

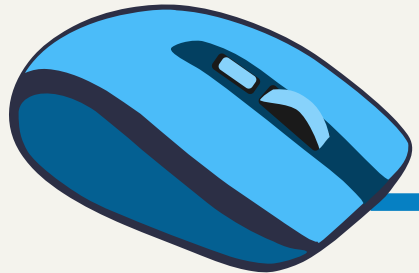


# CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

OPENCLASSROOMS

Victoire MOHEBI  
Mai 2022

# AGENDA



**Problématique**



**Mission**



**Analyse exploratoire**



**Modélisation**



**Conclusion & Piste de recherche**

# PROBLÉMATIQUE



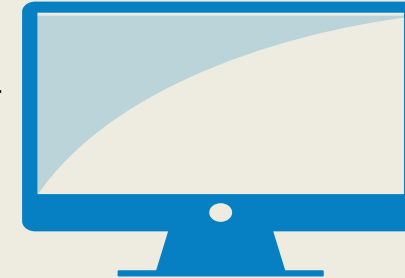
- Une plateforme d'e-commerce propose des produits à la vente
- Les données des produits incluent des descriptions textuelles et des images
- Catégories déjà renseignées pour un petit volume de produits mais le volume de produit non catégorisés est destiné à s'accroître

Est-il possible d'automatiser la classification des produits?

# MISSION

➤ **Travailler sur une base de données de 1050 produits**

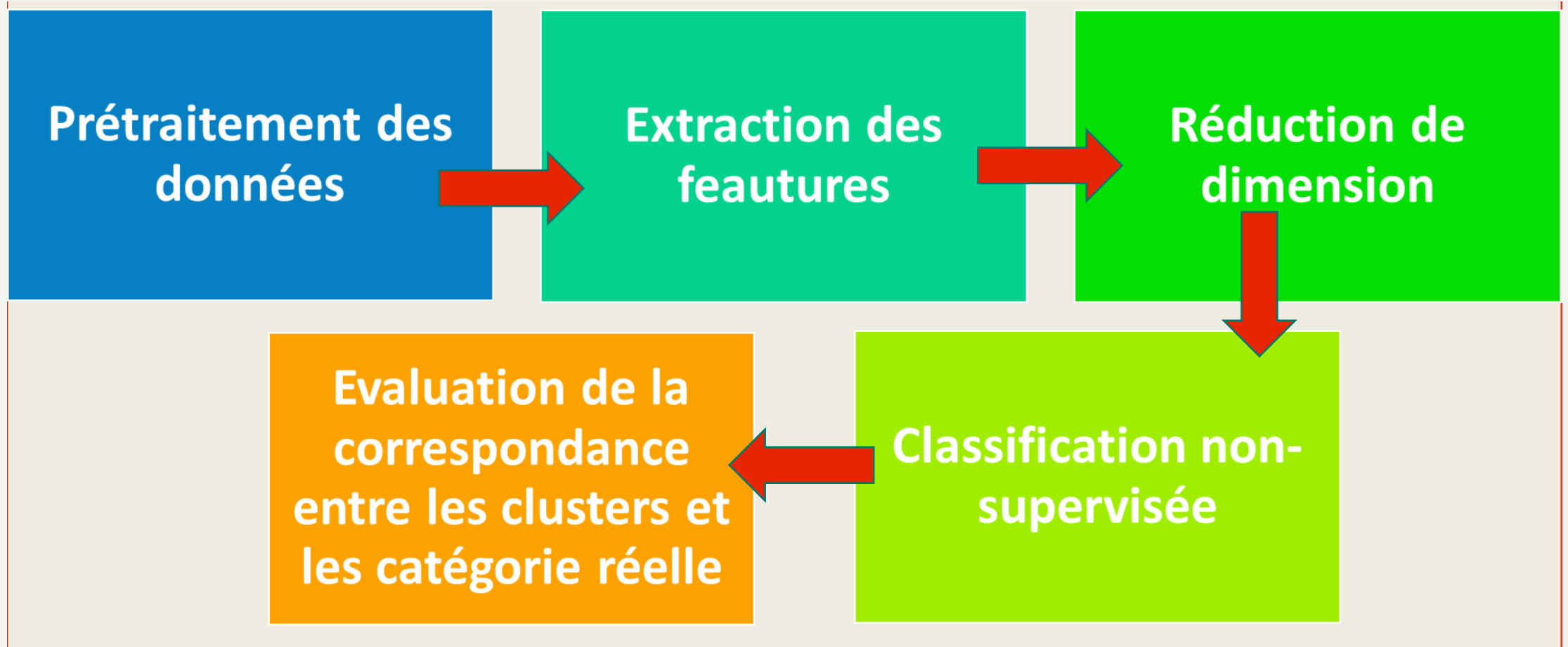
➤ **Classifier des produits de manière non-supervisée**



➤ **Evaluer le niveau de précision**

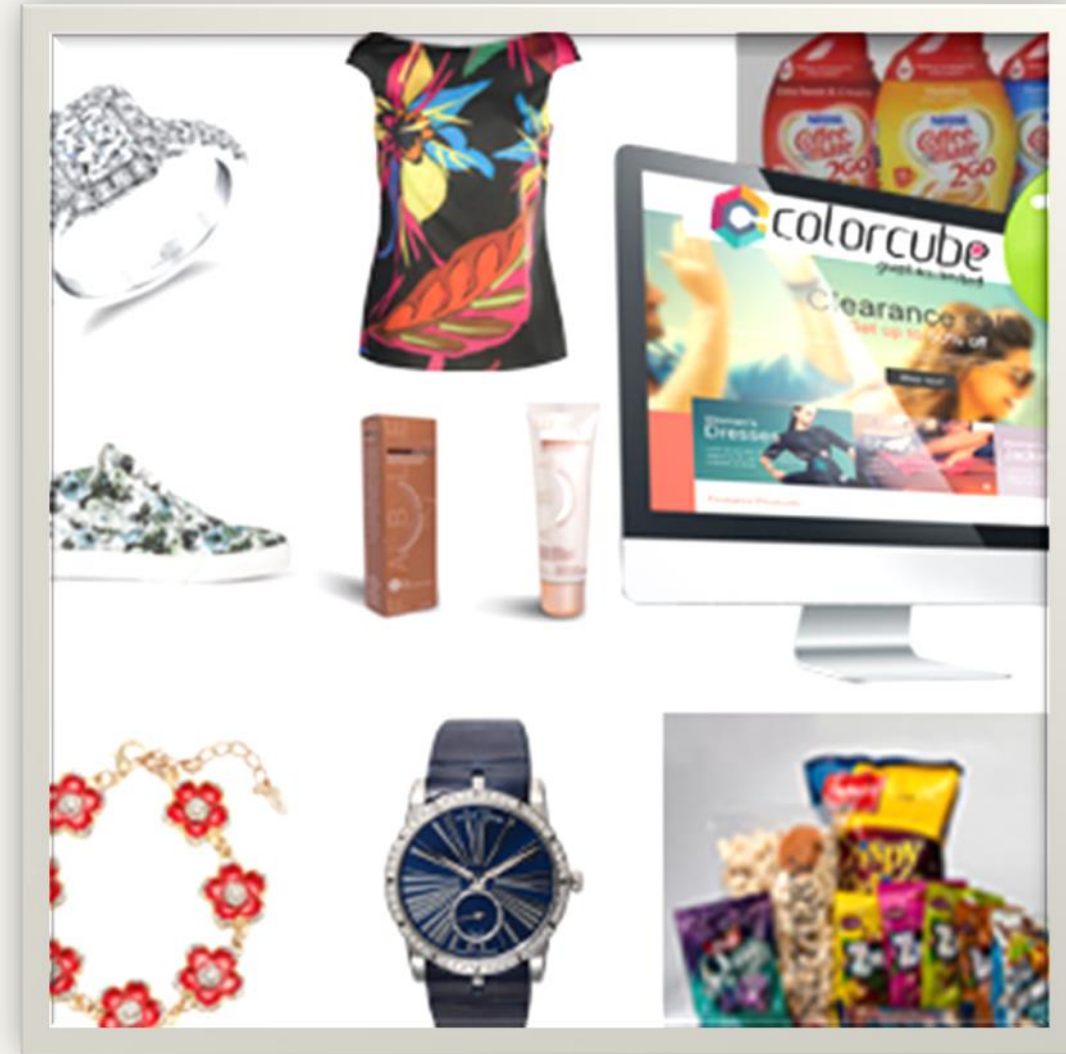
➤ **Fournir une représentation 2D des données pour illustrer les résultats**

# DÉMARCHE



# DONNÉES

- **Textuelles** : descriptions et noms des produits, de longueurs variables
- **Visuelles** : une image par produit, résolution et dimension variables



