# Classifiez automatiquement des biens de consommation

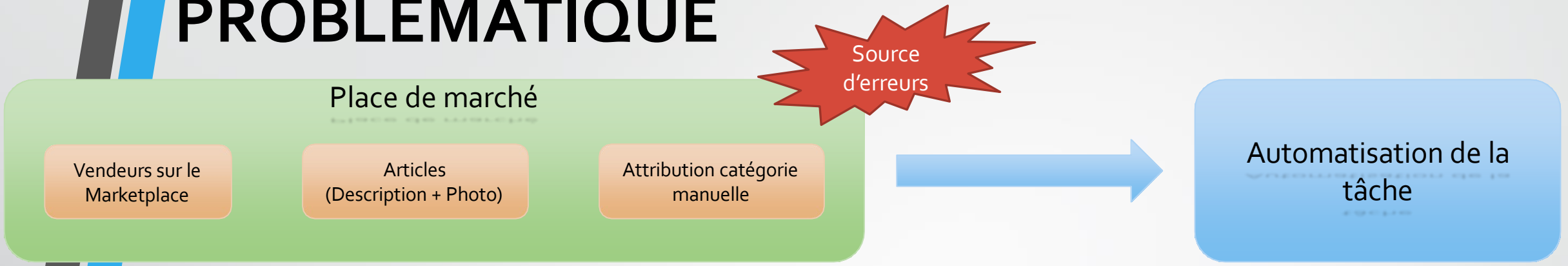

*Création le 01/10/2023*

*DAVID*

# SOMMAIRE

# I - PRESENTATION

# PROBLEMATIQUE




- **Mission :**
  - Réaliser une première **étude de faisabilité** d'un moteur de classification en se basant sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article.

- **Suggestion :**
  - Utilisation d'une API pour enrichir nos données, « Place de marché » n'ayant encore pas beaucoup d'articles en vente.

    https://rapidapi.com/edamam/api/edamam-food-and-grocery-database
- **Données :**
  - https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientist/Projet+-+Textimage+DAS+V2/Dataset+projet+pre%CC%81traitement+textes+images.zip

# II – PRESENTATION DES DONNEES

# DECOUVERTE DES DONNEES

**DATASET**

Uniq_id
Crawl_timestamp
Product_url
Product_name
Product_category_tree
Pid
Retail_price
Discounted_price
Image
Is_FK_Advantage_product
Description
Product_rating
Overall_rating
Brand
Product_specifications

1050 lignes
15 colonnes

**Utilisation de 3 features**

Exemple du premier article de notre jeu de données

Arborescence :

```
'["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]'
```


```
'Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.T
his curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give you
r home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright
sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light i
n your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyes
ter Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester'
```

# III – DONNÉES TEXTUELLES

# ANALYSE DES CATEGORIES

**La feature « product_category_tree » représente l'arborescence complète d'un article.**

Reprenons l'exemple de notre premier article dans le jeu de données

```
'["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]'
```

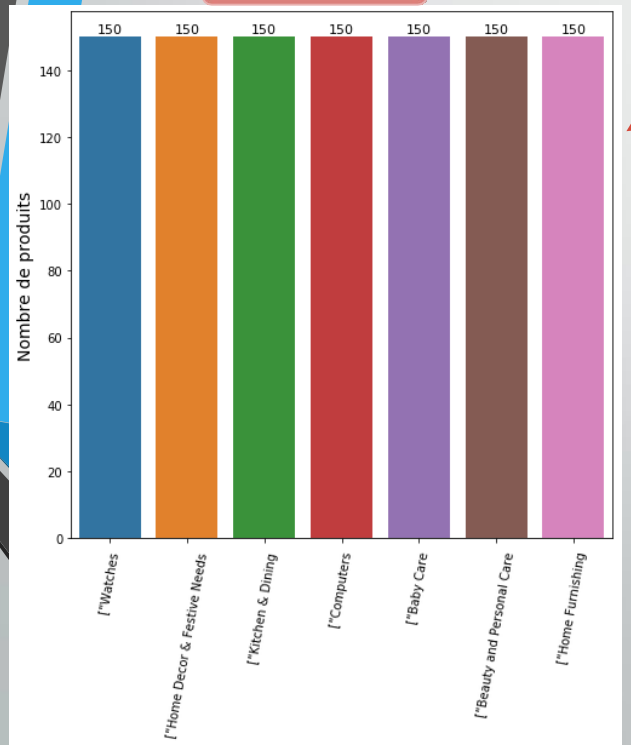
On extrait les différents nœuds de l'arborescence

```
['["Home Furnishing ', ' Curtains & Accessories', ' Curtains ', ' Elegance Polyester Multicolor Abstract Eyelet Do..."]']
```

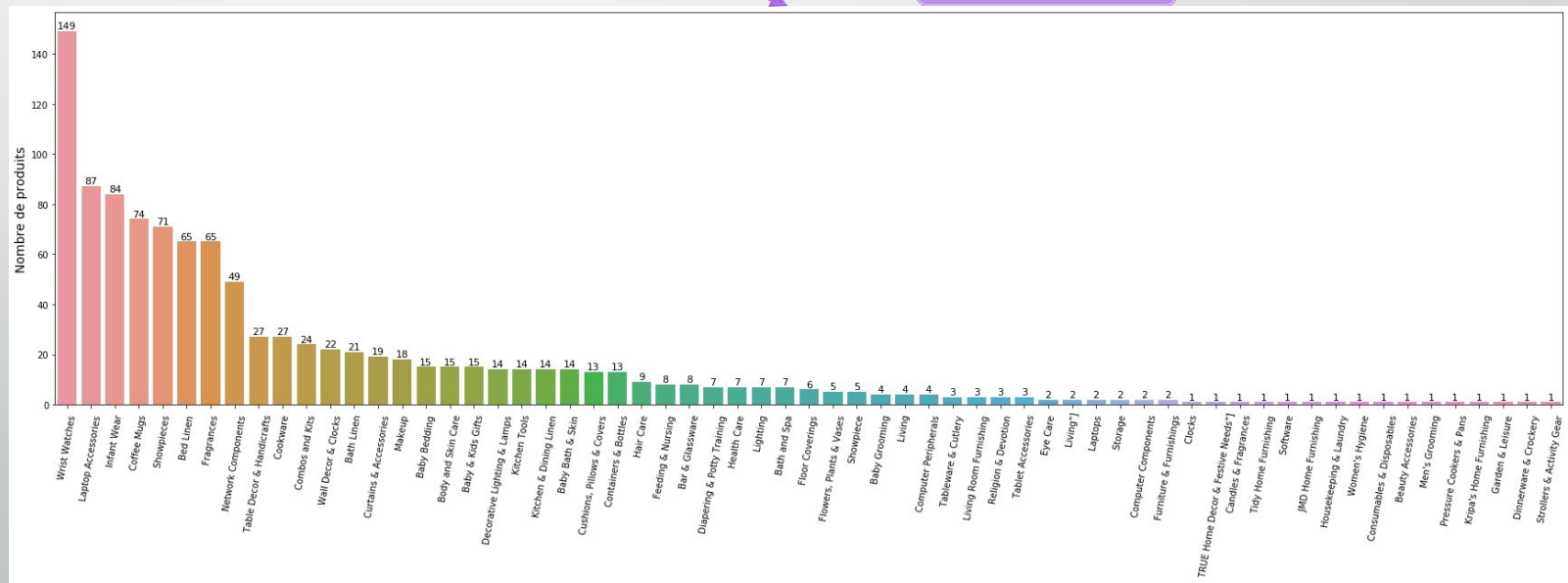
**Sous_cat_1**

**Sous_cat_2**

**7 Catégories**




**63 Catégories**



# Nettoyage données

## prétraitement du texte

```
======== PRE TRAITEMENT ========
Buy Epresent Mfan 1 Fan USB USB Fan for Rs.219 online. Epresent Mfan 1 Fan USB USB Fan at best prices with FREE shipping & cash
on delivery. Only Genuine Products. 30 Day Replacement Guarantee.

======== LOWERCASE ========
buy epresent mfan 1 fan usb usb fan for rs.219 online. epresent mfan 1 fan usb usb fan at best prices with free shipping & cash
on delivery. only genuine products. 30 day replacement guarantee.

======== TOKENIZER ========
['buy', 'epresent', 'mfan', '1', 'fan', 'usb', 'usb', 'fan', 'for', 'rs.219', 'online', '.', 'epresent', 'mfan', '1', 'fan', 'u
sb', 'usb', 'fan', 'at', 'best', 'prices', 'with', 'free', 'shipping', '&', 'cash', 'on', 'delivery', '.', 'only', 'genuine',
'products', '.', '30', 'day', 'replacement', 'guarantee', '.']

======== STOPWORDS ========
['buy', 'epresent', 'mfan', '1', 'fan', 'usb', 'usb', 'fan', 'rs.219', 'online', 'epresent', 'mfan', '1', 'fan', 'usb', 'usb',
'fan', 'prices', 'free', 'shipping', 'cash', 'delivery', 'genuine', 'products', '30', 'day', 'replacement', 'guarantee']

======== LEMMATISATION ========
['buy', 'epresent', 'mfan', '1', 'fan', 'usb', 'usb', 'fan', 'rs.219', 'online', 'epresent', 'mfan', '1', 'fan', 'usb', 'usb',
'fan', 'price', 'free', 'shipping', 'cash', 'delivery', 'genuine', 'product', '30', 'day', 'replacement', 'guarantee']
```

## Tf Idf

| | mot | tfidf |
|---|---|---|
| 9 | fan | 0.637175 |
| 23 | usb | 0.454869 |
| 8 | epresent | 0.388319 |
| 14 | mfan | 0.388319 |
| 0 | 219 | 0.194159 |
| 18 | prices | 0.071857 |
| 3 | best | 0.071346 |
| 16 | online | 0.054523 |
| 24 | with | 0.053976 |
| 2 | at | 0.050738 |

TF-IDF (Term Frequency-Inverse Document Frequency) : Le TF-IDF combine les deux mesures, TF et IDF, pour évaluer l'importance d'un terme dans un document par rapport à l'ensemble de la collection. Il est calculé en multipliant le TF par l'IDF. Ainsi, un terme aura un score TF-IDF élevé dans un document s'il est fréquent dans ce document tout en étant rare dans la collection, ce qui indique son importance spécifique à ce document.

