AMELIOREZ LE PRODUIT IA DE VOTRE START-UP

Présentation “Projet 6” chez “OPENCLASSROOM”

Jaoid KRAIRI

(Décembre 2021)

SOMMAIRE

Compréhension de la problématique métier :
1/ Rappel du contexte

Rappel du contexte

- ✓ Améliorer sa plateforme avec une nouvelle fonctionnalité de collaboration,
- ✓ Détecter les sujets d'insatisfaction,
- ✓ Labelliser automatiquement les photos postées,
- ✓ L'objectif.

Présentation de la problèmatique :
2/ Problèmatique

- Pas assez de données sur la plateforme Avis Restau,
- S’assurer de la possibilité de collecter de nouvelles données.

Description des jeux de données :
1/ Les documents mise à ma disposition

❑ Les documents fourni sont composés de six fichiers Json plus un dossier image :

- ✓ yelp_academic_dataset_business.json
- ✓ yelp_academic_dataset_review.json
- ✓ yelp_academic_dataset_checkin.json
- ✓ yelp_academic_dataset_tip.json
- ✓ yelp_academic_dataset_user.json
- ✓ photos.json
- ✓ dossier photos

❑ Les trois documents utiles plus le dossier image sont :

- ✓ **yelp_academic_dataset_business.json,**
- ✓ **yelp_academic_dataset_review.json,**
- ✓ **photos.json,**
- ✓ **Un dossier photos.**

Description des jeux de données :
2/ Importer sur API Yelp les avis clients des restaurants de Londres.

❑ S’assurer de la possibilité de collecter de nouvelles données :

- ✓ Collecter les informations relatives à environ 2032 restaurants de Londres,
- ✓ Créer un compte sur API Yelp et une application,
- ✓ Inscription à la version beta du développeur,
- ✓ 2 variables fixes comportant l’identité du client et la clé API,
- ✓ Effectuer sur mon notebook une requête Graphql,
- ✓ Définir le répertoire, transformé les données en fichier csv et sauvegarder.

Description des jeux de données :
3/ Définir mes 2 jeux de données(textuelles)

❑ Créer 2 jeux de données :

- ✓ Décompresser le dossier 'yelp_dataset.tar',
- ✓ Intégrer le fichier 'yelp_academic_dataset_business.json',
- ✓ Sélectionner uniquement la catégorie Restaurants,
- ✓ Le jeu de données comporte 163 observations et 14 variables,
- ✓ Garder uniquement 2 variables et supprimer les autres,
- ✓ Intégrer le fichier 'yelp_academic_dataset_review.json' par une fonction de rèduction de dimension et de filtre.

Description des jeux de données :
4/ Assembler mes 3 jeux de données

❑ Créer un seul jeu de données d'avis d'insatisfaction client:

- ✓ 2 jeux de données créé précédemment,
- ✓ Plus le jeu de données importer d'API yelp,
- ✓ Assembler mes 2 premiers jeux de données,
- ✓ Les 2 jeux restant,
- ✓ Extraire les avis d'insatisfaction client,
- ✓ La variable 'stars' < 3.

Description des jeux de données :
5/ Extraire le jeu de données utile(image)

❑ Créer un seul jeu de donnée brute en fichiers csv:

- ✓ Le fichier est trop volumineux,
- ✓ Instancier 5 variables(datafile, img_dir, outputfile, outputfile_raw et chunksize),
- ✓ Définir une boucle for,
- ✓ Récupérer mon second fichier csv(P6_01_fichiercsv_photos_02.csv),
- ✓ Le jeu de données contient 200 000 images,
- ✓ 5 catégories (interior, food, outside, menu et drink).

Description des jeux de données :
6/ Extraire le jeu de données d'entraiment(image)

❑ Un seul jeu de donnée d'entrainement utilisable en fichiers csv:

- ✓ Créer une variable 'label_num',
- ✓ Un jeu de données d'entrainement partiel de 500 images par catégorie,
- ✓ Créer une variable 'fichiers_photo',
- ✓ Supprimer la variable 'photo_id',
- ✓ Sauvegarder mon jeu de données d'entrainement(P6_01_fichiercsv_photos_03_train.csv)

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

1/ Analyser les commentaires pour détecter les différents sujets d'insatisfaction

1-A / Nuage de mots

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

1/ Analyser les commentaires pour détecter les différents sujets d'insatisfaction

1-B / Exemple de nettoyage sur un texte

'Ordered all veggies, because of my clean eating regimen and as I returned to work from my 1hr lunch I discovered that ALL my veggies were literally swimming in OIL & GREASE!!! I was more than pissed, but knew it would be a lost effort to attempt to take anything back. I did in turn call the business and make a complaint. Wont return. #LostCustomer'

['ordered', 'all', 'veggies', 'because', 'of', 'my', 'clean', 'eating', 'regimen', 'and', 'as', 'i', 'returned', 'to', 'work', 'from', 'my', '1hr', 'lunch', 'i', 'discovered', 'that', 'all', 'my', 'veggies', 'were', 'literally', 'swimming','in', 'oil', 'grease', 'i', 'was', 'more', 'than', 'pissed', 'but', 'knew', 'it', 'would', 'be', 'a', 'lost', 'effort', 'to', 'attempt', 'to', 'take', 'anything', 'back', 'i','did','in','turn','call','the','business','and','make','a',
'complaint','wont','return','lostcustomer']

['ordered', 'all','veggie','because','of','my','clean','eating','regimen','and','a','i','returned','to', 'work', 'from', 'my', '1hr', 'lunch', 'i', 'discovered', 'that', 'all', 'my', 'veggie', 'were', 'literally', 'swimming', 'in', 'oil', 'grease', 'i', 'wa', 'more', 'than', 'pissed', 'but', 'knew', 'it', 'would', 'be', 'a', 'lost', 'effort', 'to', 'attempt', 'to', 'take', 'anything', 'back', 'i', 'did', 'in', 'turn', 'call', 'the', 'business', 'and', 'make', 'a', 'complaint', 'wont','return', 'lostcustomer']

['ordered', 'veggie', 'clean', 'eating', 'regimen', 'returned', 'work', '1hr', 'lunch', 'discovered', 'veggie', 'literally', 'swimming', 'oil', 'grease', 'pissed', 'knew', 'would', 'lost', 'effort', 'attempt', 'take', 'anything', 'back','turn', 'call', 'business', 'make', 'complaint', 'wont', 'return', 'lostcustomer']

