Classifiez automatiquement des biens de consommation

Formation Data Scientist
Projet n°6

Sommaire

Problématique

- Entreprise : **"Place de marché”**
- Activité : **marketplace e-commerce** :
  - des vendeurs proposent des articles à des acheteurs en postant une photo et une description.
- Problèmes :
  - **catégorie** d'un article donnée **manuellement** par les **vendeurs**.
  - **volume** des articles très **faible**.
- Besoins :
  - faciliter la mise en ligne de nouveaux articles pour les vendeurs
  - faciliter la recherche de produits pour les acheteurs
- Mission : Réaliser une étude de faisabilité d'un moteur de classification d'articles basé sur une image et une description

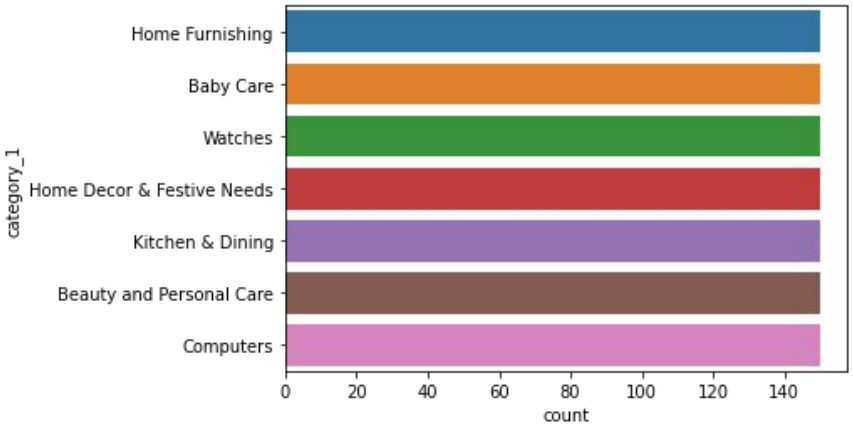
Problématique

- Méthode
  - Analyser le jeu de données en réalisant :
    - un prétraitement des images et des descriptions de produits
    - une réduction de dimension,
    - un clustering.
  - Présenter les résultats en deux dimensions.
- Contraintes
  - A minima un algorithme de type SIFT / ORB / SURF.
  - Un algorithme de type CNN Transfer Learning

Bay AT

Jeu de données textuelles

- Fichier CSV avec 1050 lignes 15 colonnes
- 3 variables nécessaires dont une créé
  - description
  - image
  - category_1*



description :
T STAR UFT-TSW-005-BK-BR Analog Watch - For Boys
Price: Rs. 399
Whether you are on your way to work or travelling abroad with family, lifestyle accessories like watches, wallets and belts help to add a touch of sophistication and class to your otherwise mundane and regular daily wear. When it all comes down to it, suave leather belts and intricately designed and finished timepieces are what separate you from the rest.
Whether you are on your way to work or travelling abroad with family, lifestyle accessories like watches, wallets and belts help to add a touch of sophistication and class to your otherwise mundane and regular daily wear. When it all comes down to it, suave leather belts and intricately designed and finished timepieces are what separate you from the rest.

image: dd0e3470a7e6ed76fd69c2da27721041.jpg

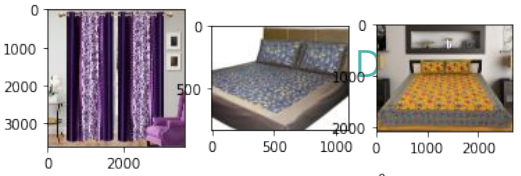
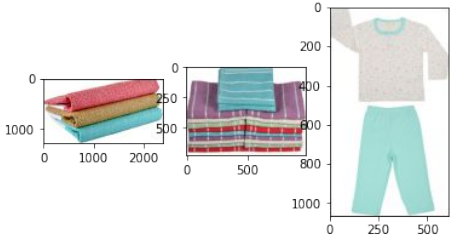
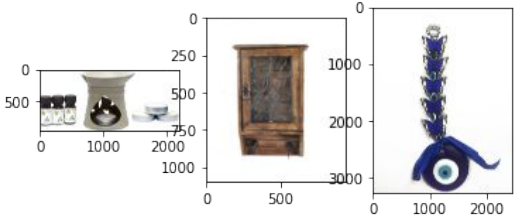
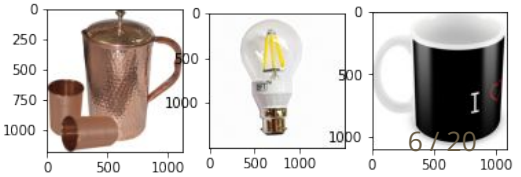
category_1: 'Watches'

