OPENCLASSROOMS

Kevin

Parcours Data Scientist

Problématique

L’entreprise

- Marketplace e-commerce

Objectifs de la marketplace

- Rencontre entre vendeurs et acheteurs
- Demande d’informations concernant les produits en vente
- Erreurs humaines éventuelles
- Développer un avantage concurrentiel et une meilleure satisfaction client

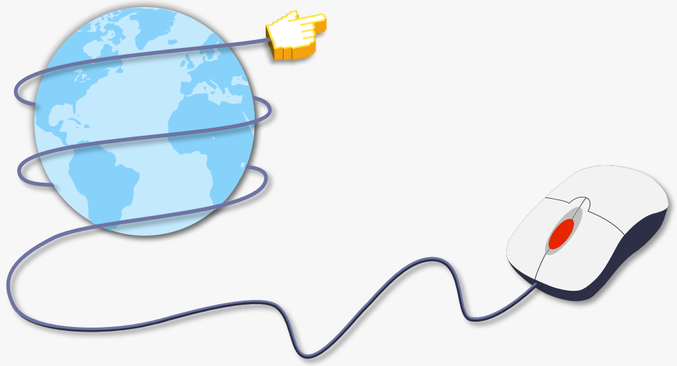
Mission

- Etudier la faisabilité de l’automatisation et sa précision
- Catégorisation via description/image
- Des contraintes dans le process…

Jeu de données

Process

Faisabilité / Recommandations



1

Jeu de données

Fichier Data

Contenu

1 fichier CSV
Dossier image

Données

1050 produits
15 variables

Variables

Nom du produit
Date sur le marché
Prix de vente
Description
Image
Arbre de Catégorie

Catégorie

7 catégories
150 produits

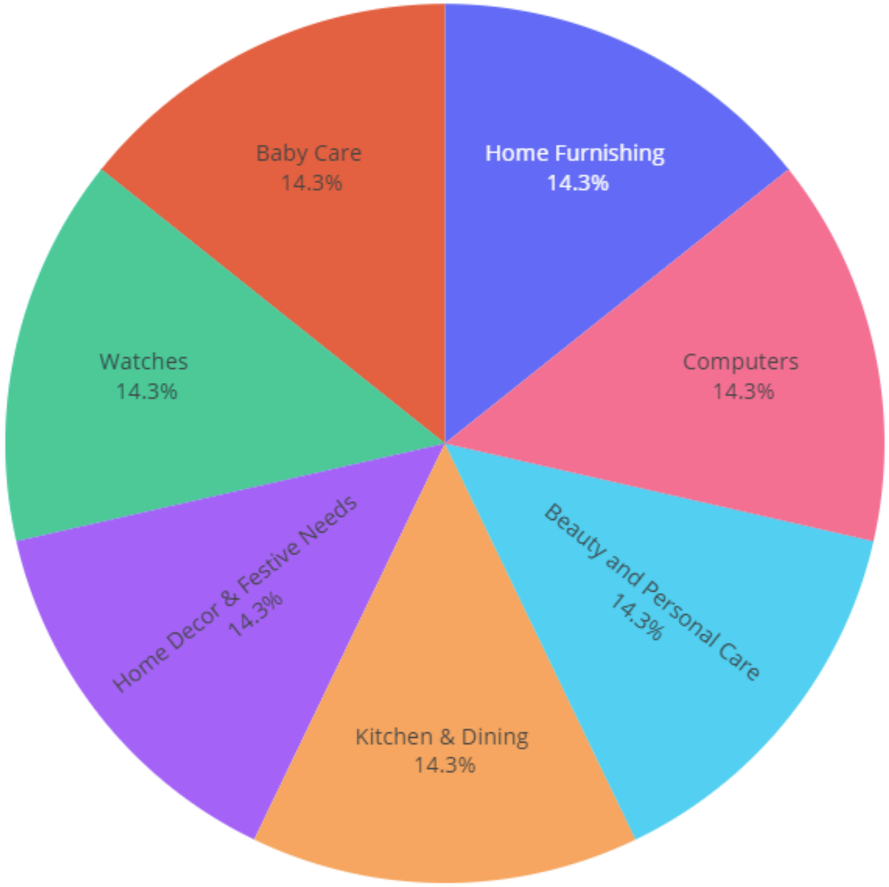
Données manquantes

Dataset de qualité

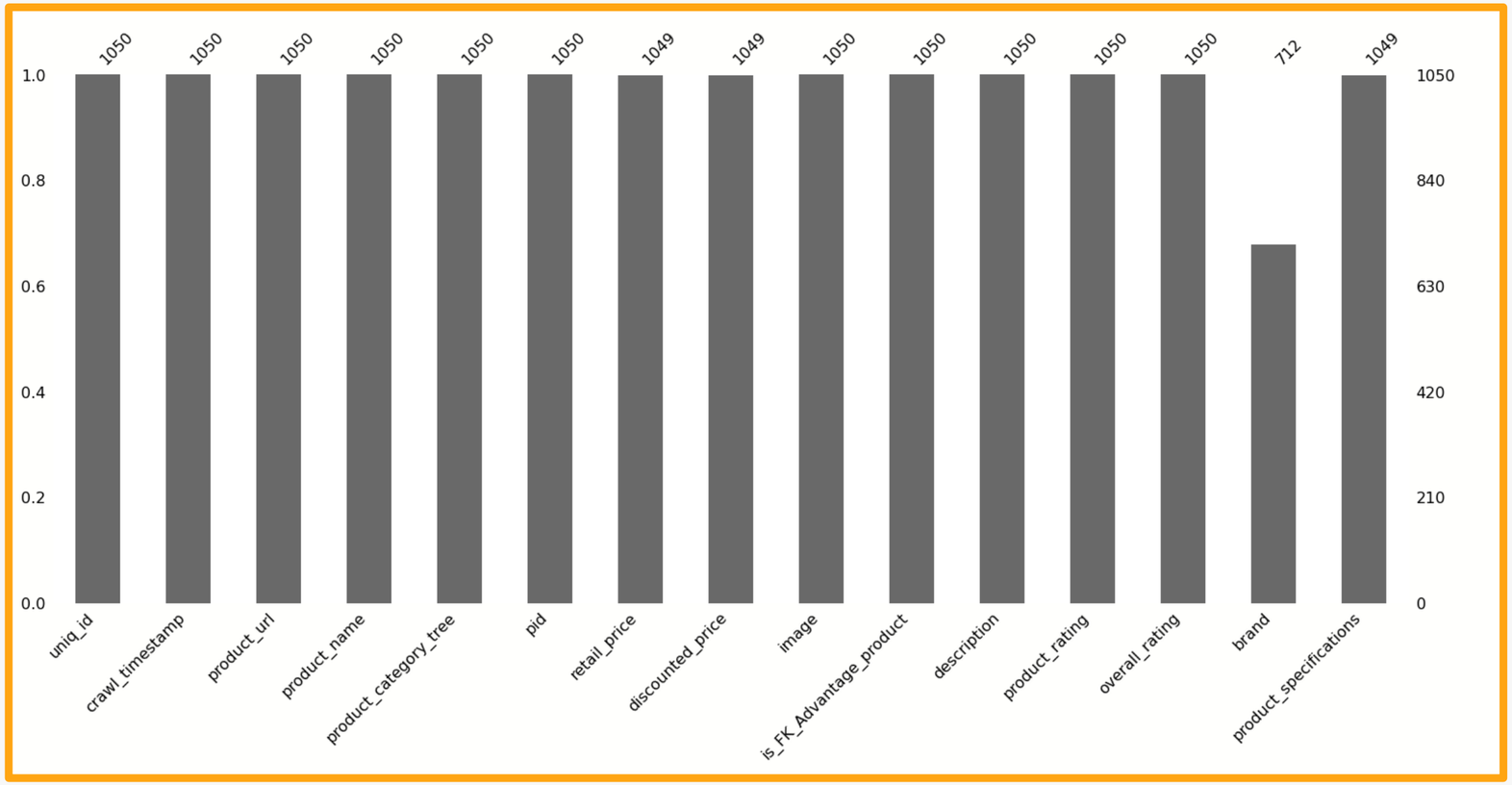
Préparation de la data

Suppression
Préparation du texte

Catégories


Baby Care
14.3%
Home Furnishing
14.3%
Watches
14.3%
Computers
14.3%
Home Decor & Festive Needs
14.3%
Beauty and Personal Care
14.3%
Kitchen & Dining
14.3%

