Analyse sémantique du corpus des Cahiers citoyens

Sami Guembour

Analyse sémantique du corpus des Cahiers citoyens

Sami GUEMBOUR[1]
(1) LASTIG, Univ Gustave Eiffel, ENSG, IGN, France
sami.guembour@ign.fr

RÉSUMÉ

Cet article présente une recherche originale qui se concentre sur une analyse sémantique du corpus des Cahiers citoyens, qui regroupe les contributions et les doléances des citoyens français déposées au niveau des mairies dans le cadre du Grand Débat National. L'article offre un état de l'art complet sur les divers travaux réalisés sur ce corpus et vise à obtenir une compréhension approfondie des thèmes émergents et des préoccupations citoyennes dans les différentes régions. Plusieurs hypothèses concernant ces travaux ont été émises, et différentes méthodes ont été proposées pour répondre à ces hypothèses, de la segmentation et du pré-traitement du corpus au calcul des vecteurs de plongement des phrases à l'aide de modèles de langues pré-entraînés, aboutissant au clustering de ces vecteurs pour construire des regroupements en fonction des problématiques abordées.

ABSTRACT

Semantic analysis of the "Cahiers citoyens" corpus

This article presents an original research focusing on a semantic analysis of the "Cahiers citoyens" (Citizen Notebooks) corpus, which compiles the contributions and grievances of French citizens submitted at the municipal level as part of the "Grand Débat National" (Grand National Debate). The article provides a comprehensive state-of-the-art review of various studies conducted on this corpus, aiming to achieve a profound understanding of emerging themes and citizen concerns in different regions. Multiple hypotheses regarding these studies have been formulated, and various methods have been proposed to address these hypotheses, ranging from corpus segmentation and pre-processing to calculating sentence embedding vectors using pre-trained language models. This culminates in clustering these vectors to construct groupings based on the addressed issues.

MOTS-CLÉS : Cahiers Citoyens - Grand débat National - Corpus - TAL - Modèle de langue - Vecteur de plongement - Classification.

KEYWORDS: Citizen Notebooks - Grand National Debate - Corpora - NLP - Language Model - Vector embedding - Clustering.

1 Introduction

La crise sociale déclenchée par le mouvement des Gilets Jaunes en France à l'automne 2018 a engendré une série de réponses institutionnelles, et plusieurs formes de participation citoyenne ont vu le jour. En décembre 2018, l'Association des Maires Ruraux de France (AMRF) [1] a lancé l'opération "Mairies Ouvertes". L'idée était de mettre des "Cahiers de doléances et de propositions" à disposition dans les mairies, offrant aux habitants une opportunité de s'exprimer librement. Ce qui devait être

1. https ://www.amrf.fr/

une action courte du 8 au 15 décembre a été prolongé en raison de son succès inattendu.

En janvier 2019, le gouvernement français a lancé le Grand Débat National (GDN) [2], offrant à la fois une plateforme numérique dématérialisée et des supports matériels, les Cahiers citoyens, disponibles dans des lieux publics. Certains Cahiers de doléances et de propositions ont maintenu leur dénomination initiale, tandis que d'autres sont transformés en Cahiers citoyens. À la clôture de la période de contribution mi-mars 2019, les Cahiers de doléances et de propositions sont enrichis par ceux des Cahiers citoyens, créant ainsi une dualité entre les expressions en ligne des citoyens via la plate-forme officielle et les contributions des Cahiers citoyens.

Cet article s'inscrit dans le contexte d'une recherche qui entre dans le cadre d'une thèse. L'objectif de cette recherche est d'analyser les Cahiers citoyens de manière sémantique et spatiale en utilisant les méthodes et outils du Traitement Automatique des Langues. Dans ce contexte, l'analyse sémantique des Cahiers citoyens consiste à examiner le contenu textuel du corpus afin de comprendre les significations et les relations sémantiques entre les termes, les phrases, et les thèmes abordés. Quant à l'analyse spatiale du corpus, elle consiste à examiner comment les caractéristiques géographiques des citoyens, telles que leur lieu de résidence ou leur origine, sont liées aux thématiques abordées dans leurs contributions. Cette initiative tire ses fondements des résultats d'un travail antérieur (Chandora, 2023), au cours duquel des clusters regroupant des phrases abordant des thématiques similaires dans les contributions ont été construits. Cependant, les résultats n'ont pas atteint la satisfaction escomptée car 91 % des phrases n'ont pas été classées, incitant ainsi à entreprendre une nouvelle démarche de recherche plus approfondie visant à améliorer le regroupement de ces phrases.

