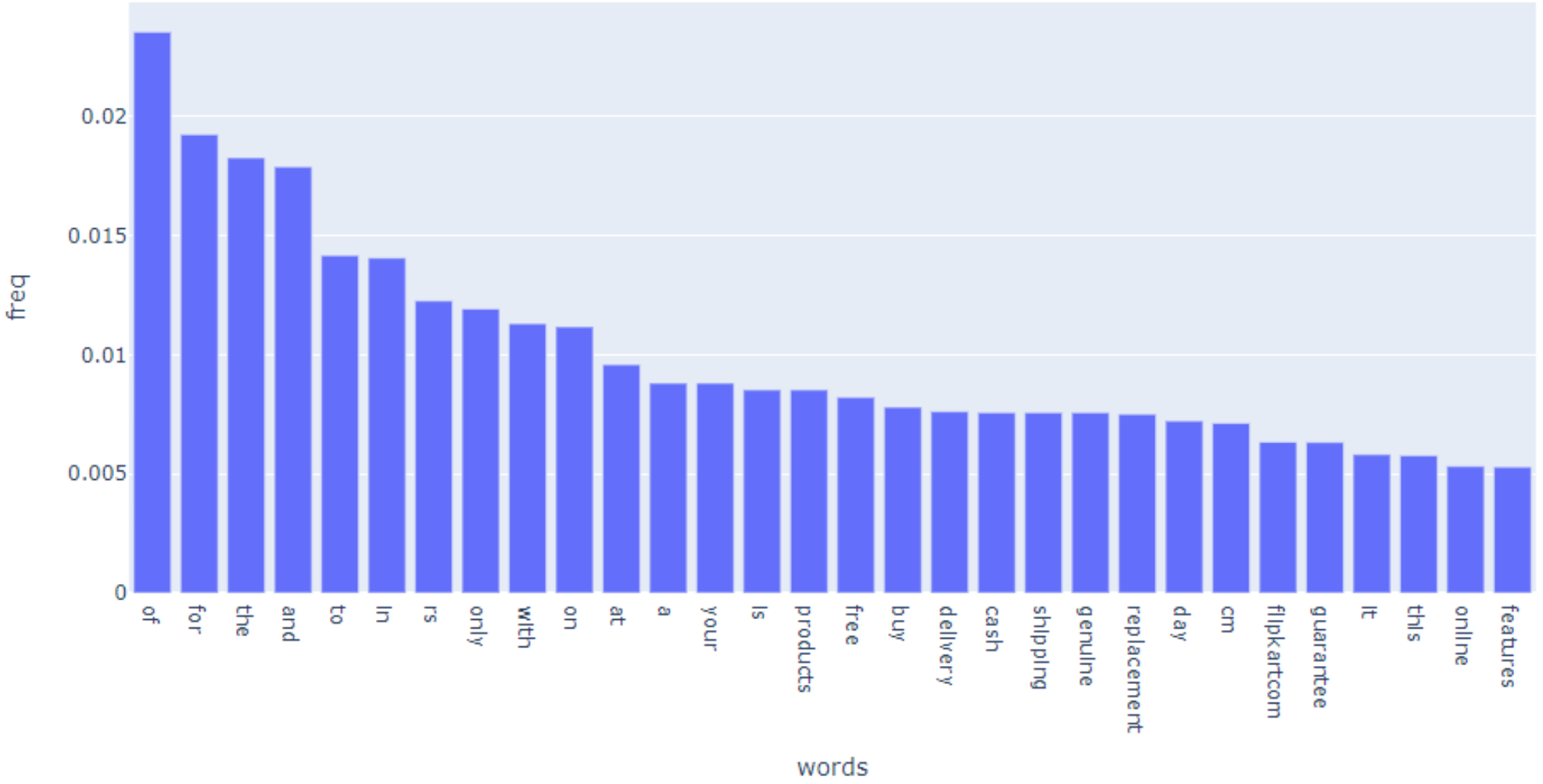
Phrase témoin

Retrait ponctuation + minuscule

Tokenization

Lemmatization

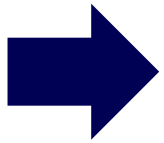
Retrait des stop words

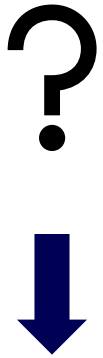


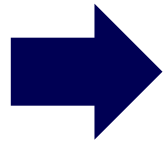
99gems, smart, otg, connection, kit, usb, usb, cable, black, 199, feature, work, microusb, smart, phone, device, ontogo, function, otg, allows, connect, usb, device, keyboard, mouse, usb, flash, drive, otg, compatible, phone, tablet, small, light, easy, carry, plug, play, compatible

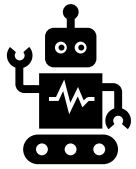
Sac de mots (bag of words)

Corpus
Pré traité









Algorithme

	Il	Faire	Beau	Moche	Être	inverse
Il faire beau	1	1	1	0	0	0
Il faire moche	1	1	0	1	0	0
Moche être inverse beau	0	0	1	1	1	1

(1050, 6121)  sparse matrix (matrice creuse)

Sac de mots (bag of words)

sparse matrix

PCA

t-SNE

Algorithmes de reduction de dimension
- Réduction des coûts
- Amélioration de l'apprentissage
- Modèle moins complexe
- Elimination de variables non pertinentes
- Combat le fléau de la dimensionalité.

TF-IDF

Term frequency *
Inverse document frequency

N-gramms

Il faire beau depuis hier
→ (il, fait), (fait, beau), (beau, depuis), (depuis, hier)

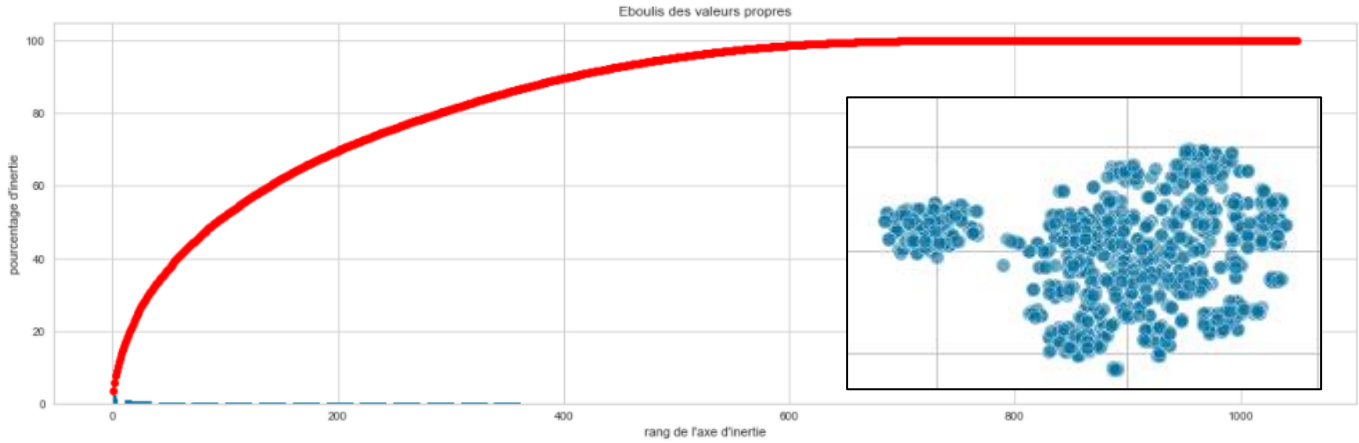
$$P(w_{i-1}|w_i) = \frac{count\ (w_i, w_{i-1})}{count\ (w_{i-1})}$$

Combinaisons des 3 techniques

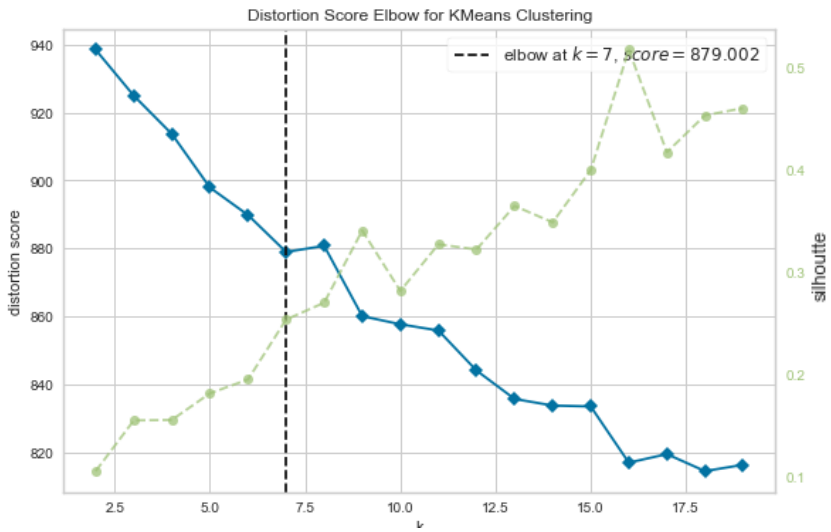
Clustering (K-Means)

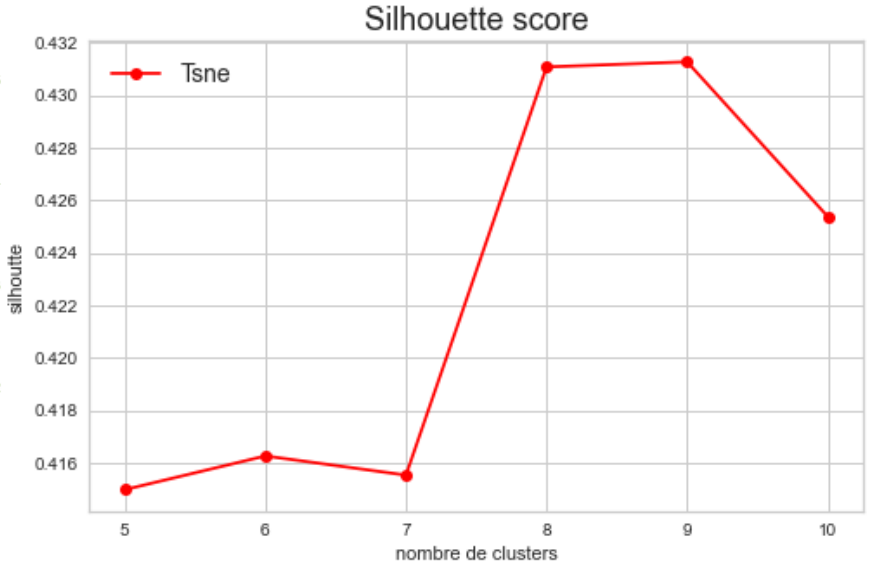
PCA suivi d'un t-SNE

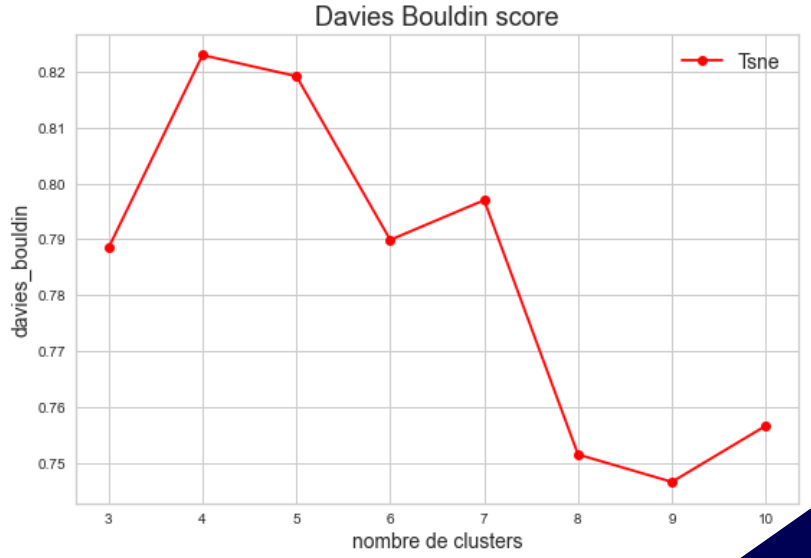
TF-IDF → Unigrams / Bigrams / Trigrams →



