Avis Restau

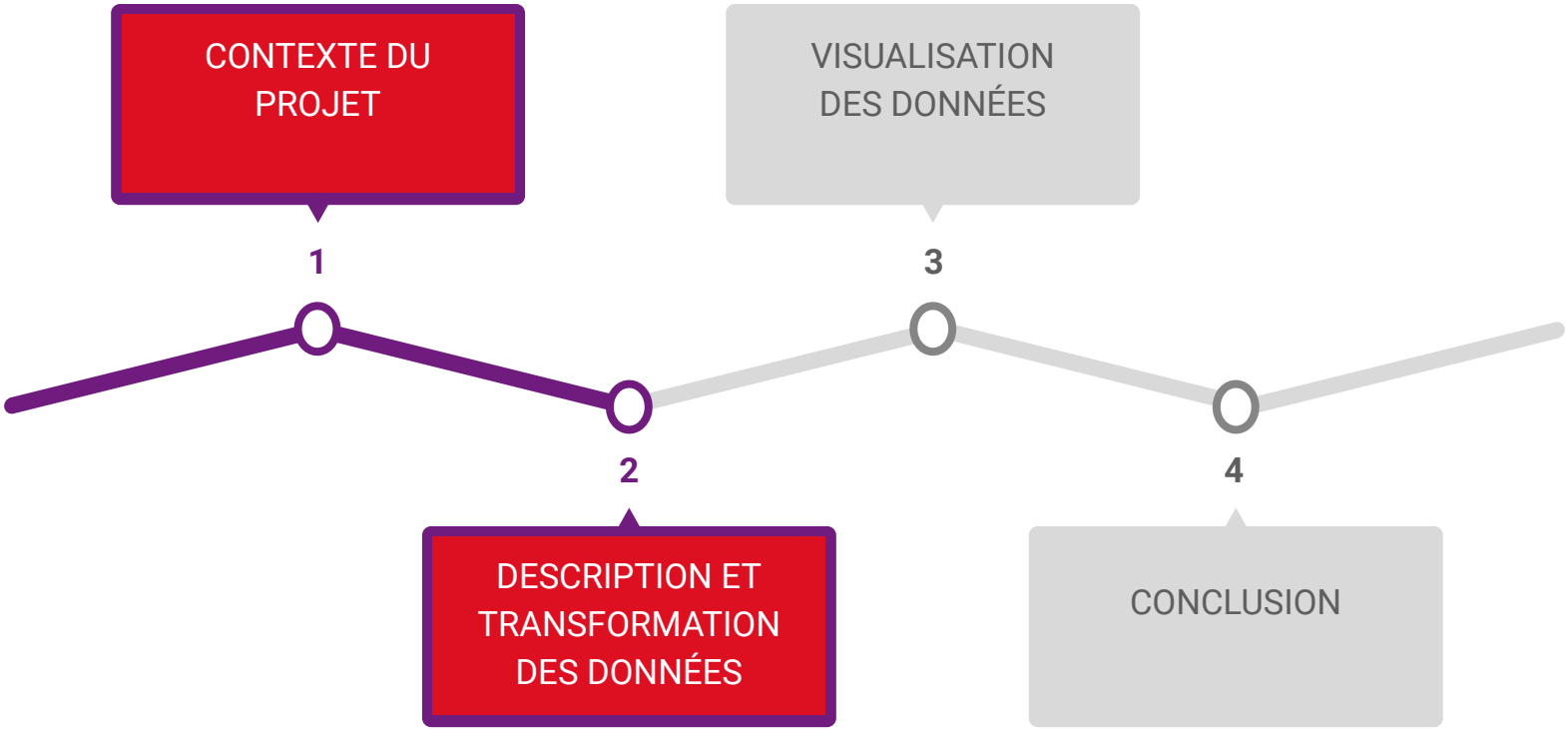
Améliorer le produit IA de Avis Restau

Moussa KIBALY

OPENCLASSROOMS

Projet 6 - Parcours Ingénieur IA

SOMMAIRE


CONTEXTE DU PROJET
1
DESCRIPTION ET TRANSFORMATION DES DONNÉES
2
VISUALISATION DES DONNÉES
3
CONCLUSION
4

1. CONTEXTE DU PROJET

La société **Avis Restau** souhaite connaître les avis postés par leur client sur leur plateforme.

Cependant, on va utiliser l'**API Yelp** qui centralise les **commentaires** et les **photos** d'un grand nombre de restaurants, dans différentes villes dans le monde.

On pourra ainsi connaître les **sujets d'insatisfaction** des clients et classifier les images postées.

2. DESCRIPTION ET TRANSFORMATION DES DONNÉES

Le jeu de données comprend un fichier principal contenant les commentaires et photos de **200 restaurants** de la ville de **Paris**.

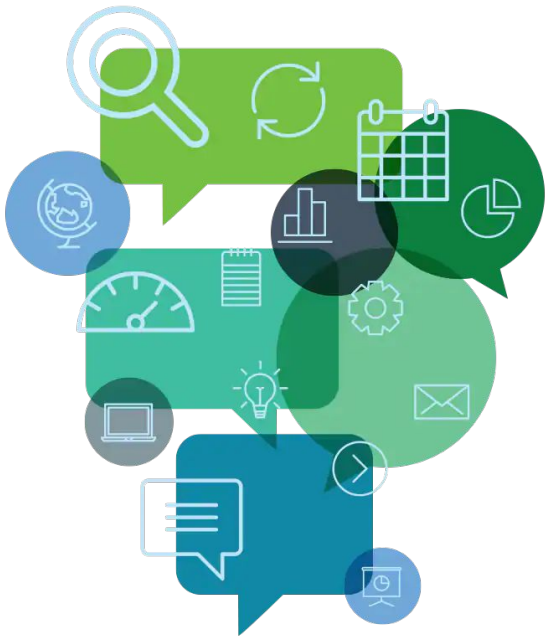
Ces données sont récupérées en envoyant des requêtes GET via l'**API Yelp**.

Les données textuelles et visuelles de ce fichier vont être pré-traitées avant l'utilisation de nos modèles.

Fichier des commentaires et images des restaurants de Paris

name	text_review	photo_1	photo_2	photo_3
Le Comptoir de la Gastronomie	Une adresse immanquable. \nLe meilleur magret...	https://s3-media2.fl.yelpcdn.com/bphoto/Je6THJ...	https://s3-media2.fl.yelpcdn.com/bphoto/Y0D70M...	https://s3-media4.fl.yelpcdn.com/bphoto/HetBW7...
Bistro des Augustins	Voilà un bar qu on aime voir à Paris !\nL acc...	https://s3-media1.fl.yelpcdn.com/bphoto/hPCZTb...	https://s3-media4.fl.yelpcdn.com/bphoto/Y5fZV7...	https://s3-media2.fl.yelpcdn.com/bphoto/mYdGVn...
L'As du Fallafel	Rien a ajouter, ils sont toujours parfaits le...	https://s3-media3.fl.yelpcdn.com/bphoto/QMNELS...	https://s3-media1.fl.yelpcdn.com/bphoto/xtLO-X...	https://s3-media4.fl.yelpcdn.com/bphoto/xlxo5a...
L'Avant Comptoir	Nous sommes complètement déstabilisé, possibl...	https://s3-media3.fl.yelpcdn.com/bphoto/mVwgxg...	https://s3-media3.fl.yelpcdn.com/bphoto/XA5QGl...	https://s3-media2.fl.yelpcdn.com/bphoto/azxglN...
Grenouilles	Nous quittons à regret cette très sympathique...	https://s3-media4.fl.yelpcdn.com/bphoto/fgVrLC...	https://s3-media3.fl.yelpcdn.com/bphoto/dyUbhh...	https://s3-media1.fl.yelpcdn.com/bphoto/pJOb8D...

