Comparaisons des métiers dans différentes villes: Lausanne, New York & Paris

Groupe Lausannuaire

Problème de recherche

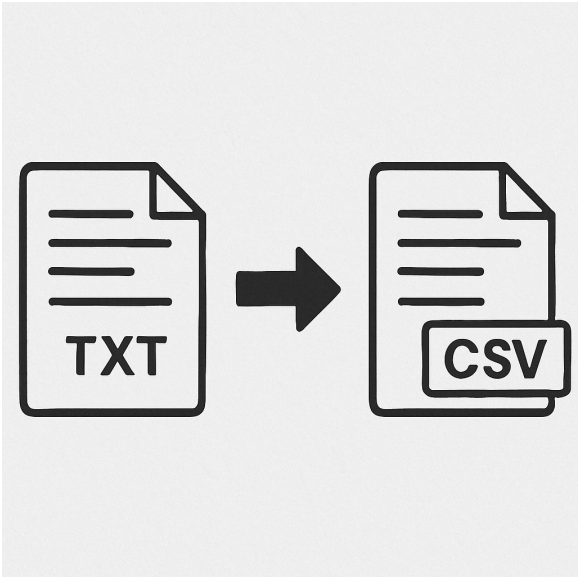
- Comment les métiers évoluent à **Lausanne** entre **1885** et **1951** ?
- Quelles différences de métiers entre **Lausanne**, **Paris** et **New York** en **1885** ?

Présentation de la source

- Annuaires de Lausanne **1885**, **1901**, **1923** et **1951**
- Annuaire de Paris **1885**
- Annuaire de New York **1885**

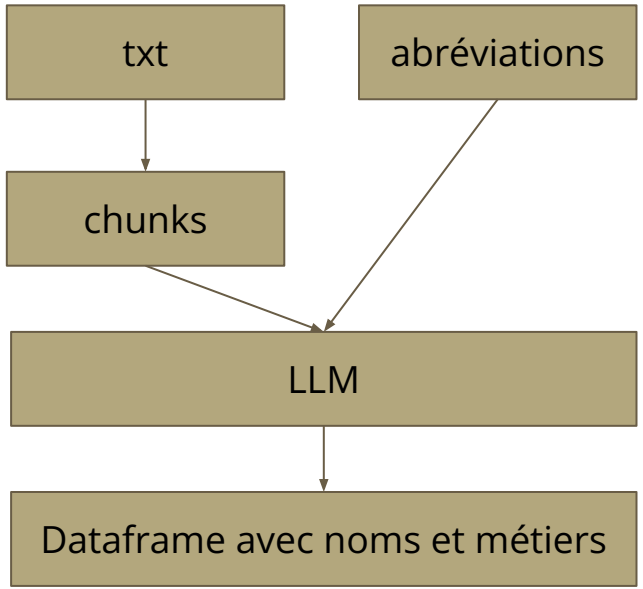
Comprenant **Nom**, **Addresse**, **Métier**,
occasionnellement des **Coordonnées géographiques**.

Méthodologie



Extraction du dump

- Création d’une pipeline pour extraire les noms et les métiers
- Appliquée à l’annuaire de New-York 1885
- performances
  - 310’746 noms dans l’annuaire
  - 261’217 noms extraits
  - 119’872 après nettoyage
  - => ~40% de l’annuaire

