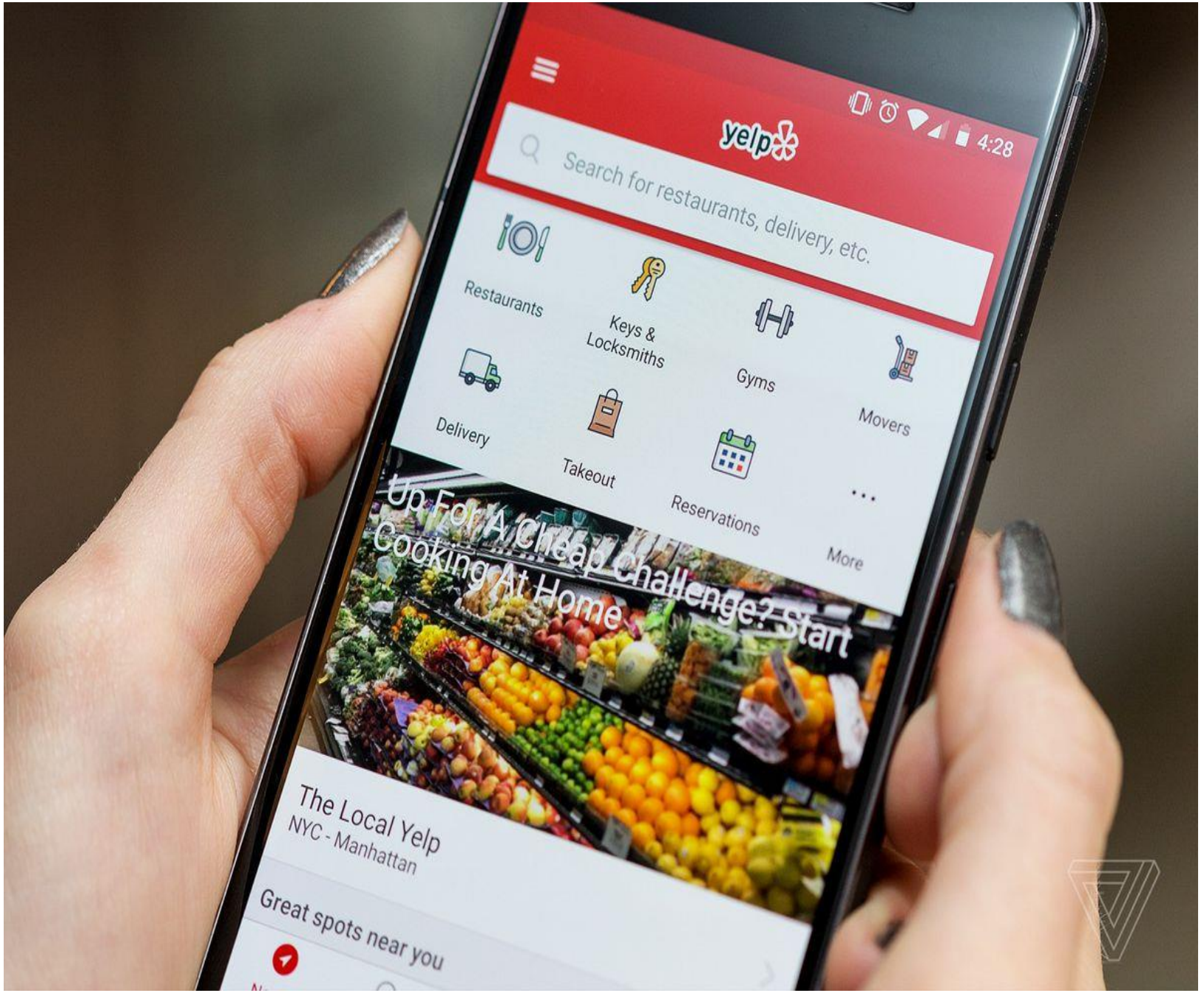
SOUTENANCE

DÉTECTION DE SUJETS D'INSATISFATION
LABELLISATION DE PHOTOS

Cédric Dietzi

Sommaire

- Contexte
- Jeu de données
- Sujets d'insatisfaction
- Labellisation automatique des photos
- Conclusion



