

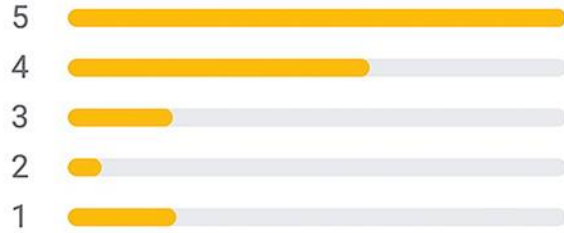
# Améliorez le produit IA de votre start-up



**Avis Restau**



## All reviews



# 4.0



277 reviews



Write a review



Sort

All

small 11

prices 11

atmosphere 9

brick oven 8

delivered 8

vodka 6

bar 5

chicken 5

carrot 4

penne 4



# Sommaire

PARTIE 1 – PROJET DE L'ENTREPRISE

PARTIE 2 – DÉTECTER LES SUJETS D'INSATISFACTION

PARTIE 3 – LABELLISER AUTOMATIQUEMENT LES PHOTOS

PARTIE 4 – COLLECTER DE NOUVELLES DONNÉES


CONCLUSION





# PARTIE 1 – PROJET DE L'ENTREPRISE

# Une nouvelle fonctionnalité !

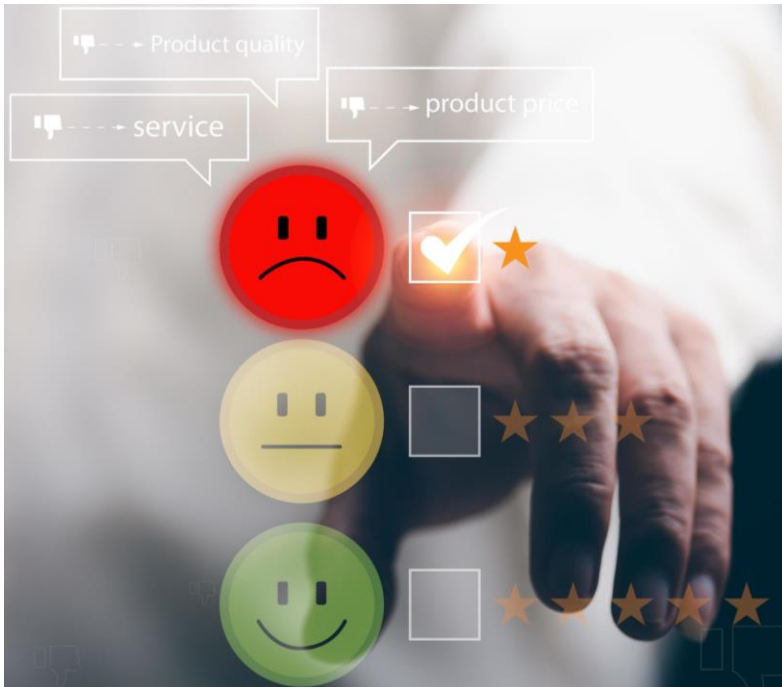
- Avis Restau lance le développement d'une nouvelle fonctionnalité de **collaboration** :
  - Poster des **commentaires**
  - Poster des **photos**
-  **opportunités** :
  - **Exploiter** ces nouvelles données
  - **Comprendre** les utilisateurs





# Opportunités

Détecter les sujets d'insatisfaction dans les commentaires négatifs



Labelliser automatiquement les photos postées sur la plateforme



Pour l'instant ➡ Étude **préalable** seulement



## PARTIE 2 – DÉTECTER LES SUJETS D'INSATISFACTION

# Méthode

- Objectif final : détection sujets mécontentement commentaires qui SERONT postés sur « Avis Restau »

➡ nous n'avons pas encore ces données pour tester la faisabilité ...

➡ utiliser une base de données existante :

- Étapes :
  - **Sélectionner** des commentaires négatifs,
  - **Prétraiter** ces données,
  - Obtenir une représentation de type « **bag-of-words** »,
  - Tester un modèle de **topic-modeling**,
  - **Visualiser** les sujets détectés !





# Sélectionner des commentaires

- Dataset des reviews : `dataYelp/yelp_academic_dataset_review.json`

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	KU_O5udG6zpxOg-VcAEodg	mh_-eMZ6K5RLWhZyISBhwa	XQfwVwDr-v0ZS3_CbbE5Xw	3	0	0	0	If you decide to eat here, just be aware it is...	2018-07-07 22:09:11
1	BiTunyQ73aT9WBnpR9DZGw	OyoGAe7OKpv6SyGZT5g77Q	7ATYjTlgM3jUlt4UM3lypQ	5	1	0	1	I've taken a lot of spin classes over the year...	2012-01-03 15:28:18
2	saUsX_uimxRICVr67Z4Jig	8g_iMtfSiwikVnbP2etR0A	YjUWPpI6HXG530lwP-fb2A	3	0	0	0	Family diner. Had the buffet. Eclectic assortm...	2014-02-05 20:30:30
3	AqPFMIeE6RsU23_auESxiA	_7bHUi9Uuf5_HHc_Q8guQ	kxX2SOes4o-D3ZQBkiMRfA	5	1	0	1	Wow! Yummy, different, delicious. Our favo...	2015-01-04 00:01:03
4	Sx8TMOWLNUJBWer-0pcmoA	bcjbaE6dDog4jkNY91ncLQ	e4Vwtrqf-wpJfwesgvdgxQ	4	1	0	1	Cute interior and owner (?) gave us tour of up...	2017-01-14 20:54:15

- Avis négatifs :  
- Problème : pas que des restaurants sur “Avis Restau” ...

➡ colonne `business_id`



# Sélectionner des commentaires

- Dataset des businesses : `dataYelp/yelp_academic_dataset_business.json`

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes	categories	hours
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426679	-119.711197	5.0	7	0	{'ByAppointmentOnly': 'True'}	Doctors, Traditional Chinese Medicine, Naturop...	None
1	mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	87 Grasso Plaza Shopping Center	Afton	MO	63123	38.551126	-90.335695	3.0	15	1	{'BusinessAcceptsCreditCards': 'True'}	Shipping Centers, Local Services, Notaries, Ma...	{'Monday': '0:0-0:0', 'Tuesday': '8:0-18:30', ...}
2	tUFrWirKiKi_TAnsVWINQQ	Target	5255 E Broadway Blvd	Tucson	AZ	85711	32.223236	-110.880452	3.5	22	0	{'BikeParking': 'True', 'BusinessAcceptsCredit...	Department Stores, Shopping, Fashion, Home & G...	{'Monday': '8:0-22:0', 'Tuesday': '8:0-22:0', ...}

- Avis négatifs :
- Problème : pas que des restaurants sur “Avis Restau” ...

