Faisabilité d’un moteur de classification basé sur des images et du texte

présentation du 13 octobre 2021

Plan de la présentation

1) Problématique métier

2) Description des données

3) Traitement des images :
   - VGG-16
   - SIFT et ORB

4) Traitement des textes :
   - TF-IDF
   - Word2Vec :
     - pré-entrainé
     - entrainé

5) Essais de faisabilité avec un classifieur SVM

6) Conclusion

Problématique du site " Place de marché "

Contexte :

- les vendeurs sur le site postent une photo et une description.
- attribution de la catégorie d’un article effectuée manuellement.
- besoin d’automatiser cette tâche.

Objectif : étudier la faisabilité d'un moteur de classification des articles en différentes catégories

Missions confiées :

- réaliser un prétraitement des images et des descriptions des produits, une réduction de dimension, puis un clustering.
- illustrer le fait que le clustering permet de regrouper des produits de même catégorie, à l’aide d’une représentation en deux dimensions.

Présentation des données

1050 articles

7 catégories (effectif de 150 articles / catégorie) :
"Beauty and Personal Care", "Computers", "Baby Care", "Home Decor", "Home Furnishing", "Kitchen & Dining", "Watches“

Exemple de données pour un article (principales variables) :

Nom variable	Valeur
product_name	Eurospa Cotton Terry Face Towel Set
product_ category_tree	**Baby Care** >> Baby Bath & Skin >> Baby Bath Towels >> …
image	64d5d4a258243731dc7bbb1eef49ad74.jpg
description	"Key Features of Eurospa Cotton Terry Face Towel Set Size: small Height: 9 inch GSM: 360,Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) Price: Rs. 299 Eurospa brings to you an exclusively designed, 100% soft cotton towels of export quality. All our products have soft texture that takes care of your skin and gives you that enriched feeling you deserve. Eurospa has been exporting its bath towels to lot of renowned brands for last 10 years and is famous for its fine prints, absorbency, softness and durability…”

Plan de la présentation

1) Problématique métier

2) Description des données

2) Traitement des images :
- VGG-16
- SIFT et ORB

3) Traitement des textes :
- TF-IDF
- Word2Vec :
  - préentrainé
  - entrainé

4) Essais de faisabilité avec un classifieur SVM

5) Conclusion

Génération de descripteurs par VGG-16

VGG-16 :
Réseau de neurones convolutif à 16 couches pour la classification d'images.
Réseau pré-entrainé sur 1.3 millions d'images.
Possibilité de l'utiliser en transfer learning.
Dernière couche : classifieur softmax qui prend un vecteur de dimension 4096 en entrée.

Génération du descripteur (4096 dimensions) d'une image :

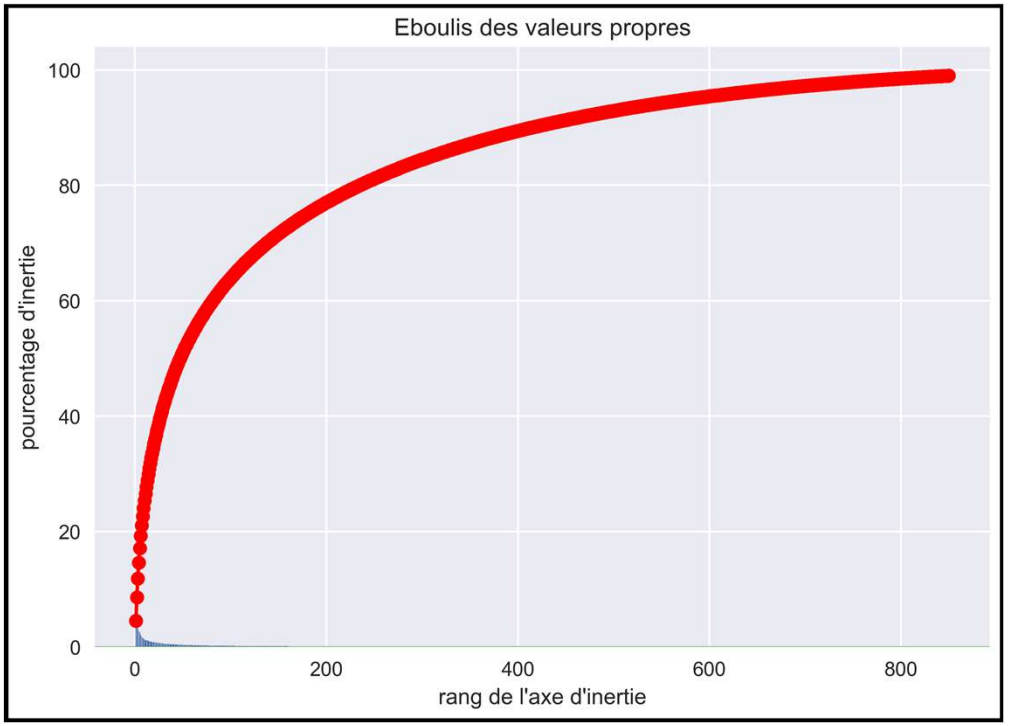
1) Prepocessing de l'image (en particulier redimensionnement au format 224*224 pixels)
2) Calcul du descripteur via le modèle préentrainé.

