AMÉLIORER LE PRODUIT IA* DE VOTRE STARTUP

Voahangy Joan ALEONARD – 01/04/2021

*Intelligence Artificielle

AGENDA DU JOUR

SCOPE PROJET ET DONNÉES

AVIS CLIENT ET INSATISFACTION

CLASSIFICATION D'IMAGES

COLLECTE DE DONNÉES VIA API

SCOPE PROJET ET DONNÉES

PROBLÉMATIQUE MÉTIER

Avis Restau

Plateforme en ligne de connexion entre restaurants et clients – les clients pouvant publier des avis et des photos.

Mission IA

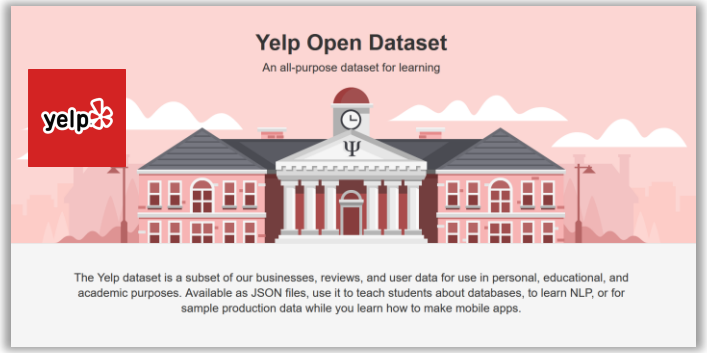
Etudier la faisabilité de:

- **Détecter les motifs d'insatisfaction ;**
- **Labelliser automatiquement les photos (5 classes) ;**
- **Collecter de nouvelles données.**

DISPONIBILITÉ DES DONNÉES

Pas assez de données sur la plateforme Avis Restau → **Utilisation d'un jeu de données existant**

https://www.yelp.com/dataset



200.000 photos
1 fichier JSON

8,6 millions d'avis
2 fichiers JSON

Outil de requêtage (API)

DESCRIPTION DES DONNÉES

68.000 avis
Filtre sur la catégorie ‘restaurants’

20.000 photos
16.000 Train
2000 Validation
2000 Test

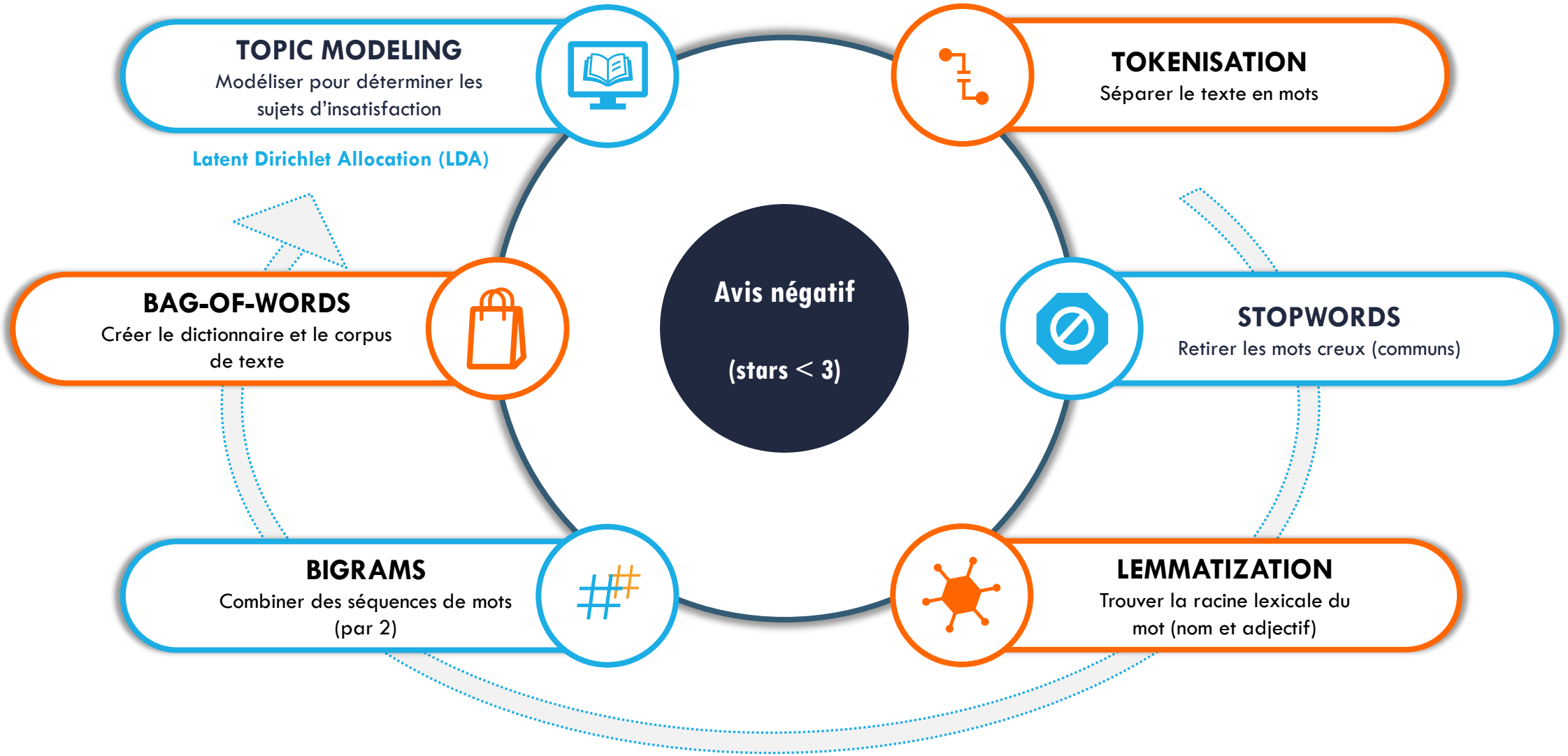
200 restaurants parisiens
3 avis par restaurant

