Classifiez automatiquement des biens de consommation

Place de marché - Projet 6

Manuel Martin - 02/2023

Classifiez automatiquement des biens de consommation

Sommaire

- Rappel de la problématique et présentation du jeu de données
- Explication des prétraitements, des extractions de features et des résultats de l’étude de faisabilité
- Résultats de la classification supervisée
- Présentation du test de l’API
- Conclusion

Rappel de la problématique et présentation du jeu de données

Rappel de la problématique et présentation du jeu de données

Problématique

- Lancement d’une marketplace d’e-commerce
- Catégorisation des articles par les vendeurs, manuellement
- Besoin d’automatiser la catégorisation
  - Moteur de classification des articles

Rappel de la problématique et présentation du jeu de données

Problématique

- Etude de faisabilité
  - Analyse des descriptions textuelles et des images des produits
  - Prétraitement, extraction de features, réduction en 2 dimensions, analyse graphique, mesure de similarité
  - Différentes approches à explorer
- Classification supervisée des images
- Collecte de produit à base de champagne via API

Rappel de la problématique et présentation du jeu de données

Présentation du jeu de données

- 1 fichier .csv : « flipkart_com-ecommerce_sample_1050 »
  - 1050 individus
  - 15 features
    - 11 features textuelles
    - 1 feature temporelles (au format texte)
    - 2 features numériques
    - 1 feature booléenne

Rappel de la problématique et présentation du jeu de données

Présentation du jeu de données

- 1 dossier Images
  - 1050 images couleur au format jpg

Explication des prétraitements, des extractions de features et des résultats de l’étude de faisabilité

Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Prétraitements

- Feature « product_category_tree »
  - Target pour évaluer le clustering
- Features « product_name » et « description »
  - Features à transformer pour réaliser le clustering
- Feature « image »
  - Servira pour identifier les fichiers image pour le clustering

Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Prétraitements

- Nettoyage des documents du corpus
  - Tokenization des mots dans chaque document
  - Nettoyage et suppression de certains mots (stop words, caractères numériques et symboles)
  - Remplacement des majuscules
  - Lematisation des mots (réduction à la racine du mot)

Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

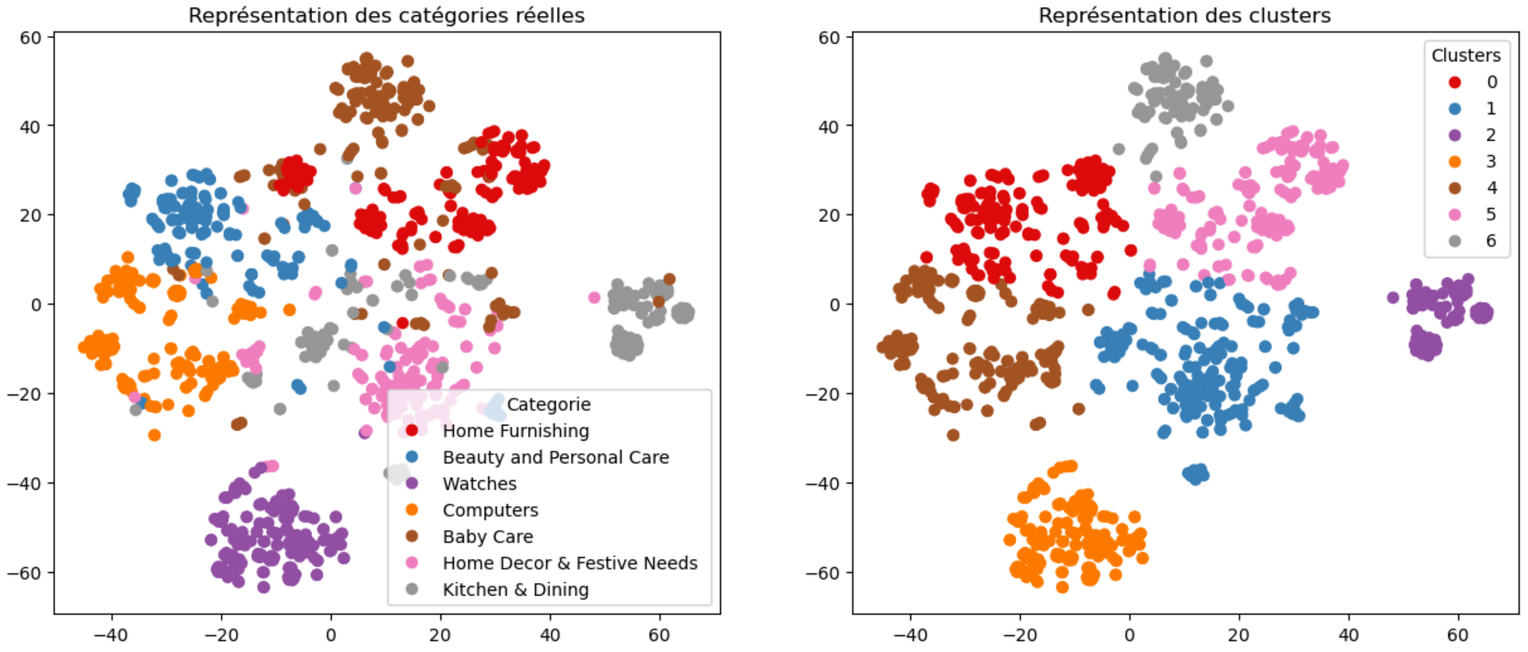
Extractions de features textuelles

- Bag of Words
  - Comptage simple
  - TF - IDF
- Différentes combinaison
  - description, product_name, description + product_name
  - 7 catégories principales
  - 63 catégories secondaires
- T-SNE + K-Means

Meilleurs résultats :
- TF-IDF
- description + product_name
- 7 catégories principales

