Skriptsprachen
Naumann
SS 2014

Aufgabenblatt (5)

Die einfachste Möglichkeit Wortformen auf eine normierte Grundform zurückzuführen (Lemmatisierung) besteht darin, die Informationen zu nutzen, die in einem Vollformlexikon vorhanden sind. Wir verwenden im folgenden zu diesem Zweck CELEX. Die Wortformen und die ihnen zugeordneten Lemmata sind in zwei verschiedenen Dateien enthalten. Die Bezüge zwischen einer Wortform und ihrem Lemma wird durch Indices hergestellt.

Aufgabe (1) [3 Punkte]

Definieren Sie die Klasse **Lemma** inklusive einer ***initialize***- und einer ***to_s***-Methode. Zum Speichern der Wortformen bzw. Lemmata werden zwei Arrays verwendet:

Beispiel

test_lemmatisator = Lemma.new([. . . , ['Maus', 32, 'n.sg'],['Mäuse', 32, 'n.pl'],['Mensch', 33, 'n.sg'],. . .], [. . . , 'Maus', 'Mensch', . . .])
puts test_lemmatisator
Der Lemmatisator enthält 45 Wortform- und 29 Lemma-Einträge.

In dem Beispiel verweist die 32 bzw. 33 auf das 32-te bzw. 33-te Element der Lemma-Arrays.

Aufgabe (2) [3 Punkte]

Definieren Sie für die Klasse **Lemma** die Methode **lemmatisiereString**, die einen Satz (String) als Argument nimmt und - soweit es möglich ist - alle darin enthaltenen Wörter lemmatisiert.

Beispiel

a = Lemmatisator.new
a.lemmatisiereString("nur ein kleiner Test kkük")

nur	: nur	X
ein	: ein	0
kleiner	: klein	o6,c0
Test	: Test	nS,dS,aS
kkük	: —	—

Aufgabe (3) [4 Punkte]

Modifizieren Sie die Methode **transcode** des Textfilters (s. Aufgabenblatt (3)) und die Methode **identily_language** des Sprachidentifikators (s. Aufgabenblatt (4)) so, dass Sie in Dateien gespeicherte Texte filtern bzw. identifizieren können.