PROJET 6 - AMÉLIOREZ LE PRODUIT IA DE VOTRE START-UP

Xiaofan LEI

ORDRE DU JOUR

- Contexte
- Structuration des données textes brutes
  - Exploration, nettoyage, normalisation
  - Méthode d'extraction des termes potentiels avec TF-IDF
  - Classification des termes avec Kmeans
- Classification des données visuelles
  - Extraction des features, filtrage du bruits
  - Détection des features(SIFT)
  - classification (SVM)
- Collecte de nouvelles données via API

CONTEXTE DE LA MISSION

- Dans une démarche d'amélioration continue, ce projet a pour mission de proposer une nouvelle fonctionnalité de collaboration pour la plateforme Avis Restau.

- Le but de ce nouvel apport consiste à détecter de manière automatique :
  - les différents aspects d'insatisfaction présents dans les commentaires
  - et les catégories des photos
- Tout en s'assurant la possibilité de collecter de nouvelles données

DÉTECTER LES SUJETS D'INSATISFACTION

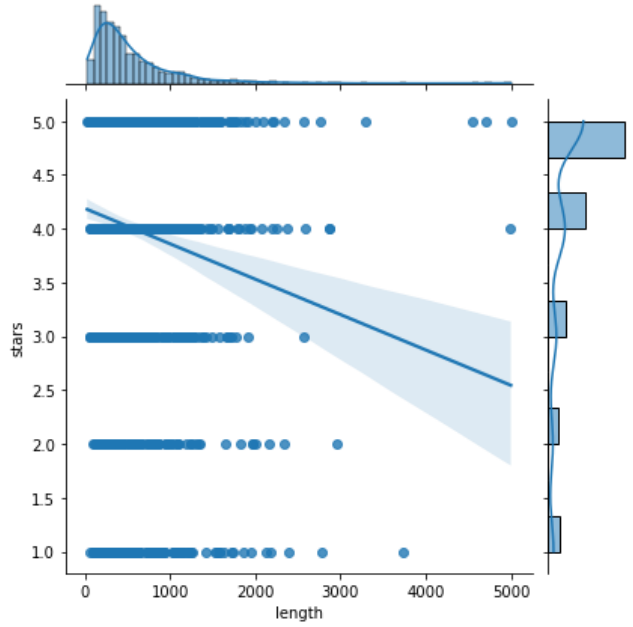
DESCRIPTION DU JEU DE DONNÉES TEXTUELLES

Jeu de données :

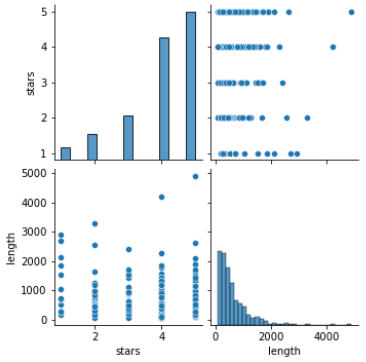
- Yelp reviews sur 1342 commentaires des 200 restaurants => moins c'est bon, plus on parle.

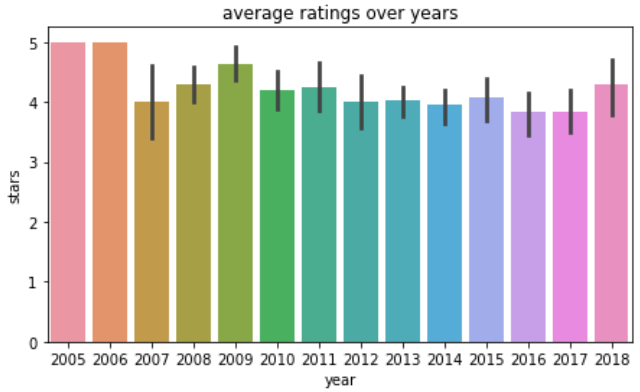
Cas d'étude : South Congress Cafe

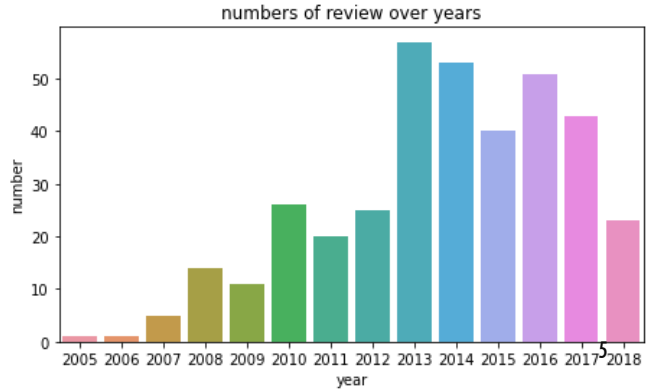
- Les clients sont relativement contents du restaurant : 87/370 commentaires sont négatifs (avec au plus 3 étoiles)
- Il semble gagner en popularité dans les débuts des années 2010.