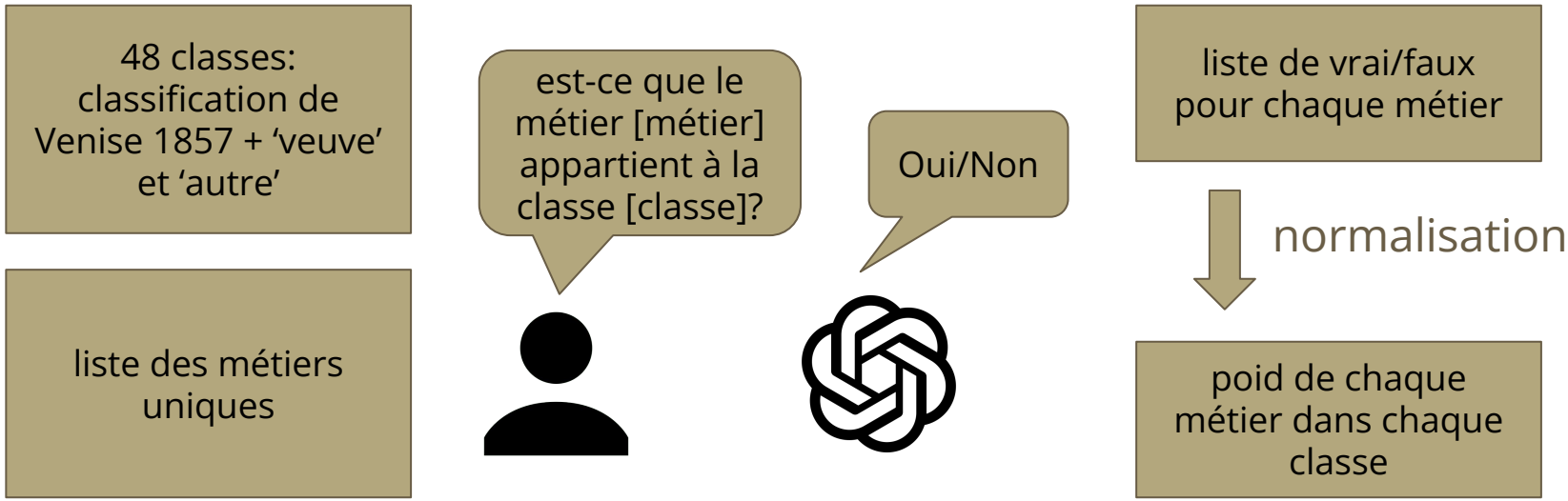


Extraction du dump: difficultés

- Hallucinations
- Tendance à oublier la 2eme colonne

Clustering des métiers

clustering des embeddings peu concluant => changement d'approche



Clustering des métiers: limitation et difficultés

- approche par clustering d'embedding non analysable
  - soit trop de classes (>600) soit trop peu (3) avec HDBSCAN
  - clusters qui ne représentent pas des groupes de métiers avec Kmeans
- problème des métiers émergents
  - nouvelles classes de métiers seront toutes mises comme "autre"
  - classification de venise pas forcément adaptée à toutes les villes
- problème des noms de métiers
  - certains métiers ont changé de nom
  - certains métiers sont abrégés dans les donnés
- classification parfois fausse
  - pas de manière simple d'évaluer si la classification d'un métier est correcte ou non

Alignement des adresses : méthode automatique

Combinaison de deux mesures

Distance de levenshtein

- proximité textuelle

Cosine distance sur les embeddings

- proximité sémantique

Alignement des adresses : réglage des paramètres

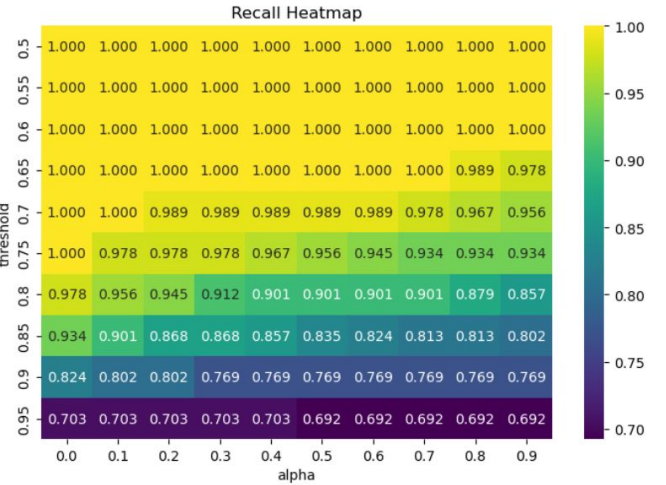
threshold (0.9) seuil d'acceptation d'une association

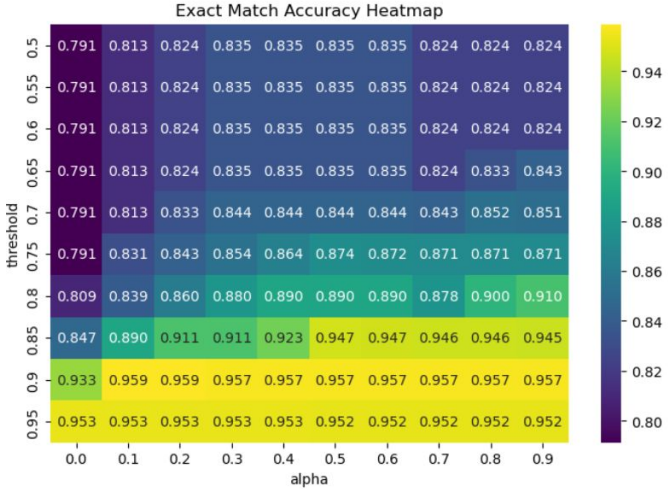
alpha (0.2) importance accordée à la distance de levenshtein