➡ colonne `business_id`



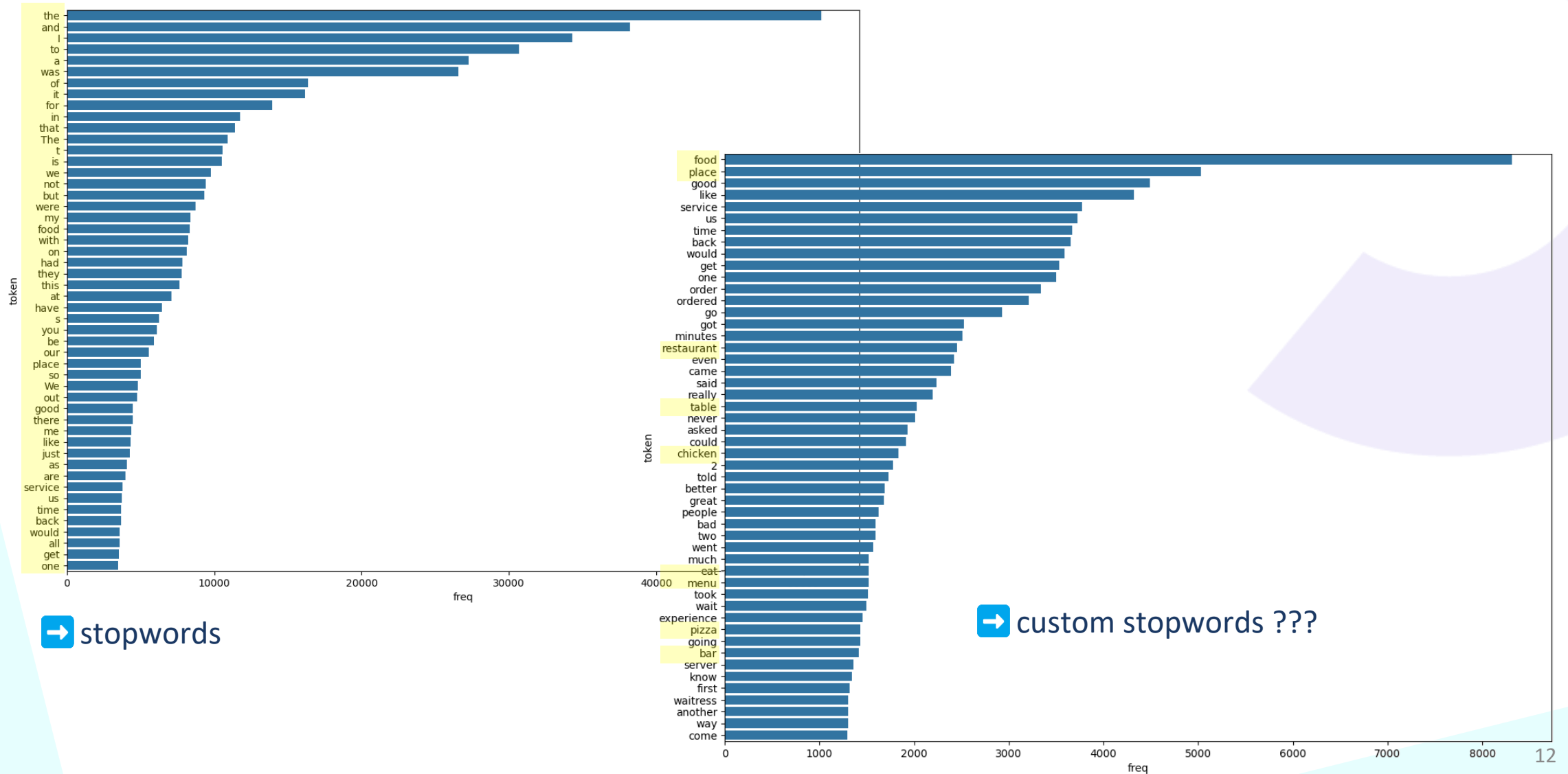
# Sélectionner des commentaires

- Dataset de 5 Go ➡ lecture par « chunk »
- Réaliser nos **filtres**
- Sélectionner quelques milliers de commentaires :

	text
1947	I came here for lunch and was very disappointe...
5136	One star for really nice service + no wait (ca...
5803	Fried rice looks like brown rice about 2 1/4 i...
3459	I agree with some of the other commenters, the...
7301	I used the mobile app, the bill was for \$31 a...
...	...
196744	Passing through Reno we stopped to get a bean ...
199082	For over an hour we were told we were next. Wa...
197735	If you're on a clock, this is not the place fo...
191503	Stopped in for dinner break as I work in the p...
191338	Ducked in to meet a local politician and talk ...



# Mots fréquents



→ stopwords

→ custom stopwords ???



## Mots avec des caractères répétés





# Cleaning – un exemple

"This was our first time trying Cal Taco, as we just moved to town and it's a 3 minute walk from our apartment. You probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars.\n\nWe ordered two tacos each, with the beans/rice/drink combo as well. We probably should have split one of the combo plates as the walk home was a bit more of a waddle. The fish tacos were decent, but I've had better. \n[http://www.yelp.com/biz\\_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=-cSzA1ONPpzJcj\\_GkQ2Eow](http://www.yelp.com/biz_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=-cSzA1ONPpzJcj_GkQ2Eow)\n\nThe fish itself didn't have much flavor so most of the taco's taste came from the toppings. I'd try a burrito next time. The beans and rice were fine, but the cheese shreds on top seemed unmelted! Throughout the whole meal they maintained their waxy shredded-cheese shape. Ah well!\n\n[http://www.yelp.com/biz\\_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=eLr4PuYuaIF1PUTVGQ8D1A](http://www.yelp.com/biz_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=eLr4PuYuaIF1PUTVGQ8D1A)\n\nOverall I thought it was fine, but a littleeee too pricey for the quality."



# Cleaning – Passer en minuscules

"This was our first time trying Cal Taco, as we just moved to town and it's a 3 minute walk from our apartment. You probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars.\n\nWe ordered two tacos each, with the beans/rice/drink combo as well. We probably should have split one of the combo plates as the walk home was a bit more of a waddle. The fish tacos were decent, but I've had better. \nhttp://www.yelp.com/biz\_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=-cSza1ONPpzJcj\_GkQ2Eow\n\nThe fish itself didn't have much flavor so most of the taco's taste came from the toppings. I'd try a burrito next time. The beans and rice were fine, but the cheese shreds on top seemed unmeltable! Throughout the whole meal they maintained their waxy shredded-cheese shape. Ah well!\n\nhttp://www.yelp.com/biz\_photos/VeFfrEZ4iWaecrQg6Eq4cg?select=elr4PuYuaif1PUTVGQ8D1A\n\nOverall I thought it was fine, but a littleeee too pricey for the quality."



"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars.\n\nwe ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. \nhttp://www.yelp.com/biz\_photos/veffrez4iwaecrQg6Eq4cg?select=-csza1onppzjcj\_gkq2eow\n\nthe fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmeltable! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well!\n\nhttp://www.yelp.com/biz\_photos/veffrez4iwaecrQg6Eq4cg?select=elr4puyuaif1putvgq8d1a\n\noverall i thought it was fine, but a littleeee too pricey for the quality."



# Cleaning – enlever les url

"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars.\n\nwe ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. \n[http://www.yelp.com/biz\\_photos/veffrez4iwaecrqg6eq4cg?select=-csza1onppzjcj\\_gkq2eow](http://www.yelp.com/biz_photos/veffrez4iwaecrqg6eq4cg?select=-csza1onppzjcj_gkq2eow)\n\nthe fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well!\n\n[http://www.yelp.com/biz\\_photos/veffrez4iwaecrqg6eq4cg?select=elr4puyuaif1putvgq8d1a](http://www.yelp.com/biz_photos/veffrez4iwaecrqg6eq4cg?select=elr4puyuaif1putvgq8d1a)\n\noverall i thought it was fine, but a littleeee too pricey for the quality."



"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars.\n\nwe ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. \n\n\nthe fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well!\n\n\noverall i thought it was fine, but a littleeee too pricey for the quality."



# Cleaning – enlever les séquences d'échappement

"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars. \n\nwe ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. \n\nthe fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well!\n\noverall i thought it was fine, but a littleeee too pricey for the quality."



"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars. we ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. the fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well! overall i thought it was fine, but a littleeee too pricey for the quality."



# Cleaning – enlever les caractères répétés

"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars. we ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. the fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well! overall i thought it was fine, but a **littleeee** too pricey for the quality."



"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars. we ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. the fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmelted! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well! overall i thought it was fine, but a **little** too pricey for the quality."





# Cleaning – nettoyage plus général

"this was our first time trying cal taco, as we just moved to town and it's a 3 minute walk from our apartment. you probably wouldn't notice the place unless you were looking for it, but it's less of a dive inside than you would think, i.e. chain-fast-food-restaurant-style chairs and tables that are anchored to the ground and commercial signs designating the soda and salsa bars. we ordered two tacos each, with the beans/rice/drink combo as well. we probably should have split one of the combo plates as the walk home was a bit more of a waddle. the fish tacos were decent, but i've had better. the fish itself didn't have much flavor so most of the taco's taste came from the toppings. i'd try a burrito next time. the beans and rice were fine, but the cheese shreds on top seemed unmeltable! throughout the whole meal they maintained their waxy shredded-cheese shape. ah well! overall i thought it was fine, but a little too pricey for the quality."