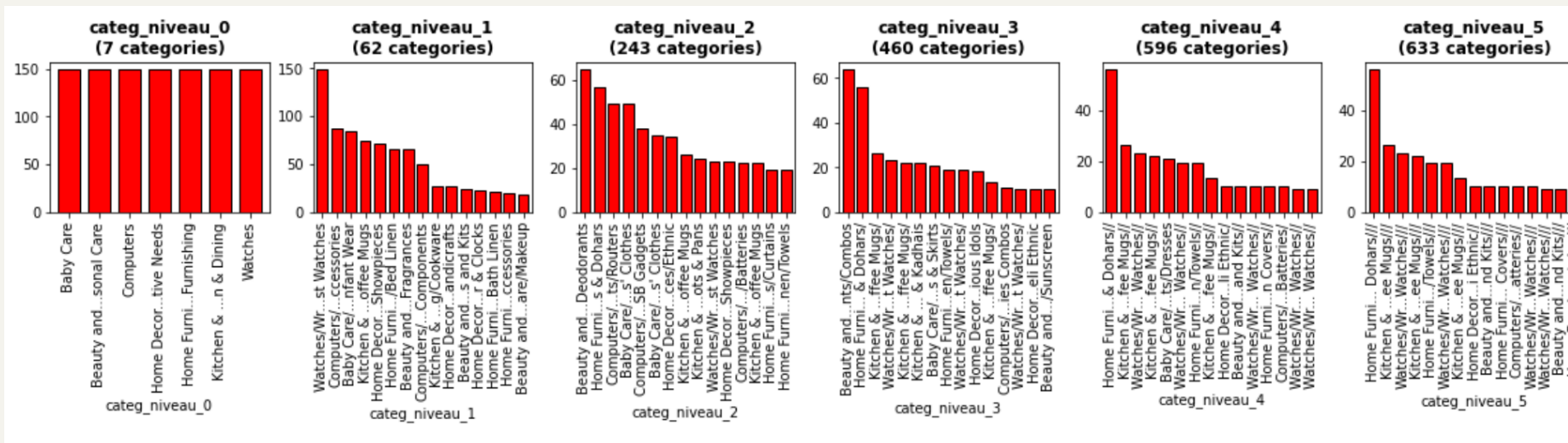
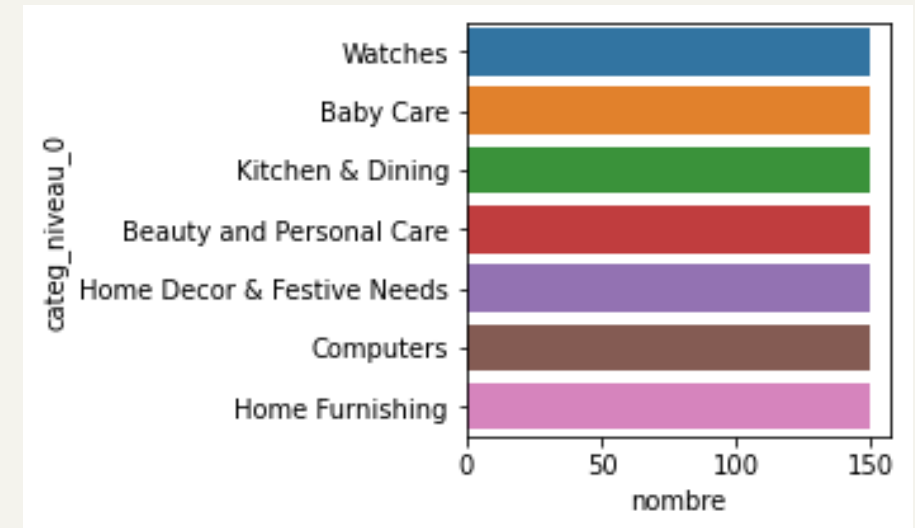
# DONNÉES TEXTUELLES





# NETTOYAGE DE DONNÉES

- 7 categories principales
- 6 sous catégories
- Choix du niveau\_0 des catégories
- 150 produits par catégories





# NORMALISATION DES DONNÉES TEXTUELLES

➤ Supprimer les ponctuation et les chiffres

➤ Supprimer les espaces blancs multiples

➤ Convertir en une seule casse

➤ Tokenization

➤ Supprimer les stopwords

➤ Lemmatization / Stemming

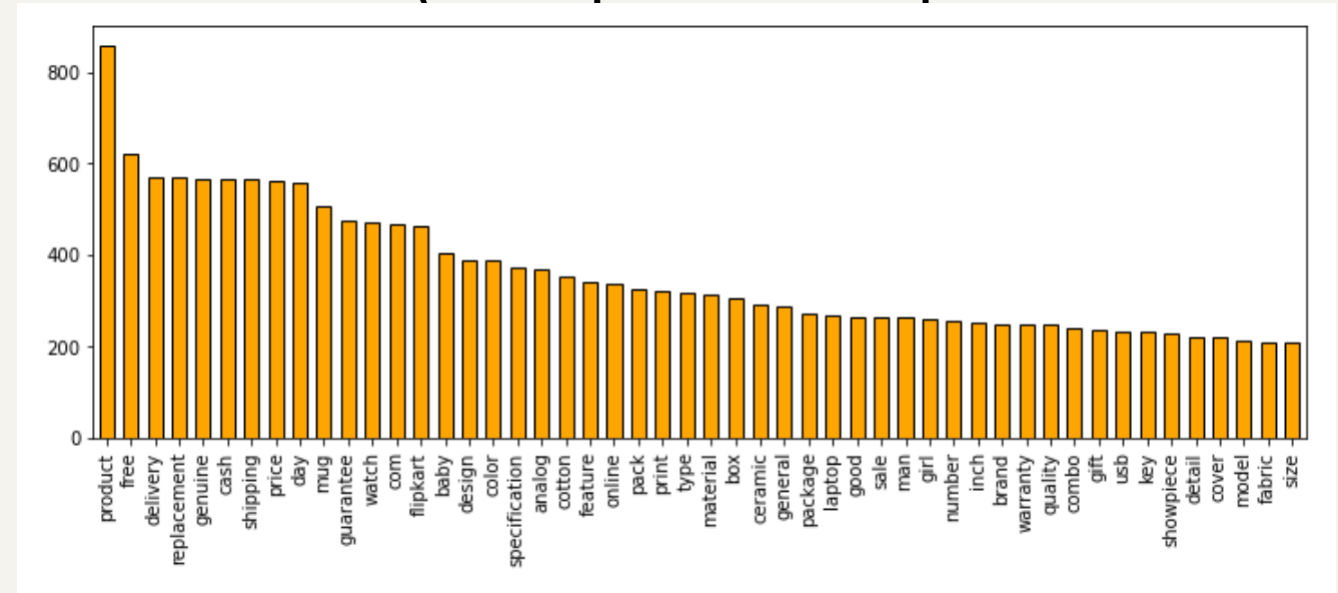
➤ POS (Part-Of-Speech) Tagging pour conserver certains tags



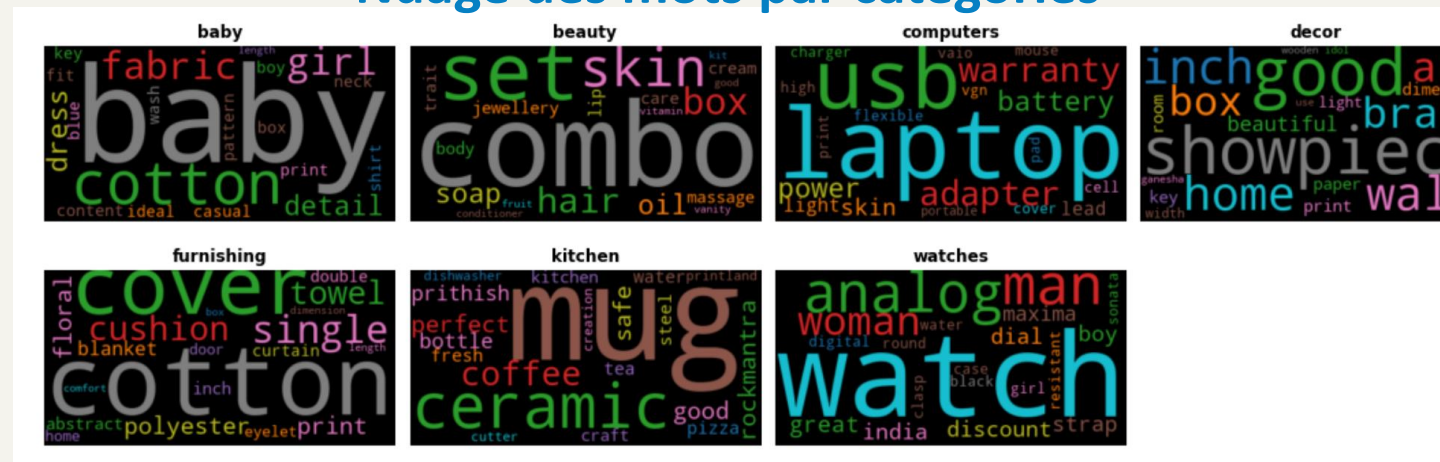
# ANALYSE EXPLORATOIRE DES DONNÉES TEXTUELLES

### Top 50 mot les plus fréquent dans le corpus (nom de produit et description

- Les mots les plus fréquents comme « *product* », « *set* » sont des « bruits »
- Ces mots sont les *stopword de domaine* : les mots non significatifs



