

Aufgabe 3

Sentimentanalyse mit der Naïve-Bayes-Methode

Laden Sie das *polarity dataset v2.0* von der Adresse www.cs.cornell.edu/people/pabo/movie-review-data herunter und packen Sie es aus. Das Korpus enthält tokenisierte Filmrezensionen, die in positive und negative Rezensionen unterteilt sind. Verschieben Sie die ersten 50 positiven und negativen Reviews in separate Verzeichnisse für Evaluierungszwecke (Testdaten). Die nächsten 50 positiven und negativen Reviews verschieben Sie in separate Verzeichnisse für Optimierungszwecke (Developmentdaten). Die übrigen Daten dienen als Trainingsdaten. (Die Developmentdaten werden Sie erst bei der nächsten Aufgabe benötigen.)

Implementieren Sie dann einen Naïve-Bayes-Klassifikator für die Sentimentanalyse, den Sie auf diesen Daten trainieren. Der Klassifikator besteht aus einem Trainingsprogramm, welches die Modell-Parameter in einer Datei speichert, und einem Anwendungsprogramm, welches die Parameterdatei einliest und die Testdateien (bzw. Developmentdateien) annotiert.

Vorüberlegungen

- Welche Teilaufgaben umfasst das Training?
- Was speichern Sie in der Parameterdatei?
- Welche Teilaufgaben umfasst das Anwendungsprogramm?

Schicken Sie das fertige Programm, die Liste der ausgegebenen Klassen und die erzielte Genauigkeit an schmid@cis.lmu.de.