Classifier automatiquement des biens de consommation

Soutenance Projet 6 Data Scientist Agnès Regaud septembre 2023

Sommaire

1.Problématique

Projet de classification à partir d'un texte ou d'une image

- Entreprise « Place de marché » marketplace e-commerce
- Actuellement catégorisation d'article manuelle
- Problème : peu fiable
- But : automatiser la tâche
- Missions :
  - ✓ étude de faisabilité d'un moteur de classification
  - ✓ classification supervisée à partir des images
  - ✓ test de collecte de produits via API

2. Prétraitements, extractions de features et résultats de l’étude de faisabilité

2.1 Analyse, Prétraitement

Fichier flipkart_com-ecommerce_sample_1050.csv

- 1050 lignes, 15 colonnes
- Pas de ligne en double
- Pas de valeur manquante pour les 3 variables pertinentes : « description », « product_category_tree » et « image »
- Catégorie du produit → « product_category_tree »

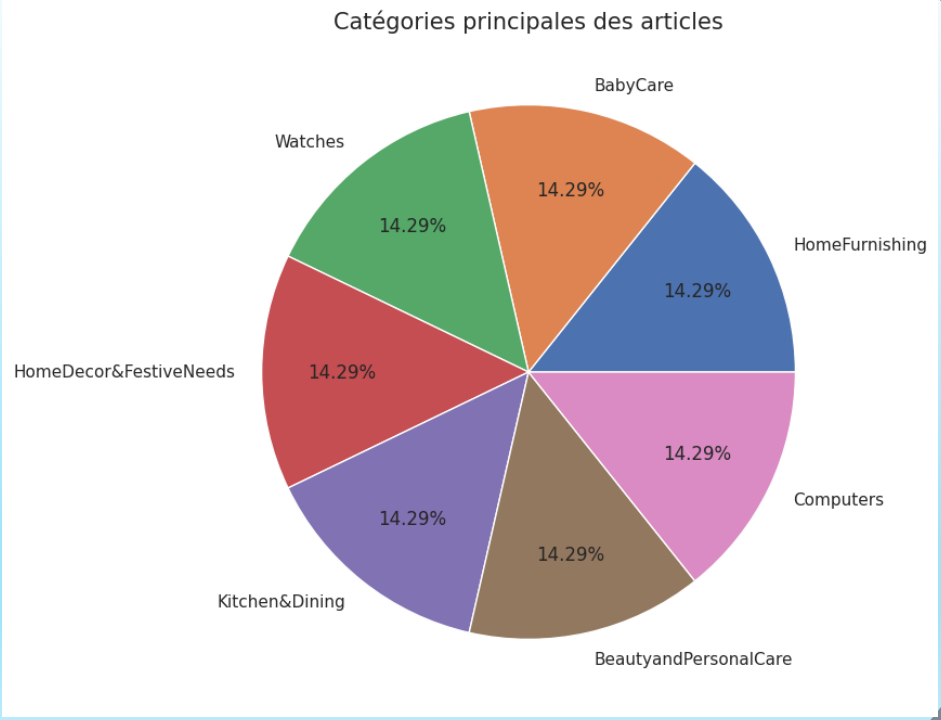
```
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   uniq_id                   1050 non-null   object
 1   crawl_timestamp           1050 non-null   object
 2   product_url               1050 non-null   object
 3   product_name              1050 non-null   object
 4   product_category_tree     1050 non-null   object
 5   pid                       1050 non-null   object
 6   retail_price              1049 non-null   float64
 7   discounted_price          1049 non-null   float64
 8   image                     1050 non-null   object
 9   is_FK_Advantage_product   1050 non-null   bool
 10  description               1050 non-null   object
 11  product_rating            1050 non-null   object
 12  overall_rating            1050 non-null   object
 13  brand                     712 non-null    object
 14  product_specifications    1049 non-null   object
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```

Dossier Images

1050 photos dont le nom est dans la colonne « image »

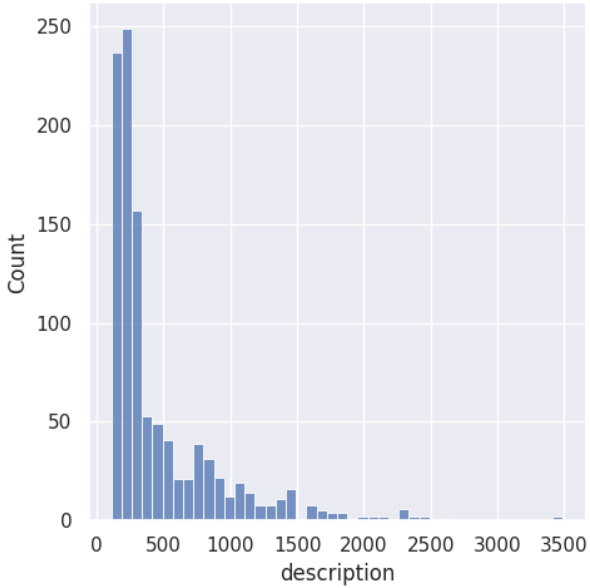
Catégories des produits

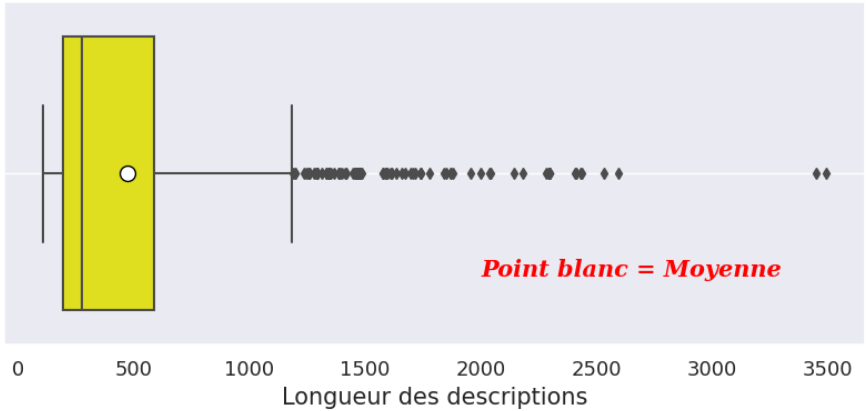
- 7 catégories principales équitablement répartie
- Nombre de niveaux de catégorie maximum : 7
- 57 articles sur 1050 ont 7 niveaux de catégorie



Variable « description »

```
count     1050.000000
mean       473.820952
std        457.910422
min        109.000000
25%        192.000000
50%        278.000000
75%        588.250000
max       3490.000000
Name: description, dtype: float64
```





Exemple :

Catégorie principale : Watches

Description : Maxima 19413PPSN FIBER COLLECTION Digital Watch - For Men - Buy Maxima 19413PPSN FIBER COLLECTION Digital Watch - For Men 19413PPSN Online at Rs.825 in India Only at Flipkart.com. Rectangular Dial, Green Strap, Water Resistant - Great Discounts, Only Genuine Products, 30 Day Replacement Guarantee, Free Shipping. Cash On Delivery!...