→ K-means







Clustering (K-Means)

Most frequent words in cluster N° 0

118	cover
299	single
106	cotton
192	cushion
110	double
43	design
2	multicolor
13	pack
1	polyester
9	floral

Literie

Most frequent words in cluster N° 1

635	laptop
518	skin
281	warranty
290	battery
1436	shape
3	print
155	quality
7	set
295	kadhai
832	cell

Related to computers

Most frequent words in cluster N° 2

3	flipkartcom
4	30
5	guarantee
0	iball
6	netis
9	airtel
12	digisol
1	wr7011a
7	wf2301
10	b310s927

Network

Most frequent words in cluster N° 3

4	watch
3	analog
18	men
145	woman
14	none
13	guarantee
6	online
12	30
10	great
11	discount

Watch

Most frequent words in cluster N° 4

52	usb
22	light
3	led
168	pizza
50	flexible
169	cutter
155	30
135	fan
134	portable
156	guarantee

Electronic accessories

Most frequent words in cluster N° 5

241	mug
1	ceramic
316	perfect
279	design
60	feature
9	ml
192	material
246	coffee
250	gift
805	adapter

Mugs

Most frequent words in cluster N° 6

111	baby
27	girl
1	cotton
120	detail
116	fabric
3	towel
26	boy
185	dress
126	1
16	general

Enfants / habillement

Most frequent words in cluster N° 7

92	showpiece
105	wall
8	best
27	1
9	30
10	guarantee
7	online
13	gift
348	brass
14	box

Name: words, dtype: object

Fait main / cadeau

Most frequent words in cluster N° 8

81	combo
191	flipkartcom
84	set
194	30
195	guarantee
196	none
216	online
137	lip
523	oxyglow
526	cream

Beauty / health

LDA

	word1	word2	word3				Word-n
Topic-1	0.024	0.012	0.014	-	-	-	0.086
Topic-2	0.026	0.186	0.164	-	-	-	0.194
Topic-3	0.018	0.112	0.192	-	-	-	0.028
|	-	-	-	-	-	-	-
Topic-K	0.128	0.144	0.084	-	-	-	0.036

\+

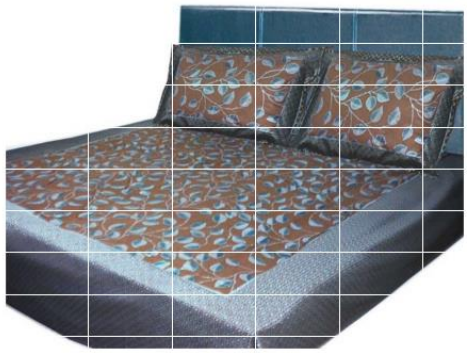
LDA loops

```
Topic 0:
skin laptop towel shape set print polyester inch cotton aroma
Topic 1:
lipstick 38 combo arabian night rythmx nail set polish avenue
Topic 2:
kadhai dlink netgear kosher showpiece wireless 30 guarantee extender wifi
Topic 3:
sticker usb vinyl led wallmantra medium light uberlyfe large cell
Topic 4:
showpiece handicraft best jewellery brass edimax guarantee 30 gift online
Topic 5:
adapter vaio smartpro warranty charger power 195v39a 75 series laptop
Topic 6:
watch analog men baby woman girl boy flipkartcom discount india
Topic 7:
intex tablet keyboard true 3d zyxel usb data inflatable card
Topic 8:
mug combo ceramic set flipkartcom prithish rockmantra perfect gift coffee
```

Topic 0 : beauty
Topic 1 : maquillage
Topic 2 : wireless / network
Topic 3 : déco ou elec ?
Topic 4 : gift / hand craft
Topic 5 : Computers
Topic 6 : Watch
Topic 8 : Computers accessories
Topic 9 : mugs

Computer vision (non supervisé)

Pré traitement image



Echelle de gris

Egalisation de l'histogramme

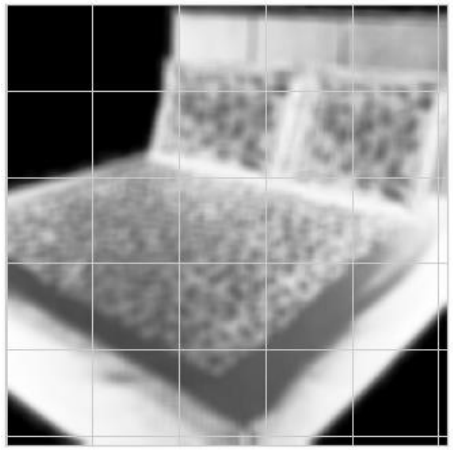
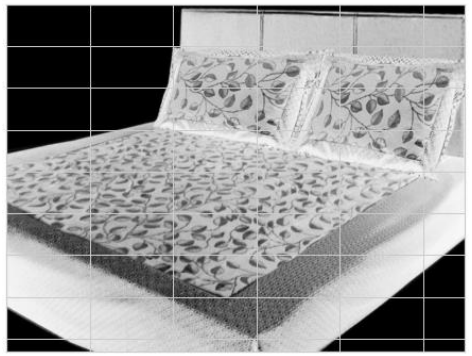
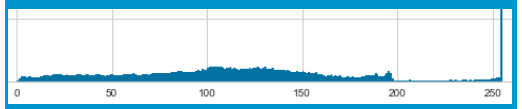
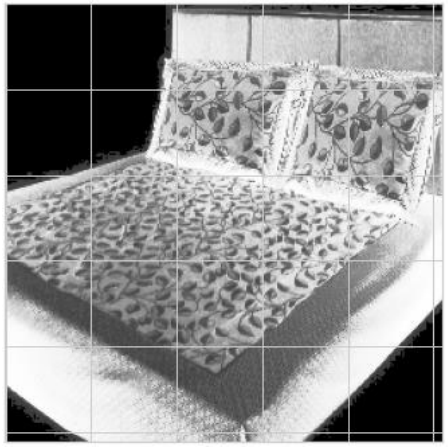


Image témoin





Redimensions



Filtre Gaussien

Sift (Scale-invariant feature transform)

Point d'intérêt → Descripteur → Matching

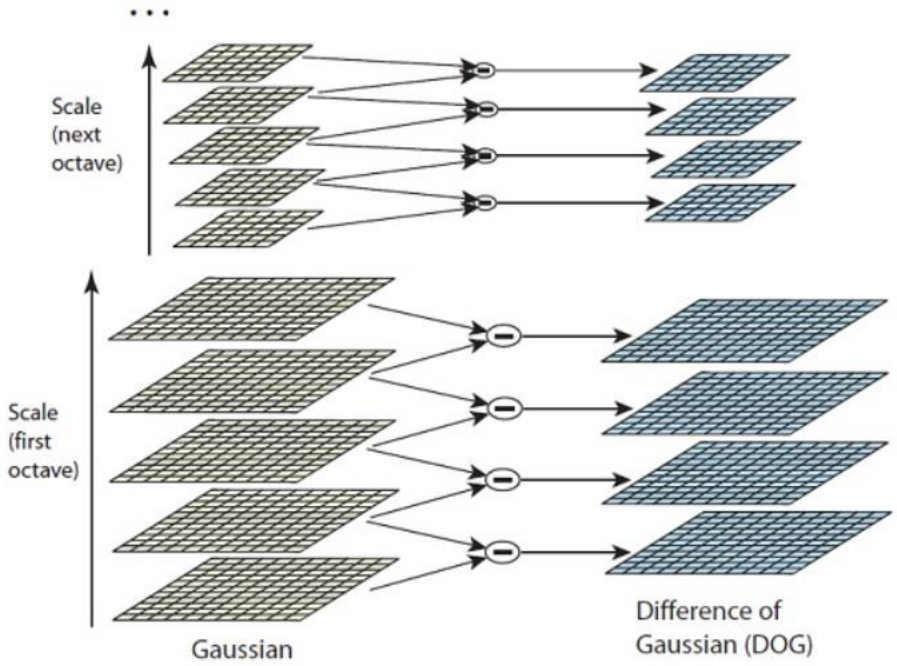
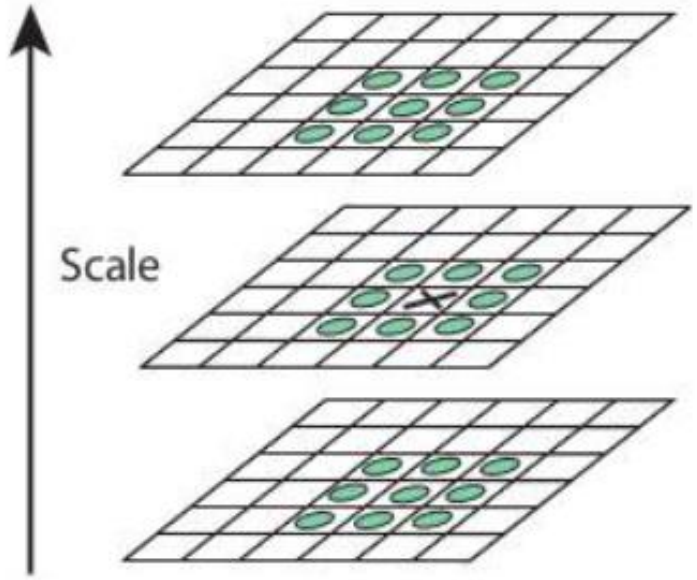


