Parcours Data Scientist | Projet 6

Classifier automatiquement des biens de consommation

Présenté par Fiacre KAKPO, le 11 sept 2021

Plan de travail détaillé

- ➔ Problématique et exploration du jeu de données
- ➔ Manipulation de données textuelles, applications des techniques de NLP pour prétraitement des données textuelles : stopwords, TF IDF / bag of words
- ➔ Réduction de dimension par LDA / NMF (données textuelles)
- ➔ Prétraitement d'images et Extraction de features (SIFT/SURF/ORB)
- ➔ Classification non supervisée (clustering des descripteurs)
- ➔ Réduction dimensionnelle avec PCA et Projection plane par t-SNE
- ➔ ARI de similarité catégories images / clusters
- ➔ Faisabilité

Problématique & Présentation du jeu de données

Problématique

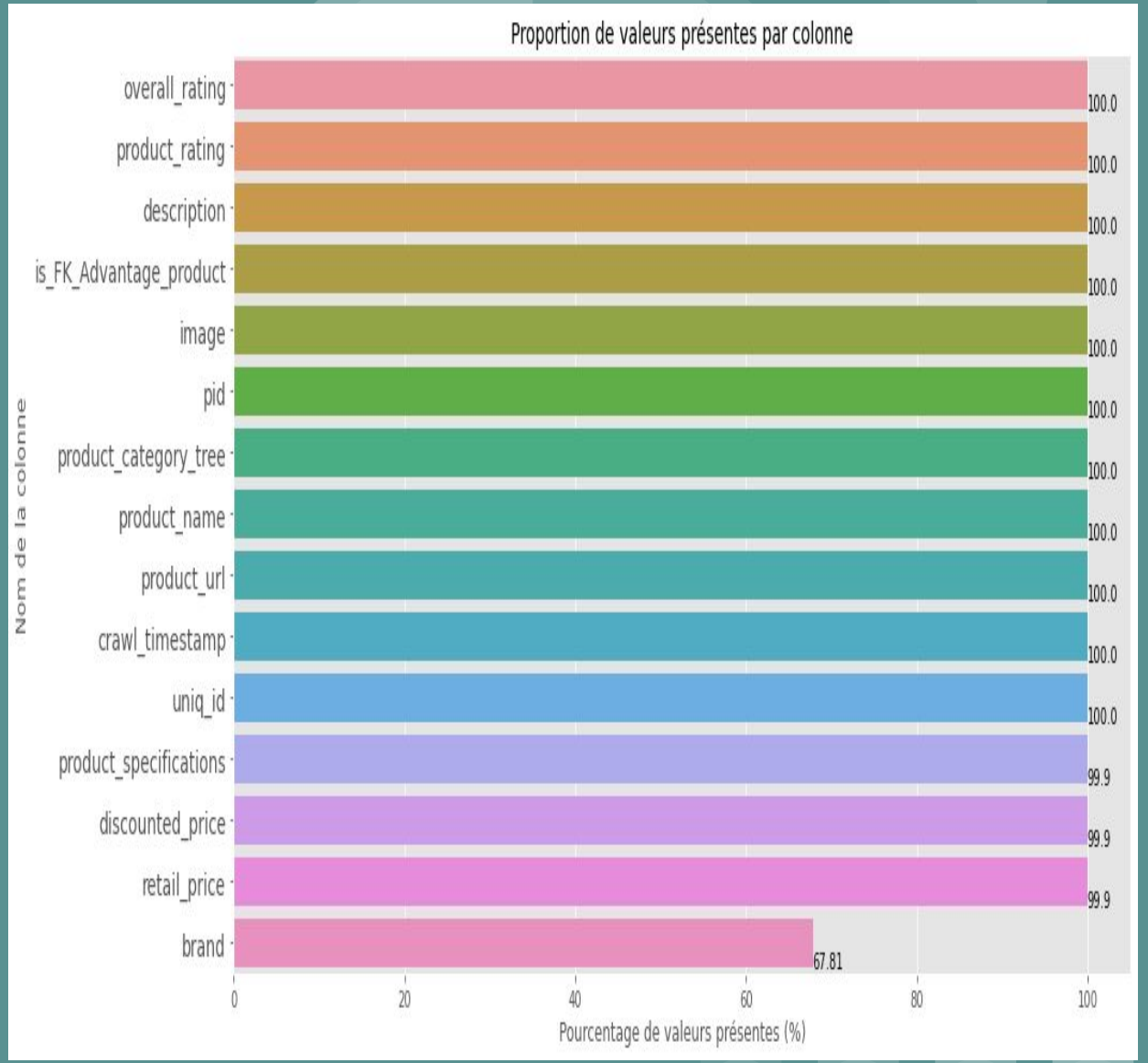
Place de marché : marketplace e-commerce

- → **vendeurs :** articles avec **photo et description**
- → **attribution manuelle** : fastidieuse et peu fiable
- → **Objectif :** automatiser l'attribution des catégories
- → **Perspectives :** faciliter la mise en ligne de nouveaux articles, faciliter la recherche de produits

Étude de faisabilité d'un moteur de classification

Présentation du jeu de données

- Stockées sur **Amazon S3** : lien de téléchargement
- Données du site d'e-commerce **flipkart.com**
- Données sur **1'050 produits**, et les photos associées
- Jeu de données au format **.csv** et images au format **.jpg**
- Jeu de données contenant **15 colonnes**
- Taille totale : environ **350 Mo**



Prétraitements : Données textuelles

Nettoyage du jeu de données

→ Suppression des champs peu pertinents

- *'crawl_timestamp'*
- *'product_url'*
- *'pid'*
- *'retail_price'*
- *'discounted_price'*
- *'is_FK_Advantage_product'*
- *'product_rating'*
- *'overall_rating'*
- *'product_specifications'*
- *Brand*
-

→ Vérification de l'intégrité des données

- Pas de « doublons »