Traitement des images (VGG-16) : vectorisation du dataset

Méthodologie pour le dataset :

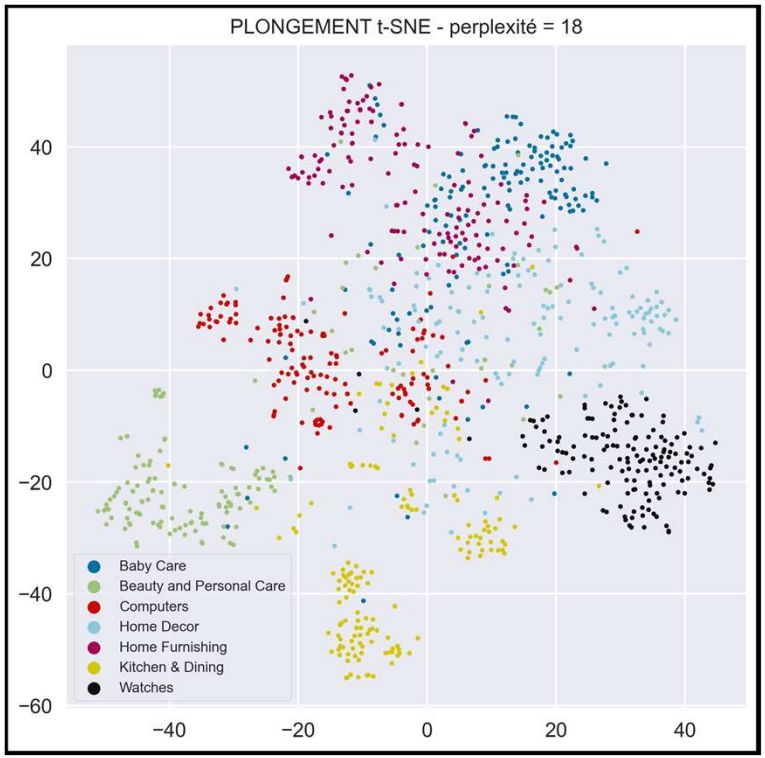
1) Pour chaque image, générer le vecteur à 4096 dimensions.
2) Réduire la dimension par PCA (99% d'inertie), après centrage-réduction.

Dimensions dataset avant réduction PCA : (1050, 4096)
Dimensions dataset après réduction PCA : (1050, 850)

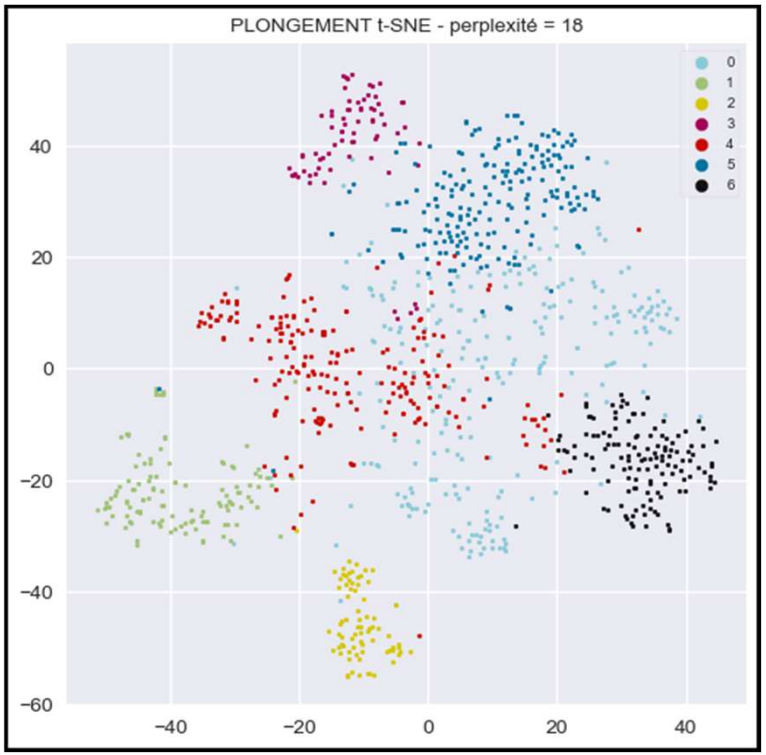


Utilisation des données vectorisées (VGG-16)

t-SNE : bonne séparation des catégories :
- Beauty and Personnal Care
- Kitchen & Dining
- Watches



Clustering par **k-means** (k=7) :
- ARI = 0.50
- Exactitude = 0.73 (après attribution des catégories)