[(11, 1), (19, 1), (60, 1), (61, 1), (62, 1), (63, 1), (64, 1), (65, 1), (66, 1), (67, 1), (68, 1), (69, 1), (70, 1), (71, 1), (72, 1), (73, 1), (74, 1), (75, 1), (76, 1), (77, 1), (78, 1), (79, 1), (80, 1), (81, 1), (82, 1), (83, 1), (84, 1), (85, 2), (86, 1), (87, 1), (88, 1)]

Texte brute

En minuscule, enlèvement des ponctuation, tokenisation du texte

Lemmatisation du texte tokeniser

Supprimer les mots vides et retirer les mots du texte lemmatiser de moins de 3 lettres

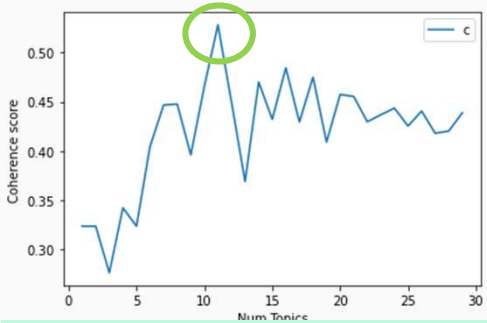
Représentation bag of words

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

1/ Analyser les commentaires pour détecter les différents sujets d'insatisfaction

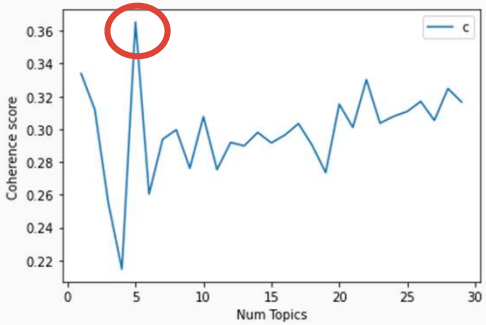
1-C / Comparer 4 modéles (LDA, LDA mallet, NMF et K-Means)

LDA



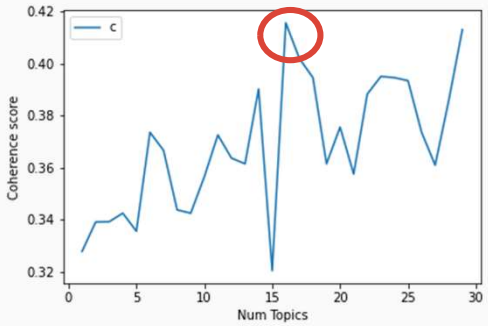
Nbre Topics = 11
valeur de cohérence de 0.5284

LDA mallet



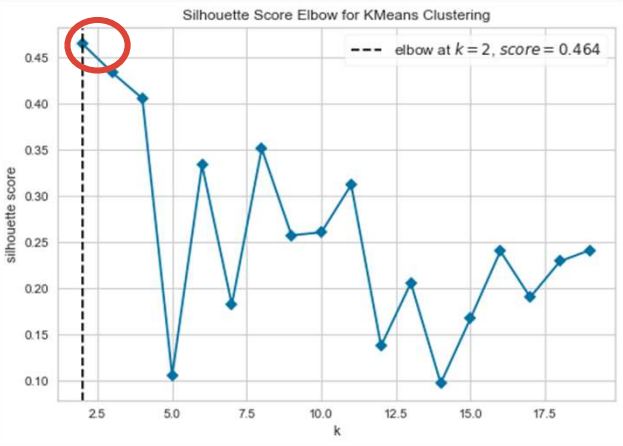
Nbre Topics = 5
valeur de cohérence de 0. 3653

NMF



Nbre Topics = 16
valeur de cohérence de 0.4156

K-Means

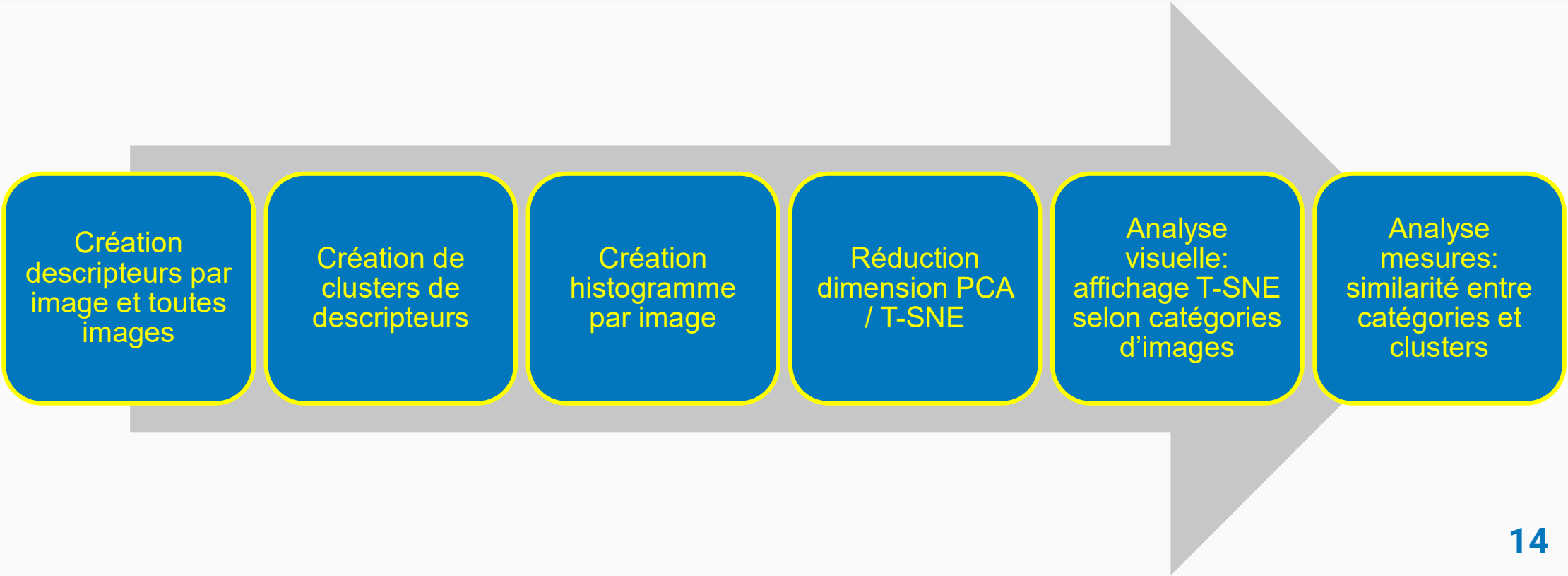


Nbre de clusters = 2
valeur de silouhette de 0.464

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1 / Démarche générale (SIFT)



Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

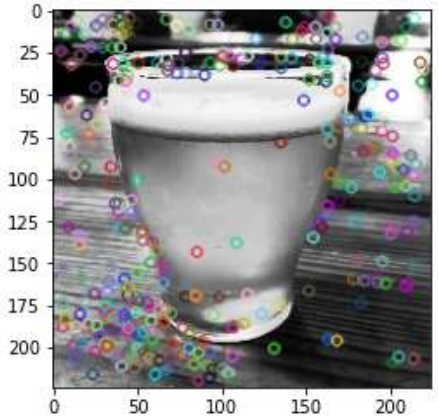
2/ Analyser les photos pour déterminer les catégories des photos

2-1-1 / Création descripteurs par image et toutes images

- 1. Pour chaque image passage en gris et equalization
- 2. Création d'une liste de descripteurs par image qui sera utilisée pour réaliser les histogrammes par image
- 3. Création d'une liste de descripteurs pour l'ensemble des images qui sera utilisée pour créer les clusters de descripteurs

```
sift_keypoints_by_img = np.asarray(sift_keypoints)
sift_keypoints_all    = np.concatenate(sift_keypoints_by_img, axis=0)
```

```
Nombre de descripteurs :  (1174107, 128)
temps de traitement SIFT descriptor :              32.83 secondes
```




```
Descripteurs :  (339, 128)

[[ 19.   7.   0. ...  37.   0.   0.]
 [ 13.  41.  56. ...   0.   0.   0.]
 [  1.   0.   0. ...  19.  35.  30.]
 ...
 [ 14.  88.  30. ...  40.  19.   0.]
 [ 16.  50.  12. ...   0.   0.   1.]
 [111.  85.   0. ...   0.   1.   2.]]
```

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1-2 / Création de clusters de descripteurs

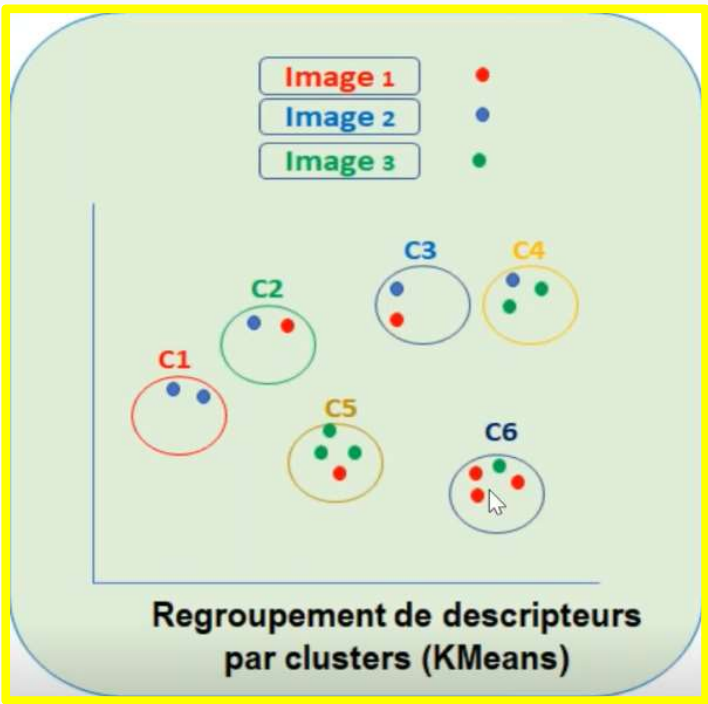
- **Utilisation de MiniBatchKMeans pour obtenir des temps de traitement raisonnables**
- **Nombre de clusters = à la racine carrée du nombre total de descripteurs**

```
# Détermination du nombre de clusters
temps1=time.time()

k = int(round(np.sqrt(len(sift_keypoints_all)),0))
print("Nombre de clusters estimés : ", k)
print("Création de",k, "clusters de descripteurs ...")

# Clustering
kmeans = cluster.MiniBatchKMeans(n_clusters=k, init_size=3*k, random_state=0)
kmeans.fit(sift_keypoints_all)

duration1=time.time()-temps1
print("temps de traitement kmeans : ", "%15.2f" % duration1, "secondes")
```

```
Nombre de clusters estimés :  1084
Création de 1084 clusters de descripteurs ...
```

```
C:\Users\JK253\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:887: UserWa
ry leak on Windows with MKL, when there are less chunks than available threads. Y
2 or by setting the environment variable OMP_NUM_THREADS=1
  warnings.warn(
```

```
temps de traitement kmeans :          131.55 secondes
```



Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

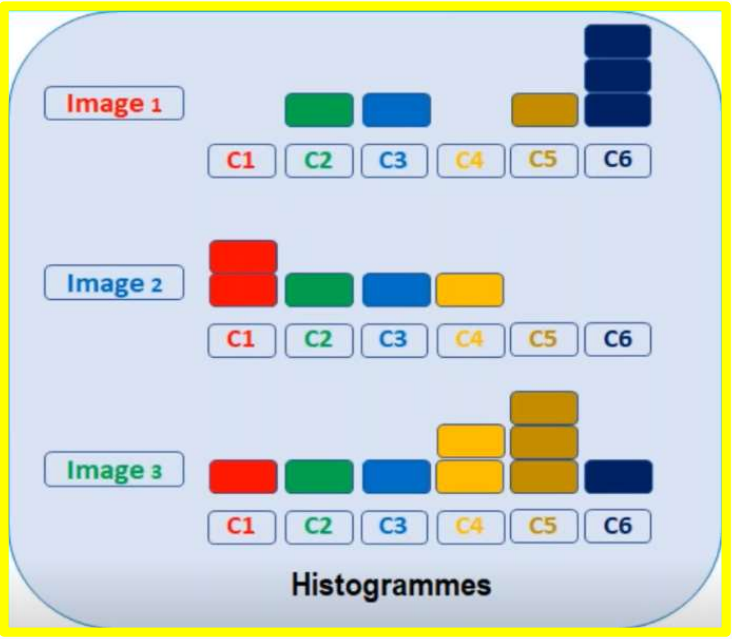
2-1-3 / Création histogramme par image

- ❑ **Pour chaque image :**

- ✓ **Prédiction des numéros de cluster de chaque descripteur**

- ✓ **Création d'un histogramme = au comptage pour chaque numéro de cluster du nombre de descripteurs de l'image**

```
def build_histogram(kmeans, des, image_num):
    res = kmeans.predict(des)
    hist = np.zeros(len(kmeans.cluster_centers_))
    nb_des=len(des)
    if nb_des==0 : print("problème histogramme image  : ", image_num)
    for i in res:
        hist[i] += 1.0/nb_des
    return hist
```



Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1-4 / Réduction dimension PCA / T-SNE

- ❑ **PCA :**
- ✓ **Création de features décorrélées entre elles**
- ✓ **Maintien d'un niveau de variance expliquée élevé (99%)**
- ✓ **Diminution de la dimension :**
  - ○ **Meilleure séparation des données via le T-SNE**
  - ○ **Réduction du temps de traitement du T-SNE**

```
Dimensions dataset avant réduction PCA :  (2500, 1084)
Dimensions dataset après réduction PCA :  (2500, 973)
```

- ❑ **T-SNE :**
- ✓ **Réduction de dimension T-SNE en 2 composantes pour affichage en 2D des images**

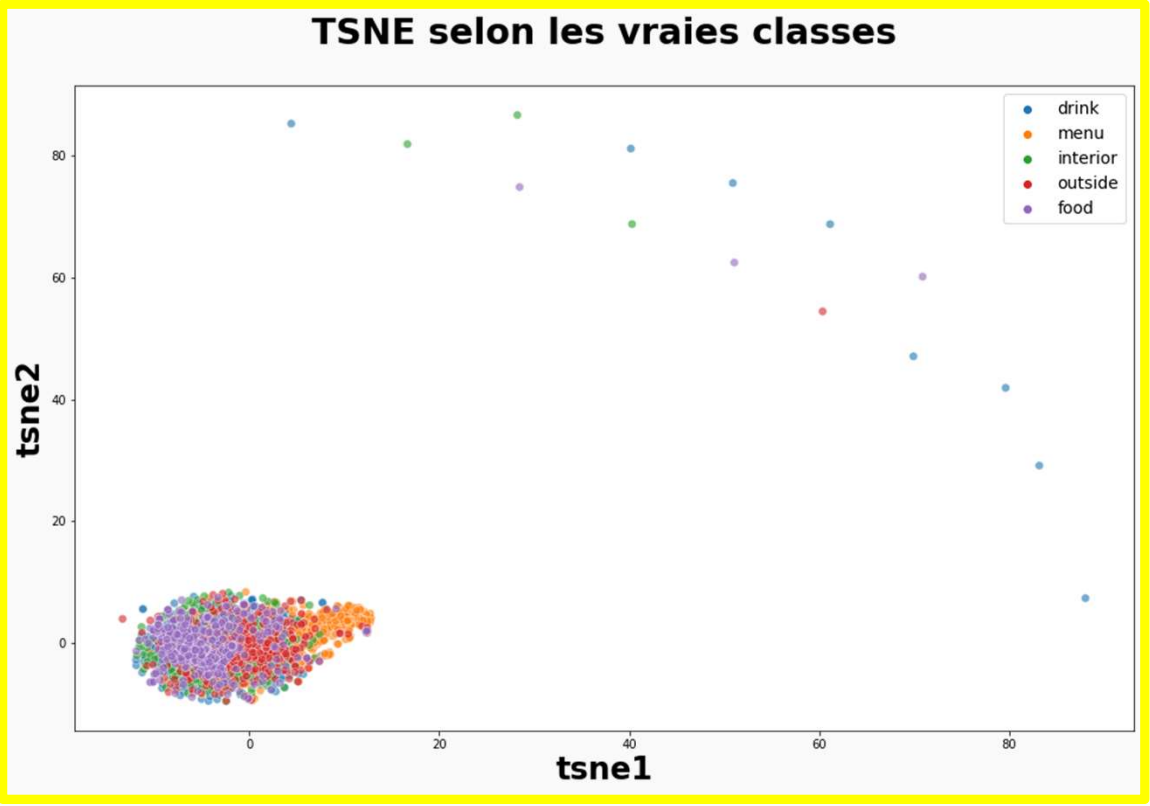
```
tsne = manifold.TSNE(n_components=2, perplexity=30,
                     n_iter=2000, init='random', random_state=6)
X_tsne = tsne.fit_transform(feat_pca)
```

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1-5 / Analyse visuelle: affichage T-SNE selon catégories d'images

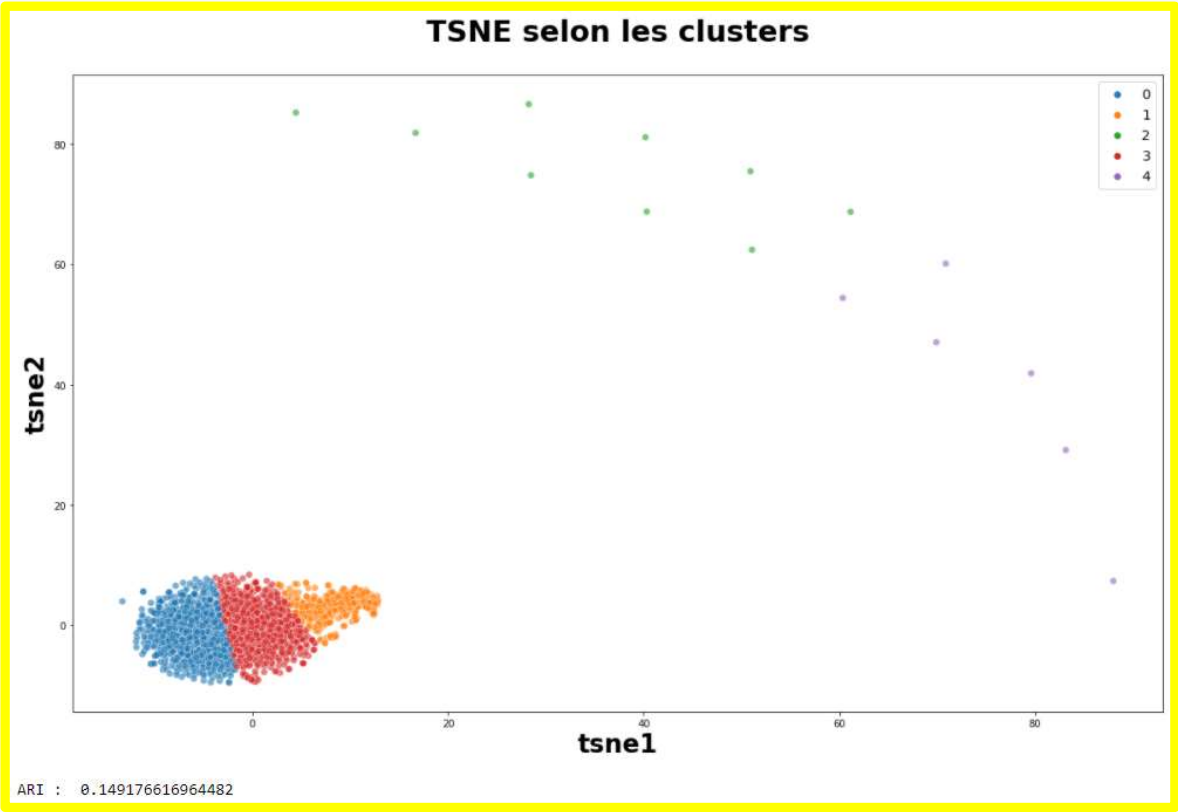
- Séparation partielle des classes food, outside et menu

