Soutenance Projet 6

Classifiez automatiquement des biens de consommation

OPENCLASSROOMS - Formation Data Scientist

Michel Blazevic

Ordre du jour

1. Introduction
    1. Rappel problématique
    2. Présentation Bases de données
2. Données textuelles
    1. Bag-of-words
    2. Word/sentence embeddings
3. Données visuelles
    1. Bag-of-visual words
    2. Transfer-learning CNN
4. Conclusion & Aller plus loin

Introduction

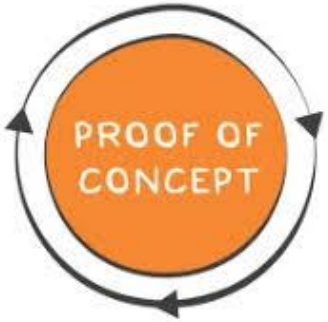
- Rappel problématique
- Descriptions des deux jeux de données

Rappel Problématique

Problématique:

- ***Place de marché***: Marketplace proposant divers produits
- Les clients déposent leur article avec une description et une photo
- Catégories de l'article ?
  Assigné à la main par le vendeur → Chronophage, pas précis…

Mission:



⇒ Etude de faisabilité d'un système de classification automatique

⇒ Découvrir les données (textuelles et visuelles)

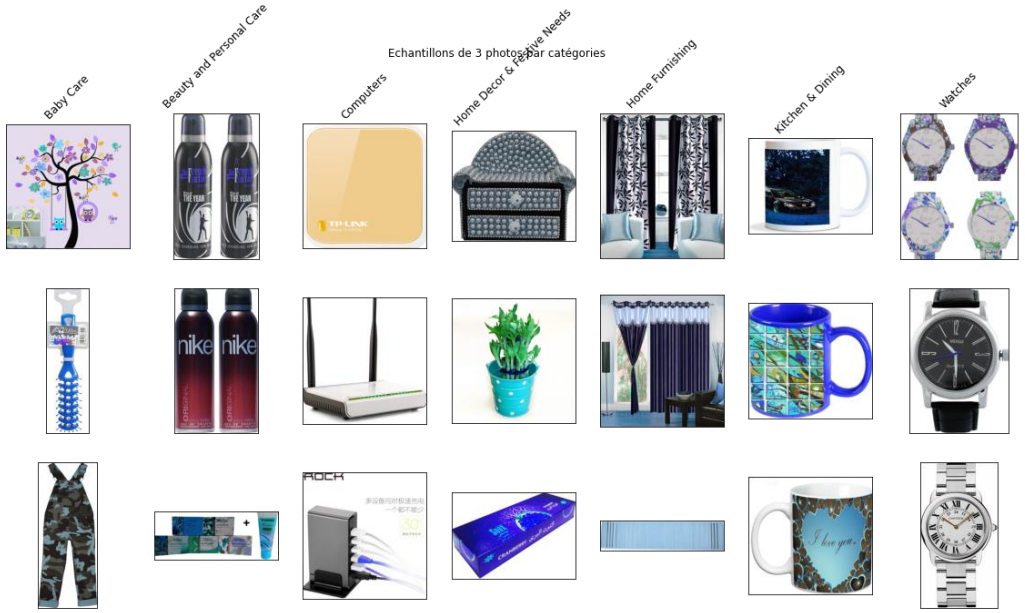
⇒ Recherche de similarité entre les produits et comparaison avec les catégories connues

Bases de données

15 variables: lien, info produits (description, marque, prix…), nom image…

1050 entrées (produits)

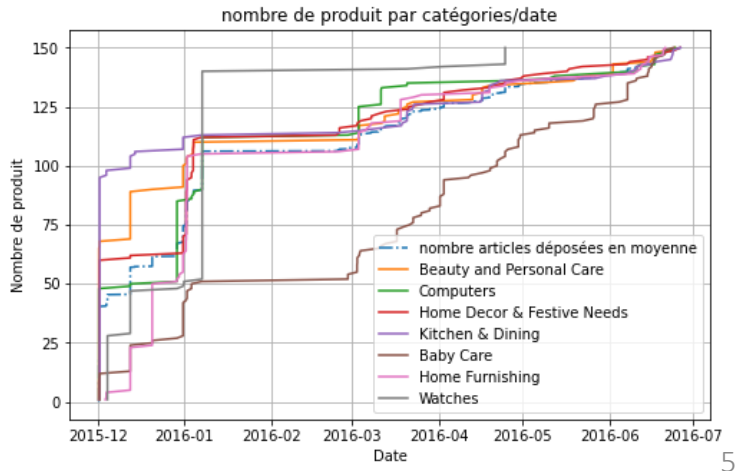
2% de valeurs manquantes
12 strings, 2floattants, 1 booléen



Catégories des produits :

=> 7 catégories principales + sous catégories

cat_1	cat_2	count
Baby Care	9	4
Beauty and Personal Care	11	5
Computers	8	4
Home Decor & Festive Needs	10	5
Home Furnishing	12	4
Kitchen & Dining	11	4
Watches	2	3



