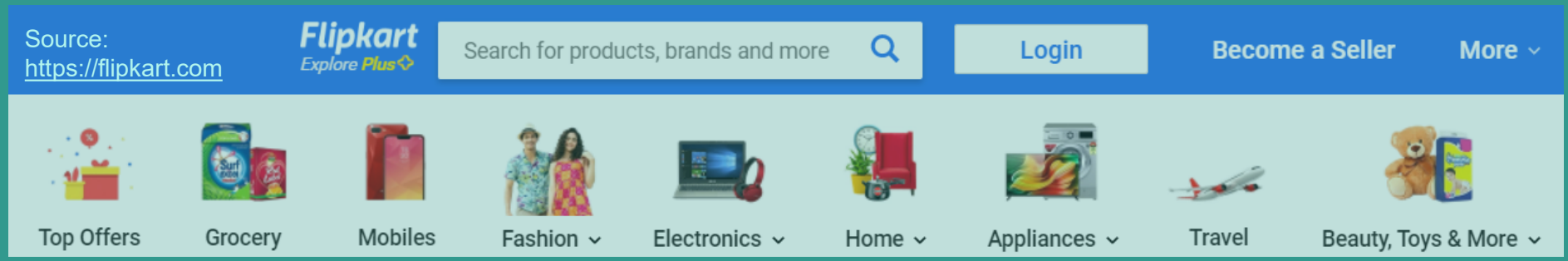


Classifiez automatiquement des biens de consommation

Projet 6 du parcours
« Data Scientist » d'OpenClassrooms

Mark Creasey

Sommaire

Classifiez automatiquement des biens de consommation

01 Présentation de la problématique

Mission - Classification automatique des biens

Etudier et démontrer

- la **faisabilité d'un moteur de classification**
- entre **catégories,**
- basé sur une **image** et une **description** pour chaque article (1050 articles)

Démarches

- **réalise un prétraitement** des descriptions des produits et des images,
- **réduction de dimension**,
- **clustering**,
- présenter graphiquement (2D),
- **calcul de similarité** entre les catégories réelles et les clusters.
- illustre que **les caractéristiques extraites** permettent de **regrouper des produits** de même catégorie.

Interprétation de la problématique

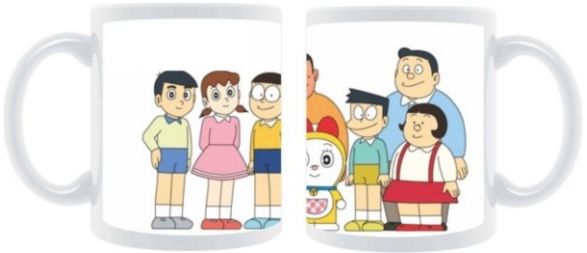
Données

- id
- category_tree
- url
- brand
- product_name
- *specifications*
- description
- price
- rating
- image

Baby Care
>> Baby & Kids Gifts
>> Decorations

Specifications of **Doraemon** **Gift Family Ceramic Coffee Mug** *Multicolour Mug - 325 ml In The Box Sales Package Ceramic Coffee Mug General Type Mug Size in Number 9 inch Size Small*

6325bf868b9040a0599f257aba42e9e0.jpg

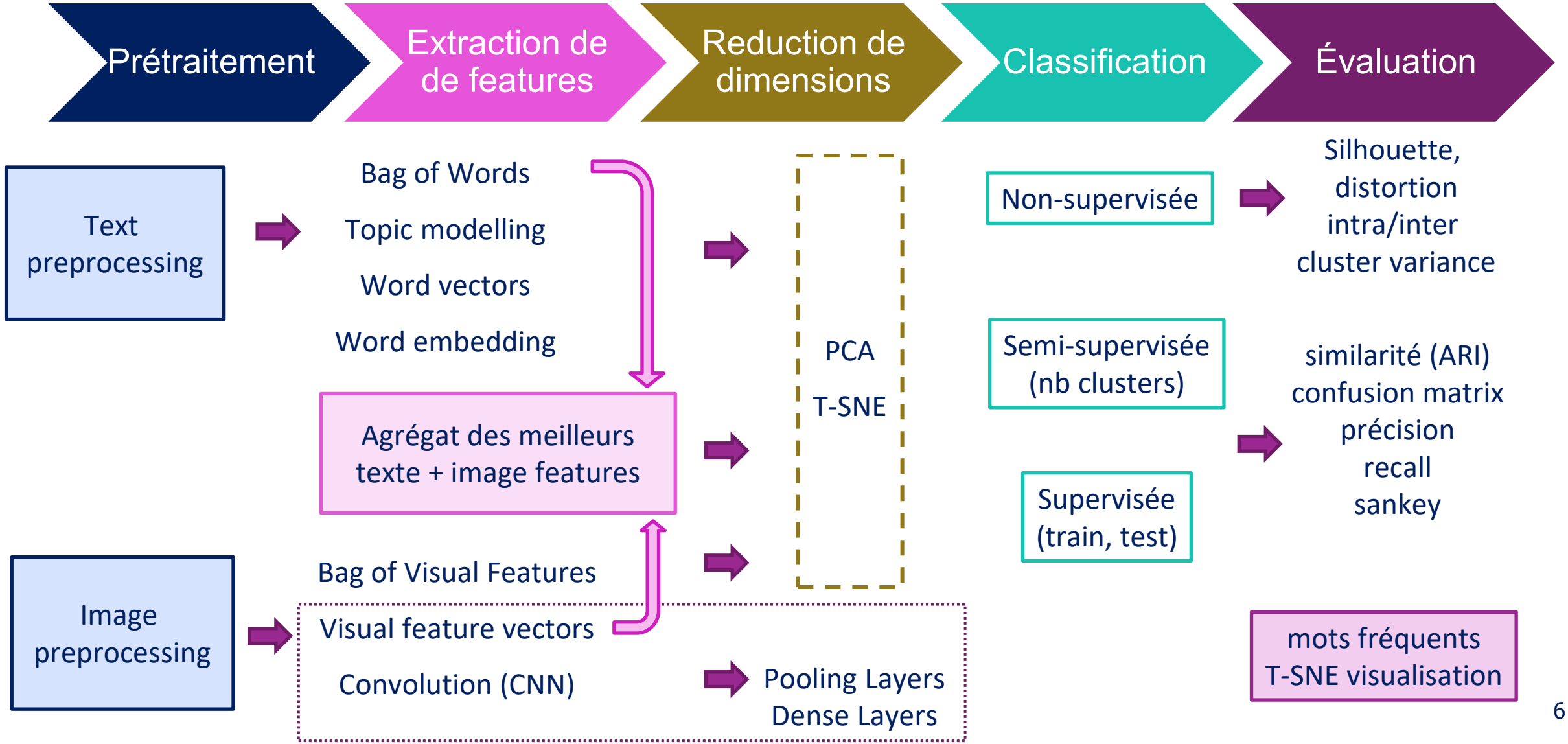


Catégorisation manuelle

- Prend du temps
- Fiabilité ?
- Catégorisation arbitraire ?
- Gestion de la croissance de produits a vendre

flipkart_com-ecommerce_sample_1050.csv

Les démarches

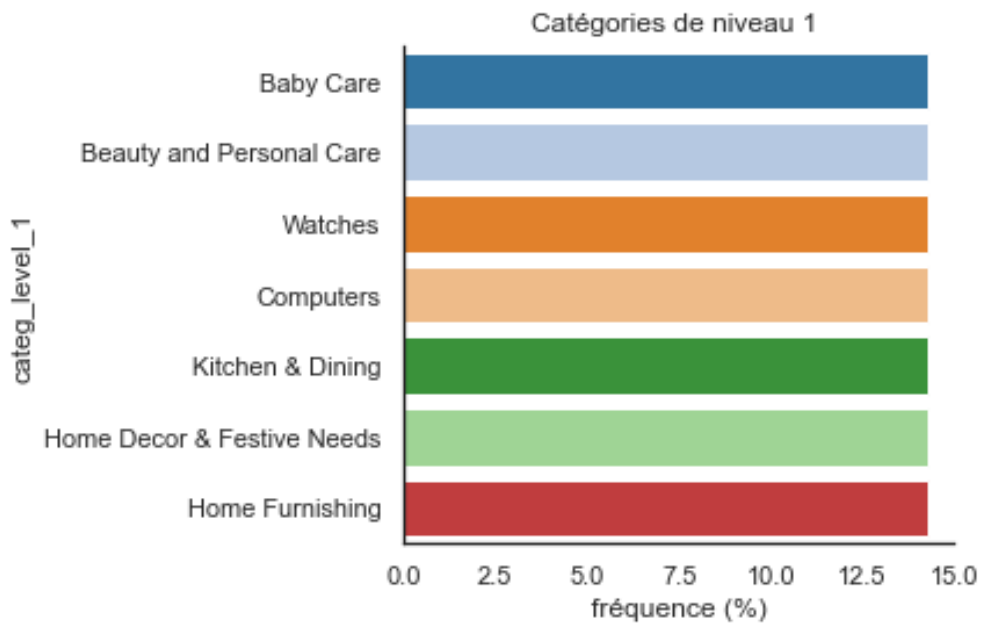


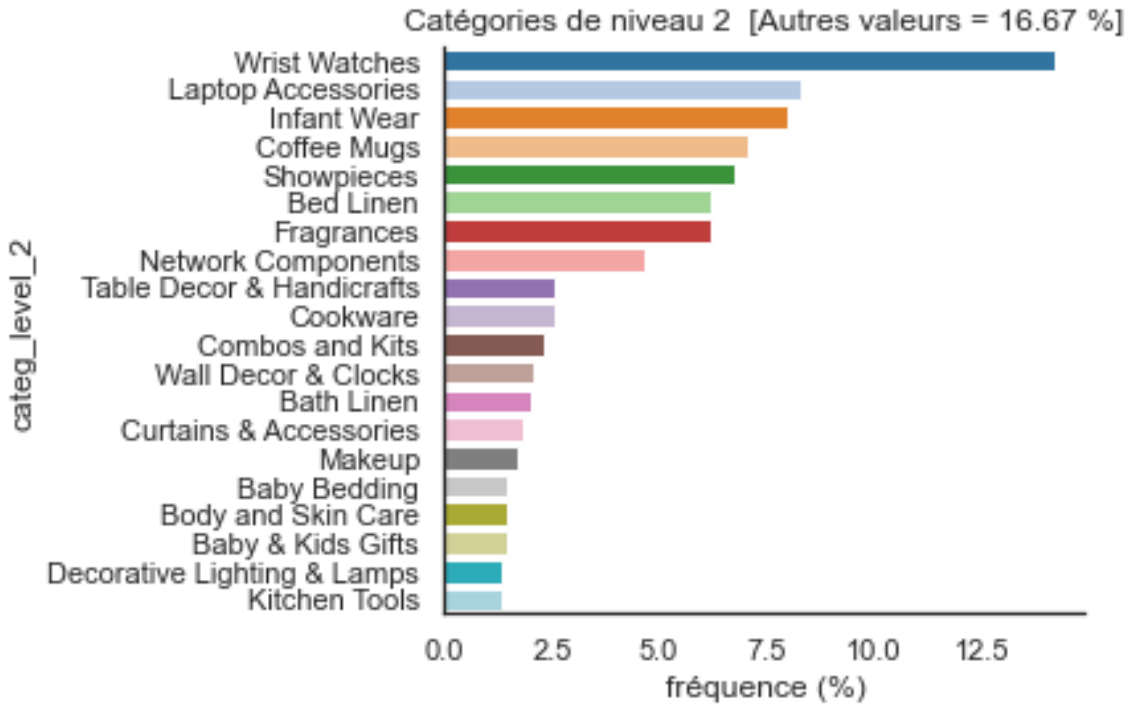
02 Classification des textes

Natural Language Processing (NLP)

Exploration des données textes - catégories

7 catégories principales





62 catégories secondaires

- Poids trop variés : catégories dominants

Exploration des catégories – bruit des mots publicitaires

- Mots fréquents (avant nettoyage)


```
STOP_PHRASES = [
'Buy', 'Only Genuine Products', '!’,
'Cash On Delivery’,
'Free Shipping’,
'30 Day Replacement Guarantee',
'Online', 'at Flipkart.com’,
'from Flipkart.com', 'Flipkart.com’,
'best prices', 'Lowest Prices',
'Great Discounts’,
'in India Only']
```

Exploration catégories – après pré-traitement

- Suppression de stopwords et stop-phrase

Prétraitement des textes

- Elimination des phrases, mots et ponctuation qui ne discrimine le produit

Buy eCraftIndia Floral Cushions Cover at Rs. 404 at Flipkart.com. Only Genuine Products. Free Shipping. Cash On Delivery!