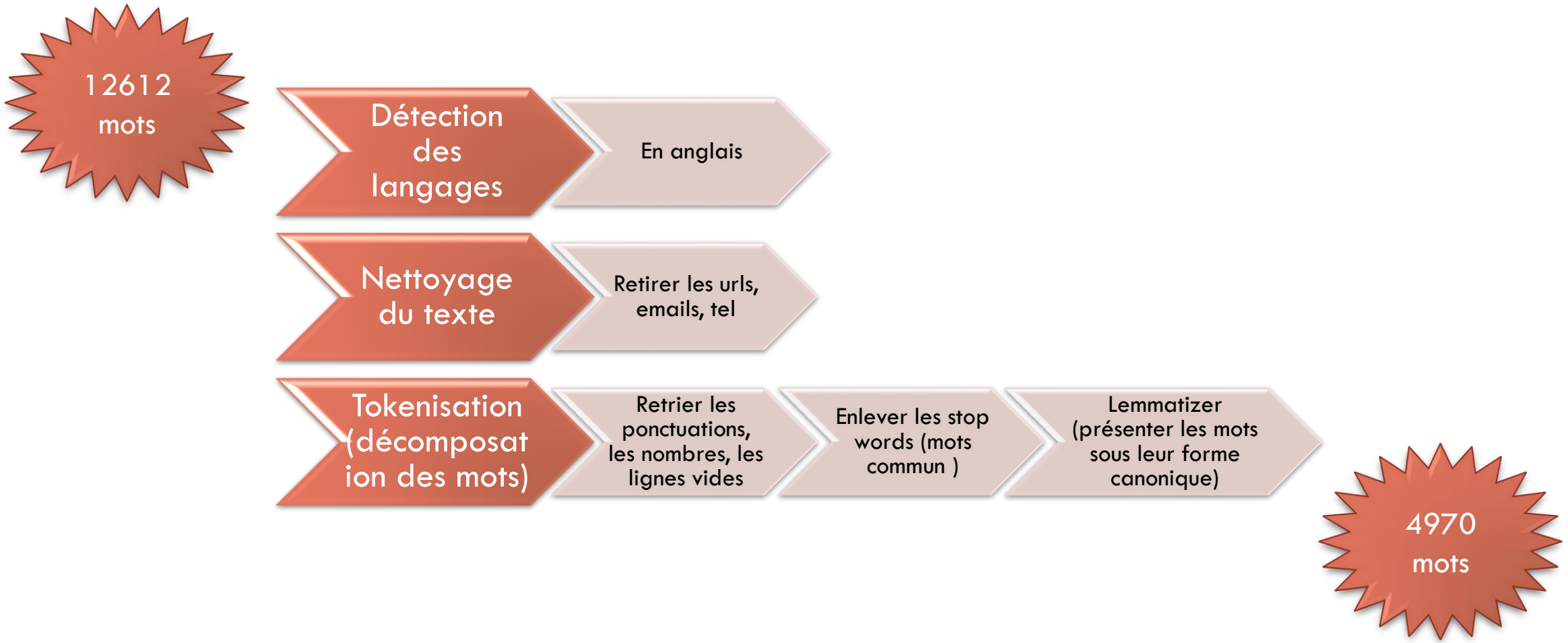
name	stars	number
South Congress Cafe	4.040541	370
Pelons Tex-Mex	3.562500	80
Breakfast At Valerie's	3.953846	65
Miller's Ale House - Orlando	3.190476	63



