Modélisation statistique de documents et interfaces humain-machine pour la recherche en sciences sociales

Sujet pour le Tremplin Recherche d'ESIEE Paris, 2024–2025

Laboratoire et équipe

Le projet se déroulera au sein de la plateforme Cortext – l'une des plateformes scientifiques de l'UGE – spécialisée dans l'analyse de données pour les sciences sociales, en particulier l'analyse de documents à travers de réseaux socio-sémantiques, cartographies géo-spatiales et autres approches algorithmiques. Cortext est hébergée par le Laboratoire Interdisciplinaire Sciences Innovations Sociétés (LISIS), une unité mixte de recherche entre CNRS, INRAE et UGE.

Tuteur

Le tuteur du projet sera le Dr. Alexandre Hannud Abdo, Ingénieur de recherche au CNRS et responsable scientifique de la plateforme Cortext.
Adresse électronique : alexandre.hannud-abdo@univ-eiffel.fr

Filières visées

Datascience et intelligence artificielle ; Informatique ; Artificial intelligence and cybersecurity.

Ouvert à des étudiants en E4 (nov-avr) avec possibilité de poursuite par un stage (mai-aoû), ou en E5 (nov-jan) avec un stage de fin d'études dans le laboratoire (fév-jul). Le sujet sera adapté selon le niveau. Possibilité de conduire une partie du stage à l'étranger à discuter.

Présentation générale

Les analyses de gros ensembles de documents et traces numériques assistées par méthodes computationnelles sont une démarche de plus en plus présente dans les recherches en sciences sociales. Davantage qu'un outil, il s'agit d'une nouvelle façon de mobiliser les concepts clés de ses disciplines pour mieux transiter entre les échelles micro et macro qui caractérisent les systèmes sociaux. La plateforme Cortext est l'une des rares infrastructures dédiées à mettre des instruments computationnels et statistiques de pointe à portée de main et au service des compétences et connaissances de chercheurs dont le métier est ancré, loin des méthodes quantitatives, dans des liens profonds avec leurs terrains à travers d'observations, ethnographies, études de cas, et participation dans la vie sociale.

Objectif du projet

Dans la continuité de travaux scientifiques menés au LISIS ces dernières années, le projet pourra se décliner selon l'intérêt de l'étudiant entre :

- L'application de modèles d'apprentissage automatique sur la base du framework de Modèles à Blocs Stochastiques bayésiens (SBMs) [1] pour la classification de graphes de documents et ses composantes hétérogènes (sociales, lexicales, organisationnelles, géographiques, catégoriques) [2]. Un objectif concret serait d'approfondir l'exploration de l'apprentissage par transfert dans ce contexte, sujet de mémoire d'un ancien étudiant[3], afin d'opérationnaliser une collaboration bidirectionnelle entre chercheurs et modèles pour l'apprentissage interactif centré sur l'humain [4].
- Le développement d'interfaces humain-modèles-données adaptées à la lecture, interprétation et interaction avec les représentations produites par ce type de modèle et approche, sous forme de cartes de réseaux hétérogènes (Fig. 1), diagrammes de blocs (Fig. 2), ou graphiques de flux [2]. Il s'agit de mobiliser critères et mesures fondées sur la théorie de l'information pour faire évoluer les interfaces résultantes de nos recherches ou implémenter de nouvelles, en travaillant leur interactivité aussi dans la perspective de faire le lien avec l'inférence de nouveaux modèles [4].

A travers son travail de recherche l'étudiant développera également des compétences dans l'usage de bibliothèques logicielles telles quelles `graph-tool`, pour l'inférence statistique de SBMs [2], et `bokeh`, pour la production d'interfaces interactives, et pourra contribuer à la bibliothèque `sashimi` sur laquelle sont fondés des méthodes d'analyse proposées par la plateforme Cortext.

