ANALYSE SÉMANTIQUE COMPUTATIONNELLE ET SPATIALISÉE DU CORPUS DES CAHIERS CITOYENS

Sami Guembour

Analyse sémantique computationnelle et spatialisée du corpus des Cahiers citoyens - Caractérisation de «corpus de petite taille» à l'aide de profils sémantiques

Sami GUEMBOUR
Univ Gustave Eiffel, ENSG, IGN, LASTIG / CAMS, EHESS

Contexte

- Déroulement du Grand Débat National (GDN) de janvier à mars 2019 en réponse aux mouvements des Gilets Jaunes (novembre 2018).
- Mise en place d'une plateforme de participation citoyenne en ligne.
- Lancement de l'opération "Mairies Ouvertes" par l'Association des Maires Ruraux de France (AMRF) et dépôt des Cahiers Citoyens (CC) au niveau des mairies , offrant aux citoyens la possibilité de s'exprimer librement.
- Production de corpus de taille considérable (Corpus GDN et Corpus CC).

Problématique

Caractérisation des corpus de petite taille se concentrant sur des problématiques spécifiques liées à des régions géographiques à l'aide de profils sémantiques.

Corpus de Travail

- Corpus des Cahiers citoyens (CC).
- 225 224 contributions citoyennes.
- Contributions manuscrites et dactylographiées.
- Géolocalisation des contributions par code postal et code INSEE.

Objectifs

- Analyse des contributions du corpus des Cahiers citoyens d'un point de vue sémantique, et spatial.
- Caractérisation et comparaison des corpus de petite taille en fonction de critères thématiques et géographiques.

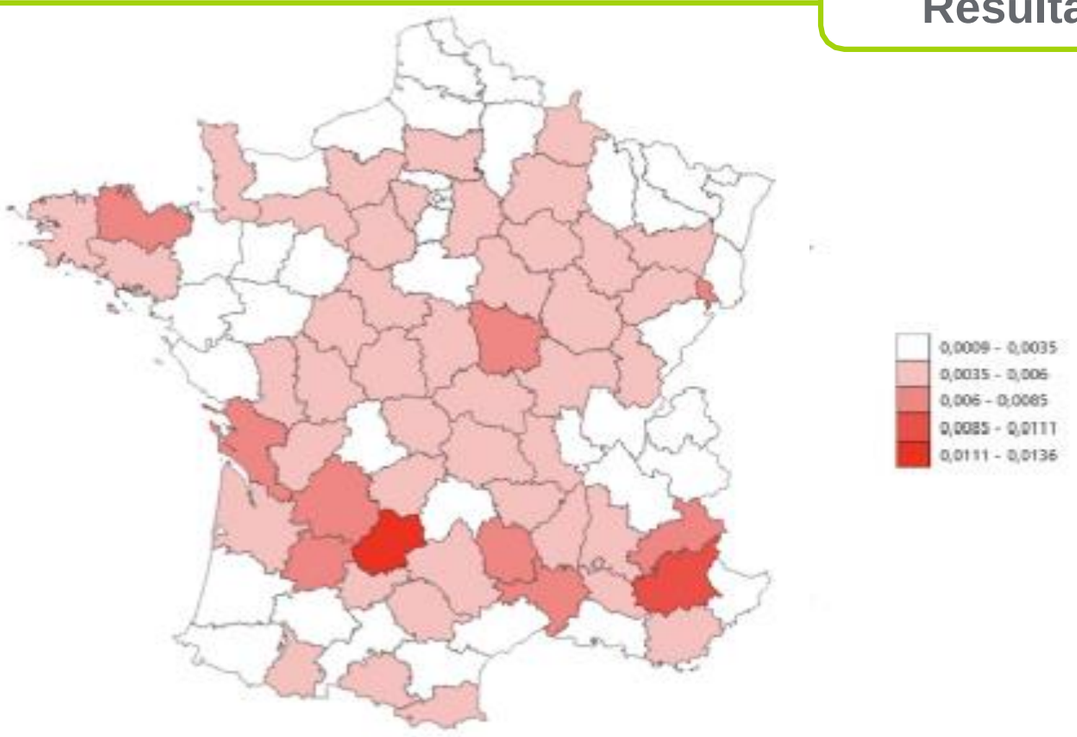
Hypothèses

1. Influence de la localisation des contributeurs sur les problématiques abordées dans les contributions.
2. Caractérisation des profils sémantiques des contributions par des mots-clés spécifiques, des réseaux lexicaux et des phrases types.
3. Variation des profils sémantiques des contributions en fonction des caractéristiques socio-démographiques des contributeurs.

Méthodes

- Segmentation des contributions en phrases à l'aide d'outils de segmentation [1] afin d'obtenir une unité d'étude plus fine et précise.
- Prétraitement des contributions en utilisant des techniques de traitement automatique des langues.
- Utilisation des modèles de langue fondés sur des réseaux de neurones [2] pour pour calculer les vecteurs de phrases et extraire les informations de contexte.
- Identification des problématiques abordées par les citoyens en effectuant un clustering sur les vecteurs de phrases.
- Construction de graphes de liens sémantiques entre les termes et les phrases abordant des problématiques similaires pour une analyse sémantique.
- Utilisation des méthodes de plongement de graphes pour extraire les liens sémantiques.

Résultats attendus



Nombre de contributions par habitant par département dans le corpus CC [3]

En Géographie :

- Représentation cartographique des problématiques abordées dans les contributions citoyennes par région géographique.
- Représentation cartographique des profils sémantiques par région géographique.

En Traitement Automatique des Langues :

- Analyses contrastives du corpus des Cahiers citoyens.
- Extraction des profils sémantiques des problématiques abordées dans les contributions.
- Construction des corpus de petites tailles.

Planning

- **1ère année :**
  - État de l'art et organisation des concepts.
  - Prétraitement et segmentation des contributions.
  - Regroupement des contributions par problématique.
- **2ème année :**
  - Caractérisation des corpus de petite taille.
  - Analyse sémantique et spatialisée des corpus de petite taille.
- **3ème année :**
  - Publication et interprétation des résultats obtenus.
  - Rédaction de la thèse.

Informations

- **Date de début de la thèse :** 01/10/2023
- **Institutions :**
  - LASTIG, IGN
  - Université Gustave Eiffel
  - CAMS, EHESS
- **Domaines :**
  - Traitement Automatique des Langues (TAL)
  - Science des données
- **Directrices de thèse :**
  - Catherine DOMINGUÈS
  - Sabine PLOUX

Bibliographie

1 HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spacy : Industrial- strength natural lan- guage processing in python. Journal of Open Source Soft- ware, 5(51).

2 MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language mode. Proceedings of the 58th Annual Meeting of the As- sociation for Computational Linguistics.

3 CHANDORA S. (2023). Fouille sémantique et spatiale dans le corpus cahiers citoyens : comparaison de méth- odes symbolique et numérique.