Données textuelles

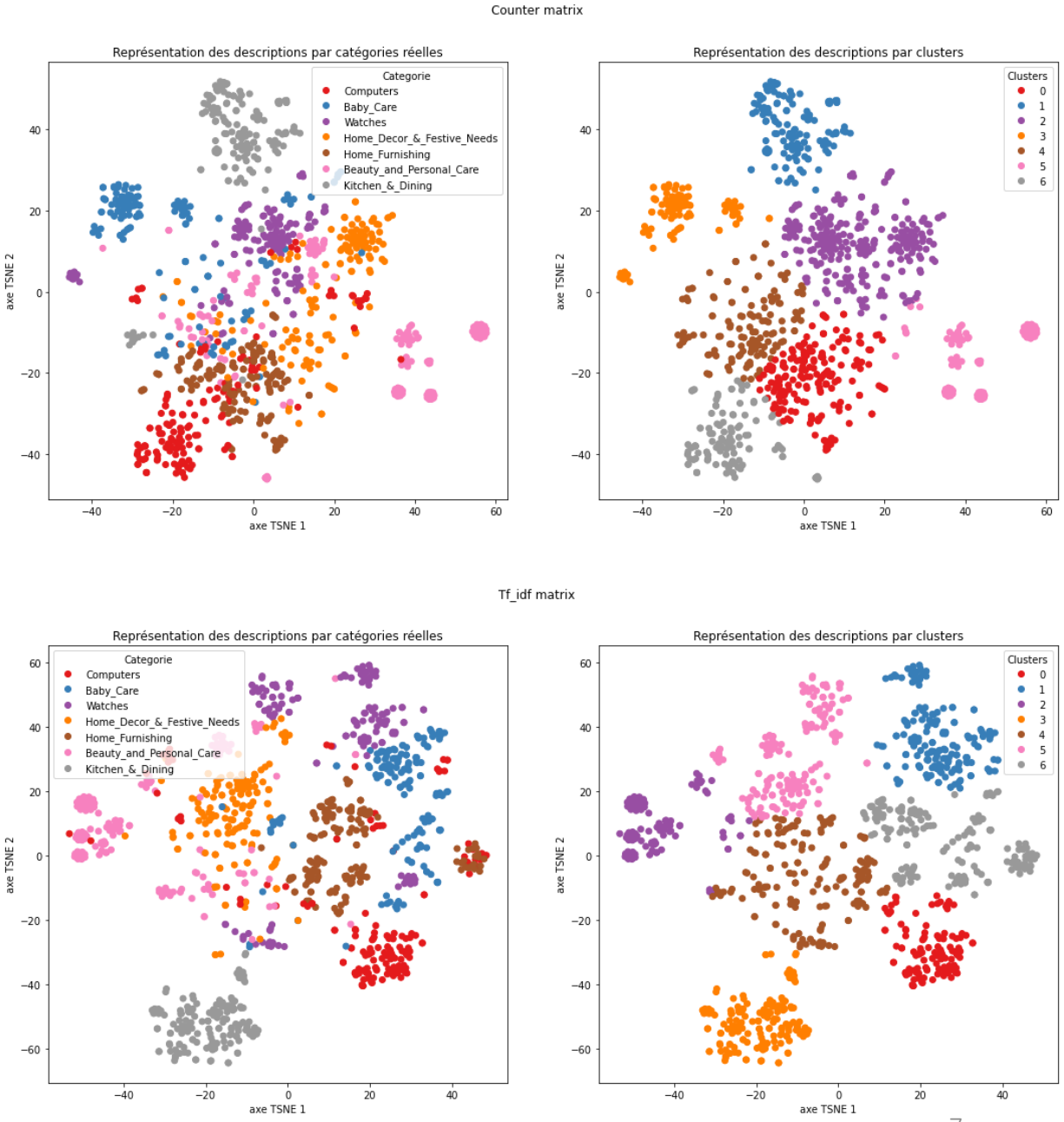
- Bag of words (Counter et Tf-idf)
- LDA et NMF
- Word embeddings:
  - Word2Vec
  - Bert
  - USE

Bag-of-words

- Bag-of-words ?
- Document-term matrix ?
- Preprocessing:
  - Minuscule
  - les stopwords et les liens web, ponctuations, nombres
  - Tokenisation et Lemmatisation
  - Garder mots >2
- Count et Tf-idf ?

(5324mots)

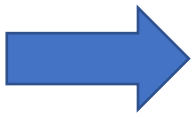
```
CountVectorizer :
-----------------
ARI :  0.402 time :  9.0

Tf-idf :
--------
ARI :  0.4011 time :  8.0
```



LDA & NMF

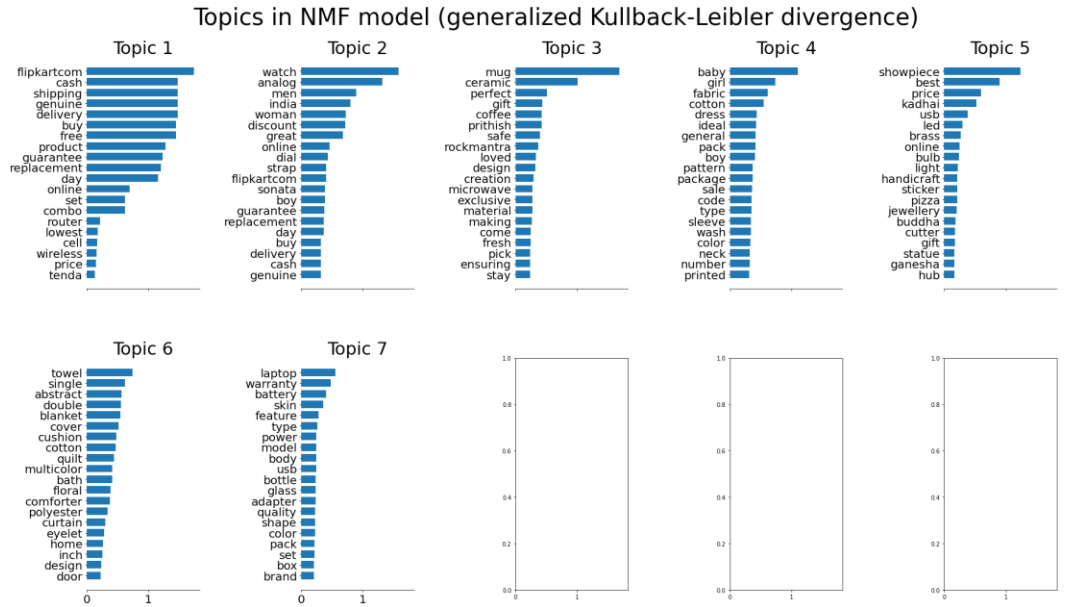
Algorithmes permettant de réunir les documents par sujet (topic)