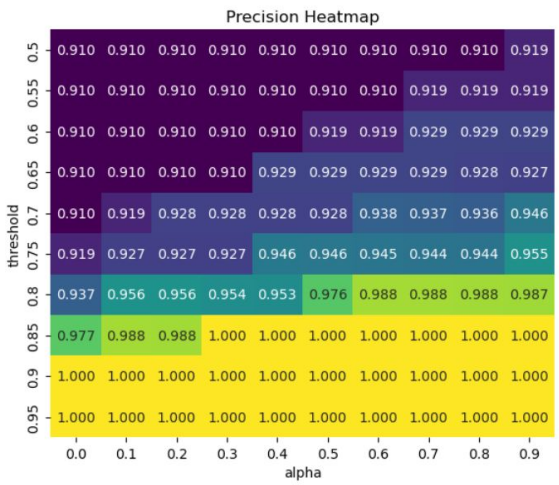
1 - alpha (0.8) importance accordée à la similarité des embeddings

Alignement des adresses : réglage des paramètres

Test sur 100 addresses alignées manuellement







Alignement des adresses : taux de réussite

Lausanne

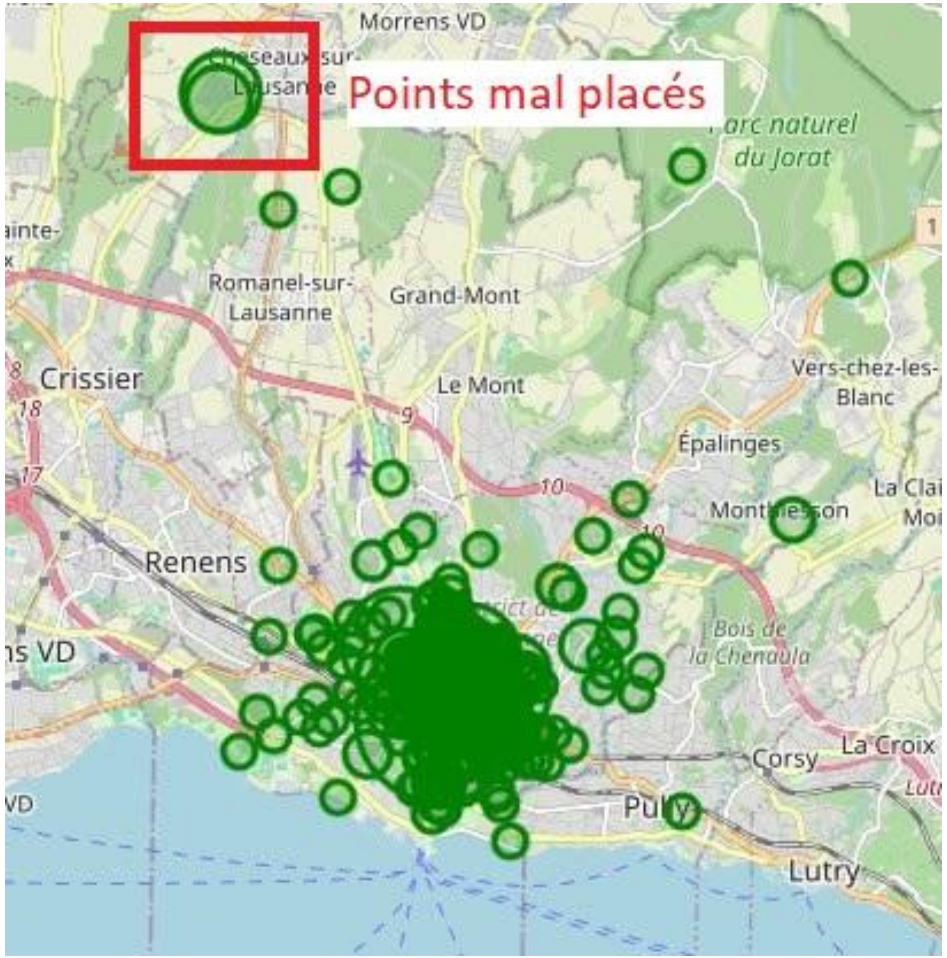
- 1885: 65%
- 1901: 45%
- 1951: 64%

Paris

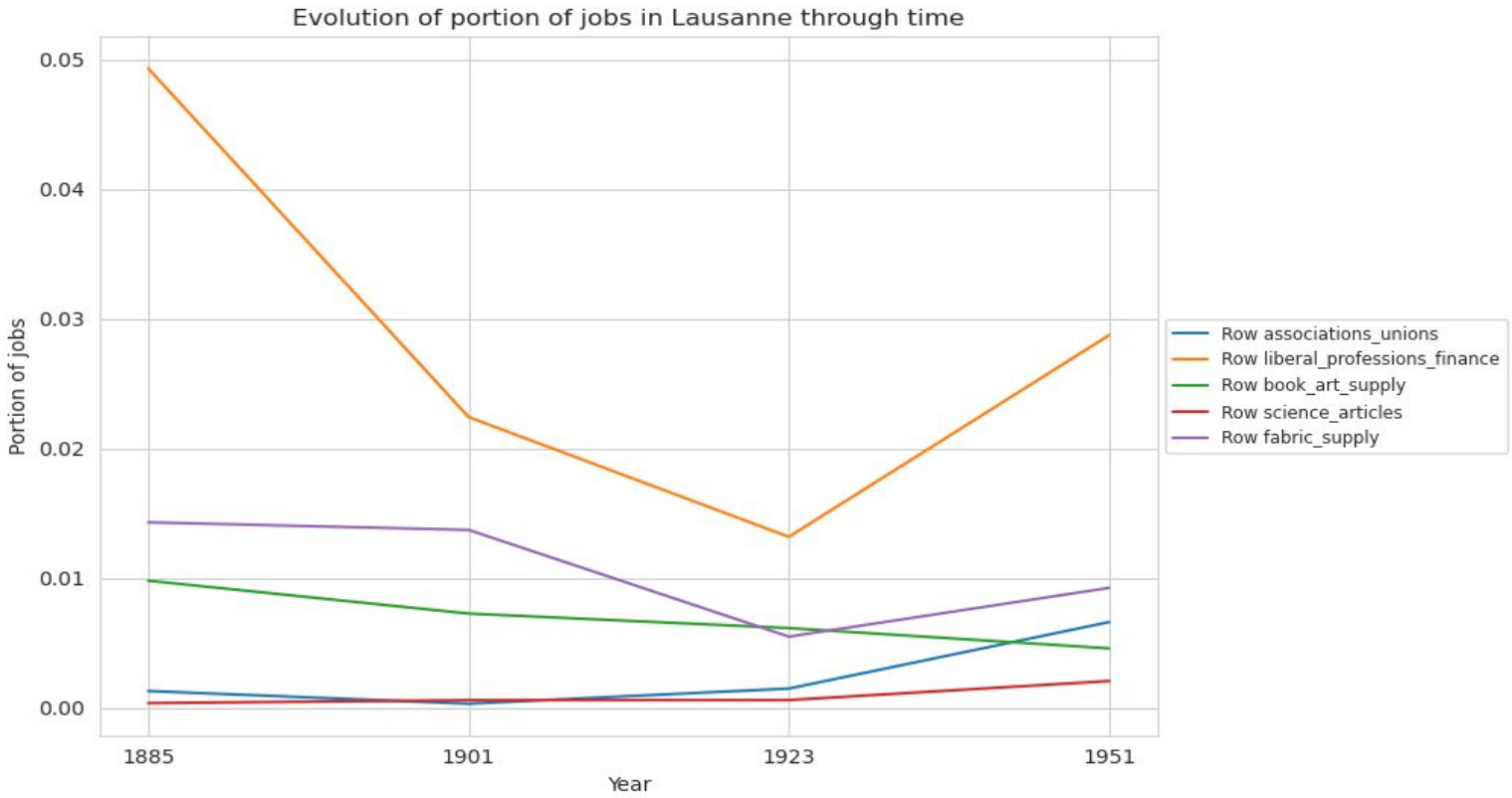
- 1885: 88%

Alignement des adresses : difficultés

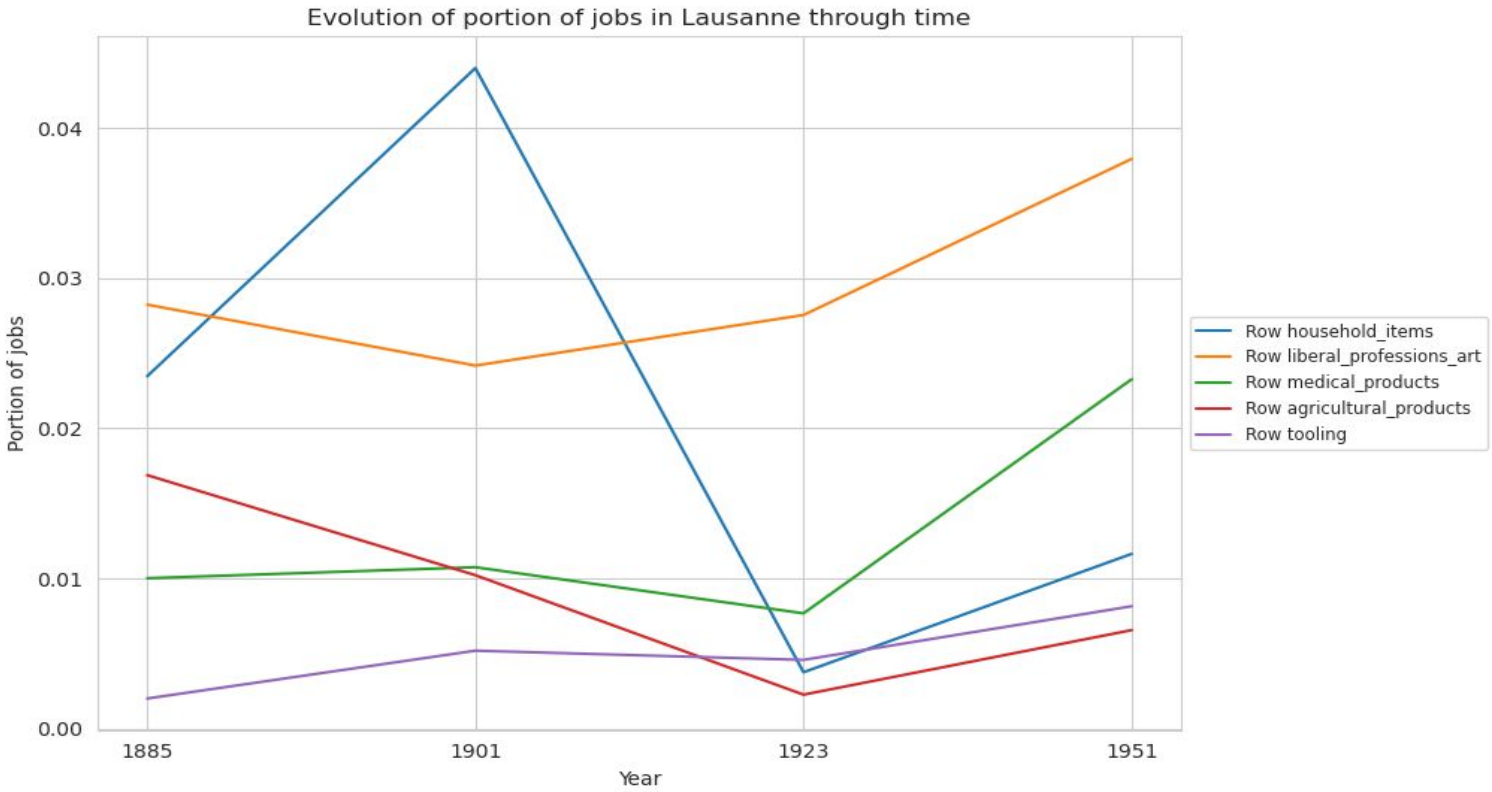
