Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

Dr. Jürg M. Stettbacher
Neugutstrasse 54
CH-8600 Dübendorf

Telefon: +41 43 299 57 23
E-Mail: dsp@stettbacher.ch

Information und Entropie

Praktikum

Version 3.12
2016-03-03

Zusammenfassung: In diesem Praktikum geht es darum, ein Programm zu schreiben, das von gegebenen Daten in einer Datei die Information und die Entropie bestimmt.

Inhaltsverzeichnis

1 Einleitung 2

2 Dateien 2

3 Aufgabe 3

4 Zusatzaufgabe 3

1 Einleitung

In der Informationstheorie betrachten wir Datenquellen oft als Zufallsvariablen und jedes Symbol, das aus der Quelle kommt, als Zufallsereignis. Sind die Auftretenswahrscheinlichkeiten bekannt, so kann für jedes Symbol die Information berechnet werden und für die Quelle der mittlere Informationsgehalt, also die Entropie.

In diesem Praktikum haben wir verschiedene Datenquellen in der Form von Dateien zur Verfügung. Es handelt sich um ASCII-Dateien (**.txt*) und jedes ASCII-Zeichen daraus stellt ein Symbol dar. Gesucht sind jeweils der Informationsgehalt für jedes Symbol einer Datei, sowie die Entropie der gesamten Datei.

2 Dateien

Die folgenden Dateien stehen Ihnen für das Bearbeiten des Praktikums zur Verfügung:

- *Entropy_template.java* (Hauptaufgabe)
- *data_1.txt* bis *data_6.txt* und *deutsch_2.txt* (Testdaten)

3 Aufgabe

Verwenden Sie die Testdaten in den Dateien *data_1.txt* bis *data_6.txt*. Schreiben Sie ein Java-Programm *Entropy.*java, das eine bestimmte Datei öffnen und zeichenweise lesen kann. Der Name der ASCII-Datei soll auf der Kommandozeile übergeben werden. Kompilieren sie Ihr Programm:

```
> javac Entropy.java
```

Führen Sie das Programm aus:

```
> java Entropy data.txt
```

Es steht eine vorbereitete Datei *Entropy_template.java* zur Verfügung, die Sie als Vorlage verwenden können. Sie brauchen dann nur an den bezeichneten Stellen Ihren Code zu ergänzen. Für die Berechnung von Information und Entropie gehen Sie so vor:

- Erstellen Sie ein Histogramm in dem Sie für alle Zeichen angeben, wie häufig sie vorgekommen sind und zählen Sie, wieviele Zeichen insgesamt in der Datei enthalten sind.
- Berechnen Sie für jedes Zeichen die Information.
- Berechnen Sie die Entropie der Quelle.
- Auf dem Bildschirm soll für jedes Zeichen ausgegeben werden
  - wie oft es vorgekommen ist,
  - was seine Information ist.
- Zudem soll auf dem Bildschirm für die Quelle ausgegeben werden
  - wie viele Zeichen total vorgekommen sind,
  - wie gross die Entropie ist.

4 Zusatzaufgabe

Verwenden Sie diesmal die Testdaten in der Datei *deutsch_2.txt*. Wir wollen nun prüfen, was geschieht, wenn die Symbole einer Quelle nicht statistisch unabhängig sind. Schreiben Sie zu diesem

Zweck ein zweites Java-Programm *Entropy2*.java, das eine ASCII Datei öffnen und zeichenweise lesen kann. Für ein gegebenes Zeichen y_0 soll das Programm die bedingte Entropie $H(X|y_0)$ ermitteln, also die Entropie jener Symbole[1] X, die auf das Symbol y_0 folgen. Wir nennen dies die *Entropie von X gegeben* y_0.

$$H(X|y_0) = \sum_{x \in \Omega} P(x|y_0) \cdot log_2 \frac{1}{P(x|y_0)} \tag{1}$$

Dabei ist Ω der Ereignisraum der Zufallsvariable X, in unserem Fall also der Zeichenvorrat der Quelle X. Der Aufruf des Java-Programms soll so aussehen:

```
> java Entropy2 deutsch_2.txt y0
```

Wählen Sie für y_0 nacheinander verschiedenen Zeichen aus dem Vorrat der Quelle *deutsch_2.txt*.

Die Entropie $H(X|Y)$ ist übrigens der Mittelwert von $H(X|y)$ über allen Symbolen y der Quelle Y. In unserem Fall also:

$$H(X|Y) = \sum_{y \in \Omega} P(y) \cdot H(X|y) = \sum_{x,y \in \Omega} P(y) \cdot P(x|y) \cdot log_2 \frac{1}{P(x|y)} = \sum_{x,y \in \Omega} P(x,y) \cdot log_2 \frac{1}{P(x|y)} \tag{2}$$

Dies wollen wir an dieser Stelle aber nicht weiter verfolgen.

Die Resultate des Programms *Entropy2* sollen mit den Resulten der ersten Aufgabe verglichen werden. Überlegen Sie sich dabei die folgenden Punkte:

- Bei welchen Daten sind Unterschiede zu erwarten?
- Bei welchem Zeichen y_0 treten die Unterschiede besonders deutlich hervor?

[1] Wir fassen dabei X als eine Zufallsvariable auf. Aus diesem Grund schreiben wir X gross.