spaCy

Andorfer, Peter; peter.andorfer@oeaw.ac.at / Schlögl, Matthias; matthias.schloegl@oeaw.ac.at

SpaCy ist eine in *Python (Cython)* geschriebene Programmbibliothek für natürliche Sprachverarbeitung (NLP). Im Gegensatz zu dem ebenfalls in *Python* implementierten NLP-Framework NLTK, das auf Forschung und Lehre fokussiert, will *spaCy* Lösungen für die Industrie bereitstellen. Dafür wird eine saubere und einfach zu verwendende API bereitgestellt, die sich auf die performante Erledigung von Standard-NLP-Aufgaben konzentriert. *SpaCy* stellt momentan Sprachmodelle für zehn Sprachen in verschiedenen Ausbaustufen zur Verfügung und erlaubt – je nach Modell – *tokenizing*, *sentence splitting*, *tagging*, *parsing*, *named entity recognition* und *word similarity calculations*. Zudem erlaubt *spaCy* die relativ einfache Erweiterung der Kernfunktionen. So können etwa in der *spaCy*-Pipeline auch externe Komponenten aufgerufen oder in der Tokenklasse *custom attributes* registriert werden (eine Anwendungsmöglichkeit, die z. B. *acdh-spacytei* nutzt).

Ein Großteil der von *spaCy* zur Verfügung gestellten Modelle, Klassen und Funktionen basiert auf *Deep Learning*-Technologien. Neben der *Python*-Bibliothek stellt *spaCy* auch Shell-Skripte zur Verfügung, mit deren Hilfe neue Modelle auf Basis eigener Trainingsdaten erstellt sowie bestehende Modelle weiter trainiert werden können.

Software:

spacy , Natural Language Toolkit (nltk), acdh-spacytei

Verweise:

NLP, Named Entity Recognition / NER, Part-of-Speech-Tagging, acdh-spacytei, Tagger

Themen:

Natural Language Processing, Software und Softwareentwicklung

Zitiervorschlag:

Andorfer, Peter; Schlögl, Matthias. 2021. spaCy. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.170