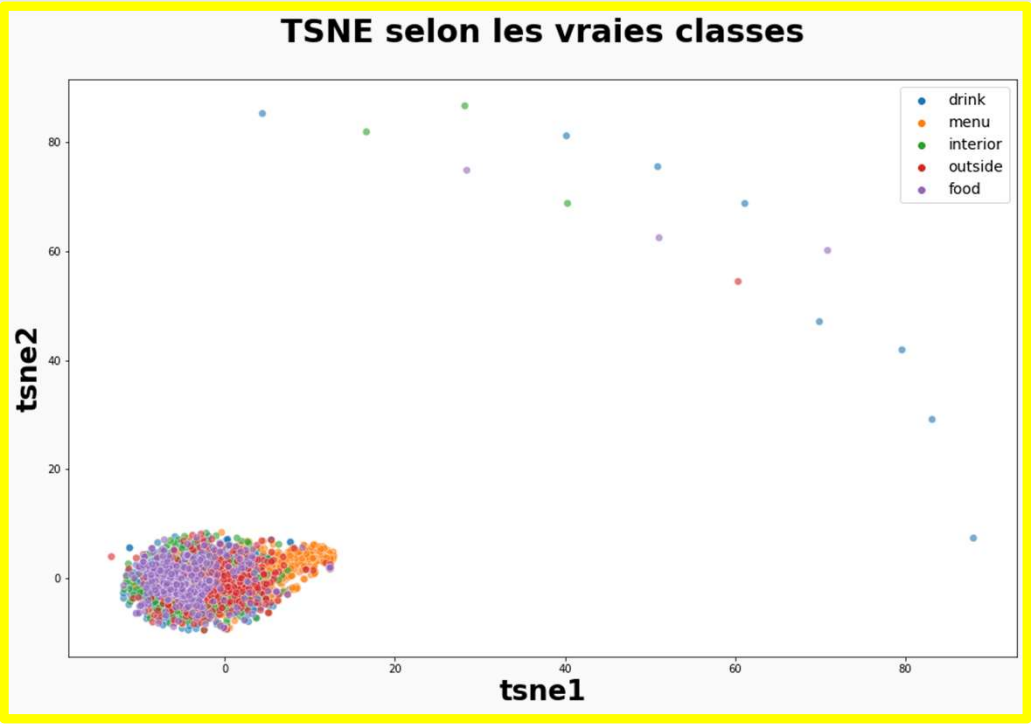


Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1-6 / Analyse mesures: similarité entre catégories et clusters classique(1/2)

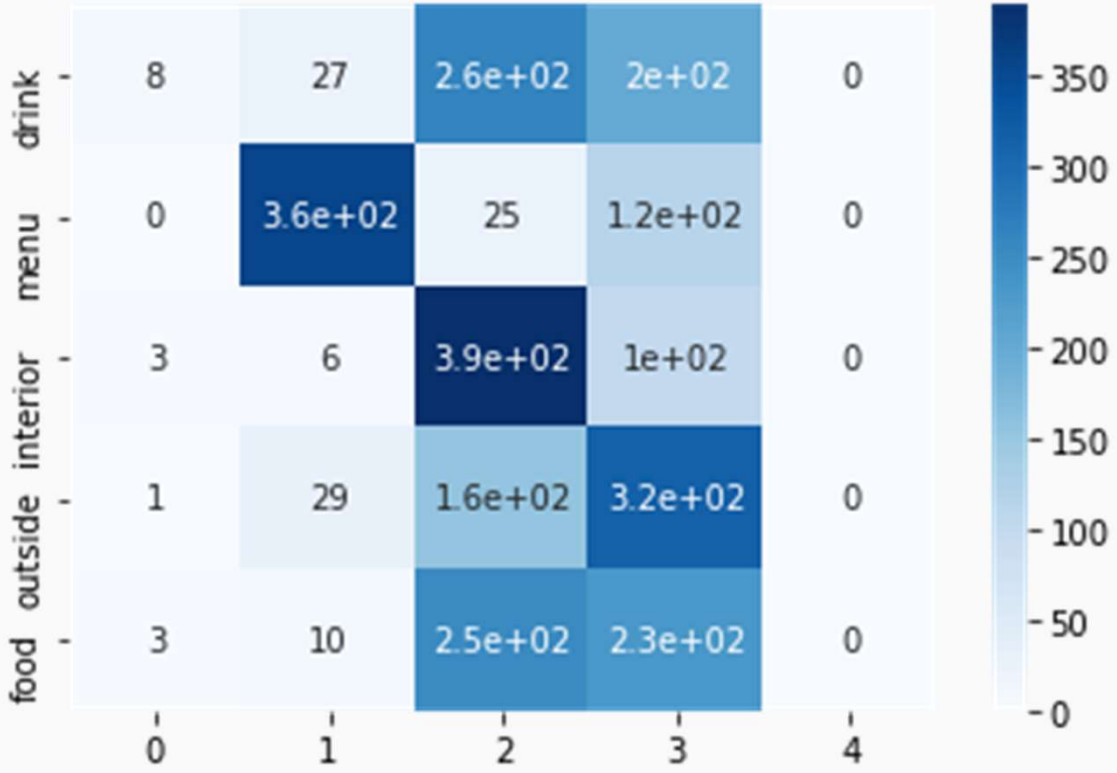




Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-1-6 / Analyse mesures: similarité entre catégories et clusters réelle(2/2)



Adjusted Randscore = 0. 149

	precision	recall	f1-score	support
0	0.53	0.02	0.03	500
1	0.83	0.72	0.77	500
2	0.36	0.78	0.49	500
3	0.33	0.63	0.43	500
4	0.00	0.00	0.00	500
accuracy			0.43	2500
macro avg	0.41	0.43	0.34	2500
weighted avg	0.41	0.43	0.34	2500