ecraftindia floral cushion cover

Key Features of Mom and Kid Baby Girl's Printed Green Top & Pyjama Set Fabric: Cotton Brand Color: Green,Mom and Kid Baby Girl's Printed Green Top & Pyjama Set Price: Rs. 309 Girls Pyjama set,Specifications of Mom and Kid Baby Girl's Printed Green Top & Pyjama Set General Details Pattern Printed Ideal For Baby Girl's Night Suit Details Fabric Cotton Type Top & Pyjama Set Neck Round Neck In the Box 1 Top & Pyjama Set

key mom kid baby girl printed green top pyjama set fabric cotton brand color greenmom kid baby girl printed green top pyjama set girl pyjama setspecifications mom kid baby girl printed green top pyjama set pattern printed ideal baby girl night suit fabric cotton top pyjama set neck round neck box top pyjama set

- Mise en minuscule
- Stop phrases
  - Elimination publicité
- Suppression des nombres
  - Elimination des prix, dimensions,..
- Tokenisation
  - (mots + ponctuation)
- Stopwords
  - mots fréquents, bas IDF
- Lemmatisation
  - stemming

Bag of Words : Count / TF-IDF vectorization

key mom kid baby girl printed green top pyjama set fabric cotton brand color greenmom kid baby girl printed green top pyjama set girl pyjama setspecifications mom kid baby girl printed green top pyjama set pattern printed ideal baby girl night suit fabric cotton top pyjama set neck round neck box top pyjama set

Description d'un produit (nettoyé)

Feature Extraction

Une **colonne pour chaque mot** du vocabulaire

- **Bag of Words (BOW) = Term Frequency**

baby	girl	grey	blue	pyjama	cotton	hair	product_name
4	5	0	0	6	2	0	Mom and Kid Baby Girl's Printed Green Top & Pyjama Set
0	0	2	0	0	1	0	Kripa's Printed Cushions Cover
4	5	3	4	5	2	0	Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set
0	2	0	0	0	0	10	Burt s Bees Hair Repair Shea And Grapefruit Deep Conditioner

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

baby	girl	grey	blue	pyjama	cotton	hair	product_name
0.25	0.31	0.00	0.00	0.66	0.12	0.00	Mom and Kid Baby Girl's Printed Green Top & Pyjama Set
0.00	0.00	0.31	0.00	0.00	0.09	0.00	Kripa's Printed Cushions Cover
0.25	0.32	0.29	0.29	0.56	0.12	0.00	Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set
0.00	0.09	0.00	0.00	0.00	0.00	0.71	Burt s Bees Hair Repair Shea And Grapefruit Deep Conditioner

Fine tuning features (semi-supervisée) :

Tester plusieurs paramètres de filtrage / prétraitement

- Uni-grams, bi-grams, ..
- min nb. Occurrences (3+)
- regex patterns (>2 chars)

```
CountVectorizer(
  analyzer='word',
  ngram_range=(1, 1),
  min_df=3,
  token_pattern='[a-zA-Z0-9]{3,}'
)
```

Bag of Words – Dimension Reduction

5100 mots (vocabulaire)

1400 features (bag-of-words)

(optionnellement)

400 composants (PCA 99% variance)

2 dimensions t-SNE

Dimension Reduction

PCA

- **Composants qui explique la plus de variance de fréquences des mots** *(réduction du bruit des mots non-commun entre produits de la même catégorie)*

T-SNE

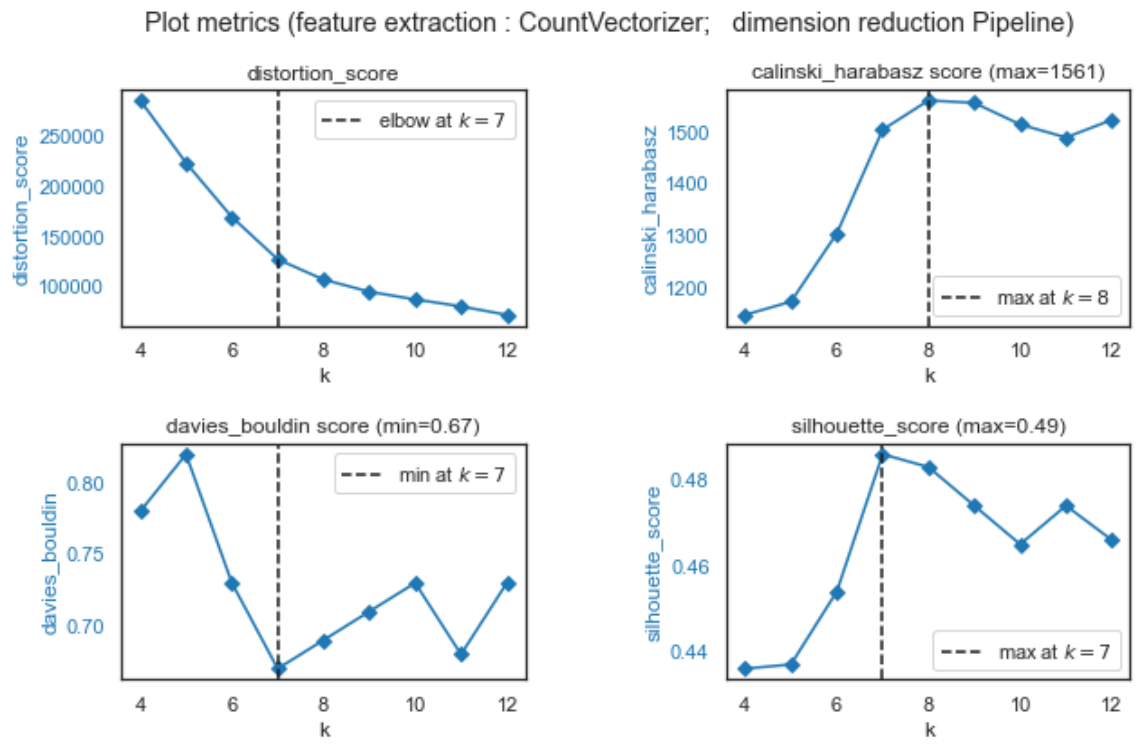
- **Phrases avec similaire fréquences des mêmes mots**

Clustering non-supervisée sur Bag-of-Words

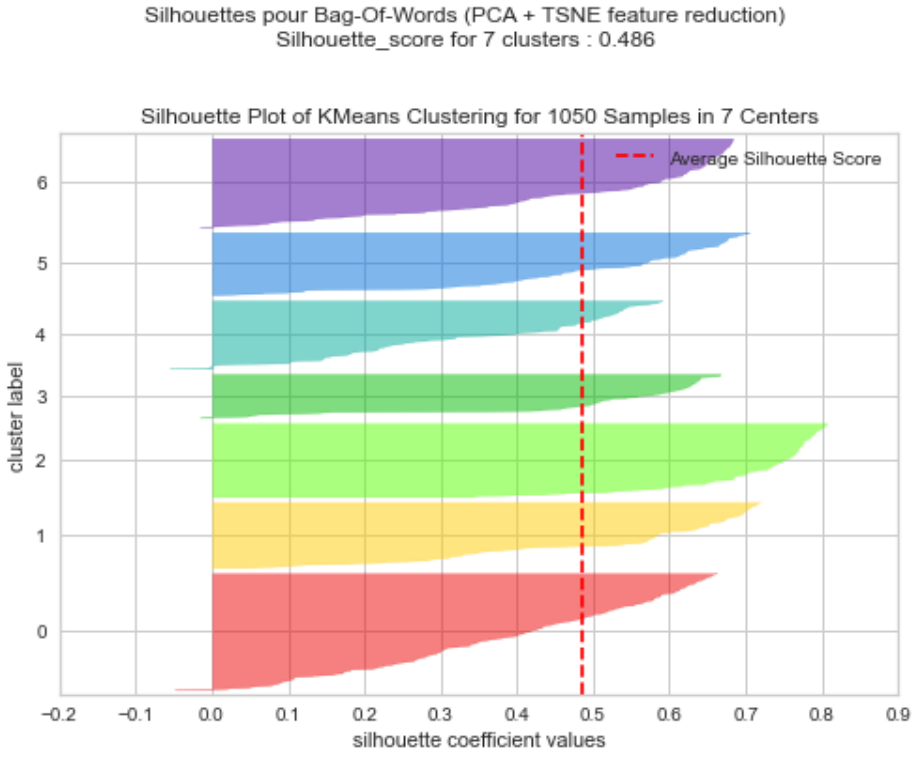
k=7

Meilleure séparation entre clusters

- Silhouette score
- Davies-Bouldin score
- Distortion curve



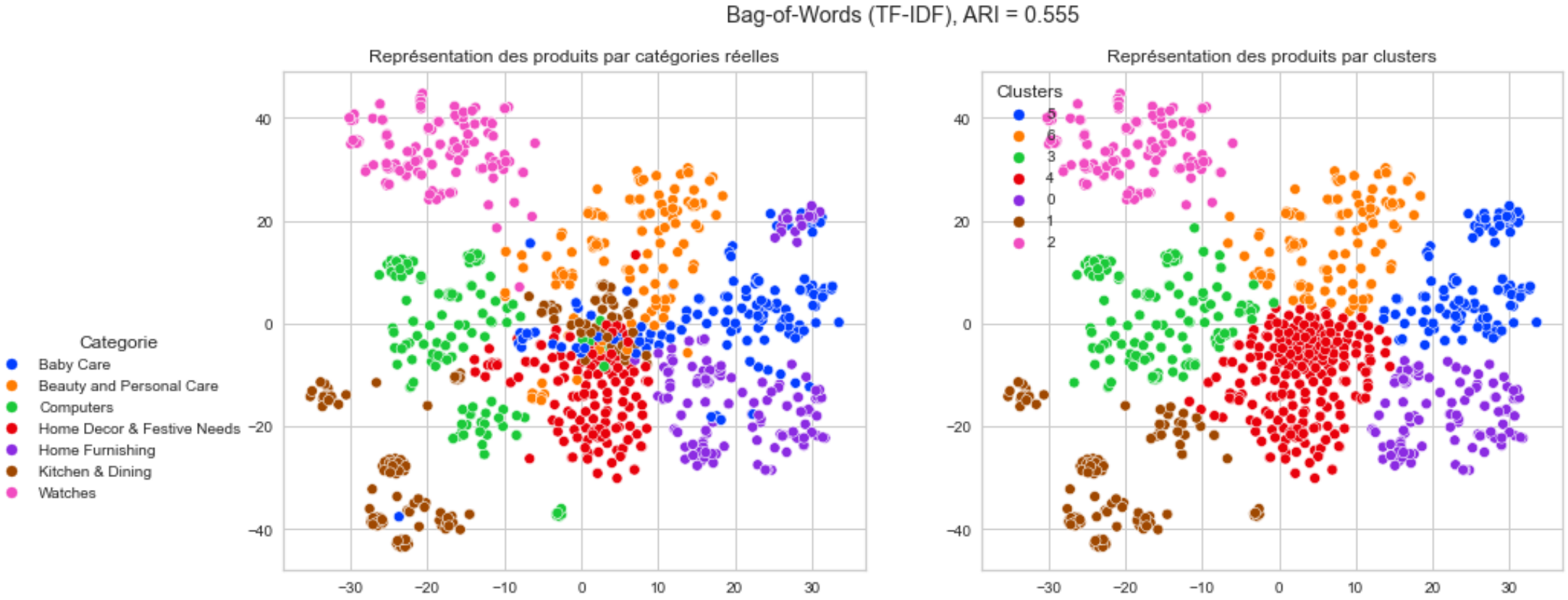
- Choix de K automatisé basé sur silhouette



Clustering semi-supervisée sur Word Frequency (Tf-Idf)

BOW (TF-IDF) : ARI = 0.55

BOW (CountVectorizer) : ARI = 0.48



Evaluation : Précision et Recall

Précision

- Quelle portion du cluster prédit sont du vrai classe ?

$$\frac{TP}{TP + FP}$$

Recall

- Quelle portion du vrai classe sont présent dans le cluster prédit ?