1. Lemmatisation : La lemmatisation est le processus de réduction des mots à leur forme de base, appelée "lemme". Par exemple, pour le mot "manger", le lemme est "mange". Cela permet de regrouper différentes formes d'un mot pour analyser ou traiter le texte de manière plus cohérente. La lemmatisation prend en compte la grammaire et la signification des mots. Elle peut également transformer les verbes conjugués en leur forme à l'infinitif.

2. Tokenisation : La tokenisation est le processus de division d'un texte en "tokens", qui sont généralement des mots ou des morceaux de mots. Un "token" est l'unité de base du texte après sa division. Par exemple, la phrase "Le chat dort." serait divisée en trois tokens : "Le", "chat", et "dort". La tokenisation est une étape essentielle pour analyser le texte, car elle permet de déterminer les unités de base pour l'analyse et le traitement ultérieurs.

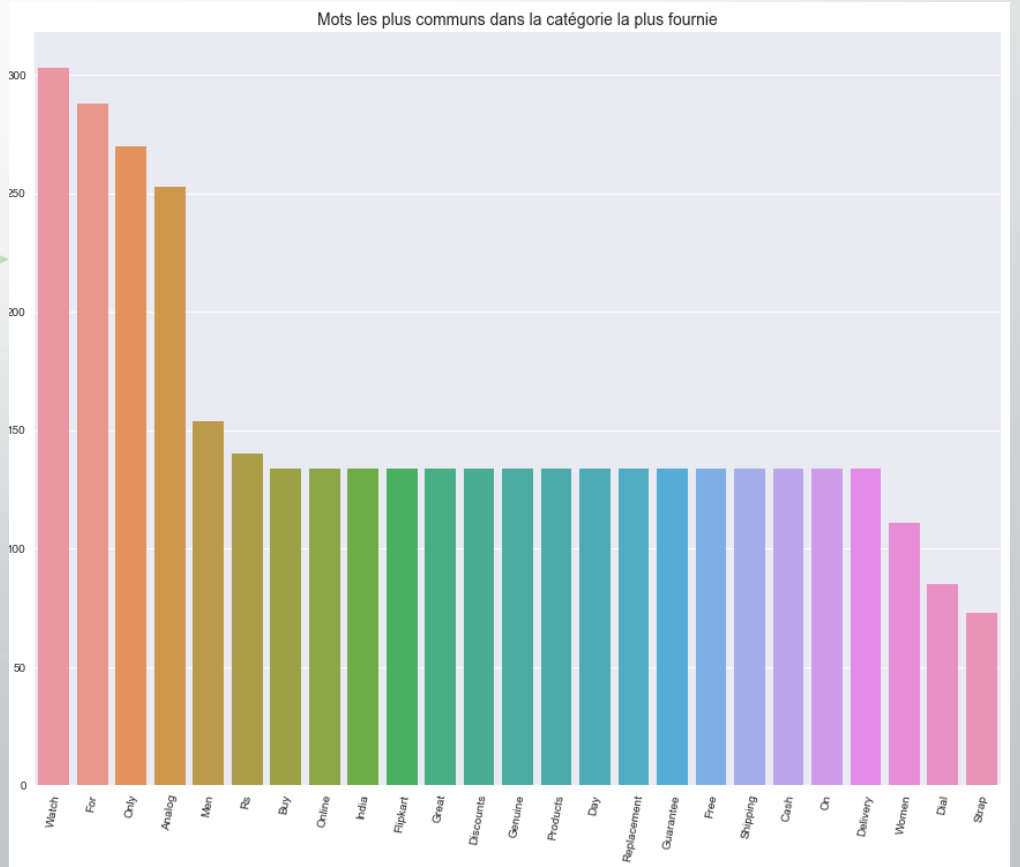
3. Suppression des "Stop Words" : Les "stop words" (mots vides) sont des mots courants qui n'apportent généralement pas beaucoup de sens au texte. Ils incluent des mots comme "le", "la", "de", "et", "un", etc. La suppression des "stop words" consiste à filtrer ces mots du texte pour réduire le bruit et améliorer la précision de l'analyse textuelle. Ces mots sont souvent omis car ils sont très fréquents et n'apportent pas d'informations distinctives.

# ANALYSE DU CORPUS

Le corpus représente l'ensemble des descriptions de notre jeu de données.

- **Analyse des mots les plus fréquents dans le corpus :**
  - Les mots les plus fréquents dans notre corpus sont des mots « généraux » qui ne décrivent pas les produits.
  - C'est normal, car ces mots doivent apparaître dans toutes les descriptions.


- **Analyse des mots les plus fréquents de la catégorie « Wrist Watches » : (montre-bracelet) (sous_cat_2) :**
  - On retrouve des mots spécifiques à la catégorie comme « watch » ou « analog ».
  - On retrouve aussi des mots qui ont peu d'importance comme « only », ou « buy ».




- Afin de bien préparer nos données pour la classification, nous allons essayer plusieurs opérations afin de choisir celle qui s'avère la plus performante pour notre corpus.

- Le critère de performance choisi est l'accuracy d'une régression logistique mise en place pour prédire la catégorie d'un article et optimisée par un RandomizeSearchCV.

# OPERATIONS TEXTUELLES

| Opération | Analyse performance<br>*Utilisation de la régression logistique pour quantifier les performances de l'opération sur le corpus. Optimisation du modèle par un RandomizeSearchCV* | Nombre de mots extraits | Commentaire |
|---|---|---|---|
| **Bag of words**<br>*Extraction simple des mots de chaque description* | Accuracy = 79% | 3777 | |
| **Bag of words amélioré**<br>*Suppression des mots qui apparaissent dans au moins 3 documents* | Accuracy = 79% | 1029 | Pas d'amélioration des performances, mais nous avons éliminé des caractéristiques inutiles, ce qui peut rendre le modèle plus interprétable. |
| **Stopwords**<br>*Suppression des mots les plus courants* | Accuracy = 79% | 939 | Pas d'amélioration des performances. Cela est dû à la taille réduite de notre jeu de données. |
| **Tf-Idf**<br>*Application d'une tf-Idf en tenant compte des 3 premières opérations* | Accuracy = 87% | 939 | Nous pouvons aussi inspecter quels mots cette méthode a déterminé comme étant les plus importants.<br>Tf-idf a pour but de trouver des mots qui permettent de distinguer des documents. Il s'agit d'une technique purement non supervisée. |
| **N-grammes**<br>*Recherche de la meilleure valeur possible pour la plage des n-grammes à partir de notre précédent tf-idf* | Accuracy = 88% | 939 | Les performances sont équivalentes en ajoutant les bigrammes et trigrammes. |
| **Racinisation** (Stemming)<br>*Représentation des mots suivant leurs racines* | Accuracy = 87% | 2342 | Ces deux techniques n'apportent pas d'amélioration avec notre jeu de données.<br>**Essai des dictionnaires contenus dans les librairies nltk.** |
| **Lemmatisation**<br>*Utilisation d'un dictionnaire de formes pour représenter les mots sous leur forme simplifiée ou verbale* | Accuracy = 87% | 2371 | |

# Word embedding – Word2Vec

**Définition :** En intelligence artificielle et en apprentissage machine, **Word2vec** est un groupe de modèles utilisé pour le plongement lexical (*word embedding*). Ce sont des réseaux de neurones artificiels à deux couches entraînés pour reconstruire le contexte linguistique des mots.
Cette technique permet de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels.

**Principe :**

- Chaque description deviendra un tableau unidimensionnel de taille 1000 et de type float32.

```
array([-1.16553428e-02, -1.02846347e-01, -1.22004480e-03, -4.45277616e-02,
       -4.24539521e-02, -3.97195108e-02, -1.94274797e-03,  1.77300535e-02,
        2.57900823e-02,  8.14499035e-02, -6.14679642e-02, -5.50374202e-02,
       -7.84587786e-02, -4.14693467e-02,  1.88364126e-02, -7.50013888e-02,
       -2.59726495e-03, -5.85863888e-02, -1.55507389e-03,  8.63027796e-02,
       -7.57160829e-03, -3.65603641e-02, -8.28375388e-03,  1.22139575e-02,
        1.29484385e-02, -2.55026761e-02, -3.34958918e-03,  5.33713959e-04,
        2.56275758e-02, -7.96539336e-03,  4.18626629e-02, -1.90516934e-02,
        3.62377353e-02, -3.32955755e-02,  2.14239024e-02, -7.01822639e-02,
        1.96042247e-02,  5.00909351e-02, -5.26802316e-02,  7.18225632e-03,
```

**Explications :**
https://openclassrooms.com/fr/courses/4470541-analysez-vos-donnees-textuelles/4855006-effectuez-des-plongements-de-mots-word-embeddings

un vecteur en data science est une structure de données linéaire qui stocke des informations essentielles pour l'analyse et la modélisation des données.

# BERT

**Définition :** **BERT, acronyme de "Bidirectional Encoder Representations from Transformers", est un modèle de traitement automatique du langage naturel (TALN) basé sur l'architecture des transformers. Conçu par Google, BERT est particulièrement remarquable pour sa capacité à comprendre le contexte d'un mot dans une phrase en examinant les mots qui l'entourent, à la fois avant et après. Cette approche bidirectionnelle améliore considérablement la compréhension du sens des mots dans un texte. BERT a révolutionné de nombreuses applications de TALN, notamment la traduction automatique, la recherche en langage naturel, la réponse aux questions et la génération de texte en produisant des représentations linguistiques de haute qualité.**

```
tensor([[-9.4021e-02, -2.9146e-02,  4.6952e-01, -3.0623e-02,  3.9867e-01,
         -1.7665e-01, -1.6236e-05,  3.5967e-01,  7.3603e-02, -9.3743e-02,
          9.8127e-03, -5.2392e-01, -3.0431e-02,  4.3858e-01, -2.9220e-01,
          4.3036e-01,  2.7268e-01,  6.3349e-02, -2.2681e-01,  2.5400e-01,
          5.7116e-01, -2.0690e-01,  3.4550e-02,  4.5737e-01,  2.3544e-01,
         -6.7621e-02,  1.1306e-01,  2.2832e-02,  1.4110e-01,  5.9371e-02,
          5.0839e-01,  8.1406e-02, -5.0256e-02, -1.2258e-01,  4.2087e-01,
         -1.3788e-01, -2.0721e-01,  4.4601e-03, -1.8902e-01,  1.5709e-01,
         -3.7609e-01, -2.8341e-01, -2.2375e-02, -3.2211e-02, -2.5398e-01,
         -1.7931e-01,  1.9198e-02,  1.2296e-01, -1.6139e-01,  2.1508e-01,
         -1.4844e-01,  2.0538e-01, -1.1922e-01, -2.6250e-01,  3.1684e-01,
          6.4786e-01,  3.5757e-02, -1.3180e-01, -3.2027e-01, -1.6303e-01,
          1.6238e-01, -2.2374e-01,  9.4102e-02, -9.6309e-02,  1.7991e-01,
          2.1243e-01, -1.9782e-02,  1.9077e-01, -5.8116e-01, -1.2853e-01,
         -4.9059e-01, -7.5758e-02,  2.3059e-01,  7.4979e-02, -6.2412e-02,
```

un "tensor" est une structure de données multidimensionnelle essentielle pour stocker et manipuler des informations.

