

Rakuten Institute of Technology

Participants :

Olga TOLSTOLUTSKA

Mohamed BACHKAT

Charly LAGRESLE

 DataScientest

Mentor : Manu POTREL

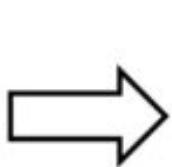
Promotion: DST Bootcamp DEC22



白いクリーマー

商品仕様 ★サイズ
8.3×7×H7.8cm ★重量
160g ★容量 240cc ★色
ほんの少しグレーが
かかった白 ★素材 日
本製磁器 ★その他 電
子レンジ,食器洗浄機
OK! オープン不可 こ
ちらはアウトレット
商品です。細かなキ
ズ、ピンホール(針
を刺したような穴) ...

Rakuten Data Challenge: extracting information from texts and images with Deep Learning to enrich the catalog of product



Material **Porcelain**
Color **White**
...

Filters

The screenshot shows a search result for 'ミルク ピッチャー' (milk pitcher) on the Rakuten website. The search bar at the top has the query 'ミルク ピッチャー'. Below the search bar, there's a promotional banner for 'Rakuten UN-LIMIT V' with text '料金そのまま4G・5Gも使える プラン 料金/月 1年無料'. The main search results area shows several items:

- Milk Pitcher (2 pieces) (White) (Outlet) Pitcher**: Price 380 yen + tax 460 yen, 3 points (1 point), 4.00 (40 reviews), 99 likes.
- Royal Doulton Sugar Bowl & Milk Pitcher 3-piece Set**: Price 3,036 yen + tax 3,643 yen, 300 points (1 point), 4.75 (14 reviews), 99 likes.
- Gorham Milk Pitcher Small 500ml 5 Colors (SAKAI WH-0)**: Price 330 yen + tax 400 yen, 3 points (1 point), 4.4 (43 reviews).
- Common Milk Pitcher 100ml Milk入 西海陶器 SAKAI WH-0**: Price 1,100 yen + tax 1,320 yen, 11 points (1 point), 4.5 (5 reviews), 39 likes.

On the left side of the results, there are filters for material ('Material'), color ('Color'), price ('Price'), and brand ('Brand'). The 'Color' filter is specifically highlighted with a blue oval around the color palette. The palette includes options for black, grey, white, red, orange, yellow, green, blue, and purple.

<https://www.rakuten.co.jp>, Jan. 2021

Description des données

- 27 variables cibles
- 84 916 observations: des données textuelles ainsi que des images .
- Pas de duplications des données
- Les données textuelles sont divisés en deux colonnes : designation et description . Elles represent un titre du produit et sa decription.
- Le titre du produit est composé de 4 à 54 mots
- La description est plus longs et contient entre 0 (certaines descriptions sont vides) et 2 068 mots
- Images : couleur, 500x500px encodées au format JPG

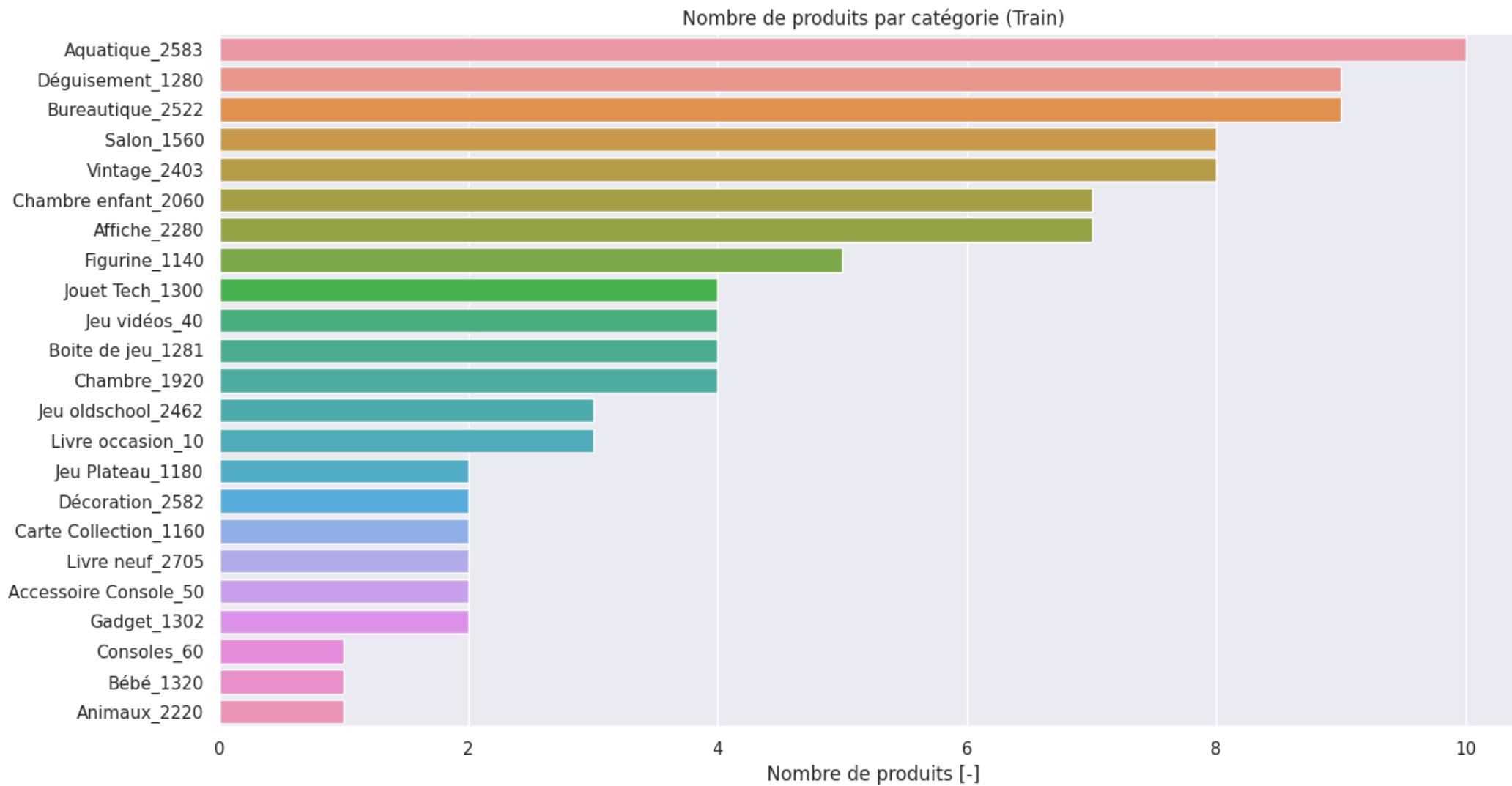
Description des données / Nombre de produits par catégorie



