

# Vorlesung “Einführung in die Computerlinguistik”

Prof. Dr. D. Klakow

## Übung zur Statistischen Sprachverarbeitung 2

Ausgabe: 5.2.2008

Abgabe: 12.2.2008

### Aufgabe 2.1

In dieser Aufgabe sollen Sie für einen vereinfachten Text von “Krieg und Frieden” die Wahrscheinlichkeit für jedes Wort durch die relativen Häufigkeiten (Häufigkeit des Worts geteilt durch die Anzahl aller Wörter im Text) schätzen. Um den Umfang der Aufgabe in Grenzen zu halten, wurden nur die häufigsten neun Wörter übernommen und alle anderen durch `OTHER_WORD` ersetzt, für das Sie allerdings auch die Wahrscheinlichkeit schätzen sollen. Nutzen Sie dazu das Perl-Skript `zipf_stat.pl`. Der Aufruf auf einem Standard-Linux System (z.B. CoLi-CIP-Pool) ist `cat tolstoy_war_and_peace.t10.dat | zipf_stat.pl`. Auf anderen Rechnern müssen Sie unter Umständen Anpassungen vornehmen.

### Aufgabe 2.2

Geben Sie die Entropie für diesen reduzierten Text an.

### Aufgabe 2.3

Bestimmen Sie einen binären Kode für alle zehn Wörter im reduzierten Text. Berechnen Sie dazu den Logarithmus zur Basis zwei der Wahrscheinlichkeiten aus Aufgabe 2.1, wie auch in der Vorlesung gemacht. Runden Sie geeignet.

### Aufgabe 2.4

Geben Sie die Länge des kodierten (reduzierten) Texts an. Gewichten Sie dazu die Länge des Kodes mit der Häufigkeiten des damit kodierten Wortes.

### Aufgabe 2.5

Auf der Webseite finden Sie auch eine reduzierte Fassung von Gullivers Reisen. Schätzen Sie auch für diesen Text die Wahrscheinlichkeiten.

### Aufgabe 2.6

Bestimmen Sie die Kullback-Leibler Divergenz mit der in der Vorlesung verwendeten Formel wobei  $p$  die Wahrscheinlichkeiten auf “Krieg und Frieden” und  $q$  die Wahrscheinlichkeiten auf “Gullivers Reisen” sind.