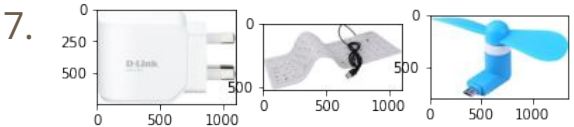
* product_category_tree : ["category_1 >> category_2 >> category_3"] : ["Watches >> Wrist Watches >> T STAR Wrist Watches"]

Jeu de données visuelles

- 1050 images au format jpeg de résolutions comprises entre 400x145 pixels et 11042x8484 pixels
  - Home Furnishing : 150
  - Baby Care : 150
  - Watches : 150
  - Home Decor & Festive Needs : 150
  - Kitchen & Dining : 150
  - Beauty and Personal Care : 150
  - Computers : 150

Total articles: 1050

1. 
2. 
3. 
4. 
5. 
6. 

7. 

Prétraitements textuels

- Nettoyage
  - **tokenization + lower case** : Suppression des chiffres et ponctuation + Passage en minuscule

T STAR UFT-TSW-005-BK-BR Analog Watch - For Boys Price: Rs. 399
Whether you are on your way to work or travelling abroad with family, lifestyle accessories like watches, wallets and belts help to add a touch of sophistication and class to your otherwise mundane and regular daily wear. When it all comes down to it, suave leather belts and intricately designed and finished timepieces are what separate you from the rest.
Whether you are on your way to work or travelling abroad with family, lifestyle accessories like watches, wallets and belts help to add a touch of sophistication and class to your otherwise mundane and regular daily wear. When it all comes down to it, suave leather belts and intricately designed and finished timepieces are what separate you from the rest.

['t', 'star', 'ufttswbkbr', 'analog', 'watch', 'for', 'boys', 'price', 'rs', 'whether', 'you', 'are', 'on', 'your', 'way', 'to', 'work', 'or', 'travelling', 'abroad', 'with', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'and', 'belts', 'help', 'to', 'add', 'a', 'touch', 'of', 'sophistication', 'and', 'class', 'to', 'your', 'otherwise', 'mundane', 'and', 'regular', 'daily', 'wear', 'when', 'it', 'all', 'comes', 'down', 'to', 'it', 'suave', 'leather', 'belts', 'and', 'intricately', 'designed', 'and', 'finished', 'timepieces', 'are', 'what', 'separate', 'you', 'from', 'the', 'rest', 'whether', 'you', 'are', 'on', 'your', 'way', 'to', 'work', 'or', 'travelling', 'abroad', 'with', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'and', 'belts', 'help', 'to', 'add', 'a', 'touch', 'of', 'sophistication', 'and', 'class', 'to', 'your', 'otherwise', 'mundane', 'and', 'regular', 'daily', 'wear', 'when', 'it', 'all', 'comes', 'down', 'to', 'it', 'suave', 'leather', 'belts', 'and', 'intricately', 'designed', 'and', 'finished', 'timepieces', 'are', 'what', 'separate', 'you', 'from', 'the', 'rest']

Prétraitements textuels

- Nettoyage
  - tokenization + lower case : Suppression des chiffres et ponctuation + Passage en minuscule
  - **Stopwords cleaning :** Suppression des mots inutiles (exemple : to, you…)

['t', 'star', 'ufttswbkbr', 'analog', 'watch', 'for', 'boys', 'price', 'rs', 'whether', 'you', 'are', 'on', 'your', 'way', 'to', 'work', 'or', 'travelling', 'abroad', 'with', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'and', 'belts', 'help', 'to', 'add', 'a', 'touch', 'of', 'sophistication', 'and', 'class', 'to', 'your', 'otherwise', 'mundane', 'and', 'regular', 'daily', 'wear', 'when', 'it', 'all', 'comes', 'down', 'to', 'it', 'suave', 'leather', 'belts', 'and', 'intricately', 'designed', 'and', 'finished', 'timepieces', 'are', 'what', 'separate', 'you', 'from', 'the', 'rest', 'whether', 'you', 'are', 'on', 'your', 'way', 'to', 'work', 'or', 'travelling', 'abroad', 'with', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'and', 'belts', 'help', 'to', 'add', 'a', 'touch', 'of', 'sophistication', 'and', 'class', 'to', 'your', 'otherwise', 'mundane', 'and', 'regular', 'daily', 'wear', 'when', 'it', 'all', 'comes', 'down', 'to', 'it', 'suave', 'leather', 'belts', 'and', 'intricately', 'designed', 'and', 'finished', 'timepieces', 'are', 'what', 'separate', 'you', 'from', 'the', 'rest']