Score ARI = 0,60

Analyse via t-SNE

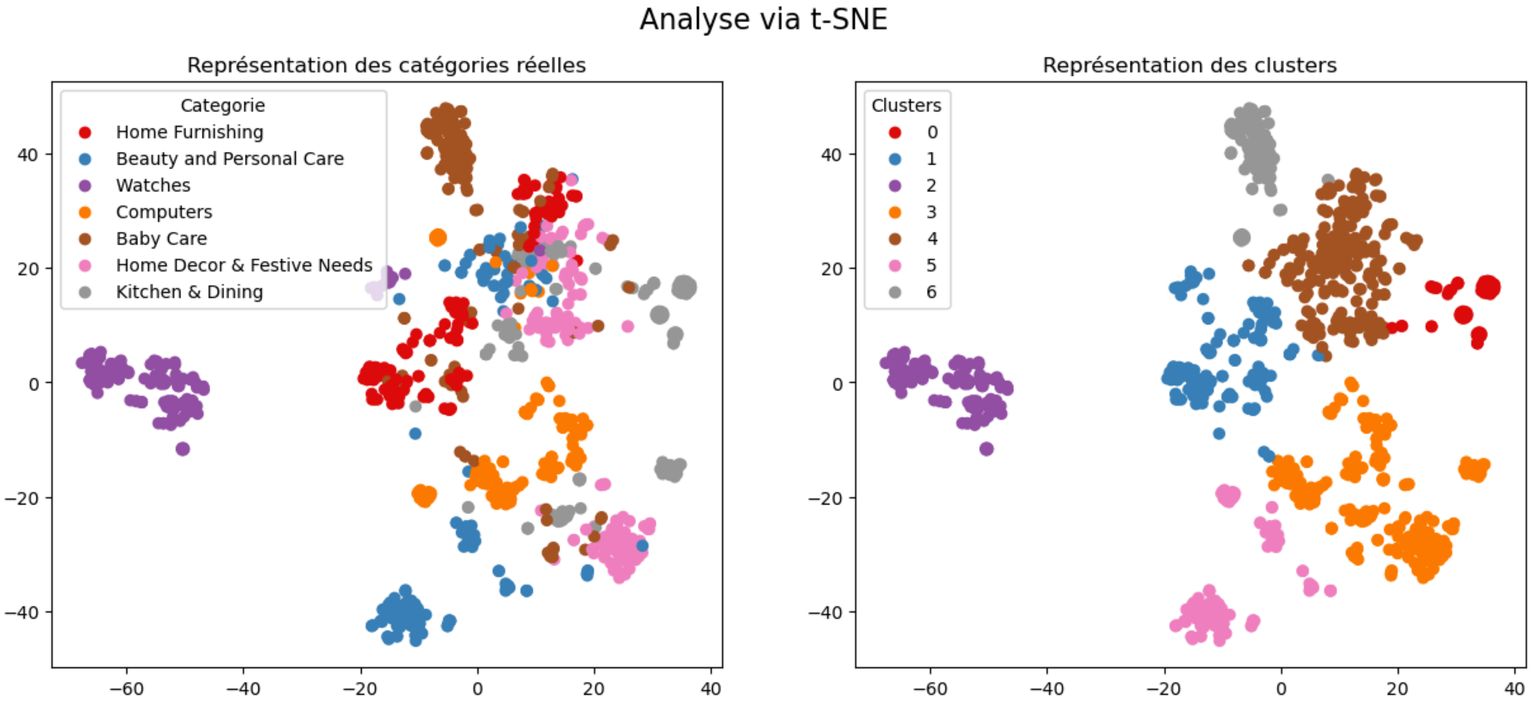


Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Extractions de features textuelles

- Plongement de mots (Word2Vec)
  - Librairie gensim
  - Utilisation de tous les mots du corpus pour le vocabulaire : 4457 mots
  - Réduction à 300 dimensions
  - Modèle TensorFlow avec une couche d'embedding
- description + product_name
- 7 catégories principales

Score ARI = 0,37

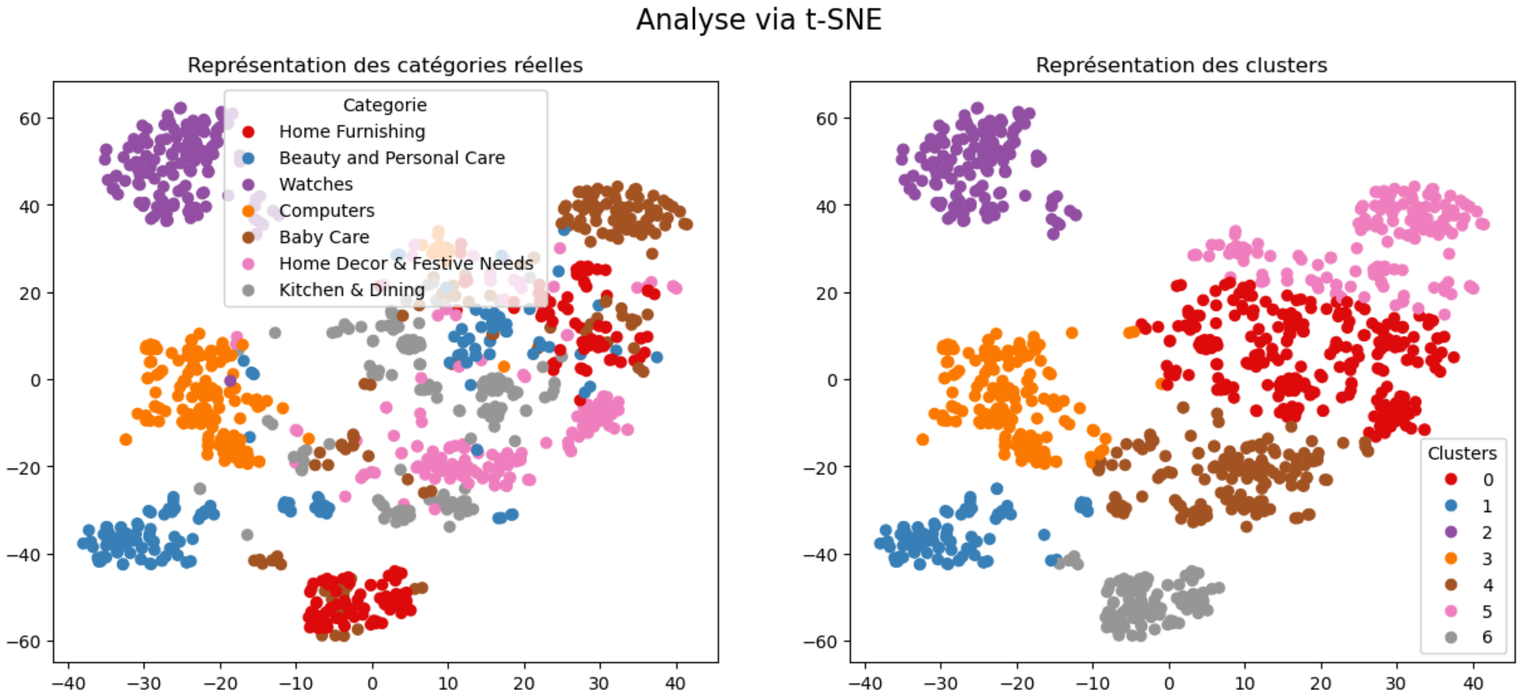


Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Extractions de features textuelles

