acdh-spacytei

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

acdh-spacytei ist ein *Python*-Package, das eine Reihe von Hilfsklassen und -funktionen zur Prozessierung von XML/TEI-kodierten Texten mit *spaCy* zur Verfügung stellt. Wie die meisten NLP-Tools ist auch *spaCy* dahingehend konzipiert, *plain text*, also nicht annotierte Volltexte, zu verarbeiten. Um also beispielsweise *Named-Entity-Recognition* (NER) in einer bereits in XML/TEI-kodierten Edition durchführen zu können, müssten die entsprechenden Textknoten extrahiert, in *plain text* konvertiert, mit dem jeweiligen Tool prozessiert und die generierten Annotationen separat vom XML/TEI-Dokument gespeichert werden, was zu einer gewissen Datenduplizierung führen würde.

acdh-spacytei hingegen prozessiert XML/TEI-kodierte Texte annotationsbewahrend. Dafür werden die XML/TEI-kodierten Dokumente tokenisiert, die Tokens in <w>-Tags geschrieben und mit IDs versehen. Die Tokenisierung wird dabei jedoch nicht von *acdh-spacytei* direkt vorgenommen, *acdh-spacytei* ruft dafür vielmehr den Webservice *xTokenizer* auf. Aus der Sequenz der Tokens wird daraufhin ein *spaCy*-Doc-Element erstellt und jedem Token-Element die <w>-Tag-IDs hinzugefügt. Dieses Doc-Element wird dann mit *spaCy* prozessiert. Die von *spaCy* vorgenommenen Anreicherungen (z. B. Lemmatisierung, POS-Tag, NER) werden als TEI-Attribute in die jeweiligen <w>-Tags zurückgeschrieben bzw. *named entities* von typisierten <rs>-Tags umfasst. Dafür nutzt *acdh-spacytei* die in *custom attributes* der *spaCy*-Token-Klasse abgelegten <w>-Tag-IDs.

Abgesehen von dem oben beschriebenen Interface zur annotationsbewahrenden Anreicherung stellt *acdh-spacytei* auch Methoden zur Verfügung, um aus bereits semantisch annotierten XML/TEI-Dokumenten Daten für das Training von NER-Modellen generieren zu können.

Software:

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

Verweise:

NLP, Named Entity Recognition / NER, Part-of-Speech-Tagging, spacy, xTokenizer, ACDH-CH

Themen:

Natural Language Processing, Annotation und Modellierung, Software und Softwareentwicklung

Zitiervorschlag:

Andorfer, Peter; Schlögl, Matthias. 2021. acdh-spacytei. 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.2