Avis Restau

Améliorez le produit IA de votre start-up

P6 - Sylvia Bankowska - Août 2022

Sommaire

1. Enjeux et objectifs
2. Détecter les sujets d’insatisfaction
3. Classification automatique d’images
4. Validation de la faisabilité
5. Synthèse

Enjeux et objectifs

Enjeux

mieux **comprendre les avis** postés

labelliser automatiquement les photos postées

Objectifs

analyser les avis => sujets d'insatisfaction

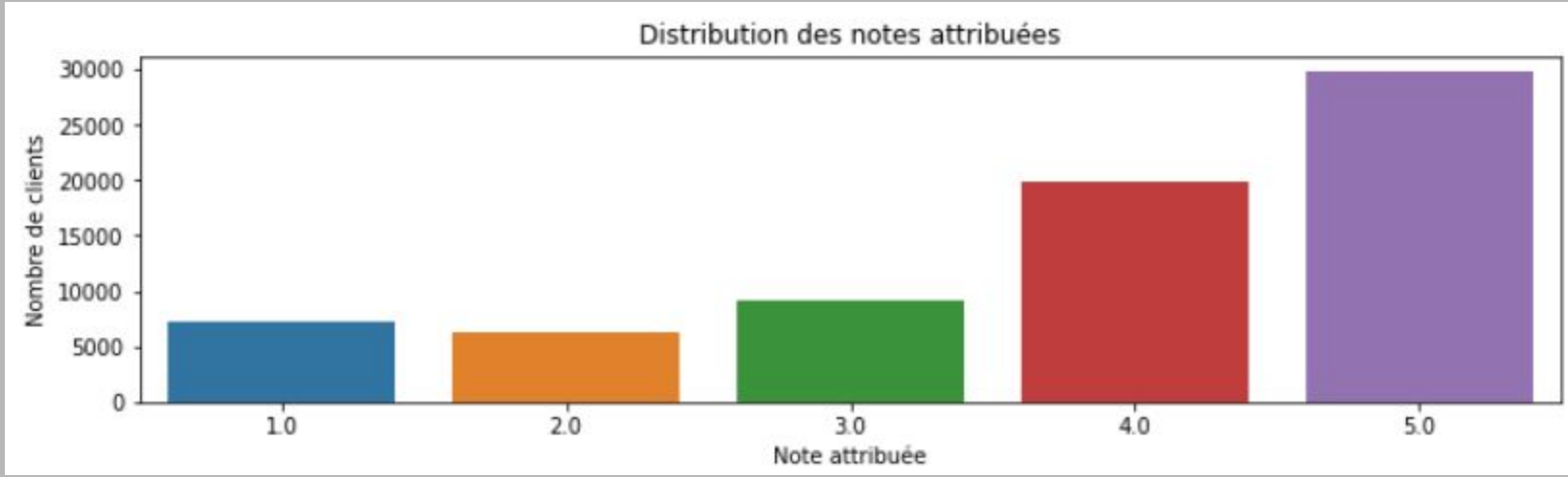
analyser les photos => étude de faisabilité : séparer les images selon la catégorie réelle