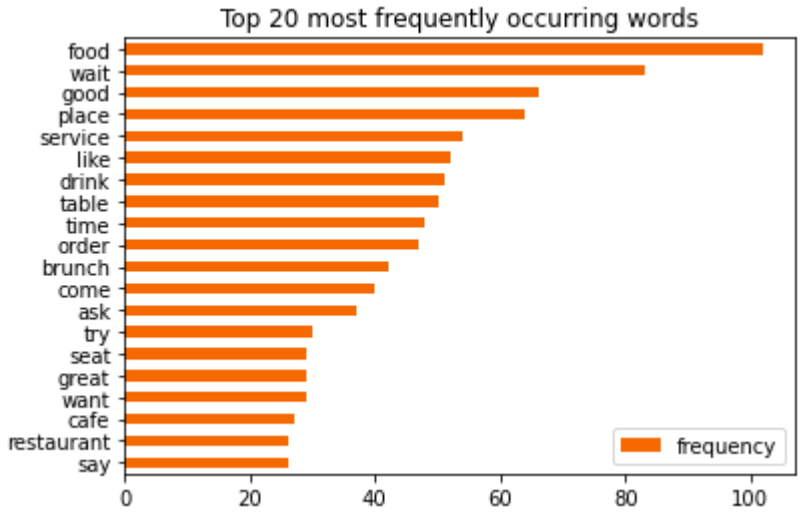




NETTOYAGE ET NORMALISATION



ANALYSE DES FRÉQUENCES



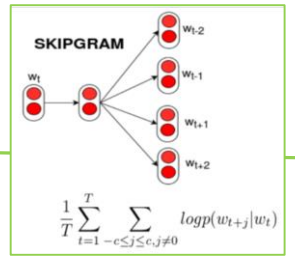
Créer une matrice sparse pour représenter les mots.

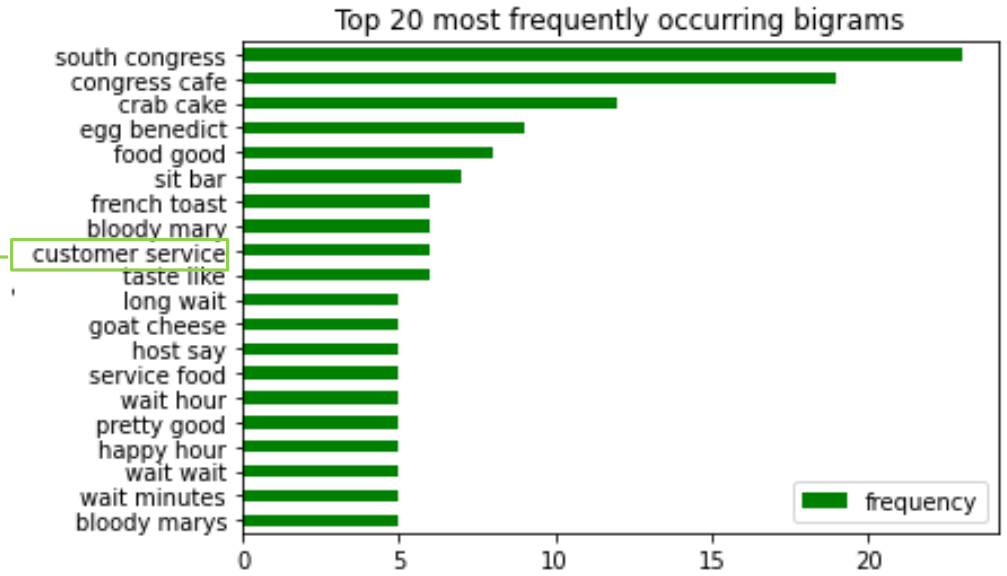
Supprimer les mots qui sont apparus dans 85% des commentaires

Calculer la fréquence des mots

VISUALISATION DES BIGRAMMES

```
[('food', 0.9466162323951721),
 ('wait', 0.9412321448326111),
 ('drink', 0.9395725131034851),
 ('like', 0.9343560338020325),
 ('take', 0.9325699806213379),
 ('cafe', 0.9304937124252319),
 ('ask', 0.9235355854034424),
 ('table', 0.9232072830200195),
 ('congress', 0.9217003583908081),
 ('good', 0.921075701713562)]
```





Word Embedding
(skip-gram)

cette architecture vise à prédire les mots du contexte étant donné un mot en entrée.

EXTRACTION DES BIGRAMMES/TRIGRAMMES AVEC TF-IDF

Term Frequency – Inverse Document Frequency est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes.

TF

IDF

Plus un terme apparaît très fréquemment dans quelques textes seulement, plus l'IDF est élevée.

TF permet de mesurer l'importance relative d'un mot dans un document

IDF : nombre d'occurrences du mot dans le document en fonction de la fréquence du mot dans le corpus.