Données manquantes



Bonne qualité

Description des produits

Tokenisation

- Chaque mot / punctuation est un token
- Pas d'espace

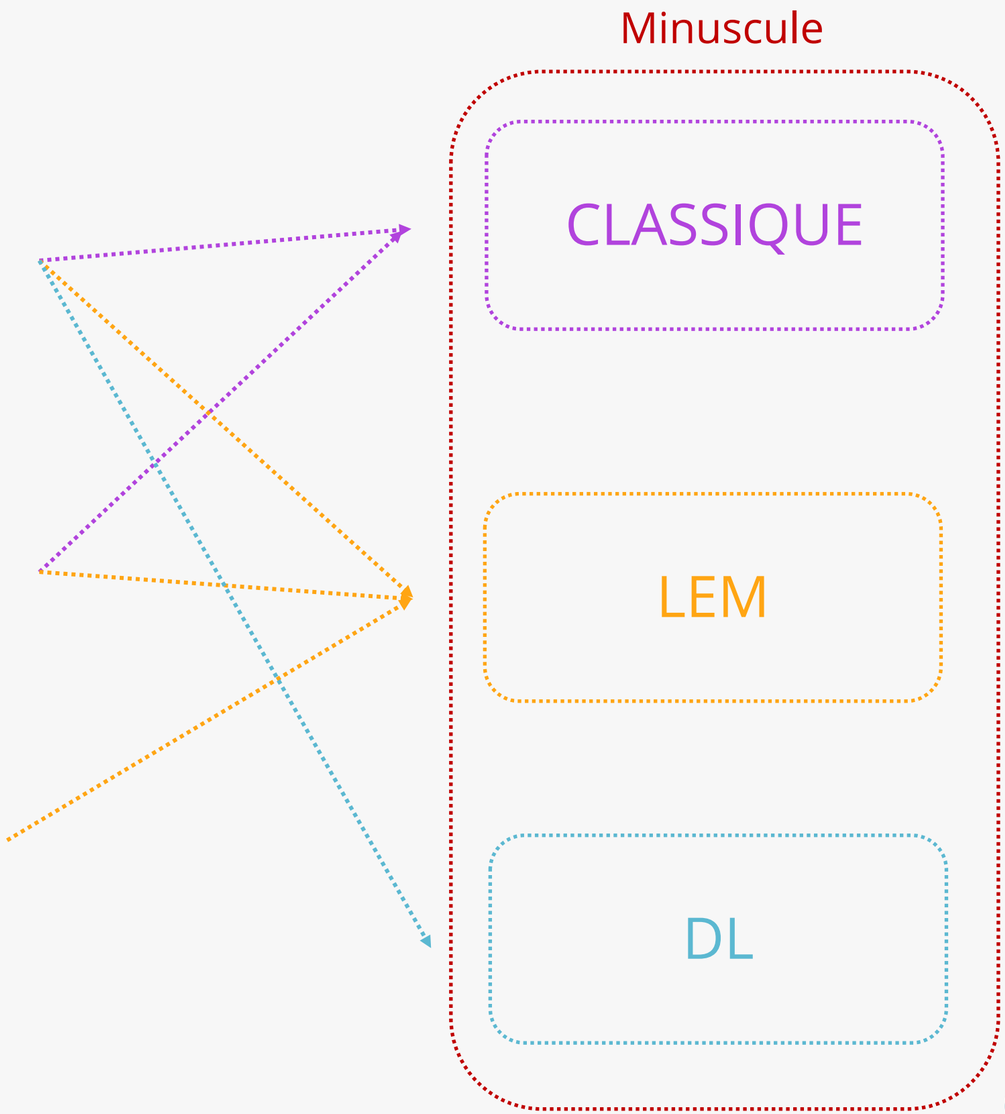
Stopwords

- Mots qui n'ont pas d'intérêt
- Déterminants, ponctuations, pronoms…

Lemmatizer

Forme canonique

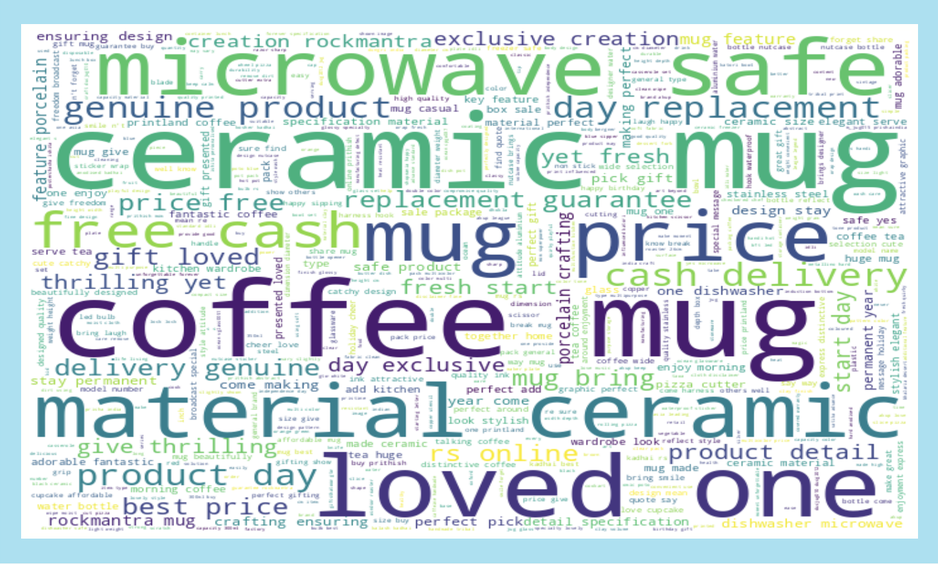
Supprime le genre / pluriel



Bag of Words

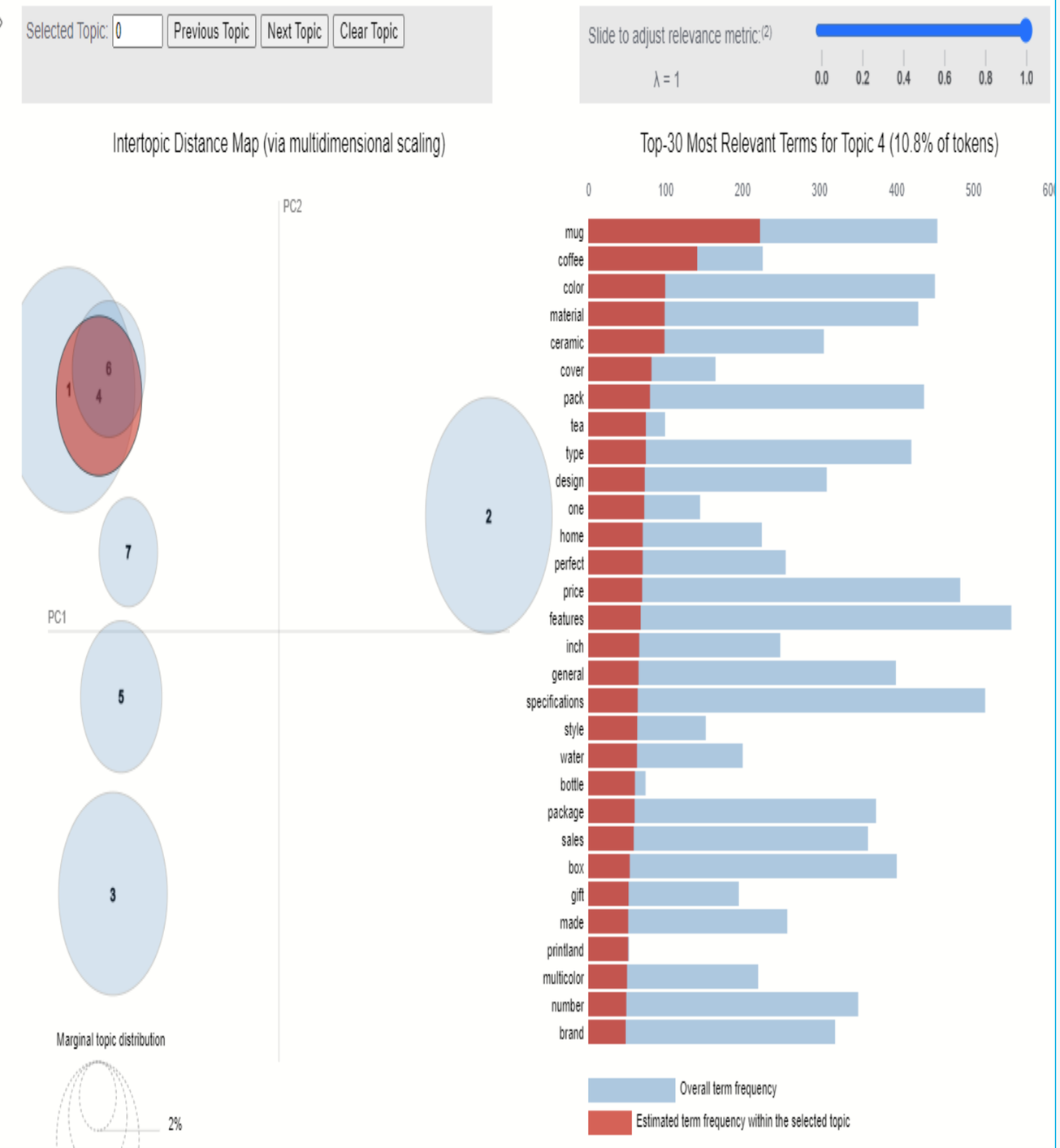
Global

Bags of Words par catégorie



Kitchen & Dining

Watches



LDA

- Déjà entrainé pour déceler des topics
- Créer des visualisations

Innovative & creative design

- Groupe 1/4/5 ont des points communs et seront peut-être plus difficiles à clusteriser.
- Groupe 2 unique

Fichier Data

Contenu

1 fichier CSV
Dossier image

Données

1050 produits
15 variables

Variables

Nom du produit
Date sur le marché
Prix de vente
Description
Image
Arbre de Catégorie

