Skriptsprachen
Naumann
WS 2016

Aufgabenblatt (7)

Für dieses und das folgende Aufgabenblatt benötigten Sie das Nokogiri-GEM.

Aufgabe (1) [5 Punkte]

Definieren Sie die Methode **extrahiere_lexikon**, die aus der über StudIP zur Verfügung gestellten Datei korpus.xml ein Frequenzwörterbuch erstellt. Verwenden Sie als Datenstruktur einen Hash, der die Wortformen als Schlüssel verwendet und für jede Wortform einen Hash mit Wortart-Frequenzangaben speichert:

Beispiel

puts extrahiere_lexikon "korpus.xml"
{"Veruntreute"=>{"VVFIN"=>1},"die"=>{"ART"=>22, "PRELS"=>5}, "AWO"=>{"NN"=>8}, ...}

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

Aufgabe (2) [8 Punkte]

Definieren Sie die Methode **extrahiere_bigramme**, die aus der über StudIP zur Verfügung gestellten Datei korpus.xml satzweise alle Bigramme extrahiert. Markieren Sie Anfang und Ende eines Satzes durch ein implizites <s>- bzw. </s>-Tag.

Verwenden Sie auch in diesem Fall einen Hash: Als Schlüssel werden in diesem Fall POS-Tags (1. Element des Bigramms) verwendet. Jedem Schlüssel wird als Wert wieder ein Hash mit POS-Tag (2. Element des Bigramms) Frequenzangaben zugeordnet

Beispiel

puts extrahiere_bigramme "korpus.xml"
{"ART"=>{"NN"=>66, "ADJA"=>20, "NE"=>3, "$("=>2, "CARD"=>2, "ADV"=>1}, ...}