AVIS RESTAU

AMÉLIORER LE PRODUIT IA DE VOTRE START-UP

Contexte

Enjeux

« Avis Restau » cherche a mettre en place une **functionalite de collaboration**

L'equipe IA doit:

- **Detecter** les sujets d'insatisfaction des clients
- **Suggerer** les photos de restaurant pertinentes

Objectifs

Implemeter des modeles de Machine Learning permettant repondre aux questions :

- Est-ce qu'on peut distinguer un commentarier **negatif** ?
- Quels son les **sujets d'insatisfaction** des commentaires clients?
- Que **representre** une photo de restaurant?

Méthode

Nous utilisons deux jeu de données fourni par Yelp:

- Le dataset Academic: donnees de 150346 restaurants, ~7 millions de commentaires et 200100 photos annotees.
- L'API GraphQL de YELP: donnees de 200 restaurants, 1 commentaire et une photo par restaurants (200 commentaires et 200 photos)

Nous sommes face a deux problemes:

- Un probleme de **Natural Language Processing (NLP):**
  - Definir si un commentaire est positif ou negatif (sentiment analysis)
  - Indentifier les sujets d'un ensamble de commentaire (topic modeling)

- Une probleme de **Computer Vision (CV):**
  - Reconnaitre le contenu d'une image et la labelliser («image labelling »): classification multi-classe

Dans les deux cas, nous allons:

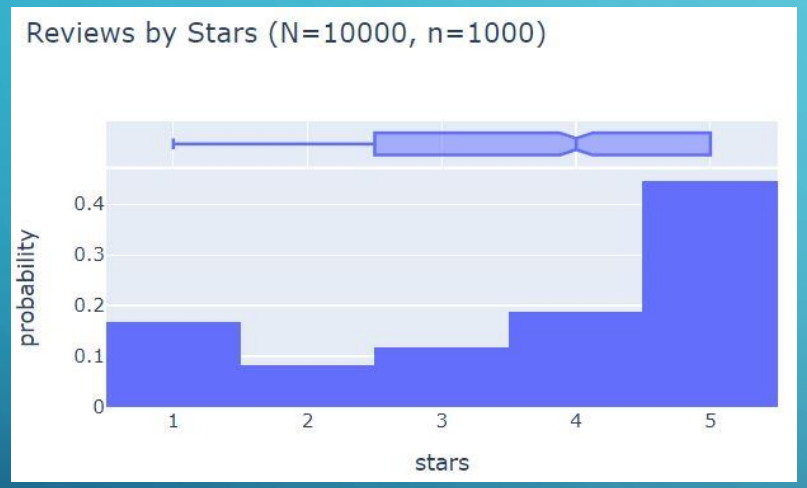
- Extraires les **features:**
  - Texte: « **Tokens** » = groupe de mots reduits a leur forme la plus simple
  - Images: **« visual words » = caracteristiques SIFT**

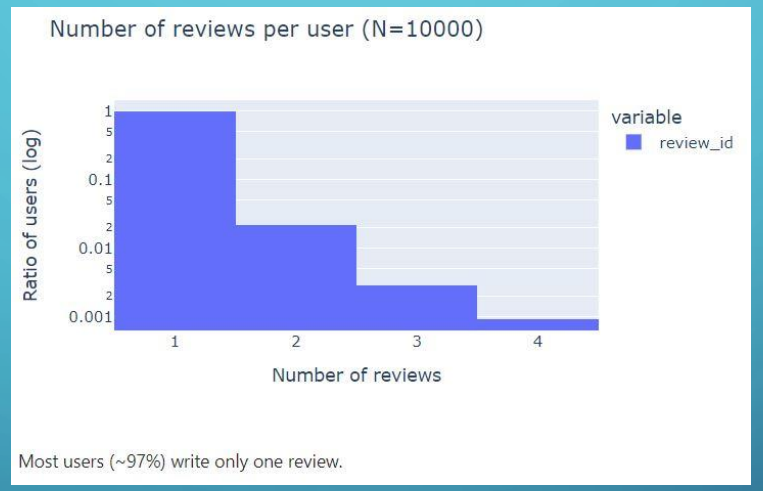
- Representer le corpus sous forme de **Bag of (visual) Words**
- **Reduire la dimension** de cette representation
- Comparer les resultats a des **modeles pre-entraines**

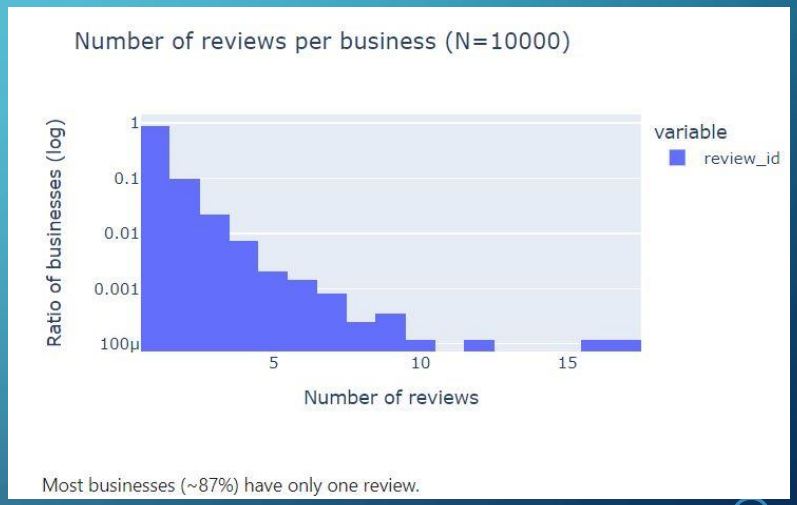
Natural language processing (NLP): Exploratory data analysis