PRÉ-TRAITEMENT DES DONNÉES TEXTUELLES

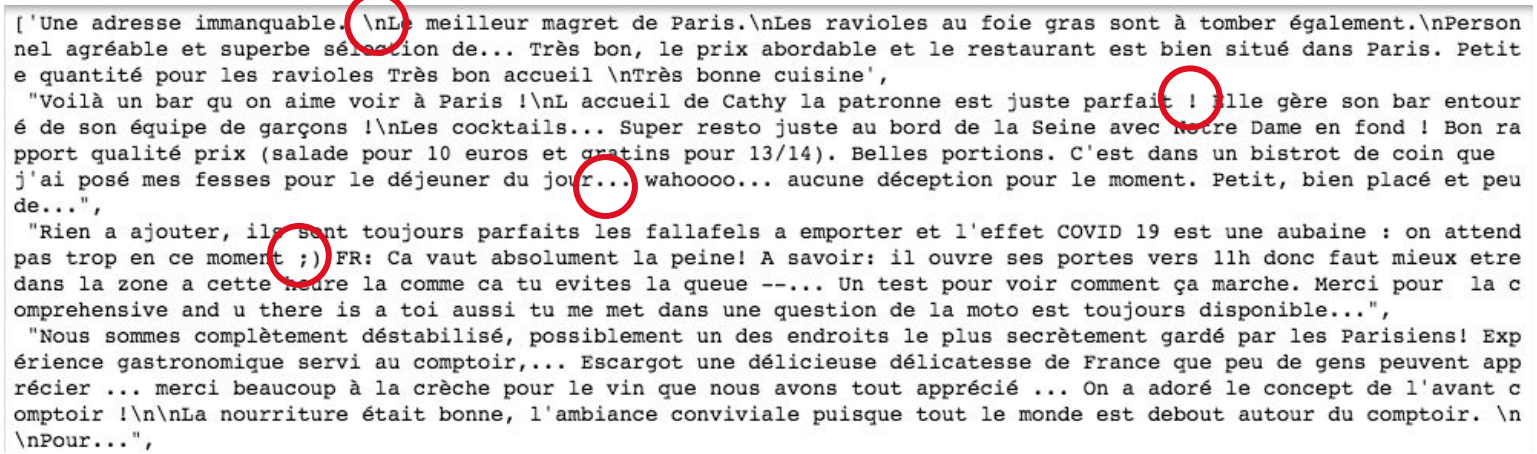


Les commentaires sont nettoyés en supprimant les signes de ponctuation et les nombres, la conversion des mots en minuscule. Les stopwords français sont aussi supprimés ainsi que certains articles définis et conjonction (le, la, les, un, une, et, ...).

Les commentaires sont normalisés en effectuant une lemmatisation (conversion des verbes à l'infinitif, nom commun au masculin singulier) afin de récupérer les sens des mots. Le stemming qui consiste à récupérer la racine des mots par la suppression des suffixes et préfixes n' pas été fait.

Commentaires avant le nettoyage.

(Les signes de ponctuation et caractères spéciaux sont nombreux)



```
['Une adresse immanquable. \nLe meilleur magret de Paris.\nLes ravioles au foie gras sont à tomber également.\nPerson
nel agréable et superbe sélection de... Très bon, le prix abordable et le restaurant est bien situé dans Paris. Petit
e quantité pour les ravioles Très bon accueil \nTrès bonne cuisine',
 "Voilà un bar qu on aime voir à Paris !\nL accueil de Cathy la patronne est juste parfait ! Elle gère son bar entour
é de son équipe de garçons !\nLes cocktails... Super resto juste au bord de la Seine avec Notre Dame en fond ! Bon ra
pport qualité prix (salade pour 10 euros et gratins pour 13/14). Belles portions. C'est dans un bistrot de coin que
j'ai posé mes fesses pour le déjeuner du jour... wahoooo... aucune déception pour le moment. Petit, bien placé et peu
de...",
 "Rien a ajouter, ils sont toujours parfaits les fallafels a emporter et l'effet COVID 19 est une aubaine : on attend
pas trop en ce moment ;) FR: Ca vaut absolument la peine! A savoir: il ouvre ses portes vers 11h donc faut mieux etre
dans la zone a cette heure la comme ca tu evites la queue --... Un test pour voir comment ça marche. Merci pour  la c
omprehensive and u there is a toi aussi tu me met dans une question de la moto est toujours disponible...",
 "Nous sommes complètement déstabilisé, possiblement un des endroits le plus secrètement gardé par les Parisiens! Exp
érience gastronomique servi au comptoir,... Escargot une délicieuse délicatesse de France que peu de gens peuvent app
récier ... merci beaucoup à la crèche pour le vin que nous avons tout apprécié ... On a adoré le concept de l'avant c
omptoir !\n\nLa nourriture était bonne, l'ambiance conviviale puisque tout le monde est debout autour du comptoir. \n
\nPour...",
```

Commentaires après le nettoyage.

(Les signes de ponctuation et caractères spéciaux ont été supprimés et les mots convertis en minuscule.)

```
['adresse immanquable meilleur magret pari raviole foie gras tomber également personnel agréable superbe sélection tr
ès bon prix abordable restaurer bien situer pari petit quantité raviole très bon accueil très bon cuisine',
 'voilà bar aime voir pari accueil cathy patronne juste parfait gérer bar entourer équipe garçon cocktail super resta
urant juste bord sein dame fond bon rapport qualité prix salad euro gratin bel portion bistrot coin poser fesse déjeu
ner jour wahoooo aucun déception moment petit bien placer peu',
 'rien avoir ajouter toujours parfait fallafel avoir emporter effet covid aubain attendre trop moment fr ca valoir ab
solument peine avoir savoir ouvrir porte vers heure donc falloir mieux etre zone avoir ce heure comme ca evite queue
test voir comment ça marche merci comprehensiv and u there is avoir aussi mettre question moto toujours disponible',
 'complètement déstabiliser possiblement endroit plus secrètement garder parisien expérience gastronomique servir com
ptoir escargot délicieux délicatesse france peu gens pouvoir apprécier merci beaucoup crèche vin tout apprécier avoir
adorer concept avant comptoir nourriture bon ambiance convivial puisque tout monde debout autour comptoir',
```

Traitement des commentaires

Effet de la fonction stemming

get_stemmed_text('adresse immanquable meilleur magret paris ravioles foie gras tomber également personnel agréable superbe sélection')

```
Out[16]: 'adress immanqu meilleur magret paris raviol foi gras tomb égal personnel agréabl superb sélect'
```

Traitement des commentaires

Effet de la fonction de lemmatisation

get_lemmatized_text('adresse immanquable meilleur magret paris ravioles foie gras tomber également personnel agréable superbe sélection')

```
Out[18]: 'adresse immanquable meilleur magret pari raviole foie gras tomber également personnel agréable superbe sélection'
```

Tokenisation et application

Top 8 des restaurants en nombre de mots des commentaires

La **Tokenisation** permet de comptabiliser le nombre de mots uniques des commentaires des restaurants

Bag of words

Matrice après vectorisation

	abord	abordable	abrite	abriter	absolument	absolut	absolute	abîme	ac	accent	...	évidemment	évite	éviter	évolution	évoque	être	île	îlot	œil
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
639	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
640	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
641	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
642	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
643	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

On peut aussi représenter le corpus en **bag-of-words** (n-grams avec n=1). La méthode compte le nombre d'apparition de chaque dans le corpus et la Vectorisation convertit les commentaires en vecteur numérique. On obtient une matrice dont chaque colonne correspond à un mot du corpus et son nombre d'apparition.