- Réseau de neurones BERT
  - Utilisation de HuggingFace
  - Modèle bert-base-uncased pré-entrainé
- description + product_name
- 7 catégories principales

Score ARI = 0,44

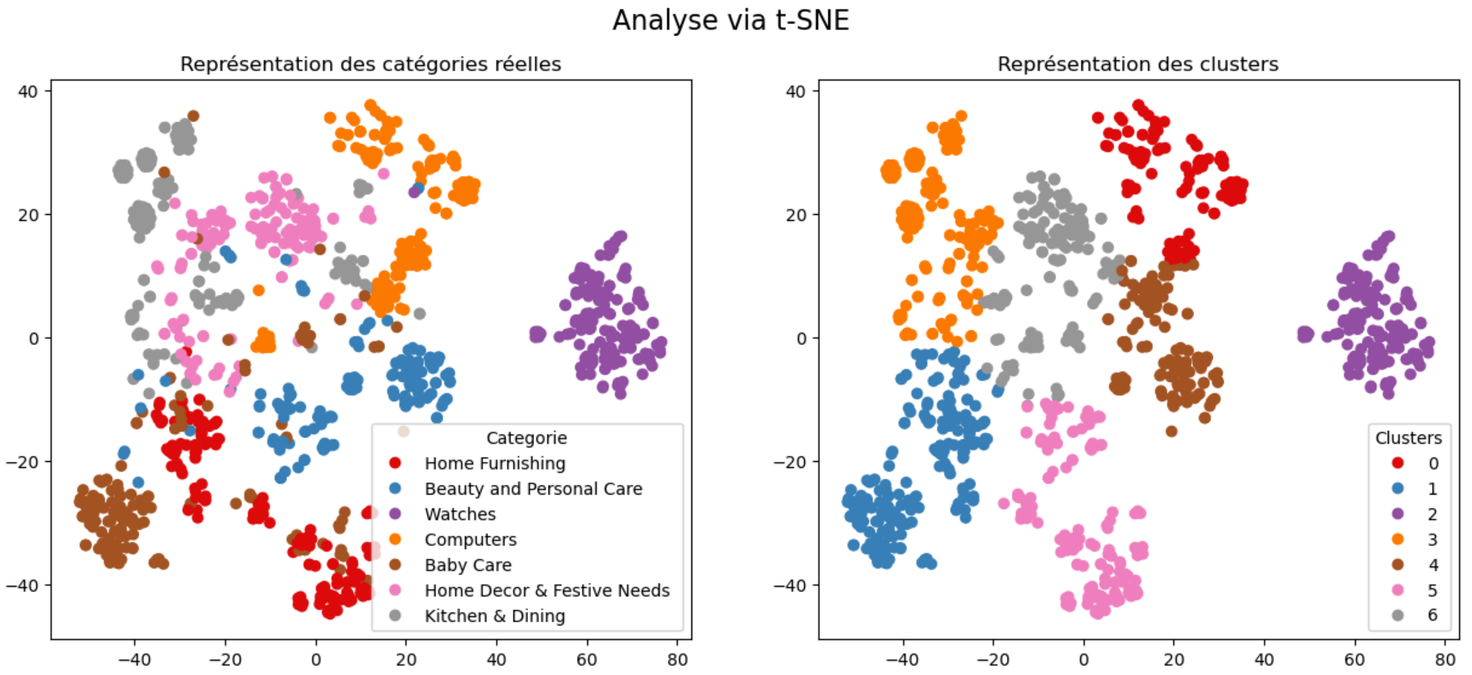


Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Extractions de features textuelles

- Modèle USE
  - Encodage de phrases pour la classification de texte
  - Modèle pré-entrainé
  - Tensorflow-hub
- description + product_name
- 7 catégories principales