Catégorie

7 catégories
150 produits

Données manquantes

Dataset de qualité

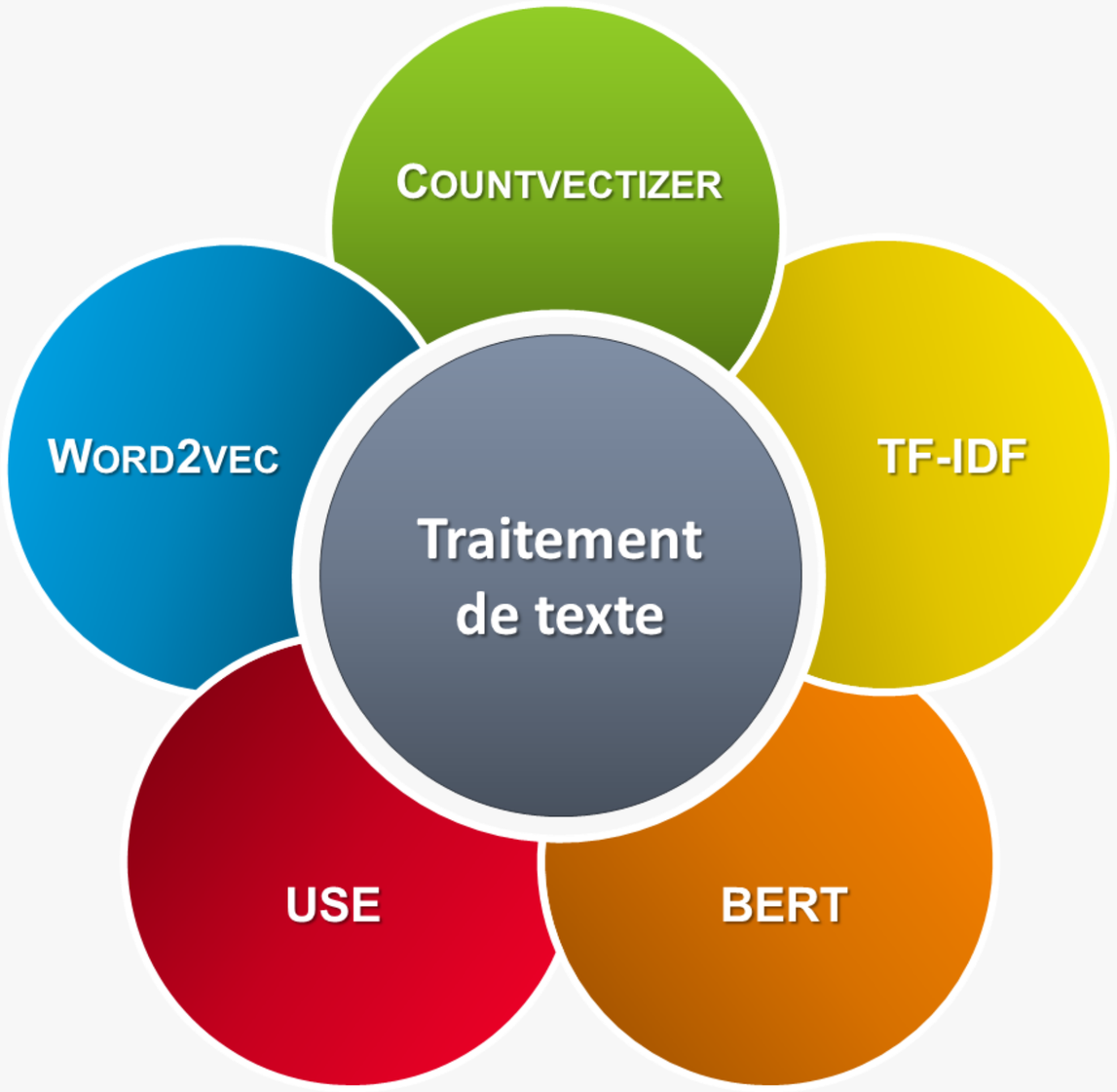
Préparation de la data

Suppression
Préparation du texte

2

Process

Procédés

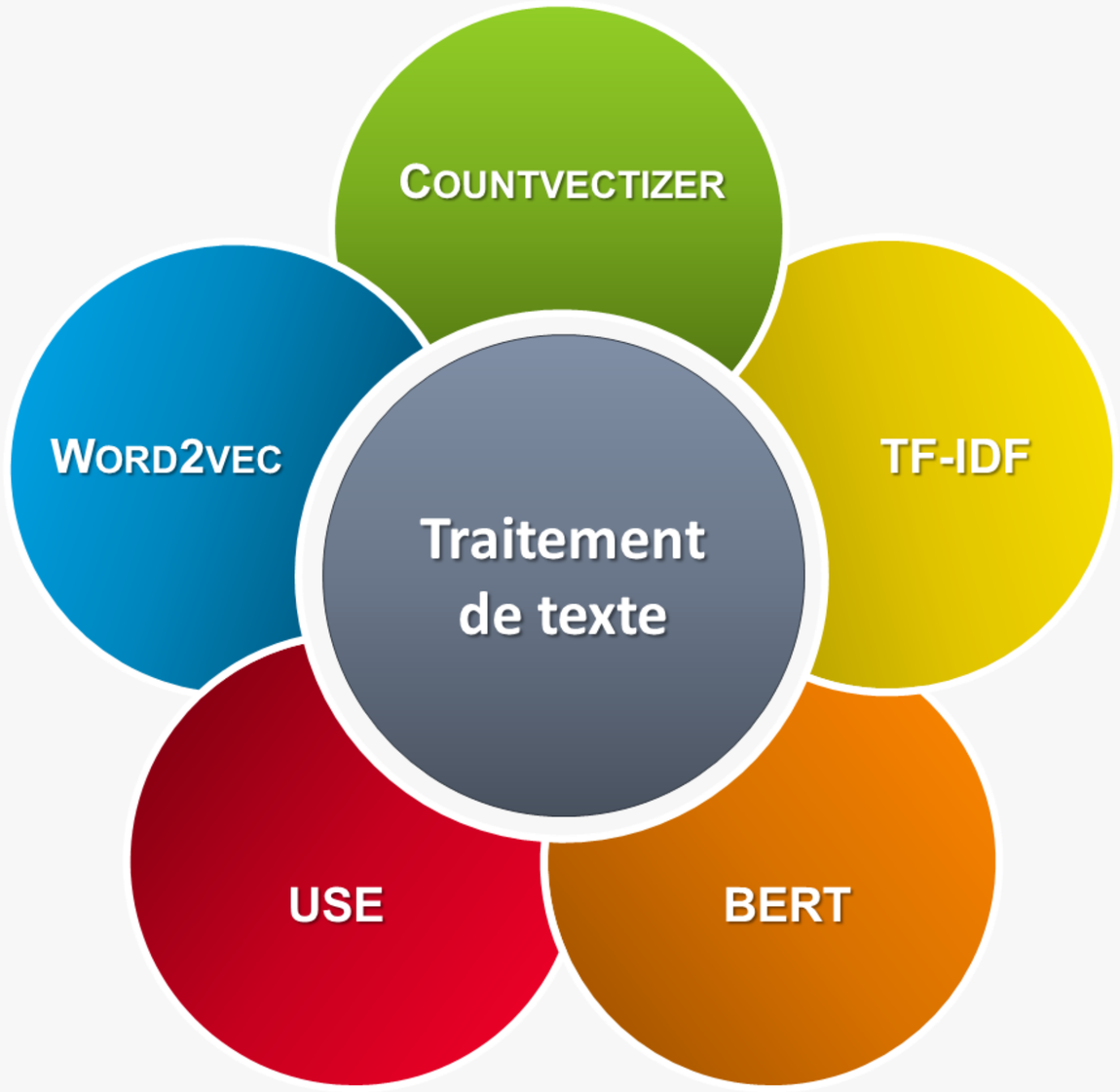


CountVectizer

Hello my name is james, this is my python notebook

	hello	is	james	my	name	notebook	python	this
0	1	2	1	2	1	1	1	1

Procédés



TF-IDF

TF = occurrences du mot / nombre de mots dans le document