TF-IDF

Matrice après vectorisation

	abord	abord bistro	abord finalement	abord goût	abordable	abordable ailleurs	abordable plus	abordable restaurer	abordable revenir	abrite	...	être tester	île	île cité	île madelein	île saint	îlot	œil	œil surtout
0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.235132	0.0	0.0	0.280631	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
639	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
640	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
641	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
642	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
643	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

La méthode **TF-IDF** (cas sur des bi-grams) affecte un poids à chaque mot en fonction de leur fréquence d'apparition relativement dans tous les commentaires.

PRÉ-TRAITEMENT DES DONNÉES VISUELLES

Les images des restaurants vont être chargées et des corrections d'image telles que **l'étirement pour l'exposition** ou **l'égalisation pour le contraste** vont être appliquées.

Cela passe notamment par la consultation de **l'histogramme des images** qui représente la répartition des pixels selon leur intensité.

Les images présentent des différences de luminosité entre certaines zone notamment avec la présence de pic pour certaines zones, il faut donc égaliser ces images pour corriger le contraste (**méthode equalize**).

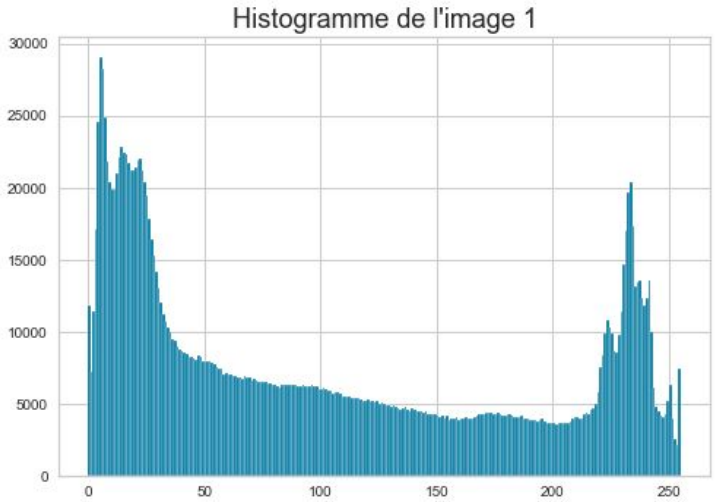
Exemples d’ image du jeu de données

Le jeu de données contient des images de plats alimentaires.

Le jeu de données contient aussi des images du décor intérieur ou extérieur de restaurant.

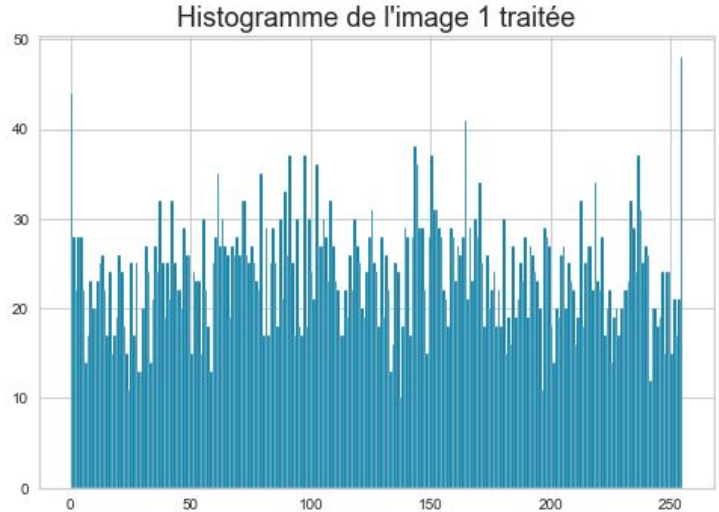
Pré-traitement des images

Image sans correction



L'histogramme met en évidence des différences de luminosité entre certaines zone notamment avec la présence de pic pour certaines zones. Il faut égaliser l'image pour corriger le contraste.

Image après correction (filtre & contraste)



L'histogramme après l'égalisation et l'application d' un filtre moyenneur (filtre de taille 3X3) montre une répartition uniforme des pixels sur tous les niveaux d'intensité.

3. VISUALISATION DES DONNÉES

Méthode **Bag of words**

Top 20 des mots les plus utilisés des commentaires quelque soit le rating. L'adjectif bon, les mots restaurant, plat sont souvent utilisés

Nombres d'occurrences des mots du dictionnaire

word	count
bon	237
très	228
avoir	148
petit	126
tout	111
restaurant	99
plat	89
restaurer	84
...	