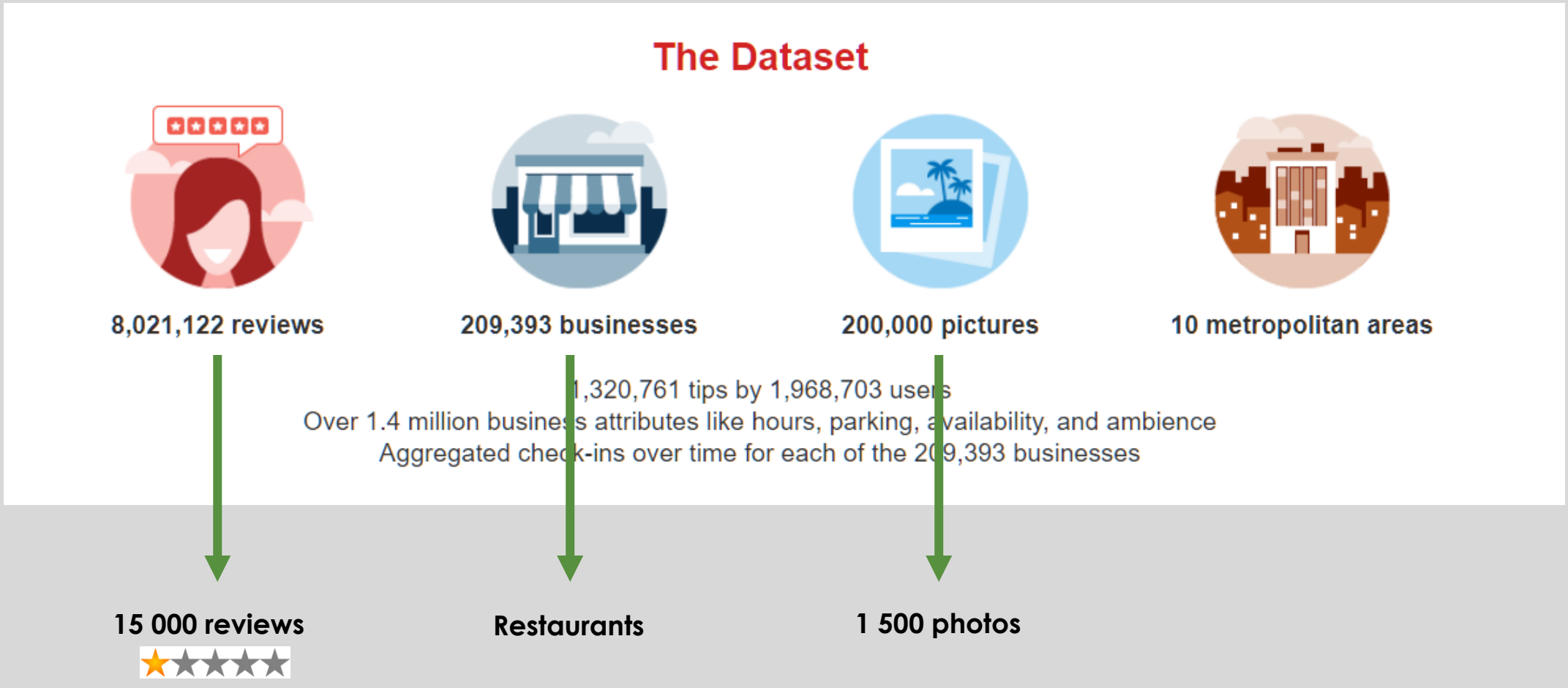
Contexte

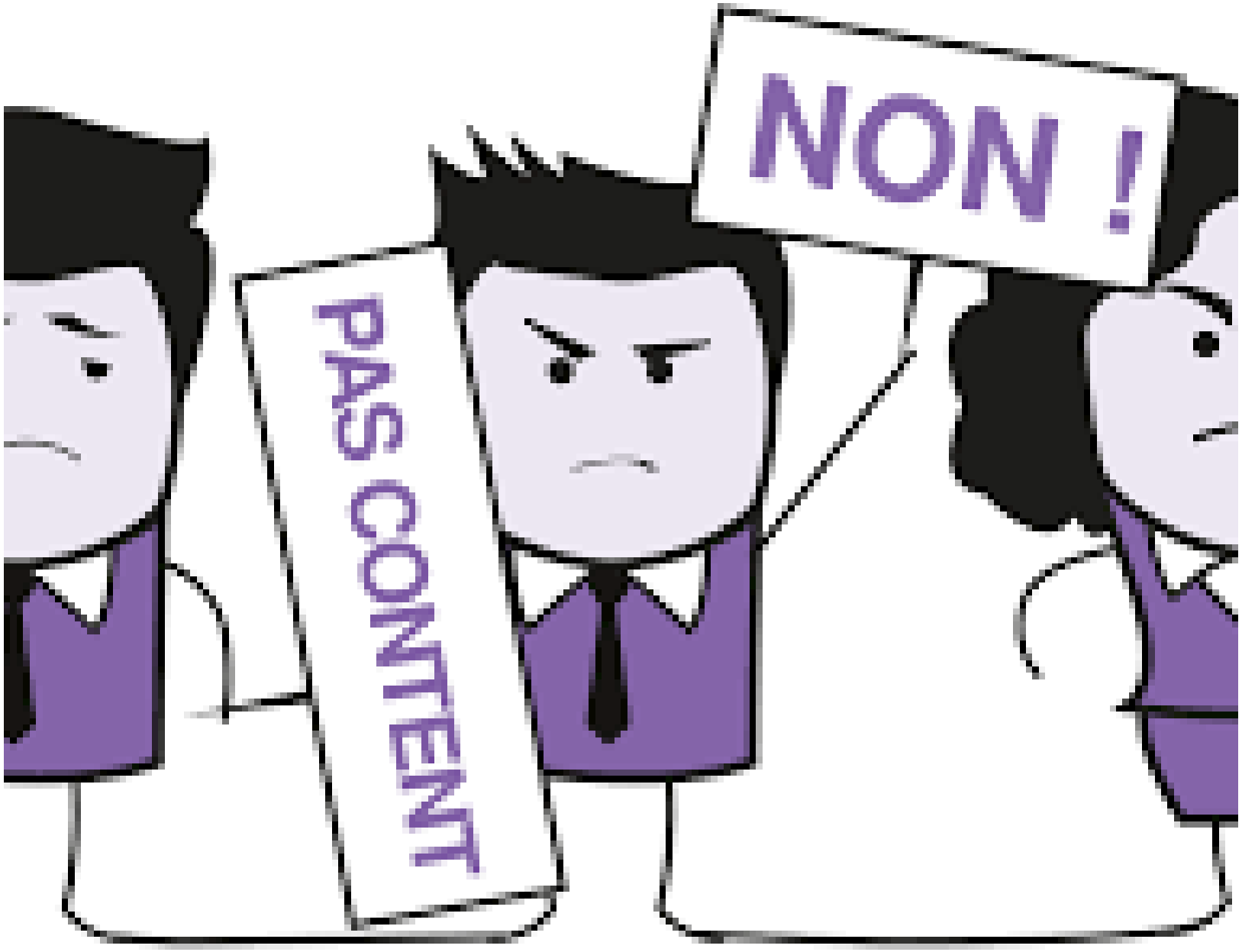
La société **Avis Restau** souhaite développer de nouvelles fonctionnalités de collaboration sur sa plateforme.

- **Détecter les sujets d'insatisfaction**
- **Labelliser automatiquement les photos**

=> Mission: faisabilité de ces fonctionnalités

Le jeu de données





Détection des sujets d'insatisfaction

Peut-on identifier:

1) Quels sont les thèmes récurrents dans les revues 1 étoile ?
2) Dans une revue en particulier, de quoi se plaint le client ?

Si on peut identifier ces 2 points, la détection de topics est faisable et pourra être améliorée

Topic Modeling: chaine de traitement

Pre-processing:

- Suppr. ponctuation
  Casse en minuscules
- Tokenisation
- Sélection des symboles alphabétiques
- Stemming ou Lemmatisation
- Sélection des noms (Part Of Speech processing)
- Suppression des stops-words
- Suppression des mots trop fréquents
- Suppression des mots trop peu fréquents