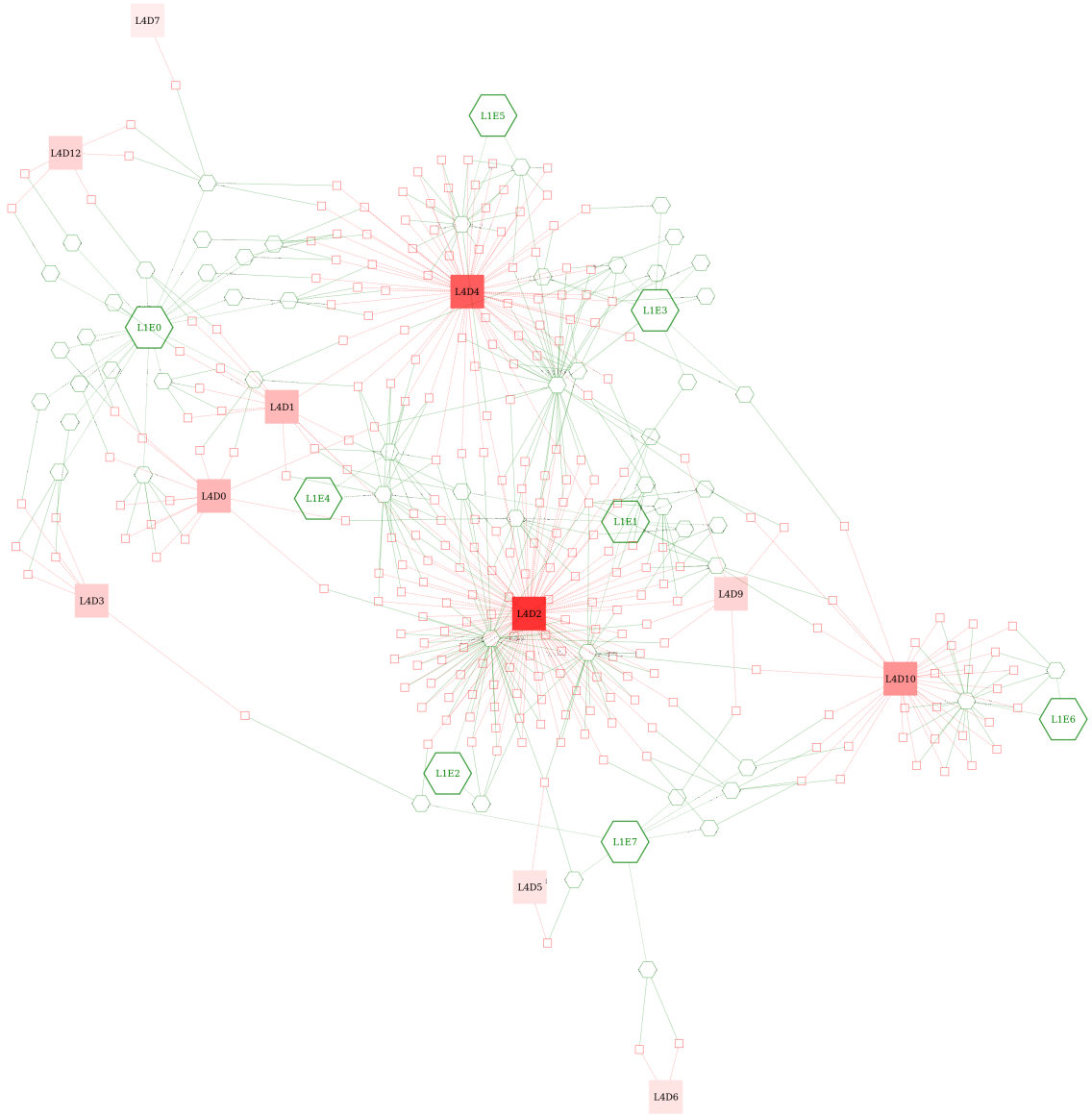

FIGURE 1 – Carte de réseau hétérogène multi-niveau. Sont présentés les acteurs humains et les documents qu'ils co-écrivent (petits nœuds en vert et rouge), et les blocs qui regroupent les éléments similaires de chaque type (grands nœuds). Les étiquettes des nœuds élémentaires ont été effacées pour ce document public.

Bibliographie

[1] T. P. PEIXOTO, « Bayesian stochastic blockmodeling, » in *Advances in Network Clustering and Blockmodeling*. Wiley-Interscience, 2018.

