CATI : une approche interactive de découverte et de classification de grands corpus de documents

Cédric Boscher, Előd Egyed-Zsigmond, Sylvie Calabretto

Université de Lyon, LIRIS UMR 5205 CNRS
prenom.nom@insa-lyon.fr

Résumé. Dans cet article, nous présentons CATI, une application web interactive d'exploration et de classification de documents. Notre application permet à des utilisateurs non-informaticiens d'explorer et classifier de grandes collections de documents pouvant contenir du texte, des images, et des méta-données telles qu'une date, un auteur, une géolocalisation, etc... CATI fournit un ensemble d'assistants de classification tels qu'un module de détection d'événements, ou encore des méthodes de clustering basées sur des images et du texte. Nous montrons que CATI permet de classifier de grands jeux de données en quelques clics, à l'aide des assistants de classification implémentés et d'assistants permettant à l'utilisateur de sélectionner des attributs méta-données pertinents pour la classification d'un jeu de données.

1 Introduction

Dans un contexte où les réseaux sociaux et la presse en ligne s'imposent comme les principales sources d'informations en ligne permettant d'analyser les tendances de l'opinion publique sur des sujets majeurs d'actualité, les enjeux de méthodes d'extraction d'information appliquées à de grandes collections de documents sont essentiels pour des acteurs de tous types, notamment économiques ou politiques. Diverses approches de fouille de texte combinées à des techniques d'apprentissage automatique permettent de cibler des mots-clés récurrents et des sujets en tendances, lesquels permettent alors de modéliser un ensemble de connaissances et de faciliter la prise de décision d'experts métiers dans un contexte précis. Cependant, ces acteurs n'intègrent pas nécessairement de profils type Data Scientist dans leurs équipes, ou alors ne disposent pas d'employés dédiés à l'annotation massive de données.

Ce papier présente une version améliorée de CATI (Bosetti et al. (2019)), une plate-forme d'assistance à la découverte et à la classification de grandes collections de documents dédiée aux experts métiers et aux utilisateurs non informaticiens. La réalisation d'un tel travail est motivée par l'intérêt potentiel qu'il représente pour différents acteurs économiques ou experts métiers, et répond au besoin d'une solution centrée utilisateur. Elle s'inscrit notamment dans le cadre des projets IDENUM [1] et LIVRONS [2], qui s'intéressent aux représentations visuelles

1. https ://imu.universite-lyon.fr/projet/idenum-identites-numeriques-urbaines/
2. https ://imu.universite-lyon.fr/projet/livrons-livraison-a-velo-representations-sociales-et-donnees-des-reseaux-sociaux-2020s