Information und Entropie
Praktikum

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

Einleitung

Wir betrachten eine Datenquelle als Strom von Symbolen x_n mit n = 0 ... N−1.

Sind die Auftretenswahrscheinlichkeiten $P(x_n)$ bekannt, so kann für jedes Symbol x_n die Information $I(x_n)$ berechnet werden und für die Quelle der mittlere Informationsgehalt, also die Entropie H.

In diesem Praktikum betrachten wir verschiedene Datenquellen. Von jeder Quelle liegt uns ein endlicher Ausschnitt ihres Ausgangs in Form einer Datei vor. Es handelt sich um ASCII-Dateien (*.txt) und jedes ASCII-Zeichen daraus stellt ein Symbol x_n dar. Gesucht sind jeweils der Informationsgehalt $I(x_n)$ für jedes Symbol x_n einer Quelle, sowie die Entropie H der Quelle.

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

- Template für Programmcode mit Array: EntropyArray_template.java
- Testdaten: source_1.txt bis source_6.txt

Die Java-Templates enthalten den grössten Teil des notwendigen Codes für das Lösen der Aufgaben.

Die Dateien source_1.txt bis source_4.txt enthalten Testdaten aus vier verschiedenen binären Quellen, wobei jedes Symbol mit einem Byte dargestellt wird. Die Dateien source_5.txt und source_6.txt enthalten deutschen Text im ASCII-Format ohne Umlaute und ohne Ziffern.

Öffnen und betrachten Sie die Testdateien erst, wenn es die Aufgabenstellung verlangt.

Aufgaben

1. Verwenden Sie das Java-Template, um ein Programm Entropy.java zu schreiben. Sie brauchen nur an den bezeichneten Stellen in der Vorlage den Code zu ergänzen. Die Stellen sind mit «TODO» markiert.

   Ziel: Das Programm soll eine bestimmte ASCII-Datei mit Testdaten öffnen und zeichenweise einlesen. Der Name der Datei ist auf der Kommandozeile als Argument anzugeben.

   Für die gelesenen Zeichen, resp. Symbole wird zuerst die Häufigkeit bestimmt, die angibt, wie oft jedes der verschiedenen Symbole vorkommt, sowie die gesamte Anzahl der Zeichen und die Menge der unterschiedlichen Symbole. Dabei sollen nur druckbare Zeichen berücksichtigt werden, das heisst, dass die Leerzeichen und Steuerzeichen zu ignoriert sind.

   Aus der Häufigkeit berechnet das Programm die Wahrscheinlichkeit $P(x_n)$, dann den Informationsgehalt $I(x_n)$ von jedem Zeichen x_n und schliesslich die Entropie H der Quelle.

   Dazu ist wird bekanntlich der Logarithmus zur Basis 2 benötigt. Die entsprechende Funktion muss programmiert werden.

2. Betrachten Sie die Ausgaben des Programms für die Testdatei source_1.txt:
   a) Wie gross ist das N der Quelle?
   b) Welche Symbole x_n produziert die Quelle?
   c) Wie gross sind die Wahrscheinlichkeiten $P(x_n)$ der Symbole?
   d) Überlegen Sie, auf welche Dateigrösse sich die Testdatei source_1.txt komprimieren lassen sollte.
   e) Verwenden Sie beispielsweise ein Kompressionsprogramm (z.B. zip/Linux oder pkzip/Windows), um die Testdatei source_1.txt zu komprimieren. Überprüfen Sie die resultierende Dateigrösse mit ihren obigen Überlegungen.
   f) Was ist das Fazit aus den Resultaten?
3. Wiederholen Sie die vorhergehende Aufgabe mit der Testdatei source_2.txt.
   a) Gibt es Unterschiede?
   b) Falls ja, wie sind die Unterschiede erklärbar?
   c) Wenn Sie eine Erklärung gefunden haben: Öffnen und betrachten Sie die beiden Testdateien in einem Texteditor. Was stellen Sie fest?
4. Nun drehen wir den Spiess um und verwenden die Testdaten source_3.txt und source_4.txt:
   a) Berechnen (resp. schätzen) Sie die Entropie H der betreffenden Quellen, indem Sie untersuchen, wie stark sich die Testdateien komprimieren lassen.
   b) Vergleichen Sie die Resultate mit der Ausgabe des Java-Programms Entropy von oben.
   c) Was stellen Sie fest? Erklären Sie allfällige Unterschiede.
   d) Wenn Sie eine Erklärung gefunden haben: Öffnen und betrachten Sie die beiden Testdateien in einem Texteditor. Was stellen Sie fest?

5. Führen Sie nun das Programm Entropie.java mit den Testdaten source_5.txt aus.
   a) Wie gross ist die Entropie H dieser Quelle?
   b) Betrachten Sie diese Testdatei. Wie vertrauenswürdig sind die Resultate, die das Programm anzeigt in Bezug auf die deutsche Sprache?
6. Führen Sie nun das Programm Entropie.java mit den Testdaten source_6.txt aus.
   a) Wie gross ist die Entropie H dieser Quelle?
   b) Angenommen nur die Buchstaben ohne Punktuationszeichen würden als Symbole gelten, wäre die wahre Entropie dann grösser oder kleiner als der berechnete Wert? Begründen Sie Ihre Antwort.
   Beachte: Die «Space» Zeichen werden im Programm nicht berücksichtigt, sie werden schon in der Funktion isPrintable(int charCode) herausgefiltert.
7. Fakultative Zusatzaufgabe:
   a) Ändern Sie das Programm, dass nur Buchstaben betrachtet werden. Ausserdem sollen ein Gross- und Kleinbuchstabe als das gleiche Symbole gelten.
   b) Erstellen Sie anhand der Testdaten source_6.txt eine nach Informationsgehalt geordnete Liste der Symbole.
   Anmerkung: Sie müssen das Problem nicht vollständig in Java lösen.