Image source : Openclassrooms



Descripteur

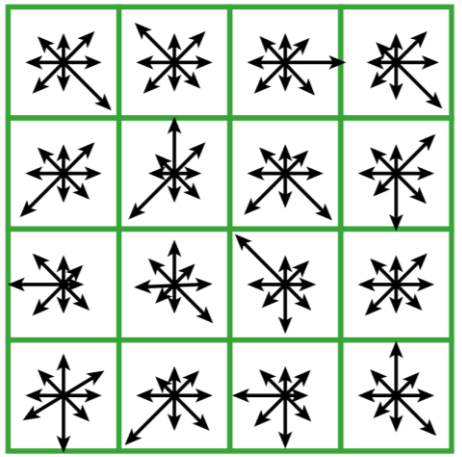
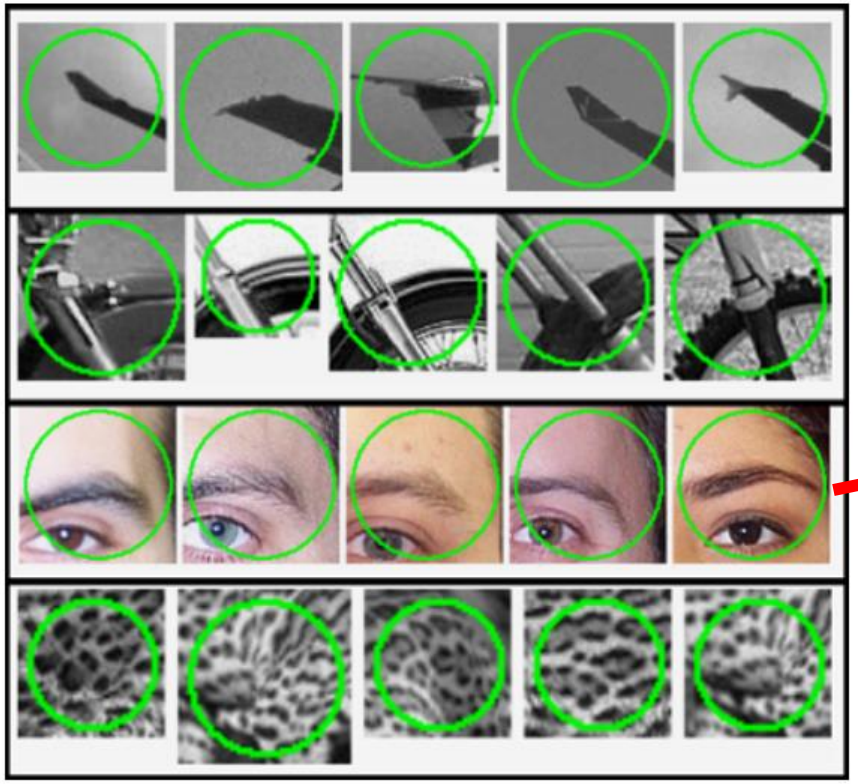


Image source : Wikipédia

Bag of visual words

features → K-means → Visual words → Histogramme

Nombre de descripteurs :
(501841, 128)



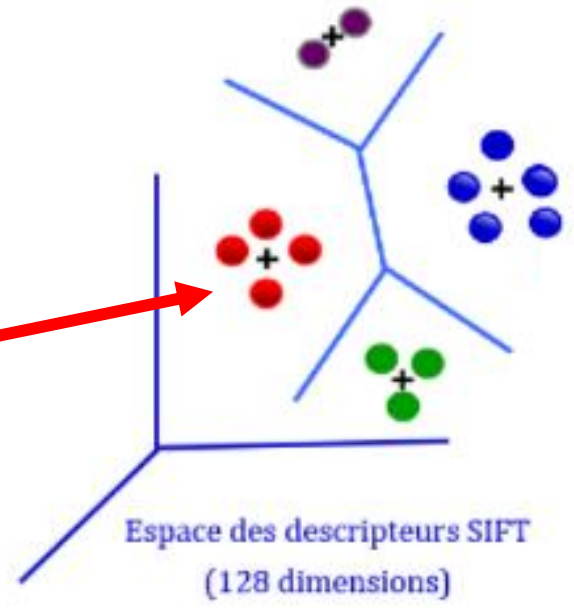
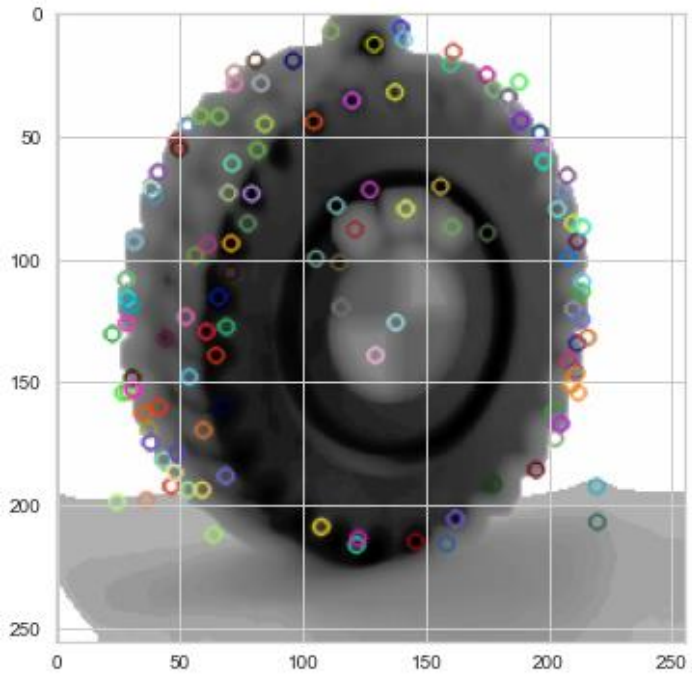