Séparation des catégories ⇒ Résultat encourageant

Plan de la présentation

1) Problématique métier

2) Description des données

2) Traitement des images :
- VGG-16
- SIFT et ORB

3) Traitement des textes :
- TF-IDF
- Word2Vec :
  - préentrainé
  - entrainé

4) Essais de faisabilité avec un classifieur SVM

5) Conclusion

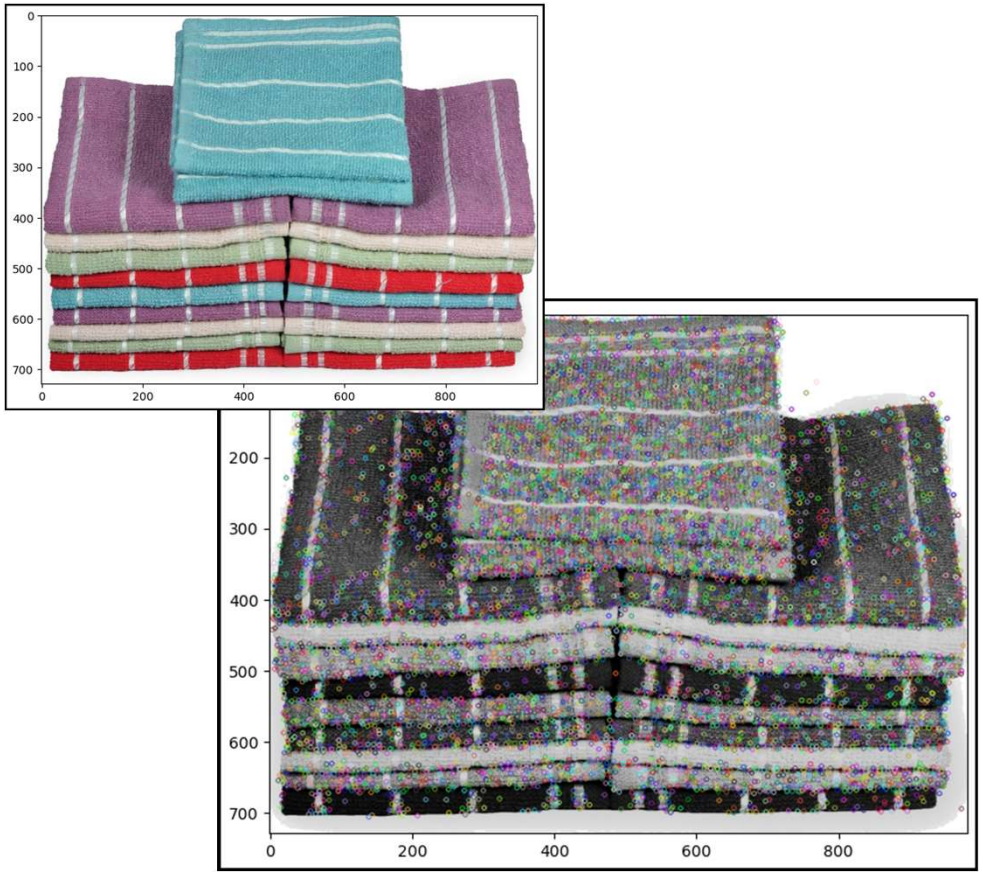
Traitement des images : génération de descripteurs par SIFT

Méthodologie pour une image :

1) Convertir en niveaux de gris
2) Uniformiser l'histogramme des gris
3) Calculer les descripteurs SIFT

⇒ obtention de vecteurs à 128 dimensions
(plusieurs milliers par image du dataset)

Exemple : lot de serviettes ⇒ 8896 descripteurs



Traitement des images (SIFT) : clustering des descripteurs du dataset

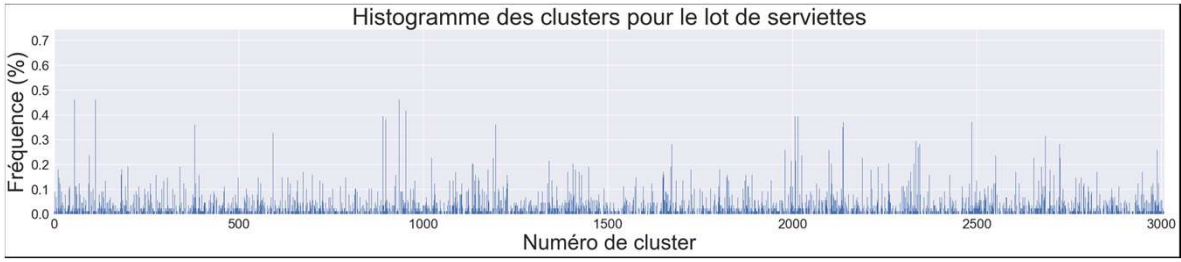
Contrairement à VGG-16, les descripteurs ne sont pas « prêts à l'emploi ».

