Soutenance Projet 6

Pierre Schifflers

Agenda

- Problématique et Objectifs
- Nettoyage des données et analyse exploratoire
- Traitement du texte + étude de faisabilité classification
- Traitement des images + étude de faisabilité classification
- Conclusions et pistes d'amélioration

Environnement

- Python 3.7.6
- JupyterLab 1.2.6
- Scikit Learn 0.22.1
- Keras 2.3.1
- Numpy 1.18.1
- Seaborn 0.10.0
- Matplotlib 3.1.3
- NLTK 3.4.5
- Pillow 7.0.0

Problématique

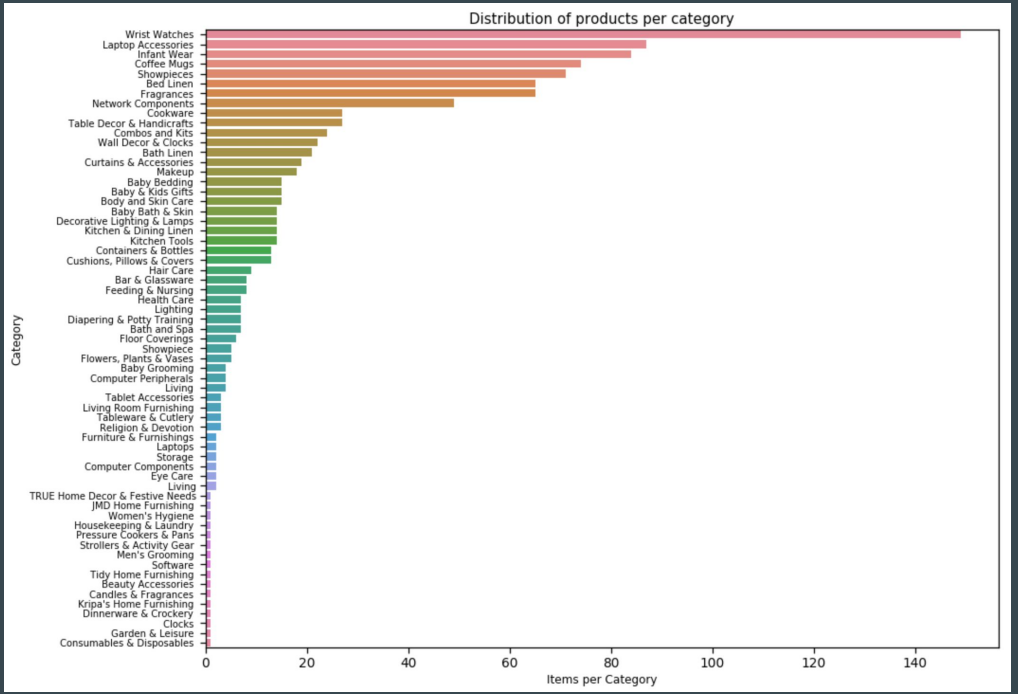
- L'entreprise 'Place de Marché' lance une marketplace e-commerce
- Des vendeurs proposent des articles
- Attribution de catégorie faite manuellement par les vendeurs
- Pour le passage à l'échelle et facilitation de mise en ligne de produits → **Automatisation**

Objectifs

- Réaliser une première étude de faisabilité d'un moteur de classification
- Analyser le jeu de données
- Réaliser un pré-traitement des images et descriptions de produits
- Réduction de dimension
- Clustering

Nettoyage et Exploration

- 1050 produits
- Expansion de la colonne “Product Category Tree”
- Sélection de la 2ème catégorie comme descripteur principal
- Distribution inégale des produits
- One-Hot Encoding des labels de catégories



Traitement du Texte

```
Maserati Time R8851116001 Analog Watch  - For Boys - B
uy Maserati Time R8851116001 Analog Watch  - For Boys
R8851116001 Online at Rs.24400 in India Only at Flipka
rt.com. - Great Discounts, Only Genuine Products, 30 D
ay Replacement Guarantee, Free Shipping. Cash On Deliv
ery!
```

