Projet 6 - Classifiez automatiquement des biens de consommation

Classification des articles d'une marketplace en les catégorisant automatiquement, sur la base de la description et de l'image fournie.

Sommaire

puis remarques & axes d'amélioration.

Problématique & Données

La problématique

Contexte : “Place de marché” souhaite lancer une marketplace sur laquelle les vendeurs peuvent proposer des produits à la vente en fournissant une photo et une description. Les vendeurs choisissent manuellement la catégorie du produit. Pour augmenter en fiabilité Linda, Lead DS, me confie la mission d’étudier la possibilité d’automatiser cette tâche d’attribution de catégories.

Objectif : Etudier la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant. Dans l’objectif de gagner en fiabilité et en temps.

Moyens : Une sélection d’articles déjà catégorisées.

Les données - un fichier CSV & images associées

Catégories	Quantité
Home Furnishing	150
Baby Care	150
Watches	150
Home Decor & Festive Needs	150
Kitchen & Dining	150
Beauty and Personal Care	150
Computers	150
Total	**1050**

Les variables à utiliser :

- ➔ product_name
- ➔ product_category_tree ⇒ category
- ➔ image ⇒ image_path (des images fournies)
- ➔ description

Taux de remplissage de 100% pour ces variables !

Prétraitement, extraction de Features sur la partie TEXTE

Prétraitement du TEXTE

- Texte = **product_name** + description

Elegance Polyester Multicolor Abstract Eyelet Door Curtain *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 interior…*

- Fonction de preprocessing :
  - Conversion en minuscules & Suppression de la ponctuation
  - Tokenisation des mots
  - Suppression des stop words & des mots de moins de 3 lettres
  - Stemming
  - Suppression des caractères non alphabétiques
  - Suppression des chaînes de texte vides

```
['eleg','polyest','multicolor','abstract','eyelet','door','curtain','key','featur','eleg','polyest','multicolor','abstract','eyelet','door','curtain','floral','curtaineleg','polyest','multicolor','abstract','eyelet','door','curtain','height','pack','pric','curtain','enhanc','look','interiorsth',...
```

- Suppression du vocabulaire des mots qui apparaissent moins de 6 fois et ceux qui apparaissent trop souvent

Les méthodes testées pour représenter le texte

MÉTHODES PLUS TRADITIONNELLES

Bag of words et TF-IDF (fréquence, stats.)

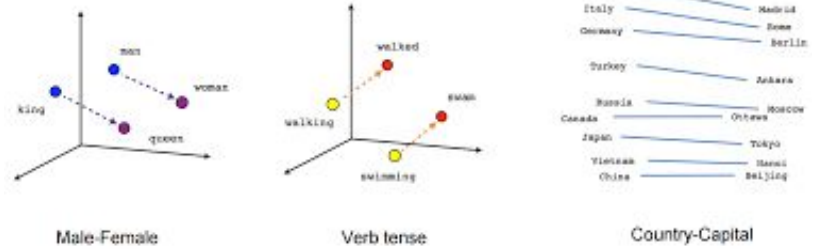
	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2								0	8
Review 3	1	1	1								1	6

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF (Term x within document y)

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

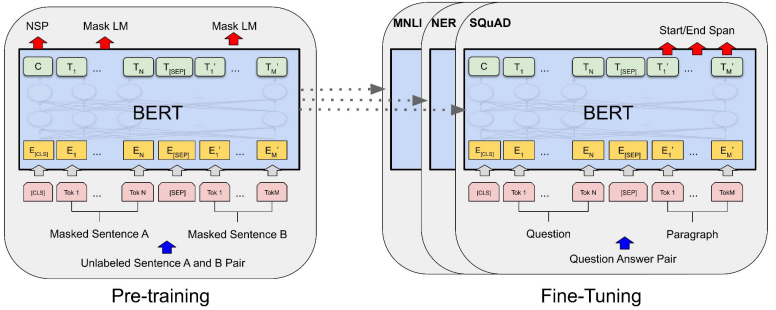
Word vectors (doc2vec, word2vec) (peu profonds



