Traitement automatique de textes par apprentissage automatique : de la détection de position ou de thèmes à la recherche de sensorialité dans des collections de documents

Christine Largeron

Si la science des données s'est d'abord focalisée sur des données de type numérique, elle s'est intéressée ensuite aux données textuelles, au fur à mesure de leur production mais surtout de leur stockage sous forme numérique et, aujourd'hui, les progrès réalisés en matière de traitement automatique de documents figurent parmi les avancées les plus importantes réalisées récemment en IA, notamment grâce aux méthodes de fouille de données et d'apprentissage automatique. Au cours de cette présentation, nous l'illustrerons à travers des tâches variées telles que la détection de position d'un texte ou d'un auteur vis-à-vis d'un sujet donné (stance detection), l'extraction d'information sensorielle ou encore la recherche de thèmes (topic detection) dans des collections de documents. Nous verrons aussi que si l'apprentissage profond, via notamment les grands modèles de langage (LLM), permet de résoudre très efficacement ces tâches, il présente aussi des limites sérieuses comme les biais ou les coûts induits par l'apprentissage des modèles ou encore l'explicabilité et l'interprétabilité des résultats produits.