Vocabulaire de 278 mots

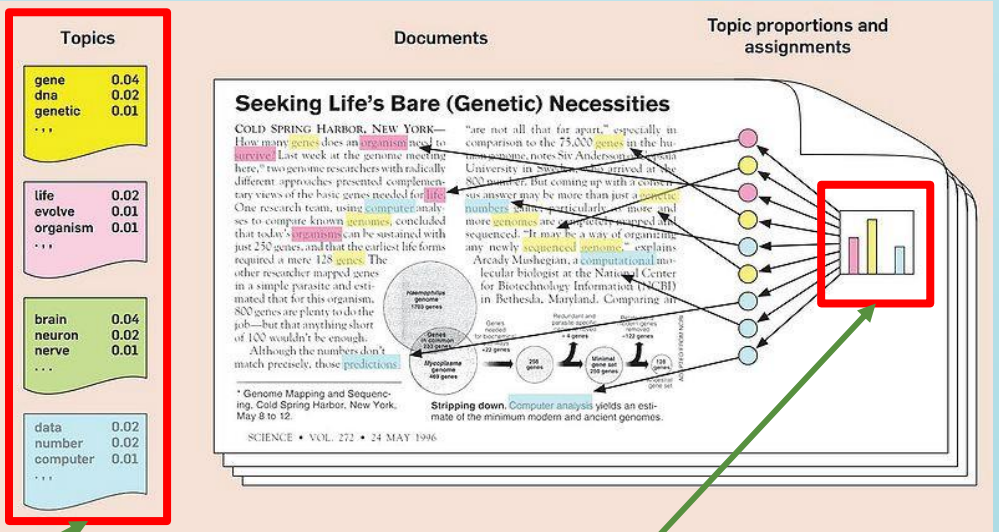
Pre-processing:
Vectorisation des documents

Vectorisation avec Countvectorizer()

Bag of Words

Topics Modeling

Latent Dirichlet Allocation = LDA



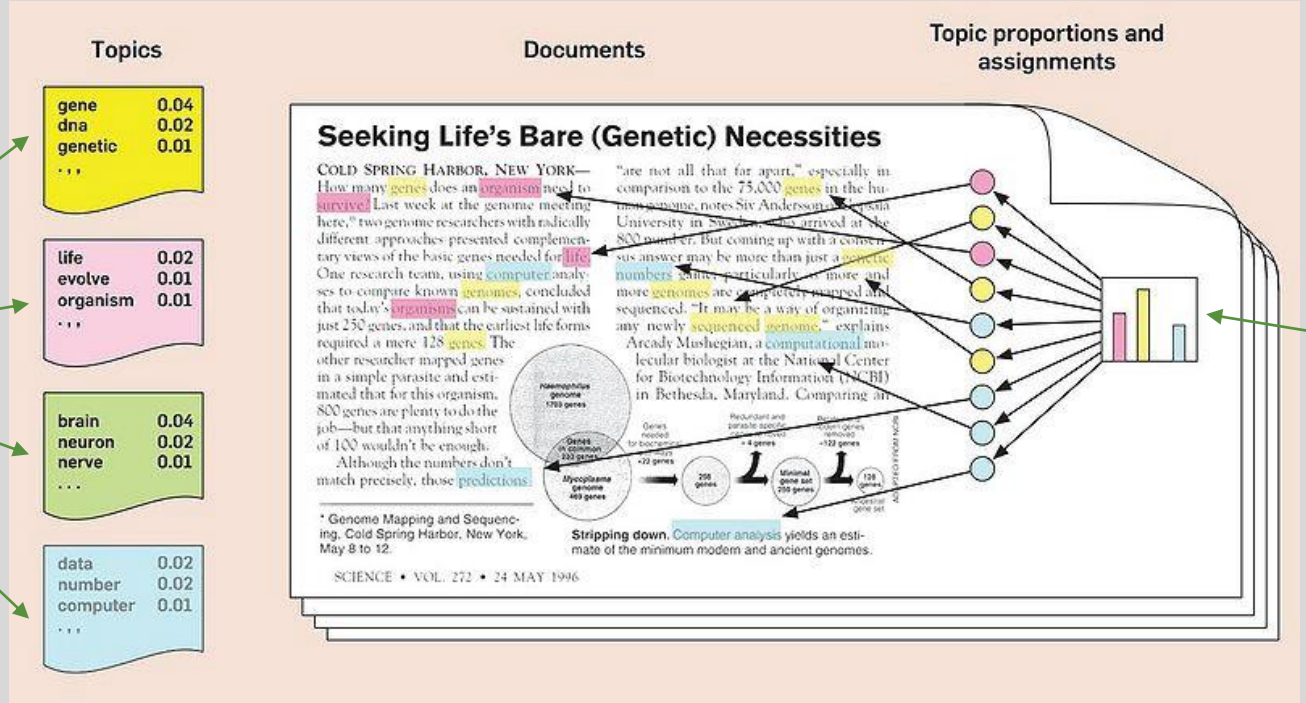
Quels sont les thèmes récurrents dans les revues 1 étoile ?