Éclatement de la variable catégorie

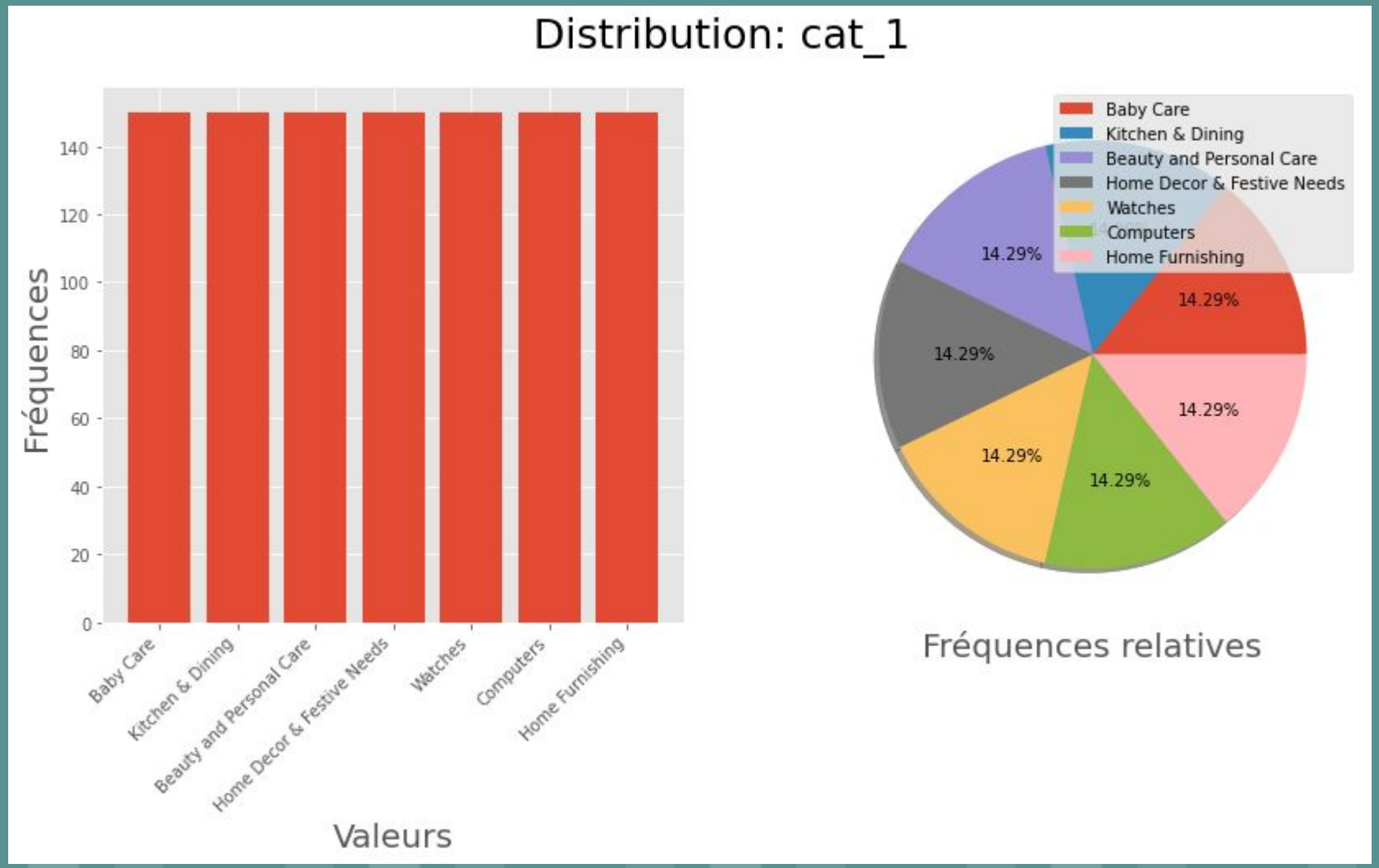
Nettoyage de la variable Catégorie en trois sous-catégorie

Exemple :

["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Y..."]

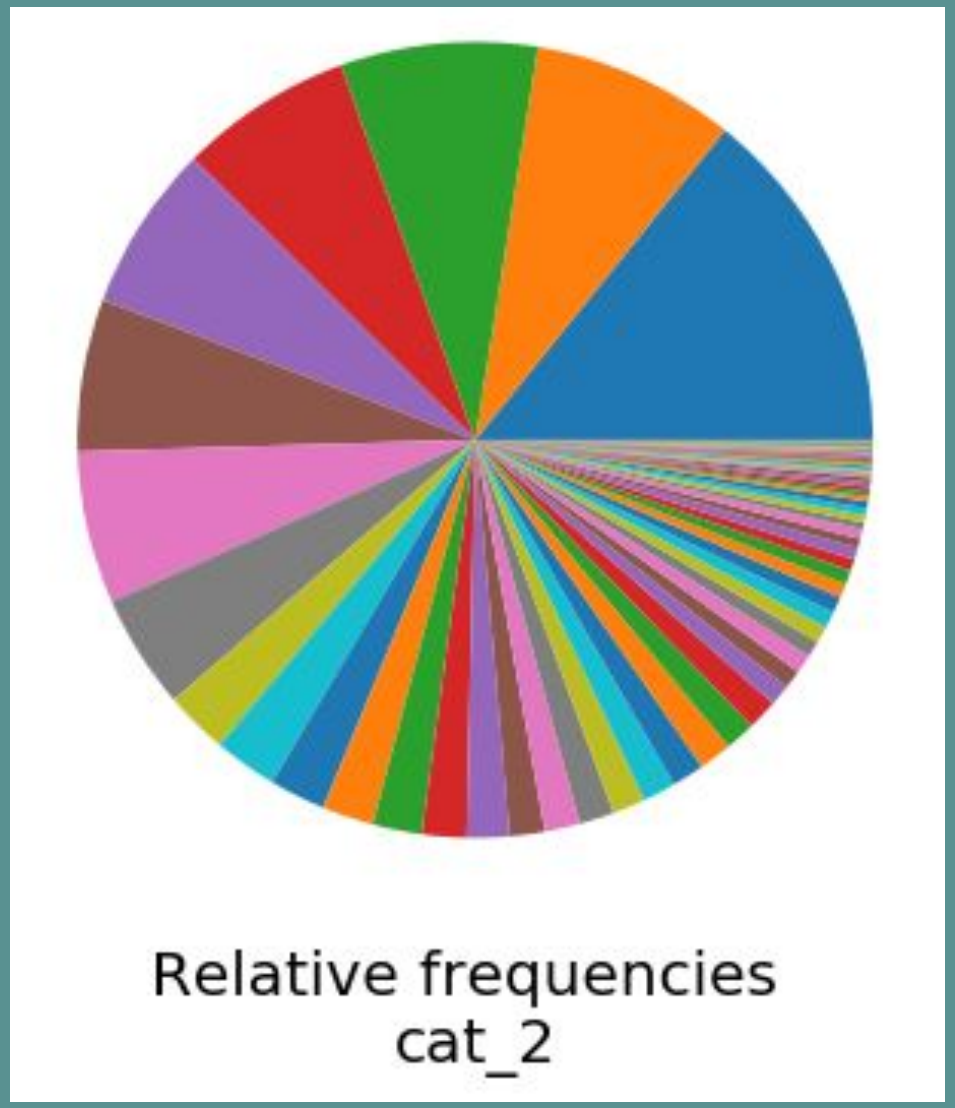
Eléments	cat_1	cat_2	cat_3
Product_category_tree	'Baby Care'	'Baby Bath & Skin'	'Baby Bath Towels'
Modalités	7	62	242
Remplissage	100 %	100 %	100 %