Statistiques :

- Sampling academic dataset a 10000 commentaires







- 45% des utilisatuers donnent 5 etoiles
- 98% des utilisatuers laissent un seule commentaire
- 87% des restaurants ont un seule commentaire

NLP: Prétraitement et analyse des données textuelles

Prétraitement:

- **Lowercase:** transforme le mots en miniscule
- **Punctuation:** Elimine toutes les pontuations
- **Contraction:** transfor contractions
- **Tokenize:** Transforme un texte en liste de tokens

Cleaning:

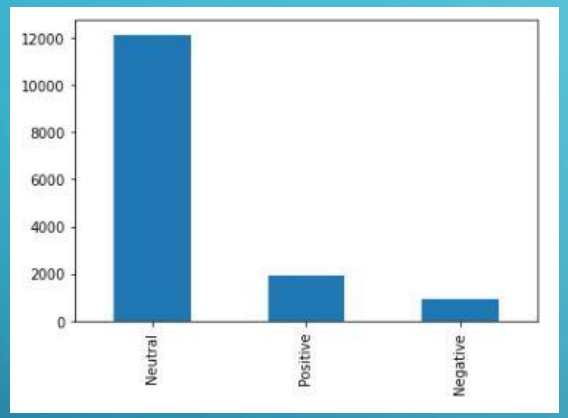
- **Stopwords:** elimine les tokens appartenant a une liste de mots commun ne portant pas de sens
- **Conserver la racine du mot**
  - **Stemming:** supprimer les prefixes et suffixes pour eliminer les variations d'un meme mot.
  - **Lemmatisation:** modifier le mot pour retrouver sa fome de base

- **Vectorizer:** transforme un liste de tokens en vecteur de « features »
  - **CountVectorizer:** simple comptage du nombre d'occurance des tokens par document. Token prequent dans un document => important

- Chaque feaures est de tres grande dimension => nous appliqouns un clustering pour reagruper les features similaires.
- Nous trasnformons notre representation de words en Bag of Words.
- Nous essayons le «topic modeling» et le «word embedding» technique pour recuperer les tematiques.

NLP: Tokenizers et sentiment analysis

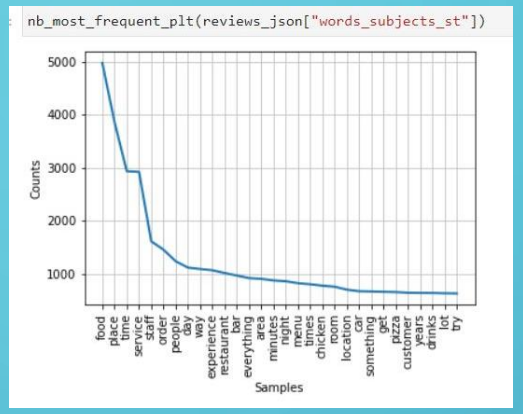
Apres tre-traitement, nous avos obtenue les tokens du text de commentaires.



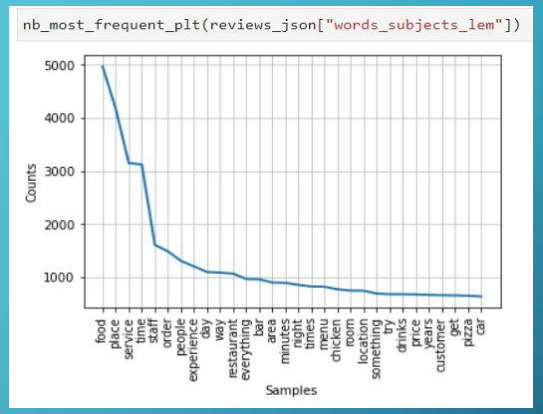
En utilisant **VADER,** nous avos realise un sentiment analysis

Nombre de mots

Stemming

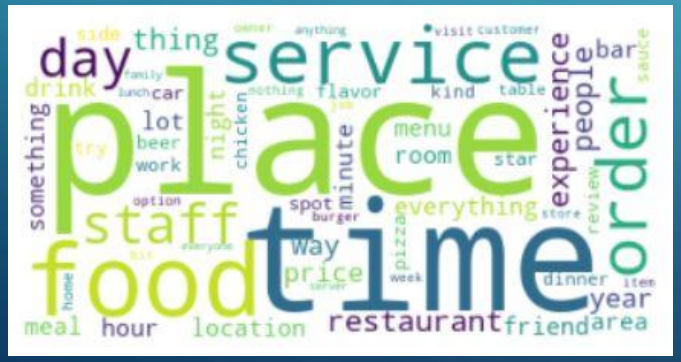


Lemmatization



Nuage de mots

Stemming



Lemmatization

NLP: Vectorization

A partir du vectors cree, nous avons genere:

