Sujet de stage :

Inférence de systèmes de vote et d'accord de principe, applications au traitement automatique des langues

Antoine Lejay (IECL & équipe-projet Inria PASTA)
Lionel Lenotre (IRIMAS, équipe-projet Inria PASTA & UMR Archimède)

Si le traitement automatique des langues permet de comprendre un discours, d'en répertorier les thématiques ou encore de saisir certaines subtilités telles qu'une ironie latente, elle ne permet pas de produire une analyse de discours conforme aux canons en Sciences Humaines et Sociales (SHS), en particulier en histoire et historiographie. *In extenso*, il ne permet pas de tirer des conclusions et de proposer des interprétations plausibles d'un discours, encore moins de quantifier l'incertitude autour de ces interprétations.

Le projet Apollon (Archimède, CNRS, IECL, Inria, Université de Haute-Alsace, Université de Lorraine, Université de Pavie), réunissant un groupement scientifique pluridisciplinaire composé de probabilistes, statisticiens, historiens de l'antiquité et philologues, cherche à avancer vers une solution de ce problème d'analyse fine de discours. Ce problème est crucial et requiert des avancées en apprentissage statistique. Parmi les défis identifiés, nous trouvons un problème de nature théorique d'inférence de consensus.

Les méthodes d'apprentissage moderne permettent de plonger un texte dans des espace vectoriels de grande dimension sur lesquels des distances peuvent être utilisée pour caractériser les relations entre les mots et établir des relations d'ordre. Notons que des relations d'ordres de nature plus générale peuvent également être utilisée. En pratique, un mot va émettre des préférences hiérarchiques à l'encontre d'un ensemble de mots, établissant une forme de vote pour une liste de candidats. En rassemblant un paquet de mots, par exemple des synonymes, et leurs listes de vote, alors il est nécessaire de se restreindre à un certain nombres de règles précises pour obtenir un ordre global ou un consensus comme le montre le théorème d'impossibilité d'Arrow. Or un consensus est observable puisque l'on peut regarder comment un paquet de synonymes interagit globalement avec un ensemble de mots avec la distance d'un ensemble à un autre. Par suite, d'après le théorème de Gibbard-Satterhwaite, il pourrait exister une règle sensible au vote tactique. Et c'est cette règle que nous souhaiterons inférer puisqu'elle refléterait la pensée profonde de l'auteur du texte.

Le but de ce stage est de préciser les différentes notions de relation d'ordre que l'on peut mettre en place dans des espaces de très grandes dimensions ou les ensembles de points forment des variétés complexes, en particulier la notion de listes de préférences d'un ensemble par rapport à un autre ensemble. Il vise ensuite à établir un état des lieux précis des différents résultats tels les théorèmes d'Arrow ou de Gibbard-Satterhwaite qui sont applicables aux relations d'ordre. Enfin, il s'agit de mettre au point une méthode d'inférence des consensus, compromis, accord, choix tactique, pour lesquels les éléments d'un ensemble ont opté. Ce stage mélange donc statistique en grande dimension, théorie du choix sociale et traitement automatique des langues.

Les candidats attendus sont des étudiants en seconde année de Master en statistique ayant un intérêt pour les problèmes en grande dimension. Des connaissances en théorie des jeux (pour les problèmes de consensus) ou en logique (pour les questions d'ordre) sera un plus.

Lieu. Le stage se déroulera à l'Institut Élie Cartan de Lorraine, site de Nancy. Le stagiare sera rattaché à l'équipe PASTA du centre Inria de l'Université de Lorraine.

Durée. De 4 à 6 mois.

Pour candidater. Nous vous remercions par avance d'envoyer un curriculum vitæainsi qu'un relevé de notes à lionel.lenotre@inria.fr et antoine.lejay@inria.fr.