Extração de Citação de Localidades em Textos Técnico-Científicos em Língua Portuguesa[1]

Clara Mattos Medeiros[2,4] e Maria Fernanda Moura[3]

[1]Trabalho realizado com apoio financeiro da Embrapa, CNPq e Faped. [2]Graduanda em Engenharia da Computação, UNICAMP, Campinas, SP [3]Pesquisadora da Embrapa Agricultura Digital, Campinas, SP. [4]mattosclara22@gmail.com

Resumo - A proposta deste trabalho é selecionar e utilizar um reconhecedor de entidades nomeadas (REN) para a língua portuguesa a fim de extrair metadados de citações a localizações geográficas brasileiras em publicações técnico-científicas, especialmente as do domínio agropecuário no subtema pastagens. Primeiramente identificaram-se algumas ferramentas REN de domínio público que suportem a língua portuguesa; foram experimentadas: a TopExtract, da Embrapa; Apache OpenNLP; FreeLing; Stanza, da Stanford; NLTK – Natural Language Toolkit; e a biblioteca SpaCy. Devido a alguns problemas de manutenção desses softwares, especialmente no que toca aos treinamentos para a língua portuguesa, optou-se pela SpaCy. A SpaCy, uma biblioteca de software de código aberto para processamento avançado de linguagem natural, escrita em Python, mostrou-se bastante adequada para localizar entidades nomeadas em textos na língua portuguesa, uma vez que possui um corpus bem anotado e, principalmente, é uma biblioteca que tem recebido constante evolução. Além disso permite que sejam realizados novos treinamentos mediante anotação de corpus. Assim, para casos nos quais a SpaCy cometia alguns erros de classificação, tais como classificar cultivares como localidades, por exemplo “BRS Tarumaxi”, e outros erros similares, a estratégia adotada consistiu na identificação das regras de citação a localidades de interesse e a consequente personalização do reconhecedor de entidades nomeadas da biblioteca SpaCy, isto é, selecionar e anotar um novo córpus para treinamento do REN da SpaCy. Os experimentos conduzidos, com os novos treinamentos da SpaCy, mostram uma revocação média de 0,92 e uma precisão média de 0,95, permitindo aceitar que a identificação das localidades nos textos seja bastante confiável.

Termos para indexação: reconhecimento de entidades nomeadas; processamento de linguagem natural; SpaCy; geolocalização; mineração de textos

Citation from Localities Extraction in Technical-Scientific Texts in the Portuguese Language

Abstract - The purpose of this work is to select and use a named entity recognizer (NER) for the Portuguese language in order to extract metadata from citations to Brazilian geographic locations in technical-scientific publications, especially those in the agricultural domain in the sub-theme pastures. First, some public domain NER tools that support the Portuguese language were identified; tried out: TopExtract, from Embrapa; Apache OpenNLP; FreeLing; Stanford’s Stanza; NLTK – Natural Language Toolkit; and the SpaCy library. Due to some maintenance problems of these software, especially with regard to training for the Portuguese language, SpaCy was chosen. SpaCy, an open source software library for advanced processing of natural language, written in Python, proved to be quite suitable for locating named entities in texts in Portuguese, since it has a well-annotated córpus and, mainly, is a library that has been constantly evolving. In addition, it allows new training to be carried out by means of body annotation. Thus, for cases in which SpaCy made some classification errors, such as classifying cultivars as localities, for example “BRS Tarumaxi”, and other similar errors, the strategy adopted consisted of identifying the citation patterns for localities of interest and the consequent customizing the named entity recognizer from the SpaCy library, that is, selecting and annotating a new córpus for training the SpaCy NER. The experiments carried out with the new SpaCy trainings show an average recall of 0.92 and an average accuracy of 0.95, allowing us to accept that the identification of localities in the texts is quite reliable.

Index terms: recognition of named entities, natural language processing, SpaCy, geolocation, text mining.