# IV – DONNEES VISUELLES

# ANALYSE IMAGES

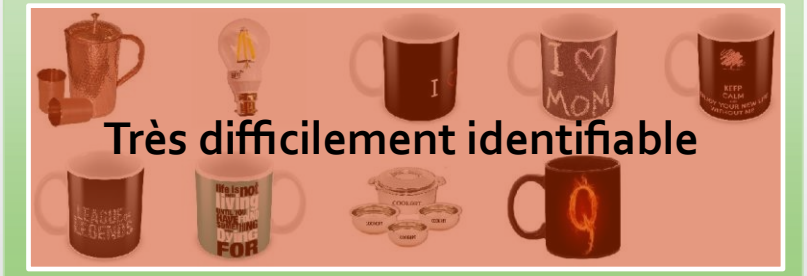
## Home furnishing


## Baby care


## Watches


## Kitchen & Dining


## Decor & Festive needs




## Computers


## Beauty and Personal care


- **Après analyses :**
  - Des images de tailles différentes
  - Des images à recontraster pour faciliter l'extraction de features

- **Premier point de vue :**
  - En regardant simplement les images, on peut remarquer les catégories qui sont facilement identifiables ou non. La classification nous confirmera, ou non, ce point de vue.

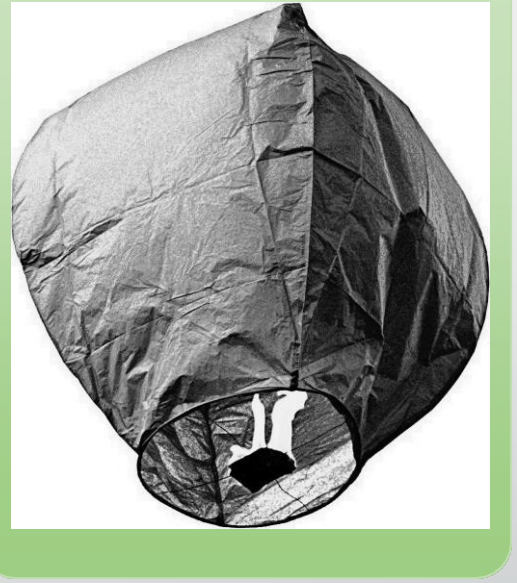
- **Modèles utilisés pour encodage des images :**
  - Algorithme ORB pour les BoVW (Bag of Visual Words)
  - Transfer Learning :
    - VGG16
    - ResNet50

# TRANSFORMATIONS IMAGES GRAYSCALE

Pour certains des algorithmes  comme ORB, nous devons convertir nos images en niveaux de gris.

**Image originale**

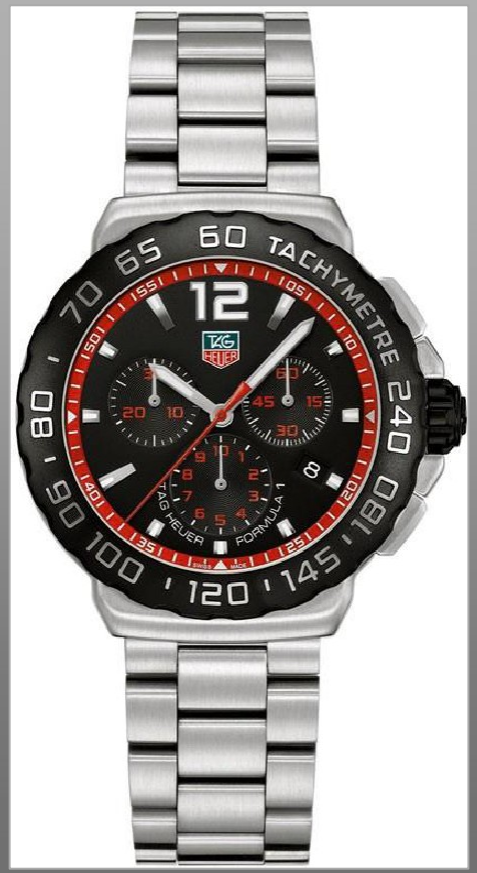
**Conversion de l'image en niveau de gris**



**<u>Amélioration de l'image avec :</u>**

- Egalisation de l'histogramme
- Réduction du bruit par un filtre médian

# TRANSFORMATIONS IMAGES COULEURS

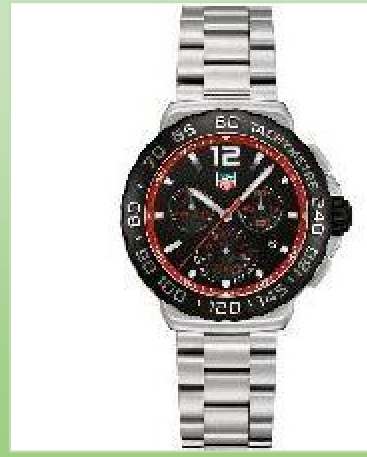


Dimensions initiales de l'image

416x784

**Transformation de la taille des images en 224x224**

VGG16 et ResNet50 sont des modèles qui acceptent des images de taille 224x224 en entrée (par défaut).



**Préservation du ratio Hauteur/Largeur des images**

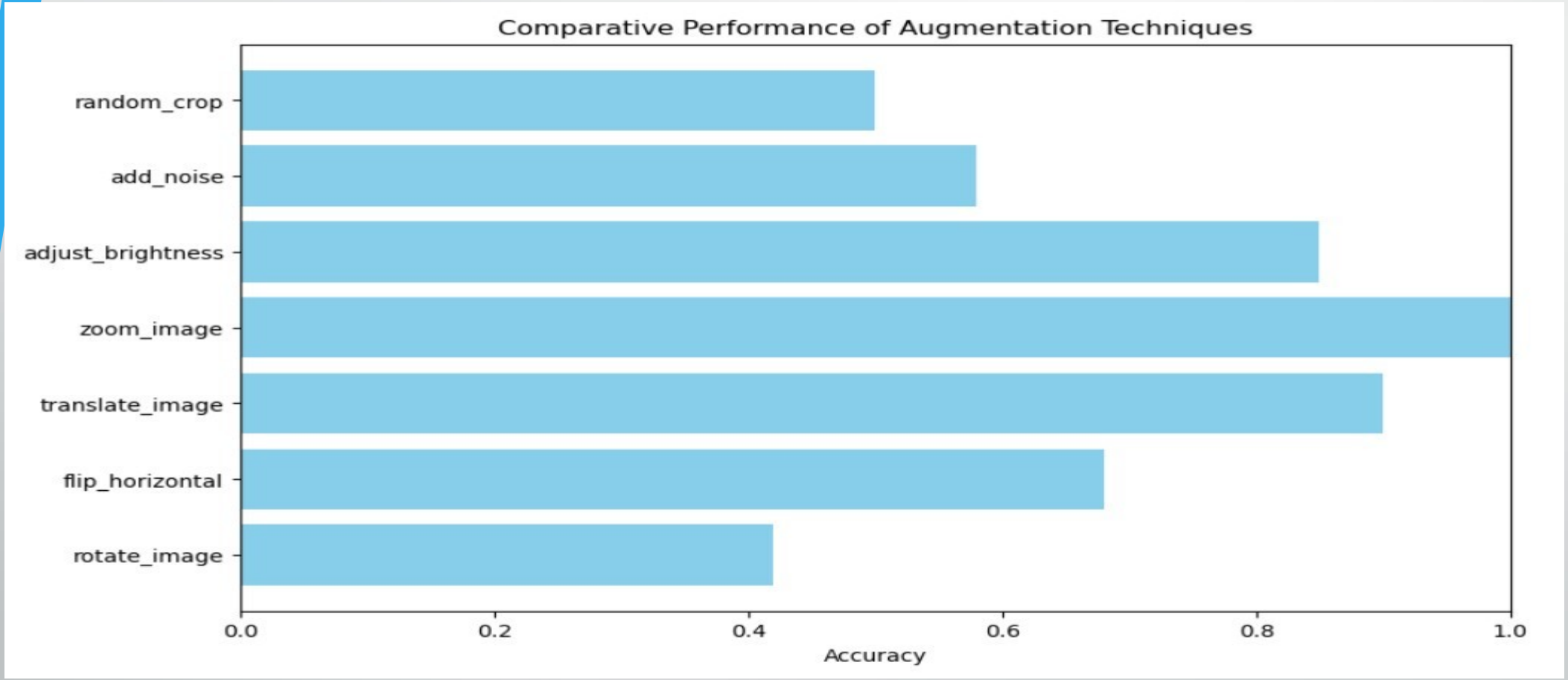
Afin de garder le ratio initial des images avec la nouvelle dimension, la hauteur ou la largeur suivant le cas, ont été comblées par de la couleur blanche.

**Amélioration du contraste**

Afin d'améliorer l'extraction de features sur les image, le contraste a été amélioré.

**Exemple d'image plus claire recontrastée**

# Comparaison des performances des techniques d’augmentation



# Bag of Visual Words

**Définition de l'algorithme ORB (Oriented FAST and Rotated BRIEF) :**

- ORB est un algorithme de detection de features dans une image. Il est plus performant que SURF et fonctionne aussi bien que SIFT tout en étant plus rapide. Il a aussi l'avantage d'être gratuit .
- ORB s'appuie sur le célèbre détecteur de keypoints FAST et sur le descripteur BRIEF. Ce sont deux techniques intéressantes en raison de leurs bonnes performances.
- Source definition : https://medium.com/@deepanshut041/introduction-to-orb-oriented-fast-and-rotated-brief-4220e8ec40cf

**Extraction de descripteurs**

ORB extrait par défaut 500 descripteurs de 32 valeurs numériques de type float par images.