Méthodologie pour le dataset :

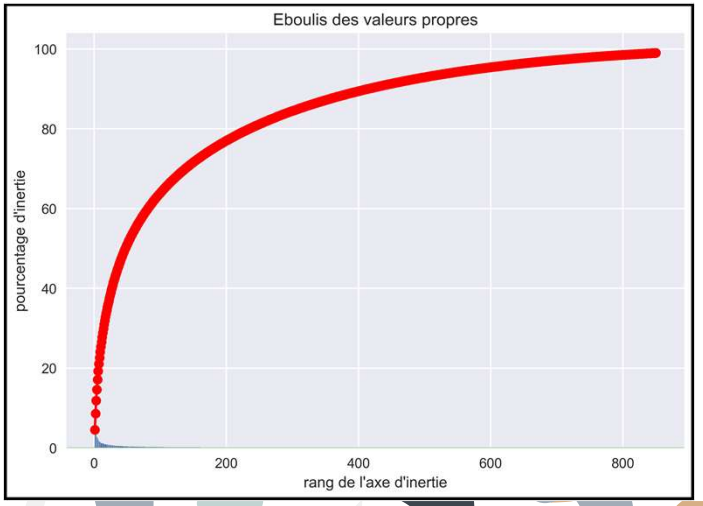
1) Partitionner les descripteurs en n clusters dans l'espace à 128 dimensions.
2) Pour chaque image : compter le nb de descripteurs dans chaque cluster.
3) Chaque image est alors décrite par un vecteur de dimension n.
4) Réduire la dimension par PCA (99% d'inertie), après centrage-réduction.

⇒ Toutes les images sont décrites par un vecteur de dimension identique (n).

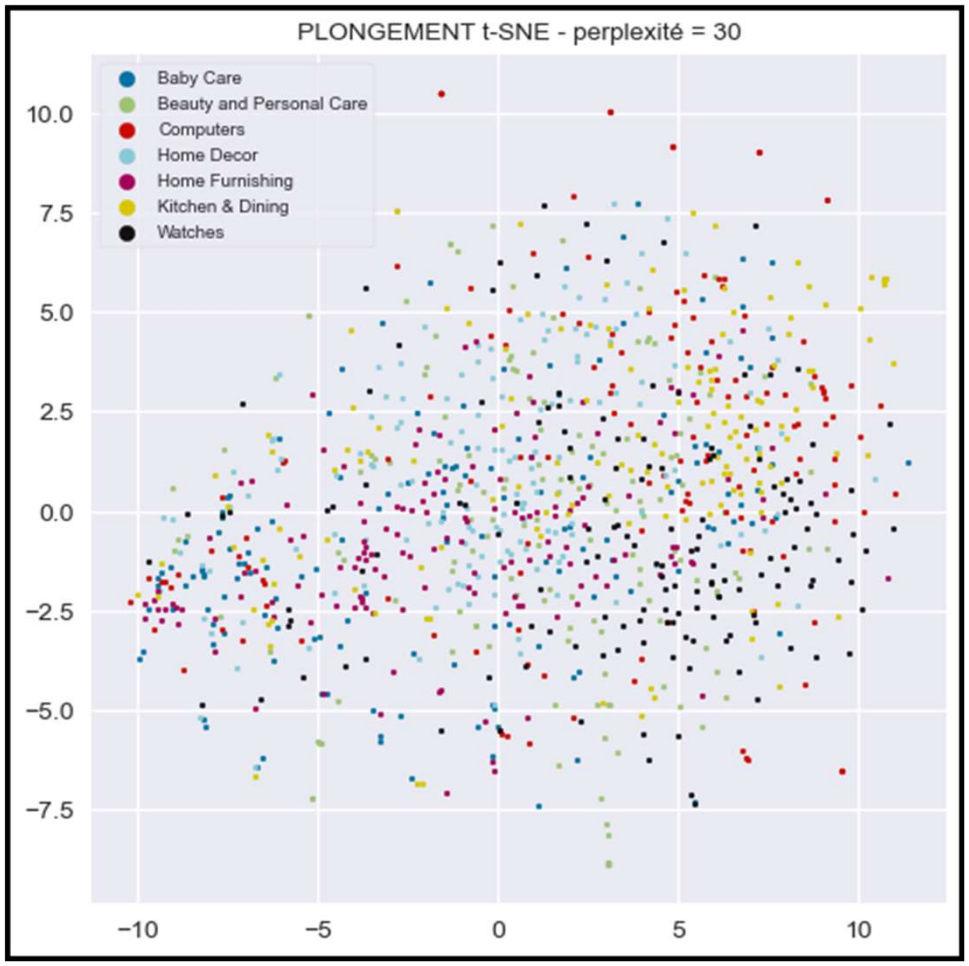
Exemple : lot de serviettes / dataset de 1050 articles
9 millions de descripteurs
Partition en n=3009 clusters (minibatch k-means).



Dimensions dataset avant réduction PCA : (1050, 3009)
Dimensions dataset après réduction PCA : (1050, 850)



Utilisation des données vectorisées (SIFT)

