De nouvelles fonctionnalités pour la start-up

Avis Restau

Contexte

Avis Restau

Avis Restau propose une solution de mise en relation des clients et des restaurants.

Cet outil permet à nos **utilisateurs** de trouver des établissements adaptés à leurs besoins et aux **restaurants** d'attirer de nouveaux clients.

Mais pour les aider et continuer sa croissance notre start-up aimerait **ajouter de nouvelles fonctionnalités de collaboration**.

Objectif du projet

Avis Restau

Suite à l'ajout de nouvelle fonctionnalités permettant aux utilisateurs de l'application de :

- **poster des avis** sous forme de commentaires
- **poster des photos** prises dans le restaurant

Il nous faut déployer deux solutions d'IA:

- **détecter les sujets d'insatisfaction** présents dans les commentaires postés
- **labelliser automatiquement les photos** postées sur la plateforme.

Modélisation thématique *(Topic Modelling)*

Avis Restau

COLLECTE

EXTRACTION DE FEATURES

Tokens = [available, capability, cea, challenge, chemistry, chimie, coating, commercial, ...]

Dictionary = {'available': 0, 'capability': 1, 'cea': 2, 'challenge': 3, 'chemistry': 4, 'chimie': 5, 'coating': 6, 'commercial': 7, ...

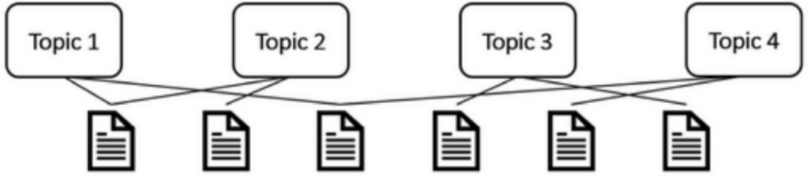
\+ Corpus (BoW)

MODELISATION

LDA / NMF / ...

Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | ...

INFERENCE



Modélisation thématique *(Topic Modelling)*

Avis Restau

Présentation du jeu de données

Pour ce problème, nous avons travaillé avec 2 fichiers source décrivant les reviews laissées et les commerces concernés.

- 150.346 commerces décrits par 14 variables
- 6.990.280 reviews décrites par 9 variables

Mais nous avons constatés qu'il y avait de nombreux types de commerces...

- 51.864 commerces restants
- 677.372 reviews sur 1 millions chargées
- 10.000 reviews choisies au hasard

Modélisation thématique *(Topic Modelling)*

Avis Restau

EDA - vérifications de base

Les vérifications des erreurs les plus fréquentes
(valeurs manquantes, doublons, outliers, erreurs de format, erreurs lexicales, contenus multiples)
puis les analyses univariées et multivariées
ont permis de déceler quelques problèmes qui
ont été corrigés avec les actions suivantes:

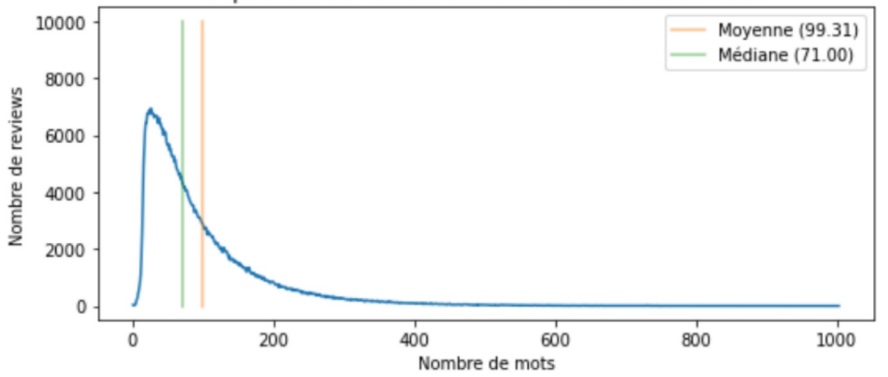
- Suppression des colonnes inutiles au projet
- Traitement des outliers
- Traitement du déséquilibre des notes