[2] A. HANNUD ABDO, J.-P. COINTET, P. BOURRET et A. CAMBROSIO, « Domain-topic models with chained dimensions : charting an emergent domain of a major oncology conference, » *Journal of the Association for Information Science and Technology*, t. 73, n° 7, p. 992-1011, juill. 2022, ISSN : 2330-1635, 2330-1643. DOI : 10.1002/asi.24606. adresse : https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24606 (visité le 13/09/2024).

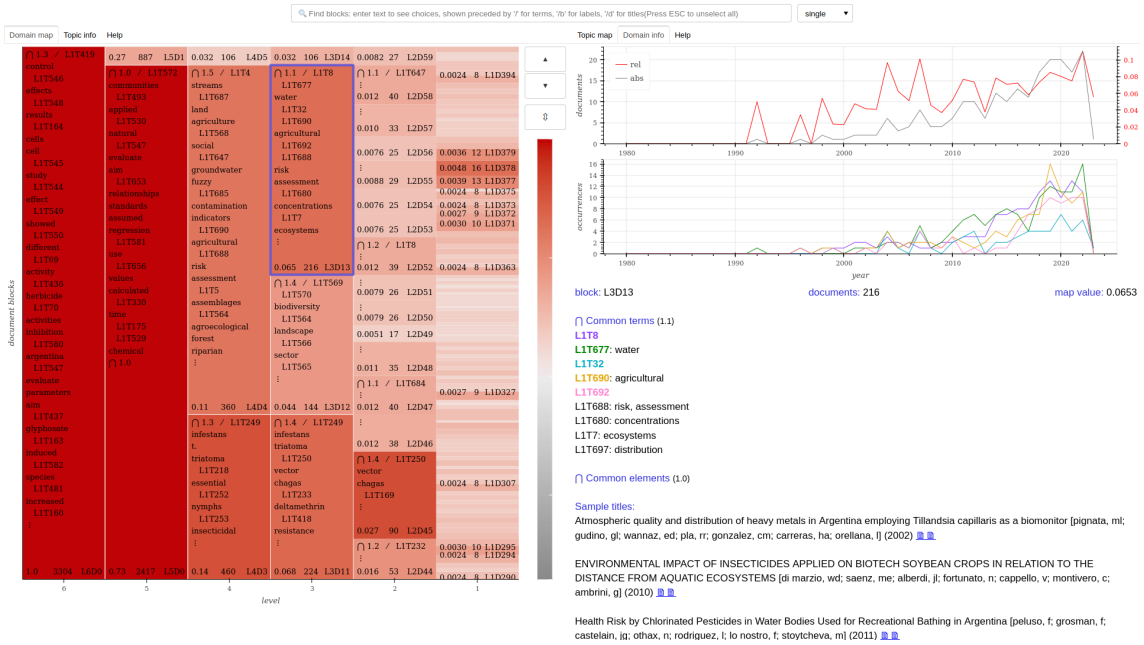

FIGURE 2 – Diagramme de blocs interactif. À gauche la partition multi-niveau d'un ensemble de documents inférée par un modèle à blocs stochastique. A droite l'évolution temporelle du bloc de niveau 3 sélectionné et des topiques qui le caractérisent, puis un échantillon de documents qu'il contient.

[3] L. D. M. ZULUAGA, « Enhancing domain-topic models with transfer learning, » 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC 2024 Dijon), 22-26 jan. 2024.

[4] E. MOSQUEIRA-REY, E. HERNÁNDEZ-PEREIRA, D. ALONSO-RÍOS, J. BOBES-BASCARÁN et Á. FERNÁNDEZ-LEAL, « Human-in-the-loop machine learning : a state of the art, » *Artificial Intelligence Review*, t. 56, n° 4, p. 3005-3054, avr. 2023, Publisher : Springer Science and Business Media LLC, ISSN : 0269-2821, 1573-7462. DOI : 10.1007/s10462-022-10246-w. adresse : https://link.springer.com/10.1007/s10462-022-10246-w (visité le 17/09/2024).