Les catégories principales

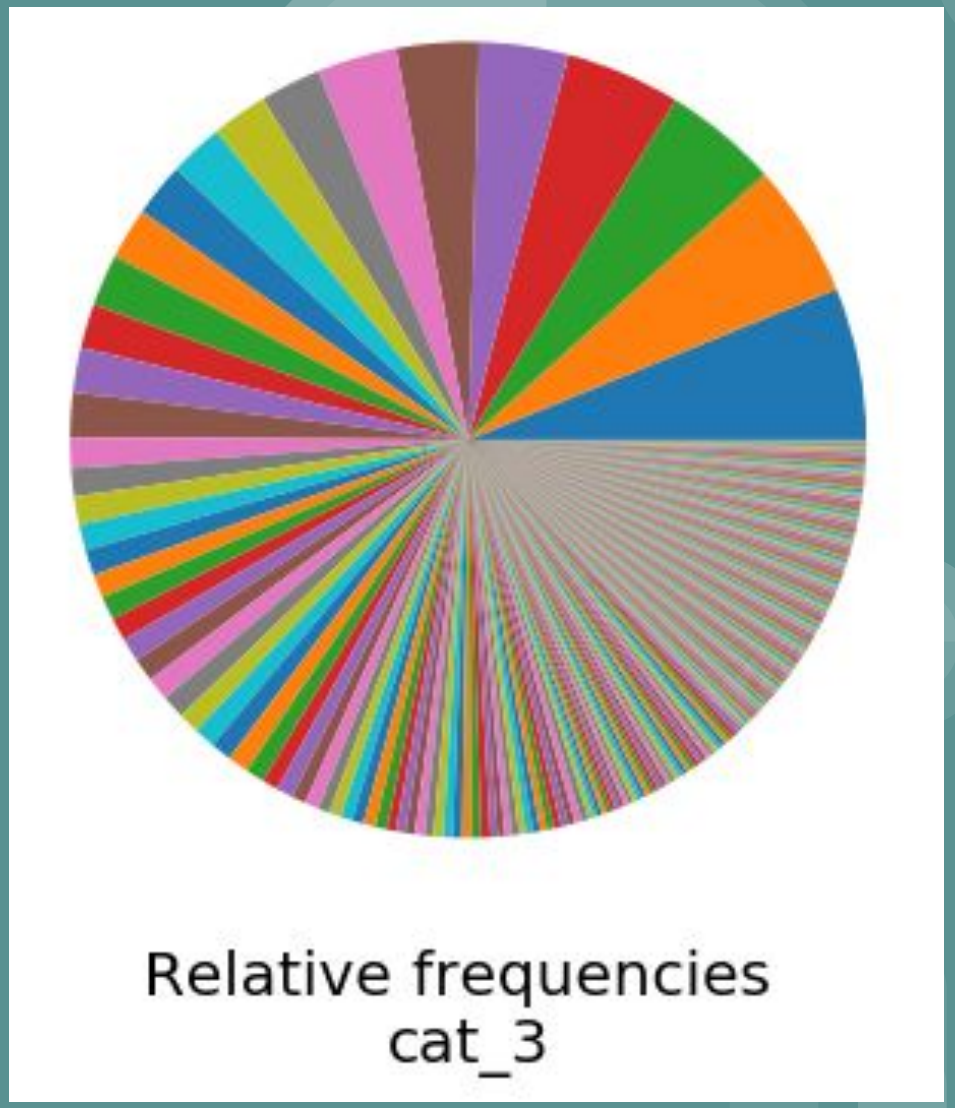


7 catégories principales

Répartition des catégories 2 et catégorie 3



62 catégories secondaires



242 catégories tertiaires

Prétraitements des données textuelles

#	Méthodes de traitement des données textuelles
#1	Encodage TF-IDF
#2	Encodage TF-IDF + Réduction NMF
#3	Encodage BOW + Réduction LDA

Mise en place du corpus ⟹ Concaténation des champs « product_name » & « description »

Encodage TF-IDF

Encodage du texte avec la méthode de pondération *tf-idf* :

Utilisation de **TfidfVectorizer()** de scikit-learn :

- **nettoyage** du texte (Nom produit + description) : accents, ponctuation, casse...
- élimination des « ***stop-words*** » et mots trop peu fréquents
- **tokenisation** en sac-de-mots (CountVectoriser)
- encodage par **tf-idf** (TfidfTransformer)

Taille du vocabulaire : 2 389 mots (+ 3'373 stop-words)
⇒ Encodage par vecteurs creux de **2389 dimensions.**

NMF (Factorisation matricielle non négative)

Encodage du texte avec le *NMF* :

1) Utilisation de **TfidfVectorizer()** de scikit-learn

2) Réduction dimensionnelle avec *Non-negative Matrix Factorisation* (**NMF**)

3) Choix de l'hyperparamètre **n_components** = **7** (principales catégories)

```
Topic 0:
watch analog men women discounts india sonata great maxima boys
Topic 1:
set combo flipkart shipping cash genuine delivery products 30 guarantee
Topic 2:
mug ceramic rockmantra coffee perfect gift loved safe prithish creation
Topic 3:
baby girl boy dress details cotton fabric neck shirt sleeve
Topic 4:
cm showpiece prices 10 brass handicrafts online 30 guarantee replacement
Topic 5:
abstract blanket single double quilts comforters multicolor raymond floral flipkart
Topic 6:
laptop battery cell hp pavilion lapguard skin shapes mouse pad
```

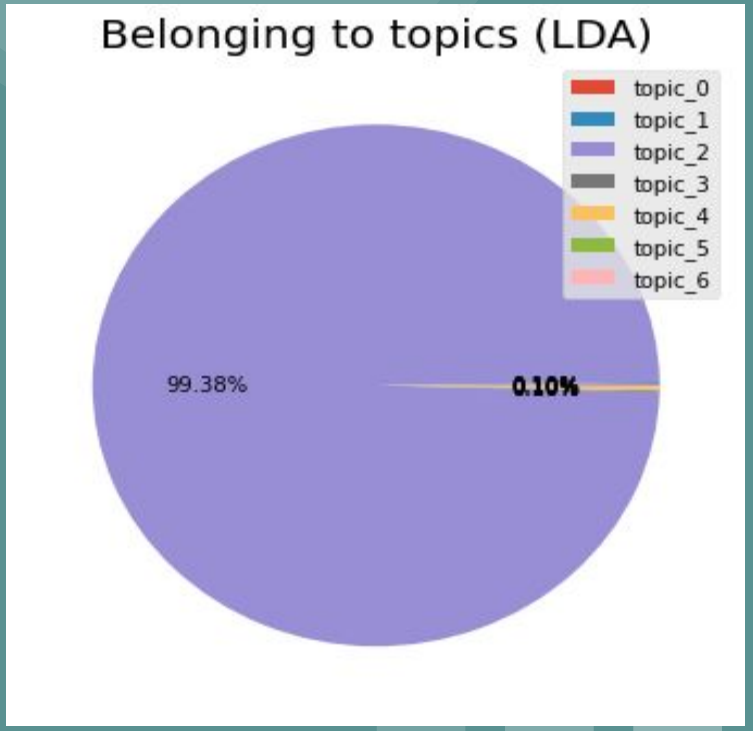
Latent Dirichlet Allocation (LDA)