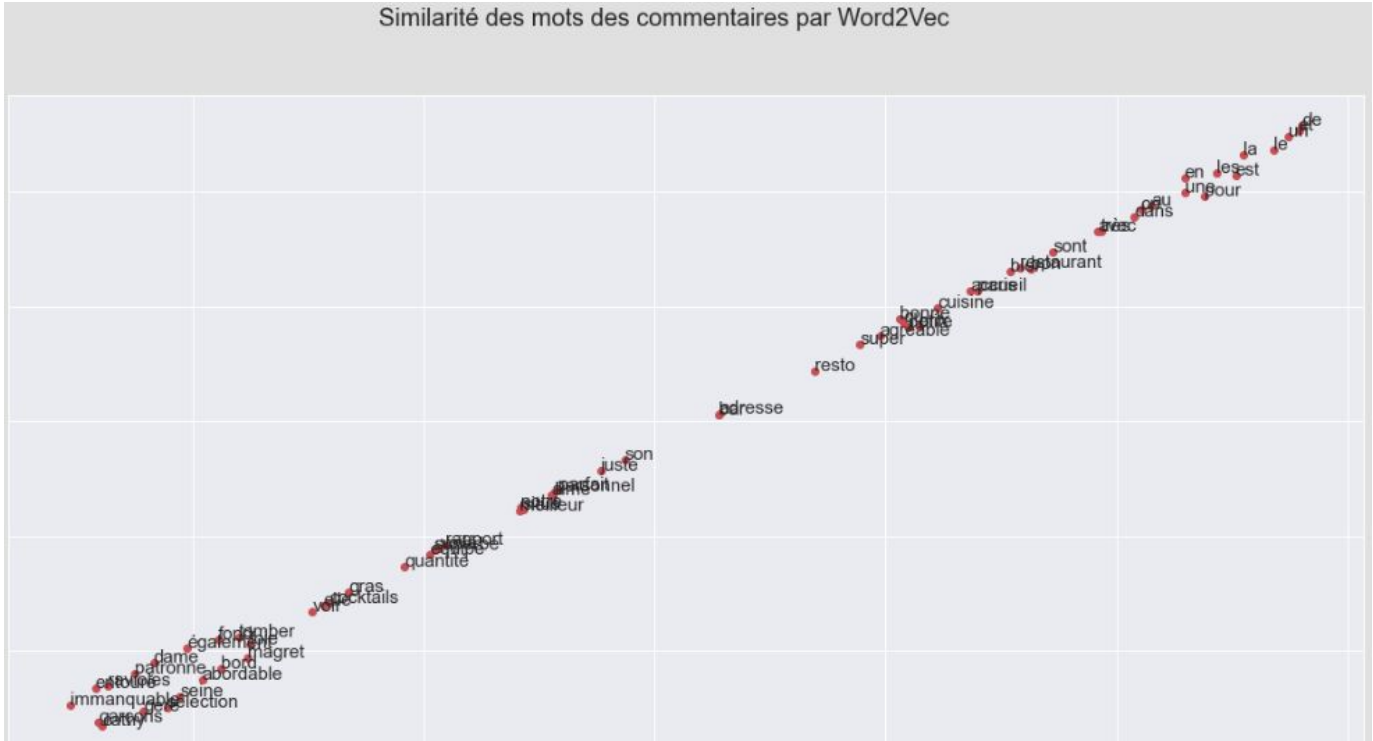
TF-IDF des mots du dictionnaire

word	count
prix	5.438616
service	4.894011
restaurer	4.785976
plat	4.751053
avoir	4.559232
bien	4.323215
très	4.201960
rien	3.489343
...	

*Pour les commentaires de **rating < 3**, les clients ne sont pas satisfaits du prix, du service, le plat ou le cadre du restaurant*

Méthode **Word2Vec sur tous les commentaires**

Cette méthode permet de rapprocher les mots par similarité. Des mots se rassemblent tels que resto-super, bonne-cuisine qui reflète la satisfaction des clients



Méthode **Word2Vec sur les commentaires de rating < 3**

Les mots ont moins de similarité et plus distribués concernant les commentaires de faible rating. Des mots se rassemblent tels que qualité-plat, qui reflète l'insatisfaction des clients