TERMES CLASSIFICATION AVEC KMEANS

```
Topic 1 : overall experience,rave review,pretty good,food good,goat cheese
_____________________________________________________________________________
Topic 2 : south congress,congress cafe,south congress cafe,crab cake,want try
_____________________________________________________________________________
Topic 3 : bloody mary,egg benedict,wait minutes,crab cake,deep fry
_____________________________________________________________________________
Topic 4 : sit bar,food mediocre,crab cake,drink order,wait wait
_____________________________________________________________________________
Topic 5 : worth wait,weekend brunch,hour half,wait hour,wait wait
_____________________________________________________________________________
```

CONCLUSION SUR LA FAISABILITÉ

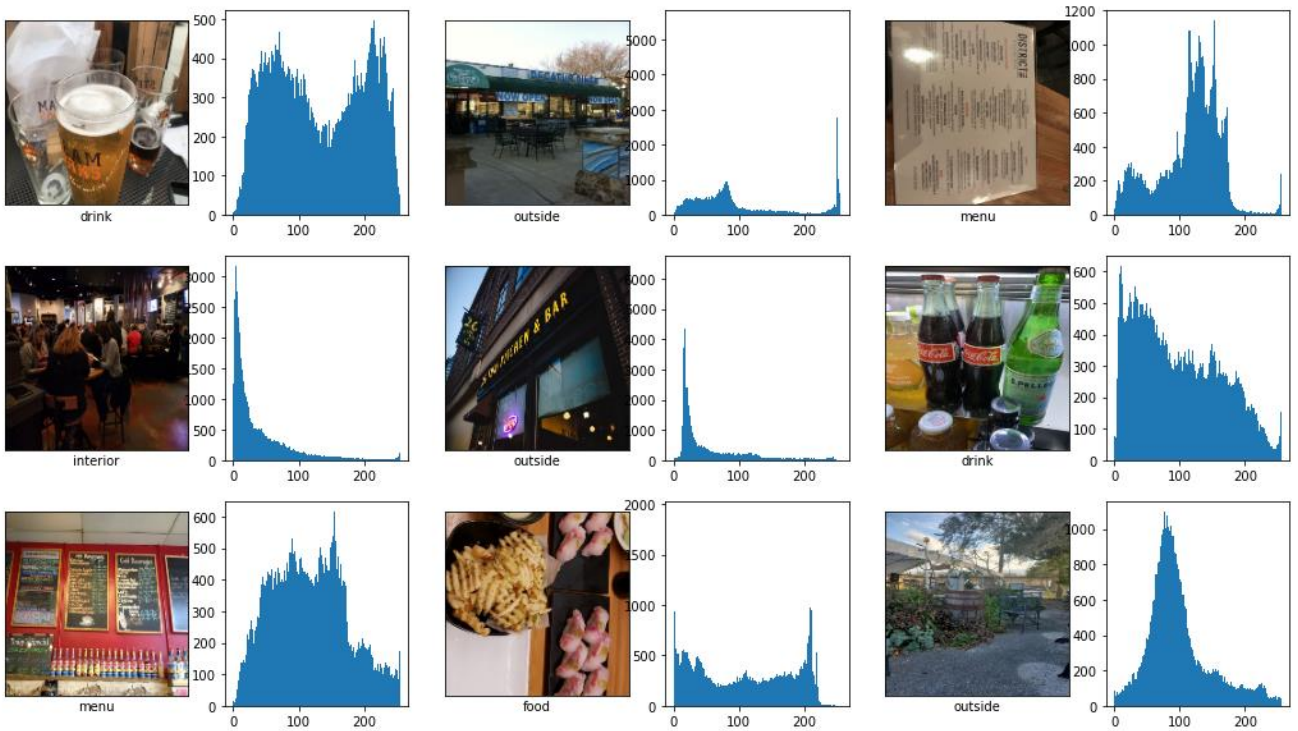
- En conclusion :
  - La méthode TF-IDF permet de ressortir les éléments sur les principaux motifs de plainte de South Congres Cafe.

LABELLISER LES PHOTOS POSTÉES

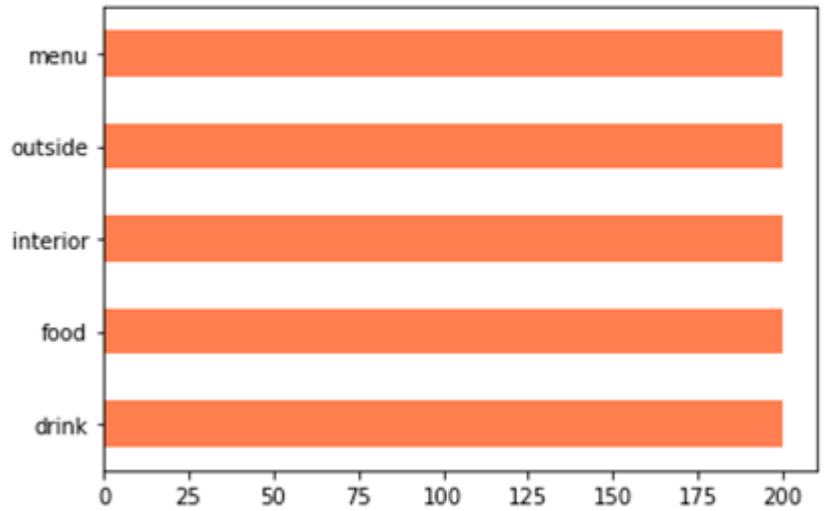
DESCRIPTION DU JEU DE DONNÉES VISUELLES

Jeu de données :