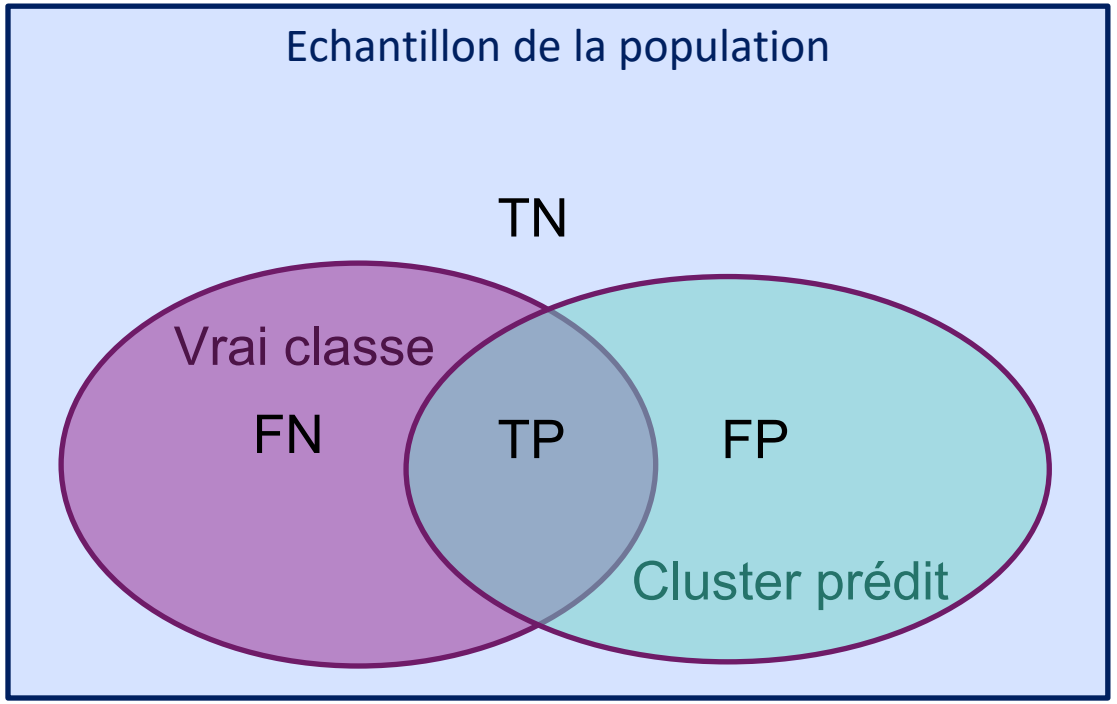
$$\frac{TP}{TP + FN}$$

F1 Score

- accuracy « equilibré » :

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

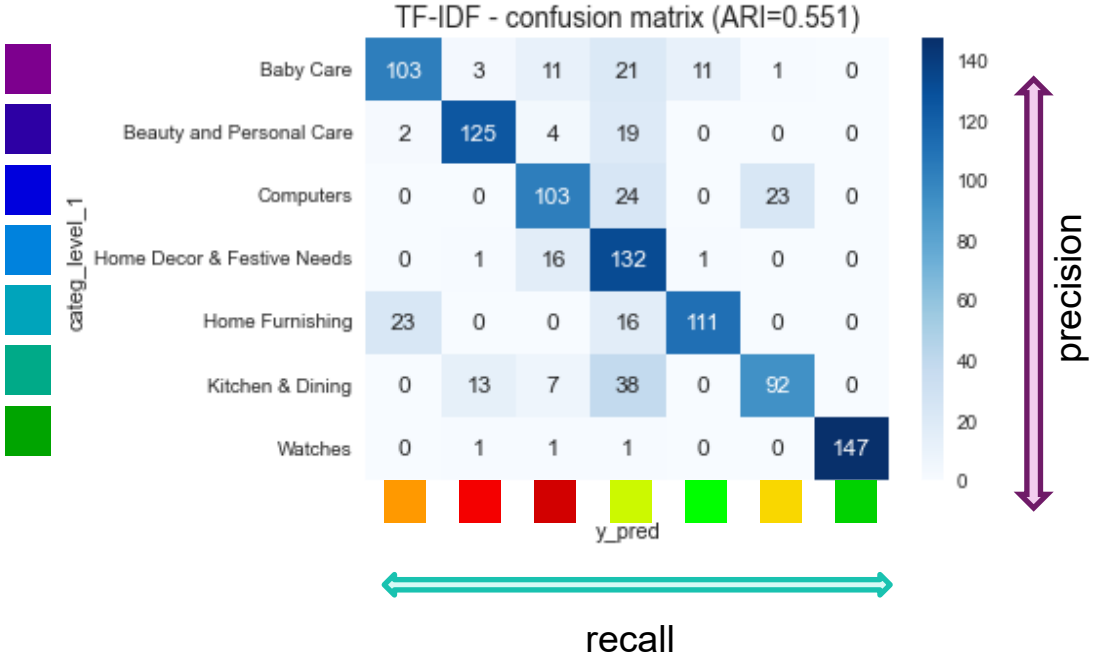
- Cluster prédit : TP + FP
- (Vrai classe : TP + FN)



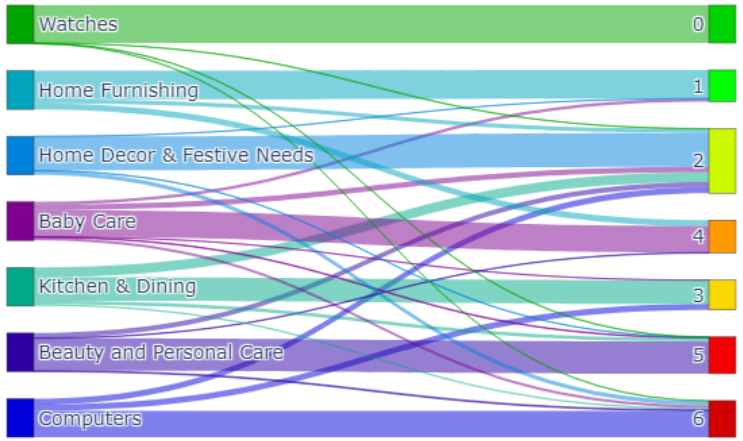
Clustering (semi-) supervisée : évaluation

- TF-IDF, ARI = 0.55

- Confusion Matrix
  - Classification Report
  - Sankey Diagramme



	$\frac{TP}{TP+FP}$ precision	$\frac{TP}{TP+FN}$ recall	f1-score	support
Baby Care	0.80	0.69	0.74	150
Beauty and Personal Care	0.87	0.83	0.85	150
Computers	0.73	0.69	0.71	150
Home Decor & Festive Needs	0.53	0.88	0.66	150
Home Furnishing	0.90	0.74	0.81	150
Kitchen & Dining	0.79	0.61	0.69	150
Watches	1.00	0.98	0.99	150
accuracy			0.77	1050
macro avg	0.80	0.77	0.78	1050
weighted avg	0.80	0.77	0.78	1050

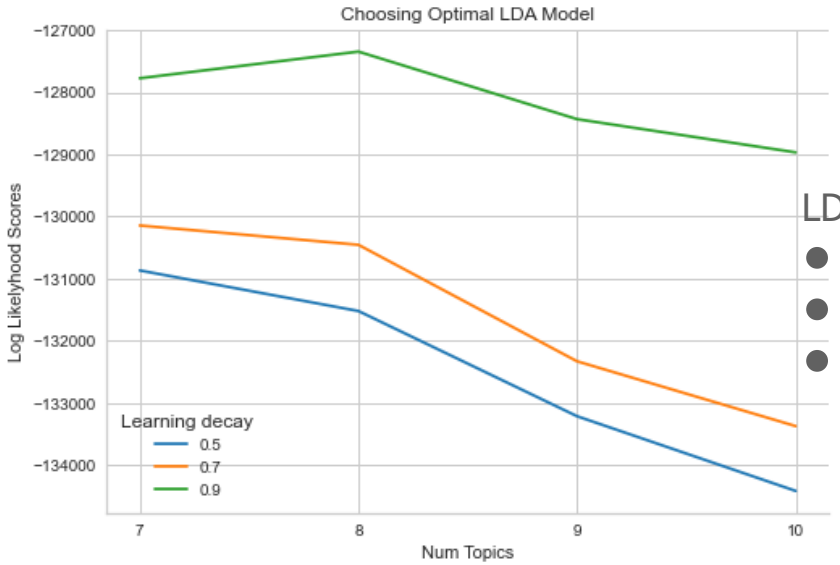


Topic Modelling (LDA)

Latent Dirichlet Allocation

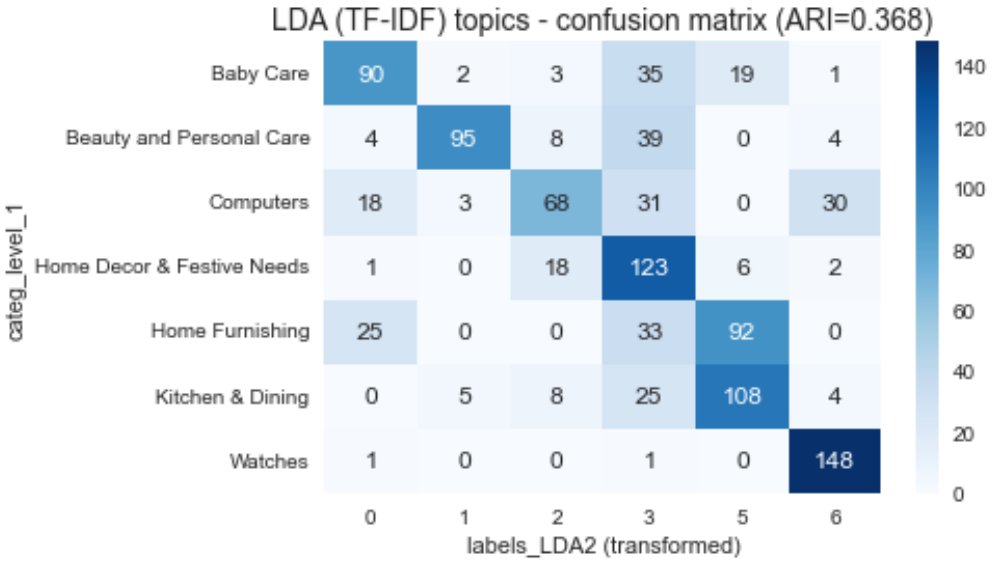
- identifie les topics
- distribution des mots par topic
- topic plus dominant pour chaque description

ARI= 0.37



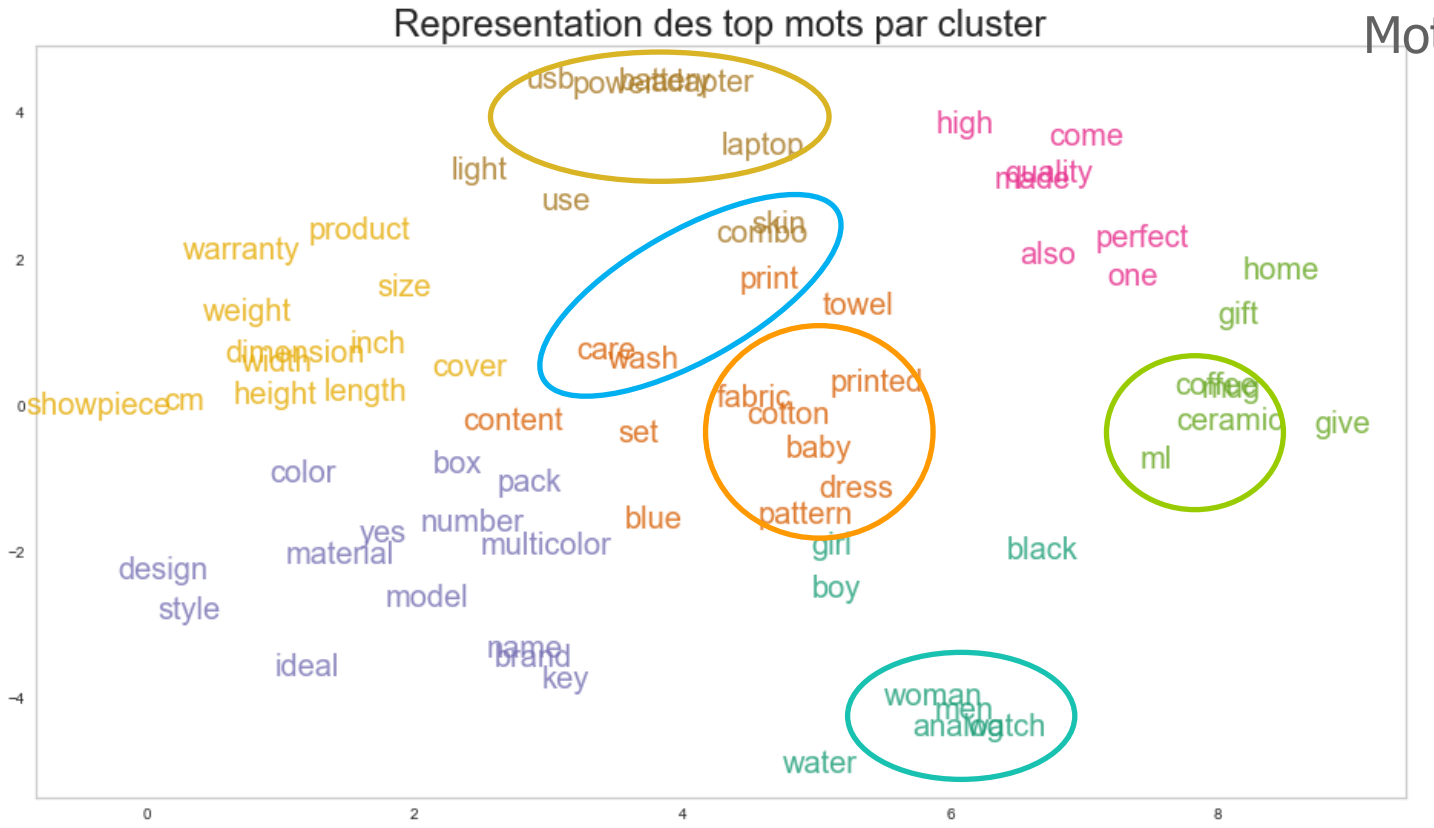
LDA - hyperparamètres

- nb topics
- learning decay
- learning offset



Word2Vec – Word vectors

- Mots qui se trouve fréquemment ensemble dans des phrases sont attribués des vecteurs similaires