Réglage de l'extraction sur 1500 descripteurs

Total descripteurs extraits :

**1517331**

**Recherche de mots visuels**

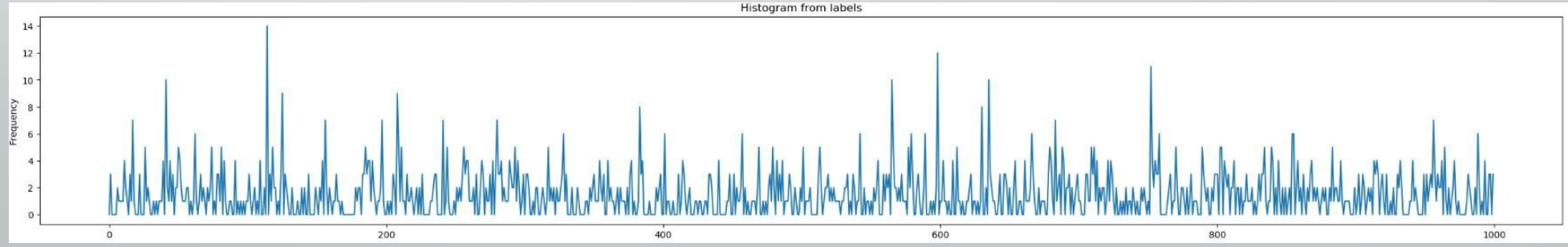
Avec un K-means réglé sur 1000 clusters, nous segmentons chaque image.

Cela nous donne un mot visuel (Visual Word) de dimension 1000 par image.

**Création de l'histogramme**

Afin de visualiser à quoi ressemble un mot visuel, nous l'avons représenté sous forme d'histogramme.

**Exemple avec la 2e image de notre jeu de données**



# Réseau de neurones convolutif (CNN)

Un CNN fonctionne un peu comme notre cerveau lorsqu'il perçoit des formes et des objets. Voici comment il fonctionne :

1. Convolution : Le CNN divise l'image en petites parties et examine chaque partie à la recherche de motifs, comme des bords, des coins ou d'autres caractéristiques.

2. Pooling : Ensuite, le CNN réduit la taille de l'image tout en préservant les caractéristiques importantes. Cela permet de simplifier l'information sans perdre la signification.

3. Réseau de neurones : Les caractéristiques apprises dans les étapes précédentes sont ensuite envoyées à un réseau de neurones traditionnel qui détermine ce qui se trouve dans l'image (un chat, un chien, etc.).

Ce qui rend les CNN si puissants, c'est qu'ils sont capables d'apprendre automatiquement à partir des données. Par exemple, au lieu de programmer explicitement ce qu'est un bord ou un coin, le CNN apprend ces concepts par lui-même en analysant de nombreuses images. Cela les rend très efficaces pour des tâches comme la reconnaissance d'images.

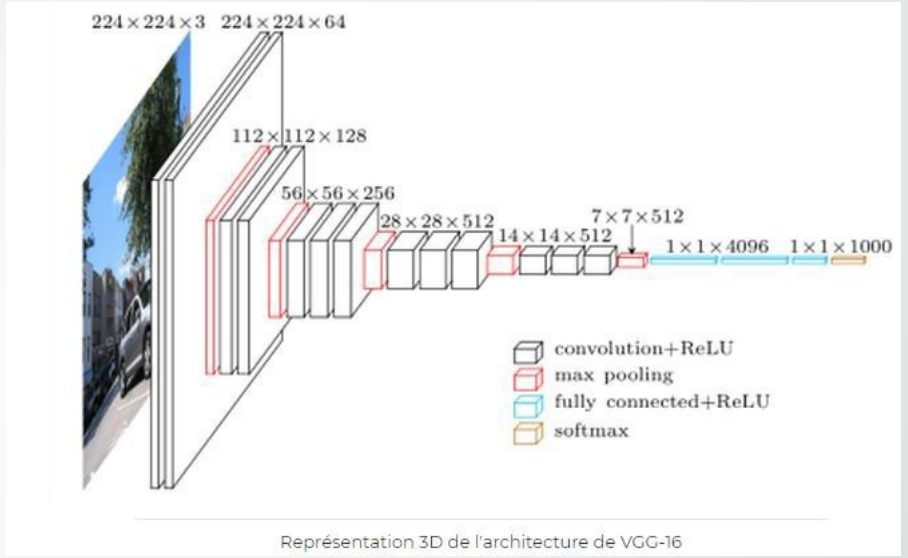
En résumé, un CNN est un type de modèle d'apprentissage profond qui est excellent pour comprendre les images et les données visuelles en général. C'est un peu comme un détective qui cherche des formes et des caractéristiques dans les images pour dire ce qu'elles représentent.

# CNN – Modèle VGG-16

**Définition du modèle VGG-16 :**

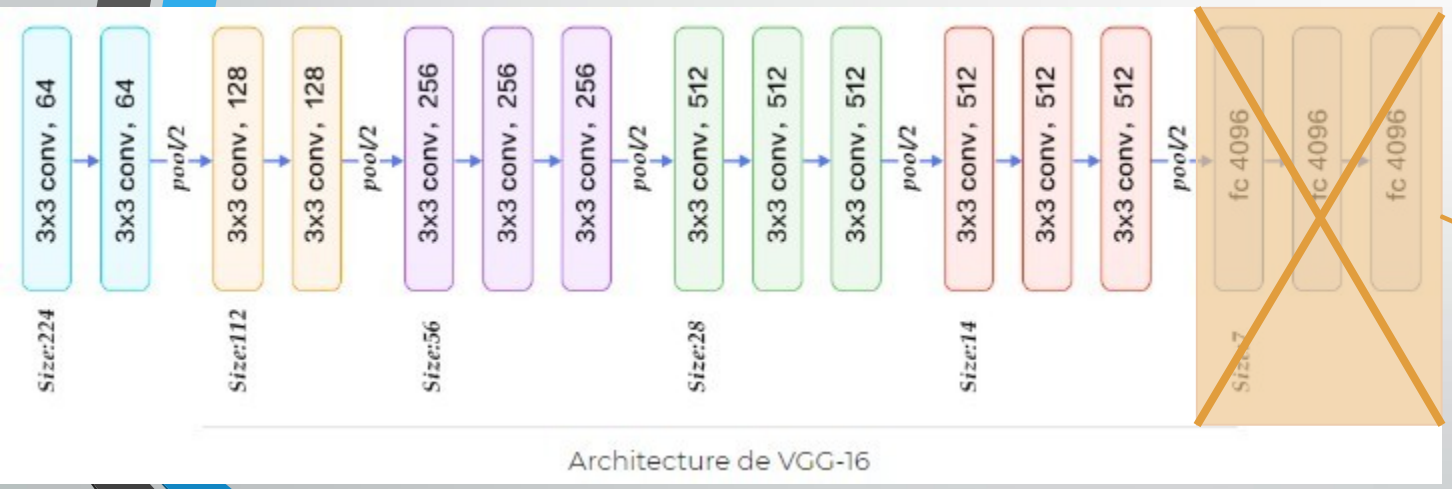
- VGG-16 est une version du réseau de neurones convolutif VGG-Net.
- Il prend en entrée une image en couleurs de dimensions 224x224 et la classifie dans l'une des 1000 classes de ImageNet. Il renvoie donc un vecteur de dimension 1000, qui contient les probabilités d'appartenance à chacune des classes.

- Source définition : https://openclassrooms.com/fr/courses/4470531-classez-et-segmentez-des-donnees-visuelles/5097666-tp-implementez-votre-premier-reseau-de-neurones-avec-keras




Représentation 3D de l'architecture de VGG-16

**Architecture VGG-16**




Architecture de VGG-16

**Transfer Learning :**

- Utilisation d'un modèle pré-entraîné sur ImageNet.
- Suppression des 3 couches fully-connected qui permettent de classifier les images avec ImageNet.
- Taille du vecteur par image : 25088

Model: "vgg16"

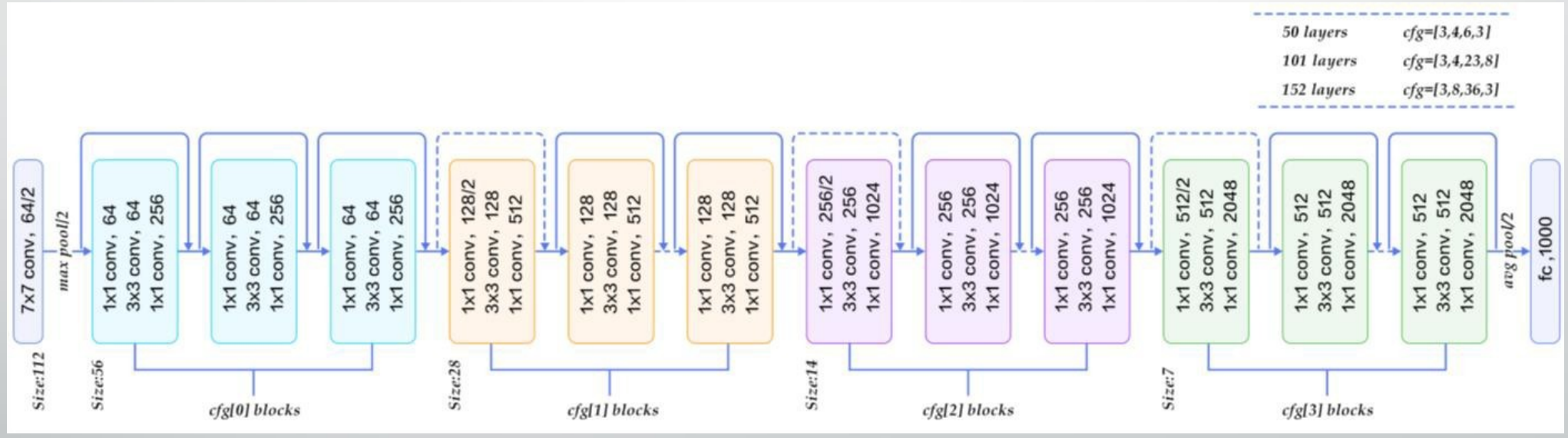
| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_1 (InputLayer) | (None, None, None, 3) | 0 |
| block1_conv1 (Conv2D) | (None, None, None, 64) | 1792 |
| block1_conv2 (Conv2D) | (None, None, None, 64) | 36928 |
| block1_pool (MaxPooling2D) | (None, None, None, 64) | 0 |
| block2_conv1 (Conv2D) | (None, None, None, 128) | 73856 |
| block2_conv2 (Conv2D) | (None, None, None, 128) | 147584 |
| block2_pool (MaxPooling2D) | (None, None, None, 128) | 0 |
| block3_conv1 (Conv2D) | (None, None, None, 256) | 295168 |
| block3_conv2 (Conv2D) | (None, None, None, 256) | 590080 |
| block3_conv3 (Conv2D) | (None, None, None, 256) | 590080 |
| block3_pool (MaxPooling2D) | (None, None, None, 256) | 0 |
| block4_conv1 (Conv2D) | (None, None, None, 512) | 1180160 |
| block4_conv2 (Conv2D) | (None, None, None, 512) | 2359808 |
| block4_conv3 (Conv2D) | (None, None, None, 512) | 2359808 |
| block4_pool (MaxPooling2D) | (None, None, None, 512) | 0 |
| block5_conv1 (Conv2D) | (None, None, None, 512) | 2359808 |
| block5_conv2 (Conv2D) | (None, None, None, 512) | 2359808 |
| block5_conv3 (Conv2D) | (None, None, None, 512) | 2359808 |
| block5_pool (MaxPooling2D) | (None, None, None, 512) | 0 |