Retirer : **ponctuations** / **stopwords** / **nombres** /

Retirer : **stopwords personnalisés** : ['restaurant', 'diner', 'dinner,bistro', 'cafe', 'dining', 'hotel', 'lunch', 'place']

Retirer : POS (Part-of-Speech) – tout ce qui n'est pas : "NOUN","ADJ"



"time cal taco town minute walk apartment dive chain fast food style chair table ground commercial sign soda salsa bar taco bean rice drink combo combo plate walk home bit waddle fish taco decent well fish flavor taco taste topping burrito time bean rice fine cheese shredding unmeltable meal waxy cheese shape overall fine little pricey quality"



# Tokenization

'time cal taco town minute walk apartment dive chain fast food style chair table ground commercial sign soda salsa bar taco bean rice drink combo combo plate walk home bit waddle fish taco decent well fish flavor taco taste topping burrito time bean rice fine cheese shredding unmeltable meal waxy cheese shape overall fine little pricey quality'



['time', 'cal', 'taco', 'town', 'minute', 'walk', 'apartment', 'dive', 'chain', 'fast', 'food', 'style', 'chair', 'table', 'ground', 'commercial', 'sign', 'soda', 'salsa', 'bar', 'taco', 'bean', 'rice', 'drink', 'combo', 'combo', 'plate', 'walk', 'home', 'bit', 'waddle', 'fish', 'taco', 'decent', 'well', 'fish', 'flavor', 'taco', 'taste', 'topping', 'burrito', 'time', 'bean', 'rice', 'fine', 'cheese', 'shredding', 'unmeltable', 'meal', 'waxy', 'cheese', 'shape', 'overall', 'fine', 'little', 'pricey', 'quality']



# Création du dictionnaire & Filtres sur la fréquence

```
[  
    ['time', 'cal', 'taco', 'town', 'minute', 'walk', 'apartment', 'dive', 'chain', 'fast', 'food', 'style', 'chair', 'table', 'ground', 'commercial', 'sign', 'soda', 'salsa', 'bar',  
     'taco', 'bean', 'rice', 'drink', 'combo', 'combo', 'plate', 'walk', 'home', 'bit', 'waddle', 'fish', 'taco', 'decent', 'well', 'fish', 'flavor', 'taco', 'taste', 'topping', 'burrito',  
     'time', 'bean', 'rice', 'fine', 'cheese', 'shredding', 'unmeltable', 'meal', 'waxy', 'cheese', 'shape', 'overall', 'fine', 'little', 'pricey', 'quality'],  
    [... Tokens du doc 2 ...],  
    [... Tokens du doc 3...],  
    ... ,  
]
```



```
{'apartment': 0, 'bar': 1, 'bean': 2, 'bit': 3, 'burrito': 4, 'cal': 5, 'chain': 6, 'chair': 7, 'cheese': 8, 'combo': 9, 'commercial': 10, 'decent': 11, 'dive': 12, 'drink': 13, 'fast': 14,  
'fine': 15, 'fish': 16, 'flavor': 17, 'food': 18, 'ground': 19, 'home': 20, 'little': 21, 'meal': 22, 'minute': 23, 'overall': 24, 'plate': 25, 'pricey': 26, 'quality': 27, 'rice': 28, 'salsa':  
29, 'shape': 30, 'shredding': 31, 'sign': 32, 'soda': 33, 'style': 34, 'table': 35, 'taco': 36, 'taste': 37, 'time': 38, 'topping': 39, 'town': 40, 'unmeltable': 41, 'waddle': 42,  
'walk': 43, 'waxy': 44, 'well': 45, ... , ..., ..., ...}
```



```
[  
    [(0, 1), (1, 1), (2, 2), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 2), (9, 2), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 2), (16, 2), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1),  
     (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 2), (29, 1), (30, 1), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 4), (37, 1), (38, 2), (39, 1), (40, 1), (41, 1), (42, 1),  
     (43, 2), (44, 1), (45, 1), ..., ..., ...],  
    [...],  
    [...],  
    ... ,  
]
```

... avec<sup>]</sup> ou sans filtres sur la fréquence

Sur les tokens rares : retirer ceux non présents dans plus de n documents

Sur les tokens fréquents : retirer ceux présents dans plus de x% des documents



# Pondération TF-IDF

[(0, 1), (1, 1), (2, 2), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 2), (9, 2), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 2), (16, 2), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 2), (29, 1), (30, 1), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 4), (37, 1), (38, 2), (39, 1), (40, 1), (41, 1), (42, 1), (43, 2), (44, 1), (45, 1), ..., ..., ...]



$tfidf = TermFrequency \times InverseDocumentFrequency$

$$TF = N_{word,doc} \qquad IDF = \log_2\left(\frac{N_{docs}}{DF}\right)$$



[(0, 0.141), (1, 0.141), (2, 0.283), (3, 0.141), (4, 0.141), (6, 0.141), (8, 0.283), (10, 0.141), (12, 0.141), (13, 0.141), (14, 0.141), (15, 0.283), (16, 0.283), (18, 0.141), (19, 0.141), (20, 0.141), (24, 0.141), (25, 0.141), (26, 0.141), (28, 0.283), (29, 0.141), (30, 0.141), (34, 0.141), (37, 0.141), (38, 0.283), (41, 0.141), (42, 0.141), (43, 0.283), (44, 0.141), ..., ..., ...]



# Prétraitement - Pipeline

- Créer des classes « enfants » de `sklearn.base.TransformerMixin` et `sklearn.base.BaseEstimator` avec nos différentes fonctions
- Intégrer les méthodes nécessaires (`__init__`, `fit`, etc.)
- Afin de créer un pipeline Scikit Learn :

Reviews  
Loader

Lower  
Case

url  
Remover

Escape  
Sequences  
Remover

Repeated  
Chars  
Corrector

Text  
Cleaner

Dict &  
Vector  
Maker



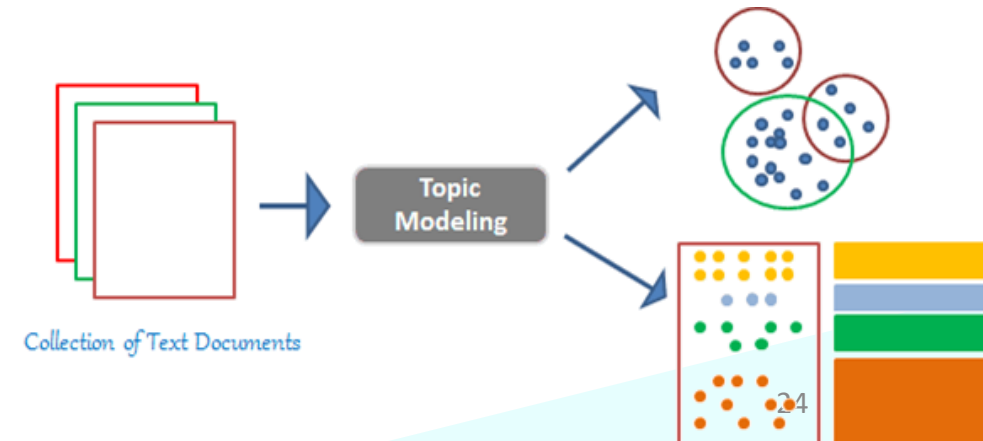


# Topic modeling – détecter sujets d'insatisfaction

- Latent Dirichlet Allocation
- Non supervisé
- Réduction de dimension :

$$\textit{Matrice}_{documents-mots} = \textit{Matrice}_{documents-topics} \cdot \textit{Matrice}_{topics-mots}$$

Nb docs X Nb mots  $\longrightarrow$  Nb docs X Nb topics



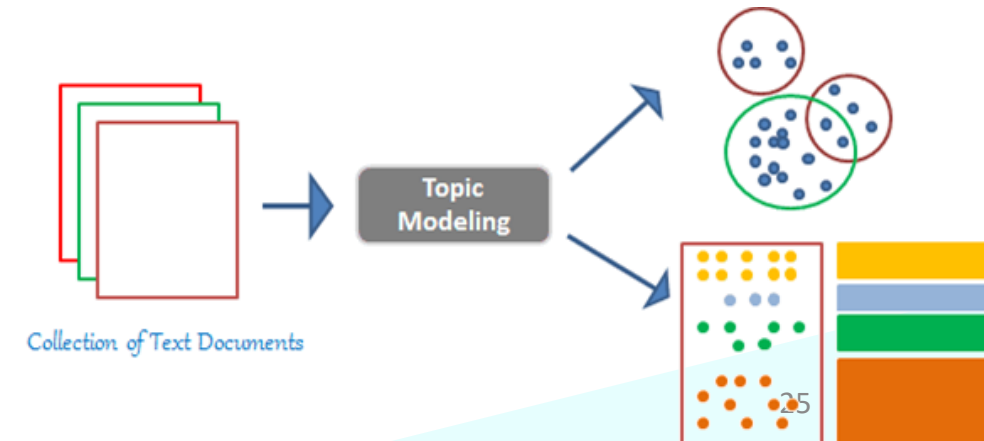
# Topic modeling – LDA

- Latent Dirichlet Allocation
- Modèle génératif qui :
  - Considère chaque document comme un mélange de topics : (.24, .36, .40)
  - Considère chaque topic comme un mélange de mots : (.02, .13, ... , ... , .05)
  - Génère ainsi chaque mot de chaque document en choisissant un topic dans (.24, .36, .40) puis en choisissant un mot dans (.02, .13, ... , ... , .05)
  - S'optimise en itérant sur chaque document et chaque mot

Topics du modèle optimisé



Sujets cachés dans le corpus !