Pré-entrainé sur skipgrams

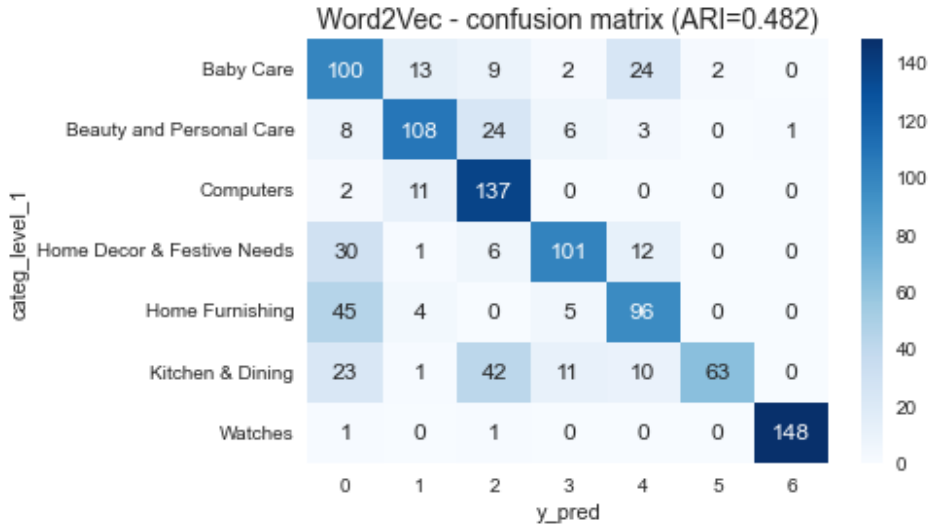
- Comme pour bi-grams, mais avec des mots pas (toujours) directement cote-a-cote

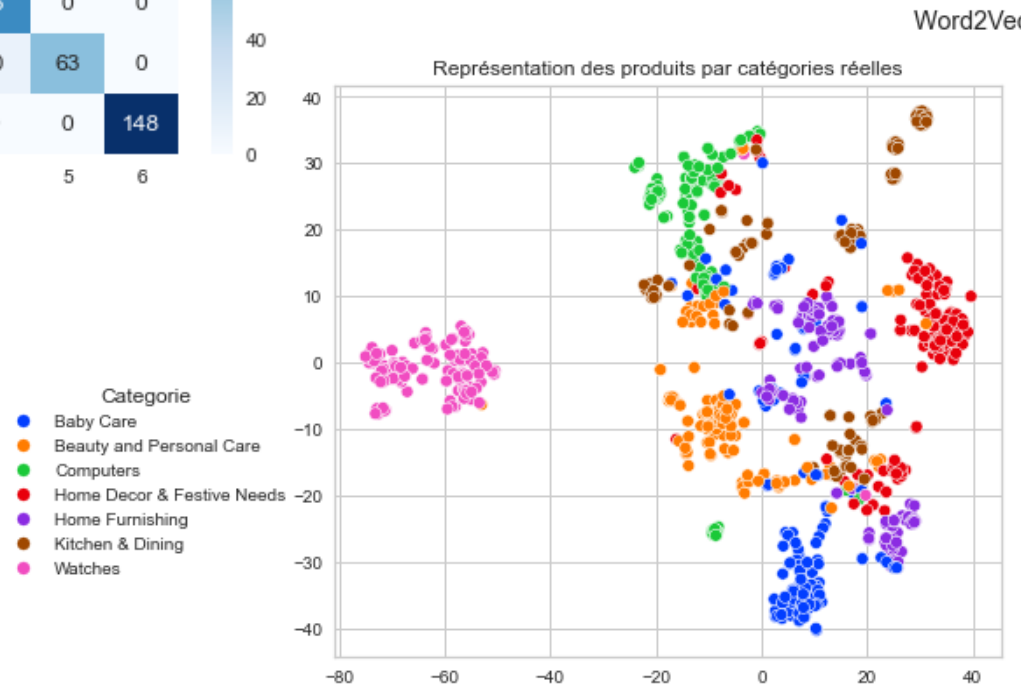
Mot cible

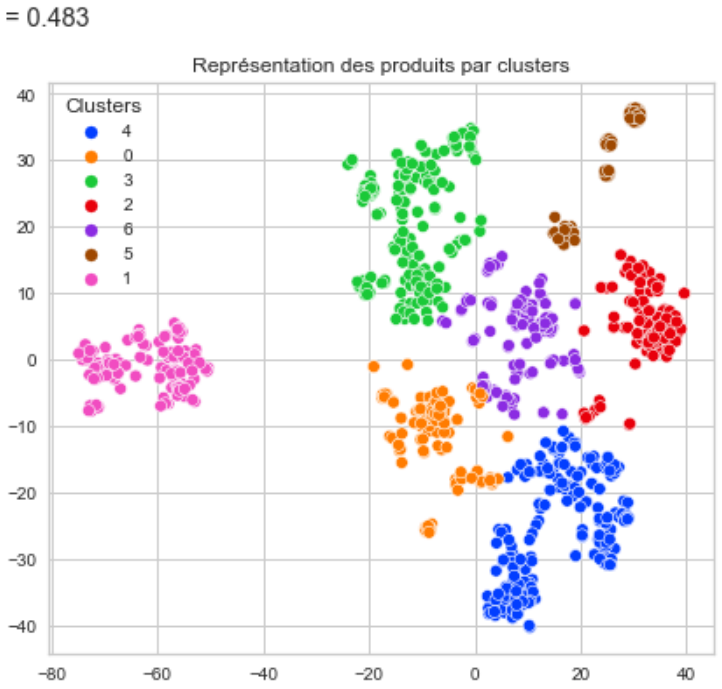
Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot

Clustering sur Word Vectors

- ARI = 0.48

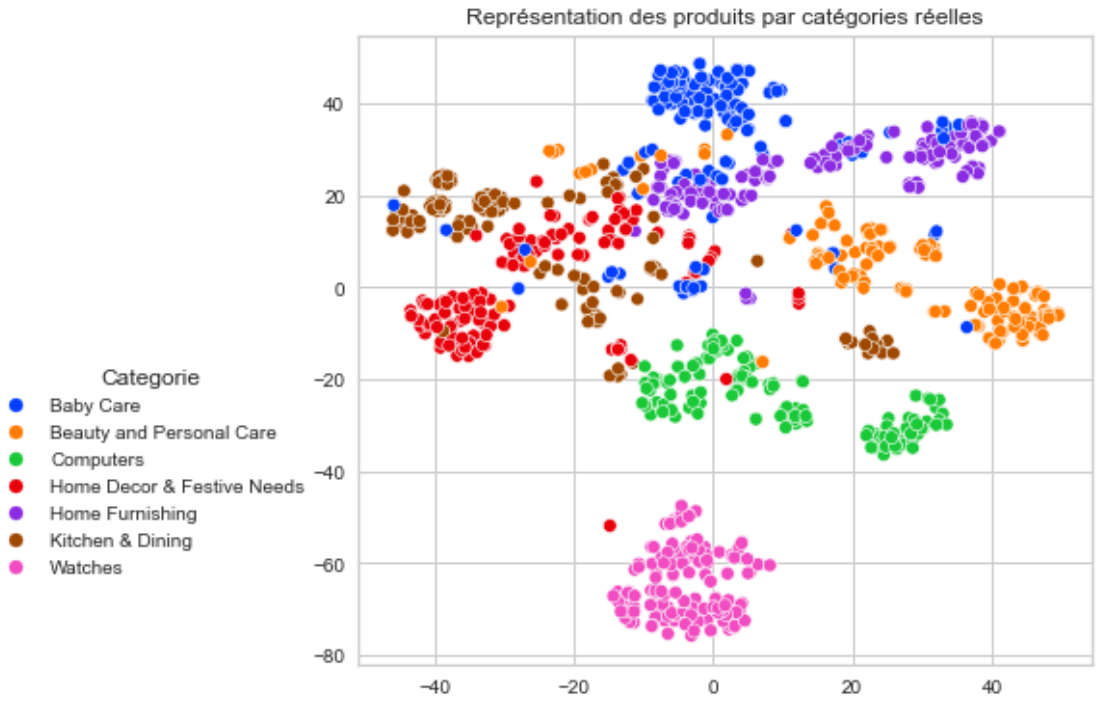






Word Embedding contextuel

- Embedding skipgrams dans des réseaux neuronal LSTM pour ajouter **du contexte** sur des mots
- permet de distinguer des homonyms



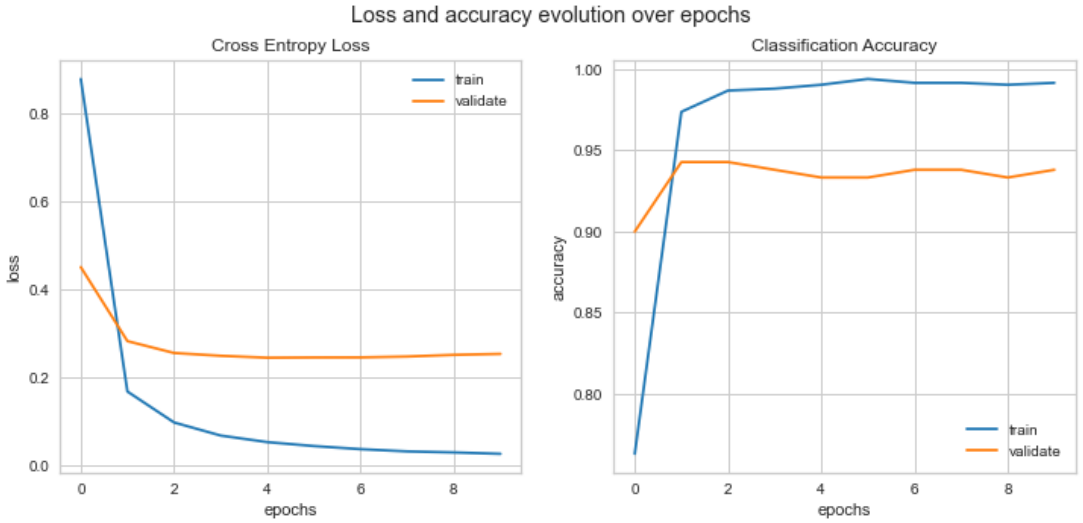
Models (pré-entrainés) utilisés

- BERT
  - Bert uncased (ARI = 0.36)
  - Bert HuggingFace (ARI = 0.41)
- USE
  - Universal Sequence Encoder
  - (ARI = 0.46)