Male-Female | Verb tense | Country-Capital

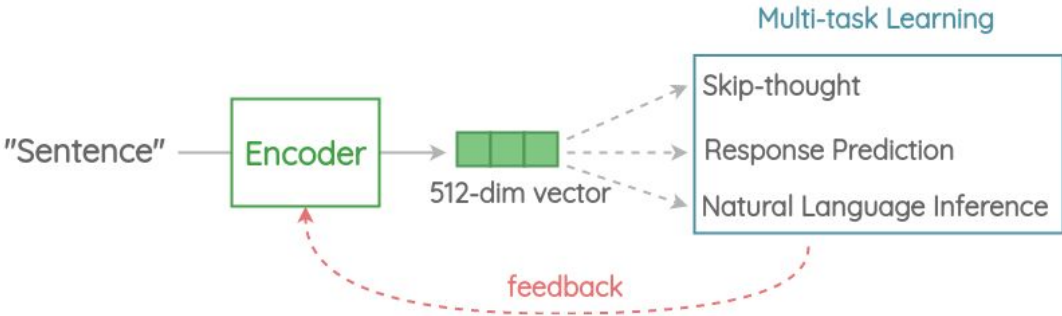
DEEP LEARNING

BERT



Pre-training | Fine-Tuning

USE

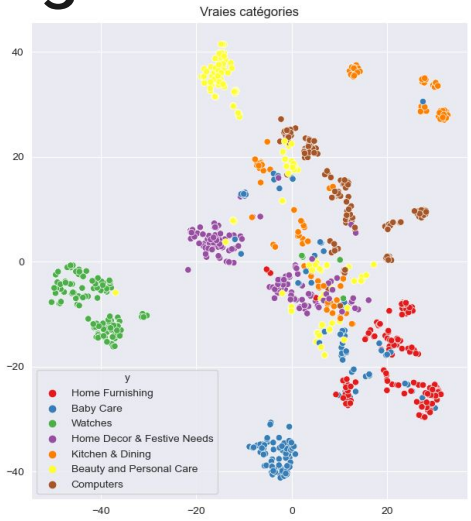


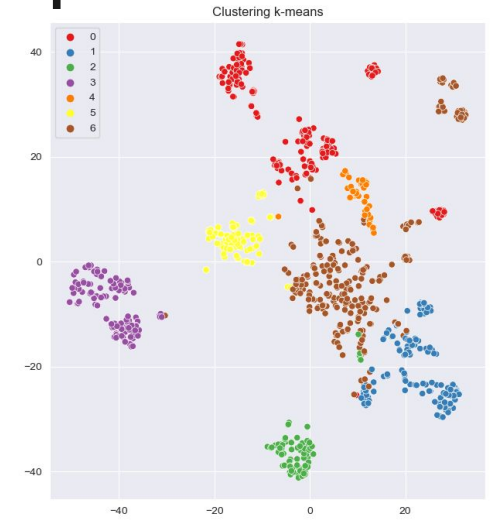
Méthodologie : Segmentation & Comparaison des résultats

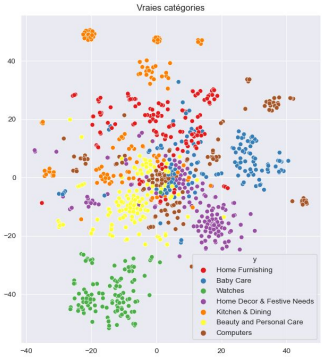
- Utilisation du feature engineering présenté
- K-means avec 7 clusters sur les features adaptés pour chaque modèle
- Réprésentation visuelle via TSNE à 2 dimensions
- Comparaison vraies catégories et segmentation
- Indice de Rand pour comparer les résultats des différentes méthodes