# Topic modeling – LDA



## Paramètres principaux :

- `numTopics`
- `alpha`
- `eta`



## Évaluation

- Cohérence
- Perplexité
- Mais surtout, interprétation humaine !

# Topic modeling – un 1<sup>er</sup> essai peu concluant...

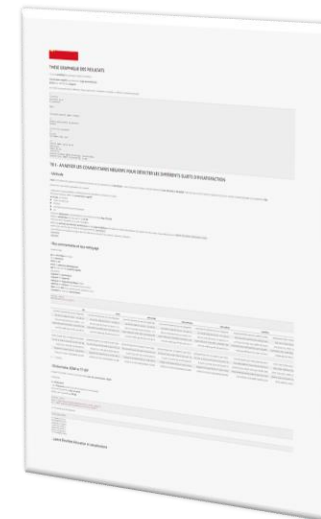
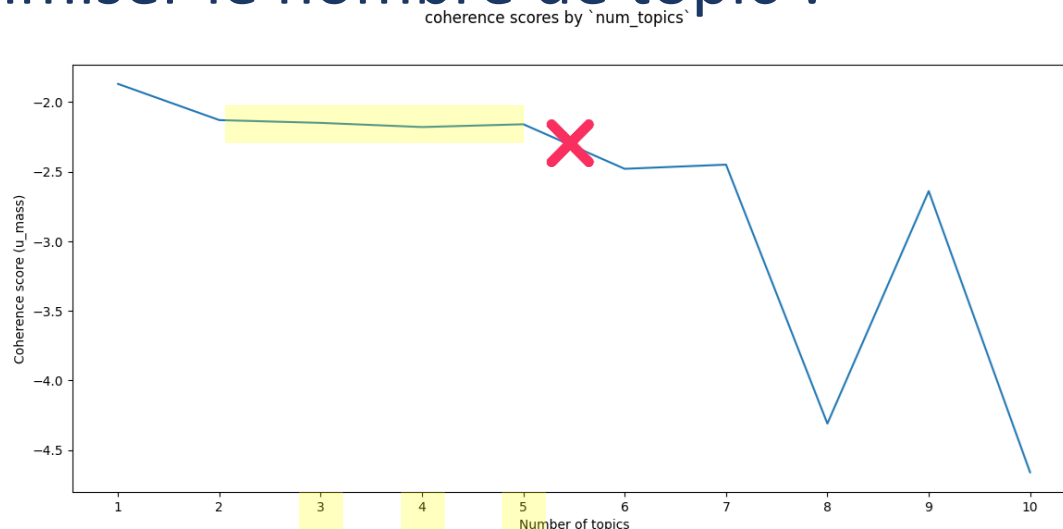
- Premier essai, peu concluant ...

LDA Topic Modeling with 3 topics  
Coherence score = -2.31



# Topic modeling – améliorer la modélisation

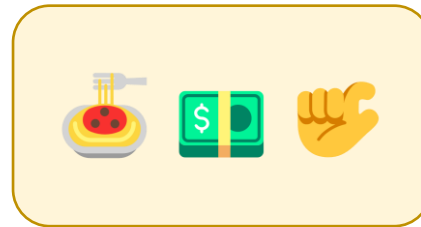
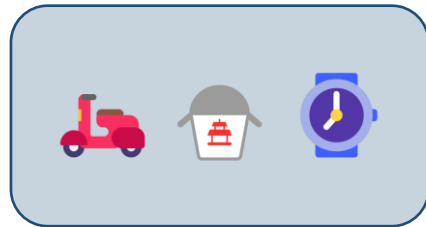
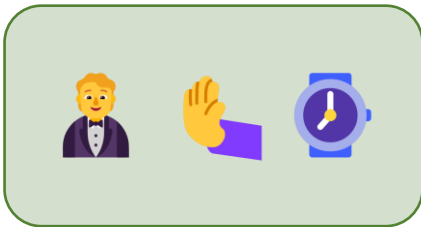
- Tester avec plus de commentaire : ~~10000~~ → 60000
- Améliorer le pré-traitement :
  - + de stopwords personnalisés : nom de plats, de boisson, de type de restaurant
  - Filtres fréquence (haut et bas)
- Optimiser le nombre de topic ?



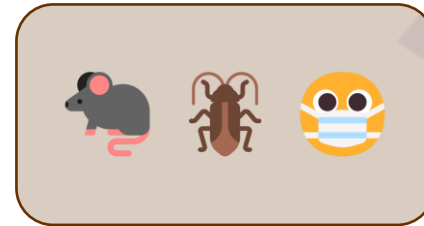
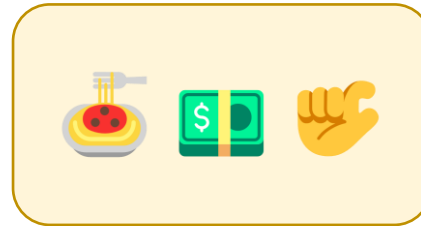
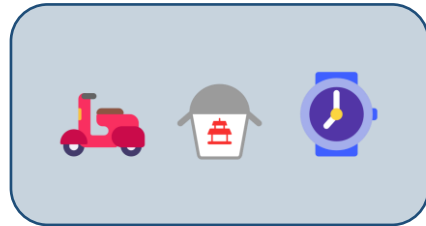
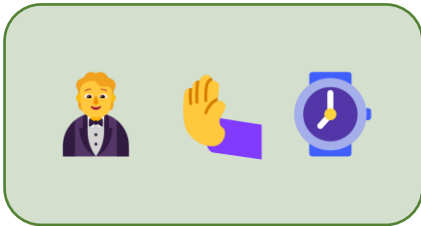


# Topic modeling – améliorer la modélisation

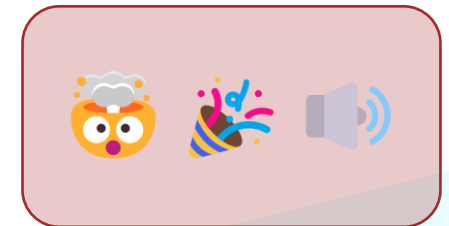
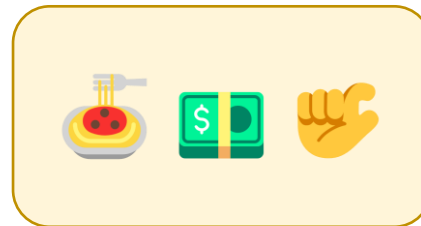
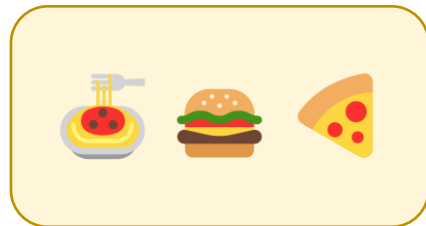
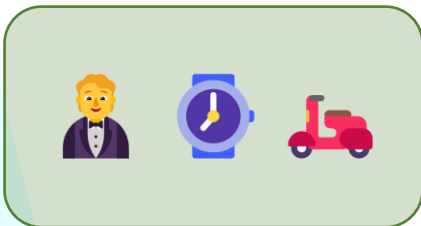
- 3 topics :



- 4 topics :



- 5 topics :



# Détection sujets mécontentement - conclusion

- Un nombre de topics idéal ?
- Faisabilité du projet :
  - 👍 : Meilleure préparation + Plus de données ➡ Topics clairs
  - 👎 : beaucoup de stopwords spécifiques / topics instables
- La suite :
  - tous les commentaires
  - optimisation hyperparamètres du pipeline (alpha, eta, stopwords, filtres, etc.)
  - tests d'autres modèles





# PARTIE 3 – LABELLISER AUTOMATIQUEMENT LES PHOTOS

# Méthode

- Photos **yelp** 
- Objectif final : Algorithme de classification supervisée.  
A utiliser ultérieurement sur les photos qui seront publiées sur « Avis Restau ».