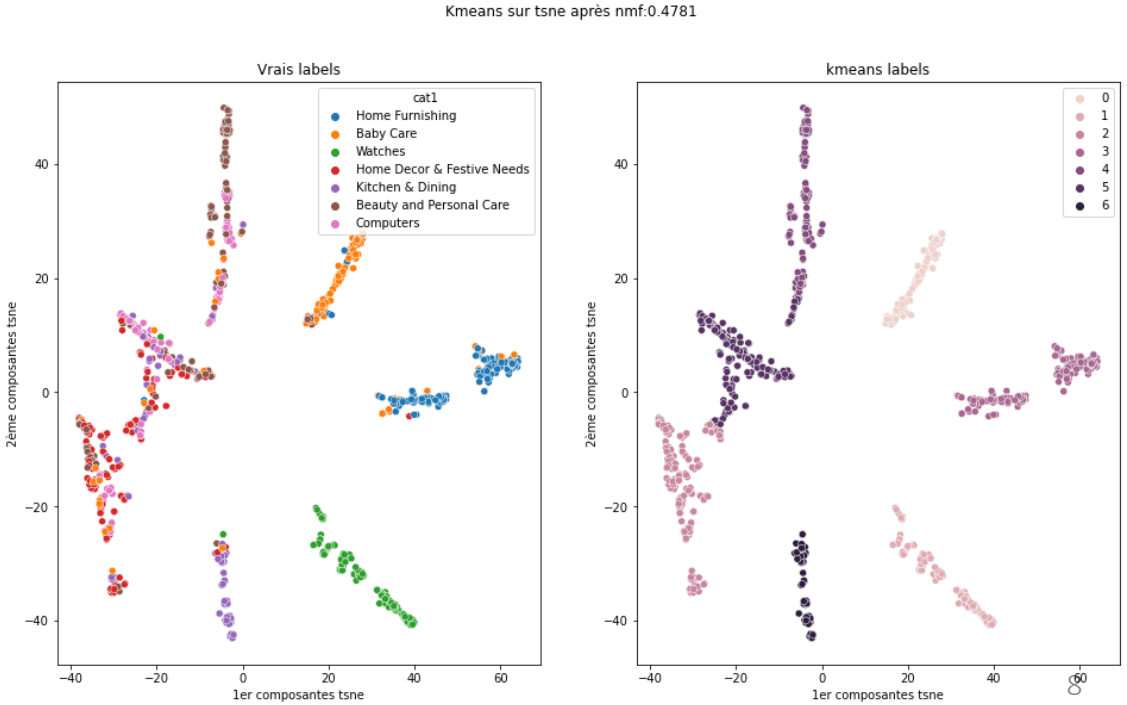
- Preprocessing:
  - Minuscule
  - les stopwords et les liens web, ponctuations, nombres
  - Tokenisation et Lemmatisation
  - Garder mots >2
- Tf-Idf (Count pour LDA) (1000mots)



LDA moins intéressant que NMF

Résultat ARI 0.47





Word Embeddings

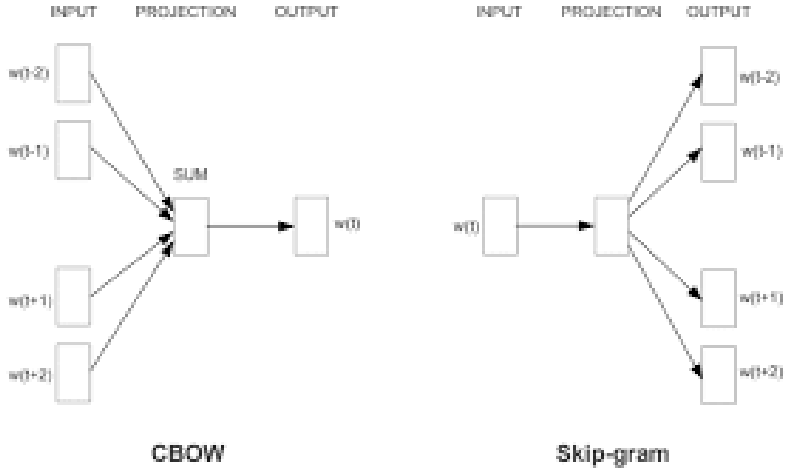
Problème du Bag-of-Words:

- Ne s'intéresse pas aux sens des mots
- Observation de la similarité de la composition de chaque document
- Donne des matrices creuses (beaucoup de zéros)

Solution sémantique:

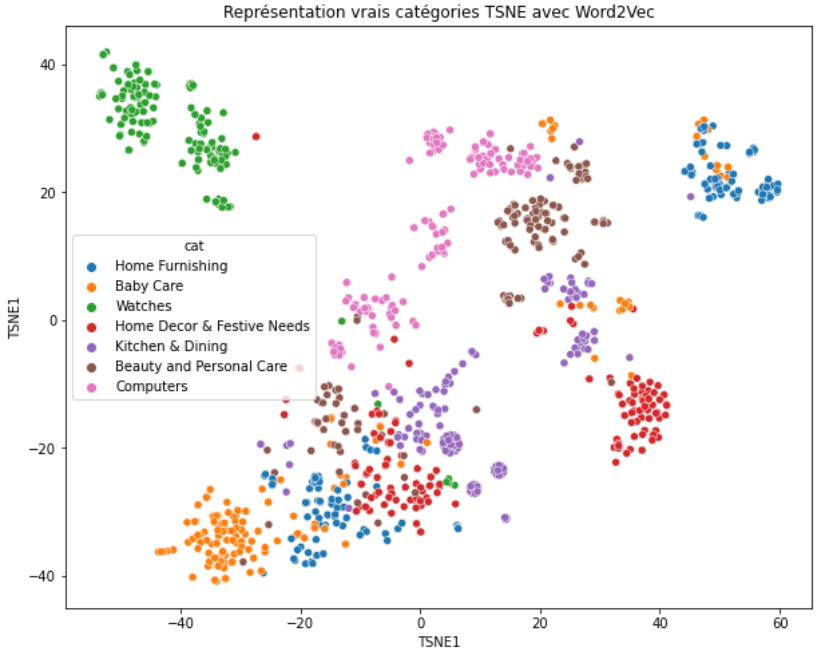
- Prendre le contexte des mots pour approcher la sémantique des mots
- Permet de voir les sujets s'ils sont proches dans l'espace grâce à un plongement de mot (word embeddings)
- Vecteurs dense et de taille similaire
- Word2Vec, GLoVe, Fasttext, Transformers (BERT), Universal Sentence Encoder…