- **Bag of Words (BOW)**

- **Term-Frequency - Inverse Document Frequency (TF-IDF)**

Depuis, nous avons essaye une reduction de dimentions pour extraer les mots plus representative de chaque cluster (frequence)

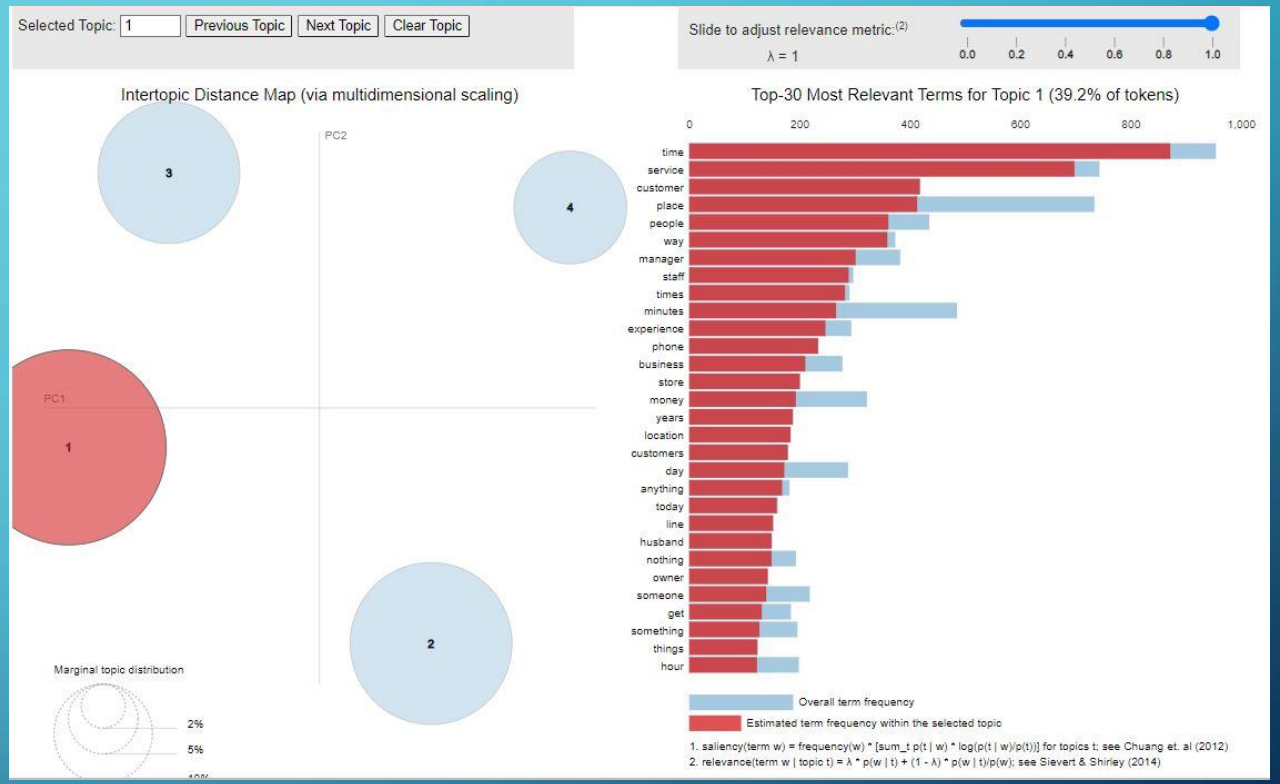
- **PCA**
- **TSNE**
- **Kmeans**

Mais, on remarque que les k-means ont du mal à différencier les différents sujets, on retrouve le mot "temps" dans chaque sujet.

NLP: Topic Modelling: LDA sur dataset

Nous observons des topics pertinents:

- Topic 1: mots liés à l'expérience du client (time, service, manager, staff, experience...)
- Topic 2: mots liés au renting (car, company, credit, days, hotel)
- Topic 3: mots liés a la norriture (food, meal, pizza, meat....)
- Topic 4: mots liés à l'environnement (price, items, bill, places, menu, prices, kids...)



Conclusion

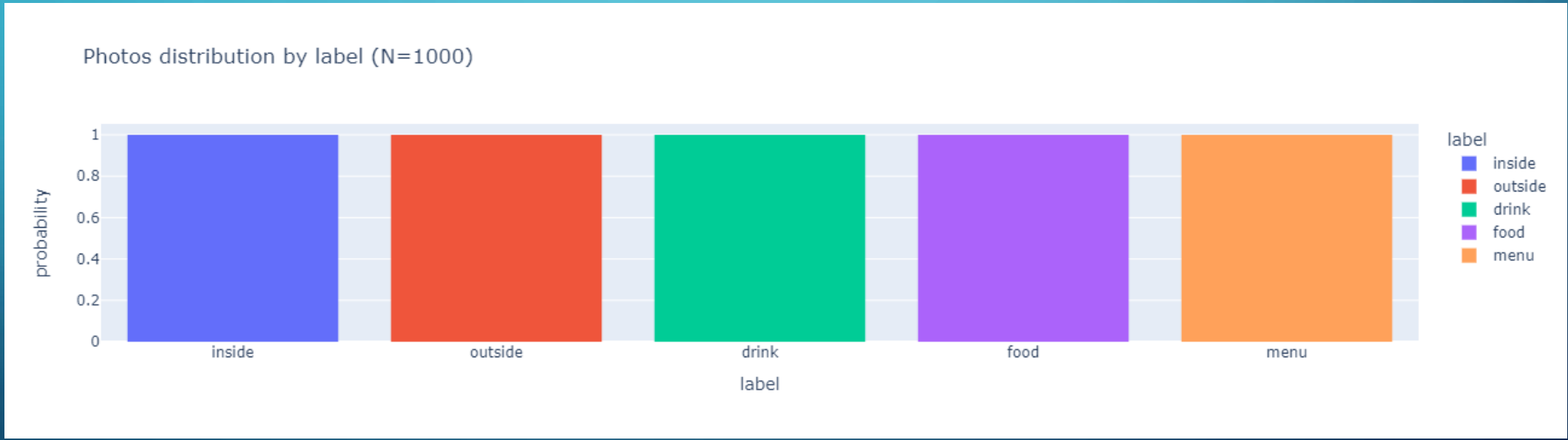
Nous avons pu resoudre les problematiques posees et nous repondre aux questions:

- **Est-ce qu'un commentaire est positif ou negatif?**
  - La utilisation du modele *sentiment_scores* de **VADER** (Valence Aware Dictionary and sEntiment Reasoner) permet de reprondre a cette question

- **Quels sont les sujets d'insatisfaction?**
  - Le «**topic modeling**» (LDA) permet d'identifier les sujets :
    - Quelite de la nourriture
    - Environnement
    - Service
    - Temps

Computer vision: Exploratory data analysis

- La dataset est compose de 200000 photos (sample a 1000)
- Chaque photo a un label: « drink», « food», « inside», « outside» ou « menu»
- Les photos sont en couleur (RGB) et vont de 150x114 px a 600x400 px.



Computer vision: Extraction des features => Bag of Visual Words

Pour chaque photo, nous allons extraires les features:

- Point d'interet: SIFT descriptors

Chaque image est representee par une liste de vecteurs (features)

Chaque feaures est de tres grande dimension => nous appliqouns un clustering pour reagruper les features similaires. Chaque cluster represente un «terme visuel» (Visual Word).

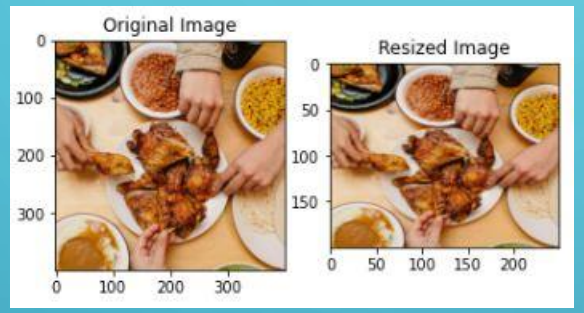
Nous trasnformons notre representation de images en Bag of Visual Words

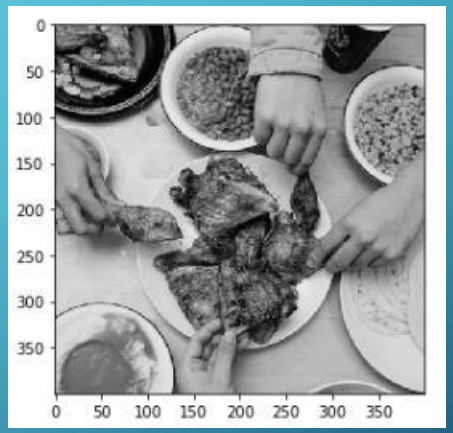
Enfin, nous appliquons une reduction de dimension afin de densifier le corpus.

Cumputer vision: Pretreteatment des images

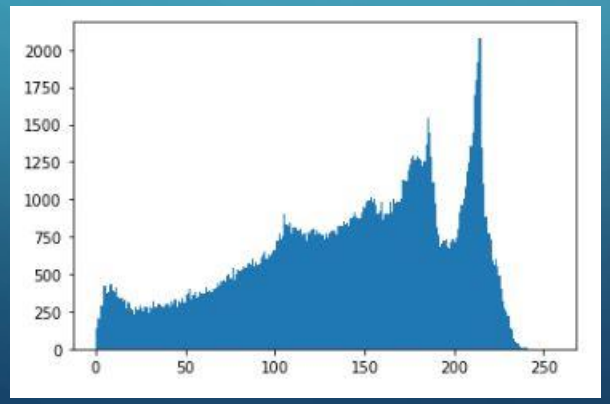
Avant de faire l'extraction de descripteurs, il faut preparer les images