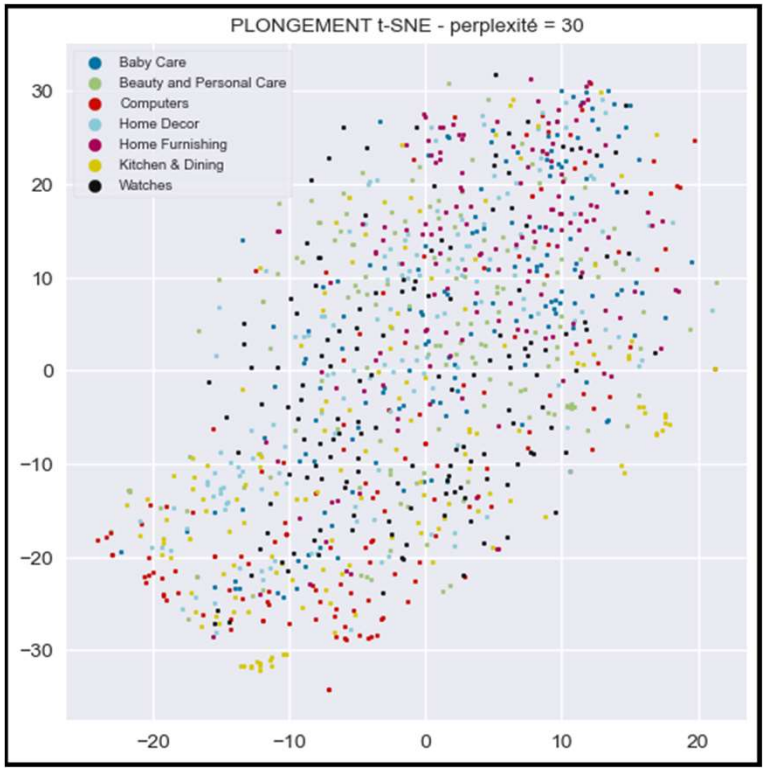


Clustering par k-means (k=7) : ARI = 0.08

Catégories mal séparées
⇒ Résultat peu encourageant

ORB : traitement d'image, vectorisation, clustering

Plus de descripteurs générés que par SIFT (27 millions).
Post-traitement analogue à celui réalisé par SIFT (5200 clusters de descripteurs).
Après PCA (99% d'inertie) : 579 dimensions.



Clustering par k-means (k=7) : ARI = 0.05

Catégories mal séparées
⇒ Résultat peu encourageant

Plan de la présentation

1) Problématique métier

2) Description des données

2) Traitement des images :
- VGG-16
- SIFT et ORB

3) Traitement des textes :
- TF-IDF
- Word2Vec :
  - préentrainé
  - entrainé

4) Essais de faisabilité avec un classifieur SVM

5) Conclusion

Traitement du texte : étape préliminaire de nettoyage

Méthodologie pour un document :

1) Tokenisation.
2) Casse : minuscule.
3) Suppression :
   - de la ponctuation,
   - des termes non alphabétiques,
   - des stop words.
4) Stemming (optionnel selon traitement)

Exemple : lot de 20 serviettes

Extrait du document originel :

“Key Features of Eurospa Cotton Terry Face Towel Set Size: small Height: 9 inch GSM: 360,Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) …”

Document après nettoyage + stemming:

'key', 'featur', 'eurospa', 'cotton', 'terri', 'face', 'towel', 'set', 'size', 'small', 'height', 'inch', 'gsm', 'eurospa', 'cotton', 'terri', 'face', 'towel', 'set', 'piec', 'face', 'towel', 'set', 'assort'

Traitement du texte : TF-IDF

Méthodologie TF-IDF pour le dataset :

1) Conversion du corpus de documents (tokens) en matrice de comptage de tokens ("bag of words").
2) Transformation TF-IDF.

⇒ Toutes les documents sont décrits par un vecteur de dimension identique.

3) Réduction de dimension optionnelle (perte de la correspondance des tokens).

Exemple : lot de serviettes / dataset de 1050 documents

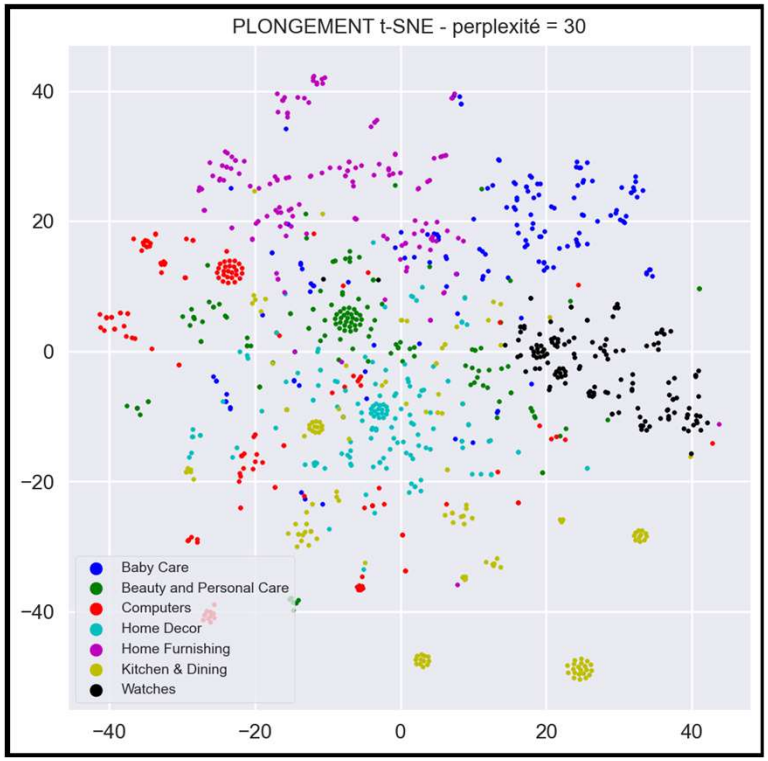
Utilisation de TfidfVectorizer pour la réalisation du traitement (document frequency ≥ 10)
→ vecteur de dimension 542

Lot de serviettes : 68 éléments non nuls dans le vecteur creux de dimension 542.