"I've been there several times. I've experienced the following at each visit:\n1. Stale baked goods,\n2. Expensive, mediocre coffee\n3. Indifferent service, approaching surly."

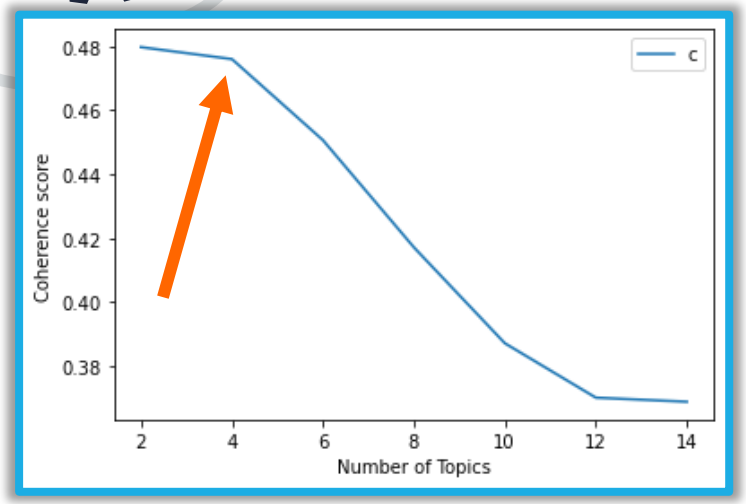
"Horrible- Unclean, rancid, foul odors, unkempt, and poor service. The restaurant was nearly empty, and the cashier didn't even look up to say hello as my husband stood hungrily grinning at the counter. Needless to say, I talked him into driving up the road to the other location. This one is unacceptable."

AVIS CLIENT ET INSATISFACTION

DÉMARCHE DE TRAITEMENT DES AVIS CLIENT



LATENT DIRICHLET ALLOCATION (LDA)