Détecter les sujets d’insatisfaction

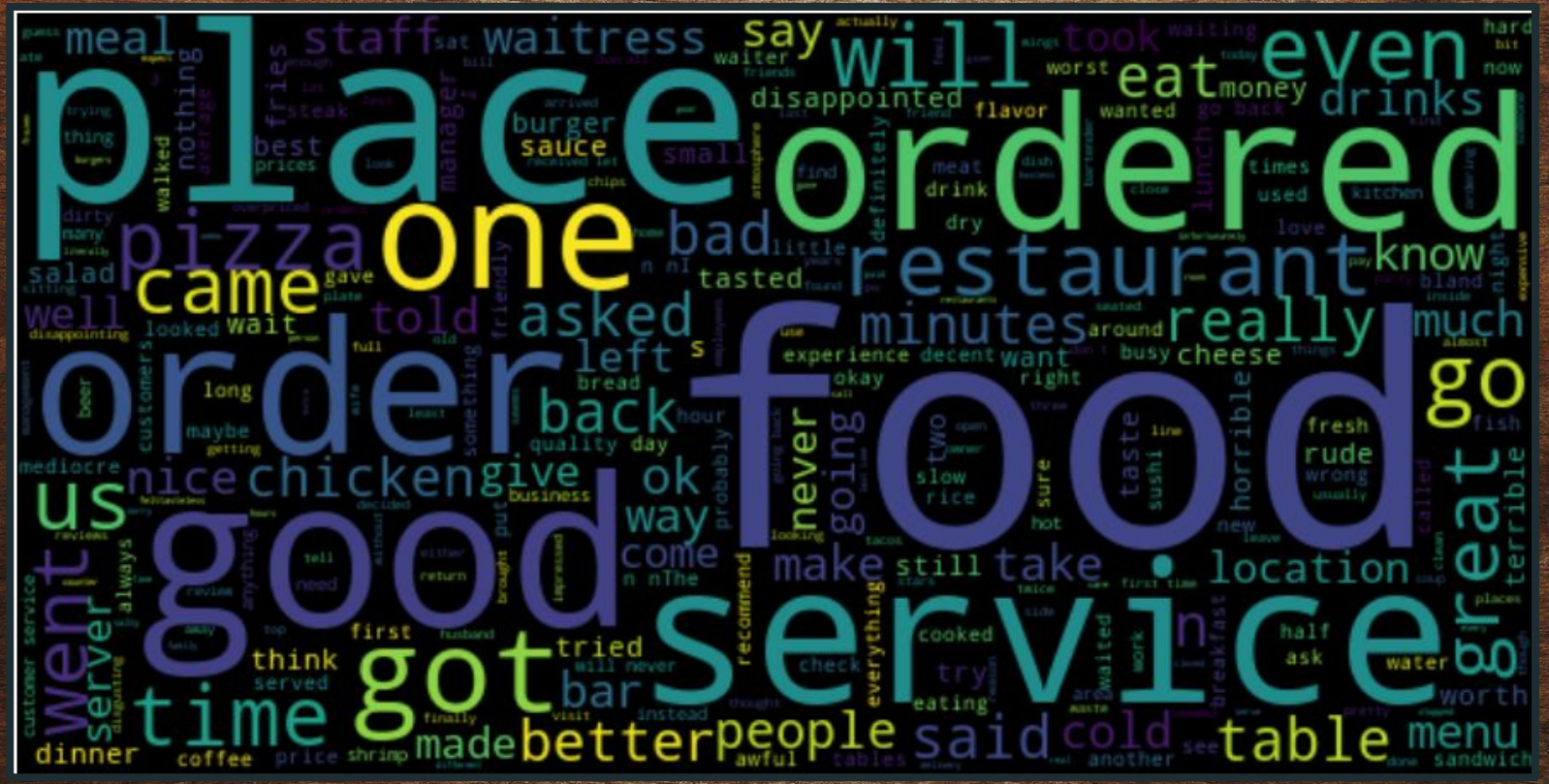
Dataset

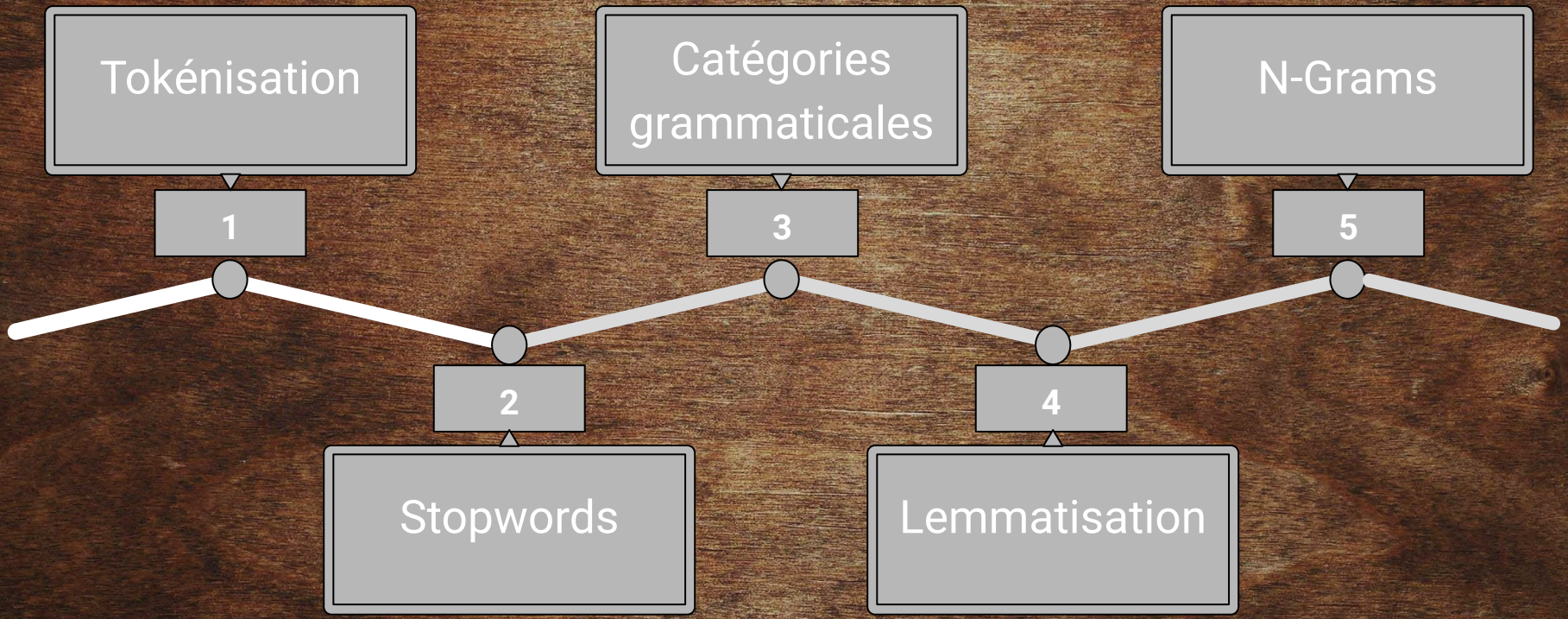
Jeu de données d'origine :
- **100 000 avis clients**

Echantillon :
- **1 et 2 étoiles : 13 536 avis**
- **avis trop longs : 9946 avis**



Avant le prétraitement

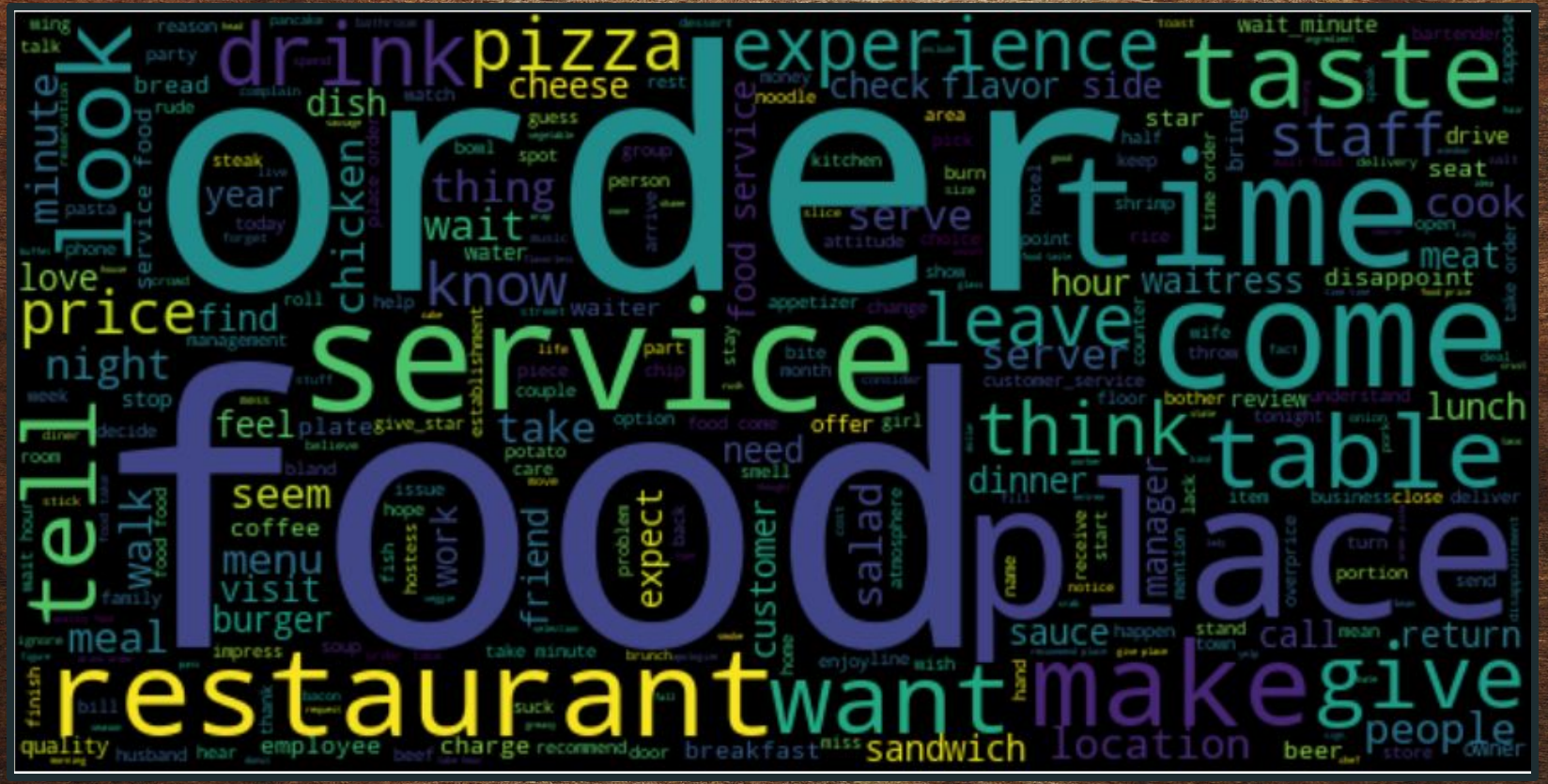




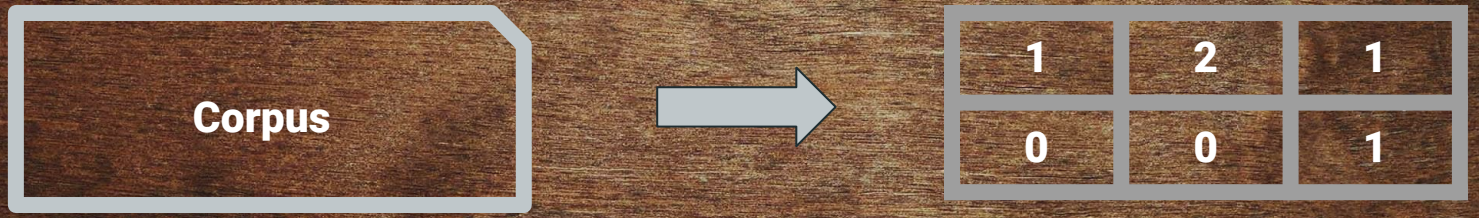

```
Avant :     Waited several minutes waiting to order. I was the only car waiting. An employee saw pull out didn't
Après :     ['wait', 'minute', 'wait', 'order', 'car', 'wait', 'employee', 'see', 'pull', 'say']

Avant :     Went there at 4am and there was only one waitress. I don't go to Denny's often but I didn't remember
Après :     ['go', 'waitress', 'denny', 'remember', 'food', 'get', 'menu', 'waitress', 'come', 'table']
```

