Albert-Ludwigs-Universität
Institut für Informatik
Prof. Dr. S. Albers
S. Schmelzer, M. Schmidt

Freiburg i. Br., den 11. Mai 2005

Software-Praktikum „Datenkompression“ SS 05 Aufgabenblatt 3

Bearbeitung: bis 8.6.2005

Michael Burrows und David Wheeler beschreiben in ihrer Arbeit „A Blocksorting Lossless Data Compression Algorithm“ ein Verfahren, mit dem sich die Zeichen eines Textes so umordnen lassen, dass sich die Lokalität einzelner Zeichen signifikant erhöht, d.h. dass sie mehrfach unmittelbar hintereinander stehen. Dabei teilen sie den Text in gleich große Blöcke ein, innerhalb deren sie die Zeichen umordnen. Wir nennen im Folgenden die Anzahl der Zeichen in einem solchen Block die *Blockgröße*.

Erweitern Sie in Ihrem Programm den Benutzerdialog, der bei Auswahl des Punktes `Einstellungen` im Menü `BSTW` erscheint, um eine aus drei Blockgrößenalternativen `1024 Byte`, `4096 Byte` und `65536 Byte` bestehende Auswahlknopfgruppe sowie um zwei Schaltflächen `Übernehmen` und `Abbrechen`. Bei der Auswahlknopfgruppe soll anfangs die Möglichkeit `1024 Byte` ausgewählt sein. Betätigung einer der beiden Schaltflächen führt zur Beendigung des Dialogs, wobei `Übernehmen` den ausgewählten Wert übernimmt und `Abbrechen` die Blockgröße auf ihren bisherigen Wert zurücksetzt – bei erneutem Aufruf des Dialogs muss also wieder der bisherige Wert als ausgewählt gekennzeichnet sein.

Ferner sollen, abhängig davon, ob im Menü `BSTW` das Ankreuzfeld `BW` ausgewählt ist oder nicht, die Punkte `Komprimieren` und `Dekomprimieren` dieses Menüs unterschiedliche Reaktionen bewirken. Ist das Feld nicht ausgewählt, so behalten die beiden Punkte ihre in Aufgabenblatt 2 beschriebene Funktionalität. Im Falle ausgewählten Feldes ist bei `Komprimieren` die zu komprimierende Datei in Blöcke gemäß des unter `Einstellungen` ausgewählten Wertes zu unterteilen, auf jeden dieser Blöcke die Burrows-Wheeler-Transformation anzuwenden und die hieraus resultierende Zeichenfolge mit dem BSTW-Verfahren zu komprimieren. Zur Unterscheidung von der bisherigen Kompression wird beim neuen Kompressionsverfahren die Endung „`.BW`“ an den Namen der zu komprimierenden Datei angehängt. Entsprechend ist bei Auswahl des Punktes `Dekomprimieren` ein Benutzerdialog zu öffnen, der nur Dateien mit Endung „`.BW`“ anbietet. Im Hauptfensterfeld der gerade durchgeführten Aktion sollen in Abhängigkeit des Ankreuzfeldzustands unterschiedliche Ausgaben erfolgen.

Schreiben Sie eine Klasse für die Burrows-Wheeler-Transformation. Diese soll bei Übergabe eines Blocks die zugehörige umgeordnete Zeichenfolge sowie den Zeilenindex zurückliefern. Wird der Klasse ein Block und ein Zeilenindex übergeben, so ist hieraus die ursprüngliche Zeichenfolge zu rekonstruieren.

Da die neue Kompression von der gewählten Blockgröße abhängt, müssen wir diese in der komprimierten Datei ablegen. Hierzu kodieren wir die Blockgröße auf genau die gleiche Weise, wie wir bei der BSTW-Kompression den Index des aktuellen Zeichens kodieren. Diese Bitfolge schreiben wir an den Anfang der komprimierten Datei. Hierauf folgen unmittelbar die Kompressionen der einzelnen Blöcke, wobei wir zunächst den jeweiligen Zeilenindex in gewohnter Kodierung und dann die Kodierung der umgeordneten Zeichenfolge in die Datei wegschreiben. Am Ende jedes Blocks füllen wir die Kodierung auf volle Byte auf, bevor wir mit der Kompression des nächsten Blocks beginnen. Beim Übergang von einem Block zum nächsten setzen wir das Wörterbuch *nicht* zurück.

Ergänzen Sie Ihr UML-Klassendiagramm um die durch die Burrows-Wheeler-Transformation und die zugehörigen Einstellungen hinzugekommenen Komponenten.

Hinweis: Auf unserer Praktikumsseite gibt es eine `tar`-Datei mit Beispielinstanzen. Diese umfasst folgende Dateien:

- `abraca.txt`
- `abraca.txt.bwt`
- `abraca.txt.BW`
- `buergschaft1.txt`
- `buergschaft1.txt.bwt`
- `buergschaft1.txt.BW`
- `alotofa.txt`
- `alotofa.txt.BW`
- `schneew.txt`
- `schneew1024.txt.BW`
- `schneew4096.txt.BW`
- `schneew65536.txt.BW`

Die `.txt`-Dateien sind die Originaldateien. Der Zeilenindex und die Zeichenfolgen nach der Burrows-Wheeler-Transformation stehen in den `.bwt`-Dateien. Die `.BW`-Dateien sind schließlich die Dateien nach vollständiger Kompression.