Classification supervisée des textes

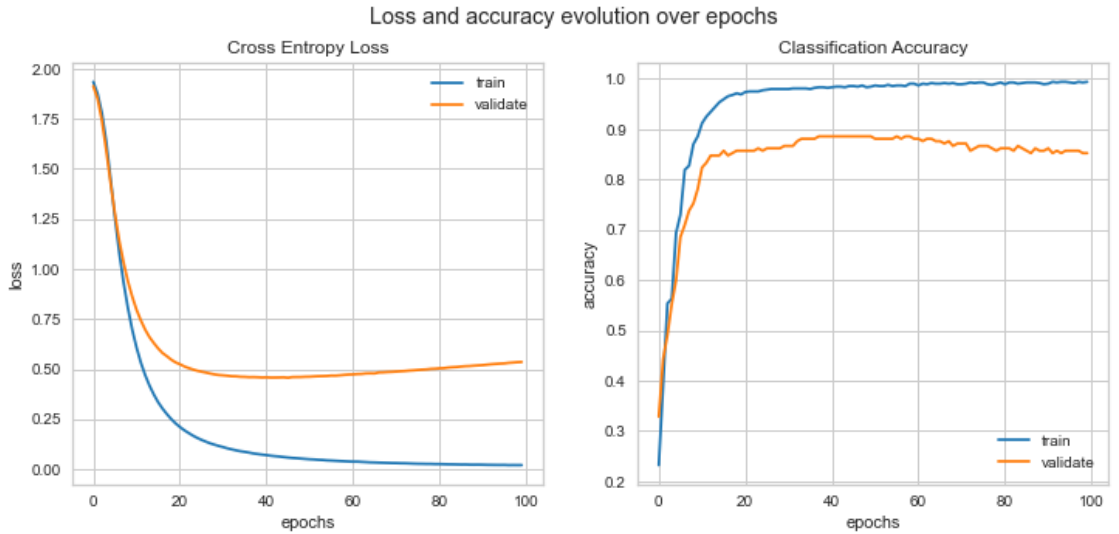
- TF-IDF (ARI=0.86)

	precision	recall	f1-score
Watches	0.88	0.88	0.88
Home Furnishing	0.91	0.94	0.93
Baby Care	0.94	0.97	0.96
Computers	0.96	0.83	0.89
Kitchen & Dining	0.96	0.96	0.96
Home Decor & Festive Needs	0.93	1.00	0.96
Beauty and Personal Care	1.00	1.00	1.00
accuracy			0.94
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94



- Keras.Embedding (ARI=0.69)

	precision	recall	f1-score
Watches	0.79	0.81	0.80
Home Furnishing	0.94	0.88	0.91
Baby Care	0.88	0.88	0.88
Computers	0.81	0.76	0.79
Kitchen & Dining	0.84	0.81	0.82
Home Decor & Festive Needs	0.77	0.92	0.84
Beauty and Personal Care	0.93	0.90	0.91
accuracy			0.85
macro avg	0.85	0.85	0.85
weighted avg	0.86	0.85	0.85



Sommaire – classification textes

Semi-Supervisée

Model	ARI
TF-IDF	0.55
BOW (PCA+TSNE)	0.48
LDA topics (BOW)	0.33

Model	ARI
Word2Vec	0.48
Universal Sequence Encoder	0.46
BERT Huggingface	0.41
BERT base uncased	0.36

Supervisée

- Overfitting sur le jeu d'entrainement
- Attention: seulement nos propres vocabulaires → biais vers les produits et catégories existants

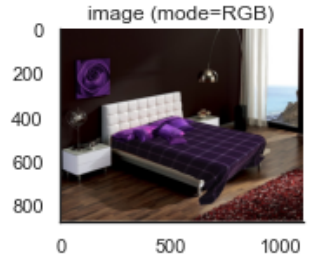
Model	ARI
CNN sur TF-IDF vectors	0.86
Keras WordEmbedding	0.69

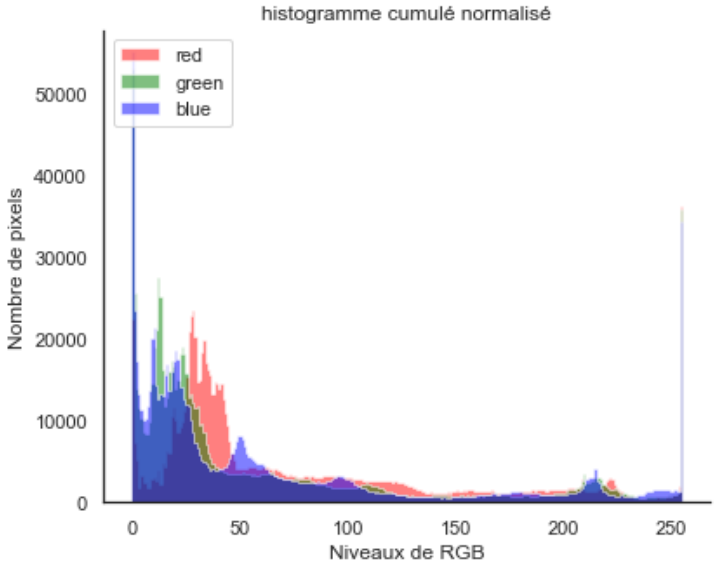
03. Classification des images

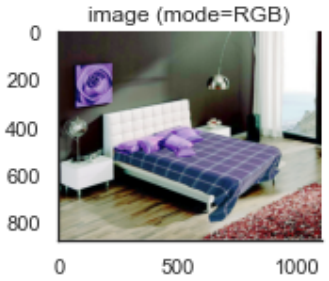
Computer Vision (CV)

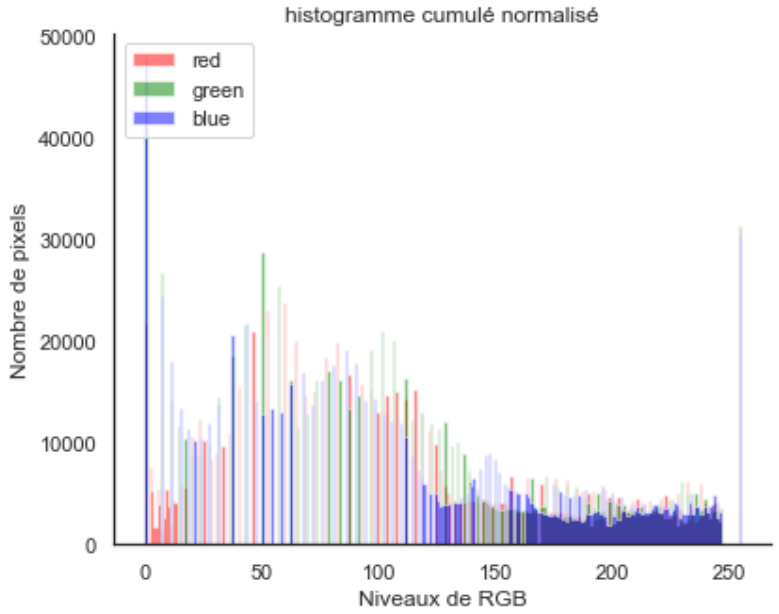
Pré-traitement des images

- Redimensionnement
- Couleur → Gris
- Exposition
- Egalisation de contrast
- Filtrage de bruit
- Changement en forme carré
- Normalisation de values entre -1 et 1



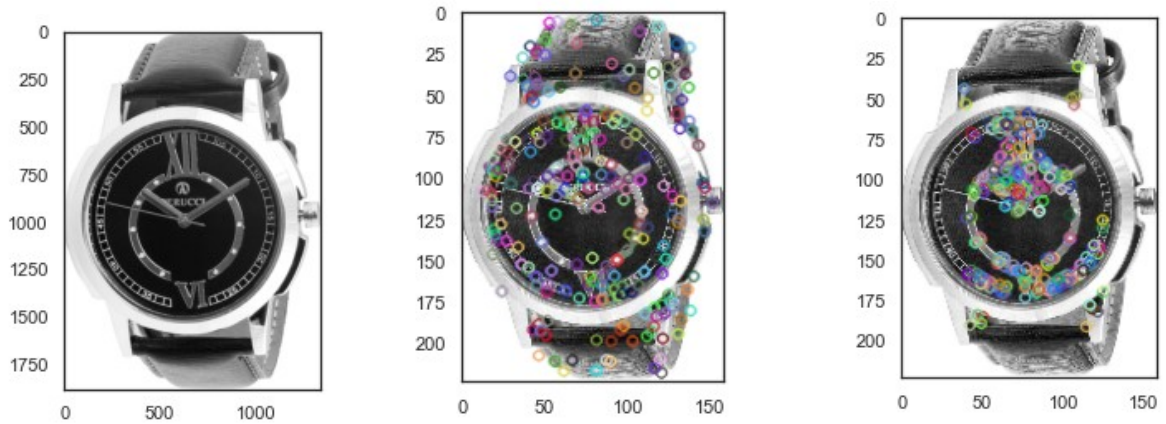






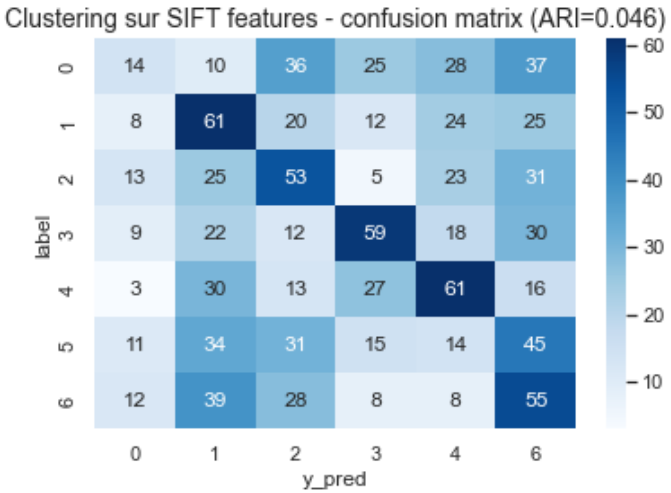
SIFT / ORB bag of visual words

- SIFT (**S**cale-**I**nvariant **F**eature **T**ransform)
- ORB (**O**riented **F**AST and **R**otated BRIEF)