Encodage du texte avec *LDA* :

1) ***Term-frequency*** sous la forme d'un sac-de-mot avec **CountVectorizer()**

2) Utilisation de **LatentDirichletAllocation()** de sklearn avec 7 « sujets »

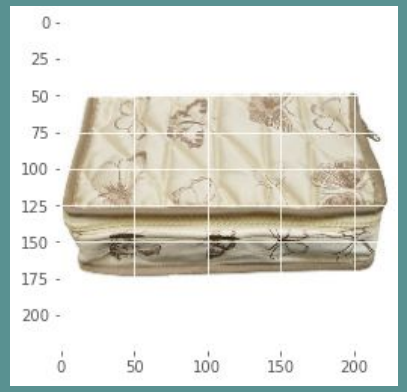
```
Topic 0:
warranty adapter product laptop cm sheet cotton bedsheet replacement features
Topic 1:
usb light led cover inch power flexible table lamp port
Topic 2:
cm pack design sticker color box material polyester features model
Topic 3:
products free delivery shipping cash genuine buy 30 day guarantee
Topic 4:
mug ceramic coffee perfect mugs hair towel gift material rockmantra
Topic 5:
single skin quilts cream blanket comforters kit abstract soap products
Topic 6:
laptop baby skin print set shapes cotton pad girl combo
```

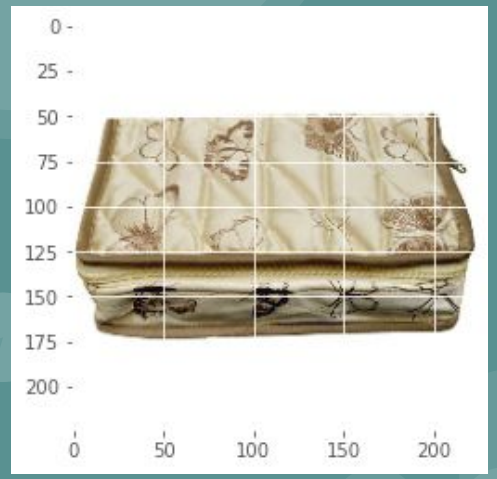


Prétraitements : Données visuelles

Prétraitement des données visuelles

Pré-traitement générique des données visuelles :

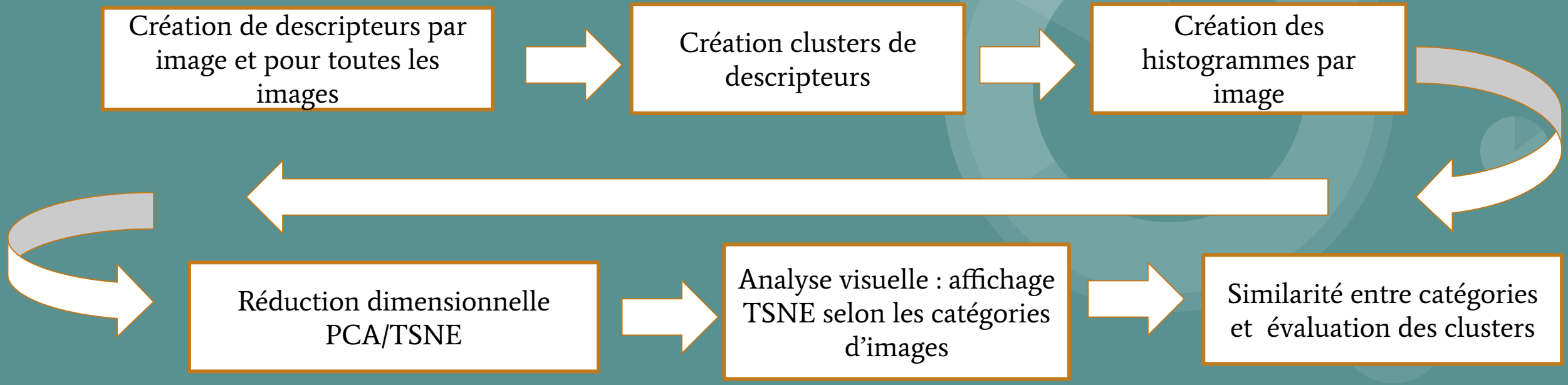




Redimensionnement 224 X 224

Contraste et Luminosité

Démarche



#	Méthode de traitement (extraction descripteurs) des données visuelles
#1	Encodage « *Bags-of-Visual-Words* » (BoVW) avec ORB

Bag of features

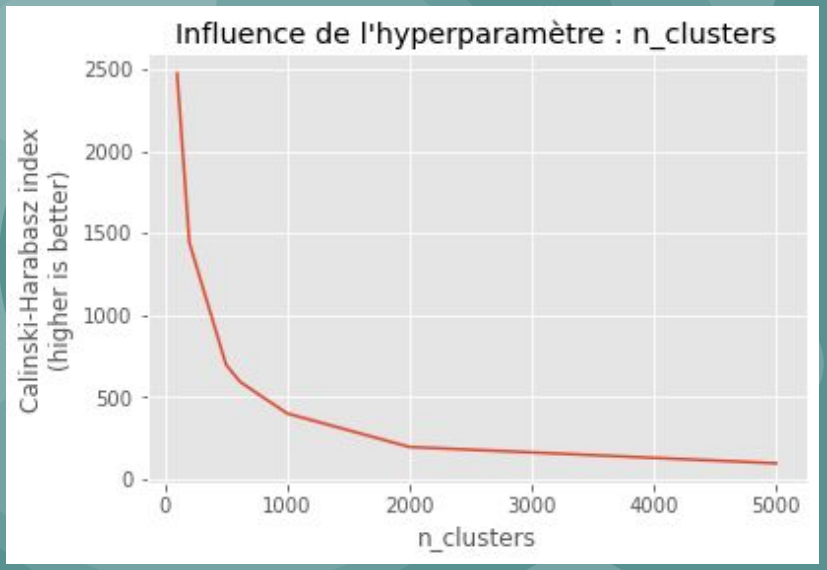
Création du “Sac de mots visuels”