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-2 / CNN Transfer Learning

- ❑ **Mettre en œuvre une démarche de pré-traitement des images basée sur du CNN Transfer Learning :**

- ✓ **Utilisation du modèle VGG-16 de deep learning pré-entraîné sur des millions d'images**

- ✓ **La stratégie utilisée est l'extraction de features, sans entraînement, faire un predict pour chaque image afin de créer des features**

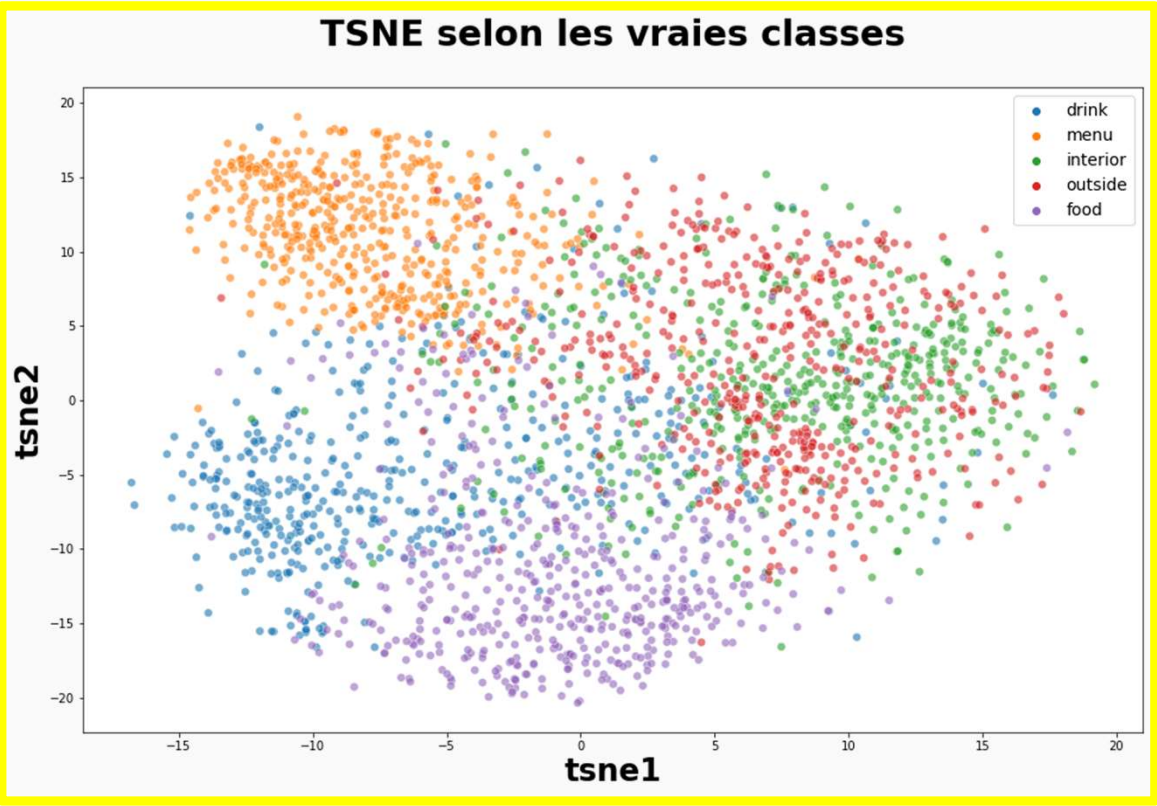
- ✓ **Réaliser la même démarche de réduction de dimension effectuée précédemment avec le SIFT**

Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-2-1- / Analyse visuelle: affichage T-SNE selon catégories d'images

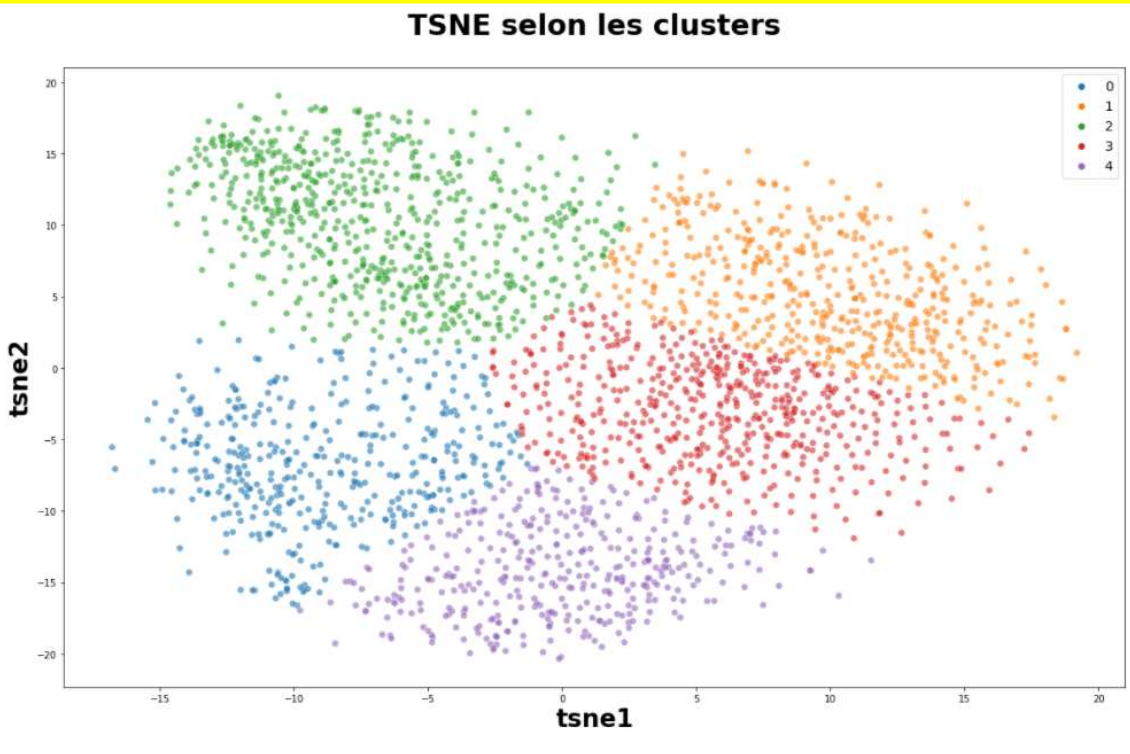
- Séparation partiel des classes food, drink, interior, outside et menu