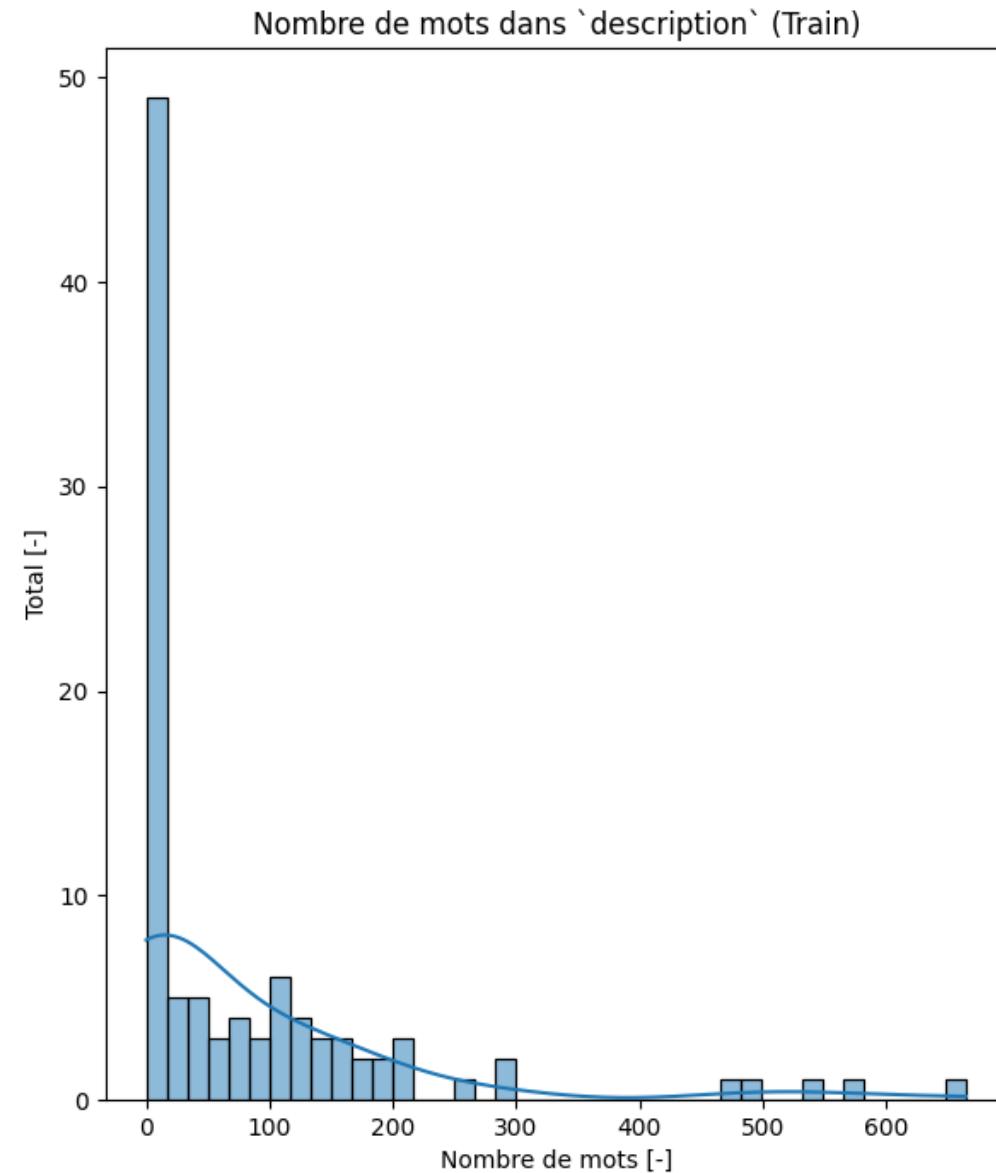
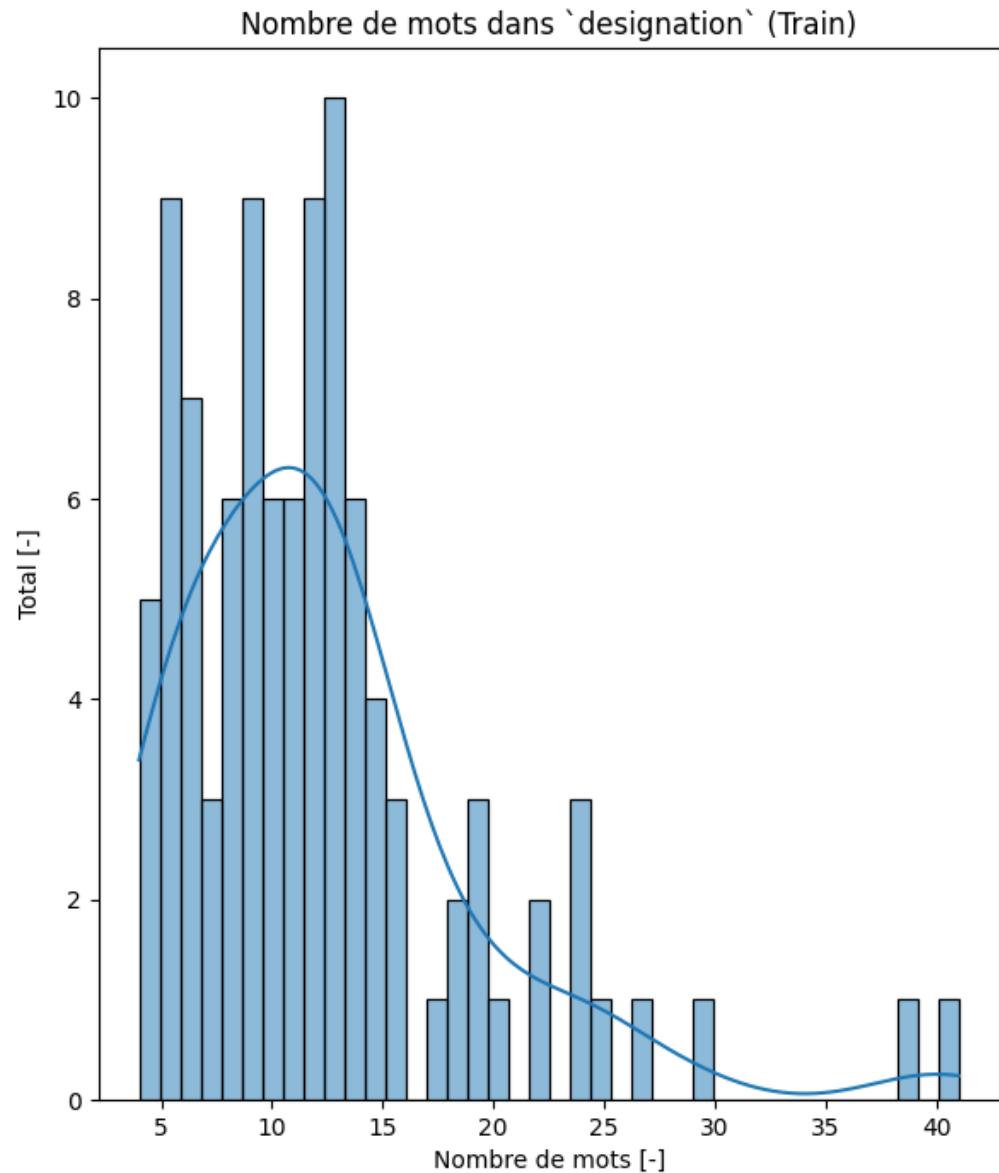
Description des données /Les catégories et leurs descriptions