Dans une revue en particulier, de quoi se plaint le client ?

Topic Modeling: paramétrage du LDA

1) Nombre de topics (sujets)

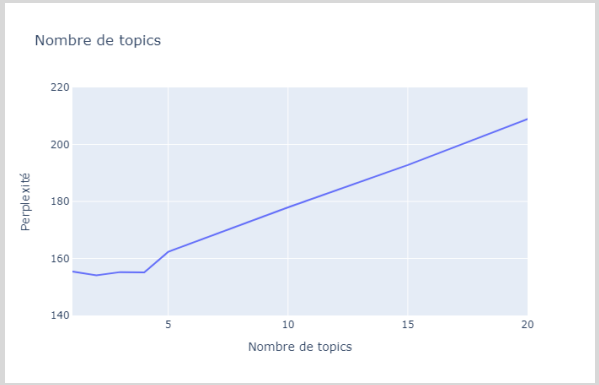
2) Hypothèses sur la distribution du vocabulaire dans chaque topic

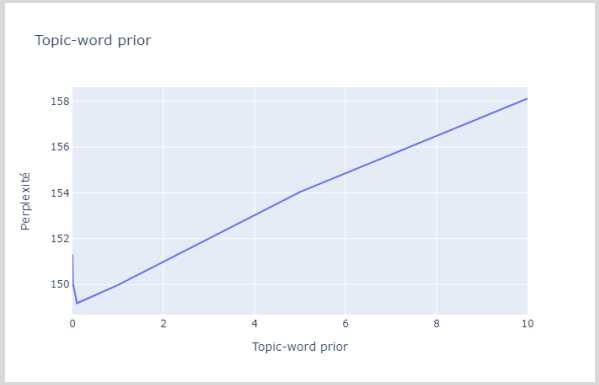


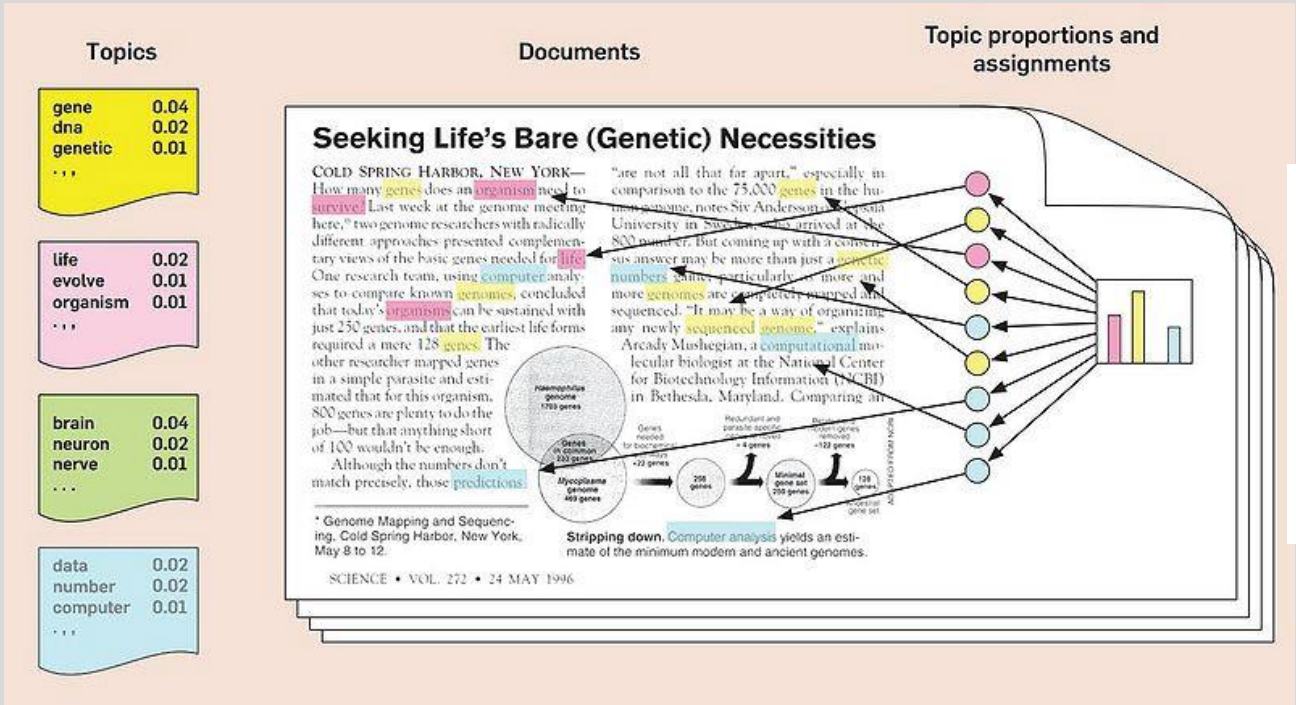
3) Hypothèses sur la distribution des topics dans chaque document