Le plan de ce papier s'articule autour de plusieurs sections. La section 2 se consacre à la définition et à la présentation du corpus des Cahiers citoyens. La section 3 présente les travaux déjà entrepris en matière d'analyse des Cahiers citoyens. Dans la section 4, nous abordons la construction des hypothèses qui sous-tendent notre approche, ainsi que la définition des objectifs. La section 5 détaille la méthodologie que nous adopterons pour analyser le corpus et vérifier les hypothèses formulées. Enfin, la section 6 conclut l'article et évoque les principales attentes.

2 Définition du corpus de travail

Le corpus utilisé dans cette étude, désigné sous le nom de Cahiers citoyens (CC), rassemble des contributions provenant des habitants de diverses communes. Il s'agit des contenus rédigés par les citoyens et déposés au niveau des mairies pour exprimer leurs préoccupations. Elles ont été collectées à partir de divers supports d'expression fournis par les mairies participantes, incluant des carnets d'écoliers, des courriers électroniques, et des supports papiers avec des thèmes prédéfinis. Les Cahiers citoyens contiennent des contributions variées, allant de textes manuscrits ou dactylographiés à des courriers électroniques directement adressés aux mairies, des dossiers comportant parfois des pièces jointes, ainsi que des pétitions collectives dactylographiées. Plus de 16 000 communes ont participé à cette initiative. Ces contributions diverses sont associées à un code INSEE facilitant la localisation des communes. La consultation de ces Cahiers est soumise à une dérogation accordée par les Archives Nationales [3], accompagnée d'une clause de protection des données, en raison du caractère privé des informations qu'ils renferment.

2. https ://www.gouvernement.fr/le-grand-debat-national
3. https ://www.archives-nationales.culture.gouv.fr/

Le processus de construction du corpus a débuté par la collecte des Cahiers des différentes mairies, qui ont été transmis aux préfectures pour numérisation, générant ainsi un corpus de fichiers image. Ces fichiers images ont ensuite été envoyés à la Bibliothèque nationale de France (BnF), qui, en utilisant des outils d'OCR, a converti les fichiers en format texte, créant ainsi un corpus textuel où chaque fichier représente un ou plusieurs cahiers localisés. La BnF a également fait appel à trois prestataires pour vérifier la transcription automatique, la vérification portant notamment sur le découpage des cahiers en contributions, les métadonnées de chaque contribution, ainsi que sur leur contenu textuel. La concaténation de ces fichiers textuels a abouti à la création d'un fichier au format CSV, désormais appelé Corpus CC.

Le tableau 1 fournit des statistiques descriptives détaillées sur le corpus CC. La tokenisation des contributions s'est effectuée à l'aide de l'outil NLTK (Bird *et al.*, 2009).

TABLE 1 – Statistiques descriptives du corpus CC

Nombre total de contributions	225 224
Nombre total de tokens dans le corpus	55 838 490
Nombre de codes postaux uniques	5 551
Nombre moyen de tokens par code postal	10 059
Nombre de codes INSEE uniques	16 421
Nombre moyen de tokens par code INSEE	3 400
Nombre de dates de réception uniques	85

3 Travaux antérieurs

La recherche sur le Grand Débat National et les Cahiers citoyens a été marquée par des défis d'accès et des approches variées. Contrairement au GDN, dont le corpus est accessible en open data, l'accès aux Cahiers citoyens est complexe en raison du Règlement Général sur la Protection des Données (RGPD). De ce fait, plusieurs travaux se sont concentrés sur le corpus GDN et sur des débats alternatifs : Entendre la France [4] (une application Messenger inspirée des questions du GDN) et le Vrai Débat [5] (une plate-forme contestataire créée par des Gilets Jaunes), diversifiant les méthodes employées.