1. Description originale

```
maserati time r8851116001 analog watch  - for boys - b
uy maserati time r8851116001 analog watch  - for boys
r8851116001 online at rs.24400 in india only at flipka
rt.com. - great discounts, only genuine products, 30 d
ay replacement guarantee, free shipping. cash on deliv
ery!
```

2. Lower Case

```
maserati time r8851116001 analog watch  - for boys - b
uy maserati time r8851116001 analog watch  - for boys
r8851116001 online at rs. in india only at flipkart.co
m. - great discounts, only genuine products,  day repl
acement guarantee, free shipping. cash on delivery!
```

3. Remove isolated digits

```
['maserati', 'time', 'r8851116001', 'analog', 'watch',
'for', 'boys', 'buy', 'maserati', 'time', 'r885111600
1', 'analog', 'watch', 'for', 'boys', 'r8851116001',
'online', 'at', 'rs', 'in', 'india', 'only', 'at', 'fl
ipkart', 'com', 'great', 'discounts', 'only', 'genuin
e', 'products', 'day', 'replacement', 'guarantee', 'fr
ee', 'shipping', 'cash', 'on', 'delivery']
```

4. Tokenization

```
['maserati', 'time', 'r8851116001', 'analog', 'watch',
'boys', 'buy', 'maserati', 'time', 'r8851116001', 'ana
log', 'watch', 'boys', 'r8851116001', 'online', 'rs',
'india', 'flipkart', 'com', 'great', 'discounts', 'gen
uine', 'products', 'day', 'replacement', 'guarantee',
'free', 'shipping', 'cash', 'delivery']
```

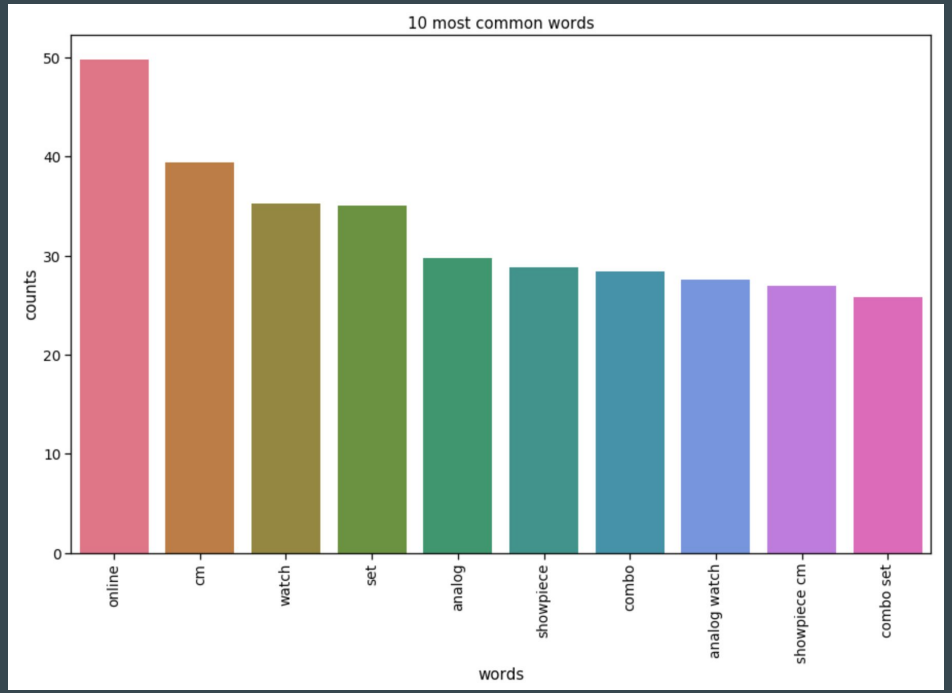
5. Remove stopwords

```
['maserati', 'time', 'r8851116001', 'analog', 'watch',
'boys', 'maserati', 'time', 'r8851116001', 'analog',
'watch', 'boys', 'r8851116001', 'online']
```

6. Remove corpus stopwords and single characters

```
['maserati', 'time', 'r8851116001', 'analog', 'watch',
'boy', 'maserati', 'time', 'r8851116001', 'analog', 'w
atch', 'boy', 'r8851116001', 'online']
```