 **ici : étude faisabilité seulement**

- Étapes :
  - **Sélectionner** quelques photos
  - **Pré-traiter** les photos
  - **Extraire** des features (via le SIFT)
  - Créer un **dictionnaire de *visual words*** et une matrice de ***bag of visual words***
  - Réduire la **dimension**
  - **Faisabilité** via un algorithme de **clustering** (Kmeans)
  - **ARI** clusters VS classes
  - **Visualisation**



# Sélectionner des photos

Yelp photo - some examples

drink



food



inside



menu



outside



# Nuance de gris

Yelp photos examples - gray scale conversion

drink



food



inside



menu



outside





# Égalisation d'histogramme

Yelp photos examples - equalized

drink



food



inside



menu



outside





# Filtrage du bruit

Yelp photos examples - gaussian blur

drink



food



inside



menu



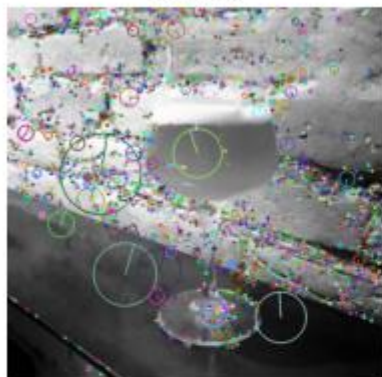
outside



# Extraction descripteur SIFT

Yelp photos examples - SIFT descriptors

drink



food



inside



menu



outside

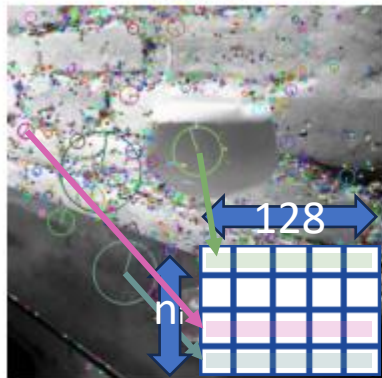




# Dictionnaire de *visual words*

Yelp photos examples - SIFT descriptors

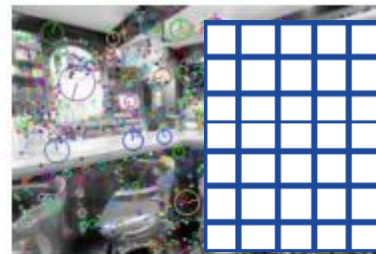
drink



food



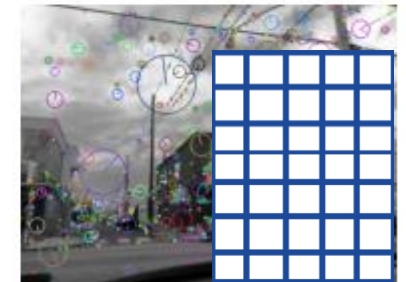
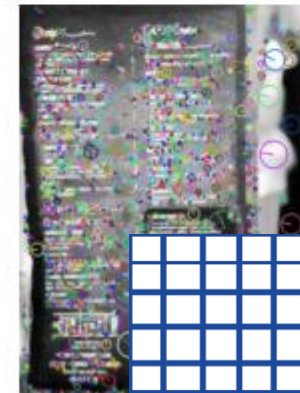
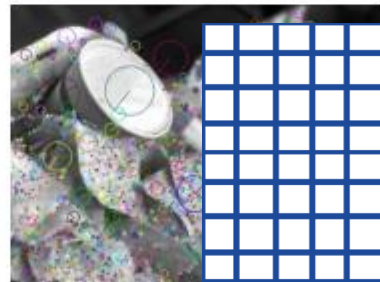
inside