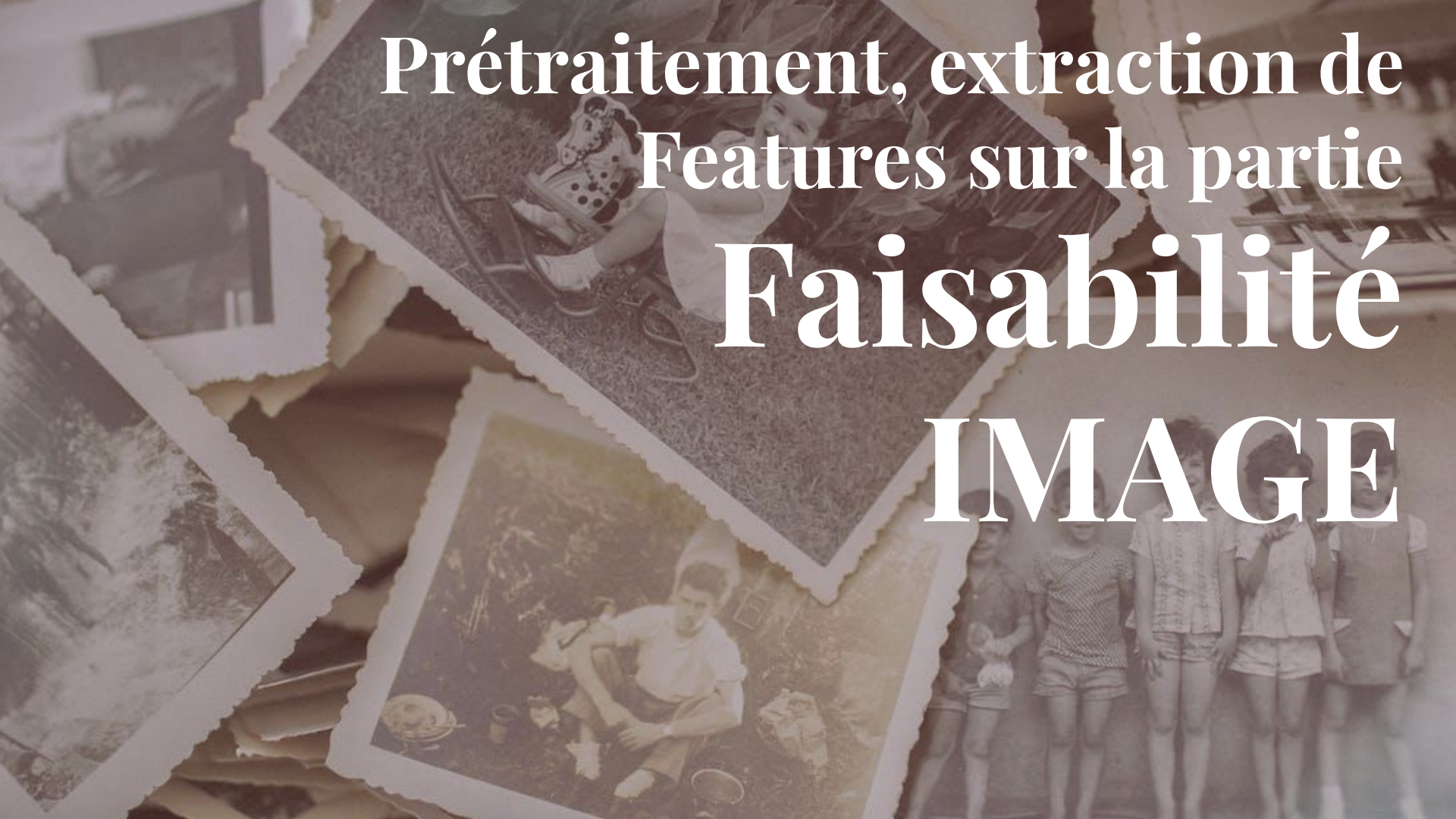
Méthodologie : Segmentation & Comparaison des résultats