7. Lemmatization



Traitement du Texte

- TF/IDF avec unigrammes et bigrammes
- 777 features
- Max_df = 0.99
- Min_df = 0.01

1	2	...	weight	weight height	weight kg	well	well know	white	wide	wide selection	width	width cm
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.092693	0.054538

Classification non-supervisée

Latent Dirichlet Allocation

- On aperçoit déjà une cohérence au sein de certains topics (4), centré autour des catégories où l'on a plus de données (Wrist Watches, Fragrances, Bed Linen, Showpieces)
- On aperçoit aussi certains groupes plus hétérogènes (0)
- On aperçoit aussi certains stopwords du corpus tels que 'online'

```
Topics found via LDA with TF/IDF Vectorizer:

Topic #0
usb coffee mug tea coffee mug home usb usb port original stylish

Topic #1
baby girl baby girl detail fabric cotton sticker baby boy boy cushion

Topic #2
showpiece showpiece cm cm cm online online handicraft gift ganesha rockmantra table

Topic #3
kadhai brass kadhai online online range item decorative wireless cm brass rajasthan

Topic #4
watch analog analog watch men watch men discount woman watch woman online discount online

Topic #5
mug ceramic ceramic mug glass adapter power ml feature specification skin

Topic #6
towel router lowest spf wireless bath cotton bath towel cotton bath line

Topic #7
cm inch width black design dimension sale polyester height feature

Topic #8
blanket single double abstract quilt comforter quilt comforter abstract single multicolor led

Topic #9
combo set combo set set online online skin laptop wild set combo shape
```

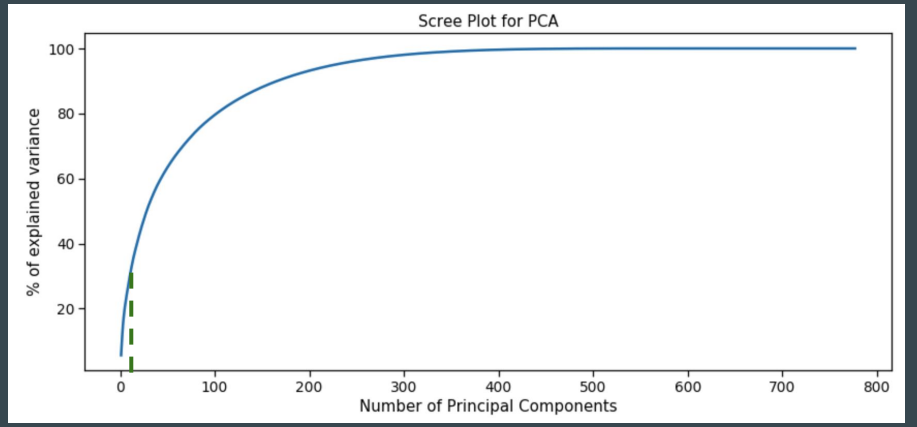
topic_lda	category_lvl_0	category_lvl_1	
0	Baby Care	Baby & Kids Gifts	1
		Baby Grooming	1
		Diapering & Potty Training	1
	Computers	Laptop Accessories	9
		Network Components	18
	Home Furnishing	Bed Linen	3
	Kitchen & Dining	Bar & Glassware	1
		Coffee Mugs	11
		Containers & Bottles	2

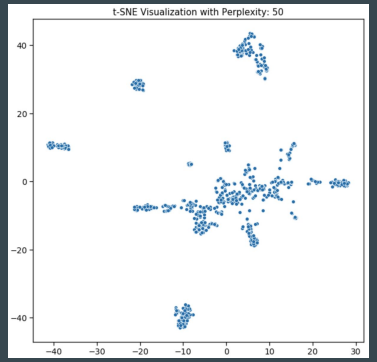
4	Coffee Mugs	10
	Laptop Accessories	7
	Makeup	1
	Wrist Watches	107