IDF = nombre de documents / nombre de documents où apparait le mot

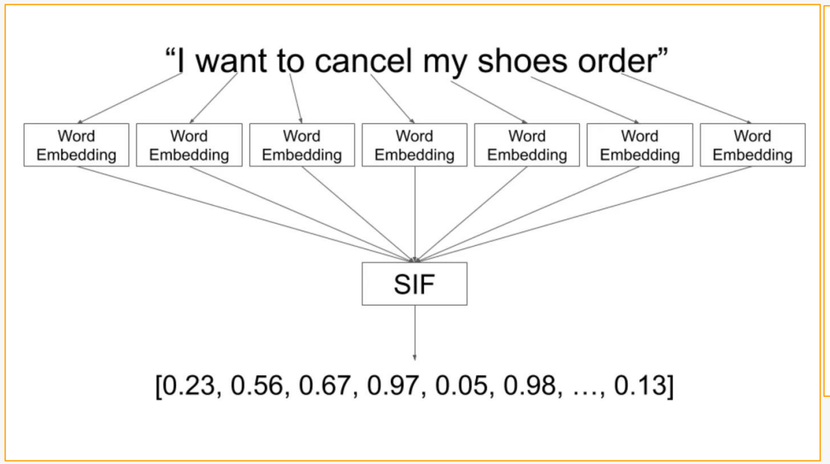
TF-IDF = TF * IDF

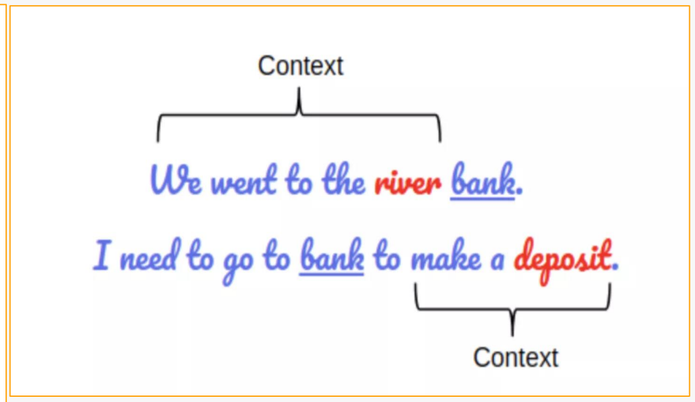
Procédés

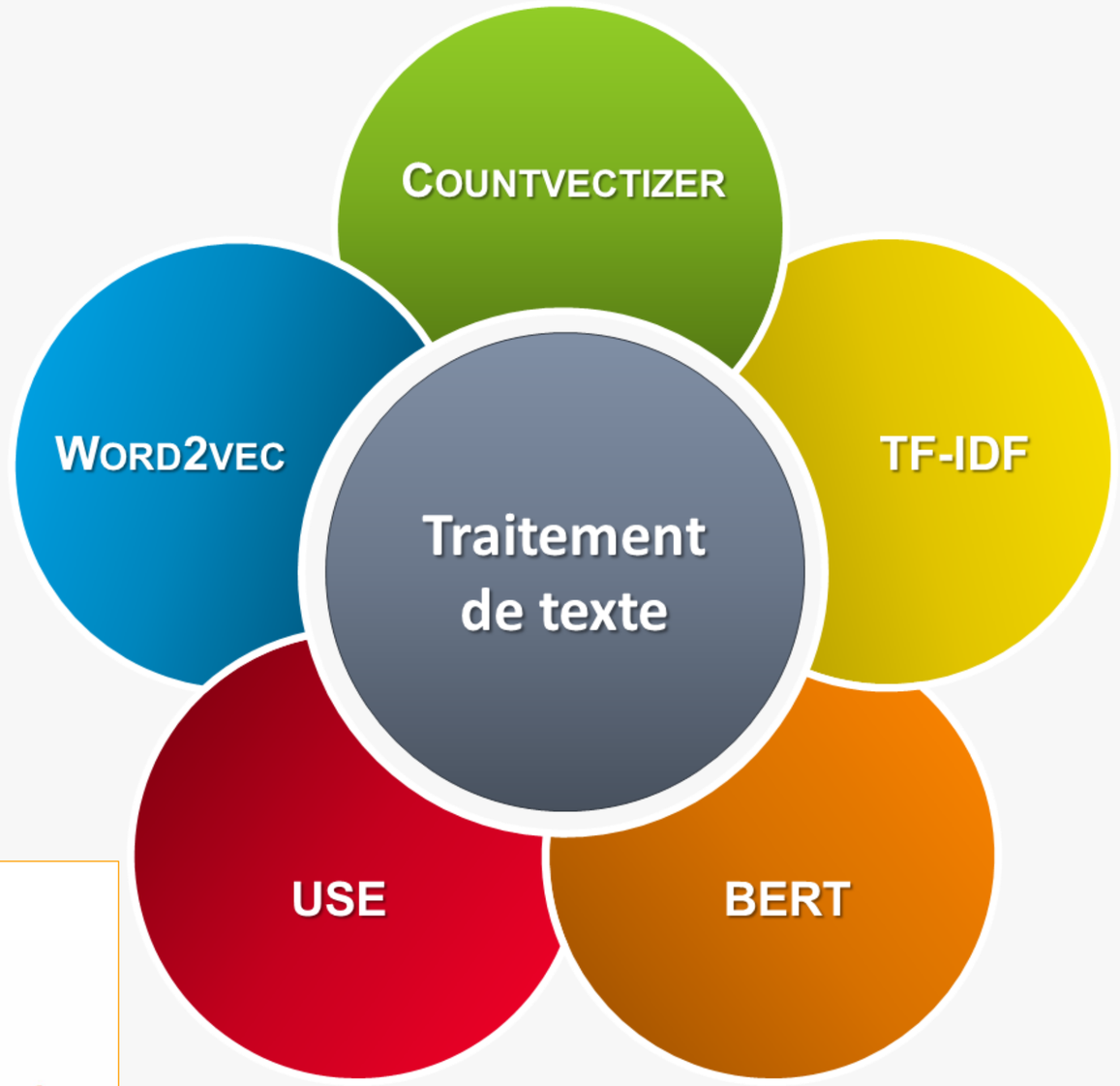
BERT

Transformation des phrases en vecteurs

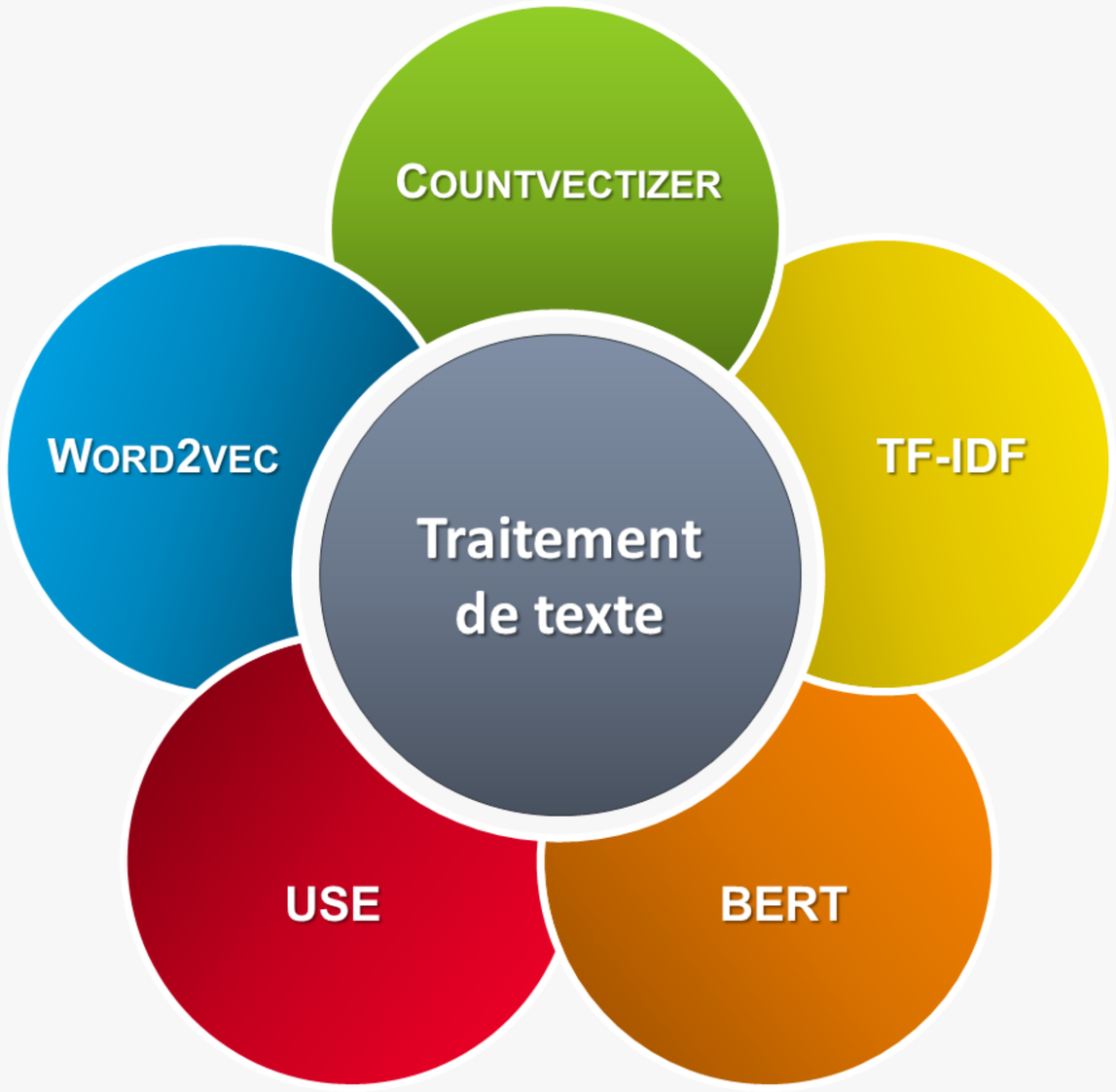
1050 vecteurs de taille 768







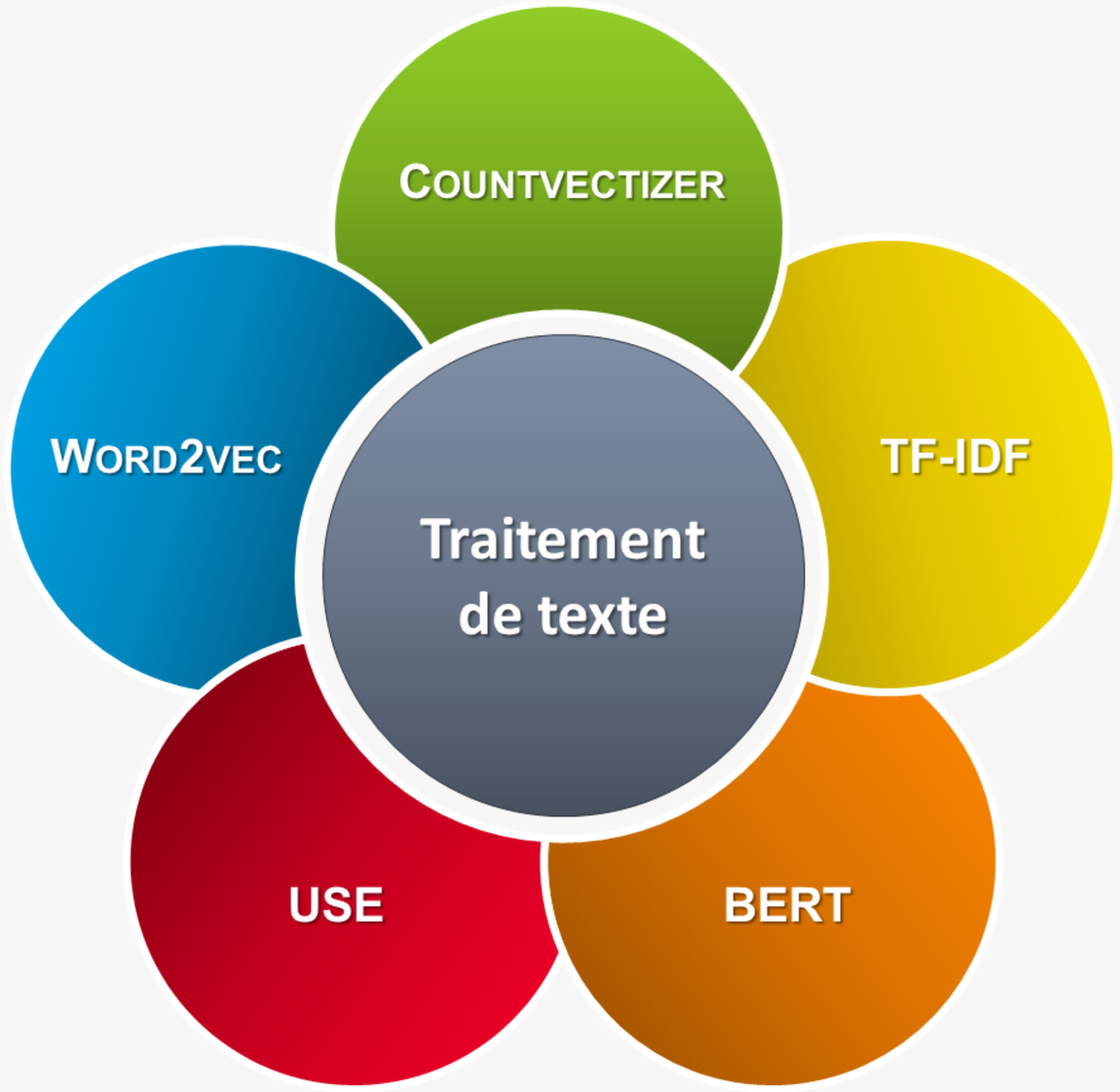
Procédés



USE

Transformation des phrases en vecteurs (512 dimensions)