2.2 Extractions de features texte
a) Préprocessing

- Suppression mots rares (utilisés une seule fois)
- Minuscule
- Tokenisation : «`RegexpTokenizer(r"\w+")`»
- Stopwords de NLTK
- Mots de plus de 3 lettres
- Caractères alphabétiques
- **PorterStemmer**
- Mots anglais seulement dans dictionnaire NLTK (préalablement passés au Stemmer)
- **Résultat :**

```
Nombre de tokens : 45937, et nombre de tokens uniques : 2166
```

- **Exemple :**

```
Index de l'article : 773 Longueur de la description originale : 129
'Buy Ireeya Abstract Single Coral Blanket Blue at Rs. 529 at Flipkart.com. Only
Genuine Products. Free Shipping. Cash On Delivery!'
Longueur de la description cleanée : 76
'buy abstract singl coral blanket blue genuin product free ship cash deliveri'
```

2.2 Extractions de features texte
b) Text Processing Models

Après nettoyage des descriptions (étape précédente)

Modèles	Définition	Avantage ou inconvénient
Bag of words	- Chaque document → vecteur de la taille du vocabulaire du corpus - Comptabilise la fréquence d'apparition des tokens trouvés dans le corpus - Matrice composée de l'ensemble de ces N documents qui forment le corpus.	- Matrice creuse
TF-IDF	- Même principe de base - Pondère cette fréquence par un indicateur de similarité	- Matrice creuse - Les mots rares ont plus de poids donc de sens