Dans (Ploux *et al.*, 2021), les auteurs ont opté pour une analyse sémantique des corpus GDN, Entendre la France et Vrai Débat en utilisant des réseaux lexicaux. Leur méthodologie, fondée sur l'identification de "cliques" via des calculs de co-occurrences, a révélé des variations thématiques en fonction de la taille des communes, apportant une perspective intéressante sur les dynamiques territoriales.

Le point de départ de l'analyse du corpus CC est la synthèse réalisée par (Berger *et al.*, 2019), commandée par le gouvernement, mais critiquée pour ses objectifs politiques et son opacité méthodologique. L'agence Cognito Consulting [6] a joué un rôle central dans cette analyse. La méthode, fondée sur une cartographie sémantique, a identifié des "clusters lexicaux", mais le manque de transparence quant à l'algorithme et la rapidité d'exécution soulèvent des interrogations sur la qualité des résultats.

4. https ://www.entendrelafrance.fr/
5. https ://levraidebat.org
6. https ://www.cognito.fr/

(Ray, 2023) a utilisé des modèles d'extraction de sujets tels que BERTopic (Grootendorst, 2022) et LDA (Latent Dirichlet Allocation) (Blei *et al.*, 2003) pour analyser les contributions du corpus des Cahiers citoyens, cherchant à comparer les résultats avec l'analyse précédente de Cognito et les catégories du corpus GDN. Cette approche offre une nouvelle perspective sur la diversité des thèmes abordés dans les contributions citoyennes.

(Monnier, 2023) a réalisé une étude approfondie sur la thématique éolienne à partir des Cahiers citoyens. Son travail s'inscrit dans une analyse transversale en sciences sociales, combinant des approches linguistiques et géographiques. L'analyse a ciblé trois départements où la thématique éolienne était prépondérante, permettant une analyse des contributions en fonction des caractéristiques naturelles et sociales spécifiques à chaque territoire. La production de cartes a été utilisée pour visualiser de manière spatialisée les extractions textuelles.

Dans son analyse exploratoire du corpus des Cahiers citoyens, présentée lors de la Journée d'étude "Cahiers citoyens 2019 : approches croisées" [7], Ploux a utilisé des méthodes fréquentielles qui s'appuient sur les variations de fréquence des mots par rapport à un corpus de référence. Ses observations ont révélé que les contributeurs abordaient des sujets différents en fonction de la taille de leur commune. Une analyse fréquentielle a également conduit à l'observation que la taille des communes était inversement corrélée au nombre des contributions de la commune. De plus, Ploux a formulé l'hypothèse que les sujets abordés dans les Cahiers citoyens seraient plus variés, voire différents, de ceux du corpus du GDN. Cette divergence s'expliquerait par le format libre des Cahiers citoyens, contrairement au GDN qui était structuré autour de quatre thèmes prédéfinis avec des questions associées, encadrant ainsi les productions des auteurs.

Le travail de (Bendinelli, 2023) se distingue par son approche fine et approfondie axée sur l'analyse d'un seul cahier provenant de la commune de Dole dans le Jura. Cette étude s'appuie sur une combinaison de disciplines, notamment la linguistique, les approches communicationnelles et sémiotiques de l'écrit, ainsi que l'analyse du discours outillée. Bien que ces méthodes offrent une compréhension approfondie du contenu et des nuances des contributions, il est important de noter qu'elles présentent un caractère peu automatisé. De ce fait, leur applicabilité à l'ensemble du corpus est limitée.

Enfin, (Chandora, 2023) a proposé une nouvelle approche dans l'exploration du corpus CC. Elle s'attache à une double perspective, alliant une analyse sémantique approfondie à une évaluation de la répartition géographique des préoccupations citoyennes. Son étude s'articule autour d'une analyse du vocabulaire complet et de la distribution géographique des contributions, révélant des thèmes et des caractéristiques propres au corpus CC. La fouille sémantique s'appuie sur deux méthodes distinctes. La première, le clustering à partir de plongements de phrases, identifie des propositions de contributeurs, tout en soulignant la pertinence de l'unité de la phrase pour l'exploration d'un corpus. Cependant, la qualité de la segmentation en phrases est impactée par la nature du corpus. La deuxième méthode, utilisant des automates à états finis, permet d'extraire un nombre plus important de séquences textuelles pour les propositions identifiées. La représentation spatiale des propositions citoyennes est explorée, combinant des techniques de TAL et des représentations cartographiques. Les résultats indiquent que les différences thématiques observées entre les propositions sont globalement peu marquées, ne semblant pas être spécifiques à des zones géographiques particulières en France. Cependant, ses résultats montrent que les citoyens les plus actifs, ayant le plus contribué aux Cahiers citoyens, résident plutôt dans de petites communes lorsqu'on rapporte les résultats au nombre d'habitants.