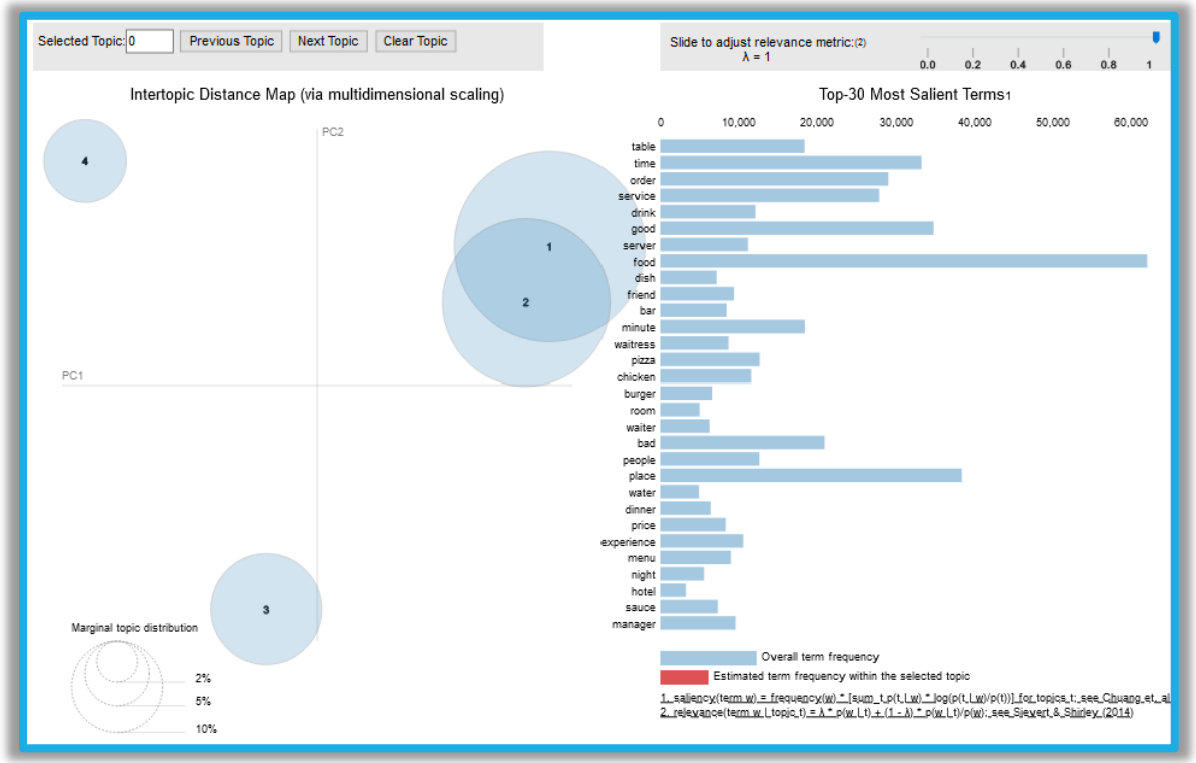
Algorithme d'apprentissage non supervisé qui tente de découvrir la proportion de rubriques partagées par des documents au sein d'un corpus de texte

Nombre de rubriques devant être spécifié par l'utilisateur

Rubriques apprises par le modèle sous la forme d'une distribution de probabilité sur les mots rencontrés