	Nom du modèle	indice_rand
0	Bag of Words	0.13546
1	TFIDF	0.32654
2	Word2Vec	0.39308
3	BERT bis	0.32155
4	USE bis	0.31724





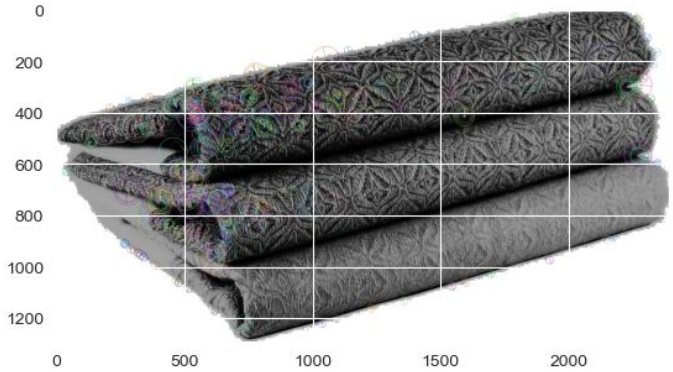




Prétraitement, extraction de Features sur la partie Faisabilité IMAGE

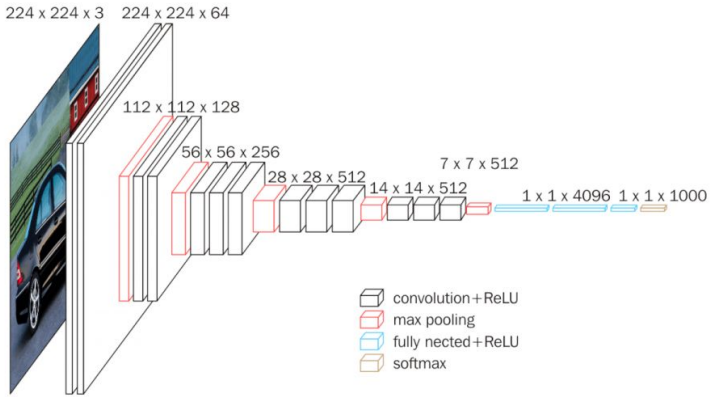
2 Méthodes, modèles et preprocessing associé

MÉTHODE TRADITIONNELLE



- Resize
- Nuances de gris
- Egalisation des histogrammes
- descripteurs d'image SIFT
- créations de clusters de descripteurs (625)
- Création histogramme par image

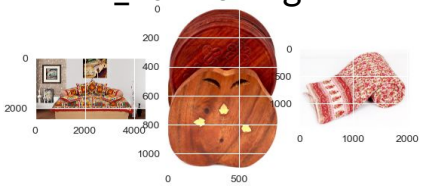
VGG16 PRE ENTRAÎNE



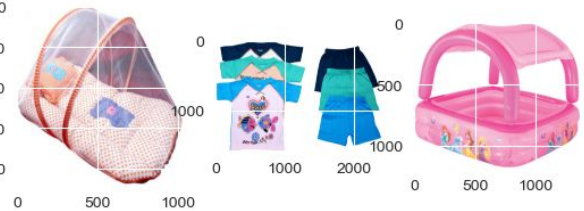
- Resize
- Turn into array
- Expand dimension
- Fonction de preprocessing dédiée à VGG16