7. https ://geographie-cites.cnrs.fr/cahiers-citoyens-2019-approches-croisees/

4 Hypothèses et objectifs

L'hypothèse principale des travaux concernant les Cahiers citoyens est que les problématiques abordées dans les contributions dépendent de la localisation des contributeurs, et que les caractéristiques géographiques et socio-démographiques (telles que l'âge, le genre, le niveau d'éducation, l'appartenance sociale, etc.) de ces derniers peuvent influencer les thématiques qu'ils abordent. Comme explicité dans la section 1, cette recherche vise à améliorer les résultats obtenus par une étude antérieure (Chandora, 2023). Au cours de celle-ci, des groupes ont été formés en fonction des thèmes abordés dans les contributions à l'aide d'un clustering qui utilise les phrases comme unités d'étude. Les contributions ont été segmentées en phrases à l'aide de l'outil Unitex [8], ensuite les vecteurs de plongement de ces phrases ont été calculés avec le modèle de langue *sentence-camembert-base* [9] (Martin *et al.*, 2020; Nils Reimers, 2019). Pour former les groupes de phrases à partir de ces vecteurs, l'algorithme de *Fast Clustering* [10] a été choisi. Cependant, les résultats de ce clustering se sont avérés décevants, ne classant que 9% des phrases, ce qui ne permet pas de prétendre une analyse sémantique complète.

Dans le but d'atteindre les objectifs de cette recherche, des améliorations des résultats du clustering des phrases en augmentant le nombre de phrases classées ont été envisagées. Pour ce faire, des hypothèses remettant en question le choix des trois méthodes suivantes dans (Chandora, 2023) ont été émises :

— **Segmentation des phrases :** La première hypothèse s'oriente vers l'outil de segmentation choisi pour découper les contributions en phrases. Elle avance que Unitex n'est peut-être pas l'outil le mieux adapté au corpus des Cahiers citoyens, étant donné que le niveau de langue dans les contributions est varié et que leur typographie ainsi que leur syntaxe ne sont pas fiables, puisqu'elles proviennent de diverses catégories sociales . Cette non-conformité entraîne une segmentation imprécise et de qualité médiocre.
— **Modèle de langue :** Cette hypothèse se focalise sur le modèle *sentence-camembert-base* utilisé pour calculer les vecteurs de phrases, et estime qu'il ne garantit pas la meilleure représentation des phrases. Cela signifie que des phrases sémantiquement similaires ne sont pas suffisamment proches dans l'espace vectoriel, et que les distances qui les séparent ne sont pas minimales. Ceci pourrait conduire à des dispersions éloignées des phrases abordant des thématiques similaires dans cet espace vectoriel.
— **Algorithme de clustering :** La dernière hypothèse porte sur la sélection de l'algorithme de clustering et la configuration des hyperparamètres. Elle suggère que l'algorithme *Fast Clustering* ne facilite pas une agrégation efficace et complète des phrases abordant les mêmes thématiques. Il est également possible que les valeurs des hyperparamètres fixées ne soient pas appropriées, compromettant ainsi la capacité de l'algorithme à proposer un regroupement optimal, et entraînant par conséquent la dispersion de phrases abordant des thématiques similaires dans des clusters distincts.

8. https ://unitexgramlab.org/
9. https ://huggingface.co/dangvantuan/sentence-camembert-base
10. https ://www.sbert.net/examples/applications/clustering/README.html

5 Méthodologie

Les méthodes suivantes ont été proposées pour répondre aux hypothèses émises dans la section 4. L'objectif est d'obtenir des groupes de phrases regroupant, dans les mêmes clusters, celles traitant des mêmes sujets et thématiques, à travers l'application d'un nouveau processus de clustering, tout en proposant une nouvelle segmentation des contributions en phrases, un pré-traitement et un nettoyage des phrases résultantes, ainsi qu'un nouveau modèle pour calculer leurs vecteurs.