Word2Vec	- Word embeddings - perceptrons linéaires simples avec une seule couche cachée - Représentation du mot dans un espace qui le positionne en fonction des mots adjacents (distance statistique) - Corpus compressé → dictionnaire de vecteurs denses	- Dimensions inférieures (20-100) - Matrice dense - Corpus beaucoup plus grand

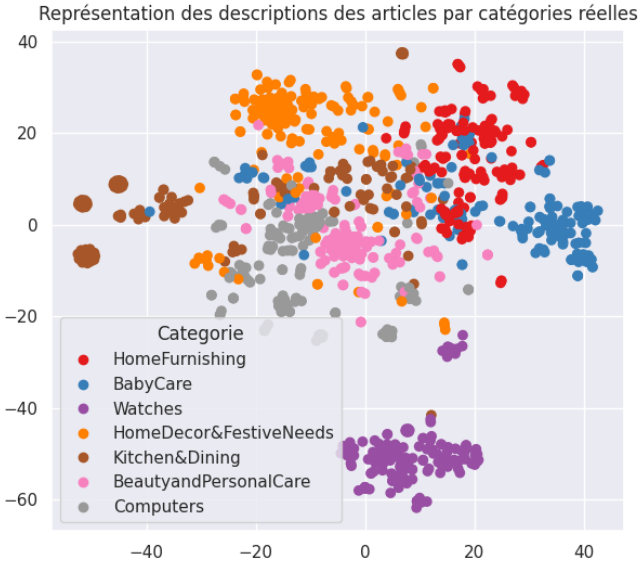
2.2 Extractions de features texte
b) Text Processing Models - suite

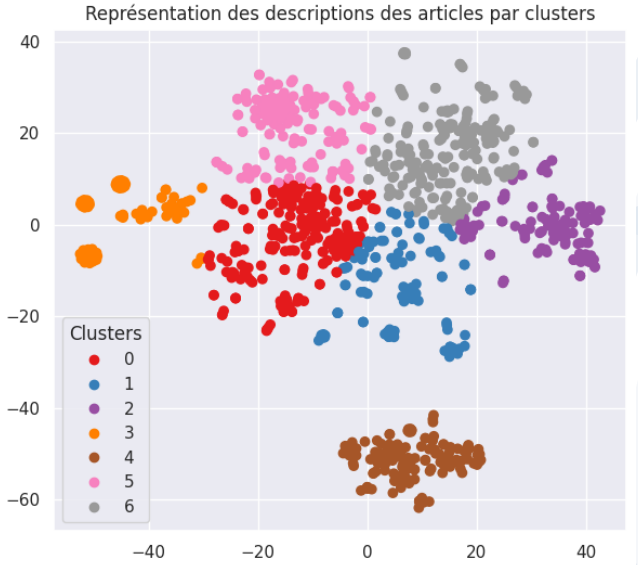
Après nettoyage des descriptions

Modèles	Définition	Avantage ou inconvénient
BERT (Bidirectional Encoder Representations from Transformers) model_type = 'bert-base-uncased'	- Sentence embedding - **Transformer → réseaux de neurones** spécifique - Petit nombre constant d'étapes - Relations entre les mots de la phrase pour prédire un mot qu'il masque - Modèle pré-entraîné (non-supervisé) à choisir selon domaine - Transfert Learning	- Meilleure compréhension - 512 tokens maximum à la fois en entrée
USE (Universal Sentence Encoder)	- Sentence embedding - Entraîné sur Wikipédia, contenu web - Calcule la représentation vectorielle du texte - Détermine la similarité cosinus entre la représentation vectorielle respective de deux textes - Présente à l'utilisateur les textes par ordre de similarité décroissante	- De bons résultats pour classifier

2.2 Extractions de features texte
b) Text Processing Models – Résultats et meilleur modèle

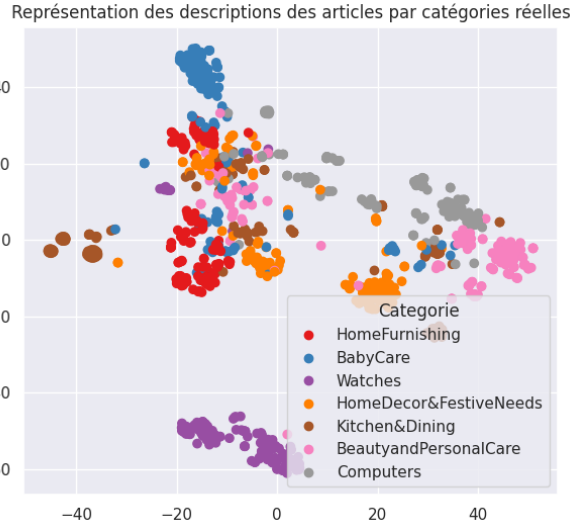
	Modèles	ARI	Time
0	Bag of word	0.4214	10.0
1	TF-IDF	0.4188	7.0