- Traitement des fichiers
- Paramètres de l’algorithme
- Erreurs dans les fichiers de référence
- Numéros de rues manquants



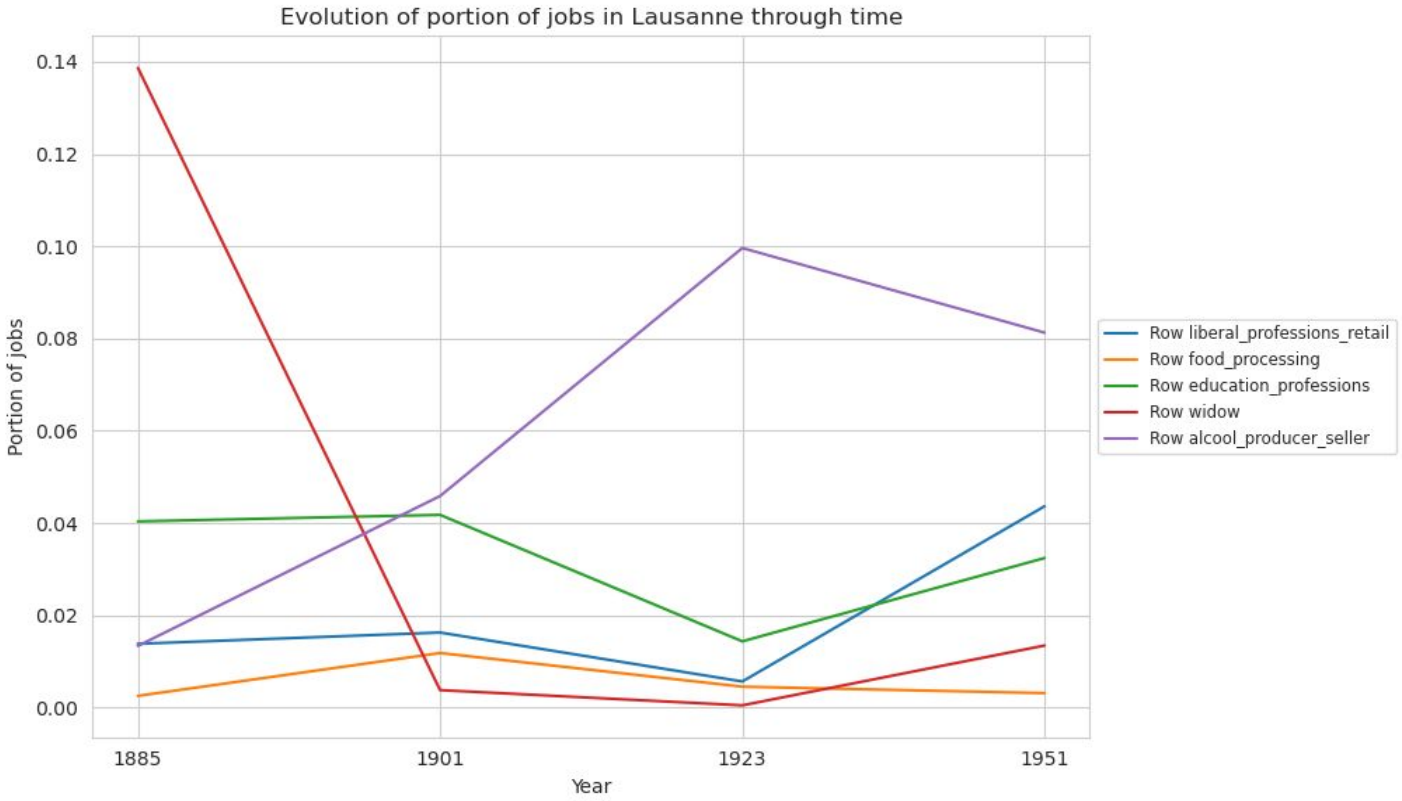
Observations sur l'évolution de Lausanne



