

Aufgabe 5

Training eines LC-CRF-Wortart-Taggers

Für diese Aufgabe haben Sie zwei Wochen Zeit.

Laden Sie Trainingsdaten für das Wortart-Tagging von `www.cis.lmu.de/~schmid/lehre/Experimente/data/Tiger.zip` herunter und dekomprimieren Sie sie. Jede Zeile enthält ein Wort und ein Wortart-Tag. Auf jeden Satz folgt eine Leerzeile. Die Daten sind in Trainingsdaten, Testdaten und Development-Daten aufgeteilt.

Trainieren Sie auf den Daten einen LC-CRF-Tagger.

Vorüberlegungen

- Welche Merkmale verwenden Sie am besten?
- Welche Teilaufgaben umfasst das Training?
- Welche Datenstrukturen verwenden Sie?
- Was speichern Sie in der Parameterdatei?
- Wie vermeiden Sie Underflow?
- Wie können Sie das Training beschleunigen?

Schicken Sie das fertige Programm an `schmid@cis.lmu.de`.