## Nuage des mots par catégories



# EXTRACTION DES FEATURES

## Représentation vectorielle du texte

- Limiter au 7 clusters qui correspondent aux 7 catégories
- Evaluer le model en comparant les clusters avec les catégories réelles
- Score : *ARI (Adjusted Rand Index)*
- Algorithmes choisies :
  - Kmeans
  - Latent Dirichlet allocation (Topic modeling)



# MÉTHODE UTILISÉE POUR LA REPRÉSENTATION VECTORIELLE DU TEXTE

**BoW**

**BagOfWords**

Comptage simple du nombre d'apparition du mot

**TFIDF**

**TFIDF**

Fréquence du mot dans le document par rapport à sa fréquence dans le corpus

**W2V**

**Word2Vec**

Technique de word embedding basée sur réseau de neurone à deux couches

**USE**

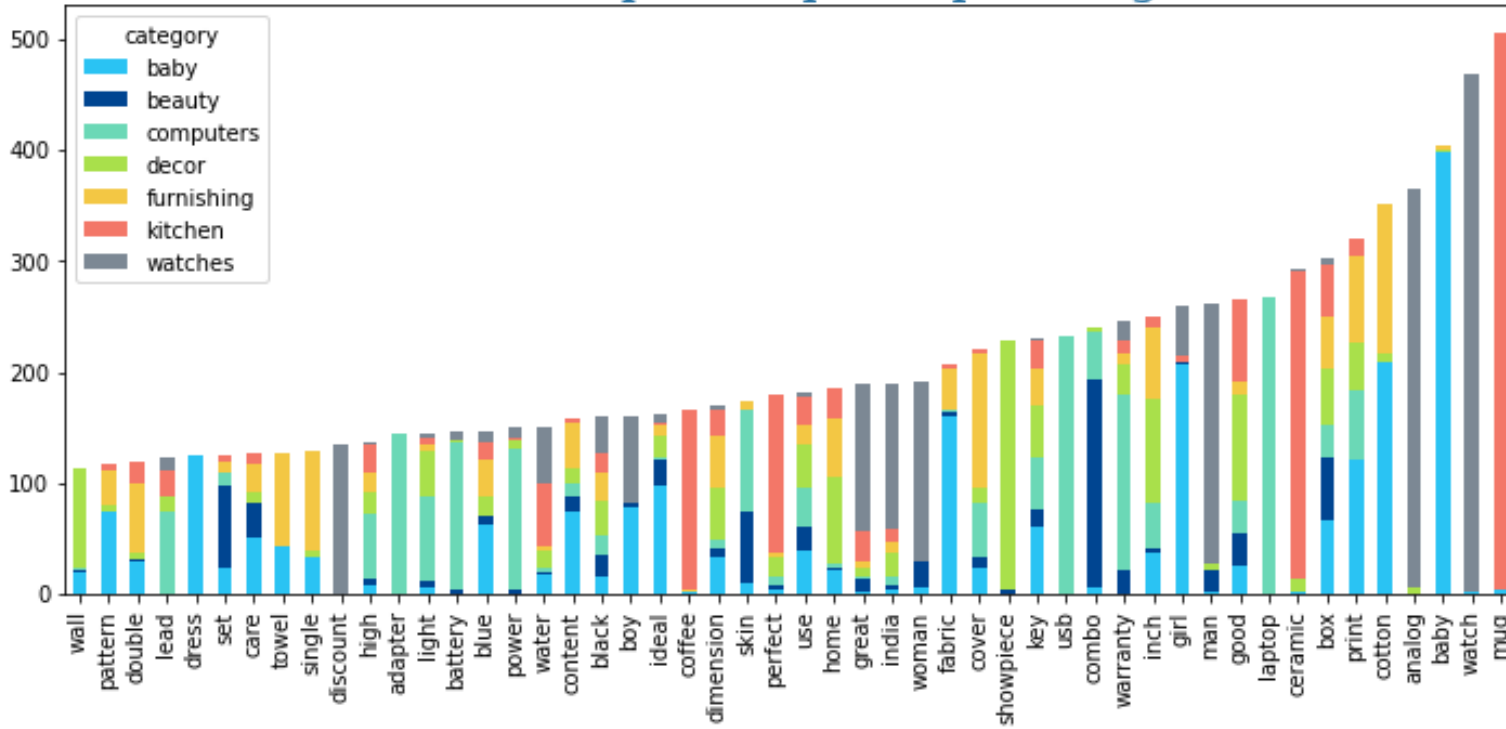
**Universal Sentence Encoder**

Réseaux de neurones pré-entraînés



# Document-term matrix :BAG OF WORDS

Les mots les plus fréquents par catégorie



- Des mots communs entre certains des catégories
- Ces mots peuvent engendrer des bruits pour le modèle

# BAG OF WORDS

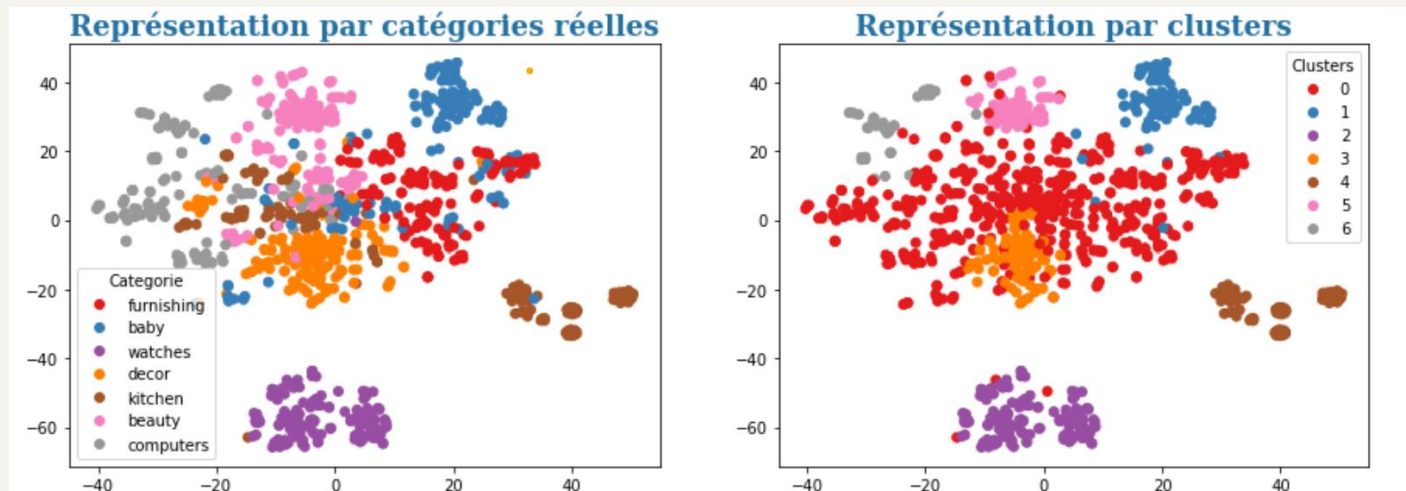
- Réduction nombre de vecteurs de 752 à 364 par SVD en conservant plus de 0.99% de variance
- Implémenter *kmeans* (k=7)