Procédés



Word2Vec

4646 vecteurs de taille 300

Bank :

an organization where people and businesses can invest or borrow money, change it to foreign money, etc., or a building where these services are offered:

sloping raised land, especially along the sides of a river:

Process

Application des algorithmes de transformation

TSNE / Pas de reduction

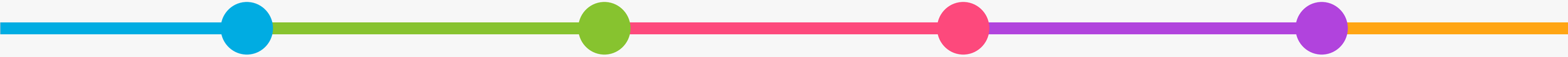
Etape 1

Kmeans : 7 Clusters

ARI

K-means

Etape 3



Etape 2

Stopwords / Liste perso

Liste perso : Flipkart, shipping ...

Apporte peu d'intérêt aux visuels

Etape 4

Visualisation

Scatterplot avec les categories réelles

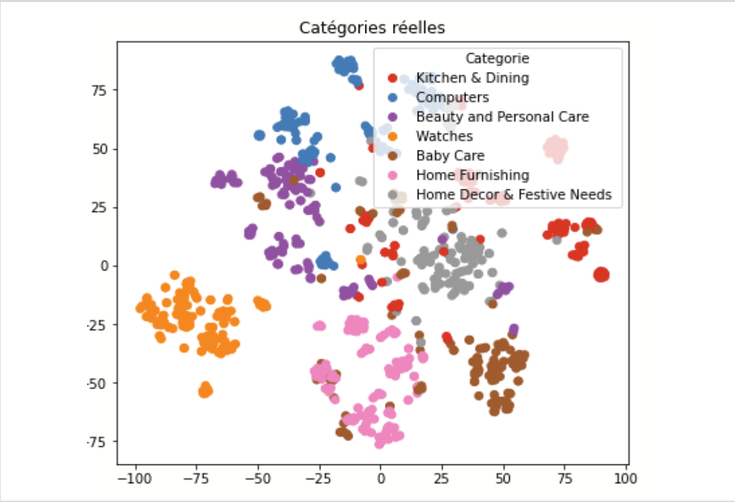
Scatterplot K-means

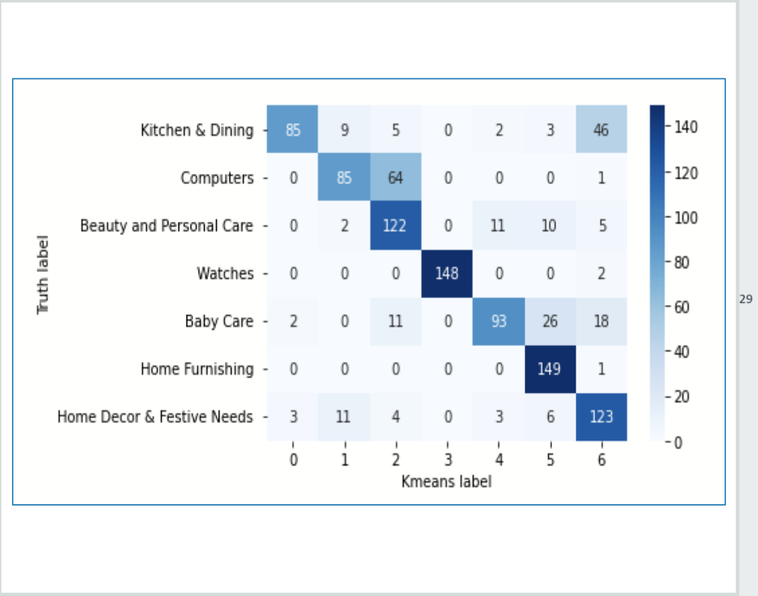
Matrice de confusion

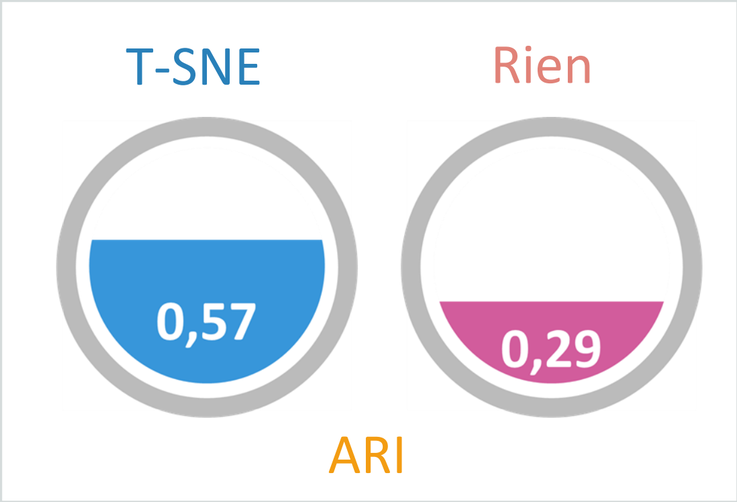
Camember

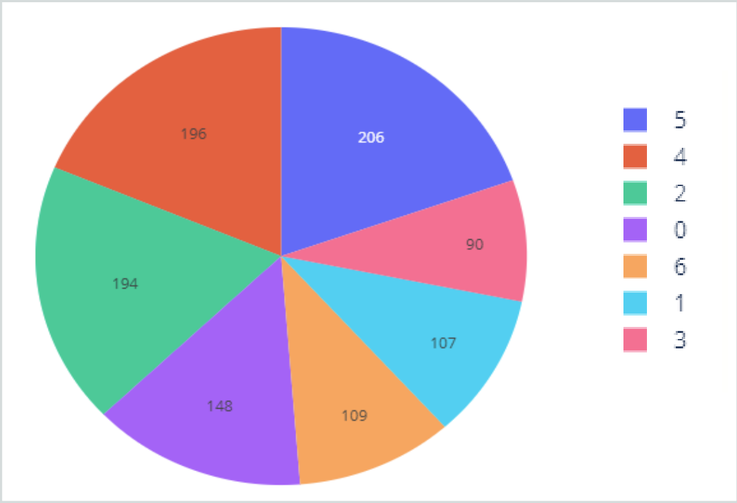
RÉSULTAT

TF-IDF









5 Points importants

Résultats du texte

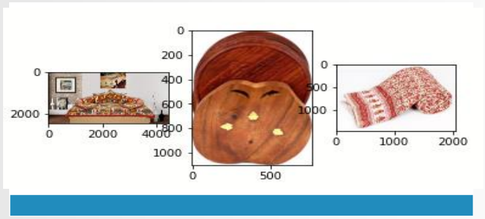
Nom	ARI
TF-IDF sans liste perso(TSNE)	0.5727
TF-IDF liste perso(TSNE)	0.5061
CountVectizer sans liste perso(TSNE)	0.4344
CV liste perso(TSNE)	0.3772
BERT sans liste perso(TSNE)	0.375
BERT liste perso(TSNE)	0.375
Word2Vec avec liste perso (TSNE)	0.36
USE sans liste perso (TSNE)	0.35
USE avec liste perso (TSNE)	0.35
USE avec liste perso	0.33
USE sans liste perso	0.33
TF-IDF sans liste perso	0.2943
Word2Vec avec liste perso (TSNE)	0.2803
BERT liste perso	0.2712
BERT sans liste perso	0.2712
TF-IDF liste perso	0.2622
Word2Vec avec liste perso	0.2077
Word2Vec avec liste perso	0.1995
CountVectizer sans liste perso	0.0555
CV liste perso	0.0549

1. T-SNE Efficace
2. Liste personnalisée contre-productive pour le clustering
3. Les méthodes de comptage > DeepLearning
4. Autre méthode effectuée : Nom + description
5. TF-IDF Meilleur