1) Extraction des **descripteurs** (**ORB**) à 32 dimensions

- Pour chaque image, passage en redimensionnement et modification du contraste et de la brillance
- création d'une liste de descripteurs par image qui sera utilisée pour réaliser les histogrammes
- création d'une liste de descripteurs pour l'ensemble des images qui sera utilisée pour créer les clusters de descripteurs

Segmentation des descripteurs

2) **Segmentation** des descripteurs avec **k-means** (MiniBatchKMeans)
k = int(round(np.sqrt(Nombre total de descripteurs dans l'échantillon : 378 866)),0)) ====> 616

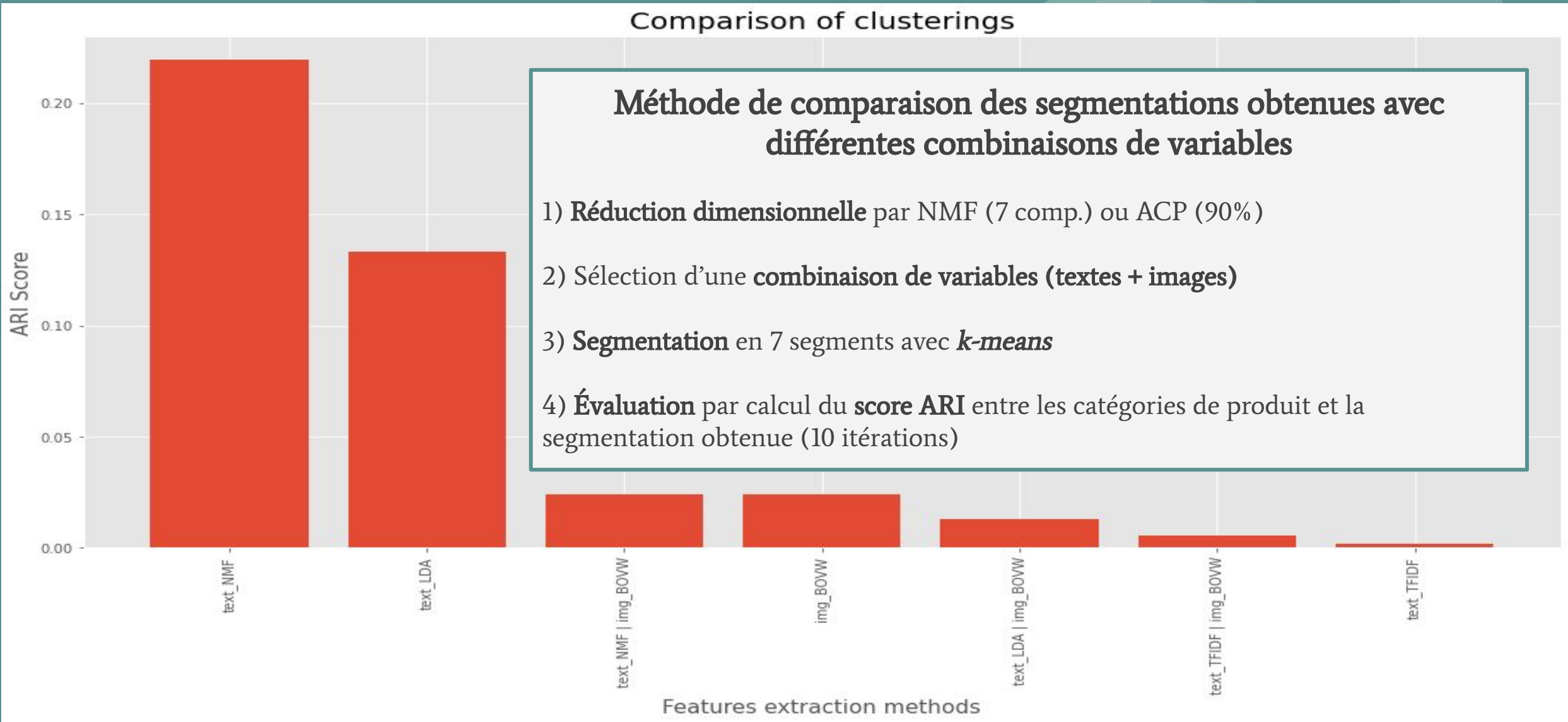


3) **Optimisation** du nombre de clusters « mots visuels ») avec GridSearch :

3) Création des **histogrammes** de « mots visuels »

```
Best hyperparameters: {'n_clusters': 100}
Best Silhouette score: 2474.3752807116916
Training time: 12.898325441000011
```

Evaluation des clusters



Méthode de comparaison des segmentations obtenues avec différentes combinaisons de variables

1) **Réduction dimensionnelle** par NMF (7 comp.) ou ACP (90%)