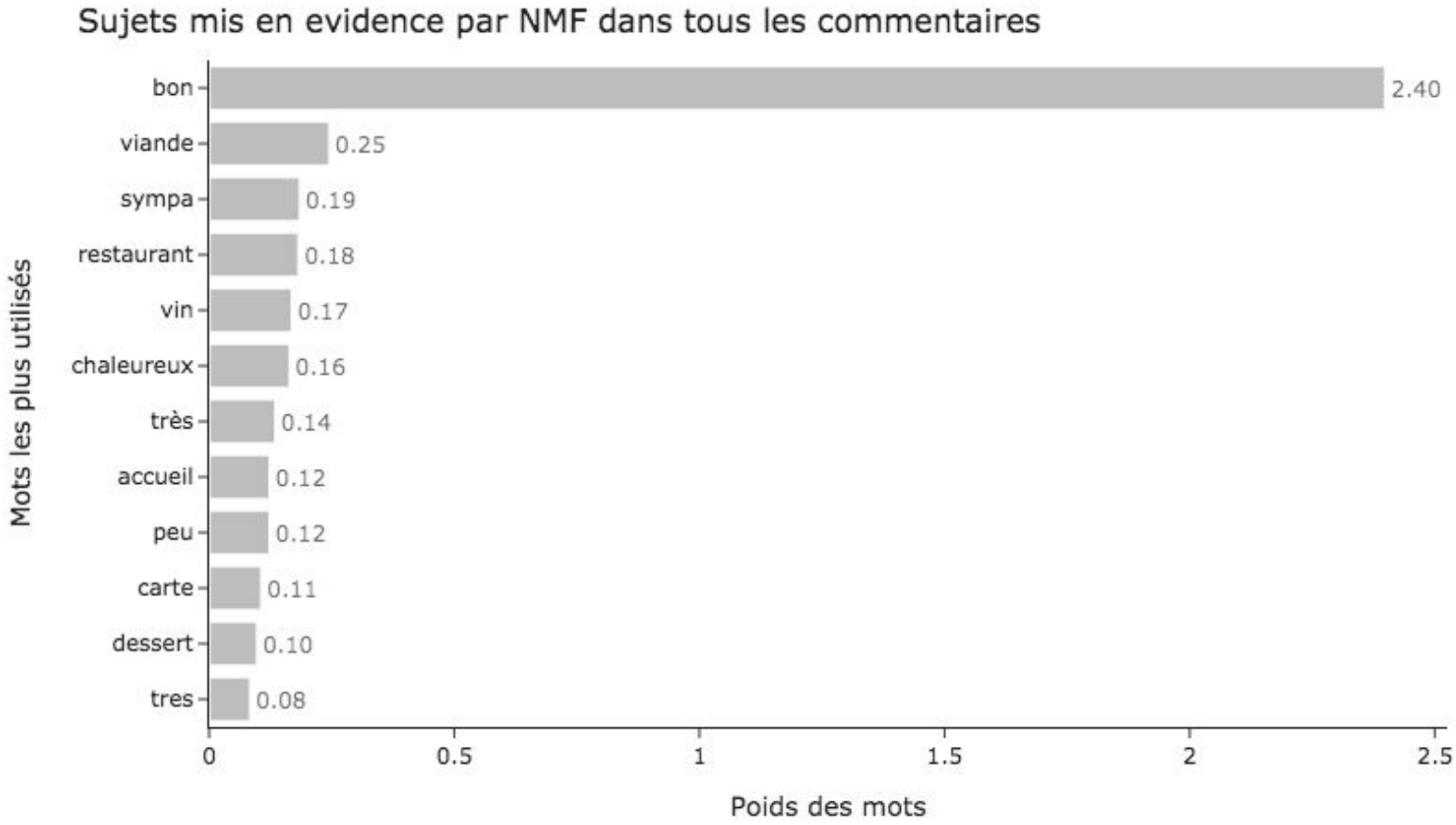
Visualisation des textes

Word Cloud des mots les plus fréquents dans les commentaires

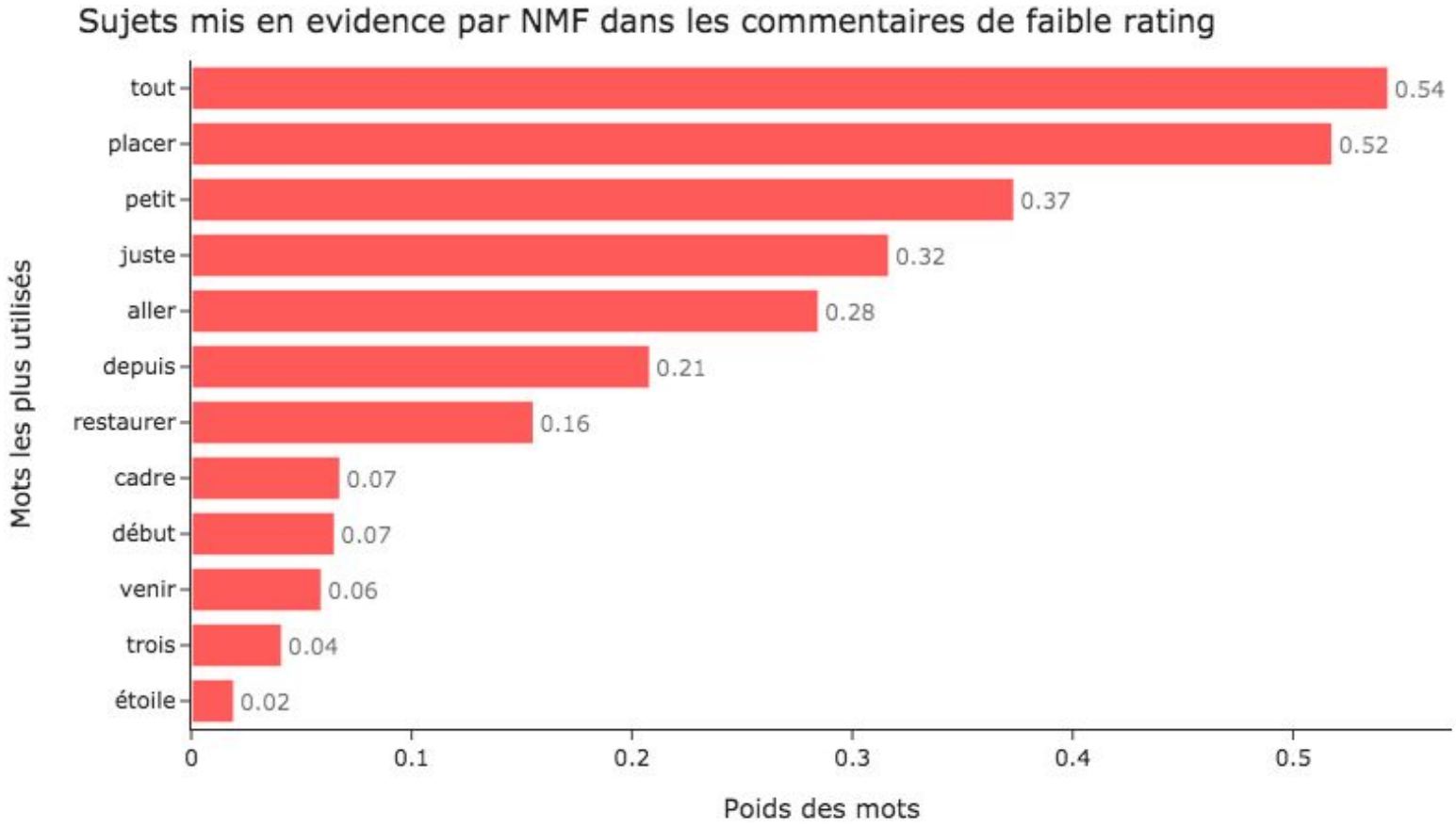
Visualisation des textes

Word Cloud des mots les plus utilisés dans le commentaires négatifs

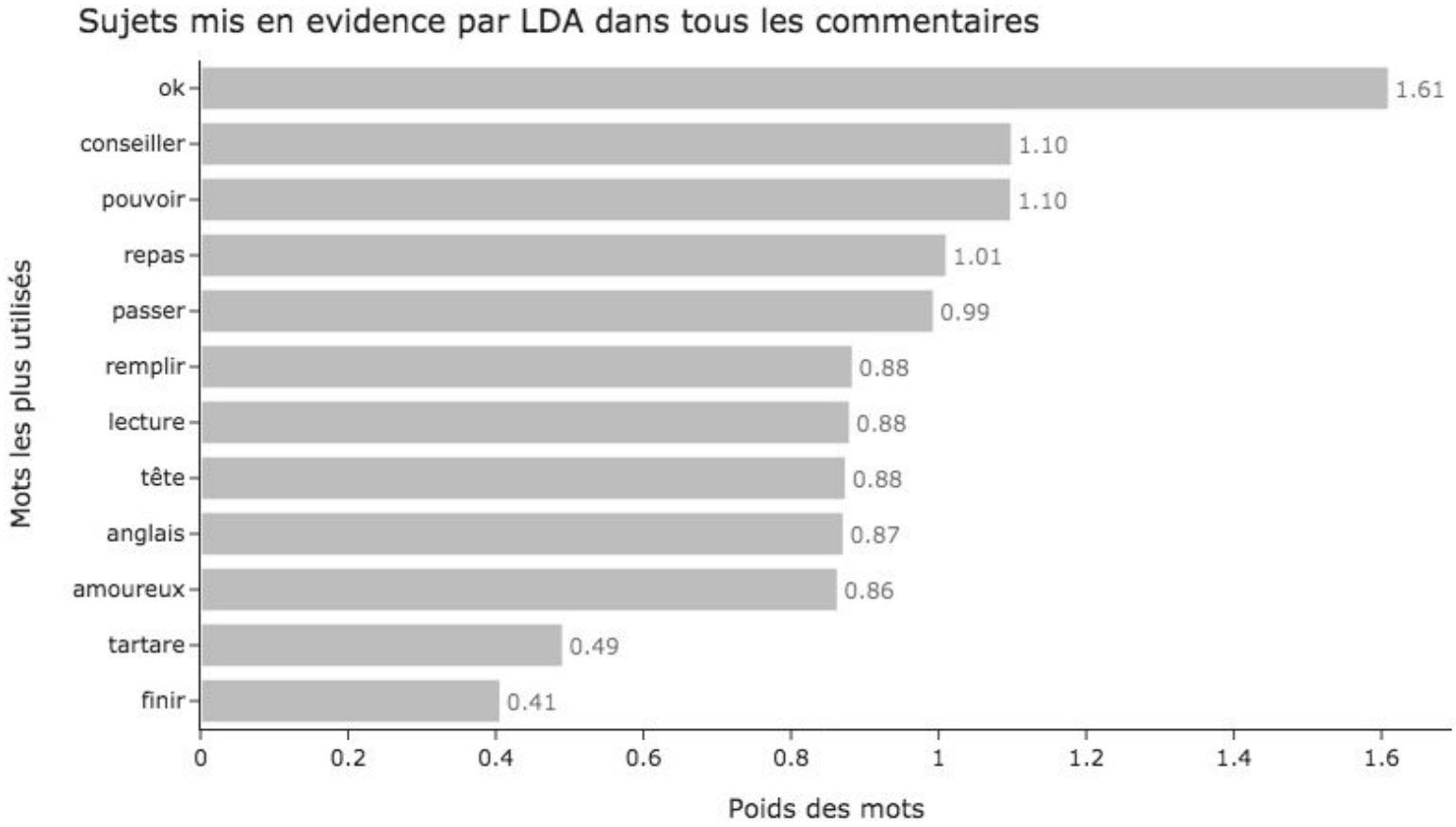
Topic Modelling par NMF



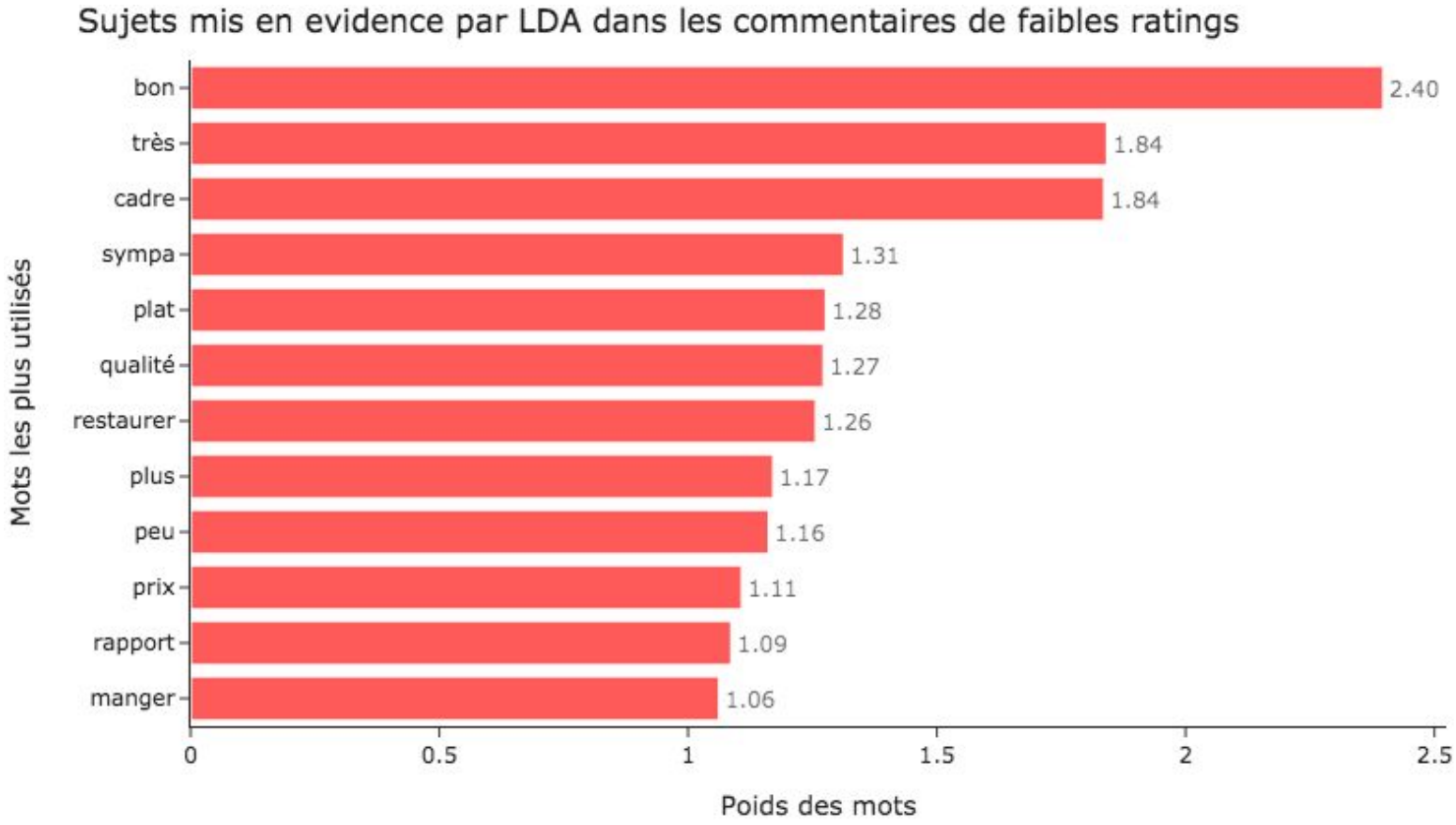
Topic Modelling par NMF



Topic Modelling par LDA



Topic Modelling par LDA