PRODUITS

IMAGES

Meubles



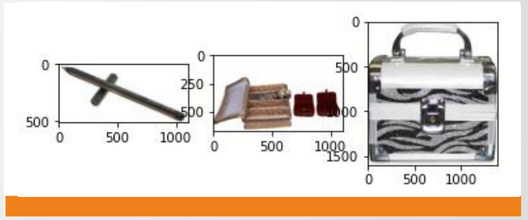
Soins pour bébé

Montres

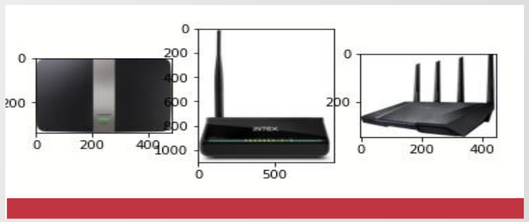
Décorations

Cuisine

Produits de beauté et soins personnels

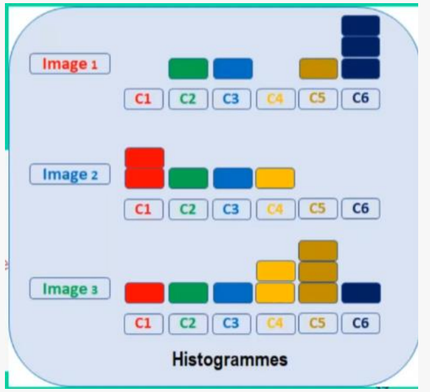


Ordinateurs



SIFT

Simple timeline



Etape 1

Descripteurs

Descripteurs = Décrit une image à partir d'un certain nombre de vecteurs

Etape 2

Préprocess

Passage au gris

Normaliser la luminosité

Améliorer le contraste

Etape 3

Traitement des images

Listes de descripteurs

517351 descripteurs de taille 128

Etape 4

Entrainement Kmeans

Choix des clusters :

Racine carré des descripteurs = 719

Label * 10 = 70

Etape 5

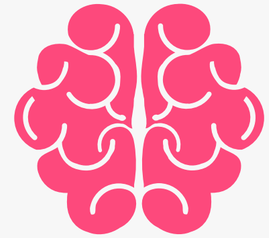
Feature img

Création des histogrammes

Test sans reduction de dimension / PCA (99% variance : 498 dim) / T-SNE

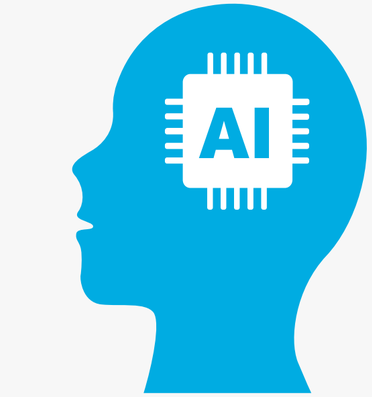
ARI
0,07

Transfert Learning



1. **Pré-entrainés**
   Reconnaissance d'image
2. **En libre téléchargement**
3. **Transfert**
4. **Couches**
   Flexible
5. **Adapté aux nouveaux problèmes**
6. **Modèles choisis**
   VGG16 / VGG19 / RESNET / InceptionV3

Test des modèles



Redimensionné : Image, ligne, colonne, channel

VGG16

77% : compas magnétique

13% : montre

VGG19

64% : montre

19% : montre digitale

RESNET

100% : Chien

InceptionV3

99% : chaussure

Extraction de features

Couches convolutionnelles

Proches des input

Reconnaissent des caractéristiques “faibles”

Convolutionnelles

Milieu

Couches du milieu

Caractéristiques abstraites plus complexes

Classification

Couches proches du résultat

Interprète les résultats

Process

Couches du modèle

2 couches avant la fin

Redimension

Format précis en 4 dimensions :

Echantillon – Lignes – Colonne - Channel

Features

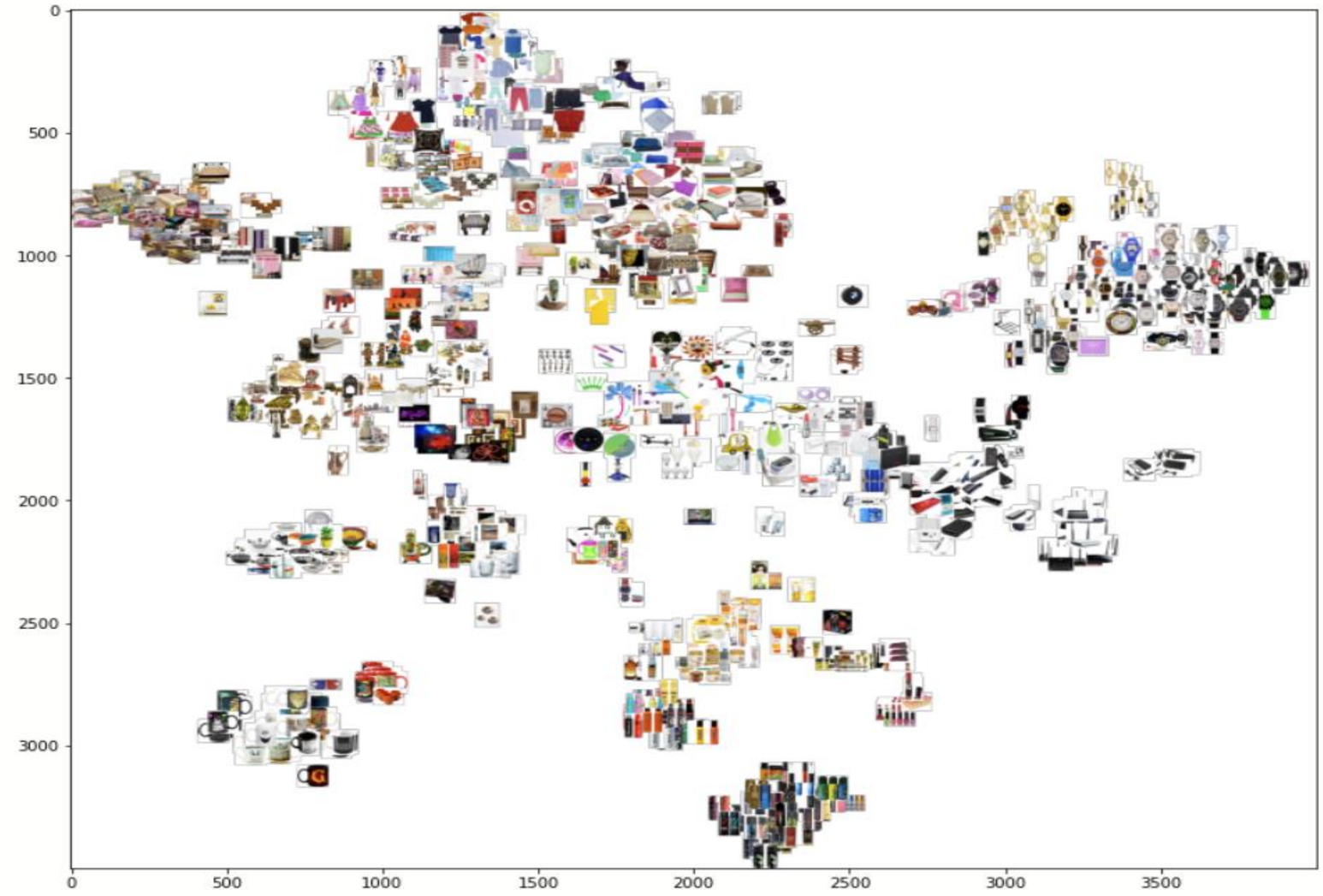
Dimensions :

VGG : (1050,1,4096) | RESNET/Inception : (1050,1,2048)

