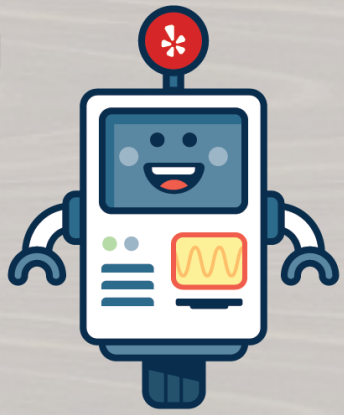
Présentation de l'appel à projets

Téléchargement des données

Bases de données volumineuses

API



- **Détecter** les sujets d'insatisfaction

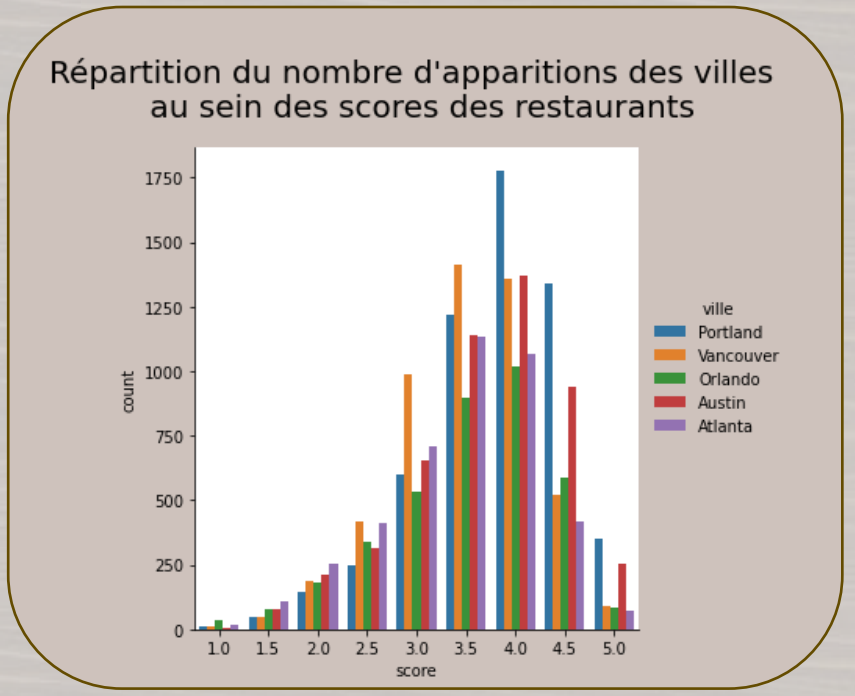
Environnement de travail

Librairies python spécialisées importées :

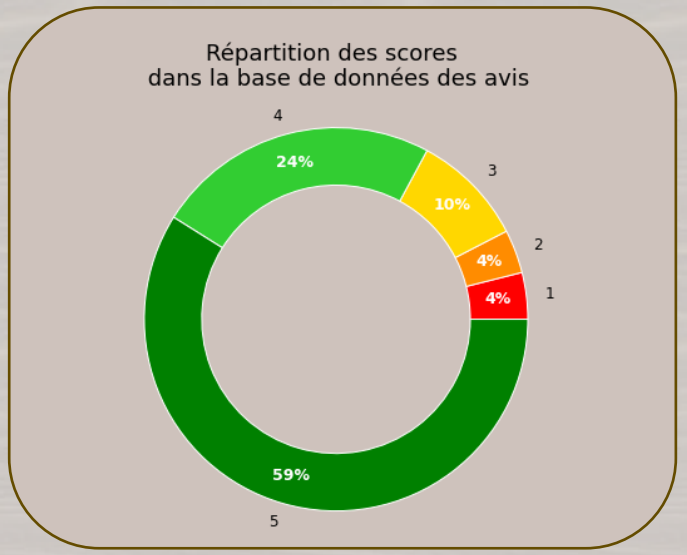
- Scikit-learn
- OpenCV
- SIFT
- Keras
- TensorFlow