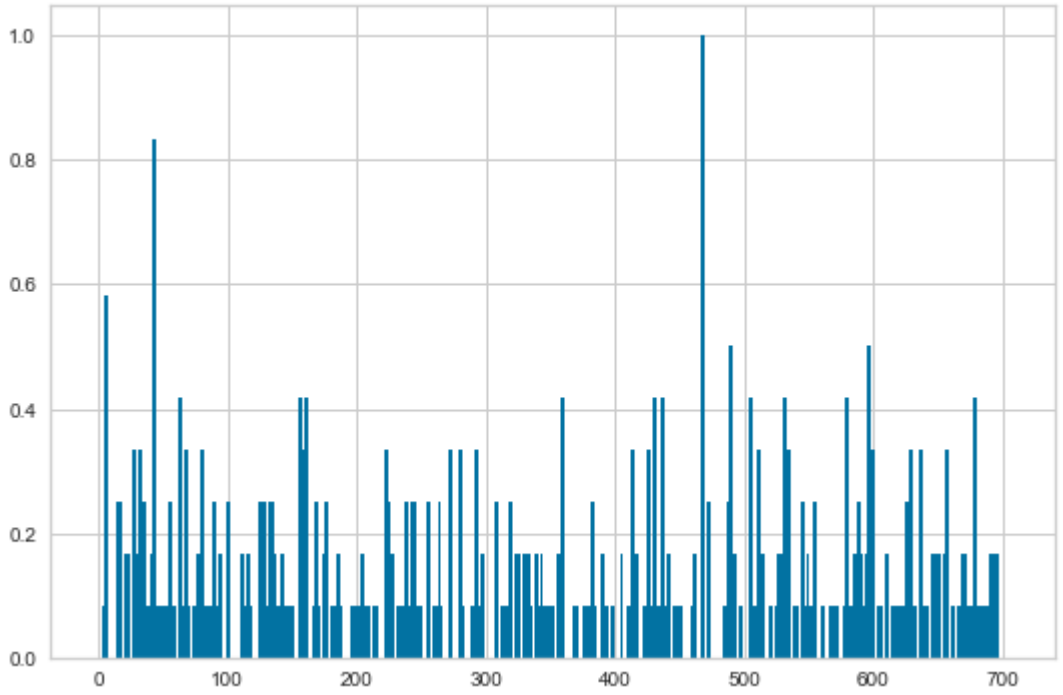
Image source : Openclassrooms

Bag of visual words

features → K-means → Visual words → Histogramme



Features d'un pneu



Histogramme de l'image

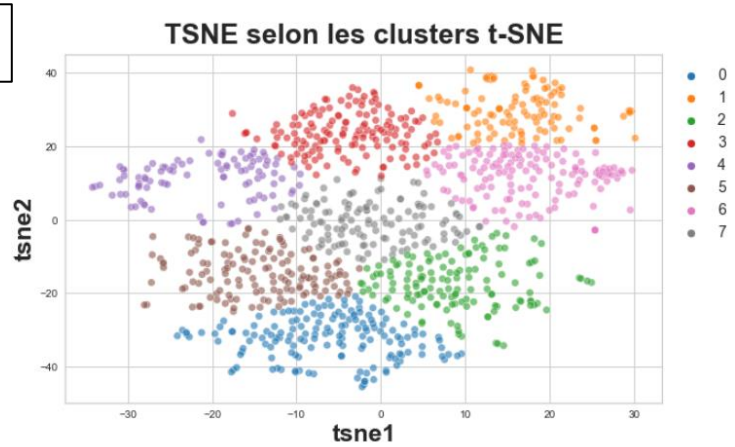
SIFT

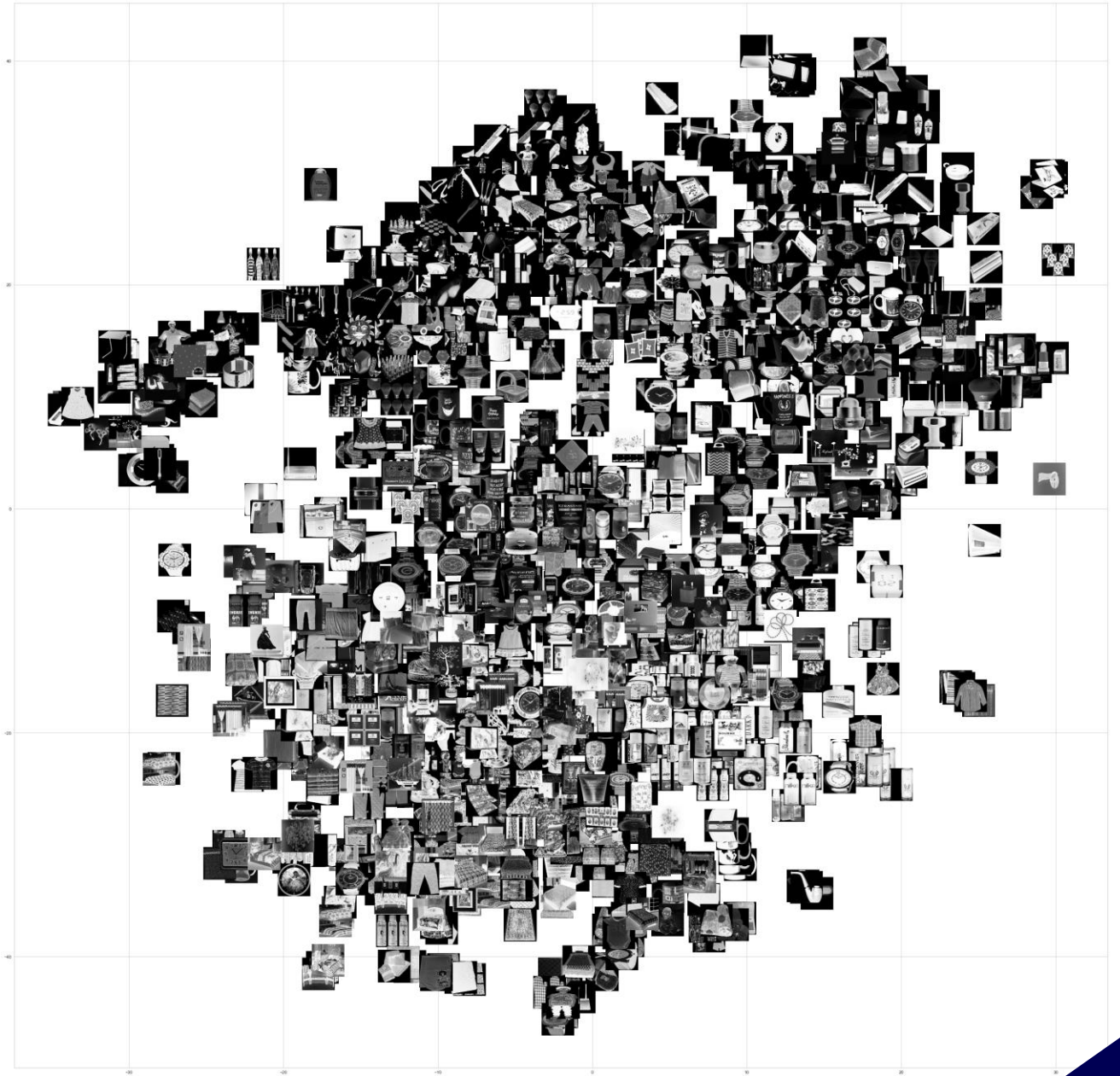
Bag of V_words

PCA

Dimensions avant PCA : (1050, 708)
Dimensions après PCA : (1050, 493)

t-SNE





« Semi » supervisé

product_category_tree
["Home Decor & Festive Needs >> Decorative Lighting & Lamps >> Floor Lamp >> Nutcase Floor Lamp >> Nutcase Multicolor Column Floor Lamp (31 cm)"]

7 catégories

Furnishing	150
Baby	150
Watches	150
Decor	150
Kitchen	150
Beauty	150
Computers	150

63 catégories

Watches Wrist Watches	149
Computers Laptop Accessories	87
Baby Care Infant Wear	84
Kitchen & Dining Coffee Mugs	74
Home Decor & Festive Needs Showpieces	71
...	
Home Decor & Festive Needs Candles & Fragrances	1
Kitchen & Dining Consumables & Disposables	1
Home Furnishing JMD Home Furnishing	1
Home Decor & Festive Needs TRUE Home Decor & Festive Needs]	1
Beauty and Personal Care Beauty Accessories	1

Pré traitement du texte

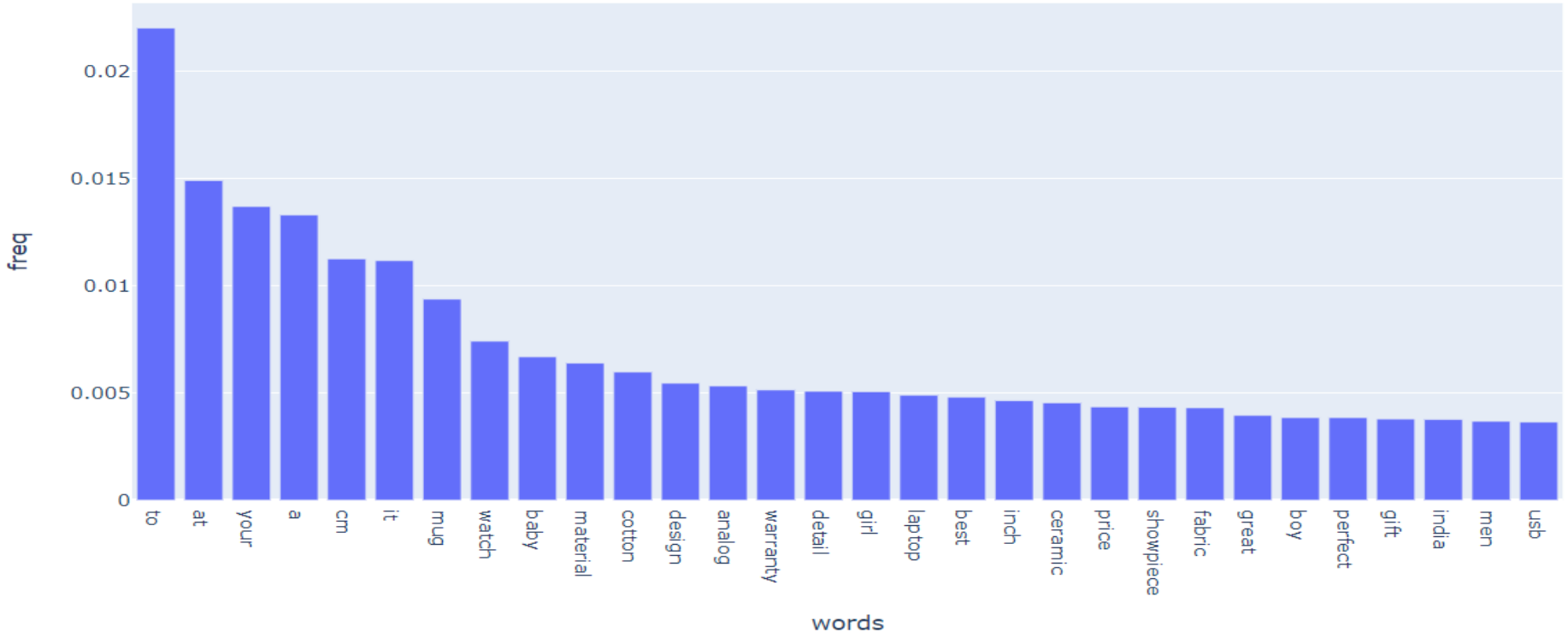
Phrase témoin

Retrait ponctuation + minuscule

Tokenization

Lemmatization

Retrait des stop words



99gems, smart, otg, connection, kit, usb, usb, cable, black, price, r, 199, feature, work, on, microusb, smart, phone, device, ontogo, function, otg, allows, you, to, connect, usb, device, such, a, keyboard, mouse, and, usb, flash, drive, to, your, otg, compatible, phone, or, tablet, small, and, light, easy, to, carry, plug, and, play, compatible

99gems, smart, otg, connection, kit, usb, usb, cable, 199, work, microusb, smart, phone, device, ontogo, function, otg, allows, to, connect, usb, device, such, keyboard, mouse, usb, flash, drive, to, your, otg, compatible, phone, tablet, small, light, to, carry, plug, play, compatible

Stop words et catégories

7 catégories

Furnishing	150
Baby	150
Watches	150
Decor	150
Kitchen	150
Beauty	150
Computers	150

Sélection selon la représentation dans les catégories

words

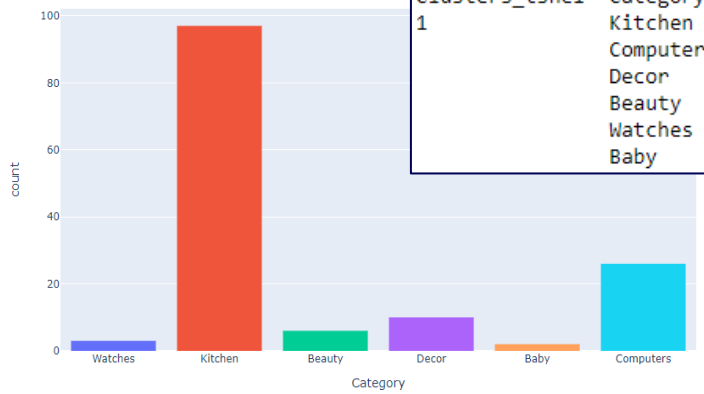
	Words	Furnishing	Baby	Watches	Decor	Kitchen	Beauty	Computers
0	of	17.459	21.694	1.946	18.890	18.374	11.563	10.074
1	for	2.722	15.632	21.284	15.003	14.236	13.050	18.074
2	the	10.809	15.368	1.691	22.500	23.015	9.485	17.132
3	and	8.258	11.862	2.477	19.294	27.853	13.288	16.967
4	to	8.436	9.763	2.180	14.597	38.009	6.351	20.664
5	in	15.658	18.348	14.697	16.427	13.641	8.357	12.872
6	only	9.009	6.306	30.631	9.122	6.532	18.806	19.595
7	product	10.801	7.085	15.796	15.099	14.983	18.235	18.002
8	with	7.601	12.589	0.594	25.416	21.378	11.520	20.903
9	on	9.856	8.894	16.827	17.788	16.226	13.942	16.466

---

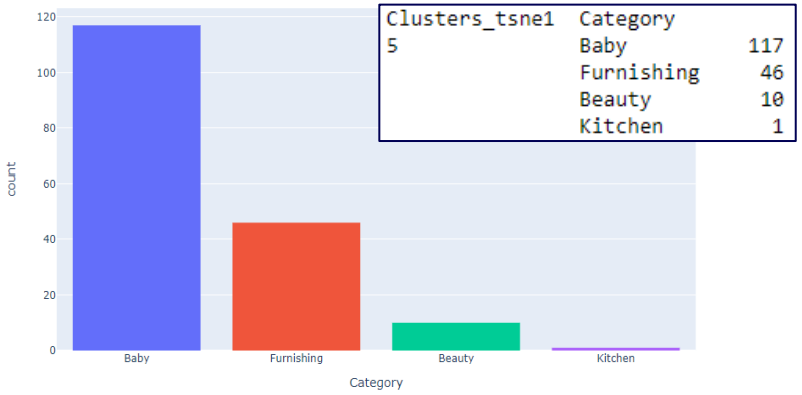
the number of calculated stopwords is : 109

Clustering (K-Means)

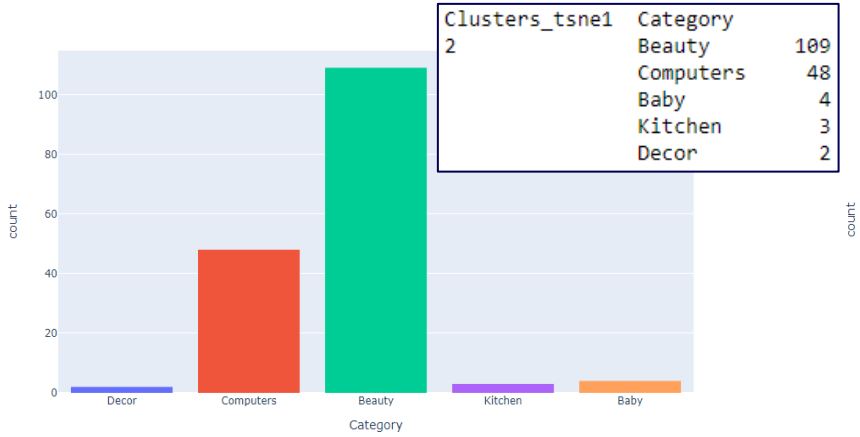
Clusters_tsne1	Category	
1	Kitchen	97
	Computers	26
	Decor	10
	Beauty	6
	Watches	3
	Baby	2