Catégorie	Description	Catégorie	Description	Catégorie	Description
10	Livre d'occasion	1301	Chaussette	2462	Jeu oldschool
40	Jeu Console	1302	Gadget	2522	Bureautique
50	Accessoire Console	1320	Bébé	2582	Décoration
60	Tech	1560	Salon	2583	Aquatique
1140	Figurine	1920	Chambre	2585	Soin et Bricolage
1160	Carte Collection	1940	Cuisine	2705	Livre neuf
1180	Jeu Plateau	2060	Chambre enfant	2905	Jeu PC
1280	Déguisement	2220	Animaux		
1281	Boite de jeu	2280	Affiche		
1300	Jouet Tech	2403	Revue		

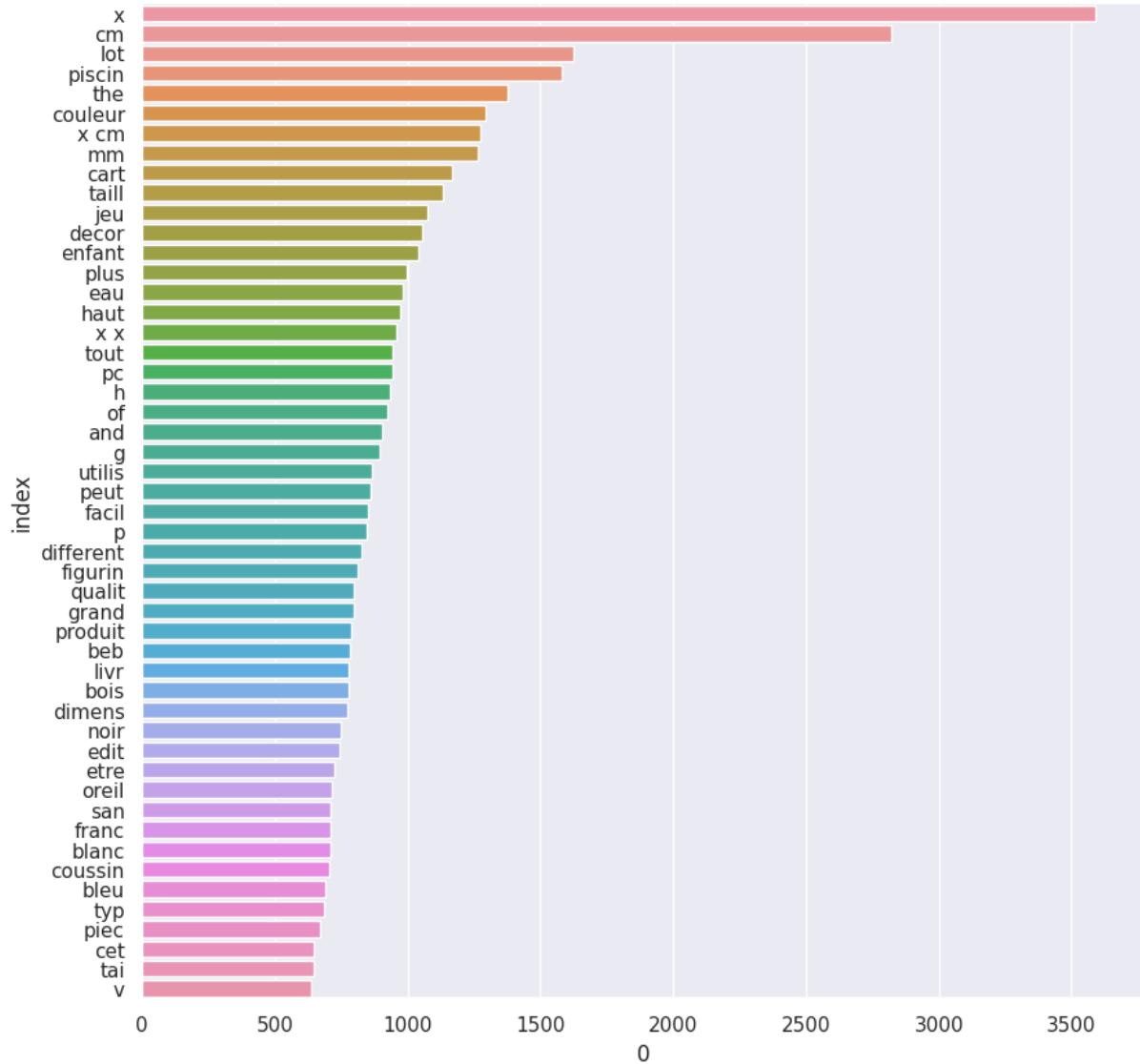
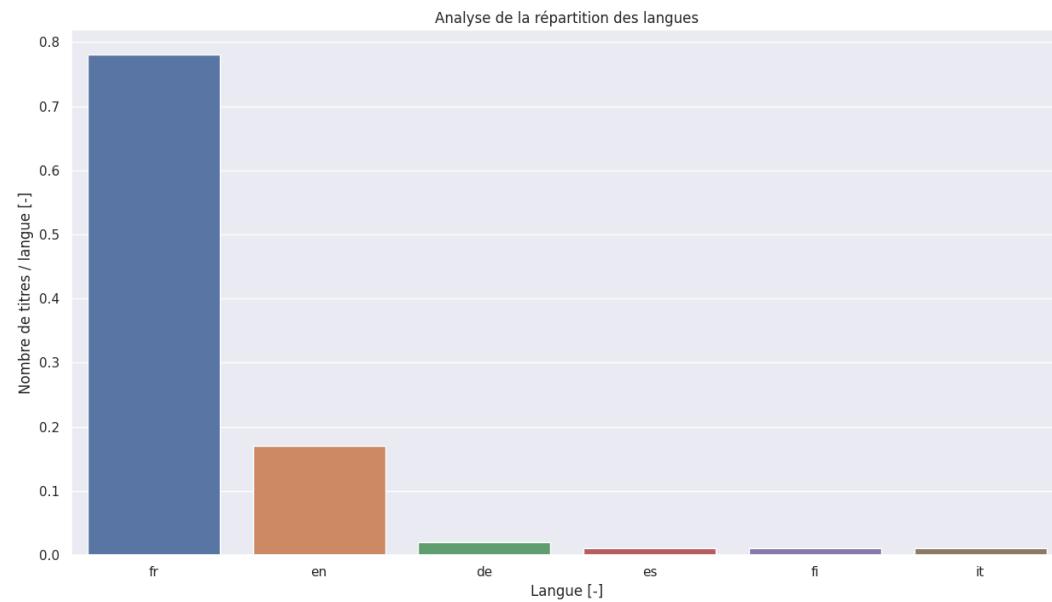
Exploration des données / Target