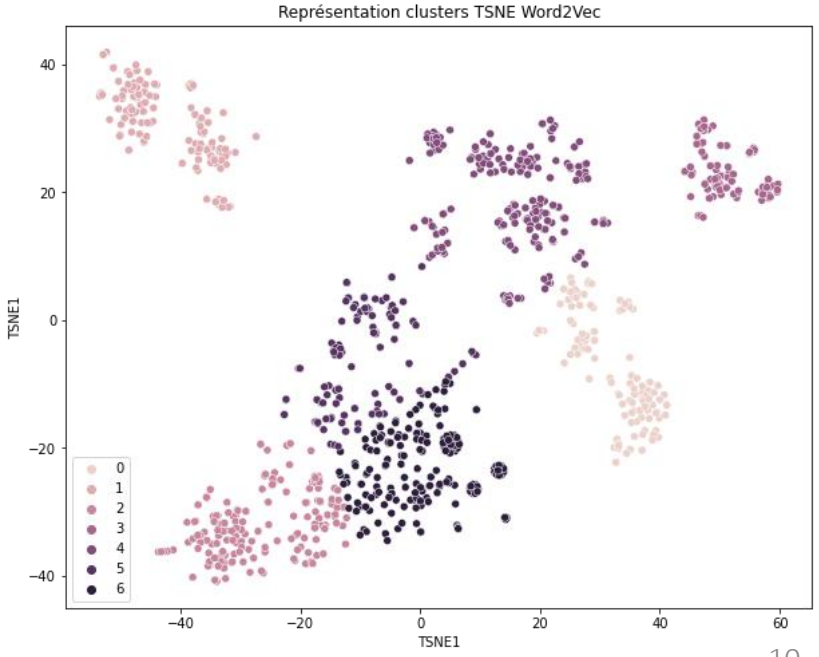
Word2Vec



- Utilisation d'un modèle pré entraîné: word2vec-google-news-300
- Permet de « vectoriser » des tokens sous des vecteurs de longueurs 300