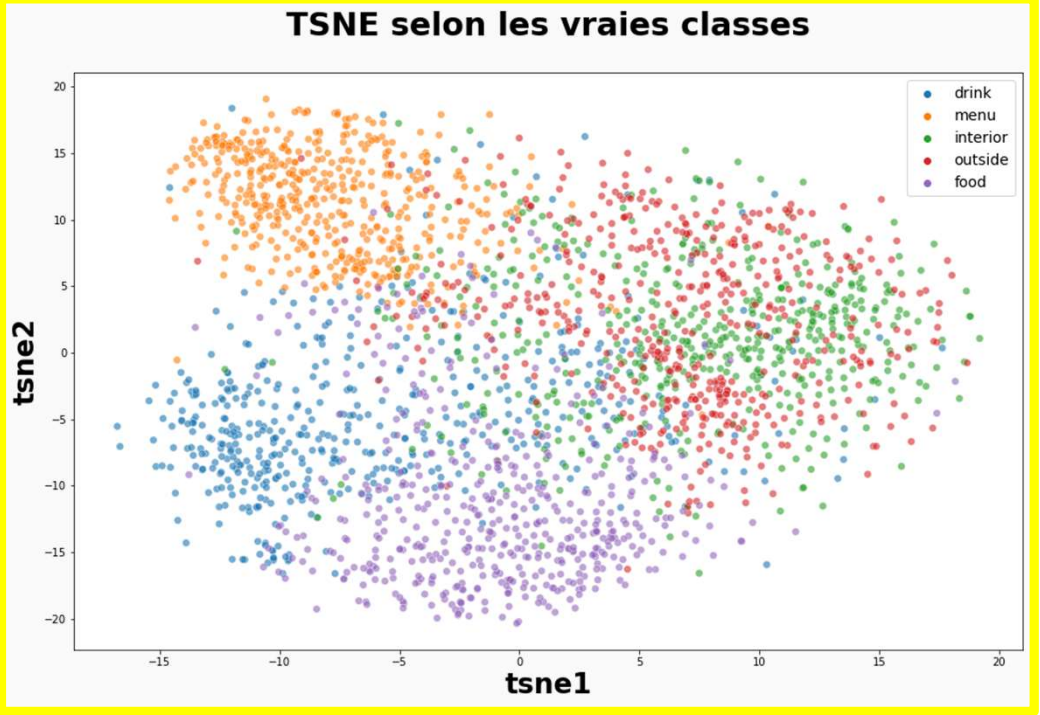
Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-2-2- / Analyse mesures: similarité entre catégories et clusters classique(1/2)



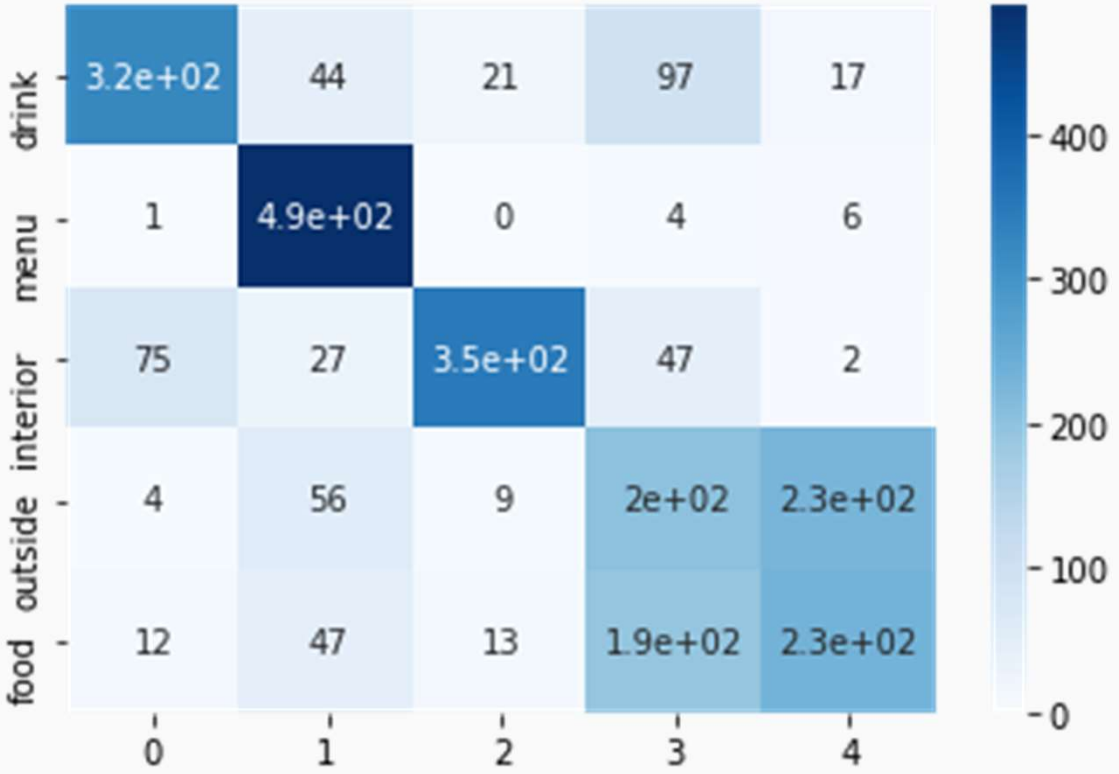
ARI : 0.4095655464016818



Visualisations graphiques (à l'aide d'une page web générée grâce au package Voilà)

2/ Analyser les photos pour déterminer les catégories des photos

2-2-3 / Analyse mesures: similarité entre catégories et clusters réelle(2/2)



Adjusted Randscore = 0. 409

	precision	recall	f1-score	support
0	0.78	0.64	0.70	500
1	0.74	0.98	0.84	500
2	0.89	0.70	0.78	500
3	0.37	0.41	0.39	500
4	0.48	0.47	0.47	500
accuracy			0.64	2500
macro avg	0.65	0.64	0.64	2500
weighted avg	0.65	0.64	0.64	2500

Conclusion sur la faisabilité

1/ Analyser les commentaires pour détecter les différents sujets

Mon choix : Modèle LDA avec 11 topics

```
'Ordered all veggies, because of my clean eating regimen and as I returned to work from my 1hr lunch I discovered that ALL my veggies were literally swimming in OIL & GREAS
E!!! I was more than pissed, but knew it would be a lost effort to attempt to take anything back. I did in turn call the business and make a complaint. Wont return. #LostCus
tomer'
```