Clusters_tsne1	Category	
5	Baby	117
	Furnishing	46
	Beauty	10
	Kitchen	1

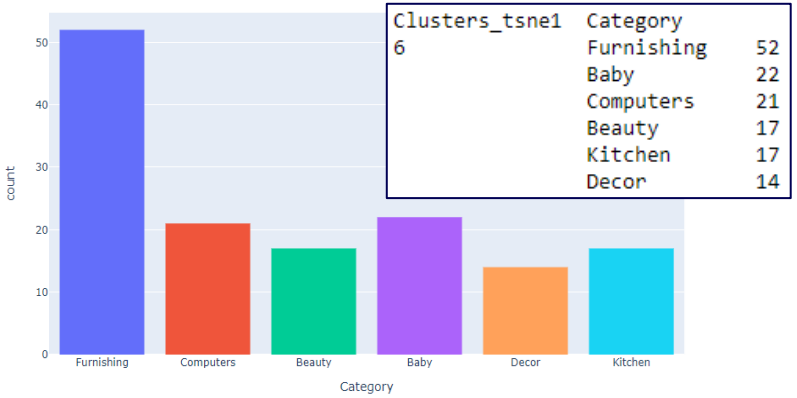


Clusters_tsne1	Category	
2	Beauty	109
	Computers	48
	Baby	4
	Kitchen	3
	Decor	2

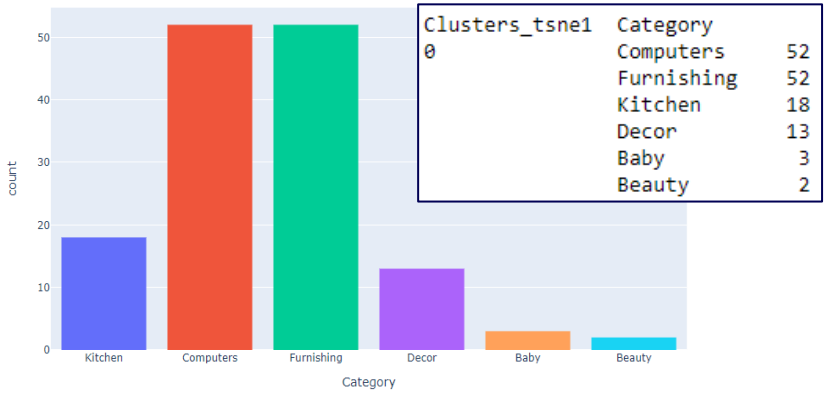


Clusters_tsne1	Category	
4	Watches	147
	Decor	3
	Beauty	1

Clusters_tsne1	Category	
6	Furnishing	52
	Baby	22
	Computers	21
	Beauty	17
	Kitchen	17
	Decor	14



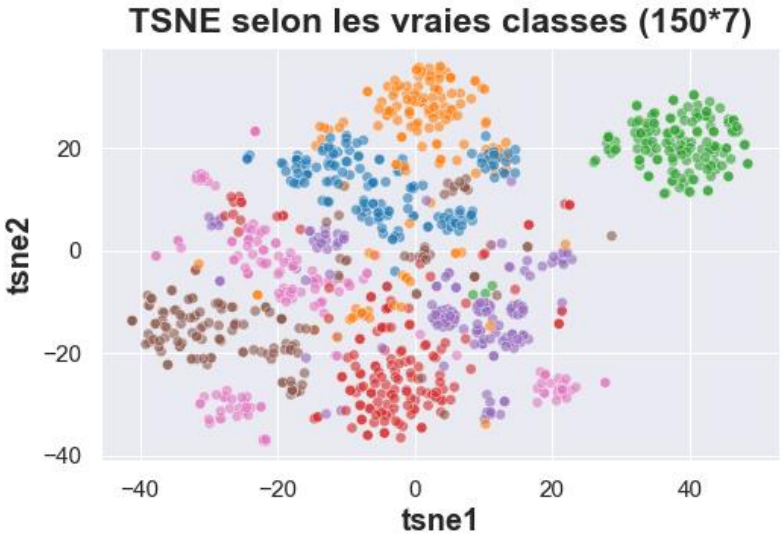
Clusters_tsne1	Category	
0	Computers	52
	Furnishing	52
	Kitchen	18
	Decor	13
	Baby	3
	Beauty	2



Clusters_tsne1	Category	
3	Decor	108
	Kitchen	14
	Beauty	5
	Computers	3
	Baby	2

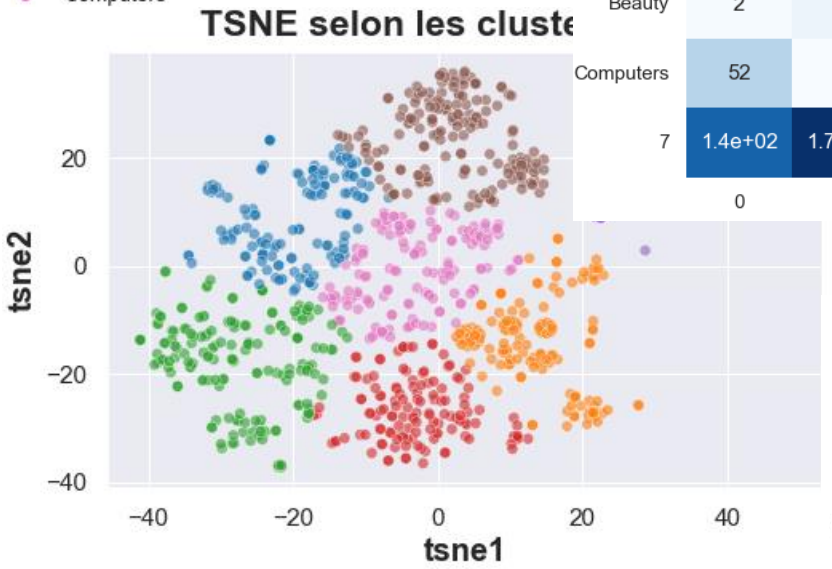
22

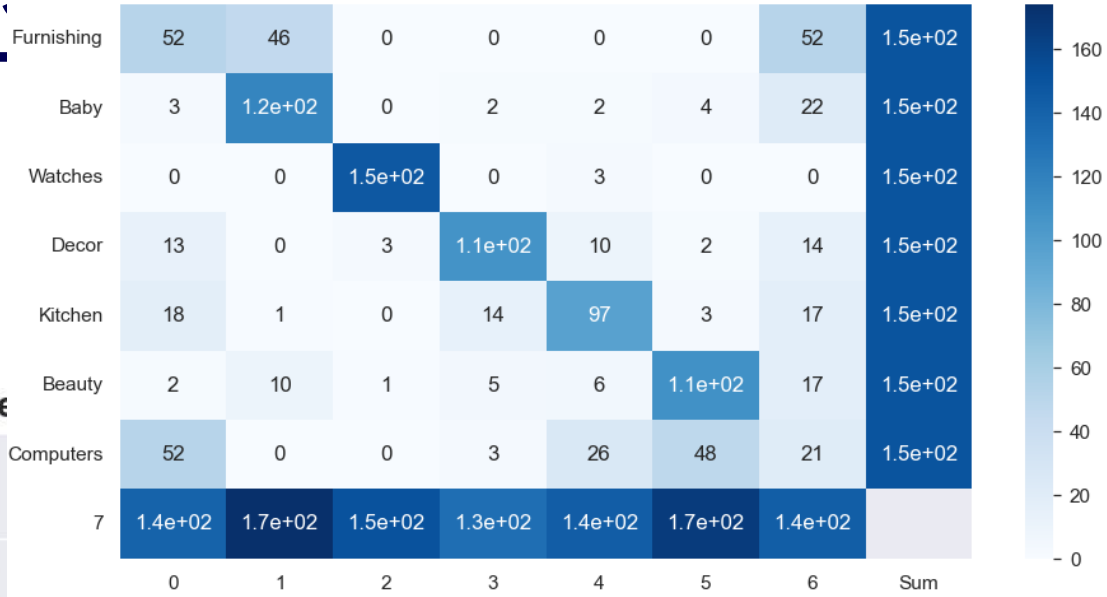
Clustering (K-Means)

