Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX

Constant Mathieu

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX

Mathieu Constant[1]
(1) Université de Lorraine, CNRS, ATILF, 44 avenue de la Libération, 54000 Nancy, France
Mathieu.Constant@univ-lorraine.fr

RÉSUMÉ

La plateforme ISTEX (https://www.istex.fr) permet d'accéder à une large base d'archives scientifiques comptant plus de 25 millions de documents de tous les grands domaines scientifiques. Les documents incluent non seulement les métadonnées mais aussi le texte plein, et ont été prétraités de manière homogène pour faciliter leur traitement automatique. Dans cet exposé, nous présenterons une initiative pour favoriser les travaux de recherche en TAL et fouille de textes autour de ces données. En particulier, nous présenterons les travaux en cours pour la construction d'un jeu de données permettant d'apprendre et d'évaluer des modèles pour différentes tâches comme l'extraction de mots-clés, l'identification du domaine scientifique ou la génération de résumés. Nous avons circonscrit notre ensemble de travail aux publications d'ISTEX dont il existe une version open-access, grâce à la collaboration de l'INIST, avec l'objectif de constituer un jeu de données ouvert. Ce filtrage a permis d'identifier un ensemble multilingue de 3 millions de documents environ, dont une très large majorité en anglais mais couvrant une dizaine de langues différentes. Dans cet exposé, nous décrirons en particulier les traitements réalisés à l'ATILF pour éliminer les documents non pertinents ou bruités (ex. mauvaise qualité d'OCR), et ainsi ne garder que les documents de qualité.

ABSTRACT

Construction of a dataset of scientific publications for NLP and text mining from ISTEX.

The ISTEX platform (https://www.istex.fr) provides access to a large database of scientific archives with more than 25 million documents from all major scientific fields. The documents include not only the metadata but also the plain text, and have been pre-processed in a homogeneous way to facilitate their automatic processing. In this talk, we will present an initiative to promote research work in NLP and text mining around these data. In particular, we will present the work in progress for the construction of a dataset to learn and evaluate models for different tasks like keyword extraction, scientific domain identification, abstract generation. We have limited our working set to ISTEX publications having an open-access version, thanks to the collaboration of INIST, with the objective of constituting an open dataset. This filtering made it possible to identify a multilingual set of approximately 3 million documents, the vast majority of which are in English but covering around ten languages. In this talk, we will describe in particular the processing carried out at ATILF to eliminate irrelevant or noisy documents (e.g. poor OCR quality), and thus keep only documents of good quality.

MOTS-CLÉS : Publications scientifiques, construction de jeu données, traitement automatique des langues, fouille de textes.

KEYWORDS: Scientific publications, dataset construction, natural language processing, text mining.