5.1 Nouvelle segmentation en phrases

La première hypothèse formulée remettait en question l'outil de segmentation en phrases utilisé et la qualité de son découpage. Afin d'améliorer la qualité du découpage en phrases, différentes approches de segmentation ont été explorées, impliquant le test de plusieurs méthodes alternatives. Parmi celles-ci, on trouve : NLTK (Bird *et al.*, 2009), SparkNLP (Kocaman & Talby, 2021) , et Spacy (Honnibal *et al.*, 2020) avec ses quatre modèles (*fr_core_news_sm*, *fr_core_news_md*, *fr_core_news_lg*, et *fr_dep_news_trf*).

Une évaluation manuelle des résultats de segmentation de ces méthodes est réalisée sur des extraits du corpus des Cahiers citoyens. À cette fin, ces extraits ont été segmentés manuellement, et les résultats des différentes méthodes ont été évalués en comparaison avec cette segmentation manuelle. Cette évaluation [11] a démontré que le modèle *fr_dep_news_trf* de Spacy, qui est fondé sur les transformeurs, permet d'obtenir la meilleure segmentation en phrases sur le corpus des Cahiers citoyens.

5.2 Pré-traitement et nettoyage des phrases

La segmentation des contributions en phrases a révélé que certaines d'entre elles sont formatées [12] (comme : "*Je vous prie d'agréer, Monsieur, mes sincères salutations*", "*Mardi 18 décembre 2018*"). Ainsi, un pré-traitement s'avère nécessaire pour identifier ces phrases formatées, afin de ne pas les inclure dans le processus de clustering. Cela permettra d'éviter une perte de temps et de ressources mémoire lors du calcul des vecteurs de phrases. De plus, ce pré-traitement permettra de réduire la consommation de mémoire et d'accélérer l'obtention des clusters via l'algorithme de clustering sélectionné.

La méthode préconisée pour le pré-traitement repose sur deux éléments. En premier lieu, il s'agit de mener une recherche syntaxique ciblant certaines structures linguistiques. En outre, une alternative consiste à mettre en œuvre un processus de clustering regroupant ces phrases formatées en fonction de leur similarité sémantique (par exemple, regrouper dans un même cluster toutes les dates et dans un autre cluster les formules de politesse).

11. La description de cette évaluation ne fait pas partie de cet article.

12. Dans cet article, nous appelons une phrase formatée une phrase qui ne porte aucune information sur les sujets et les problématiques abordés par les citoyens dans leurs contributions. Ces phrases comprennent des éléments tels que la date des contributions, les noms des contributeurs, le destinataire, les formules de politesse, etc.

5.3 Calcul des vecteurs de phrases

Afin de mettre à l'épreuve la deuxième hypothèse, qui suggère que le modèle *sentence-camembert-base*, utilisé pour calculer les vecteurs des phrases, ne garantit pas la meilleure représentation des phrases, une comparaison a été effectuée avec un autre modèle à architecture plus large, le *sentence-camembert-large* [13]. Le but de cette comparaison est de vérifier si le modèle avec l'architecture large permet d'obtenir une meilleure représentation des phrases que celui de base.

La comparaison entre les deux modèles repose sur l'évaluation des similarités cosinus [14] retournées par chacun d'eux sur trois catégories de phrases extraites du corpus, dont les sujets sont très abordés dans les Cahiers citoyens, et utilisées dans (Chandora, 2023) pour comparer le modèle de base à d'autres modèles multilingues. Ces trois catégories de phrases sont :

- Des phrases contenant à la fois des formes développées et des sigles d'appellation ("Rétablir l'ISF", "Rétablir l'impôt sur la fortune", "Rétablir l'impôt de solidarité sur la fortune" et "Rétablir l'APL").
- Des paraphrases ("Il faut diminuer le train de vie de l'État", "Il faut réduire le train de vie de l'État", "Il faut revoir à la baisse le train de vie de l'État", "Il faut diminuer les dépenses de l'État", et "Il faut revoir à la baisse les dépenses de l'État").
- Des phrases contradictoires sur la thématique et sur la polarité ("Vive les gilets jaunes", "Honte aux gilets jaunes", "Je soutiens les gilets jaunes", "Je suis contre les gilets jaunes" et "Je suis contre le glyphosate").