Méthodologie : Segmentation & Comparaison des résultats

Home_Furnishing



Baby_Care



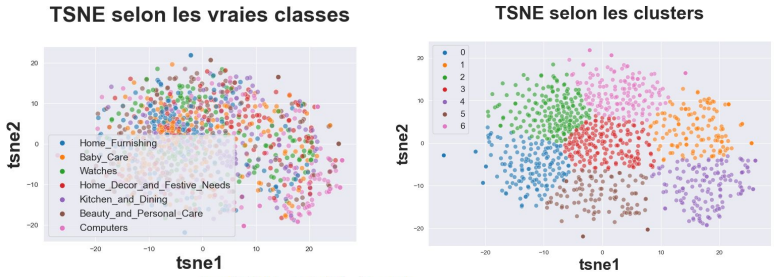
Watches

- Preprocessing des images selon méthode
- Utilisation des features dégagées selon chacune des méthodes
- Réduction de dimension PCQ n_components=0.99
- T-SNE à n = 2
- K-means avec 7 clusters sur les features le T-SNE
- Représentation visuelle via TSNE à 2 dimensions
- Comparaison vraies catégories et segmentation
- Indice de Rand, Accuracy, Matrice de confusion

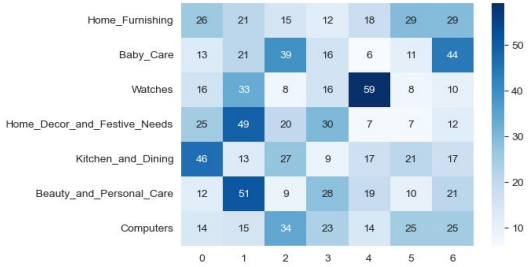
Méthodologie : Segmentation & Comparaison des résultats

MÉTHODE TRADITIONNELLE

0.03

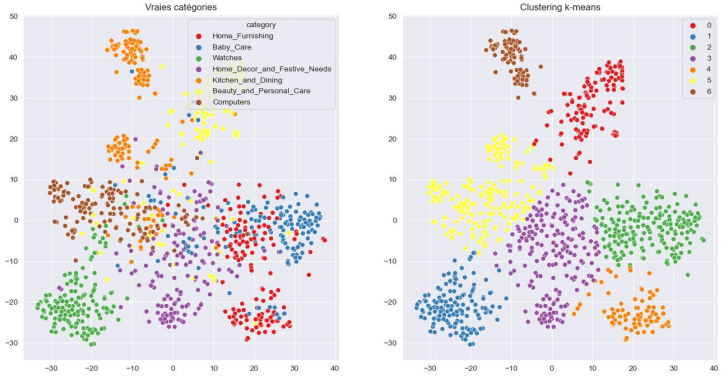


	precision	recall	f1-score
Baby_Care	0.17	0.17	0.17
Beauty_and_Personal_Care	0.10	0.14	0.12
Computers	0.05	0.05	0.05
Home_Decor_and_Festive_Needs	0.22	0.20	0.21
Home_Furnishing	0.12	0.11	0.12
Kitchen_and_Dining	0.09	0.07	0.08
Watches	0.16	0.17	0.16
accuracy			0.13
macro avg	0.13	0.13	0.13
weighted avg	0.13	0.13	0.13

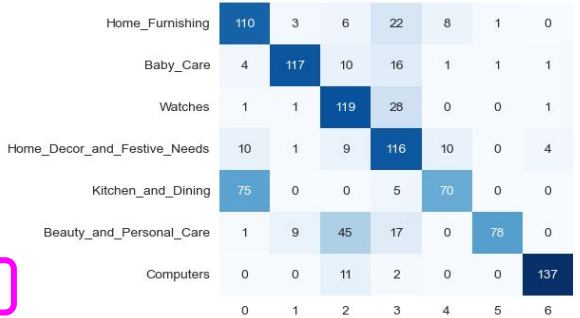


VGG16 PRE ENTRAÎNE

ARI 0.48



	precision	recall	f1-score
Baby_Care	0.55	0.73	0.63
Beauty_and_Personal_Care	0.89	0.78	0.83
Computers	0.59	0.79	0.68
Home_Decor_and_Festive_Needs	0.56	0.77	0.65
Home_Furnishing	0.79	0.47	0.59
Kitchen_and_Dining	0.97	0.52	0.68
Watches	0.96	0.91	0.94
accuracy			0.71
macro avg	0.76	0.71	0.71
weighted avg	0.76	0.71	0.71