Exploration des données / Text

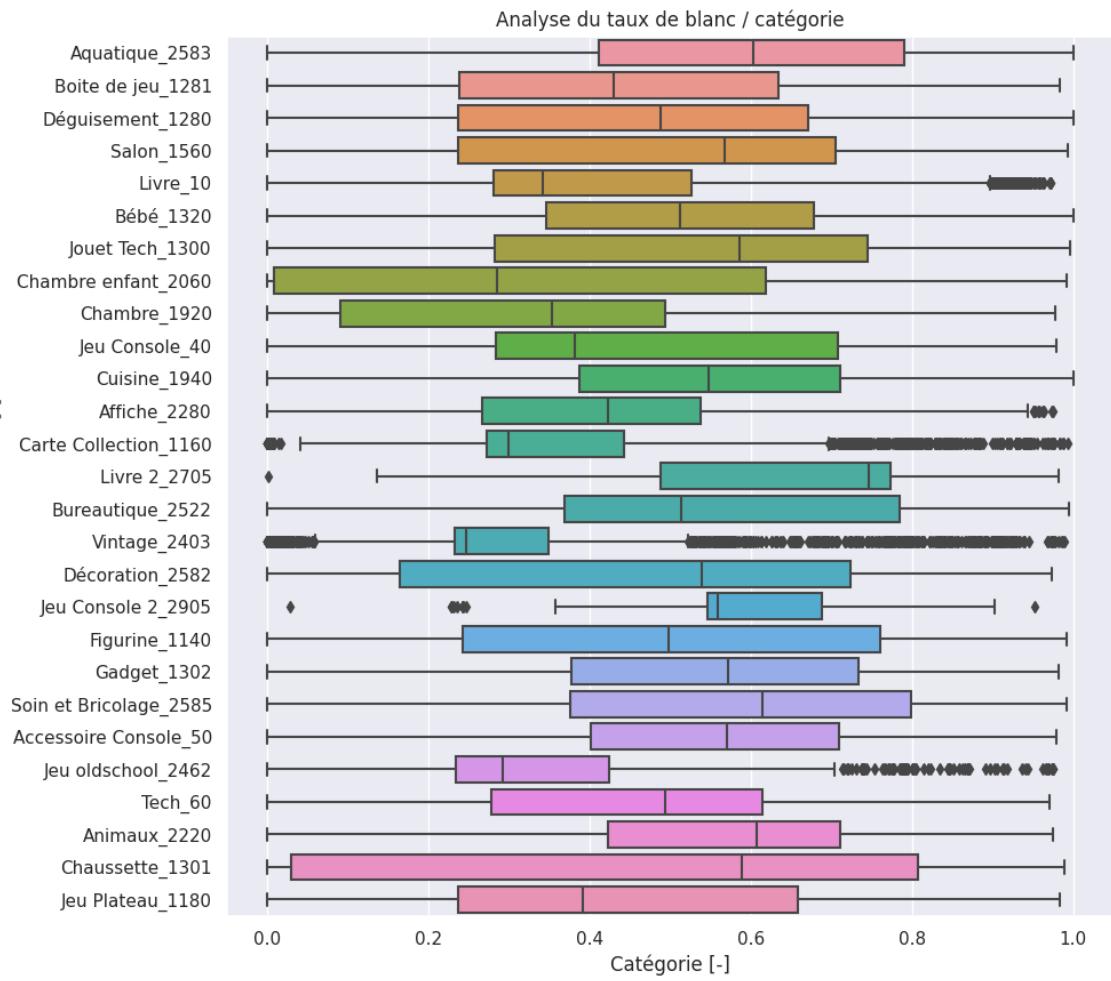


Exploration des données / Text

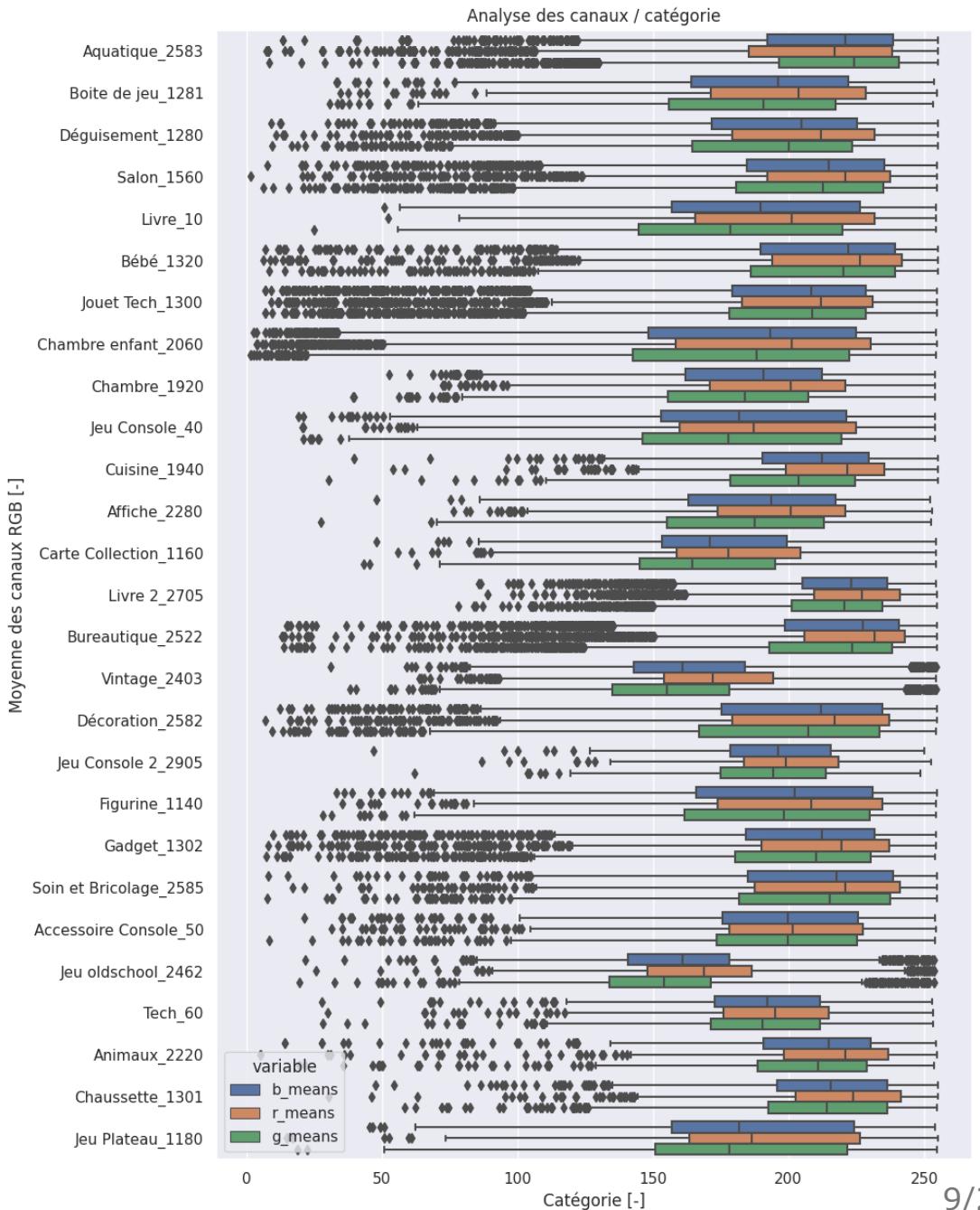


Exploration des données / Images

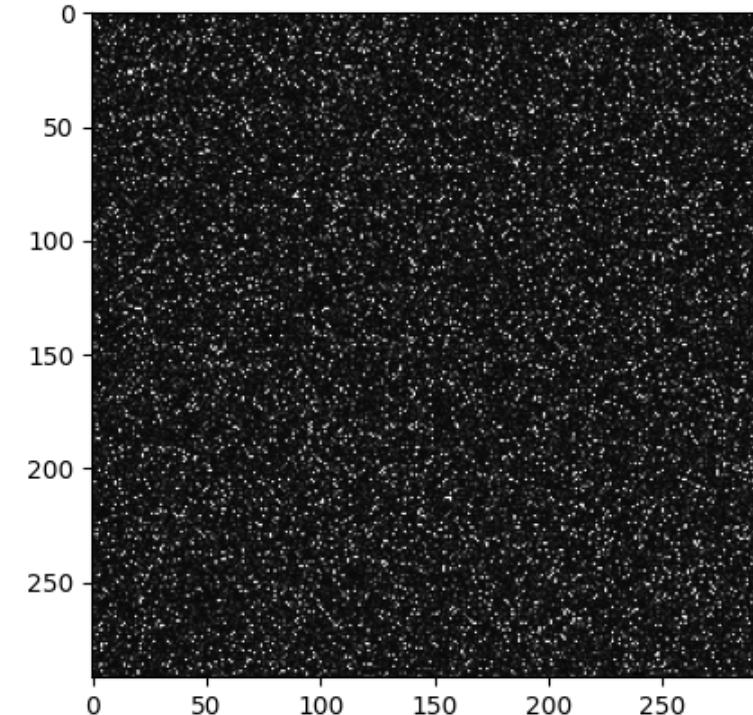
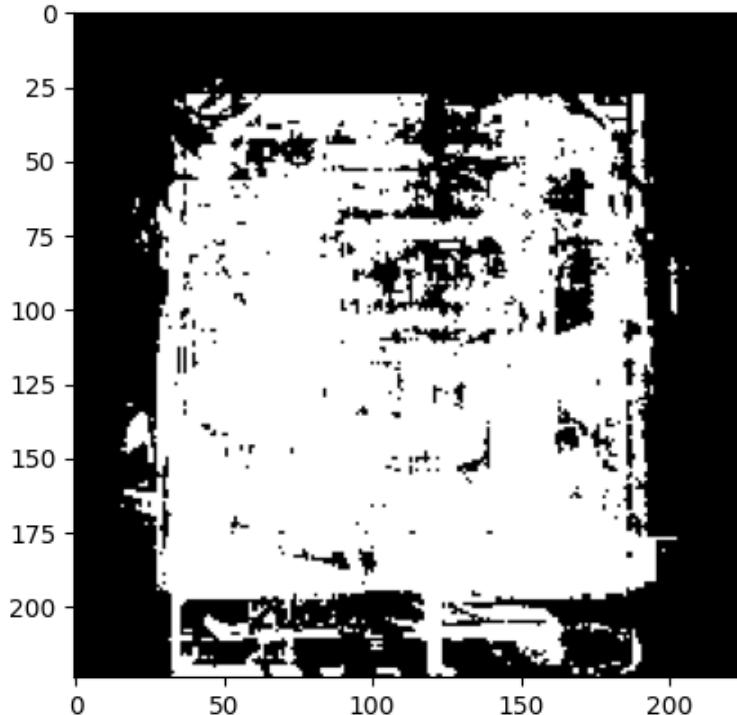
Taux de blanc [-]



Moyenne des canaux RGB [-]



Exploration des données / Images



Les étapes suivantes sont exécutées via des pipelines afin de transformer les données textuelles :

- Fusion de deux colonnes `designation` et `description` dans `text`
- Détection la langue dans `text` et ensuite traduire les textes vers le français
- Nettoyage de la colonne `text` : supprimer les caractères spéciaux, supprimer la ponctuation et etc.
- Suppression des *stopwords*
- Extraction de la racine des mots
- Vectorisation du texte via un `Tokenizer`

L'étape de vectorisation varie en fonction du modèle utilisé par la suite.

Préparation des données / Text

Les étapes suivantes sont exécutées via des pipelines afin de transformer les données textuelles :

- Fusion de deux colonnes `designation` et `description` dans `text`
- Détection la langue dans `text` et ensuite traduire les textes vers le français

Etape	Résultat
Suppression les balises html	<p> Blabla
Suppression de la ponctuation et caractères spéciaux	Bla bla
Suppression les valeurs numériques	Bla bla
Suppression les majuscules	Bla bla
Suppression des <i>stopwords</i>	Bla bla
Extraction de la racine des mots	Bla bla
Vectorisation du texte via un <code>Tokenizer</code>	[86, 1, 2 ...]

