Skriptsprachen
Naumann
WS 2014

Aufgabenblatt (7)

Für den Tagger benötigen wir ein Lexikon, das wir aus einer Korpusdatei extrahieren können. Wir verwenden zu diesem Zweck das TuebaDZ-Korpus.

Aufgabe (1) [5 Punkte]

Definieren Sie für die Klasse **Tagger** die Methode ***extrahiereTuebadzSaetze***, die zwei Datei- bzw. Pfadnamen als Argumente nimmt. Das erste Argument bezeichnet die Korpusdatei, aus der die relevanten Informationen gelesen werden; das zweite Argument die Datei, in der sie gespeichert werden sollen.

Relevant sind alle Zeilen, die zwischen einer Zeile, die mit #BOS beginnt und einer Zeile, die mit #EOS beginnt, stehen und nicht selbst mit dem Zeichen '#' beginnen.

Aufgabe (2) [5 Punkte]

Definieren Sie für die Klasse **Tagger** die Methode **generiereFrequenzlexikon**, die aus einer Datei, die mit Hilfe der Methode ***extrahiereTuebadzSaetze*** generiert wurde, ein Frequenzlexikon generiert, in dem für ein Wort X

- jede morphosyntaktische Lesart (POS-Tag) samt
- Merkmalen und
- Frequenz

gespeichert wird.