Modélisation thématique *(Topic Modelling)*

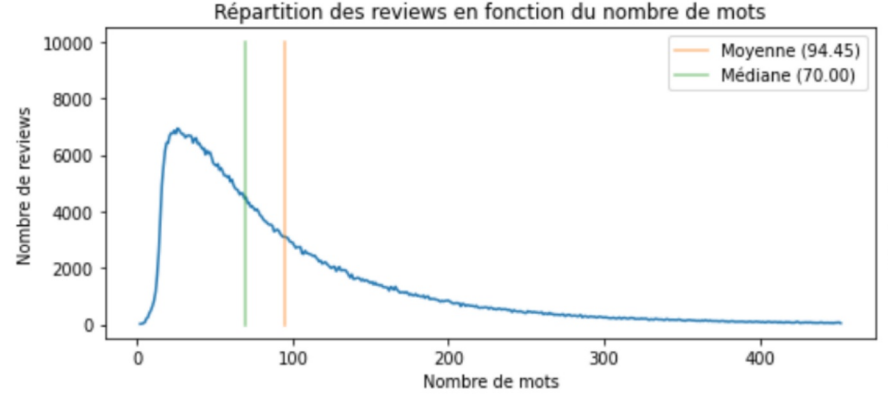
Avis Restau

EDA - Nombre de mots / reviews > outliers

AVANT



APRÈS

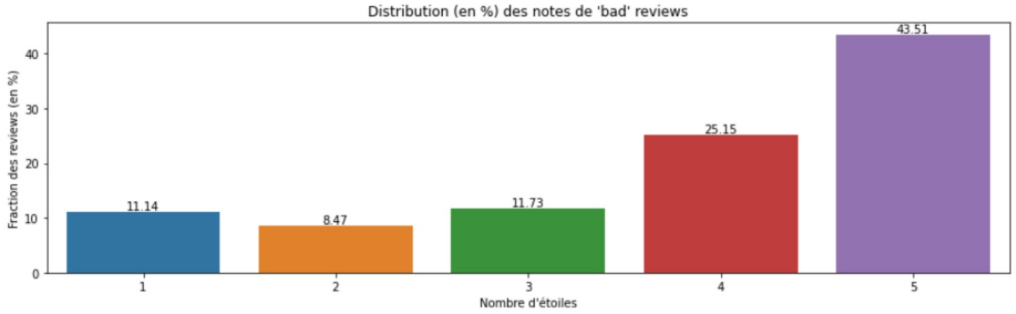


Modélisation thématique *(Topic Modelling)*

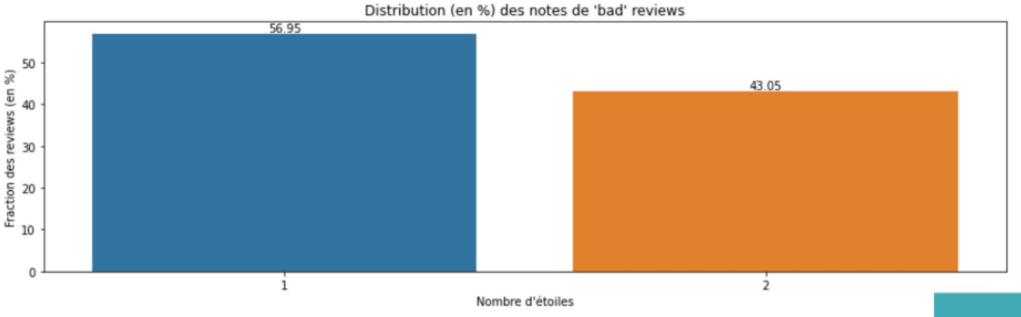
Avis Restau

EDA - Répartition des notes > binarisation

AVANT



APRÈS



Modélisation thématique *(Topic Modelling)*

Avis Restau

Pre-processing & mise en application > démo

Yelp API
+
Streamlit

Classification d'images

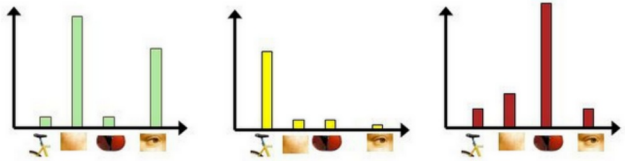
Avis Restau

COLLECTE

EXTRACTION DE FEATURES

Visual-words

Bag of VW



SIFT
ORB

python
numpy
pandas

MODELISATION

Image Classifier

k-means / Convolutional Neural Network / ...