ImageDataGenerator:

- streaming per batch : les images sont transmises sous de batchs ce qui évite de traiter l'ensemble des données d'un coup
- augmentation de données via les transformation appliqués
- rédimensionnement en taille 224x224
- application de la fonctionne `preprocess_input` spécifique pour chaque modèle

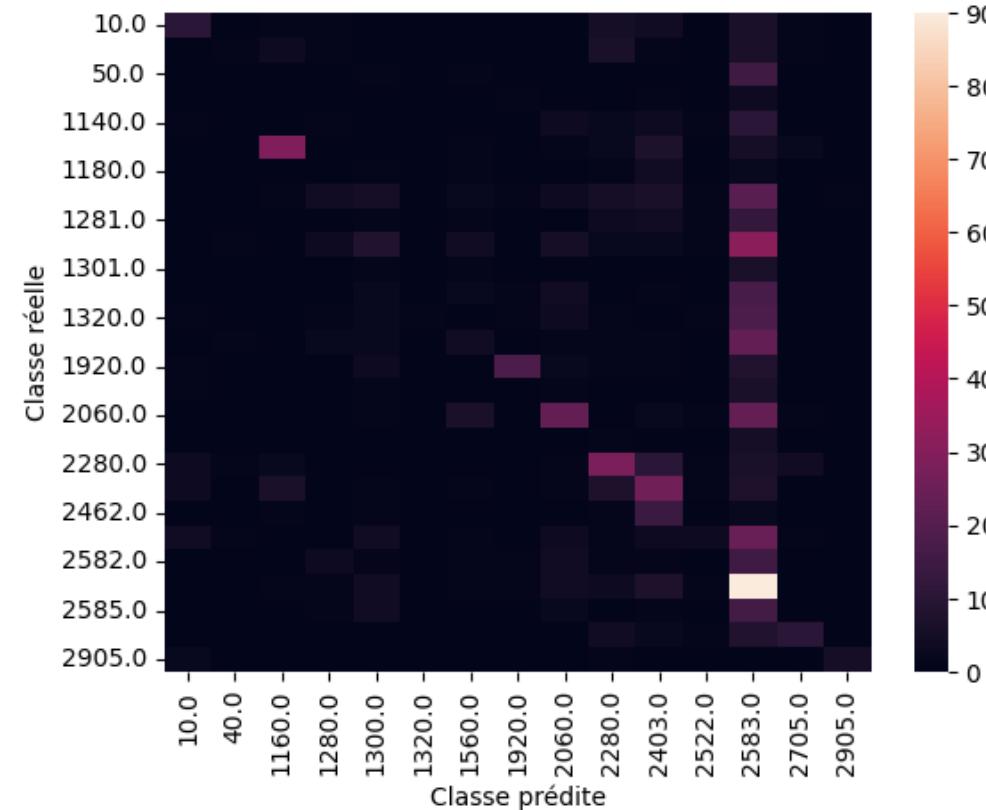


Les modèles / Machine Learning / Text

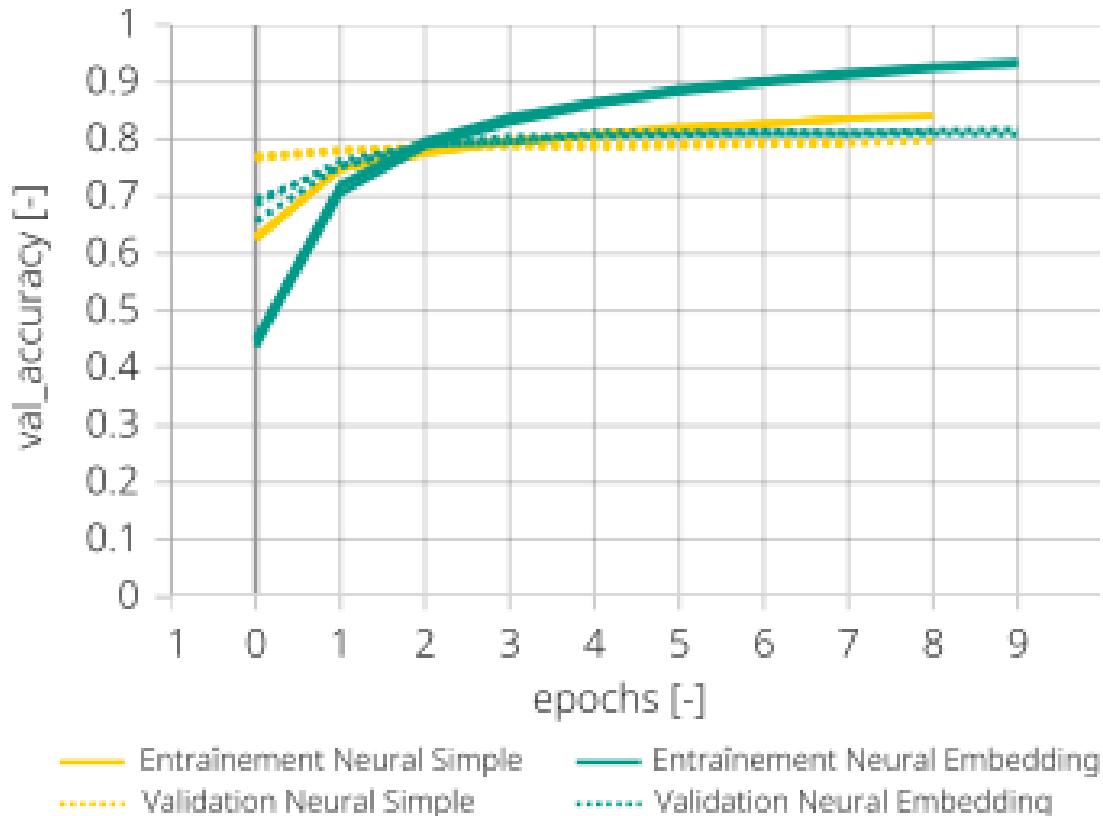
- Catégorie 10 (Livre d'occasion) souvent confondue avec 2705 (Livre neuf) et 2403 (Revue)
- Catégorie 40 (Jeu console) souvent confondue avec 10 (Livre occasion) et 2462 (Jeu oldschool)
- Catégorie 1280 (Déguisement) souvent confondue avec 1281 (Boîte de jeu) et 1140 (Figurine)

	logistic_regression	random_forest	kneighbours	decision_tree
10	0.449	0.472	0.260	0.376
40	0.543	0.595	0.320	0.499
50	0.730	0.770	0.518	0.618
60	0.875	0.891	0.814	0.823
1140	0.670	0.685	0.530	0.586
1160	0.877	0.865	0.731	0.836
1180	0.391	0.475	0.381	0.420
1280	0.638	0.603	0.449	0.534
1281	0.507	0.482	0.322	0.410
1300	0.899	0.866	0.729	0.890
1301	0.880	0.849	0.847	0.793
1302	0.754	0.758	0.603	0.671
1320	0.699	0.674	0.579	0.576
1560	0.793	0.746	0.636	0.660
1920	0.893	0.905	0.856	0.828
1940	0.821	0.791	0.648	0.669
2060	0.731	0.739	0.620	0.664
2220	0.744	0.751	0.488	0.632
2280	0.790	0.812	0.577	0.764
2403	0.720	0.734	0.604	0.671
2462	0.722	0.768	0.604	0.709
2522	0.891	0.846	0.763	0.755
2582	0.697	0.671	0.525	0.571
2583	0.964	0.926	0.900	0.919
2585	0.728	0.676	0.522	0.540
2705	0.643	0.644	0.321	0.545
2905	0.950	0.977	0.045	0.963
weighted F1-Score	0.771	0.760	0.619	0.695