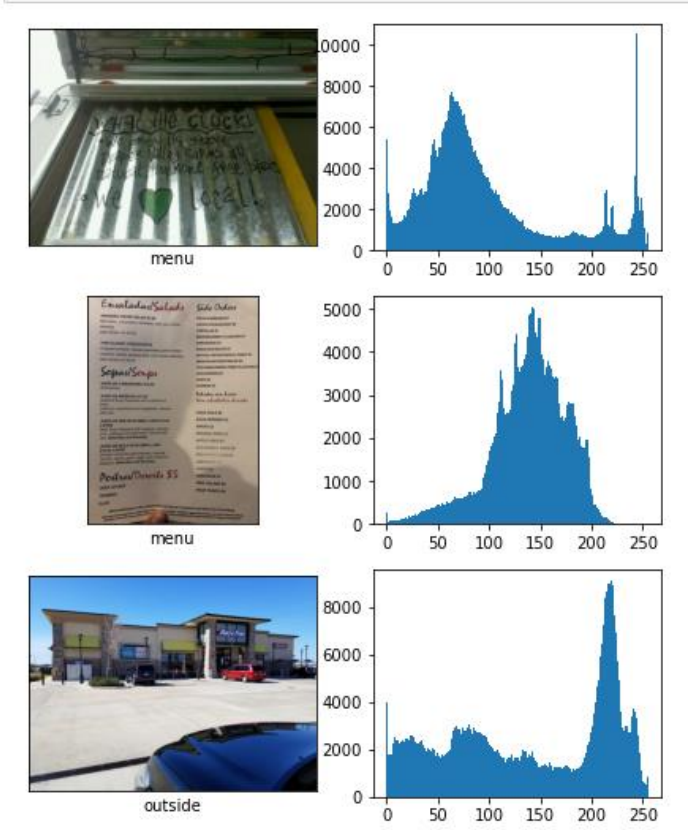
1000 photos de 5 catégories ont été sélectionnées pour ce test.




```
class_names: ['drink', 'food', 'interior', 'menu', 'outside']
total training photos: 1000
```