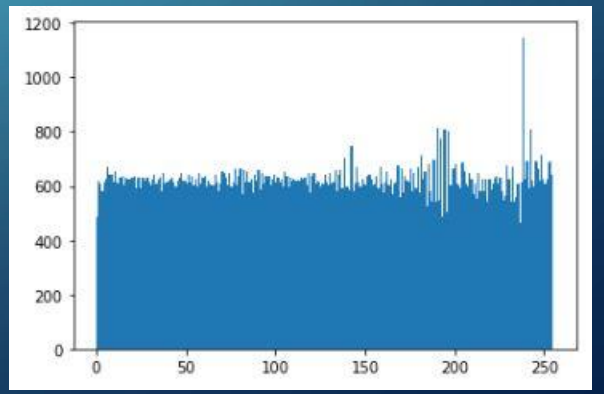
Example des images et labels





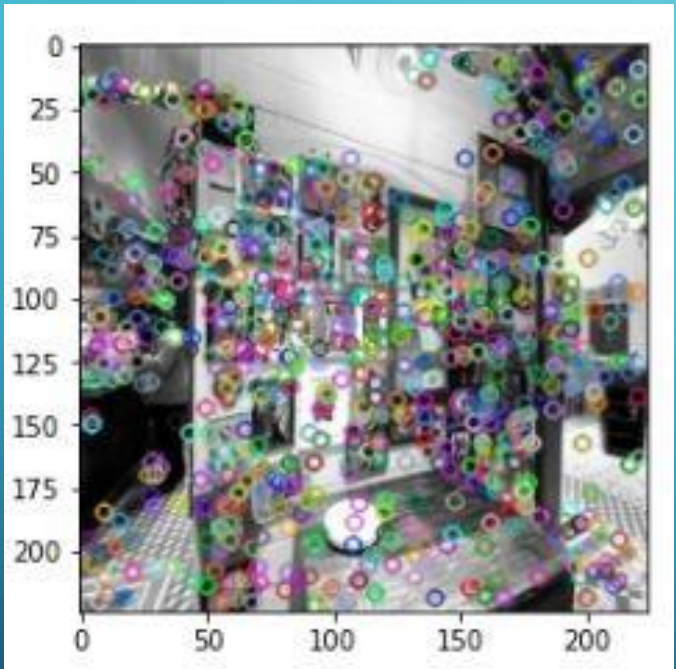
Égalisation du niveau de gris de l'image.





SIFT descripteurs

La scale-invariant feature transform (SIFT), est un algorithme utilisé dans le domaine de la vision par ordinateur pour détecter et identifier les éléments similaires entre différentes images numériques.



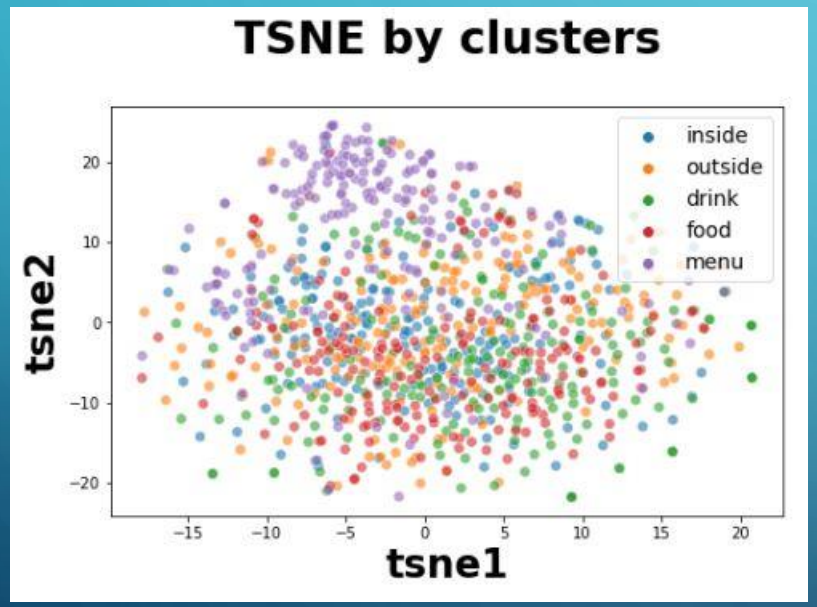

```
Number of descriptors :  (425598, 128)
processing time SIFT descriptor :           14.23 seconds
```