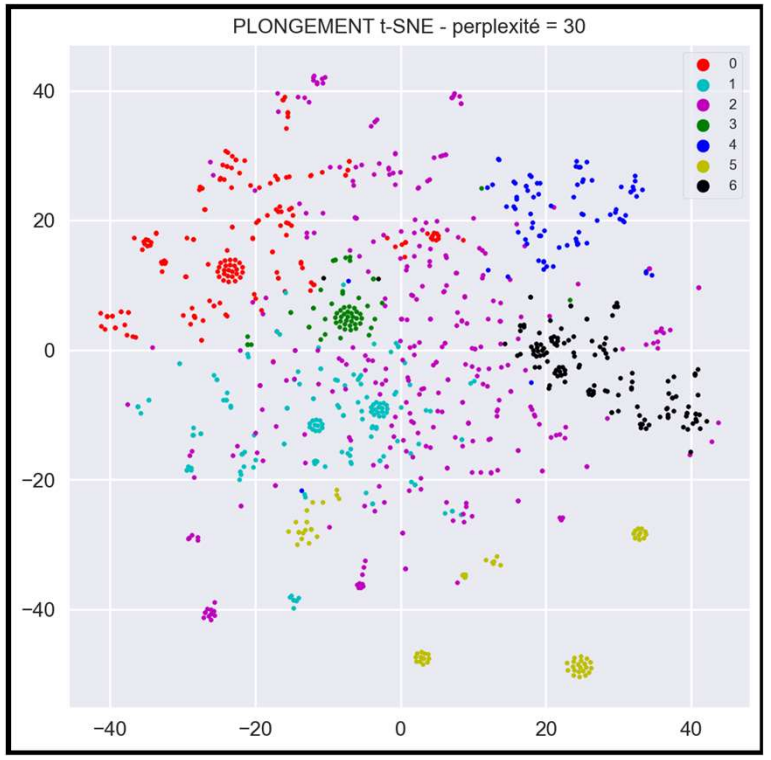
Dimension réduite par PCA (99% d'inertie) :
- avant réduction PCA : (1050, 542)
- après réduction PCA : (1050, 379)

Traitement du texte : TF-IDF

t-SNE : bonne séparation des catégories :
- Baby Care
- Home Furnishing
- Watches



Clustering par **k-means** (k=7) :
- ARI = 0.27
- Exactitude = 0.55



⇒ Résultat encourageant

Plan de la présentation

1) Problématique métier

2) Description des données

2) Traitement des images :
- VGG-16
- SIFT et ORB

3) Traitement des textes :
- TF-IDF
- Word2Vec :
  - préentrainé
  - entrainé

4) Essais de faisabilité avec un classifieur SVM

5) Conclusion

Traitement du texte avec Word2Vec pré-entrainé

Word2Vec
Réseau de neurones entraîné pour analyser le contexte linguistique des mots.

Utilisation du réseau pré-entrainé avec 'Google News dataset'
Entrainement sur 100 miliards de mot.
Contient un dictionnaire de 3 millions de mots, vectorisés en dimension 300.

Méthodologie pour vectoriser notre dataset :

Pour chaque document :

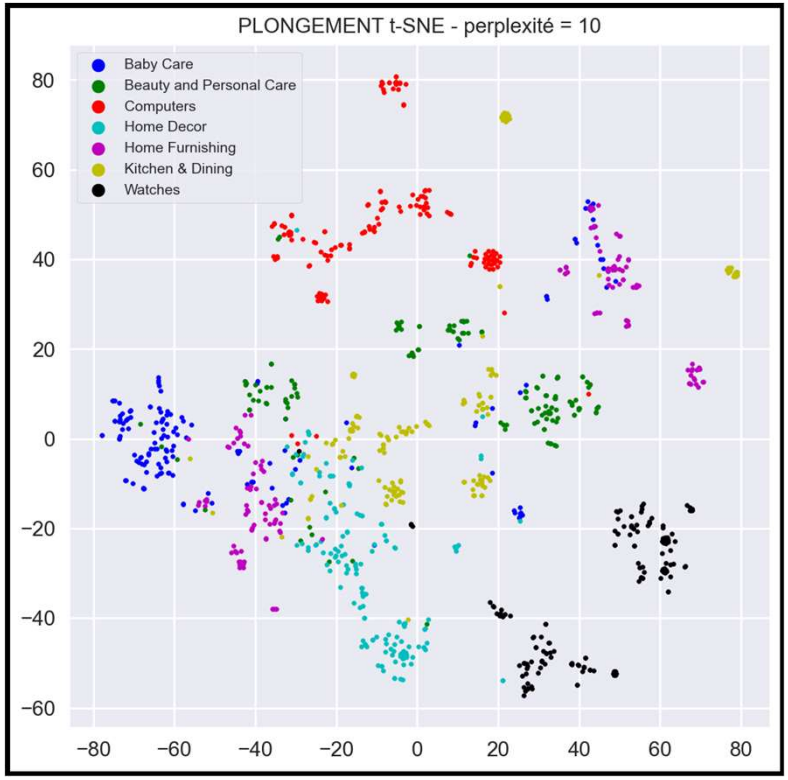
1) les tokens (sans stemming) sont convertis en vecteur de dimension 300 (si disponible dans le dictionnaire Word2Vec),
2) la moyenne des vecteurs est calculée ⇒ obtention d'un vecteur pour le document.