# CNN – Modèle ResNet50

**Définition du modèle ResNet50 :**

- VGG-Net présente deux inconvénients majeurs :
  - Il est lent à entraîner
  - Les poids de l'architecture de réseau sont importants (en termes de disque / bande passante), cela peut rendre le déploiement de VGG fastidieux
- Le réseau ResNet est plutôt une forme « d'architecture exotique » qui repose sur des modules de micro- architecture.
- Le terme micro-architecture désigne l'ensemble des «blocs de construction» utilisés pour construire le réseau.
- Un ensemble de blocs de construction de micro-architecture est composé de couches standards (Convolution, Pool, etc.)

- Source définition : https://www.pyimagesearch.com/2017/03/20/imagenet-vggnet-resnet-inception-xception-keras/




**Transfer Learning :**

- Utilisation d'un modèle pré-entraîné sur ImageNet.
- Taille du vecteur par image : 100352.

# V – CLASSIFICATION & CLUSTERING

**Deux méthodes proposées :**

- Classification multi classes : Cela nous permettra d'analyser les performances de classification d'un nouvel article.
- Clustering : Cela nous permettra d'analyser la qualité de notre feature engineering.

# PREPARATION DES DONNEES

| Itération | Description | Taille vecteur initale | Taille vecteur totale après concaténation | ACP (90% d'infos gardées) | NMF (Réduction à 7 composants) | Taille vecteur après réduction dimensionnelle |
|---|---|---|---|---|---|---|
| 1 | Tf-Idf | 2371 | 2371 | Non | Oui | 7 |
| 2 | BoVW | 1000 | 1000 | Non | Oui | 7 |
| 3 | VGG-16 | 25088 | 25088 | Non | Oui | 7 |
| 4 | ResNet50 | 100352 | 100352 | Non | Oui | 7 |
| 5 | Texte : Tf-Idf<br>Image : BoVw | 2371<br>1000 | 3371 | Oui | Non | 431 |
| 6 | Texte : Tf-Idf Image : BoVw | 2371<br>1000 | 3371 | Non | Oui | 7 |
| 7 | Texte : Tf-Idf Image : VGG-16 | 2371<br>25088 | 27459 | Oui | Non | 584 |
| 8 | Texte : Tf-Idf Image : VGG-16 | 2371<br>25088 | 27459 | Non | Oui | 7 |
| 9 | Texte : Word2Vec Image : BoVW | 1000<br>1000 | 2000 | Oui | Non | 448 |
| 10 | Texte : Word2Vec Image : VGG-16 | 1000<br>25088 | 26088 | Oui | Non | 584 |
| 11 | Texte : Word2Vec<br>Image : VGG-16 | 1000<br>25088 | 26088 | Non | Oui sur VGG-16 | 1007 |
| 12 | Texte : Tf-Idf Image : ResNet50 | 2371<br>100352 | 102723 | Oui | Non | 665 |
| 13 | Texte : Tf-Idf Image : ResNet50 | 2371<br>100352 | 102723 | Non | Oui | 7 |
| 14 | Texte : Word2Vec<br>Image : ResNet50 | 1000<br>100352 | 101352 | Non | Oui sur ResNet50 | 10007 |

# CLASSIFICATION MULTICLASSES

**Mise en place de modèles de classification :**

- Mesure de performance par l'accuracy : Le pourcentage des prédictions correctes sur le total des prédictions.
- classification accuracy = correct predictions / total predictions
- Nous avons 7 classes (nos categories ici), et l'accuracy nous donne une idée globale de la performance des modèles, mais nous ne savons pas si toutes les classes sont prédites de la même manière, ou si une ou deux sont négligées par le modèle.
- Nous allons donc mettre en place une matrice de confusion, ce qui nous montrera les bonnes et mauvaises predictions pour chaque classe.

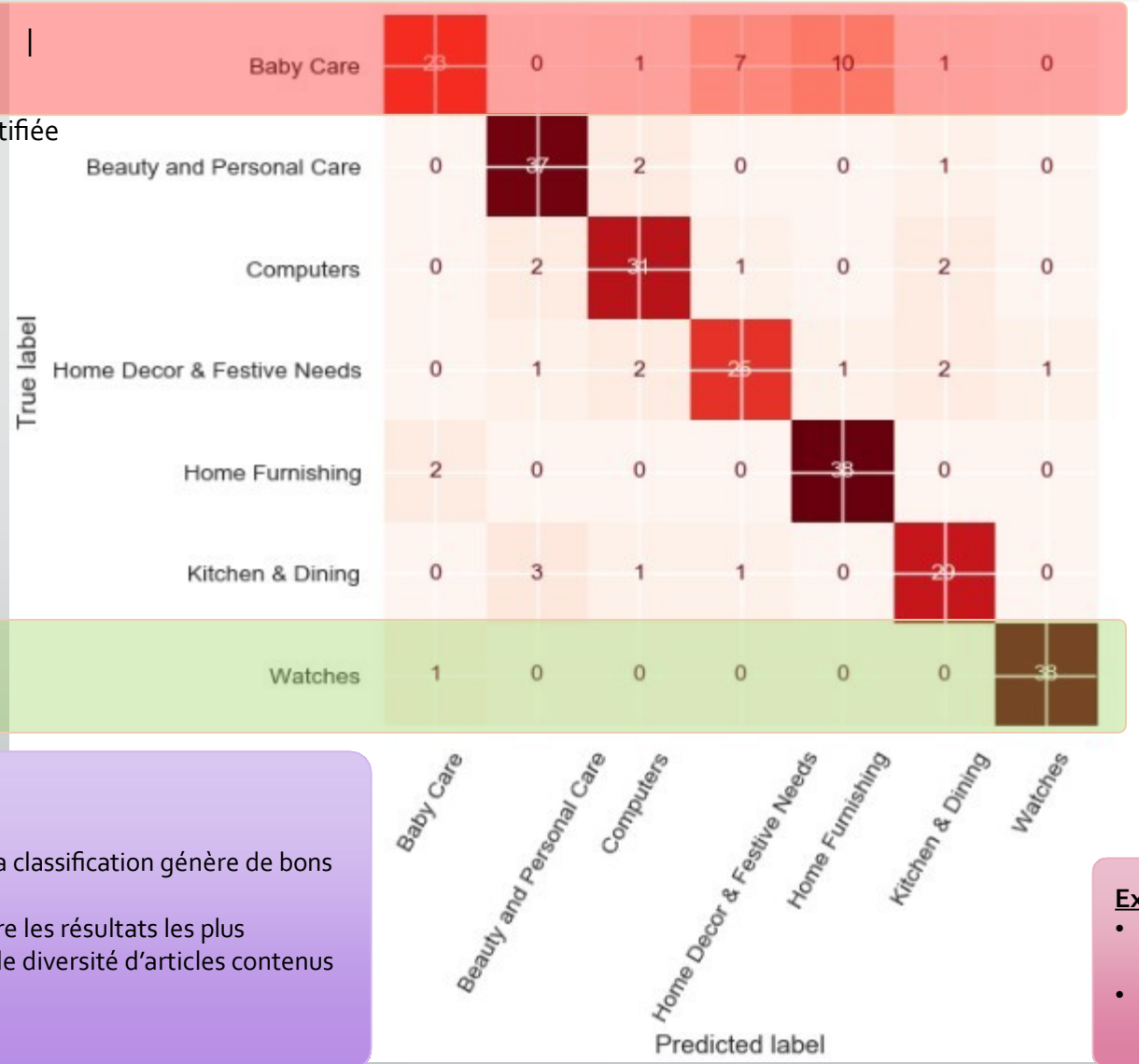
| iteration | bayes | tree | svm | knn |
|---|---|---|---|---|
| 1 | 0.49 | 0.33 | 0.54 | 0.79 |
| 2 | 0.18 | 0.19 | 0.22 | 0.21 |
| 3 | 0.56 | 0.35 | 0.62 | 0.61 |
| 4 | 0.72 | 0.33 | 0.77 | 0.75 |
| 5 | 0.4 | 0.27 | 0.78 | 0.57 |
| 6 | 0.21 | 0.17 | 0.25 | 0.24 |
| 7 | 0.51 | 0.32 | 0.8 | 0.66 |
| 8 | 0.56 | 0.35 | 0.62 | 0.61 |
| 9 | 0.27 | 0.21 | 0.36 | 0.33 |
| 10 | 0.52 | 0.32 | 0.8 | 0.66 |
| 11 | 0.62 | 0.38 | 0.83 | 0.71 |
| 12 | 0.46 | 0.41 | 0.85 | 0.76 |
| 13 | 0.72 | 0.33 | 0.77 | 0.75 |
| 14 | 0.62 | 0.38 | 0.84 | 0.76 |

# MATRICE DE CONFUSION

Notre échantillon de test contient 263 articles.
Meilleure itération retenue : Textes : Tf-IDF  Images : ResNet50 | Réduction dimensionnelle : NMF

Moins bien identifiée

Mieux identifiée




| Précision par catégorie | |
|---|---|
| 42 articles | Précision à 55% |
| 40 articles | Précision à 93% |
| 36 articles | Précision à 86% |
| 32 articles | Précision à 78% |
| 40 articles | Précision à 95% |
| 34 articles | Précision à 85% |
| 39 articles | Précision à 97% |
| 263 articles | |

**Analyse :**

- On remarque que globalement la classification génère de bons résultats.
- La catégorie « Baby Care » génère les résultats les plus mauvais. Cela est dû a une grande diversité d'articles contenus dans celle-ci.