k-means

CNN

INFÉRENCE

Classe 1 | Classe 2 | Classe 3 | Classe 4

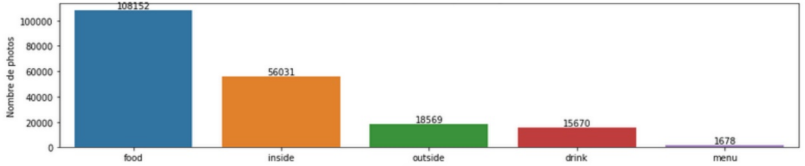
Classification d'images

Avis Restau

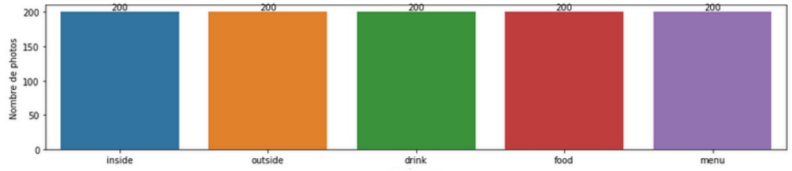
Présentation du jeu de données

Pour ce problème, nous avons travaillé avec un fichier source donnant des informations sur les photos collectées par les utilisateurs de Yelp.

- 200.100 photos associées à 4 variables
- 5 catégories



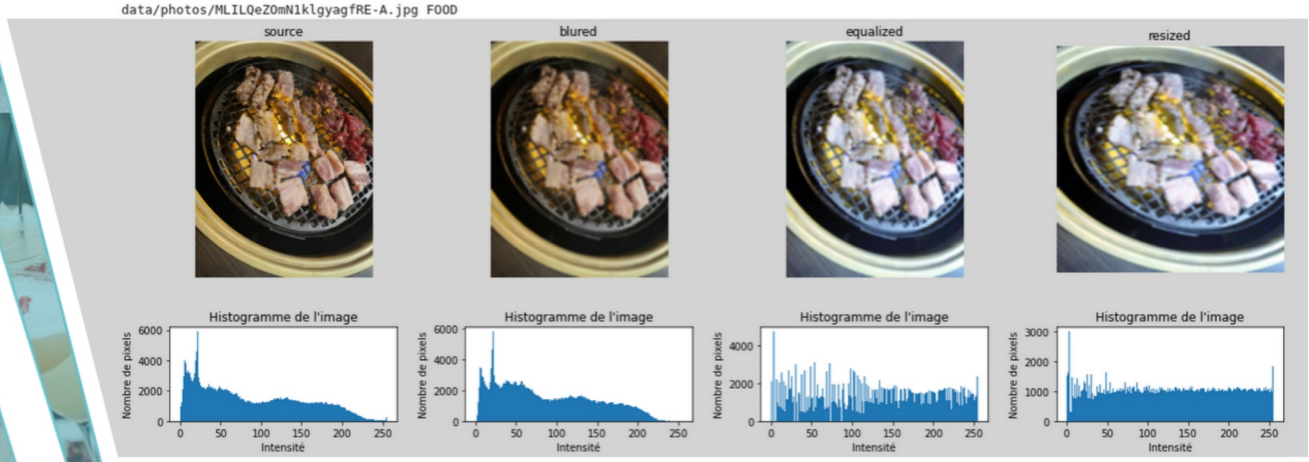
- 200 * 5 photos sélectionnées



Classification d'images

Avis Restau

Pré-traitement des images



(1) (2) (3)

1. suppression du bruit
2. égalisation de l'histogramme
3. redimentionnement

Classification d'images

Avis Restau

Feature extraction SIFT & CNN > démo

Axes d'amélioration

Avis Restau

- **Essayer avec plus de données**
  (pour voir si l'on peut améliorer le modèle de classification et jusqu'à quel point ou pour essayer de nouvelles distributions de topics)
- **Essayer d'autres algorithmes / architectures**
  (EfficientNet, ResNet pour les CNN, Latent Semantic Analysis pour les topics)
- **Essayer des tests statistiques**
  (Hopkins statistic, Distance distribution)
- **Surveiller les données collectées avec les nouvelles fonctionnalités**
  (pour éviter les problèmes de data-drift ou de concept-drift)

Merci de m'avoir écouté, évalué et conseillé.

Avis Restau