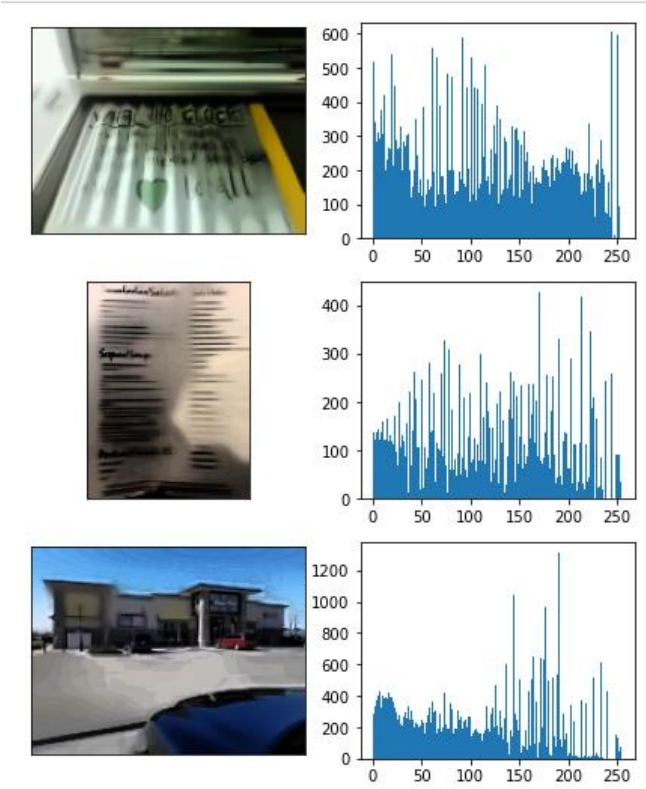
NORMALISATION D'IMAGE



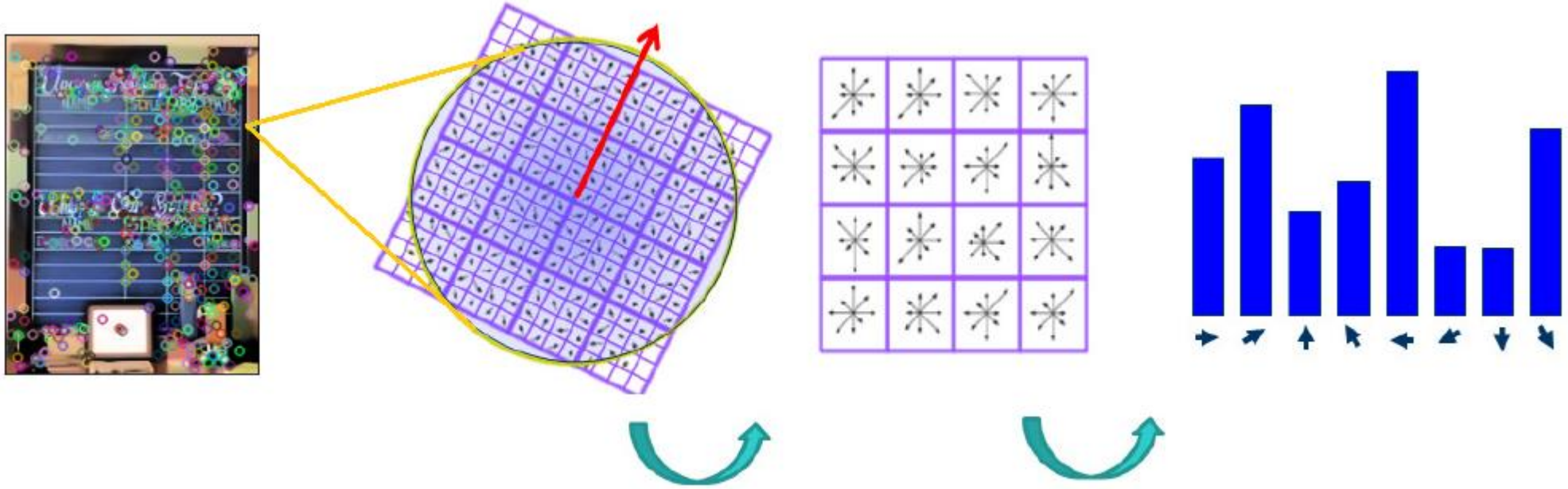
Réduire la taille de l'image

Filtrer les bruits par patchs (Non-local means)

Égaliser l'histogram me



DECTECTION DES FEATURES VIA SIFT



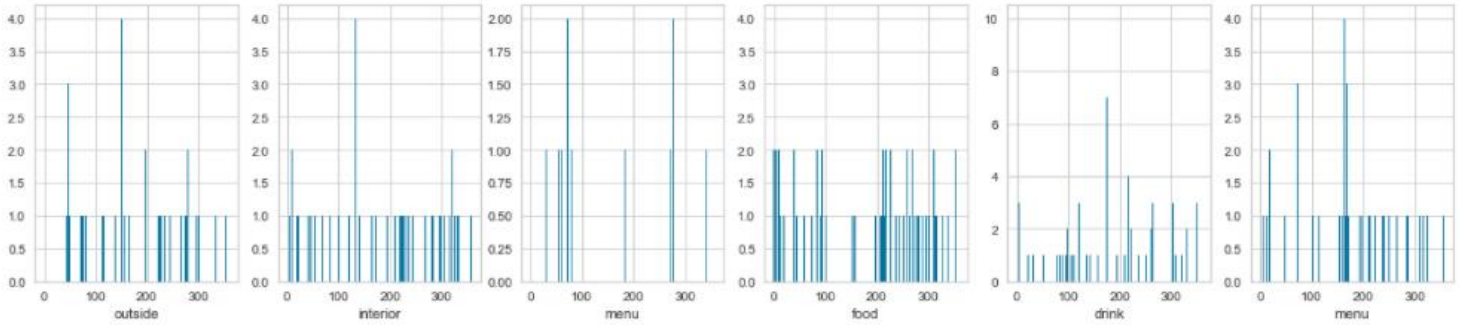
Algorithme Scale-Invariant Feature Transform (transformation de caractéristiques visuelles invariante à l'échelle)

- **Détection des points d'intérêt** dont le rayon est proportionnel à son échelle caractéristique
- **Génération des descripteurs** : qui présentent de nombreuses propriétés d'invariance (rotation, échelle, illumination)