Classification: maison

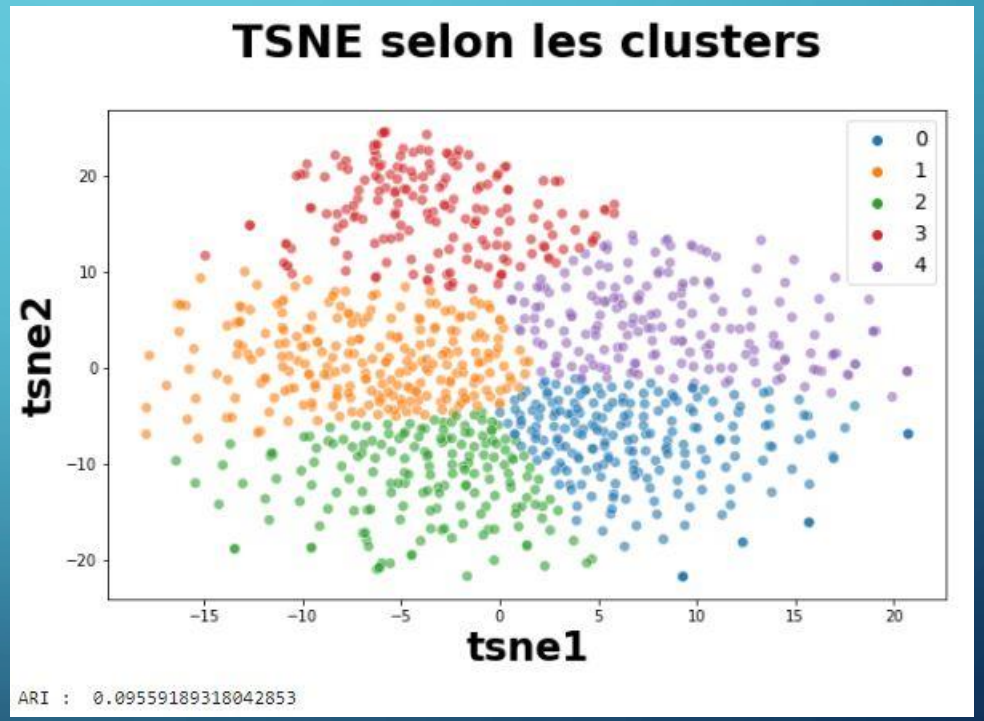
Analyse multidimensionnelle

ACP / T-SNE

```
Dataset dimensions before PCA reduction  :  (1000, 652)
Dataset dimensions after PCA reduction  :  (1000, 537)
```



ACP / T-SNE + Kmeans



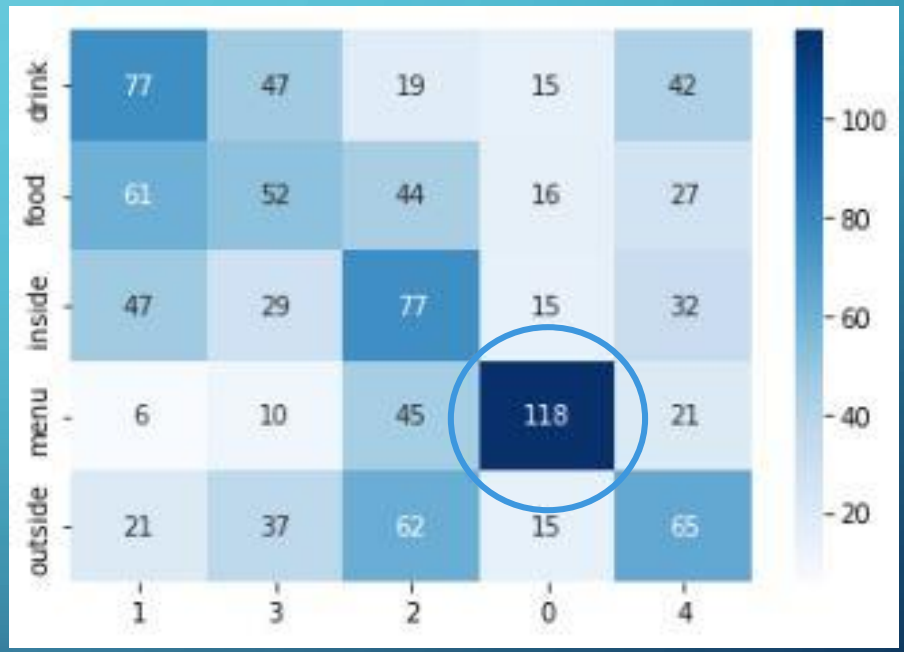
ARI : 0.09559189318042853

Classification: maison

- **F1-score = 0,39**

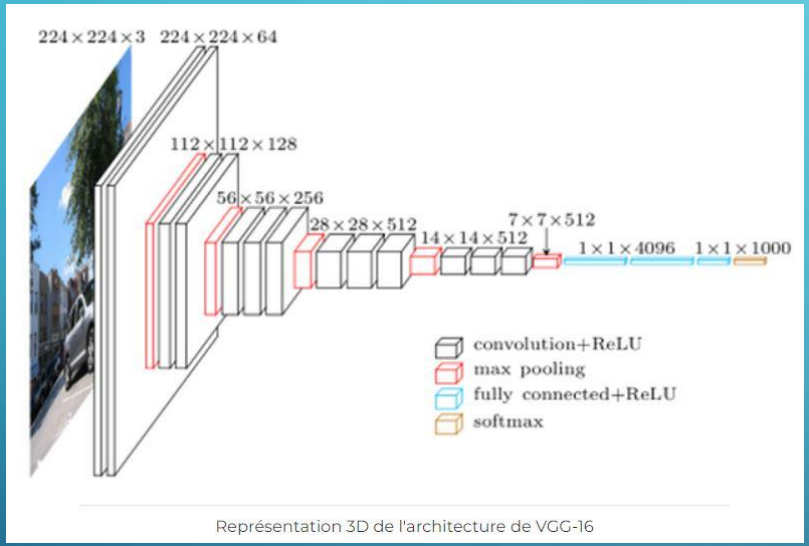
```
Correspondance des clusters :  [0 2 1 3 4]
[[ 77  47  19  15  42]
 [ 61  52  44  16  27]
 [ 47  29  77  15  32]
 [  6  10  45 118  21]
 [ 21  37  62  15  65]]
```

	precision	recall	f1-score	support
0	0.36	0.39	0.37	200
1	0.30	0.26	0.28	200
2	0.31	0.39	0.34	200
3	0.66	0.59	0.62	200
4	0.35	0.33	0.34	200
accuracy			0.39	1000
macro avg	0.40	0.39	0.39	1000
weighted avg	0.40	0.39	0.39	1000



Classification: transfer learning

Ici, nous reutilisons un modele pre-entraine (VGG16) que nous adaptons a notre probleme de classification (« transfer leaning » - « feature extraction»).