Classifier	Acc.	Precision weighted	Recall weighted	F1 weighted
LogReg	0.18	0.16	0.18	0.16
RF	0.12	0.04	0.12	0.04
KNN	0.18	0.16	0.18	0.16
SVC	0.18	0.17	0.18	0.17
GradBoost	0.09	0.08	0.09	0.06

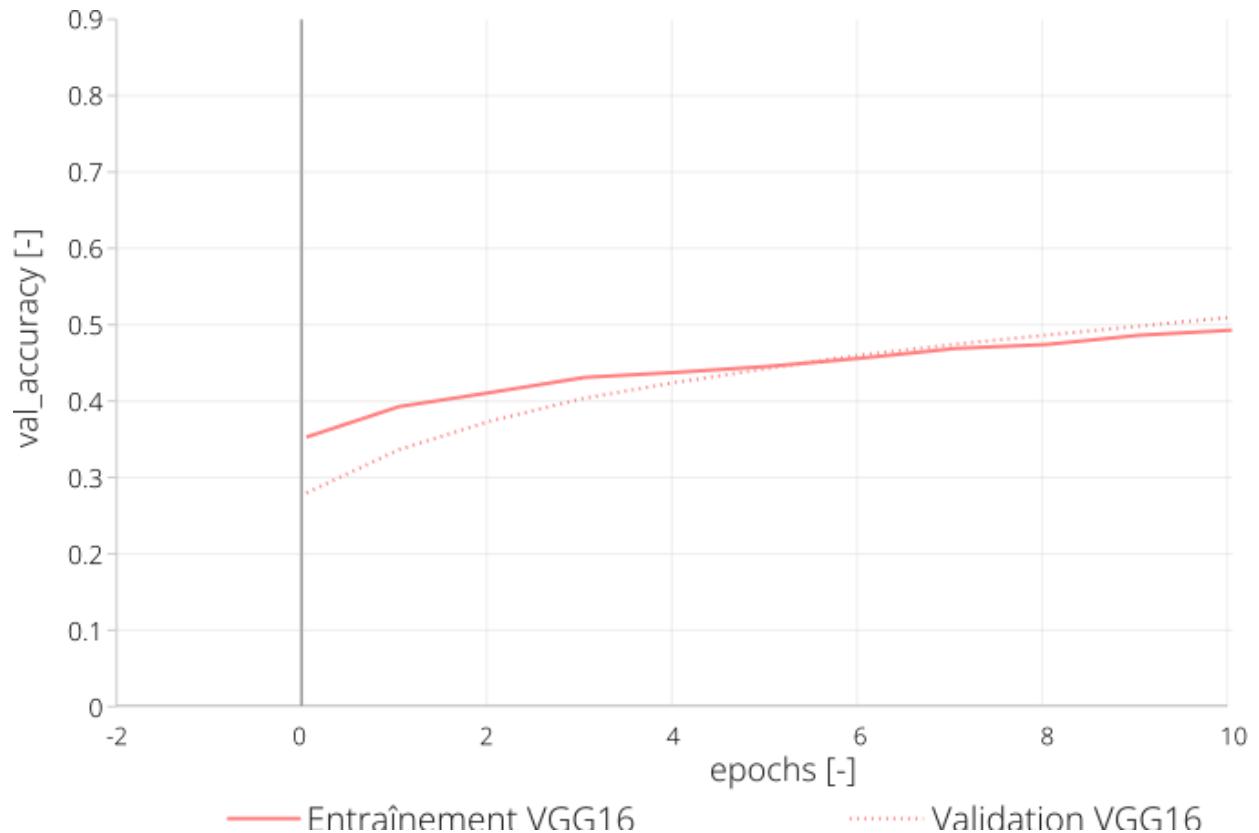


Les modèles / Deep learning / Text



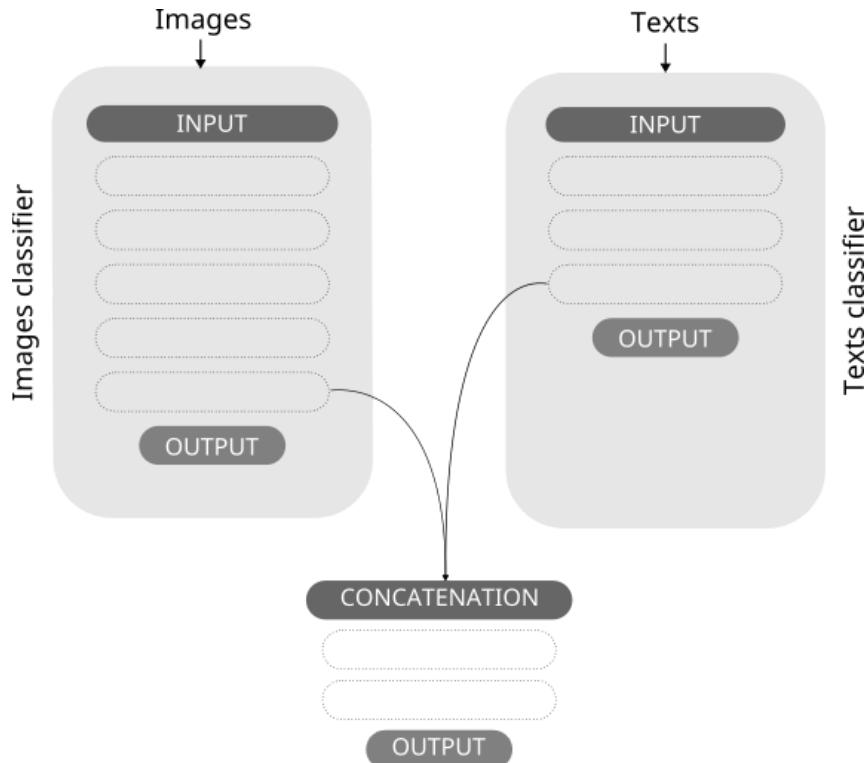
	neural_simple	neural_embedding
10	0.486	0.535
40	0.575	0.661
50	0.795	0.822
60	0.872	0.878
1140	0.707	0.731
1160	0.892	0.938
1180	0.445	0.577
1280	0.648	0.708
1281	0.549	0.582
1300	0.906	0.951
1301	0.888	0.957
1302	0.794	0.816
1320	0.729	0.783
1560	0.803	0.819
1920	0.899	0.908
1940	0.815	0.900
2060	0.764	0.785
2220	0.701	0.809
2280	0.829	0.818
2403	0.737	0.750
2462	0.735	0.764
2522	0.910	0.928
2582	0.730	0.742
2583	0.964	0.974
2585	0.749	0.795
2705	0.678	0.691
2905	0.957	0.941
weighted F1-Score	0.790	0.819