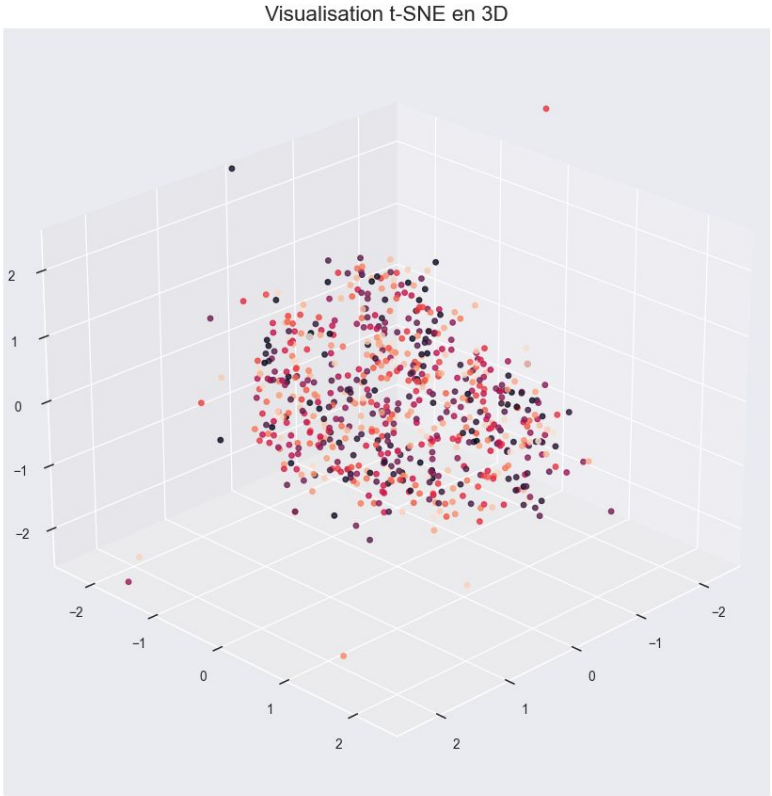


DONNÉES VISUELLES

Similarité des images par la méthode t-SNE en 2D

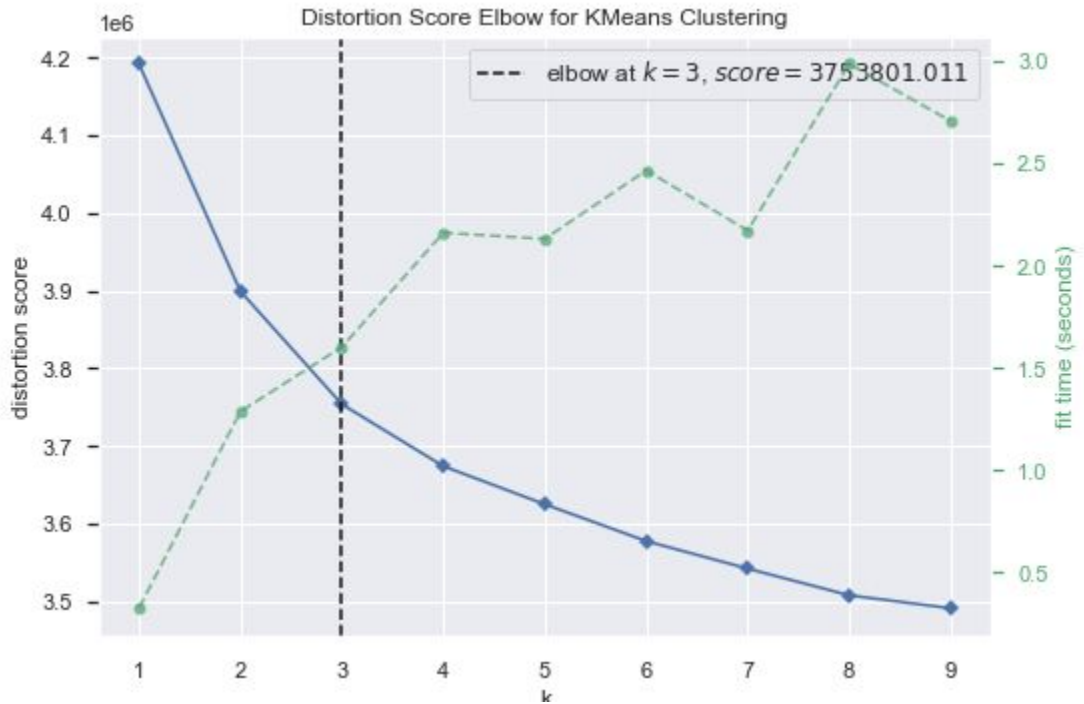
La notion de localité et de similarité est mise en évidence par la méthode t-SNE. Notamment par la formation de groupes d'image ressemblantes telles que des soupes au sommet de la visualisation, des desserts au centre, des images du décor intérieur des restaurants au centre gauche