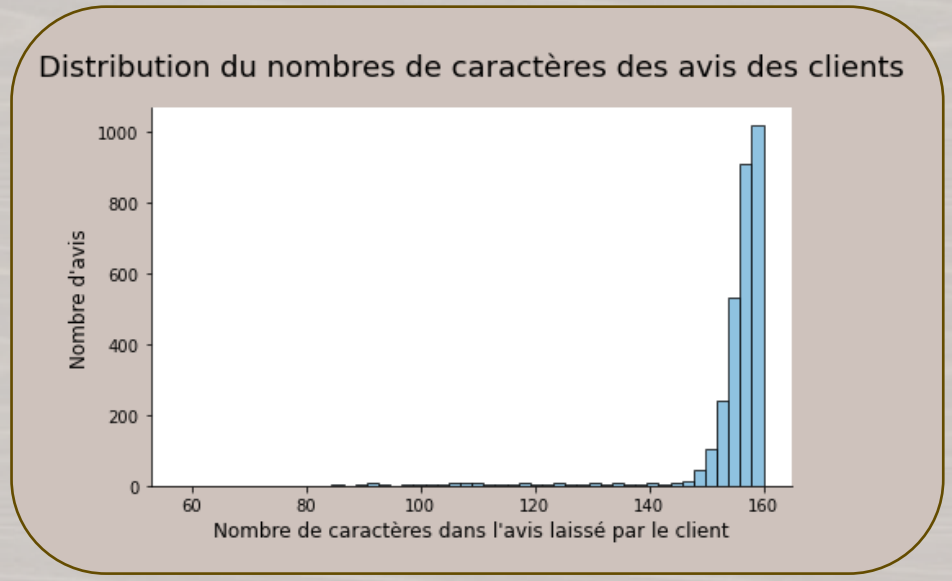
Description du jeu de données académique : les restaurants

Extraction sélective des données



3000 avis récoltés





Description, exploration et transformation du jeu de données

Extraction d'informations à partir des différentes bases de données fournies par Yelp
- Extraction de grandes quantités d'informations

Extraction d'informations à partir de l'API Yelp
- Extraction de données particulières correspondante au besoin défini
- Sauvegarde des données recueillies

Exploration des données et transformation des variables
Evaluation de la pertinence des variables

Prétraitements des données textuelles : nettoyage et transformation

"This is as close to **dining** in Italy as you'll find in New England.\r\nChef Paolo is bringing exquisite, authentic Genovese **cuisine** to the North Shore.\r\nIf you're looking for fine dining in a fun, **casual** atmosphere run, don't walk to Prides Osteria in Beverly."

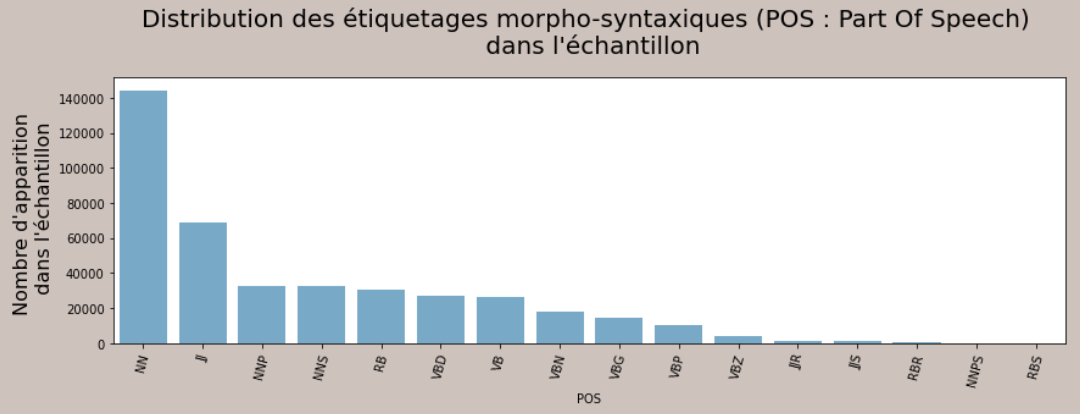
```
Transformation " supprimer fin de lignes " effectuée
Transformation " remplacer les prix " effectuée
Transformation " remplacer les nombres " effectuée
Transformation " supprimer les espaces redondants " effectuée
Transformation " remplacer les liens " effectuée
Transformation " remplacer les I'm par I am " effectuée
Transformation " remplacer les 'll par will' " effectuée
Transformation " remplacer les 'd par would' " effectuée
Transformation " supprimer les 's' " effectuée
```

```
Transformation " tokenisation des phrases " effectuée
Transformation " tokenisation des mots " effectuée
Transformation " étiquetages morpho-syntaxique " effectuée
Transformation " lemmatisation " effectuée
Transformation " stop POS " effectuée
Transformation " stop words " effectuée
Transformation " mots les plus fréquents " effectuée
```