Model	Accuracy	Val accuracy
VGG16	0.50	0.49
ResNet	0.16	0.18
MobileNet	0.87	0.47



Les modèles / Deep learning / Fusion

	neural_simple	neural_embedding
10	0.486	0.535
40	0.575	0.661
50	0.795	0.822
60	0.872	0.878
1140	0.707	0.731
1160	0.892	0.938
1180	0.445	0.577
1280	0.648	0.708
1281	0.549	0.582
1300	0.906	0.951
1301	0.888	0.957
1302	0.794	0.816
1320	0.729	0.783
1560	0.803	0.819
1920	0.899	0.908
1940	0.815	0.900
2060	0.764	0.785
2220	0.701	0.809
2280	0.829	0.818
2403	0.737	0.750
2462	0.735	0.764
2522	0.910	0.928
2582	0.730	0.742
2583	0.964	0.974
2585	0.749	0.795
2705	0.678	0.691
2905	0.957	0.941
weighted F1-Score	0.790	0.819



	fusion_concat_embedding_50
10	0.624
40	0.702
50	0.824
60	0.898
1140	0.741
1160	0.941
1180	0.545
1280	0.680
1281	0.566
1300	0.949
1301	0.943
1302	0.812
1320	0.817
1560	0.820
1920	0.896
1940	0.891
2060	0.781
2220	0.827
2280	0.832
2403	0.783
2462	0.806
2522	0.923
2582	0.735
2583	0.971
2585	0.771
2705	0.805
2905	0.925
accuracy	0.827
macro avg	0.808
weighted avg	0.827

Analyse du meilleur modèle

Pas d'impacte sur les performances réduites du modèle d'image.

- Toutes les catégories dépassent le score de 54% et
- Une catégorie sur trois dépasse le score de 90%

Le modèle concaténé s'aide du modèle d'image pour catégoriser les produits où le modèle de texte sous-performait :

- La catégorie 1080 (Jeu Plateau) gagne 25 points
- La catégorie 2705 (Livre neuf) gagne 23 points

réalité	10	40	50	60	1140	1160	1180	1280	1281	1300	1301	1302	1320	1560	1920	1940	2060	2220	2280	2403	2462	2522	2582	2583	2585	2705	2905
10	0.60	0.04	0.00	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.08	0.00	0.00	0.00	0.00	0.00	0.11	0.00	
40	0.06	0.72	0.04	0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.01	0.00	
50	0.00	0.03	0.84	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.01	0.00	0.00	
60	0.00	0.01	0.02	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.00	
1140	0.02	0.03	0.00	0.00	0.79	0.01	0.04	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.01	0.00	
1160	0.00	0.01	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1180	0.02	0.02	0.01	0.00	0.05	0.00	0.76	0.01	0.03	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.02	0.01	
1280	0.01	0.02	0.01	0.00	0.09	0.01	0.01	0.62	0.11	0.05	0.00	0.02	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	
1281	0.03	0.04	0.00	0.00	0.02	0.03	0.05	0.18	0.54	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.02	
1300	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1301	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1302	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.03	0.00	0.00	0.77	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.00	0.00	
1320	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.01	0.00	0.00	0.01	0.81	0.02	0.02	0.01	0.02	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.00	
1560	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.04	0.00	0.04	0.00	0.00	0.00	0.01	0.03	0.00	0.03	0.00	0.00	
1920	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.94	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	
1940	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.04	0.00	0.01	0.86	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	
2060	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.02	0.05	0.06	0.00	0.73	0.00	0.00	0.01	0.00	0.01	0.02	0.00	0.03	0.00	0.00	
2220	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.03	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00	
2280	0.03	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.06	0.00	0.00	0.00	0.00	0.00	0.02	
2403	0.06	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.76	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.00	
2462	0.00	0.08	0.04	0.02	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2522	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.91	0.00	0.00	0.02	0.00	0.00	0.00	
2582	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.02	0.00	0.05	0.01	0.00	0.00	0.01	0.71	0.02	0.08	0.00	0.00	
2583	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.97	0.01	0.00	0.00	0.00	
2585	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.02	0.83	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	
2705	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
2905	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.95	

Limites

Le projet est un projet mêlant de l'analyse de texte et du traitement d'images : des notions poussées de deep-learning sont nécessaires à la compréhension et l'implémentation de telles techniques.

De nombreux limites sont apparus tout au long de ce projet :

- L'accès à des ressources de calcul de type GPU ou TPU nous a été quasi impossible, notamment via Google Collab.
- L'accès aux 84 916 images, stockées dans un Google Drive et nécessaires à l'entraînement du modèle d'images, était érattique : de nombreuses coupures de ce lien entre Google Drive et Google Collab ont entraîné ici aussi une grande perte de temps et une grande frustration.
- Le traitement des 84916 images nécessite d'utilisation de générateurs. Ces derniers sont à customiser manuellement afin de permettre une gestion en batch des données textuelles et d'images pour le modèle de fusion.
- La création d'un modèle de fusion a été une tâche ardue, principalement pour la gestion des entrées sous forme de générateurs.

“

Nous continuons de croire que le monde numérique a le potentiel d'améliorer la vie de chacun d'entre nous. Oubliez la peur. Adoptez l'optimisme.

Hiroshi Mikitani – Fondateur et CEO de Rakuten



Perspectives

- Ajout d'autres modèles au modèle de fusion.
- Ajout d'autres modèles au modèle de fusion.
- Uniformisation des données dans le code.
Actuellement, des dataframes Pandas, des tableaux Numpy, des générateurs d'images fonctionnent ensemble. Tout pourrait être géré autour d'un seul type de données, comme les `tf.data.Dataset`.
- Changement de la couche d'embedding ou création d'un modèle parallèle. Le modèle de texte par exemple pourrait être doté d'une couche d'embedding pré-entraînée, par exemple celle issue de CamemBERT.





Le projet **Rakuten** a été très intéressant, car complexe et faisant appel à des notions avancées mêlant le traitement de textes et le traitement d'images.

L'exploration de données, le travail de groupe, les différentes implémentations et sprints ont fait de ce projet un projet répondant, nous l'espérons, aux besoins d'une entreprise.