Observations sur l'évolution de Lausanne

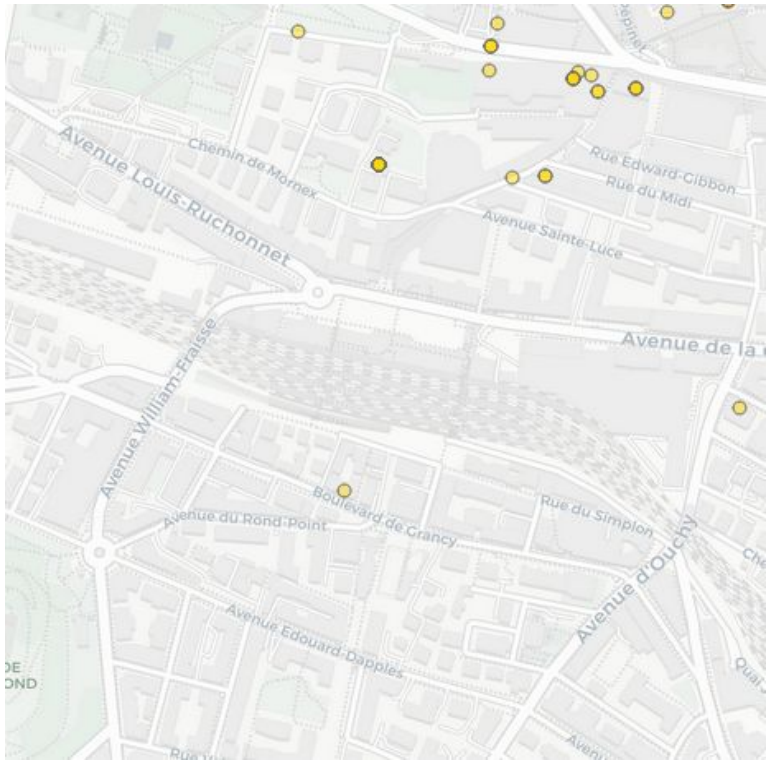


Observations sur l'évolution de Lausanne

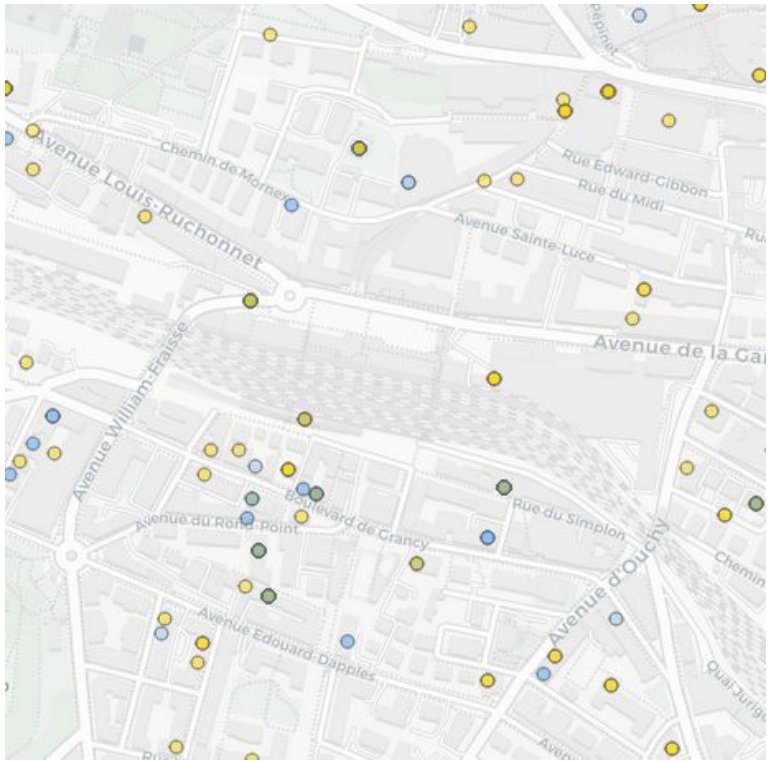


Observations sur l'évolution de Lausanne

Hospitality lodging



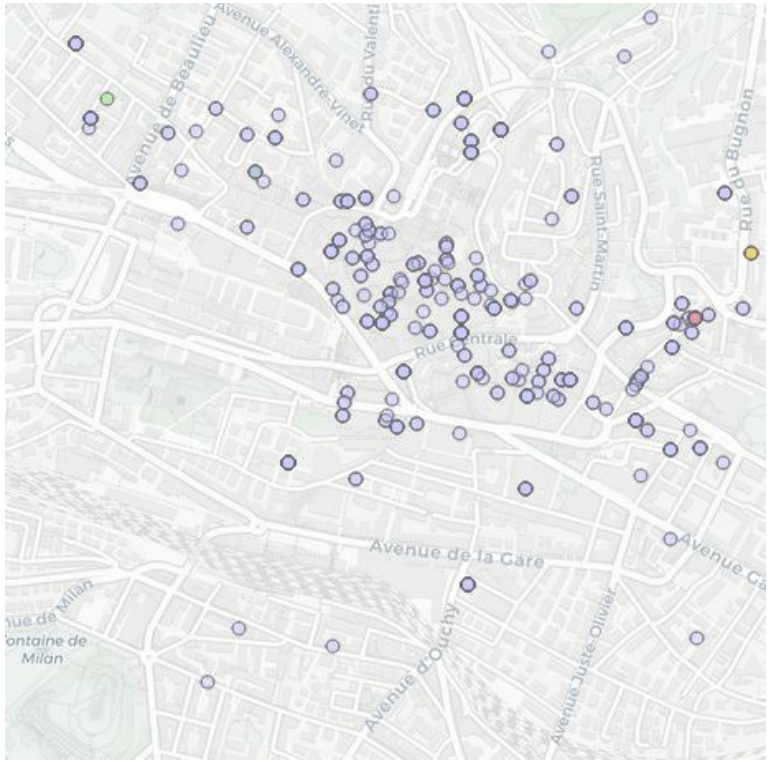
1885



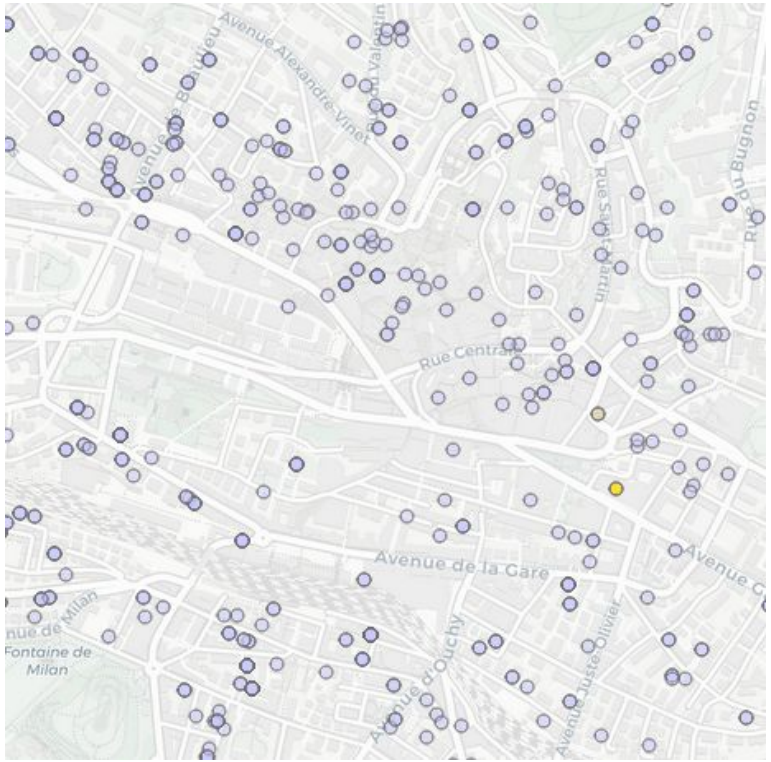
