Ingénieur·e d'études

Informatique, géomatique et humanités numériques
CDD 12 mois

Contexte

Le poste s'inscrit dans le cadre du projet GÉODE[1] (« Encyclopedic GEOgraphical DiscoursE: Writing about Geography in France from the Enlightenment to the Age of Wikipedia », 2020-2024) financé par le LabEx ASLAN dont l'objectif est l'étude des changements majeurs survenus dans les discours géographiques dans les encyclopédies françaises éditées en France entre 1751 (parution du premier tome de l'Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers de Diderot et d'Alembert) et nos jours (Wikipédia, version 2018). Notre démarche interdisciplinaire réunit des spécialistes de linguistique (D. Vigier, ICAR), d'informatique / TAL (L. Moncla, LIRIS), de géographie / géomatique (T. Joliveau, EVS), d'histoire des idées et d'humanités numériques (K. McDonough). En nous appuyant sur les outils et méthodologies de classification semi-supervisée des textes, de génération de modèles de langues et de repérage automatique des routines discursives, nous souhaitons extraire les informations géographiques à partir des textes pour étudier les changements survenus dans les discours géographiques encyclopédiques.

L'ingénieur·e sera intégré·e à l'équipe de recherche DM2L du laboratoire LIRIS qui développe ses recherches autour des méthodes d'apprentissage automatique, de fouille de données, d'analyse de graphes et de traitement automatique du langage naturel. La personne recrutée sera amenée à collaborer avec les différents membres de l'équipe projet GEODE impliquant les laboratoires LIRIS, EVS et ICAR.

Missions

La personne recrutée interviendra sur le volet traitement et analyse de données géographiques. L'objectif de la mission est de développer une méthode de désambiguïsation des toponymes (noms de lieux) combinant des méthodes de TAL et de raisonnement spatial.

La mission s'appuie sur des travaux déjà démarrés par les différents partenaires du projet (McDonough et al., 2019 ; Moncla et al., 2019 ; Vigier et al., 2020). Les données disponibles sont issues d'un traitement automatique d'annotation des entités nommées et des informations géographiques telles que les coordonnées géographiques et les relations spatiales. Le repérage de ces informations est réalisé grâce à la librairie Python Perdido[2] (Moncla & Gaio, 2023). On s'intéresse en particulier aux articles de l'Encyclopédie de Diderot et d'Alembert (1751-1772) décrivant des lieux. Certains articles contiennent des coordonnées géographiques (latitude / longitude) mais qui peuvent être incorrectes ou incomplètes. L'objectif sera de proposer une solution pour corriger et/ou compléter ces coordonnées en utilisant les informations de contexte présentes dans le texte de l'article (i.e., autres noms de lieux, relations spatiales : topologiques, distances, orientations, etc.). Pour les articles qui ne contiennent pas de coordonnées (environ 70% des articles géographiques), l'objectif sera d'utiliser les informations contextuelles extraites du texte pour filtrer et classer les résultats provenant de l'interrogation de gazetiers (geocoding) afin de résoudre le problème d'homonymie. La méthode proposée devra donc s'appuyer sur l'extraction des informations contextuelles par des méthodes de TAL et sur leur interprétation par des méthodes de raisonnement spatial. Les algorithmes proposés pourront être hybrides et implémenter à la fois des approches symboliques (Gaio & Moncla, 2019) et statistiques (Fize et al., 2021).

Liste des tâches :

1. Développer une méthode de désambiguïsation des toponymes vedettes d'articles encyclopédiques (en combinant des approches de TAL et de raisonnement spatial).
2. Proposer un framework d'évaluation et de comparaison avec les approches de l'état de l'art. Cette tâche sera réalisée avec le soutien d'un stagiaire de M2.
3. Intégrer la solution développée au sein de la librairie Python Perdido.

Profil recherché

- Diplôme : Master (ou équivalent) en Informatique
- Compétences :
  - Compétences en informatique : programmation (Python), bases de données
  - Des connaissances en TAL et en géomatique seront appréciées
  - Langues : bon niveau en français requis
- Qualités personnelles : l'ingénieur-e devra faire preuve d'aptitude relationnelles pour le travail en équipe, de qualités de rigueur scientifique, d'autonomie et d'esprit d'initiative

Information pratiques

- Durée du CDD : 12 mois (prolongation possible jusqu'à 16 mois)
- Début du contrat : entre septembre 2023 et janvier 2024
- Salaire : suit la grille ingénieur d'étude du CNRS : entre 1685 € et 1892 € net par mois selon expérience.
- Lieu de travail : LIRIS, Bâtiment Blaise Pascal, Campus La Doua, INSA Lyon, Villeurbanne.
- Affectation : Laboratoire LIRIS UMR CNRS 5205
- Contacts :
  - Ludovic Moncla (ludovic.moncla@insa-lyon.fr)
  - Thierry Joliveau (thierry.joliveau@univ-st-etienne.fr)
  - Denis Vigier (denis.vigier@ens-lyon.fr)
- La candidature doit se faire sur le portail emploi du CNRS https://emploi.cnrs.fr/Offres/CDD/UMR5205-SYLOUD-003/Default.aspx **au plus tard** le **24 mai 2023**. Entretiens de recrutement courant mai / juin pour une **prise de fonction entre septembre 2023 et janvier 2024**.

Références

- Gaio, M., & Moncla, L. (2019), *Geoparsing and geocoding places in a dynamic space context*. In The semantics of Dynamic Space in French: Descriptive, experimental and formal studies on motion expression. In Human Cognitive Processing: Vol. 66 (Michel Aurnague and Dejan Stosic, pp. 354-386). John Benjamins Publishing Company.
- Fize, J., Moncla, L., & Martins, B. (2021). Deep learning for toponym resolution: Geocoding based on pairs of toponyms. ISPRS International Journal of Geo-Information, 10(12), 818.
- McDonough, K., Moncla, L. & Van de Camp, M. (2019), *Named entity recognition goes to old regime France : geographic text analysis for early modern French corpora*. International Journal of Geographical Information Science (IJGIS), 33 (12), 25 pages
- Moncla, L., McDonough, K., Vigier, D., Joliveau T., & Brenon, A. (2019), *Toponym Disambiguation in Historical Documents Using Network Analysis of Qualitative Relationships.* In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, 4 pages, Chicago, IL, USA
- Moncla, L. & Gaio, M. (2023), *Perdido : librairie Python pour le geoparsing et le geocoding de textes en français.* Extraction et Gestion des Connaissances (EGC'2023), Lyon, France.
- Vigier, D., Moncla, L., Brenon, A., McDonough, K., & Joliveau T. (2020), *Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres} (1751-1772).* In: 7th Congrès Mondial de Linguistique Française (CMLF), Montpellier, France

[1] https://geode-project.github.io/
[2] https://github.com/ludovicmoncla/perdido