CRÉATION DES HISTOGRAMMES DE FEATURES

Grouper les features par Kmeans (k=racine carrée du nombre de features SIFT)

créer les histogrammes de features basant sur les centres de kmeans

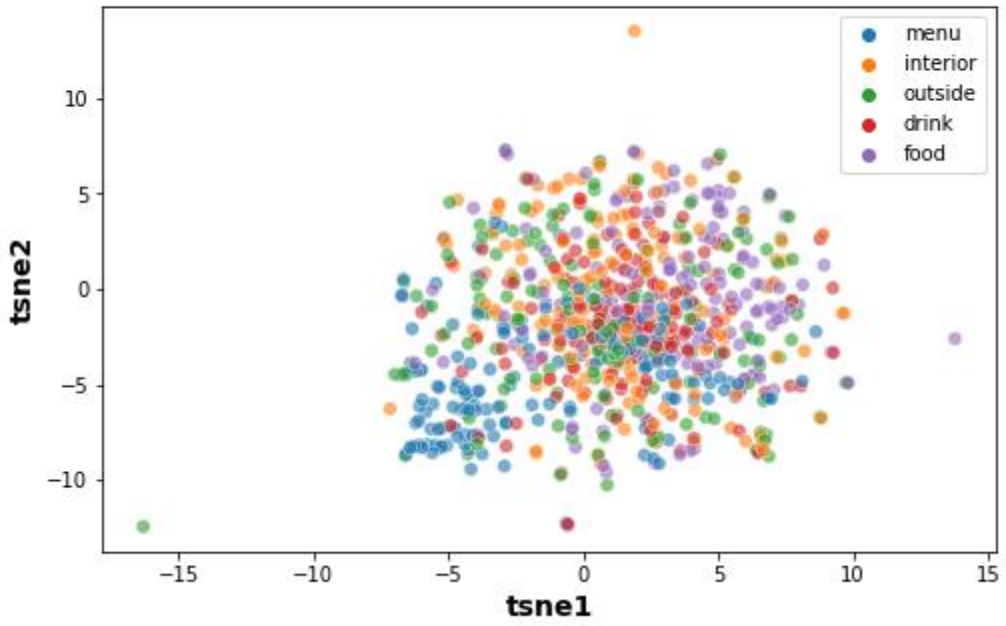


CLASSIFICATION DES IMAGES

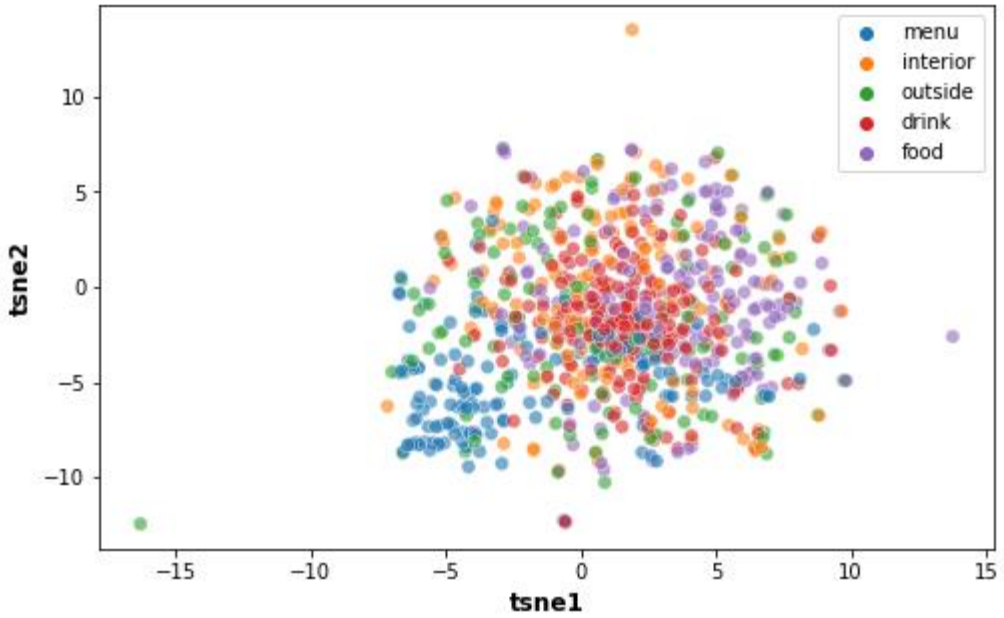
Reduction de dimensions en conservant 99% d'information

```
Dimensions dataset before PCA :  (800, 357)
Dimensions dataset after PCA :  (800, 308)
```