Score ARI = 0,46



Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

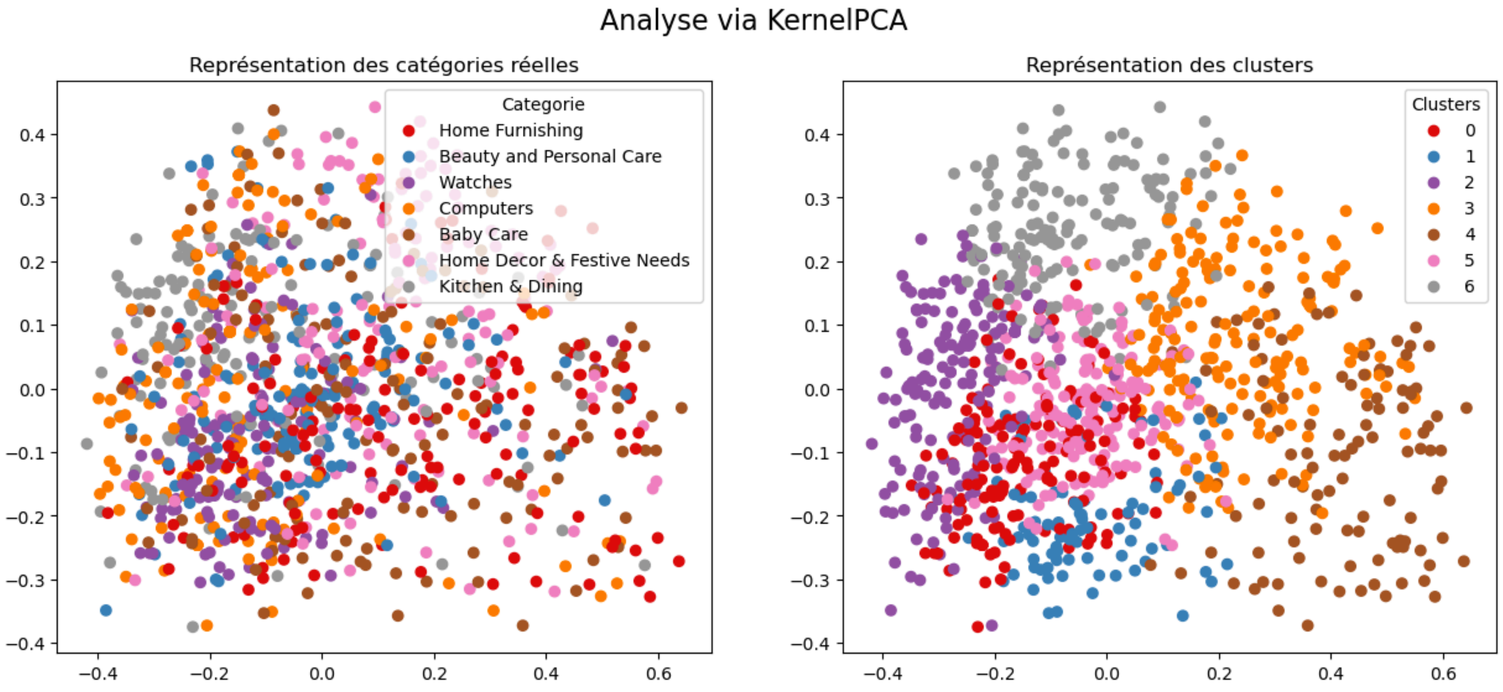
Extractions de features visuelles

- Algorithme SIFT
  - Prétraitement des images (passage en gris et equalisation)
  - Création d'une liste de descripteurs
  - Création de clusters de descripteurs
  - Création des features des images
- Utilisation des 1050 images
- 7 catégories principales
- Utilisation de T-SNE et Kernel PCA (noyau Cosine)

Meilleurs résultats :

- Kernel PCA

Score ARI = 0,1



Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

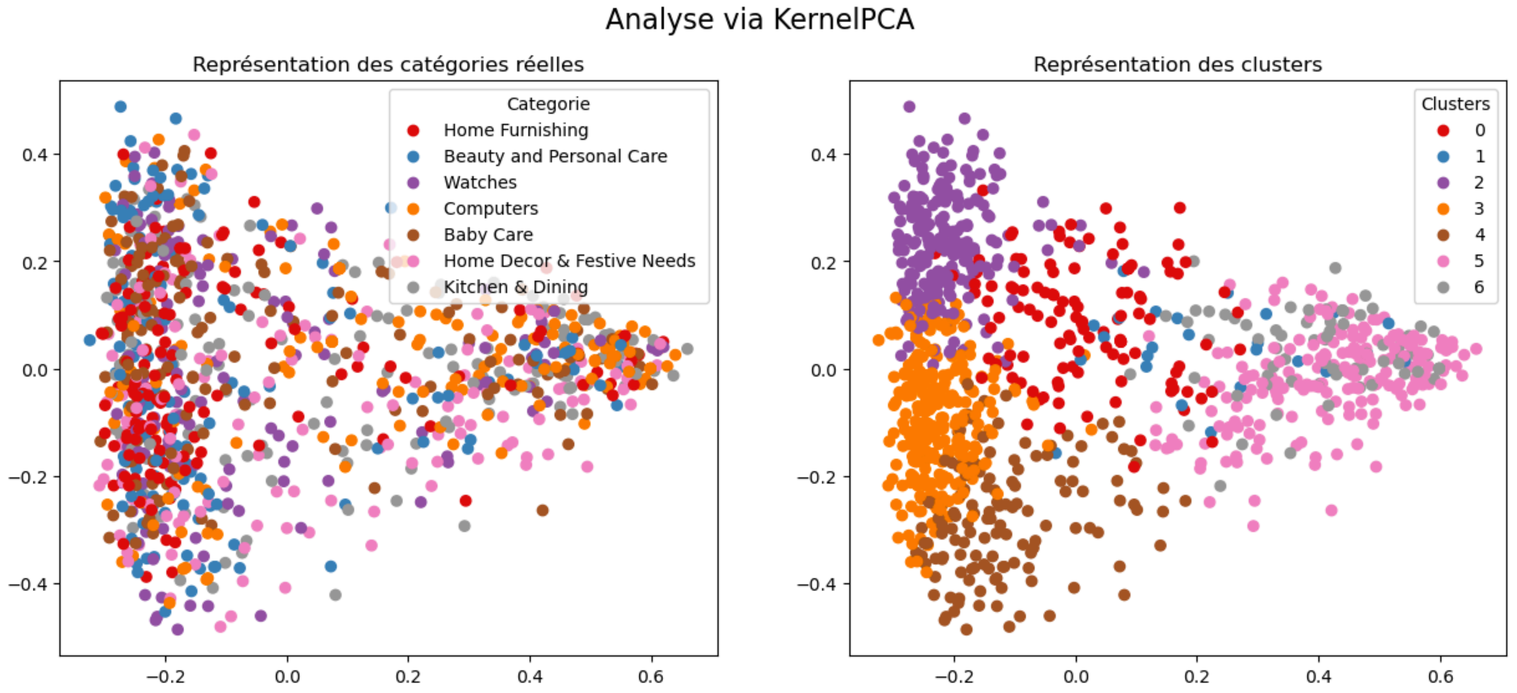
Extractions de features visuelles

- Algorithme ORB
  - Prétraitement des images (passage en gris et equalisation)
  - Création d'une liste de descripteurs
  - Création de clusters de descripteurs
  - Création des features des images
- Utilisation des 1050 images
- 7 catégories principales
- Utilisation de T-SNE et Kernel PCA (noyau Cosine)

Meilleurs résultats :