Après le prétraitement



BOW



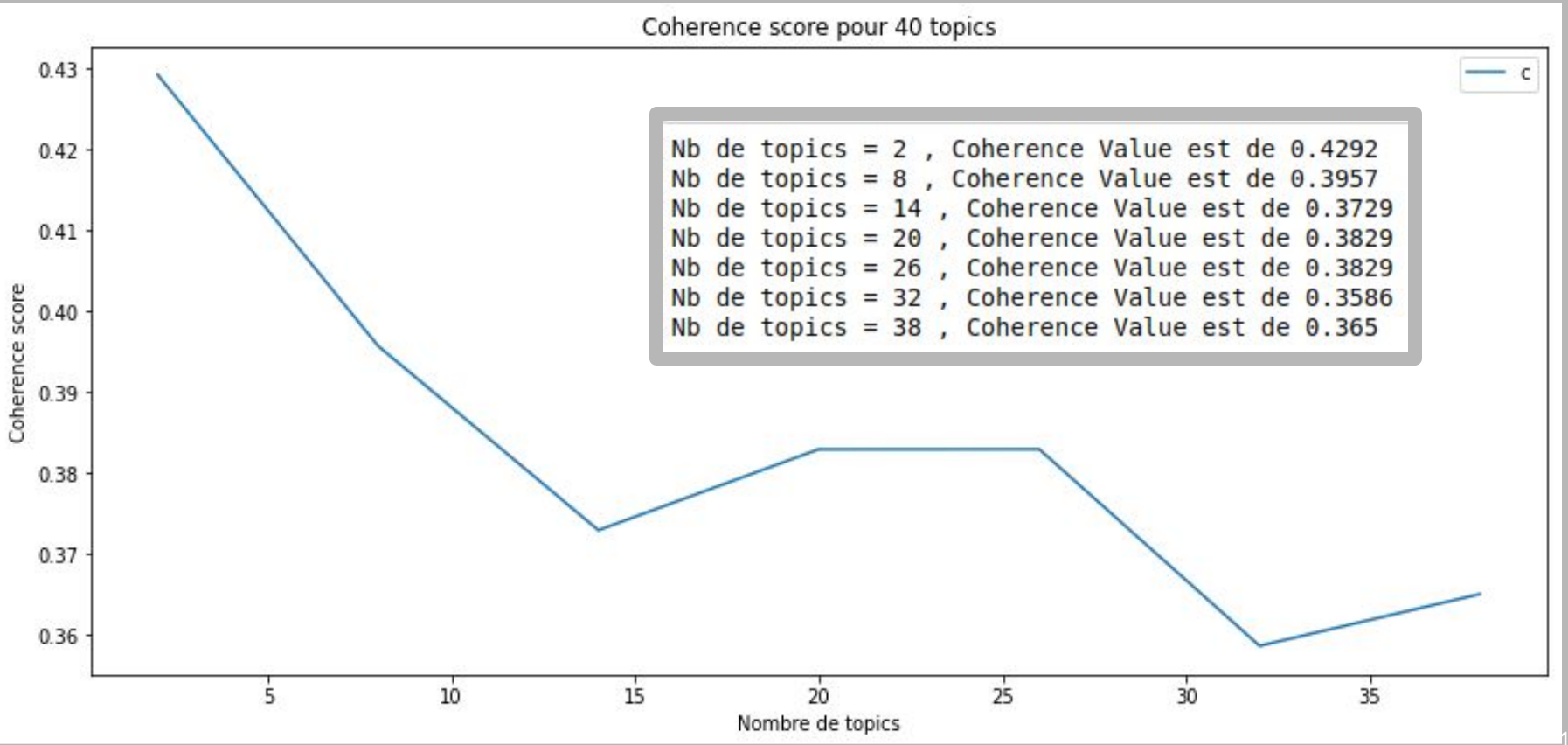
9946 documents à traiter

8909 tokens dans le dictionnaire

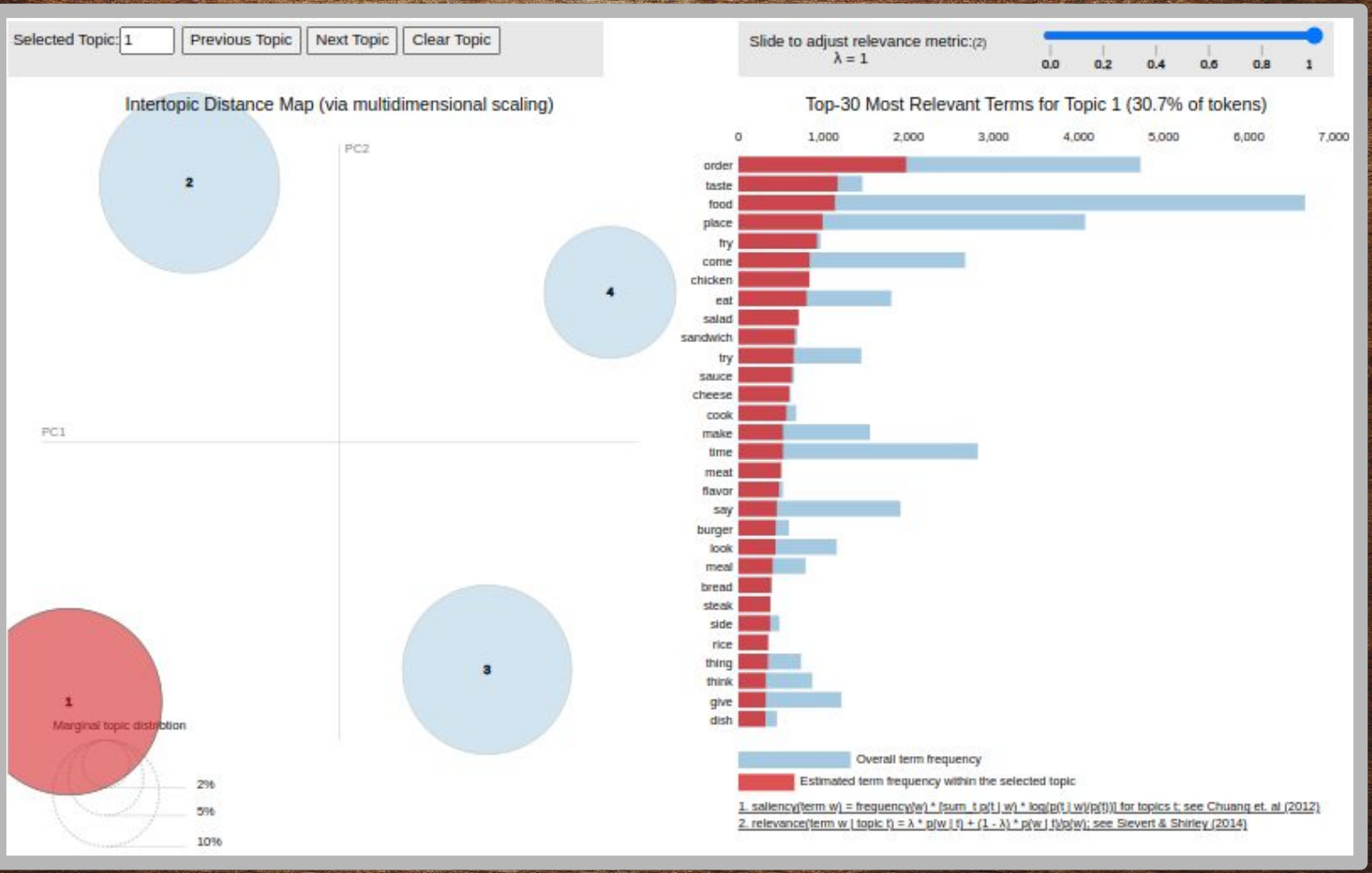
```
Le mot 30 ("waitress") apparaît 2 fois.
Le mot 119 ("hostess") apparaît 1 fois.
Le mot 219 ("friend") apparaît 1 fois.
Le mot 249 ("talk") apparaît 1 fois.
Le mot 259 ("seem") apparaît 1 fois.
Le mot 398 ("fiance") apparaît 1 fois.
Le mot 399 ("fry") apparaît 1 fois.
```

```
[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 2), (9, 2), (10, 1)]]
```