ARI = 0.27

Correspondance entre les catégories réelle et les cluster

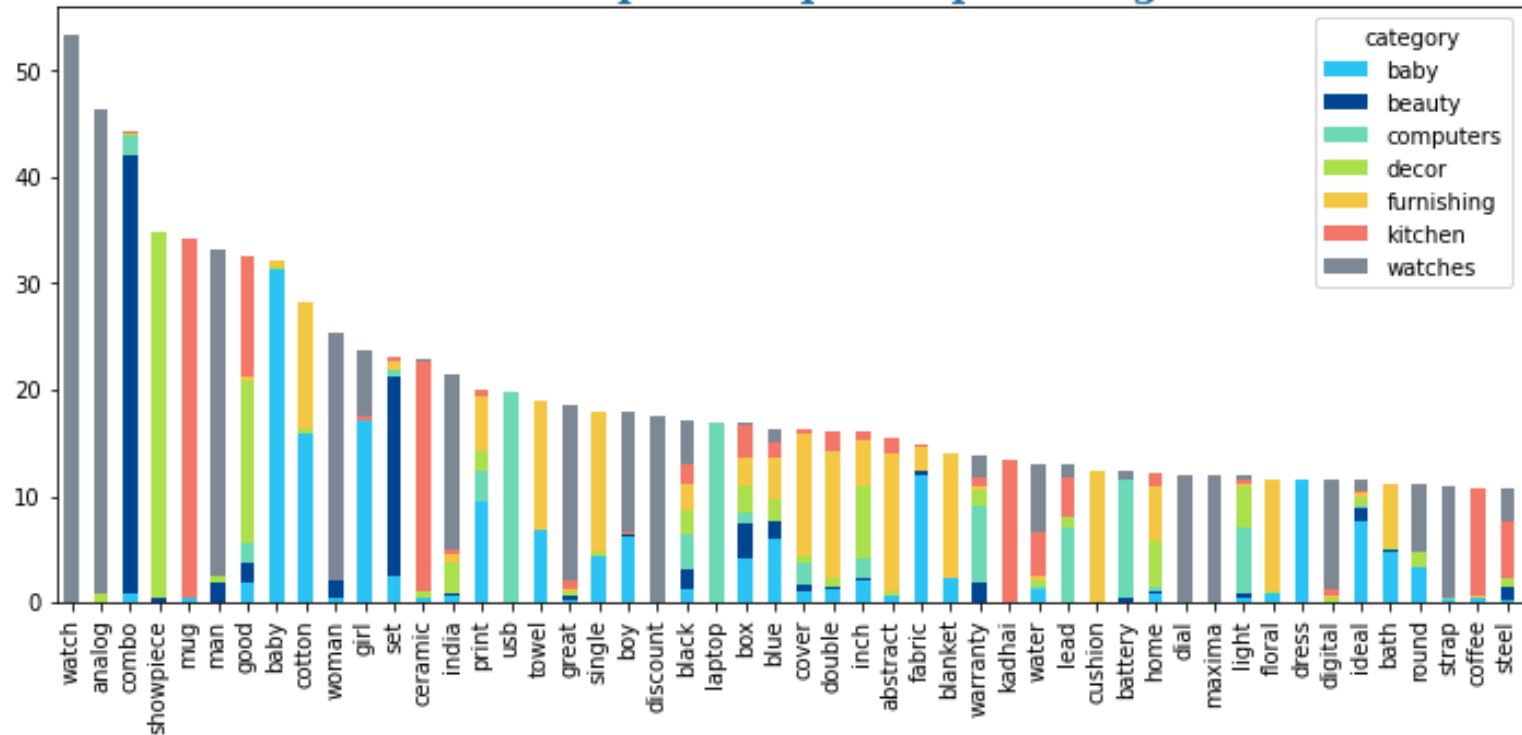


Représentation 2D



# Document-term matrix : TFIDF

Les mots les plus fréquents par catégorie



- Des mots communs entre certains des catégories sont moins nombreux que dans le document\_term matrix de BoW
- Cela peuvent engendrer des bruits pour le modèle

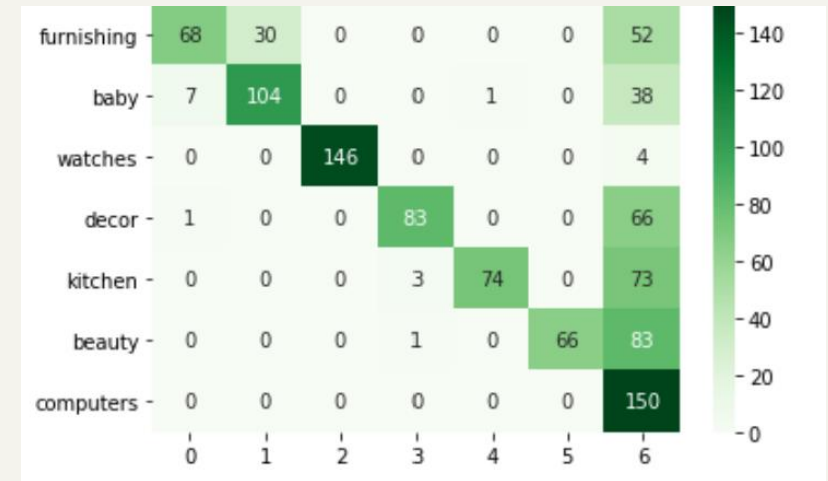


# TFIDF

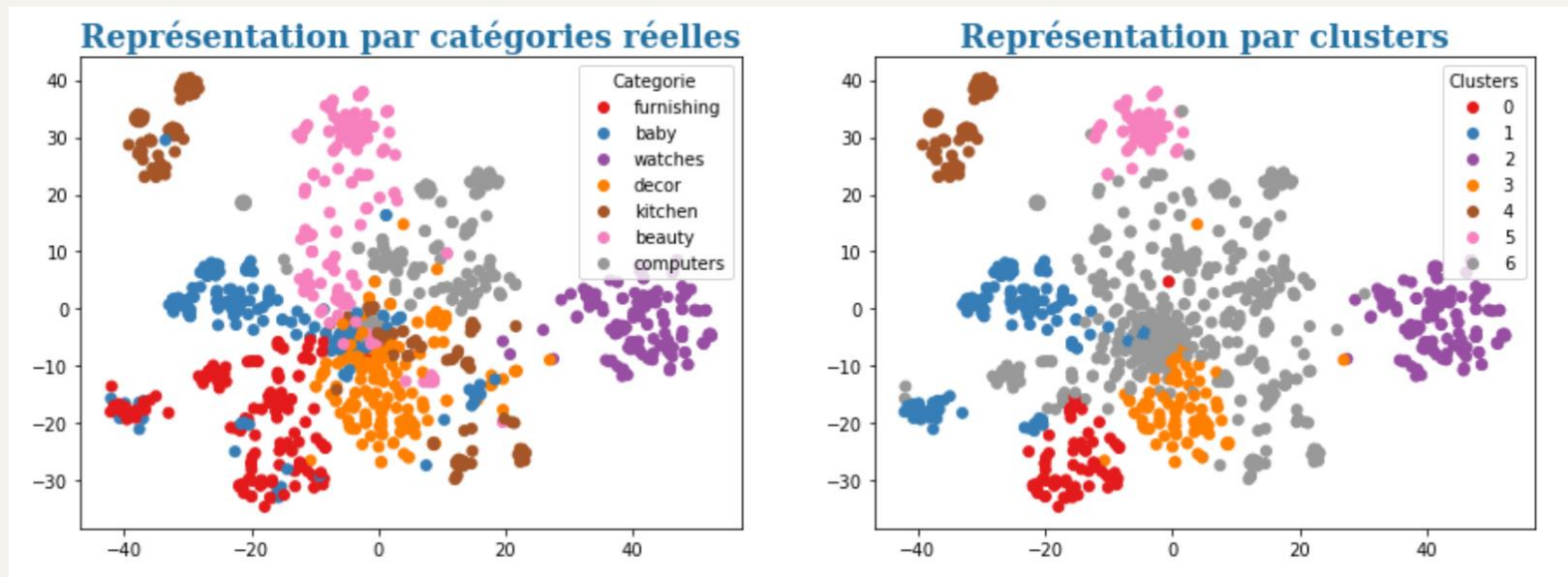
## Correspondance entre les catégories réelles et les clusters

- Réduction de nombre de vectors de 752 à 429 par SVD en conservant plus de 0.99% de variance
- Implémenter *kmeans* ( $k=7$ )