- Kernel PCA

Score ARI = 0,03



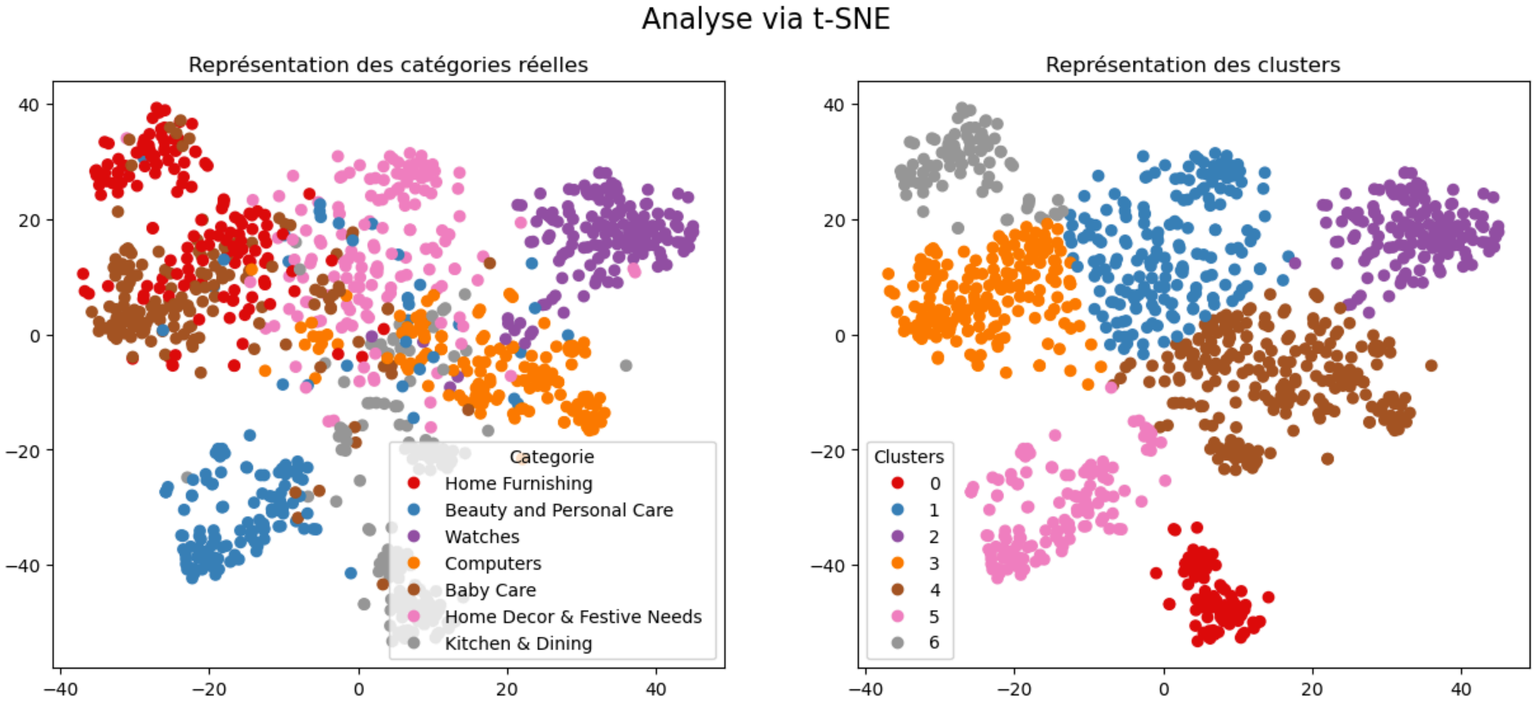
Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Extractions de features visuelles

- CNN Transfer Learning
  - Modèle VGG16 pré-entrainé
  - Redimensionnement des images en 224 x 224
  - Normalisation des valeurs de chaque pixel
- 2 configurations
  - Sans couche fully connected
  - Avec 2 couches fully connected
- Utilisation des 1050 images
- 7 catégories principales

Meilleurs résultats :
- Avec 2 couches fully connected

Score ARI = 0,52



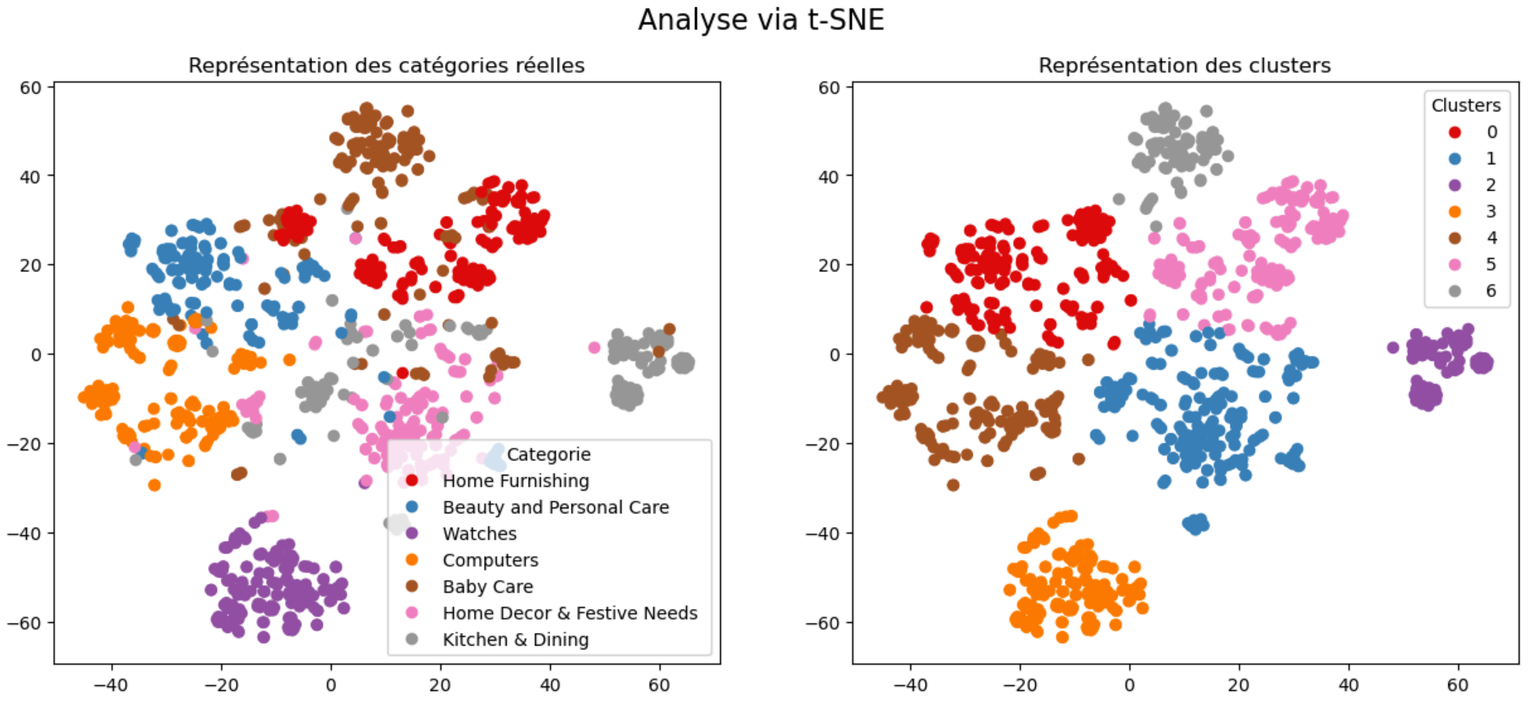
Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Résultats de l'étude de faisabilité