2	Word2Vec	0.3384	7.0
3	Bert	0.3233	6.0
4	USE	0.4352	7.0

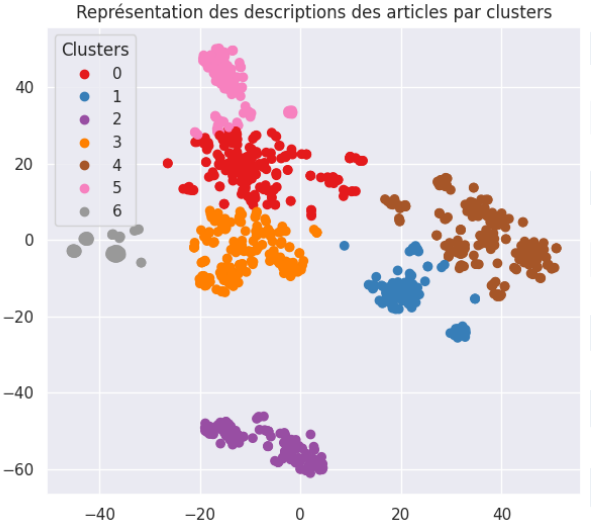


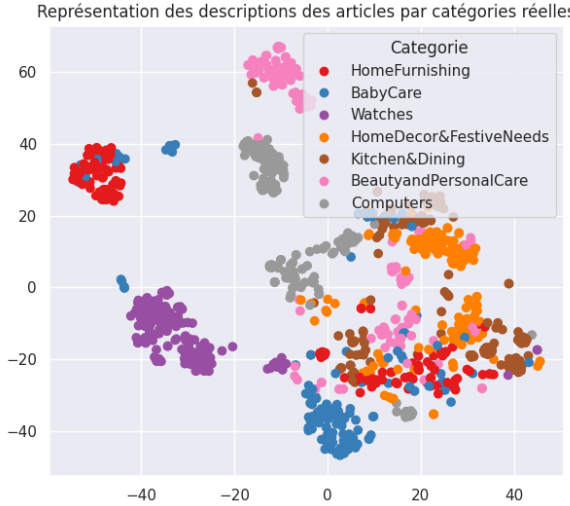


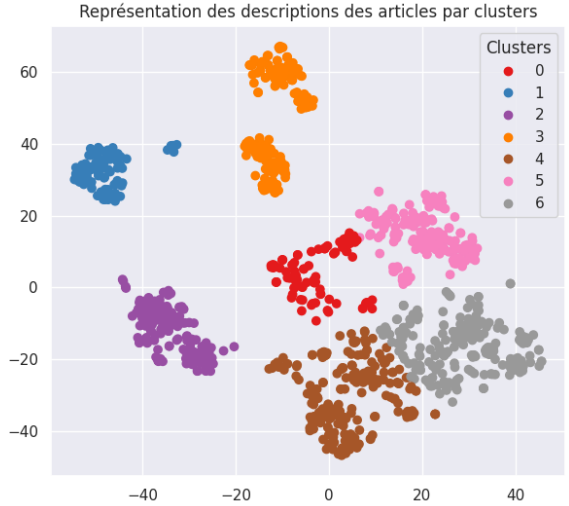
→ CountVectorizer
ARI = 0,4214

Word2Vec →
ARI = 0,3384

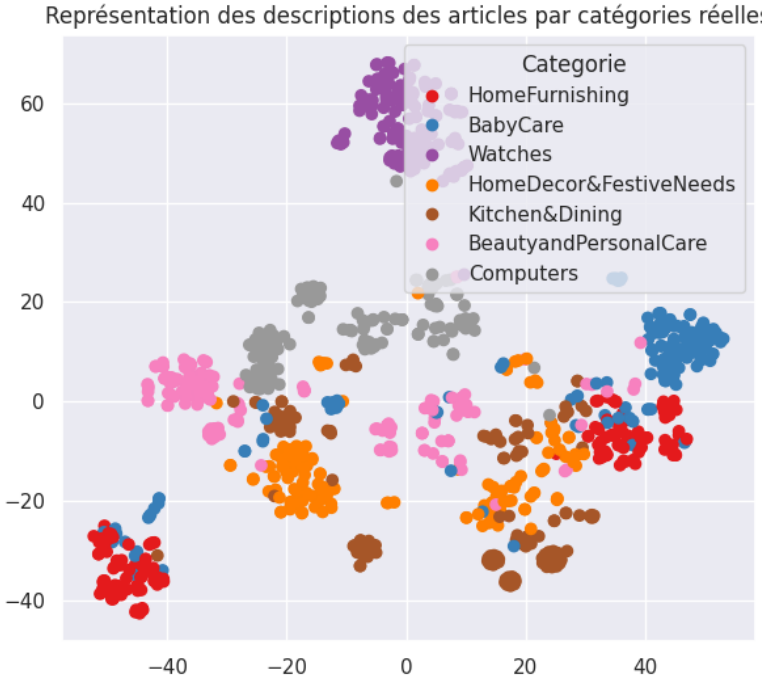


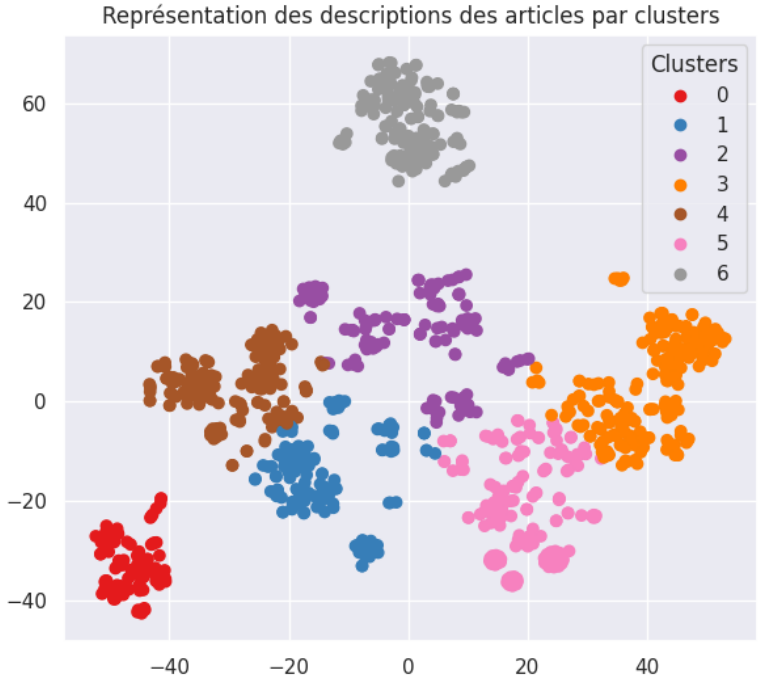






→ BERT
ARI = 0,3233

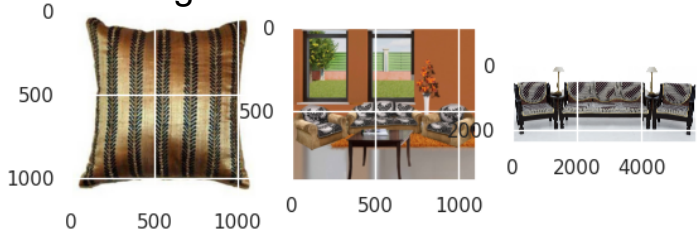




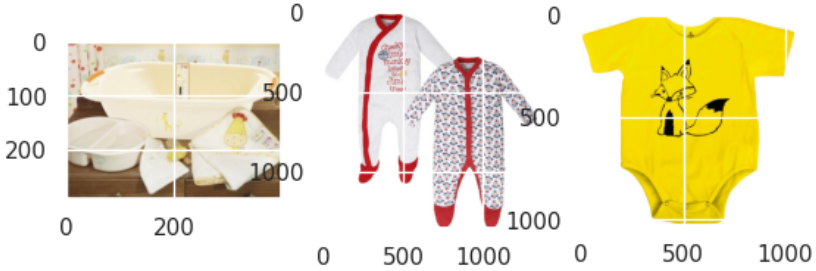
USE
ARI = 0,4352

2.3 Extractions de features images
a) Exemples d'image par label

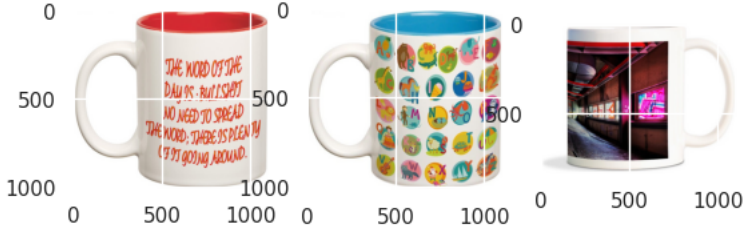
HomeFurnishing :



BabyCare :



Kitchen&Dining :

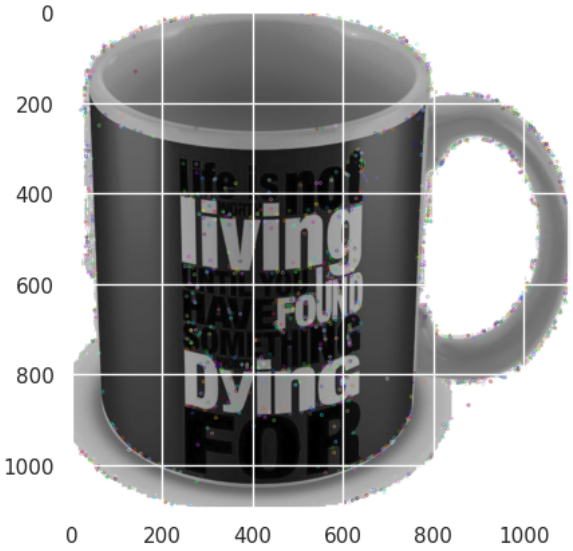


Watches :

2.3 Extractions de features images
b) Préprocessing

- Premier traitement du contraste : openCV,
- Retraitement d'images : passage en gris, filtrage du bruit, égalisation, floutage

Exemple :



L'image contient 1287 descripteurs
Chaque descripteur est un vecteur de longueur 128

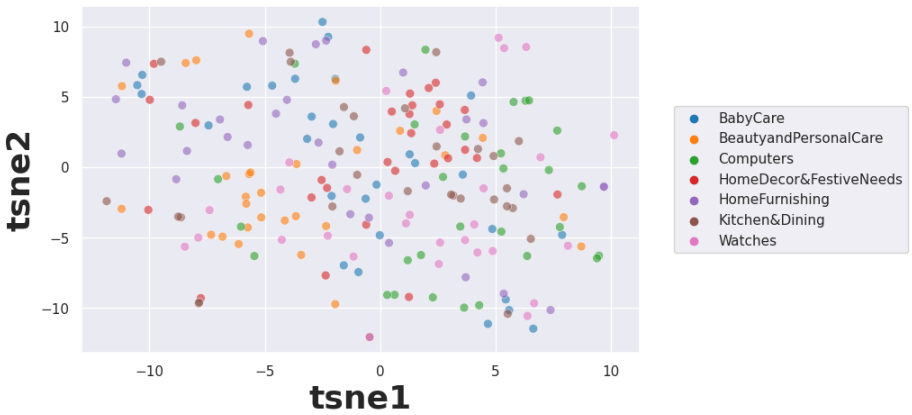
2.3 Extractions de features images
c) Image Processing Models

Après nettoyage des images (étape précédente)

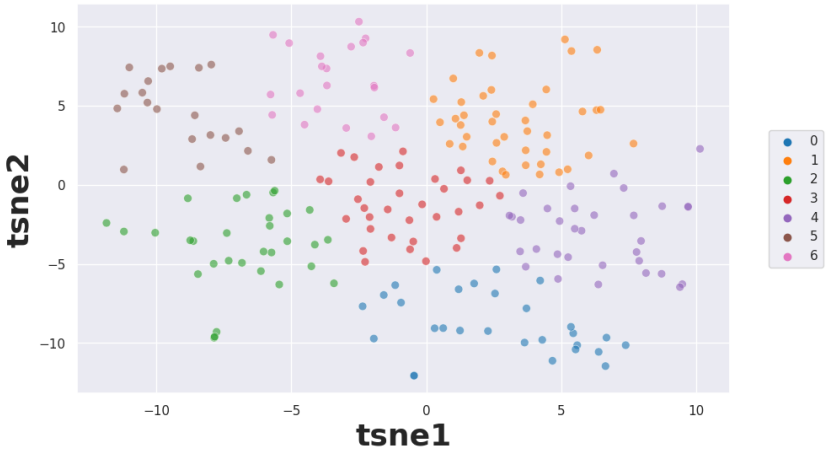
SIFT

- Utilisation d'un échantillon de 210 images réparties par groupes de 30 sur chaque label
- Nombre de descripteurs : `103421`
- Nombre de clusters estimés : `322` pour MiniBatchKMeans
- Features d'une image = Histogramme d'une image = Comptage pour une image du nombre de descripteurs par cluster