Similarité des images par la méthode t-SNE en 3D



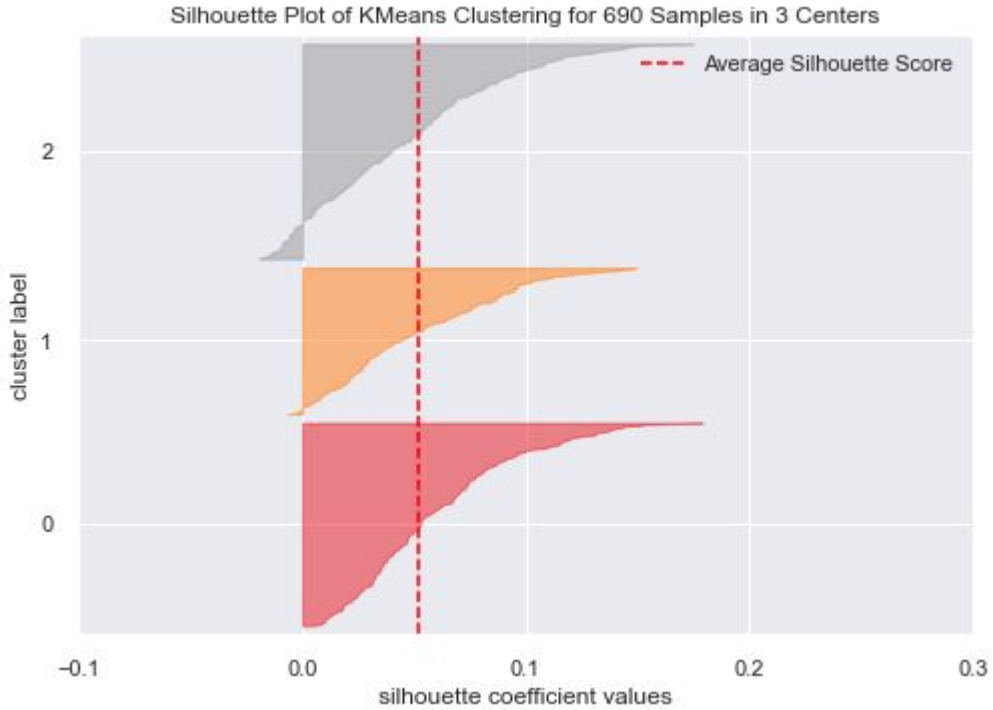
La visualisation t-SNE permet d'observer la similarité des images des restaurants

Labellisation automatique des images par KMeans



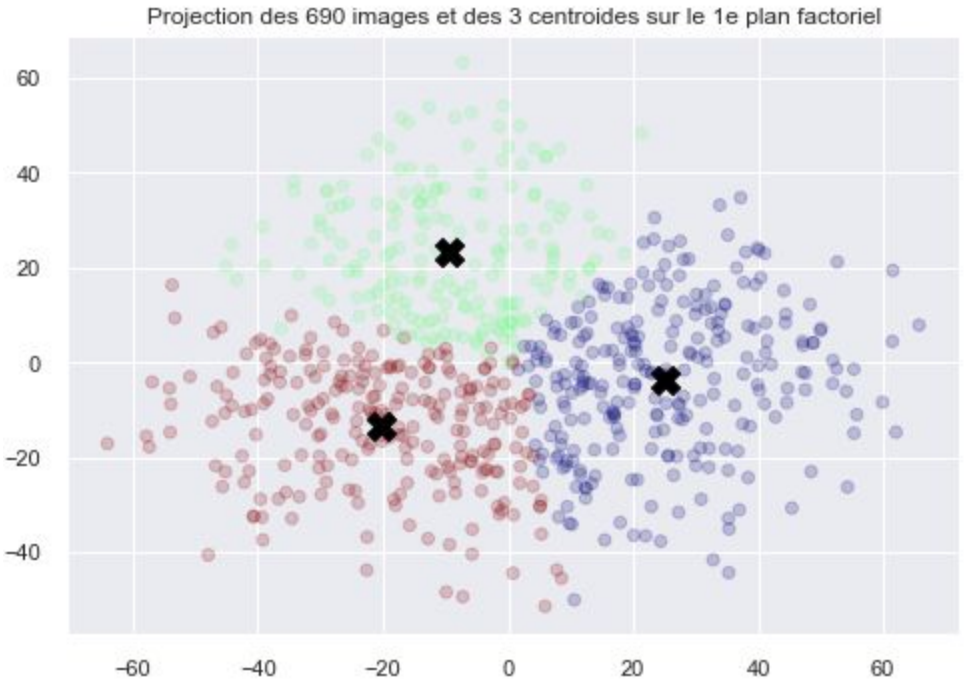
La méthode dite du “coude” nous indique que le nombre de clusters optimal est k=3. On peut définir 3 types d’image des restaurants