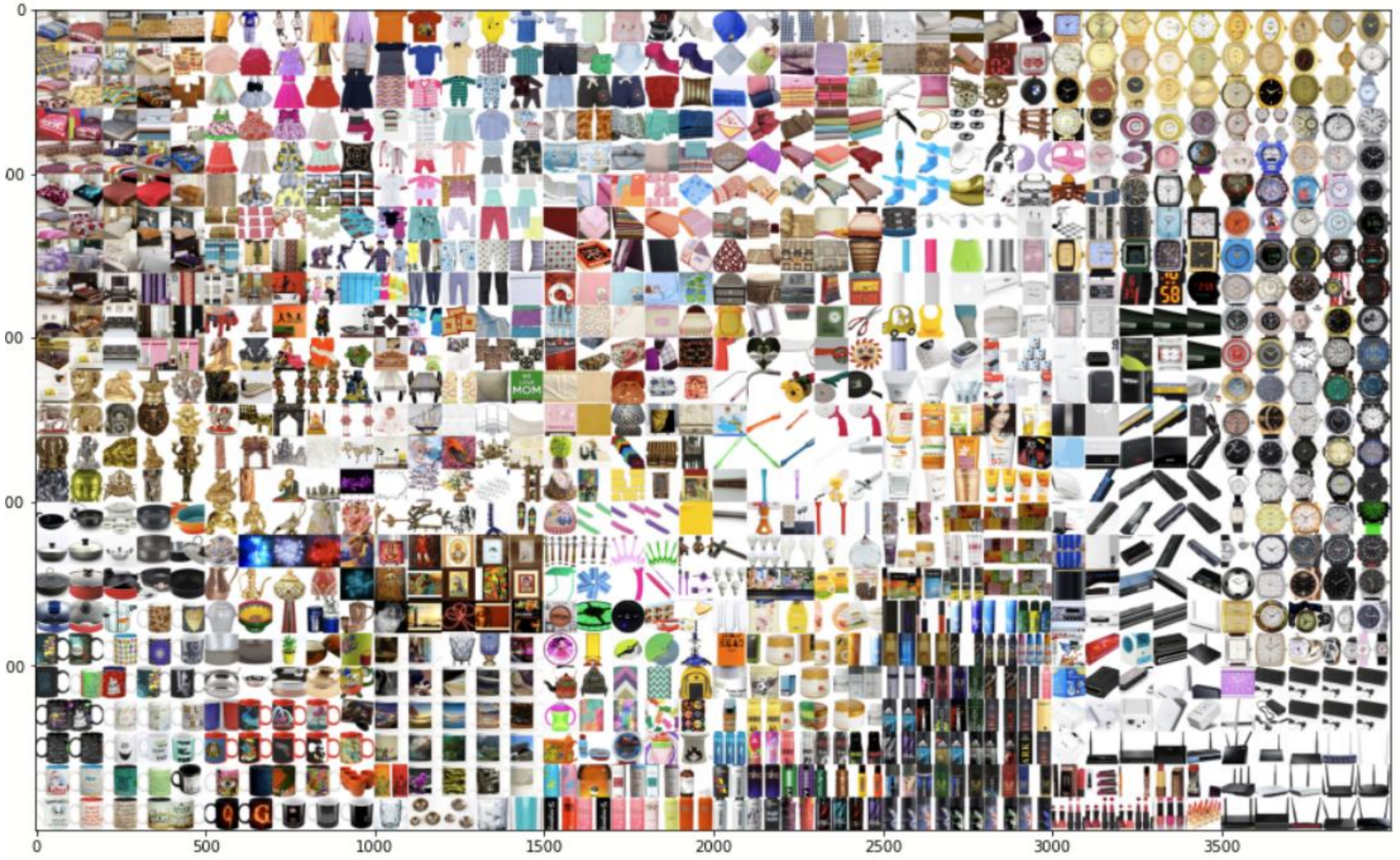
Différents modèles

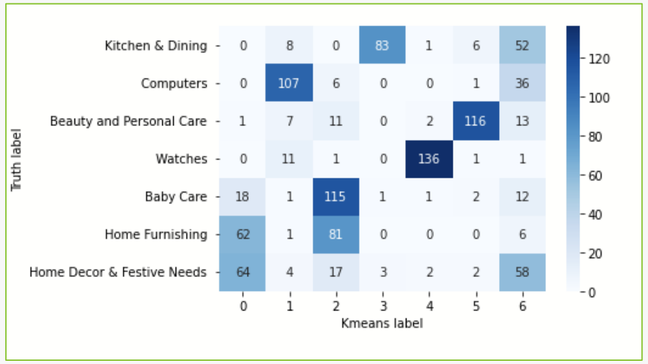
VGG16	VGG19	RESNET50	InceptionV3
1050,4096	1050,4096	1050,2048	1050, 2048
Dimensions PCA : 803	Dimensions PCA : 796	Dimensions PCA : 44	Dimensions PCA : 248
Rien : 0,466	Rien : 0,4759	Inférieur à 0,07	Inférieur à 0,06
ACP/TSNE : 0,4597	ACP/TSNE : 0,4464		
ACP : 0,4639	ACP : 0,4753		

Catégories


0
500
1000
1500
2000
2500
3000
0
500
1000
1500
2000
2500
3000
3500

Catégories


0
00
00
00
00
0
500
1000
1500
2000
2500
3000
3500


Kitchen & Dining 0 8 0 83 1 6 52
Computers 0 107 6 0 0 1 36
Beauty and Personal Care 1 7 11 0 2 116 13
Watches 0 11 1 0 136 1 1
Baby Care 18 1 115 1 1 2 12
Home Furnishing 62 1 81 0 0 0 6
Home Decor & Festive Needs 64 4 17 3 2 2 58
Truth label
0 1 2 3 4 5 6
Kmeans label
120
100
80
60
40
20
0

3

Faisabilité / Recommandations

Meilleurs algorithmes

Texte

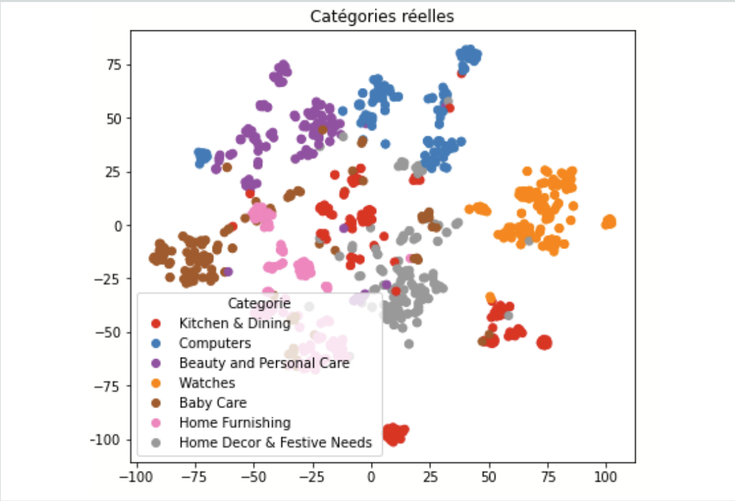
Nom	ARI
TF-IDF sans liste perso(TSNE)	0.5727
TF-IDF liste perso(TSNE)	0.5061
CountVectizer sans liste perso(TSNE)	0.4344
CV liste perso(TSNE)	0.3772
BERT sans liste perso(TSNE)	0.375
BERT liste perso(TSNE)	0.375
Word2Vec avec liste perso (TSNE)	0.36
USE sans liste perso (TSNE)	0.35
USE avec liste perso (TSNE)	0.35
USE avec liste perso	0.33
USE sans liste perso	0.33
TF-IDF sans liste perso	0.2943
Word2Vec avec liste perso (TSNE)	0.2803
BERT liste perso	0.2712
BERT sans liste perso	0.2712
TF-IDF liste perso	0.2622
Word2Vec avec liste perso	0.2077
Word2Vec avec liste perso	0.1995
CountVectizer sans liste perso	0.0555
CV liste perso	0.0549

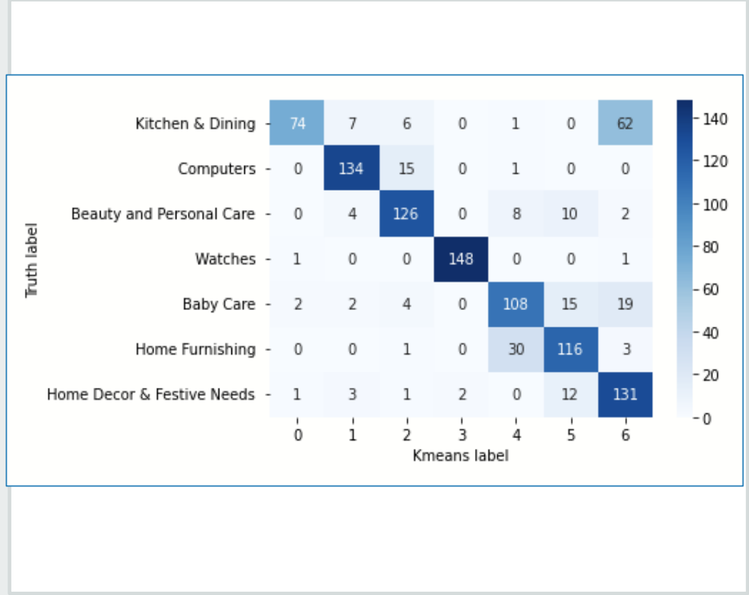
Image

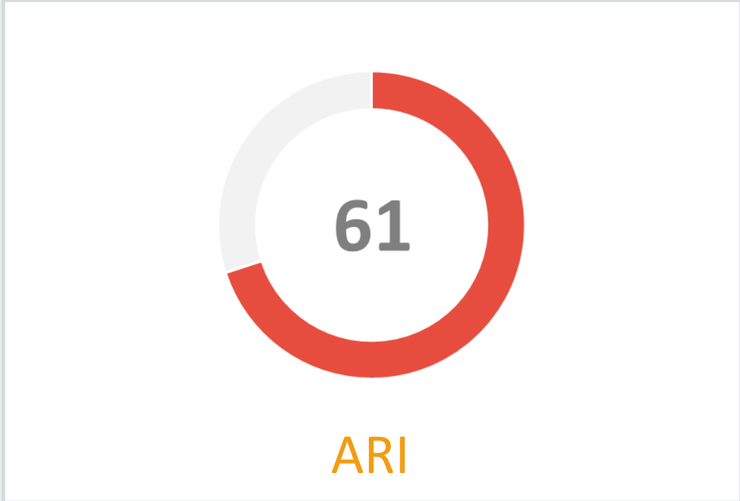
Nom	ARI
VGG19	0.4759
VGG19 (ACP)	0.4753
VGG16	0.466
VGG16 (ACP)	0.4639
VGG16 (TSNE)	0.4608
VGG16 (ACP/TSNE)	0.4597
VGG19 (ACP/TSNE)	0.4464
VGG19 (TSNE)	0.4345
SIFT max clusters (TSNE)	0.0713
SIFT max clusters	0.0663
Inception (TSNE)	0.0632
Inception(ACP/TSNE)	0.0531
SIFT clusters restreints	0.0525
SIFT clusters restreints (TSNE)	0.0518
Inception	0.0405
Inception(ACP)	0.0405
RESNET(TSNE)	0.0366
RESNET (APC/TSNE)	0.0343
RESNET	0.0169
RESNET (ACP)	0.0168

