Perdido : librairie Python pour le geoparsing et le geocoding de textes en français

Ludovic Moncla*, Mauro Gaio**

* Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, F-69621
ludovic.moncla@insa-lyon.fr
https://ludovicmoncla.github.io
** Université de Pau et des Pays de l'Adour, LMAP, UMR 5142, Pau, France
mauro.gaio@univ-pau.fr

Résumé. Cet article présente la librairie Python Perdido pour le geoparsing et le geocoding de textes en français. Nous présentons l'architecture générale de l'outil Perdido composée de trois couches : back-office, API et librairie Python. Nous détaillons les méthodes utilisées pour le développement de la chaîne de traitement et des différentes tâches (reconnaissance et classification des entités nommées et résolution des toponymes). Enfin, nous présentons les différentes fonctionnalités de la librairie Python et la façon de l'utiliser. La librairie est développée comme une surcouche faisant appel aux services de l'API et permet de manipuler, visualiser et exporter les résultats du geoparsing et du geocoding. Un notebook [1] Jupyter décrit, sous la forme d'un tutoriel, l'ensemble des fonctionnalités implémentées dans la librairie.

1 Introduction

Cet article présente la librairie Python Perdido [2] pour le geoparsing et le geocoding de textes en français. Le geoparsing est une tâche très importante en recherche d'information géographique (Jones et Purves, 2008) et plus largement en Traitement Automatique des Langues (TAL). Elle se décompose en deux sous-tâches : (1) la reconnaissance et la classification d'entités nommées et d'informations spatiales (ou *geotagging*) et (2) la résolution de toponymes (ou *geocoding*). De nombreuses définitions de la notion d'entités nommées existent, mais de manière assez générale nous pouvons définir la tâche de reconnaissance d'entités nommées comme l'action de repérer et de catégoriser dans un texte les mots ou groupes de mots (le plus souvent des noms propres ou descriptions définies), permettant d'identifier un objet de manière stable et non ambiguë (Nouvel et al., 2015). Dans le cas du geoparsing, nous nous intéressons plus spécifiquement au repérage d'informations spatiales (ou géographiques) c'est-à-dire d'éléments du texte faisant référence à un lieu, une localisation (absolue ou relative) ou encore un déplacement. On parle alors de geotagging. En complément, le geoparsing comprend également l'étape de résolution des entités nommées (ou *entity linking*), qui dans ce cas peut

1. https://github.com/ludovicmoncla/demo-perdido-egc-2023
2. https://github.com/ludovicmoncla/perdido