['star', 'ufttswbkbr', 'analog', 'watch', 'boys', 'price', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'belts', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'comes', 'suave', 'leather', 'belts', 'intricately', 'designed', 'finished', 'timepieces', 'separate', 'rest', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'belts', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'comes', 'suave', 'leather', 'belts', 'intricately', 'designed', 'finished', 'timepieces', 'separate', 'rest']

Prétraitements textuels

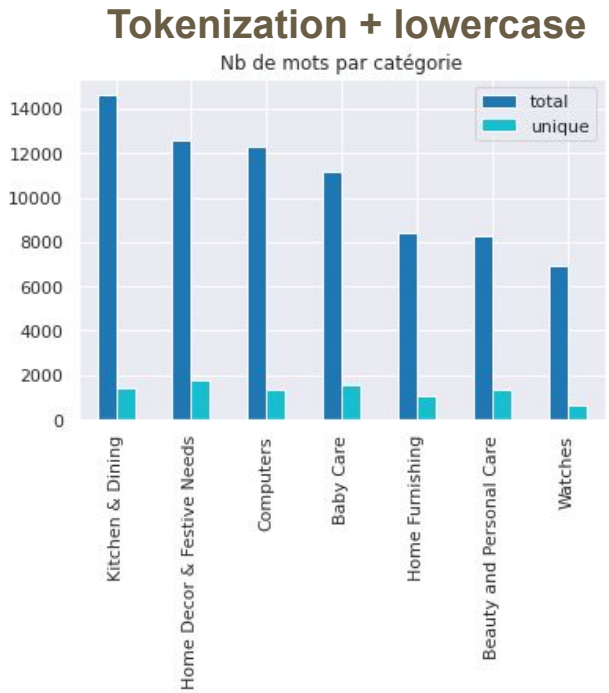
- Nettoyage
  - tokenization + lower case : Suppression des chiffres et ponctuation + Passage en minuscule
  - Stopwords cleaning : Suppression des mots inutiles (exemple : to, you...)
  - **Lemmatization :** Passage en racine des mots (exemple : is, are => be)

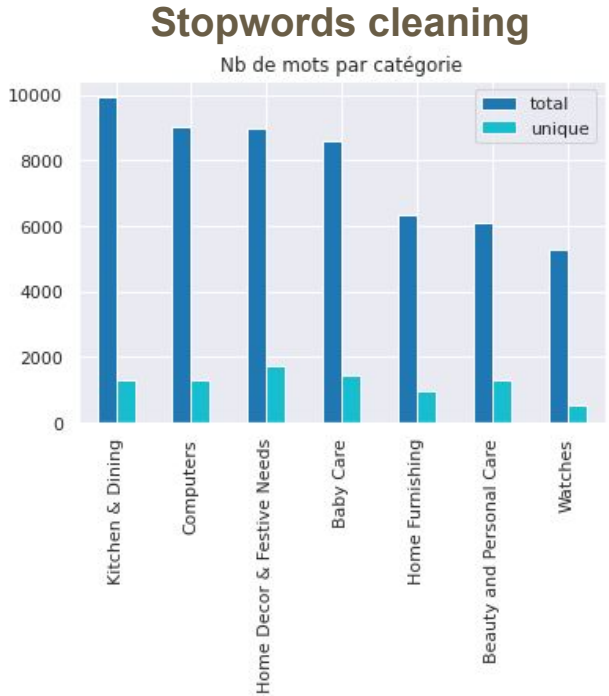
['star', 'ufttswbkbr', 'analog', 'watch', 'boys', 'price', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'belts', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'comes', 'suave', 'leather', 'belts', 'intricately', 'designed', 'finished', 'timepieces', 'separate', 'rest', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessories', 'like', 'watches', 'wallets', 'belts', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'comes', 'suave', 'leather', 'belts', 'intricately', 'designed', 'finished', 'timepieces', 'separate', 'rest']