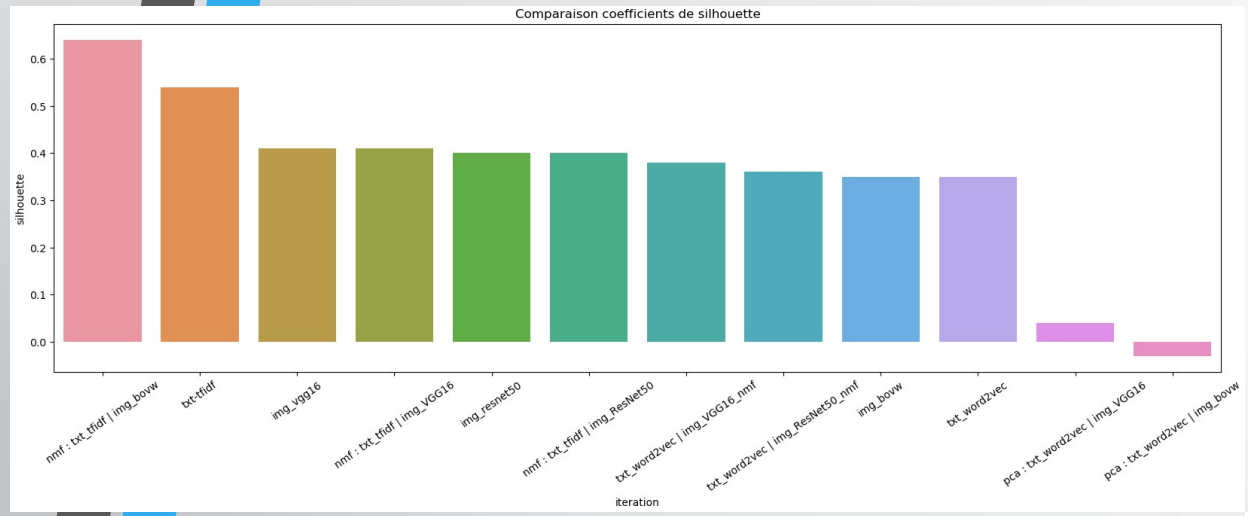
**Exemple d'ajout de produit :**

- Montre : Le produit a 97% de chance d'être classé dans la catégorie adéquate.
- Un produit pour bébé : Le produit a 1 chance sur 2 d'être mal classé.

# CLUSTERING

**Construction clustering :**
Le clustering a été effectué par un modèle K-means optimisé par GridSearchCV.
Le modèle a été paramétré pour un clustering en 7 clusters.

ARI mesure à quel point les groupes créés par un algorithme de clustering correspondent aux groupes réels, tandis que le Coefficient de Silhouette mesure la qualité des groupes eux-mêmes, en évaluant à quel point les objets d'un même groupe sont similaires les uns aux autres et différents des objets des autres groupes. Plus ces mesures sont proches de 1, meilleures sont les performances de votre algorithme de clustering.




**Coefficient de silhouette :**

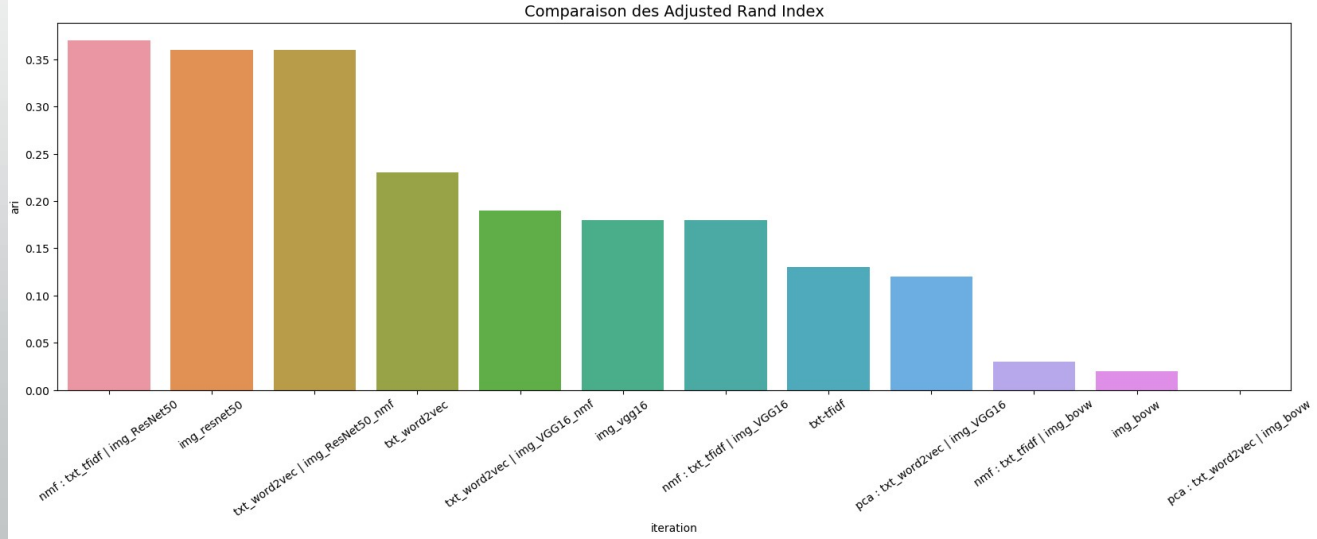
- Le coefficient de silhouette permet de mesurer la densité (homogénéité) et la séparation des clusters.
- Il est compris entre -1 et 1, sachant que 1 correspond à la plus grande qualité du clustering.

**Adjusted Rand Index (ARI) :**

- L'ARI permet d'évaluer si les 7 groupes formés par le k-means correspondent à nos 7 catégories initiales (sous_cat_1).
- IL est compris entre 0 et 1, sachant que 1 est un clustering qui correspond exactement à nos catégories initales.

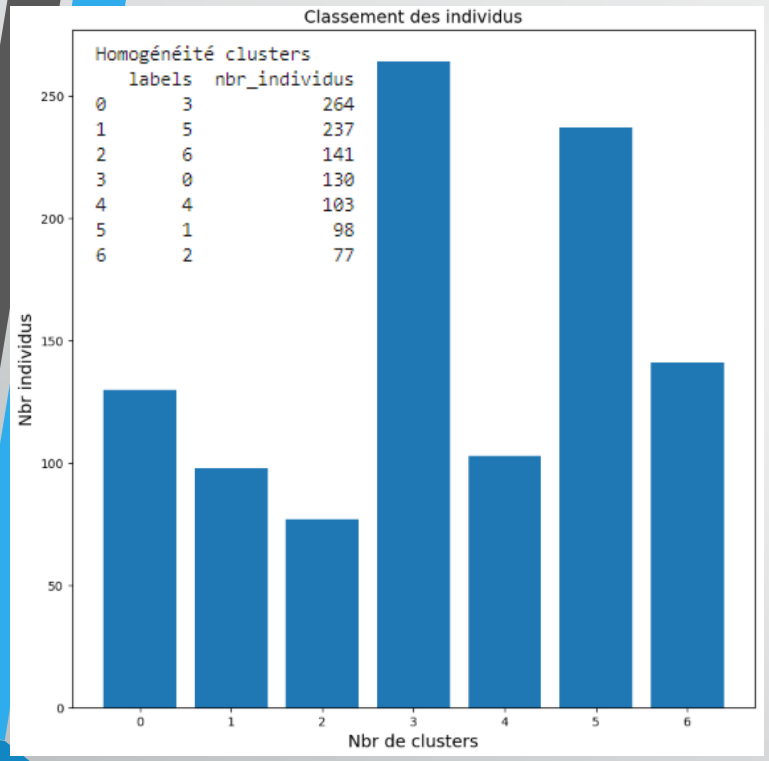
Ici, c'est l'ARI qui nous intéresse le plus. Nous allons donc analyser la meilleure itération sur cette mesure :