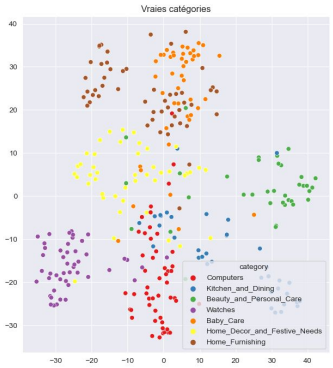


Classification supervisée IMAGE

3 méthodes testées

Modèle SANS data augmentation sans entraînement

- Preprocessing VGG16 cf faisabilité
- PCA n=0,99
- TSNE n=2
- Train & Test Split (0.3)
- Knn plus proches voisins
- Accuracy = 0.80

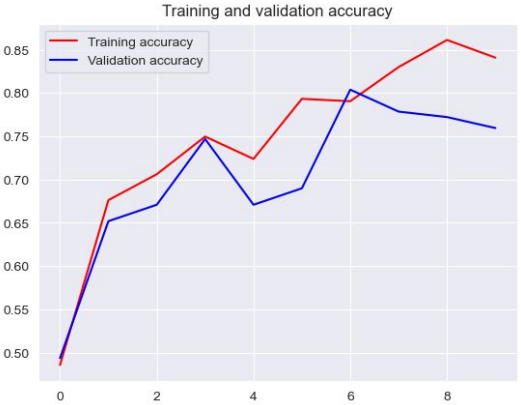


Modèle SANS data augmentation avec entraînement

- Train, Test, Validation Split (0.7, 0.15, 0.15)
- Preprocessing VGG16 cf faisabilité
- Label encoding
- Chargement de VGG16 et conservation des top layers pour entraînement des couches de classification
- Accuracy = 0.78

Modèle AVEC data augmentation avec entraînement

- Train, Test, Validation Split (0.7, 0.15, 0.15)
- Création de 2 datagen pour train_generator & valid_generator (label encoding & preprocessing inclus)
- Chargement de VGG16 et conservation des top layers pour entraînement des couches de classification
- Accuracy = 0.63



Quelle méthode choisir ?

Modèle SANS data augmentation sans entraînement

- Preprocessing VGG16 cf faisabilité
- PCA n=0,99
- TSNE n=2
- Train & Test Split (0.3)
- Knn plus proches voisins
- Accuracy = 0.80

Problème de dimensionnalité, selon le nombre de données qui alimenteraient le modèle

Beaucoup d'étapes

Modèle SANS data augmentation avec entraînement

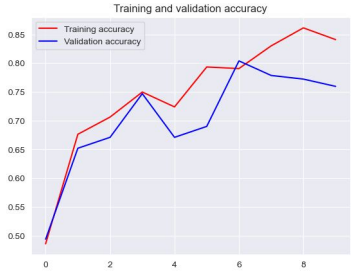
- Train, Test, Validation Split (0.7, 0.15, 0.15)
- Preprocessing VGG16 cf faisabilité
- Label encoding
- Chargement de VGG16 et conservation des top layers pour entraînement des couches de classification
- Accuracy = 0.78