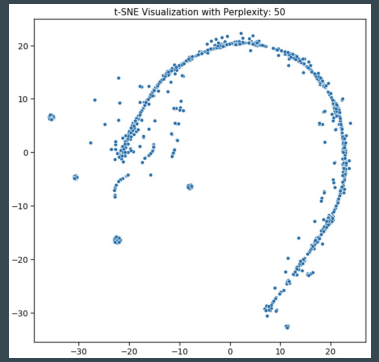
Classification non-supervisée

Réduction de Dimensions

- Avec 10 composantes principales, nous expliquons déjà **30%** de la variance
- Visualisations de cette réduction avec un noyau linéaire et RBF avec t-SNE



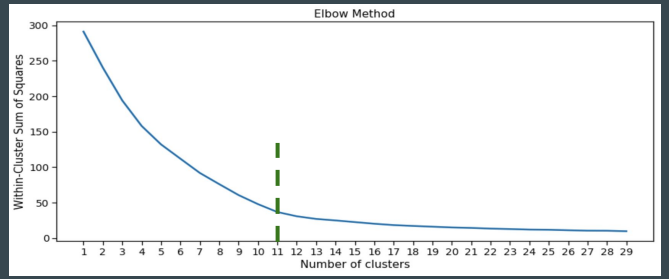


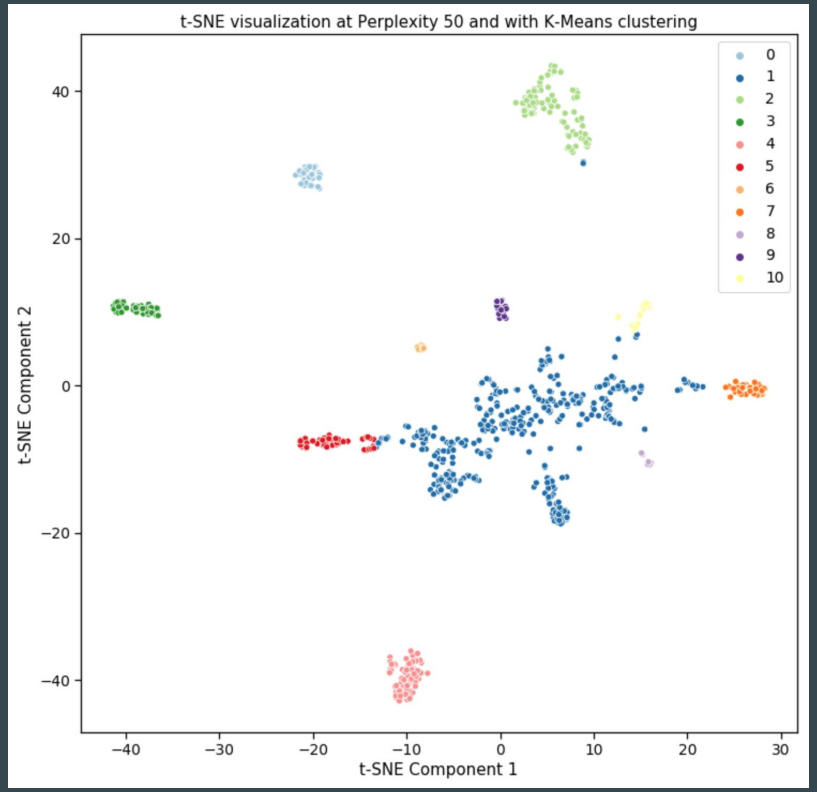


Classification non-supervisée

K-Means

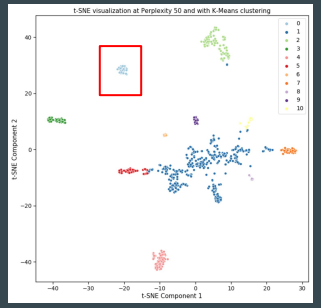
- Coude à 11 clusters
- Silhouette Score maximisé (0.62)
- Davies Bouldin Index minimisé (0.5)
- Présence de clusters éloignés et dense mais présence d'un cluster beaucoup plus large que le reste





