Skriptsprachen
Naumann
WS 2014

Aufgabenblatt (6)

Aufgabe (1) [3 Punkte]

Definieren Sie die Klasse **Tagger** inklusive einer ***to_s***-Methode. Eine spezielle ***initialize***-Methode ist nicht notwendig, da wir zum Initialisieren der Instanzenvariablen spezifische Methoden definieren werden. Es werden vier Instanzenvariablen verwendet:

- @tags
  im Trainingskorpus verwendetes Tag-Set;
- @tagSequenzen
  die Tagfolgen für alle Sätze des Trainingskorpus;
- @bigramme
  Bigramm-Tabelle auf Grundlage des Trainingskorpus und
- @trigramme
  Trigramm-Tabelle auf Grundlage des Trainingskorpus.

Aufgabe (2) [3 Punkte]

Definieren Sie für die Klasse **Tagger** die Methode **extrahiereTagSet**, die einen Dateinamen (String) als Argument nimmt und aus der spezifizierten (XML-)Datei die Namen aller verwendeten Tags extrahiert und in der Instanzenvariable @tags (Array ohne Duplikate) speichert.

Aufgabe (3) [4 Punkte]

Definieren Sie für die Klasse **Tagger** die Methode **extrahiereTagSequenzen**, die einen Dateinamen (String) als Argument nimmt und aus der spezifizierten (XML-)Datei für jeden Satz die Folge der verwendeten POS-Tags extrahiert und in der Instanzenvariable @tagSequenzen (Array von Arrays) speichert.

Hinweis: Sie können voraussetzen, dass die Korpusdateien die in der Beispieldatei verwendete Struktur haben.