⇒ Toutes les documents sont décrits par un vecteur de dimension 300.

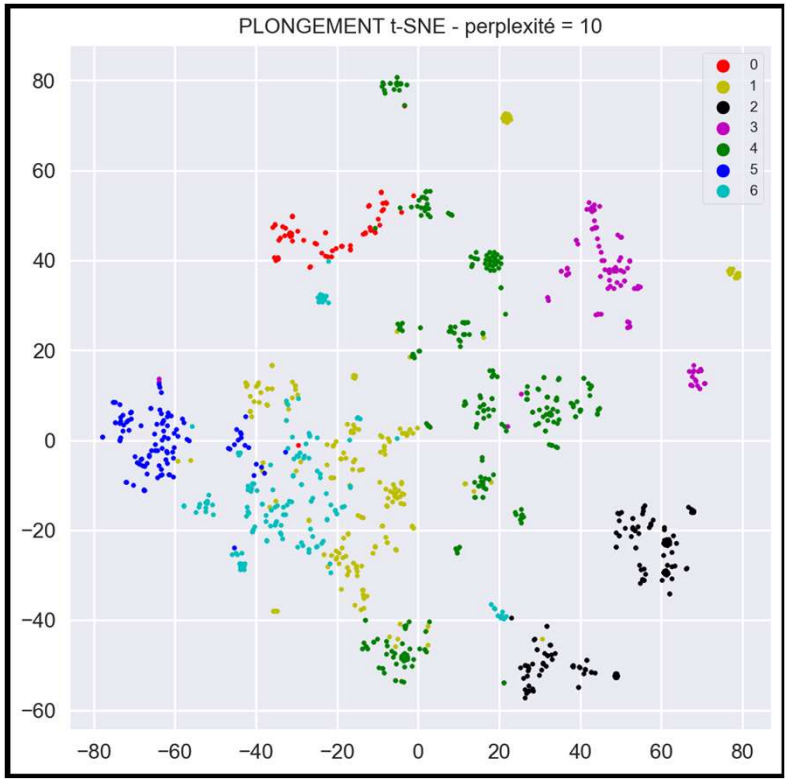
Réduction de dimension optionnelle (perte de la correspondance des tokens).
Dimension réduite par PCA (99% d'inertie) :
- avant réduction PCA : (1050, 300)
- après réduction PCA : (1050, 223)

Traitement du texte avec Word2Vec pré-entrainé

t-SNE : bonne séparation des catégories :
- Baby Care
- Computers
- Watches



Clustering par **k-means** (k=7) :
- ARI = 0.31
- Exactitude = 0.55



⇒ Résultat encourageant

Traitement du texte avec Word2Vec entrainé sur notre dataset

Entrainement du réseau Word2Vec
Entrainement sur les 1050 documents.
Choix de diminuer le nombre de dimensions à 10.
Obtention d'un dictionnaire de 1379 mots, vectorisés en dimension 10.

Méthodologie pour vectoriser notre dataset :

Idem méthodologie précédente (avec Word2Vec pré-entrainé) :

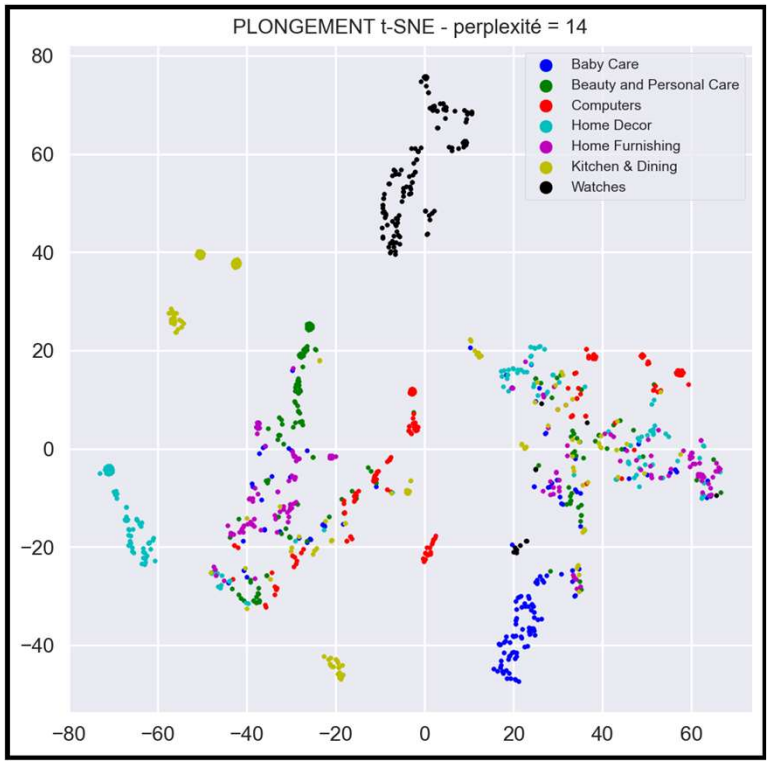
⇒ Toutes les documents sont décrits par un vecteur de dimension 10.