- Temps de traitement SIFT descriptor : `219.35 secondes`
- Temps de traitement MiniBatchKmeans : `1.94 secondes`
- Temps de création des histogrammes : `0.24 secondes`
- Réduction de dimension PCA à 99% : `(210, 322)` → `(210, 146)`
- Temps de réduction de dimension T-SNE pour affichage en 2D : `1.01 secondes`

TSNE selon les vraies classes



TSNE selon les clusters

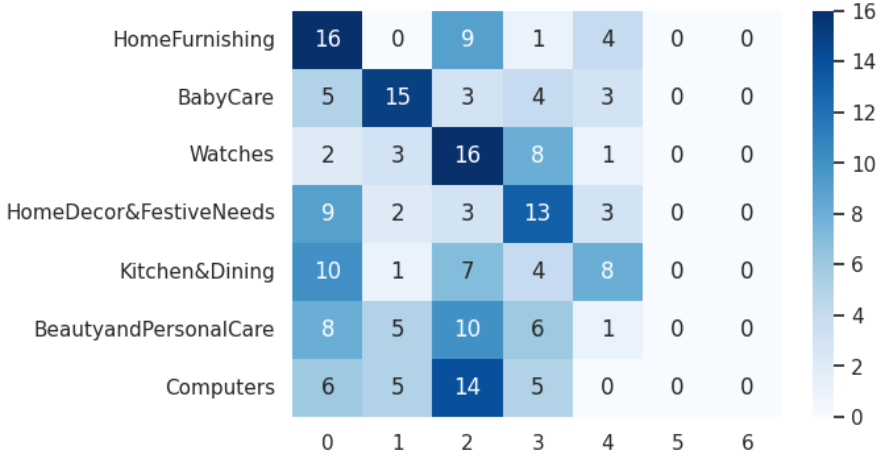


ARI = 0.0574

Analyse visuelle négative :

- L'analyse graphique montre visuellement qu'il n'est pas réalisable de séparer automatiquement les images selon leurs vraies classes avec SIFT
- Autre approche à trouver

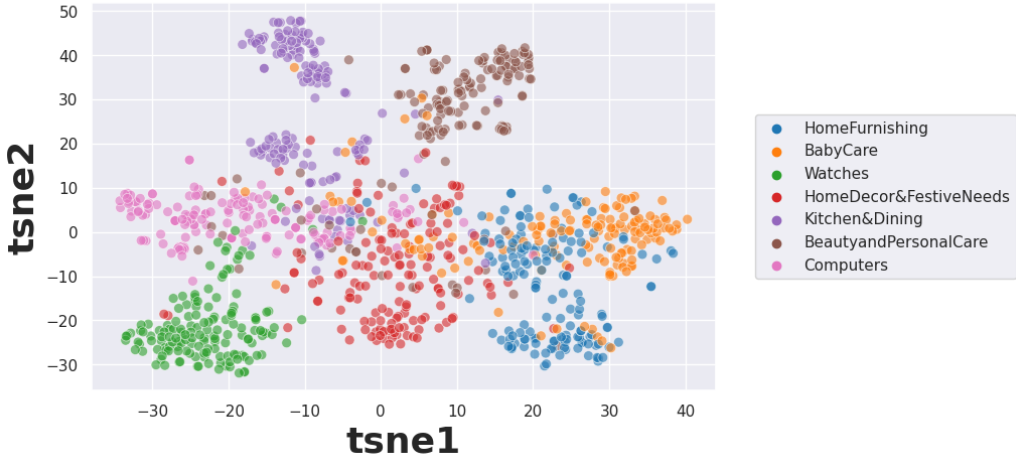
Analyse par classe :



CNN – Transfert Learning

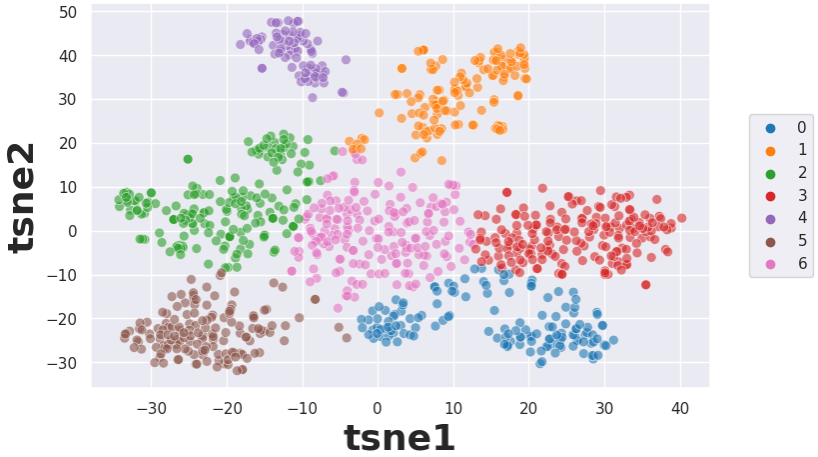
- Modèle VGG16 – Temps de création du modèle : `17.53 secondes`
- Création des features images : `(1050, 4096)` – Temps : `452.92 secondes`
- Réduction de dimension PCA à 99% : `(1050, 4096)` → `(1050, 803)`
- Réduction de dimension T-SNE – Temps : `7.81 secondes`
- K-Means 7 clusters – Temps : `0.96 secondes`
- Affichage en 2D :

TSNE selon les vraies classes



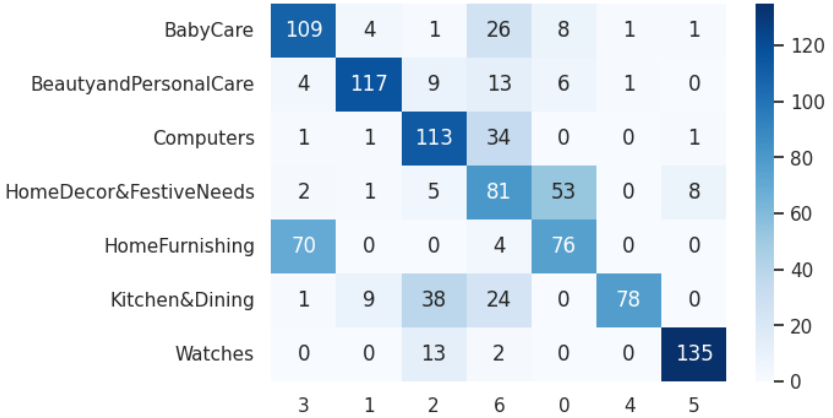
Analyse visuelle positive

TSNE selon les clusters



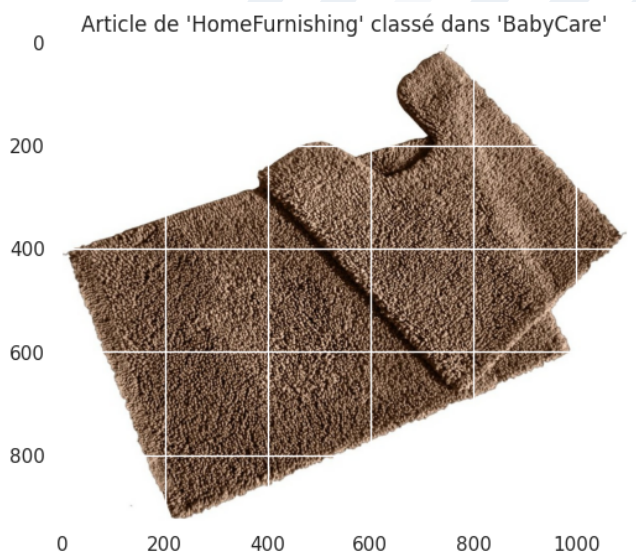
`ARI : 0.4572`

Heatmap de correspondance entre les labels réels et les numéros de cluster avec CNN



<u>Commentaire</u> :