- Meilleur résultat pour l'analyse textuelle
  - Bag of word avec TF-IDF

Meilleurs résultats :
- TF-IDF
- description + product_name
- 7 catégories principales

Score ARI = 0,60



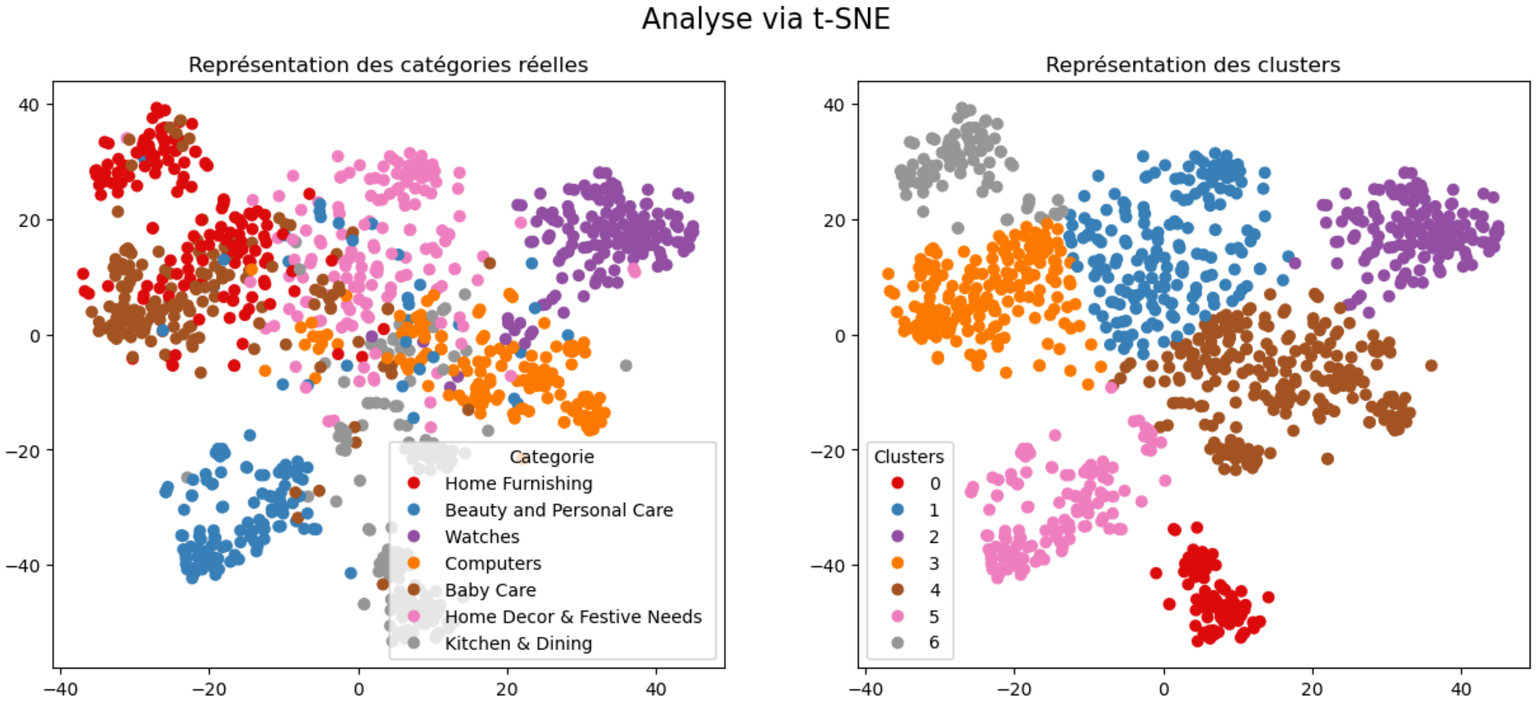
Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Résultats de l'étude de faisabilité

- Meilleur résultat pour l'analyse visuelle
  - VGG16 avec 2 couches fully connected