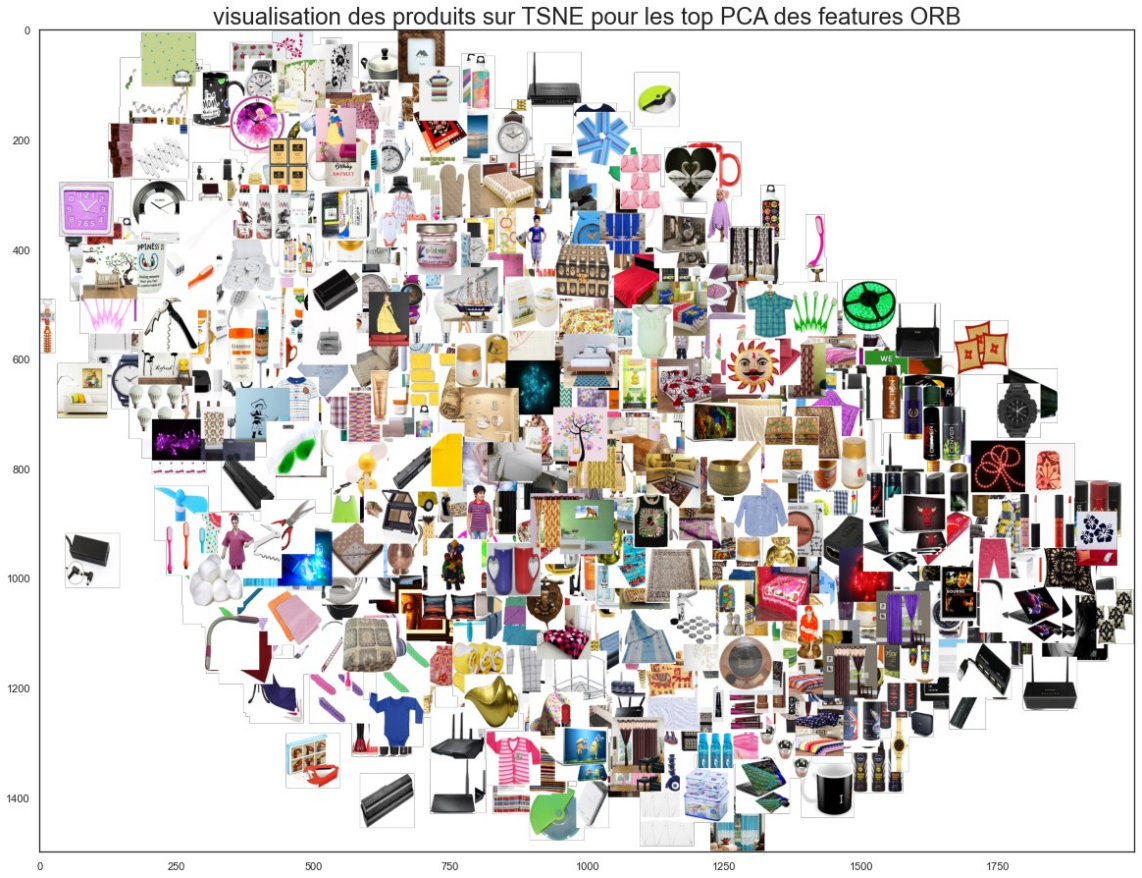


Descripteurs SIFT

Descripteurs ORB

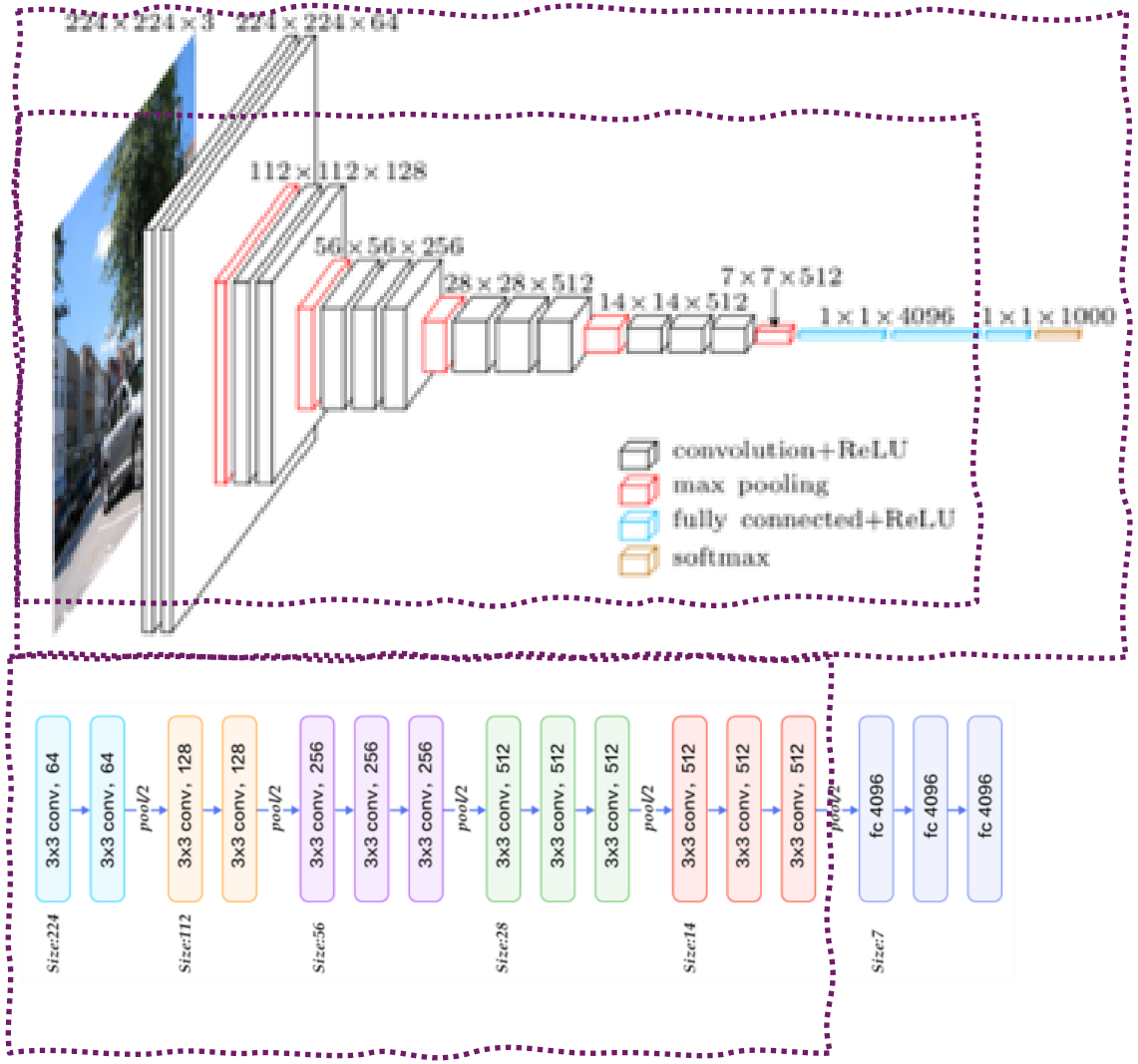


- SIFT, ARI = 0.04
- ORB, ARI= 0.05
- Pas concluant



VGG16 CNN pre-entrainé (ImageNet)

- 16 couches



1. Classification semi-supervisée
   - Extraire les 1000 labels de la dernière couche
   - feature vectors shape [n x 1000]
   - PCA → tSNE → Kmeans

2. Classification non-supervisée
   - Extraire features moins 2 couches
   - feature vecteurs shape [n x 4096]
   - PCA → tSNE → Kmeans

3. Classification supervisée
   - **Transfer learning**
   - Remplacer les dernier 4 couches du VGG16
     a) Extraction des features de le dernière couche
     a. Fine Tuning

VGG-16 non-supervisée (1000 features)

coffee_mug (98%)

AKUP life-is-not-living Ceramic Mug

bath_towel (100%)

Sathiyas Cotton Bath Towel

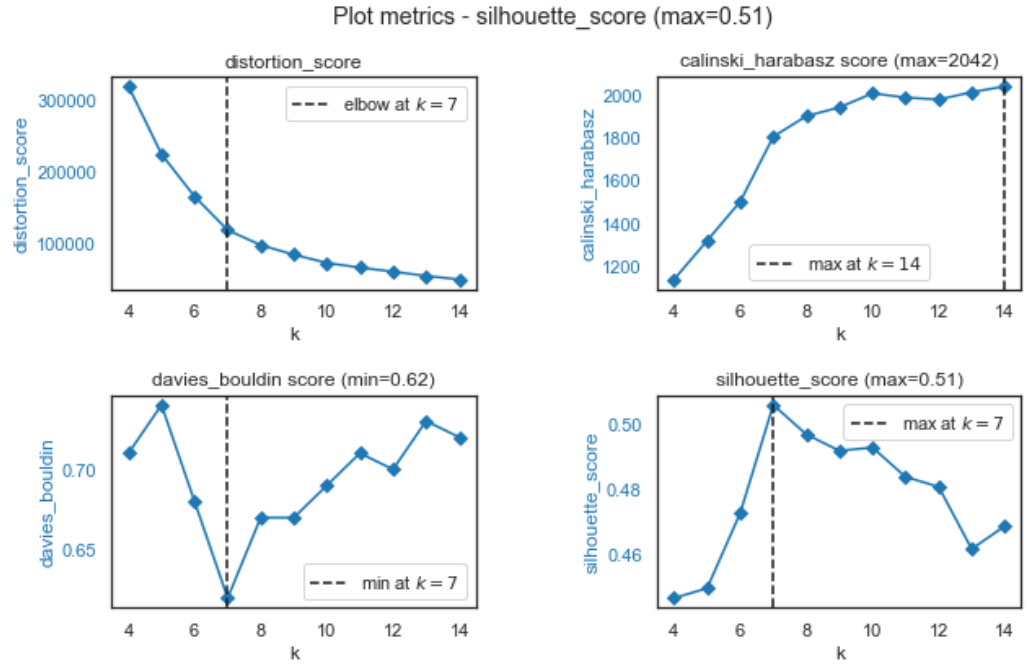
digital_watch (82%)

Youth Digital Watch - For Men, Boys

1. Vérification de l'applicabilité du modèle à nos données

2. Feature Extraction
   - VGG16.predict(image) → **1000 features**
   - Probabilités d'être dans chacun des 1000 classes ImageNet

3. Reduction de dimension : PCA + TSNE



4. Clustering non supervisée **k=7**

   Meilleure séparation entre clusters
   - Silhouette score
   - Davies-Bouldin score
   - Distortion curve

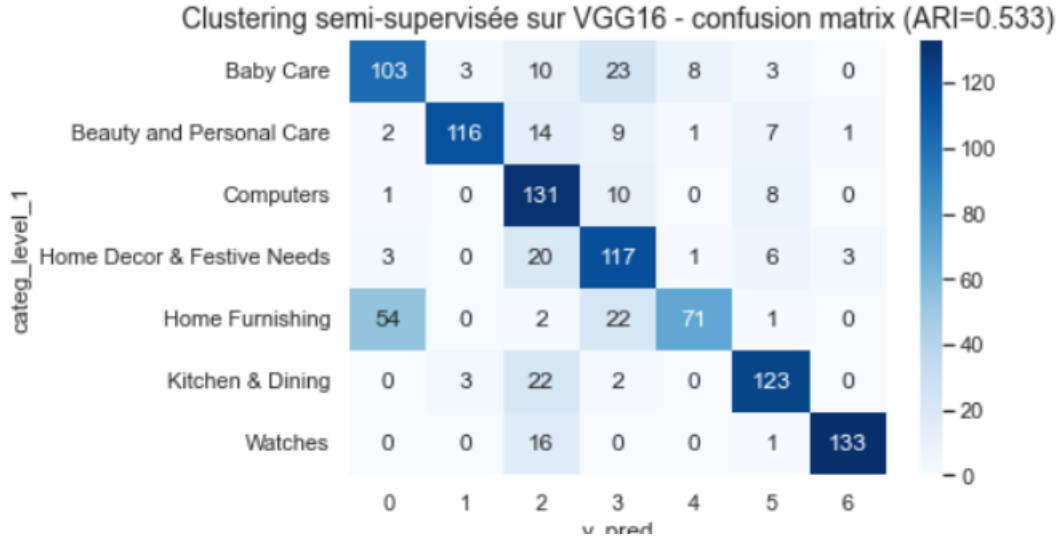
5. Comparaison avec catégories fournies

ARI = 0.38

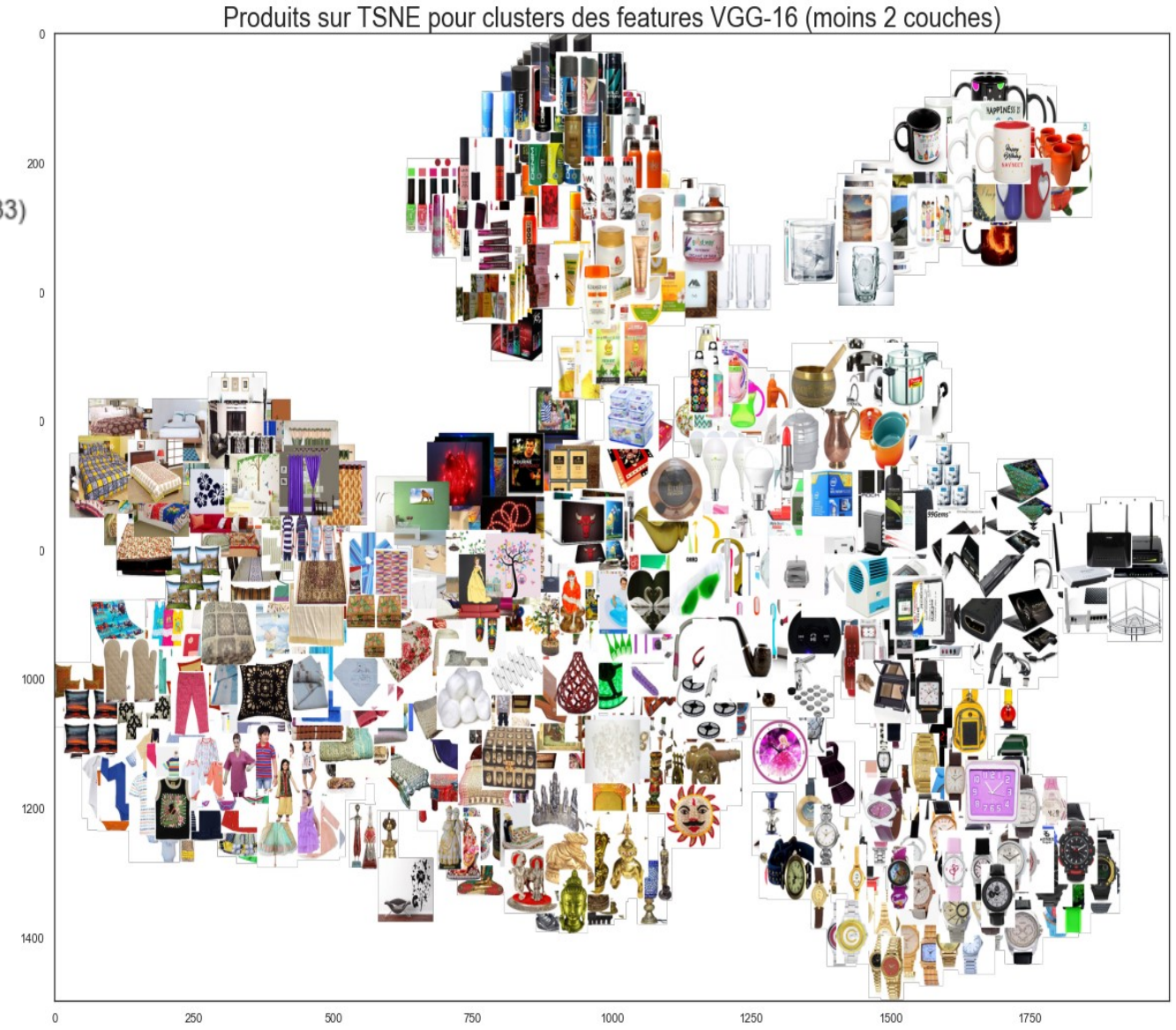
VGG-16 semi-supervisée (4096 features)