1951

Observations sur l'évolution de Lausanne

Transports



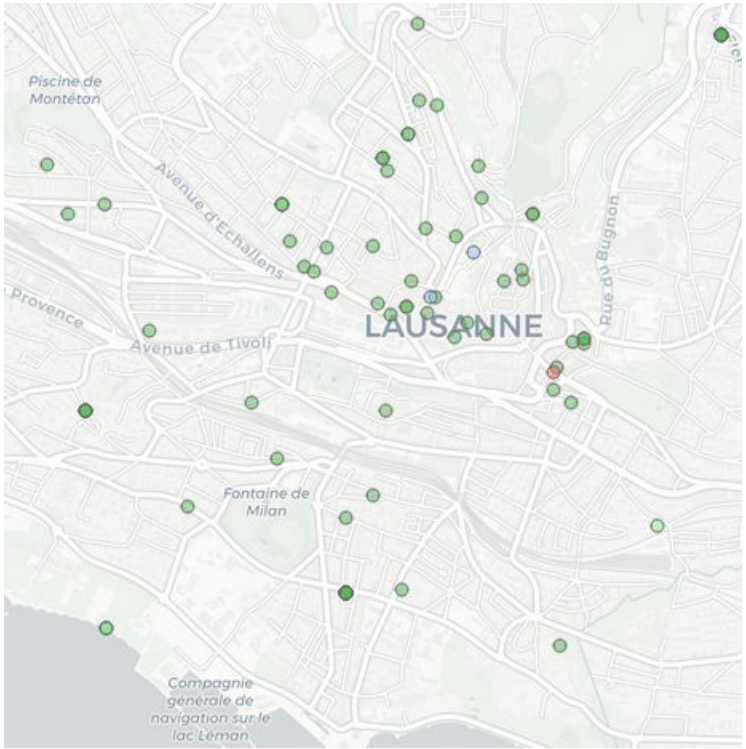
1885



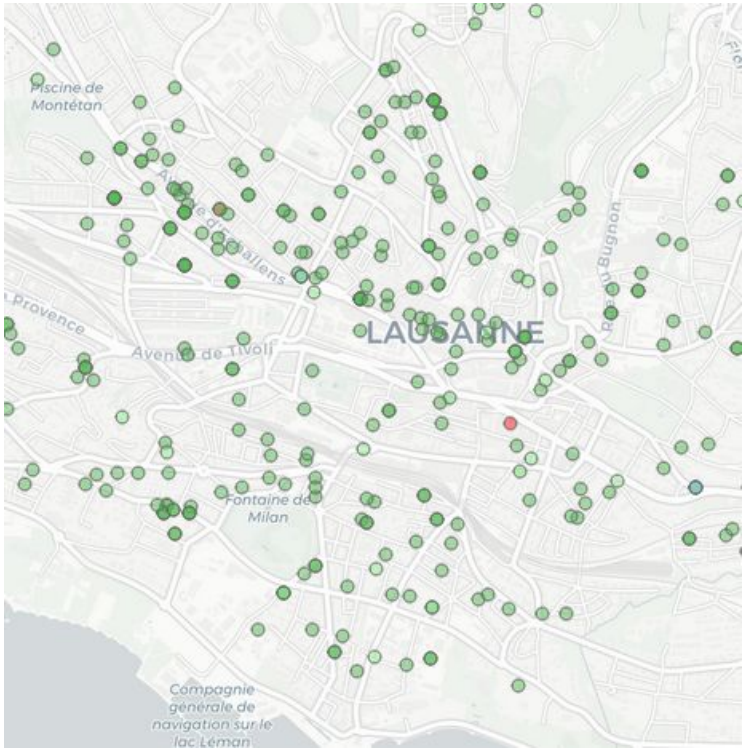
1951

Observations sur l'évolution de Lausanne

Métier agricole



1885



1951

Comparaison des villes en 1885

- D’avantage d’associations/syndicats à Lausanne et Paris qu’à New York
- Part plus importante des professions du transport à Lausanne
- Domaine du luxe / vêtements plus important à Paris

Présentation du site

[https://projects.lausannetimemachine.ch/student-project-2025-lausannuaire/]

Merci de votre attention

Questions ?