ARI : 0.4158

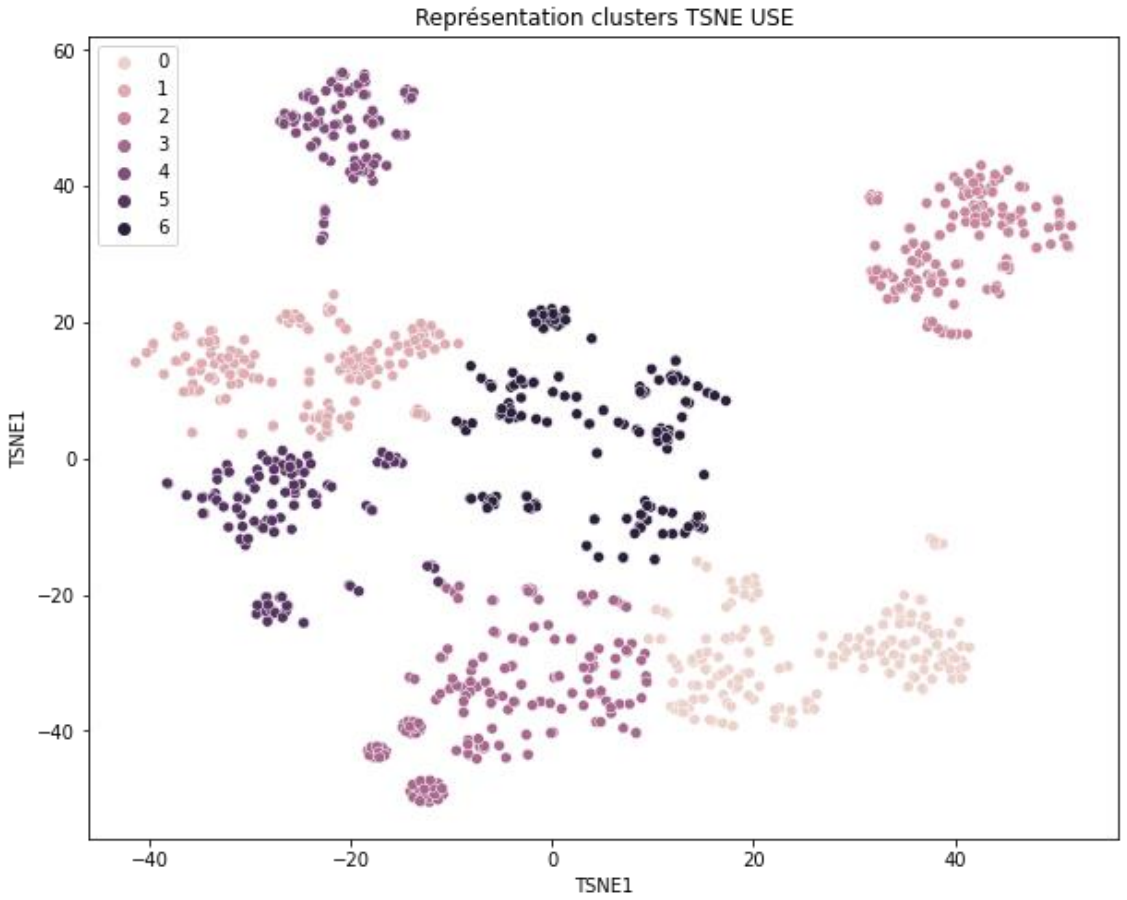


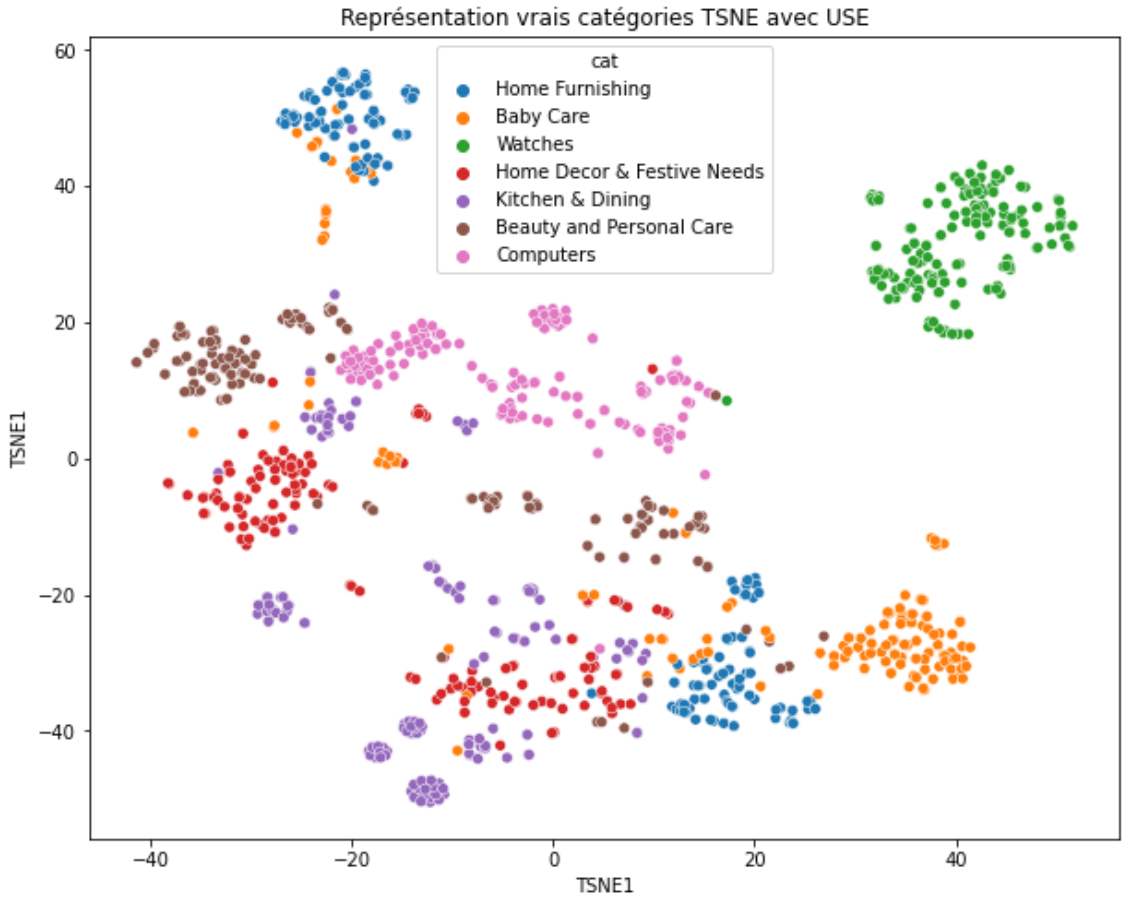


Universal Sentence Encoder

Méthode avec un embeddings de taille 512
Accessible avec tensorflow_hub

ARI: 0.4526



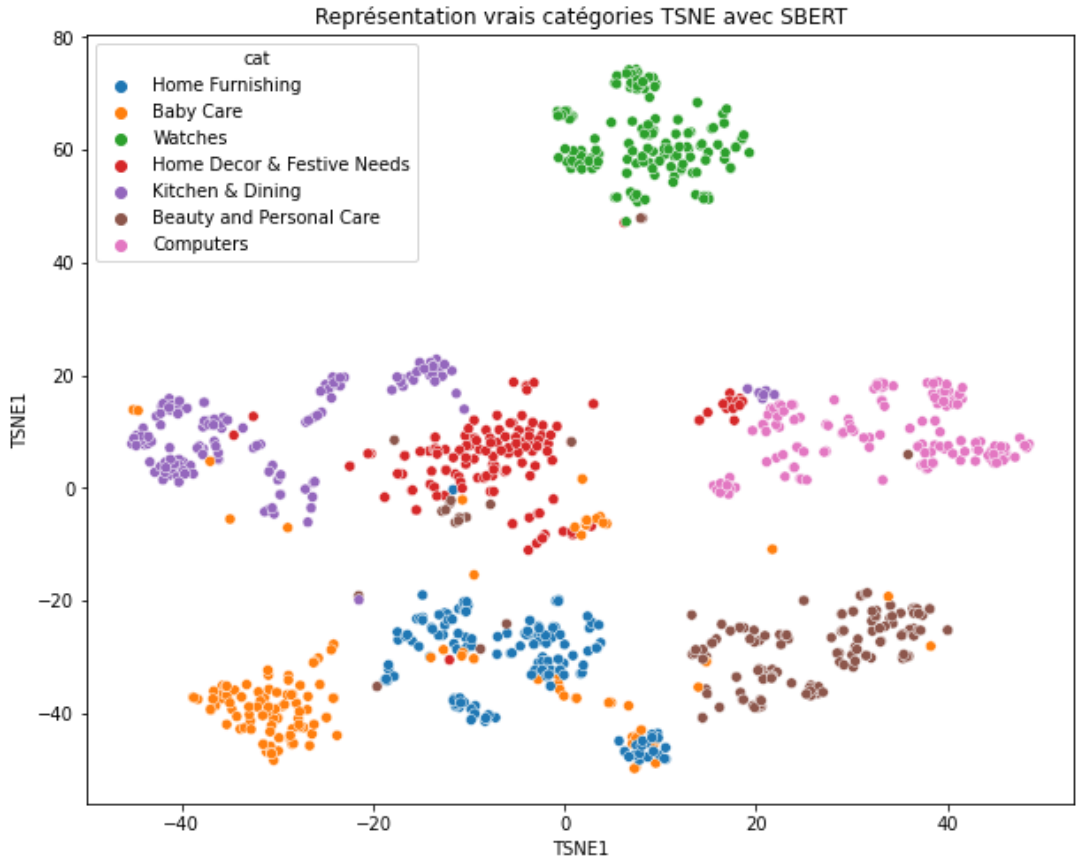


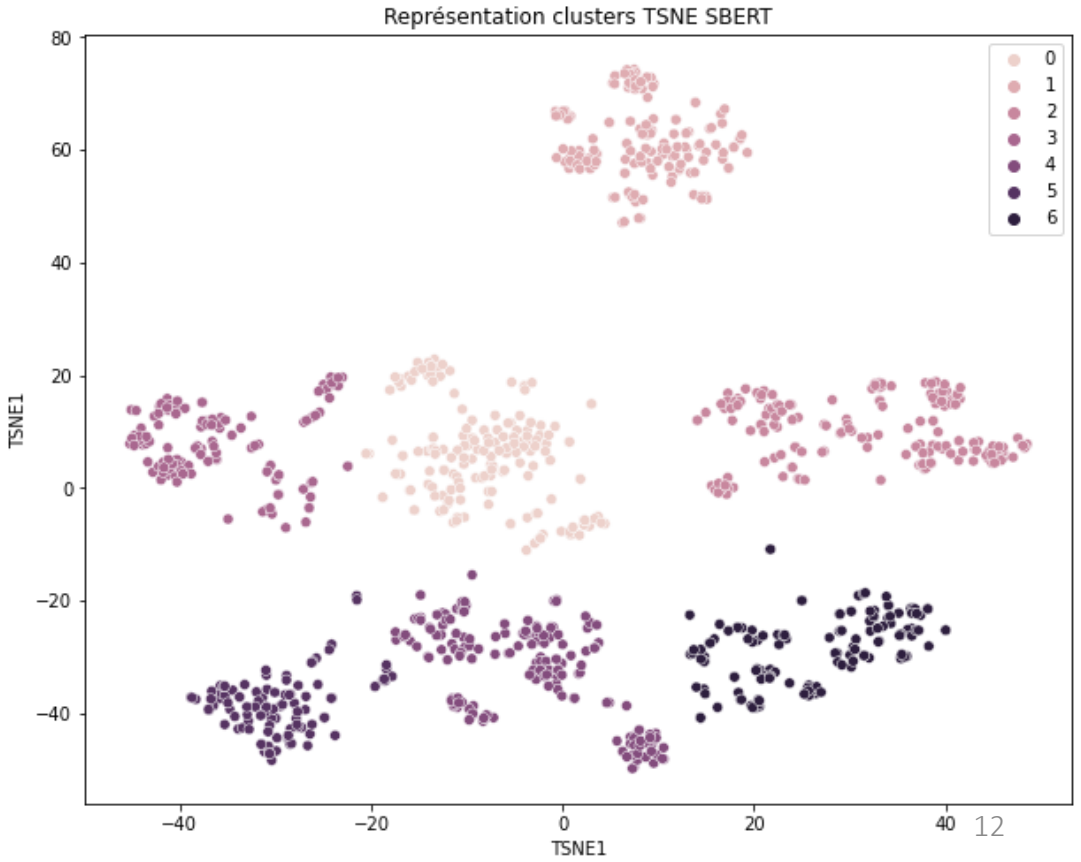
BERT

Plongement en vecteur 768

Utilisation sentence_transformers de huggingface

ARI : 0.7391





Données visuelles

- SIFT: établissement des descripteurs

- Transfer-learning, deux tests: un avec pattern détecté et un autre avec la liste des différentes catégories par un réseau déjà entrainé et une analyse avec réduction de dimension