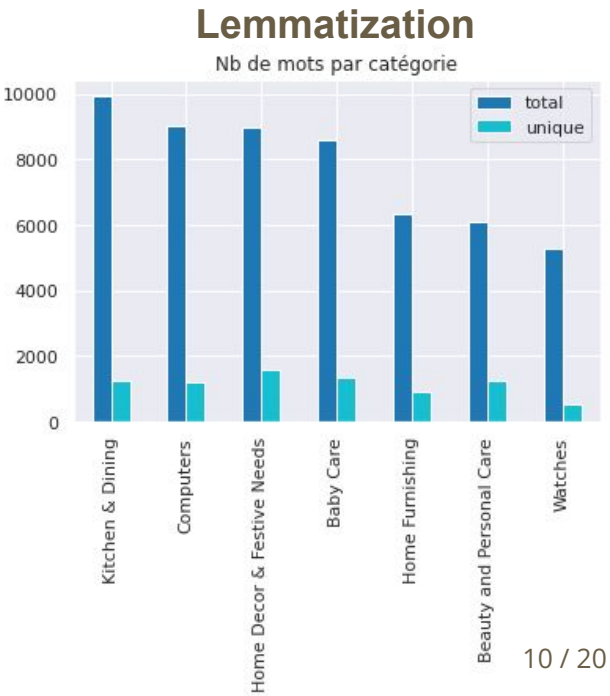
['star', 'ufttswbkbr', 'analog', 'watch', 'boy', 'price', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessory', 'like', 'watch', 'wallet', 'belt', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'come', 'suave', 'leather', 'belt', 'intricately', 'designed', 'finished', 'timepiece', 'separate', 'rest', 'whether', 'way', 'work', 'travelling', 'abroad', 'family', 'lifestyle', 'accessory', 'like', 'watch', 'wallet', 'belt', 'help', 'add', 'touch', 'sophistication', 'class', 'otherwise', 'mundane', 'regular', 'daily', 'wear', 'come', 'suave', 'leather', 'belt', 'intricately', 'designed', 'finished', 'timepiece', 'separate', 'rest']

Prétraitements textuels

- Nettoyage : résultats







Prétraitements textuels

- Nettoyage : résultats
  - Top 10 des mots les plus employés

Tokenization + lowercase	Stopwords cleaning	Lemmatization
'of': 1753	'products': 635	'product': 861 ↑
'for': 1433	'free': 612	'free': 612
'the': 1360	'buy': 581	'buy': 581
'and': 1332	'delivery': 567	'cm': 567 ↑
'to': 1055	'genuine': 564	'delivery': 567
'in': 1047	'shipping': 564	'genuine': 564
'rs': 913	'cash': 564	'shipping': 564
'only': 888	'replacement': 559	'cash': 564
'with': 842	'day': 538	'price': 559 ↑
'on': 832	'cm': 531	'replacement': 559

Prétraitements textuels

- Fréquence des mots
  - TF-IDF : “Term Frequency — Inverse Document Frequency”
    - évaluer l'importance d'un terme contenu dans description, relativement à une collection
      - Fit => {'key': 2451, 'feature': 1667, 'elegance': 1457, ...} => 5153 features
      - Fit_transform=>1050x5153 matrix
      - Réduction par T-SNE=> 1050x2 matrix
      - Coloriage par category_1

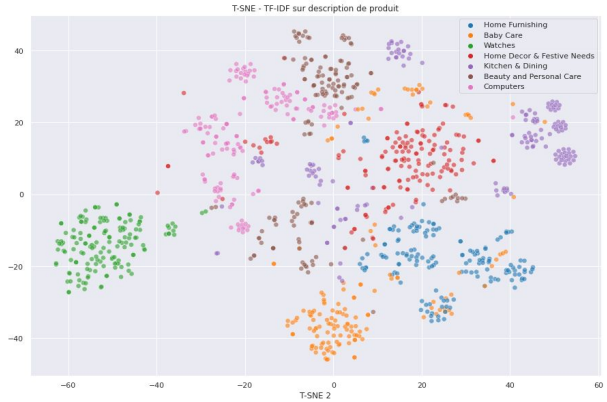
Prétraitements textuels

- Classification
  - modélisation des sujets avec des méthodes non supervisées (entrée : tfidf.fit_transform 1050x5153 matrix, sortie : 1050x7 matrix)