Coherence Score



LDA


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (30.7% of tokens)
0 1,000 2,000 3,000 4,000 5,000 6,000 7,000
order
taste
food
place
fry
come
chicken
eat
salad
sandwich
try
sauce
cheese
cook
make
time
meat
flavor
say
burger
look
meal
bread
steak
side
rice
thing
think
give
dish
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

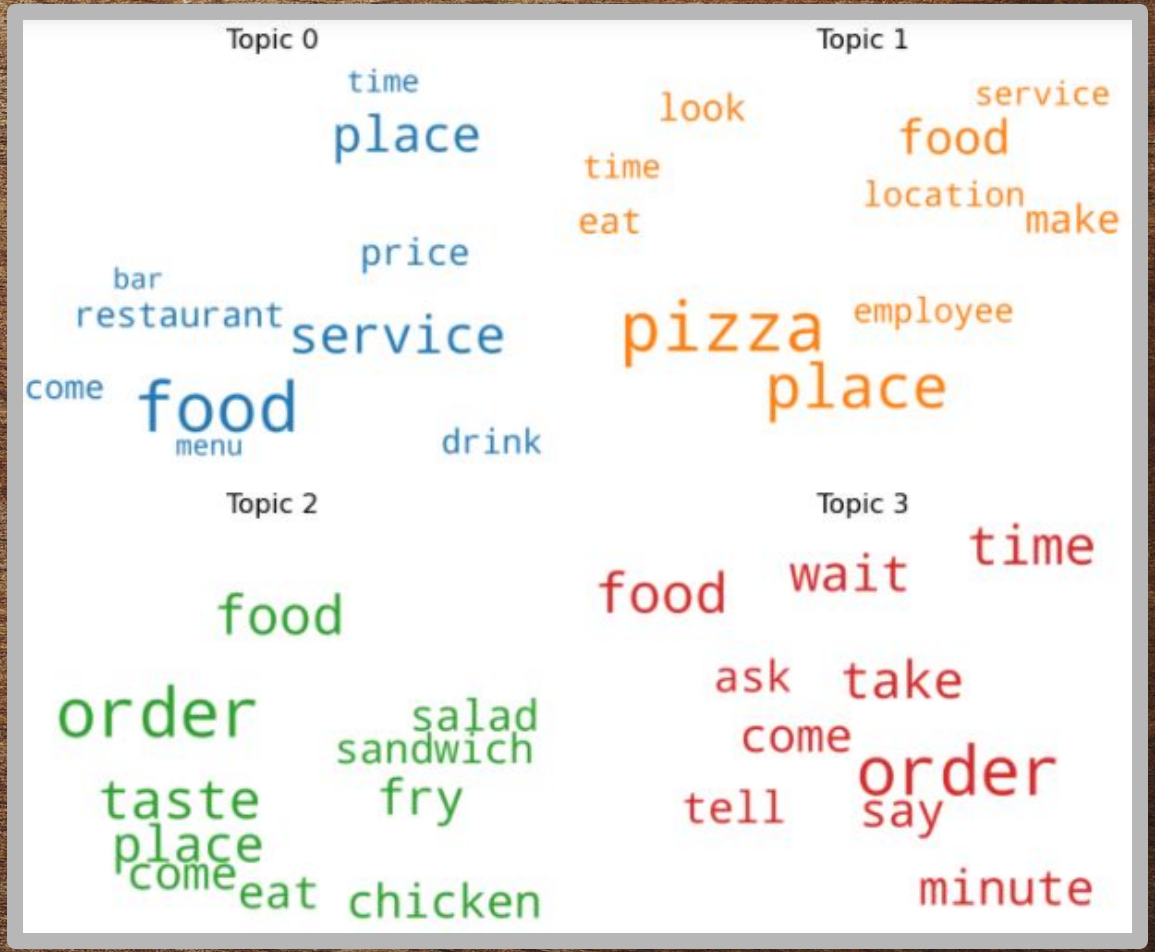
Topics identifiés

```
[(0,
  '0.067*"food" + 0.034*"service" + 0.031*"place" + 0.015*"price" + '
  '0.014*"restaurant" + 0.013*"drink" + 0.013*"come" + 0.011*"time" + '
  '0.009*"bar" + 0.009*"menu"'),
 (1,
  '0.039*"pizza" + 0.030*"place" + 0.016*"food" + 0.011*"make" + 0.010*"look"
  '+ 0.009*"eat" + 0.008*"location" + 0.008*"employee" + 0.008*"time" + '
  '0.007*"service"'),
 (2,
  '0.032*"order" + 0.019*"taste" + 0.018*"food" + 0.016*"place" + 0.015*"fry"
  '+ 0.014*"come" + 0.013*"chicken" + 0.013*"eat" + 0.011*"salad" + '
  '0.011*"sandwich"'),
 (3,
  '0.045*"order" + 0.027*"food" + 0.026*"time" + 0.024*"take" + 0.023*"wait" +
  '0.021*"minute" + 0.020*"say" + 0.020*"come" + 0.017*"tell" + 0.017*"ask"')]
```