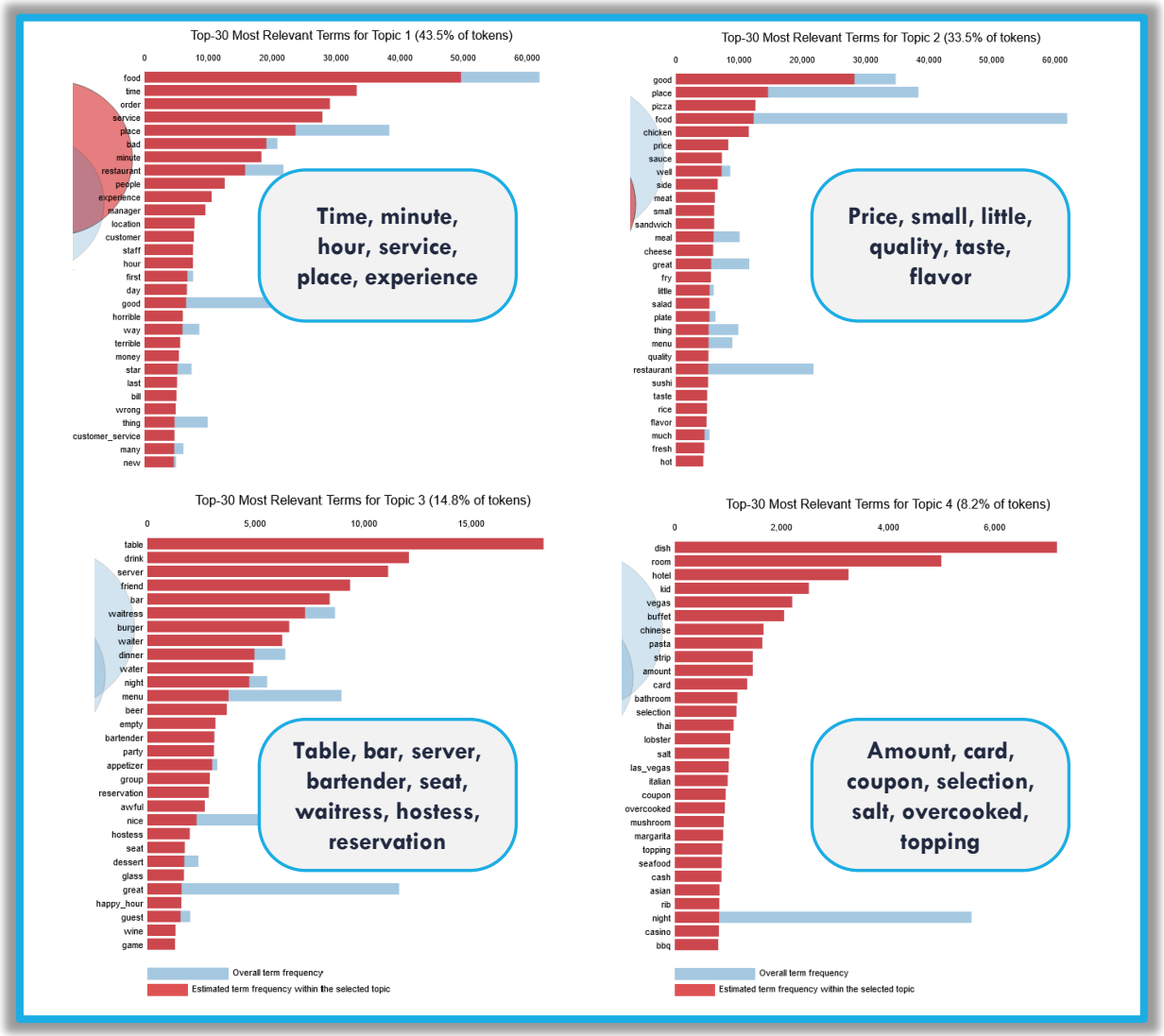
Forme de réduction de la dimensionnalité, par la réduction de la taille du vocabulaire en un nombre k de rubriques spécifié par l'utilisateur

EXTRACTION DES TOPICS



4 motifs d'insatisfaction:

- Délai dans le service ;
- Rapport qualité/prix médiocre, absence de goût;
- Incompétence du personnel
- Inadéquation entre les promotions et la qualité.