Classification non-supervisée

- Les clusters n'ont pas l'air de particulièrement bien représenter les catégories
- Beaucoup de catégories se retrouvent au sein d'un même cluster
- Aucune catégorie ne domine particulièrement les clusters



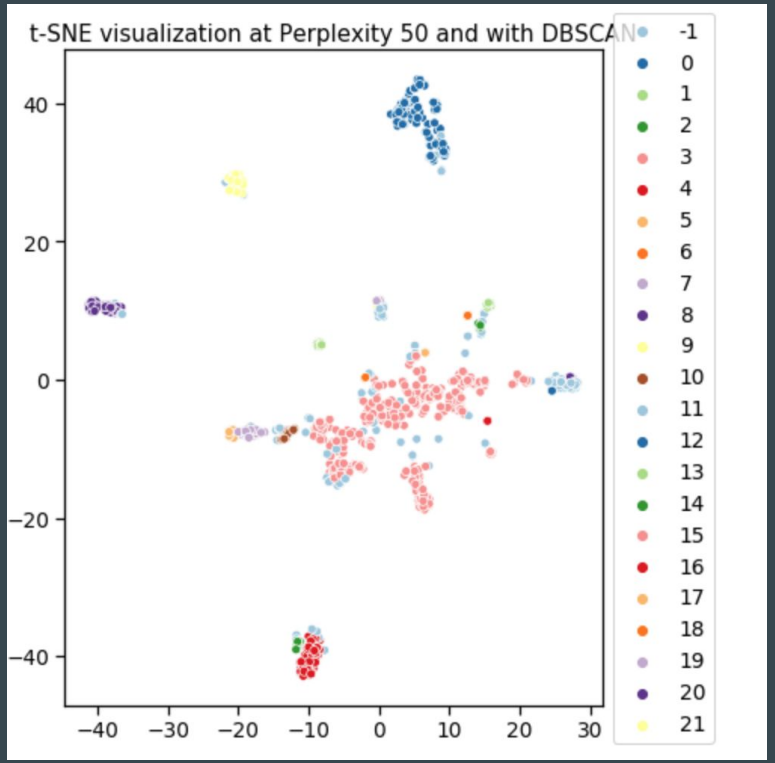
cluster_labels	category_lvl_1
0	21
1	49
2	38
3	26
4	29
5	25
6	12
7	24
8	14
9	20
10	20

cluster_labels	category_lvl_1	category_lvl_0
0	Bath Linen	2
	Bath and Spa	1
	Coffee Mugs	4
	Combos and Kits	1
	Cookware	2
	Curtains & Accessories	2
	Cushions, Pillows & Covers	1
	Diapering & Potty Training	1
	Dinnerware & Crockery	1
	Floor Coverings	1
	Fragrances	4
	Infant Wear	9
	Kitchen & Dining Linen	1
	Kitchen Tools	1
	Laptop Accessories	6
	Network Components	2
	Showpieces	4
	Storage	1
	TRUE Home Decor & Festive Needs	1
	Table Decor & Handicrafts	1
	Wrist Watches	8