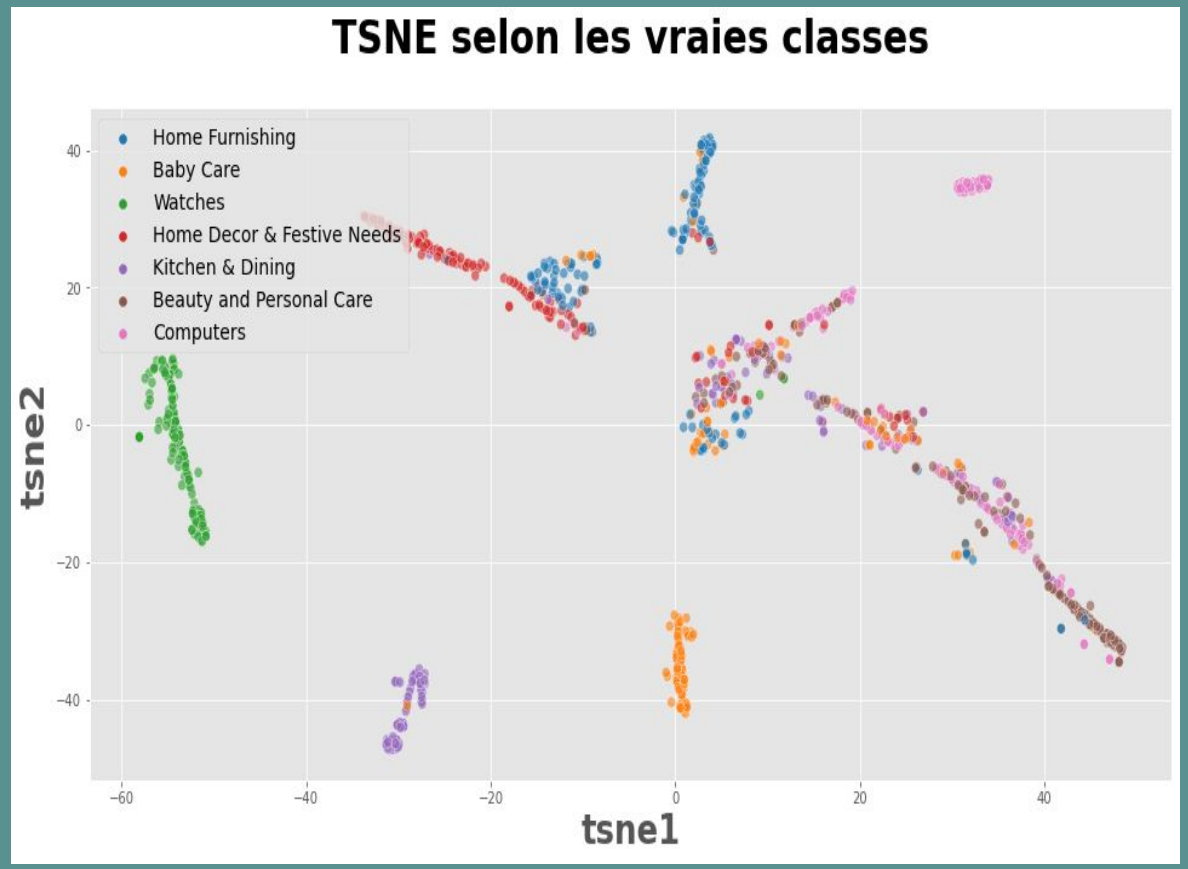
2) Sélection d'une **combinaison de variables (textes + images)**

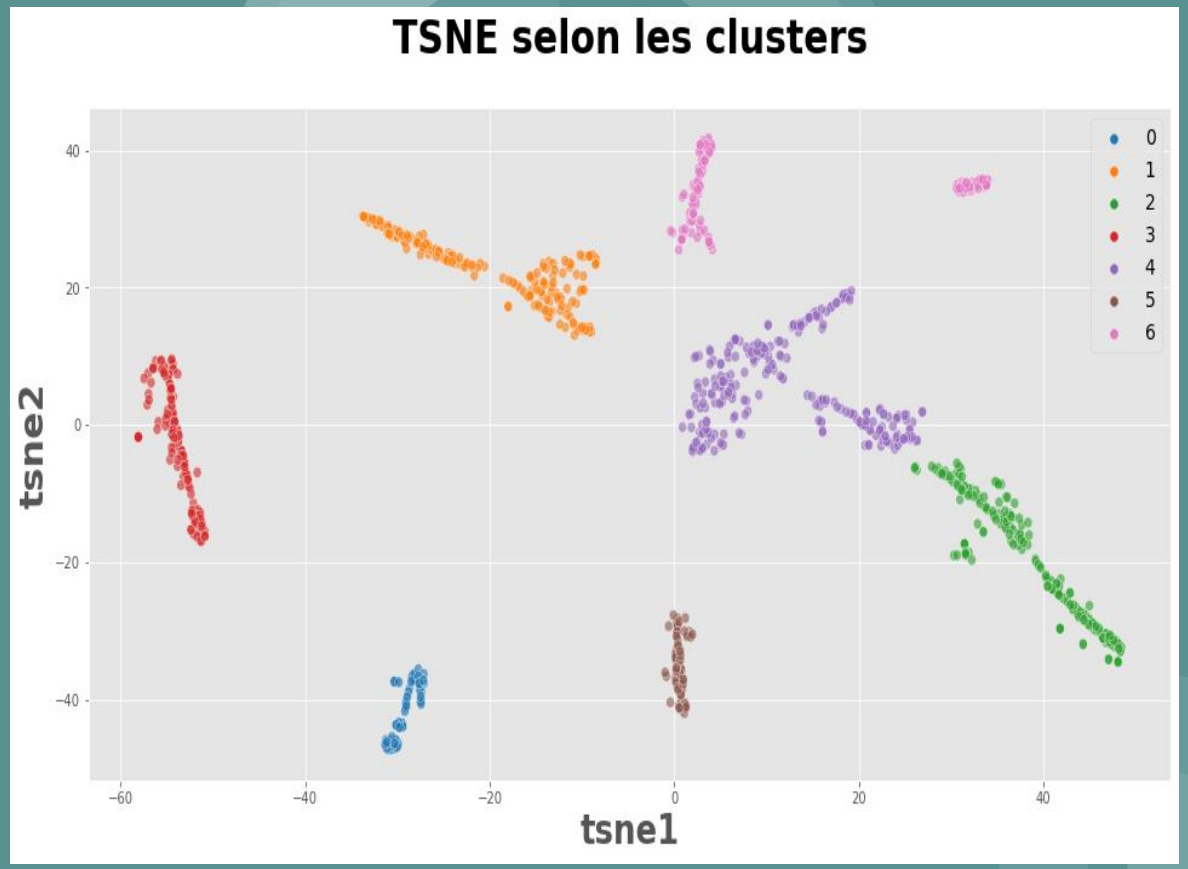
3) **Segmentation** en 7 segments avec ***k-means***

4) **Évaluation** par calcul du **score ARI** entre les catégories de produit et la segmentation obtenue (10 itérations)