Word Cloud

Documents par topic

	Nb Docs	% Docs
2.0	3137	0.3154
3.0	2786	0.2801
0.0	2657	0.2671
1.0	1366	0.1373



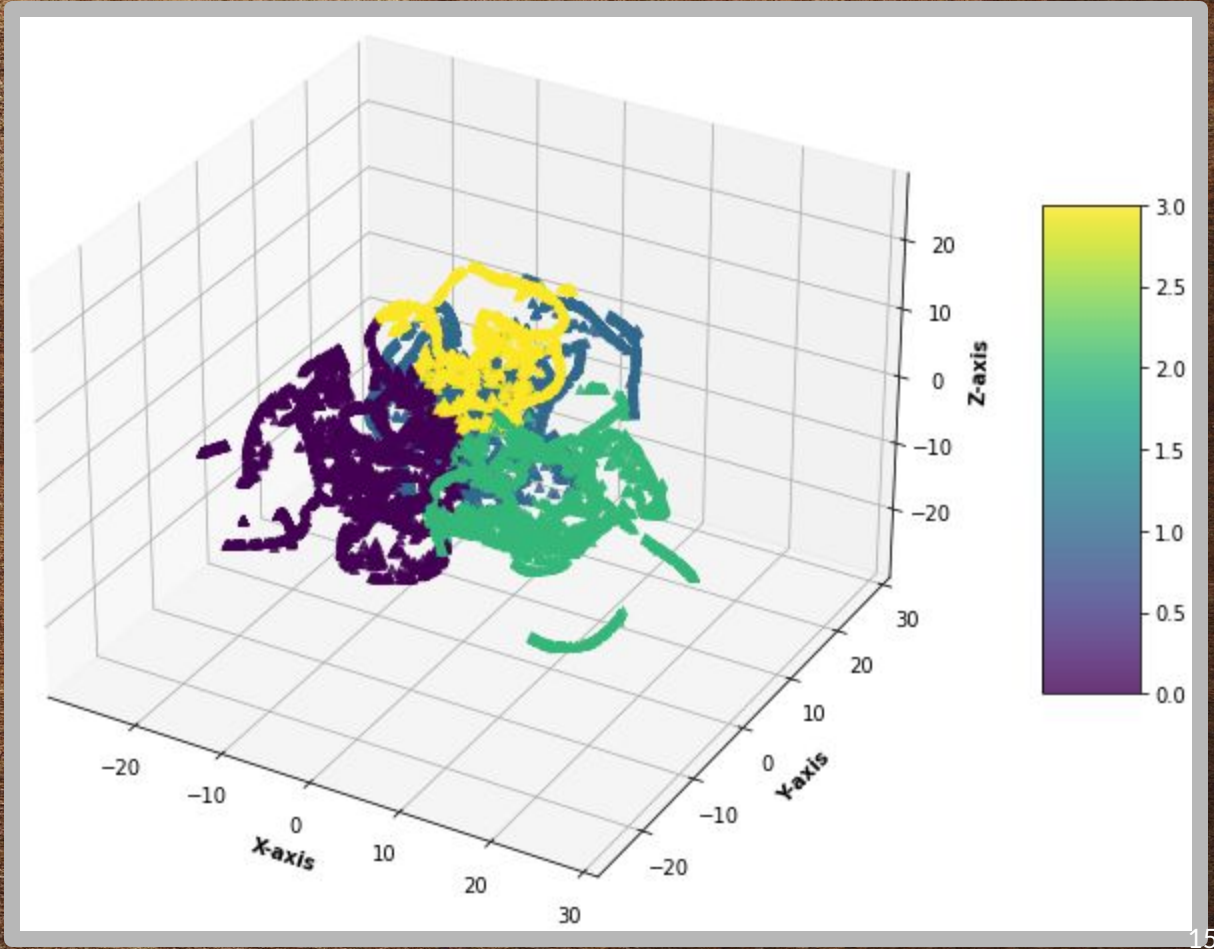
T-SNE

Cluster 0 :
order, taste, food

Cluster 1 :
order, food, time

Cluster 2 :
food, service, place

Cluster 3 :
pizza, place, food

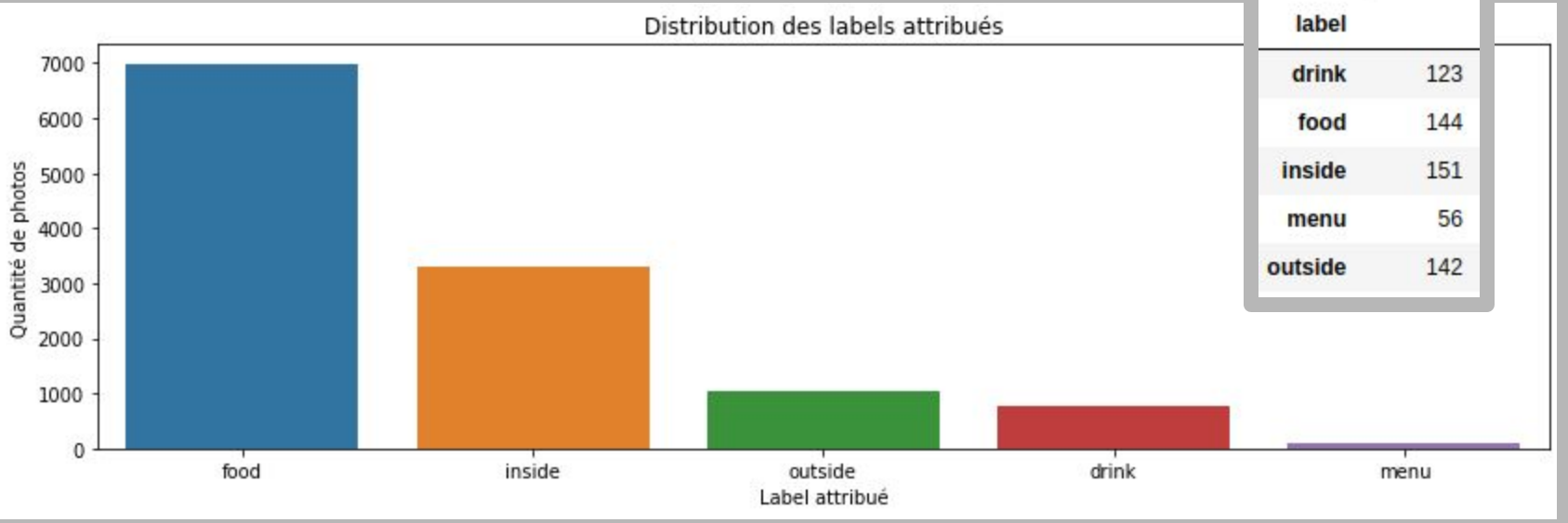


Classification automatique d’images

Dataset

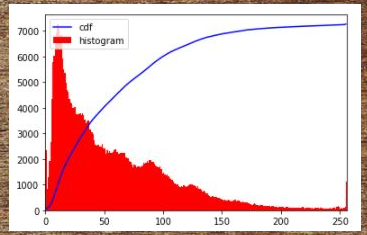
Jeu de données d'origine :
- ➢ **200 000 photos**