- Moins bien prédites sont : "Kitchen&Dining", "BeautyandPersonalCare", "HomeDecor&FestiveNeeds"
- Les articles de "HomeFurnishing" sont particulièrement difficiles à identifier, ici 70 individus (presque autant que le nombre d'articles bien catégorisés) sont classés dans le groupe "BabyCare"



3. Classification supervisée

Etapes

- Partage du dataframe en deux : train/test (0,7-0,3)

train :

	image_path	label_name	label
0	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
1	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
2	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
3	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
4	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
...	...	...	...
730	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
731	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
732	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
733	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
734	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6

735 rows × 3 columns

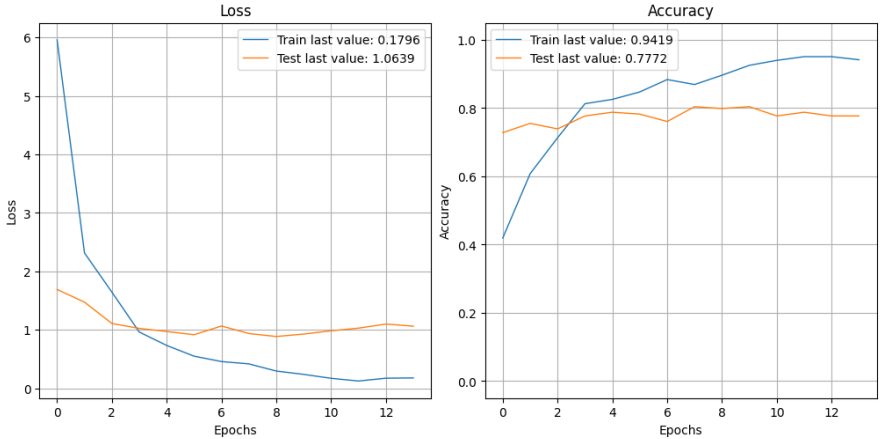
test :

	image_path	label_name	label
0	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	HomeFurnishing	4
1	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
2	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
3	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
4	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Watches	6
...	...	...	...