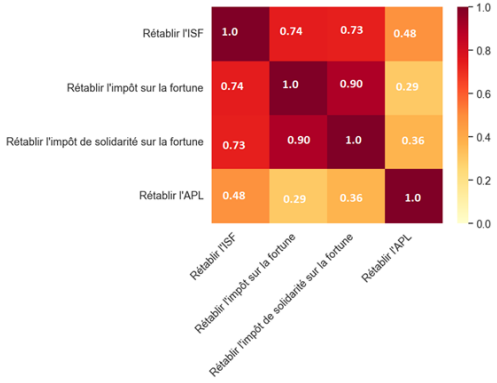
Pour les deux premières catégories de phrases, le modèle fournissant la plus grande similarité cosinus est considéré comme le meilleur, tandis que pour la dernière catégorie, étant donné qu'il s'agit de phrases contradictoires, le modèle renvoyant la plus faible similarité cosinus est considéré comme le meilleur.

La figure 1 illustre les scores de similarité cosinus retournés par les deux modèles entre les phrases de la première catégorie sous forme de matrices de confusion. Ces scores indiquent que le modèle *sentence-camembert-large* est plus efficace pour détecter la similarité entre une phrase contenant une forme développée et une autre contenant des sigles d'appellation, affichant des scores de similarité cosinus plus élevés que le modèle *sentence-camembert-base*. Les matrices de confusion de la figure 2 montrent également les scores de similarité cosinus retournés par les deux modèles mais entre les phrases de la deuxième catégorie. Ces comparaisons ont démontré que le modèle large surpasse le modèle de base dans l'identification de similarités entre des paraphrases, en renvoyant des scores plus élevés. Enfin, la figure 3 affiche les scores de similarité cosinus retournés par les deux modèles entre les phrases de la troisième catégorie. Ces scores sont plus bas pour le modèle large par rapport au modèle de base, suggérant que, même pour ce type de phrases, le modèle large offre de meilleures performances que le modèle de base.

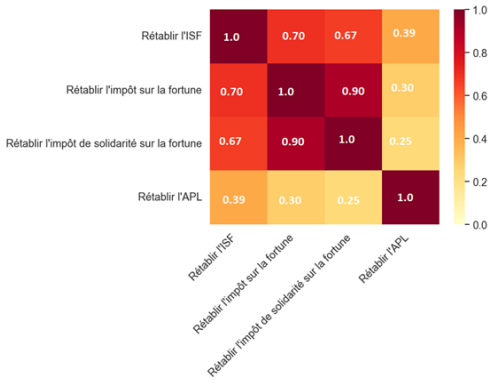
Ainsi, ces comparaisons confirment que le modèle *sentence-camembert-large* est plus adapté pour calculer les vecteurs des phrases du corpus des Cahiers citoyens, assurant une meilleure représentation, et garantissant que les phrases similaires soient plus étroitement projetées dans l'espace vectoriel, tandis que les phrases non similaires se trouvent à une plus grande distance dans cet espace.

13. https ://huggingface.co/dangvantuan/sentence-camembert-large

14. La similarité cosinus est une mesure statistique utilisée pour évaluer la similitude entre deux vecteurs dans un espace multidimensionnel. Dans le contexte de la similarité sémantique entre les phrases, elle mesure l'angle cosinus entre les vecteurs représentant ces phrases. Une similarité plus élevée indique une proximité sémantique accrue entre les phrases.

