MATHEMATISCHES INSTITUT
DER UNIVERSITÄT MÜNCHEN
Dr. W. Spann, G. Kleen

WS 2022/2023

Programmieren II für Studierende der Mathematik

Aufgabe 8

Das LZW-Verfahren (Lempel/Ziv/Welch) dient zur Kompression von Dateien. Die Grundidee besteht darin, Teilzeichenketten in einer Tabelle abzulegen und anstelle der Teilzeichenkette den zugehörigen Tabellenindex zu übertragen. Die Tabelle hat eine feste Zahl von Einträgen (z.B. 4096), in den (z.B. 256) ersten Einträgen ist der zugrundeliegende Zeichensatz – ein Zeichen pro Eintrag – abgelegt. Die Tabelle wird während des Komprimierens bzw. Dekomprimierens im Hauptspeicher aus den Daten erzeugt.

Kompressionsalgorithmus

1. Tabelle mit dem Zeichensatz initialisieren.
2. (i) Längste in der Tabelle enthaltene Teilzeichenkette im Eingabestrom (Bez. W) bestimmen.
   (ii) Statt W den Tabellenindex von W ausgeben.
   (iii) Falls die Tabelle noch freie Einträge enthält und auf W ein Zeichen (Bez. z) folgt, Wz in die Tabelle eintragen.
   (iv) 2. wiederholen, dabei mit dem auf W folgenden Zeichen starten.

Bsp.: Zeichensatz: `A`,`B`,`C`; Tabellenlänge: 10
Eingabe: `ACBACBBACCCAB`
Ausgabe: `0213 15 29 01`
Tabelle: `A B C AC CB BA ACB BB BAC CC`

Dekompressionsalgorithmus

1. Tabelle mit dem Zeichensatz initialisieren.
2. (i) Index i lesen, $W(i)$ ausgeben.
   (ii) Falls die Tabelle noch freie Einträge enthält und auf i im Eingabestrom ein weiterer Index j folgt:
   Wenn W(j) belegt ist, dann $W(i)$ und das erste Zeichen von $W(j)$ in die Tabelle eintragen; sonst $W(i)$ und das erste Zeichen von $W(i)$ in die Tabelle eintragen.
   (iii) 2. mit dem auf i folgenden Index wiederholen.

(a) Programmieren Sie den Kompressionsalgorithmus, geben Sie jedoch die erzeugten 12-Bit-Tabellenindizes als vierstellige Dezimalzahlen mit führenden Nullen und jeweils einem trennenden Leerzeichen zeilenweise (max. 80 Zeichen/pro Zeile) in eine Textdatei aus. In der Standardfehlerausgabe soll die Anzahl der Zeichen der Eingabedatei erscheinen. Außerdem soll die Anzahl der Zeichen bestimmt und ausgegeben werden, die die komprimierte Datei enthalten würde. Verwenden Sie aus Effizienzgründen für die Tabelle möglichst *keinen* Vektor vom Datentyp `vector<string>`, sondern die Umkehrabbildung (`map<string,int>`).

(b) Fügen Sie den Dekompressionsalgorithmus hinzu.

(c*) Ändern Sie das Programm so ab, dass die Tabellenindizes in eine Binärdatei (2 12-Bit-Tabellenindizes $\widehat{=}$ 3 Byte) ausgegeben bzw. von dort gelesen werden.

Bearbeitungszeitraum: bis Donnerstag, 12.01.2023, 16^{00}