Classification non-supervisée

DBSCAN

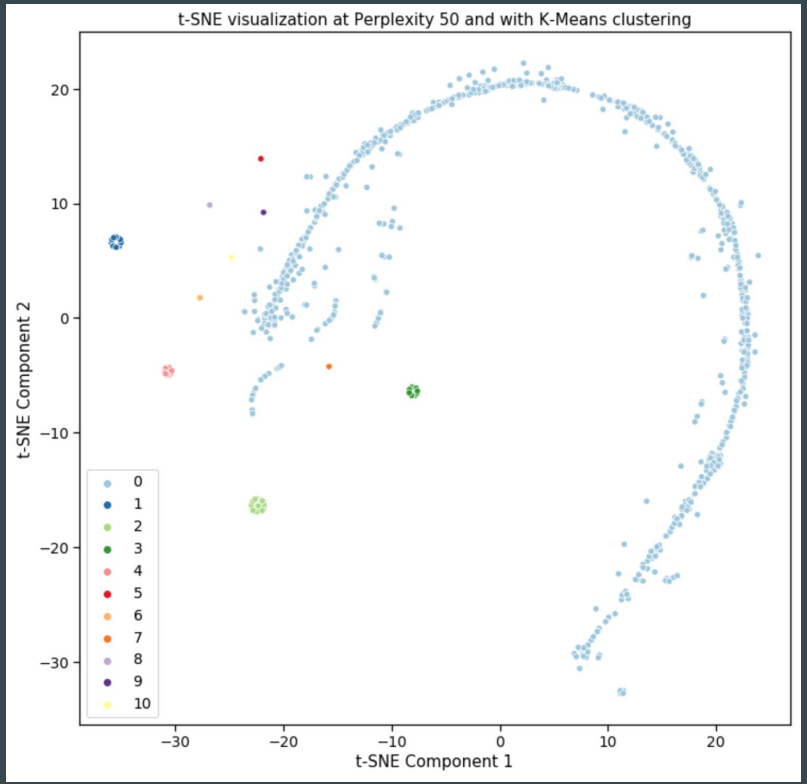
- Résultats encore moins concluants
- Silhouette Score: 0.38
- Davies Bouldin: 1.1



Classification non-supervisée

K-Means

- kPCA non concluant
- Un très grand cluster disproportionné



Classification Supervisée

Model	Training Accuracy	Test Accuracy
Naive Bayes	71,5%	66,2%
SVC	98,6%	80,5%
Logistic Regression	98,6%	**81,0%**
Neural Network (2 layers)	98,6%	79,1%

- Globalement des bons résultats mais risque d'overfitting

Faisabilité de Classification

- Sur base de peu de données, les algorithmes de classification supervisés donnent déjà de bons résultats (80%)
- La classification non-supervisée montre que certains clusters se dégagent mais ceux-ci ne représentent pas les catégories de produits. De plus, une majorité des données reste dans un cluster central, ce qui n'est pas concluant.
- La méthode LDA donne quelques topics représentatifs mais surtout dans les catégories où nous avons plus de données

Traitement des Images

Image d'origine

Grayscale

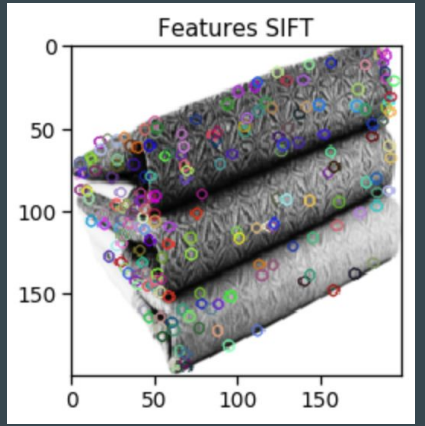
Equalize

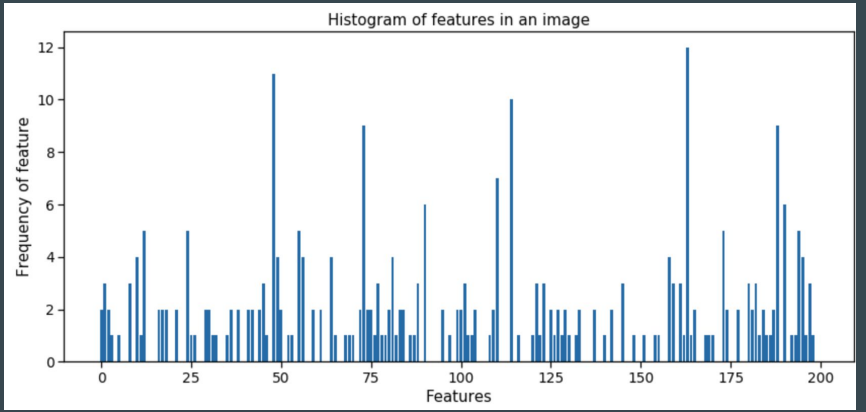
Filter (Gaussian Blur)

Resize
(200x200)

Extraction de Features

- Utilisation de SIFT pour détecter les features dans chaque image: **245,154 features**
- Groupement de features en ‘bag of visual words’ avec K-Means (200 features)
- Création d’un histogramme de features pour chaque image