Représentation 3D de l'architecture de VGG-16

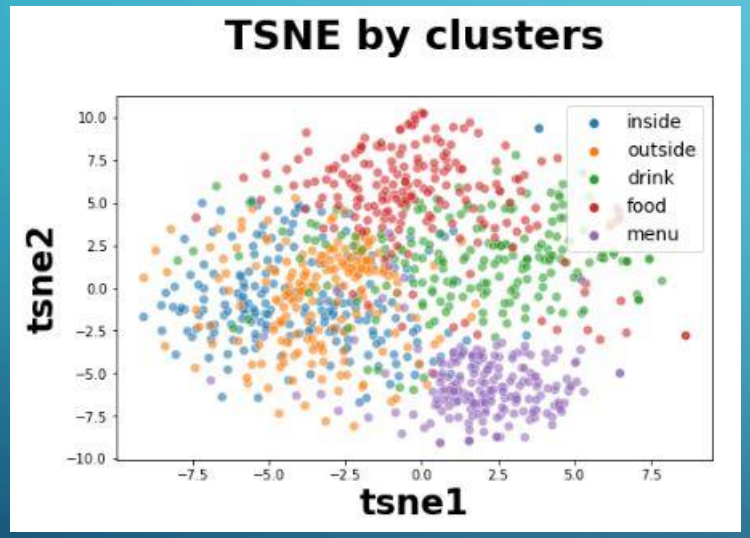
```
Model: "vgg16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_4 (InputLayer)         [(None, 224, 224, 3)]     0

block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792

block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928

block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0

block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856

block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584

block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0

block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168

block3_conv2 (Conv2D)        (None, 56, 56, 256)       590080

block3_conv3 (Conv2D)        (None, 56, 56, 256)       590080

block3_pool (MaxPooling2D)   (None, 28, 28, 256)       0

block4_conv1 (Conv2D)        (None, 28, 28, 512)       1180160

block4_conv2 (Conv2D)        (None, 28, 28, 512)       2359808

block4_conv3 (Conv2D)        (None, 28, 28, 512)       2359808

block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0

block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808

block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808

block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808

block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0

=================================================================
Total params: 14,714,688
Trainable params: 0
Non-trainable params: 14,714,688
```

Classification: transfer learning

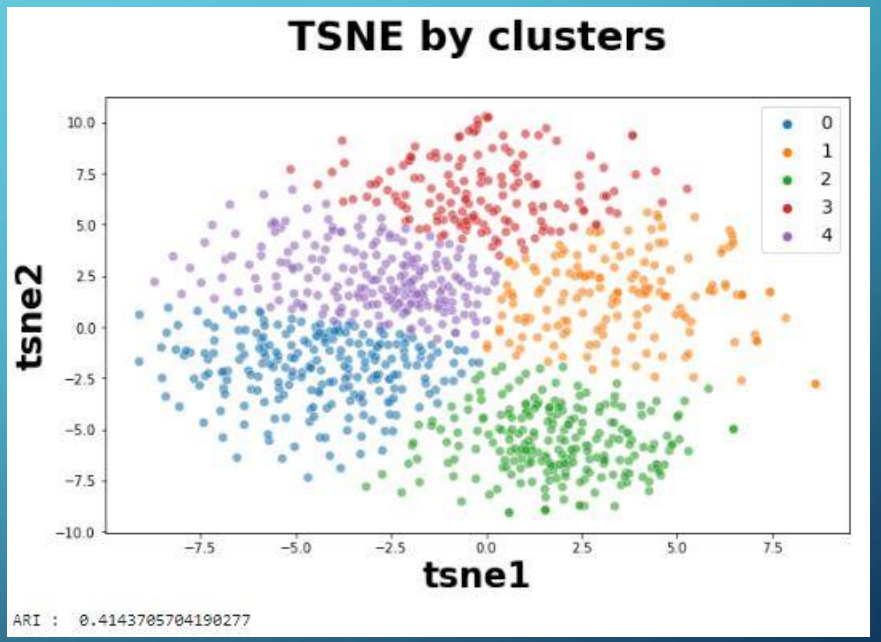
Analyse multidimensionnelle

ACP / T-SNE

```
Dataset dimensions before PCA reduction :  (1000, 25089)
Dataset dimensions after PCA reduction :  (1000, 925)
```