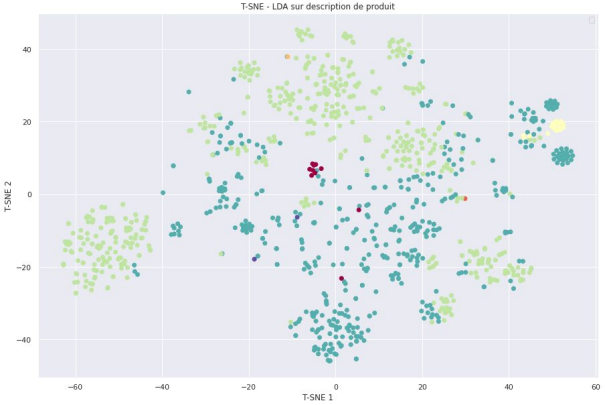
    - **LDA : ”Latent Dirichlet Allocation”**
    - topic 0: stage install tshirtspecifications abkl_grn_grn_grn wetdry potency dependsview front traveller nihar
    - topic 1: unbreakable flaunt shot app avi upasana kushies suttle brought bottlepatented
    - topic 2: cutlery highest mental shoo visual snack modernistic floralina serf eye
    - topic 3: afterspecifications forever sunlast micro bezel demandposterchacha duty cycle n taklon
    - topic 4: loose bergner hence night mathematics tip kth anytime polo blanket
    - topic 5: de vgncrelsmartpro serve one leggingsbabyoye geforce afterspecifications sle organic wallpaper
    - topic 6: sized spiritual lushomes vgncrglsmartpro carriage tucked catarrh hd airtexdongli inputoutput

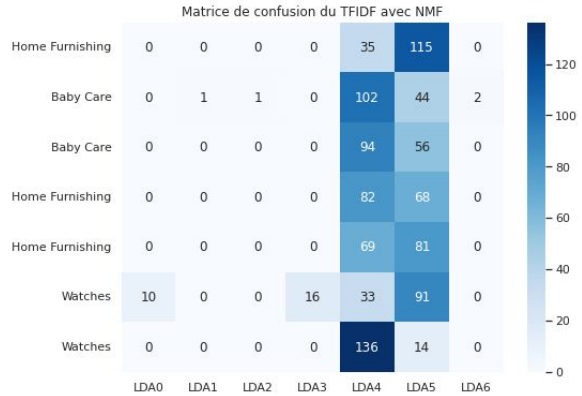
    - **NMF : ”Negative Matrix Factorisation”**
    - topic 0: loose polo hence night bergner mathematics blanket tip kth dosa
    - topic 1: afterspecifications forever micro demandposterchacha sunlast mesh bezel ability malhar face
    - topic 2: antimicrobial vary workspecifications frame extravagant error friendly strain vapor theme
    - topic 3: one even antique hop sanganeri melangespecifications highlighter value review equinox
    - topic 4: serve qp giorgio lantern vase spiciness geforce vintage leggingsbabyoye half
    - topic 5: balanced afterspecifications forever chicken onesspecifications forlarge gen wrought prisha superheroes
    - topic 6: high lenco comforter xbluetooth loose jacquard night hence bergner tip