- Texte : Tf-Idf | Image : ResNet50 | Réduction dimensionnelle : NMF



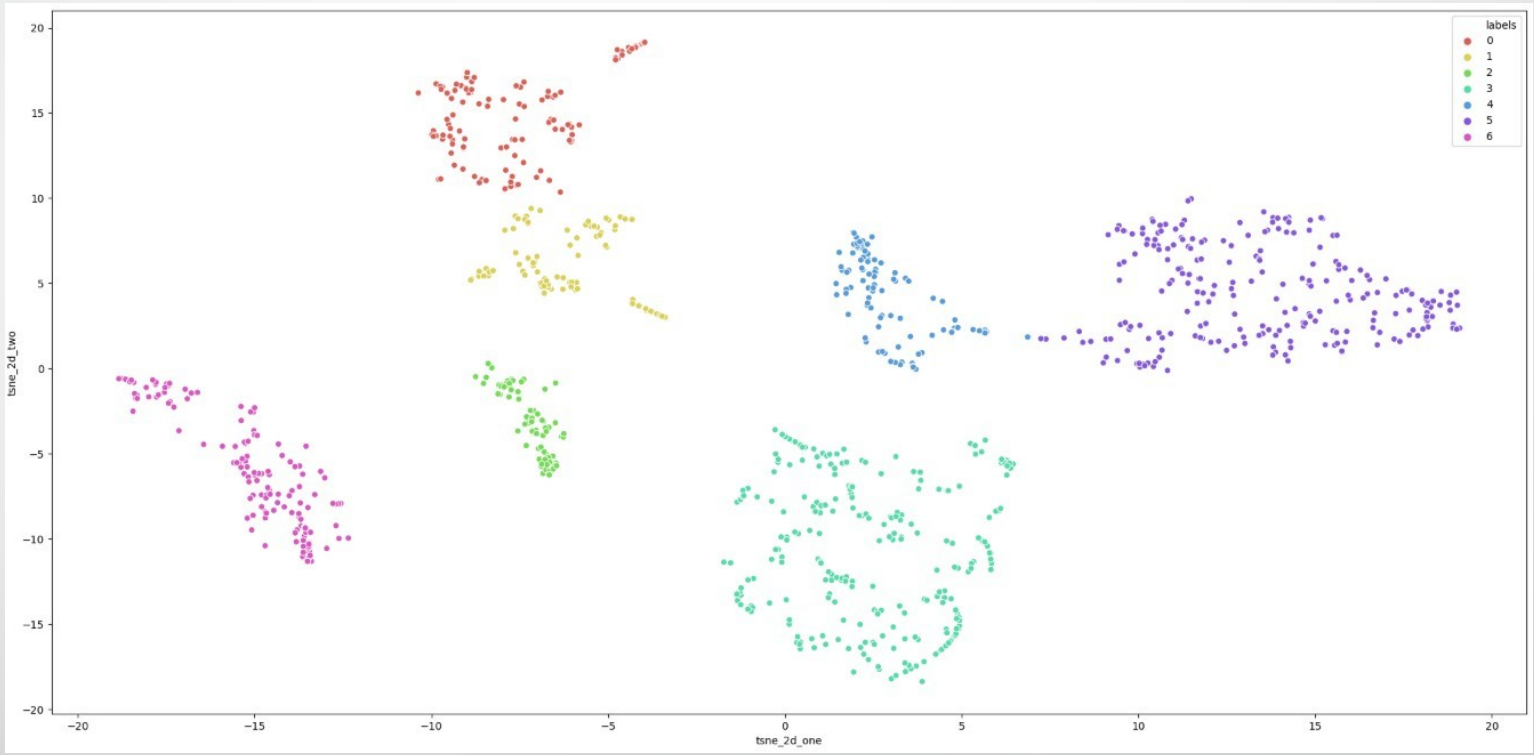
# ANALYSE MEILLEUR CLUSTERING

## Homogénéité clusters




- Pour rappel, nos 7 catégories initiales comportent 150 articles chacune.
- Ici nous avons des clustering assez bien répartis, mais qui ne correspondent pas à nos categories initiales.

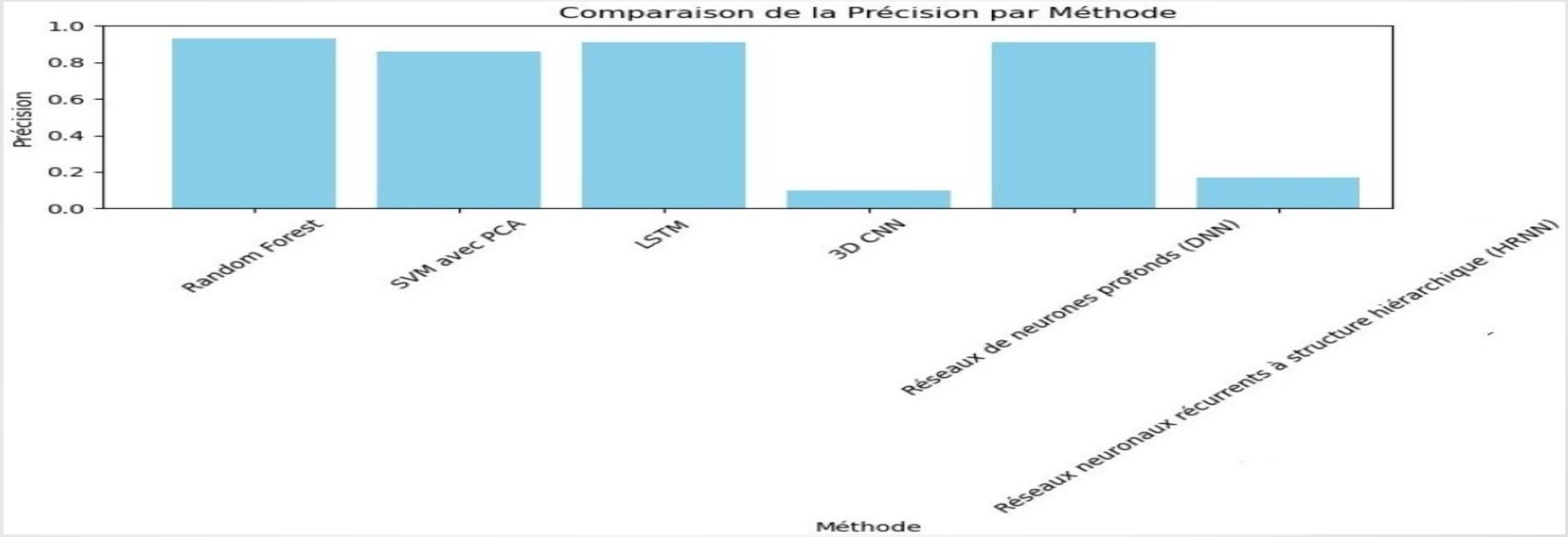
## TSNE



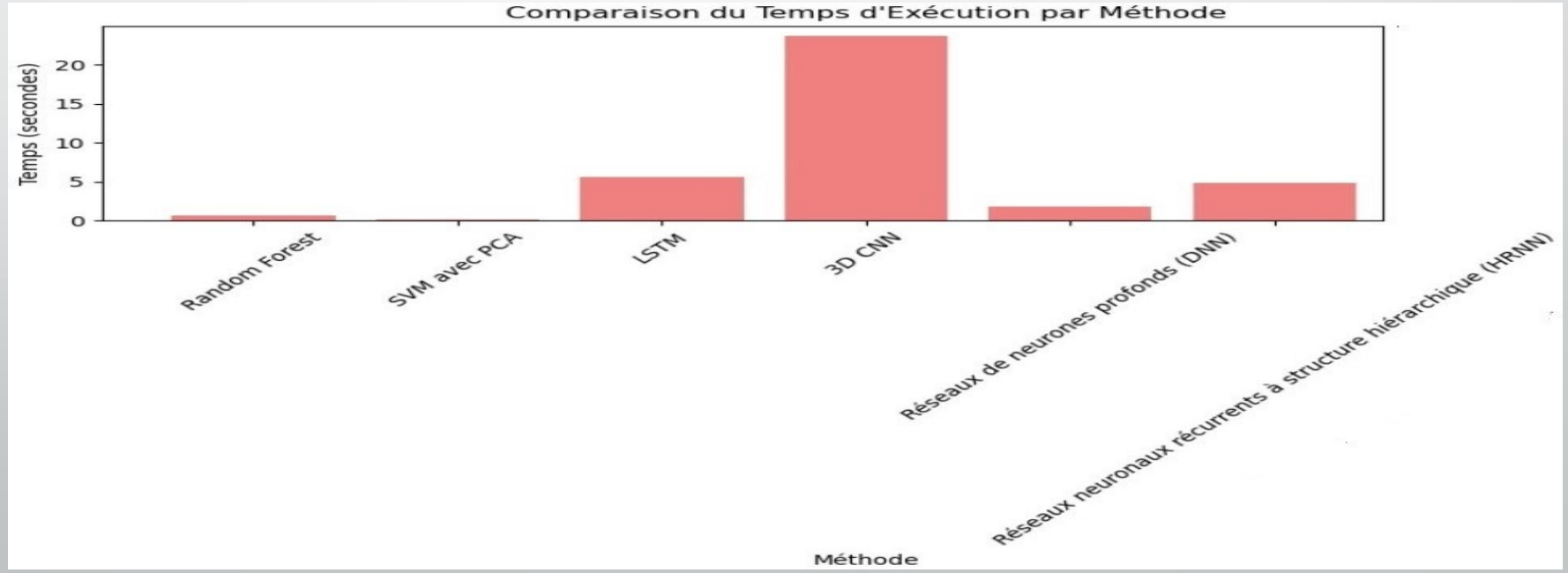

- La TSNE nous montre bien des clusters assez denses, et espacés.