CLASSIFICATION D'IMAGES

DÉMARCHE DE TRAITEMENT DES IMAGES

Labellisation des photos en 5 catégories:
food, drink, menu, inside, outside

1 dossier par set

Création des Train/ Test/ Validation sets

1 dossier par label dans chaque set

Structure de stockage des images

Data Augmentation

Keras:
ImageDataGenerator

Transfer Learning

Modèle pré-entraîné:
VGG-16

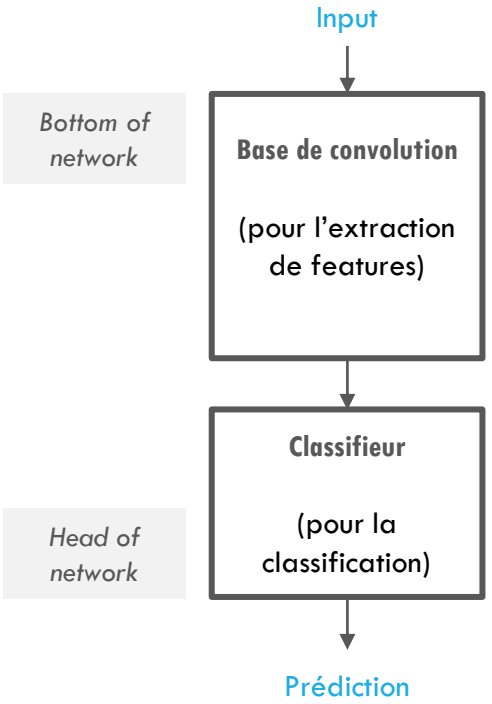
Feature Extraction

Fine-Tuning partiel

Analyse de performance et visualisation des erreurs

APPRENTISSAGE PAR TRANSFERT (*TRANSFER LEARNING*)

Architecture simplifié d'un réseau de neurones convolutionnel (CNN)

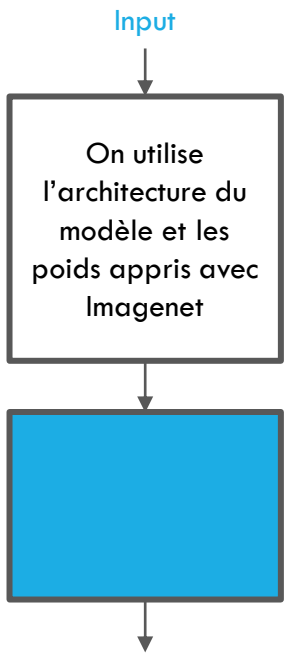


Stratégies de transfer learning

Feature extraction

Stratégie 1

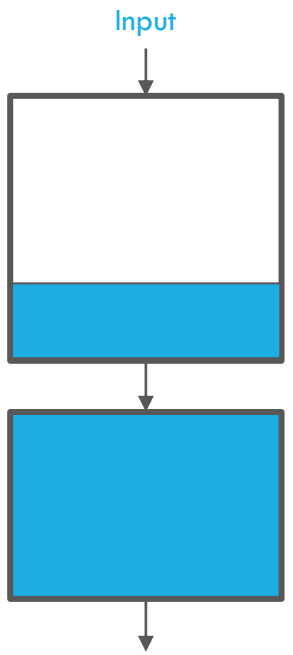
Figer la base convolutionnelle



Fine-tuning (partiel)

Stratégie 2

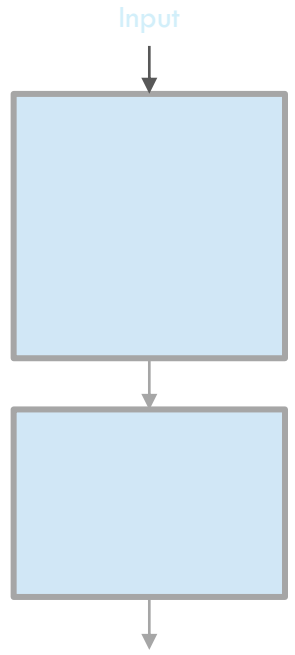
Entraîner certains couches de la base

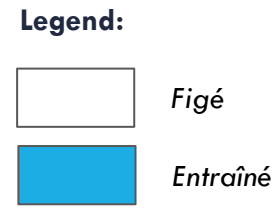


Fine-tuning (total)