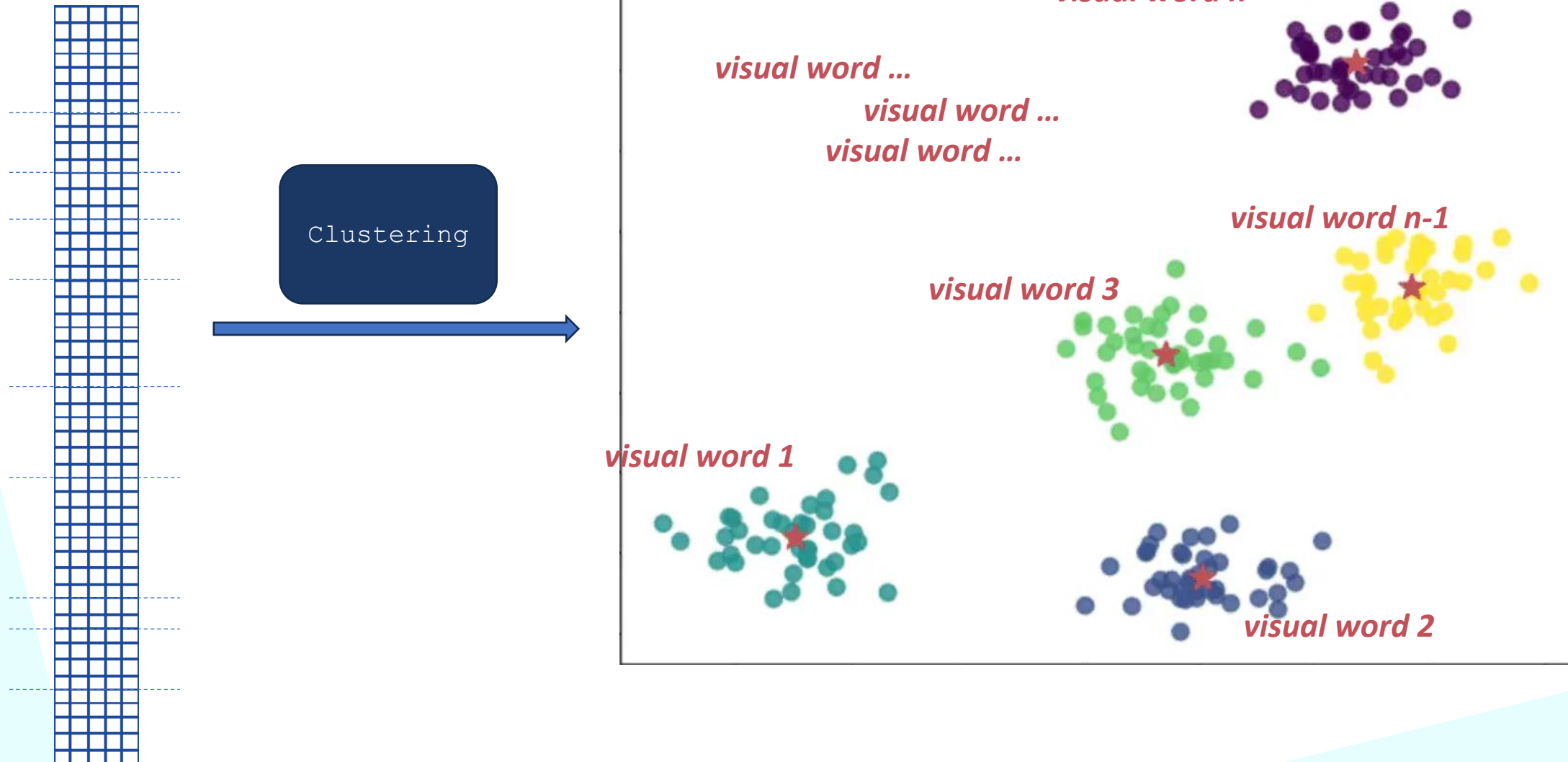
menu



outside



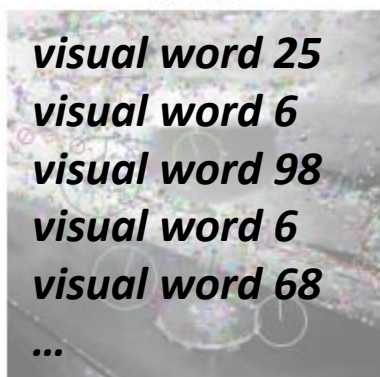
# Dictionnaire de *visual words*



# Matrice *bags-of-visual-words*

Yelp photos examples - SIFT descriptors

drink



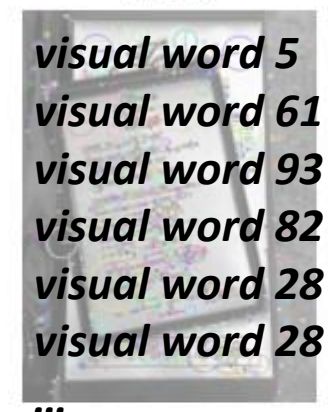
food



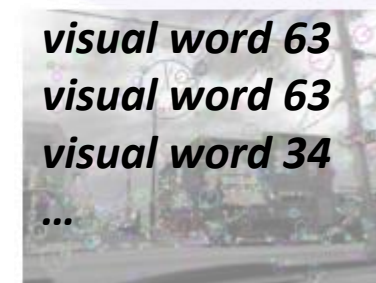
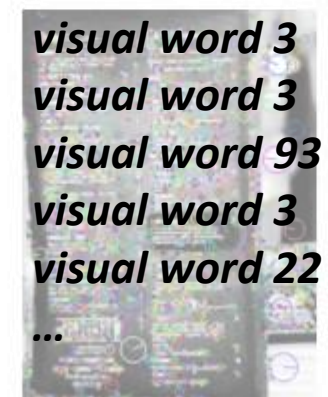
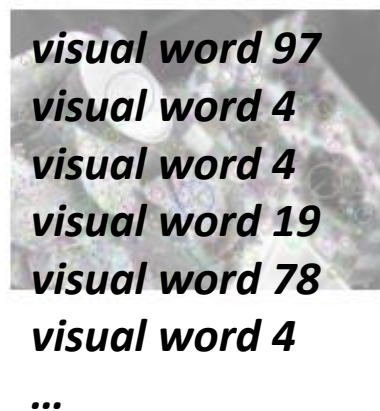
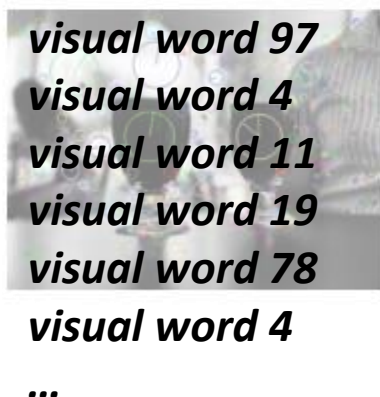
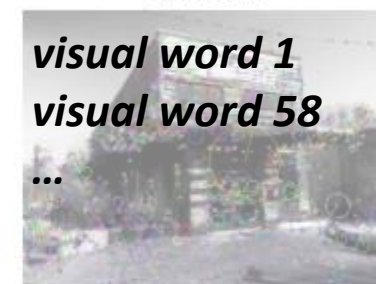
inside



menu



outside



# Matrice *bags-of-visual-words*

	0	1	2	3	4	5	6	7	8	9	...	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079
0	5	1	1	1	2	0	2	0	0	0	...	0	1	0	1	2	1	0	0	7	2
1	0	0	0	0	1	2	1	1	0	0	...	1	1	0	0	0	2	1	0	0	1
2	0	0	0	0	0	2	1	0	0	1	...	0	0	1	0	0	0	0	1	1	0
3	0	1	0	0	1	1	3	0	13	2	...	1	2	0	0	0	2	0	1	1	0
4	0	0	0	0	0	0	1	0	2	1	...	0	0	3	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
745	2	0	0	0	1	0	2	0	1	3	...	2	1	1	0	3	1	3	0	1	0
746	1	0	2	0	1	0	0	2	1	0	...	3	1	3	0	2	0	0	0	1	1
747	0	0	0	0	0	0	0	0	4	0	...	2	0	0	1	2	2	0	1	0	0
748	1	0	0	0	0	4	6	3	3	2	...	2	2	1	2	1	1	1	2	1	4
749	2	4	0	0	1	1	4	3	12	2	...	2	2	4	1	0	5	5	4	1	2





# Pondération TF-IDF

	0	1	2	3	4	5	...	1076	1077	1078	1079
0	0.065749	0.013844	0.013844	0.014338	0.028015	0.000000	...	0.000000	0.000000	0.071480	0.030146
1	0.000000	0.000000	0.000000	0.000000	0.038308	0.048584	...	0.028936	0.000000	0.000000	0.041223
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.072744	...	0.000000	0.037969	0.041815	0.000000
3	0.000000	0.031250	0.000000	0.000000	0.031619	0.020050	...	0.000000	0.020931	0.023050	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...
745	0.035463	0.000000	0.000000	0.000000	0.018888	0.000000	...	0.042801	0.000000	0.013770	0.000000
746	0.023517	0.000000	0.049517	0.000000	0.025051	0.000000	...	0.000000	0.000000	0.018262	0.026957
747	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.030942	0.000000	0.000000
748	0.026915	0.000000	0.000000	0.000000	0.000000	0.072721	...	0.021656	0.037957	0.020901	0.123405
749	0.025719	0.054153	0.000000	0.000000	0.013698	0.008686	...	0.051734	0.036271	0.009986	0.029481





# ACP

	C0	C1	C2	C3	C4	C5	...	C606	C607	C608
<b>0</b>	-0.079165	-3.800291	14.066756	0.984657	5.144575	8.648161	...	-0.214824	0.246502	-0.510977
<b>1</b>	4.668302	-1.241076	-3.295023	2.710035	-2.393686	1.030828	...	-0.003246	-0.325203	-0.115126
<b>2</b>	-0.501951	-5.224448	-5.325085	3.149939	1.008314	-1.202440	...	-0.036203	-0.078368	-0.545426
<b>3</b>	-7.326880	-9.606032	-3.030050	-0.454922	-1.654513	-1.160068	...	0.653721	-0.202591	0.492679
<b>4</b>	1.945519	-2.702972	0.182178	1.762309	-2.828546	6.268726	...	0.226559	0.037979	0.339935
...	...	...	...	...	...	...	...	...	...	...
<b>745</b>	-0.748908	3.689911	-3.130025	-4.271580	-1.739386	-4.707257	...	0.116614	-0.029192	0.164529
<b>746</b>	-0.979159	-2.489993	-2.394485	-3.009860	-3.376783	5.438454	...	-0.429891	0.129443	-0.421667
<b>747</b>	7.316390	-7.549401	-3.281177	-5.320557	7.607325	-0.349459	...	0.317693	0.220423	0.127270
<b>748</b>	-4.471304	-6.847694	-9.406762	-4.260997	2.071330	2.058631	...	-0.533770	0.029683	0.137277
<b>749</b>	-10.251017	-3.028337	4.381681	-1.659601	3.600789	-2.654735	...	-0.747806	0.086976	0.010088



# Prétraitement - Pipeline

- Créer des classes « enfants » de `sklearn.base.TransformerMixin` et `sklearn.base.BaseEstimator` avec nos différentes fonctions
- Intégrer les méthodes nécessaires (`__init__`, `fit`, etc.)
- Afin de créer un pipeline Scikit Learn :

Data  
Loader

sampler

deleter

sampler  
2

Images  
Loader

Gray  
Converter

equalizer

Gaussian  
Blurer

SIFT  
extractor

BOVW  
maker

TFIDF  
weighter

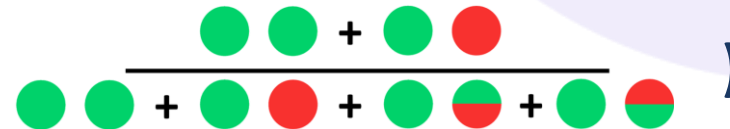
PCA  
reducer



# Études de faisabilité - clustering

- Pertinent de lancer un projet de modèle de classification automatique ?
- ➡ clustering  $\approx$  4 classes
- ➡ Kmeans avec `n_cluster = 4`
- Comment comparer ?