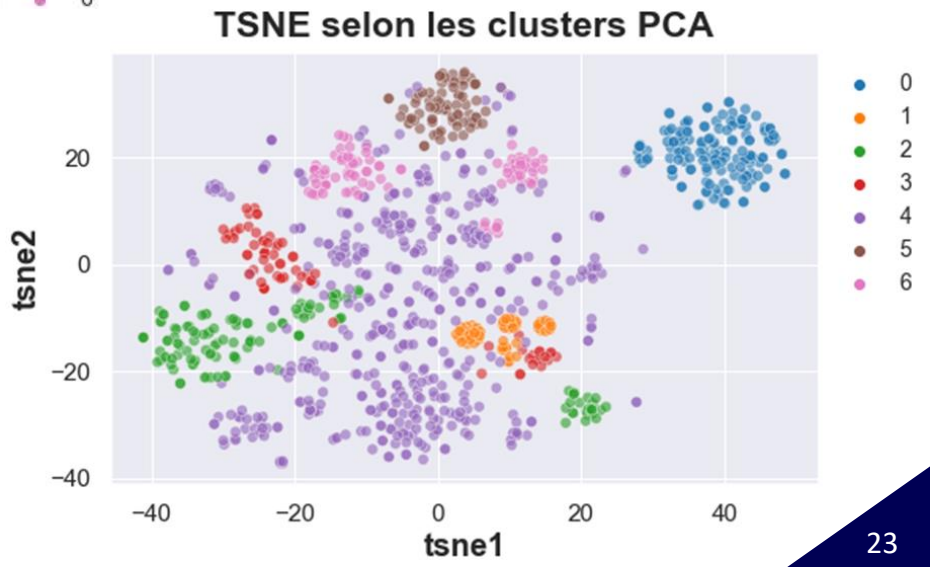


ARI = 0.53

Unigrams

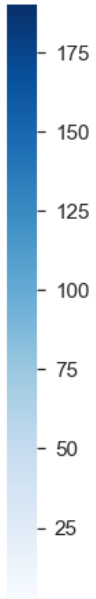




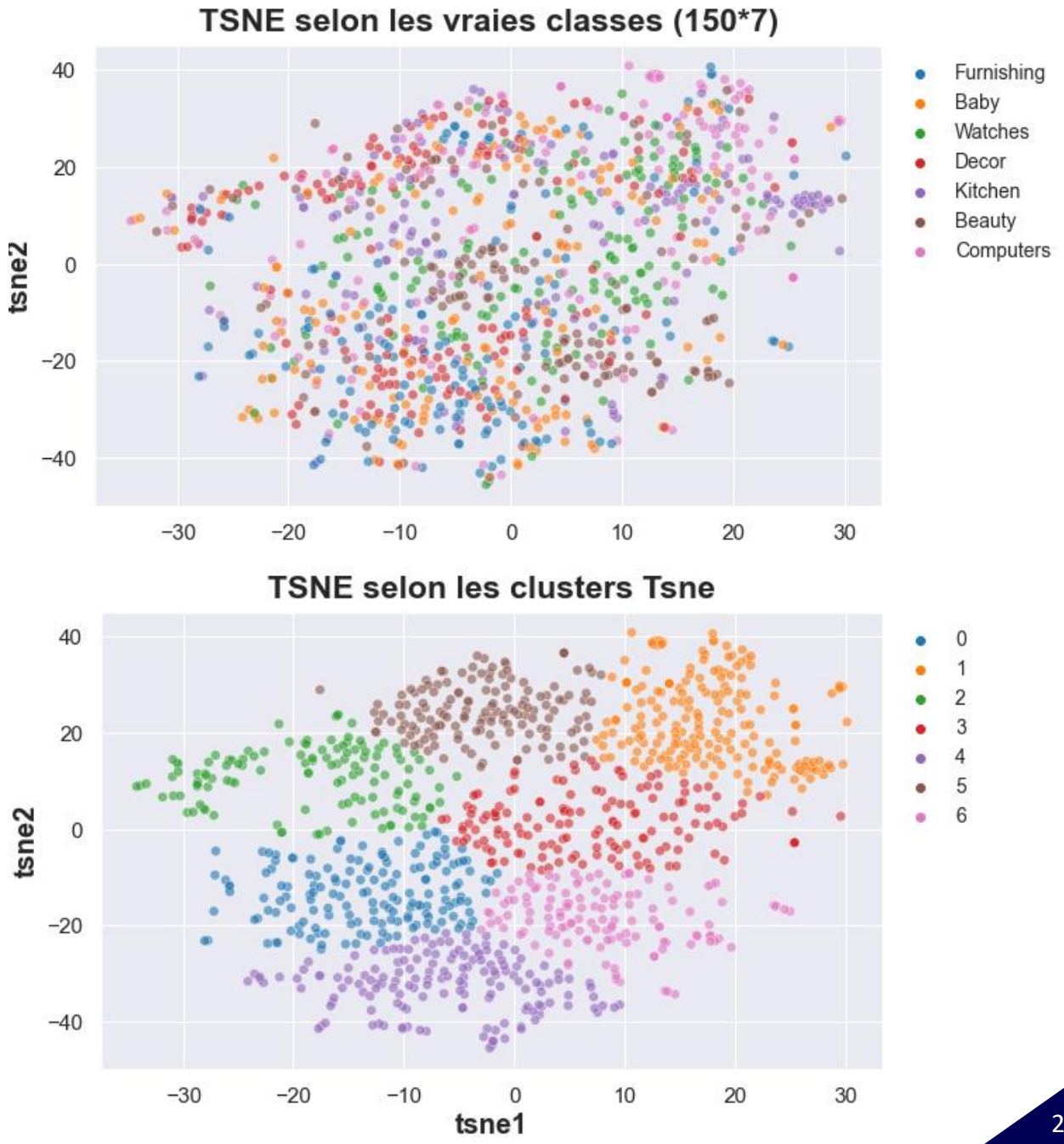


SIFT

	0	1	2	3	4	5	6	Sum
Furnishing	58	32	13	14	10	17	6	1.5e+02
Baby	43	24	14	23	16	13	17	1.5e+02
Watches	7	26	49	11	29	13	15	1.5e+02
Decor	29	24	3	35	13	18	28	1.5e+02
Kitchen	6	18	13	25	48	10	30	1.5e+02
Beauty	8	21	30	15	14	50	12	1.5e+02
Computers	7	15	24	24	60	5	15	1.5e+02
7	1.6e+02	1.6e+02	1.5e+02	1.5e+02	1.9e+02	1.3e+02	1.2e+02	

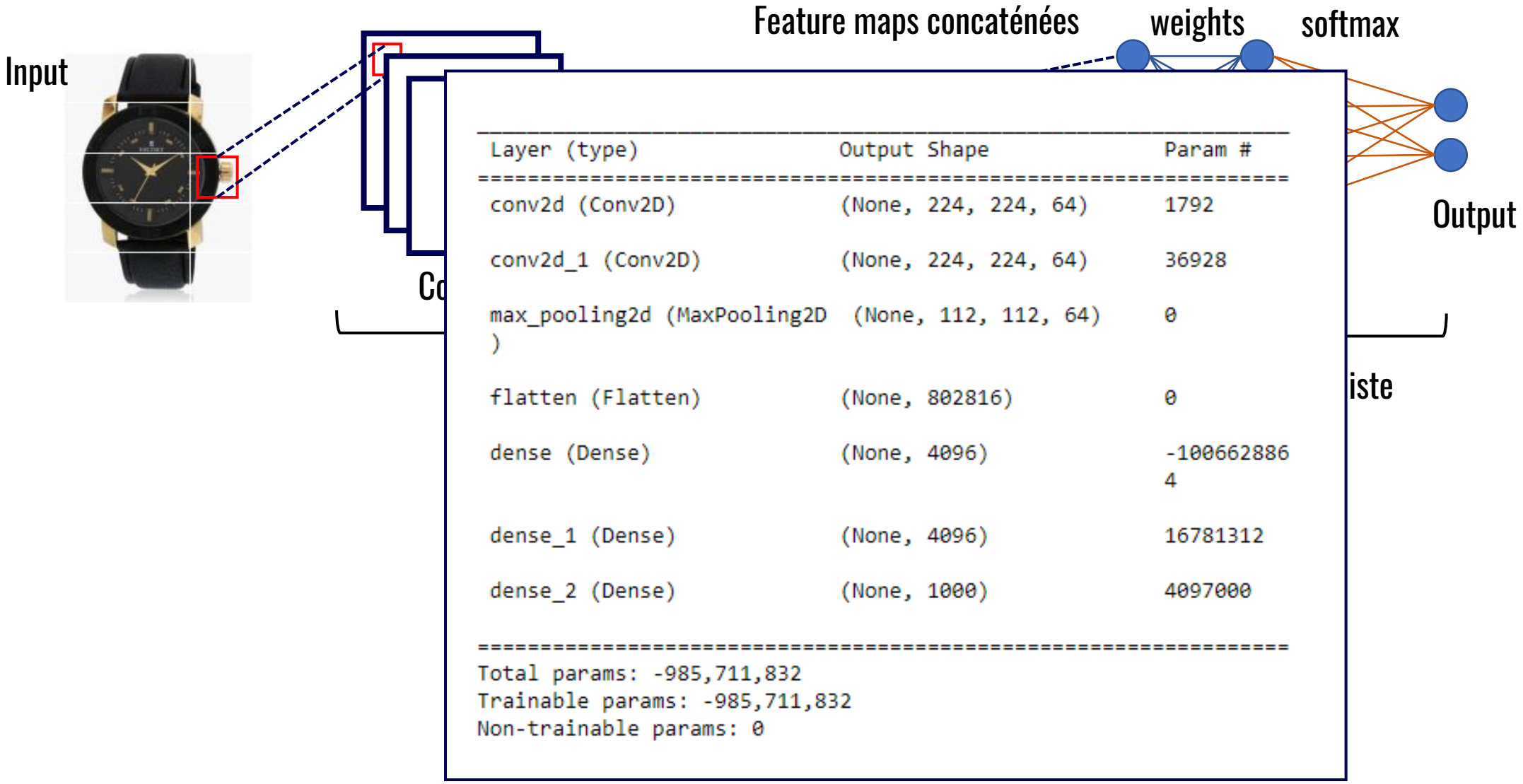


ARI_SIFT = 0.059
ARI_ORB = 0.037



Réseau de Neurones

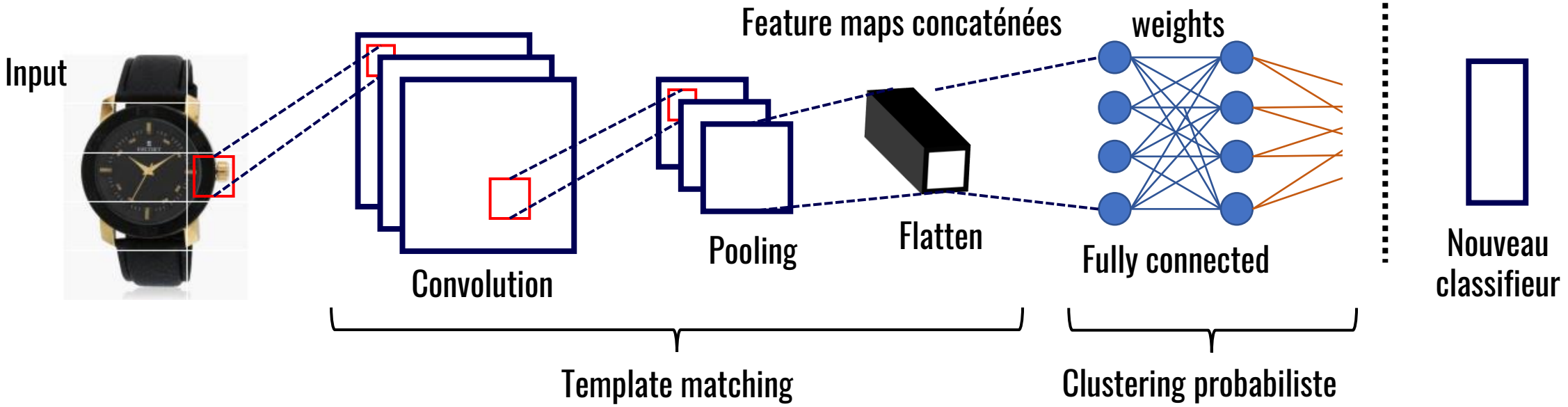
Convolutional Neural Network



Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 224, 224, 64)	1792
conv2d_1 (Conv2D)	(None, 224, 224, 64)	36928
max_pooling2d (MaxPooling2D)	(None, 112, 112, 64)	0
flatten (Flatten)	(None, 802816)	0
dense (Dense)	(None, 4096)	-1006628864
dense_1 (Dense)	(None, 4096)	16781312
dense_2 (Dense)	(None, 1000)	4097000

Total params: -985,711,832
Trainable params: -985,711,832
Non-trainable params: 0