Algorithme SIFT

Scale-invariant feature transform

Trouver des descripteurs (zones caractéristiques):

- ✓ Invariant aux changements d'échelle
- ✓ Invariant aux rotations
- ✓ Invariant aux translations
- ✗ Sensible aux changements de luminosité
- ✗ Sensible aux points de vue 3D

-> chaque descripteurs est un vecteur dimension 1x128

Représentation descripteurs

image sans rien

image en niveau de gris

histogram du niveau de gris

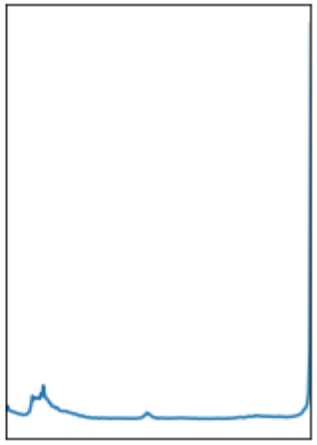
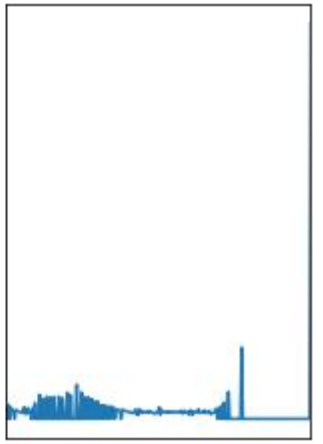