310	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	HomeFurnishing	4
311	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	Computers	2
312	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
313	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0
314	/content/drive/MyDrive/Colab Notebooks/OC/Proj...	BabyCare	0

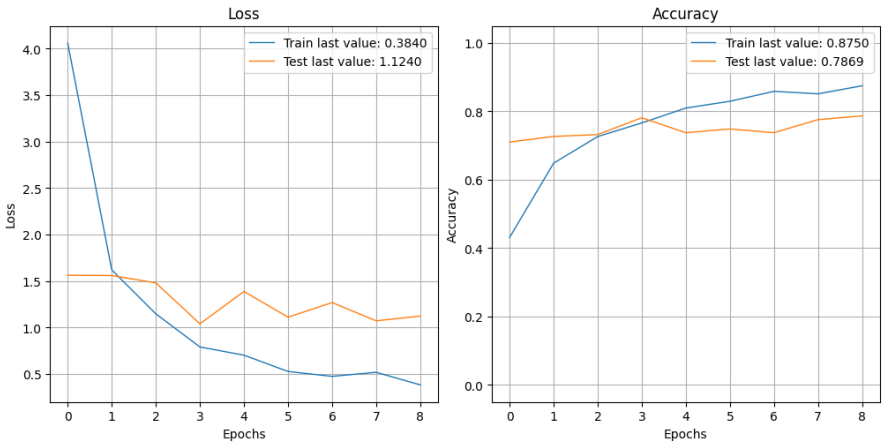
315 rows × 3 columns

- Dans Train : partie Eval à 0,25
- Enregistrement des images train/test dans 2 répertoires et par label
- Classification supervisée des images : 4 approches :
  - ❖ Approche simple par préparation initiale de l'ensemble des images avant classification supervisée (VGG16 – Imagenet)
  - ❖ Approche par data generator : data augmentation, les images sont directement récupérées à la volée dans le répertoire des images
  - ❖ Approche récente Tensorflow.org par DataSet, sans data augmentation
  - ❖ Approche par DataSet, avec data augmentation intégrée au modèle : layer en début de modèle

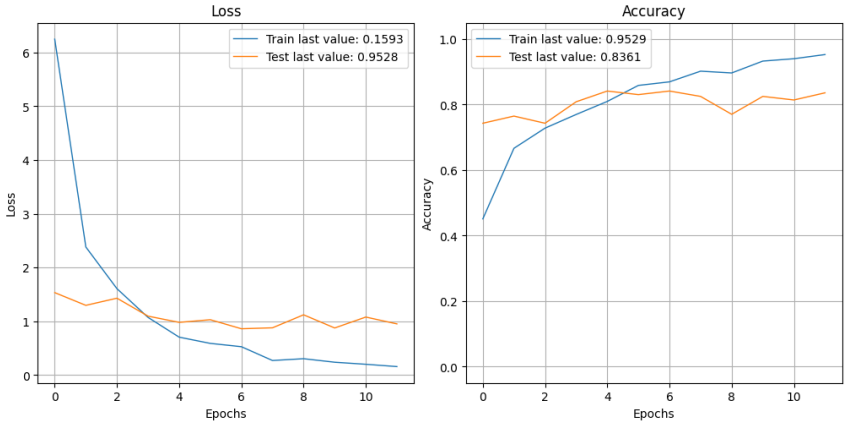
Approche simple VGG16



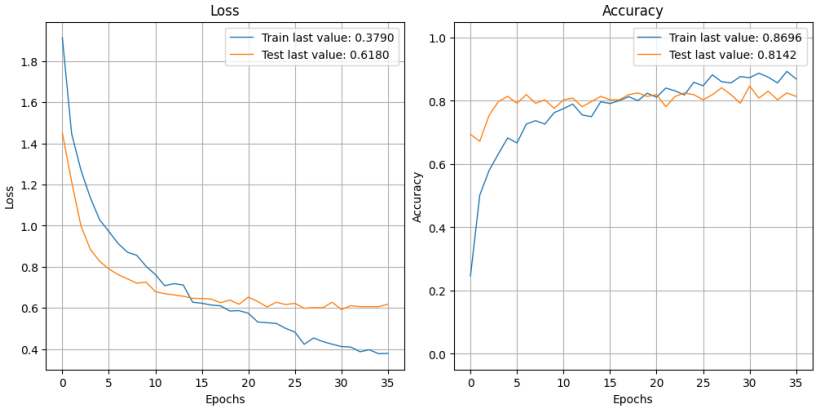
Approche ImagedataGenerator



Approche nouvelle par dataset sans data augmentation



Approche par dataset avec data augmentation intégrée



Résultats

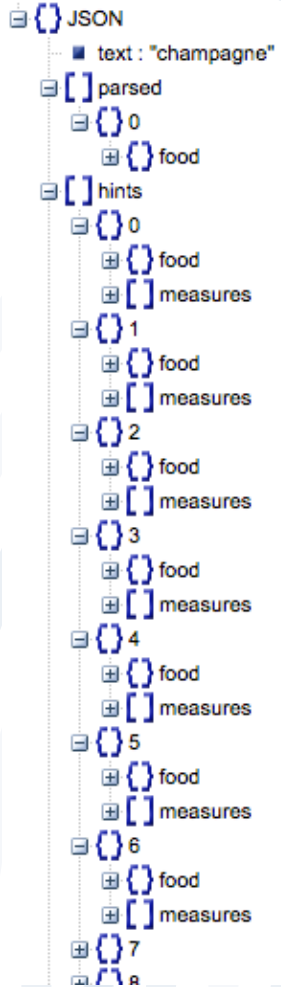
Scores epoch optimal pour chaque modèle testé :

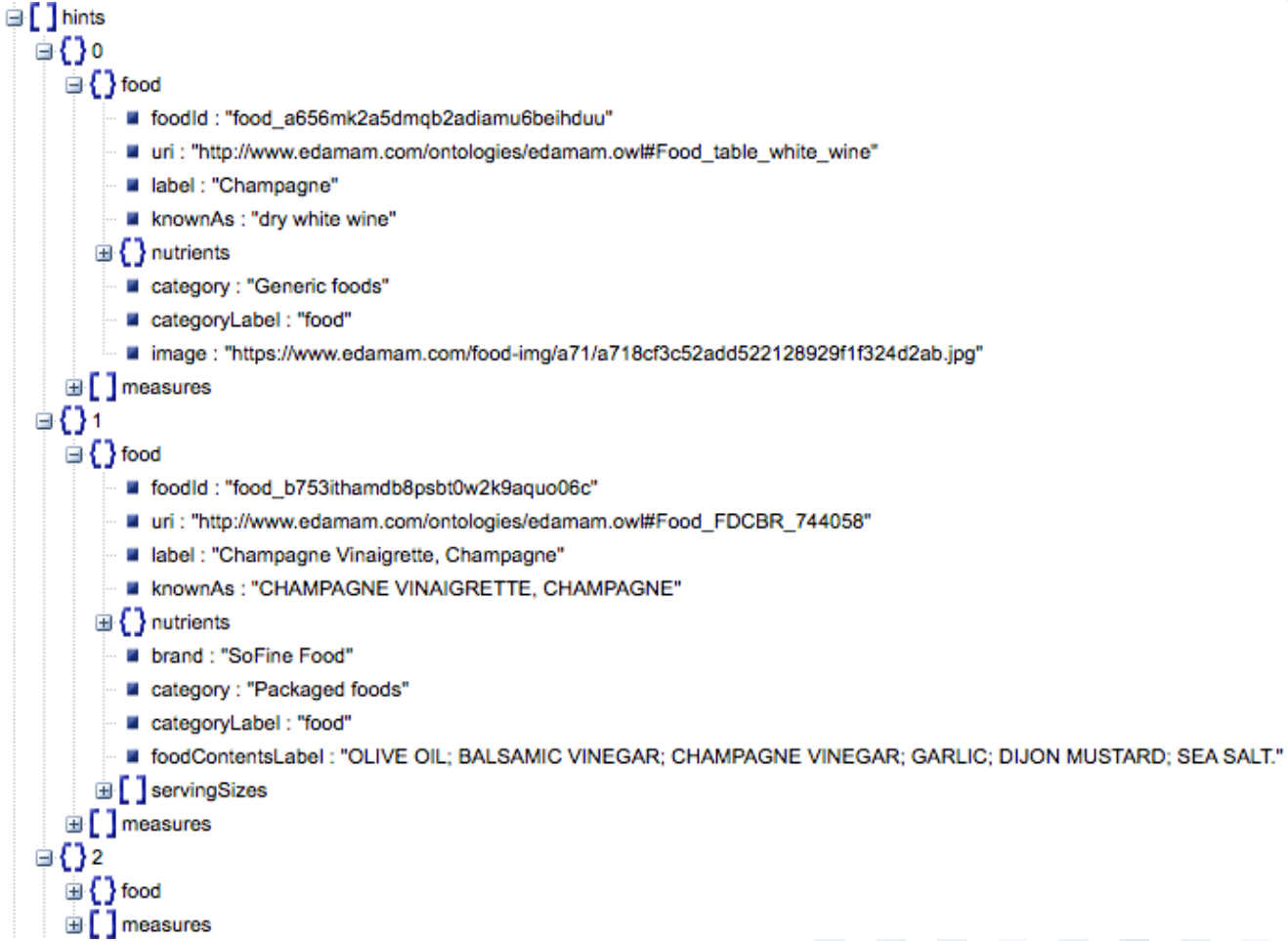
	Modèle	Validation Accuracy	Test Accuracy	Temps de création du modèle en secondes	Temps d'entraînement du modèle en secondes
0	VGG16	0.798913	0.822222	0.462533	17.016821
1	IDG	0.754098	0.800000	0.445700	186.266010
2	SansDA	0.841530	0.831746	0.447957	49.726068
3	DAIntégrée	0.846995	0.834921	0.767045	148.539452

4. Test de l’API

Etapes

- Création d'un compte sur rapidapi
- Récupération de la clé
- Requête pour obtenir les données via l'API avec filtre : « ingr » = « champagne » et champs demandés
- Arborescence du fichier retourné grâce à jsonviewer.stack.hu :




hints
0
food
foodId : "food_a656mk2a5dmqb2adiamu6beihduu"
uri : "http://www.edamam.com/ontologies/edamam.owl#Food_table_white_wine"
label : "Champagne"
knownAs : "dry white wine"
nutrients
category : "Generic foods"
categoryLabel : "food"