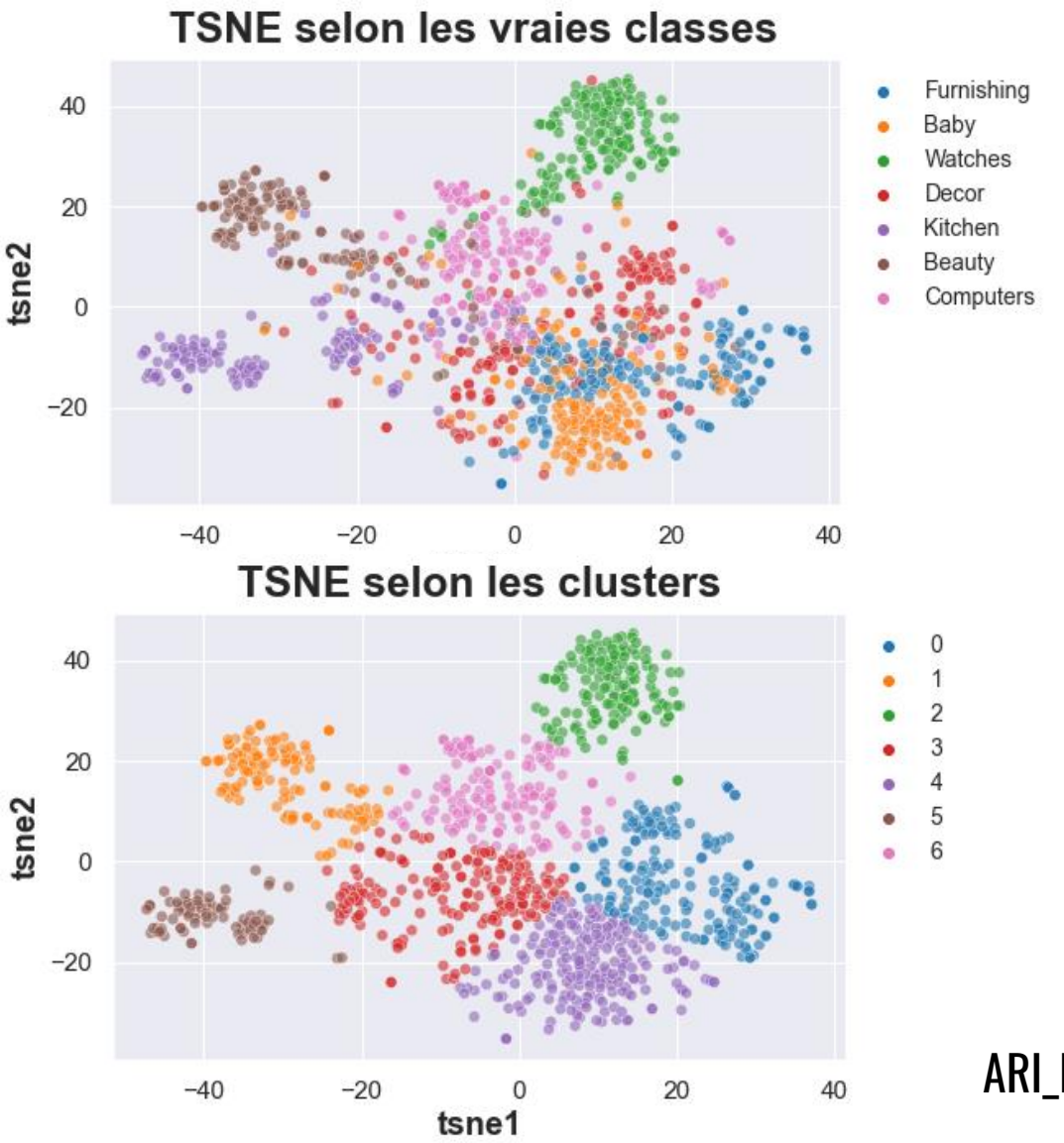
Convolutional Neural Network



Transfert Learning

Stratégie : Extraction des features
On utilise les features du réseau pré entrainé pour représenter les images de notre problème.

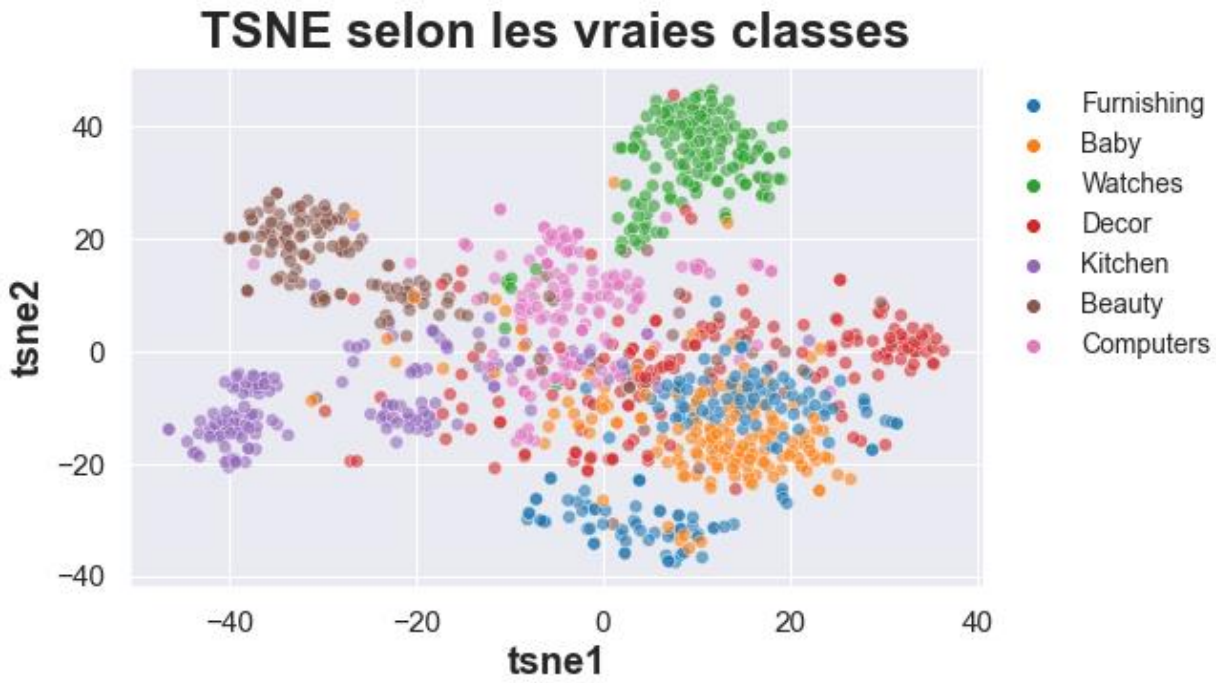
CNN



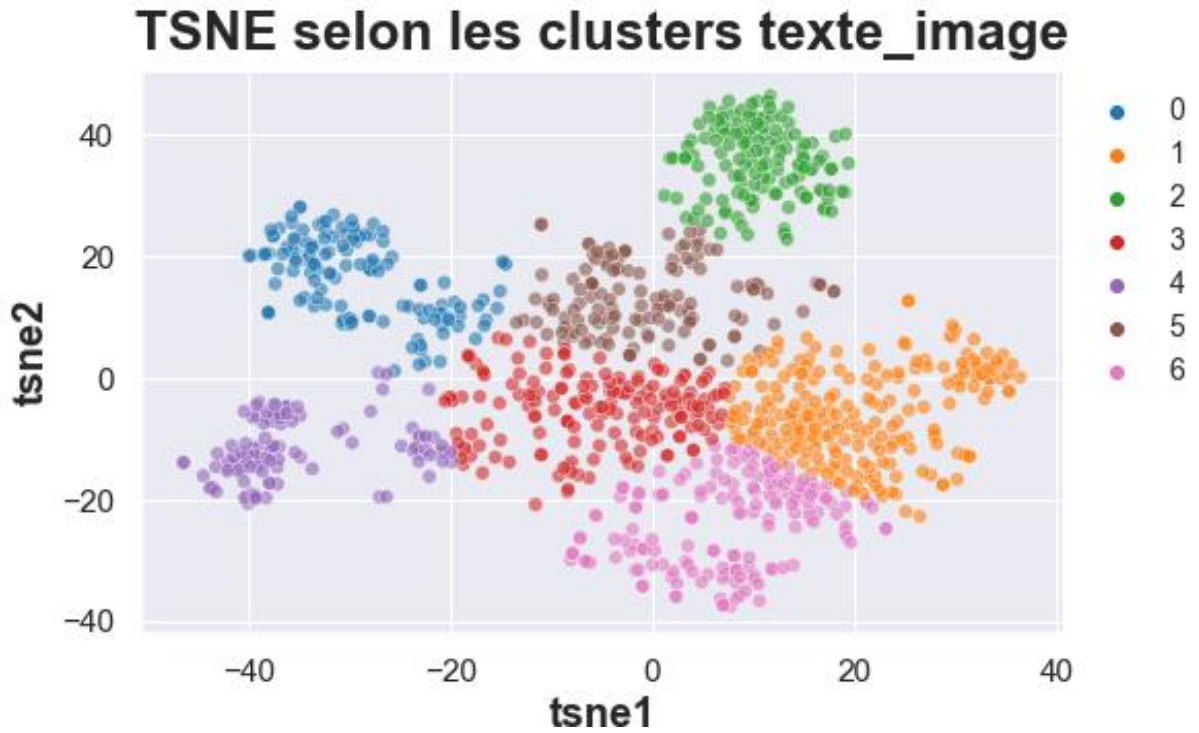
ARI_ResNet50 = 0,36

ARI_VGG16 = 0,42

K-means mixte



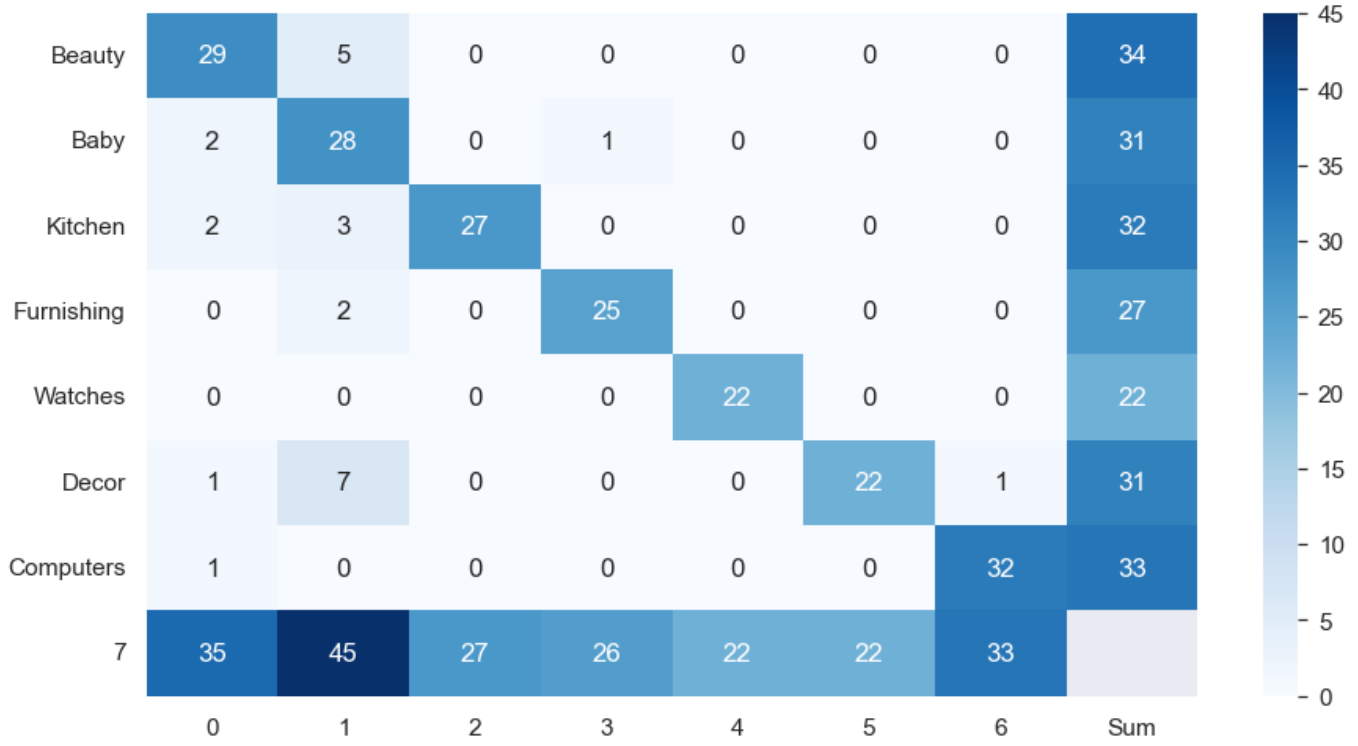
ARI_mixte = 0,48



NLP supervisé

SVM

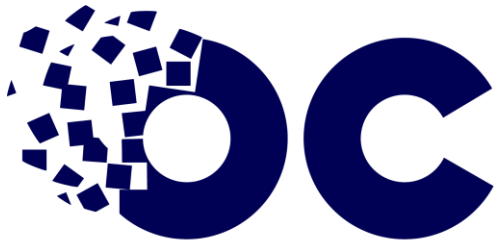
F1 Score : 0.8908511325790812
Accuracy : 0.8809523809523809
ARI : 0.7259871233207815