image après égalisation

histogram après égalisation

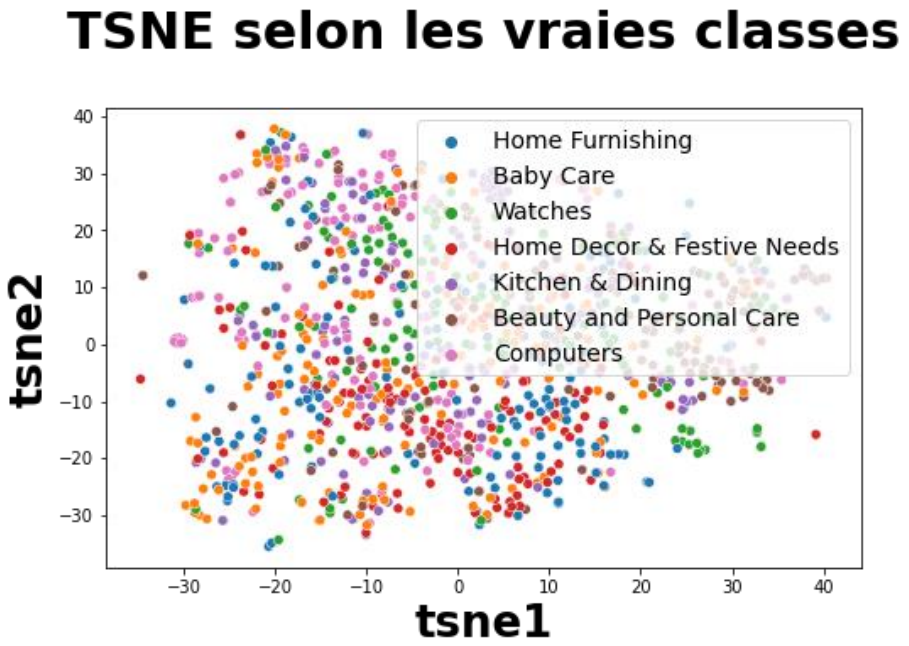


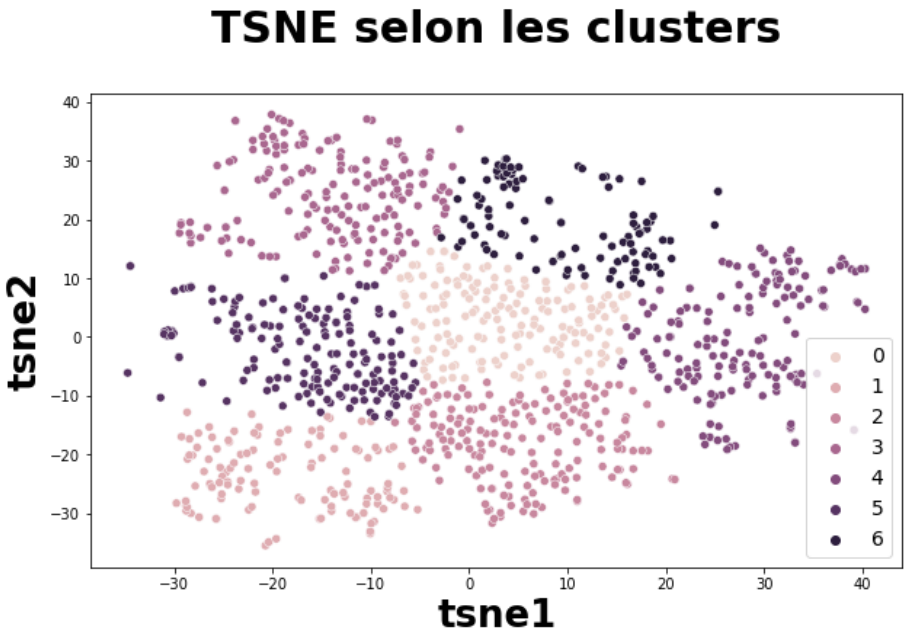
Algorithme SIFT

Scale-invariant feature transform

- Recherche des descripteurs pour chaque image
- Faire un clustering de descripteurs avec $\sqrt{n}$ où n nombre total de descripteurs
- Création d'un histogramme par image de bag-of-visual-word
- Réduction de dimension par PCA (560=>425) en gardant 99% de l'info puis réduction TSNE sur 2 composantes

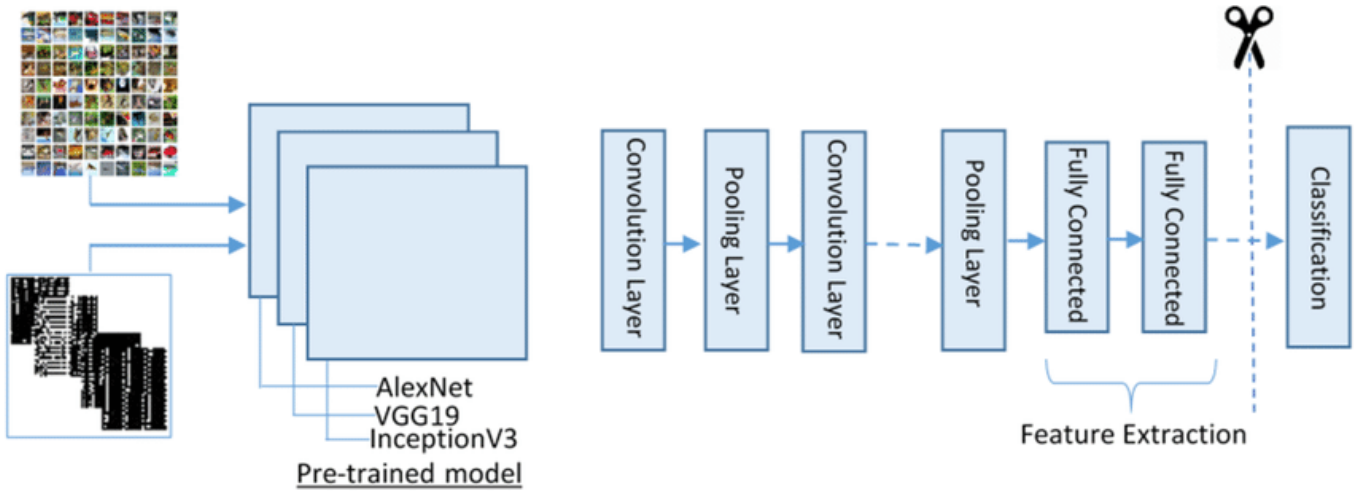
ARI : 0.0622





Transfer Learning

VGG16 & InceptionResNetV2

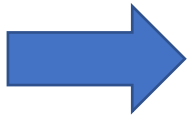


- Les entrées du modèles sont de taille (299,299,3) et (224,224,3)
- Modèles entrainés avec ImageNet (14 M images labelisées) avec 1000 catégories
- On garde toutes les couches entrainées sauf la dernière
- Les dernières couches donnent accès à une extraction de features