Mesure de la performance du clustering avec le coefficient silhouette



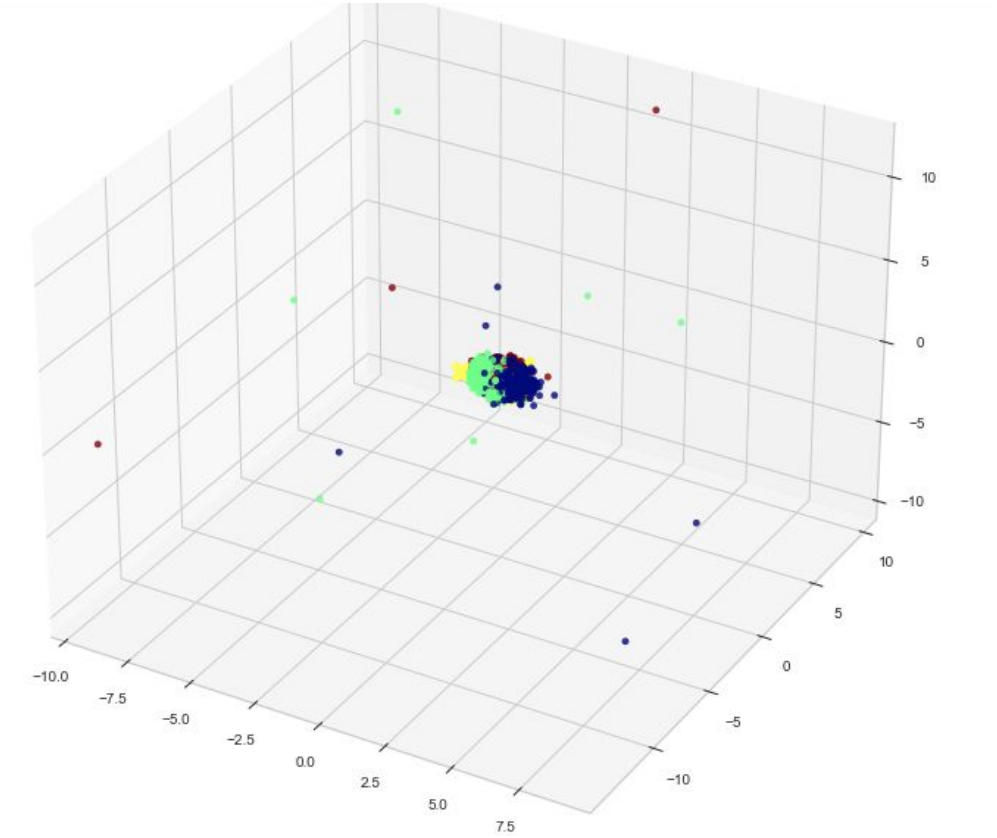
La mesure de la performance est donnée par le coefficient de silhouette (compris entre -1 et 1) pour chaque cluster. Dans notre cas, les clusters sont bien définis et leur coefficient de silhouette est supérieur à la moyenne 0.05

Labellisation automatique des images par KMeans



Les photos sont groupés selon 3 clusters différents après une ACP

Visualisation t-SNE des 3 clusters en 3 dimensions



Les photos sont groupés selon 3 clusters différents après une projection t-SNE sur 3 composantes

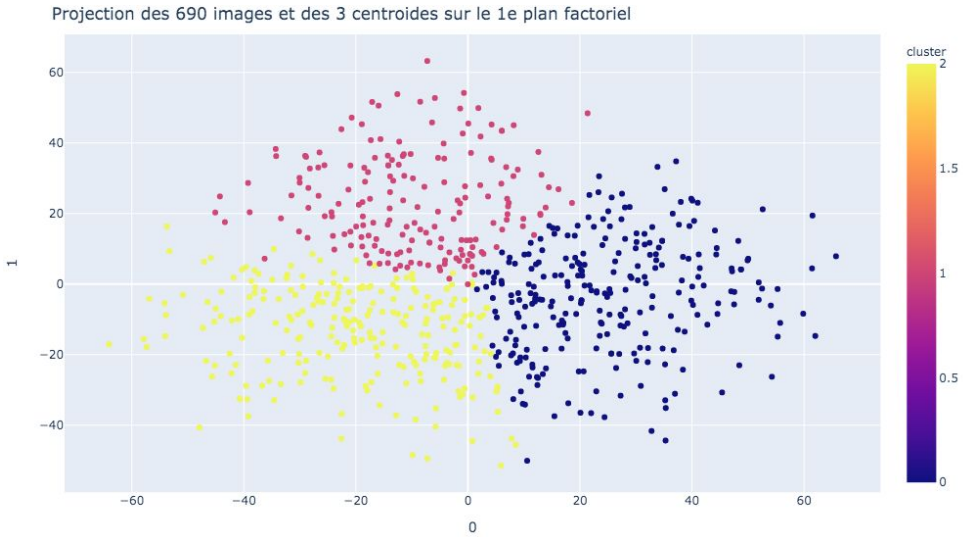
4. CONCLUSION

Sujet d'insatisfaction et de satisfaction

```
('décevoir', -1.0080453304164525)
('cadre', -0.837899740433186)
('contre', -0.8304509187571294)
('prix', -0.8228163478478543)
('ramen', -0.7569189690783168)
('repas', -0.7124168605315605)
('vraiment', -0.7119351600135643)
('fois', -0.6835460888822105)
('commentaire', -0.6663246062043506)
('étoile', -0.6495575478171531)
```

```
('bon', 0.7249239295146709)
('ambiance', 0.581637372067532)
('cuisine', 0.5773216002281969)
('lieu', 0.5762395436232107)
('petit', 0.5297740021336907)
('restaurant', 0.5187382448595316)
('manger', 0.4430821251676288)
('délicieux', 0.4277965406562085)
('meilleur', 0.4145837453747906)
('donc', 0.4118122082048165)
```

Labellisation des images



Définition et interprétation des clusters

Le cluster 1






Le cluster 1 regroupe en majorité des images du décor intérieur

Définition et interprétation des clusters

Le cluster 2






Le cluster 2 regroupe en majorité des images d'aliments de forme circulaire et des assiettes qui présentent la même forme

Définition et interprétation des clusters

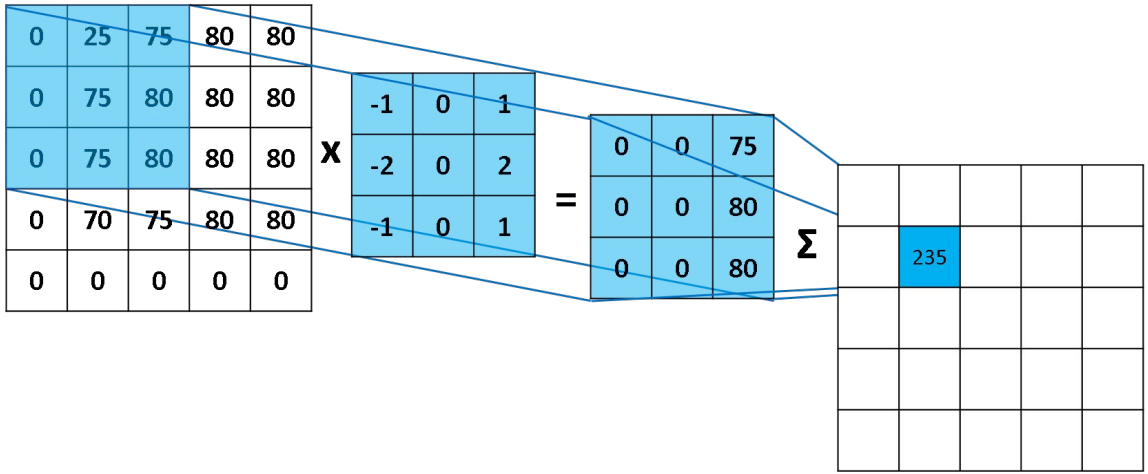
Le cluster 3






Le cluster 3 regroupe en majorité des images contenant des objets en grand nombre (bouteilles, assiettes)

Les réseaux de neurones convolutifs (CNN)



Les réseaux de neurones convolutifs aussi appelés CNN ou ConvNet sont les modèles les plus performants pour la classification d'images. Ils extraient et apprennent automatiquement les features des images

Définition et interprétation des classes obtenues

La classe 1






La **classe 1** obtenue par transfert learning de l'architecture **VGG 16** regroupe en majorité des images du **décor intérieur**

Définition et interprétation des classes obtenues

La classe 2






La **classe 2** regroupe en majorité des images de **plats alimentaires**

Définition et interprétation des classes obtenues

La classe 3





La **classe 3** regroupe en majorité des images de **plats ou aliments de forme circulaire**