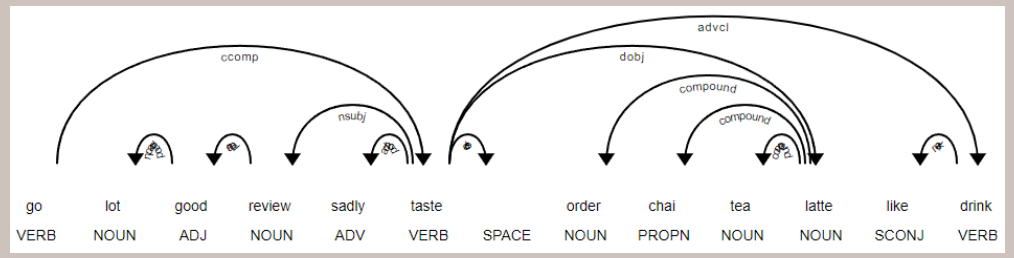
Pipeline
Nettoyage
Transformation

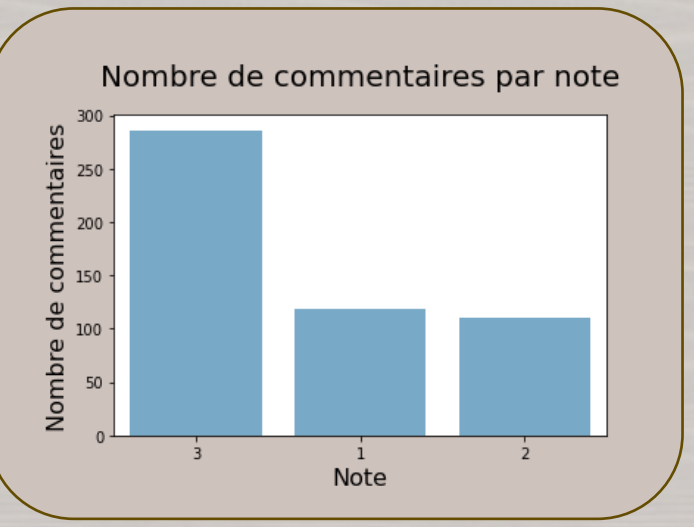
'close din italy find new england chef paolo bring exquisite authentic genovese cuisine north shore look fine din fun casual atmosphere run walk prides osteria beverly'

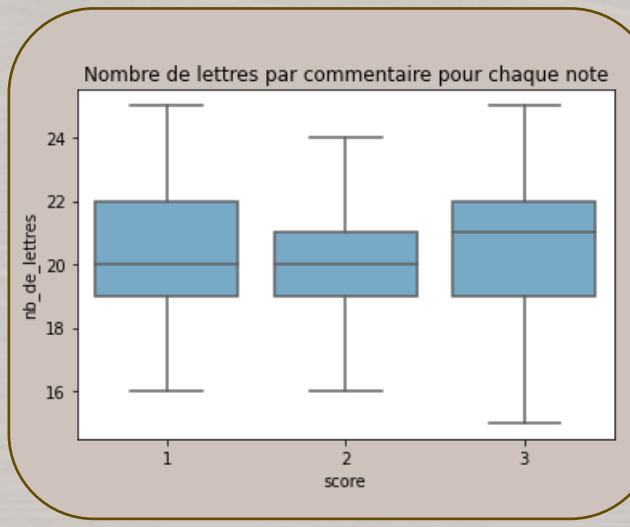
Prétraitements des données textuelles : statistiques et visualisations



'Went here because there were lots of good reviews. Sadly it was not up to taste. \nWe ordered the chai tea latte and I would say it was just like drinking...'