ACP / T-SNE + Kmeans



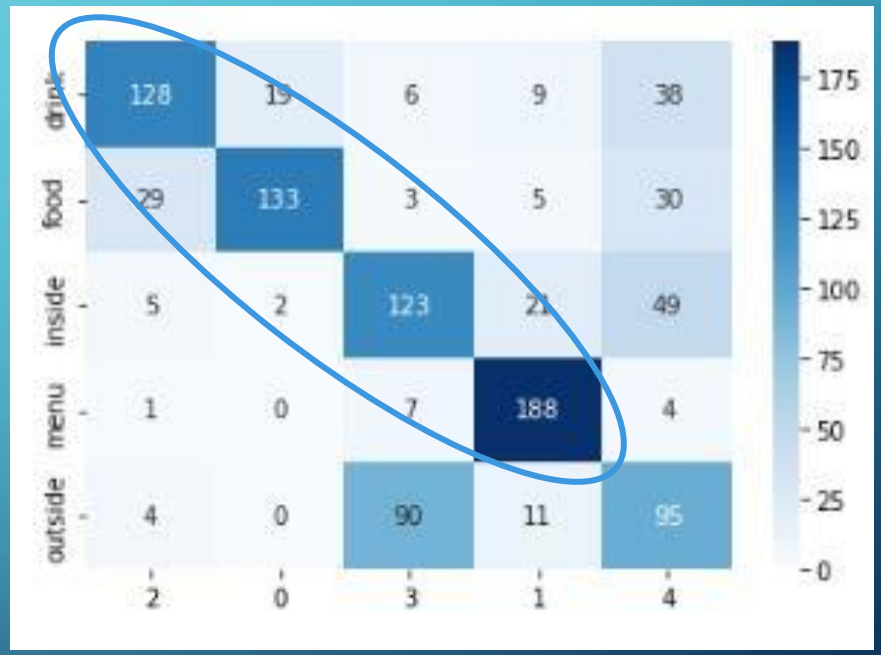

```
ARI :  0.4143705704190277
```

Classification: transfer learning

Ce modele est tres largement **superieur** a notre modele maison

- **F1-score = 0,67 (+71%)**

```
Correspondance des clusters :  [2 0 3 1 4]
[[128  19   6   9  38]
 [ 29 133   3   5  30]
 [  5   2 123  21  49]
 [  1   0   7 188   4]
 [  4   0  90  11  95]]

              precision    recall  f1-score   support

           0       0.77      0.64      0.70       200
           1       0.86      0.67      0.75       200
           2       0.54      0.61      0.57       200
           3       0.80      0.94      0.87       200
           4       0.44      0.47      0.46       200

    accuracy                           0.67      1000
   macro avg       0.68      0.67      0.67      1000
weighted avg       0.68      0.67      0.67      1000
```



Conclusion:

- **Nous avons resoudre les problematiques posees et nous pouvons repondre aux questions:**
  - Que **representre** une photo de restaurant?
    - Une image peut être représentée par un ensemble de features (par example descripteurs) qui, une fois convertis en vecteurs, permettent de réduire la dimension et de procéder à un regroupement pour prédire le label correspondant.
    - En utilisant un modele pre-entraine (VGG16) que l'on adapte (transfer learning) a notre objetif, nous pouvons labellise nos images avec une assez haute precision.

Resquest API YELP

- récupérer uniquement les champs nécessaires,
- stocker les résultats dans un fichier exploitable (par exemple CSV).

Let's export the API data in csv

```
df.to_csv(r'C:\\Users\\ezequ\\proyectos\\openclassrooms\\Projet_6\\data\\P6_01_filecsv\\restaurants_v2.csv')
```

```
df.head(2)
```

	id	name	photos	rw_id	rw_text	rw_rating
0	-0iLH7iQNYtoURciDpJf6w	Le Comptoir de la Gastronomie	[https://s3-media2.fl.yelpcdn.com/bphoto/czh2l...	r3bxlJ2ekrp8UPseAj2wjQ	What an amazing Bistro!! First off, even if yo...	5
1	lU9_wVOGBKjfqTTpAXpKcQ	Bistro des Augustins	[https://s3-media4.fl.yelpcdn.com/bphoto/b95P0...	pjgcPURNS2PvAQ6a0LOShw	You cannot beat this spot for food, drinks, an...	5

Let's import the AP data (csv)

```
df = pd.read_csv(r'C:\\Users\\ezequ\\proyectos\\openclassrooms\\Projet_6\\data\\P6_01_filecsv\\restaurants_v2.csv')
```

```
...
...
```

```
df.head()
```

	id	name	photos	rw_id	rw_text	rw_rating
0	-0iLH7iQNYtoURciDpJf6w	Le Comptoir de la Gastronomie	['https://s3-media2.fl.yelpcdn.com/bphoto/czh2...	r3bxlJ2ekrp8UPseAj2wjQ	What an amazing Bistro!! First off, even if yo...	5
1	lU9_wVOGBKjfqTTpAXpKcQ	Bistro des Augustins	['https://s3-media4.fl.yelpcdn.com/bphoto/b95P...	pjgcPURNS2PvAQ6a0LOShw	You cannot beat this spot for food, drinks, an...	5
2	ctP4c3mwVO5oOzLl48LtuQ	Les Antiquaires	['https://s3-media3.fl.yelpcdn.com/bphoto/aBwa...	tA-ci4UFiGVOmzsdRyEKfA	French onion soup - check!\nFresh oysters - ch...	5
3	KggnM_Z4wOa_JExunaaWHg	Le Temps des Cerises	['https://s3-media1.fl.yelpcdn.com/bphoto/g3Aa...	U8G3f8ITSlqBzET8DVO8bw	Foods were delicious, great ambiance, and help...	5
4	pztzge22A_c_BfzLHCmaMw	Le Bistro du Périgord	['https://s3-media2.fl.yelpcdn.com/bphoto/tc_w...	g84iMTg-NcSiUqcBgI-r8g	Had a delicious meal in such a cute and homey ...	5

NLP: Word embedding

Nous utilisons la method de word embedding **Word2Vec et FastText** afin de decouvrir les relations entre tokens

Grace a l dusieme method, nous pouvons observer les termes proches.