Résultats textuels

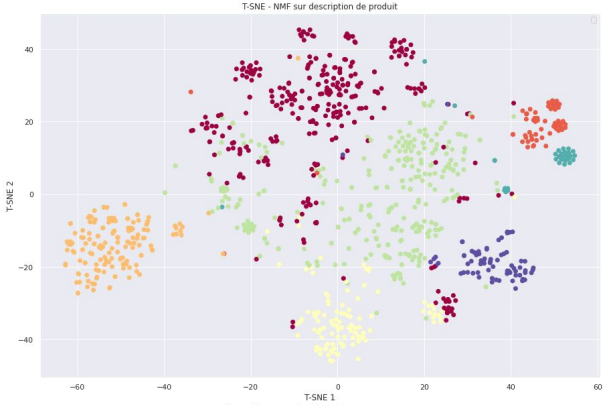


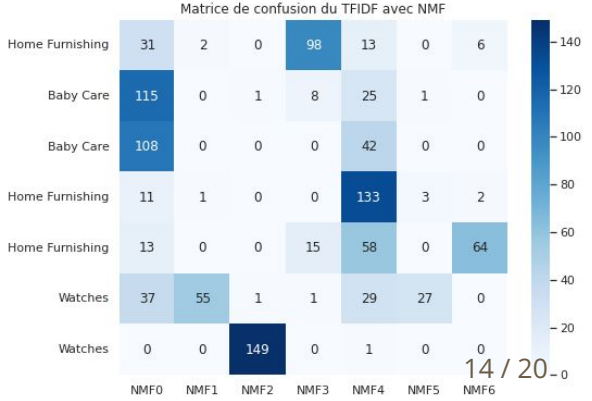
LDA
Score ARI : 0.0547



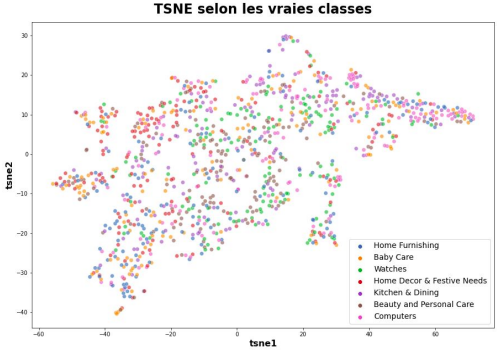


NMF
Score ARI : 0.3527



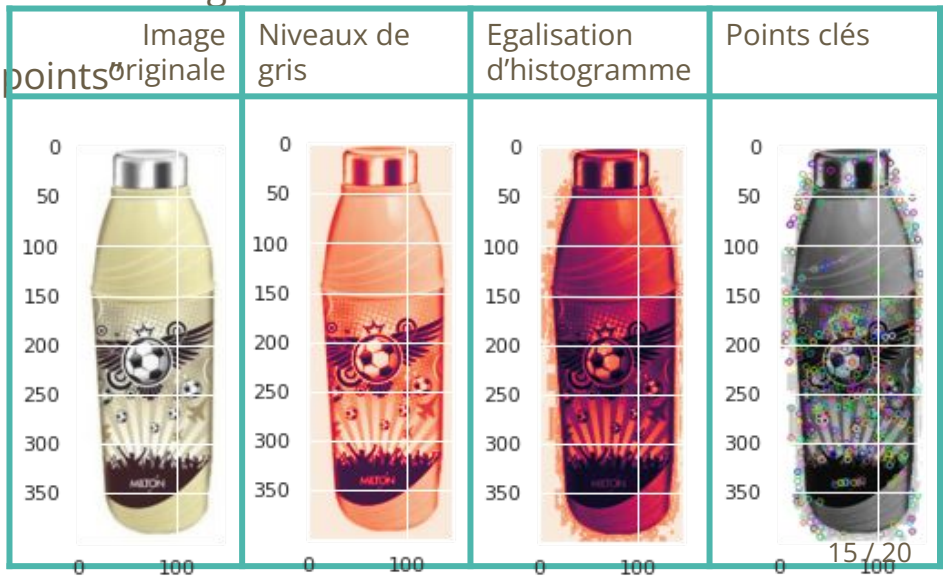


Prétraitements visuels



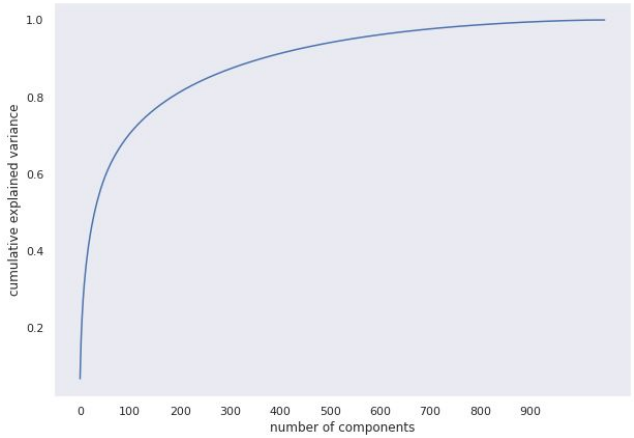
- SIFT : Scale Invariant Fourier Transform
  - Traitement en niveaux de gris : chaque pixel est représenté par un niveau de lumière
  - Egalisation d'histogramme : amélioration du contraste
  - Détection des "key points" : reconnaissance de l'image avec échelle et orientation différentes
  - Descripteurs : caractéristiques des "key points"
  - Features d'images : histogrammes
    - répartition des orientations
  - Réduction de dimensions 1050x1016
    - PCA puis t-SNE

descripteurs : (602, 128)
[[0. 0. 1. ... 6. 3. 3.]
[45. 0. 0. ... 1. 0. 0.]
[121. 6. 1. ... 0. 0. 0.]
...
[14. 2. 3. ... 70. 0. 0.]
[186. 1. 0. ... 0. 0. 0.]
[96. 137. 0. ... 1. 5. 3.]]