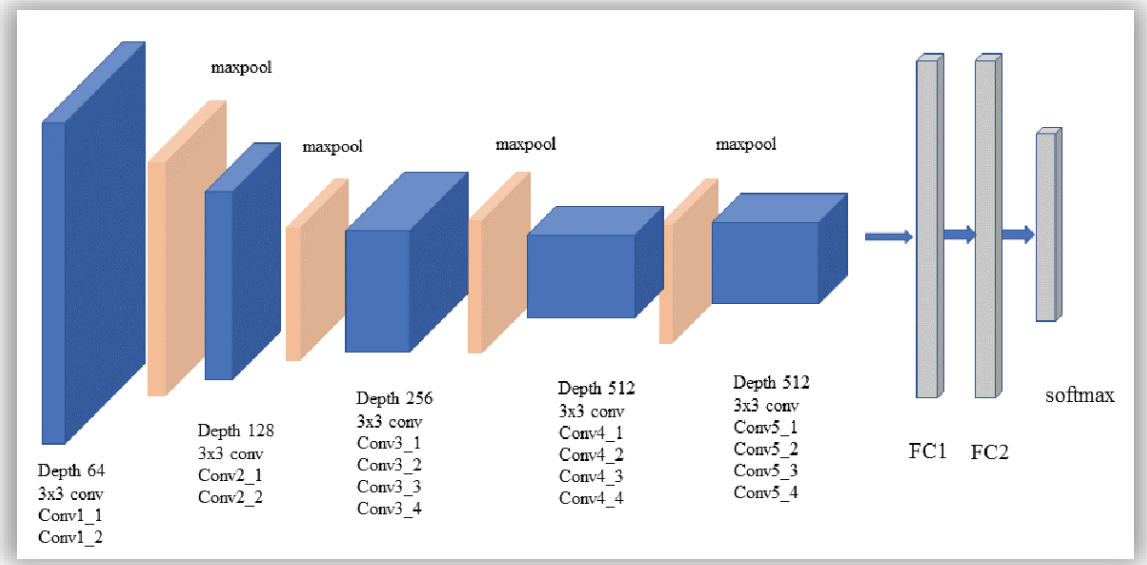
Stratégie 3

Entraîner le modèle entier





MODÈLE PRÉ-ENTRAINÉ : VGG-16



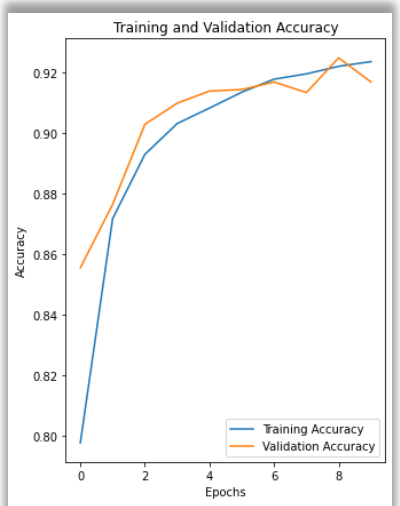
Réseau de neurones convolutionnel (CNN) composé de 16 couches d'apprentissages

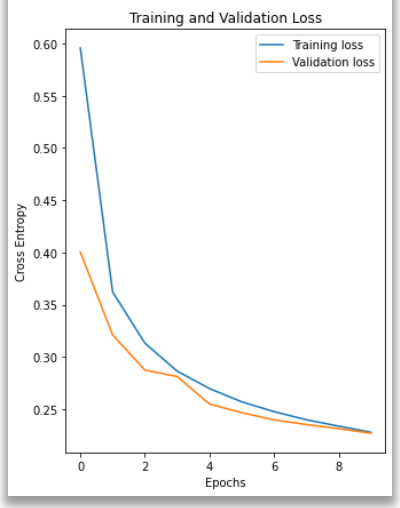
Près de 139 millions de paramètres entraînables

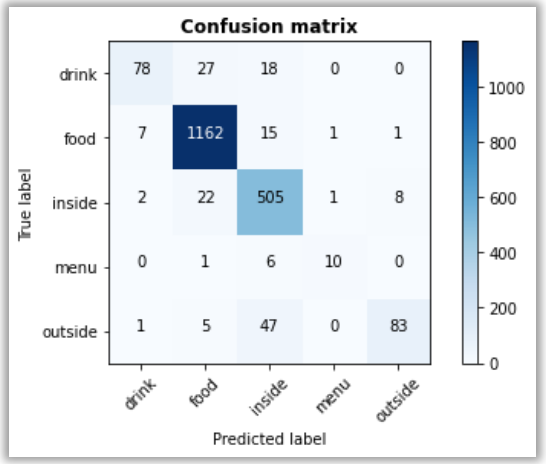
Initié par Oxford group et gagnant de la compétition ImageNet 2014 (*ILSVRC : ImageNet Large Scale Visual Recognition Challenge)*