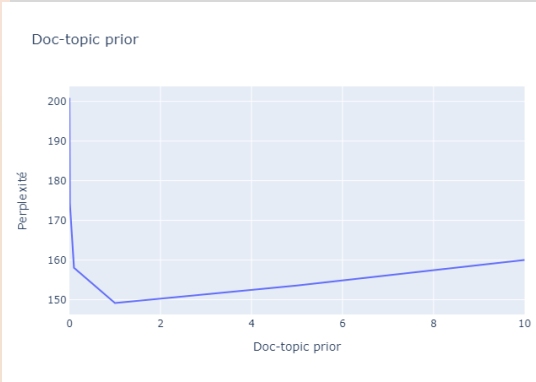
Mesure de performance utilisée pour déterminer les meilleurs paramètres: perplexité

Topic Modeling: paramétrage du LDA





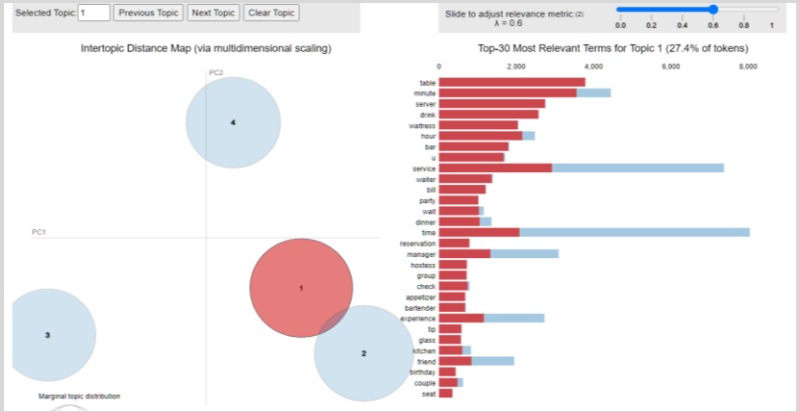




Mesure de performance utilisée pour déterminer les meilleurs paramètres: perplexité

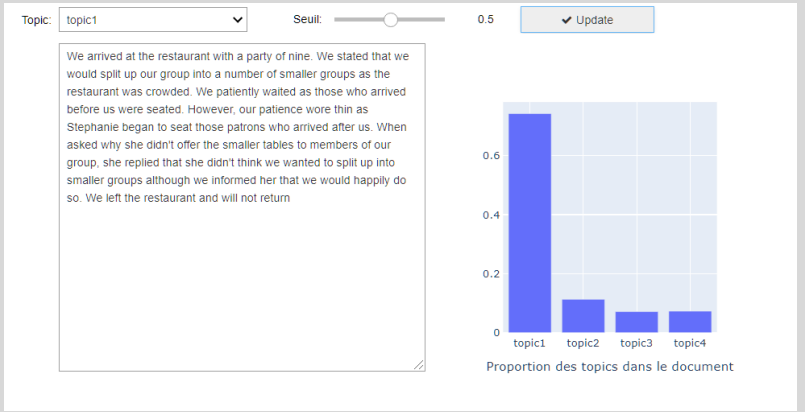
Topic Modeling: démo des résultats

Quels sont les thèmes récurrents dans les revues 1 étoile ?



