Skriptsprachen
Naumann
WS 2015

Aufgabenblatt (6)

Der große Umbau beginnt...

Aufgabe (1) [3 Punkte]

Überführen Sie die Klassen, die wir für den Tokenizer und den Lemmatisator entwickelt hatten in zwei Module. Welche Änderungen sind erforderlich? Worauf ist zu achten? Welche Methoden werden überflüssig?

Aufgabe (2) [3 Punkte]

Erweitern Sie das Modul Lemma durch eine Methode **generiere_lexikon**, die die Celex-Dateien einliest und eine Datei erzeugt, die für jede Celex-Wortform eine Zeile der Form:

$$Wortforn \backslash Wortart \backslash Lemma \backslash Merkmale$$

enthält. Beachten Sie, dass es eine neue Datei (gsl) gibt, die die Wortartinformation für jedes Lemma enthält. Die Wortarten werden durch Zahlen zwischen 1 und 10 kodiert. Verwenden Sie bitte folgende symbolische Bezeichner: 1 = 'N', 2 = 'ADJ', 3 = 'Q/N', 4 = 'V', 5 = 'ART', 6 = 'PRO', 7 = 'ADV', 8 = 'PRP', 9 = 'KON', 10 = 'ITJ'.

Aufgabe (3) [4 Punkte]

Erweitern Sie das Modul außerdem um die Methode **lemmatisiere_tokenliste**, die eine Tokenliste und optional eine Zahl (0 bzw. 1) als Argument nimmt und als Wert eine modifizierte Tokenliste liefert, in der

(a) alle im Lexikon enthaltenen Wortformen durch das zugeordnete *Lemma* (2. Argument: 0) ersetzt werden;

(b) alle im Lexikon enthaltenen Wortformen durch *Lemma/ Wortart* (2. Argument: 1) ersetzt werden bzw.

(c) jedes Token, das nur aus einem Zeichen besteht und jede im Lexikon enthaltene Wortform, die eine der folgenden Kategorien angehört, durch die Zeichenkette '***' ersetzt wird (2. Argument: nicht spezifiziert): 'KON','ITJ','ART','PRP','PRO','Q/N'.