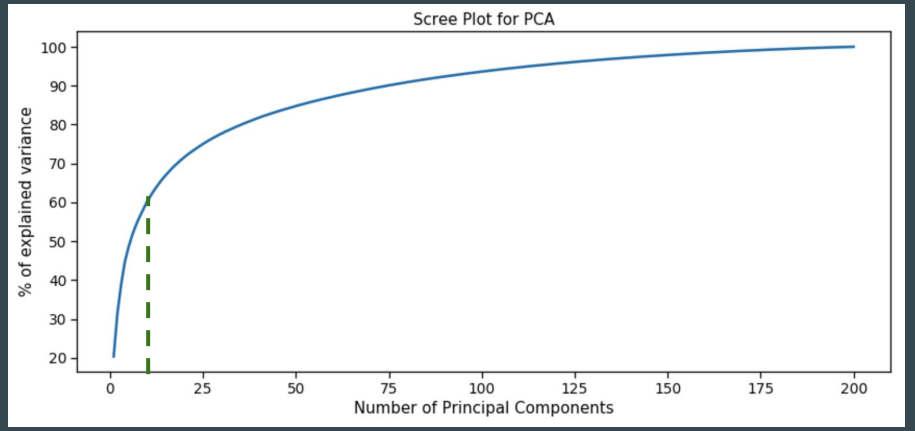


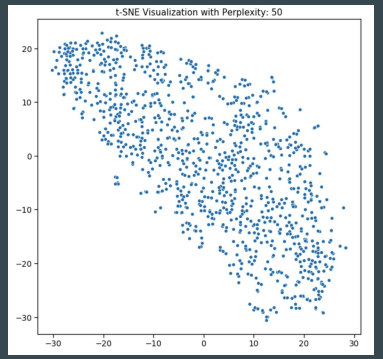


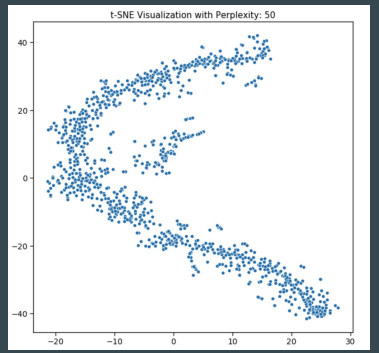
Classification non-supervisée

Réduction de Dimensions

- Avec 10 composantes principales, nous expliquons déjà **60%** de la variance
- Visualisations de cette réduction avec un noyau linéaire et poly



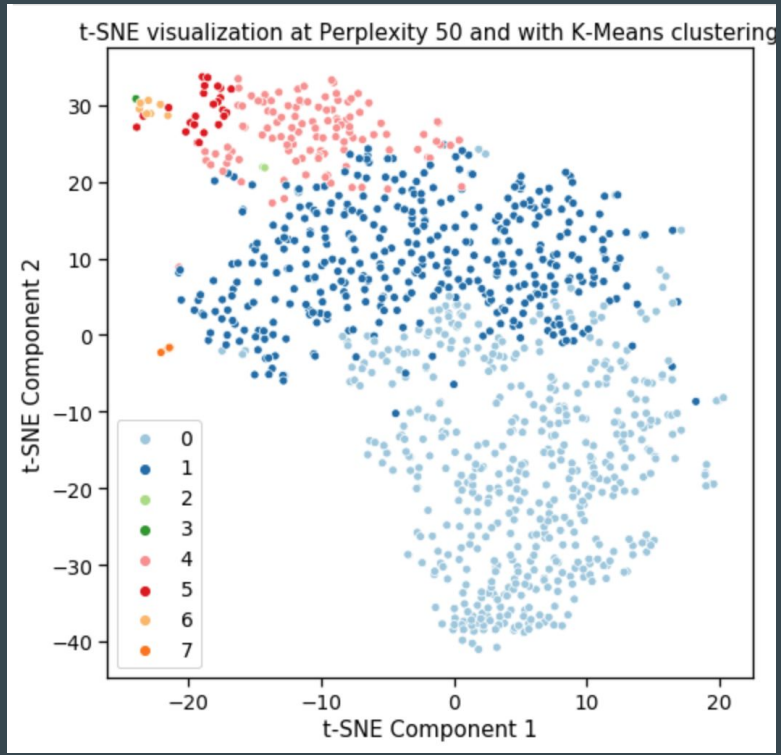




Classification non-supervisée

K-Means

- Coude à 8 clusters
- Pas de présence de clusters distincts et clusters non-denses
- Même situation avec le kPCA