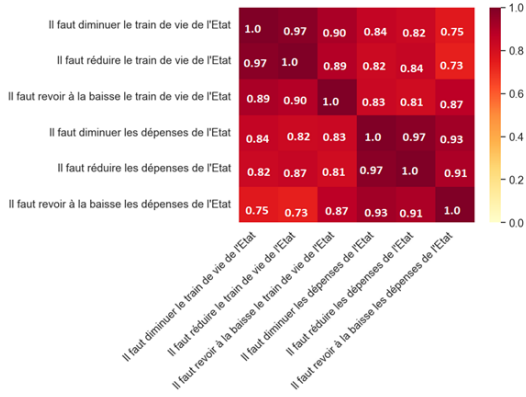


sentence-camembert-large

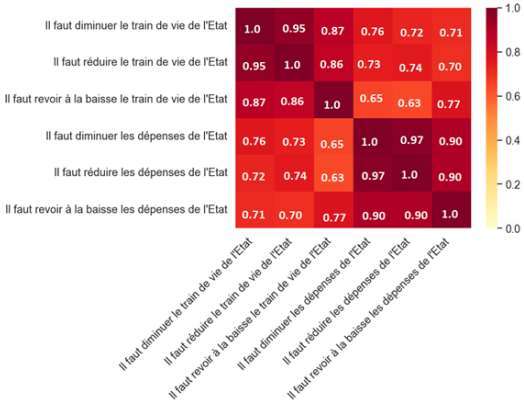


sentence-camembert-base

FIGURE 1 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des phrases contenant à la fois des formes développées et des sigles d'appellation



sentence-camembert-large



sentence-camembert-base

FIGURE 2 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des paraphrases

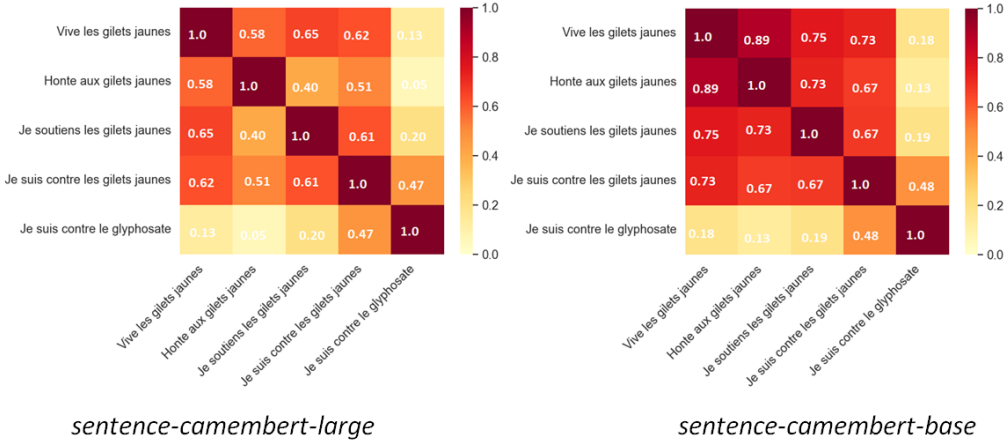


FIGURE 3 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des phrases contradictoires

5.4 Clustering des phrases

La dernière hypothèse, qui permet d'interroger les résultats du clustering, se focalise sur le choix de l'algorithme et la configuration des hyperparamètres. Pour valider cette hypothèse, la méthode proposée s'emploie à explorer d'autres algorithmes de clustering, tels que K-means (Jin & Han, 2010), DBSCAN (Ester *et al.*, 1996), Classification Ascendante Hiérarchique (CAH) (Cecil C. Bridges, 1966), ainsi que des modèles de topic modeling, tels que Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) ou BERTopic (Grootendorst, 2022), avec différentes configurations d'hyperparamètres. Ces algorithmes seront appliqués sur les vecteurs de phrases calculés avec le modèle *sentence-camembert-large*, et celui qui permettra d'obtenir les meilleurs regroupements sera sélectionné [15]. Le critère principal pour évaluer les performances des algorithmes sera le nombre de phrases classées.

6 Conclusion

Dans cet article, une méthode permettant d'effectuer une analyse sémantique des contributions citoyennes du corpus des Cahiers citoyens a été proposée afin d'identifier les différents sujets abordés par les citoyens en vue de les exploiter pour réaliser une étude spatiale du corpus. Les résultats de (Dominguès & Jolivet, 2024) seront également exploités dans cette étude spatiale. La démarche énoncée dans ce papier présente les différentes études et analyses effectuées sur ce corpus, remettant en question certaines méthodes qui n'ont pas donné les résultats souhaités. Elle émet ainsi des hypothèses sur les raisons de ces échecs et propose des solutions alternatives.