ARI = 0.32



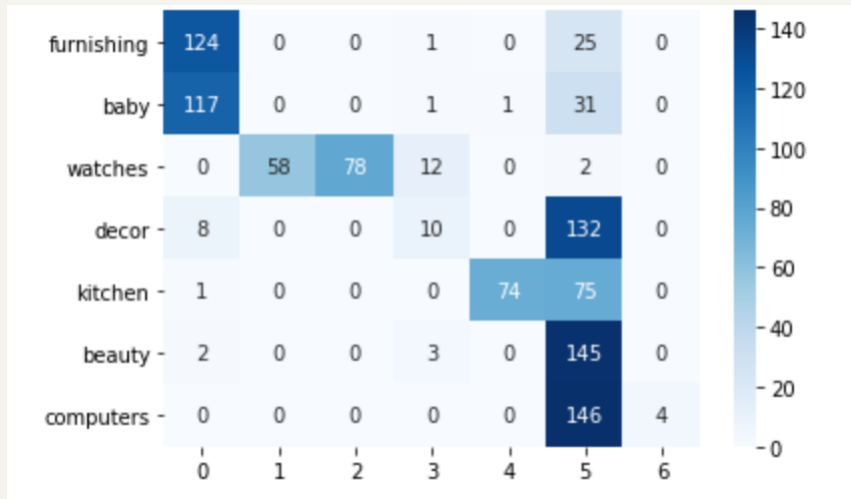
## Représentation 2D



# WORD2VEC

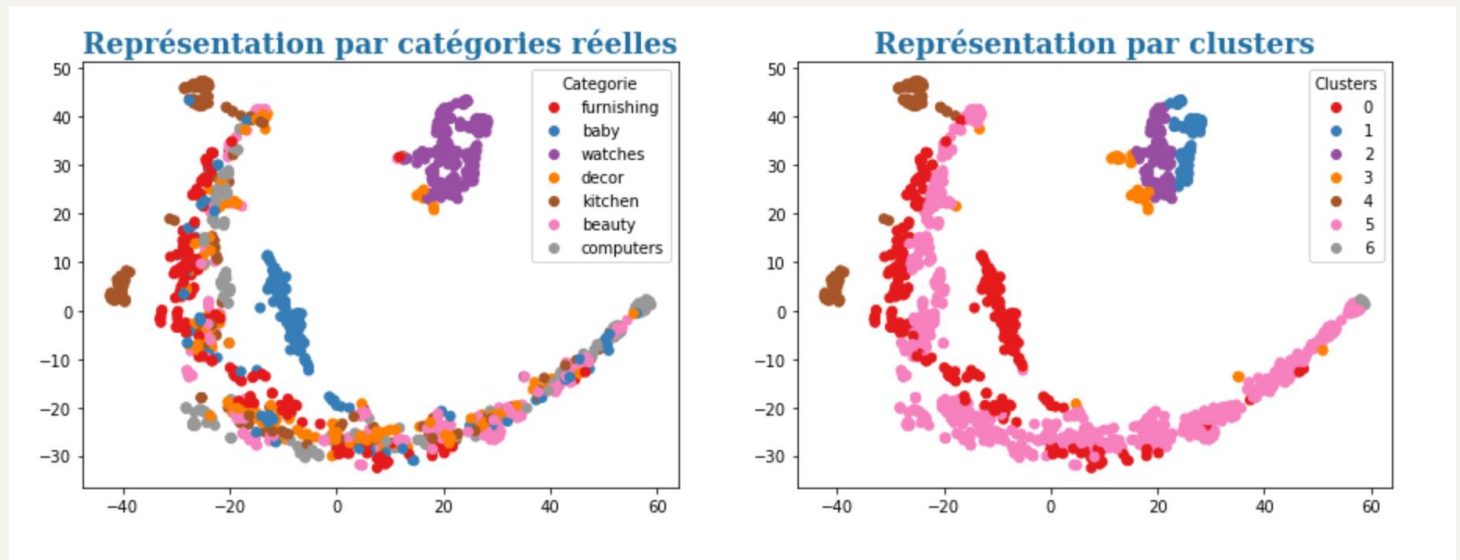
## Sans réduction de dimension

Correspondance entre les catégories réelle et les cluster



ARI : 0.12

Représentation 2D



# Topic Modeling avec LDA

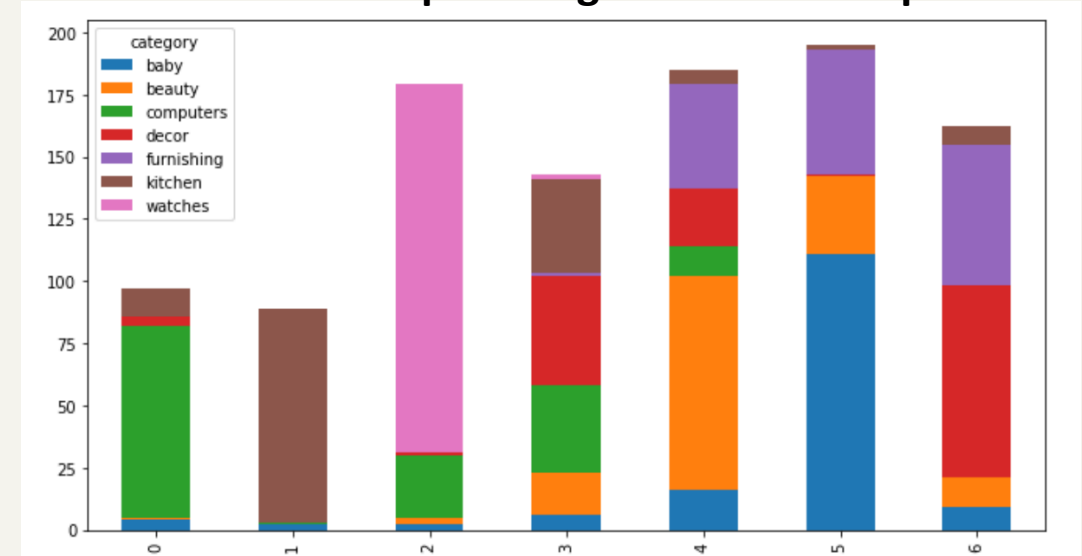
(Matrice BagOfWord)

## Correspondance entre les catégories réelle et les cluster



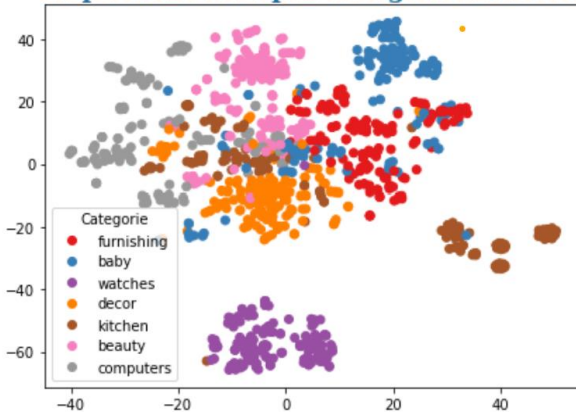
ARI = 0.37

## Part de chaque catégorie dans les topics

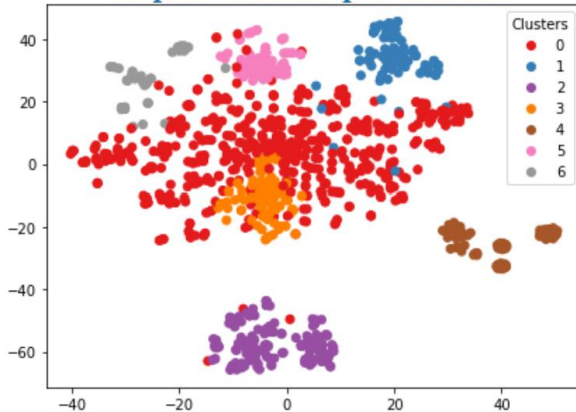


## Représentation 2D

### Représentation par catégories réelles



### Représentation par clusters



Topic #0: usb adapter power light lead warranty laptop charger bottle vgn vaio portable fan flexible smartpro  
Topic #1: mug ceramic coffee perfect rockmantra safe tea prithish pizza creation kitchen printland cutter dishwasher fresh  
Topic #2: watch analog man laptop woman india great discount battery dial strap boy digital maxima resistant  
Topic #3: box use wall warranty bowl showpiece key dimension glass brass clean surface beautiful art place  
Topic #4: skin combo inch print sticker laptop pad mouse shape cover easy set warranty vinyl wall  
Topic #5: baby cotton girl fabric polyester eyelet comfort print ideal curtain aroma towel blue dress box  
Topic #6: showpiece single good home blanket abstract wallmantra steel double stainless piece pot art quilt statue

- - Le topic 0 peut correspondre à « *computer* »
- - Le topic 1 peut correspondre à « *kitchen* »
- - Le topic 3 peut correspondre à « *watch* »
- - Le topic 6 peut correspondre à « *décor* »
- - Les autres topics ne sont pas clairement identifiés