L'entraînement n'améliore pas les résultats

Bien que bons, ils sont peut être liés à un mauvais choix des données

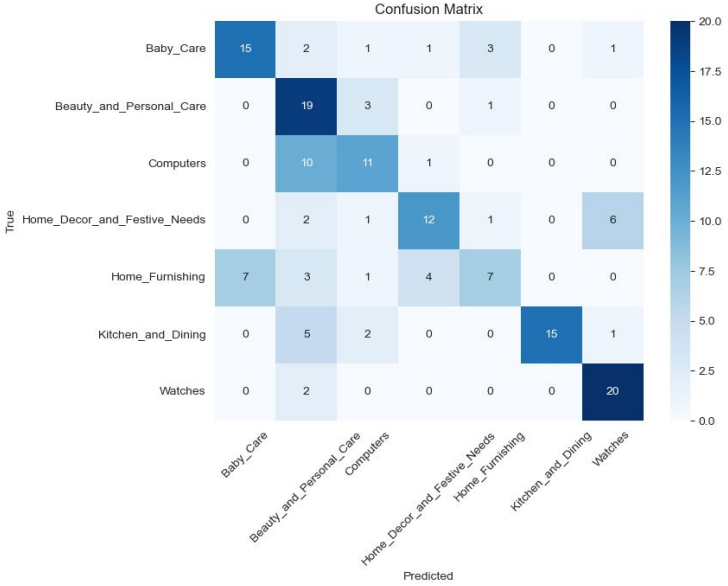
Modèle AVEC data augmentation avec entraînement

- Train, Test, Validation Split (0.7, 0.15, 0.15)
- Création de 2 datagen pour train_generator & valid_generator (label encoding & preprocessing inclus)
- Chargement de VGG16 et conservation des top layers pour entraînement des couches de classification
- Accuracy = 0.63



L'entraînement améliore les résultats

Rapport sur la méthode choisie



Classification Report:

	precision	recall	f1-score	support
Baby_Care	0.68	0.65	0.67	23
Beauty_and_Personal_Care	0.44	0.83	0.58	23
Computers	0.58	0.50	0.54	22
Home_Decor_and_Festive_Needs	0.67	0.55	0.60	22
Home_Furnishing	0.58	0.32	0.41	22
Kitchen_and_Dining	1.00	0.65	0.79	23
Watches	0.71	0.91	0.80	22
accuracy			0.63	157
macro avg	0.67	0.63	0.63	157
weighted avg	0.67	0.63	0.63	157

Exemples d'erreurs :

Prédit Watches à la place de Kitchen_and_Dining :

Prédit Beauty_and_Personal_Care à la place de Kitchen_and_Dining :

Prédit Home_Furnishing à la place de Baby_Care :

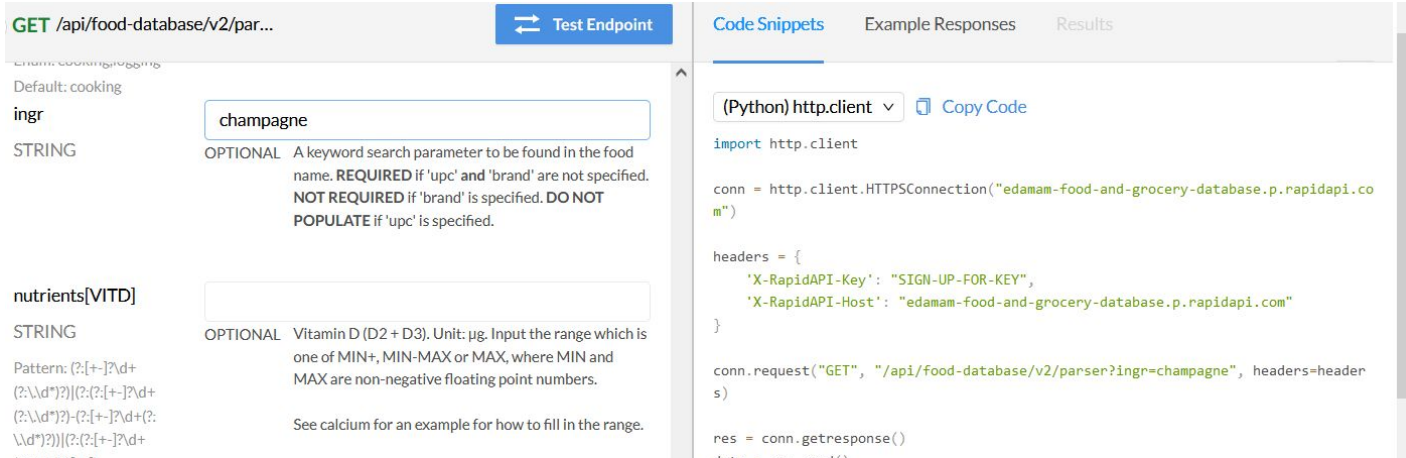

Test de l'API

Collecte de données

Contexte : Elargir notre gamme de produits, en particulier dans l'épicerie fine.