Prétraitements visuels

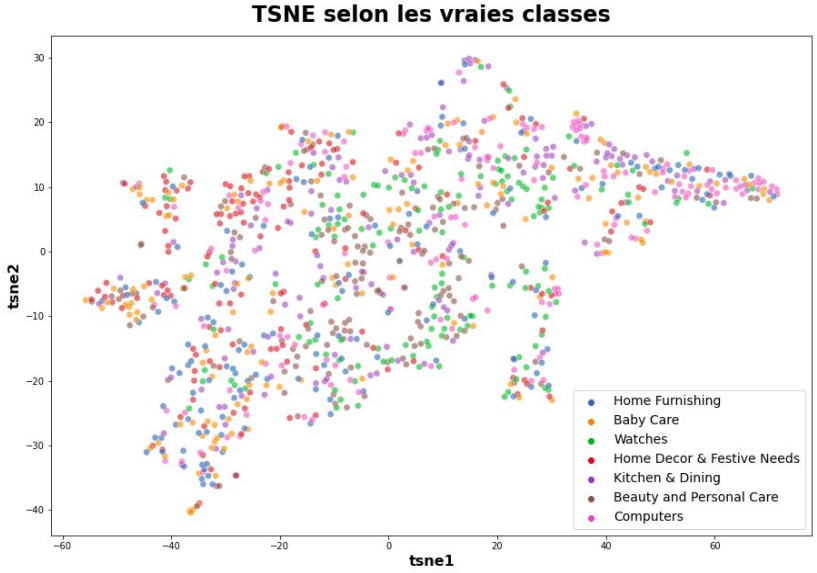
- CNN transfer learning : Transfer Learning with Convolutional Neural Networks
  - Entrainement du modèle VGG16 sur les images + prediction
  - Réduction de dimensions par PCA et T-SNE de 1050x4096 en 1050x2
  - Classification KMeans sur dimensions réduites
  - Correspondance manuelle clusters et category_1

