Sprachtechnologie

2. Übung

Prof. Tobias Scheffer
Paul Prasse
Michael Großhans
Sebastian Arzt
Philipp Schmidt

Sommer 2014

Ausgabe am: 28.04.14
Besprechung am: 05.05.14

Aufgabe 1 *N-Gramme*

Gegeben sei das folgende Trainingskorpus: *„Ich habe Hunger und Durst. Aber ich habe kein Essen und kein Getränk dabei. Ausserdem habe ich kein Geld. Ich muss zuerst Geld besorgen.“* und das Testkorpus: *„Ich habe kein Geld.“*.

- Bestimmen Sie die Vokabulargröße k des Trainingskorpus.
- Bestimmen Sie den ML- und MAP-Schätzer für die bedingten Wahrscheinlichkeiten $P(w_t|w_{t-1})$ unter der Annahme der Priorparameter $\alpha_i = 2$ mit Hilfe des Trainingskorpus.
- Geben mit Hilfe der jeweils geschätzen Parametern die Gesamtwahrscheinlichkeit für das Testkorpus an.

Aufgabe 2 *Autovervollständigung*

Autovervollständigungssysteme wie sie beispielsweise bei Texteditoren Anwendung finden, sollen dem Benutzer Wörter vorschlagen bevor sie vollständig eingeben wurden sind. Angenommen Sie haben die Aufgabe eine derartig intelligente Eingabehilfe zu entwickeln. Diskutieren Sie Lösungsansätze für diese Aufgabenstellung.

Aufgabe 3 *T9*

Die Eingabehilfe T9 ermöglicht die Texteingabe auf Zifferntastaturen. Diskutieren Sie, wie Sie T9 mit Hilfe eines N-Gramm-Modells auf Buchstabenebene implementieren können. Das System soll stets das wahrscheinlichste Wort für die getippten Ziffern ermitteln. Wie rechenaufwändig ist Ihre Lösung?

Zusatzaufgabe *Binomialverteilung*

Beim Münzwurf-Experiment stellt sich die Frage, wie oft in einer Reihe von Münzwürfen Kopf geworfen wurde. Die Anzahl N_K von beobachteteten Kopfwürfen ist dabei eine binomialverteilte Zufallsvariable (analog ist N_Z die Anzahl von beobachteten Zahlwürfen). Wie in der Vorlesung gezeigt, besitzt diese Binomialverteilung einen Parameter θ, der aus Daten gelernt werden kann. Bestimmen sie die ML-Schätzung für die Binomialverteilung, also genau den Parameter θ, der die Liklihood der Binomialverteilung, gegeben durch:

$$\mathcal{L}(\theta) = \binom{N_K + N_Z}{N_K} \theta^{N_K} (1-\theta)^{N_Z}$$

maximiert.
Hinweis: Leiten Sie die Funktion $\ln \mathcal{L}(\theta)$ ab.