Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Toolbox : une chaîne de traitement de corpus pour les humanités numériques

Johanna Cordova[1,2] Yoann Dupont[1] Ljudmila Petkovic[1] James Gawley[1]
Motasem Alrahabi[1] Glenn Roe[1,3]
(1) ObTIC - Sorbonne Université, 4 place Jussieu, 75005 Paris, France
(2) ERTIM / CERLOM, INALCO, 2 rue de Lille, 75007 Paris, France
(3) CELLF – UMR 8599 (CNRS – Sorbonne Université), 1 rue Victor Cousin, 75230 Paris Cedex 05, France
prenom.nom@sorbonne-universite.fr

RÉSUMÉ
Le projet Toolbox propose une chaîne de traitement pour la manipulation et le traitement de corpus textuels incluant la numérisation (OCR/HTR), la conversion au format TEI, la fouille de texte (reconnaissance d'entités nommées) et la visualisation de données. Les fonctionnalités sont accessibles via une interface en ligne qui sert de surcouche graphique à des scripts développés par nos soins ou utilisant des outils externes. Elles permettent d'automatiser les tâches élémentaires de traitement de corpus pour les chercheurs en humanités numériques. Cet outil est ouvert aux contributions externes.

ABSTRACT
Toolbox : a corpus processing pipeline for digital humanities.

The Toolbox project aims to develop a pipeline for the manipulation, processing, and analysis of textual corpora that includes digitisation (OCR/HTR), conversion to TEI-XML format, text mining (named entity recognition) and data visualisation. These functionalities are accessible via an online interface that serves as a graphical overlay to programming scripts we have previously developed or based on external tools. The Toolbox thus allows for the automation of several basic corpus processing tasks for researchers in the digital humanities. This tool is open to external contributions.

MOTS-CLÉS : Humanités numériques, TEI, OCR, reconnaissance des entités nommées.

KEYWORDS: Digital Humanities, TEI, OCR, named entity recognition.

1 Motivations

Pour le traitement de leurs corpus textuels, les chercheurs en sciences humaines ont besoin d'un certain nombre d'outils pour des tâches classiques telles que la numérisation (OCR/HTR), la conversion en un format compatible avec l'édition numérique (XML-TEI), la fouille de texte et la visualisation des données. Si de nombreux outils sont proposés pour exécuter certaines de ces tâches individuellement, il n'existe actuellement pas de chaîne de traitement en accès libre qui les centralise. Ainsi, mettre en œuvre chacune de ces tâches peut s'avérer chronophage et nécessiter la collaboration de chercheurs de plusieurs domaines d'expertise. Nous proposons donc une solution qui réponde au double besoin de centraliser et d'automatiser les traitements de corpus textuels, et d'en permettre l'usage à des chercheurs non experts en informatique.

2 Présentation de la plateforme

Nous proposons une interface Web basée sur du code HTML et Javascript, et enrichie de l'application Flask [1]. L'interface propose un ensemble de fonctionnalités pour le traitement de fichiers de l'utilisateur. Ces fonctionnalités se basent sur des scripts Python disponibles indépendamment et essentiellement développés par nos soins. La plateforme comporte pour le moment trois volets de traitements de données textuelles, détaillés ci-dessous. Cette liste est vouée à être progressivement enrichie selon les besoins. Pour chaque tâche, l'utilisateur doit seulement téléverser le(s) document(s) à traiter, préciser quelques paramètres au besoin, et récupérer le fichier de sortie généré.

Acquisition de corpus Le volet acquisition de corpus comporte les fonctionnalités ci-dessous :

- ***Scraping* de corpus** à partir de Wikisource [2]. L'utilisateur peut choisir de créer un corpus aléatoire à partir de textes Wikisource, ou de préciser des URLs à moissonner. Il est possible de régler la taille du corpus à collecter ou le pourcentage à extraire à partir d'un texte donné.
- **Numérisation et transcription** de documents (OCR/HTR) via des instances Kraken [3] ou Tesseract (Smith, 2007) selon les modèles utilisés.
- **Correction d'erreurs** des sorties de reconnaissance automatique de caractères.
- **Conversion de fichiers** (.txt ou .odt) vers un format XML-TEI (Pytlik Zillig, 2009) ; formatage de fichiers pour le calcul d'accord inter-annotateurs.

Fouille de texte Actuellement, pour le volet de la fouille de texte, la plateforme propose la reconnaissance des entités nommées. L'indexation de corpus via une interface de fouille textuelle est également proposée. Le premier module permet d'annoter en entités nommées des fichiers au format XML-TEI. Il est possible d'annoter avec SpaCy (Honnibal & Montani, 2017) ou Flair (Akbik *et al.*, 2018) en sélectionnant le modèle de son choix. La deuxième possibilité permet d'indexer un corpus textuel sur l'interface Obvie [4] à partir de fichiers respectant le format d'entrée.

Visualisation Ce volet renvoie vers des outils développés en interne : 1) Tanagra, une interface web pour l'identification, la géolocalisation et la cartographie des noms de lieux dans les textes ; 2) Minerva, qui permet de visualiser le lexique et les sentiments autour des entités nommées. Si ces dernières sont désambiguïsées via Wikidata, il est possible de les regrouper selon une propriété de la base (par exemple le genre pour les personnes).

Scripts et documentation Les scripts ayant servi au développement de l'application Toolbox sont disponibles sur Github [5] et exécutables individuellement en ligne de commandes. À terme, ces scripts seront remaniés pour être déployés en librairie Python. Les outils pourront ainsi être manipulés avec davantage de paramètres par les utilisateurs les plus experts. Le Github contient également de la documentation pour chaque tâche couverte par la chaîne de traitement, avec un état de l'art des outils disponibles pour celles-ci. La Toolbox a donc vocation a être un guide global pour de nombreuses tâches d'intérêt dans les humanités numériques.

1. https://flask.palletsprojects.com/en/2.1.x/
2. https://fr.wikisource.org/wiki/Wikisource:Accueil
3. https://kraken.re/master/index.html
4. https://obvil.huma-num.fr/obvie/
5. https://github.com/obtic-scai/Toolbox

Références

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018*, p. 1638–1649.

HONNIBAL M. & MONTANI I. (2017). spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, **7**(1), 411–420.

PYTLIK ZILLIG B. L. (2009). TEI analytics : converting documents into a TEI format for cross-collection text analysis. *Literary and Linguistic Computing*, **24**(2), 187–192.

SMITH R. (2007). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, p. 629–633 : IEEE.