# UNIVERSAL SENTENCE ENCODER (USE)

Correspondance entre les catégories réelle et les clusters

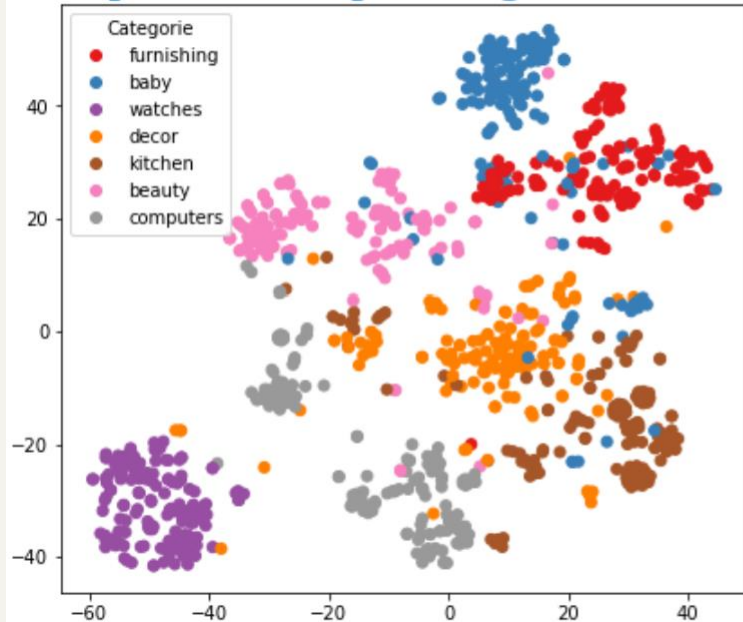
- Réduction nombre de vectors de 512 à 359 par SVD en conservant plus de 0.99% de variance
- Implémenter *kmeans* ( $k=7$ )

ARI = 0.49

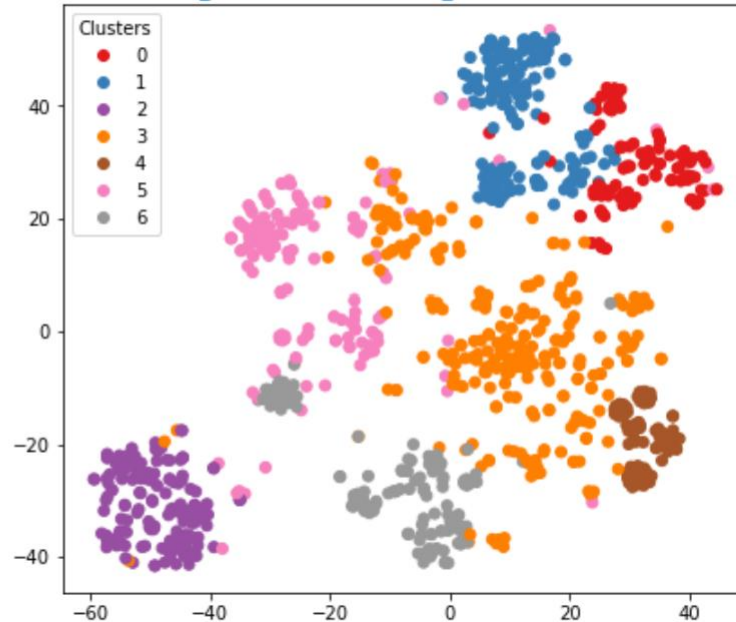
furnishing	98	46	0	2	0	4	0
baby	7	108	0	27	1	6	1
watches	0	0	145	2	0	3	0
decor	0	1	2	122	0	23	2
kitchen	0	0	0	61	74	14	1
beauty	0	2	0	59	0	87	2
computers	0	0	0	4	0	23	123
	0	1	2	3	4	5	6

Représentation 2D

Représentation par catégories réelles

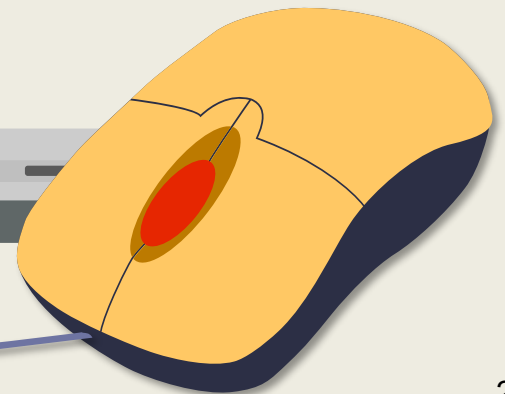


Représentation par clusters



# DONNÉES VISUELLES

Place de marché



# DÉMARCHE



**Prétraitement des images**

1

**Création d'une liste des descripteurs**


2

**Création des clusters des descripteurs**

3

**Création des factures des images : BOVW**

4



**Evaluer la similarité entre catégories réelles et les clusters**

7

**Création des clusters à partir de T-SNE**

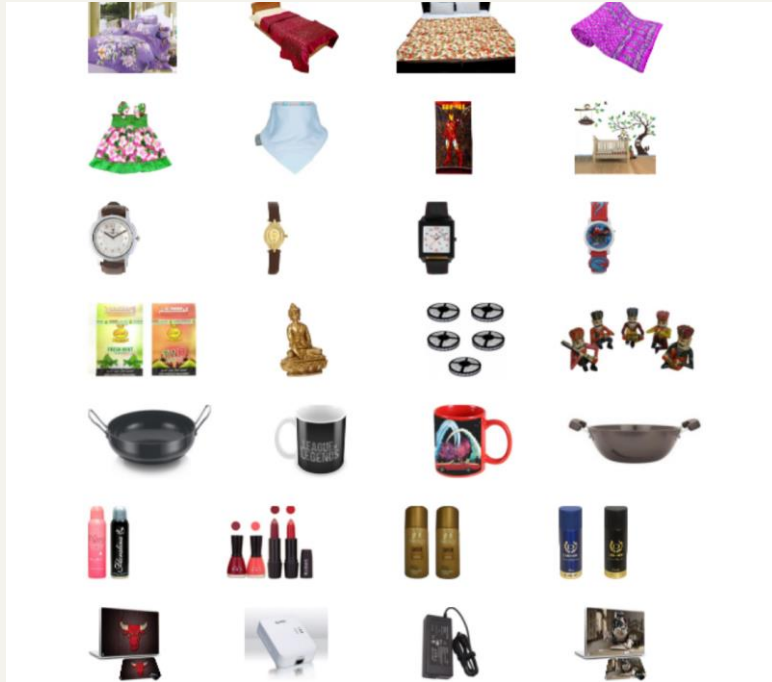
6

**Réduction de la dimensionnalité  
PCA , T-SNE**

5

# TRAITEMENT DES IMAGES

## Exemple des image par catégorie



## Prétraitements effectués

- Grayscale
- Redimensionner
- Ajuster le contraste (CLAHE)
- Egaliser les histogrammes
- Réduire les bruits avec filtrage gaussien (sigma = 0.5)

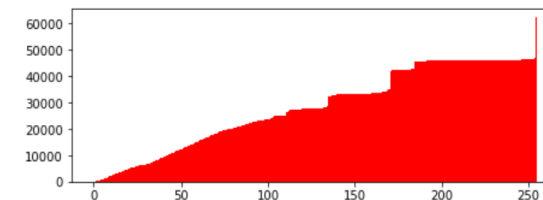
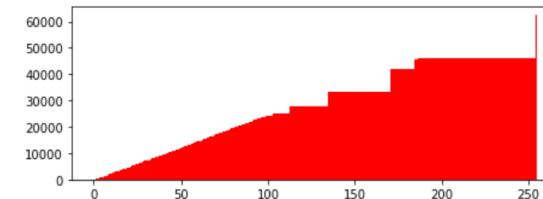
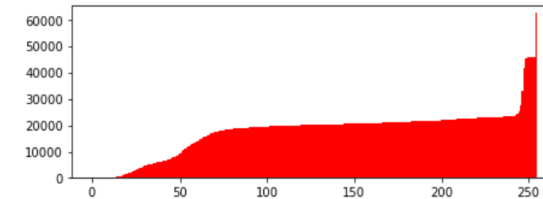
Adjustement des contrast



Equalized



filtre gaussien

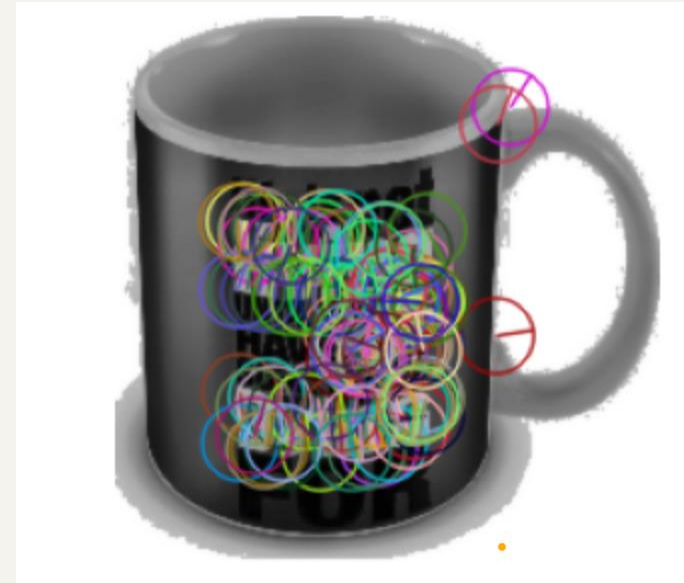




# DESCRIPTEURS DE CHAQUE IMAGE

Algorithme « ORB » (Oriented FAST and rotated BRIEF)

