Modèles de représentation textuels et méthodes d'apprentissage adaptés à l'identification d'auteurs

Christine Largeron*, Jordan Fréry*, Mihaela Juganaru-Mathieu*

* Université Jean Monnet, Saint-Etienne, France
christine.largeron@univ-st-etienne.fr

Qui n'a pas dit un jour, en écoutant un morceau de musique : "Tiens, on dirait untel" ? Sur la base d'un extrait, on peut en effet identifier directement le compositeur ou l'interprète même si on ne connaît pas forcément le morceau. De même, pour des documents textuels, ce problème d'authentification consiste à décider si un texte donné a été écrit par l'auteur d'un autre groupe de documents. Pour le résoudre, la fouille de texte peut s'avérer très utile. Ainsi par exemple, pour authentifier une élégie de Shakespeare, des techniques telles que le comptage exclusif des mots et la prise en compte de mots rares ont été employées avec succès. Mais le domaine de la littérature n'est pas le seul concerné et, l'identification d'auteurs peut aussi être utile dans d'autres applications comme par exemple, dans le domaine juridique pour l'authentification de testament, dans le cadre d'investigations anticriminelles ou antiterroristes pour déterminer la provenance d'une demande de rançon ou de posts émis sur des forums du Darweb ou encore, en marketing pour le profiling des auteurs de blogs ou de commentaires sur le Web. Pour résoudre ce problème il est nécessaire de représenter de façon appropriée les documents afin de pouvoir les comparer. Toutefois, nous pensons qu'il est illusoire de rechercher une empreinte pour l'auteur d'un texte, qui serait unique au même titre qu'une empreinte digitale. Nous pensons qu'il faut utiliser divers espaces de représentation pour les textes à analyser, selon la langue d'origine ou encore le genre ou la qualité du document. Dans cette communication, nous décrirons de tels modèles de représentation et montrerons comment, après avoir formalisé l'identification d'auteur comme un problème de classement, nous pouvons le résoudre en faisant appel à des méthodes de comptage ou d'apprentissage automatique. Nous présentons également des expérimentations réalisées sur des corpus variés (textes littéraires courts ou longs, articles de presse ou publications, blogs) proposés dans le cadre de la compétition PAN-CLEF. Les résultats obtenus confirment l'intérêt de ces approches tant du point de vue de leur performance que des temps de traitement.