Traitement des données textuelles : Bag of Words – TF-IDF

```
bow_vectorizer = CountVectorizer(min_df=20, max_df=50, ngram_range=(2, 2))
```

Seuil de fréquence minimum

Seuil de fréquence maximum

Bigrammes

mot	nombre
resort fee	78
storm crow	71
love love	69
orlando meats	66
credit card	63
grill pork	62
pork belly	61
cl oz	59
miller ale	57
passion fruit	57

	« love, resort, fee, love, burger »	« love, view, love, chicken, orlando, good, meats, sorry, credit, card, serious, fee»
Resort	1	0
Fee	1	1
Love	2	4
Orlando	0	1
Credit	0	1

Term frequency-inverse document frequency

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

- $TFIDF_{t,d,D}$: Importance d'un terme t dans un document d
- $TF_{t,d}$: Fréquence d'un terme t dans un document d
- $IDF_{t,D}$: Importance du terme t dans l'ensemble des documents D

	« love, resort, fee, love, burger »	« love, view, love, chicken, orlando, good, meats, sorry, credit, card, serious,»
Resort	1/5 * log (2 / 1) = 0,06	0
Fee	0	0
Love	2/5 * log (2 / 2) = 0	0
Orlando	0	0,03
Credit	0	0,03

Détection des sujets : Réduction de dimensions NMF et LSA

NMF

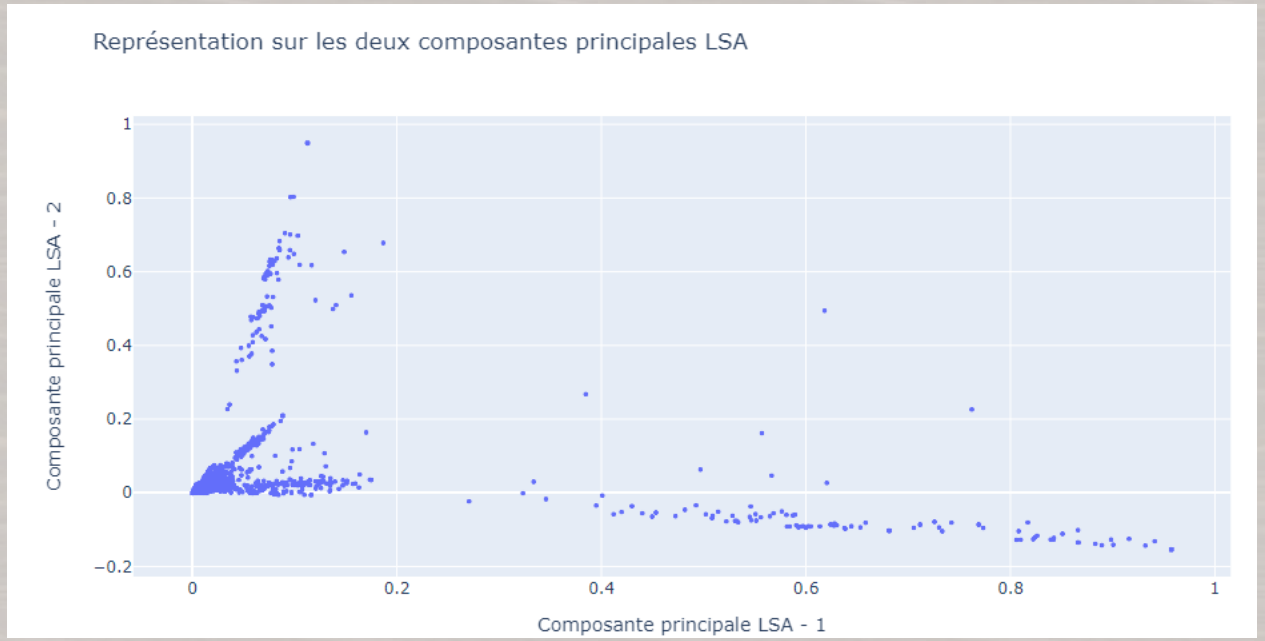
```
Topic 0:
[('lobster roll', 4.91), ('barking crab', 0.63), ('tourist trap', 0.42), ('clam chowder', 0.37), ('crab cake', 0.34), ('fish ch
ip', 0.15), ('crab leg', 0.12), ('new england', 0.07), ('sub par', 0.04), ('stay away', 0.04)]
Topic 1:
[('somewhere else', 4.73), ('save money', 0.2), ('spend money', 0.13), ('barking crab', 0.05), ('call ahead', 0.05), ('year ol
d', 0.05), ('sub par', 0.04), ('front desk', 0.04), ('tourist trap', 0.04), ('medium rare', 0.04)]
Topic 2:
[('nothing special', 4.68), ('ok nothing', 0.45), ('krispy kreme', 0.13), ('high hope', 0.1), ('co worker', 0.1), ('walk door',
0.06), ('short rib', 0.05), ('dan dan', 0.05), ('pay attention', 0.04), ('anytime soon', 0.04)]
Topic 3:
[('credit card', 4.45), ('pay bill', 0.11), ('charge extra', 0.08), ('resort fee', 0.08), ('water glass', 0.07), ('hot pot', 0.
07), ('front desk', 0.05), ('walk away', 0.04), ('call manager', 0.04), ('anything else', 0.04)]
Topic 4:
[('waste money', 4.37), ('anywhere else', 0.11), ('tourist trap', 0.08), ('house blues', 0.05), ('need anything', 0.04), ('new
york', 0.04), ('luke warm', 0.04), ('walk away', 0.03), ('large group', 0.03), ('beef rib', 0.03)]
```