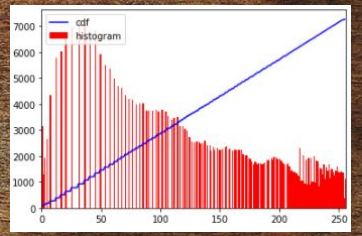
Echantillon :
- ➢ **5 labels**
- ➢ **plus de 100 photos par label**



label	photo_id
drink	123
food	144
inside	151
menu	56
outside	142

Prétraitement





Grayscale

Egalisation

Débruitage

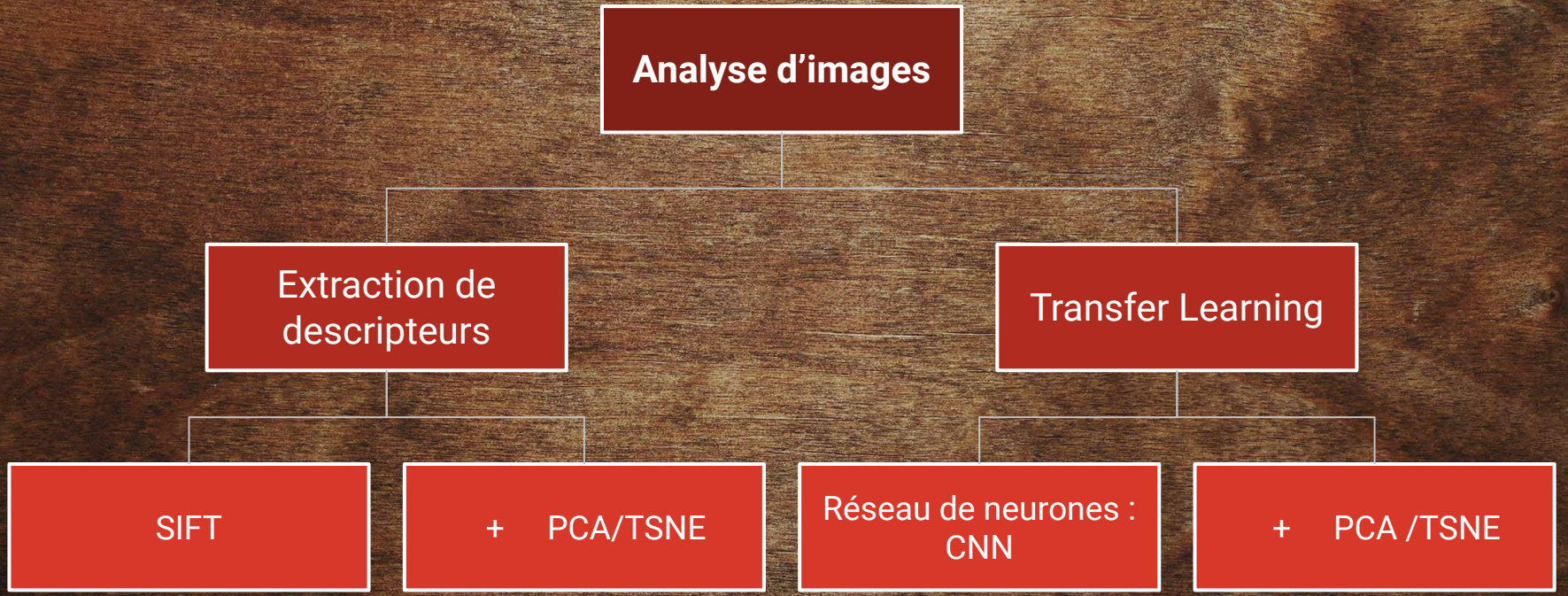
Image d'origine

Niveaux de gris avec 'rgb2gray'

Photo égalisée

Photo débruitée

Approches


Analyse d'images
Extraction de descripteurs
Transfer Learning
SIFT
+ PCA/TSNE
Réseau de neurones : CNN
+ PCA /TSNE

SIFT : descripteurs

Descripteurs par image

Descripteurs : (815, 128)

Descripteurs pour l'ensemble des images

Bag of visual words

1. Stockage des keypoints par image
2. Concaténation des keypoints de toutes les images

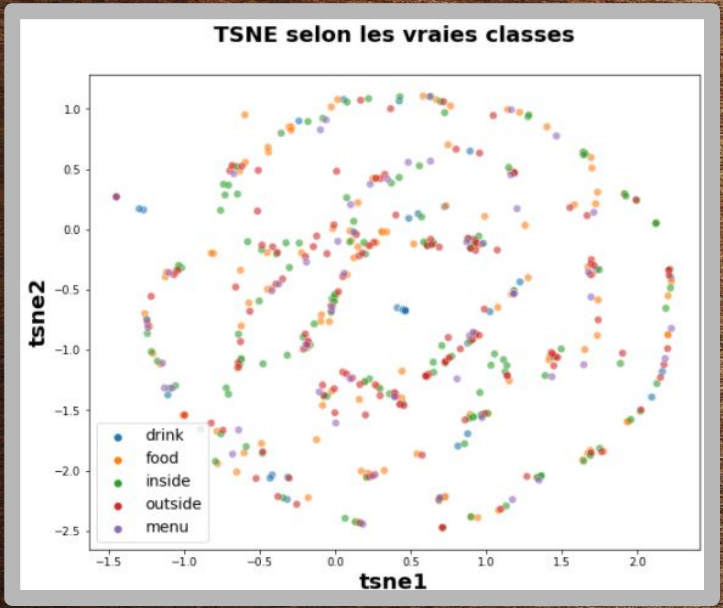
➢ Nombre d'images prétraitées : 616
➢ Nombre total de descripteurs : 692 384