Meilleurs résultats :
- Avec 2 couches fully connected

Score ARI = 0,52

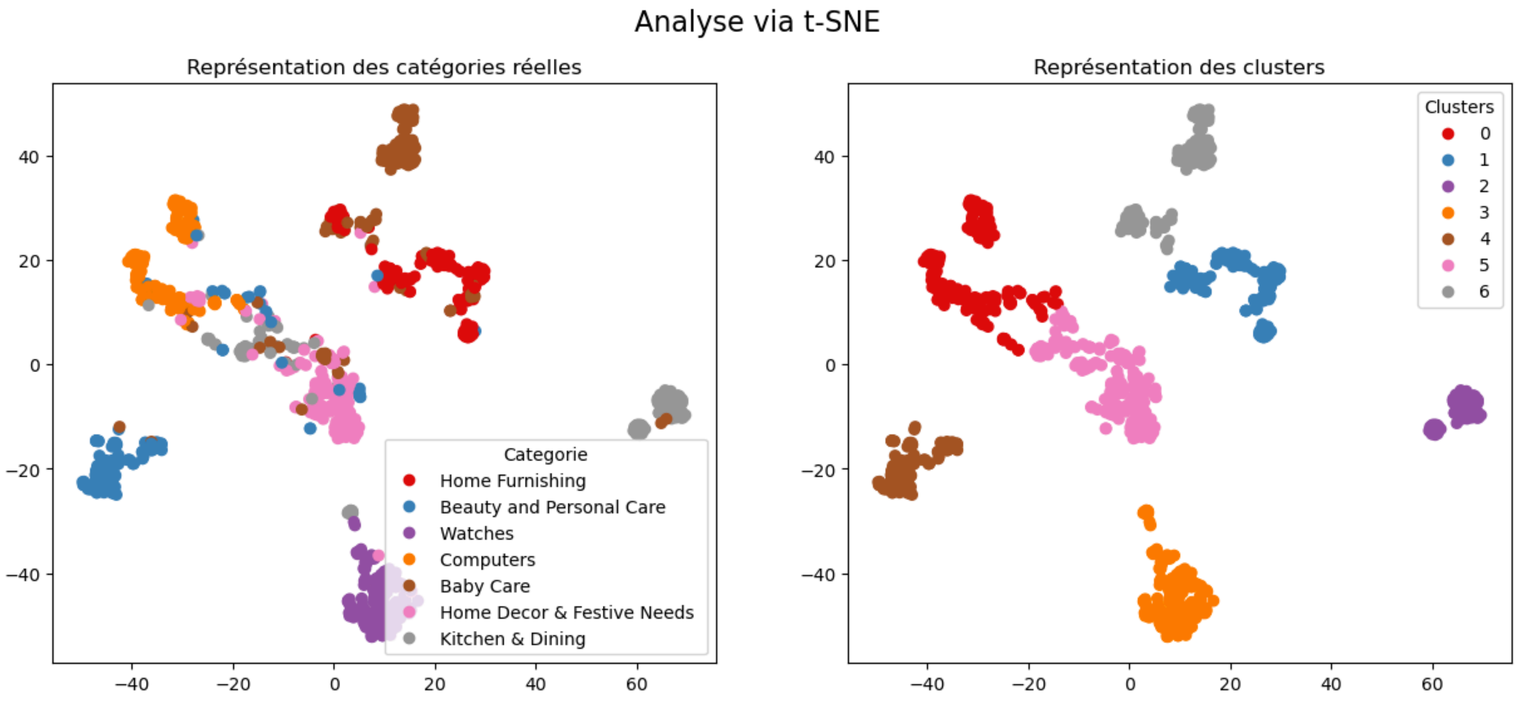


Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité

Résultats de l'étude de faisabilité

- Regroupement des 2 clusterings
  - Bag of Words
  - VGG16
- T-SNE et K-Means des 2 précédents T-SNE et K-Means

Score ARI = 0,63



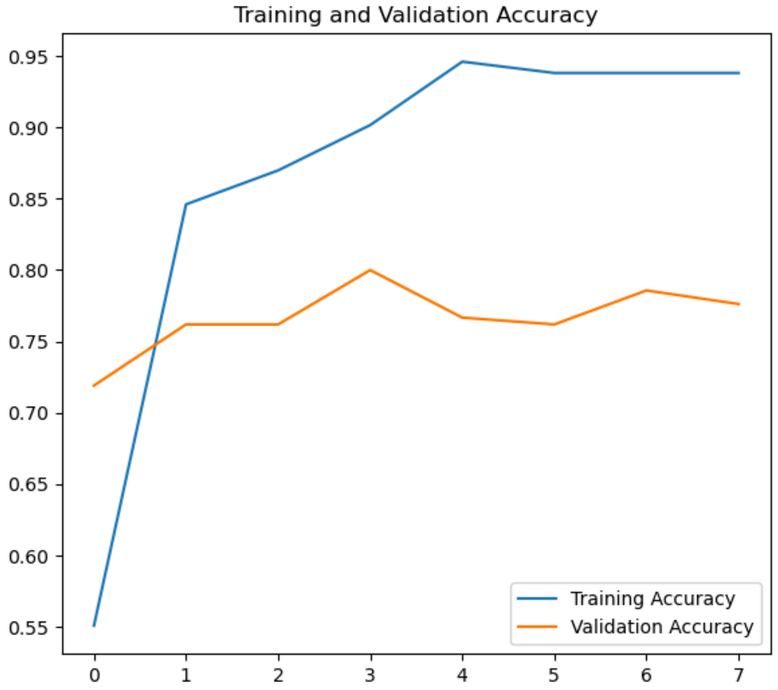
Résultats de la classification supervisée

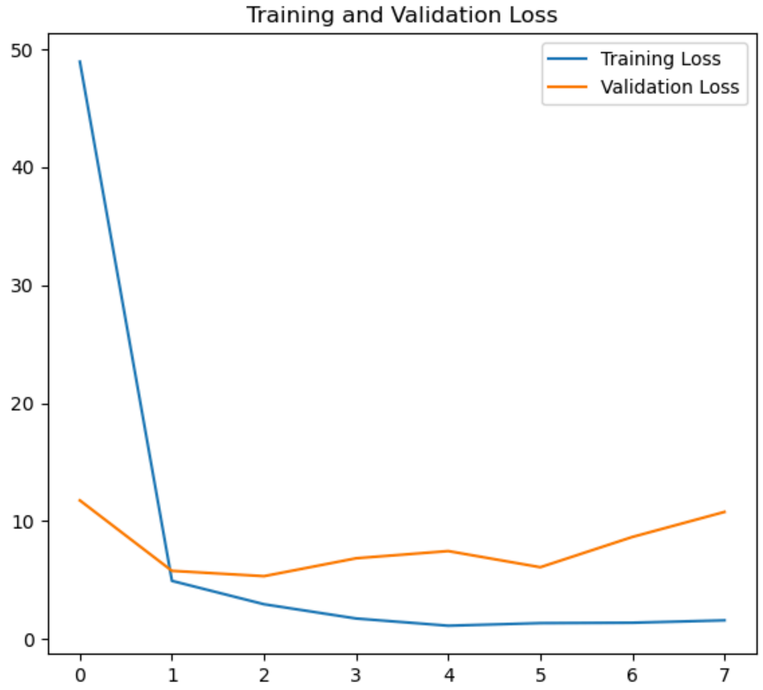
Résultats de la classification supervisée

Sans Data Augmentation

Training Accuracy = 1
Validation accuracy = 0,78
Test accuracy = 0,78

- VGG16 pré-entrainé
  - Sans couche fully connected
  - Ajout de couches à entrainer
    - Flatten
    - Dense
    - Dropout
    - Prediction
- Entrainement sur 840 images
- Validation sur 210 images
- Test sur 210 images