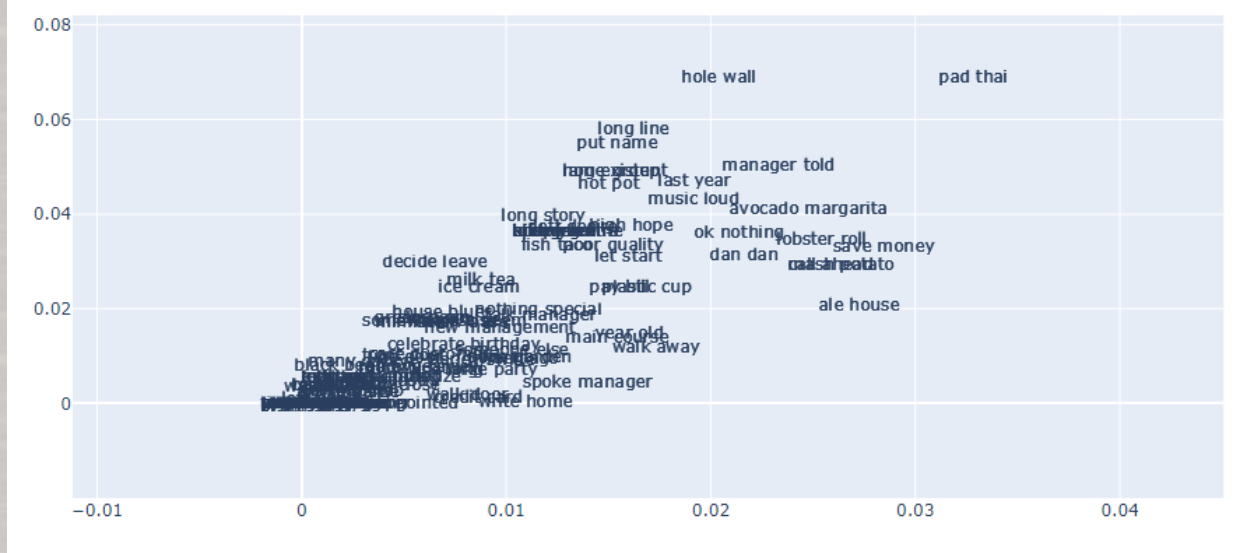
LSA

```
Topic 0:
[('lobster roll', 0.96), ('barking crab', 0.15), ('somewhere else', 0.11), ('tourist trap', 0.1), ('nothing special', 0.09),
('clam chowder', 0.08), ('crab cake', 0.08), ('fish chip', 0.04), ('save money', 0.04), ('crab leg', 0.03)]
Topic 1:
[('somewhere else', 0.95), ('nothing special', 0.21), ('save money', 0.07), ('credit card', 0.07), ('waste money', 0.06), ('spe
nd money', 0.05), ('front desk', 0.04), ('year old', 0.03), ('barking crab', 0.03), ('ok nothing', 0.03)]
Topic 2:
[('nothing special', 0.95), ('ok nothing', 0.1), ('credit card', 0.05), ('co worker', 0.04), ('high hope', 0.04), ('krispy krem
e', 0.03), ('waste money', 0.03), ('save money', 0.03), ('year ago', 0.03), ('walk away', 0.03)]
Topic 3:
[('credit card', 0.98), ('walk away', 0.07), ('waste money', 0.07), ('front desk', 0.06), ('hot pot', 0.05), ('long line', 0.0
4), ('anything else', 0.04), ('pay bill', 0.03), ('charge extra', 0.03), ('resort fee', 0.03)]
Topic 4:
[('waste money', 0.97), ('walk away', 0.09), ('save money', 0.06), ('house blues', 0.06), ('tourist trap', 0.06), ('stay away',
0.04), ('everyone else', 0.04), ('anywhere else', 0.04), ('write review', 0.04), ('barking crab', 0.03)]
```