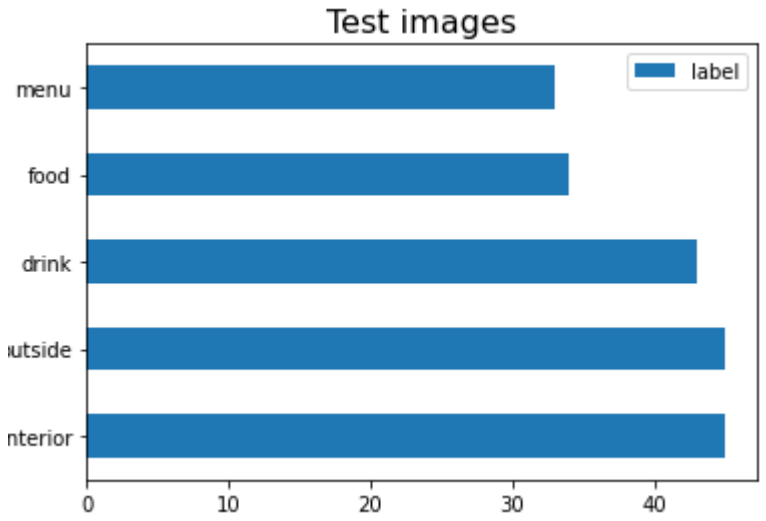
TSNE by category



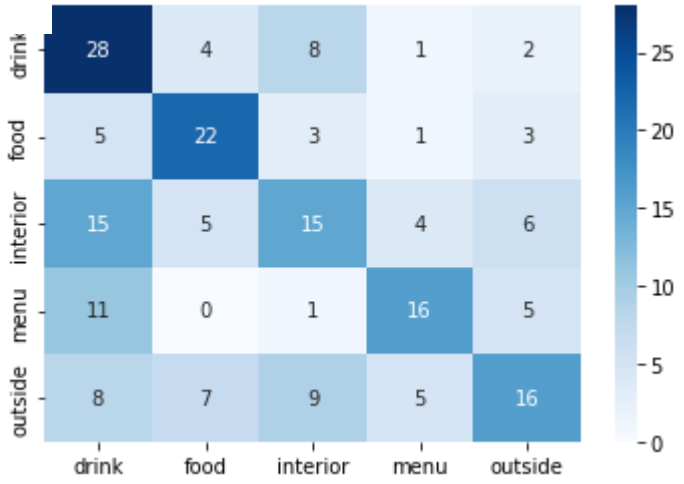
SVM classification



PRÉDICTION SUR DE NOUVELLES IMAGES



Matrice de confusion



Rapport de classification

Test score: 0.48

Classification report for Classifier:

SVC():

	precision	recall	f1-score	support
drink	0.42	0.65	0.51	43
food	0.58	0.65	0.61	34
interior	0.42	0.33	0.37	45
menu	0.59	0.48	0.53	33
outside	0.50	0.36	0.42	45
accuracy			0.48	200
macro avg	0.50	0.49	0.49	200
weighted avg	0.49	0.48	0.48	200

CONCLUSION SUR LA FAISABILITÉ

- La méthode SIFT est confirmée par le test pour sa pertinence dans l'extraction des features. Son insensibilité aux transformations géométriques permet à un algorithme de classification de labelliser de nouvelles images.
- D'autres méthodes plus récentes ont démontrées une meilleure performance, tel le CNN, elles ne rentrent pas dans le cadre de cette étude.

COLLECTER DE NOUVELLES DONNÉES VIA L'API YELP

APPROCHE AVEC GRAPHQL API

S'authentifier → Construire le framwork → Construire le client → Construire le contenu de la requête en incluant les identifications des 200 restaurants → Récupérer le résultat de la requête dans un tableau → Enregistrer le table dans un fichier csv

Exemple d'une requête:

```
{  reviews(business: "yx3Ab0V39RIFJ7iPC9WIXg") {
        review {
            rating
            text}
    }
}
```

	business_id	rating	text
0	yx3Ab0V39RIFJ7iPC9WIXg	4	Located just two blocks from the Texas State C...
1	yx3Ab0V39RIFJ7iPC9WIXg	5	I visit Lavazza at least every other week. The...
2	yx3Ab0V39RIFJ7iPC9WIXg	2	Drink=4 stars.\n\nCustomer service= 1.5-2 star...

CONCLUSION SUR LA FAISABILITÉ

- Le test est effectué avec GraphQL.

- Par apport aux APIs classiques, Yelp GraphQL propose une plus grande flexibilité en termes de requêtes ponctuels et spécifiques.

QUESTIONS?