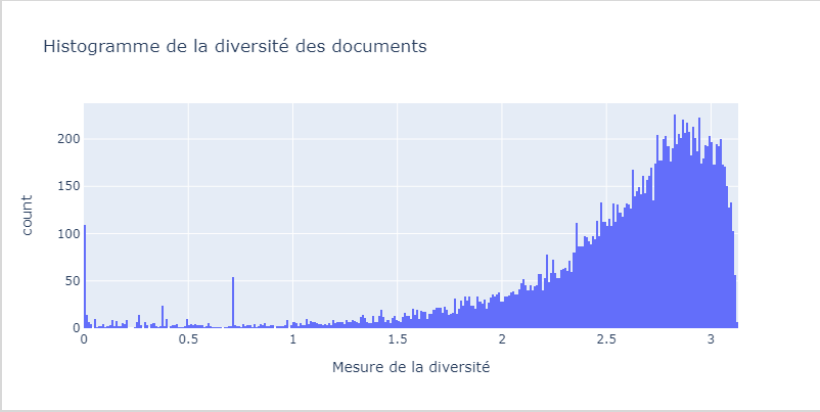
Visualisation par Multi-Dimensional Scaling qui vise à conserver les distances entre points.

Dans une revue en particulier, de quoi se plaint le client ?



En général, un document parle-t-il d'un seul sujet ou de plusieurs sujets très éloignés ?

D'après "Text-Based Measures of Document Diversity - Bache, Newman, Smyth



Labellisation automatique des photos

Peut-on clustériser facilement les photos ?

Si oui, c'est qu'une labellisation automatique est possible.

2 approches testées:

- SIFT + Bag of visual words
- Convolutional neural network

Labellisation automatique: chaine de traitement

Pré-traitement:

Echantillonnage de 300 photos par label	
Conversion noir et blanc	S
Redimensionnement	C
Filtrage du bruit	S
Egalisation	S

S: Bag of Visual Words (Sift)

C: transfer learning (CNN)

Photos (300 par label)

Features engineering:

- SIFT + Visual Bag of Words
- Transfer learning from a CNN

Features

Clustering:

- Réduction de dimension: PCA
- Réduction de dimension: TSNE
- K-Means

Labellisation automatique: SIFT + Bag of Words

Calcul du vocabulaire

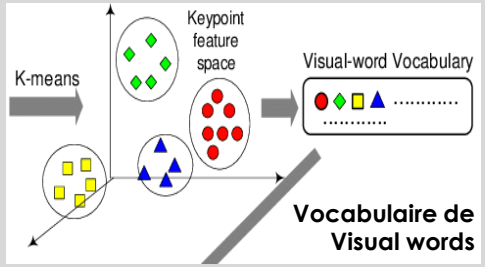
SIFT: calcul des descripteurs de toutes les photos d'une classe

Suppression des outiers

K-Means:

⇒ calcul des visual words associés aux descripteurs

= centroïdes des clusters de descripteurs



1500 photos
1,9 M descripteurs
1285 Visual Words

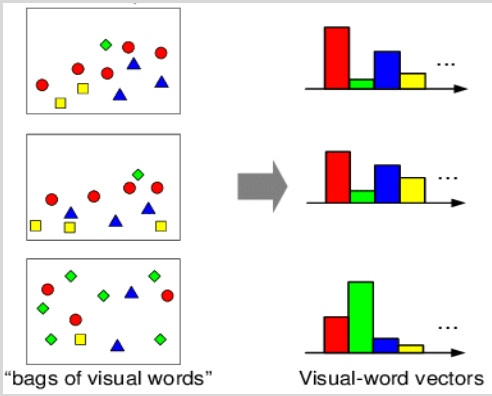
Clustering

SIFT: calcul des descripteurs de chaque photos

Pour chaque document:

Bag of visual Words: association des descripteurs aux visual words

Histogramme



Features: (1500, 1 285)

Labellisation automatique: CNN

1500 Photos

Model: "functional_1"

Layer (type)	Output Shape	Param #
input_1 (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
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312

Total params: 134,260,544
Trainable params: 0
Non-trainable params: 134,260,544

Features: (1500, 4096)

Labellisation automatique : démo des résultats

SIFT ne sépare correctement que les menus qui sont des photos plus caractéristiques que les autres.

CNN permet de très bien séparer les différentes classes de photos.

- ✓ Faisabilité de la détection de sujets d'insatisfaction
- ✓ Faisabilité de la labellisation automatique