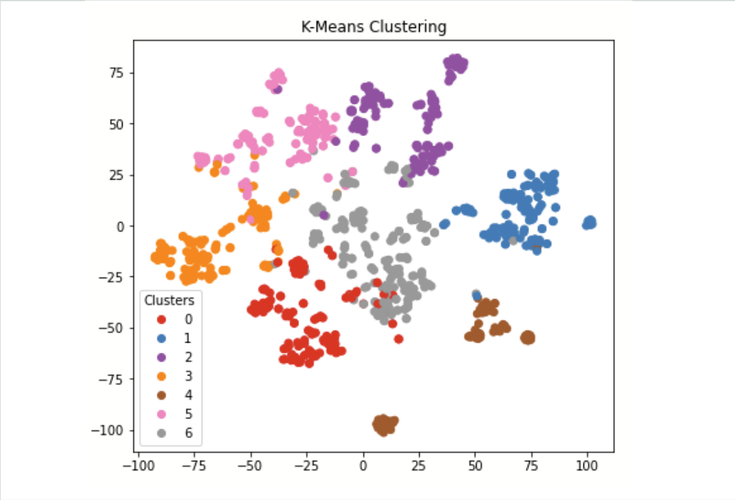
RÉSULTAT

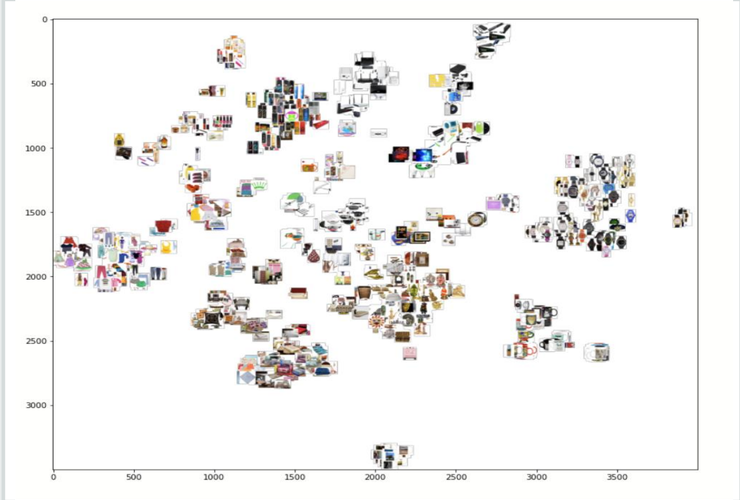
Clustering final







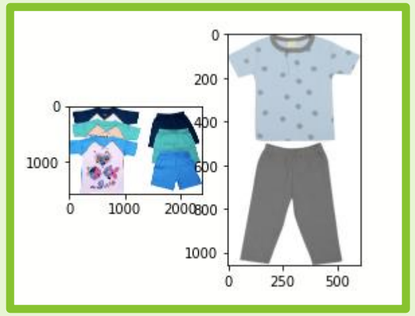




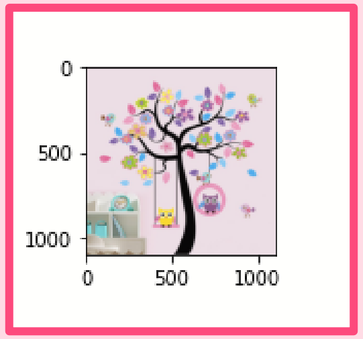
Comparaison

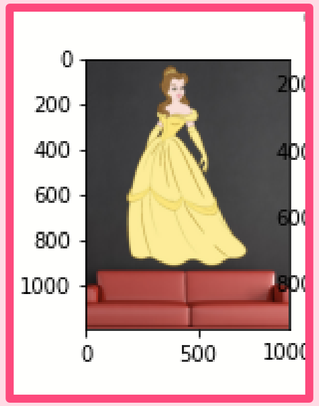
Catégorie BabyCare

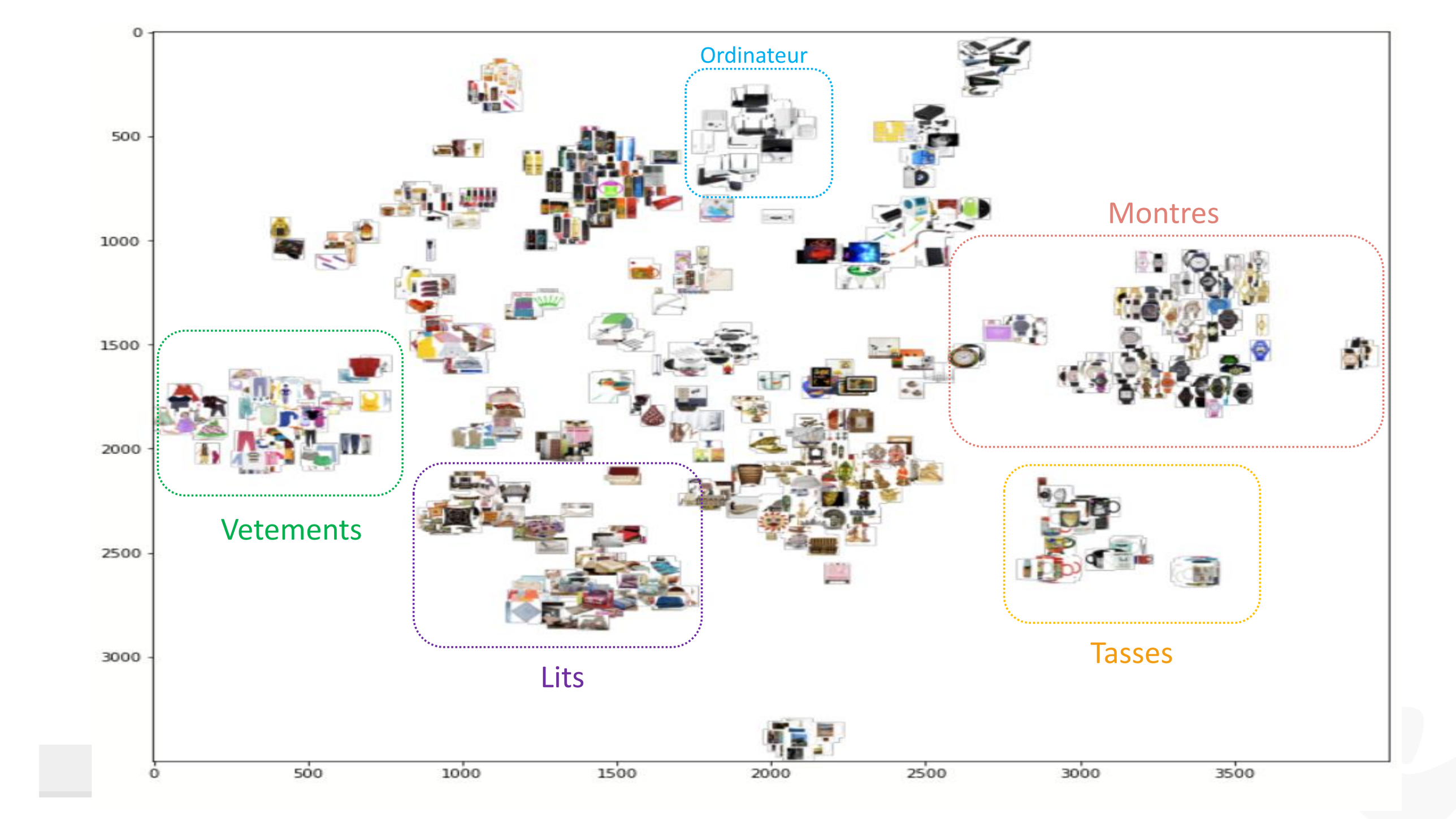
Bons clusters

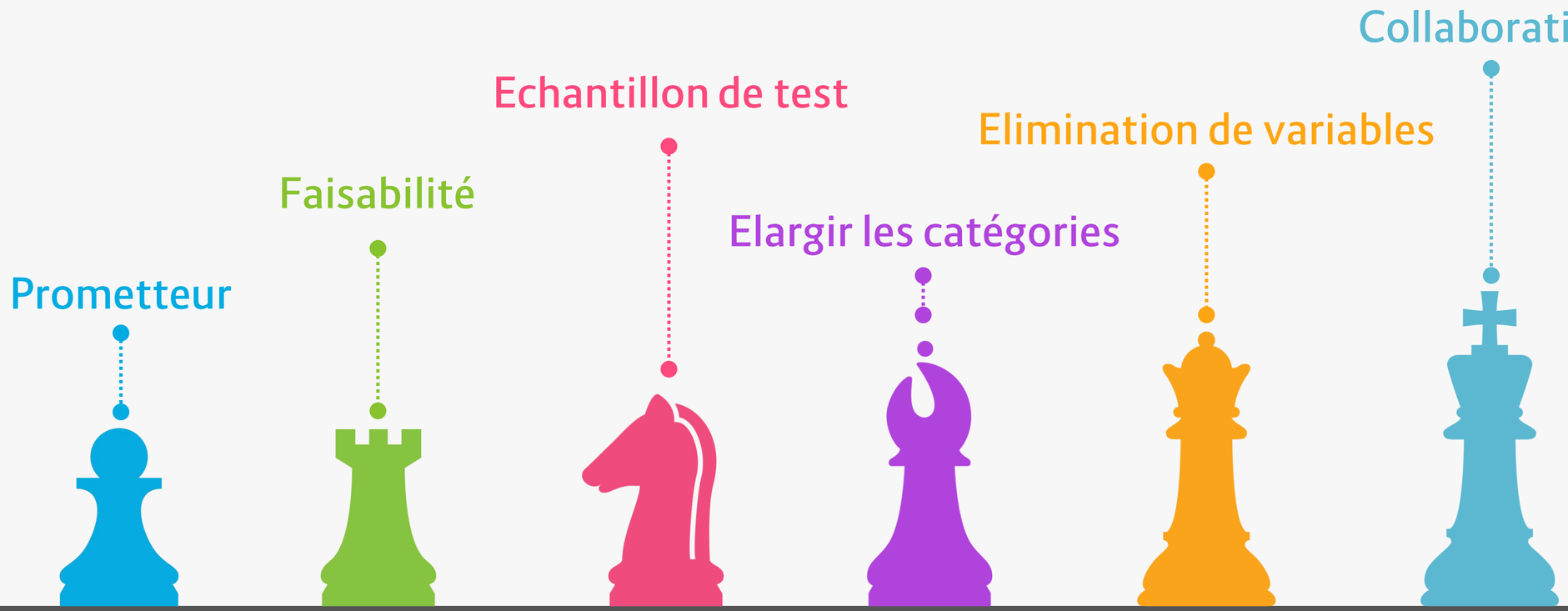


Mauvais clusters






Ordinateur
Montres
Vetements
Lits
Tasses
0
500
1000
1500
2000
2500
3000
3500



Strategies pour le **Success**