- Version « ajustée » de l'indice de Rand (

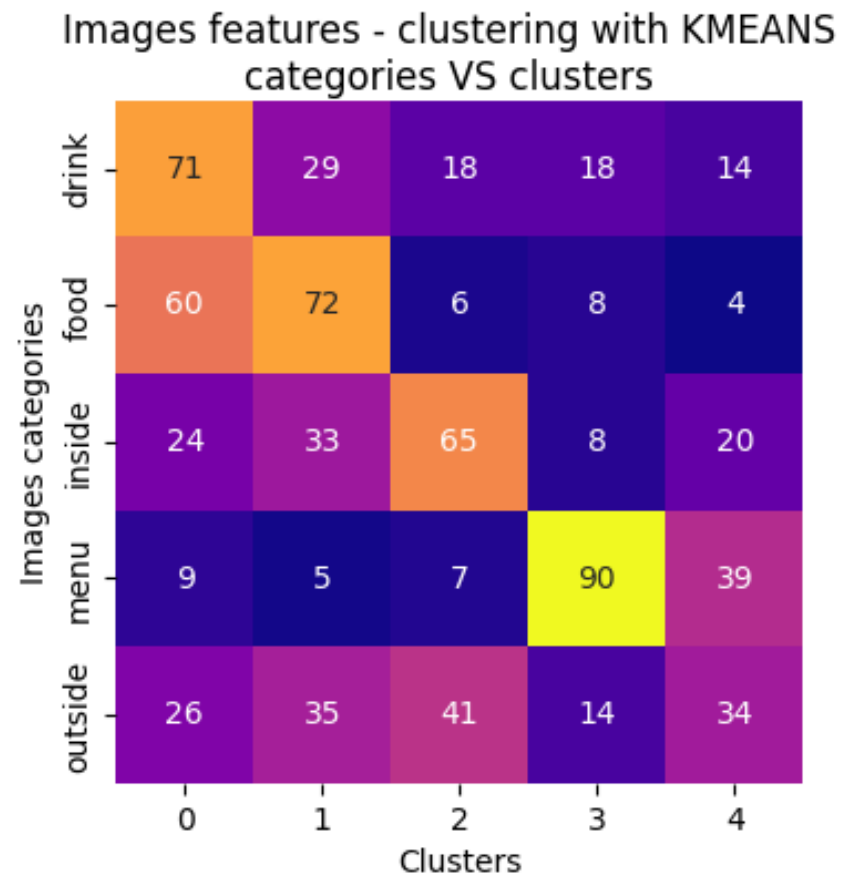


- ARI : 
$$\frac{RI - E(RI)}{\text{Max}(RI) - E(RI)}$$

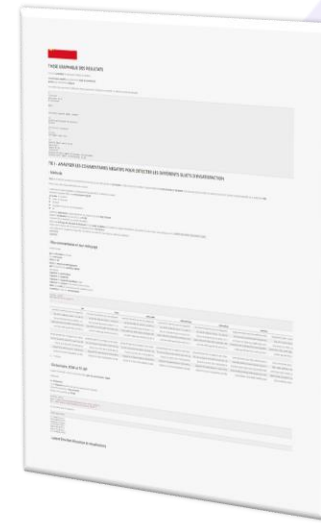


# Études de faisabilité – résultat avec le SIFT

- Des résultats mitigés ...

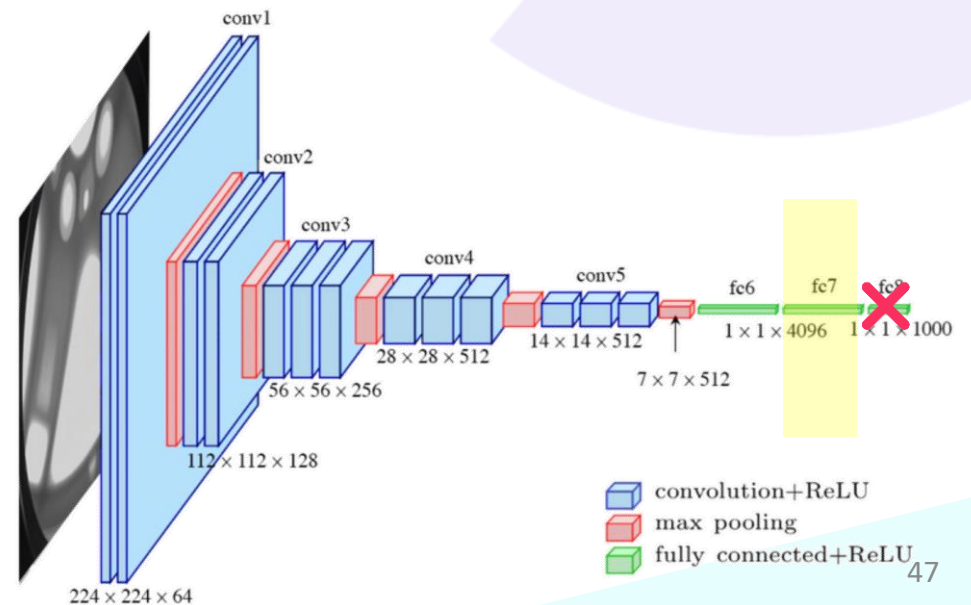


ARI = 0.14



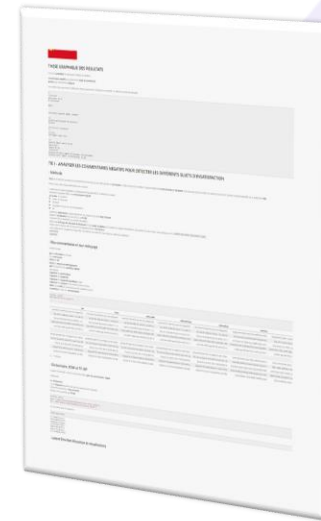
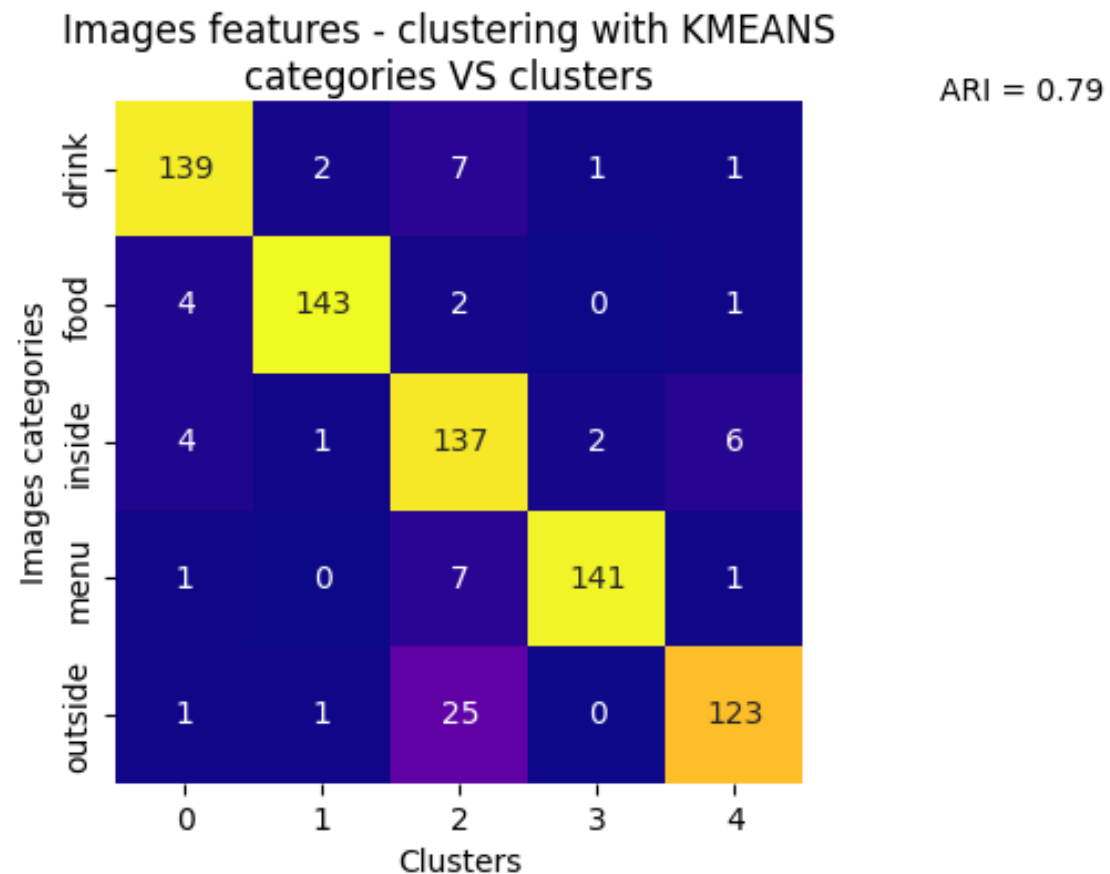
# Faire mieux : Transfer Learning VGG16