Détection des sujets : Représentation de données de grandes dimensions : LSA

Projection sur composantes principales





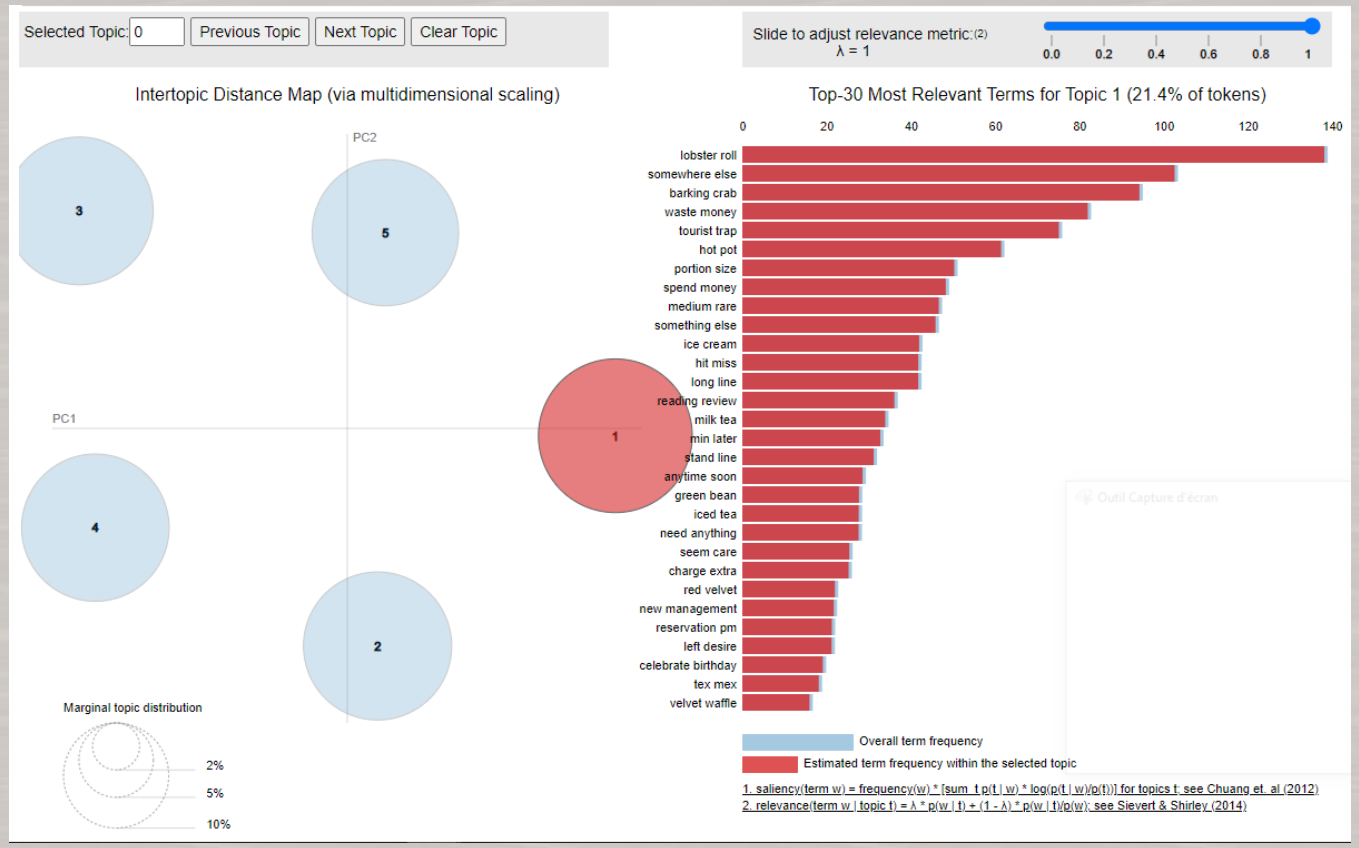
Détection des sujets : Détermination du nombre optimal de sujets : LDA