Best ARI results: 21 % with text_NMF

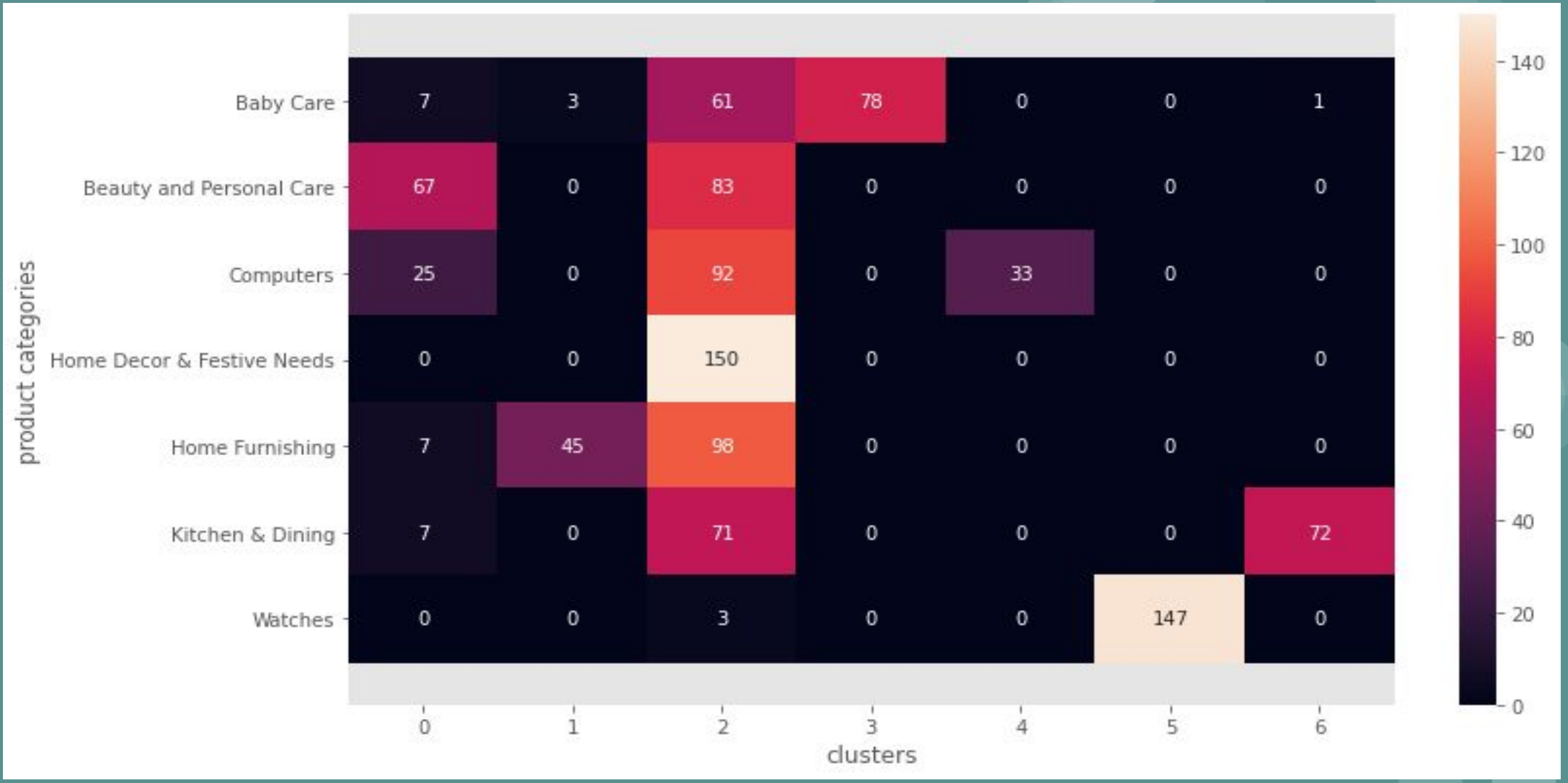
Projection plane par t-SNE - Projections selon les vraies classes





Résultats de la segmentation

Matrice de confusion

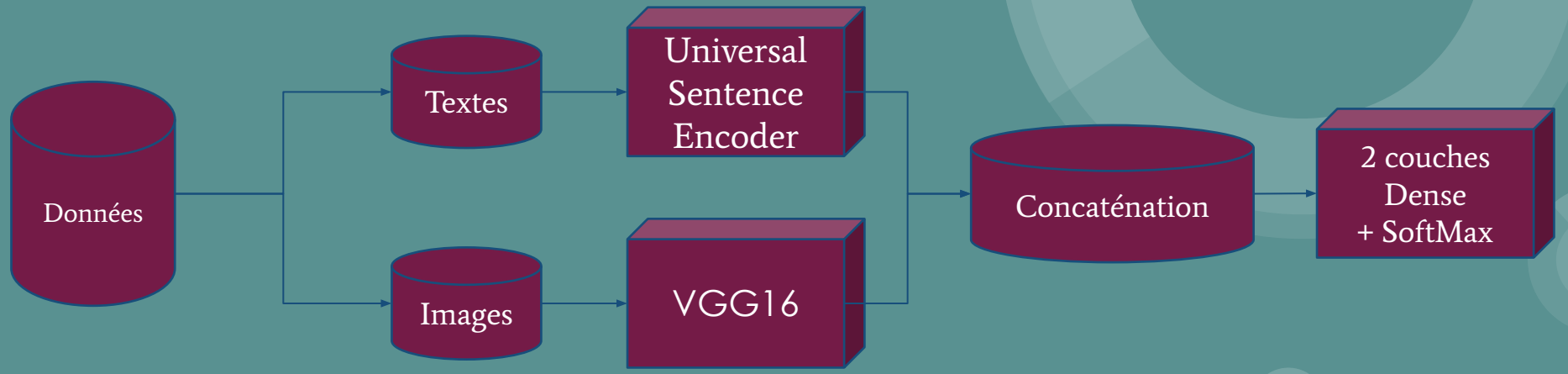


Conclusions

- Plus de données
- Les plongements de mots (ou word embeddings)
- Réseau neuronal avec apprentissage conjoint sur texte et images

  Utilisation d'un réseau de neurone convolutif en s'appuyant avec une méthode de Transfer learning, notamment le VGG16 qui est un réseau neuronal convolutif entraîné sur un sous-ensemble du jeu de données ImageNet, une collection de plus de 14 millions d'images appartenant à 22 000 catégories.