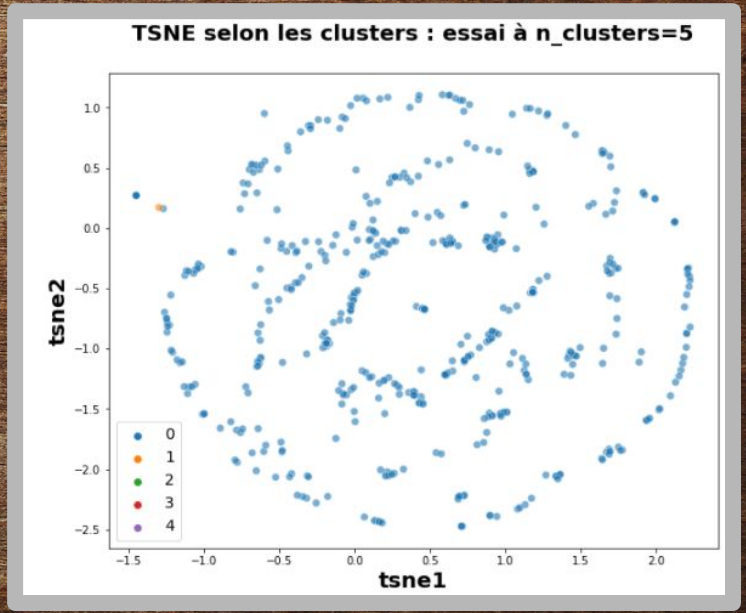
SIFT : clustering

MiniBatchKMeans avec nombre de k=832

Création de features d'images

Réduction de dimensions : PCA + clustering + T-SNE

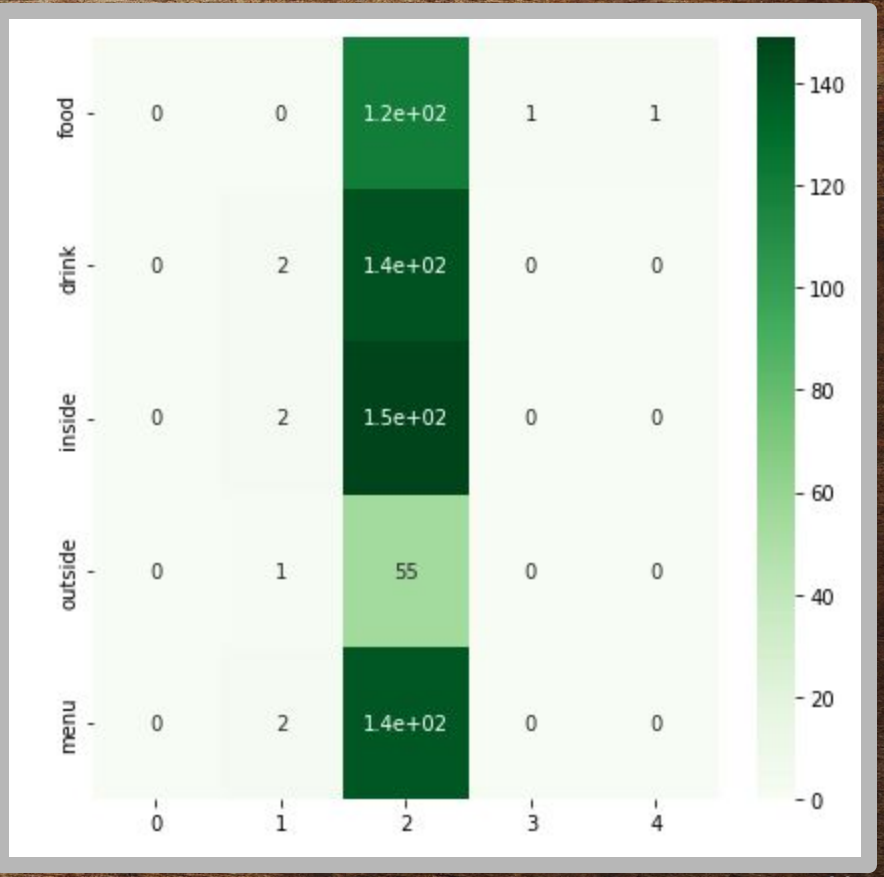




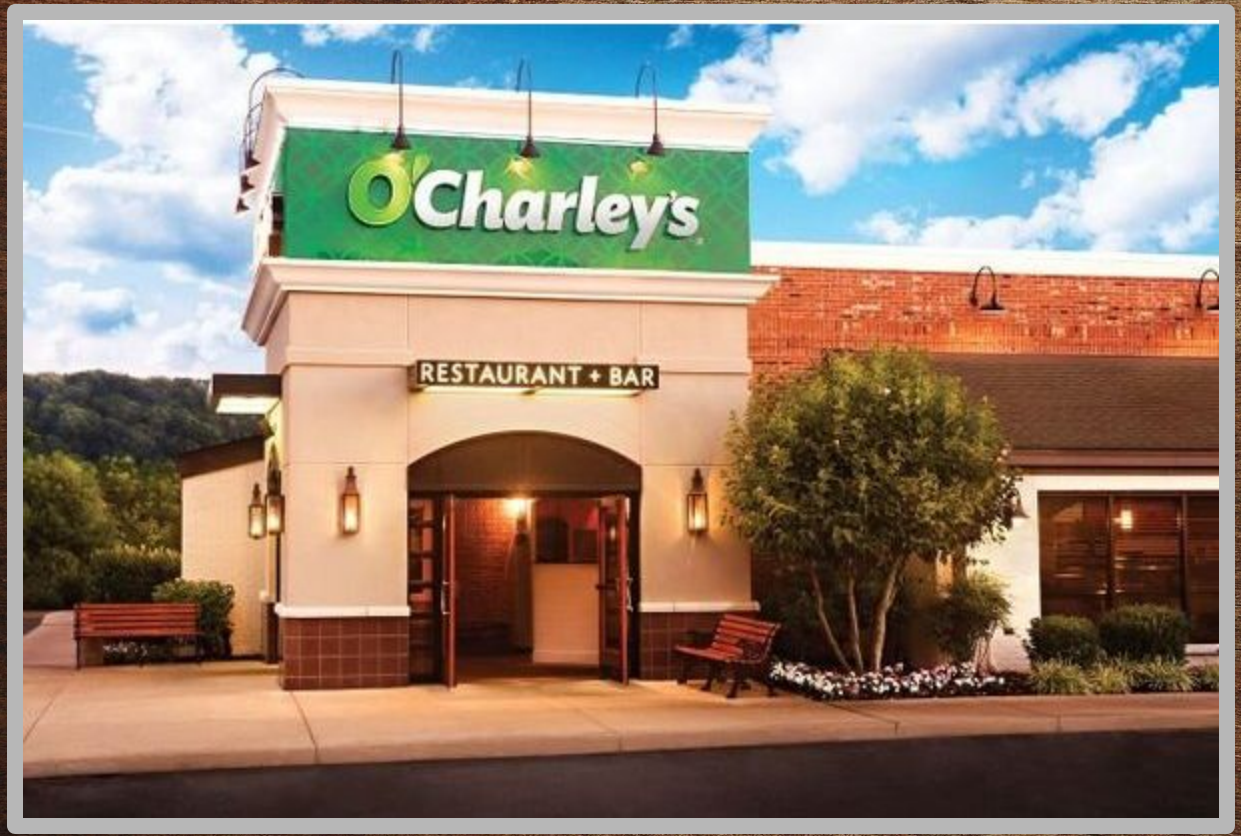
SIFT : résultats

- Score ARI : 0,0006
- Accuracy : 25%
- pas de classe bien identifiée

	precision	recall	f1-score	support
0	0.00	0.00	0.00	123
1	0.29	0.01	0.03	144
2	0.25	0.99	0.39	151
3	0.00	0.00	0.00	56
4	0.00	0.00	0.00	142
accuracy			0.25	616
macro avg	0.11	0.20	0.08	616
weighted avg	0.13	0.25	0.10	616



CNN : VGG16



Réseau pré-entraîné

1000 catégories d'ImageNet

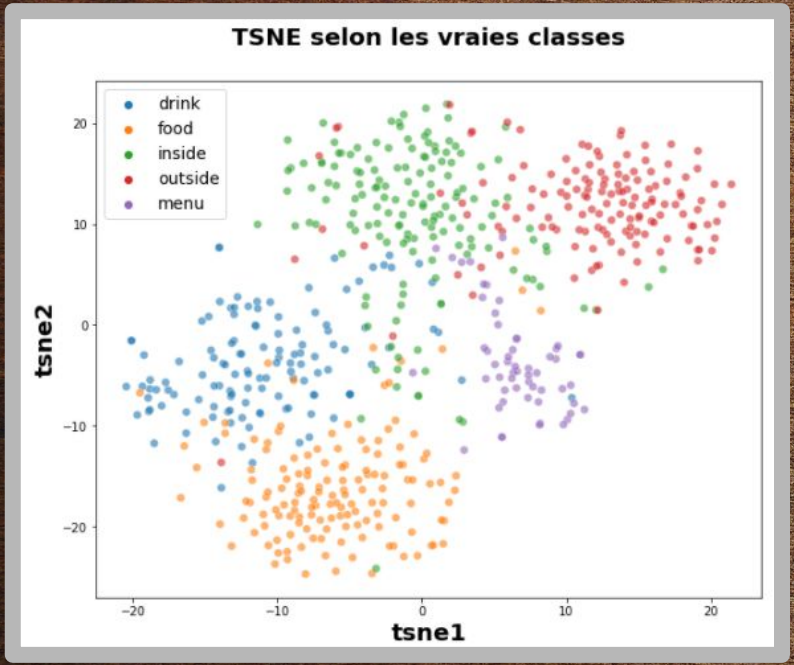
1. “cinéma” : 80%
2. “restaurant” : 3,8%
3. “mosquée” : 1,9%