## Illustration des descripteur d'une image



On obtient 368902  
descripteurs pour 1050  
images

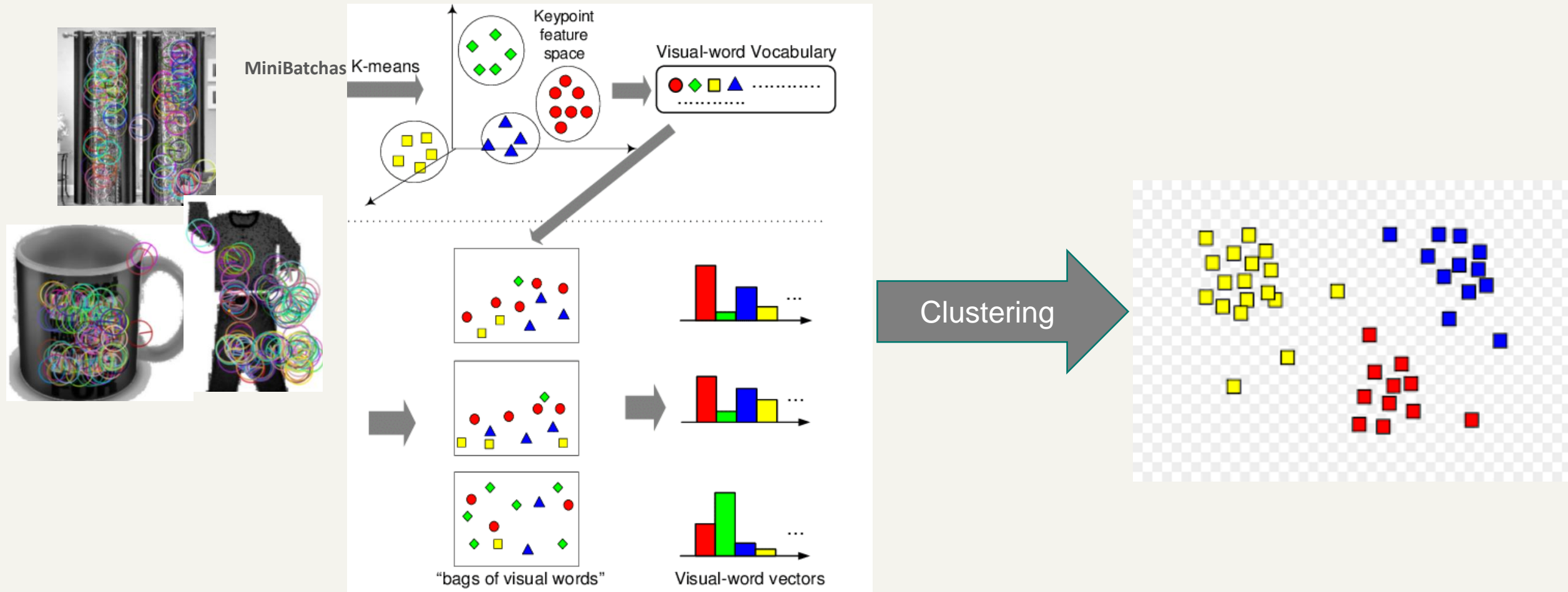
Environ 351  
descripteurs par image

Chaque descripteur est  
un vecteur de longueur  
de 32

Le nombre des  
descripteurs des  
images ne sont  
pas identique

# REGROUPEMENT DES DESCRIPTEURS

## Création de « Bag Of Visual Words »



# CRÉATION DES CLUSTERS À PARTIR DE T-SNE

- Réduire les vecteurs à 383 dimensions par PCA en conservant 95% de la variance
- Réduction TSNE en 2D

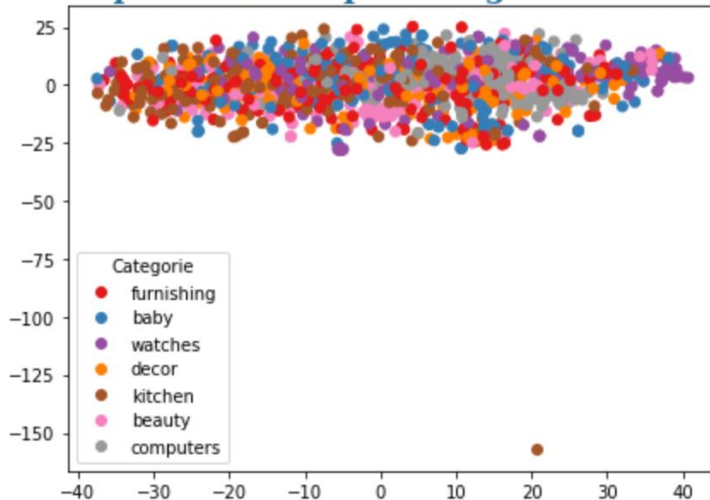
ARI : 0.04

Correspondance entre les catégories réelles et les clusters

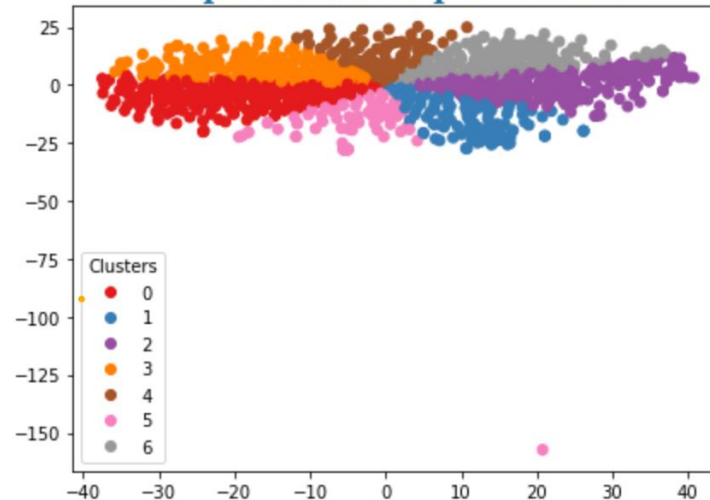
furnishing	41	22	23	26	13	12	13
baby	26	30	18	29	22	7	18
watches	16	9	55	21	9	14	26
decor	39	27	33	30	7	5	9
kitchen	56	12	7	28	24	14	9
beauty	41	9	23	19	5	31	22
computers	2	11	57	13	10	5	52