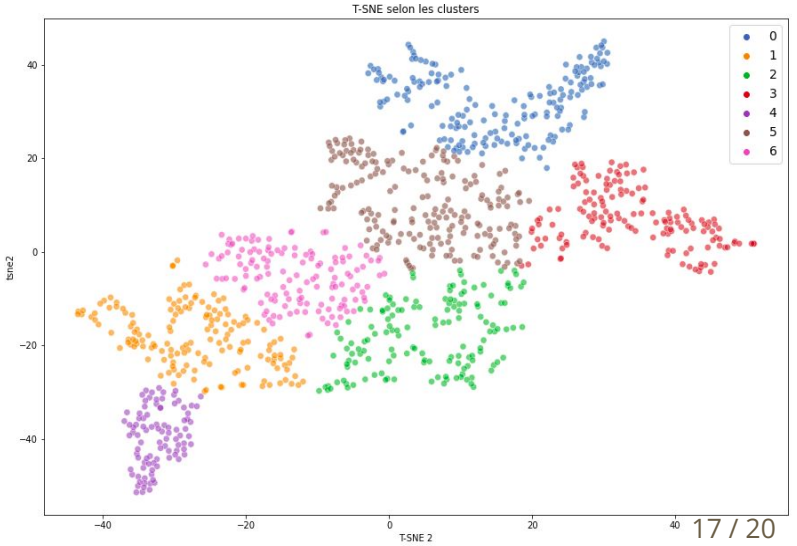


Résultats visuels

- SIFT

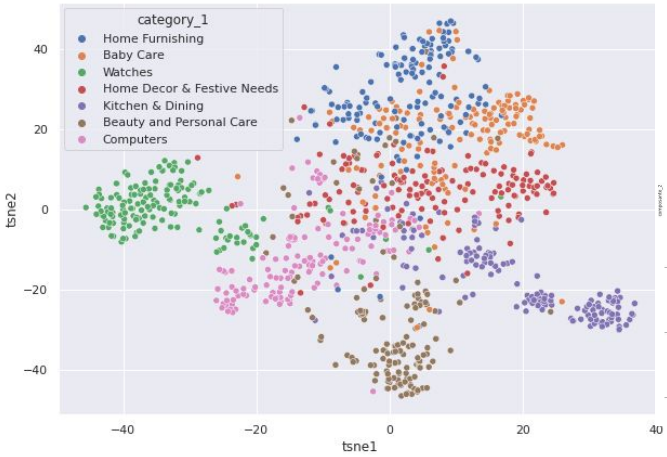
SIFT
Score ARI : 0.0457

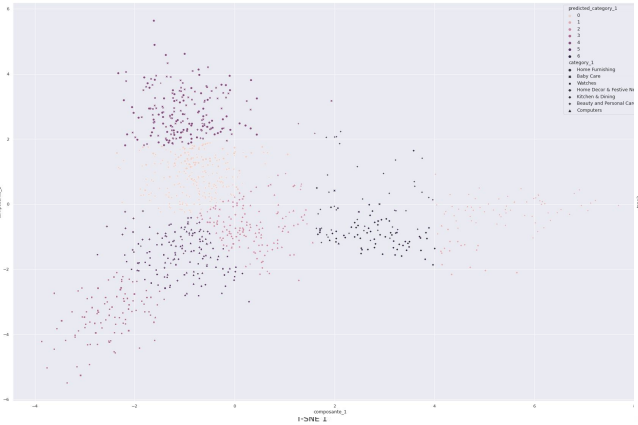




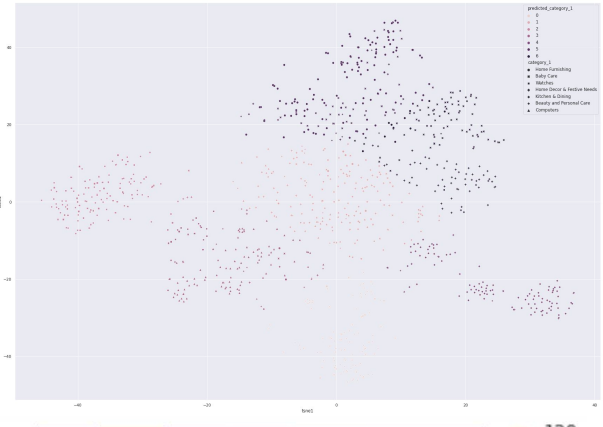
Résultats visuels

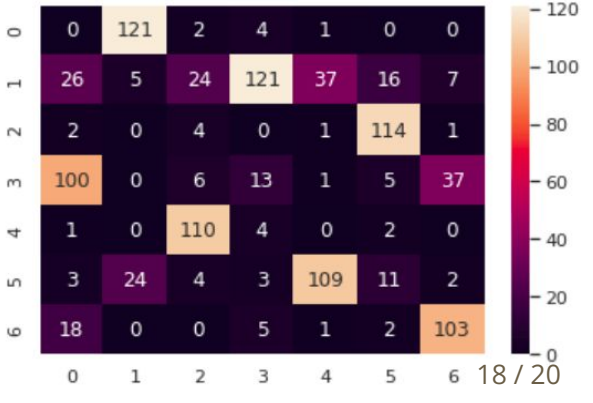
VGG16 - Kmeans - PCA
Score ARI : 0.2919





VGG16 - Kmeans - T-SNE
Score ARI : 0.4756





Faisabilité du moteur de classification

- Moteur de classification faisable par un apprentissage supervisé
  - Modèle linéaire de classification de type régression logistique
  - Modèle de classification ensembliste de type forêt aléatoire
  - Modèle de réseau de neurones de classification de type perceptron multicouches (MLP)
- Recommandations
  - Prétraitement textuel
    - correction orthographique
    - ajout / détection de tags
  - Prétraitement visuel
    - détection de flou

Conclusion

- Natural Language Processing
  - Solutions très variées,
  - Bibliothèques bien fournies
  - Champs d'application diversifiés
- Classification d'images
  - Solutions très différentes à mettre en oeuvre
  - Utilisation en fonction des besoins et des contraintes