Transfer Learning

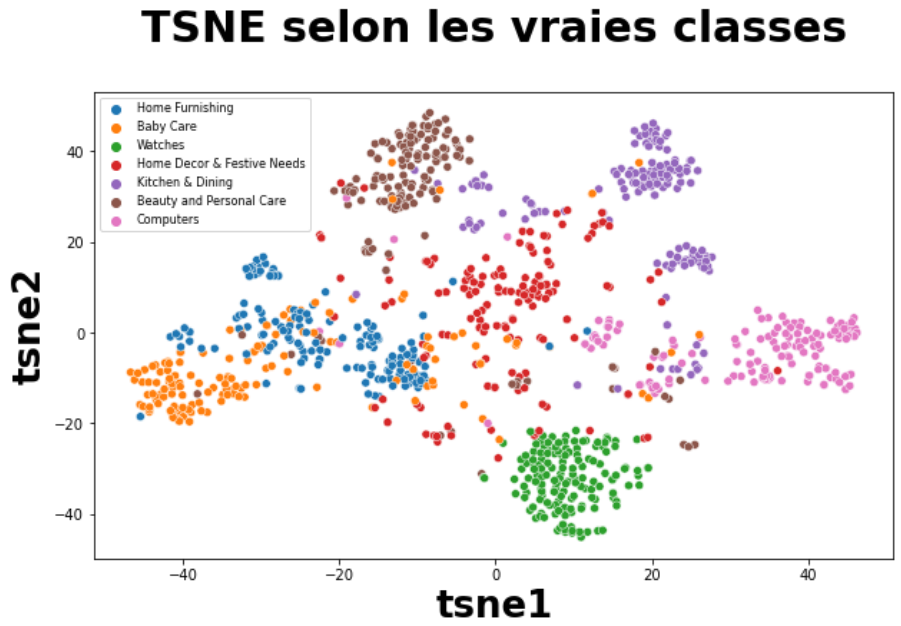
VGG16 & InceptionResNetV2

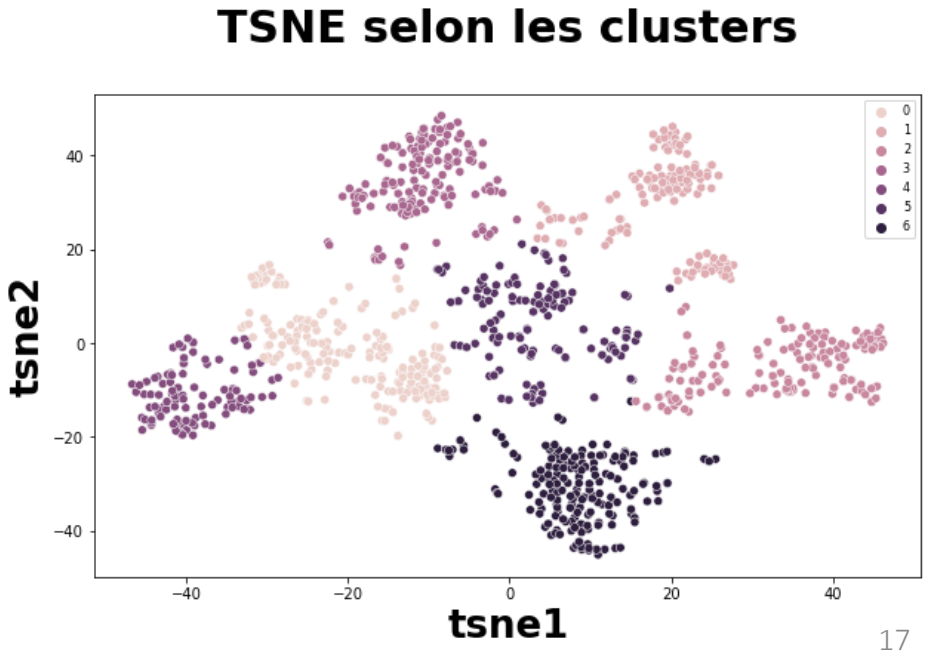
- Loader chaque image avec la taille d'entrée du CNN
- Passer chaque image dans le CNN et récupérer le vecteur contenant l'extraction de features
- Réduction de la dimension avec PCA et T-SNE
- Clustering avec Kmeans sur les 7 catégories souhaités
- Observation et calcul ARI entre classe réelle et cluster



ARI VGG16: 0.4592

ARI InceptionResNetV2: 0.5974





Conclusion

- Conclusion sur le travail réalisé
- Aller plus loin

Conclusion

- Découverte des données de type textuelle et visuelle
- Utilisation de différentes méthode d’analyse de caractéristique
- Etude de faisabilité pour une approche non supervisée de classification
- Obtention de résultat concluant sur la faisabilité avec une comparaison entre des clusters non supervisés et les labels réels

Pour aller plus loin

- Continuer de faire différents preprocessing sur les données:
  - Filtrage sur les données visuelles…
  - Choix plus poussé de stopwords, différentes tokenisations…

- Continuer d'explorer les différents modèles avec les variantes (différents BERT, CNN etc…)
- Compléter le jeu de données:
  - Recherche de document textuelle lié à la vente d'article pour créer un corpus plus grand et entraîner un modèle avec plus de données
  - Même avec les images

Merci!
Des questions ?