- Extraction des features 2 couches avant la derniere couche Softmax

ARI = 0.53

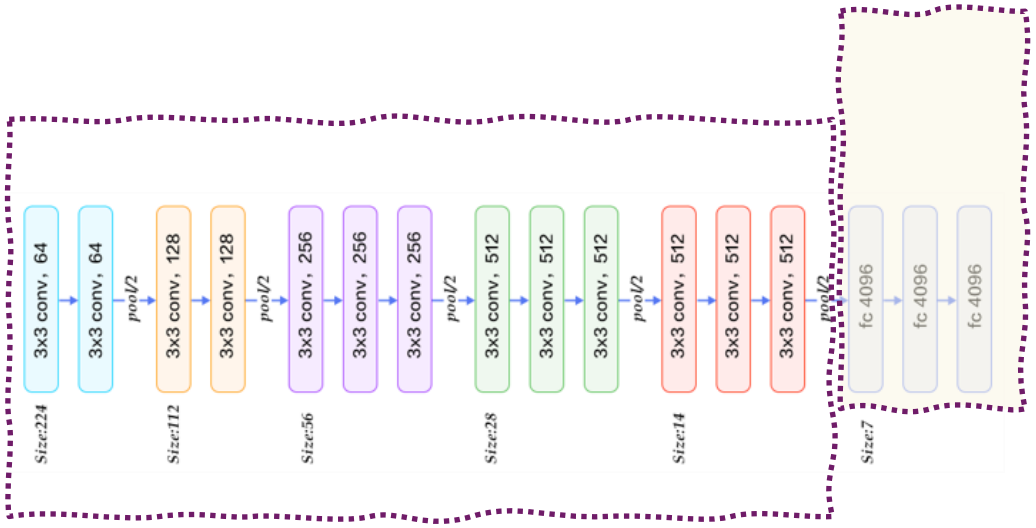


	precision	recall	f1-score	support
Baby Care	0.63	0.69	0.66	150
Beauty and Personal Care	0.95	0.77	0.85	150
Computers	0.61	0.87	0.72	150
Home Decor & Festive Needs	0.64	0.78	0.70	150
Home Furnishing	0.88	0.47	0.61	150
Kitchen & Dining	0.83	0.82	0.82	150
Watches	0.97	0.89	0.93	150



VGG-16 supervisée (Transfer Learning)

- Remplacement des derniers couches « fully connected » avec nos propres couches de pooling/flatten, dropout, et fully connected



Base model

Etape 1

- Extraction des features des couches de convolution de VGG16 pré-entrainé sur des images ImageNet (base_model)

Etape 2

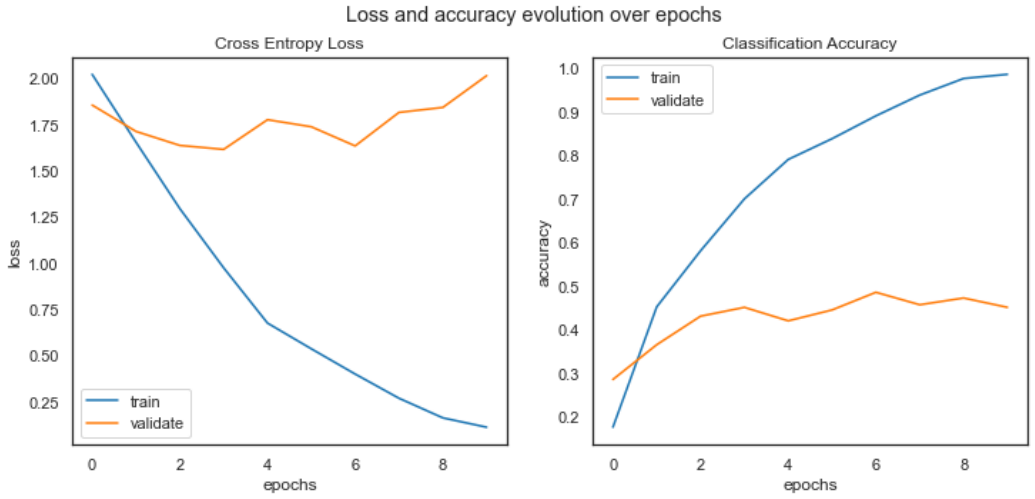
- Fine tuning sur nos images

Régularisation pour éviter « overfitting »

Overfitting : Régularisation

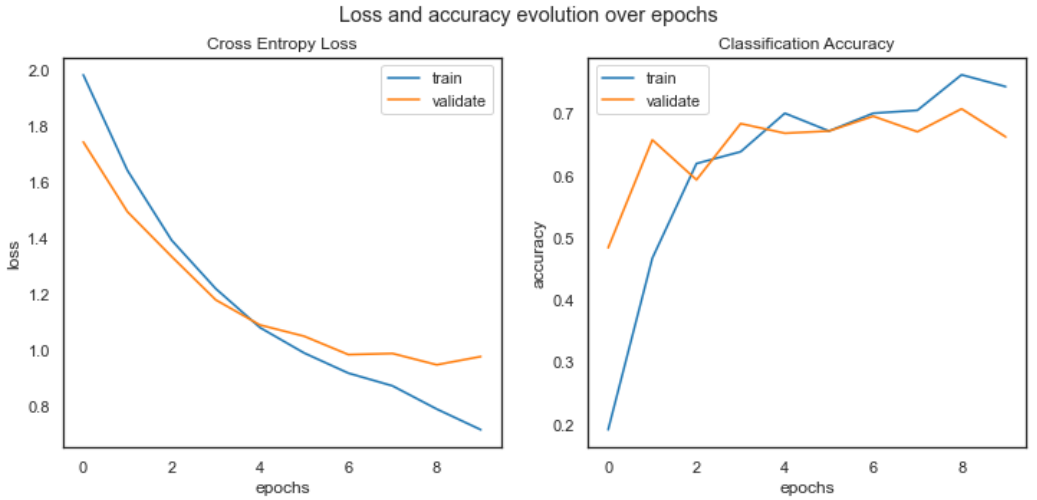
Overfitting

- Les poids des features de training empêche d'identifier les images similaires dans les données test



Régularisation

- Dropout Layers (pendant training)
- Augmentation des images

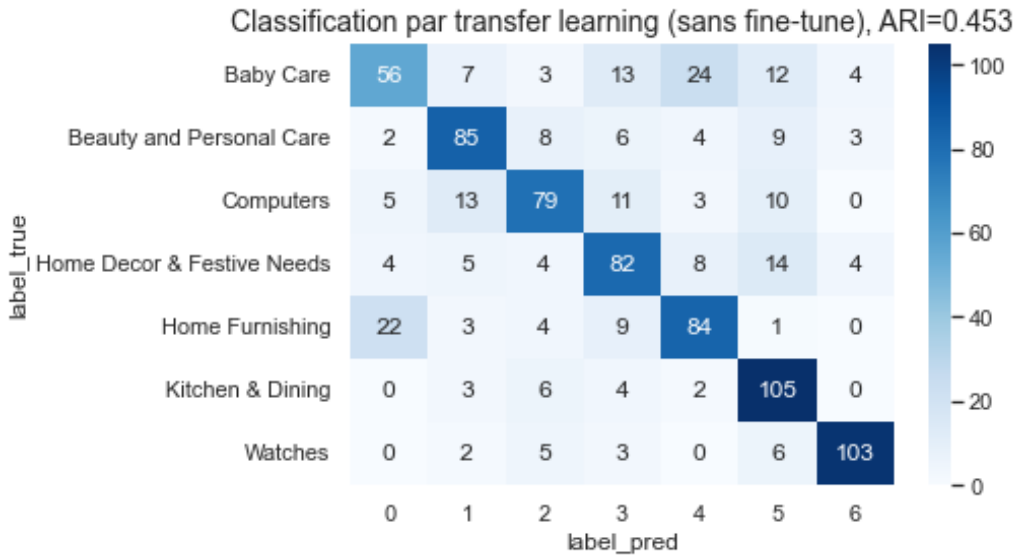


Transfer Learning – VGG16 (pre-entrainé sur ImageNet)

Transfer Learning (Extract Features)

- ARI = 0.45

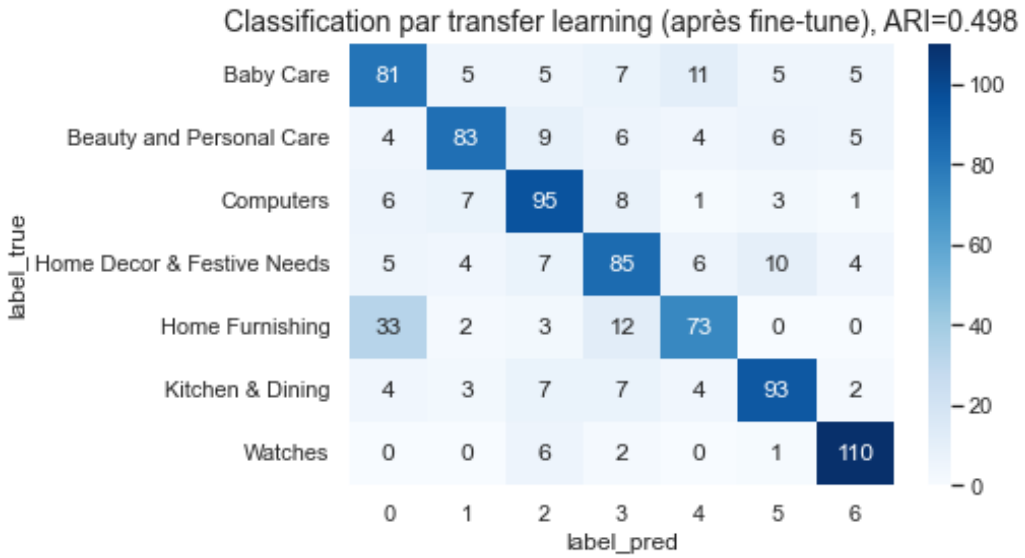
	precision	recall	f1-score	support
Baby Care	0.63	0.47	0.54	119
Beauty and Personal Care	0.72	0.73	0.72	117
Computers	0.72	0.65	0.69	121
Home Decor & Festive Needs	0.64	0.68	0.66	121
Home Furnishing	0.67	0.68	0.68	123
Kitchen & Dining	0.67	0.88	0.76	120
Watches	0.90	0.87	0.88	119



Transfer Learning (Fine Tune)

- ARI = 0.50

	precision	recall	f1-score	support
Baby Care	0.61	0.68	0.64	119
Beauty and Personal Care	0.80	0.71	0.75	117
Computers	0.72	0.79	0.75	121
Home Decor & Festive Needs	0.67	0.70	0.69	121
Home Furnishing	0.74	0.59	0.66	123
Kitchen & Dining	0.79	0.78	0.78	120
Watches	0.87	0.92	0.89	119