image : "https://www.edamam.com/food-img/a71/a718cf3c52add522128929f1f324d2ab.jpg"
measures
1
food
foodId : "food_b753ithamdb8psbt0w2k9aquo06c"
uri : "http://www.edamam.com/ontologies/edamam.owl#Food_FDCBR_744058"
label : "Champagne Vinaigrette, Champagne"
knownAs : "CHAMPAGNE VINAIGRETTE, CHAMPAGNE"
nutrients
brand : "SoFine Food"
category : "Packaged foods"
categoryLabel : "food"
foodContentsLabel : "OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR; GARLIC; DIJON MUSTARD; SEA SALT."
servingSizes
measures
2
food
measures

On obtient le dataframe suivant avec les champs demandés :

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

Que l'on sauvegarde sous forme csv : `Regaud_Agnès_4_fichier_csv_extraction_api_092023.csv`

5. Conclusion et RGPD

Norme RGPD sur ce projet :

Les 5 grands principes des règles de protection des données personnelles sont les suivants :

- **Le principe de finalité** : le responsable d'un fichier ne peut enregistrer et utiliser des informations sur des personnes physiques que dans un but bien précis, légal et légitime ;
- **Le principe de proportionnalité et de pertinence** : les informations enregistrées doivent être pertinentes et strictement nécessaires au regard de la finalité du fichier ;
- **Le principe d'une durée de conservation limitée** : il n'est pas possible de conserver des informations sur des personnes physiques dans un fichier pour une durée indéfinie. Une durée de conservation précise doit être fixée, en fonction du type d'information enregistrée et de la finalité du fichier ;
- **Le principe de sécurité et de confidentialité** : le responsable du fichier doit garantir la sécurité des informations qu'il détient. Il doit en particulier veiller à ce que seules les personnes autorisées aient accès à ces informations ;
- **Les droits des personnes.**

Source CNIL

Sur de projet, les images et descriptions ne concernent pas des données personnelles. Il n'y a aucune contrainte de propriété intellectuelle sur les données et les images. De plus, je n'ai utilisé que les données dont j'avais besoin.

SOURCE du projet :

https://rapidapi.com/edamam/api/edamam-food-and-grocery-database