La méthode proposée cherche à regrouper les contributions citoyennes en fonction des thématiques

15. Cette étape est en cours de réalisation et le critère du choix de l'algorithme offrant le meilleur regroupement est en cours d'étude.

abordées, en appliquant un clustering sur les phrases du corpus des Cahiers citoyens. Cet article décrit diverses démarches permettant d'obtenir les meilleurs regroupements, tout en proposant une méthode de segmentation des contributions en phrases, de pré-traitement et nettoyage des phrases, de calcul des vecteurs des phrases, et d'utilisation d'algorithmes de clustering.

Étant donné que ce travail est toujours en cours de réalisation, de nouvelles hypothèses et méthodes peuvent être envisagées en fonction des résultats obtenus et de l'évolution de cette recherche, qui, comme indiqué précédemment, s'inscrit dans le cadre d'une thèse.

Remerciements

Je tiens à exprimer ma profonde gratitude envers Catherine Dominguès pour son soutien et ses conseils précieux tout au long de la rédaction de cet article. Son expertise et ses suggestions ont grandement enrichi le contenu, et ses corrections attentives ont contribué à améliorer la clarté et la cohérence du texte. Je suis particulièrement reconnaissant pour ses critiques constructives qui ont permis d'affiner les idées et de renforcer l'argumentation.

De même, je souhaite exprimer mes sincères remerciements à Salomé Chandora pour sa générosité en mettant à disposition le code nécessaire à la génération des matrices de confusion illustrées dans les figures.

Enfin, je tiens à remercier Sabine Ploux pour ses différents conseils et orientations durant les réunions de ma thèse.

Références

BENDINELLI M. (2023). Sens et matérialités d'un cahier citoyen : le cas de la ville de Dole (Jura). *Mots. Les langages du politique*, **131**, 145–169.

BERGER R., BLUENOVE & COGNITO (2019). *Analyse des contributions libres : Cahiers citoyens, courriers et emails, comptes-rendus des réunions d'initiative locale*. Rapport interne.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

CECIL C. BRIDGES J. (1966). Hierarchical cluster analysis. *Psychological Reports*, **18**(3), 851–854. DOI : 10.2466/pr0.1966.18.3.851.

CHANDORA S. (2023). *Fouille sémantique et spatiale dans le corpus Cahiers citoyens : comparaison de méthodes symbolique et numérique*. Mémoire de master, Institut National des Langues et Civilisations Orientales, LASTIG - Univ Gustave Eiffel - ENSG - IGN, France.

DOMINGUÈS C. & JOLIVET L. (2024). Analyse textométrique et spatialisée des cahiers citoyens. In *JADT 2024 : 17th International Conference on Statistical Analysis of Textual Data*.

ESTER M., KRIEGEL H.-P., SANDER J., XU X. *et al.* (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, p. 226–231.

GROOTENDORST M. (2022). Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spacy : Industrial-strength natural language processing in python. *Journal of Open Source Software*, **5**(51), 2456. DOI : 10.5281/zenodo.1212303.

JIN X. & HAN J. (2010). *K-Means Clustering*, In C. SAMMUT & G. I. WEBB, Éds., *Encyclopedia of Machine Learning*, p. 563–564. Springer US : Boston, MA. DOI : 10.1007/978-0-387-30164-8_425.

KOCAMAN V. & TALBY D. (2021). Spark NLP : natural language understanding at scale. *CoRR*, **abs/2101.10848**.

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 model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MONNIER M. (2023). *L'analyse spatiale des Cahiers citoyens appliquée au thème de l'écologie*. Mémoire de master, École des hautes études en sciences sociales.

NILS REIMERS I. G. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *https ://arxiv.org/abs/1908.10084*.

PLOUX S., GENAY M. & PLOUX-CHILLÈS L. (2021). Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes. *Humanités numériques*, (4). DOI : 10.4000/revuehn.2655, HAL : hal-03511103.

RAY M. (2023). *Analyse du corpus Cahiers citoyens*. Mémoire de master, Université Paris Cité.