Sommaire – classification images

Semi-Supervisée

Model	ARI
SIFT	0.05
ORB	0.04

Model	ARI
VGG-16 pré-entrainé	0.48
VGG-16 (moins 2 couches)	0.46

Supervisée

- CNN: seulement nos propres images → overfitting des produits existants

Model	ARI
Transfer Learning	0.45
Transfer Learning + Fine tune	0.50

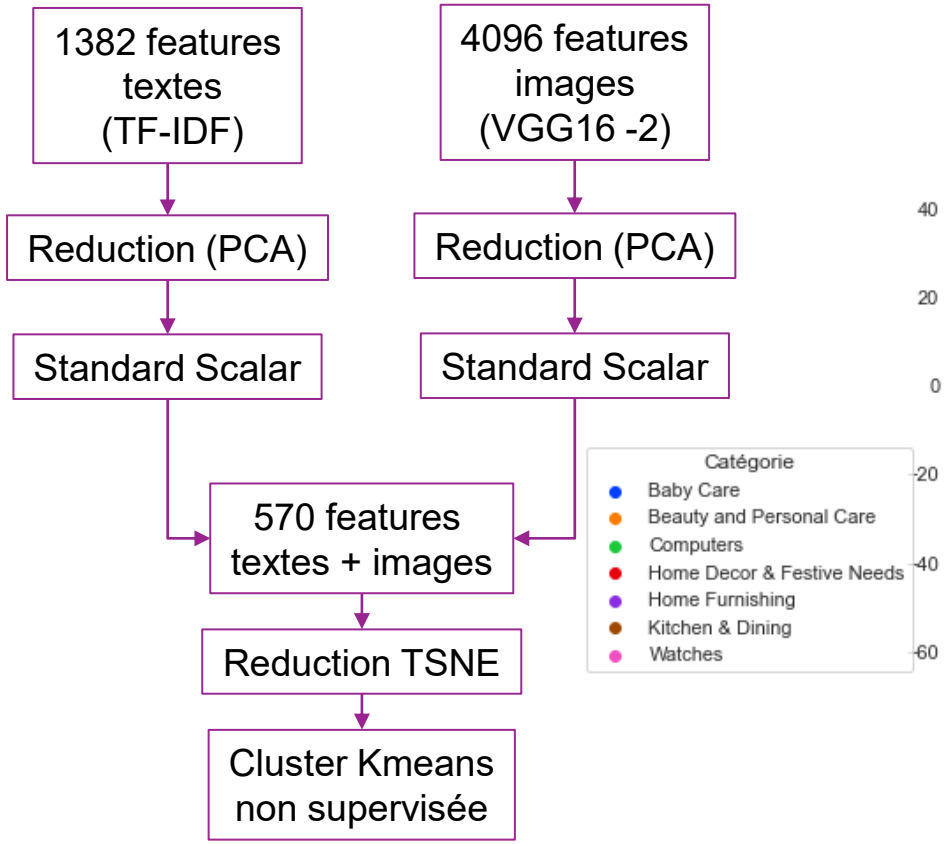
04 Texte + Images

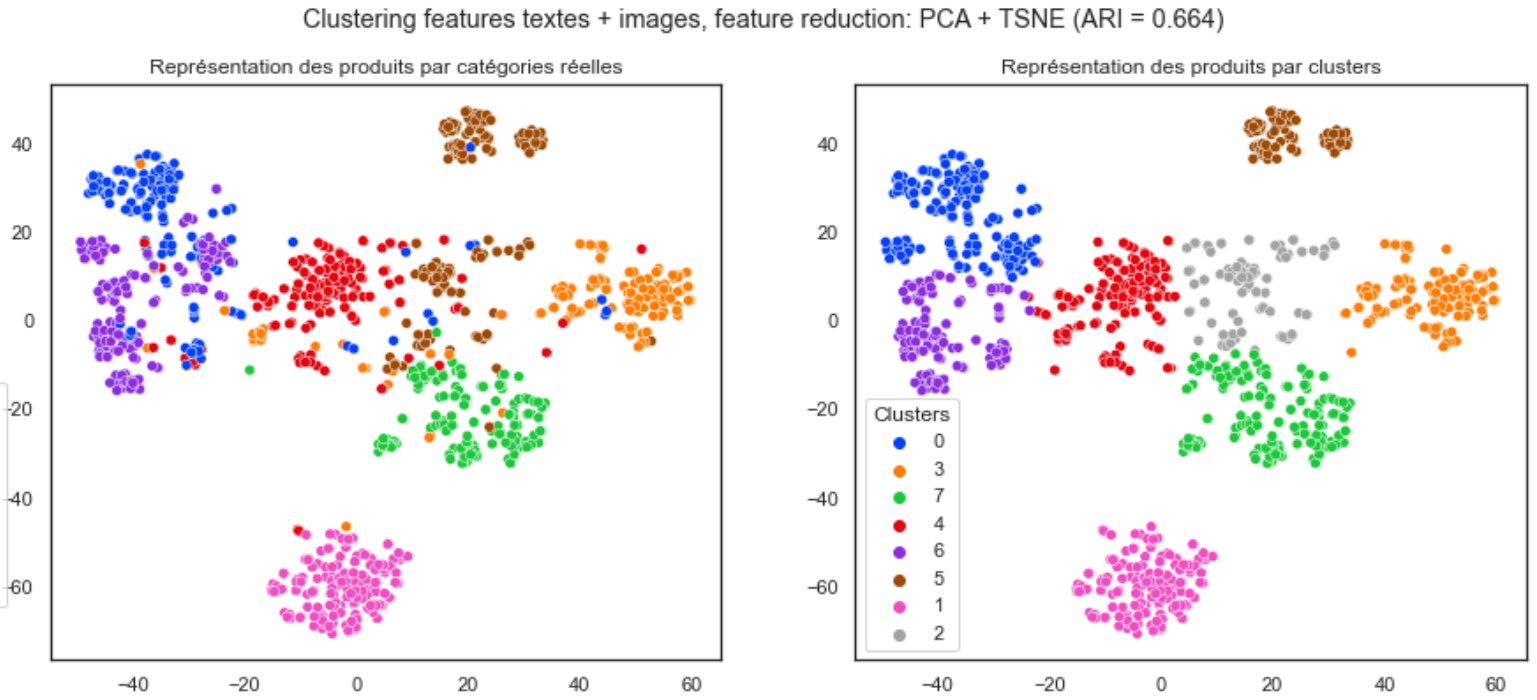
Clustering non-supervisée sur features des textes et images

ARI = 0.70

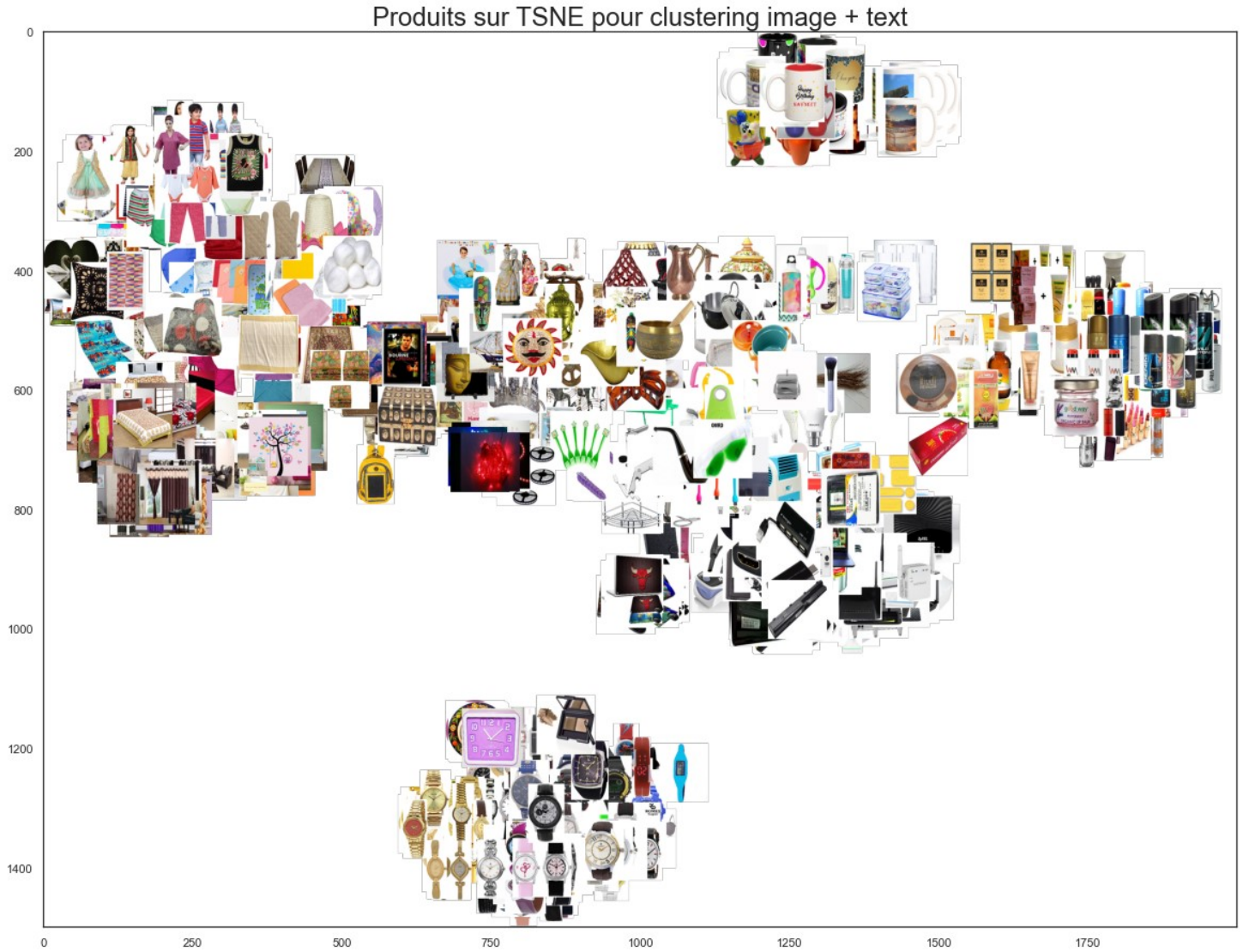
K = 8

(attribution des clusters 2 et 5 à Kitchen & Dining)



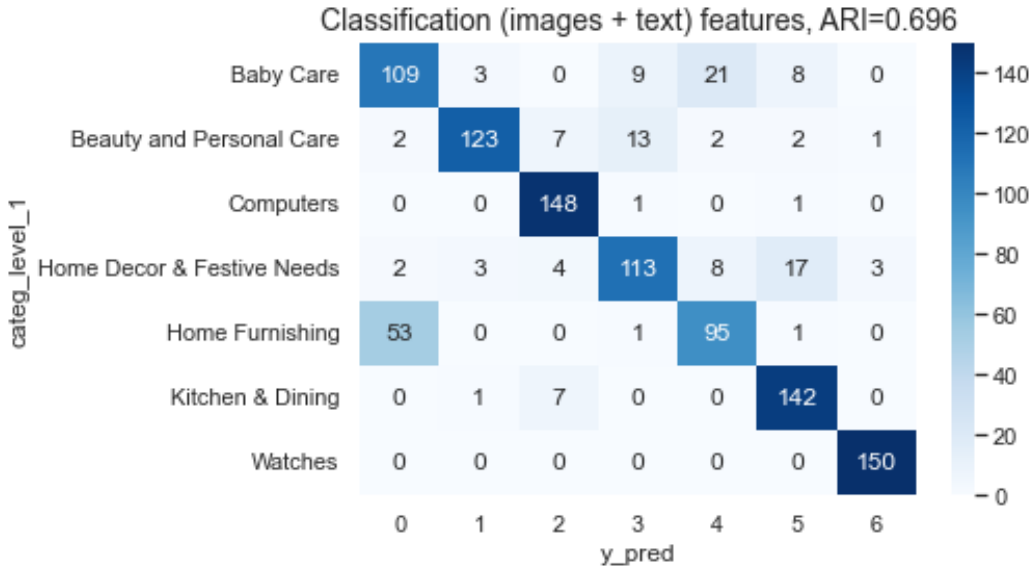


Clustering Images + Textes



- ARI = 0.70
- Accuracy = 0.84

	precision	recall	f1-score	support
Baby Care	0.66	0.73	0.69	150
Beauty and Personal Care	0.95	0.82	0.88	150
Computers	0.89	0.99	0.93	150
Home Decor & Festive Needs	0.82	0.75	0.78	150
Home Furnishing	0.75	0.63	0.69	150
Kitchen & Dining	0.83	0.94	0.88	150
Watches	0.97	1.00	0.99	150



05 Conclusion et améliorations à faire

Conclusions

Classification (Non/Semi)-Supervisée

- Meilleur segmentation en 7 groupes de produits
- Catégories « Watches » et « Mugs » très distincts – facile à classifier
- Catégories « Baby Care » et « Home Furnishing » contient produits similaires

Meilleur model	ARI	Accuracy
Textes : TF-IDF	0.55	0.77
Images : VGG16 pré-entrainé, features de moins 2 couches	0.53	0.76
Textes + Images (combinaison des features des 2 models ci-dessus)	0.70	0.84

Classification Supervisée

- Utiliser la classification supervisée pour proposer des catégories quand la classification par image ou par texte n'est pas prédit avec confiance

Meilleur model	ARI	Accuracy
Textes : CNN sur TF-IDF vectors	0.86	0.94
Images : VGG16 Transfer Learning + Fine tune	0.50	0.73

Améliorations à faire

Textes

- Prétraitement – supprimer les descriptions en double

- Classification Supervisée
  - WordEmbedding avec Universal Sequence Encoder ?

Images

- Regularization
  - Augmentation des images
  - Plus d'echantillons

- Transfer Learning
  - Tuning des hyperparamètres
  - Optimisation de choix des couches

Textes + Images

- Construction d'un model CNN pour classification supervisée

Questions

images: Mark Creasey

- mrcreasey@gmail.com

- Merci !