- Étapes :
  - **Sélectionner** quelques photos
  - **Pré-traiter** les photos :
    - Les charger à la **taille 224 x 224**
    - Les convertir en **arrays**
    - Les **rassembler** en 1 seul array
    - Les convertir en **BGR**
    - **Centrer** chaque canal de couleur
- VGG16 **pré-entraîné** :
  - **Supprimer** dernière couche *fully-connected* (celle contenant la fonction d'activation de classification *SoftMax*)
  - 4096 features complexes
- ACP



# Études de faisabilité – résultat avec le Transfer Learning

- Des performances plus prometteuses



# Analyse photos - conclusion

- Faisabilité du projet :
  - Features qui permettent bien de différencier les classes **avec le Transfer Learning**
  - La suite :
    - Entraîner un **modèle de classification automatique** pour l'appliquer aux futures photos déposées sur « Avis Restau »
    - Avez bien plus de photos
    - En utilisant, comme ici, directement le VGG16 comme un extracteur de features indépendant ?
    - Ou « fine-tuning » total ou partiel VGG16 ?
    - Autre modèle ?







# PARTIE 4 – COLLECTER DE NOUVELLES DONNÉES

# API - Méthode

- Objectif final : s'assurer de la possibilité de collecter de nouvelles données.

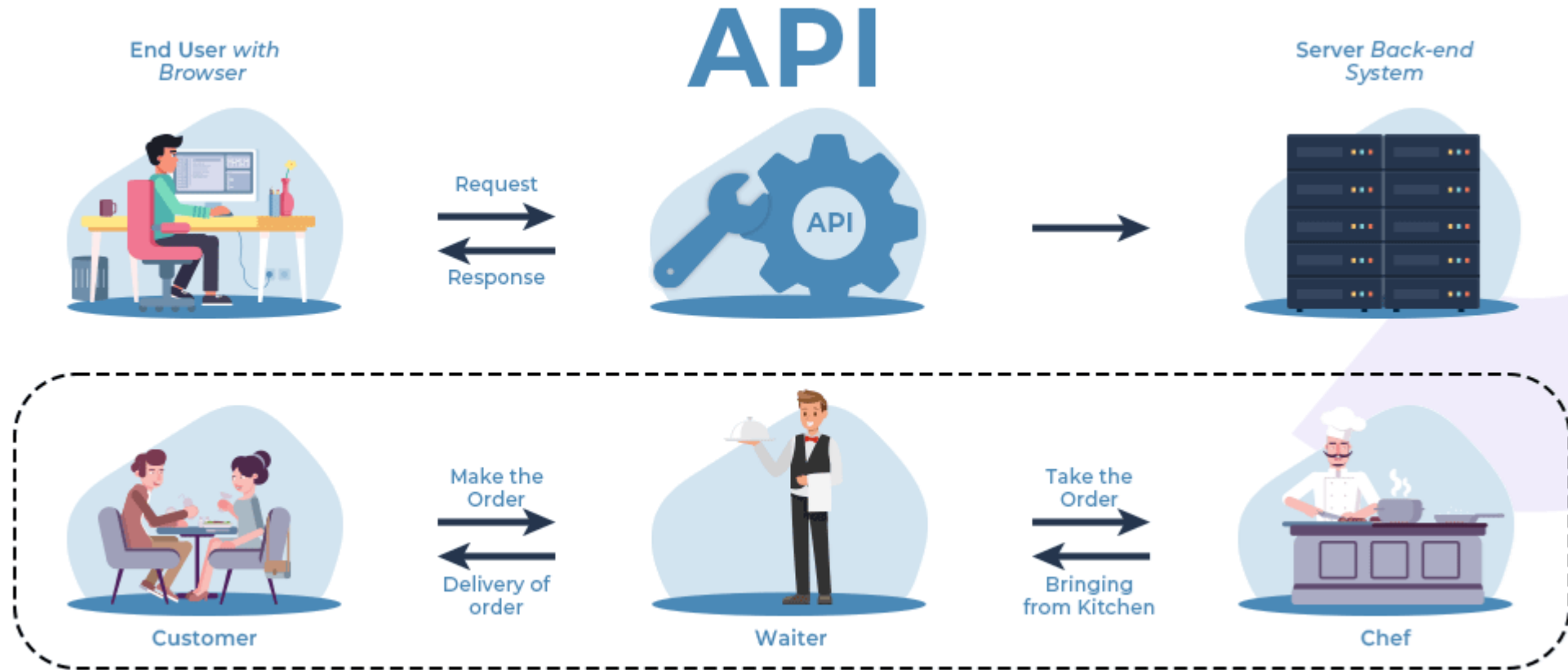
➡ **ici** : utiliser l'API **yelp**  pour récupérer :

- l'ID de 200 restaurants tous situés dans la même ville
- les commentaires clients associés

- Étapes :
  - Créer un **compte** développeur sur
  - S'inscrire au *Developer Beta program*
  - Utiliser le langage de requête  GraphQL



# API - Fonctionnement



# API – Nos requêtes

- Point de terminaison : *businesses/search* pour appliquer des filtres :
- Recherche :
  - `location` : « santa barbara »
  - `term` : « restaurants »
  - `limit` : 200 → impossible car limité à 50...
  - `offset` : ~~2~~ 50, 100, 150]
- Champs business à récupérer :
  - `id`
  - `reviews` : limitées à 3
    - `rating`
    - `text`
  - `name`



# API - Résultat

	restaurant_name	restaurant_id	review_rating	review_text
0	Santo Mezcal	nYPzsOjvida-ne7swSPHpA	5	This was my first time here and I was beyond i...
1	Santo Mezcal	nYPzsOjvida-ne7swSPHpA	5	Santo Mezcal is a beautiful restaurant in the ...
2	Santo Mezcal	nYPzsOjvida-ne7swSPHpA	5	Such a delicious dinner location. Thankfully w...
3	Brophy Bros - Santa Barbara	U3grYFleu6RgAAQgdriHww	5	Excellent setting and view -- fantastic food ...
4	Brophy Bros - Santa Barbara	U3grYFleu6RgAAQgdriHww	5	Amazing place! This bar/restaurant is located ...
...	...	...	...	...
592	Oak Park Market + Eatery	LMm555z8BRxDxgkpF_TEug	4	I was excited that everything wasn't priced ve...
593	Oak Park Market + Eatery	LMm555z8BRxDxgkpF_TEug	4	This place is pretty good. I got a chocolate p...
594	Su's Bowl	MiMRIsXOpHJMSTDAuw8s1w	5	Very good chicken fried rice and xiao long bao...
595	Su's Bowl	MiMRIsXOpHJMSTDAuw8s1w	5	Fulfilled my craving for good asian food! The ...
596	Su's Bowl	MiMRIsXOpHJMSTDAuw8s1w	5	I always doordash from here, but I decided to ...





# CONCLUSION

All reviews

5

4

3

2

1

4.0

★

★

★

★

★

277 reviews

Write a review

Sort

All

atmosphere 9

brick oven 8

delivered 8

vodka 6

bar 5

chicken 5

carrot 4

penne 4

merci