LDA : Latent Dirichlet Allocation

```
Paramètres du meilleur modèle LDA :  {'n_components': 5}
Meilleur score de vraisemblance logarithmique pour le modèle LDA :  -23704.37
Perplexité du modèle LDA sur les données entraînées :  720.74
```

```
Topic 0:
[('somewhere else', 116.02), ('nothing special', 112.76), ('write review', 85.94), ('front desk', 76.84), ('stay away', 73.52),
('next door', 53.19), ('large party', 51.38), ('large group', 48.1), ('co worker', 47.93), ('fish chip', 47.63)]
Topic 1:
[('waste money', 92.71), ('save money', 78.21), ('barking crab', 71.4), ('ice cream', 47.46), ('hit miss', 47.23), ('long lin
e', 47.22), ('pork belly', 36.94), ('answer phone', 35.48), ('stand line', 35.27), ('main course', 35.27)]
Topic 2:
[('lobster roll', 156.21), ('walk away', 81.85), ('house blues', 71.1), ('sub par', 66.26), ('year old', 63.98), ('spend mone
y', 54.61), ('someone else', 52.78), ('pad thai', 49.4), ('everyone else', 48.99), ('clam chowder', 48.47)]
Topic 3:
[('credit card', 105.24), ('hot pot', 69.45), ('crab cake', 56.18), ('medium rare', 52.73), ('something else', 51.86), ('glass
wine', 46.21), ('potato salad', 41.52), ('mash potato', 40.63), ('high end', 39.66), ('let start', 38.94)]
Topic 4:
[('tourist trap', 84.94), ('year ago', 77.16), ('anything else', 67.4), ('portion size', 56.88), ('high hope', 50.36), ('high e
xpectation', 48.75), ('bottom line', 45.18), ('read review', 44.99), ('chip salsa', 43.26), ('write home', 41.08)]
```

LDA (associé à t-SNE) visualisation pyLDAvis



Compétences

- Pré-traitement expliqués et automatisés
  Etablissement des Bags of Words
  Détermination des features des textes

- Réduction de dimensions (NMF - LSA)
  Visualisation des données de grande dimension (LSA)

- Détection des sujets avec NMF, LSA et LDA pour le corpus académique et pour le corpus via l'API