Représentation en 2D

Représentation par catégories réelles



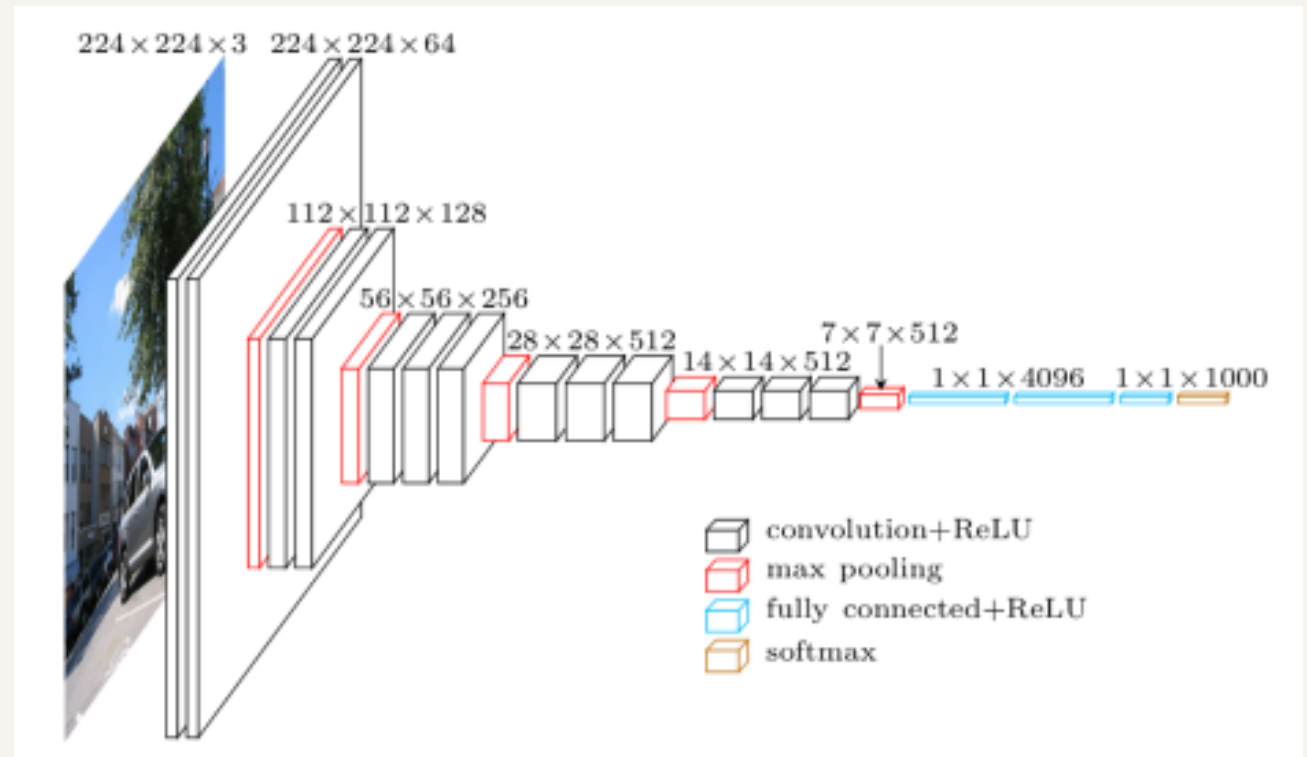
Représentation par clusters



# DEEP LEARNING : UTILISATION D'UN MODÈLE PRÉENTRAÎNÉ

- Utilisation d'un réseaux de neurones pré-entraîné sur la base de données ImageNet
- Supprimer la dernière couche « fully connected »
- pour l'extraction des feautres
- Bag-of-visual-words de 25088 dimensions,
- Réduction à 915 dimensions par PCA (99% de variance).

Architecture d'un modèle VGG16



# CLUSTERING À PARTIR DE T-SNE

Correspondance entre les catégories réelle et les cluster

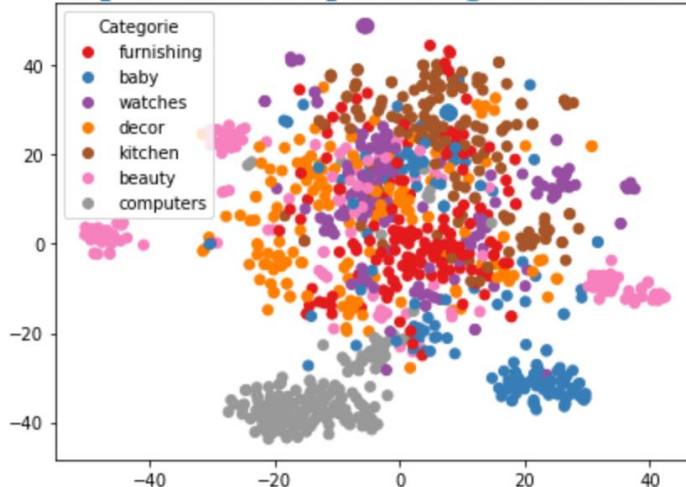
- Réduire les vecteurs de 25088 à 915 dimensions par PCA en conservant 99% de la variance
- Réduction en 2D par T-SNE

ARI\_ VGG16: 0.21

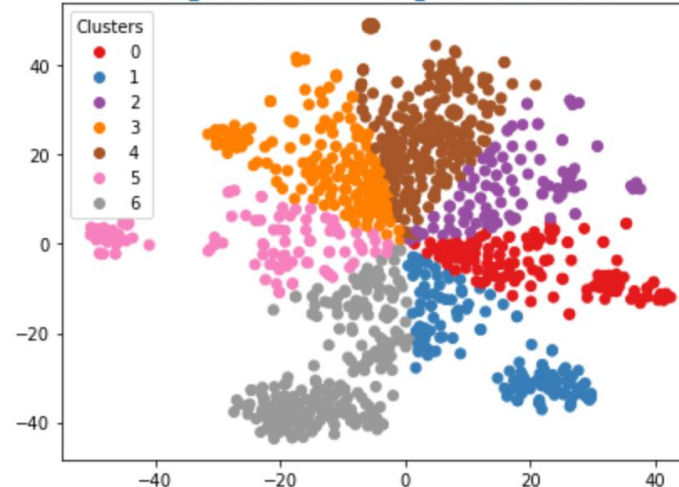
furnishing	31	20	22	15	30	10	22
baby	10	84	6	9	33	2	6
watches	11	13	32	43	28	11	12
decor	18	5	8	39	22	38	20
kitchen	23	0	32	13	79	3	0
beauty	43	7	1	39	7	44	9
computers	0	1	1	4	4	0	140
	0	1	2	3	4	5	6

Représentation 2D

Représentation par catégories réelles

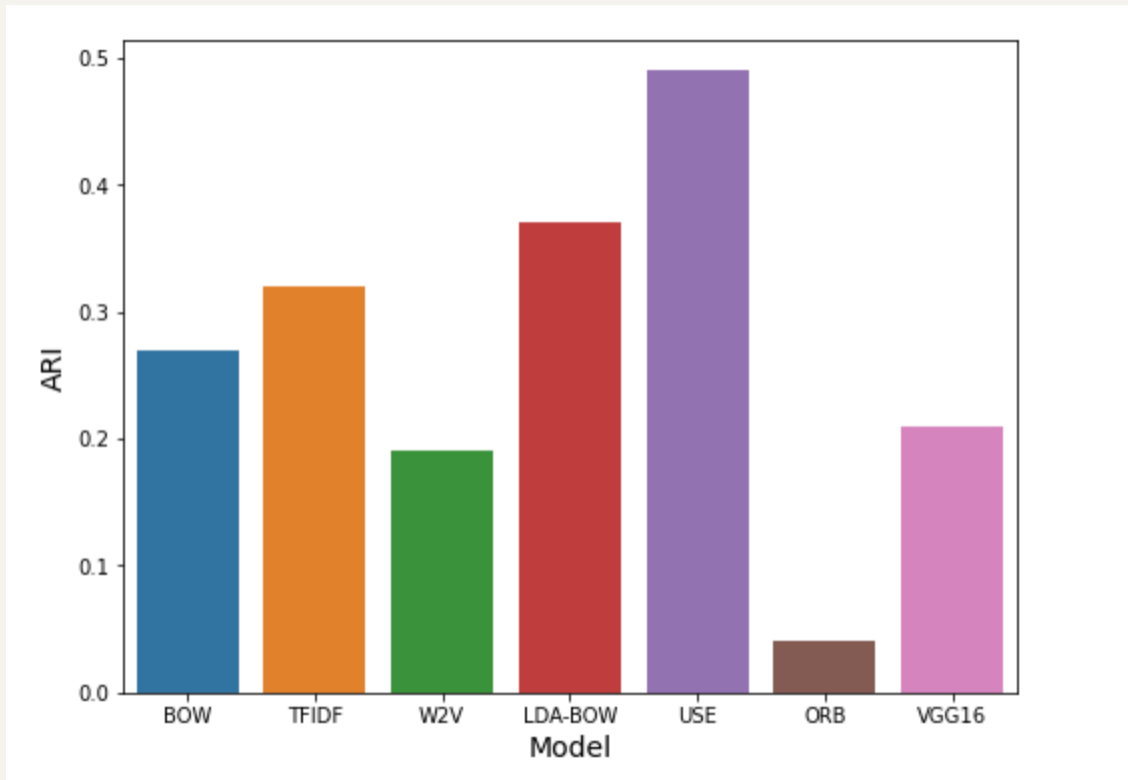


Représentation par clusters



# COMPARAISON DES MODÈLE

Le score ARI pour les modèles choisis



- Modèles basés sur les données textuelles ont les meilleurs
- Modèle Universal Sentence Encoder basée sur le *deep learning* a le meilleur score

# CONCLUSION

- **Analyse des données textuelles et visuelles**
- **Extraction des *features***
  - NLP** : BoW, WordEmbedding, encodage de phrase
  - Images** : Pixels bruts, *BoVW* (ORB), Extraction de *features* (CNN *Transfer-VGG16*)
- **Données textuelles nous permet de créer des modèle de clustering**
- **Identification des produit est difficile à catégoriser**
- **Cela est dû à la présence de termes communs entre certains des textes , lorsqu'ils sont vectorisés, obtiennent des valeurs égales pour certaines dimensions.**
- **Faisabilité de la classification automatique non-supervisé**
  - Possibilité d'atteindre un ARI jusqu'à 0.49 entre clusters et catégories réelle avec les données textuelles





# PISTE DE RECHERCHE



Fusionner les features  
extraites des données  
textuelles et visuelles

Effectuer une classification  
supervisée sur les features  
fusionnée en calculant la  
probabilité d'appartenance  
d'un produit à une catégorie

Pondérer la probabilité du  
volet NLP et du *computer  
vision* pour obtenir le  
meilleur score



**Merci de votre attention !**