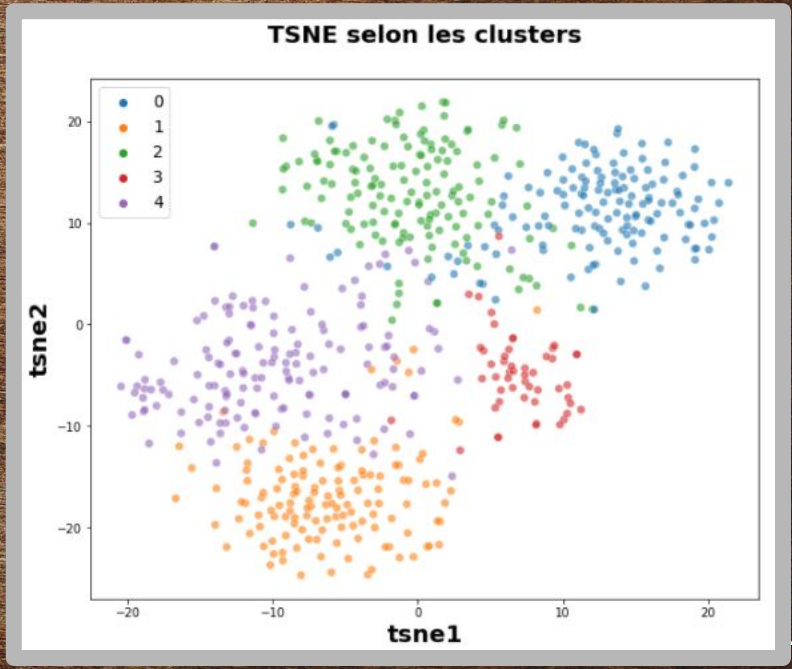
CNN Transfer Learning

Suppression des couches FC

Réseau VGG16 : programme d' extraction de features

Réduction de dimensions : PCA + clustering + TSNE

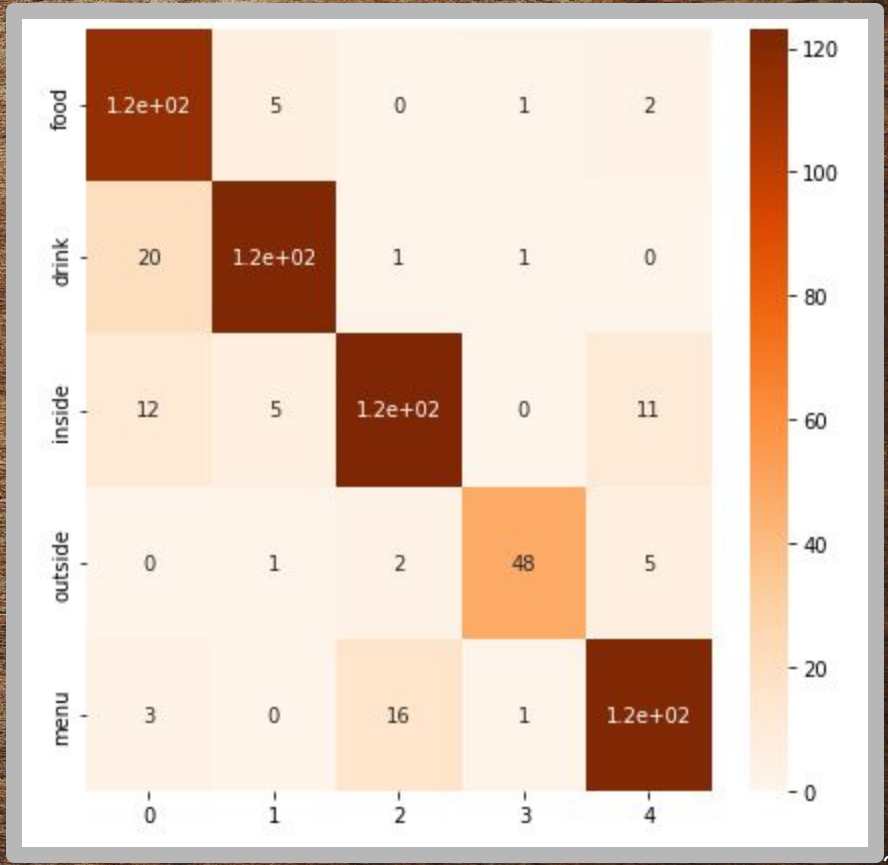




CNN Transfer Learning : résultats

- Score ARI : 0,67
- Accuracy : 86%
- 1 classe moins bien identifiée : “outside”

	precision	recall	f1-score	support
0	0.77	0.93	0.84	123
1	0.92	0.85	0.88	144
2	0.87	0.81	0.84	151
3	0.94	0.86	0.90	56
4	0.87	0.86	0.87	142
accuracy			0.86	616
macro avg	0.87	0.86	0.87	616
weighted avg	0.87	0.86	0.86	616



Validation de la faisabilité

Collecte de données : 600 avis & 200 photos

```
reviews.head()
```

	review_id	business_id	stars	text
0	gKnt3x8FFTduKx_UWakMVA	V7IXZKBDzScDeGB8JmnzSA	3	When you look up places to eat in New York City, Katz's Deli is one of the first places to come ...
1	G4wHamc9tvvKhmk9C0UonQ	V7IXZKBDzScDeGB8JmnzSA	5	Have love me some cats deli I moved to Virginia there is no one cat tap your pastrami sandwich a...
2	Bvf1ZXm4B2EdsAB4ht58Iw	V7IXZKBDzScDeGB8JmnzSA	5	Food was amazing - ordered the pastrami, Ruben, fries and, Matzo ball soup. Pastrami was the fam...
3	cnBXCpBlqXXbujF6caVkCg	44SY464xDHbvOcjDzRbKkQ	5	SO GOOD. Also fast service, super clean, very nice environment. Ramen was some of the best I've ...
4	f4gpIr9QnH-QFyWnRz8CQg	44SY464xDHbvOcjDzRbKkQ	5	Ippudo is my all-time favorite ramen spot- both in NYC and in Japan!\n\nMy fiancé and I stopped ...

```
photos.head()
```

	photo_url	business_id
0	https://s3-media1.fl.yelpcdn.com/bphoto/mrIdx2pZ3pR2UlqjKsSMZA/o.jpg	V7IXZKBDzScDeGB8JmnzSA
1	https://s3-media1.fl.yelpcdn.com/bphoto/zF3EgqHCk7zBUwD2B3WTEA/o.jpg	44SY464xDHbvOcjDzRbKkQ
2	https://s3-media1.fl.yelpcdn.com/bphoto/MYnXprCKOS0JlpQJRMOR7Q/o.jpg	xEnNFXtMLDF5kZDxfaCJgA
3	https://s3-media4.fl.yelpcdn.com/bphoto/xM4eGRjk_EfSc1V8MdkRXw/o.jpg	0CjK3esfpFcxIopebzjFxA
4	https://s3-media1.fl.yelpcdn.com/bphoto/d0XSKED0U0sTgFWhCQdY7w/o.jpg	4yPqqJDJOQX69gC66YUDkA

Image via l'API Yelp

Conclusions

► **Détection des sujets d'insatisfaction : DONE**

➢ définir les sources du mécontentement des clients

► **Labellisation automatique des photos : DONE**

➢ très bons résultats avec CNN Transfer Learning

➢ classification supervisée ultérieure est envisageable

► **Collecte de nouvelles données : DONE**

Questions ?