Réseau neuronal avec **apprentissage conjoint** sur texte et images



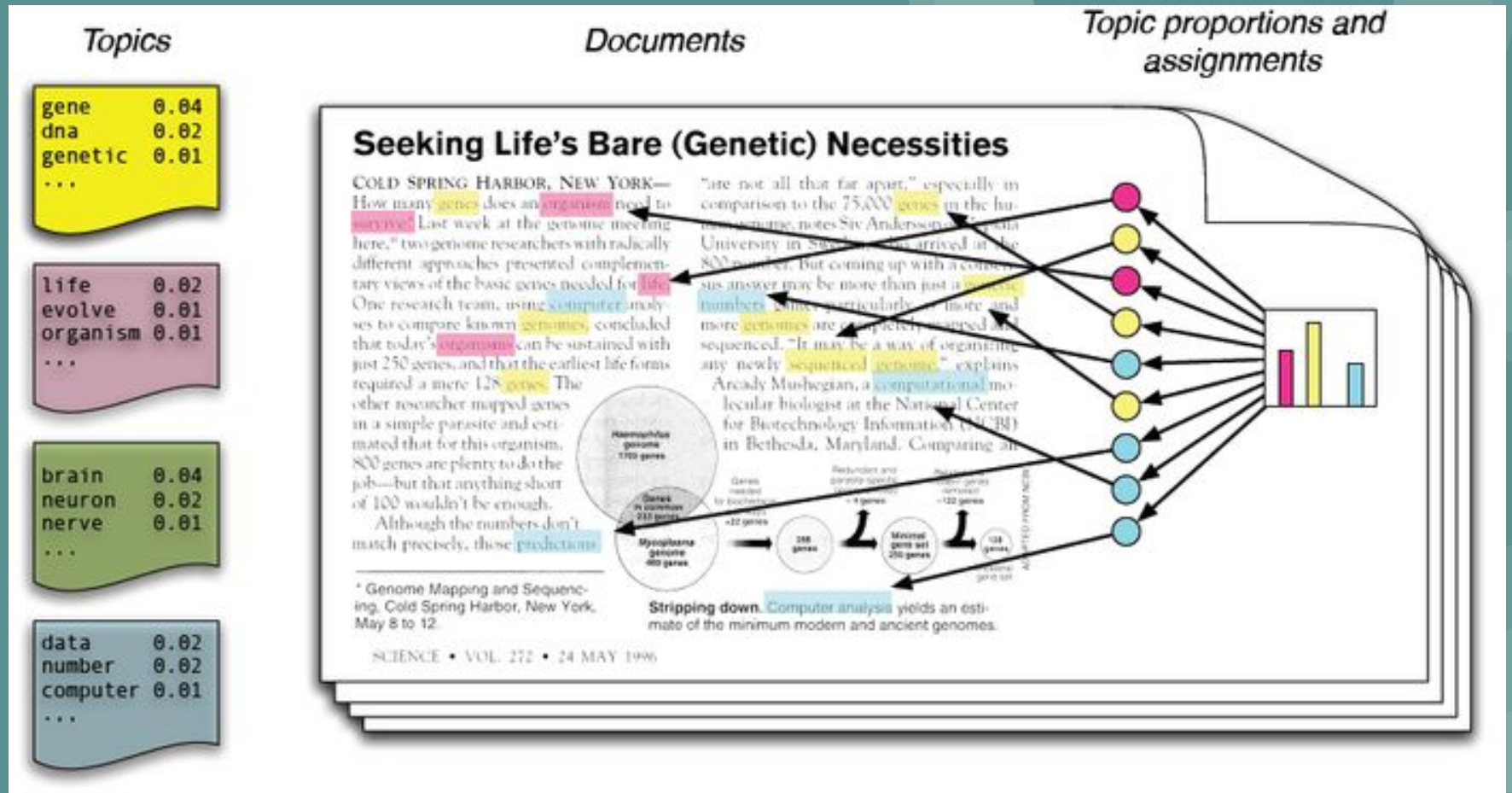
Merci !!!!

Annexes

ANNEXE 1

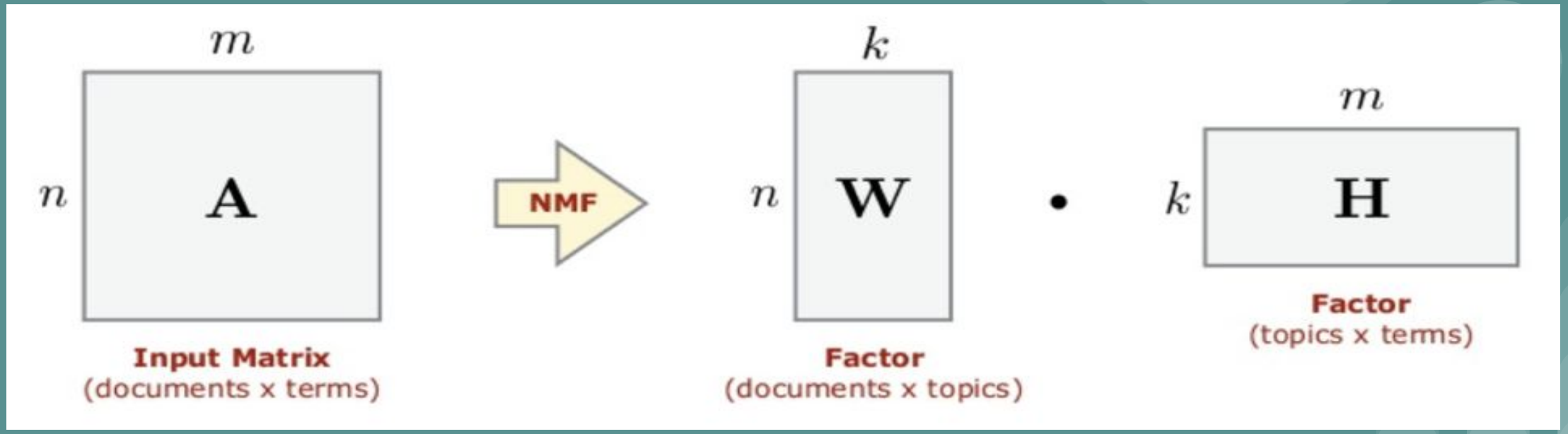
LDA (latent Dirichlet Allocation) est un algorithme de *modélisation de sujets* (non-supervisé).

Cet algorithme se base sur les deux hypothèses suivantes :
1) Les documents sont des distributions de probabilités sur des sujets latents.
2) Les sujets sont des distributions de probabilité sur les mots.



ANNEXE 2

NMF (*non-négative matrix factorization*) est un algorithme de réduction dimensionnelle dans laquelle une matrice **A** est approximativement factorisée en deux matrices **W** (basis document matrix) et **H** (weight matrix), avec la particularité que ces trois matrices n'ont pas d'éléments négatifs.



ANNEXE 3

Indice de Calinski-Harabasz (Critère de rapport des variances)

Pour *k* groupes, l'indice *s* de Calinski-Harabasz est le rapport de la dispersion inter-groupes moyenne et de la dispersion intra-groupe :

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$$

- où B_k est la matrice de dispersion inter-groupes : $B_k = \sum_{q=1}^{k} n_q (c_q - c)(c_q - c)^T$
- et W_k est la matrice de dispersion intra-groupe : $W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$

avec *N* le nombre de points de données, C_q l'ensemble des points dans le groupe *q*, c_q le centroïde du groupe *q*, c le centroïde du groupe *E* et n_q le nombre de points dans le groupe *q*.

ANNEXE 4

Indice de Rand Ajusté (ARI) L'*Adjusted Rand Index* (ARI) est la normalisation de RI qui permet de comparer deux partitions de nombres de classes différentes.

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

- RI : indice de Rand : proportion de paires de points qui sont groupés de la même façon dans les deux partitions.
- E(RI) : espérance de l'indice de Rand (pour une partition aléatoire)
- max(RI) : indice de Rand maximal qui pourrait être obtenu étant donné le nombre de classes distincts.