Objectif : Tester la collecte de produits à base de "champagne" et fournir une extraction des 10 premiers produits dans un fichier ".csv", contenant pour chaque produit les données suivantes : foodId, label, category, foodContentsLabel, image.

Moyens : via l'API EDAMAM.



Le fichier CSV

A	B	C	D	E	F	G	H
	foodId	label	category	foodContentsLabel	image		
0	food_a656mk2a5dmqb2adi	Champagne	Generic foods	None	https://www.edamam.com/foo		
1	food_a656mk2a5dmqb2adi	Champagne	Generic foods	None	https://www.edamam.com/foo		
2	food_b753ithamdb8psbt0w	Champagne Vinaigrette, Cham	Packaged foods	OLIVE OIL; BALSAMIC VINEG	None		
3	food_b3dyababjo54xobm6r	Champagne Vinaigrette, Cham	Packaged foods	INGREDIENTS: WATER; CANO	https://www.edamam.com/foo		
4	food_a9e0ghsamvoc45bwa	Champagne Vinaigrette, Cham	Packaged foods	CANOLA AND SOYBEAN OIL;	None		
5	food_an4jjueaucpus2a3u1n	Champagne Vinaigrette, Cham	Packaged foods	WATER; CANOLA AND SOYBE	None		
6	food_bmu5dmkazwuvpaa5p	Champagne Dressing, Champa	Packaged foods	SOYBEAN OIL; WHITE WINE (	https://www.edamam.com/foo		
7	food_alpl44taoyv11ra0lic1q	Champagne Buttercream	Generic meals	sugar; butter; shortening; var	None		
8	food_byap67hab6evc3a0f9v	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; (	None		
9	food_am5egz6aq3fpjlaf8xpl	Champagne Truffles	Generic meals	butter; cocoa; sweetened co	None		


RGPD

RGPD

Les grands principes :

- Licéité, loyauté et transparence : données collectées de manière légale, transparente et avec le consentement des utilisateurs.
- Limitation des finalités
- Minimisation des données
- Exactitude des données : maintenues à jour et exactes.
- Limitation de la conservation

L'application au projet :

Ici : photos & descriptions sur une marketplace

Seulement des données non personnellement identifiables ?

Des précautions à prendre :

- consentement des utilisateurs sur l'utilisation des données,
- consigne de ne pas mettre de photos identifiables (ex visage, identité ...)
- ni de description avec des données personnelles (ex numéro de téléphone, adresse, identité...)
- suppression des images et texte une fois que les modèles sont entraînés (indépendamment de leur présence sur le site)

Mail de Linda concernant la PI :

Linda

PS : J'ai bien vérifié qu'il n'y avait aucune contrainte de propriété intellectuelle sur les données et les images.

Conclusions & Remarques

Ce que nous soumettons :

- La méthode de classification supervisée avec data augmentation avec entraînement du modèle semble être la plus prometteuse. Il faudrait étendre le test à un plus grand nombre de e données pour être sûr.

Si c'était à refaire :

- Pour le test split, j'utiliserais des images que j'ai extraites à l'aide de web scraping depuis Google Images, afin d'obtenir de nouvelles données et d'utiliser la totalité de mon jeu de données, qui est de taille très réduite.

Point de vigilance :

- Il faut effectuer de la veille régulièrement sur les nouvelles méthodes d'analyse d'image

=> Slide : Veille et Proof of Concept d'une récente technique ?

Le plus important : j'ai gagné en compétences