LE FINE-TUNING DONNE LE MEILLEUR RÉSULTAT

Feature Learning



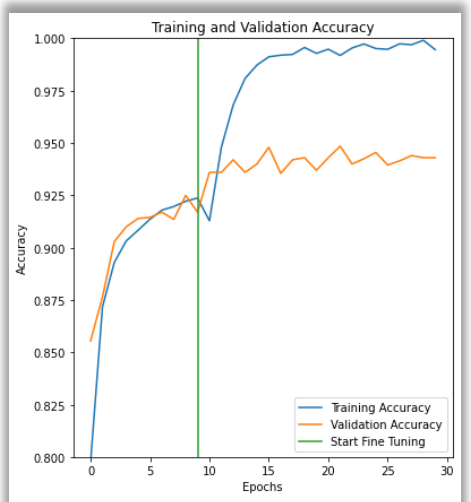


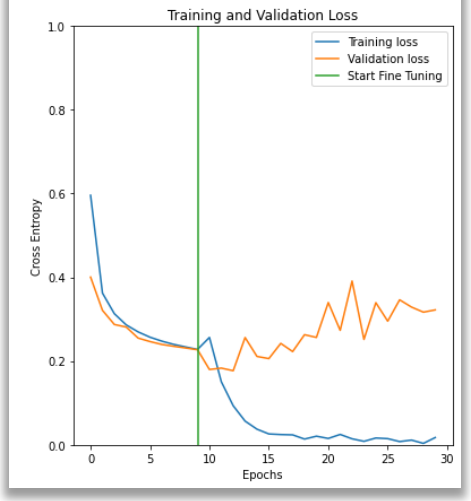


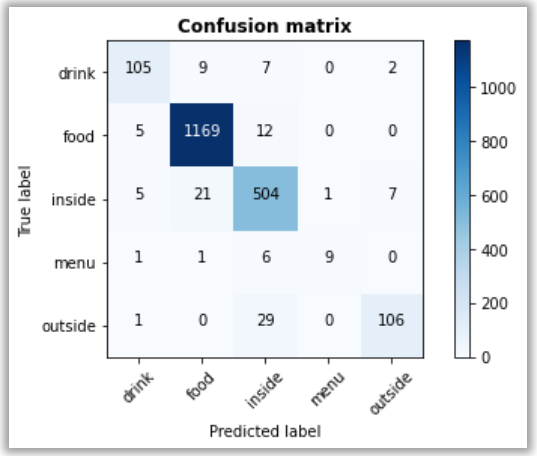
	precision	recall	f1-score	support
0	0.89	0.63	0.74	123
1	0.95	0.98	0.97	1186
2	0.85	0.94	0.89	538
3	0.83	0.59	0.69	17
4	0.90	0.61	0.73	136
accuracy			0.92	2000
macro avg	0.89	0.75	0.80	2000
weighted avg	0.92	0.92	0.91	2000

	vggFE
Loss	**0.248**
Accuracy	**0.919**

Fine Tuning



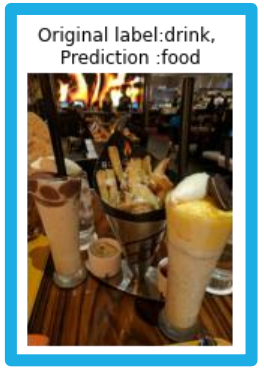




	precision	recall	f1-score	support
0	0.90	0.85	0.88	123
1	0.97	0.99	0.98	1186
2	0.90	0.94	0.92	538
3	0.90	0.53	0.67	17
4	0.92	0.78	0.84	136
accuracy			0.95	2000
macro avg	0.92	0.82	0.86	2000
weighted avg	0.95	0.95	0.95	2000

	vggFT
Loss	**0.396**
Accuracy	**0.946**

EXEMPLE D'ERREURS DE CLASSIFICATION



Vraie erreur de classification

Choix entre plusieurs labels

Erreur de label à l'origine ?

COLLECTE DE DONNÉES VIA API

API YELP : COLLECTE DE DONNÉES

Création compte
(avec clé API)

Filtre sur
Restaurants, Paris

Multiples requêtes
(résultats limités à 50)

Sauvegarde
fichier CSV

SYNTHÈSE

VISUALISATION DANS UNE PAGE WEB

voilà

Author : V.Joan Aléonard
Last update : 30 March 2021

Avis Restau

AVIS RESTAU

Avis Restau is a website which connects clients and restaurants:

- The clients can publish reviews and photos of the restaurants;
- The restaurants can leverage from customer's feedbacks for improvement.