Conclusion

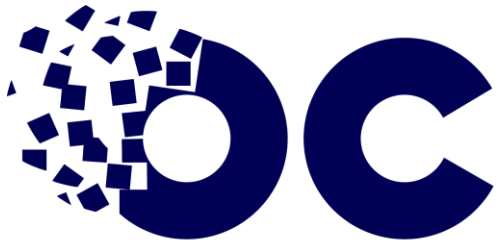
- V0 = Etudier la faisabilité : faite
- Non supervisé : 9 groupes
- « Semi » supervisé : NLP : ARI = 0,53
  CV → SIFT : 0,059 / CNN : 0,42
- Option supervisé : ARI NLP = 0,72

- Pistes d'améliorations…



Fin

Merci de votre attention !



Convolutional Neural Network)

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 224, 224, 64)	1792
conv2d_1 (Conv2D)	(None, 224, 224, 64)	36928
max_pooling2d (MaxPooling2D)	(None, 112, 112, 64)	0
flatten (Flatten)	(None, 802816)	0
dense (Dense)	(None, 4096)	-1006628864
dense_1 (Dense)	(None, 4096)	16781312
dense_2 (Dense)	(None, 1000)	4097000

Total params: -985,711,832
Trainable params: -985,711,832
Non-trainable params: 0

Stop words et catégories

7 catégories

Furnishing	150
Baby	150
Watches	150
Decor	150
Kitchen	150
Beauty	150
Computers	150

words

	Words	Furnishing	Baby	Watches	Decor	Kitchen	Beauty	Computers
0	of	17.459	21.694	1.946	18.890	18.374	11.563	10.074
1	for	2.722	15.632	21.284	15.003	14.236	13.050	18.074
2	the	10.809	15.368	1.691	22.500	23.015	9.485	17.132
3	and	8.258	11.862	2.477	19.294	27.853	13.288	16.967
4	to	8.436	9.763	2.180	14.597	38.009	6.351	20.664
5	in	15.658	18.348	14.697	16.427	13.641	8.357	12.872
6	only	9.009	6.306	30.631	9.122	6.532	18.806	19.595
7	product	10.801	7.085	15.796	15.099	14.983	18.235	18.002
8	with	7.601	12.589	0.594	25.416	21.378	11.520	20.903
9	on	9.856	8.894	16.827	17.788	16.226	13.942	16.466

---

the number of calculated stopwords is : 109

NMF non sup / sup

```
Topic 0:
watch analog men woman discount india great sonata maximum flipkartcom
Topic 1:
combo set flipkartcom guarantee 30 online denver deodorant playboy adidas
Topic 2:
mug ceramic coffee prithish printland perfect gift bring ml happy
Topic 3:
baby girl dress boy fabric cotton sleeve neck printed ideal
Topic 4:
showpiece best guarantee 30 online handicraft brass buddha kadhai exotic
Topic 5:
laptop battery cell hp lapguard pavilion skin shape warranty rega
Topic 6:
towel bath cotton set flipkartcom soft face hand terry nkp
Topic 7:
abstract single blanket quilt comforter double multicolor raymond floral cushion
Topic 8:
rockmantra mug ceramic crafting toodishwasher thrilling porcelain permanent ensuring stay
```

```
Topic 0:
watch analog men discount india great woman dial strap sonata
Topic 1:
mug ceramic perfect rockmantra coffee gift prithish loved design safe
Topic 2:
baby girl fabric dress cotton boy sleeve neck ideal pattern
Topic 3:
showpiece cm best brass buddha handicraft statue gift ganesha exotic
Topic 4:
combo denver 350 399 paris deodorant ice wild stone gift
Topic 5:
towel bath cotton hand face terry 100 yellow linen 299
Topic 6:
laptop battery cell hp skin pavilion lapguard 00 warranty shape
```

Topic 0 : Watches
Topic 1 : Kitchen
Topic 2 : Baby
Topic 3 : Decor
Topic 4 : Furnishing (by default)
Topic 5 : Beauty
Topic 6 : Computers

Transfert Learning

Entraîner un réseau de neurones convolutif → très coûteux (dépend du nombre de couche)

La solution → Le Transfer Learning (ou apprentissage par transfert)

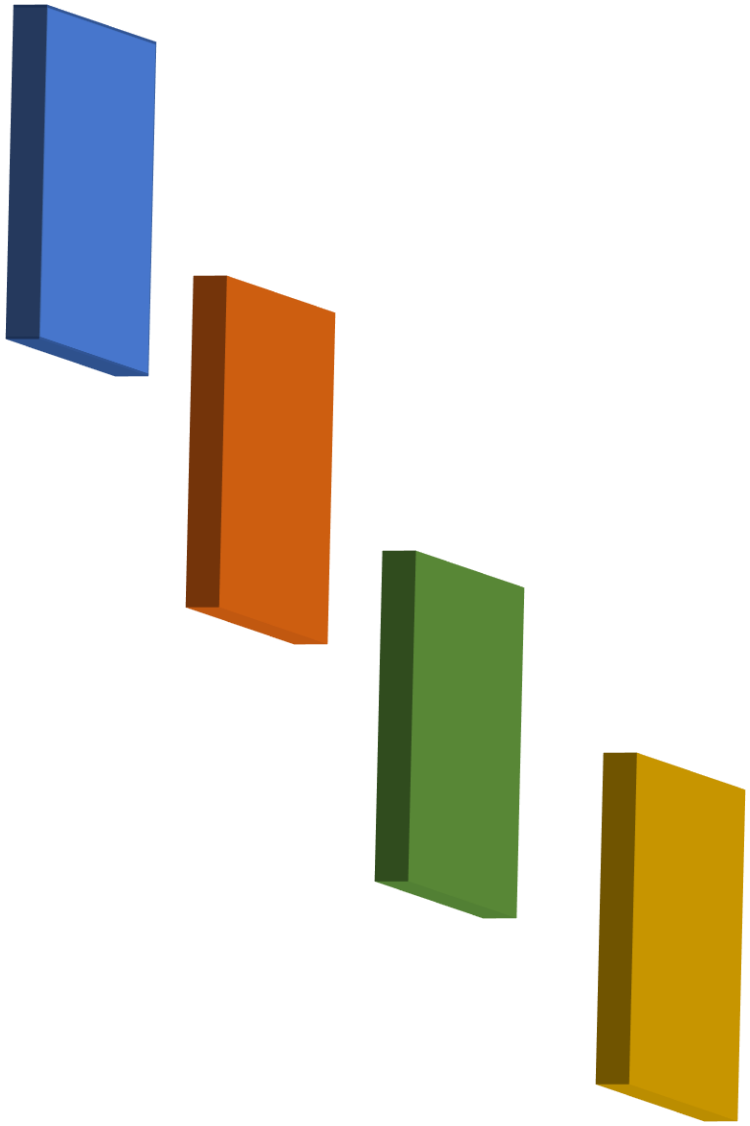
Avantages : Entrainement accéléré / Eviter le sur-apprentissage

Stratégie #1 : fine-tuning total
On remplace la dernière couche fully-connected par un classifieur adapté au nouveau problème (SVM, régression logistique...) et initialisé de manière aléatoire. Toutes les couches sont ensuite entraînées sur les nouvelles images.
Conditions : Base de donnée d'image grande (pas d'overfitting)

Stratégie #2 : extraction des features
On utilise les features du réseau pour représenter les images de notre problème.
Retrait de la dernière couche fully-connected, les autres paramètres restent fixé.
Conditions : Base de donnée d'image petite et similaire aux images de pré-entraînement.

Stratégie #3 : fine-tuning partiel (Mélange de #1 et #2)
On remplace la dernière couche par un classifieur initialisé aléatoirement, et on fixe les paramètres de certaines couches du réseau.
Conditions : Base de donnée d'image petite et différentes des images de pré-entraînement.

CNN



Couche de convolution (première couche ++)

Filtrage par convolution
Pour chaque paire (image, filtre) → carte d'activation (feature map)
Plus la valeur est élevée, plus la correspondance spot image / feature est grande

Couche de pooling (souvent placé entre 2 couches de convolutions)
Entrée → feature maps → opération de pooling → feature maps
Intérêt : Réduire le nombre de paramètres et de calculs dans le réseau
Améliore l'efficacité du réseau
Evite le sur apprentissage

Couche de correction ReLu (Rectified Linear Units)
[-∞ , 0 [→ 0

Couche fully connected (dernière couche)
Classification (logistique ou softmax)
Entrée → vecteur → application poids linéaire + fonction d'acitvation → vecteur de taille N (avec N = nbr de classes)

Stop words et catégories

7 catégories

```
Furnishing     150
Baby           150
Watches        150
Decor          150
Kitchen        150
Beauty         150
Computers      150
```

Sélection selon la représentation dans les catégories

Words	Furnishing	Baby	Watches	Decor	Kitchen	Beauty	Computers
flipkartcom	74	32	134	0	5	159	68
mug	0	4	0	0	445	0	0
watch	0	2	349	0	0	0	4
baby	2	316	0	1	0	1	0
cotton	99	182	0	5	0	0	0
analog	0	0	249	6	0	0	0
warranty	9	0	17	27	12	23	158
detail	0	180	0	10	45	8	0
girl	0	197	37	0	2	6	0
laptop	0	0	0	0	0	0	234
ceramic	0	2	1	10	204	0	0
showpiece	0	0	0	204	0	3	0
fabric	33	162	0	0	5	4	2
men	1	2	158	0	0	15	0
usb	0	0	0	0	0	0	174