# ANALYSE MEILLEURE MÉTHODE

**Précisions**



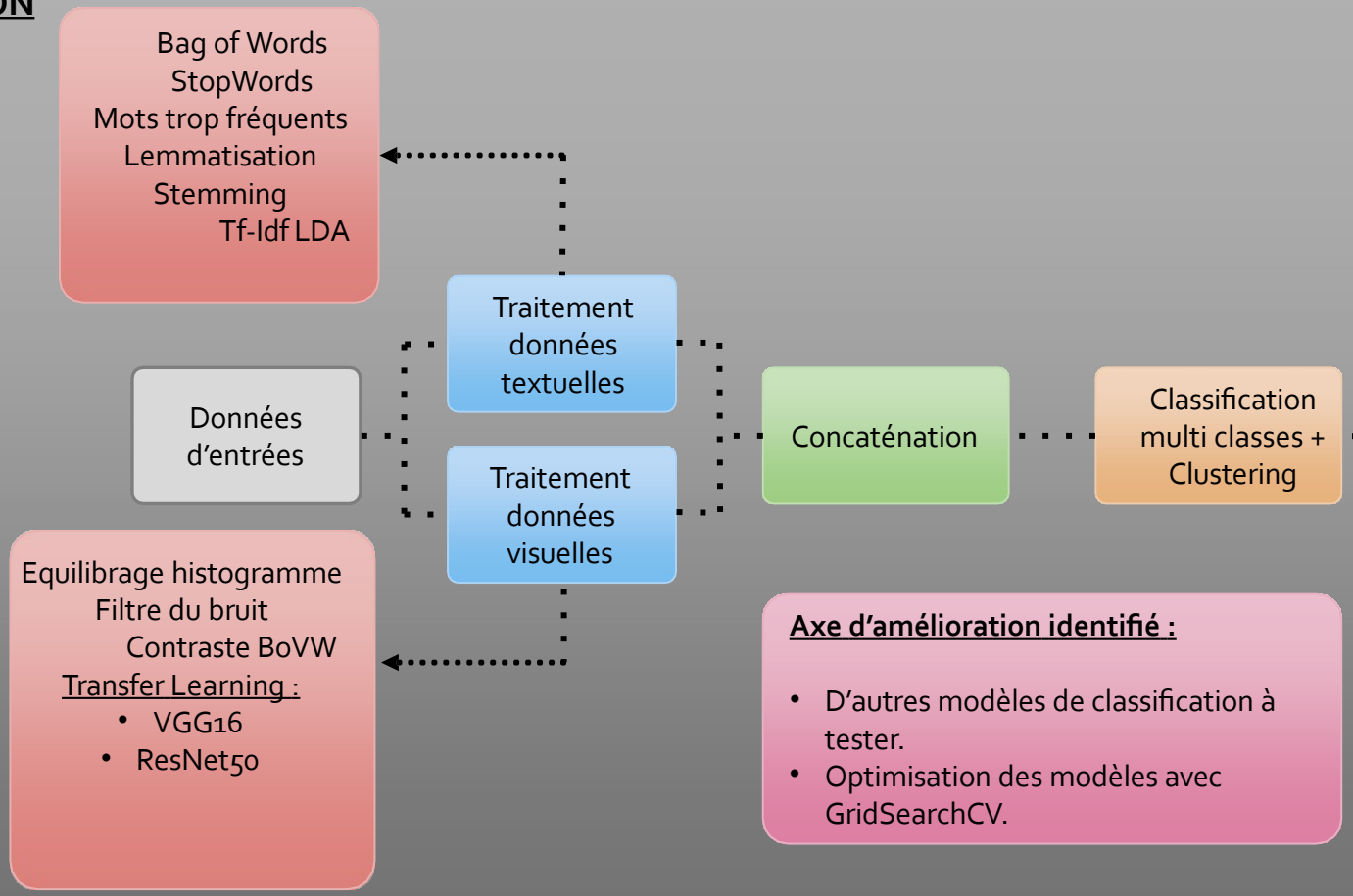

**Temps d'éxecution**



# VI – CONCLUSION

# RESUME

**CLASSIFICATION**




**Conclusion sur l'étude de faisabilité d'un moteur classification**

- Utilisation de réseaux neuronaux pour avoir de meilleurs résultats.
- Il faut densifier la base de données pour améliorer les performances
- D'autres modèles de réseaux neuronaux sont à essayer, notamment Universal Sentence Encoder de Google dont les performances sont très bonnes.

. . . . .

# API

**Fonctionnement de l'API :**

- Le fonctionnement se fait en plusieurs étapes :
  1. La première étape est de s'inscrire sur le site
  2. Ensuite, nous devons identifier la fonction ou argument de base que nous allons utiliser.
  3. Nous devons envoyer une requête à l'API.

## Requete envoyé à l'api

```
api_url = "https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"
headers = {
  'X-RapidAPI-Key': "82e7f76641msh8efe42731f3e34bp1164ddjsn1d61cc9d0dc1",
  'X-RapidAPI-Host': "edamam-food-and-grocery-database.p.rapidapi.com"  }
```

## Parametre de la requete

```
params = {
    "ingr": "champagne",
    "nutrition-type": "cooking",
    "category[0]": "generic-foods",
    "health[0]": "alcohol-free"  }
```

## Sélection des colonnes

```
selected_columns = ["foodId", "label", "category", "categoryLabel", "knownAs", "image"]
filtered_data = filter_and_select_data(api_data, selected_columns)
```

## Résultat de la requête(récupération des données)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | foodId | label | category | categoryLabel | knownAs | image |
| 2 | food_a656mk2a5dmqb2adiamu6beihduu | Champagne | Generic foods | food | dry white wine | https://www.edamam.com/food-img/a71/a718cf3c52add522128929f1f324d2ab.jpg |
| 3 | food_b753ithamdb8psbt0w2k9aquo06c | Champagne Vinaigrette, Champagne | Packaged foods | food | CHAMPAGNE VINAIGRETTE, CHAMPAGNE | |
| 4 | food_b3dyababjo54xobm6r8jzbghjgqe | Champagne Vinaigrette, Champagne | Packaged foods | food | CHAMPAGNE VINAIGRETTE, CHAMPAGNE | https://www.edamam.com/food-img/d88/d88b64d97349ed062368972113124e35.jpg |
| 5 | food_a9e0ghsamvoc45bwa2ybsa3gken9 | Champagne Vinaigrette, Champagne | Packaged foods | food | CHAMPAGNE VINAIGRETTE, CHAMPAGNE | |
| 6 | food_an4jjueaucpus2a3u1ni8auhe7q9 | Champagne Vinaigrette, Champagne | Packaged foods | food | CHAMPAGNE VINAIGRETTE, CHAMPAGNE | |
| 7 | food_bmu5dmkazwuvpaa5prh1daa8jxs0 | Champagne Dressing, Champagne | Packaged foods | food | CHAMPAGNE DRESSING, CHAMPAGNE | https://www.edamam.com/food-img/ab2/ab2459fc2a98cd35f68b848be2337ecb.jpg |
| 8 | food_alpl44taoyv11ra0lic1qa8xculi | Champagne Buttercream | Generic meals | meal | Champagne Buttercream | |
| 9 | food_byap67hab6evc3a0f9w1oag3s0qf | Champagne Sorbet | Generic meals | meal | Champagne Sorbet | |
| 10 | food_am5egz6aq3fpjlaf8xpkdbc2asis | Champagne Truffles | Generic meals | meal | Champagne Truffles | |
| 11 | food_bcz8rhiajk1fuva0vkfmeakbouc0 | Champagne Vinaigrette | Generic meals | meal | Champagne Vinaigrette | |
| 12 | food_a79xmnya6togreaeukbroa0thhh0 | Champagne Chicken | Generic meals | meal | Champagne Chicken | |
| 13 | food_aoxaf73b3o0igebpj6wjga6kqhco | Strawberry Champagne | Generic meals | meal | Strawberry Champagne | |
| 14 | food_ax1n26waalpd9cbc64bjob7pw6hg | Champagne Jelly | Generic meals | meal | Champagne Jelly | |
| 15 | food_b4va8u0bb6pf74akh2rtcb3llna9 | Champagne Punch | Generic meals | meal | Champagne Punch | |
| 16 | food_a4j8wm8ayflf13b45t3c3bk9w4ek | Champagne Sangria | Generic meals | meal | Champagne Sangria | |
| 17 | food_bw7gtgxbnn7nbwa62ppwpar9ljc1 | Champagne Cotton Candy, Champagne | Packaged foods | food | CHAMPAGNE COTTON CANDY, CHAMPAGNE | |
| 18 | food_bu12urpbtuo9v6b4jpvk2a1fh4hh | Champagne Simply Dressed Vinaigrette, Champagne | Packaged foods | food | CHAMPAGNE SIMPLY DRESSED VINAIGRETTE, CHAMPAGNE | https://www.edamam.com/food-img/736/736a3e27a63d799d4073d2774075b509.png |
| 19 | food_bba727vaimolf0b8stgoibx7ujei | Champagne Cake | Generic meals | meal | Champagne Cake | |
| 20 | food_a6mj2obbqy38soat01vrxaqnvvet | Champagne Cupcakes | Generic meals | meal | Champagne Cupcakes | |
| 21 | food_aj3tbbpbl068bhagn76uubtzyzyv | Champagne Vinegar | Packaged foods | food | CHAMPAGNE VINEGAR | |

# RGPD

## 1 - NE COLLECTEZ QUE LES DONNÉES VRAIMENT NÉCESSAIRES POUR ATTEINDRE VOTRE OBJECTIF

Les données sont collectées pour un but bien déterminé et légitime et ne sont pas traitées ultérieurement de façon incompatible avec cet objectif initial.
Le principe de finalité limite la manière dont vous pourrez utiliser ou réutiliser ces données dans le futur et évite la collecte de données « au cas où ».
Le principe de minimisation limite la collecte aux seules données strictement nécessaires à la réalisation de votre objectif.

## 2 - SOYEZ TRANSPARENT

Les individus doivent conserver la maîtrise des données qui les concernent. Cela suppose qu'ils soient clairement informés de l'utilisation qui sera faite de leurs données dès leur collecte. Les données ne peuvent en aucun cas être collectées à leur insu. Les personnes doivent également être informées de leurs droits et des modalités d'exercice de ces droits.

## 3 - ORGANISEZ ET FACILITEZ L'EXERCICE DES DROITS DES PERSONNES

Vous devez organiser des modalités permettant aux personnes d'exercer leurs droits et répondre dans les meilleurs délais à ces demandes de consultation ou d'accès, de rectification ou de suppression des données, voire d'opposition, sauf si le traitement répond à une obligation légale (par exemple, un administré ne peut s'opposer à figurer dans un fichier d'état civil). Ces droits doivent pouvoir s'exercer par voie électronique à partir d'une adresse dédiée.

## 4 - FIXEZ DES DURÉES DE CONSERVATION

Vous ne pouvez pas conserver les données indéfiniment.

Elles ne sont conservées en « base active », c'est-à-dire la gestion courante, que le temps strictement nécessaire à la réalisation de l'objectif poursuivi. Elles doivent être par la suite détruites, anonymisées ou archivées dans le respect des obligations légales applicables en matière de conservation des archives publiques.

## 5 - SÉCURISEZ LES DONNÉES ET IDENTIFIEZ LES RISQUES

Vous devez prendre toutes les mesures utiles pour garantir la sécurité des données : sécurité physique ou sécurité informatique, sécurisation des locaux, armoires et postes de travail, gestion stricte des habilitations et droits d'accès informatiques. Cela consiste aussi à s'assurer que seuls les tiers autorisés par des textes ont accès aux données.
Ces mesures sont adaptées en fonction de la sensibilité des données ou des risques qui peuvent peser sur les personnes en cas d'incident de sécurité.

## 6 - INSCRIVEZ LA MISE EN CONFORMITÉ DANS UNE DÉMARCHE CONTINUE

La conformité n'est pas gravée dans le marbre et figée.
Elle dépend du bon respect au quotidien par les agents, à tous les niveaux, des principes et mesures mis en œuvre
Vérifiez régulièrement que les traitements n'ont pas évolué, que les procédures et les mesures de sécurité mises en place sont bien respectées et adaptez-les si besoin.

# QUESTIONS - REPONSES