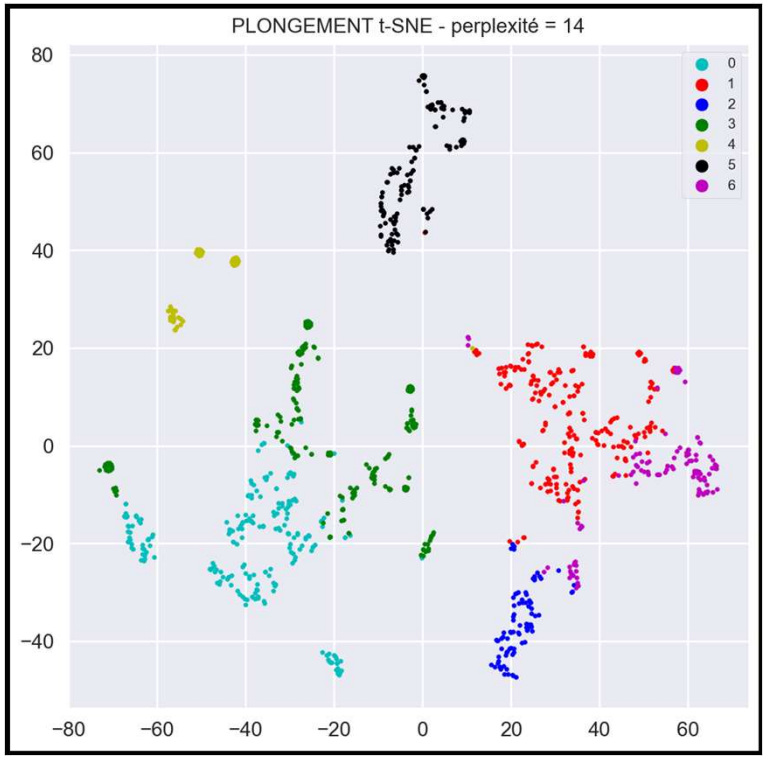
Réduction de dimension optionnelle (perte de la correspondance des tokens).
Dimension réduite par PCA (99% d'inertie) :
- avant réduction PCA : (1050, 10)
- après réduction PCA : (1050, 6)

Traitement du texte avec Word2Vec entrainé sur notre dataset

t-SNE : bonne séparation d'une seule catégorie :
- Watches

Clustering par **k-means** (k=7) :
- ARI = 0.21
- Exactitude = 0.43





⇒ Métriques moins bonnes que pour le modèle pré-entrainé, mais fort potentiel sur dataset de plus grande taille

Plan de la présentation

1) Problématique métier

2) Description des données

2) Traitement des images :
- VGG-16
- SIFT et ORB

3) Traitement des textes :
- TF-IDF
- Word2Vec :
  - préentrainé
  - entrainé

4) Essais de faisabilité avec un classifieur SVM

5) Conclusion

Essais de faisabilité avec un classifieur SVM

Méthodologie de test d'un classifieur SVM :

1) Données réduites en dimensions par PCA
2) Validation croisée 100 plis
3) Entrainement du modèle SVM (hyperparamètres par défaut de sci-kit learn)
4) Calcul du score d'exactitude sur les 100 plis

Essais de faisabilité avec un classifieur SVM

Pré-traitement des données	ARI (k-means, k=7)	Exactitude du SVM	Conclusion : pré-traitement pertinent ?
VGG-16	0.50	**0.84**	**Oui**
SIFT	0.08	0.50	**Non**
ORB	0.05	0.47	**Non**
TF-IDF	0.27	**0.89**	**Oui**
Word2Vec pré-entrainé	0.31	**0.93**	**Oui**
Word2Vec entrainé	0.21	**0.77**	**Oui (à confirmer par + de données)**

Conclusion

Différentes méthodes de pré-traitement testées comparativement

Images : VGG-16 >> SIFT > ORB

Données textes : Word2Vec > TF-IDF
Fort potentiel d'amélioration d'un modèle basé sur **Word2Vec** avec plus de données

Résultats issus des données pré-traitées : **textes > images**

Essais préliminaires encourageants quant au développement d'un moteur de classification basé sur l'image ou la description des articles

Perspectives :
- Nombreux réseaux de neurones pré-entrainés alternatifs à VGG-16 pour la description d'image
- Nombreux modèles pour la description de documents alternatifs à Word2Vec
- Possibilité de combiner les vecteurs issus des vectorisation d'images et vectorisation de textes