Classification non-supervisée

- Les clusters ne représentent pas bien les catégories
- Beaucoup de catégories se retrouvent au sein d'un même cluster
- Aucune catégorie ne domine particulièrement les clusters
- 3 clusters sont beaucoup plus larges que les autres

cluster_labels_images	category_lvl_1
0	52
1	48
2	2
3	1
4	30
5	13
6	5
7	5

5	Baby Grooming	1
	Bath Linen	2
	Bed Linen	3
	Cookware	2
	Fragrances	1
	Garden & Leisure	1
	Hair Care	1
	Infant Wear	2
	Lighting	1
	Men's Grooming	1
	Network Components	1
	Showpieces	1
	Wrist Watches	7

Classification Supervisée

Model	Training Accuracy	Test Accuracy
SVC	98,6%	32,4%
kNN	100%	25,2%
Logistic Regression	98,0%	**33,8%**

- Résultats non concluants avec overfitting

Transfer Learning

- Utilisation de VGG16 comme modèle de Transfer Learning

Coffee Mug (76,13%)

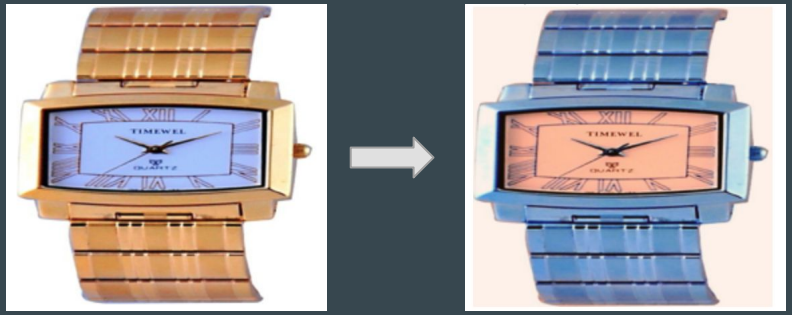
Magnetic Compass (82,90%)

Transfer Learning

- Remplacement de dernière couche par une couche fully-connected contenant les 63 catégories d'objets.
- Paramétrage de la dernière couche uniquement
- Utilisation des images couleur avec fonction pre-processing de Keras
- Redimensionnement des images à 224x224

Layer (type)	Output Shape	Param #
vgg16 (Model)	(None, 1000)	134260544
dense_3 (Dense)	(None, 63)	63063

Total params: 134,323,607
Trainable params: 63,063
Non-trainable params: 134,260,544



Transfer Learning

Model	Training Accuracy	Test Accuracy
SVC	98,6%	32,4%
kNN	100%	25,2%
Logistic Regression	98,0%	33,8%
Transfer Learning	**51,4%**	**46,2%**

En utilisant le Transfer Learning, nous réduisons fortement l'overfitting et atteignons des performances supérieures (+13pp)

Conclusions

- La création d'un moteur de classification basé sur l'image et la description est envisageable **dans un cadre supervisé (catégories de produits définies à l'avance)**
- Avec peu de données, nous atteignons déjà des résultats encourageants
- Les performances devraient augmenter avec davantage de données
- La classification non-supervisée ne donne pas de résultats concluants et ne semble pas permettre de faire une classification automatique reflétant les catégories de produits.

Pistes d'amélioration

- Assemblage des données textuelles et image
- Utilisation du Transfer Learning sur le texte (BERT)
- Elargissement du nombre de données d'entraînement labellisées
- Utilisation d'une couche Dropout pour réduire l'overfitting dans le texte
- Utilisation de catégories plus larges