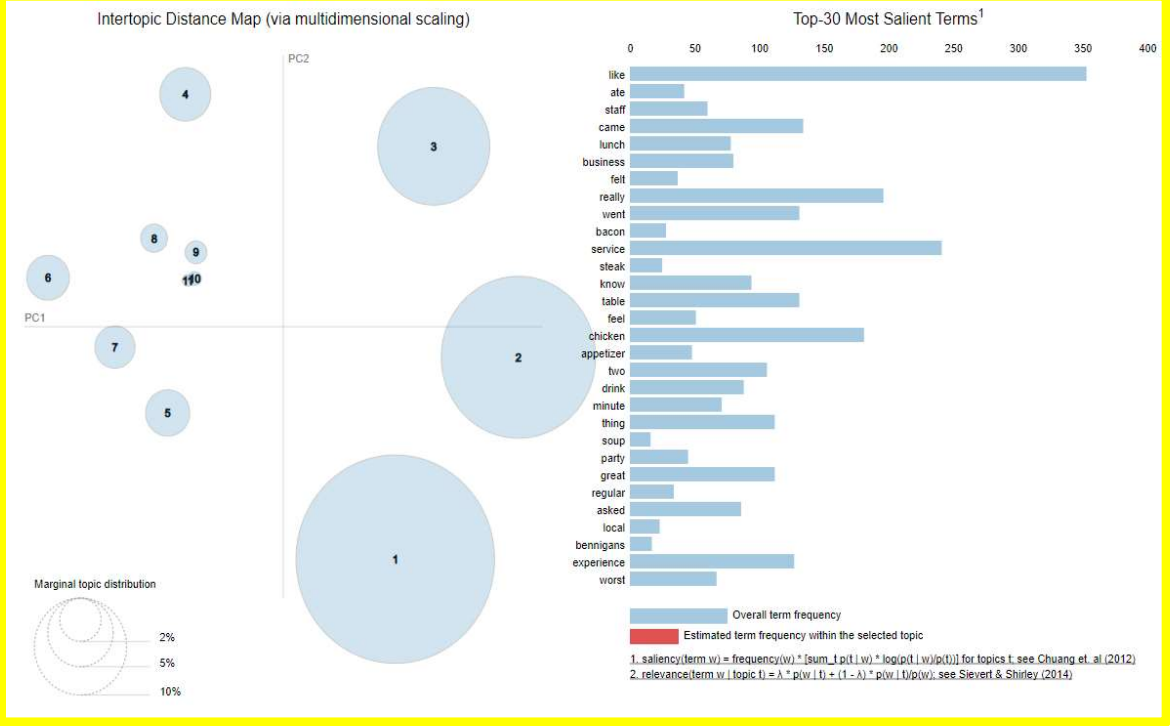
```
# Description du texte dans l'espace des topics
doc_topics_test = model_LDA_choisie.get_document_topics(test_text_bow)
print(doc_topics_test)
```

```
[(0, 0.16274282), (1, 0.35151204), (4, 0.222328), (7, 0.02554437), (8, 0.21546227)]
```

Avis d'insatisfaction clients

Prédire le sujet dont parle client

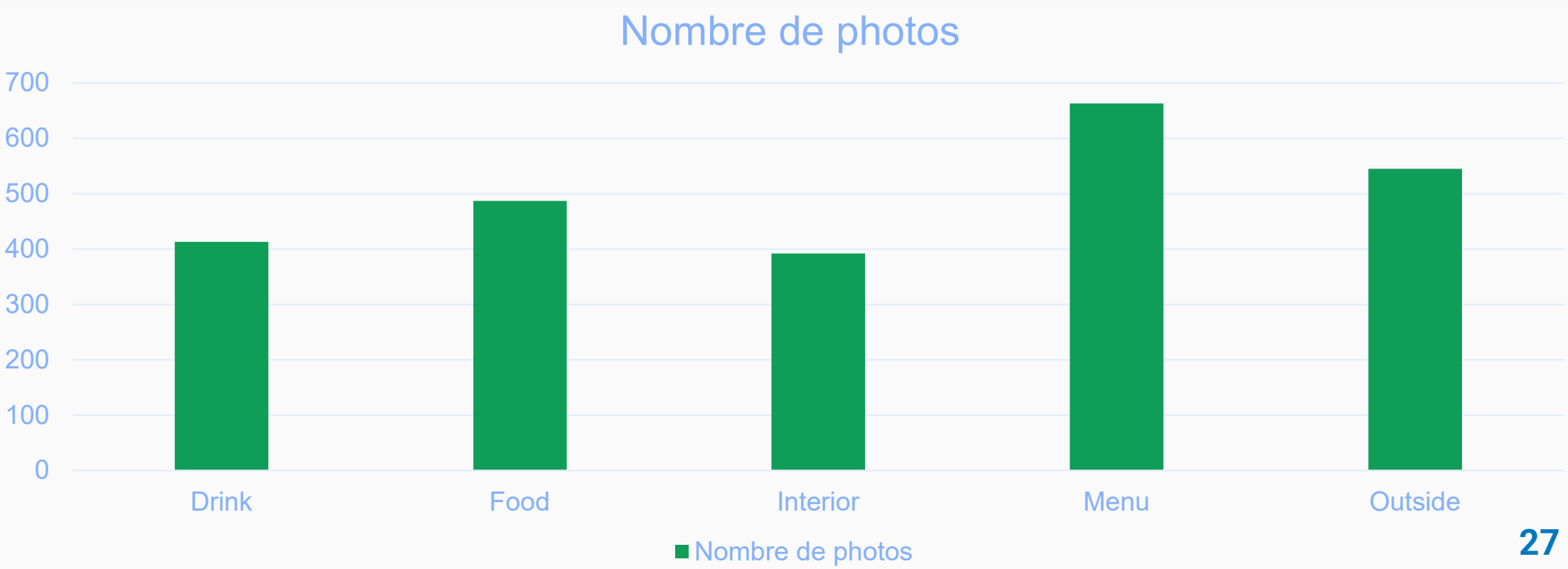
1. au service de restauration
2. au lieu où l'on mange
3. au site internet compliqué à comprendre
4. au personnel mal organisé
5. à la nourriture mauvaise
6. au comportement du personnel
7. au livraison avec un où plusieurs défaut sur le produit
8. au gaspillage de nourriture
9. à l'ambiance dans le restaurant
10. au menu bourratif
11. au menu trop légé



Conclusion sur la faisabilité

2/ Analyser les photos pour déterminer les catégories des photos

Mon choix : CNN Transfer Learning(VGG-16)



REMERCIEMENT

Merci de m’avoir écouter

REPONDRE AUX QUESTIONS