As Data scientist for Avis Restau, our mission is summarized as follows:

Task	Objective	Available notebook
Analyze customers' reviews	**Detect customer's insatisfaction topics**	Notebook N°1
Treat photo posted by customers	**Classify automatically customer's photos**	Notebook N°2
Collect new data throuhg API	**Enrich our database**	Notebook N°3

This **web page** is dedicated the presentation of the **project feasibility study**.

RÉCAPITULATIF

Mission IA

Etudier la faisabilité de:

- **Détecter les motifs d'insatisfaction ;**
- **Labelliser automatiquement les photos (5 classes) ;**
- **Collecter de nouvelles données.**

~4 motifs d'insatisfaction détectés

95% des photos correctement classifiées

Données collectées sur 200 restaurants

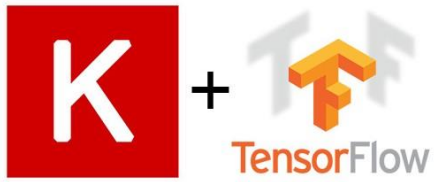
CONCLUSION

	RESULTATS	AXES D'AMELIORATION
TEXTES	› Interprétation pas toujours évidente quant à l'analyse des motifs d'insatisfaction › Filtre du dataset sur la catégorie 'restaurants' inclus également les hôtels	› Filtre sur les avis en anglais uniquement (exclure le texte en français) › Analyse des avis positifs pour améliorer la compréhension du corpus de la restauration et éventuellement, étendre les stopwords
PHOTOS	› Bonne précision sur la classification d'images grâce au transfer learning, fondé sur du fine-tuning, et peu de pré-processing › Un projet mené à bien malgré le manque de ressources (puissance de calcul)	› Des labellisations initiales pouvant être erronées, du moins sujettes à interprétation › Test sur d'autres modèles (ResNet), complexifier le modèle (ajout de layers) et/ou équilibrer les classes pour voir si on peut encore améliorer le modèle

QUESTIONS / RÉPONSES

ANNEXES

KERAS ET SES MODÈLES PRÉ-ENTRAINÉS



Keras Applications sont des modèles de Deep Learning disponibles avec des poids (*weights*) pré-entraînés.

Ces modèles peuvent être utilisés pour la prédiction, l'extraction de features et le fine-tuning.

Les poids sont téléchargés automatiquement lors de l'instanciation d'un modèle.

Available models

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-
EfficientNetB0	29 MB	-	-	5,330,571	-
EfficientNetB1	31 MB	-	-	7,856,239	-
EfficientNetB2	36 MB	-	-	9,177,569	-
EfficientNetB3	48 MB	-	-	12,320,535	-
EfficientNetB4	75 MB	-	-	19,466,823	-
EfficientNetB5	118 MB	-	-	30,562,527	-
EfficientNetB6	166 MB	-	-	43,265,143	-
EfficientNetB7	256 MB	-	-	66,658,687	-

RÉFÉRENCES

- Yelp : dataset, yelp pour les développeurs, Get Yelp API key
- NLP : Alexis Perrier, lemmatization, analyticsvidhya.com - topic modeling, machinelearningplus.com - topic modeling
- NLTK : documentation Spacy : documentation, Gensim : documentation, cohérence des topics, LDA modeling
- Réseaux de neurones convolutionnels (CNN) : wikipedia, Stanford-edu/shervine, datascientest.com,
- Keras : Image classification from scratch, transfer learning, modèles pré-entrainés
- Tensorflow : classification d'images
- Mapper les labels des images à leurs classes : https://deeplizard.com/learn/video/pZoy_j3YsQg

Ce document a été produit dans le cadre de la soutenance du projet n°6 du parcours Ingénieur IA d'OpenClassrooms :
« Améliorez le produit IA de votre start-up »

Mentor : Thierno DIOP
Evaluateur : Bertrand BEAUFILS

OPENCLASSROOMS
En partenariat avec
CentraleSupélec Microsoft