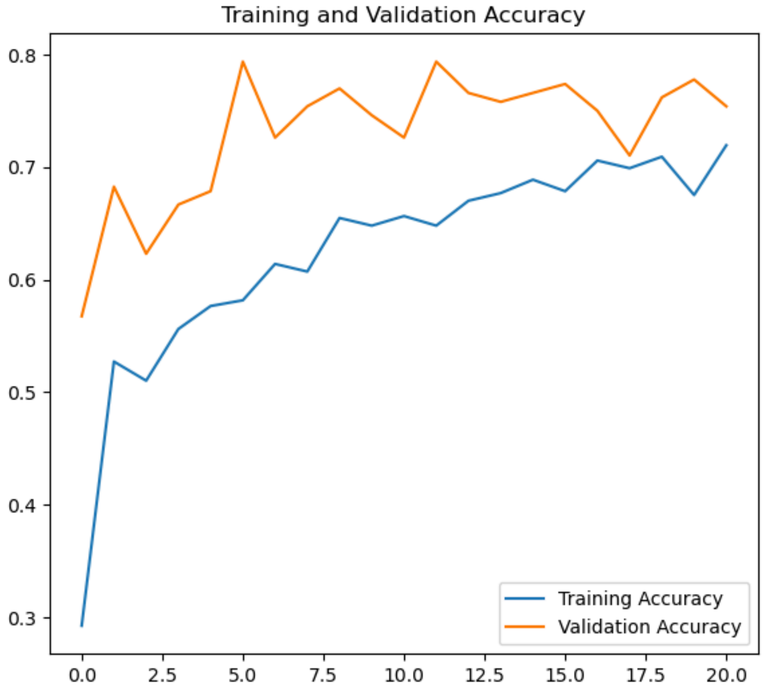
Modèle sur-entrainé

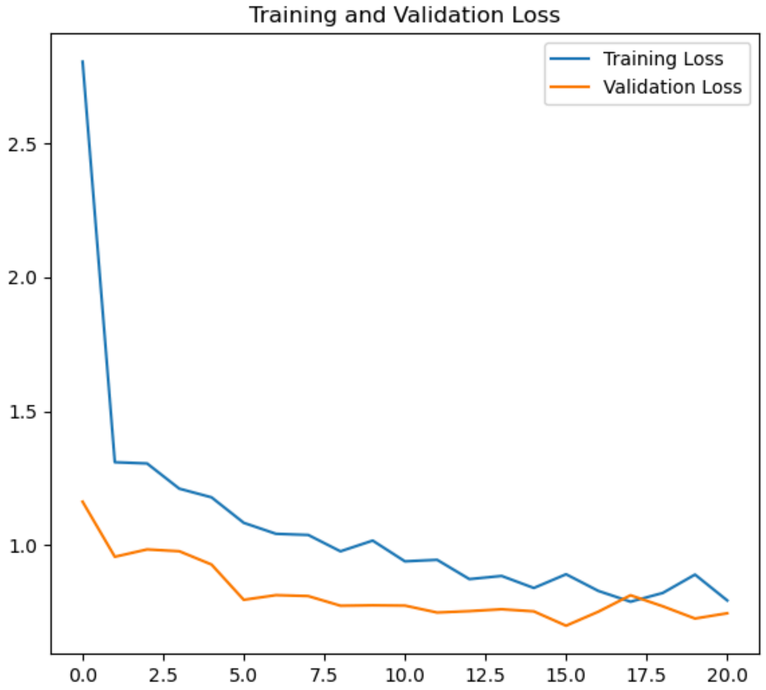
Résultats de la classification supervisée

Avec Data Augmentation

Training Accuracy = 0,82
Validation accuracy = 0,75
Test accuracy = 0,77

- VGG16 pré-entrainé
  - Sans couche fully connected
    - Même couches ajoutées en sortie que précédemment
  - Ajout d'une couche de data augmentation
    - RandomFlip
    - RandomRotation
    - RandomZoom
- Entrainement sur 840 images
- Validation sur 210 images
- Test sur 210 images





Présentation du test de l’API

Présentation du test de l'API

- S'inscrire sur rapidapi.com pour obtenir une clé d'authentification
- S'inscrire sur edamam.com pour créer une app et obtenir les codes de l'app
- Exécuter le script Python
- Renseigner sa clé d'authentification et ses codes d'app
- Un fichier .csv se créé avec les 10 premières entrées relatives au mot « champagne » avec les features demandées

extraction_10_premiers_produits_champagne

	food.foodId	food.label	food.category	food.foodContentsLabel	food.image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods		https://www.edamam.com/food-img/a71/a718cf3c52add522128929f1f324d2ab.jpg
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR; GARLIC; DIJON MUSTARD; SEA SALT.	
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINEGAR; SUGAR; OLIVE OIL; SALT; DRIED GARLIC; DRED SHALLOTS; BLACK PEPPER; XANTHAN GUM; SPICE	https://www.edamam.com/food-img/d88/d88b64d97349ed062368972113124e35.jpg
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS SULFITES); WATER; VINEGARS (CHAMPAGNE AND WHITE WINE); SUGAR; SALT; MUSTARD SEED; MONOSODIUM GLUTAMATE; GARLIC*; ONION*; SPICE; XANTHAN GUM; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; CHIVES*; TAMARIND; NATURAL FLAVOR.	
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS SULFITES); VINEGARS (CHAMPAGNE AND WHITE WINE); SUGAR; SALT; MUSTARD SEED; MONOSODIUM GLUTAMATE; GARLIC*; ONION*; SPICE; XANTHAN GUM; POTASSIUM SORBATE ADDED TO MAINTAIN FRESHNESS; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; CHIVES*; TAMARIND.	
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFITES); WATER; WHITE WINE VINEGAR; SUGAR; SALT; SPICES (INCLUDING MUSTARD SEED); MONOSODIUM GLUTAMATE; GARLIC*; ONION*; XANTHAN GUM; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; VINEGAR; CORN SYRUP; CARAMEL COLOR; CHIVES*; NATURAL FLAVOR; TAMARIND.	https://www.edamam.com/food-img/ab2/ab2459fc2a98cd35f68b848be2337ecb.jpg
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne; milk	
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	
8	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanilla extract; champagne; powdered sugar	
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; shallot; honey; Salt; pepper	

Conclusion

Conclusion

- Etude de faisabilité
  - Meilleur modèle textuel : Bag of Words (TF-IDF)
  - Meilleur modèle Visuel : VGG16 (2 couches fully connected)
  - Score ARI de 0,63
- Classification supervisée des images via VGG16, avec data augmentation
- Script API pour collecte des données fonctionnel

Merci